Towards More Practical Stochastic Gradient MCMC in Differential Privacy

Bai Li[1] **Changyou Chen**[2] **Hao Liu**[3] **Lawrence Carin**[1]

[1]Duke University [2]University at Buffalo, SUNY [3]California Institute of Technology

Abstract

Concerns related to data security and confidentiality have been raised when applying machine learning to real-world applications. Differential privacy provides a principled and rigorous privacy guarantee for machine learning models. While it is common to inject noise to design a model satisfying a required differential-privacy property, it is generally hard to balance the trade-off between privacy and utility. We show that stochastic gradient Markov chain Monte Carlo (SG-MCMC) – a class of scalable Bayesian posterior sampling algorithms – satisfies strong differential privacy, when carefully chosen stepsizes are employed. We develop theory on the performance of the proposed differentially-private SG-MCMC method. We conduct experiments to support our analysis, and show that a standard SG-MCMC sampler with minor modification can reach state-of-the-art performance in terms of both privacy and utility on Bayesian learning.

1 Introduction

Utilizing large amounts of data has helped machine learning algorithms achieve significant success in many real-world applications. However, such work also raises privacy concerns. For example, a diagnostic system based on machine learning algorithms may be trained on a large quantity of patient data, such as medical images. It is important to protect training data from adversarial attackers (Shokri et al., 2017). However, even the most widely-used machine learning algorithms may implicitly memorize the training data (Papernot et al., 2016), meaning that the learned model parameters implicitly contain information that could violate the privacy of training data. Such algorithms may be readily attacked (Fredrikson et al., 2015).

The above potential model vulnerability can be addressed by differential privacy (DP), a general notion of algorithm privacy (Dwork, 2008; Dwork et al., 2006). This approach is designed to provide a strong privacy guarantee for general learning procedures, such as statistical analysis and machine learning algorithms, that involve private information.

Among the popular machine learning models, Bayesian inference has realized significant success recently, due to its capacity to leverage expert knowledge and manifest uncertainty estimates. Notably, the recently developed stochastic gradient Markov chain Monte Carlo (SG-MCMC) technique enables scalable Bayesian inference for large datasets. While there have been many extensions of SG-MCMC, little work has been directed at studying the privacy properties of such algorithms. Specifically, Wang et al. (2015) showed that an SG-MCMC algorithm with appropriately chosen stepsizes preserves differential privacy. In practice, however, their analysis requires the stepsize to be extremely small to limit the risk of violating privacy. Such a small stepsize is not practical for sampling models with non-convex posterior distribution landscapes, which is the most common case in recent machine learning models. More details of this issue are discussed in Section 3.1.

On the other hand, Abadi et al. (2016) introduced a new privacy-accounting method, which allows one to keep better track of the privacy loss (defined in Section 2.1) for iterative algorithms. Further, they proposed a differentially-private stochastic gradient descent (DP-SGD) method for training machine learning models privately. Although they showed a significant improvement in calculating the privacy loss, there is no theory showing that their DP-SGD has a guaranteed performance under privacy constraints.

In this paper we show that using SG-MCMC for sam-

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

pling large-scale machine learning models is sufficient to achieve differential privacy with small privacy budgets. Specifically, we combine the advantages of the aforementioned works, and prove that SG-MCMC methods naturally satisfy the definition of differential privacy, even without changing their default stepsize and numbers of iterations, thus allowing both good utility and privacy in practice.

2 Preliminaries

We denote an input database with N data points as $X = (\mathbf{d}_1, \ldots, \mathbf{d}_N) \in \mathcal{X}^N$, where $\mathbf{d}_i \in \mathcal{X}$. The parameters of a model are denoted as $\boldsymbol{\theta} \in \mathbb{R}^r$, *e.g.*, the weights of a deep neural network.

2.1 Differential Privacy

The concept of DP was proposed by Dwork (2008) to describe the privacy modeling property of a randomized mechanism (algorithm) on two adjacent datasets. Here two datasets X and X' are called adjacent if they only differ by one record, *e.g.*, $\mathbf{d}_i \neq \mathbf{d}_i'$ for some i, where $\mathbf{d}_i \in X$ and $\mathbf{d}_i' \in X'$.

Definition 1 (Differential Privacy) *For any pair of adjacent datasets X and X', a randomized mechanism $\mathcal{M} : \mathcal{X}^N \to \mathcal{Y}$ mapping from data space to its range $\mathcal{Y}$ satisfies (ϵ, δ)-differential privacy if for all measurable $\mathcal{S} \subset range(\mathcal{M})$ and all adjacent X and X', we have*

$$Pr(\mathcal{M}(X) \in \mathcal{S}) \leq e^{\epsilon} Pr(\mathcal{M}(X') \in \mathcal{S}) + \delta$$

where $Pr(e)$ denotes the probability of event e, and ϵ and δ are two positive real numbers that indicate the loss of privacy. When $\delta = 0$, we say $\mathcal{M}$ has ϵ-differential privacy.

Differential privacy places constraints on the difference between the output distributions of two adjacent inputs X and X' by a random mechanism. If we assume that X and X' only differ by one record $\mathbf{d}_i$, by observing the output, any outside attackers are not able to recognize whether the output has resulted from X and X', as long as ϵ and δ are small enough (making these two probabilities close to each other). Thus, the existence of the record $\mathbf{d}_i$ is protected. Since the record in which the two datasets differ by is arbitrary, the privacy protection is applicable for all records. To better describe the randomness of $\mathcal{M}$'s output with inputs X and X', we define the privacy loss below.

Definition 2 (Privacy Loss) *Given a randomized mechanism $\mathcal{M}$ and a pair of adjacent datasets X and X', let* aux *denote any auxiliary input independent of X or X'. For an outcome $o \in \mathcal{Y}$ from the mechanism $\mathcal{M}$, the privacy loss at o is defined as:*

$$c(o; \mathcal{M}, \mathsf{aux}, X, X') \triangleq \log \frac{Pr[\mathcal{M}(\mathsf{aux}, X) = o]}{Pr[\mathcal{M}(\mathsf{aux}, X') = o]}$$

It can be shown that the (ϵ, δ)-DP is equivalent to the tail bound of the distribution of its corresponding privacy loss random variable (Abadi et al., 2016) (see Theorem 1 in the next section), thus this random variable is an important tool for quantifying the privacy loss of a mechanism.

2.2 Moments Accountant Method

To achieve differential privacy, random noise is introduced to hide the existence of a particular data point. For example, Laplace and Gaussian mechanisms (Dwork et al., 2014) add *i.i.d.* Laplace random noise and Gaussian noise, respectively, to a finite vector. While a large amount of noise makes an algorithm differentially private, it may sacrifice the utility of the algorithm. Therefore, in such paradigms, it is important to calculate the smallest amount of noise that is required to achieve a certain level of differential privacy.

The moments accountant method proposed in (Abadi et al., 2016) keeps track of a bound on the moments of the random variables defined below. As a result, it allows one to calculate the amount of noise needed to ensure the privacy loss under a given threshold.

Definition 3 (Moments Accountant) *Let $\mathcal{M} : \mathcal{X}^N \to \mathcal{Y}$ be a randomized mechanism, and let X and X' be a pair of adjacent datasets. Let* aux *denote any auxiliary input that is independent of both X and X'. The moments accountant parameterized by $\lambda > 0$ is defined as $\alpha_{\mathcal{M}}(\lambda) \triangleq \max_{\mathsf{aux}, X, X'} \alpha_{\mathcal{M}}(\lambda; \mathsf{aux}, X, X')$, where $\alpha_{\mathcal{M}}(\lambda; \mathsf{aux}, X, X') \triangleq \log \mathbb{E}[\exp(\lambda c(\mathcal{M}, \mathsf{aux}, X, X'))]$ is the log of the moment generating function at λ.*

Theorem 1 (Abadi et al. (2016))
[Composability] *Suppose that $\mathcal{M}$ consists of a sequence of adaptive mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_k$ where $\mathcal{M}_i : \prod_{j=1}^{i-1} \mathcal{Y}_j \times \mathcal{X} \to \mathcal{Y}_i$, and $\mathcal{Y}_i$ is the range of the ith mechanism, i.e., $\mathcal{M} = \mathcal{M}_k \circ \cdots \circ \mathcal{M}_1$, with $\circ$ the composition operator. Then, for any λ, we have*

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

where the input for $\alpha_{\mathcal{M}_i}$ is defined as all $\alpha_{\mathcal{M}_j}$'s outputs, $\{o_j\}$, for $j < i$; and $\alpha_{\mathcal{M}}$ takes $\mathcal{M}_i'$s output, $\{o_i\}$ for $i < k$, as the auxiliary input.

[Tail bound] *For any $\epsilon > 0$, the mechanism $\mathcal{M}$ is (ϵ, δ)-DP for $\delta = \min_{\lambda > 0} \exp\left(\alpha_{\mathcal{M}}(\lambda) - \lambda\epsilon\right)$.*

For the remainder of this paper, for simplicity we only consider mechanisms that output a real-valued vector. That is, $\mathcal{M} : \mathcal{X}^N \rightarrow \mathbb{R}^p$. Using the properties above, the following lemma about the moments accountant has been proven in (Abadi et al., 2016):

Lemma 2 *Suppose that $f : \mathcal{X}^N \rightarrow \mathbb{R}^p$ with $\|f(.)\|_2 \leq 1$. Let $\sigma \geq 1$ and J is a mini-batch sample with sampling probability q, i.e., $q = \frac{\tau}{N}$ with minibatch size of τ. If $q < \frac{1}{16\sigma}$, for any positive real number $\lambda \leq \sigma^2 \ln \frac{1}{q\sigma}$, the mechanism $\mathcal{M}(X) = \sum_{i \in J} f(\mathbf{d}_i) + N(0, \sigma^2 I)$ satisfies*

$$\alpha_{\mathcal{M}}(\lambda) \leq \frac{q^2 \lambda(\lambda+1)}{(1-q)\sigma^2} + O(q^3)$$

Remark 1 *Since q is often a small number, we use the approximate bound $\alpha_{\mathcal{M}}(\lambda) \leq \frac{q^2\lambda(\lambda+1)}{\sigma^2}$ in the rest of this paper. In our experiments, the exact bound is numerically calculated based on the code from Abadi et al. (2016)*

2.3 Stochastic Gradient Markov Chain Monte Carlo

SG-MCMC is a family of scalable Bayesian sampling algorithms, developed recently to generate approximate samples from a posterior distribution $p(\boldsymbol{\theta}|X)$, with $\boldsymbol{\theta}$ a model parameter vector. They are discretized numerical approximations of continuous-time Itô diffusions (Chen et al., 2015; Ma et al., 2015), whose stationary distributions are designed to coincide with $p(\boldsymbol{\theta}|X)$. Formally, an Itô diffusion is written as

$$\mathrm{d}\boldsymbol{\Theta}_t = F(\boldsymbol{\Theta}_t)\mathrm{d}t + g(\boldsymbol{\Theta}_t)\mathrm{d}\mathcal{W}_t \ , \tag{1}$$

with t the time index; $\boldsymbol{\Theta}_t \in \mathbb{R}^p$ represents the full variables in a system, where typically $\boldsymbol{\Theta}_t \supseteq \boldsymbol{\theta}_t$ (thus $p \geq r$) is an augmentation of the model parameters; and $\mathcal{W}_t \in \mathbb{R}^p$ is p-dimensional Brownian motion. Functions $F : \mathbb{R}^p \rightarrow \mathbb{R}^p$ and $g : \mathbb{R}^p \rightarrow \mathbb{R}^{p \times p}$ are assumed to satisfy the Lipschitz continuity condition (Ghosh, 2011). For example, the stochastic gradient Langevin dynamic (SGLD) algorithm defines $\boldsymbol{\Theta} = \boldsymbol{\theta}$, and $F(\boldsymbol{\Theta}_t) = -\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$, $g(\boldsymbol{\Theta}_t) = \sqrt{2}\,\mathbf{I}_r$, where $U(\boldsymbol{\theta}) \triangleq -\log p(\boldsymbol{\theta}) - \sum_{i=1}^N \log p(\mathbf{d}_i\,|\boldsymbol{\theta})$ denotes the unnormalized negative log-posterior, and $p(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$. which defines $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \mathbf{q})$, and $F(\boldsymbol{\Theta}_t) = \begin{pmatrix} \mathbf{q} \\ -B\,\mathbf{q} - \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) \end{pmatrix}$, $g(\boldsymbol{\Theta}_t) = \sqrt{2B}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{pmatrix}$ for a scalar $B > 0$; $\mathbf{q}$ is an auxiliary variable known as the momentum (Chen et al., 2014; Ding et al., 2014). Similar formulae can be defined for other SG-MCMC algorithms, such as the stochastic gradient thermostat (Ding et al., 2014), and other variants with Riemannian information geometry (Patterson and Teh, 2013; Ma et al., 2015; Li et al., 2016).

To make the algorithms, for example SGLD, scalable in a large-datasetting, *i.e.*, when N is large, an unbiased version of $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$ is calculated with a random subset of the full data, denoted $\nabla_{\boldsymbol{\theta}} \tilde{U}(\boldsymbol{\theta})$ and defined as $\nabla_{\boldsymbol{\theta}} \tilde{U}(\boldsymbol{\theta}) = \nabla \log p(\boldsymbol{\theta}) + \frac{N}{\tau} \sum_{\mathbf{d}_i \in J} \log p(\mathbf{d}_i\,|\boldsymbol{\theta})$, where J is a random minibatch of the data with size τ (typically $\tau \ll N$).

Algorithm 1 SGLD with Differential Privacy

```
Require: Data $X$ of size $N$, size of mini-batch $\tau$, number of iterations $T$, prior $p(\boldsymbol{\theta})$, privacy budget $\epsilon_0, \delta_0$, gradient norm bound $L$. A decreasing/fixed-step-size sequence $\{\eta_t\}$. Set $t = 1$.
for $t \in [T]$ do
    Take a random sample $J_t$ with sampling probability $q = \tau/N$. For each $i$ in $J_t$:
    Calculate $g_t(\mathbf{d}_i) \leftarrow \nabla \log \ell(\boldsymbol{\theta}_t|\,\mathbf{d}_i)$
    Clip norm: $\tilde{g}_t(\mathbf{d}_i) \leftarrow g_t(\mathbf{d}_i)/\max\left(1, \frac{\|g_t(\mathbf{d}_i)\|_2}{L}\right)$
    Sample each coordinate of $\mathbf{z}_t$ iid from $N(0, \frac{\eta_t}{N})$
    Update $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta_t\left(\frac{\nabla \log p(\boldsymbol{\theta})}{N} + \frac{1}{\tau}\sum_{i \in J_t} \tilde{g}_t(\mathbf{d}_i)\right) + \mathbf{z}_t$
    Return $\boldsymbol{\theta}_{t+1}$ as a posterior sample (after burn in).
end for
Compute the overall privacy cost $(\epsilon, \delta)$ using the moment accountant method. Ensure $\epsilon \leq \epsilon_0$ and $\delta \leq \delta_0$.
Output $\boldsymbol{\theta}_{T+1}$.
```

We typically adopt the popular Euler method to solve the continuous-time diffusion by an η-time discretization (stepsize being η). The Euler method is a first-order numerical integrator, thus inducing an $O(\eta)$ numerical error (Chen et al., 2015). Algorithm 1 illustrates the application of the SGLD algorithm with the Euler integrator for differential privacy, which is almost the same as the original SGLD, except that there is a gradient norm clipping in Step 4 of the algorithm. The norm-clipping step ensures that the computed gradients satisfy the Lipschitz condition, a common assumption on loss functions in a differential-privacy setting (Song et al., 2013; Bassily et al., 2014; Wang et al., 2015). The reasoning is intuitive: since differential privacy requires the output to be non-sensitive to any changes on an arbitrary data point, it is thus crucial to bound the impact of a single data point to the target function. The Lipschitz condition is easily met by clipping the norm of a loss function, a common technique for gradient-based algorithms to prevent gradient explosion (Pascanu et al., 2013). Note the clipping is introduced only for practical reasons. The Lipschitz property is typically assumed in SG-MCMC for the feasibility of theoretical analysis Chen et al. (2015), thus *no clipping* is needed under the Lipschitz assumption. Consequently, the only difference between our DP version of SGLD and standard SGLD is the choice of stepsize sequence, necessary to maintain the DP property. More details are discussed in Section 3.2.

3 Privacy Analysis for Stochastic Gradient Langevin Dynamics

We first develop theory to prove Algorithm 1 is (ϵ, δ)-DP under a certain condition. Our theory shows a significant improvement of the differential privacy obtained by SGLD over the most related work by Wang et al. (2015). To study the estimation accuracy (utility) of the algorithm, the corresponding mean square error estimation bounds are then proved under such differential-privacy settings.

3.1 Stepsize bounds for differentially-private SGLD

Previous work on SG-MCMC has shown that an appropriately chosen decreasing stepsize sequence can be adopted for an SG-MCMC algorithm (Teh et al., 2016; Chen et al., 2015). For the sequence in the form of $\eta_t = O(t^{-\alpha})$, the optimal value is $\alpha = \frac{1}{3}$ in order to obtain the optimal mean square error bound (defined in Section 3.2). Consequently, we first consider $\eta_t = O(t^{-1/3})$ in our below analysis, where the constant of the stepsize can be specified with parameters of the DP setting, shown in Theorem 3. The differential privacy property under a fixed stepsize is also discussed subsequently.

Theorem 3 *If we let the stepsize decrease at the rate of $O(t^{-1/3})$, there exist positive constants c_1 and c_2 such that given the sampling probability $q = \tau/N$ and the number of iterations T, for any $\epsilon < c_1 q^2 T^{2/3}$, Algorithm 1 satisfies (ϵ, δ)-DP as long as η_t satisfies:*

1. $\eta_t \leq \frac{N}{L^2}$
2. $\eta_t > \frac{q^2 N}{256 L^2}$
3. $\eta_t < \frac{\epsilon^2 N t^{-1/3}}{c_2^2 L^2 T^{2/3} \log(1/\delta)}$.

Remark 2 *In practice, the first condition is easy to satisfy, as $\frac{N}{L^2}$ is often much larger than the stepsize, especially in a large-data setting (N is large). The second condition is also easy to satisfy with properly chosen L and q, and we verify this condition in our experiments. In the rest of this section, we only focus on the third condition as an upper bound to the stepsize.* It is now clear that with optimal decreasing stepsize sequence (in terms of MSE defined in Section 3.2), Algorithm 1 maintains (ϵ, δ)-DP. There are other variants of SG-MCMC which use fixed stepsizes. We show in Theorem 4 that in this case, the algorithm still satisfies (ϵ, δ)-DP.

Theorem 4 *Under the same setting as Theorem 3, but using a fixed-stepsize $\eta_t = \eta$, Algorithm 1 satisfies (ϵ, δ)-DP whenever the stepsize satisfies i) and ii) in Theorem 3, as well as $\eta < \frac{\epsilon^2 N}{c^2 L^2 T log(1/\delta)}$ for another constant c.*

In (Wang et al., 2015), the authors proved that the SGLD method is (ϵ, δ)-DP if the stepsize η_t is small enough to satisfy $\eta_t < \frac{\epsilon^2 N}{128 L^2 T \log(2.5T/\delta) \log(2/\delta)}$. This bound is relatively small compared to ours (explained below), thus it is not practical in real applications. To address this problem, Wang et al. (2015) proposed the Hybrid Posterior Sampling algorithm, that uses the One Posterior Sample (OPS) estimator for the "burn-in" period, followed by the SGLD with a small stepsize to guarantee the differential privacy property. We note that for complicated models, especially with non-convex target posterior landscapes, such an upper bound for the stepsize still brings practical problems, even with the OPS. One issue is that the Markov chain will mix very slowly with a small stepsize, leading to highly correlated samples.

By contrast, our new upper bound for the stepsize in Theorem 3, $\eta_t < \frac{\epsilon^2 N t^{-1/3}}{c_2^2 L^2 T^{2/3} \log(1/\delta)}$, improves the bound in (Wang et al., 2015) by a factor of $T^{1/3} \log(T/\delta)$ at the first iteration. Note the constant c_2^2 in our bound is empirically smaller than 128 (see the calculating method in Section C of the SM), thus still giving a larger bound overall.

To provide intuition on how our bound compares with that in (Wang et al., 2015), consider the MNIST dataset with $N = 50,000$. If we set $\epsilon = 0.1$, $\delta = 10^{-5}$, $T = 10000$, and $L = 1$, our upper bound for decreasing stepsize can be calculated as $\eta_t < 0.103$, consistent with the default stepsize when training MNIST (Li et al., 2016). More importantly, our theory indicates that using SGLD with the default stepsize $\eta_t = 0.1$ is able to achieve (ϵ, δ)-DP with a small privacy loss for the MNIST dataset. As a comparison, Wang et al. (2015) gives a much smaller upper bound of $\eta_t < 1.54 \times 10^{-6}$, which is too small too be practically used. More detailed comparisons for these two bounds is given in Section 4.1, when considering experimental results. Finally, note that as in (Wang et al., 2015), our analysis can be easily extended to other SG-MCMC methods such as SGHMC (Chen et al., 2014) and SGNHT (Ding et al., 2014). We do not specify the results here, for conciseness.

3.2 Utility Bounds

The above theory indicates that, with a smaller stepsize, one can manifest an SG-MCMC algorithm that preserves more privacy, *e.g.*, $(0, \delta)$-DP in the limit of zero stepsize. However, this does not mean one can choose arbitrarily small stepsizes, because this would hinder the exploration of the parameter space, leading to slow mixing and potentially worse generalization. We investigate utility bounds w.r.t. mixing (how a sample estimate approximates the true posterior) and a generalization property (how a specific sample generalizes to unseen data for optimization) of the differentially-private SG-MCMC.

Mixing bound with deceasing stepsizes Following standard settings for SG-MCMC (Chen et al., 2015; Vollmer et al., 2016), we use the *mean square error* (MSE) under a target posterior distribution to measure the estimation accuracy for a Bayesian model. Specifically, our utility goal is to evaluate the *posterior average* of a test function $\phi(\boldsymbol{\theta})$, defined as $\bar{\phi} \triangleq \int \phi(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta}$, with a posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$. The posterior average is typically infeasible to compute, thus we use the *sample average*, $\hat{\phi}_T \triangleq \frac{1}{\sum_t \eta_t} \sum_{t=1}^{T} \eta_t \phi(\boldsymbol{\theta}_t)$, to approximate $\bar{\phi}$, where $\{\boldsymbol{\theta}_t\}_{t=1}^T$ are the samples from an SG-MCMC algorithm. The MSE we desire is defined as $\mathbb{E}\left(\hat{\phi}_T - \bar{\phi}\right)^2$. We impose the same assumptions on an SG-MCMC algorithm as in previous work (Vollmer et al., 2016; Chen et al., 2015), which are detailed in Section D of the SM. We assume both the corresponding Itô diffusion (in terms of its coefficients) and the numerical method of an SG-MCMC algorithm to be well behaved.

Proposition 5 *Under Assumption 1 in the SM, the MSE of SGLD with a decreasing stepsize sequence $\{\eta_t < \frac{\epsilon^2 N t^{-1/3}}{c_2^2 L^2 T^{2/3} \log(1/\delta)}\}$ as in Theorem 3 is bounded, for a constant C independent of $\{\eta, T, \tau\}$ and a bounded constant Γ_M depending on $U(\cdot)$ (see the proof for details), as $\mathbb{E}\left(\hat{\phi}_L - \bar{\phi}\right)^2 \leq C\left(\frac{2}{3}\left(\frac{N}{\tau} - 1\right) N^2 \Gamma_M T^{-1} + \frac{1}{3\tilde{\eta}_0} + 2\tilde{\eta}_0^2 T^{-2/3}\right)$, where $\tilde{\eta}_0 \triangleq \frac{\epsilon^2}{c_2^2 L^2 \log(1/\delta)}$.*

The bound in Proposition 5 indicates how the MSE decreases w.r.t. the number of iterations T and other parameters. It is consistent with standard SG-MCMC, leading to a similar convergence rate. Interestingly, we can also derive the optimal bounds w.r.t. the privacy parameters. For example, the optimal value for $\tilde{\eta}_0$ when fixing other parameters can be seen as $\tilde{\eta}_0 = O\left(T^{2/9}\right)$. Consequently, we have $\epsilon^2 = O\left(L^2 T^{2/9} \log(1/\delta)\right)$ in the optimal MSE setting. Different from the bound of standard SG-MCMC (Chen et al., 2015), when considering a (ϵ, δ)-DP setting, the MSE bound induces an asymptotic bias term of $\frac{1}{3\tilde{\eta}_0}$ as long as $\frac{\log(1/\delta)}{\epsilon^2}$ does not approach zero.

Mixing bound with a fixed stepsize We also wish to study the MSE under the fixed-step-size case. Consider a general situation, *i.e.*, $\eta_t = \eta$, for which Chen et al. (2017) has proved the following MSE bound for a fixed steps size, rephrased in Proposition 6.

Proposition 6 *With the same Assumption as Proposition 5, the MSE of SGLD is bounded as*[*]*:*

$$\mathbb{E}\left(\hat{\phi}_L - \bar{\phi}\right)^2 \leq C\left(\frac{(\frac{N}{\tau} - 1)N^2\Gamma_M}{T} + \frac{1}{T\eta} + \eta^2\right) .$$

Furthermore, the optimal MSE w.r.t. the stepsize η is bounded by

$$\mathbb{E}\left(\hat{\phi}_L - \bar{\phi}\right)^2 \leq C\left(\frac{(\frac{N}{\tau} - 1)N^2\Gamma_M}{T} + T^{-2/3}\right) ,$$

with the optimal stepsize being $\eta = O(T^{-1/3})$.

From Proposition 6, the optimal stepsize, *i.e.*, $\eta = O(T^{-1/3})$, is of a lower order than both our differential-privacy-based algorithm ($\eta = O(T^{-1})$) and the algorithm in Wang et al. (2015), *i.e.*, $\eta = O(T^{-1} \log^{-1} T)$. This means that for T large enough, both ours and the method in (Wang et al., 2015) might not run on the optimal stepsize setting. A remedy for this is to increase the stepsize at the cost of increasing privacy loss. Because for the same privacy loss our stepsizes are typically larger than in (Wang et al., 2015), our algorithm is able to obtain both higher approximate accuracy and differential privacy. Specifically, to guarantee the desired differential-privacy property as stated in Theorem 4, we substitute a stepsize of $\eta = \frac{\epsilon^2 N}{c^2 L^2 T log(1/\delta)}$ into the MSE formula in Lemma 6. Consequently, the MSE is bounded by $\mathbb{E}\left(\hat{\phi}_L - \bar{\phi}\right)^2 \leq C\left(\frac{(\frac{N}{\tau}-1)N^2\Gamma_M}{T} + \frac{c_2^2 L^2 \log\frac{1}{\delta}}{\epsilon^2 N} + \frac{\epsilon^4 N^2}{c_2^4 L^4 T^2 \log^2(1/\delta)}\right)$, which is smaller than that in the method of Wang et al. (2015).

Generalization error bound In terms of generalization error, our objective is to minimize $U(\boldsymbol{\theta})$ in an infinite-sized dataset, *i.e.*, minimizing $\mathcal{F}(\boldsymbol{\theta}) \triangleq \mathbb{E}_P[\log p(\mathbf{d}\,|\boldsymbol{\theta})]$, where P is the unknown probability

[*] With a slight abuse of notation, the constant C is independent of $\{\eta, T, \tau\}$, but might be different from that in Proposition 5.

law of the data. Let $\mathcal{F}^* \triangleq \inf_{\boldsymbol{\theta}} \mathcal{F}(\boldsymbol{\theta})$, and $\hat{\boldsymbol{\theta}}_T$ be the final sample returned by our DP-SGLD. We investigate the generalization ability in terms of the expected excess risk: $\mathbb{E}\mathcal{F}(\hat{\boldsymbol{\theta}}_T) - \mathcal{F}^*$, where the expectation is taken over the stochasticity of the algorithm. Note different from (Raginsky et al., 2017), which uses a tempered version of SGLD for optimization, it still make sense to use our proposed DP-SGLD for optimization as our algorithm is a special case of tempered-SGLD with the temperature set to 1. In the following, we show that it is possible to use our proposed DP-SGLD for optimization, whose generalization error can be bounded.

Following the techniques presented in (Raginsky et al., 2017), with some standard assumptions detailed in Section F of the SM, we can derive a generalization-error bound for the proposed DP-SGLD, where we only consider the impact of the fixed-stepsize η, the total number of iterations T and the dataset size N.

Proposition 7 *Under Assumption 2 in Section F, for a positive ω small enough and satisfying $\omega \geq -\eta^{1/4}\log(\omega)$*[†] *such that $T = A\log^5\frac{1}{\omega}/\omega^4$ for some constant A independent of ω, and $\eta \leq \min\left\{\left(\frac{\omega}{\log(1/\omega)}\right)^4, \frac{\epsilon^2 N}{c^2 L^2 T \log(1/\delta)}\right\}$, the generalization error is bounded as*

$$\mathbb{E}\mathcal{F}(\hat{\boldsymbol{\theta}}_T) - \mathcal{F}^* \leq O\left(T^{1/5}\omega^{4/5} + \omega + \frac{1}{N}\right) =$$

$$O\left(W^{1/5}(\frac{4}{5A}T) + \exp\left\{-W^{1/5}(\frac{4}{5A}T)\right\} + \frac{1}{N}\right) ,$$

where $W(\cdot)$ is the Lambert W function (Corless et al., 1996).

Proposition 7 seems to indicate that the generalization error grows w.r.t. the number of iterations at a rate of $T^{1/5}$ when T is large. However, ω would become small as T grows. Consequently, one should choose an appropriate ω so that the terms $T^{1/5}\omega^{4/5}$ and ω in the bound reach a balance, achieving a minimum bound. Proposition 7 also indicates that there is always a nonzero gap in the bound of $\mathbb{E}\mathcal{F}(\hat{\boldsymbol{\theta}}_T) - \mathcal{F}^*$, even if we have infinite data.

4 Experiments

We test the proposed differentially-private SG-MCMC algorithms by considering several tasks, including logistic regression and deep neural networks, and compare with related Bayesian and optimization methods in terms of both algorithm privacy and utility. We first verify the stepsize bounds presented in Theorems 3 and 4.

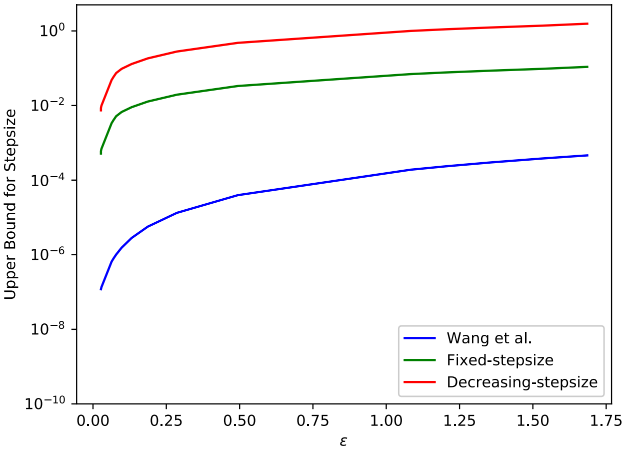

Figure 1: Upper bounds for fixed-stepsize and decreasing-stepsize (first step) with DP loss ϵ, as well as the upper bound from (Wang et al., 2015).

4.1 Stepsize Upper Bound

We compare our upper bound for the stepsize in Section 3.1 with the bound of Wang et al. (2015). Section C in the SM describes how to calculate the bound, which denotes the largest stepsize allowed to preserve (ϵ, δ)-DP.

In this simulation experiment, we use the following setting: $N = 50,000$, $T = 10,000$, $L = 1$, and $\delta = 10^{-5}$. We vary ϵ from 0.02 to 1.7 for different differential-privacy settings, for both ours (fixed and decreasing-stepsize cases) and the bound in (Wang et al., 2015), with results in the left plot in Figure 1. It is clear that our bounds give much larger stepsizes than from (Wang et al., 2015) at the same privacy loss, *e.g.*, 10^{-1} vs. 10^{-4}. Our stepsizes appear to be much more practical in real applications.

In the rest of our experiments, we focus on using the decreasing-stepsize SGLD as it gives a better MSE bound, as shown in Proposition 5. For the parameters in our bounds, *i.e.*, $(N, T, \epsilon, \delta, L)$, the default setting is often chosen to be $\delta = O(1/N)$ and $T = O(N)$; L is typically selected from a range such as $L \in \{0.1, 1, 10\}$. In this experiment, we investigate the sensitivity of our proposed upper bound w.r.t. N and L when fixing other parameters. The results are shown in the right plot in Figure 1, from which we observe that our proposed stepsize bound is stable in terms of the data size N, and is approximately proportional to $1/L$. Such a conclusion is not a direct implication from the upper bound formula in Theorem 3, as the constant c_2 also depends on $(N, T, \epsilon, \delta, L)$. The result also indicates a rule for choosing stepsizes in practice by using our upper bound, which fall into the range of $(10^{-4}, 0.1)$. When using such stepsizes, we observe that the standard

[†] The specific range is given in Section F in the SM.

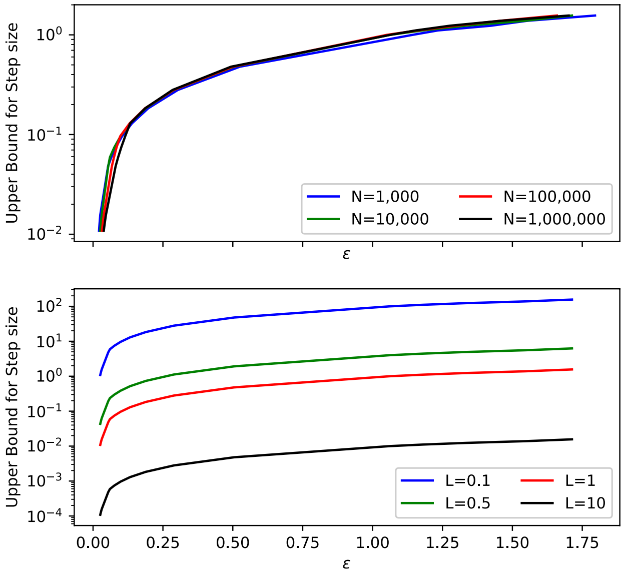

Figure 2: Stepsize upper bounds for $N = 10^3, 10^4, 10^5, 10^6$ with fixed $L = 1$ (top), and $L = 0.1, 0.5, 1.0, 10.0$ with fixed $N = 10^4$ (bottom). In both simulations, we let $\delta = 1/N$ and $T = N$.

SGLD automatically preserves (ϵ, δ)-DP even when ϵ is small.

4.2 Logistic Regression

In the remaining experiments, we compare our proposed differentially-private SGLD (DP-SGLD) with other methods. The Private Aggregation of Teacher Ensembles (PATE) model is proposed in (Papernot et al., 2016) for differentially private training of machine learning models. PATE takes advantage of the moment accountant method for privacy-loss calculation, and uses a knowledge-transfer technique via semi-supervised learning, to build a teacher-student-based model. This framework first trains multiple teachers with private data; these teachers then differential privately release aggregated knowledge, such as label assignments on several public data points, to multiple students. The students then use the released knowledge to train their models in a supervised-learning setting, or they can incorporate unlabeled data in a semi-supervised-learning setting. In (Papernot et al., 2018), the authors proposed improved analysis, named Confident-GNMax, on the PATE model, which gives the state-of-the-art privacy and performance balance. As the semi-supervised setting requires a large amount of non-private unlabeled data for training, which are not always available in practice, for fair comparison, we only consider supervised setting in this experiment.

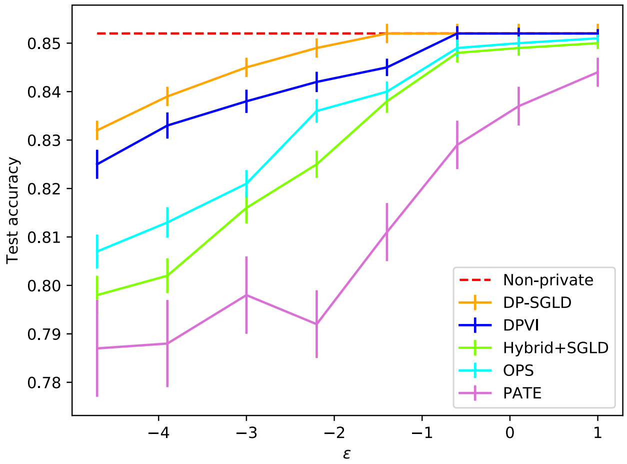

Figure 3: Test accuracies on a classification task based on Bayesian logistic regression for DPVI, One-Posterior Sample (OPS), Hybrid Posterior sampling based on SGLD, Confident-GNMax and our proposed DP-SGLD, considering different choices of privacy loss ϵ. The non-private baseline is obtained by standard SGLD.

We compare DP-SGLD with Confident-GNMax, the Hybrid Posterior Sampling algorithm (Wang et al., 2015), and recently proposed differentially private variational inference (DPVI) (Jälkö et al., 2016) on the Adult dataset from the UCI Machine Learning Repository (Lichman, 2013), for a binary classification task with Bayesian logistic regression, under the DP setting. We fix $\delta = 10^{-4}$, and compare the classification accuracy while varying ϵ. We repeat each experiment ten times, and report averages and standard deviations, as illustrated in Figure 3.

Our proposed DP-SGLD achieves a higher accuracy compared to other methods and is close to the baseline with plain SGLD. In fact, when $\epsilon \approx 0.08$ or above, our DP-SGLD becomes the standard SGLD, therefore has the same test accuracy as the baseline. Note that Confident-GNMax obtains the worst performance in this experiment. This might be because under a supervised setting with small ϵ and only labeled data, the students are restricted to use an extremely small amount of training data.

4.3 Deep Neural Networks

We compare our methods with Confident-GNMax (CGNMax) (Papernot et al., 2018) and the DP-SGD (Abadi et al., 2016) for training deep neural networks under DP settings. We use two datasets: (i) the standard MNIST dataset for handwritten digit recognition, consisting of 60,000 training examples and 10,000 testing examples (LeCun and Cortes, 2010); and (ii) the Street View House Number (SVHN) dataset, which contains 600,000 32×32 RGB images of printed digits obtained from pictures of house number in street view (Netzer et al.). We use the same network structure as for the Confident-GNMax model, which contains two stacked convolutional layers and one fully connected layer with ReLUs for MNIST, and two more convolutional layers for SVHN. We use standard Gaussian priors for the weights of the DNN. For the MNIST dataset, the standard SGLD is considered with stepsize $\eta_t = 0.3$, batch size 128, number of epochs 20, and $L = 0.3$. This setting satisfies (ϵ, δ)-DP for $\epsilon = 0.99$ and $\delta = 10^{-5}$. For the SVHN dataset, the standard SGLD with stepsize $\eta_t = 0.1$ satisfies (ϵ, δ)-DP for $\epsilon = 2.97$ and $\delta = 10^{-6}$ when we set $L = 5$. The test accuracies are shown in Table 1. In practice, we found keeping a constant stepsize instead of decreasing yields better privacy and utility balance.

Table 1: Test accuracies on MNIST and SVHN.

Dataset	Methods	ϵ	δ	Accuracy
MNIST	Non-Private			99.34%
	DP-SGD	0.5	10^{-5}	90.00%
	DP-SGD	8.0	10^{-5}	97.00%
	CGNMax	1.97	10^{-5}	98.51%
	DP-SGLD	**0.99**	10^{-5}	**99.21%**
SVHN	Non-Private			92.80%
	CGNMax	4.96	10^{-6}	91.62%
	DP-SGLD	**2.97**	10^{-6}	**91.89%**

It is shown that SGLD obtains better test accuracy than the state-of-the-art differential privacy methods, remarkably with much less privacy loss.

Application to generative-adversarial-network (GAN) training Our analysis also sheds lights on how SG-MCMC methods help improve the generalization for training generative models. For example, in (Saatchi and Wilson, 2017), a Bayesian GAN model trained with SGHMC is proposed and shows promising performance in avoiding mode-collapse problem. According to Arora et al. (2017), mode collapse is potentially due to weak generalization. As the connection between differential privacy and generalization of a model has been well acknowledged (Wang et al., 2016), it may imply Bayesian GAN moderates the mode-collapse problem, because SGHMC naturally leads to better generalization through DP. We perform additional experiments with GAN to verify our conjecture. Our experiment suggests under the same differential privacy setting ($\epsilon = 0.2, \delta = 10^{-5}$), GAN trained by SGHMC achieves 98.3% accuracy on the semi-supervised learning task with 100 labeled data on MNIST, outperforming the one trained by DP-SGD that achieves 90.8%.

5 Related Work

There are a number of papers dealing with differentially-private stochastic gradient based methods. For example, Song et al. (2013) proposed a differentially-private SGD algorithm, which requires a large amount of noise when mini-batches are sampled randomly. The theoretical performance of noisy SGD is studied in (Bassily et al., 2014) for the special case of convex loss functions. Therefore, for a non-convex loss function, a common setting for many machine learning models, there are no theoretical guarantees on performance. In (Abadi et al., 2016), another differentially private SGD was proposed, requiring a smaller variance for added Gaussian noise, yet it still did not provide theoretical guarantees on utility. On the other hand, the standard SG-MCMC has been shown to be able to converge to the target posterior distribution in theory. In this paper, we discuss the effect of our modification for differential privacy on the performance of the SG-MCMC, which endows theoretical guarantees on the bounds for the mean squared error of the posterior mean.

Bayesian modeling provides an effective framework for privacy-preserving data analysis, as posterior sampling naturally introduces noise into the system, leading to differential privacy (Dimitrakakis et al., 2014; Wang et al., 2015). In (Foulds et al., 2016), the privacy for sampling from exponential families with a Gibbs sampler was studied. In (Wang et al., 2015) a comprehensive analysis was proposed on the differential privacy of SG-MCMC methods. As a comparison, we have derived a tighter bound for the amount of noise required to guarantee a certain differential privacy, yielding a more practical upper bound for the stepsize.

6 Conclusion

Previous work on differential privacy has modified existing algorithms, or has built complicated frameworks that sacrifice performance for privacy. In some cases the privacy loss may be relatively large. This paper addresses a privacy analysis for SG-MCMC, a standard class of methods for scalable Bayesian posterior sampling. We have significantly relaxed the condition for SG-MCMC methods being differentially private, compared to previous works. Our results indicate that standard SG-MCMC methods have strong privacy guarantees for problems of large scale. In addition, we have proposed theoretical analysis on the estimation performance of differentially private SG-MCMC methods. Our results show that even when there is a strong privacy constraint, the differentially private SG-MCMC still endows a guarantee on the model performance. Our experiments have shown that standard SG-MCMC methods achieve both state-of-the-art utility and strong privacy compared with related methods on multiple tasks, such as logistic regression and deep neural networks.

Acknowledgments This research was supported in part by DARPA, DOE, NIH, NSF and ONR. We thank Ruiyi Zhang for providing the code base.

References

Martín 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. ACM, 2016.

S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *ICML*, 2017.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds. *arXiv preprint arXiv:1405.7085*, 2014.

P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. In *ICLR*, 2017.

C. Chen, N. Ding, and L. Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *NIPS*, 2015.

C. Chen, W. Wang, Y. Zhang, Q. Su, and L. Carin. A convergence analysis for a class of practical variance-reduction stochastic gradient mcmc. (arXiv:1709.01180), 2017.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1683–1691, Bejing, China, 22–24 Jun 2014. PMLR.

R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the lambertw function. *Advances in Computational Mathematics*, (5):329–359, 1996.

Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin IP Rubinstein. Robust and private bayesian inference. In *International Conference on Algorithmic Learning Theory*, pages 291–305. Springer, 2014.

N. Ding, Y. Fang, R. Babbush, C. Chen, R. D. Skeel, and H. Neven. Bayesian sampling using stochastic gradient thermostats. In *NIPS*, 2014.

Cynthia Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19. Springer, 2008.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. Springer, 2006.

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

James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. *arXiv preprint arXiv:1603.07294*, 2016.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333. ACM, 2015.

A. P. Ghosh. *Backward and Forward Equations for Diffusion Processes.* Wiley Encyclopedia of Operations Research and Management Science, 2011.

Joonas Jälkö, Onur Dikmen, and Antti Honkela. Differentially private variational inference for non-conjugate models. *arXiv preprint arXiv:1610.08749*, 2016.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

C. Li, C. Chen, D. Carlson, and L. Carin. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *AAAI*, 2016.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Y. A. Ma, T. Chen, and E. B. Fox. A complete recipe for stochastic gradient MCMC. In *NIPS*, 2015.

J. C. Mattingly, A. M. Stuart, and M. V. Tretyakov. Construction of numerical time-average and stationary measures via Poisson equations. *SIAM J. NUMER. ANAL.*, 48(2):552–577, 2010.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318, 2013.

S. Patterson and Y. W. Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *NIPS*, 2013.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Nonconvex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *COLT*, 2017.

Y. Saatchi and A. G. Wilson. Bayesian GAN. In *NIPS*, 2017.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 3–18. IEEE, 2017.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 245–248. IEEE, 2013.

Y. W. Teh, A. H. Thiery, and S. J. Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *JMLR*, (17):1–33, 2016.

S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh. Exploration of the (Non-)Asymptotic bias and variance of stochastic gradient Langevin dynamics. *JMLR*, 2016.

Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2493–2502, 2015.

Yu-Xiang Wang, Jing Lei, and Stephen E Fienberg. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of erm principle. *Journal of Machine Learning Research*, 17(183): 1–40, 2016.

A Proof of Theorem 3

We first prove Algorithm 1 is (ϵ, δ)-DP if we change the variance of $\mathbf{z}_t$ to be $\sigma_t^2 = \frac{c_2^2 L^2 T^{2/3} t^{1/3} \log(1/\delta)}{\epsilon^2 N^2} \eta_t^2 I$ for some constant c_2.

It is easy to see that SGLD in Algorithm 1 consists of a sequence of updates for the model parameter $\boldsymbol{\theta}$. Each update corresponds to a random mechanism $\mathcal{M}_i$ defined in Theorem 1, thus we will first derive the moments accountant for each iteration. In each iteration, the only data access is $\sum_{i\in J_t} \tilde{g}_t(\mathbf{d}_i)$ in Step 6. Therefore, in the following, we only focus on the interaction between $\sum_{i\in J_t} \tilde{g}_t(\mathbf{d}_i)$ and the noise $\mathbf{z}_t$, which is essentially[‡] $\frac{\eta_t}{\tau} \sum_{i\in J_t} \bar{g}_t(\mathbf{d}_i) + \mathbf{z}_t$, where $\bar{g} = \tau/\eta_t * \tilde{g}$.

To simplify the notation, we let $\tilde{\eta}^2 = \frac{\sigma_t^2 \tau^2}{L^2 \eta_t^2 t^{1/3}}$, and the variance of $\mathbf{z}_t$ can be rewritten as $\sigma_t^2 = (\tilde{\eta}^2 L^2 \eta_t^2 t^{1/3}/\tau^2) I^{\S}$. Then we have:

$$\begin{aligned}\frac{\eta_t}{\tau}\sum_{i\in J_t} \bar{g}_t(\mathbf{d}_i) + \mathbf{z}_t &= \frac{\eta_t}{\tau}\left(\sum_{i\in J_t} \bar{g}_t(\mathbf{d}_i) + N(0, (\sigma_t^2\tau^2/\eta_t^2)I)\right) \\ &= \frac{\eta_t L}{\tau}\left(\frac{1}{L}\sum_{i\in J_t} \bar{g}_t(\mathbf{d}_i) + N(0, \tilde{\eta}^2 t^{1/3} I)\right)\end{aligned}$$

If we use the notations from Lemma 2 and let $f(\mathbf{d}_i) = \frac{1}{L}\hat{g}_t(\mathbf{d}_i)$ and $\sigma^2 = \tilde{\eta}^2 t^{1/3}$, we can calculate the upper bound for the log moment of the privacy loss random variable for the t^{th} iteration to be

$$\alpha(\lambda) \leq t^{-1/3} q^2 \lambda(\lambda+1)/\tilde{\eta}^2$$

as long as the conditions in Lemma 2 are satisfied, that is $\tilde{\eta}^2 t^{1/3} \geq 1$ and the mini-batch sampling probability $q < \frac{1}{16\tilde{\eta} t^{1/6}}$. Later we will derive the corresponding bounds in terms of η_t.

Using the composability property of the moments accountant in Theorem 1, over T iterations, the log moment of the privacy loss random variable is bounded by

$$\alpha(\lambda) \leq \sum_{t=1}^{T} (t^{-1/3}) q^2 \lambda(\lambda+1)/\tilde{\eta}^2 \ .$$

According to the tail bound property in Theorem 1, δ is the minimum of $\exp\left(\alpha_{\mathcal{M}}(\lambda) - \lambda\epsilon\right)$ w.r.t. λ. To guarantee (ϵ, δ)-DP, it suffices that

$$\sum_{t=1}^{T} (t^{-1/3}) q^2 \lambda(\lambda+1)/\tilde{\eta}^2 \leq \lambda\epsilon/2, \quad \exp(-\lambda\epsilon/2) \leq \delta \ , \tag{2}$$

We also require that our choice of parameters satisfies Lemma 2. Consequently, we have

$$\lambda \leq \tilde{\eta}^2 t^{1/3} \log(1/q\tilde{\eta}^2 t^{1/3}) \leq \tilde{\eta}^2 \log(1/q\tilde{\eta}^2) \tag{3}$$

Since $\sum_{t=1}^{T} t^{-1/3} = O(T^{2/3})$, we can use a similar technique[¶] as in Abadi et al. (2016) to find explicit constants c_1 and c_2 such that when $\epsilon = c_1 q^2 T^{2/3}$ and $\tilde{\eta} = c_2 \frac{q\sqrt{T^{2/3}\log(1/\delta)}}{\epsilon}$, the conditions (2) (3) are satisfied. If we plug in $\tilde{\eta}$ and q, we have proved that Algorithm 1 is (ϵ, δ)-DP when $\mathbf{z}_i \sim N(0, \frac{c_2^2 L^2 T^{2/3} t^{1/3} \log(1/\delta)}{\epsilon^2 N^2} \eta_t^2 I)$.

For the second step of the proof, we prove that Algorithm 1 is (ϵ, δ)-DP when the original variance of $\mathbf{z}_t$ is used, *i.e.*, $\sigma_t^2 = \frac{\eta_t}{N}$. This is straightforward because when $\eta_t < \frac{\epsilon^2 N t^{-1/3}}{c_2^2 L^2 T^{2/3} \log(1/\delta)}$ we have $\frac{c_2^2 L^2 T^{2/3} t^{1/3} \log(1/\delta)}{\epsilon^2 N^2} \eta_t^2 < \eta_t/N$ as long as the stepsize η_t is positive. Adding more noise decreases the privacy loss. To satisfy (ϵ, δ)-DP, it suffices to set the variance of $\mathbf{z}_i$ as η_t/N, which gives the original Algorithm 1, a variant of the standard SGLD algorithm with decreasing stepsize. This finishes the proof for the third condition in Theorem 3.

Now we prove the first and second conditions in Theorem 3. Lemma 2 requires that $\sigma \geq 1$ and $q < \frac{1}{16\sigma}$, where $\sigma^2 = \tilde{\eta}^2 t^{1/3}$ by definition. This is equivalent to $\tilde{\eta}^2 t^{1/3} \geq 1$ and $q < \frac{1}{16\tilde{\eta} t^{1/6}}$. If we plug in the formula $\eta_t = \frac{N}{t^{1/3}\tilde{\eta}^2 L^2}$, this simplifies to $\eta_t \leq \frac{N}{L^2}$ and $\eta_t > \frac{q^2 N}{256 L^2}$. This completes the proof.

B Proof of Theorem 4

Claim: Under the same setting as Theorem 3, but using a fixed-stepsize $\eta_t = \eta$, Algorithm 1 satisfies (ϵ, δ)-DP whenever $\eta < \frac{\epsilon^2 N}{c^2 L^2 T log(1/\delta)}$ for another constant c.

Proof The only change of the proof for fixed stepsize is that the expression for the variance of the Gaussian noise $\mathbf{z}_t$ becomes $\sigma_t^2 = \eta_0^2 L^2 \eta_t^2/\tau^2$ for fixed stepsize. We still apply Theorem 1 and Lemma 2 to find the required conditions for (ϵ, δ)-DP:

$$T q^2 \lambda^2/\eta_0^2 \leq \lambda\epsilon/2$$

$$\exp(-\lambda\epsilon/2) \leq \delta, \lambda \leq \eta_0^2 \log(1/q\eta_0)$$

[‡] In this paper, we only consider the case for which we choose priors that do not depend on the data, as is common in the Bayesian setting.

[§] Later we will show the optimal decreasing ratio for the stepsize is $t^{1/3}$.

[¶] Further explained in Section C of the SM.

Using the method described in the previous section, one can find c_3 and c_4 such that when $\epsilon = c_3 q^2 T$ and $\eta_0 = c_4 \frac{q\sqrt{\log(1/\delta)}}{\epsilon}$ satisfy the above conditions. Then if we plug in η_0 and q, and compare it to η/N, it is easy to see Algorithm 1 satisfies (ϵ, δ)-DP when $\eta < \frac{\epsilon^2 N}{c_4^2 L^2 T \log(1/\delta)}$. ■

C Calculating Constants in Moment Accountant Methods

For calculating the constants c_1 and c_2, which is a part of the moment accoutant method, we refer to `https://github.com/tensorflow/models/tree/master/research/differential_privacy/privacy_accountant` [‖] as an implimentation of the moment accountant method. A comprehensive description for the implimentation can be found int Abadi et al. (2016).

This code allows one to calculate the corresponding $\epsilon(\delta)$ given $\delta(\epsilon), q, T, \eta_0$ using numerical integration. Once $\epsilon(\delta)$ is determined, it is easy to calculate c_1 and c_2 for evaluating the upper bound for the stepsize.

D Assumptions on SG-MCMC Algorithms

For the diffusion in (1), we first define the generator $\mathcal{L}$ as:

$$\mathcal{L}\psi \triangleq \frac{1}{2}\nabla\psi \cdot F + \frac{1}{2} g(\boldsymbol{\theta}) g(\boldsymbol{\theta})^* : D^2\psi \ , \tag{4}$$

where ψ is a measurable function, $D^k\psi$ means the k-derivative of ψ, $*$ means transpose. $\mathbf{a}\cdot\mathbf{b} \triangleq \mathbf{a}^T\mathbf{b}$ for two vectors $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{A} : \mathbf{B} \triangleq \mathrm{trace}(\mathbf{A}^T\mathbf{B})$ for two matrices $\mathbf{A}$ and $\mathbf{B}$. Under certain assumptions, there exists a function, ϕ, such that the following Poisson equation is satisfied Mattingly et al. (2010):

$$\mathcal{L}\psi = \phi - \bar{\phi} \ , \tag{5}$$

where $\bar{\phi} \triangleq \int \phi(\boldsymbol{\theta})\rho(\mathrm{d}\boldsymbol{\theta})$ denotes the model average, with ρ being the equilibrium distribution for the diffusion (1), which is assumed to coincide with the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$. The following assumptions are made for the SG-MCMC algorithms (Vollmer et al., 2016; Chen et al., 2015).

Assumption 1 *The diffusion* (1) *is ergodic. Furthermore, the solution of* (5) *exists, and the solution functional ψ satisfies the following properties:*

- *ψ and its up to 3th-order derivatives $\mathcal{D}^k\psi$, are bounded by a function $\mathcal{V}$, i.e., $\|\mathcal{D}^k\psi\| \leq C_k\mathcal{V}^{p_k}$ for $k = (0, 1, 2, 3)$, $C_k, p_k > 0$.*
- *The expectation of $\mathcal{V}$ on $\{\mathbf{x}_l\}$ is bounded: $\sup_l \mathbb{E}\mathcal{V}^p(\mathbf{x}_l) < \infty$.*
- *$\mathcal{V}$ is smooth such that $\sup_{s\in(0,1)} \mathcal{V}^p\left(s\,\mathbf{x} + (1-s)\,\mathbf{y}\right) \leq C\left(\mathcal{V}^p\left(\mathbf{x}\right) + \mathcal{V}^p\left(\mathbf{y}\right)\right),\ \forall \mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^m, p \leq \max\{2p_k\}$ for some $C > 0$.*

E Proof of Proposition 5

Claim: Under Assumption 1 in the section D, the MSE of SGLD with a decreasing stepsize sequence $\{\eta_t < \frac{\epsilon^2 N t^{-1/3}}{c_2^2 L^2 T^{2/3} \log(1/\delta)}\}$ as in Theorem 3 is bounded, for a constant C independent of $\{\eta, T, \tau\}$ and a constant Γ_M depending on T and $U(\cdot)$, as $\mathbb{E}\left(\hat{\phi}_L - \bar{\phi}\right)^2$

$$\leq C\left(\frac{2}{3}\left(\frac{N}{n} - 1\right) N^2\Gamma_M T^{-1} + \frac{1}{3\tilde{\eta}_0} + 2\tilde{\eta}_0^2 T^{-2/3}\right) .$$

where $\tilde{\eta}_0 \triangleq \frac{\epsilon^2}{c_2^2 L^2 \log(1/\delta)}$.

Proof

First, we adopt the MSE formula for the decreasing-step-size SG-MCMC with Euler integrator (1-st order integrator) from Theorem 5 of Chen et al. (2015), which is written as

$$\mathbb{E}\left(\hat{\phi}_L - \bar{\phi}\right)^2 \leq C\left(\sum_{t=1}^{T} \frac{\eta_t^2}{S_T^2}\mathbb{E}\left\|\Delta V_t\right\|^2 + \frac{1}{S_T} + \frac{(\sum_{t=1}^T \eta_t^2)^2}{S_T^2}\right) , \tag{6}$$

where $S_T \triangleq \sum_{t=1}^T \eta_t$, and ΔV_t is a term related to $\tilde{g}_t$, which, according to Theorem 3 of Chen et al. (2017), can be simplified as

$$\begin{aligned} &\mathbb{E}\left|\Delta V_l\right|^2 \\ =&\frac{(N-\tau)N^2}{\tau}\left(\frac{1}{N^2}\sum_{i,j}\mathbb{E}\boldsymbol{\alpha}_{li}^T\boldsymbol{\alpha}_{lj} - \frac{2}{N(N-1)}\sum_{i\leq j}\mathbb{E}\boldsymbol{\alpha}_{li}^T\boldsymbol{\alpha}_{lj}\right) \\ \triangleq&\frac{(N-\tau)N^2}{\tau}\Gamma_t \ , \end{aligned} \tag{7}$$

where $\boldsymbol{\alpha}_{li} = \nabla_{\boldsymbol{\theta}} \log p(\mathbf{d}_i \,|\boldsymbol{\theta}_\ell)$.

Let $\Gamma_M \triangleq \max_t \Gamma_t$. Substituting (7) into (6), we have

$$\mathbb{E}\left(\hat{\phi}_L - \bar{\phi}\right)^2 \leq \tag{8}$$

$$C\left(\frac{\sum_t^T \eta_t^2}{\left(\sum_t^T \eta_t\right)^2}\left(\frac{N}{\tau} - 1\right) N^2\Gamma_M + \frac{1}{\sum_t^T \eta_t} + \frac{\left(\sum_t^T \eta_t^2\right)^2}{\left(\sum_t^T \eta_t\right)}\right)$$

[‖] This is under the Apache License, Version 2.0

Now, if we assume $\tilde{\eta}_0 = \frac{\epsilon}{c_2^2 L^2 \log(1/\delta)}$, then we rewrite $\eta_t = \eta_0 t^{-1/3} T^{-2/3}$.

Note $\sum_t^T t^p \approx \frac{1}{p+1} T^{p+1}$. Plug this into the bound in (8), we have:

$$
\begin{aligned}
&\mathbb{E}\left(\hat{\phi}_L - \bar{\phi}\right)^2 \leq \\
&C\left(\frac{\sum_t^T \eta_t^2}{\left(\sum_t^T \eta\right)^2}\left(\frac{N}{\tau} - 1\right) N^2 \Gamma_M + \frac{1}{\sum_t^T \eta_t} + \frac{\left(\sum_t^T \eta_t^2\right)^2}{\left(\sum_t^T \eta_t\right)^2}\right) \\
\leq &C\left(\frac{2}{3}\left(\frac{N}{\tau} - 1\right) N^2 \Gamma_M T^{-1} + \frac{1}{3\tilde{\eta}_0} + 2\tilde{\eta}_0^2 T^{-2/3}\right)
\end{aligned}
$$

■

F Generalization Bound

Following Raginsky et al. (2017), we need to make the following assumptions to derive our generalization bound. Actually, some of these assumptions are related to Assumption 1. Interested readers are encouraged to refer Section 9 of Vollmer et al. (2016) for details.

Assumption 2 *Assume the likelihood function satisfies:*

A.1 *Let $\boldsymbol{\theta}_0$ be the initial value. There exists $A, L \geq 0$ such that*

$$
|\log p(\mathbf{d}\,|\boldsymbol{\theta}_0)| \leq A, \qquad \|\nabla_{\boldsymbol{\theta}} \log p(\mathbf{d}\,|\boldsymbol{\theta}_0)\| \leq L, \quad \forall\, \mathbf{d}
$$

A.2 *For some $M > 0$, $\forall\, \mathbf{d}_1, \mathbf{d}_2$*

$$
\|\nabla_{\boldsymbol{\theta}} \log p(\mathbf{d}_1\,|\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \log p(\mathbf{d}_2\,|\boldsymbol{\theta})\| \leq M\,\|\mathbf{d}_1 - \mathbf{d}_2\|
$$

A.3 *For some $m > 0$ and $b \geq 0$,*

$$
\langle \mathbf{d}, \nabla_{\boldsymbol{\theta}} \log p(\mathbf{d}\,|\boldsymbol{\theta})\rangle \geq m\,\|\mathbf{d}\|^2 - b, \quad \forall\, \mathbf{d}, \boldsymbol{\theta}
$$

A.4 *There exists a constant $\Delta \in [0, 1)$, such that, for each $\mathbf{d}$ and $\forall \boldsymbol{\theta}$*

$$
\mathbb{E}\left[\|\nabla_{\boldsymbol{\theta}} \log p(\mathbf{d}\,|\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})\|\right] \leq 2\Delta\left(M^2\|\boldsymbol{\theta}\|^2 + B^2\right)
$$

A.5 *Let p_0 be the distribution density of the initial $\boldsymbol{\theta}$,*

$$
\kappa_0 \triangleq \log \int e^{\|\boldsymbol{\theta}\|^2} p_0(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} < \infty
$$

In Raginsky et al. (2017), the inversed temperature parameter β is required to be larger than or equal to $\max\{1, 2/m\}$. In our setting, $\beta = 1$. Consequently, we require $\frac{2}{m} \leq 1$, which is $m \geq 2$. Thus **A.3** of the above assumption turns into

$$
\langle \mathbf{d}, \nabla_{\boldsymbol{\theta}} \log p(\mathbf{d}\,|\boldsymbol{\theta})\rangle \geq 2\,\|\mathbf{d}\|^2 - b, \quad \forall\, \mathbf{d}, \boldsymbol{\theta}
$$

Furthermore, in Proposition 7, the interval of the small constant ω is

$$
\omega \in \left(0, \min\{\frac{m}{4M^2}, e^{\Omega(\lambda_*/(r+1))}\}\right) , \tag{9}
$$

where λ_* is the *uniform spectral gap* defined as

$$
\begin{aligned}
\lambda_* \triangleq \inf_{\mathbf{d} \in \mathbf{X}} \inf &\left\{\frac{\int \|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta})\|^2 \mathrm{d}\pi}{\int \|g(\boldsymbol{\theta})\|^2 \mathrm{d}\pi} : g \in \mathcal{C}^1(\mathbb{R}^r) \cap \mathcal{L}^2(\pi),\right. \\
&\left. g \neq 0, \int g(\boldsymbol{\theta}) \mathrm{d}\pi = 0\right\} ,
\end{aligned}
$$

where π is the stationary probability measure of the diffusion defined on the training data. $\frac{1}{\lambda_*}$ might scale exponentially w.r.t. the dimension r of $\boldsymbol{\theta}$ in general, but also can be made dimension-free, for example, in the entropy-SGD objective Chaudhari et al. (2017).

Proof [Sketch Proof of Proposition 7] First, from Theorem 1 in Raginsky et al. (2017), for ω satisfying (9), taking the inversed temperature parameter β to be 1, we have the generalization error bound

$$
\begin{aligned}
&\mathbb{E}\mathcal{F}(\hat{\boldsymbol{\theta}}_T) - \mathcal{F}^* \\
\leq &O\left(\frac{(r+1)^2}{\lambda_*}\left(\Delta^{1/4} \log \frac{1}{\omega} + \omega\right) + \frac{(r+1)^2}{\lambda_* N} + r \log 2\right) ,
\end{aligned}
$$

provided $T = \Omega\left(\frac{(r+1)}{\lambda_* \omega^4} \log^5 \frac{1}{\omega}\right)$ and $\eta \leq \left(\frac{\omega}{\log(1/\omega)}\right)^4$. Here $O(\cdot)$ and $\Omega(\cdot)$ hide dependence on the parameters A, L, m, b, M, κ_0. Together with the stepsize condition to preserve DP in Theorem 4, we get that the stepsize should satisfies $\eta \leq \min\left\{\left(\frac{\omega}{\log(1/\omega)}\right)^4, \frac{\epsilon^2 N}{c^2 L^2 T \log(1/\delta)}\right\}$.

Further hiding dependency on r, Δ and λ_* (as we only wants to investigate the bound w.r,t. T, η and N), we have

$$
\mathbb{E}\mathcal{F}(\hat{\boldsymbol{\theta}}_T) - \mathcal{F}^* \leq O\left(\log \frac{1}{\omega} + \omega + \frac{1}{N}\right) . \tag{10}
$$

Since $T \propto \frac{1}{\omega^4} \log^5 \frac{1}{\omega}$, we can simplify the above equation as

$$
\mathbb{E}\mathcal{F}(\hat{\boldsymbol{\theta}}_T) - \mathcal{F}^* \leq O\left(T^{1/5} \omega^{4/5} + \omega + \frac{1}{N}\right) .
$$

To represent the bound without ω, let $x = \frac{1}{\omega}$, $m = x^4$.

From $T = A\frac{1}{\omega^4}\log^5\frac{1}{\omega}$ we have

$$\begin{aligned}
&T = Ame^{\frac{4}{5}m}, \ \ \Rightarrow \frac{4}{5A}T = \frac{4}{5}me^{\frac{4}{5}m}\\
\Rightarrow &\frac{4}{5}m = W(\frac{4}{5A}T)\\
\Rightarrow &\omega = \exp\left\{-\left(\frac{5}{4}W(\frac{4}{5A}T)\right)^{1/5}\right\}\\
\Rightarrow &\log\frac{1}{\omega} = \left(\frac{5}{4}W(\frac{4}{5A}T)\right)^{1/5}
\end{aligned}$$

Substituting the formulas for ω and $\log\frac{1}{\omega}$ into (10) and omitting constants independent of T results in the corresponding bound specified in Proposition 7. ■