The Privacy Analysis of the Differential Private Stochastic Gradient Descent

Fabrizio Boninsegna

***Abstract*—As data privacy concerns become increasingly paramount, the development and analysis of privacy-preserving machine learning algorithms have emerged as a critical area of research. This paper provides an introduction to the diverse techniques and methodologies employed in the privacy analysis of the Differential Private Stochastic Gradient Descent (DP-SGD) algorithm, a pivotal tool in the realm of privacy-preserving machine learning.**

To guarantee private learning, DP-SGD simply injects controlled noise into the gradient updates during model training. Nonetheless, the privacy analysis of the algorithm still gives some challenges. Standard differential private composition techniques are not useful for DP-SGD as it requires the addition of noise many times in the mini batch setting. To solve these problems, many techniques have been developed that make full use of the noise distribution and the mathematical properties of the loss function. In particular, we will delve into the *Moments Accountant* method, the *Privacy Amplification by Iteration* and *Langevin Dynamics*, by exposing these techniques under a common notation.

I. Introduction

Machine Learning and Deep Learning suffer from privacy issues, as noted by Fredrikson et al. who demonstrated a black box model-inversion attack that recovers images from a facial recognition system [FJR15]. Moreover, an adversary with full knowledge of the training mechanism and access to the model's parameters is more likely to reconstruct part of the training dataset due to memorization [HVY+22]. Differential privacy emerges as a powerful tool in this context, providing a framework to use data for training deep learning models while safeguarding individual privacy.

II. Preliminaries of Differential Privacy

Differential privacy is a mathematical concept that ensures the outputs of a computation (like a deep learning model's predictions) are not significantly affected by any single data point. In simple terms, it guarantees that the removal or addition of one individual's data does not substantially change the outcome of a data analysis or the behavior of a model. The privacy guarantees of a differential private mechanism are measured by the *privacy budget* ε and by the *failure probability* δ, which measures the probability to violet the privacy guarantees. In the literature it is considered accepted a privacy budget $\varepsilon \in (0, 10)$ and a $\delta = O(1/n)$ where n is the dimension of the dataset considered.

Definition II.1 ((ε, δ)-differential privacy [DR+14])**.** A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ε, δ)-differential privacy if for any two adjacent inputs d,$d' \in \mathcal{D}$ and for any subset of outputs $S \subseteq \mathcal{R}$ it holds that

$$\Pr[\mathcal{M}(d) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(d') \in S] + \delta \tag{1}$$

Another way to define differential privacy is by using the *privacy loss* random variable.

Definition II.2 (Privacy loss random variable [DR16])**.** The privacy loss of a randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ is a random variable defined as the log-likelyhood ratio of two distinct inputs

$$\xi(y; \mathcal{M}, x, x') = \log \left(\frac{\Pr[\mathcal{M}(x) = y]}{\Pr[\mathcal{M}(x') = y]} \right) \tag{2}$$

The relation between privacy loss and (ε, δ)-DP is the following

Theorem II.1 (Privacy loss and (ε, δ)-DP)**.** If a randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ has a privacy loss such that

$$\Pr[\sup_{y \in \mathcal{R}} \xi(y; \mathcal{M}, x, x') \geq \varepsilon] \leq \delta \qquad \forall x \sim x',$$

then $\mathcal{M}$ is (ε, δ)-DP.

A common paradigm for approximating a deterministic real-valued function $f : \mathcal{D} \to \mathbb{R}^d$ with a differential private mechanism is via *additive noise* calibrated to f's *sensitivity* $S_f^{(p)}$, which is defined as the maximum absolute ℓ_p distance $S_f^{(p)} = \sup_{d \sim d'} ||f(d) - f(d')||_p$ computed on adjacent inputs. The injection of Gaussian noise satisfies differential privacy

Theorem II.2 (Gaussian Mechanism in the High Privacy Regime [DR+14])**.** Let $\varepsilon \in (0, 1)$ be arbitrary. For $c^2 > 2\ln(1.25/\delta)$, the Gaussian Mechanism $\mathcal{M}(f(x)) = f(x) + \mathcal{N}(0, \sigma^2 \mathbb{1}_d)$ with parameter $\sigma \geq c S_f^{(2)}/\varepsilon$ is (ε, δ)-differential private.

The Gaussian Mechanism is still (ε, δ)-differential private for large value of ε, but the computation of σ is more complicated (Theorem 8 in [BW18]).

A crucial property of differential privacy is that it is closed under adaptive composition. Therefore, we can use many differential private mechanism on the same private dataset, using adaptively the results of each mechanism, and still get a differential private release.

Theorem II.3 (Advanced Composition [DR+14])**.** For all $\varepsilon, \delta, \delta' \geq 0$, the adaptive composition of k differential private mechanism with the same parameter ε, δ is $(\varepsilon', k\delta + \delta')$-differential private mechanism with:

$$\varepsilon' = \sqrt{2k \ln(1/\delta')}\varepsilon + k\varepsilon(e^{\varepsilon} - 1). \tag{3}$$

In the high privacy regime $\varepsilon < 1$ we have that $\varepsilon' = 2\sqrt{2k\ln(1/\delta')}\varepsilon$.

When a differential private mechanism gets inputs on a random subset of the dataset, which is usually the case for stochastic gradient descent, the privacy is amplified by a constant factor.

Theorem II.4 (Privacy Amplification by Subsampling [BBG18])**.** If $\mathcal{M}$ is an (ε, δ)-DP mechanism, then the subsample mechanism $\mathcal{M} \circ \mathcal{S}$, where S selects a random sample of $\mathcal{D}$ using Possion sampling with probability q, is $(\log(1 + q(e^{\varepsilon} - 1)), q\delta)$-DP.

In particular, in the high privacy regime $\varepsilon < 1$ we have that $\mathcal{M} \circ \mathcal{S}$ is $(O(q\varepsilon), O(q\delta))$-DP.

III. Preliminaries of Stochastic Gradient Descent

Deep neural networks define parameterized functions from inputs to outputs as compositions of many neural network layers. These parameters are ”trained” on data by minimizing an objective function called loss. More precisely, the loss $\mathcal{L}(\theta, D)$ on parameters $\theta \in \mathbb{R}^d$ is the average of the loss over the training examples $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, so $\mathcal{L}(\theta, D) = \frac{1}{n}\sum_i \ell(\theta, \mathbf{x}_i)$. For complex networks, a standard practice to find the parameters which minimize the loss is by stochastic gradient descent. At each step one forms a batch B of random examples and computes the batch loss gradient $g_B = \frac{1}{|B|}\sum_i \nabla_\theta \ell(\theta, x_i)$ as an estimation of the true gradient $\nabla_\theta \mathcal{L}(\theta)$. Then θ is updated following the gradient direction $-g_B$ towards the local minimum.

To render the training phase differential private the standard procedure is to inject Gaussian noise into the batch gradient loss $\tilde{g}_B = g_B + \mathcal{N}(0, \sigma^2(S^{(2)}_{g_B})^2\mathbb{1}_d)$. Notice that we insert the sensitivity of the batch gradient loss directly in the variance of the Gaussian distribution for convenience, as this allows us to consider any function with sensitivity one. The sensitivity of the gradient can be computed using the Lipschitz properties of the loss, or simply by clipping the gradient on a maximum ℓ_2 norm. A comprehensive introduction the the loss sensitivity can be found in Appendix A.

A standard implementation of DP-SGD with gradient clipping was introduced by [ACG+16] and it is described in Algorithm 1.

Algorithm 1 DP-SGD: Differentially private stochastic gradient descent

```
Require: Examples $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, loss function $\ell(\theta, \mathbf{x})$,
    learning rate $\eta$, noise scale $\sigma^2$, sampling probability $q$,
    gradient norm bound $C$, number of iteration $T$.

    Initialize $\theta_0$ randomly
    for t $\in [T]$ do
        Sample random mini batch
        Take a random sample $B_t$ using Poisson sampling
        with probability $q$.
        Compute gradient
        For each $i \in B_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_\theta \ell(\theta_t, x_i)$
        Clip gradient
        // this fixes the sensitivity of the gradient
        $\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{||\mathbf{g}_t(x_i)||_2}{C}\right)$
        Add noise
        $\tilde{\mathbf{g}_t} \leftarrow \frac{1}{B}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, 4\sigma^2 C^2 \mathbb{1}_d)\right)$
        Descent
        $\theta_{t+1} \leftarrow \theta_t - \eta\tilde{\mathbf{g}}_t$
    end for
    return $\theta_T$ and compute the overall privacy cost $(\varepsilon, \delta)$
```

A. Standard Advanced Composition Result

By choosing $\sigma \geq \sqrt{2\ln(1.25/\delta)}/\varepsilon$ and $\varepsilon \in (0, 1)$, one application of Gaussian noise in stochastic gradient descent is satisfies (ε, δ)-differential privacy. Using privacy amplification by sub-sampling [BBKN14] we get a $(q\varepsilon, q\delta)$-differential private algorithm for $q = O(1/n)$ (to ensure a random batch of constant size), then by advanced composition of T iterations we ends up with

$$(2\sqrt{2T\ln(1/\delta')}q\varepsilon,\ qT\delta + \delta')\text{-differential private.}$$

Let's set the standard deviation in order to get a constant (ε, δ)-differentially private algorithm. By neglecting constant factor we have that for

$$\sigma \geq \Omega\left(\frac{q\sqrt{T\log(1/\delta)\log(T/\delta)}}{\varepsilon}\right) \tag{4}$$

we get an overall (ε, δ)-differentially private algorithm. This is the most simple result that we can get using standard properties of differential privacy. This result is independent by the noise distribution we are using, as long it satisfies (ε, δ)-DP, and by the loss properties.

IV. Moments Accountant

This is the standard privacy analysis implemented in most libraries for differential private deep learning [YSS+21]. It is a new technique for differential private composition that takes into account the shape of the probability density function used for generating noise. It is based on bounding the moment generating function of the privacy loss random variable. Consider the privacy loss $\xi(y; \mathcal{M}, x, x')$ defined in definition II.2 for adjacent datasets x and x'. For the mechanism $\mathcal{M}$, we define the λ^{th} moment $\alpha_{\mathcal{M}}(\lambda; x, x')$ as the log of the moment generating function evaluated at value λ:

$$\alpha_{\mathcal{M}}(\lambda; x, x') := \log \mathbb{E}_{y \sim \mathcal{M}(x)}[\exp(\lambda\xi(y; \mathcal{M}, x, x'))].$$

The quantity of interest is the maximum λ^{th} moment over all possible adjacent inputs

$$\alpha(\lambda) := \max_{x, x'} \alpha_{\mathcal{M}}(\lambda; x, x'). \tag{5}$$

The moment accountant theorem upper bounds $\alpha_{\mathcal{M}}(\lambda)$ for all λ, where $\mathcal{M}$ is an adaptive composition of k mechanisms. Moreover, it tells how to compute δ for any $\varepsilon > 0$.

Theorem IV.1 (Moments Accountant [ACG+16])**.** Let $\alpha_{\mathcal{M}}(\lambda)$ be defined as (5). Then

1) **Composability.** Suppose that a mechanism $\mathcal{M}$ consists of a sequence of adaptive mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_k$.

Then, for any λ

$$\alpha_{\mathcal{M}}(\lambda) \leq \sum_{i=1}^{k} \alpha_{\mathcal{M}_i}(\lambda)$$

2) **Tail-bound.** For any $\varepsilon > 0$, the mechanism $\mathcal{M}$ is (ε, δ)-differentially private for

$$\delta = \min_{\lambda} \exp(\alpha_{\mathcal{M}}(\lambda) - \lambda\varepsilon).$$

In particular, for the Gaussian noise the authors in [ACG+16] demonstrated an asymptotic upper bound for the maximum log moment generating function

$$\alpha(\lambda) \leq \frac{q^2\lambda(\lambda+1)}{(1-q)\sigma^2} + O\left(\frac{q^3}{\sigma^3}\right),$$

which allows to demonstrate a better bound for the T adaptive composition of Gaussian mechanisms.

Theorem IV.2 (Moment Accountant for Gaussian Mechanism [ACG+16])**.** There exist constant c_1 and c_2 so that given the sampling probability $q = O(1/n)$ (with n dimension of the private dataset) and the number of steps T, for any $\varepsilon < c_1 q^2 T$ the k adaptive composition of T Gaussian mechanisms on a 1-sensitivity function is (ε, δ)-differentially private for any $\delta > 0$ if we choose

$$\sigma \geq c_2 \frac{q\sqrt{T \log(1/\delta)}}{\varepsilon}$$

Asymptotically, with the moments accountant method we save a $\sqrt{\log(T/\delta)}$ factor in the standard deviation.

This new composition technique makes full use of the shape of the noise distribution, allowing to get smaller privacy budget. However, the technique is not as practical as the standard advanced composition, as it requires the study of the moment generating function of the privacy loss. This inspires the development of Rényi differential privacy.

V. Rényi Differential Privacy

Based on the work done by [ACG+16] with the moments accountant, and the new definition of concentrated differential privacy [DR16] (based on sub-Gaussian distributions) and zero-concentrated differential privacy [BS16] (based on Rényi divergence, but it requires a linear bound for all positive moments), I. Mironov [Mir17] developed the concept of Rènyi differential privacy based on the concept of Rényi divergence.

Definition V.1 (Rényi divergence)**.** The Rényi divergence of order α between two distribution μ and ν is

$$D_\alpha(\mu||\nu) = \frac{1}{\alpha - 1} \log \left[\int \left(\frac{\mu(x)}{\nu(x)} \right)^\alpha \nu(x) \mathrm{d}x \right]$$

Definition V.2 (Rényi differential privacy [Mir17])**.** A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ is said to have ε-Rényi differential privacy of order α, or (α, ε)-RDP for short, if for any adjacent $x, x' \in \mathcal{D}$ and $\alpha \in (1, \infty)$ it holds that

$$D_\alpha(\mathcal{M}(x)||\mathcal{M}(x')) \leq \varepsilon,$$

A bound on the Rényi divergence is a bound on the log moment generating function of the privacy loss. Indeed, we can rewrite (IV) as

$$\begin{aligned}\alpha_{\mathcal{M}}(\lambda; x, x') &= \log \mathbb{E}_{y\sim\mathcal{M}(x)} \left[\left(\frac{\Pr[\mathcal{M}(x) = y]}{\Pr[\mathcal{M}(x') = y]} \right)^\lambda \right] \\ &= \log \mathbb{E}_{y\sim\mathcal{M}(x')} \left[\left(\frac{\Pr[\mathcal{M}(x) = y]}{\Pr[\mathcal{M}(x') = y]} \right)^{\lambda+1} \right],\end{aligned}$$

where in the last equality we performed a change of measure in the expectation value. Recall the definition of Rényi differential privacy

$$\mathrm{D}_\alpha(\mathcal{M}(x)||\mathcal{M}(x')) = \tfrac{1}{\alpha-1} \log \mathbb{E}_{y\sim\mathcal{M}(x')} \left[\left(\tfrac{\Pr[\mathcal{M}(x)=y]}{\Pr[\mathcal{M}(x')=y]} \right)^\alpha \right],$$

it is clear the similarity between the upper bound on the Rényi divergence with the upper bound of the log moment generating function of the privacy loss. Indeed, we have

$$\mathcal{D}_{\lambda+1}(\mathcal{M}(x)||\mathcal{M}(x')) = \frac{1}{\lambda} \alpha_{\mathcal{M}}(\lambda; x, x').$$

The Rényi differential privacy definition is then a well suited mathematical tool for the moments accountant method.

A. Properties of the Rényi differential privacy

Rényi differential privacy behaves well under adaptive composition

Proposition V.1 (Adaptive Composition)**.** Consider k mechanisms satisfying (α, ε)-RDP. The k adaptive composition of these mechanisms satisfies $(\alpha, k\varepsilon)$-RDP.

As the Rényi differential privacy is built from the generating function of the privacy loss, the moment accountant composition applies fundamentally. Indeed, this composition is a direct application of the composability theorem IV.1

It is also relatively easy to turn Rényi differential privacy into standard differential privacy

Proposition V.2 (From RDP to (ε, δ)-DP)**.** If $\mathcal{M}$ is an (α, ε)-RDP mechanism, it also satisfies $(\varepsilon + \frac{\log(1/\delta)}{\alpha-1}, \delta)$-differential privacy for any $0 < \delta < 1$.

However, this transition from RDP to DP is not optimal [ZDW22].

The main advantage of using Rényi differential privacy is that it offers a nice and simple analysis for the Gaussian mechanism.

Proposition V.3. The Gaussian mechanism on μ-sensitivity functions satisfies $(\alpha, \frac{\alpha\mu^2}{2\sigma^2})$-RDP.

Usually, many theorems in differential privacy are stated for 1-sensitivity function. For the Gaussian mechanism is easy to transform μ-sensitive function into 1-sensitive one. It is sufficient to rescale the noise variance σ^2

Proposition V.4. The μ re-scaled Gaussian mechanism

$$\mathcal{M}_G(f(D)) = f(D) + \mathcal{N}(0, \sigma^2\mu^2\mathbb{1}_d),$$

is $\left(\alpha, \frac{\alpha}{2\sigma^2}\right)$ for μ-sensitivity functions.

The privacy amplification by subsampling property for Rényi differential privacy was study in [WBK19]. Their theorem 9 offer a full non asymptotic view of this property and it is quite complicated. Still, for large standard deviation, small α and small $q = O(1/n)$, the amplification factor from $(\alpha, \alpha/2\sigma^2)$-RDP is

$$\left(\alpha, O\left(\frac{q^2\alpha}{\sigma^2}\right)\right) - \text{RDP}$$

A precise analytical and numerical evaluation of the Sample Gaussian Mechanism is given in [MTZ19],

B. Rényi Differential Privacy for the Stochastic Gradient Descent

By applying standard composition of Rényi differential privacy for the Gaussian mechanism, and the asymptotic privacy amplification by subsampling in [WBK19] we obtain the same asymptotic result of the Moment Accountant method, hence a standard deviation lower bound

$$\sigma \geq \Omega\left(\frac{q\sqrt{T\log(1/\delta)}}{\varepsilon}\right).$$

The proof is in the Appendix B1.

Still, this amplification holds only in the setting of private sub-sample and mostly important, only in the high privacy regime, hence for $\varepsilon \ll 1$ or $\sigma \gg 1$. In the next section we present a new technique that allows to relax this assumption on the privacy regime.

VI. Privacy Amplification by Iteration

Privacy amplification by subsampling depends crucially on the sample's randomness and secrecy, which in some scenario are difficult to ensure, like in the distributed machine learning setting. More importantly, we only get amplification in the high privacy regime not allowing a privacy budget $\varepsilon > 1$. Practically, this is a problem as usually DP mechanism in the high privacy regime has poor utility due to the privacy-utility trade off.

So far we did not use the fact that we are only interested in releasing the last iteration of the DP-SGD, meaning that we can keep the intermediate iterations secrets. The composition theorems for Rényi differential privacy (but also advanced composition) does not require the secrecy of the intermediate steps, so in principle we could release all the intermediate parameters updates θ_t of the model without changing the privacy budget. This seems excessive, as we only release the last update θ_T. *Can we get better privacy by requiring the secrecy of the intermediate steps?*

The authors in [FMTT18] showed that by keeping the intermediate steps private and releasing just the last parameters of the iteration, we can get a new amplification theorem that gives results similar to the amplification by sampling techniques, without requiring Poisson sampling nor high privacy regime.

The intuition is the following, consider this noise iterative mechanism

$$X_{t+1} = \psi_{t+1}(X_t) + \mathcal{N}(0, \sigma^2 \mathbb{1}_d).$$

Let's consider the identity case where $\phi_t = \mathbb{1}$ for all steps $t \in \{0, \dots, T\}$ and fix the sensitivity $||X_0 - X_0'|| \leq 1$. Each mechanism is $(\alpha, \frac{\alpha}{2\sigma^2})$-RDP and a simple composition would lead to a T adaptive $(\alpha, \frac{\alpha}{2\sigma^2}T)$-RDP. However, let's consider the final X_T and X_T' random variables. As the variance of the sum of independent Gaussian random variable increases linearly we have

$$X_T = X_0 + \mathcal{N}(0, T\sigma^2 \mathbb{1}_d)$$
$$X_T' = X_0' + \mathcal{N}(0, T\sigma^2 \mathbb{1}_d).$$

If we release only X_T then we can bound the Rényi divergence of X_T and X_T'. As $||X_0 - X_0'|| \leq 1$ we have that

$$D_\alpha(X_T || X_T') \leq D_\alpha(\mathcal{N}(0, T\sigma^2\mathbb{1}_d) || \mathcal{N}(1, T\sigma^2\mathbb{1}_d)) = \frac{\alpha}{2T\sigma^2}.$$

Interestingly, by releasing only X_T we obtain a privacy budget that shrinks with the iteration instead of increasing linearly. The authors in [FMTT18] showed that this identity case is actually the worst case of a more general theorem valid for *contractive Noise Iteration*. In the following we will introduce the main definition and concept that are necessary to understand the main theorem of privacy amplification by iteration [Theorem 22, [FMTT18]].

Definition VI.1 (Contraction). For a Banach space $(\mathcal{Z}, ||\cdot||)$, a function $\psi : \mathcal{Z} \to \mathcal{Z}$ is said to be contractive if it is 1-Lipschitz. Namely, for all $x, y \in \mathcal{Z}$,

$$||\psi(x) - \psi(y)|| \leq ||x - y||.$$

Definition VI.2 (Contractive Noisy Iteration (CNI)). Given an initial random state $X_0 \in \mathcal{Z}$, a sequence of contractive functions $\psi_t : \mathcal{Z} \to \mathcal{Z}$, and a sequence of noise distributions $\{\xi_t\}$, we define the Contractive Noisy Iteration (CNI) by the following update rule:

$$X_{t+1} := \psi_{t+1}(X_t) + Z_{t+1}.$$

where Z_{t+1} is drawn independently from ξ_{t+1}. For brevity, we will denote the random variable output by this process after T steps as $\text{CNI}_T(X_0, \{\psi_t\}, \{\xi_t\})$.

To interpolate the metric distance on the inputs at $t = 0$ with an information theoretic divergence at $t = T$ the authors in [FMTT18] defined a new divergence called *Shifted Rényi divergence*. The theorem they stated is quite general and complicated as they use this new concept of shifted Rényi divergence and magnitude of noise for shifted distribution. Here we state only the theorem for contractive Gaussian noise iterations. A full version of the theorem can be found in Appendix C.

Theorem VI.1 (Privacy Amplification by Iteration for Gaussian Noise). Let X_T and X_T' denote the output of $\text{CNI}_T(X_0, \{\psi_t\}, \{\xi_t\})$ and $\text{CNI}_T(X_0, \{\psi_t'\}, \{\xi_t\})$, for $\xi_t \sim \mathcal{N}(0, \sigma^2\mathbb{1}_d)$. Let $s_t := \sup_x ||\psi_t(x) - \psi_t'(x)||$. Let $a_1, \dots, a_T$ be a sequence of reals and let $z_t := \sum_{i \leq t} s_i - \sum_{i \leq t} a_i$. If $z_t \geq 0$ for all t and $z_T = 0$, then

$$D_\alpha(X_T || X_T') \leq \frac{\alpha}{2\sigma^2} \sum_{i=1}^{T} a_t^2.$$

A. Application to DP-SGD on fixed batches

In order to have contractive maps we need to assume that the loss function is convex and smooth, indeed a well known theorem of convex optimization states that for convex and β-smooth loss, the stochastic gradient descent is a contraction for a learning rate $\eta < 2/\beta$. It is important to note, as the authors did, that the standard DP-SGD uses non convex losses as generally a clip mechanism does not preserve convexity. Therefore, the analysis of DP-SGD with privacy amplification by iteration works only if we remove the clip and use β-smooth and convex losses. Still, we need a bound on the gradient to tune the differential private noise, hence the sensitivity. According to this need, we assume also L-Lipschitzness for the losses, this will ensure an upper bound on the total gradient sensitivity (see Appendix A.2).

In [FMTT18] the authors provided an analysis for the stochastic gradient descent on convex, smooth and Lipshitz losses for the 1-mini batch realization, hence only batches of size one which are practical in the distributed setting. In this review we perform an analysis of the same mechanism but with fixed batches of size B, trying to match as much as possible the initial DP-SGD.

Let's fix two adjacent databases $D = \{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n\}$ and $D' = \{\mathbf{x}_1, \dots, \mathbf{x}'_i, \dots, \mathbf{x}_n\}$, differing at i-th position. As the privacy amplification by iteration does not require Poisson sampling, we group the databases in $Q = n/B$ batches of size B getting $D = \{B_1, \dots, B_\kappa, \dots, B_Q\}$ and $D' = \{B_1, \dots, B'_\kappa, \dots, B_Q\}$, where $B_j = \{\mathbf{x}_{(j-1)B+1}, \dots, \mathbf{x}_{jB}\}$. These two grouped databases differ in the batch at position κ. The algorithm to analyze is written in Algorithm 2

We need to study the contraction ψ

$$\psi_t(\theta) = \theta - \frac{\eta}{B} \sum_{x_\ell \in B_{(t \mod Q)}} \nabla \ell(\theta, x_\ell).$$

In particular, we need to compute the distance with the same contraction on the adjacent dataset at every iteration t. Following the theorem notation we have that $s_t = \sup_\theta |\psi_t(\theta) - \psi'_t(\theta)|$, hence

$$s_t = \sup_\theta \Big| \frac{\eta}{B} \sum_{x_\ell \in B_{(t \mod Q)}} \nabla \ell(\theta, x_\ell) + \\ - \frac{\eta}{B} \sum_{x_\ell \in B'_{(t \mod Q)}} \nabla \ell(\theta, x_\ell) \Big|.$$

Algorithm 2 DP-SGD: fixed batches

```
Require: Examples $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ divided in $Q$ non-intersecting batches $\{B_1, \dots, B_Q\}$ each of size $B$, $L$-Lipschitz loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \ell(\theta, x_i)$, learning rate $\eta$, noise scale $\sigma^2$.

Initialize $\theta_0$ randomly
for $j \in [T]$ do
    for $i \in [Q]$ do
        $t \leftarrow Qj + i$
        Compute gradient
        For each $x_\ell \in B_i$, compute $\mathbf{g}_t(x_\ell) \leftarrow \nabla_\theta \ell(\theta_t, x_\ell)$
        Add noise
        $\tilde{\mathbf{g}}_\mathbf{t} \leftarrow \frac{1}{B}\big(\sum_\ell \mathbf{g}_t(x_\ell) + \mathcal{N}(0, 4\sigma^2 L^2 \mathbf{I}_d)\big)$
        Descent
        $\theta_{t+1} \leftarrow \theta_t - \eta \tilde{\mathbf{g}}_t$
    end for
end for
return $\theta_T$, and compute the overall privacy cost
```

The batches differs of an examples only in the batch at position κ, hence

$$s_t = \begin{cases} 0 & \text{for } t = \kappa \mod Q \\ \frac{2\eta L}{B} & \text{for } t \neq \kappa \mod Q \end{cases}.$$

According to this bound we can define the following sequence

$$a_t = \begin{cases} 0 & \text{if } 0 \geq t < \kappa \\ \frac{2\eta L}{BQ} & \text{if } \kappa \leq t < Q(T-1) + \kappa \\ \frac{2\eta L}{B(T-\kappa+1)} & \text{if } Q(T-1) + \kappa \leq t \leq QT \end{cases}$$

As $Q > 1$ we have that $z_t = \sum_{i \leq t} s_t - \sum_{i \leq t} a_t \geq 0$. Moreover, we have that

$$z_T = \frac{2\eta L}{B} T - \frac{2\eta L}{BQ}(Q(T-1)) + \\ - \frac{2\eta L}{B(Q-\kappa+1)}(QT - Q(T-1) - \kappa + 1) = 0.$$

Therefore, we can use the particular case of the theorem without computing any shifted Rényi divergence. Let's write explicitly the contractive noise iteration

$$\theta_{t+1} = \psi_{t+1}(\theta_t) + \mathcal{N}(0, \frac{4\sigma^2\eta^2L^2}{B^2} \mathbb{1}_d),$$

it is important to consider the full variation of the inserted noise. Let's now apply the theorem

$$\begin{aligned} D_\alpha(\theta_{QT}, \theta'_{QT}) &\leq \frac{\alpha B^2}{8\sigma^2\eta^2L^2} \sum_{t=1}^{QT} a_t^2 \\ &\leq \frac{\alpha}{2\sigma^2}\left[\frac{1}{Q^2}\left(Q(T-1) + \frac{1}{Q-\kappa+1}\right)\right] \\ &\leq \frac{\alpha}{2\sigma^2 Q^2} K \end{aligned}$$

Where at the end we explicitly wrote the total number of iteration $K = QT$. For $q = 1/Q = n/B$ we get the same result of privacy amplification by sub sampling: $(\alpha, O(q^2\alpha/\sigma^2))$-RDP. Interestingly, now this result is valid for any privacy regime and for any batches size, as it is only required to not share intermediate steps.

Beside the privacy amplification by iteration theorem, this article laid the foundation to a series of new analysis under the *Hidden State Assumption*, going beyond standard composition techniques. In the last section we will explore a new technique which uses this hidden state assumption with a particular property of the loss function, the *strong-convexity* property, to remove the linear dependence T from the Rényi privacy budget.

Algorithm 3 Noisy Gradient Descent (DP-GD)

```
Require: Dataset of examples $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, loss function $\mathcal{L}(\theta; \mathbf{x})$, learning rate $\eta$, noise variance $\sigma^2$, initial parameter vector $\theta_0$, total gradient sensitivity $S_g$.

  for $k \in [T]$ do
      $\nabla\mathcal{L}(\theta_k; D) \leftarrow \sum_{i=1}^{n} \nabla\ell(\theta_k; \mathbf{x}_i)$
      $\theta_{k+1} \leftarrow \theta_k - \eta\nabla\mathcal{L}(\theta_k, D) + \sqrt{2\eta}\mathcal{N}(0, \sigma^2\mathbb{1}_d)$
  end for
  return $\theta_T$
```

VII. Langevin Dynamics and Privacy Convergence in the Hidden State Model

An important step forward in the analysis of DP-SGD was done by R. Chourasia and J. Ye in [CYS21]. They proposed to analyze DP-SGD as the dynamics of two *coupled stochastic process*. In particular, they constructed a pair of continuous-time Langevin diffusion [SN14] that fit the discrete noisy update of the gradient descent. Using the Fokker-Plank equation [RR96] they computed a bound on the Rényi divergence under the hidden state assumption. The high level intuition is to interpreted the gradient updates as a stochastic process with a stationary distribution, and so to relate privacy loss with the mixing time of the stochastic process[1]. A short introduction on Langevin dynamics and Fokker Plank equation can be found in Appendix D. The main result was a privacy convergence for arbitrary updates under the assumption of *strong convex* loss function. The authors concentrated their effort in studying the full batch differential private gradient descent described in Algorithm 3. Notice that the noise in the update rule is different. The usual way is

$$\theta_{k+1} = \theta_k - \frac{\eta}{n}(\nabla L(\theta_k, D) + \mathcal{N}(0, \sigma^2 S_g^2 \mathbb{1}_d))$$

So the results in [CYS21] have to be rescaled with this transformation

$$\sigma^2 \mapsto \frac{\eta S_g^2}{2n^2}\sigma^2, \tag{6}$$

in order to be compatible with the other results in this paper.

The following theorem is, indeed, a restatement of the main result in [CYS21].

Theorem VII.1 (Privacy Guarantees for noisy gradient descent [CYS21])**.** Let $\mathcal{L}(\theta, \mathbf{x})$ be a λ-strongly convex, and β-smooth loss function on closed convex set $\mathcal{C}$, with a finite total gradient sensitivity S_g, then the noisy gradient descent algorithm with start parameter $\theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{\eta S_g^2 \sigma^2}{\lambda n^2}\mathbb{1}_d))$, and step size $\eta < \frac{1}{\beta}$, and number of epochs T, satisfies (α, ε)-RDP with

$$\varepsilon = \frac{4}{\lambda\eta}\frac{\alpha}{2\sigma^2}(1 - e^{-\lambda\eta T/2})$$

This analysis works only for the full batch gradient descent update, indeed the privacy amplification by sub sampling is not present in Theorem VII.1, still, it is surprising that the privacy budget convergences for large value of iteration. However, is not clear how the coefficient $1/\lambda\eta$ behaves. We know that $\eta < 1/\beta$ and $\lambda < \beta$ (as β and λ are relatively the highest and the lower eigenvalues of the Hessian matrix of the convex function), so $\tau = 1/\lambda\eta > 1$. Regarding the noise variance, in order to get a (ϵ, δ)-DP mechanism we have to pose

$$\sigma \geq \Omega\left(\frac{\sqrt{\tau \log(1/\delta)}}{\varepsilon}\right) \qquad \text{for } \tau = \frac{1}{\lambda\eta}.$$

The stochastic gradient descent algorithm using the approach developed in [CYS21] was studied in [RBP22]. However it does not introduce the amplification by sampling technique and so it does not improve the privacy bound.

We now introduce the techniques of coupled Langevin diffusion

A. The Coupled Langevin Diffusion Techniques

Let's write the gradient descent update for two neighboring datasets. We write $\nabla\mathcal{L}(\theta, D) = \nabla\mathcal{L}_D(\theta)$

$$\begin{cases} \theta_{k+1} = \theta_k - \eta\nabla\mathcal{L}_D(\theta_k) + \sqrt{2\eta\sigma^2}Z_k \\ \theta'_{k+1} = \theta'_k - \eta\nabla\mathcal{L}_{D'}(\theta'_k) + \sqrt{2\eta\sigma^2}Z_k \end{cases} \quad \text{with } Z_k \sim \mathcal{N}(0, \mathbb{1}_d),$$

In the original paper [CYS21] the authors introduced also a projection step into a convex parameters set, we avoided to introduced it as it is not an essential ingredient. This discrete un-coupled stochastic process can be interpolate by a continuous coupled stochastic process on time $\eta k \leq t \leq \eta(k+1)$ by introducing an auxiliary random variable Θ_t defined $\Theta_{\eta\kappa} = \theta_\kappa$ at $t = \eta k$ and for $\eta k < t \leq \eta(k+1)$:

$$\begin{cases} \Theta_t = \Theta_{\eta k} - \eta U_1(\Theta_{\eta k}) - (t - \eta k)U_2(\Theta_{\eta k}) + \sqrt{2(t-\eta k)\sigma^2}Z_k \\ \Theta'_t = \Theta'_{\eta k} - \eta U_1(\Theta'_{\eta k}) + (t - \eta k)U_2(\Theta'_{\eta k}) + \sqrt{2(t-\eta k)\sigma^2}Z_k \end{cases},$$

where

$$U_1(\theta) = \frac{1}{2}(\nabla\mathcal{L}_D(\theta) + \nabla\mathcal{L}_{D'}(\theta)),$$

$$U_2(\theta) = \frac{1}{2}(\nabla\mathcal{L}_D(\theta) - \nabla\mathcal{L}_{D'}(\theta)),$$

So at $t = \eta(k+1)$ we have $\Theta_t = \theta_{k+1}$. In the following we will make use of Wiener processes and Fokker Plank equation, see Appendix D for a short introduction. With this transformation we can write the stochastic differential equation for the updates

$$\begin{cases} d\Theta_t = -U_2(\Theta_k) + \sqrt{2\sigma^2}dW_t \\ d\Theta'_t = -U_2(\Theta'_k) + \sqrt{2\sigma^2}dW_t \end{cases},$$

where W_t is the d-dimensional Wiener process. This coupled stochastic process models the one step discrete process and it is called *tracing process*. Therefore, we can think of the full discrete process of T updates as T tracing processes which are described by continuous time updates.

With this formulation we can now model the tracing process using the Fokker Plank formulation, hence for $\eta k < t \leq \eta(k+1)$ the evolution of the conditional probability density function $p_{t|\eta k}(\theta|\theta_k) = p(\Theta_t = \theta|\Theta_{\eta k} = \theta_k)$ is

$$\begin{cases} \dfrac{\partial p_{t|\eta k}(\theta|\theta_k)}{\partial t} = \nabla \cdot (p_{t|\eta k}(\theta|\theta_k)U_2(\theta_k)) + \sigma^2\nabla^2 p_{t|\eta k}(\theta|\theta'_k) \\ \dfrac{\partial p'_{t|\eta k}(\theta|\theta_k)}{\partial t} = -\nabla \cdot (p'_{t|\eta k}(\theta|\theta'_k)U_2(\theta_k)) + \sigma^2\nabla^2 p'_{t|\eta k}(\theta|\theta'_k) \end{cases}. \tag{7}$$

[1] The mixing time is the time necessary for a stochastic process to reach its stationary distribution

The condition probability is necessary as the tracing process is not defined at $t = \eta k$, so it is necessary to condition the probability for $\Theta_{\eta k} = \theta_k$, to guarantee that the continuous process fits the underlying discrete process.

By taking the expectation over probability density function $p_{\eta k}(\theta_k)$ and $p'_{\eta k}(\theta'_k)$ on both sides of equation 7 we obtain the partial differential equation that models the evolution of (unconditioned) probability density function in the coupled tracing diffusions.

$$\begin{cases} \dfrac{\partial p_t(\theta)}{\partial t} = \nabla \cdot (p_t(\theta)V_t(\theta)) + \sigma^2 \nabla^2 p_t(\theta) \\ \dfrac{\partial p'_t(\theta)}{\partial t} = \nabla \cdot (p'_t(\theta)V'_t(\theta)) + \sigma^2 \nabla^2 p'_t(\theta) \end{cases} \tag{8}$$

where $V_t(\theta) = -V'_t = \mathbb{E}_{\theta_k \sim p_{\eta k|t}}[U_2(\theta_k)|\theta]$. The main advantage of this formulation is that we can now make full use of an interesting results in [CYS21] that bounds the rate of the Rényi privacy loss

Lemma VII.1 (Rate of Rényi privacy loss [CYS21])**.** Given coupled diffusion in equation 8 and $S_v = \max_\theta ||V_t(\theta) - V'_t(\theta)||\ \forall t \geq 0$, the Rènyi privacy loss rate at any $t \geq 0$ is upper bounded by

$$\frac{\partial D_\alpha(\Theta_t||\Theta'_t)}{\partial t} \leq \frac{1}{\gamma}\frac{\alpha S_v^2}{4\sigma^2} - (1-\gamma)\sigma^2\alpha \frac{I_\alpha(\Theta_t||\Theta'_t)}{E_\alpha(\Theta_t||\Theta'_t)} \tag{9}$$

where $\gamma > 0$ is a tuning parameter, I_α is the Rènyi information and E_α is the moment generating function of the ratio Θ_t/Θ'_t computed at α.

a) Observation: Notice that for $\gamma = 1$ we obtain after integration

$$D_\alpha(\Theta_T||\Theta'_T) \leq \frac{\alpha S_v^2}{4\sigma^2}T.$$

By setting $T = \eta K$, $\Theta_T = \theta_K$, changing sigma as in (6) we get

$$D_\alpha(\Theta_T||\Theta'_T) \leq \frac{\alpha S_v^2}{2\sigma^2}\frac{n^2}{S_g^2}$$

Then by using the result $S_v^2 \leq \frac{S_g^2}{n^2}$ (Lemma 4 in [CYS21]) we get the standard bound for the Gaussian Mechanism $D_\alpha(\Theta_T||\Theta'_T) \leq \frac{\alpha}{2\sigma^2}$.
Interestingly, with this formulation we have a negative additive term which is $-\frac{1}{2}\sigma^2\alpha\frac{I_\alpha}{E_\alpha} < 0$. This term, if properly handled, will make the privacy converge.

Handle the ratio between I_α/E_α is complicated and requires knowing the pdf of θ at any time. The authors used a recent result found in [VW19] to lower bound the ratio I_α/E_α using c-Log Sobolev Inequality [Gro75]. The new bound of equation 9 can be rewritten as

$$\frac{\partial D(\alpha,t)}{\partial t} \leq \frac{1}{\gamma}\frac{\alpha S_v^2}{4\sigma^2} - 2(1-\gamma)\sigma^2 c\left[\frac{D(\alpha,t)}{\alpha} + (\alpha-1)\frac{\partial D(\alpha,t)}{\partial \alpha}\right],$$

where $D(\alpha,t) = D_\alpha(\Theta_t||\Theta'_t)$. The authors solved this PDE by computing an upper bound for the solution for each K tracing diffusion process. The overall solution can be found by iteratively compose the upper bound of each tracing diffusion, obtaining

$$D_\alpha(\Theta_{\eta K}||\Theta'_{\eta K}) \leq \frac{\alpha S_g^2}{2c\sigma^4 n^2}(1 - e^{-\sigma^2 c\eta K})$$

Without entering in the details, the authors in [CYS21] demonstrated that the coupled continuous stochastic diffusion in equation 8 satisfies the c-Log Sobolev Inequality if the loss is λ-strong convex, getting the main result (after an opportune transformation in the variance) stated in Theorem VII.1.

For the first time, by using some properties of the loss function, it was demonstrated that some differential private algorithm can add noise indefinitely without increasing the privacy budget, under the *Hidden State Assumption*.

A subsequent work introduced also the amplification by sub sampling in this analysis [YS22].

B. Langevin Dynamics with Privacy Amplification by Sub-Sampling

The fact that privacy budget converges to a constant for a large number of iteration is a breakthrough result, as deep learning model usually needs a large number of epochs to be trained. However, it converges to a huge constant $O(\alpha/\sigma^2)$. Privacy amplification by sub-sampling gives a factor $O(1/n^2)$ to the overall privacy budget, but standard analysis did not show converges yet, getting $O(\alpha T/(\sigma^2 n^2))$. So for $T \ll n^2$, privacy amplification by sub-sampling still is the best choice yet.

The subsequent work [YS22] studied the coupled Langevin diffusion with privacy amplification by subsampling. In particular, the authors demonstrated a privacy amplification theorem for random batch construct with sampling with replacement of examples. This is a better approach than Poisson sampling for two reason: the random batches created have a fixed size b, so computationally the sampling runs in $O(b)$ times instead of $O(n)$. Their result is quite complicated as it offers a recursive technique of how computing the privacy budget. We expose it here for completeness, following the notation of DP-SGD in this article (hence applying the transformation in equation 6 with $n \mapsto b$ as the algorithm samples a batch of fixed size b).

Theorem VII.2 (Recursive amplification by sampling without replacement)**.** If the loss function $\ell(\theta; x)$ is λ-strongly convex, β-smooth, and if its gradient has finite ℓ_2-sensitivity S_g. Then DP-SGD under sampling without replacement and stepsize $\eta < \frac{2}{\lambda+\beta}$ satisfies (α, ε)-RDP with

$$\varepsilon \leq \frac{1}{\alpha - 1}\log(S_K^0(\alpha))$$

where the terms $S_k^j(\alpha)$ for $k = 0, \ldots, K-1$ and $j = 0, \ldots, n/b - 1$ are recursively computed by

$$S_0^0(\alpha) = 1$$

$$S_k^{j+1}(\alpha) = \frac{b}{n}e^{\frac{(\alpha-1)\alpha}{2\sigma^2}} \cdot S_k^j(\alpha) + \left(1 - \frac{b}{n}\right) \cdot S_k^j(\alpha)^{(1-\eta\lambda)^2}$$

$$S_{k+1}^0(\alpha) = S_k^{n/b}(\alpha)$$

VIII. Conclusions

In this article we presented the differential private stochastic gradient descent algorithm, the state of the art privacy technology to train machine learning and deep learning models. We presented three different techniques developed in this context to analyze better the privacy properties of the algorithm. The moments accountant was the first new privacy analysis technique developed for the DP-SGD algorithm. It allows to reduce the privacy budget than would have been computed with standard techniques, by leveraging on the Gaussian properties of the noise distribution. However, it's applicability is bounded to the high privacy regime and Poisson sampling.

The second techniques exposed is the privacy amplification by iteration, which does not require neither high privacy regime nor Poisson sampling, but only the hidden state assumption, hence to not release intermediate updates. Nonetheless, it requires additional assumption on the loss, such as smoothness and lipschitzness.

Both the previous techniques give a linear increase of the privacy budget with number of epochs, which is undesirable especially for deep learning application. The Langevin dynamics techniques solved this problem by using ad additional property of the loss, the strong convexity. All these techniques were exposed using the same notations highlighting the deficiencies and the improvements.

Further techniques were developed that are not described in this article such as the privacy amplification by iteration with privacy amplification developed in [AT22], or the use of Hockey-stick divergence instead of Rényi divergence to study DP-SGD as done in [AD23]. This field is indeed an active research area as on important question needs to be answered in order to an adoption for general deep learning algorithm: *Does the privacy budget converges at any number of epochs with no assumption on the loss function?*

References

[ACG⁺16] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[AD23] Shahab Asoodeh and Mario Diaz. Privacy loss of noisy stochastic gradient descent might converge even for non-convex losses. *arXiv preprint arXiv:2305.09903*, 2023.

[AT22] Jason Altschuler and Kunal Talwar. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. *Advances in Neural Information Processing Systems*, 35:3788–3800, 2022.

[BBG18] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in neural information processing systems*, 31, 2018.

[BBKN14] Amos Beimel, Hai Brenner, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. *Machine learning*, 94:401–437, 2014.

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[BW18] Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 394–403. PMLR, 2018.

[CYS21] Rishav Chourasia, Jiayuan Ye, and Reza Shokri. Differential privacy dynamics of langevin diffusion and noisy gradient descent. *Advances in Neural Information Processing Systems*, 34:14771–14781, 2021.

[DR⁺14] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[DR16] Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[FJR15] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[FMTT18] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 521–532. IEEE, 2018.

[Gro75] Leonard Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

[HVY⁺22] Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing training data from trained neural networks. *Advances in Neural Information Processing Systems*, 35:22911–22924, 2022.

[Mir17] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[MTZ19] Ilya Mironov, Kunal Talwar, and Li Zhang. R\'enyi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

[RBP22] Théo Ryffel, Francis Bach, and David Pointcheval. Differential privacy guarantees for stochastic gradient langevin dynamics. *arXiv preprint arXiv:2201.11980*, 2022.

[RR96] Hannes Risken and Hannes Risken. *Fokker-planck equation*. Springer, 1996.

[SN14] Issei Sato and Hiroshi Nakagawa. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In *International Conference on Machine Learning*, pages 982–990. PMLR, 2014.

[VW19] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

[WBK19] Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1226–1235. PMLR, 2019.

[YS22] Jiayuan Ye and Reza Shokri. Differentially private learning needs hidden state (or much faster convergence). *Advances in Neural Information Processing Systems*, 35:703–715, 2022.

[YSS⁺21] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User-friendly differential privacy library in pytorch. *arXiv preprint arXiv:2109.12298*, 2021.

[ZDW22] Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal accounting of differential privacy via characteristic function. In *International Conference on Artificial Intelligence and Statistics*, pages 4782–4817. PMLR, 2022.

Appendix

A. Loss Function Properties

For any data record $\mathbf{x} \in \mathcal{X}$, a loss function is $\mathcal{L}(\theta, \mathbf{x}) : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$, where $\mathcal{C} \subseteq \mathbb{R}^d$ is usually a closed convex set. The gradient of the loss with respect to the parameter θ is defined as $\nabla\mathcal{L}(\theta; \mathbf{x})$

Definition A.1 (Lipschitz continuity)**.** A function $\mathcal{L}(\theta, \mathbf{x})$ is L-Lipschitz continuous if for all $\theta, \theta' \in \mathcal{C}$ and $\mathbf{x} \in \mathcal{X}$,

$$|\mathcal{L}(\theta, \mathbf{x}) - \mathcal{L}(\theta', \mathbf{x})| \leq L||\theta - \theta'||_2 \tag{10}$$

The Lipschitzness condition bounds the sensitivity of the gradient loss

Proposition A.1 (Lipschitz sensitivity)**.** The gradient sensitivity $S^{(2)}_{\nabla\mathcal{L}}$ of an L-Lipschitz loss is $S^{(2)}_{\nabla\mathcal{L}} \leq 2L$.

Proof: We need to bound the l_2 gradient sensitivity for all parameters $\theta \in \mathcal{C}$

$$S^{(2)}_{\nabla\mathcal{L}} = \sup_{\mathbf{x}\sim\mathbf{x}'} ||\nabla_\theta\mathcal{L}(\theta, \mathbf{x}) - \nabla_\theta\mathcal{L}(\theta, \mathbf{x}')||_2, \tag{11}$$

where $\mathbf{x} \sim \mathbf{x}'$ indicates two neighbor datasets. From the L-Lipschitzness condition we can bound the l_2 gradient norm

$$||\nabla_\theta\mathcal{L}(\theta, \mathbf{x})||_2 = \frac{|\mathcal{L}(\theta + d\theta, \mathbf{x}) - \mathcal{L}(\theta, \mathbf{x})|}{||d\theta||_2} \leq L.$$

Using the triangle inequality on 11 we get

$$S^{(2)}_{\nabla\mathcal{L}} \leq \sup_{\mathbf{x}\sim\mathbf{x}'} ||\nabla_\theta\mathcal{L}(\theta, \mathbf{x})||_2 + ||\nabla_\theta\mathcal{L}(\theta, \mathbf{x}')||_2 \leq 2L$$

■

Proposition A.2 (Lipschitz sensitivity of the total gradient loss)**.** For L-Lipschitz loss function, the total gradient loss $L_B(\theta) = \sum_{x\in B} \nabla_\theta\mathcal{L}(\theta, x)$ over a batch of any size B has an ℓ_2 sensitivity

$$S^{(2)}_{L_B(\theta)} \leq 2L \tag{12}$$

Proof:

$$\begin{aligned} S^{(2)}_{L_B(\theta)} &= ||\sum_{x\in B} \nabla_\theta\mathcal{L}(\theta, x) - \sum_{x\in B'} \nabla_\theta\mathcal{L}(\theta, x)||_2 \\ &= ||\nabla_\theta\mathcal{L}(\theta, x_\kappa) - \nabla_\theta\mathcal{L}(\theta, x'_\kappa)||_2 \leq 2L, \end{aligned}$$

where in the first equality we used the fact that the two batches comes from adjacent dataset, so they can differ by at most one element, which is in generally at κ position. The last inequality comes from Proposition A.1. ■

Definition A.2 (Smoothness)**.** Differentiable function $\mathcal{L}(\theta, \mathbf{x})$ is β-smooth over $\mathcal{C}$ if for all $\theta, \theta' \in \mathcal{C}$ and $\mathbf{x} \in \mathcal{X}$

$$||\nabla\mathcal{L}(\theta, \mathbf{x}) - \mathcal{L}(\theta', \mathbf{x})|| \leq \beta||\theta - \theta'||_2 \tag{13}$$

Definition A.3 (Strong Convexity)**.** Differentiable function $\mathcal{L}(\theta, \mathbf{x})$ is λ-strongly convex if for all $\theta, \theta' \in \mathcal{C}$ and $\mathbf{x} \in \mathcal{X}$

$$\mathcal{L}(\theta', \mathbf{x}) \geq \mathcal{L}(\theta, \mathbf{x}) + \nabla\mathcal{L}(\theta, \mathbf{x})^T(\theta - \theta') + \frac{\lambda}{2}||\theta' - \theta||_2 \tag{14}$$

B. Rényi Differential privacy

1) Privacy Analysis of DP-SGD: Consider the application of Possion sampling with probability $q = O(1/n)$, where n is the number of the examples, to construct a random batch B. In the high privacy regime, privacy amplification by subsampling for the Gaussian mechanism leads to a $(\alpha, c\frac{q^2\alpha}{\sigma^2})$-RDP mechanism, for some constant $c > 0$. The adaptive composition of T Gaussian mechanisms satisfies $(\alpha, \frac{cTq^2\alpha}{\sigma^2})$-RDP. Using proposition V.2, for any $0 < \delta < 1$ we get an (ε, δ)-DP mechanism with

$$\varepsilon = \frac{cTq^2\alpha}{\sigma^2} + \frac{\log(1/\delta)}{\alpha - 1}. \tag{15}$$

We search the best divergence order α^* by differentiating the privacy parameter

$$\frac{\partial}{\partial\alpha}\left[\frac{cTq^2\alpha}{\sigma^2} + \frac{\log(1/\delta)}{\alpha - 1}\right] = \frac{cTq^2}{\sigma^2} - \frac{\log(1/\delta)}{(\alpha - 1)^2}.$$

By posing the derivative equal to zero we obtain

$$\alpha^* = 1 + \sqrt{\frac{\sigma^2\log(1/\delta)}{cTq^2}}.$$

We insert α^* in equation 15

$$\varepsilon = \frac{cTq^2}{\sigma^2} + 2\sqrt{\frac{cTq^2\log(1/\delta)}{\sigma^2}}. \tag{16}$$

We set a new variable $x = \sqrt{cTq^2/\sigma^2}$ and solve Equation 16 for x

$$x^2 + 2\sqrt{\log(1/\delta)}x - \varepsilon = 0.$$

The solution of the equation above is

$$\begin{aligned} x &= \sqrt{\log(1/\delta) + \varepsilon} - \sqrt{\log(1/\delta)} \\ &= \frac{\varepsilon}{\sqrt{\log(1/\delta) + \varepsilon} + \sqrt{\log(1/\delta)}} \leq \frac{\varepsilon}{2\sqrt{\log(1/\delta)}}. \end{aligned}$$

By inserting $x = q\sqrt{cT}/\sigma$ we get

$$\sigma \geq \Omega\left(\frac{q\sqrt{T\log(1/\delta)}}{\varepsilon}\right)$$

2) RDP of Sample Gaussian Mechanism: We present here a non asymptotic results for the Sample Gaussian Mechanism, which is the Gaussian mechanism applied to a random batch constructed using Poisson sampling.

Definition A.4 (Sample Gaussian Mechanism (SGM))**.** Let f be a function mapping subsets of $\mathcal{D}$ to $\mathbb{R}^d$. We define the Sample Gaussian mechanism (SGM) parameterized with the sampling rate $0 < q \leq 1$ and the noise $\sigma > 0$ as

$$\begin{aligned} \mathrm{SG}_{q,\sigma}(\mathcal{D}) :=& f(\{x : x \in \mathcal{D} \text{is sampled with probability } q\}) \\ &+ \mathcal{N}(0, \sigma^2\mathbb{1}_d), \end{aligned}$$

where each element of $\mathcal{D}$ is sampled independently at random with probability q without replacement, and $\mathcal{N}(0, \sigma^2\mathbb{1}_d)$ is spherical d-dimensional Gaussian noise with per-coordinate variance σ^2.

[2] The authors found that the smoothness assumption can be effectively removed by convolving the loss with a Gaussian distribution

Article	σ for (ε, δ)-DP	Techniques	Loss assumption	Deficiencies
[ACG+16]	$\Omega\left(\frac{\sqrt{T \ln(1/\delta)}}{n\varepsilon}\right)$	Moments Accountant Privacy amplification by sub-sampling	No assumption on the loss Sensitivity using clipping	Poisson sampling for the batches High privacy regime
[Mir17]	$\Omega\left(\frac{\sqrt{T \ln(1/\delta)}}{n\varepsilon}\right)$	Standard Composition of RDP Privacy amplification by sub-sampling	No assumption on the loss Sensitivity using clipping	Poisson sampling for the batches High privacy regime
[FMTT18]	$\Omega\left(\frac{\sqrt{T \ln(1/\delta)}}{n\varepsilon}\right)$	Privacy amplification by iteration	convex L-Lipshitz β-smooth	The update step has to be a contraction
[CYS21]	$\Omega\left(\frac{\sqrt{\ln(1/\delta)}}{\varepsilon}\right)$	Coupled Langevin diffusion	β-smooth L-Lipschitz λ-strong convex	Strong convexity assumption

Table I
RECAP OF THE DIFFERENT TECHNIQUES USED TO ANALYZE DP-SGD.

Symbol	Meaning
$\mathcal{D}$	Database of example
$\mathcal{D}'$	Database of example adjacent to $\mathcal{D}$
$\mathbf{x}$ or x	Particular example in $\mathcal{D}$
B	Batch of examples of $\mathcal{D}$
b	Size of the batch B
S_f	ℓ_2 sensitivity of the function f
θ	Parameters of a model
d	Dimension of the model parameters
$\ell(\theta, x)$	Loss function with parameter θ evaluated on example x
$L(\theta, B) = \sum_{x \in B} \ell(\theta, x)$	Total loss function with parameter θ evaluated on batch B
$\mathcal{L}(\theta, x) = \frac{1}{B} L(\theta, B)$	Average total loss function
η	Learning rate for the gradient descent
L	Lipshitz constant for the loss
β	Smoothness constant for the loss
λ	Strong convexity constant for the loss
C	Clipping constant for the loss
T	Number of iteration, hence number of noise insertion
ε	Rènyi or standard privacy budget
δ	Probability of uncontrolled breach in standard DP
$\xi(\mathcal{M}, x, x')$	Privacy loss random variable of the mechanism $\mathcal{M}$ computed on dataset x and x'
α	Rényi differential privacy order
$D_\alpha(\mu \Vert \nu)$	Rényi divergence between two distribution μ and ν
$E_\alpha(\mu \Vert \nu)$	α-th moment of likelihood ration between distribution μ and ν
$I_\alpha(\mu \Vert \nu)$	Rényi information of distribution ν and ν

Table II
TABLE OF NOTATIONS

Theorem A.1 ([MTZ19])**.** If $q \leq \frac{1}{5}$, $\sigma \geq 4$, and α satisfy

$$1 < \alpha \leq \frac{1}{2}\sigma^2 L - 2\log\sigma,$$
$$\alpha \leq \frac{\frac{1}{2}\sigma^2 L^2 - \log 5 - 2\log\sigma}{L + \log(q\alpha) + \frac{1}{2\sigma^2}},$$

where $L = \log\left(1 + \frac{1}{q(\alpha-1)}\right)$, then SGM applied to a function of ℓ_2-sensitivity 1 satisfies (α, ε)-RDP where

$$\varepsilon = \frac{2q^2\alpha}{\sigma^2}$$

C. Privacy Amplification by Iteration

Definition A.5 (Shifted Rényi Divergence)**.** Let μ and ν be distributions defined on a Banach space $(\mathcal{Z}, ||\cdot||)$. For parameters $z \geq 0$ and $\alpha \geq 1$, the z-shifted Rényi divergence between μ and ν is defined as

$$D_\alpha^{(z)}(\mu||\nu) := \inf_{\mu' : W_\infty(\mu, \mu') \leq z} D_\alpha(\mu'||\nu).$$

Where $W_\infty(\mu, \nu)$ is the ∞-Wasserstein distance between the two distributions

$$W_\infty(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu, \nu)} \operatorname{ess} \sup_{(x,y) \in \gamma} ||x - y||.$$

It is the essential supremum distance of the infimum coupling $\Gamma(\mu, \nu)$ over the Banach space.

This divergence is called *shifted* as it is upper-bounded by the Rényi divergence of a shifted distribution

$$D_\alpha^{(||\mathbf{x}||)}(\mu||\nu) \leq D_\alpha(\mu * \mathbf{x}||\nu)$$

Note that $\mu * \mathbf{x}$ is the distribution of $U +$ where $U \sim \mu$.

The final ingredient is the magnitude of noise for shifted distributions

Definition A.6 (Magnitude of Noise for Shifted Distribution)**.** For a noise distribution ξ over a Banach space $(\mathcal{Z}, ||\cdot||)$ we measure the magnitude of noise by considering the function that for $a > 0$, measures the largest Rényi divergence of order α between ξ and the same distribution ξ shifted by a vector of length at most a

$$R_\alpha(\xi, a) := \sup_{\mathbf{x}:||\mathbf{x}||\leq a} D_\alpha(\xi * \mathbf{x}||\xi).$$

In particular, for Gaussian noise we have that

$$R_\alpha(\mathcal{N}(0, \sigma^2 \mathbb{1}_d), a) = \frac{\alpha a^2}{2\sigma^2}$$

We are now ready to state the complete privacy amplification by iteration theorem.

Theorem A.2 (Privacy Amplification by Iteration [FMTT18]**).** Let X_T and X'_T denote the output of $\text{CNI}_T(X_0, \{\psi_t\}, \{\xi_t\})$ and $\text{CNI}_T(X_0, \{\psi'_t\}, \{\xi_t\})$. Let $s_t := \sup_x ||\psi_t(x) - \psi'_t(x)||$. Let $a_1, \dots, a_T$ be a sequence of reals and let $z_t := \sum_{i\leq t} s_i - \sum_{i\leq t} a_i$. If $z \geq 0$ for all t, then

$$D_\alpha^{(z_T)}(X_T||X'_T) \leq \sum_{i=1}^{T} R_\alpha(\xi_t, a_t).$$

In particular, if $z_T = 0$, we have

$$D_\alpha(X_T||X'_T) \leq \sum_{i=1}^{T} R_\alpha(\xi_t, a_t).$$

D. Langevin Diffusion and Fokker Plank Equation

A Langevin diffusion process in $\mathbb{R}^d$ with noise variance σ^2 is described by the following stochastic differential equation (SDE)

$$d\boldsymbol{\theta}_t = -\mathbf{f}(\boldsymbol{\theta}_t, t)dt + \sqrt{2\sigma^2} d\mathbf{W}_t$$

Where $\mathbf{f}(\boldsymbol{\theta}_t, t)$ is the *drift* function, and $\mathbf{W}_t$ is the standard d-dimensional Wiener process. A Wiener process can be seen as the integral of a white noise Gaussian process, so the random variable $d\mathbf{W}_t \sim \sqrt{dt}\mathcal{N}(0, \mathbb{1}_d)$. The Fokker Plank equation characterizes the evolution of the probability density of the random variable θ_t. Indeed, if $p(\theta, t)$ indicates the pdf of θ at time t, and θ_t is Langevin diffused, then the Fokker Plank equation states that

$$\frac{\partial p(\boldsymbol{\theta}, t)}{\partial t} = -\nabla_{\boldsymbol{\theta}} \Big[f(\boldsymbol{\theta}, t) p(\theta, t) \Big] + \sigma^2 \nabla^2 p(\boldsymbol{\theta}, t).$$

It is worth mentioning that the above equation is valid for Ito's stochastic calculus. The stationary distribution of the process (for $\partial p(\boldsymbol{\theta}, t)/\partial t = 0$) is the Gibbs distribution

$$\pi(\boldsymbol{\theta}) \propto e^{\frac{\mathbf{f}(\boldsymbol{\theta}, t)}{\sigma^2}}.$$

E. Sampling techniques

We will consider two different sampling mechanism to create the batch. Given a dataset of examples $\mathcal{D} = \{x_1, \dots, x_n\}$ we construct a batch B using:

- **Poisson Sampling:** each element in $\mathcal{D}$ is sampled with probability q

  $$B = \{x : x \in \mathcal{D} \text{ sampled with probability } q\}$$

  This sampling technique creates batches of different size with expected size $\mathbb{E}[b] = qN$.

- **Sample without replacement:** sample b different examples from $\mathcal{D}$

  $$B = \{x : x \in \mathcal{D} \text{ sampled without replacement}\}.$$

This technique produces batches of fixed size b and it is a practical alternative to Poisson sampling.