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 2.4) 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-
plling 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 $X = (d_1, \ldots, d_N) \in X^N$, where $d_i \in X$. The parameters of a model are denoted as $\theta \in \mathbb{R}^r$, 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., $d_i \neq d'_i$ for some $i$, where $d_i \in X$ and $d'_i \in X'$.

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

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

where $Pr(e)$ denotes the probability of event $e$, and $\epsilon$ and $\delta$ are two positive real numbers that indicate the loss of privacy. When $\delta = 0$, we say $M$ has $\epsilon$-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 $d_i$, by observing the output, any outside attackers are not able to recognize whether the output has resulted from $X$ and $X'$, as long as $\epsilon$ and $\delta$ are small enough (making these two probabilities close to each other). Thus, the existence of the record $d_i$ 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 $M$’s output with inputs $X$ and $X'$, we define the privacy loss below.

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

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

It can be shown that the $(\epsilon, \delta)$-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 $M : X^N \to Y$ be a randomized mechanism, and let $X$ and $X'$ be a pair of adjacent datasets. Let $\text{aux}$ denote any auxiliary input that is independent of both $X$ and $X'$. The moments accountant parameterized by $\lambda > 0$ is defined as $\alpha_M(\lambda) \equiv \max_{\text{aux}, X, X'} \alpha_M(\lambda; \text{aux}, X, X')$, where $\alpha_M(\lambda; \text{aux}, X, X') \equiv \log \mathbb{E}[\exp(\lambda c(M, \text{aux}, X, X'))]$ is the log of the moment generating function at $\lambda$.

Theorem 1 (Abadi et al., 2016)
[Composability] Suppose that $M$ consists of a sequence of adaptive mechanisms $M_1, \ldots, M_k$ where $M_i : \prod_{j=1}^{i-1} Y_j \times X \to Y_i$, and $Y_i$ is the range of the $i$th mechanism, i.e., $M = M_1 \circ \cdots \circ M_k$, with $\circ$ the composition operator. Then, for any $\lambda$, we have

$$
\alpha_M(\lambda) \leq \sum_{i=1}^k \alpha_M_i(\lambda)
$$

where the input for $\alpha_M_i$ is defined as all $\alpha_M_i$’s outputs, $\{o_j\}$, for $j < i$; and $\alpha_M_i$ takes $M_i$’s output, $\{o_i\}$, for $i < k$, as the auxiliary input.

[Tail bound] For any $\epsilon > 0$, the mechanism $M$ is $(\epsilon, \delta)$-DP for $\delta = \min_{\lambda > 0} \exp(\alpha_M(\lambda) - \lambda \epsilon)$.
For the remainder of this paper, for simplicity we only consider mechanisms that output a real-valued vector. That is, \( M : \mathcal{X}^N \rightarrow \mathbb{R}^p \). 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 \( \|f(\cdot)\|_2 \leq 1 \). Let \( \sigma \geq 1 \) and \( J \) is a mini-batch sample with sampling probability \( q_i \), \( q = \frac{T}{N} \) with minibatch size of \( T \).

If \( q \leq \frac{\sigma^2}{16N} \), for any positive real number \( \lambda \leq \sigma^2 \ln \frac{1}{q\tau} \), the mechanism \( \mathcal{M}(X) = \sum_{i \in J} f(d_i) + N(0, \sigma^2 I) \) 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 \( p(\theta|X) \), with \( \theta \) 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 \( p(\theta|X) \).

Formally, an Itô diffusion is written as

\[
d\theta_t = F(\theta_t)dt + g(\theta_t)dW_t,
\]

with \( t \) the time index; \( \theta_t \in \mathbb{R}^p \) represents the full variables in a system, where typically \( \Theta_t \supseteq \theta_t \) (thus \( p \geq r \)) is an augmentation of the model parameters; and \( W_t \in \mathbb{R}^p \) 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 \( \theta = \theta_t \), and \( F(\theta_t) = -\nabla \log p(\theta) \), \( \theta_{t+1} = \sqrt{2\tau} \mathcal{N} \).\( \log p(\theta) \) denotes the unnormalized negative log-posterior, and \( p(\theta) \) is the prior distribution of \( \theta \), which defines \( \Theta(\theta, q) \), and \( F(\theta_t) = \left( -B_q - \nabla \log p(\theta) \right) \), \( g(\theta_t) = \sqrt{2B} \) for a scalar \( B > 0 \). \( \mathcal{N} \) is an auxiliary variable known as the momentum (Chen et al. 2014; Ding et al. 2014). Similar formulæ 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 \( \nabla \log p(\theta) \) is calculated with a random subset of the full data, denoted \( \nabla \log p(\theta) \) and defined as \( \nabla \log p(\theta) = \nabla \log p(\theta) + \frac{\lambda}{\sqrt{N}} \sum_{i \in J} \log p(d_i|\theta) \), where \( J \) is a random minibatch of the data with size \( \tau \) (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 \( p(\theta) \), privacy budget \( \epsilon_0, \delta_0 \), gradient norm bound \( L \). A decreasing/fixed-step-size sequence \( \{\eta_i\} \). Set \( t = 1 \):

1: \textbf{for } \( t \in [T] \) \textbf{do}
2: \hspace{1em} Take a random sample \( J_t \) with sampling probability \( q = \tau/N \).
3: \hspace{1em} For each \( i \) in \( J_t \):
4: \hspace{2em} Compute the overall privacy cost \( \tau + \log \frac{\|g(d_i)\|_2}{L} \)
5: \hspace{1em} Sample each coordinate of \( z_t \) iid from \( N(0, \frac{\eta_t}{N}) \)
6: \hspace{1em} Update \( \theta_{t+1} = \theta_t + \eta_t \left( \nabla \log p(\theta_t) + \frac{1}{\tau} \sum_{i \in J_t} \log p(d_i|\theta_t) \right) + z_t \)
7: \hspace{1em} Return \( \theta_{T+1} \) as a posterior sample (after burn in).
8: \textbf{end for}
9: Compute the overall privacy cost \( (\epsilon, \delta) \) using the moment accountant method. Ensure \( \epsilon \leq \epsilon_0 \) and \( \delta \leq \delta_0 \).
10: Output \( \theta_{T+1} \).

We typically adopt the popular Euler method to solve the continuous-time diffusion by an \( \eta \)-time discretization (stepsize being \( \eta \)). 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 guarantees that 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 (Pascual et al. 2013). Note the clipping is introduced only for practical reasons. The Lipschitz property is typically assumed in SG-MCMC for the
It is now clear that with optimal decreasing stepsize sequence (in terms of MSE defined in Section 3.2), Algorithm 1 maintains $(\epsilon, \delta)$-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 $(\epsilon, \delta)$-DP.

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

In (Wang et al., 2015), the authors proved that the SGLD method is $(\epsilon, \delta)$-DP if the stepsize $\eta_t$ is small enough to satisfy $\eta_t < \frac{T^{1/3} \log(1/\delta)}{c_2 N L^2 T \log(1/\delta)}$. 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, $\eta_t < \frac{c^2 N L^{-1/3}}{c_2 T L^{2/3} T \log(1/\delta)}$, 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 $c_2$ 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 (L et al., 2016). More importantly, our theory indicates that using SGLD with the default stepsize $\eta_t = 0.1$ is able to achieve $(\epsilon, \delta)$-DP with a small privacy loss for the MNIST dataset. As a comparison, Wang et al. (2015) gives a much smaller upper bound of $\eta_t < 1.54 \times 10^{-11}$, which is too small too be practically used. More detailed comparisons for these two bounds is given in Section 4.4 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 3 Privacy Analysis for Stochastic Gradient Langevin Dynamics

We first develop theory to prove Algorithm 1 is $(\epsilon, \delta)$-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 $\epsilon < c_1 q 2^{T/2}$, Algorithm 1 satisfies $(\epsilon, \delta)$-DP as long as $\eta_t$ satisfies:

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

**Remark 2** In practice, the first condition is easy to satisfy, as $\frac{N}{T^2}$ 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.
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., $(\epsilon, \delta)$-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 decreasing 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 $\phi(\theta)$, defined as $\bar{\phi} = \int \phi(\theta)p(\theta|D)d\theta$, with a posterior distribution $p(\theta|D)$. The posterior average is typically infeasible to compute, thus we use the sample average, $\hat{\phi}_T = \frac{1}{T^\dagger} \sum_{t=1}^T \eta_t \phi(\theta_t)$, to approximate $\bar{\phi}$, where $\{\theta_t\}_{t=1}^T$ are the samples from an SG-MCMC algorithm. The MSE we desire is defined as $\mathbb{E}(\hat{\phi}_T - \bar{\phi})^2$. 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 [4] in the SM, the MSE of SGLD with a decreasing stepsize sequence $\{\eta_t \leq \frac{\epsilon^2 N_t^{-1/3}}{c^2 L^2 T^{2/3} \log(1/\delta)}\}$ as in Theorem [4] is bounded, for a constant $C$ independent of $\{\eta_t, T, \tau\}$ and a bounded constant $\Gamma_M$ depending on $U(\cdot)$ (see the proof for details), as $\mathbb{E}(\hat{\phi}_L - \bar{\phi})^2 \leq C \left( \frac{2}{3} \left( \frac{N}{2^2} - 1 \right) N^2 \Gamma_M T^{-1} + \frac{1}{\eta_0} + 2\eta_0^2 T^{-2/3} \right)$, where $\eta_0 \equiv \frac{\epsilon^2}{c^2 L^2 \log(1/\delta)}$.

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 $\eta_0$ when fixing other parameters can be seen as $\bar{\eta}_0 = O(T^{2/9})$.

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 $(\epsilon, \delta)$-DP setting, the MSE bound induces an asymptotic bias term of $\frac{1}{\eta_0}$ 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 step size, rephrased in Proposition 6.

Proposition 6 With the same Assumption as Proposition [4], the MSE of SGLD is bounded as
\[
\mathbb{E}(\hat{\phi}_L - \bar{\phi})^2 \leq C \left( \frac{(N - 1) N^2 \Gamma_M}{T} + 1 T \eta^2 + \eta^2 \right).
\]

Furthermore, the optimal MSE w.r.t. the stepsize $\eta$ is bounded by
\[
\mathbb{E}(\hat{\phi}_L - \bar{\phi})^2 \leq C \left( \frac{(N - 1) N^2 \Gamma_M}{T} + T^{-2/3} \right),
\]
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 ($\eta = O(T)$) 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 step sizes 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 $\eta = \frac{\epsilon^2 N_t^{-1/3}}{c^2 L^2 \log(1/\delta)}$ into the MSE formula in Lemma [6]. Consequently, the MSE is bounded by $\mathbb{E}(\hat{\phi}_L - \bar{\phi})^2 \leq C \left( \frac{(N - 1) N^2 \Gamma_M}{T} + \frac{\epsilon^2 N_t^{-1/3}}{c^2 L^2 T^{2/3} \log(1/\delta)} \right)$, 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 $U(\theta)$ in an infinite-sized dataset, i.e., minimizing $\mathcal{F}(\theta) \equiv \mathbb{E}_p[\log p(d|\theta)]$, where $P$ is the unknown probability
law of the data. Let $F^* = \inf_{\theta} F(\theta)$, and $\hat{\theta}_T$ be the final sample returned by our DP-SGLD. We investigate the generalization ability in terms of the expected excess risk: $\mathbb{E}F(\theta_T) - F^*$, 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 2 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 $\eta$, the total number of iterations $T$ and the dataset size $N$.

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

$$\mathbb{E} F(\theta_T) - F^* \leq O\left(T^{1/5} \omega^{4/5} + \omega + \frac{1}{N}\right) = O\left(W^{1/5} \left(\frac{4}{5A} T\right) + \exp\left\{-W^{1/5} \left(\frac{4}{5A} T\right)\right\} + \frac{1}{N}\right),$$

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, $\omega$ would become small as $T$ grows. Consequently, one should choose an appropriate $\omega$ so that the terms $T^{1/5} \omega^{4/5}$ and $\omega$ 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 $\mathbb{E} F(\theta_T) - F^*$, 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.

---

†The specific range is given in Section F in the SM.
SGLD automatically preserves $(\epsilon, \delta)$-DP even when $\epsilon$ 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 differentially 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.

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 $\epsilon$. 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 $\epsilon$ 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 $32 \times 32$ 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 $(\epsilon, \delta)$-DP for $\epsilon = 0.99$ and $\delta = 10^{-5}$. For the SVHN dataset, the standard SGLD with stepsize $\eta_t = 0.1$ satisfies $(\epsilon, \delta)$-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>
<thead>
<tr>
<th>Dataset</th>
<th>Methods</th>
<th>$\epsilon$</th>
<th>$\delta$</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>Non-Private</td>
<td>$0.5$</td>
<td>$10^{-3}$</td>
<td>$99.34%$</td>
</tr>
<tr>
<td></td>
<td>DP-SGD</td>
<td>$8.0$</td>
<td>$10^{-3}$</td>
<td>$97.00%$</td>
</tr>
<tr>
<td></td>
<td>CGNMax</td>
<td>$1.97$</td>
<td>$10^{-3}$</td>
<td>$98.51%$</td>
</tr>
<tr>
<td></td>
<td>DP-SGLD</td>
<td>$0.99$</td>
<td>$10^{-3}$</td>
<td>$99.21%$</td>
</tr>
<tr>
<td>SVHN</td>
<td>Non-Private</td>
<td>$4.96$</td>
<td>$10^{-6}$</td>
<td>$91.62%$</td>
</tr>
<tr>
<td></td>
<td>CGNMax</td>
<td>$2.97$</td>
<td>$10^{-6}$</td>
<td>$91.89%$</td>
</tr>
</tbody>
</table>

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 light 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 model-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 util-
ity 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


A Proof of Theorem 3

We first prove Algorithm 1 is $(\epsilon, \delta)$-DP if we change the variance of $z_t$ to be $\sigma_t^2 = \frac{c_2^2 L^2 T^{2/3} t^{1/3} \log(1/\delta) }{\epsilon^2 N^2} \eta_t^2 T$ 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 $\theta$. Each update corresponds to a random mechanism $M_i$ defined in Theorem 4, thus we first derive the moments accountant for each iteration. In each iteration, the only data access is $\sum_{t \in I_i} g_t(d_i)$ in Step 6. Therefore, in the following, we only focus on the interaction between $\sum_{t \in I_i} g_t(d_i)$ and the noise $z_t$, which is essentially $\frac{\eta_t}{\tau} \sum_{t \in I_i} g_t(d_i) + z_t$, where $\tilde{g} = \frac{\tau}{\eta_t} \star \hat{g}$.

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

$$\frac{\eta_t}{\tau} \sum_{t \in I_i} g_t(d_i) + z_t = \frac{\eta_t}{\tau} \left( \sum_{t \in I_i} \hat{g}_t(d_i) + N(0, (\sigma_t^2 T) I) \right) = \frac{\eta_t L}{\tau} \left( \sum_{t \in I_i} \hat{g}_t(d_i) + N(0, \tilde{g}^2 T I) \right)$$

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

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

as long as the conditions in Lemma 2 are satisfied, that is $\tilde{g}^2 t^{1/3} \geq 1$ and the mini-batch sampling probability $q < 1 / \log^2(1/\delta)$. Later we will derive the corresponding bounds in terms of $\eta_t$.

Using the composability property of the moments accountant in Theorem 4 over $T$ iterations, the log moment of the privacy loss random variable is bounded by

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

According to the tail bound property in Theorem 1, $\delta$ is the minimum of $\exp(\alpha_M(\lambda) - \lambda \epsilon)$ w.r.t. $\lambda$. To guarantee $(\epsilon, \delta)$-DP, it suffices that

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

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

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

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 $\tilde{g} = c_2 \sqrt{T^{2/3} \log(1/\delta)}$, the conditions (2) (3) are satisfied. If we plug in $\tilde{g}$ and $q$, we have proved that Algorithm 1 is $(\epsilon, \delta)$-DP when $z_t \sim N(0, \frac{c_2^2 L^2 T^{2/3} t^{1/3} \log(1/\delta)}{\epsilon^2 N^2} \eta_t^2 T I)$.

For the second step of the proof, we prove that Algorithm 1 is $(\epsilon, \delta)$-DP when the original variance of $z_t$ is used, i.e., $\sigma_t^2 = \frac{\eta_t}{N}$. This is straightforward because when $\eta_t < \frac{c_2^2 N \epsilon^{-1/3}}{\epsilon^2 N^2 T^{2/3} \log(1/\delta)}$ 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 $\eta_t$ is positive. Adding more noise decreases the privacy loss. To satisfy $(\epsilon, \delta)$-DP, it suffices to set the variance of $z_t$ as $\eta_t / N$, 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 $\sigma^2 = \tilde{g}^2 t^{1/3}$ by definition. This is equivalent to $\tilde{g}^2 t^{1/3} \geq 1$ and $q < \frac{1}{16 \sigma^2}$.

If we plug in the formula $\eta_t = \frac{N}{4 \tau^{1/2} \tilde{g}^2 T}$, this simplifies to $\eta_t \leq \frac{N}{17}$ and $\eta_t > \frac{9}{256 L T}$. 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 $(\epsilon, \delta)$-DP whenever $\eta < \frac{c_2^2 N T^{2/3} \log(1/\delta)}{\epsilon^2 N^2}$ for some constant $c$.

Proof The only change of the proof for fixed stepsize is that the expression for the variance of the Gaussian noise $z_t$ becomes $\sigma_t^2 = \eta_t^2 L^2 / \tau^2$ for fixed stepsize. We still apply Theorem 4 and Lemma 2 to find the required conditions for $(\epsilon, \delta)$-DP:

$$Tq^2 \lambda^2 \eta_t^2 / \eta_0^2 \leq \lambda \epsilon/2 , \exp(-\lambda \epsilon/2) \leq \delta .$$

\footnote{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 $\eta_0 = c_4 q^2 \log(1/\delta)$ satisfy the above conditions. Then if we plug in $\eta_0$ and $q$, and compare it to $\eta/N$, it is easy to see Algorithm \ref{Alg:SGD} satisfies $(\epsilon, \delta)$-DP when $\eta < \frac{c_3^2}{T L^2 T^2 \log(1/\delta)}$.

\section{Calculating Constants in Moment Accountant Methods}

For calculating the constants $c_1$ and $c_2$, which is a part of the moment accountant method, we refer to https://github.com/tensorflow/models/tree/master/research/differential_privacy/privacy_accountant as an implementation of the moment accountant method. A comprehensive description for the implementation can be found in 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.

\section{Assumptions on SG-MCMC Algorithms}

For the diffusion in \eqref{eq:diffusion}, we first define the generator $\mathcal{L}$ as:

$$
\mathcal{L}_\psi \hat{=} \frac{1}{2} \nabla \psi \cdot F + \frac{1}{2} g(\theta) g(\theta)^* : D^2 \psi ,
$$

where $\psi$ is a measurable function, $D^k \psi$ means the $k$-derivative of $\psi$, * means transpose. $A \cdot B \hat{=} a^T b$ for two vectors $a$ and $b$, $A : B \hat{=} \text{trace}(A^T B)$ for two matrices $A$ and $B$. Under certain assumptions, there exists a function, $\hat{\phi}$, such that the following Poisson equation is satisfied: \cite{mattingly2010}:\footnote{This is under the Apache License, Version 2.0}

$$
\mathcal{L}_\psi \hat{=} \hat{\phi} - \hat{\phi} ,
$$

where $\hat{\phi} \hat{=} \int \phi(\theta) \rho(d\theta)$ denotes the model average, with $\rho$ being the equilibrium distribution for the diffusion \eqref{eq:diffusion}, which is assumed to coincide with the posterior distribution $\rho(\theta|\mathcal{D})$. The following assumptions are made for the SG-MCMC algorithms \cite{vollmer2016gae, chen2015private}.

\begin{assumption}
The diffusion \eqref{eq:diffusion} is ergodic. Furthermore, the solution of \eqref{eq:poisson} exists, and the solution functional $\psi$ satisfies the following properties:

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

\section{Proof of Proposition \ref{prop:privacy}}

Claim: Under Assumption \ref{assump:diffusion} in the section \ref{sec:assumptions} the MSE of SG-LD with a decreasing stepsize sequence $\{\eta_\tau = \frac{c_3^2}{T L^2 T^2 \log(1/\delta)}\}$ as in Theorem \ref{thm:privacy} is bounded, for a constant $C$ independent of $\{\eta, T, \tau\}$ and a constant $\Gamma_M$ depending on $T$ and $U(\cdot)$, as $E \left( \frac{1}{\eta} \sum_{i=1}^{T} \mathcal{V}(S_i) \right)^2 \leq C \left( \frac{(N - 1)}{\eta T^2} \left( \frac{N^2 \Gamma_M T^{-1} + 1}{3} + 2 \eta^2 T^{-2/3} \right) \right)$.

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

$$
E \left( \frac{1}{\eta} \sum_{i=1}^{T} \mathcal{V}(S_i) \right)^2 \leq C \left( \frac{(N - 1)}{\eta T^2} \left( \frac{N^2 \Gamma_M T^{-1} + 1}{3} + 2 \eta^2 T^{-2/3} \right) \right),
$$

where $S_T \hat{=} \sum_{i=1}^{T} \eta_i$, and $\Gamma_i$ is a term related to $\phi$, which, according to Theorem 3 of Chen et al. (2017), can be simplified as

$$
E |\Delta V_i|^2 = \frac{(N - 1) N^2}{\tau} \left( \frac{1}{N^2} \sum_{i, j} E \alpha_i^T \alpha_{ij} - \frac{2}{N(N - 1)} \sum_{i < j} E \alpha_i^T \alpha_{ij} \right) \Delta (N - 1) N^2 \Gamma_i ,
$$

where $\alpha_{ij} = \nabla \phi \theta \log p(d_i | \theta_i)$. Let $\Gamma_M \hat{=} \max_i \Gamma_i$. Substituting \eqref{eq:gamma} into \eqref{eq:mse}, we have

$$
E \left( \frac{1}{\eta} \sum_{i=1}^{T} \mathcal{V}(S_i) \right)^2 \leq C \left( \frac{(N - 1)}{\eta T^2} \left( \frac{N^2 \Gamma_M T^{-1} + 1}{3} + 2 \eta^2 T^{-2/3} \right) \right).
$$
Now, if we assume \( \tilde{\eta}_0 = \frac{c}{L^2 \log(1/\theta)} \), then we rewrite \( \eta_t = \eta_0 t^{-1/3} T^{-2/3} \).

Note \( \sum_t^{T_T} \approx \frac{1}{p+1} T_T^{p+1} \). Plug this into the bound in (\ref{inequality}), we have:

\[
\mathbb{E} \left( \hat{\phi}_L - \hat{\phi} \right)^2 \leq C \left( \frac{\sum_t^T \eta_t^2}{\sum_t^T \eta_t^2} \left( \frac{N}{\tau} - 1 \right) N^2 \Gamma M + \frac{1}{\sum_t^T \eta_t} + \left( \frac{\sum_t^T \eta_t^2}{\sum_t^T \eta_t} \right)^2 \right) \leq C \left( \frac{2}{3} \left( \frac{N}{\tau} - 1 \right) N^2 \Gamma M T^{-1} + \frac{1}{3\eta_0} + 2\eta_0^2 T^{-2/3} \right)
\]

\[\text{ }\]

\[\text{ }\]

\[
\text{F Generalization Bound}
\]

Following \cite{Raginsky2017}, we need to make the following assumptions to derive our generalization bound. Actually, some of these assumptions are related to Assumption \ref{assumption}. Interested readers are encouraged to refer to Section 9 of \cite{Vollmer2016} for details.

**Assumption 2** Assume the likelihood function satisfies:

\(\text{A.1 Let } \theta_0 \text{ be the initial value. There exists } A, L \geq 0 \text{ such that}
\)

\[
\left\| \log p(d|\theta_0) \right\| \leq A, \quad \left\| \nabla_{\theta} \log p(d|\theta_0) \right\| \leq L, \quad \forall d
\]

\(\text{A.2 For some } M > 0, \forall d_1, d_2
\)

\[
\left\| \nabla_{\theta} \log p(d_1|\theta) - \nabla_{\theta} \log p(d_2|\theta) \right\| \leq M \left\| d_1 - d_2 \right\|
\]

\(\text{A.3 For some } m > 0 \text{ and } b > 0,
\)

\[
\langle d, \nabla_{\theta} \log p(d|\theta) \rangle \geq m \left\| d \right\|^2 - b, \quad \forall d, \theta
\]

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

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

\(\text{A.5 Let } p_0 \text{ be the distribution density of the initial } \theta,
\)

\[
\kappa_0 \triangleq \log \int e^{\|\theta\|^2} p_0(\theta) d\theta < \infty
\]

In \cite{Raginsky2017}, the inversed temperature parameter \( \beta \) 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

\[
\langle d, \nabla_{\theta} \log p(d|\theta) \rangle \geq 2 \left\| d \right\|^2 - b, \quad \forall d, \theta
\]

Furthermore, in Proposition 7 \( \text{the interval of the small constant } \omega \) is

\[
\omega \in \left( 0, \min\{ \frac{m}{4M^2 \log^2(1/\theta)} - 1 \} \right)
\]

\[\text{where } \lambda_* \text{ is the uniform spectral gap defined as}
\]

\[
\lambda_* \triangleq \inf_{d \in \mathbb{X}} \inf_{\theta \neq 0} \left\{ \int \frac{\| \nabla_{\theta} g(\theta) \|^2 d\pi}{\int \| g(\theta) \|^2 d\pi} : g \in C^1(\mathbb{R}^r) \cap L^2(\pi), \right. \]

\[
\left. \exists \varphi \in \mathbb{F}, \varphi \neq 0, \int g(\theta) d\pi = 0 \right\},
\]

where \( \pi \) is the stationary probability measure of the diffusion defined on the training data. \( \frac{2}{m} \) might scale exponentially w.r.t. the dimension \( r \) of \( \theta \) in general, but also can be made dimension-free, for example, in the entropy-SGD objective \cite{Chaudhari2017}.

**Proof** [Sketch Proof of Proposition 7] First, from Theorem 1 in \cite{Raginsky2017}, for \( \omega \) satisfying (9), taking the inversed temperature parameter \( \beta \) to be 1, we have the generalization error bound

\[
\mathbb{E} F(\hat{\theta}_T) - F^* \leq O \left( \frac{\left( r + 1 \right)^2}{\lambda_*} \left( \Delta^{1/4} \log \frac{1}{\omega} + \omega \right) + \left( r + 1 \right)^2 \lambda_\infty N - r \log 2 \right)
\]

provided \( T = \Omega \left( \frac{(r+1)^5}{\lambda_*} \right) \) and \( \eta \leq \left( \frac{\omega \log(1/\omega)}{\log(1/\theta)} \right)^4 \).

Here \( O(\cdot) \) and \( \Omega(\cdot) \) hide dependence on the parameters \( A, L, m, b, M, \kappa_0 \). Together with the stepsizes condition to preserve DP in Theorem 4, we get that the stepsize should satisfies \( \eta \leq \min \left( \left( \frac{\omega \log(1/\omega)}{\log(1/\theta)} \right)^4, \frac{e^2 L^2 T \log(1/\theta)}{\kappa_0} \right) \).

Further hiding dependency on \( r, \Delta \) and \( \lambda_* \) (as we only want to investigate the bound w.r.t. \( T, \eta \) and \( N \)), we have

\[
\mathbb{E} F(\hat{\theta}_T) - F^* \leq O \left( \log \frac{1}{\omega} + \omega + \frac{1}{N} \right).
\]

(10)

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

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

To represent the bound without \( \omega \), let \( x = \frac{1}{\omega}, m = x^4 \).
From $T = A \frac{1}{\omega} \log^5 \frac{1}{\omega}$ we have
\[
T = A me^{\frac{4}{5} m}, \quad \Rightarrow \quad \frac{4}{5} T = \frac{4}{5} me^{\frac{4}{5} m}
\]
\[
\Rightarrow \frac{4}{5} m = W\left(\frac{4}{5} T\right)
\]
\[
\Rightarrow \omega = \exp\left\{-\left(\frac{5}{4} W\left(\frac{4}{5} T\right)\right)^{\frac{1}{5}}\right\}
\]
\[
\Rightarrow \log \frac{1}{\omega} = \left(\frac{5}{4} W\left(\frac{4}{5} T\right)\right)^{\frac{1}{5}}
\]

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