Differentially Private Federated Learning: An Information-Theoretic Perspective

Shahab Asoodeh [1] **Flavio P. Calmon** [1]

Abstract

In this work, we propose a new technique for deriving the differential privacy parameters in the context of federated learning when only the last update is publicly released. In this approach, we interpret each iteration as a Markov kernel and quantify its impact on privacy parameters via the contraction coefficient of a certain f-divergence that underlies differential privacy. To do so, we generalize the well-known Dobrushin's ergodicity coefficient, originally defined in terms of total variation distance, to a family of f-divergences.

1. Introduction

Federated Learning (McMahan et al., 2016) refers to algorithms for aggregating multiple noisy model local updates from distributed users. In the prototypical setting, users compute their local gradients on their local data and send them to the central aggregator (uplinks). All these local updates are then aggregated to a centralized update which is then sent back to users (downlinks). This iterative distributed algorithm has recently gained attention due to its natural parallelization and storage efficiency. Although users hold on to their local data during each iteration and only gradient are transmitted, It is easy to compromise the privacy of users (Fredrikson et al., 2015; Melis et al., 2018).

Following the *de facto* standard of differential privacy (DP) in large-scale model fitting (Bassily et al., 2014; 2019; Chaudhuri & Mishra, 2006; Chaudhuri & Monteleoni, 2009; Chaudhuri et al., 2011; Duchi et al., 2013; Jain & Thakurta, 2014; Jain et al., 2012; Smith et al., 2017; Song et al., 2013; Talwar et al., 2015; Thakurta & Smith, 2013; Wang et al., 2017; Wu et al., 2017), we study the differentially private federated learning under two assumptions. First, we assume that users communicate over encrypted channels with a *trusted* aggregator. This assumption is justified by secure multiparty computation (Bonawitz et al., 2017) at the cost of higher communication and computation complexity. Second, we assume that the aggregator releases the model parameters only after a certain number of iterations and hide all intermediate updates. Augenstein et al. (2020) recently studied the same setting where, after T iterations, the last model parameters are used to generate synthetic data for data inspection purposes. This assumption is also in line with the recent frameworks of "privacy amplification by iteration" (Feldman et al., 2018) and "privacy amplification by post-processing" (Balle et al., 2019a). However, our work differs in that we allow for subsampling and adopt the approximate DP as the measure of privacy as opposed to (Balle et al., 2019a; Feldman et al., 2018) where iteration (or post-processing) was the only source of randomness (i.e., no subsampling) and the privacy was given in terms of Rényi differential privacy.[1] The main technical difference relies on the fact that Rényi divergence $D_\alpha(\mu\|\nu)$ for distributions μ and ν and $\alpha \geq 1$ is *not* jointly convex in μ and ν; whereas approximate DP is given in terms of a certain divergence (to be defined in next section) that is jointly convex.

In Section 2, we revisit different information-theoretic definitions including strong data processing inequality, E_ε-divergence, and contraction coefficient of Markov kernels under f-divergences and also generalize Dobrushin's ergodicity's coefficient. In Section 3, we turn to the privacy analysis of federated learning algorithms.

2. Information Theory Preliminaries

In this section, we first provide some preliminaries from information theory, in particular, the contraction coefficient of Markov kernels under general f-divergences. Then, we provide a closed-form expression for the contraction coefficient of kernels under a certain f-divergence which underlies differential privacy.

[1]School of Engineering and Applied Science, Harvard University. {shahab, flavio}@seas.harvard.edu. Parts of the results in this work were accepted in International Symposium on Information Theory (ISIT'20).

International Workshop on Federated Learning for User Privacy and Data Confidentiality, July 17-18, 2020.

[1]It must be pointed out that Rényi differential privacy guarantee can be converted into (ε, δ)-DP, according to (Abadi et al., 2016). However, as shown in (Asoodeh et al., 2020) our approach yields tighter bounds for ε and δ than what would be obtained by converting the results in (Feldman et al., 2018) to (ε, δ)-DP.

Given a convex function $f : [0, \infty) \to \mathbb{R}$ with $f(1) = 0$, the f-divergence (Ali & Silvey, 1966; Csiszár, 1967) between two probability measures μ and ν is defined as

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

This includes several popular measures: KL-divergence, χ^2-divergence, and total variation distance TV are f-divergences for $f(t) = t\log(t)$, $f(t) = (t-1)^2$, and $f(t) = \frac{1}{2}|t-1|$, respectively.

It was observed by Olmedo and Barthe (2013) that the definition of differential privacy can be expressed in terms of a certain f-divergence. Let $\mathcal{X}^n$ be the set of all possible datasets of size n, where each entry takes values in $\mathcal{X}$. A pair of datasets $x \in \mathcal{X}^n$ and $x' \in \mathcal{X}^n$ are neighboring (denoted by $x \sim x'$) if they differ in exactly one entry. A randomized mechanism $\mathcal{M}$ acts on each $x \in \mathcal{X}^n$ and generates a random variable with distribution $\mathcal{M}_x$. Mechanism $\mathcal{M}$ is said to be (ε, δ)-DP if we have

$$\sup_{x\sim x'} \mathsf{E}_\varepsilon(\mathcal{M}_x\|\mathcal{M}_{x'}) \le \delta, \tag{1}$$

where E_ε-divergence is the f-divergence associated with $f(t) = (t - e^\varepsilon)_+ \max\{0, t - e^\varepsilon\}$. The fact that (1) is equivalent to the typical definition of DP becomes clear once one observes the following equivalent forms of E_ε-divergence: for any probability measures μ and ν on some arbitrary set $\mathcal{Y}$, we have

$$\begin{aligned}
\mathsf{E}_\varepsilon(\mu\|\nu) &= \int_{\mathcal{Y}} \big(\mathrm{d}(\mu - e^\varepsilon\nu)(y)\big)_+ && (2)\\
&= \sup_{A\subset\mathcal{Y}} [\mu(A) - e^\varepsilon\nu(A)] \\
&= \frac{1}{2}\int |\mathrm{d}\mu - e^\varepsilon \mathrm{d}\nu| - \frac{1}{2}(e^\varepsilon - 1) && (3)\\
&= \mu\big(\log\frac{\mathrm{d}\mu}{\mathrm{d}\nu} > \varepsilon\big) - e^\varepsilon\nu\big(\log\frac{\mathrm{d}\mu}{\mathrm{d}\nu} > \varepsilon\big). && (4)
\end{aligned}$$

The E_ε-divergence representation of DP was used in (Balle & Wang, 2018; Balle et al., 2018; 2019b;c; Wang et al., 2018) to prove new privacy results or simplify the proofs of existing results. The following lemma gives the E_ε-divergence between multivariate Gaussian distributions with the same variance. It can be proved essentially similar to Lemma 6 in (Balle & Wang, 2018).

Lemma 1. *For $m_1, m_2 \in \mathbb{R}^d$ and $\sigma > 0$, let $\mathcal{N}_1$ and $\mathcal{N}_2$ denote $\mathcal{N}(m_1, \sigma^2\mathrm{I})$ and $\mathcal{N}(m_2, \sigma^2\mathrm{I})$, respectively. Then, we have*

$$\mathsf{E}_\varepsilon(\mathcal{N}_1\|\mathcal{N}_2) = \mathsf{Q}\Big(\frac{\varepsilon}{\kappa} - \frac{\kappa}{2}\Big) - e^\varepsilon\mathsf{Q}\Big(\frac{\varepsilon}{\kappa} + \frac{\kappa}{2}\Big),$$

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

This example demonstrates that E_ε-divergence is symmetric for Gaussian distributions with the same variance, i.e., it depends on the their means only through their ℓ_2 distance. Define

$$\begin{aligned}
\theta_\varepsilon(r) &\coloneqq \mathsf{E}_\varepsilon(\mathcal{N}(r, \mathrm{I})\|\mathcal{N}(0, \mathrm{I}) \\
&= \mathsf{Q}\left(\frac{\varepsilon}{r} - \frac{r}{2}\right) - e^\varepsilon\mathsf{Q}\left(\frac{\varepsilon}{r} + \frac{r}{2}\right). && (5)
\end{aligned}$$

With this definition in our disposal, we can write

$$\mathsf{E}_\varepsilon(\mathcal{N}_1\|\mathcal{N}_2) = \mathsf{E}_\varepsilon(\mathcal{N}_2\|\mathcal{N}_1) = \theta_\varepsilon\Big(\frac{\|m_1 - m_2\|}{\sigma}\Big).$$

The prominent properties of E_ε-divergence are as follows:

- $0 \le \mathsf{E}_\varepsilon(\mu\|\nu) \le \mathsf{TV}(\mu, \nu)$ for any $\varepsilon > 0$. The upper bound is equality if and only if $\varepsilon = 0$,
- $\varepsilon \mapsto \mathsf{E}_\varepsilon(\mu\|\nu)$ is continuous and strictly decreasing on $(0, \mathsf{TV}(\mu, \nu)]$,
- $\mathsf{E}_\varepsilon(\mu\|\nu)$ decreases by post-processing (data-processing inequality),
- $(\mu, \nu) \mapsto \mathsf{E}_\varepsilon(\mu\|\nu)$ is convex.

The last two properties are shared by all f-divergences. In particular, any f-divergence satisfies the data processing inequality, i.e., $D_f(\mu\mathsf{K}\|\nu\mathsf{K}) \le D_f(\mu\|\nu)$ for any convex function f and any Markov kernel[2] K, where $\mu\mathsf{K}$ denotes the push-forward of μ by K, i.e., $\mu\mathsf{K} = \int \mu(\mathrm{d}y)\mathsf{K}(y)$. This inequality is typically strict for non-trivial kernels and many interesting f-divergences. To account for this, Ahlswede and Gács (1976) studied the *strong* data processing inequality and defined the notion of *contraction coefficient* $\eta_f(\mathsf{K})$ of K under f-divergence as

$$\eta_f(\mathsf{K}) \coloneqq \sup_{\substack{\mu,\nu:\\ D_f(\mu\|\nu)\neq 0}} \frac{D_f(\mu\mathsf{K}\|\nu\mathsf{K})}{D_f(\mu\|\nu)}. \tag{6}$$

It is worth noting that for total variation distance the contraction coefficient $\eta_{\mathsf{TV}}(\mathsf{K})$ appears to be introduced twenty years earlier by Dobrushin (1956) under the name of *ergodicity coefficient*. Interestingly, Dobrushin proved that the supremum in the definition of $\eta_{\mathsf{TV}}(\mathsf{K})$ can be restricted to point masses:

$$\eta_{\mathsf{TV}}(\mathsf{K}) = \sup_{y_1, y_2\in\mathcal{Y}} \mathsf{TV}(\mathsf{K}(y_1), \mathsf{K}(y_2)). \tag{7}$$

This two-point characterization has been instrumental in studying strong ergodicity of Markov processes (Dobrushin,

[2]Here, by Markov kernel K we simply mean a family of probability distribution $\mathsf{K}(y)$ for each $y \in \mathcal{Y}$. We ignore the measurability issues in the formal definition of Markov kernel for simplicity.

1956; Kontorovich & Raginsky, 2017), the uniqueness of Gibbs measures (Georgii, 2011), contraction of mutual information (and generalized mutual information) in a Markov chain (Polyanskiy & Wu, 2016; Xu & Raginsky, 2015), and distributed estimation (Xu & Raginsky, 2015).

Since E_ε-divergence generalizes total variation distance, Dobrushin's ergodicity coefficient can be generalized by contraction coefficient of E_ε-divergence:

$$\eta_\varepsilon(\mathsf{K}) := \sup_{\substack{\mu,\nu \\ \mathsf{E}_\varepsilon(\mu\|\nu)\neq 0}} \frac{\mathsf{E}_\varepsilon(\mu\mathsf{K}\|\nu\mathsf{K})}{\mathsf{E}_\varepsilon(\mu\|\nu)}. \tag{8}$$

The following theorem establishes a two-point characterization for η_ε similar to the Dobrushin's characterization in (7).

Theorem 1. *For any $\varepsilon \geq 0$, we have*

$$\eta_\varepsilon(\mathsf{K}) = \sup_{y_1,y_2\in\mathcal{Y}} \mathsf{E}_\varepsilon(\mathsf{K}(y_1)\|\mathsf{K}(y_2)). \tag{9}$$

Proof. Given two probability measures μ and ν defined on $\mathcal{Y}$, define $\phi(y) = (\mu(y) - e^\varepsilon\nu(y))_+$ and $\phi'(y) = (-\mu(y) + e^\varepsilon\nu(y))_+$ for any $y \in \mathcal{Y}$. Note that since $\frac{1}{2}\|\mu - e^\varepsilon\nu\|_1 = \mathsf{E}_\varepsilon(\mu\|\nu) + \frac{1}{2}(e^\varepsilon - 1)$ and $\|\phi\|_1 = \mathsf{E}_\varepsilon(\mu\|\nu)$, it follows that $\|\phi'\|_1 = \mathsf{E}_\varepsilon(\mu\|\nu) + e^\varepsilon - 1$. Letting E_ε denote $\mathsf{E}_\varepsilon(\mu\|\nu)$ for brevity, we can write

$$\begin{aligned}
\|\mu\mathsf{K} - e^\varepsilon\nu\mathsf{K}\|_1 &= \|(\mu(\mathrm{d}y) - e^\varepsilon\nu(\mathrm{d}y))\mathsf{K}\|_1 \\
&= \int_{\mathcal{Y}} \left| \int_{\mathcal{X}} (\mu(\mathrm{d}y) - e^\varepsilon\nu(\mathrm{d}y))\mathsf{K}(y) \right| \\
&= \int_{\mathcal{Y}} \left| \int_{\mathcal{X}} \phi(\mathrm{d}y)\mathsf{K}(y) - \int_{\mathcal{X}} \phi'(\mathrm{d}y')\mathsf{K}(y') \right| \\
&= \int_{\mathcal{Y}} \left| \|\phi\| \int_{\mathcal{X}} \frac{\phi(\mathrm{d}y)}{\|\phi\|}\mathsf{K}(y) - \|\phi'\| \int_{\mathcal{X}} \frac{\phi'(\mathrm{d}y')}{\|\phi'\|}\mathsf{K}(y') \right| \\
&= \int_{\mathcal{Y}} \left| \|\phi\| \left(\int \frac{\phi'(\mathrm{d}y')}{\|\phi'\|} \right) \int_{\mathcal{X}} \frac{\phi(\mathrm{d}y)}{\|\phi\|}\mathsf{K}(y) \right. \\
&\quad \left. - \|\phi'\| \left(\int \frac{\phi(\mathrm{d}y)}{\|\phi\|}\mathrm{d}y \right) \int_{\mathcal{X}} \frac{\phi'(\mathrm{d}x')}{\|\phi'\|}\mathsf{K}(y') \right| \\
&\leq \max_{x,x'} \int_{\mathcal{Y}} \left| \|\phi\|\mathsf{K}(y) - \|\phi'\|\mathsf{K}(y') \right| \\
&= \max_{x,x'} \int_{\mathcal{Y}} \left| \mathsf{E}_\varepsilon\mathsf{K}(y) - (\mathsf{E}_\varepsilon + e^\varepsilon - 1)\mathsf{K}(y') \right| \\
&\leq \mathsf{E}_\varepsilon \max_{x,x'} \int_{\mathcal{Y}} \left| \mathsf{K}(y) - e^\varepsilon\mathsf{K}(y') \right| + (e^\varepsilon - 1)(1 - \mathsf{E}_\varepsilon).
\end{aligned}$$

In light of (3), the above implies

$$\mathsf{E}_\varepsilon(\mu\mathsf{K}\|\nu\mathsf{K}) \leq \mathsf{E}_\varepsilon(\mu\|\nu) \max_{y,y'} \mathsf{E}_\varepsilon(\mathsf{K}(y)\|\mathsf{K}(y')),$$

and hence $\eta_\varepsilon(\mathsf{K}) \leq \max_{y,y'} \mathsf{E}_\varepsilon(\mathsf{K}(y)\|\mathsf{K}(y'))$. Now we show that this inequality is indeed an equality. Fix $y_1 \neq y_2 \in \mathcal{Y}$ and $\delta \in (0,1)$. Define $\mu_\delta = \bar{\delta}\mathbb{I}_{\{y_0\}} + \delta\mathbb{I}_{\{y_1\}}$ and $\nu_\delta = (\bar{\delta}e^{-\varepsilon})\mathbb{I}_{\{y_0\}} + (1-\bar{\delta}e^{-\varepsilon})\mathbb{I}_{\{y_2\}}$ where $\bar{\delta} := 1-\delta$, $y_0 \notin \{y_1, y_2\}$ and $\mathbb{I}_{\{\cdot\}}$ is the indicator function. It is easy to verify that $\mathsf{E}_\varepsilon(\mu_\delta\|\nu_\delta) = \delta$. We also have $\mu_\delta\mathsf{K} = \bar{\delta}\mathsf{K}(y_0) + \delta K(y_1)$ and $\nu_\delta\mathsf{K} = (\bar{\delta}/e^\varepsilon)\mathsf{K}(y_0) + (1-\bar{\delta}/e^\varepsilon)\mathsf{K}(y_2)$. Hence, by (2),

$$\begin{aligned}
\mathsf{E}_\varepsilon(\mu_\delta\mathsf{K}\|\nu_\delta\mathsf{K}) &= \delta \int_{\mathcal{Y}} \left[\mathrm{d}(\mathsf{K}(y_1) - e^{\tilde{\varepsilon}}\mathsf{K}(y_2))(y)\right]_+ \\
&= \delta\mathsf{E}_{\tilde{\varepsilon}}(\mathsf{K}(y_1)\|\mathsf{K}(y_2)),
\end{aligned}$$

where $\tilde{\varepsilon} := \log(1 + \frac{e^\varepsilon - 1}{\delta})$. Therefore, we obtain that

$$\eta_\varepsilon(\mathsf{K}) \geq \frac{\mathsf{E}_\varepsilon(\mu_\delta\mathsf{K}\|\nu_\delta\mathsf{K})}{\mathsf{E}_\varepsilon(\mu_\delta\|\nu_\delta)} = \mathsf{E}_{\tilde{\varepsilon}}(\mathsf{K}(y_1)\|\mathsf{K}(y_2)).$$

By continuity of $\varepsilon \mapsto \mathsf{E}_\varepsilon(\mu||\nu)$, we obtain from above

$$\eta_\varepsilon(\mathsf{K}) \geq \lim_{\delta\to 1} \mathsf{E}_{\tilde{\varepsilon}}(\mathsf{K}(y_1)\|\mathsf{K}(y_2)) = \mathsf{E}_\varepsilon(\mathsf{K}(y_1)\|\mathsf{K}(y_2)).$$

Since y_1 and y_2 are arbitrary, the desired result follows. □

This theorem has an important implication: Gaussian kernels defined as $\mathsf{K}(y) = \mathcal{N}(y, \sigma^2\mathrm{I})$ for $y \in \mathbb{R}^d$ and some $\sigma > 0$ has a trivial contraction coefficient, i.e., $\eta_\varepsilon(\mathsf{K}) = 1$. However, if y is restricted to a bounded subset of $\mathbb{R}^d$, then $\eta_\varepsilon(\mathsf{K}) < 1$, as indicated by the following lemma.

Lemma 2. *Let $\mathcal{Y} \subset \mathbb{R}^d$ be a bounded set. For the Markov kernel specified by $\mathsf{K}(y) = \mathcal{N}(y, \sigma^2\mathrm{I})$ for $y \in \mathcal{Y}$ and $\sigma > 0$, we have*

$$\eta_\varepsilon(\mathsf{K}) = \theta_\varepsilon\Big(\frac{\|\mathcal{Y}\|}{\sigma}\Big),$$

where $\|\mathcal{Y}\| := \max_{y_1,y_2\in\mathcal{Y}} \|y_1 - y_2\|$.

The proof is a rather straightforward application of Theorem 1 and Lemma 1 and is hence omitted. The constraint that the input of Gaussian kernels must be bounded is not restrictive in machine learning and is satisfied in many practical algorithms. For instance, each iteration of the *projected* noisy stochastic gradient descent with Gaussian noise (see e.g., (Balle et al., 2019b; Bassily et al., 2014; 2019; Chaudhuri et al., 2011; Song et al., 2013; Wu et al., 2017)) can be viewed as a Gaussian kernel whose input (and output) are values from a compact set. Such kernels are called *projected* Gaussian kernels. We focus on this particular kernel in the next section.

3. Federated Learning

In our federated learning model, n distributed users send their updates of a shared model to a *trusted* aggregator. At each iteration, m number of users are chosen uniformly *without replacement*. Then, each of users selected computes a local update, randomizes it via a Gaussian kernel, and then sends it to the aggregator. The aggregator aggregates all these local updates, projects it onto ℓ_2-ball of fixed radius ρ and then sends the global update back to users. For

notational simplicity, we assume $m = qn$ and since the subsampling is performed without replacement, the total number of iteration is $T = \frac{n}{m} = \frac{1}{q}$. This procedure is described in Algorithm 1. The model we investigate differs from the typical settings studied in literature in that here the aggregator is expected to publicly display the model parameters only after the Tth iteration. This model is conceptually similar to the recent work of Augenstein et al. (2020) where the final model parameters were used to generate the synthetic data for the purpose of data inspection under privacy constraint.

3.1. Warm-Up: Batches of Size 1

Suppose n users, each with local data x_i, $i \in [n] \coloneqq \{1, \dots, n\}$, are to communicate over an encrypted communication channel to a trusted party and send their local update to shared model *one at a time*, i.e., $m = 1$. Although this setting may not be practical, it illuminates the proof technique employed for the general setting (i.e., $m \geq 1$).

Let $\pi \in \mathcal{S}_n$ be a random permutation map and $\mathcal{S}_n$ is the symmetric group on $[n]$. The federated learning algorithm iterates as follows:

- The aggregator samples the initial parameter W_0 in $\mathsf{ball}(\rho)$, the ℓ_2 ball of radius ρ in $\mathbb{R}^d$, according to a distribution μ_0 and sends it to user $\pi(1)$.
- User $\pi(1)$ uses W_0 and her local data $x_{\pi(1)}$ to compute the update $\tilde{W}_1 \coloneqq \eta \nabla \ell(W_0, x_{\pi(1)}) + \eta\sigma Z_1$, where $Z_1 \sim \mathcal{N}(0, \mathrm{I})$. This update is then sent back to the aggregator.
- Upon receipt of $\tilde{W}_1$, the aggregator computes $W_1 = \mathsf{proj}_\rho(W_0 - \tilde{W}_1)$, where $\mathsf{proj}_\rho(\cdot)$ denotes the projection operator onto $\mathsf{ball}(\rho)$. Then W_1 is sent to user $\pi(2)$.
- Continue the above procedure until all n users send the aggregator their updates (i.e., $T = n$ is the number of iterations). The aggregator releases W_T.

To obtain the privacy guarantee of this algorithm, we model each iteration as a projected Gaussian Markov kernel. Let K_t be the Markov kernel associated with the map $w \mapsto \mathsf{proj}_\rho\left(\Psi_t(w) - \eta\sigma Z_t\right)$ for $t \in [T]$, where

$$\Psi_t(w) \coloneqq w - \eta \nabla \ell(w, x_{\pi(t)}), \tag{10}$$

and Z_t is the standard Gaussian random variable. More precisely, $\mathsf{K}_t(w) = \mathsf{proj}_\rho(\mathcal{N}(\Psi_t(w), \eta^2\sigma^2\mathrm{I}))$. It is clear from Lemma 2 that $\eta_\varepsilon(\mathsf{K}_t) < 1$ for all $\varepsilon \geq 0$ and $\rho < \infty$. Notice that the tth iteration can be equivalently expressed by K_t whose input is W_{t-1} and output is W_t (See Fig 1). Letting μ_{t-1} denote the distribution of W_{t-1}, we therefore have $W_t \sim \mu_{t-1}\mathsf{K}_t$.

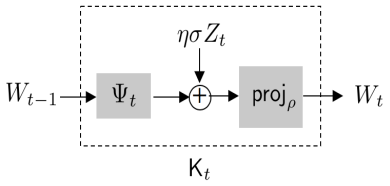

Figure 1. Iteration t can be viewed as a Markov kernel that is composed Ψ_t defined in (10), Gaussian noise addition and then projection operator onto $\mathsf{ball}(\rho)$.

Now consider a pair of neighboring datasets x and x' that differ in the ith entry (i.e., $x_i \neq x'_i$ and $x_j = x'_j$ for $j \in [n] \backslash \{i\}$) and let μ_t and μ'_t be the distributions of the W_t when algorithm runs on x and x', respectively. Let $t = \pi^{-1}(i)$ (or equivalently $\pi(t) = i$). Clearly, $\mu_j = \mu'_j$ for all $j \in [t-1]$. Also, $\mu_t = \mu_{t-1}\mathsf{K}_t$ and $\mu'_t = \mu_{t-1}\mathsf{K}'_t$ where K'_t is the Markov kernel associated with the map $w \mapsto \mathsf{proj}_\rho\left(\Psi'_t(w) - \eta\sigma Z_t\right)$, where

$$\Psi'_t(w) \coloneqq w - \eta \nabla \ell(w, x'_i).$$

In light of (1), one concludes the algorithm is (ε, δ)-DP if $\mathsf{E}_\varepsilon(\mu_T \| \mu'_T) \leq \delta$, for all $i \in [n]$. By definition, we have

$$\begin{aligned}\mathsf{E}_\varepsilon(\mu_T \| \mu'_T) &\leq \mathsf{E}_\varepsilon(\mu_{T-1} \| \mu'_{T-1})\eta_\varepsilon(\mathsf{K}_T) \\ &\leq \mathsf{E}_\varepsilon(\mu_{T-2} \| \mu'_{T-2})\eta_\varepsilon(\mathsf{K}_T)\eta_\varepsilon(\mathsf{K}_{T-1}).\end{aligned}$$

Applying this for $T - t$ times, we obtain

$$\begin{aligned}\mathsf{E}_\varepsilon(\mu_T \| \mu'_T) &\leq \mathsf{E}_\varepsilon(\mu_t \| \mu'_t) \prod_{j=t+1}^{T} \eta_\varepsilon(\mathsf{K}_j) \\ &= \mathsf{E}_\varepsilon(\mu_{t-1}\mathsf{K}_t \| \mu_{t-1}\mathsf{K}'_t) \prod_{j=t+1}^{T} \eta_\varepsilon(\mathsf{K}_j) \qquad (11)\end{aligned}$$

Consequently, the computation of δ boils down to computing the contraction coefficient of projected Gaussian kernels and E_ε-divergence between mixture of projected Gaussian distributions with the same variance. The former can be tackled via Lemma 2. The latter, however, involves Jensen's inequality (recall that $(\mu, \nu) \mapsto \mathsf{E}_\varepsilon(\mu \| \nu)$ is convex), the data processing inequality (to get rid of the projection operator) and Lemma 1. We will elaborate further in the next section where we prove the main result.

3.2. Batch of size m

Here we assume at each iteration, the aggregator shares the global update with m users. In this setting, $T = \frac{n}{m}$ and, in lieu of permutation, we define a mapping which assigns each $i \in [n]$ to a single batch.

Algorithm 1 Federated learning with a trusted aggregator

Input: Dataset $\{x_1, \dots, x_n\} \in \mathbb{R}^{nd}$, learning rate η, batch size m, noise variance σ^2, initial distribution μ_0
Choose $W_0 \sim \mu_0$
for $t = 1$ **to** T **do**
  Take batch $\mathsf{B}_t \subset [n]$ of size m uniformly without replacement
  Local update: $W_{t-1}^j = \eta\big[\nabla \ell(W_{t-1}, x_j) + Z_t^j\big]$, $\forall j \in \mathsf{B}_t$ and $Z_t^j \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$
  Upload: W_{t-1}^j is sent to aggregator
  Model aggregation: aggregator updates the model parameter as $W_t = \mathsf{proj}_\rho(W_{t-1} - \frac{1}{m}\sum_{j\in\mathsf{B}_t} W_t^j)$
end for
Output: W_T

Theorem 2. *Let the loss function $w \mapsto \ell(w, x)$ be convex, L-Lipschitz and β-smooth for all $x \in \mathcal{X}$ and also $\eta \le \frac{2}{\beta}$. Then Algorithm 1 is (ε, δ)-DP for $\varepsilon \ge 0$ and*

$$\delta = \frac{m}{n}\theta_\varepsilon\Big(\frac{2L}{\sqrt{m}\sigma}\Big)\frac{1-\theta_\varepsilon\Big(\frac{2\rho\sqrt{m}}{\eta\sigma}\Big)^{\frac{n}{m}}}{1-\theta_\varepsilon\Big(\frac{2\rho\sqrt{m}}{\eta\sigma}\Big)},$$

where θ_ε is defined in (5).

Proof. Consider two neighboring datasets $x = \{x_1, \dots, x_n\}$ and $x' = \{x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n\}$. Let μ_t and μ_t' be the distribution of W_t the output of the tth iteration when running on x and x', respectively. To derive δ for any given ε, we need to compute $\mathsf{E}_\varepsilon(\mu_T \| \mu_T')$. Let $\pi : [n] \to [T]$ specifies an assignment of users to each batch, i.e., $\pi(i) = t$ if $i \in \mathsf{B}_t$. Note that $\mu_t = \mu_t'$ for $t < \pi(i)$. We now identify each iteration with a projected Markov kernel. At iteration t, the aggregator generates

$$W_t = \mathsf{proj}_\rho\Big(W_{t-1} - \frac{\eta}{m}\sum_{j\in\mathsf{B}_t}\nabla\ell(W_{t-1}, x_j) - \tilde{\sigma} Z_t\Big),$$

where Z_t is now standard Gaussian random variable and $\tilde{\sigma}^2 := \frac{\eta^2\sigma^2}{m}$. Hence, iteration t can be realized by K_t a projected Markov kernel associated with the mapping $w \mapsto \mathsf{proj}_\rho(\Psi_t(w) - \tilde{\sigma} Z_t)$ where

$$\Psi_t(w) = w - \frac{\eta}{m}\sum_{j\in\mathsf{B}_t}\nabla\ell(w, x_j).$$

Notice that K_t receives W_{t-1} and generates W_t both taking values in $\mathsf{ball}(\rho)$. Due to the strong data processing inequality (see (11)) and convexity of $(\mu, \nu) \mapsto \mathsf{E}_\varepsilon(\mu\|\nu)$ for any $\varepsilon \ge 0$, we can write

$$\begin{aligned}\mathsf{E}_\varepsilon(\mu_T\|\mu_T') &\le \sum_{t=1}^T \Pr(\pi(i) = t)\mathsf{E}_\varepsilon(\mu_t\|\mu_t')\prod_{j=t+1}^T \eta_\varepsilon(\mathsf{K}_j)\\ &= q\sum_{t=1}^T \mathsf{E}_\varepsilon(\mu_t\|\mu_t')\prod_{j=t+1}^T\eta_\varepsilon(\mathsf{K}_j) \qquad (12)\end{aligned}$$

To compute a bound for δ, it thus suffices to compute $\eta_\varepsilon(\mathsf{K}_j)$ for $j \in [T]$ and $\mathsf{E}_\varepsilon(\mu_t\|\mu_t')$ for $t \in [T]$. We begin by computing $\eta_\varepsilon(\mathsf{K}_j)$ for $j \in [T]$ as follows

$$\begin{aligned}\eta_\varepsilon(\mathsf{K}_j) &= \sup_{w_1, w_2 \in \mathsf{ball}(\rho)} \mathsf{E}_\varepsilon(\mathsf{K}_j(w_1)\|\mathsf{K}_j(w_2))\\ &\le \sup_{w_1, w_2 \in \mathsf{ball}(\rho)} \mathsf{E}_\varepsilon(\mathcal{N}(\Psi_j(w_1), \tilde{\sigma}^2\mathrm{I})\|\mathcal{N}(\Psi_j(w_2), \tilde{\sigma}^2\mathrm{I})) \qquad (13)\\ &= \sup_{w_1, w_2 \in \Psi_j(\mathsf{ball}(\rho))} \mathsf{E}_\varepsilon(\mathcal{N}(w_1, \tilde{\sigma}^2\mathrm{I})\|\mathcal{N}(w_2, \tilde{\sigma}^2\mathrm{I}))\\ &= \theta_\varepsilon\Big(\frac{\Psi_j(\mathsf{ball}(\rho))}{\tilde{\sigma}}\Big) \qquad (14)\\ &\le \theta_\varepsilon\big(\frac{2\rho}{\tilde{\sigma}}\big) \qquad (15)\\ &= \theta_\varepsilon\Big(\frac{2\rho\sqrt{m}}{\eta\sigma}\Big) \qquad (16)\end{aligned}$$

where the inequality in (13) is due to the data processing inequality:

$$\begin{aligned}&\mathsf{E}_\varepsilon(\mathsf{proj}_\rho(\mathcal{N}(\Psi_j(w_1), \tilde{\sigma}^2\mathrm{I}))\|\mathsf{proj}_\rho(\mathcal{N}(\Psi_j(w_2), \tilde{\sigma}^2\mathrm{I})))\\ &\qquad \le \mathsf{E}_\varepsilon(\mathcal{N}(\Psi_j(w_1), \tilde{\sigma}^2\mathrm{I})\|\mathcal{N}(\Psi_j(w_2), \tilde{\sigma}^2\mathrm{I})).\end{aligned}$$

Also, the equality in (14) follows from Lemma 1 and (5), and finally, the inequality in (15) follows from the following two facts: (1) Since the loss functions $w \mapsto \ell(w, x)$ is convex and β-smooth for all $x \in \mathcal{X}$, then $w \mapsto w - \nabla\ell(w, x)$ is contractive for $\eta \le \frac{2}{\beta}$ (see e.g., Prop 18 in (Feldman et al., 2018)) and so is $w \mapsto \Psi_j(w)$; and (2) The map $r \mapsto \theta_\varepsilon(r)$ is increasing.

Next, we compute $\mathsf{E}_\varepsilon(\mu_t\|\mu_t')$. Note that

$$\mu_t = \int_{\mathsf{ball}(\rho)} \mu_{t-1}(\mathrm{d}y)\mathsf{K}_t(y).$$

Since $\pi(i) = t$, data point $x_i' \in \mathsf{B}_t$. For this batch, we define

$$\Psi_t'(w) := w - \frac{\eta}{m}\Big[\nabla\ell(w, x_i') + \sum_{j\in\mathsf{B}_t\setminus\{i\}}\nabla\ell(w, x_j)\Big],$$

and the corresponding Markov kernel K_t' associated with $w \mapsto \mathsf{proj}_\rho(\Psi_t'(w) - \tilde{\sigma} Z_t)$. It follows that

$$\mu_t' = \int_{\mathsf{ball}(\rho)} \mu_{t-1}(\mathrm{d}y)\mathsf{K}_t'(y).$$

The convexity of $(\mu, \nu) \mapsto \mathsf{E}_\varepsilon(\mu\|\nu)$ implies

$$\mathsf{E}_\varepsilon(\mu_t\|\mu_t') \le \int \mathsf{E}_\varepsilon(\mathsf{K}_t(y)\|\mathsf{K}_t'(y))\mu_{t-1}(\mathrm{d}y) \tag{17}$$

$$\le \int \mathsf{E}_\varepsilon(\mathcal{N}(\Psi_t(y), \tilde{\sigma}^2\mathrm{I})\|\mathcal{N}(\Psi_t'(y), \tilde{\sigma}^2\mathrm{I})\mu_{t-1}(\mathrm{d}y) \tag{18}$$

$$= \int \theta_\varepsilon\Big(\frac{\|\Psi_t(y) - \Psi_t'(y)\|}{\tilde{\sigma}}\Big)\mu_{t-1}(\mathrm{d}y) \tag{19}$$

$$\le \theta_\varepsilon\big(\frac{2L\eta}{m\tilde{\sigma}}\big) \tag{20}$$

$$\le \theta_\varepsilon\Big(\frac{2L}{\sqrt{m}\sigma}\Big) \tag{21}$$

where (17) follows from Jensen's inequality, (18) follows from the data processing inequality, (19) follows from Lemma 1, and finally (20) is due to Lemma 2 as follows: for any $y \in \mathsf{ball}(\rho)$

$$\|\Psi_t(y) - \Psi_t'(y)\| = \|\frac{\eta}{m}(\nabla\ell(y, x_i) - \nabla\ell(y, x_i'))\|$$
$$\le \frac{2L\eta}{m}$$

where the inequality is due to the fact that $w \mapsto \ell(w, x)$ is L-Lipschitz for all $x \in \mathcal{X}$ and hence $\|\nabla\ell(w, x)\| \le 2L$.

Plugging (16) and (21) into (12), we obtain the

$$\mathsf{E}_\varepsilon(\mu_T\|\mu_T') \le q\theta_\varepsilon\Big(\frac{2L\sqrt{m}}{\sigma}\Big)\sum_{t=1}^{T}\theta_\varepsilon^{T-t}\Big(\frac{2\rho\sqrt{m}}{\eta\sigma}\Big)$$
$$= q\theta_\varepsilon\Big(\frac{2L\sqrt{m}}{\sigma}\Big)\frac{1 - \theta_\varepsilon\Big(\frac{2\rho\sqrt{m}}{\eta\sigma}\Big)^{\frac{n}{m}}}{1 - \theta_\varepsilon\Big(\frac{2\rho\sqrt{m}}{\eta\sigma}\Big)}.$$

□

Notice that if the cost function is strongly convex, then Theorem 2 can be improved as, in this case, $w \mapsto w - \nabla\ell(w, x)$ is contractive with Lipschitz constant strictly smaller than 1 (see, e.g., Theorem 3.12 in (Bubeck, 2015)). In Fig. 2, we demonstrate the privacy parameters obtained from Theorem 2 for different sub-sampling rates $q = 0.1, 0.2, 0.3$. As illustrated in this figure, the more users are involved in each iteration, the higher the privacy guarantee is.

4. Conclusion

In this work, we introduce a new approach for computing differential privacy (DP) parameters via contraction coefficient of Markov kernels under a certain f-divergence, namely E_ε-divergence. In this approach, we interpret federated learning algorithm as a composition of several Markov kernels and express the DP privacy parameters as the product of contraction coefficients of such kernels. The main assumption is that the algorithm releases the model update only after a certain number of iterations are passed; thus no composition theorems are required. The proof technique relies on a technical theorem that establishes a close-form expression for the contraction coefficient of general Markov kernels under E_ε-divergence.

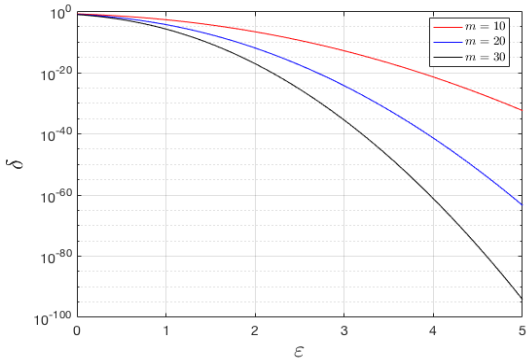

Figure 2. Differential privacy parameters of Algorithm 1 for different sub-sampling rates according to Theorem 2. The parameters of algorithm are as follows: $\eta = 0.5, L = 1, \rho = 1, \sigma = 1.5, n = 100$.

This approach can be adapted to study the the more typical scenario where the model updates get released after each iteration. The privacy analysis in this case amounts to deriving the contraction coefficient of a Markov kernel that is obtained by tensor product of all T kernels, i.e., a kernel with T-tuple input and output, under E_ε-divergence.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proc. ACM SIGSAC CCS*, pp. 308–318, 2016. doi: 10.1145/2976749.2978318.

Ahlswede, R. and Gács, P. Spreading of sets in product spaces and hypercontraction of the markov operator. *Ann. Probab.*, 4(6):925–939, 12 1976. doi: 10.1214/aop/1176995937.

Ali, S. M. and Silvey, S. D. A general class of coefficients of divergence of one distribution from another. *Journal of Royal Statistics*, 28:131–142, 1966.

Asoodeh, S., Diaz, M., and Calmon, F. P. Privacy amplification of iterative algorithms via contraction coefficients. 2020. URL https://arxiv.org/abs/2001.06546.

Augenstein, S., McMahan, H. B., Ramage, D., Ramaswamy, S., Kairouz, P., Chen, M., Mathews, R., and y Ar-

cas, B. A. Generative models for effective ml on private, decentralized datasets. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgaRA4FPH.

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

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

Balle, B., Barthe, G., Gaboardi, M., and Geumlek, J. Privacy amplification by mixing and diffusion mechanisms. In *NeurIPS*, pp. 13277–13287. 2019a.

Balle, B., Barthe, G., Gaboardi, M., Hsu, J., and Sato, T. Hypothesis testing interpretations and Rényi differential privacy. *ArXiv*, abs/1905.09982, 2019b.

Balle, B., Bell, J., Gascón, A., and Nissim, K. The privacy blanket of the shuffle model. 11693:638–667, 2019c. doi: 10.1007/978-3-030-26951-7_22.

Barthe, G. and Olmedo, F. Beyond differential privacy: Composition theorems and relational logic for f-divergences between probabilistic programs. In *Proc. ICALP*, pp. 49–60, 2013. ISBN 978-3-642-39211-5.

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

Bassily, R., Feldman, V., Talwar, K., and Guha Thakurta, A. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems 32*, pp. 11282–11291. Curran Associates, Inc., 2019.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pp. 1175–1191, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349468. doi: 10.1145/3133956.3133982. URL https://doi.org/10.1145/3133956.3133982.

Bubeck, S. *Convex Optimization: Algorithms and Complexity*, volume 8. Foundations and Trends in Machine Learning, 2015.

Chaudhuri, K. and Mishra, N. When random sampling preserves privacy. In *Advances in Cryptology - CRYPTO 2006*, pp. 198–213. Springer Berlin Heidelberg, 2006.

Chaudhuri, K. and Monteleoni, C. Privacy-preserving logistic regression. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems*, pp. 289–296. 2009.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

Csiszár, I. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

Dobrushin, R. L. Central limit theorem for nonstationary markov chains. I. *Theory Probab. Appl.*, 1(1):65–80, 1956.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *Proc. of IEEE Foundations of Computer Science (FOCS)*, 2013.

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

Fredrikson, M., Jha, S., and Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pp. 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450338325. doi: 10.1145/2810103.2813677. URL https://doi.org/10.1145/2810103.2813677.

Georgii, H. *Gibbs Measures and Phase Transitions*. De Gruyter Studies in Mathematics. De Gruyter, 2011. ISBN 9783110250329. URL https://books.google.com/books?id=Xl_NocttwvAC.

Jain, P. and Thakurta, A. G. (near) dimension independent risk bounds for differentially private learning. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 476–484, Bejing, China, 22–24 Jun 2014. PMLR.

Jain, P., Kothari, P., and Thakurta, A. Differentially private online learning. In Mannor, S., Srebro, N., and Williamson, R. C. (eds.), *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 24.1–24.34, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. URL http://proceedings.mlr.press/v23/jain12.html.

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

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2016.

Melis, L., Song, C., Cristofaro, E. D., and Shmatikov, V. Inference attacks against collaborative learning. *ArXiv*, abs/1805.04049, 2018.

Polyanskiy, Y. and Wu, Y. Dissipation of information in channels with input constraints. *IEEE Trans. Inf. Theory*, 62(1):35–55, Jan 2016. ISSN 0018-9448. doi: 10.1109/TIT.2015.2482978.

Smith, A., Thakurta, A., and Upadhyay, J. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 58–77, 2017.

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 245–248, 2013.

Talwar, K., Thakurta, A., and Zhang, L. Nearly-optimal private lasso. In *International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pp. 3025–3033, Cambridge, MA, USA, 2015. MIT Press.

Thakurta, A. G. and Smith, A. Differentially private feature selection via stability arguments, and the robustness of the lasso. In Shalev-Shwartz, S. and Steinwart, I. (eds.), *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 819–850, Princeton, NJ, USA, 12–14 Jun 2013. PMLR.

Wang, D., Ye, M., and Xu, J. Differentially private empirical risk minimization revisited: Faster and more general. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 2719–2728, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P. Subsampled Rényi differential privacy and analytical moments accountant. In *AISTAT*, volume 89, pp. 1226–1235, 16–18 Apr 2018.

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

Xu, A. and Raginsky, M. Converses for distributed estimation via strong data processing inequalities. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 2376–2380, 2015.