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 $\sigma_t^2 = (\tilde{\eta}^2 L^2 \eta_t^2 t^{1/3} / \tau^2) I$[§]. 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 $\mathbf{b}$, $\mathbf{A} : \mathbf{B} \triangleq \text{trace}(\mathbf{A}^T\mathbf{B})$ 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.,* $\|\mathcal{D}^k\psi\| \le C_k\mathcal{V}^{p_k}$ *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* $\sup_{s\in(0,1)} \mathcal{V}^p\left(s\,\mathbf{x} + (1-s)\,\mathbf{y}\right) \le C\left(\mathcal{V}^p(\mathbf{x}) + \mathcal{V}^p(\mathbf{y})\right)$*,* $\forall \mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^m, p \le \max\{2p_k\}$ *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 $\mathbb{E}\left(\hat{\phi}_L - \bar{\phi}\right)^2$

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

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

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

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

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

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

$$\begin{aligned}
&\mathbb{E}\left(\hat{\phi}_L - \bar{\phi}\right)^2 \le \\
&C\left(\frac{\sum_t^T \eta_t^2}{\left(\sum_t^T \eta_t\right)^2}\left(\frac{N}{\tau} - 1\right) N^2\Gamma_M + \frac{1}{\sum_t^T \eta_t} + \frac{\left(\sum_t^T \eta_t^2\right)^2}{\left(\sum_t^T \eta_t\right)}\right)
\end{aligned} \tag{8}$$

[‖] 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

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

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. ■