Differentially Private Learning Needs Hidden State (Or Much Faster Convergence)

Anonymous Author(s)
Affiliation
Address
email

Abstract

Differential privacy analysis of randomized learning algorithms typically relies on composition theorems, where the implicit assumption is that the internal state of the iterative algorithm is revealed to the adversary. However, by assuming that the internal state of the algorithm is not revealed, recent works prove a smaller privacy bound for noisy gradient descent (on strongly convex smooth loss functions), compared with composition bounds. In this paper, we significantly improve privacy analysis under the hidden state assumption. We enable taking advantage of privacy amplification by sub-sampling and randomized post-processing, and extend the analysis to "shuffle and partition" and "sample without replacement" stochastic minibatch gradient descent schemes. We prove that, in these settings, our privacy bound is substantially smaller than the composition bounds, notably after many epochs of training (required for learning from high-dimensional data). So, unless the DP algorithm converges fast, our privacy analysis shows that differentially private learning benefits from hidden state privacy analysis. We also empirically show that, given a fixed privacy budget, our analysis enables DP training of image classification models with a better prediction accuracy, compared with prior work.

1 Introduction

Machine learning models leak sensitive information about their training data. To protect user privacy, the widely-used differentially private training algorithm, DP-SGD [1], adds carefully calibrated noise in each gradient update for the model parameters. This added noise guarantees that the models trained on two neighboring datasets are indistinguishable in probability distributions. To quantify the indistinguishability, differential privacy analysis *bounds* the (moment of) likelihood ratio between a pair of models trained on any two neighboring datasets. Improving this privacy analysis is crucial for obtaining a higher (train and test) accuracy of the output model, under a constrained privacy budget.

For a vast majority of learning tasks (such as private image classification), composition-based privacy analysis of DP-SGD [1] is the mainstream analysis method, mainly due to its simplicity and broad applicability. However, private learning via DP-SGD usually suffers from a slow convergence in empirical prediction accuracy [10], under which the total privacy loss grows very large due to composing it over a large number of iterations. This, in turn, results in overestimating the magnitude of required noise for differentially private training, which worsens the privacy-accuracy trade-off.

To alleviate this problem, one popular direction is to design new variants of the DP-SGD algorithm, that converge faster, i.e. require a smaller number of iterations to reach a stable training accuracy. This includes works that derive privacy preserving variants of fast optimization algorithms, such as performing DP-SGD with momentum [22], adaptive gradient clipping [2, 32], adaptive selection of step-size and noise scale [3], and using pre-trained or hand-crafted features [36]. All these approaches aim for a *faster training convergence* for the DP learning algorithm, such that the composition

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

analysis is applied to a smaller number of iterations. This is to reduce the high performance cost of overestimating the privacy loss (and the required noise scale) due to DP composition.

Another line of work focuses on directly improving the privacy analysis of the DP-SGD algorithm. To this end, prior work [18, 5, 15] suggests that hiding the internal state of noisy (S)GD, therefore analyzing the privacy bound for releasing only the last iterate results in a more accurate estimation of privacy loss, which can substantially reduce the performance cost of differential privacy. By applying the last-iterate privacy analysis [18], Feldman et al. [19] design new differentially private algorithms that achieve theoretically optimal excess risk with linear runtime for convex optimization, which is better than the prior privacy utility trade-off for the noisy SGD algorithm [7]. However, the analysis applies only to one single epoch of training, thus it is unclear how such training performs in practical learning settings (which require multiple epochs of training). Chourasia et al. [15] derive a strong privacy guarantee for multiple epoch training with hidden-state noisy GD, under strongly convex loss function. However, the analysis is limited to computationally expensive GD [20], and even if we attempt to directly apply it to stochastic mini-batch training, the analysis fails to model the privacy amplification by post-processing [18, 5] and mini-batch sub-sampling [11, 23, 7, 1, 4]. Thus, the bound under SGD learning becomes very large (as we also show in Figure 1).

Thus, under the hidden-state assumption over many epochs of training, the challenge is to compute a small privacy bound for differentially private stochastic gradient descent algorithms by taking advantage of privacy amplifications due to sub-sampling and randomized updates over mini-batches.

Contributions. In this paper, we tackle this challenge, and bridge the gaps in the prior work on hidden-state (last-iterate) privacy analysis. We model the privacy dynamics of noisy SGD, and show the amplification for hidden-state privacy bounds due to stochastic mini-batches, and demonstrate the empirical utility gain of our bound. **(i)** For multi-epoch noisy *stochastic mini-batch* gradient descent under *"shuffle and partition"* and *iterative "sampling without replacement"*, we prove new converging privacy bounds (for strongly convex smooth loss function) that significantly improves over prior bounds [4, 28, 15]. Our proof relies on our new bounds for the privacy amplification by post-processing (Lemma 4.2) and sub-sampling (Theorems 5.2 and 5.3). **(ii)** We evaluate the privacy accuracy trade-off of practical learning tasks under our privacy dynamics analysis. We empirically show that for learning tasks that require many epochs of training, where DP-SGD [1] has a slow convergence (e.g., logistic regression models trained from scratch), our privacy bound enables achieving a much higher prediction accuracy than what the composition-based analysis brings about. For tasks that converge within a few epochs (e.g., private fine-tuning of pre-trained models), our privacy bound can cause a slightly better privacy accuracy trade-off. [1]

Our results show that, to obtain high accuracy, it is crucial to apply our last-iterate privacy dynamics analysis for learning tasks with slow to moderate convergence. Alternatively, it is not empirically so costly to use composition-based privacy analysis when the training process converges very quickly.

2 Preliminaries

Definition 2.1 (Differential Privacy [16, 17])**.** A randomized algorithm $\mathcal{A}$ is (ε, δ)-differentially private if for any neighboring datasets D, D', and for all possible event S in the output space of $\mathcal{A}$,

$$P\left(\mathcal{A}(D) \in S\right) \leq e^{\varepsilon} \cdot P\left(\mathcal{A}(D') \in S\right) + \delta \tag{1}$$

where we say D, D' are *neighboring* if they are of the same size and differ in at most one data record.

Definition 2.2 ((α, ε)-Rényi DP [27])**.** A randomized algorithm $\mathcal{A}$ is said to satisfy (α, ε)-Rényi differential privacy (or (α, ε)-Rényi DP for short), if for any neighboring datasets D and D',

$$R_\alpha(\mathcal{A}(D) \| \mathcal{A}(D')) \leq \varepsilon, \text{ where } R_\alpha(\mu \| \nu) = \frac{1}{\alpha - 1} \log \mathbb{E}_{\theta \sim \nu}\left[\left(\frac{\mu(\theta)}{\nu(\theta)}\right)^{\alpha}\right] \tag{2}$$

where $\mathcal{A}(D)$ ($\mathcal{A}(D')$) denote the distribution of output given input dataset D (D'), and $R_\alpha(\mu \| \nu)$ is the *Rényi divergence* [33] of order $\alpha > 1$ for two distributions with density $\mu(\theta)$ and $\nu(\theta)$ on $\mathbb{R}^d$.

Definition 2.3 (log-Sobolev Inequality [37])**.** A distribution ν over $\mathbb{R}^d$ satisfies the log-Sobolev inequality (LSI) with constant c if for all smooth function $g : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}_{\theta \sim \nu}\left[g(\theta)^2\right] < \infty$,

$$\mathbb{E}_{\theta \sim \nu}\left[g(\theta)^2 \log\left(g(\theta)^2\right)\right] - \mathbb{E}_{\theta \sim \nu}\left[g(\theta)^2\right] \cdot \log \mathbb{E}_{\theta \sim \nu}\left[g(\theta)^2\right] \leq \frac{2}{c} \mathbb{E}_{\theta \sim \nu}\left[\|\nabla g(\theta)\|^2\right]. \tag{3}$$

[1] As shown after Theorem 5.2, our privacy bound is smaller than composition after $O(1/(\lambda\eta) + n/b)$ epochs.

Algorithm 1 $\mathcal{A}_{\text{Noisy-mBGD}}$: Noisy (Stochastic) Minibatch Gradient Descent

Input: Dataset $D = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$. Parameter space $\theta \in \mathbb{R}^d$. Loss function $\ell(\theta; \mathbf{x})$. Stepsize η. Noise standard deviation σ. Minibatch size b. Initial parameters θ_0^0 sampled from an arbitrary distribution $p_0(\theta)$.
Batch Generation: obtain minibatches B_k^j of size b, for $j = 0, \cdots, n/b-1$; for epochs $k = 0, \cdots, K-1$.
  ▷ **If "Shuffle and partition"**: randomly partition n data indices into $B^0, \cdots, B^{n/b-1}$, let $B_k^j = B^j$. [2]
  ▷ **If Sample without replacement**: resample b different indices from $\{1, \cdots, n\}$ to obtain every B_k^j.
for epoch $k = 0, 1, \cdots, K-1$ **do**
  for iteration $j = 0, 1, \cdots, n/b-1$ **do**
    $\theta_k^{j+1} = \theta_k^j - \eta \cdot g\left(\theta_k^j; B_k^j\right) + \sqrt{2\eta\sigma^2} \cdot \mathcal{N}(0, \mathbb{I}_d) \quad \text{where} \quad g\left(\theta_k^j; B_k^j\right) = \frac{1}{b}\sum_{i \in B_k^j} \nabla \ell(\theta_k^j; \mathbf{x}_i)$
  $\theta_{k+1}^0 = \theta_k^{n/b}$
Output: θ_K^0

3 An overview of the problem and our approach

We analyze differential privacy loss of the noisy (stochastic) minibatch gradient descent Algorithm 1 under two minibatch generation variants: **(1)** "shuffle and partition" (widely implemented in privacy libraries [42, 21]); **(2)** "sample without replacement" (analyzed extensively in prior results [11, 23, 4, 39] despite its computational cost). Our ultimate goal is to prove a *worst-case* upper bound of the Rényi divergence between distributions of *last-iterate* parameters θ_K^0 and ${\theta'}_K^0$ trained on *any* two neighboring datasets, that is *tighter* than composition and enables a *better* privacy accuracy trade-off.

Our starting point is a tight converging hidden-state privacy dynamics analysis in Chourasia et al. [15] for noisy gradient descent. To extend this GD analysis to the stochastic minibatch setting, a natural solution is to conceptually view the updates that *involve* the (sensitive) differing data record (between neighboring datasets) as gradient descent on a *smaller*, iteratively *changing* dataset of size b (i.e., the size of a minibatch). This simple extension results in the following naive privacy baseline. [3]

Theorem 3.1 (Naive Privacy Dynamics Baseline [15])**.** *If the loss function $\ell(\theta; \mathbf{x})$ is λ-strongly convex and β-smooth, and its gradient has ℓ_2-sensitivity S_g, then Algorithm 1 under "shuffle and partition" and step-size $\eta < \frac{2}{\lambda+\beta}$ satisfies (α, ε)-Rényi DP with $\varepsilon \leq \frac{\alpha S_g^2}{\lambda\sigma^2 b^2}(1 - e^{-\lambda\eta K/2})$.*

However, as also shown in Figure 1, this naive privacy dynamics baseline converges *slowly* to a *huge* constant, which is significantly worse than the bounds derived by composition [28, 1] after a large number of epochs. This is because this naive baseline *fails to* capture the privacy amplification due to the stochasticity of minibatch sub-sampling, and the iterative noisy updates which amplify the privacy of preceding minibatches (referred to as amplification by post-processing). In this paper, however, we show how to compute a much tighter differential privacy bound for SGD, under hidden-state assumption, while taking advantage of these privacy amplifications.

Our approach is as follows. We decompose the distribution of last iterate parameters in Algorithm 1 as a mixture of conditional distributions $p(\theta_K^0) = \sum_B p(B) \cdot p(\theta_K^0 | B)$, given any possible minibatch sequence B. We first analyze the Rényi privacy loss for a fixed minibatch sequence while modeling privacy amplification by randomized post-processing (i.e., how much the randomized gradient update improve the privacy due to updates on preceding mini-batches). We then quantify the privacy amplification by subsampling under stochastic minibatches.

Privacy amplification by randomized post-processing. When a (fixed) minibatch only contains the indices of shared data records between two neighboring datasets, then the deterministic minibatch gradient descent mapping does not cause any privacy loss. Addition of the Gaussian noise that follows this deterministic update, however, serves as randomized post-processing which *decreases* the Rényi divergence between the two processes. To compute this privacy amplification, we pre-

[2] For simplicity of presentation, we assume b divides n. If n/b is not an integer, and if the algorithm ignores the last $n - \lfloor n/b \rfloor \cdot b$ data points, then our privacy dynamics bound holds by replacing n/b with $\lfloor n/b \rfloor$.

[3] A recent concurrent work [34] also follows this direct approach to extend the GD analysis [15] to SGLD setting, and proves a similar bound to our naive privacy dynamics baseline Theorem 3.1. However, there is a slight difference between [34, Corollary 3.3] and our Theorem 3.1, due to the flawed assumption in [34, Lemma 3.4] that the LSI constant proved in Chourasia et al. [15] (that *only* holds for GD process) also holds for SGLD process (that takes the form of a more complicated mixture distribution). See more details in Appendix B.

cisely model the change of parameter distributions with the Fokker-Planck equation (for diffusion process with zero drift).

Hidden-state privacy amplification by sub-sampling. Although privacy amplification by a *single* sub-sampling operation is well-studied [11, 23, 7, 1, 4], to the best of our knowledge, no prior bounds are applicable to *hidden-state* iterative resampling. This is because under the hidden-state assumption, the number of mixing components in the last-iterate parameter distribution grows exponentially with the number of epochs, which makes the mixture distribution very difficult to analyze. In this paper, we first carefully couple the mixture components between the processes associated with neighboring datasets. We use the joint convexity of exponentiated Rényi divergence to bound how much the Rényi divergence between mixture distributions for the pair of processes is smaller than its worst case.

4 Improved privacy dynamics for noisy (fixed) minibatch gradient descent

In this section, we quantify the privacy amplification by randomized post-processing in Algorithm 1, during iterations that do not access the sensitive differing data between neighboring datasets. We then combine it with Rényi DP composition during the remaining iterations to prove a converging privacy bound.

4.1 Privacy amplification by randomized post-processing (additive Gaussian noise)

We first explain the key lemma that proves exponentially decaying Rényi privacy loss under additive Gaussian noise post-processing, when the parameter distributions satisfy the log-Sobolev inequality. This is a new bound for the well-known "privacy amplification by iteration" phenomenon [18, 5].

Lemma 4.1. *Let μ, ν be two distributions on $\mathbb{R}^d$. Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a measurable mapping on $\mathbb{R}^d$. We denote $\mathcal{N}(0, 2t\sigma^2 \cdot \mathbb{I}_d)$ to be the standard Gaussian distribution on $\mathbb{R}^d$ with covariance matrix $2t\sigma^2 \cdot \mathbb{I}_d$. We denote $p_t(\theta)$ and $p'_t(\theta)$ to be the probability density functions for the distributions $f_\#(\mu) * \mathcal{N}(0, 2t\sigma^2\mathbb{I}_d)$ and $f_\#(\nu) * \mathcal{N}(0, 2t\sigma^2\mathbb{I}_d)$ respectively, where $f_\#(\mu), f_\#(\nu)$ denote the push forward distributions of μ, ν under mapping f. Then if μ and ν satisfy log-Sobolev inequality with constant c, and if the mapping f is L-Lipschitz, then for any order $\alpha > 1$,*

$$\frac{\partial}{\partial t} R_\alpha \left(p_t(\theta) \| p'_t(\theta)\right) \leq -c_t \cdot 2\sigma^2 \cdot \left(\frac{R_\alpha(p_t(\theta)\|p'_t(\theta))}{\alpha} + (\alpha - 1) \cdot \frac{\partial}{\partial \alpha} R_\alpha(p_t(\theta)\|p'_t(\theta)) \right), \quad (4)$$

where $c_t = \left(\frac{L^2}{c} + 2t\sigma^2\right)^{-1}$ is the log-Sobolev inequality constant for distributions $p_t(\theta)$ and $p'_t(\theta)$.

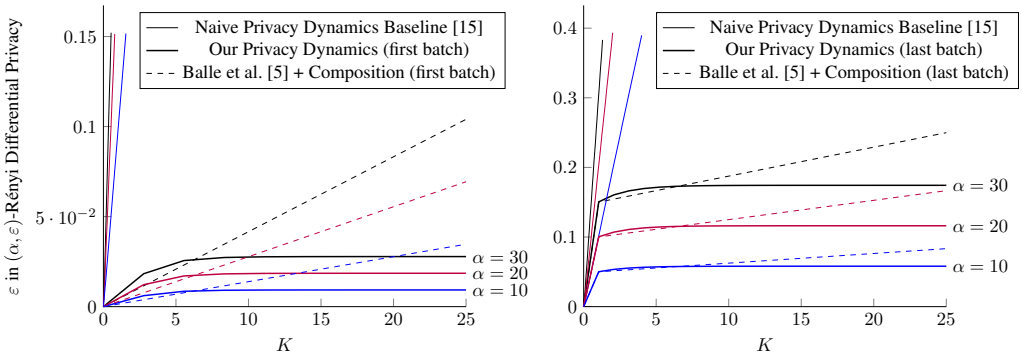
Proof Sketch. The proof starts by modelling Gaussian noise as a diffusion process with *zero* drift. We then bound the rate of Rényi divergence with the LSI constant for the process (following prior works [37, 15]). However, instead of assuming a fixed LSI constant c for all intermediate distributions, we prove a more precise LSI constant c_t that changes with t. Complete proof is in Appendix C.1. □

The partial differential inequality in Lemma 4.1 quantifies the amplification of Rényi privacy loss under additive Gaussian noise. By solving Equation (4) on $t \in [0, \eta]$, we prove the following lemma that models recursive privacy dynamics during one step of noisy minibatch gradient descent.

Lemma 4.2. *Let D, D' be an arbitrary pair of neighboring datasets that differ in the i_0-th data point (i.e. $x_{i_0} \neq x'_{i_0}$). Let B_k^j be a fixed minibatch used (in iteration j of epoch k) in Algorithm 1, which contains b indices sampled from $\{1, \cdots, n\}$. We denote θ_k^j and ${\theta'}_k^j$ as the intermediate parameters in Algorithm 1 on input datasets D and D', respectively. If the distributions of θ_k^j and ${\theta'}_k^j$ satisfy log-Sobolev inequality with a constant c, and if the minibatch GD mapping $f(\theta) = \theta - \eta \cdot \frac{1}{b} \cdot \sum_{i \in B_k^j} \ell(\theta; \mathbf{x}_i)$ is L-Lipschitz for parameters θ, then the following recursive bound for Rényi divergence holds.*

$$\frac{R_\alpha(\theta_k^{j+1} \| {\theta'}_k^{j+1})}{\alpha} \leq \begin{cases} \frac{R_{\alpha'}(\theta_k^j \| {\theta'}_k^j)}{\alpha'} \cdot \left(1 + \frac{c \cdot 2\eta\sigma^2}{L^2}\right)^{-1} & \text{if } i_0 \notin B_k^j \\ \frac{R_\alpha(\theta_k^j \| {\theta'}_k^j)}{\alpha} + \frac{\eta S_g^2}{4\sigma^2 b^2} & \text{if } i_0 \in B_k^j \end{cases} \text{ with } \alpha' = \frac{\alpha - 1}{1 + \frac{c \cdot 2\eta\sigma^2}{L^2}} + 1. \quad (5)$$

Proof Sketch. When $i_0 \notin B_k^j$, we apply Lemma 4.1 on the deterministic minibatch gradient descent update f, and solve Equation (4) to obtain recursive Rényi privacy bound. This proof is similar to



(a) first batch (b) last batch

Figure 1: Rényi privacy loss of noisy (fixed) minibatch gradient descent over K epochs, which goes over a fixed sequence of minibatch partition in each epoch. We show the privacy loss for data points in the first minibatch B^0 and the last minibatch $B^{n/b-1}$ of each pass. We show the ε in the (q, ε)-RDP guarantee derived by our privacy dynamics analysis (bold lines), the naive privacy dynamics baseline (thin lines), and the privacy amplification by mixing and diffusion analysis [5] combined with composition (dashed line). We evaluate under the following setting: RDP order $\alpha \in \{10, 20, 30\}$; λ-strongly convex loss function with $\lambda = 1$; β-smooth loss function with $\beta = 4$; total gradient sensitivity $S_g = 4$; size of the data set $n = 50$; step-size $\eta = 0.02$; noise variance $\sigma^2 = 4$, minibatch size $b = 2$. We use Theorem 4.3 for our privacy dynamics; We use Theorem 3.1 for naive privacy dynamics baseline; We use [5, Theorem 5] to obtain Balle et al. [4] + Composition (details are in Appendix C.6).

that of [15, Theorem 2], however, dealing with the new LSI constant c_t that changes with t introduces an additional technical difficulty. When $i_0 \in B_k^j$, we use composition theorem to prove the additive recursion in the second row of Equation (5). The complete proof is in Appendix C.2. □

Comparison with prior amplification by post-processing bound. Prior to this work, for convex and smooth loss functions, Feldman et al. [18] derive a tight bound for the amplification by post-processing in noisy stochastic gradient descent. In Appendix C.4, we show that our new recursive amplification bound could recover this known *tight* bound for convex smooth loss functions, while using a different proof. For (a more restrictive setting of) *one* epoch of noisy SGD on *strongly convex* loss function, Balle et al. [5] further improve the bound in Feldman et al. [18] via a careful coupling-based approach. However, we do not see an easy way to extend the bound in Balle et al. [4] to *multiple* epochs except by using Rényi DP composition (which would give a linearly worsening Rényi DP bound with the number of epochs). By contrast, our recursive amplification bound easily applies to multiple epochs (under hidden state assumption, and without requiring composition over the epochs), and enables converging privacy dynamics for Algorithm 1 on strongly convex smooth loss functions (as discussed in Section 4.2).

4.2 Improved privacy dynamics for (fixed) minibatch gradient descent

By using the recursive privacy bound Lemma 4.2 and the non-overlapping property of minibatch partitions of the dataset, we prove position-dependent Rényi DP bounds for Algorithm 1 as follows.

Theorem 4.3 (Privacy dynamics under strongly convex smooth loss)**.** *Conditioned on a fixed sequence of partitioned minibatches $B^0, \cdots, B^{n/b-1}$ in Line 3, if the loss function is λ-strongly convex, β-smooth and its gradient has ℓ_2-sensitivity S_g, then running Algorithm 1 for $K \geq 1$ epochs with step-size $\eta < \frac{2}{\lambda+\beta}$, satisfies (α, ε)-Rényi DP for data points in the batch B^{j_0}, with*

$$\varepsilon \leq \varepsilon_0^{\lfloor \frac{n}{2b} \rfloor}(\alpha) \cdot \frac{1 - (1-\eta\lambda)^{2\cdot(K-1)\cdot(n/b - \lfloor \frac{n}{2b} \rfloor)}}{1 - (1-\eta\lambda)^{2\cdot(n/b - \lfloor \frac{n}{2b} \rfloor)}} + \varepsilon_0^{n/b - j_0}(\alpha) \tag{6}$$

where $\varepsilon_0^j(\alpha) = \frac{\alpha\eta S_g^2}{4\sigma^2 b^2} \cdot (1-\eta\lambda)^{2\cdot(j-1)} \cdot \frac{1}{\sum_{s=0}^{j-1}(1-\eta\lambda)^{2s}}$ for any $j = 1, \cdots, \frac{n}{b}$ (we assume $\frac{n}{b} \geq 2$).

Proof Sketch. We first prove that the distribution of parameters θ_k^j satisfies LSI with a constant c_k^j that depends on k, j. Then we plug the LSI constants into Lemma 4.2 and prove a recursive privacy bound for data points in each batch B^{j_0}. Finally, by carefully choosing which recursion to use and solve, we obtain the privacy bound in the theorem statement. The complete proof is in Appendix C.5. □

The above new privacy bound Theorem 4.3 quantifies the privacy amplification during iterations that only access the shared data point between neighboring datasets, and it is significantly smaller than the naive privacy dynamics baseline derived from Chourasia et al. [15] (that does not capture this additional privacy amplification). In Figure 1, we observe that for data points in the first batch B^0, our new privacy dynamics bound is smaller by a multiplicative factor of approximately n/b (where b/n is the batch sampling ratio), and the improvement for data points in the last-batch is also significant.

Bound improvement compared with Balle et al. [5]. In the first epoch (when $K = 1$), our privacy dynamics bound (based on recursive scheme) in Theorem 4.3 is of same order as the privacy amplification bound Balle et al. [5, Theorem 5], as shown in Figure 1. However, for training with multiple epochs, Balle et al. [5] use a coupling-based approach, and we do not see an easy way to extend their analysis except by using Rényi DP composition [27] over the epochs, which gives a linearly accumulating Rényi DP bound for multiple epochs (as K increases). On the contrary, as shown in Figure 1, our improved privacy dynamics bound converges to a constant, thus significantly improves over the Rényi DP composition of Balle et al. [5, Theorem 5].

5 Privacy dynamics for noisy (stochastic) minibatch gradient descent

In this section, we further improve over the position-dependent privacy dynamics analysis in Section 4, by incorporating the effect of amplification of privacy loss due to the stochastic minibatch sampling. We first prove the following lemma by the convexity of f-divergence with $f(x) = x^\alpha$.

Lemma 5.1 (Joint convexity of scaled exponentiation of Rényi divergence)**.** *Let $\mu_1, \cdots, \mu_m$ and $\nu_1, \cdots, \nu_m$ be distributions over R^d. Then for any RDP order $\alpha \geq 1$, and any coefficients $p_1, \cdots, p_m \geq 0$ that satisfy $p_1 + \cdots + p_m = 1$, the following inequality holds.*

$$e^{(\alpha-1)\cdot R_\alpha(\sum_{j=1}^m p_j\mu_j \| \sum_{j=1}^m p_j\nu_j)} \leq \sum_{j=1}^m p_j \cdot e^{(\alpha-1)\cdot R_\alpha(\mu_j\|\nu_j)} \tag{7}$$

We provide a detailed proof for Lemma 5.1 in Appendix D. This Lemma is our main tool for quantifying the additional privacy amplification under stochastic minibatches in the rest of the section.

5.1 Privacy dynamics under shuffle and partition

The "shuffle and partition" batch generation scheme is commonly used for practical DP-SGD implementations in privacy libraries [21, 42]. By using Theorem 4.3 and new bounds for the privacy amplification by shuffling in Algorithm 1, we prove the following Rényi DP guarantee.

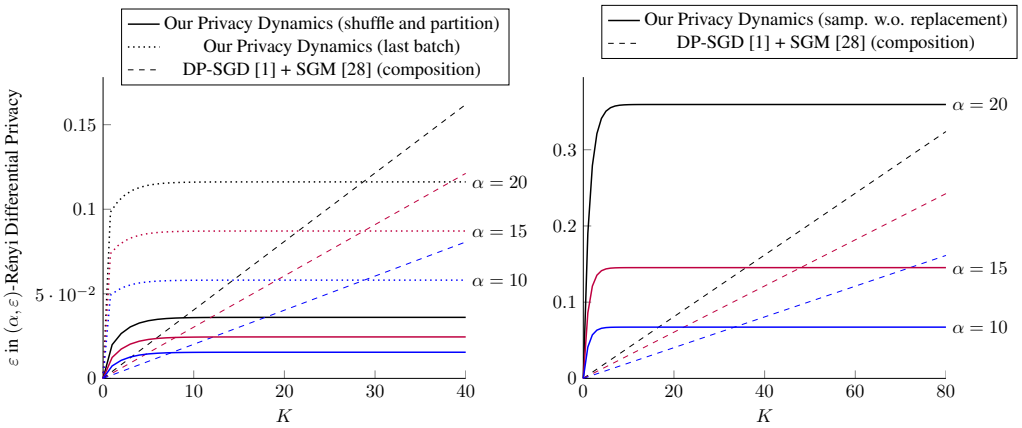
Theorem 5.2 (Privacy dynamics under "shuffle and partition")**.** *If the loss function $\ell(\theta; x)$ is λ-strongly convex, β-smooth, and if its gradient has finite ℓ_2-sensitivity S_g, then for $\frac{n}{b} \geq 2$, running Algorithm 1 for $K \geq 1$ epochs with step-size $\eta < \frac{2}{\lambda+\beta}$, under "shuffle and partition" batch generation scheme, satisfies (α, ε)-Rényi DP with*

$$\varepsilon \leq \varepsilon_0^{\lfloor \frac{n}{2b} \rfloor}(\alpha) \cdot \frac{1-(1-\eta\lambda)^{2\cdot(K-1)\cdot(n/b-\lfloor \frac{n}{2b} \rfloor)}}{1-(1-\eta\lambda)^{2\cdot(n/b-\lfloor \frac{n}{2b} \rfloor)}} + \frac{1}{\alpha-1} \cdot \log\left(\underset{0\leq j_0<n/b}{Avg} e^{(\alpha-1)\varepsilon_0^{n/b-j_0}(\alpha)} \right) \tag{8}$$

where $\varepsilon_0^j(\alpha) = \frac{\alpha\eta S_g^2}{4\sigma^2 b^2} \cdot (1-\eta\lambda)^{2\cdot(j-1)} \cdot \frac{1}{\sum_{s=0}^{j-1}(1-\eta\lambda)^{2s}}$ for any $j = 1, \cdots, \frac{n}{b}$ (we assume $\frac{n}{b} \geq 2$).

The above Rényi DP bound (8) in Theorem 5.2 is always smaller than the bound in Theorem 4.3 given the worst-case fixed minibatch sequence (i.e., when the differing data point is in the last minibatch of each pass). Therefore, Theorem 5.2 quantifies the privacy amplification due to shuffling, when compared to the worst-case Rényi DP bound among all possible minibatch sequences. We illustrate this amplification in Figure 2 (a). Moreover, our Rényi DP bound in Theorem 5.2 is strictly smaller than composition-based privacy bound for DP-SGD [1, 28] after $4 \cdot \frac{1}{\lambda\eta} + 4 \cdot \frac{n}{b}$ epochs. This is because the first term in Equation (178) is strictly smaller than the composition-based privacy bound after $4 \cdot \frac{1}{\lambda\eta}$ epochs, and the second term in Equation (178) is strictly smaller than the composition-based privacy bound after $4 \cdot (n/b)$ epochs (as we explain with more details in Appendix D.1).

[4] When $\alpha - 1$ and $\varepsilon_0^{n/b-j_0}(\alpha)$ are large, the second term in Equation (8) might overflow. In the experiments, we compute this log-sum-exp term with the shifted approximation $\varepsilon_0^1 + \log \underset{0\leq j_0<n/b}{Avg} e^{(\alpha-1)\left(\varepsilon_0^{n/b-j_0}(\alpha) - \varepsilon_0^1(\alpha)\right)}$.



(a) shuffle and partition

(b) (re)sample without replacement

Figure 2: Rényi privacy loss of noisy (stochastic) minibatch gradient descent over K epochs. We show ε in the (q, ε)-RDP guarantee derived by our privacy dynamics bound under "shuffle and partition" (bold lines, left plot), our privacy dynamics bound under sampling without replacement (bold lines, right plot), our privacy dynamics bound for data points in the last batch $B^{n/b-1}$ (thin dashed lines), and the baseline composition-based bound for DP-SGD [1, 28] (thin lines). We evaluate under the following setting: RDP order $\alpha \in \{10, 20, 30\}$; λ-strongly convex loss function with $\lambda = 1$; β-smooth loss function with $\beta = 4$; finite total gradient sensitivity $S_g = 4$; size of the data set $n = 50$; step-size $\eta = 0.02$; noise variance $\sigma^2 = 4$, batch size $b = 2$. The expressions for computing the privacy bounds are: Privacy Dynamics (shuffle and partition): Theorem 5.2 [4]; Privacy Dynamics (samp. w.o. replacement): Theorem 5.3; Privacy Dynamics (last batch): Theorem 4.3 under $j_0 = n/b - 1$; Composition: derived from Section 3.3. of Mironov et al. [28] that approximately equals $\frac{b}{n} \cdot \frac{\alpha \eta S_g^2}{4\sigma^2 b^2} \cdot K$.

5.2 Privacy dynamics under sampling without replacement

We now similarly analyze privacy dynamics for another variant of noisy minibatch gradient descent (Algorithm 1), where we iteratively (re)sample b indices without replacement from $\{1, \cdots, n\}$ to obtain each minibatch. That is, the minibatches used in different iterations (of one epoch) may overlap.

Theorem 5.3 (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 Algorithm 1 under sampling without replacement and stepsize $\eta < \frac{2}{\lambda+\beta}$ satisfies (α, ε)-Rényi DP guarantee with*

$$\varepsilon \leq \frac{1}{\alpha - 1} \log \left(S_K^0(\alpha) \right) \tag{9}$$

where the terms $S_k^j(\alpha)$ for $k = 0, \cdots, K-1$ and $j = 0, \cdots, n/b - 1$ are recursively computed by $S_0^0(\alpha) = 1$; $S_k^{j+1}(\alpha) = \frac{b}{n} \cdot e^{\frac{(\alpha-1)\alpha\eta S_g^2}{4\sigma^2 b^2}} \cdot S_k^j(\alpha) + (1 - \frac{b}{n}) \cdot S_k^j(\alpha)^{(1-\eta\lambda)^2}$; and $S_{k+1}^0(\alpha) = S_k^{n/b}(\alpha)$.

In Figure 2 (b), we compare the above Rényi DP bound Theorem 5.3 with the baseline composition-based privacy bound for DP-SGD [1, 27]. We observe that for RDP order $\alpha = 10, 15$, our privacy dynamics bound significantly improves over the baseline composition-based bound after 50 epochs. However, this improvement is smaller for large RDP order $\alpha = 20$. Moreover, our privacy bound under "sample without replacement" in Figure 2 (b) is larger than our privacy bound under "shuffle and partition" in Figure 2 (a), which suggests room for future improvement of the bound in Theorem 5.3.

6 Evaluating privacy accuracy trade-off under improved privacy dynamics

In this section, we evaluate regularized logistic regression in a multi-class setting (with per-class bias). We compare the optimal accuracy of (ε, δ)-DP models under our privacy dynamics bound, against the best known prior results [36] under DP-SGD [1, 28], for the same privacy budget ε and δ.

Experiment setting. We run experiments on image classification tasks. The training dataset that we take as input is $D = (\mathbf{z}_1, \cdots, \mathbf{z}_n)$, where each data record $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$ consists of the d-dimentional data input feature vector $\mathbf{x}_i \in \mathbb{R}^d$, and the label vector $\mathbf{y}_i \in \{0, 1\}^c$ (in one-hot encoding). To train a

Table 1: **Test accuracy of logistic regression model trained on ScatterNet features under our privacy dynamics bound in Corollary E.3, compared with prior results [36] under DP-SGD** on MNIST, Fashion-MNIST, and CIFAR-10. We fix the privacy budget for all the trained models at $(\varepsilon = 3, \delta = 10^{-5})$, and show the mean and standard deviation across five runs. The hyperparameters are optimized within each setup to faciliate optimal accuracy. Feature clipping worsens the test accuarcy of trained model (compared to gradient clipping). Despite this negative effect, the test accuracy under privacy dynamics (with feature clipping) is consistently higher than that under unregularized DP-SGD (with momentum). Full hyperparameters are in Appendix E.7.

Analysis	DP-SGD [1] + momentum			Dynamics	Non-DP
DP Notion	Add/Remove a record [5]			Change a record	-
Regularizer	No		Yes	Yes	No
Clipping	Gradient	Feature+Gradient	Feature+Gradient	Feature+Gradient	No
MNIST	98.47±0.03%	98.36±0.06%	98.33±0.08%	**98.85±0.05%**	99.3±0.0%
FMNIST	89.64±0.19%	87.71±0.21%	87.76±0.22%	**89.87±0.18%**	91.5±0.0%
CIFAR-10	66.78±0.27%	65.87±0.20%	65.71±0.32%	**69.51±0.23%**	71.1±0.0%
Reference	[36]	-	-	This Paper	[36]

differentially private classifier, we use Algorithm 1 under "shuffle and partition", which is equivalent to practical implementations of DP-SGD [1] (as explained in Appendix E.1). To satisfy the necessary conditions for our privacy dynamics bound, we use regularized logistic regression to ensure strong convexity, use feature clipping to ensure smoothness, and use gradient clipping to enforce finite gradient sensitivity (described in Appendix E.2, E.3 and E.4). The formulas for these privacy bounds, are in Appendix E.5. Our code is provided in the supplemental material.

Hyperparameter tuning. In hyperparameter tuning, multiple accesses to the same private training dataset are required for training models under different hyperparameters, thus contributing to possibly higher privacy cost than the privacy bound for a single run of training algorithm. However, most prior works [1, 36] do not account for this additional privacy cost due to hyperparameter tuning. Therefore, for a fair comparison with Tramèr and Boneh [36], we also do not account the privacy cost due to hyperparameter tuning. Moreover, recent works [24, 30] also have proposed algorithms that reduce the privacy cost for hyperparameter tuning to a constant times the privacy cost of a single run.

6.1 Experiments on ScatterNet features

ScatterNet features. We follow Tramèr and Boneh [36] and similarly train regularized logistic regression models from scratch on hand-crafted ScatterNet [9, 29] features (extracted via data-independent scale invariant feature transformation [25]). This is because, to the best of our knowledge, Tramèr and Boneh [36] provides the best known accuracy for logistic regression models trained from scratch. More details could be found in Section 2.1 and Appendix C.1 in Tramèr and Boneh [36].

Slow convergence of DP-SGD for training from scratch. Tramèr and Boneh [36] show that differentially private linear models trained on hand-crafted ScatterNet features improve the current state-of-the art test accuracy for end-to-end differentially private CNN models (both under DP-SGD analysis). We reproduce their test accuracy results for MNIST, Fashion-MNIST, and CIFAR-10 in the following Table 1. We observe that due to the restriction from the composition-based privacy bound, their training algorithm could only run for a limited number of epochs, e.g. 60 epochs for training on CIFAR-10. However, the accuracy of the trained model continues to grow even after 60 epochs (discussed in detail in Appendix E.6), indicating that insufficient number of allowed running epochs of DP-SGD restricts its accuracy for training differentially private model from scratch.

Improved accuracy under privacy dynamics Analysis. Our tighter privacy bound in Corollary E.3 allows the noisy minibatch gradient to run for significantly more epochs than DP-SGD, within the same fixed privacy budget. In Table 1, we show that our privacy dynamics bound enables in a better privacy accuracy trade-off for (regularized) logistic regression on hand-crafted features, than that in Tramèr and Boneh [36] (which, to the best of our knowledge, is also the best reported test accuracy for differentially private image classification models trained from scratch in prior works).

[5] The "change a record" DP notion is different from the "add/remove a record" DP notion. However, (ε, δ)-DP under "add/remove a record" can be easily translated to $(2\varepsilon, \delta)$-DP under "change a record".

Table 2: **Test accuracy of model obtained by differentially privately fine-tuning the last layer of pretrained models under our privacy dynamics bound in Corollary E.3, compared with DP-SGD results reproduced from Tramèr and Boneh [36].** We fix the privacy budget for all the trained models at $(\varepsilon = 2, \delta = 10^{-5})$, and show the mean and standard deviation across five random runs. The hyperparameters are optimized within each setup to faciliate optimal accuracy. The test accuracy under privacy dynamics is only slightly higher than that under DP-SGD (with momentum). The full hyperparameters are in Appendix E.8.

Analysis	DP-SGD [1] + momentum	Dynamics	Non-DP
DP Notion	Add/Remove a record [5]	Change a record	-
Clipping	Gradient	Feature + Gradient	No
ResNeXt-29 (CIFAR-100)	79.79±0.22%	**80.09±0.08%**	84%
SIMCLR v2 (unlabelled ImageNet)	92.36±0.11%	**92.71±0.11%**	95.3%
Reference	[36]	This paper	[41] and [13]

6.2 Experiments on Transfer Learning

In this section, we empirically analyze the complementing setting of transfer learning, which typically only requires a few epochs to reach a stable accuracy. We use CIFAR-10 as the target private dataset, and differentially privately fine-tune *the last layer* of the source pre-trained model (trained on another dataset), which is equivalent to training a logistic regression model on top of frozen second-to-last layer features of the source pre-trained model.

Source pre-trained model. The problem of privately fine-tuning a source model (trained from a different/public dataset) to a private target training dataset is well-studied [1, 31, 36]. The high quality features learned from pre-trained model usually enable a better privacy accuracy trade-off of the target learning task, than training from scratch. We follow the setups in Tramèr and Boneh [36] (which to the best our knowledge, attain the highest accuracies for fine-tuned models in the literature) and experiment with two non-private pre-trained models: **(1)** a ResNeXt-29 [41] model trained on the CIFAR-100 dataset; **(2)** a SimCLR [13] model trained on unlabeled ImageNet dataset.

Fast convergence of DP-SGD for fine-tuning. In our experiments, we observe that DP-SGD reaches stable accuracy within 50 epochs, for fine-tuning pre-trained models on CIFAR-10. This is much faster than training from scratch that typically requires a few times more epochs to converge.

Similar accuracy under privacy dynamics analysis. Transfer learning tasks need relatively small number of iterations to reach convergence. However, our privacy dynamics bound is only smaller than composition-based bound after $O(\frac{1}{\lambda\eta} + \frac{n}{b})$ epochs (shown after Theorem 5.2), which suggests that composition may be tighter if the training only requires a few epochs. Therefore, in Table 2, we observe that our privacy dynamics analysis only enables slightly better privacy accuracy trade-off.

7 Conclusion

We prove a novel converging last-iterate privacy bound for noisy stochastic mini-batch gradient descent on strongly convex smooth loss functions. Our bound substantially improves the prior privacy dynamics bound for noisy GD [15], by proving novel bounds for the additional privacy amplification (by post-processing and subsampling) during training with stochastic mini-batches. We evaluate the privacy accuracy trade-off for various learning tasks under our improved privacy dynamics bound, and compare it with that under DP-SGD. We show that when the training process has a moderate to slow convergence, our privacy dynamics analysis enables better accuracy than that under composition privacy bound for DP-SGD. This shows that to achieve the optimal privacy accuracy trade-off, differentially private learning needs to be evaluated by a last-iterate privacy bound, unless it has a very fast convergence (under which, in practice, the cost of composition bound is not significant).

Future Work. There is room for improving our privacy amplification bound Theorem 5.2 under smaller batch size b, because under *one* sampling step, the best known amplification rate is as small as $O(\frac{b^2}{n^2})$ [28]. This suggests possibility for a better *hidden-state* privacy amplification bound (without using composition) for multiple iterations of sampling.

References

[1] 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.

[2] Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. *Advances in Neural Information Processing Systems*, 34, 2021.

[3] Hilal Asi, John Duchi, Alireza Fallah, Omid Javidbakht, and Kunal Talwar. Private adaptive gradient methods for convex optimization. In *International Conference on Machine Learning*, pages 383–392. PMLR, 2021.

[4] 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.

[5] Borja Balle, Gilles Barthe, Marco Gaboardi, and Joseph Geumlek. Privacy amplification by mixing and diffusion mechanisms. *arXiv preprint arXiv:1905.12264*, 2019.

[6] Jean-Baptiste Bardet, Nathaël Gozlan, Florent Malrieu, and Pierre-André Zitt. Functional inequalities for gaussian convolutions of compactly supported measures: Explicit bounds and dimension dependence. *Bernoulli*, 24(1):333–353, 2018.

[7] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

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

[9] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

[10] Zhiqi Bu, Hua Wang, Qi Long, and Weijie J Su. On the convergence of deep learning with differential privacy. *arXiv preprint arXiv:2106.07830*, 2021.

[11] Kamalika Chaudhuri and Nina Mishra. When random sampling preserves privacy. In *Annual International Cryptology Conference*, pages 198–213. Springer, 2006.

[12] Hong-Bin Chen, Sinho Chewi, and Jonathan Niles-Weed. Dimension-free log-sobolev inequalities for mixture distributions. *Journal of Functional Analysis*, 281(11):109236, 2021.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[14] Xiangyi Chen, Steven Z Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: A geometric perspective. *Advances in Neural Information Processing Systems*, 33:13773–13782, 2020.

[15] 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.

[16] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[17] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

[18] 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.

[19] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

[20] Jonas Geiping, Micah Goldblum, Phillip E Pope, Michael Moeller, and Tom Goldstein. Stochastic training is not necessary for generalization. *arXiv preprint arXiv:2109.14119*, 2021.

[21] Google et al. Tensorflow privacy, 2021.

[22] Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. *arXiv preprint arXiv:2103.00039*, 2021.

[23] Ninghui Li, Wahbeh Qardaji, and Dong Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, pages 32–33, 2012.

[24] Jingcheng Liu and Kunal Talwar. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 298–309, 2019.

[25] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[26] Siwei Lyu. Interpretation and generalization of score matching. *arXiv preprint arXiv:1205.2629*, 2012.

[27] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

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

[29] Edouard Oyallon and Stéphane Mallat. Deep roto-translation scattering for object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2865–2873, 2015.

[30] Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy. *arXiv preprint arXiv:2110.03620*, 2021.

[31] Nicolas Papernot, Steve Chien, Shuang Song, Abhradeep Thakurta, and Ulfar Erlingsson. Making the shoe fit: Architectures, initializations, and tuning for learning with privacy. 2019.

[32] Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X Yu, Sashank J Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019.

[33] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562. University of California Press, 1961.

[34] Théo Ryffel, Francis Bach, and David Pointcheval. Differential privacy guarantees for stochastic gradient langevin dynamics. *arXiv preprint arXiv:2201.11980*, 2022. URL `https://arxiv.org/abs/2201.11980v2`.

[35] Inder Jeet Taneja and Pranesh Kumar. Relative information of type s, csiszár's f-divergence, and information inequalities. *Information Sciences*, 166(1-4):105–125, 2004.

[36] Florian Tramèr and Dan Boneh. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations*, 2020.

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

[38] Feng-Yu Wang and Jian Wang. Functional inequalities for convolution probability measures. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 52, pages 898–914. Institut Henri Poincaré, 2016.

[39] 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.

[40] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[41] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[42] 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.

[43] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

[44] David Zimmermann. Logarithmic sobolev inequalities for mollified compactly supported measures. *Journal of Functional Analysis*, 265(6):1064–1083, 2013.

[45] David Zimmermann. Elementary proof of logarithmic sobolev inequalities for gaussian convolutions on $\mathbb{R}$. In *Annales Mathématiques Blaise Pascal*, volume 23, pages 129–140, 2016.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1.
   (b) Did you describe the limitations of your work? [Yes] See Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 4 and 5.
   (b) Did you include complete proofs of all theoretical results? [Yes] See supplemental material.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Table 1 and 2.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See supplemental material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See supplemental material.
   (b) Did you mention the license of the assets? [Yes] See supplemental material.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See supplemental material.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]