Differential Privacy Guarantees for Stochastic Gradient Langevin Dynamics

Théo Ryffel, Francis Bach, David Pointcheval

Differential Privacy Guarantees for Stochastic Gradient Langevin Dynamics

Théo Ryffel [1] **Francis Bach** [1] **David Pointcheval** [2]

Abstract

We analyse the privacy leakage of noisy stochastic gradient descent by modeling Rényi divergence dynamics with Langevin diffusions. Inspired by recent work on non-stochastic algorithms, we derive similar desirable properties in the stochastic setting. In particular, we prove that the privacy loss converges exponentially fast for smooth and strongly convex objectives under constant step size, which is a significant improvement over previous DP-SGD analyses. We also extend our analysis to arbitrary sequences of varying step sizes and derive new utility bounds. Last, we propose an implementation and our experiments show the practical utility of our approach compared to classical DP-SGD libraries.

1. Introduction

Differential privacy (Dwork et al., 2014) for machine learning is a promising approach to reduce exposure of training datasets when releasing machine learning models. The privacy leakage from these models can be quantified using Rényi differential privacy (Mironov, 2017) which models it through the divergence of the distributions of two models trained on datasets that only differ in one item. The intuition behind is that a model whose behavior is sensitive to the presence or absence of a single individual is likely to memorize information about specific individuals, which can then be uncovered using several types of attacks like membership inference attacks (Shokri et al., 2017).

The most standard approaches to training neural networks with differential privacy are derived from Abadi et al. (2016)'s method of differentially private stochastic gradient descent (DP-SGD). DP-SGD is an attractive method as it closely mimics classic SGD training of neural networks, and applies to almost all architectures. It therefore enjoys easy adoption from data scientists and has been integrated in popular libraries like Opacus (Yousefpour et al., 2021). Differential privacy of the whole training mechanism is computed using the strong composition theorem (Dwork et al., 2010), which states that the privacy leakage modeled through standard (ϵ, δ)-differential privacy grows approximately in $\sqrt{K}$ for high privacy regimes, where K is the number of iterations. This is a strong limitation of DP-SGD in real world applications since training for a large number of iterations would lead to a prohibitive privacy bound.

Recently, Chourasia et al. (2021) have proposed a novel analysis of the differential privacy dynamics of Langevin diffusion, resulting in a new differentially private noisy gradient descent algorithm (DP-GLD). This method notably guarantees that under strongly convex and smooth objectives, the privacy leakage can be bounded by a constant, which allows for an unlimited number of model updates. The key difference with DP-SGD is that the model is assumed to be hidden during training and released only once the training is over. This is a setting that we typically encounter while training a model using multi-party computation (Knott et al., 2021; Wagh et al., 2021; Ryffel et al., 2022) where only the final version of the model is visible. However, the algorithm and the privacy analysis provided by Chourasia et al. (2021) only addresses full gradient descent (DP-GLD) which is impractical for large datasets and makes its adoption by data scientists or standard differential privacy libraries less probable.

Contributions. We provide a *stochastic* version of the noisy gradient descent algorithm (DP-SGLD) and build a privacy analysis based on Langevin diffusion. We prove that DP-SGLD achieves similar privacy and utility guarantees than DP-GLD, including exponential convergence of the privacy bound, and we extend the analysis to the case where the step size is not constant. More specifically:

- We introduce DP-SGLD, a stochastic version of the DP-GLD algorithm studied by Chourasia et al. (2021), and we show that it achieves the same privacy guarantees including exponentially fast convergence.
- We show that DP-SGLD achieves similar utility than

[1]INRIA, Département d'informatique de l'ENS, ENS, CNRS, PSL University, Paris, France [2]Département d'informatique de l'ENS, ENS, CNRS, PSL University, INRIA, Paris, France. Correspondence to: Théo Ryffel <theo.ryffel@ens.fr>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

DP-GLD up to a term due to using stochastic estimates of the gradient. We also relax assumptions on the step size η in utility theorems of Chourasia et al. (2021) to only verify $\eta \leq \frac{1}{2\beta}$ instead of $\eta \leq \frac{\lambda}{2\beta^2}$, where β is the smoothness constant and λ the strong-convexity parameter, thus obtaining the classical scaling from convex optimization.

- We extend our analysis of DP-SGLD to non-constant step sizes and derive utility bounds when the step size is parametrized as $\eta_k = \frac{1}{2\beta+\lambda k/2}$ and removes the term due to stochastic training.
- Last, we provide an implementation of DP-SGLD[1] and an experimental evaluation where we train differentially private logistic regressions on several datasets. We show that DP-SGLD achieves higher experimental accuracy than DP-SGD on these tasks, and that it almost closes the gap with non-private training. We also show that standard DP-SGD does not benefit from training deeper networks on these tasks, which aligns with the conclusions drawn by Tramèr & Boneh (2021).

Note that our algorithm ends up being similar to the one of Welling & Teh (2011), except that we do not try to construct samples from the posterior distribution but instead to derive privacy guarantees.

2. Preliminaries

Let us first recall the standard definition of (ϵ, δ)-differential privacy, as introduced by Dwork et al. (2014) :

Definition 2.1 ((ϵ, δ)-differential privacy)**.** A randomized algorithm $\mathcal{A} : \mathcal{D} \mapsto \mathbb{R}^d$ satisfies (ϵ, δ)-differential privacy if for any neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, i.e., datasets that only differ in one item, and any subset $S \in \mathbb{R}^d$, the distribution of $\mathcal{A}$ satisfies:

$$\mathrm{P}[\mathcal{A}(\mathcal{D})] \leq e^{\epsilon}\mathrm{P}[\mathcal{A}(\mathcal{D}')] + \delta.$$

An alternative notion, coined as *Rényi differential privacy* has been proposed by Mironov (2017), which is more suited to studying composition mechanisms, but can be converted back to standard (ϵ, δ)-differential privacy.

Definition 2.2 (Rényi differential privacy)**.** A randomized algorithm $\mathcal{A} : \mathcal{D} \mapsto \mathbb{R}^d$ satisfies (α, ε)-Rényi differential privacy if for any neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, the α Rényi divergence satisfies $R_\alpha(\mathcal{A}(\mathcal{D}) \,||\, \mathcal{A}(\mathcal{D}')) \leq \varepsilon$, where:

$$R_\alpha(\mathcal{A}(\mathcal{D}) \,||\, \mathcal{A}(\mathcal{D}')) = \frac{1}{\alpha-1}\log \mathop{\mathbb{E}}_{\theta\sim\mathcal{A}(\mathcal{D}')}\left[\left(\frac{\mu_{\mathcal{A}(\mathcal{D})}(\theta)}{\mu_{\mathcal{A}(\mathcal{D}')}(\theta)}\right)^{\alpha}\right],$$

and where $\mu_{\mathcal{A}}$ denotes the density $\mathcal{A}$.

[1]The code is provided in the supplementary material.

Conversion from Rényi differential privacy to (ϵ, δ)-differential privacy is given by the following proposition:

Proposition 2.1 (From Rényi to (ϵ, δ)-differential privacy (Mironov, 2017))**.** If $\mathcal{A}$ satisfies (α, ε)-Rényi differential privacy, it also satisfies (ϵ, δ)-differential privacy for any $0 < \delta < 1$ with

$$\epsilon = \varepsilon + \frac{\log(1/\delta)}{\alpha - 1}$$

3. Privacy analysis of noisy stochastic gradient descent

We use the same notations as Chourasia et al. (2021). Let $\mathcal{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be a dataset of size n, with $\mathbf{x}_i \in \mathbb{R}^p$. Let $\ell(\theta, \mathbf{x})$ be the loss function of a learning algorithm parametrized by $\theta \in \mathcal{C}$ on an input x, where $\mathcal{C}$ is a closed convex set of $\mathbb{R}^d$. $\Pi_{\mathcal{C}}$ denotes the orthogonal projection onto $\mathcal{C}$. We denote by $\mathcal{L}_{\mathcal{D}}(\theta)$ the global empirical loss of the model, and by $\mathcal{L}_{\mathcal{B}_k}(\theta)$ the estimated empirical loss computed on the batch $\mathcal{B}_k$ of size $|\mathcal{B}_k| = m$.

$$\mathcal{L}_{\mathcal{D}}(\theta) = \frac{1}{n}\sum_{\mathbf{x}\in\mathcal{D}} \ell(\theta, \mathbf{x}) \qquad \mathcal{L}_{\mathcal{B}_k}(\theta) = \frac{1}{m}\sum_{\mathbf{x}\in\mathcal{B}_k} \ell(\theta, \mathbf{x}).$$

Algorithm 1 $\mathcal{A}_{\mathrm{DP-SGLD}}$: Noisy Stochastic Gradient Descent

Input: Dataset $\mathcal{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, loss function ℓ, step size $\{\eta_k\}_{k\geq 0}$, noise variance σ^2 and initial parameter $\theta_0 \in \mathcal{C}$
for $k = 0, \ldots, K-1$ **do**
  Sample batch $\mathcal{B}_k$ from $\mathcal{D}$ with replacement
  Compute $\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k) = \frac{1}{m}\sum_{\mathbf{x}\in\mathcal{B}_k} \nabla\ell(\theta_k, \mathbf{x})$
  $\theta_{k+1} = \Pi_{\mathcal{C}}(\theta_k - \eta_k\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k) + \sqrt{2\eta_k}\,\mathcal{N}(0, \sigma^2\mathbf{I}_d))$
end for
Output: θ_K

We analyse the privacy loss of the DP-SGLD algorithm given in Algorithm 1 which implements noisy stochastic gradient descent.

Let θ_k and θ'_k denote the parameters at the k-th iteration of $\mathcal{A}_{\mathrm{DP-SGLD}}$ on neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, respectively. Stating that $\mathcal{D}$ and $\mathcal{D}'$ are *neighbors* means that they only differ by one $\mathbf{x}_{i_0}$, for some index i_0. Batch $\mathcal{B}_k$ is sampled with replacement from $\mathcal{D}$, meaning that a subset of size m is chosen at random from $\mathcal{D}$ and then replaced at the end of the k-th iteration. Hence, $\mathbf{x}_{i_0}$ can only appear once in $\mathcal{B}_k$, with probability m/n. We denote by Θ_{t_k} and Θ'_{t_k} the corresponding random variables associated with θ_k and θ'_k. We abuse notation to also denote their probability density functions by Θ_{t_k} and Θ'_{t_k}. Our objective is to model and analyze the dynamics of differential privacy of this algorithm

and to compare it to the ones of the DP-GLD algorithm described by Chourasia et al. (2021), which implements full noisy gradient descent. To this aim, we use the theoretical framework and constructions they provide to analyze the privacy loss of releasing the output θ_K of the algorithm, assuming private internal states (i.e., $\theta_1, \ldots, \theta_{K-1}$).

More precisely, the proof strategy goes as follows: we first model the transition from any step k to the next step $k+1$ in DP-SGLD using a diffusion process and derive the evolution equation of the distribution Θ_t during $k < t < k+1$. We use the theoretical results of Chourasia et al. (2021) to establish the evolution of the Rényi divergence of two distributions based on neighboring datasets during $k < t < k+1$. Finally, we compute a bound on the global Rényi divergence for the K-step DP-SGLD process.

3.1. Tracing diffusion for DP-SGLD

To analyze the privacy loss of DP-SGLD, which is a discrete-time stochastic process, we first interpolate each discrete update from θ_k to θ_{k+1} with a piecewise continuously differentiable diffusion process. Given step size $\{\eta_k\}_{k\geq 0}$, variance noise σ^2 and initial parameter vector θ_0, the k-th discrete update in Algorithm 1 is:

$$\begin{aligned} &\theta_{k+1} = \Pi_{\mathcal{C}}(\theta_k - \eta_k \nabla \mathcal{L}_{\mathcal{B}_k}(\theta_k) + \sqrt{2\eta_k\sigma^2}\mathbf{Z}_K) \\ &\text{with } \mathbf{Z}_k \sim \mathcal{N}(0, \mathbf{I}_d), \mathcal{B}_k \overset{\$}{\leftarrow} \mathcal{D}, \end{aligned} \tag{1}$$

where $\mathcal{B}_k \overset{\$}{\leftarrow} \mathcal{D}$ means that batch $\mathcal{B}_k$ is sampled with replacement from $\mathcal{D}$, and where the loss $\mathcal{L}_{\mathcal{B}_k}(\theta_k)$ is the estimated empirical loss function over batch $\mathcal{B}_k$. This discrete jump can be interpolated with the following random process $\{\Theta_t\}_{t_k \leq t \leq t_{k+1}}$, where $t_k = \sum_{i=1}^{k} \eta_i$,

$$\begin{cases} \Theta_{t_k} = \theta_k \\ \Theta_{t_k+\Delta t} - \left(\Theta_{t_k} - \eta_k \sum_{i=1, i\neq i_0}^{m} \dfrac{\nabla \ell(\Theta_{t_k}, \mathbf{x}_i)}{m}\right) \\ \qquad = -\Delta t \dfrac{\nabla \ell(\Theta_{t_k}, \mathbf{x}_{i_0})}{m} + \sqrt{2\Delta t\, \sigma^2}\mathbf{Z}_k,\ 0 < \Delta t < \eta_k \\ \Theta_{t_{k+1}} = \Pi_{\mathcal{C}}(\lim_{\Delta t \to \eta_k} \Theta_{t_k+\Delta t}),\ 0 < \Delta t < \eta_k, \end{cases} \tag{2}$$

where $\mathbf{Z}_k \sim \mathcal{N}(0, \mathbf{I}_d)$, $\mathcal{B}_k$ is sampled with replacement from $\mathcal{D}$, i_0 refers to the index in $\mathcal{B}_k$ of the data item which differs between $\mathcal{D}$ and $\mathcal{D}'$ if it is part of $\mathcal{B}_k$, otherwise i_0 is chosen at random.

We can compute from (2) $\lim_{\Delta t \to \eta_k} \Theta_{t_k+\Delta t} = \theta_k - \eta_k \nabla \mathcal{L}_{\mathcal{B}_k}(\theta_k) + \sqrt{2\eta_k\sigma^2}\mathbf{Z}_K$. Therefore by the update equation (1), we see that the random variable $\Theta_{t_{k+1}} = \Pi_{\mathcal{C}}(\lim_{\Delta t \to \eta_k} \Theta_{t_k+\Delta t})$ has the same distribution as the parameter θ_{k+1} at $k+1^{\text{th}}$ step of DP-SGLD.

We now differentiate (2) over time t for $\{\Theta_t\}_{t_k < t < t_{k+1}}$ and we derive the following stochastic differential equation.

$$d\Theta_t = -\frac{\nabla \ell(\Theta_{t_k}, \mathbf{x}_{i_0})}{m} dt + \sqrt{2\sigma^2} d\mathbf{W}_t, \tag{3}$$

where $d\mathbf{W}_t \sim \sqrt{t}\mathcal{N}(0, \mathbf{I}_d)$ describes the Wiener process on $\mathbb{R}^d$, and i_0 is chosen as such:

$$\begin{cases} \{\mathbf{x}_{i_0}\} = \{\mathbf{x}_i, \mathbf{x}_i \in \mathcal{B}_k, \mathbf{x}_i \notin \mathcal{B}'_k\} & \text{if } \mathcal{B}_k \neq \mathcal{B}'_k \\ \mathbf{x}_{i_0} \overset{\$}{\leftarrow} \mathcal{B}_k & \text{if } \mathcal{B}_k = \mathcal{B}'_k, \end{cases}$$

where we assume without loss of generality that two neighboring batches $\mathcal{B}_k$ and $\mathcal{B}'_k$ are indexed so as to be equal on all indices but one.

This shows that $\{\Theta_t\}_{t_k < t < t_{k+1}}$ is a diffusion process and we repeat the construction for $k = 0, 1, \ldots$ to define a piecewise continuous diffusion process $\{\Theta_t\}_{t\geq 0}$ whose distribution at time $t = t_k$ is consistent with θ_k. We refer to this process as the tracing diffusion for DP-SGLD.

Definition 3.1 (Coupled tracing diffusions (Chourasia et al., 2021))**.** Let $\Theta_0 = \Theta'_0$ be two identically distributed random variables. We refer to the stochastic processes $\{\Theta_t\}_{t\geq 0}$ and $\{\Theta'_t\}_{t\geq 0}$ defined by (2) as coupled tracing diffusions processes for DP-SGLD under loss function $\ell(\theta, \mathbf{x})$ on neighboring datasets $\mathcal{D}$, $\mathcal{D}'$ differing in i_0-th data point.

The Rényi divergence $R_\alpha(\Theta_{t_K} \,||\, \Theta'_{t_K})$ reflects the Rényi privacy loss of Algorithm 1 with K steps. Conditioned on observing θ_k and on sampling $\mathcal{B}_k$ from $\mathcal{D}$, the process $\{\Theta_t\}_{t_k < t < t_{k+1}}$ is a Langevin diffusion along a constant vector field $\nabla \ell(\theta_k, \mathbf{x}_{i_0})$ for duration η_k. Therefore, the conditional probability distribution $p_{t|t_k}(\theta|\theta_k)$ follows the following Fokker-Planck equation, where the notation $p_{t|t'}(\theta|\theta')$ represents the conditional probability density function $p(\Theta_t = \theta | \Theta_{t'} = \theta')$:

$$\frac{\partial p_{t|t_k}(\theta|\theta_k)}{\partial t} = \nabla \cdot \left(p_{t|t_k}(\theta|\theta_k) \frac{\nabla \ell(\theta_k, \mathbf{x}_{i_0})}{m} \right) + \sigma^2 \Delta p_{t|t_k}(\theta|\theta_k).$$

By taking expectations over the distribution $p_{t_k}(\theta_k)$ on both sides, we get the partial differential equation that models the evolution of probability distribution $p_t(\theta)$ in the tracing diffusion.

Lemma 3.1. For the SDE (3), the equivalent Fokker-Planck equation at time $t_k < t < t_{k+1}$ is

$$\frac{\partial p_t(\theta)}{\partial t} = \nabla \cdot \left(p_t(\theta) \mathop{\mathbb{E}}_{\theta_k \sim p_{t_k|t}} \left[\frac{\nabla \ell(\theta_k, \mathbf{x}_{i_0})}{m} | \theta, \mathcal{B}_k \right] \right) + \sigma^2 \Delta p_t(\theta).$$

Using this distribution evolution equation, we model the privacy dynamics in the tracing diffusion process. This process is similar to Langevin diffusion under conditionally expected loss function $\nabla \mathcal{L}_{\mathcal{B}_k}(\theta) = \mathbb{E}_{\theta_k \sim p_{t_k|t}} \left[\dfrac{\nabla \ell(\theta_k, \mathbf{x}_{i_0})}{m} \middle| \theta, \mathcal{B}_k \right]$.

3.2. Privacy erosion in tracing (Langevin) diffusion

The analysis of the privacy erosion in tracing langevin diffusion is detailed by Chourasia et al. (2021) but we provide here the key results that we will apply to our algorithm. Privacy erosion refers to the continuous increase of the privacy loss over time as more data is accessed to compute the gradient of the loss in the update equation.

Consider a Langevin diffusion process modelled through the following Fokker-Planck equation:

$$\frac{\partial p_t(\theta)}{\partial t} = \nabla \cdot (p_t(\theta) \nabla \mathcal{L}_{\mathcal{B}_k}(\theta)) + \sigma^2 \Delta p_t(\theta).$$

Definition 3.2 (Log-Sobolev inequality (Gross, 1975))**.** The distribution of a variable Θ satisfies the Log-Sobolev Inequality for a constant c, or is c-LSI, if for all functions $f : \mathbb{R}^d \to \mathbb{R}$ with continuous ∇f and $\mathbb{E}[f(\Theta)^2] < \infty$, it satisfies

$$\mathbb{E}[f(\Theta)^2 \log f(\Theta)^2] - \mathbb{E}[f(\Theta)^2] \log \mathbb{E}[f(\Theta)^2] \leq \frac{2}{c} \mathbb{E}[\|\nabla f(\Theta)\|_2^2].$$

Lemma 3.2 (Dynamics for Rényi privacy loss under c-LSI (Chourasia et al., 2021))**.** Assuming that Θ is c-LSI and S_g is the sensitivity of the loss gradient, the dynamics of the Rényi privacy loss can be modelled as such, where $R(\alpha, t)$ represents the α Rényi divergence $R_\alpha(\Theta_t \mid\mid \Theta'_t)$:

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

for some γ which can be arbitrarily fixed.

The initial privacy loss satisfies $R(\alpha, 0) = 0$ as $\Theta_0 = \Theta'_0$. The solution $R(\alpha, t)$ for this equation increases with time $t \geq 0$, which models the erosion of Rényi privacy loss in Langevin diffusion. Intuitively, the c-LSI condition which provides the negative dependence $\frac{\partial R(\alpha,t)}{\partial t}$ with respect to $R(\alpha, t)$, can be interpreted as a sufficient condition to ensure that the the Rényi privacy loss is bounded, which is further formalized in Theorem 3.1.

3.3. Privacy guarantee for DP-SGLD

We now extend these results to the tracing diffusion for DP-SGLD. In addition, we do not make the assumption that the step size is constant and use a sequence $\{\eta_k\}_{k \geq 0}$ instead. We first bound the gradient sensitivity of the conditionally expected loss $\nabla \mathcal{L}_{\mathcal{B}_k}(\theta) = \mathbb{E}_{\theta_k \sim p_{t_k|t}} \left[\frac{\nabla \ell(\theta_k, \mathbf{x}_{i_0})}{m} | \theta, \mathcal{B}_k \right]$.

Lemma 3.3 (Sensitivity)**.** Let $\ell(\theta, \mathbf{x})$ be an L-Lipschitz loss function on closed convex set $\mathcal{C}$, then for coupled tracing diffusions $\{\Theta_t\}_{t \geq 0}$ and $\{\Theta'_t\}_{t \geq 0}$ for DP-SGLD on neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, and noise variance σ^2, the gradient sensitivity of conditionally expected loss $\nabla \mathcal{L}_{\mathcal{B}_k}(\theta) = \mathbb{E}_{\theta_k \sim p_{t_k|t}} \left[\frac{\nabla \ell(\theta_k, \mathbf{x}_{i_0})}{m} | \theta, \mathcal{B}_k \right]$ is bounded by:

$$\left\| \mathop{\mathbb{E}}_{\theta_k \sim p_{t_k|t}} \left[\frac{\nabla \ell(\theta_k, \mathbf{x}_{i_0})}{m} | \theta, \mathcal{B}_k \right] - \mathop{\mathbb{E}}_{\theta'_k \sim p'_{t_k|t}} \left[\frac{\nabla \ell(\theta'_k, \mathbf{x}'_{i_0})}{m} | \theta, \mathcal{B}_k \right] \right\|_2 \leq \frac{2L}{n},$$

where n is the size of the dataset $\mathcal{D}$ and $\mathcal{D}'$, and m is the size of the batch $\mathcal{B}_k$ and $\mathcal{B}'_k$.

The proof is provided in Appendix on page 10.

We now substitute the sensitivity S_g with $\frac{2L}{n}$ in equation (4) modelling the Rényi privacy loss dynamics of tracing diffusion at $t_k \leq t < t_{k+1}$ under c-LSI condition:

$$\frac{\partial R(\alpha,t)}{\partial t} \leq \frac{1}{\gamma} \frac{\alpha L^2}{n^2 \sigma^2} - 2(1-\gamma)\sigma^2 c \left[\frac{R(\alpha,t)}{\alpha} + (\alpha - 1) \frac{\partial R(\alpha,t)}{\partial \alpha} \right]. \tag{5}$$

Following the methodology of Chourasia et al. (2021), we solve this PDE under $\gamma = \frac{1}{2}$ and derive the RDP guarantee for the DP-SGLD algorithm.

Theorem 3.1 (RDP for DP-SGLD under c-LSI)**.** Let $\ell(\theta, \mathbf{x})$ be an L-Lipschitz loss function on a closed convex set $\mathcal{C}$. Let $\{\Theta_t\}_{t \geq 0}$ be the tracing diffusion for $\mathcal{A}_{\mathrm{DP-SGLD}}$ under loss function $\ell(\theta, \mathbf{x})$ on dataset $\mathcal{D}$. If Θ_t satisfies c-LSI throughout $0 \leq t \leq \sum_{k=1}^K \eta_k$, then $\mathcal{A}_{\mathrm{DP-SGLD}}$ satisfies (α, ε)-Rényi Differential Privacy for

$$\varepsilon = \frac{2\alpha L^2}{cn^2\sigma^4} (1 - e^{-\sigma^2 c \sum_{k=1}^K \eta_k}).$$

The complete proof is provided in Appendix on page 10.

Sketch of proof. We introduce $t_k = \sum_{i=1}^k \eta_i$ for $k \geq 0$. The idea of the proof is to bound $R(\alpha, t_K)$, the Rényi divergence after K updates in DP-SGLD, with a function of $R(\alpha, t_0)$. This is first done by considering the k-th update, and proving the following equation for some constants a_1, a_2

$$R(\alpha, \lim_{t \to t_{k+1}^-} t) - \frac{a_2}{a_1}\alpha \leq (R(\alpha, \lim_{t \to t_k^+} t) - \frac{a_2}{a_1}\alpha) e^{-a_1 \eta_{k+1}},$$

together with

$$R(\alpha, \lim_{t \to t_k^+} t) \leq R(\alpha, \lim_{t \to t_k^-} t).$$

This allows to bound $R(\alpha, \lim_{t \to t_{k+1}^-} t)$ with a function of $R(\alpha, \lim_{t \to t_k^-} t)$ and the final bound follows by recursivity, by noting that $R(\alpha, t_0^-) = 0$ since coupled tracing diffusions have the same start parameter.

This theorem guarantees that under the c-LSI condition, the privacy loss converges during the noisy SGD process if $\lim_{K \to \infty} \sum_{k=1}^K \eta_k = \infty$.

In particular, the case where the step size is constant is straightforward:

Corollary 3.1 (RDP for DP-SGLD under c-LSI with constant step-size)**.** Let Θ_t be defined as in Theorem 3.1. If $\mathcal{A}_{\mathrm{DP-SGLD}}$ has constant step size η and if Θ_t satisfies c-LSI throughout $0 \leq t \leq \eta K$, then $\mathcal{A}_{\mathrm{DP-SGLD}}$ satisfies (α, ε) Rényi Differential Privacy for

$$\varepsilon = \frac{2\alpha L^2}{cn^2\sigma^4}(1 - e^{-c\sigma^2\eta K}).$$

In addition, Chourasia et al. (2021) show that the c-LSI condition is satisfied in DP-GLD with constant step size, for loss functions that are Lipschitz, strongly convex and smooth, with appropriate conditions on the algorithm parameters and initialization. We derive an equivalent lemma for DP-SGLD with varying step size.

Lemma 3.4 (LSI for DP-SGLD)**.** If loss function $\ell(\theta, \mathbf{x})$ is λ-strongly convex and β-smooth over a closed convex set $\mathcal{C}$, then the coupled tracing diffusion processes $\{\Theta_t\}_{t\geq 0}$ and $\{\Theta'_t\}_{t\geq 0}$ for DP-SGLD with step size $\{\eta_k\}_{k\geq 0}$ satisfying $\eta_k < \frac{1}{\beta}$ for $k \geq 0$, and with initial distribution $\Theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\mathbf{I}_d))$, satisfy c-LSI for $t \geq 0$ with $c = \frac{\lambda}{2\sigma^2}$.

The proof of this lemma is exactly the same as Lemma 5 of Chourasia et al. (2021), where n needs to be replaced with the batch size m and η with η_k.

We immediately derive the following bound on the Rényi privacy loss for DP-SGLD.

Theorem 3.2 (Privacy guarantee for DP-SGLD)**.** Let $\ell(\theta, \mathbf{x})$ be an L-Lipschitz, λ-strongly convex and β-smooth loss function on closed convex set $\mathcal{C}$, then $\mathcal{A}_{\mathrm{DP-SGLD}}$ with start parameter $\theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\mathbf{I}_d))$ and step-size $\eta < \frac{1}{\beta}$ satisfies (α, ε)-Rényi differential privacy with

$$\varepsilon = \frac{4\alpha L^2}{\lambda n^2\sigma^2}(1 - e^{-\frac{\lambda}{2}\sum_{k=1}^{K}\eta_k}).$$

The case where the step size is constant follows:

Corollary 3.2 (Privacy Guarantee for DP-SGLD with constant step-size)**.** With $\ell(\theta, \mathbf{x})$ and $\mathcal{A}_{\mathrm{DP-SGLD}}$ defined as in Theorem 3.2, and with constant step size $\eta < \frac{1}{\beta}$, $\mathcal{A}_{\mathrm{DP-SGLD}}$ satisfies (α, ε)-Rényi Differential Privacy with

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

We also provide the case where the step size is defined as $\eta_k = \frac{1}{2\beta + \lambda k/2}$, which is further analyzed in the next section.

Corollary 3.3 (Privacy Guarantee for DP-SGLD with decreasing step-size)**.** With $\ell(\theta, \mathbf{x})$ and $\mathcal{A}_{\mathrm{DP-SGLD}}$ defined as in Theorem 3.2, and with step size $\eta_k = \frac{1}{2\beta + \lambda k/2}$, $\mathcal{A}_{\mathrm{DP-SGLD}}$ satisfies (α, ε)-Rényi Differential Privacy with

$$\varepsilon = \frac{4\alpha L^2}{\lambda\sigma^2 n^2}(1 - e^{-\log(1+\frac{\lambda K}{4\beta})}) = \frac{4\alpha L^2}{\lambda n^2\sigma^2}\frac{\lambda K}{4\beta + \lambda K}.$$

Discussion. Let us consider that η in Corollary 3.2 is set as $\eta = \frac{1}{2\beta}$, so that it can be compared to Corollary 3.3 and also matches the maximum upper bound for which we derive utility guarantees is the next section. In the regime where K is small (compared to $\frac{\beta}{\lambda}$), both corollaries have equivalent bounds on ε. Indeed, in the fixed step size setting we have with $\eta = \frac{1}{2\beta}$:

$$\varepsilon = \frac{4\alpha L^2}{\lambda n^2\sigma^2}(1 - e^{-\lambda K/4\beta}) \sim_{K \ll \frac{\beta}{\lambda}} \frac{\alpha L^2 K}{\beta n^2\sigma^2},$$

while in the decreasing step size setting we have

$$\varepsilon = \frac{4\alpha L^2}{\lambda n^2\sigma^2}\frac{\lambda K}{4\beta + \lambda K} \sim_{K \ll \frac{\beta}{\lambda}} \frac{\alpha L^2 K}{\beta n^2\sigma^2}.$$

In particular, ε reaches the baseline composition analysis from Abadi et al. (2016) up to a factor 2 : $\varepsilon' = \frac{\alpha L^2}{n^2\sigma^2} \cdot \eta K = \frac{\alpha L^2 K}{2\beta n^2\sigma^2}$.

In the regime where K is sufficiently large (compared to $\frac{\beta}{\lambda}$), both the fixed η and decreasing η_k settings also reach the same bound on ε, equal to

$$\varepsilon \sim_{K \gg \frac{\beta}{\lambda}} \frac{4\alpha L^2}{\lambda n^2\sigma^2}.$$

As a side note, we notice that we can consider the unconstrained regularized version of the problem: $\widetilde{\mathcal{L}_{\mathcal{D}}}(\theta) = \mathcal{L}_{\mathcal{D}}(\theta) + \frac{\lambda}{2}\|\theta\|_2^2$ and derive equivalent properties. In this scenario, we no longer need the strong convexity assumption on $\mathcal{L}_{\mathcal{D}}(\theta)$. In addition we can use the optimality of θ^* for $\widetilde{\mathcal{L}_{\mathcal{D}}}$ to derive two equations :

$$\mathcal{L}_{\mathcal{D}}(\theta^*) + \frac{\lambda}{2}\|\theta^*\|_2^2 \leq \mathcal{L}_{\mathcal{D}}(0), \qquad c\nabla\mathcal{L}_{\mathcal{D}}(\theta^*) + \lambda\theta^* = 0.$$

Each of these provides a bound on $\|\theta^*\|_2$, by using respectively the positivity of $\mathcal{L}_{\mathcal{D}}$ and the lipschitzness of $\mathcal{L}_{\mathcal{D}}$,

$$\|\theta^*\|_2 \leq \Big(\frac{2\mathcal{L}_{\mathcal{D}}(0)}{\lambda}\Big)^{1/2}, \quad \|\theta^*\|_2 \leq \frac{L}{\lambda},$$

which can be used as bounds for the radius of the convex $\mathcal{C}$ we project onto, so that we still actually end up solving the unconstrained problem.

4. Utility analysis for noisy stochastic gradient descent

Differential privacy is known for setting a trade-off between privacy and utility. To assess the utility of the noisy stochastic gradient descent algorithm $\mathcal{A}_{\mathrm{DP-SGLD}}$, we measure two quantities, the worst case excess empirical risk

$$\max_{\mathcal{D}\in\mathcal{X}^n} \mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K) - \mathcal{L}_{\mathcal{D}}(\theta^*)],$$

and the worst case average empirical risk

$$\max_{\mathcal{D}\in\mathcal{X}^n} \mathbb{E}[\frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\mathcal{D}}(\theta_k) - \mathcal{L}_{\mathcal{D}}(\theta^*)], \tag{6}$$

where θ_K is the output of the randomized algorithm $\mathcal{A}_{\mathrm{DP-SGLD}}$ on $\mathcal{D}$ during K iterations, θ^* is the solution to the standard non-noisy GD algorithm and the expectation is taken over the randomness of the algorithm.

4.1. Fixed step size η

We propose a bound on the worst case excess empirical risk when the learning rate is fixed and satisfies $\eta < \frac{1}{2\beta}$.

Lemma 4.1 (Empirical risk for smooth and strongly convex loss)**.** Let $\ell(\theta, \mathbf{x})$ be an L-Lipschitz, λ-strongly convex and β-smooth loss function on closed convex set $\mathcal{C}$, $\mathcal{A}_{\mathrm{DP-SGLD}}$ be parameterized with step-size $\eta < \frac{1}{2\beta}$ and start parameter $\theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\mathbf{I}_d))$, then the empirical risk of $\mathcal{A}_{\mathrm{DP-SGLD}}$ is bounded by

$$\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K) - \mathcal{L}_{\mathcal{D}}(\theta^*)] \leq \frac{\beta}{2}\mathbb{E}[\|\theta_0 - \theta^*\|_2^2]e^{-\lambda\eta K} + \frac{\beta\eta\xi^2}{2\lambda} + \frac{\beta d\sigma^2}{\lambda}, \tag{7}$$

where θ^* is the minimizer of $\mathcal{L}_{\mathcal{D}}(\theta)$ in $\mathcal{C}$ and $\xi^2 = \mathbb{E}[\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta^*)\|_2^2]$.

The proof is provided in Appendix on page 11.

Sketch of proof. First, we recursively bound $\|\theta_{k+1} - \theta^*\|_2^2$ as a function of $\|\theta_k - \theta^*\|_2^2$, using the definition of θ_{k+1}. Then, we take the expectation with respect to $\mathcal{B}_k$ and use co-coercivity of the gradients, and take the expectation again to derive a recursive relationship between $\mathbb{E}[\|\theta_{k+1} - \theta^*\|_2^2]$ and $\mathbb{E}[\|\theta_k - \theta^*\|_2^2]$. Last, we express the empirical risk $\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K) - \mathcal{L}_{\mathcal{D}}(\theta^*)]$ as a function of $\mathbb{E}[\|\theta_K - \theta^*\|_2^2]$.

This shows that under Lipschitz smooth strongly convex loss function, the empirical risk of $\mathcal{A}_{\mathrm{DP-SGLD}}$ decreases as the iterations increase, and reaches a constant factor which is the sum of a term directly related to the variance of the noise σ^2 added at each iteration and another term which represents the error due to the SGD process, which decreases with the learning rate η.

Lemma 4.2 (Empirical risk for smooth and strongly convex loss, independent of θ)**.** Let ℓ and $\mathcal{A}_{\mathrm{DP-SGLD}}$ be defined as in Lemma 4.1, then the empirical risk is bounded by

$$\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K) - \mathcal{L}_{\mathcal{D}}(\theta^*)] \leq \frac{2\beta L^2}{\lambda^2}e^{-\lambda\eta K} + \frac{\beta\eta\xi^2}{2\lambda} + \frac{\beta d\sigma^2}{\lambda},$$

where $\xi^2 = \mathbb{E}[\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta^*)\|_2^2]$.

Proof. Since we have $\|\mathcal{C}\| \leq \frac{2L}{\lambda}$, we can bound $\mathbb{E}[\|\theta_0 - \theta^*\|_2^2] \leq \frac{4L^2}{\lambda^2}$, as $\theta_0, \theta^* \in \mathcal{C}$. □

Combining now Lemma 4.2 and Theorem 3.2, we derive the utility of $\mathcal{A}_{\mathrm{DP-SGLD}}$ under (α, ε)-Rényi differential privacy.

Theorem 4.1 (Utility bound for (α, ε)-Rényi differential privacy)**.** Let $\ell(\theta, \mathbf{x})$ be an L-Lipschitz, λ-strongly convex and β-smooth loss function on closed convex set $\mathcal{C}$, then $\mathcal{A}_{\mathrm{DP-SGLD}}$ with start parameter $\theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\mathbf{I}_d))$ and constant step-size $\eta = \frac{1}{2\beta}$, satisfies (α, ε) Rényi differential privacy and

$$\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K) - \mathcal{L}_{\mathcal{D}}(\theta^*)] = O\left(\frac{\alpha\beta d L^2}{\varepsilon\lambda^2 n^2}\right) + \frac{\xi^2}{4\lambda},$$

where σ^2 and K are set as such:

$$\sigma^2 = \frac{4\alpha L^2}{\varepsilon\lambda n^2}, \quad K = \frac{2\beta}{\lambda}\log\left(\frac{\varepsilon n^2}{\alpha d}\right).$$

The complete proof is provided in Appendix on page 13.

Theorem 4.2 (Utility bound for (ϵ, δ)-differential privacy)**.** With the same conditions as is Theorem 4.1, for $\epsilon \leq 2\log(1/\delta)$ and $\delta > 0$, $\mathcal{A}_{\mathrm{DP-SGLD}}$ satisfies (ϵ, δ) differential privacy and

$$\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K) - \mathcal{L}_{\mathcal{D}}(\theta^*)] = O\left(\frac{\beta d L^2 \log(1/\delta)}{\epsilon^2\lambda^2 n^2}\right) + \frac{\xi^2}{4\lambda},$$

where σ^2 and K are set as such:

$$\sigma^2 = \frac{8L^2(\epsilon + 2\log(1/\delta)}{\epsilon^2\lambda n^2}$$
$$K = \frac{2\beta}{\lambda}\log\left(\frac{\epsilon^2 n^2}{4\log(1/\delta)d}\right).$$

The complete proof is provided in Appendix on page 13.

As a side note, arguments of the proof of Lemma 4.1 (like the co-coercivity of the gradients) can be reused to improve the upper bound on η from $\frac{\lambda}{2\beta^2}$ to $\frac{1}{2\beta}$ in Theorem 4 of Chourasia et al. (2021) on the utility of DP-GLD. We provide experimental evidence in the next section that this factor $\frac{\lambda}{\beta}$ is non-negligible.

4.2. Decreasing step size η_k

To remove the $\frac{\xi^2}{4\lambda}$ term which is due to using stochastic gradient descent, we follow the approach from Roux et al. (2012) and propose to bound the worst case average empirical risk (6) when the step size is decreasing and follows:

$$\eta_k = \frac{1}{2\beta + \frac{\lambda k}{2}}, \; k \geq 0.$$

Lemma 4.3 (Empirical risk for smooth and strongly convex loss with decreasing learning rate)**.** Let $\ell(\theta, \mathbf{x})$ be an L-Lipschitz, λ-strongly convex and β-smooth loss function on closed convex set $\mathcal{C}$, $\mathcal{A}_{\mathrm{DP-SGLD}}$ be parameterized

with *decreasing* step-size $\eta_k = \frac{1}{2\beta+\lambda k/2}$ and start parameter $\theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\mathbf{I}_d))$, then the average empirical risk of $\mathcal{A}_{\mathrm{DP-SGLD}}$ is bounded by

$$\mathbb{E}[\frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\mathcal{D}}(\theta_k) - \mathcal{L}_{\mathcal{D}}(\theta^*)] \leq \frac{2\beta}{K}\,\mathbb{E}[\|\theta_0 - \theta^*\|_2^2] + \frac{4\xi^2}{K\lambda}\log\left(1+\frac{\lambda K}{4\beta}\right) + 2d\sigma^2, \quad (8)$$

where θ^* is the minimizer of $\mathcal{L}_{\mathcal{D}}(\theta)$ in $\mathcal{C}$.

The proof is provided in Appendix on page 14.

The term $\frac{\xi^2}{\lambda}$ decreases roughly in $\frac{1}{K}\log(K)$. The term $d\sigma^2$ still appears as in Lemma 4.1 but without the $\frac{\beta}{\lambda}$ factor, that we observe to be quite significant in Section 5.

We then use this lemma to derive the following utility bound under (α, ε)-Rényi differential privacy:

Theorem 4.3 (Utility bound for (α, ε)-Rényi differential privacy with decreasing learning rate)**.** Let $\ell(\theta, \mathbf{x})$ be an L-Lipschitz, λ-strongly convex and β-smooth loss function on closed convex set $\mathcal{C}$, then $\mathcal{A}_{\mathrm{DP-SGLD}}$ with start parameter $\theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\mathbf{I}_d))$ and *decreasing* step-size $\eta_k = \frac{1}{2\beta+\lambda k/2}$, satisfies (α, ε) Rényi differential privacy and

$$\mathbb{E}[\frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\mathcal{D}}(\theta_k) - \mathcal{L}_{\mathcal{D}}(\theta^*)] = O\left(\frac{\alpha d L^2}{\varepsilon\lambda n^2}\right),$$

where σ^2 and K are set as such:

$$\sigma^2 = \frac{4\alpha L^2}{\varepsilon\lambda n^2}, \quad K = \max\left(\frac{\beta}{\lambda}\frac{\varepsilon n^2}{\alpha d}, \frac{\lambda}{\beta}\left(\frac{\varepsilon n^2}{\alpha d}\right)^2\right).$$

The complete proof is provided in Appendix on page 14.

Compared to previous Theorem 4.1, we improve the utility bound by a factor $\frac{\beta}{\lambda}$ which is non negligible in practice. However, the number of iterations K can now evolve either in n^2 or in n^4 in the regime where $\frac{\varepsilon n^2}{\alpha d} > \left(\frac{\beta}{\lambda}\right)^2$.

Theorem 4.4 (Utility bound for (ϵ, δ)-differential privacy with decreasing learning rate)**.** With the same conditions as is Theorem 4.3, for $\epsilon \leq 2\log(1/\delta)$ and $\delta > 0$, $\mathcal{A}_{\mathrm{DP-SGLD}}$ satisfies (ϵ, δ) differential privacy and

$$\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K) - \mathcal{L}_{\mathcal{D}}(\theta^*)] = O\left(\frac{dL^2\log(1/\delta)}{\epsilon^2\lambda n^2}\right),$$

where σ^2 and K are set as such:

$$\sigma^2 = \frac{8L^2(\epsilon + 2\log(1/\delta))}{\epsilon^2\lambda n^2}$$
$$K = \max\left(\frac{\beta}{\lambda}\frac{\epsilon^2 n^2}{4\log(1/\delta)d}, \frac{\lambda}{\beta}\left(\frac{\epsilon^2 n^2}{4\log(1/\delta)d}\right)^2\right).$$

The complete proof is provided in Appendix on page 15.

5. Experiments: application to logistic regression

We now propose an experimental evaluation of DP-SGLD and compare it to DP-SGD on a classification task using logistic regression on two vision datasets, CIFAR10 and Pneumonia, a dataset of chest X-ray images of pediatric pneumonia published by Kermany et al. (2018). Details about the datasets and models are available in Appendix D.

The task consists of pre-training a model (here AlexNet or ResNet18) on a dataset that will be considered *public* (here we take CIFAR100 and Imagenet). Then, all layers of the model are freezed except for the last one which is retrained from scratch using a softmax loss function on a *private* dataset (here CIFAR10 or Pneumonia). This corresponds to logistic regression and some regularization is added to guarantee strong convexity. Pre-training provides generic feature maps learned on a public dataset which improves the task accuracy without any compromise on the privacy.

First, we formalize this setting and provide the smoothness and convexity constants for logistic regression. Second, we report accuracy achieved with DP-SGLD and compare it to DP-SGD for constant and decreasing step size η.

5.1. Smoothness and convexity of logistic regression

For clarity, we replace the generic parameter θ with the single matrix $\mathbf{W}$ that it represents for logistic regression.

The loss with regularization writes $\ell(\mathbf{W}, \mathbf{x}) = \log\left(\boldsymbol{\sigma}(\mathbf{W}\mathbf{x})\right)_y + \lambda\|\mathbf{W}\|_2^2$ where C is the number of classes, $y \in [1, C]$ is the label of sample $\mathbf{x}$, $\mathbf{W} \in \mathbb{R}^{C\times p}$ and $\boldsymbol{\sigma} : \mathbb{R}^C \mapsto \mathbb{R}^C$ is the sigmoid function (not to be confused with the noise variance σ):

$$\left(\boldsymbol{\sigma}(\mathbf{z})\right)_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}, \; \forall i = 1, .., C\,, \mathbf{z} \in \mathbb{R}^C$$

Lemma 5.1 (Convexity and smoothness constants for regularized logistic regression)**.** Let $\ell(\mathbf{W}, \mathbf{x})$ be defined as above. Then ℓ is λ-strongly convex and β-smooth, with

$$\beta = \frac{1}{2}\lambda_{\mathsf{max}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right) + \lambda$$

where λ_{max} refers to the maximum eigenvalue.

The proof is provided in Appendix on page 15.

5.2. Experimental utility of logistic regression

We compare our DP-SGLD algorithm with the standard DP-SGD from Abadi et al. (2016) implemented in Opacus and with the baseline SGD without DP on several vision tasks. In particular, we study the case where the step size

is constant $\eta = \frac{1}{2\beta}$ and where it is decreasing as follows : $\eta_k = \frac{1}{2\beta + \lambda k/2}$. To be able to provide somewhat comparable results, all methods (DP-SGLD, DP-SGD and No-DP) use the same step size, number of training epochs, privacy budget $(\epsilon, \delta) = (1.0, 10^{-5})$ when applicable and use no momentum. Other hyperparameters are tuned to provide optimal accuracy for each method and are provided in the source code included in the submission.

Results are given in Table 1 for constant step size and in Table 2 for decreasing step size. The model indicated is the feature extraction model, which is pre-trained on CIFAR100 when the task is on CIFAR10 and on Imagenet when the task is on Pneumonia. Only its last layer is re-trained using logistic regression. As the tables show, DP-SGLD outperforms standard DP-SGD for the tasks considered and considerably reduces the gap in accuracy compared to SGD without differential privacy. However, such results need to be taken cautiously before drawing conclusions since this task is strongly convex and smooth while DP-SGD also applies to non-convex tasks.

To better understand the effect of clamping DP-SGD to smooth and strongly convex tasks, we repeat the first experiment of Table 1, but instead of leveraging only the last layer for DP-SGD, we also unfreeze the last and fourth block of the ResNet18 architecture, composed notably of 5 convolutional layers. Results for DP-SGD and SGD without DP provided in Table 3 show that while SGD benefits from this fine-tuning and increases accuracy from 70.7% to 77.0%, DP-SGD does not improve and accuracy even decreases marginally from 68.0% to 67.8%. Such observation aligns with Tramèr & Boneh (2021) in the sense that basic models like logistic regression currently are competitive compared to deeper models when trained with differential privacy, which underlines the importance of studying classical tasks like training smooth and strongly convex objectives.

Last, we provide in Table 4 the experimental value of some parameters, in light with comments made after Lemma 4.3 about the value of $\frac{\beta}{\lambda}$ and after Theorem 4.3 about the dependence in n^2 or n^4 of K, depending of the ratio $\frac{\varepsilon n^2}{\alpha d} / \left(\frac{\beta}{\lambda}\right)^2$. As we show, this ratio is of magnitude 10^{-3} or less which shows that the evolution of K is quadratic in n.

6. Conclusion

We have extended the theoretical framework of Chourasia et al. (2021) to provide a differential privacy analysis of noisy stochastic gradient descent based on Langevin diffusion (DP-SGLD) with arbitrary step size. Although our experiments already show the practical utility of our results, relaxing the smoothness and strong convexity hypothesis remains an open challenge and would pave the way for wide adoption by data scientists.

Table 1. Accuracy (in %) of logistic regression using SGD with a constant learning rate.

Method	Dataset	Model	Epochs	ϵ	Acc.
DP-SGLD	CIFAR10	Resnet18	30	1.0	70.3
DP-SGD	CIFAR10	Resnet18	30	1.0	68.0
No DP	CIFAR10	Resnet18	30	-	70.7
DP-SGLD	CIFAR10	Alexnet	30	1.0	57.5
DP-SGD	CIFAR10	Alexnet	30	1.0	56.4
No DP	CIFAR10	Alexnet	30	-	57.7
DP-SGLD	Pneumonia	Resnet18	50	1.0	58.8
DP-SGD	Pneumonia	Resnet18	50	1.0	58.8
No DP	Pneumonia	Resnet18	50	-	59.3

Table 2. Accuracy (in %) of logistic regression using SGD with a decreasing learning rate.

Method	Dataset	Model	Epochs	ϵ	Acc.
DP-SGLD	CIFAR10	Resnet18	30	1.0	70.1
DP-SGD	CIFAR10	Resnet18	30	1.0	68.1
No DP	CIFAR10	Resnet18	30	-	70.2
DP-SGLD	CIFAR10	Alexnet	30	1.0	57.3
DP-SGD	CIFAR10	Alexnet	30	1.0	56.4
No DP	CIFAR10	Alexnet	30	-	57.6
DP-SGLD	Pneumonia	Resnet18	50	1.0	58.8
DP-SGD	Pneumonia	Resnet18	50	1.0	58.8
No DP	Pneumonia	Resnet18	50	-	59.3

Table 3. Accuracy (in %) when fine-tuning ResNet18.

Method	Dataset	Model	Epochs	ϵ	Acc.
DP-SGD	CIFAR10	Resnet18	30	1.00	67.8
No DP	CIFAR10	Resnet18	30	-	77.0

Table 4. Value of some parameters used for DP-SGLD.

Dataset	Model	β	$\frac{\beta}{\lambda}$	$\frac{\varepsilon n^2}{\alpha d} / \left(\frac{\beta}{\lambda}\right)^2$
CIFAR10	Resnet18	55	5.5×10^4	3.2×10^{-3}
CIFAR10	Alexnet	259	2.6×10^5	1.5×10^{-4}
Pneumonia	Resnet18	354	7.1×10^4	6.8×10^{-5}

Acknowledgments

We would like to thank Pierre Tholoniat for the helpful discussions throughout this project. We are also grateful for the long-standing support of the OpenMined community and in particular its dedicated cryptography team.

This work was supported in part by the French-German Project CRYPTO4GRAPH-AI and by PRAIRIE, the PaRis Artificial Intelligence Research InstitutE.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Böhning, D. Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, 44(1): 197–200, 1992.

Chourasia, R., Ye, J., and Shokri, R. Differential privacy dynamics of langevin diffusion and noisy gradient descent. *Advances in Neural Information Processing Systems*, 2021.

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

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Gross, L. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5): 1122–1131, 2018.

Knott, B., Venkataraman, S., Hannun, A., Sengupta, S., Ibrahim, M., and van der Maaten, L. Crypten: Secure multi-party computation meets machine learning. *Advances in Neural Information Processing Systems*, 2021.

Krishnapuram, B., Carin, L., Figueiredo, M. A., and Hartemink, A. J. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (6):957–968, 2005.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

Krizhevsky, A., Nair, V., and Hinton, G. The CIFAR-10 dataset. *online:* `cs.toronto.edu/~kriz/cifar.html`, 55, 2014.

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

Mitrinovic, D. S. and Vasic, P. M. *Analytic Inequalities*, volume 1. Springer, 1970.

Roux, N., Schmidt, M., and Bach, F. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems*, pp. 3167–3175, 2012.

Ryffel, T., Tholoniat, P., Pointcheval, D., and Bach, F. Ariann: Low-interaction privacy-preserving deep learning via function secret sharing. *Proceedings on Privacy Enhancing Technologies*, 2022.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *Symposium on Security and Privacy (SP)*, pp. 3–18, 2017.

Tramèr, F. and Boneh, D. Differentially private learning needs better features (or much more data). In *9th International Conference on Learning Representations, ICLR*, 2021.

Wagh, S., Tople, S., Benhamouda, F., Kushilevitz, E., Mittal, P., and Rabin, T. FALCON: Honest-Majority Maliciously Secure Framework for Private Deep Learning. *Proceedings on Privacy Enhancing Technologies*, 2021.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the International conference on machine learning*, pp. 681–688. Citeseer, 2011.

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., and Mironov, I. Opacus: User-friendly differential privacy library in pytorch. *arXiv preprint arXiv:2109.12298*, 2021.

A. (Proofs) Privacy analysis of noisy stochastic gradient descent

Lemma 3.3 (Sensitivity)**.** Let $\ell(\theta, \mathbf{x})$ be an L-Lipschitz loss function on closed convex set $\mathcal{C}$, then for coupled tracing diffusions $\{\Theta_t\}_{t\geq 0}$ and $\{\Theta'_t\}_{t\geq 0}$ for DP-SGLD on neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, and noise variance σ^2, the gradient sensitivity of conditionally expected loss $\nabla\mathcal{L}_{\mathcal{B}_k}(\theta) = \mathbb{E}_{\theta_k\sim p_{t_k|t}}\left[\frac{\nabla\ell(\theta_k, \mathbf{x}_{i_0})}{m}|\theta, \mathcal{B}_k\right]$ is bounded by:

$$\left\|\mathop{\mathbb{E}}_{\theta_k\sim p_{t_k|t}}\left[\frac{\nabla\ell(\theta_k, \mathbf{x}_{i_0})}{m}|\theta, \mathcal{B}_k\right] - \mathop{\mathbb{E}}_{\theta'_k\sim p'_{t_k|t}}\left[\frac{\nabla\ell(\theta'_k, \mathbf{x}'_{i_0})}{m}|\theta, \mathcal{B}_k\right]\right\|_2 \leq \frac{2L}{n},$$

where n is the size of the dataset $\mathcal{D}$ and $\mathcal{D}'$, and m is the size of the batch $\mathcal{B}_k$ and $\mathcal{B}'_k$.

Proof of 3.3. We first distinguish on the events $\mathcal{B}_k = \mathcal{B}'_k$ and $\mathcal{B}_k \neq \mathcal{B}'_k$ and then use the triangle inequality:

$$\begin{aligned}
&\Bigg\| \mathbb{E}_{\theta_k\sim p_{t_k|t}}\left[\frac{\nabla\ell(\theta_k; \mathbf{x}_{i_o})}{m}|\theta, \mathcal{B}_k\right] \\
&\qquad - \mathbb{E}_{\theta'_k\sim p'_{t_k|t}}\left[\frac{\nabla\ell(\theta'_k; \mathbf{x'}_{i_o})}{m}|\theta, \mathcal{B}_k\right]\Bigg\|_2 \\
&= \mathbb{P}[\mathcal{B}_k \neq \mathcal{B}'_k] \cdot \Bigg\| \mathbb{E}\left[\mathop{\mathbb{E}}_{\theta_k\sim p_{t_k|t}}\left[\frac{\nabla\ell(\theta_k; \mathbf{x}_{i_o})}{m}|\theta\right]\right. \\
&\qquad \left. - \mathop{\mathbb{E}}_{\theta'_k\sim p'_{t_k|t}}\left[\frac{\nabla\ell(\theta'_k; \mathbf{x'}_{i_o})}{m}|\theta\right] \mid \mathcal{B}_k \neq \mathcal{B}'_k\right]\Bigg\|_2 \\
&\leq \frac{m}{n}\cdot\frac{2L}{m} \leq \frac{2L}{n}
\end{aligned}$$

□

Theorem 3.1 (RDP for DP-SGLD under c-LSI)**.** Let $\ell(\theta, \mathbf{x})$ be an L-Lipschitz loss function on a closed convex set $\mathcal{C}$. Let $\{\Theta_t\}_{t\geq 0}$ be the tracing diffusion for $\mathcal{A}_{\mathrm{DP-SGLD}}$ under loss function $\ell(\theta, \mathbf{x})$ on dataset $\mathcal{D}$. If Θ_t satisfies c-LSI throughout $0 \leq t \leq \sum_{k=1}^K \eta_k$, then $\mathcal{A}_{\mathrm{DP-SGLD}}$ satisfies (α, ε)-Rényi Differential Privacy for

$$\varepsilon = \frac{2\alpha L^2}{cn^2\sigma^4}(1 - e^{-\sigma^2 c\sum_{k=1}^K \eta_k}).$$

Proof of Theorem 3.1. At each step k, the Rényi privacy follows the evolution equation defined in (5) (since it is a tracing diffusion) when $t_k < t < t_{k+1}$, where $t_k = \sum_{i=1}^k \eta_i$:

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

Let's $a_1 = 2(1-\gamma)\sigma^2 c$, $a_2 = \frac{1}{\gamma}\frac{\alpha L^2}{\sigma^2 n^2}$ and

$$u(t,y) = \begin{cases} \frac{R(\alpha,t)}{\alpha} - \frac{a_2}{a_1} & t_k < t < t_{k+1} \\ \frac{R(\alpha, \lim_{t\to t_k^+} t)}{\alpha} - \frac{a_2}{a_1} & t = t_k \end{cases}$$

We can rewrite the evolution equation as such:

$$\frac{\partial u}{\partial t} + a_1 u + a_1 \frac{\partial u}{\partial y} \leq 0 \ , \ t_k < t < t_{k+1}$$

with initial condition $u(t_k, y) = \frac{R(\alpha, \lim_{t\to t_k^+} t)}{\alpha} - \frac{a_2}{a_1}$.

Let now introduce $\tau = t$ and $z = t - \frac{1}{a_1} y$ and write $v(\tau, z) = u(t, y)$. The equation now writes:

$$\frac{\partial v}{\partial \tau} + a_1 v < 0$$

with initial condition $v(t_k, z) = u(t_k, -a_1(z - t_k))$

The solution of this equation is:

$$v(\tau, z) \leq v(t_k, z) e^{-a_1(\tau - t_k)}$$

which also writes

$$u(t,y) \leq u(t_k, y - a_1(t - t_k)) e^{-a_1(t-t_k)}$$

or in terms of $R(\alpha, t)$:

$$R(\alpha,t) - \frac{a_2}{a_1}\alpha \leq (R(\alpha, \lim_{t\to t_k^+}) - \frac{a_2}{a_1}\alpha) e^{-a_1(t-t_k)}$$

for $t_k < t < t_{k+1}$.

Taking the limit $t \to t_{k+1}^-$, we have

$$R(\alpha, \lim_{t\to t_{k+1}^-} t) - \frac{a_2}{a_1}\alpha \leq (R(\alpha, \lim_{t\to t_k^+} t) - \frac{a_2}{a_1}\alpha) e^{-a_1 \eta_{k+1}}$$

In addition, by the tracing diffusion process described in (2), we have

$$\lim_{t\to t_k^+} \Theta_t = \phi(\Pi_{\mathcal{C}}(\lim_{t\to t_k^-} \Theta_t))$$
$$\lim_{t\to t_k^+} \Theta'_t = \phi(\Pi_{\mathcal{C}}(\lim_{t\to t_k^-} \Theta'_t))$$

where $\phi(\theta) = \theta - \eta \sum_{i=1, i\neq i_0}^m \frac{\nabla \ell(\theta, \mathbf{x}_i)}{m}$ is a mapping which is the same for both processes, as the batches $\mathcal{B}_k$ for each distribution only differ at most on $\mathbf{x}_{i_0}$. Hence, by post-processing of Rényi divergence, we have

$$R(\alpha, \lim_{t\to t_k^+} t) \leq R(\alpha, \lim_{t\to t_k^-} t)$$

Combining the above two inequalities, we derive the following recursive equation

$$R(\alpha, \lim_{t\to t_{k+1}^-} t) - \frac{a_2}{a_1}\alpha \leq (R(\alpha, \lim_{t\to t_k^-}) - \frac{a_2}{a_1}\alpha) e^{-a_1 \eta_{k+1}}$$

Repeating this step for $k = 0, \ldots, K-1$ we have

$$R(\alpha, \lim_{t\to t_K^-} t) - \frac{a_2}{a_1}\alpha \leq (R(\alpha, \lim_{t\to 0^-} t) - \frac{a_2}{a_1}\alpha) e^{-a_1 t_K}$$

where $t_K = \sum_{i=1}^K \eta_i$. Because coupled tracing diffusions have the same start parameter, we have $R(\alpha, \lim_{t\to 0^-} t) = 0$. Moreover, since projection is a post-processing mapping we have $R(\alpha, t_K) \leq R(\alpha, \lim_{t\to t_K^-} t)$. Putting back the values of a_1 and a_2 we have:

$$R(\alpha, t_K) \leq \frac{\alpha L^2}{2\gamma(1-\gamma)c\sigma^4 n^2}(1 - e^{-2(1-\gamma)\sigma^2 c t_K})$$

Setting $\gamma = \frac{1}{2}$ suffices to prove the Rényi privacy loss bound in the theorem.

□

B. (Proofs) Utility analysis for noisy stochastic gradient descent

B.1. (Proofs) Fixed learning rate η

Lemma 4.1 (Empirical risk for smooth and strongly convex loss)**.** Let $\ell(\theta, \mathbf{x})$ be an L-Lipschitz, λ-strongly convex and β-smooth loss function on closed convex set $\mathcal{C}$, $\mathcal{A}_{\mathrm{DP-SGLD}}$ be parameterized with step-size $\eta < \frac{1}{2\beta}$ and start parameter $\theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\mathbf{I}_d))$, then the empirical risk of $\mathcal{A}_{\mathrm{DP-SGLD}}$ is bounded by

$$\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K) - \mathcal{L}_{\mathcal{D}}(\theta^*)] \leq \frac{\beta}{2}\mathbb{E}[\|\theta_0 - \theta^*\|_2^2] e^{-\lambda\eta K} + \frac{\beta\eta\xi^2}{2\lambda} + \frac{\beta d\sigma^2}{\lambda}, \quad (7)$$

where θ^* is the minimizer of $\mathcal{L}_{\mathcal{D}}(\theta)$ in $\mathcal{C}$ and $\xi^2 = \mathbb{E}[\|\nabla \mathcal{L}_{\mathcal{B}_k}(\theta^*)\|_2^2]$.

Proof of 4.1. Let's recall the noisy SGD update equation:

$$\theta_{k+1} = \Pi_{\mathcal{C}}(\theta_k - \eta \nabla \mathcal{L}_{\mathcal{B}_k}(\theta_k) + \sqrt{2\eta\sigma^2}\mathcal{N}(0, \mathbf{I}_d))$$

From the definitions of $\Pi_{\mathcal{C}}(\cdot)$ and θ^*, we have:

$$\Pi_{\mathcal{C}}(\theta^*) = \theta^*$$

Combining this facts and using the contractivity of the pro-

jection $\Pi_{\mathcal{C}}(\cdot)$, we derive:

$$\begin{aligned}
&\|\theta_{k+1}-\theta^*\|_2^2 \\
&= \left\|\Pi_{\mathcal{C}}(\theta_k-\eta\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)+\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbf{I}_d))-\theta^*\right\|_2^2 \\
&= \left\|\Pi_{\mathcal{C}}(\theta_k-\eta\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)+\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbf{I}_d))-\Pi_{\mathcal{C}}(\theta^*)\right\|_2^2 \\
&\leq \left\|\theta_k-\eta\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)+\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbf{I}_d)-\theta^*\right\|_2^2 \\
&= \left\|\theta_k-\theta^*-\eta\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)+\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbf{I}_d)\right\|_2^2 \\
&= \|\theta_k-\theta^*\|_2^2+\eta^2\left\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)\right\|_2^2+2\eta\sigma^2\left\|\mathcal{N}(0,\mathbf{I}_d)\right\|_2^2 \\
&\quad +2\langle\theta_k-\theta^*,\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbf{I}_d)\rangle \\
&\quad -2\eta\langle\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k),\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbf{I}_d)\rangle \\
&\quad -2\eta\langle\theta_k-\theta^*,\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)\rangle
\end{aligned}$$

We now take the expectation with respect to the random variable $\mathcal{B}_k$ sampled from $\mathcal{D}$, and we note that $\mathbb{E}_{\mathcal{D}}[\mathcal{L}_{\mathcal{B}_k}(\theta_k)] = \mathcal{L}_{\mathcal{D}}(\theta_k)$ since $\mathcal{L}_{\mathcal{B}_k}$ is an unbiased estimate of $\mathcal{L}_{\mathcal{D}}$.

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\|\theta_{k+1}-\theta^*\|_2^2] \\
&= \|\theta_k-\theta^*\|_2^2+\eta^2\,\mathbb{E}_{\mathcal{D}}[\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)\|_2^2] \\
&\quad +2\eta\sigma^2\,\mathbb{E}_{\mathcal{D}}[\|\mathcal{N}(0,\mathbf{I}_d)\|_2^2] \\
&\quad +2\langle\theta_k-\theta^*,\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbf{I}_d)\rangle \\
&\quad -2\eta\langle\nabla\mathcal{L}_{\mathcal{D}}(\theta_k),\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbf{I}_d)\rangle \\
&\quad -2\eta\langle\theta_k-\theta^*,\nabla\mathcal{L}_{\mathcal{D}}(\theta_k)\rangle
\end{aligned}$$

Using the classic inequality $(a+b)^2 \leq 2(a^2+b^2)$, we derive:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)\|_2^2] \\
&= \mathbb{E}_{\mathcal{D}}[\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)-\nabla\mathcal{L}_{\mathcal{B}_k}(\theta^*)+\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_*)\|_2^2] \\
&\leq 2\,\mathbb{E}_{\mathcal{D}}[\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)-\nabla\mathcal{L}_{\mathcal{B}_k}(\theta^*)\|_2^2]+2\,\mathbb{E}_{\mathcal{D}}[\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta^*)\|_2^2]
\end{aligned}$$

Furthermore, by β-smoothness of $\nabla\mathcal{L}_{\mathcal{B}_k}$, we can use the property of co-coercivity of the gradients:

$$\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)-\nabla\mathcal{L}_{\mathcal{B}_k}(\theta^*)\|_2^2 \leq \beta\langle\theta_k-\theta^*,\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)-\nabla\mathcal{L}_{\mathcal{B}_k}(\theta^*)\rangle \tag{9}$$

Taking expectation over $\mathcal{B}_k$ and using optimality of θ^*:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)-\nabla\mathcal{L}_{\mathcal{B}_k}(\theta^*)\|_2^2] \\
&\leq \beta\langle\theta_k-\theta^*,\nabla\mathcal{L}_{\mathcal{D}}(\theta_k)-\nabla\mathcal{L}_{\mathcal{D}}(\theta^*)\rangle \\
&= \beta\langle\theta_k-\theta^*,\nabla\mathcal{L}_{\mathcal{D}}(\theta_k)\rangle
\end{aligned} \tag{10}$$

Combining these elements, we derive:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\|\theta_{k+1}-\theta^*\|_2^2] \\
&= \|\theta_k-\theta^*\|_2^2+2\eta^2\,\mathbb{E}_{\mathcal{D}}[\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta_k)-\nabla\mathcal{L}_{\mathcal{B}_k}(\theta^*)\|_2^2] \\
&\quad +2\eta^2\,\mathbb{E}_{\mathcal{D}}[\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta^*)\|_2^2]+2\eta\sigma^2\,\mathbb{E}_{\mathcal{D}}[\|\mathcal{N}(0,\mathbf{I}_d)\|_2^2] \\
&\quad +2\langle\theta_k-\theta^*,\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbf{I}_d)\rangle \\
&\quad -2\eta\langle\nabla\mathcal{L}_{\mathcal{D}}(\theta_k),\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbf{I}_d)\rangle \\
&\quad -2\eta\langle\theta_k-\theta^*,\nabla\mathcal{L}_{\mathcal{D}}(\theta_k)\rangle \\
&\leq \|\theta_k-\theta^*\|_2^2+2\eta(\eta\beta-1)\langle\theta_k-\theta^*,\nabla\mathcal{L}_{\mathcal{D}}(\theta_k)\rangle \\
&\quad +2\eta^2\,\mathbb{E}_{\mathcal{D}}[\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta^*)\|_2^2]+2\eta\sigma^2\,\mathbb{E}_{\mathcal{D}}[\|\mathcal{N}(0,\mathbf{I}_d)\|_2^2] \\
&\quad +2\langle\theta_k-\theta^*,\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbf{I}_d)\rangle \\
&\quad -2\eta\langle\nabla\mathcal{L}_{\mathcal{D}}(\theta_k),\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbf{I}_d)\rangle \\
&\quad \text{using (9) and (10)}
\end{aligned}$$

Let's assume that $\eta \leq \frac{1}{2\beta}$. Therefore, $2\eta(\eta\beta-1) \leq -\eta$. By optimality of θ^* and strong convexity arguments on $\mathcal{L}_{\mathcal{D}}$, we deduce the following inequality:

$$\begin{aligned}
-\langle\theta_k-\theta^*,\nabla\mathcal{L}_{\mathcal{D}}(\theta_k)\rangle &\leq \mathcal{L}_{\mathcal{D}}(\theta_k)-\mathcal{L}_{\mathcal{D}}(\theta^*)+\frac{\lambda}{2}\|\theta_k-\theta^*\|_2^2 \\
&\leq \lambda\,\|\theta_k-\theta^*\|_2^2
\end{aligned}$$

Plugging this together, we have:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\|\theta_{k+1}-\theta^*\|_2^2] \\
&\leq (1-\eta\lambda)\,\|\theta_k-\theta^*\|_2^2+\eta^2\,\mathbb{E}_{\mathcal{D}}[\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta^*)\|_2^2] \\
&\quad +2\eta\sigma^2\,\mathbb{E}_{\mathcal{D}}[\|\mathcal{N}(0,\mathbf{I}_d)\|_2^2] \\
&\quad +2\langle\theta_k-\theta^*,\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbf{I}_d)\rangle \\
&\quad -2\eta\langle\nabla\mathcal{L}_{\mathcal{D}}(\theta_k),\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbf{I}_d)\rangle
\end{aligned}$$

We take the expectation again:

$$\begin{aligned}
\mathbb{E}[\|\theta_{k+1}-\theta^*\|_2^2] &\leq (1-\eta\lambda)\,\mathbb{E}[\|\theta_k-\theta^*\|_2^2] \\
&\quad +\eta^2\,\mathbb{E}[\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta^*)\|_2^2]+2\eta d\sigma^2
\end{aligned} \tag{11}$$

Now, by β-smoothness of $\mathcal{L}_{\mathcal{D}}(\cdot)$, we have:

$$\mathcal{L}_{\mathcal{D}}(\theta_K)-\mathcal{L}_{\mathcal{D}}(\theta^*) \leq \langle\nabla\mathcal{L}_{\mathcal{D}}(\theta^*),\theta_K-\theta^*\rangle+\frac{\beta}{2}\,\|\theta_K-\theta^*\|_2^2$$

Second, by optimality of θ^* in $\mathcal{C}$ and the fact that $\theta_K \in \mathcal{C}$, we have

$$\langle\nabla\mathcal{L}_{\mathcal{D}}(\theta^*),\theta_K-\theta^*\rangle = 0$$

Combining these two we get:

$$\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K)-\mathcal{L}_{\mathcal{D}}(\theta^*)] \leq \frac{\beta}{2}\,\mathbb{E}[\|\theta_K-\theta^*\|_2^2]$$

Using the recursive equation (11) repeatedly for $k =$

$0, \dots, K-1$, we have:

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K) - \mathcal{L}_{\mathcal{D}}(\theta^*)] \\
&\leq \frac{\beta}{2}(1-\eta\lambda)^K \, \mathbb{E}[\|\theta_0 - \theta^*\|_2^2] \\
&\qquad + \frac{\beta}{2}(\eta^2\xi^2 + 2\eta d\sigma^2) \textstyle\sum_{k=0}^{K-1}(1-\eta\lambda)^k \\
&\leq \frac{\beta}{2} e^{-\lambda\eta K} \, \mathbb{E}[\|\theta_0 - \theta^*\|_2^2] + \frac{\beta\eta\xi^2}{2\lambda} + \frac{\beta d\sigma^2}{\lambda} \\
&\text{where } \xi^2 = \mathbb{E}[\|\nabla \mathcal{L}_{\mathcal{B}_k}(\theta^*)\|_2^2]
\end{aligned}$$

□

Theorem 4.1 (Utility bound for (α, ε)-Rényi differential privacy)**.** Let $\ell(\theta, \mathbf{x})$ be an L-Lipschitz, λ-strongly convex and β-smooth loss function on closed convex set $\mathcal{C}$, then $\mathcal{A}_{\mathrm{DP-SGLD}}$ with start parameter $\theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\mathbf{I}_d))$ and constant step-size $\eta = \frac{1}{2\beta}$, satisfies (α, ε) Rényi differential privacy and

$$\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K) - \mathcal{L}_{\mathcal{D}}(\theta^*)] = O\left(\frac{\alpha\beta d L^2}{\varepsilon\lambda^2 n^2}\right) + \frac{\xi^2}{4\lambda},$$

where σ^2 and K are set as such:

$$\sigma^2 = \frac{4\alpha L^2}{\varepsilon\lambda n^2}, \quad K = \frac{2\beta}{\lambda}\log\left(\frac{\varepsilon n^2}{\alpha d}\right).$$

Proof of Theorem 4.1. First, as $\theta_0, \theta^* \in \mathcal{C}$ and $\|C\|_2 \leq \frac{2L}{\lambda}$, we have $\mathbb{E}[\|\theta_0 - \theta^*\|_2^2] \leq \frac{4L^2}{\lambda^2}$.

Then, we plug the values of σ^2 and K in Lemma 4.1:

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K) - \mathcal{L}_{\mathcal{D}}(\theta^*)] &\leq \frac{\beta}{2}\frac{4L^2}{\lambda^2} e^{-\lambda\eta K} \\
&\quad + \frac{\beta d\sigma^2}{\lambda} + \frac{\beta\eta\xi^2}{2\lambda} \\
&\leq \frac{2\beta L^2}{\lambda^2}\frac{\alpha d}{\varepsilon n^2} \\
&\quad + \frac{\beta d}{\lambda}\frac{4\alpha L^2}{\lambda\varepsilon n^2} + \frac{\beta\eta\xi^2}{2\lambda} \\
&\leq \frac{6\alpha\beta d L^2}{\varepsilon\lambda^2 n^2} + \frac{\xi^2}{4\lambda} \text{ with } \eta = \frac{1}{2\beta}
\end{aligned}$$

□

Theorem 4.2 (Utility bound for (ϵ, δ)-differential privacy)**.** With the same conditions as is Theorem 4.1, for $\epsilon \leq 2\log(1/\delta)$ and $\delta > 0$, $\mathcal{A}_{\mathrm{DP-SGLD}}$ satisfies (ϵ, δ) differential privacy and

$$\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K) - \mathcal{L}_{\mathcal{D}}(\theta^*)] = O\left(\frac{\beta d L^2 \log(1/\delta)}{\epsilon^2\lambda^2 n^2}\right) + \frac{\xi^2}{4\lambda},$$

where σ^2 and K are set as such:

$$\sigma^2 = \frac{8L^2(\epsilon + 2\log(1/\delta)}{\epsilon^2\lambda n^2}$$
$$K = \frac{2\beta}{\lambda}\log\left(\frac{\epsilon^2 n^2}{4\log(1/\delta)d}\right).$$

Proof of Theorem 4.2. By setting $\varepsilon = \frac{\epsilon}{2}$, we derive from Proposition 2.1 :

$$\alpha = 1 + \frac{2}{\epsilon}\log(1/\delta)$$

.

We use this to rewrite the results from Theorem 4.1:

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K) - \mathcal{L}_{\mathcal{D}}(\theta^*)] &\leq \frac{6\alpha\beta d L^2}{\varepsilon\lambda^2 n^2} + \frac{\xi^2}{4\lambda} \\
&= \frac{6\beta d L^2}{\lambda^2 n^2}\frac{1 + \frac{2}{\epsilon}\log(1/\delta)}{\frac{\epsilon}{2}} + \frac{\xi^2}{4\lambda} \\
&= \frac{6\beta d L^2}{\lambda^2 n^2}\frac{2\epsilon + 4\log(1/\delta)}{\epsilon^2} + \frac{\xi^2}{4\lambda} \\
&\leq \frac{6\beta d L^2}{\lambda^2 n^2}\frac{8\log(1/\delta)}{\epsilon^2} + \frac{\xi^2}{4\lambda} \\
&\quad \text{using } \epsilon \leq 2\log(1/\delta)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\sigma^2 &= \frac{4L^2\alpha}{\lambda n^2\varepsilon} \\
&= \frac{4L^2}{\lambda n^2}\frac{1 + \frac{2}{\epsilon}\log(1/\delta)}{\frac{\epsilon}{2}} \\
&= \frac{8L^2(\epsilon + 2\log(1/\delta)}{\epsilon^2\lambda n^2}
\end{aligned}$$

$$\begin{aligned}
K &= \frac{2\beta}{\lambda}\log\left(\frac{n^2}{d}\frac{\varepsilon}{\alpha}\right) \\
&= \frac{2\beta}{\lambda}\log\left(\frac{n^2}{d}\frac{\frac{\epsilon}{2}}{1 + \frac{2}{\epsilon}\log(1/\delta)}\right) \\
&= \frac{2\beta}{\lambda}\log\left(\frac{n^2}{d}\frac{\epsilon^2}{2\epsilon + 4\log(1/\delta)}\right) \\
&\leq \frac{2\beta}{\lambda}\log\left(\frac{\epsilon^2 n^2}{4\log(1/\delta)d}\right)
\end{aligned}$$

□

B.2. (Proofs) Decreasing step size η_k

Lemma 4.3 (Empirical risk for smooth and strongly convex loss with decreasing learning rate)**.** Let $\ell(\theta, \mathbf{x})$ be an L-Lipschitz, λ-strongly convex and β-smooth loss function on closed convex set $\mathcal{C}$, $\mathcal{A}_{\mathrm{DP-SGLD}}$ be parameterized with *decreasing* step-size $\eta_k = \frac{1}{2\beta + \lambda k/2}$ and start parameter $\theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\mathbf{I}_d))$, then the average empirical risk of $\mathcal{A}_{\mathrm{DP-SGLD}}$ is bounded by

$$\begin{aligned}
\mathbb{E}[\frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\mathcal{D}}(\theta_k) - \mathcal{L}_{\mathcal{D}}(\theta^*)] &\leq \frac{2\beta}{K}\,\mathbb{E}[\|\theta_0 - \theta^*\|_2^2] \\
&\quad + \frac{4\xi^2}{K\lambda}\log\left(1 + \frac{\lambda K}{4\beta}\right) + 2d\sigma^2, \quad (8)
\end{aligned}$$

where θ^* is the minimizer of $\mathcal{L}_{\mathcal{D}}(\theta)$ in $\mathcal{C}$.

Proof of 4.3. We define η_k as such

$$\eta_k = \frac{1}{2\beta + \frac{\lambda k}{2}}.$$

Note that in particular that for $k \geq 0$, $\eta_k \leq \frac{1}{2\beta}$ and hence $2\eta_k(\eta_k\beta - 1) \leq -\eta_k$.

As detailed in the proof of Lemma 4.1 we have the following:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\|\theta_{k+1} - \theta^*\|_2^2] \\
&\leq \|\theta_k - \theta^*\|_2^2 + 2\eta_{k+1}(\eta_{k+1}\beta - 1)\langle \theta_k - \theta^*, \nabla\mathcal{L}_{\mathcal{D}}(\theta_k)\rangle \\
&\quad + 2\eta_{k+1}^2\, \mathbb{E}_{\mathcal{D}}[\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta^*)\|_2^2] + 2\eta_{k+1}\sigma^2\, \mathbb{E}_{\mathcal{D}}[\|\mathcal{N}(0, \mathbf{I}_d)\|_2^2] \\
&\quad + 2\langle \theta_k - \theta^*, \sqrt{2\eta_{k+1}\sigma^2}\mathcal{N}(0, \mathbf{I}_d)\rangle \\
&\quad - 2\eta_{k+1}\langle \nabla\mathcal{L}_{\mathcal{D}}(\theta_k), \sqrt{2\eta_{k+1}\sigma^2}\mathcal{N}(0, \mathbf{I}_d)\rangle \\
&\quad \text{using (9) and (10)}
\end{aligned}$$

By taking expectation again we derive

$$\begin{aligned}
\mathbb{E}[\|\theta_{k+1} - \theta^*\|_2^2] &\leq \mathbb{E}[\|\theta_k - \theta^*\|_2^2] \\
&\quad + 2\eta_{k+1}(\eta_{k+1}\beta - 1)\, \mathbb{E}[\langle \nabla\mathcal{L}_{\mathcal{D}}(\theta_k), \theta_k - \theta^*\rangle] \\
&\quad + 2\eta_{k+1}^2\, \mathbb{E}[\|\nabla\mathcal{L}_{\mathcal{B}_k}(\theta^*)\|_2^2] + 2\eta_{k+1}\sigma^2 d \\
&\leq \mathbb{E}[\|\theta_k - \theta^*\|_2^2] \\
&\quad - \eta_{k+1}\, \mathbb{E}[\langle \nabla\mathcal{L}_{\mathcal{D}}(\theta_k), \theta_k - \theta^*\rangle] \\
&\quad + 2\eta_{k+1}(\eta_{k+1}\xi^2 + d\sigma^2)
\end{aligned}$$

By optimality of θ^* and strong convexity arguments on $\mathcal{L}_{\mathcal{D}}$, we have

$$-\langle \theta_k - \theta^*, \nabla\mathcal{L}_{\mathcal{D}}(\theta_k)\rangle \leq \mathcal{L}_{\mathcal{D}}(\theta_k) - \mathcal{L}_{\mathcal{D}}(\theta^*) + \frac{\lambda}{2}\|\theta_k - \theta^*\|_2^2$$

Hence we have

$$\begin{aligned}
&\mathbb{E}[\|\theta_{k+1} - \theta^*\|_2^2] \\
&\leq (1 - \frac{\eta_{k+1}\lambda}{2})\, \mathbb{E}[\|\theta_k - \theta^*\|_2^2] - \eta_{k+1}\, \mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_k) - \mathcal{L}_{\mathcal{D}}(\theta^*)] \\
&\quad + 2\eta_{k+1}(\eta_{k+1}\xi^2 + d\sigma^2)
\end{aligned}$$

Equivalently

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_k) - \mathcal{L}_{\mathcal{D}}(\theta^*)] \\
&\leq -\frac{1}{\eta_{k+1}}\, \mathbb{E}[\|\theta_{k+1} - \theta^*\|_2^2] + (\frac{1}{\eta_{k+1}} - \frac{\lambda}{2})\, \mathbb{E}[\|\theta_k - \theta^*\|_2^2] \\
&\quad + 2(\eta_{k+1}\xi^2 + d\sigma^2)
\end{aligned}$$

Plugging in the definition of η_k we have

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_k) - \mathcal{L}_{\mathcal{D}}(\theta^*)] \\
&\leq -(2\beta + \frac{\lambda}{2}(k+1))\, \mathbb{E}[\|\theta_{k+1} - \theta^*\|_2^2] \\
&\quad + (2\beta + \frac{\lambda}{2}k)\, \mathbb{E}[\|\theta_k - \theta^*\|_2^2] \\
&\quad + 2(\eta_{k+1}\xi^2 + d\sigma^2)
\end{aligned}$$

We derive the average empirical risk

$$\begin{aligned}
&\mathbb{E}[\tfrac{1}{K}\textstyle\sum_{k=0}^{K-1} \mathcal{L}_{\mathcal{D}}(\theta_k) - \mathcal{L}_{\mathcal{D}}(\theta^*)] \\
&= \frac{1}{K}\textstyle\sum_{k=0}^{K-1} \mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_k) - \mathcal{L}_{\mathcal{D}}(\theta^*)] \\
&= -\frac{1}{K}(2\beta + \frac{\lambda}{2}K)\, \mathbb{E}[\|\theta_K - \theta^*\|_2^2] \\
&\quad + \frac{2\beta}{K}\, \mathbb{E}[\|\theta_0 - \theta^*\|_2^2] + 2d\sigma^2 + \frac{2\xi^2}{K}\sum_{k=1}^{K}\eta_k \\
&\leq \frac{2\beta}{K}\, \mathbb{E}[\|\theta_0 - \theta^*\|_2^2] + 2d\sigma^2 + \frac{2\xi^2}{K}\sum_{k=1}^{K}\frac{1}{2\beta + \frac{\lambda k}{2}} \\
&\leq \frac{2\beta}{K}\, \mathbb{E}[\|\theta_0 - \theta^*\|_2^2] + 2d\sigma^2 + \frac{2\xi^2}{K}\int_0^K \frac{1}{2\beta + \frac{\lambda t}{2}}dt \\
&\leq \frac{2\beta}{K}\, \mathbb{E}[\|\theta_0 - \theta^*\|_2^2] + 2d\sigma^2 + \frac{4\xi^2}{K\lambda}\log\left(1 + \frac{\lambda K}{4\beta}\right)
\end{aligned}$$

□

Theorem 4.3 (Utility bound for (α, ε)-Rényi differential privacy with decreasing learning rate)**.** Let $\ell(\theta, \mathbf{x})$ be an L-Lipschitz, λ-strongly convex and β-smooth loss function on closed convex set $\mathcal{C}$, then $\mathcal{A}_{\mathrm{DP-SGLD}}$ with start parameter $\theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\mathbf{I}_d))$ and *decreasing* step-size $\eta_k = \frac{1}{2\beta + \lambda k/2}$, satisfies (α, ε) Rényi differential privacy and

$$\mathbb{E}[\frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\mathcal{D}}(\theta_k) - \mathcal{L}_{\mathcal{D}}(\theta^*)] = O\left(\frac{\alpha d L^2}{\varepsilon\lambda n^2}\right),$$

where σ^2 and K are set as such:

$$\sigma^2 = \frac{4\alpha L^2}{\varepsilon\lambda n^2}, \quad K = \max\left(\frac{\beta}{\lambda}\frac{\varepsilon n^2}{\alpha d}, \frac{\lambda}{\beta}\left(\frac{\varepsilon n^2}{\alpha d}\right)^2\right).$$

Proof of Theorem 4.3. Since $\theta_0, \theta^* \in \mathcal{C}$ and $\|C\|_2 \leq \frac{2L}{\lambda}$, we have $\mathbb{E}[\|\theta_0 - \theta^*\|_2^2] \leq \frac{4L^2}{\lambda^2}$. Additionally, by Lipschitzness of the loss we have $\xi^2 \leq L^2$.

We can use this and plug the value of σ^2 to rewrite Lemma 4.3:

$$\begin{aligned}
\mathbb{E}[\tfrac{1}{K}\textstyle\sum_{k=1}^{K}\mathcal{L}_{\mathcal{D}}(\theta_k) - \mathcal{L}_{\mathcal{D}}(\theta^*)] &\leq \frac{8\alpha d L^2}{\varepsilon\lambda n^2} + \frac{8\beta L^2}{K\lambda^2} \\
&\quad + \frac{L^2}{\beta}\frac{4\beta}{K\lambda}\log(1 + \frac{K\lambda}{4\beta})
\end{aligned}$$

Then, we use the following bound on the logarithmic function from (Mitrinovic & Vasic, 1970):

$$\forall x > -1, \frac{1}{x}\log(1 + x) \leq \frac{1}{\sqrt{1 + x}}$$

to derive:

$$\frac{L^2}{\beta}\frac{4\beta}{K\lambda}\log(1+\frac{K\lambda}{4\beta}) \leq \frac{L^2}{\beta}\frac{1}{\sqrt{1+\frac{K\lambda}{4\beta}}}$$

We can now plug the value of K.

$$\begin{aligned}\mathbb{E}[\tfrac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\mathcal{D}}(\theta_k)-\mathcal{L}_{\mathcal{D}}(\theta^*)] &\leq \frac{8\alpha d L^2}{\varepsilon\lambda n^2}+\frac{8\beta L^2}{K\lambda^2} \\ &\quad +\frac{L^2}{\beta}\sqrt{\frac{4\beta}{\lambda}\frac{\beta}{\lambda}\left(\frac{\alpha d}{\varepsilon n^2}\right)^2} \\ &\leq \frac{18\alpha d L^2}{\varepsilon\lambda n^2}\end{aligned}$$

□

Theorem 4.4 (Utility bound for (ϵ,δ)-differential privacy with decreasing learning rate)**.** With the same conditions as is Theorem 4.3, for $\epsilon \leq 2\log(1/\delta)$ and $\delta > 0$, $\mathcal{A}_{\mathrm{DP-SGLD}}$ satisfies (ϵ,δ) differential privacy and

$$\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_K)-\mathcal{L}_{\mathcal{D}}(\theta^*)] = O\left(\frac{dL^2\log(1/\delta)}{\epsilon^2\lambda n^2}\right),$$

where σ^2 and K are set as such:

$$\sigma^2 = \frac{8L^2(\epsilon+2\log(1/\delta)}{\epsilon^2\lambda n^2}$$

$$K = \max\left(\frac{\beta}{\lambda}\frac{\epsilon^2 n^2}{4\log(1/\delta)d}, \frac{\lambda}{\beta}\left(\frac{\epsilon^2 n^2}{4\log(1/\delta)d}\right)^2\right).$$

Proof of Theorem 4.4. By setting $\varepsilon = \frac{\epsilon}{2}$, we derive from Proposition 2.1 :

$$\alpha = 1+\frac{2}{\epsilon}\log(1/\delta)$$

.

We use this to rewrite the results from Theorem 4.3:

$$\begin{aligned}\mathbb{E}[\tfrac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\mathcal{D}}(\theta_k)-\mathcal{L}_{\mathcal{D}}(\theta^*)] &\leq \frac{18\alpha d L^2}{\varepsilon\lambda n^2} \\ &= \frac{18dL^2}{\lambda n^2}\frac{1+\frac{2}{\epsilon}\log(1/\delta)}{\frac{\epsilon}{2}} \\ &= \frac{18dL^2}{\lambda n^2}\frac{2\epsilon+4\log(1/\delta)}{\epsilon^2} \\ &\leq \frac{18dL^2}{\lambda n^2}\frac{8\log(1/\delta)}{\epsilon^2} \\ &\quad \text{using } \epsilon \leq 2\log(1/\delta)\end{aligned}$$

Similarly,

$$\begin{aligned}\sigma^2 &= \frac{4L^2\alpha}{\lambda n^2\varepsilon} \\ &= \frac{4L^2}{\lambda n^2}\frac{1+\frac{2}{\epsilon}\log(1/\delta)}{\frac{\epsilon}{2}} \\ &= \frac{8L^2(\epsilon+2\log(1/\delta)}{\epsilon^2\lambda n^2}\end{aligned}$$

$$\begin{aligned}\frac{\epsilon n^2}{\alpha d} &= \frac{n^2}{d}\frac{\frac{\epsilon}{2}}{1+\frac{2}{\epsilon}\log(1/\delta)} \\ &= \frac{n^2}{d}\frac{\epsilon^2}{2\epsilon+4\log(1/\delta)} \\ &\leq \frac{\epsilon^2 n^2}{4\log(1/\delta)d}\end{aligned}$$

$$\begin{aligned}K &= \max\left(\frac{\beta}{\lambda}\frac{\epsilon n^2}{\alpha d}, \frac{\lambda}{\beta}\left(\frac{\epsilon n^2}{\alpha d}\right)^2\right) \\ &\leq \max\left(\frac{\beta}{\lambda}\frac{\epsilon^2 n^2}{4\log(1/\delta)d}, \frac{\lambda}{\beta}\left(\frac{\epsilon^2 n^2}{4\log(1/\delta)d}\right)^2\right)\end{aligned}$$

□

C. (Proofs) Experiments: application to logistic regression

Lemma 5.1 (Convexity and smoothness constants for regularized logistic regression)**.** Let $\ell(\mathbf{W},\mathbf{x})$ be defined as above. Then ℓ is λ-strongly convex and β-smooth, with

$$\beta = \frac{1}{2}\lambda_{\mathsf{max}}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\Big)+\lambda$$

where λ_{max} refers to the maximum eigenvalue.

Proof of 5.1. By plugging in the definition of the sigmoid, the loss ℓ also writes:

$$\ell(\mathbf{W},\mathbf{x}) = \log\frac{\exp\big((\mathbf{W}\mathbf{x})_y\big)}{\sum_{i=1}^{C}\exp\big((\mathbf{W}\mathbf{x})_i\big)}+\lambda\left\|\mathbf{W}\right\|_2^2,$$

where $\mathbf{W}\in\mathbb{R}^{C\times p}$ also writes $[\mathbf{w}_1^\top,\ldots,\mathbf{w}_C^\top]^\top$.

Let's define for i = 1, . . . , C

$$\begin{aligned}p_i = P(y_i=1|\mathbf{x},\mathbf{W}) &= \frac{\exp\big((\mathbf{W}\mathbf{x})_i\big)}{\sum_{c=1}^{C}\exp\big((\mathbf{W}\mathbf{x})_c\big)} \\ &= \frac{\exp(\mathbf{w}_i^\top\mathbf{x})}{\sum_{c=1}^{C}\exp(\mathbf{w}_c^\top\mathbf{x})}\end{aligned}$$

We can rewrite the loss ℓ as such

$$\ell(\mathbf{W},\mathbf{x}) = \log\prod_{i=1}^{C}p_i^{y_i}+\lambda\left\|\mathbf{W}\right\|_2^2$$

where the label y is now one-hot encoded: $y = (y_0,\ldots,y_C)$.

Following the work of (Böhning, 1992), we have:

$$\nabla^2\ell(\mathbf{W},\mathbf{x}) = (\mathbf{D}(\mathbf{p})-\mathbf{p}\mathbf{p}^\top)\otimes\mathbf{x}\mathbf{x}^\top+\lambda\mathbf{I}_C$$

where $\mathbf{D}(\mathbf{p}) = \mathbf{I}_C\mathbf{p}$.

We derive

$$\nabla^2 \mathcal{L}_{\mathcal{D}}(\mathbf{W}) = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{D}(\mathbf{p}_j) - \mathbf{p}_j\mathbf{p}_j^\top) \otimes \mathbf{x}_j\mathbf{x}_j^\top + \lambda \mathbf{I}_C$$

where $\mathbf{p}_j$ corresponds to $\mathbf{p}$ conditioned with $\mathbf{x}_j$.

As shown in (Krishnapuram et al., 2005), $\nabla^2 \mathcal{L}_{\mathcal{D}}(\mathbf{W})$ satisfies the following

$$\lambda \mathbf{I}_C \preceq \nabla^2 \mathcal{L}_{\mathcal{D}}(\mathbf{W}) \preceq \frac{1}{2}\big(\mathbf{I}_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top\big) \otimes \frac{1}{n}\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_C$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times d}$. In particular, we deduce:

$$\begin{aligned}
\beta &= \lambda_{\mathsf{max}}(\nabla^2 \mathcal{L}_{\mathcal{D}}(\mathbf{W})) \\
&\leq \lambda_{\mathsf{max}}\left(\frac{1}{2}\big(\mathbf{I}_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top\big) \otimes \frac{1}{n}\mathbf{X}\mathbf{X}^\top + \right) + \lambda \\
&= \max_{\lambda_{\mathsf{eig}}, \lambda'_{\mathsf{eig}}} \lambda_{\mathsf{eig}}\left(\frac{1}{2}\big(\mathbf{I}_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top\big)\right) \lambda'_{\mathsf{eig}}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\top\right) + \lambda \\
&\leq \frac{1}{2n}\lambda_{\mathsf{max}}\left(\mathbf{X}\mathbf{X}^\top\right) + \lambda
\end{aligned}$$

□

D. Datasets and models

Datasets

We have selected for our experiments two datasets commonly used for training image classification models: CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2014), and also a dataset with healthcare data which can more closely mimic a scenario where we care about training a model on sensitive data: Pneumonia (Kermany et al., 2018).

CIFAR CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2014) both consist of 50,000 images in the training set and 10,000 in the test set. They are respectively composed of 10 and 100 different balanced classes (such as airplanes, dogs, horses, etc.) and each image consists of a colored 32×32 image. The datasets are disjoints, which allows us to pretrain our models AlexNet and Resnet18 on CIFAR-100 and consider it *public* pre-training before performing logistic regression on CIFAR10.

Pneumonia Pneumonia is a dataset of chest X-ray images of pediatric pneumonia that was published by (Kermany et al., 2018). It is composed of 5163 training and 624 test non-colored images of varying sizes. Images are divided in 3 classes: bacterial (26%), normal (48%) and viral (26%). It provides an interesting use case as it is a relatively small dataset and is composed of healthcare data.

Models

We have selected 2 models for our experimentations.

AlexNet AlexNet is the famous winner of the 2012 ImageNet ILSVRC-2012 competition (Krizhevsky et al., 2012). It has 5 convolutional layers and 3 fully connected layers and it can use batch normalization layers for stability and efficient training.

ResNet18 ResNet18 (He et al., 2016) is the runner-up of the ILSVRC-2015 competition. It is a convolutional neural network that is 18 layers deep, and has 11.7M parameters. It uses batch normalisation layers, but as only the last layer is retrained with differential privacy, we need not replace those layers with group normalisation.