Stochastic Gradient Langevin Dynamics with Variance Reduction

Zhishen Huang
Department of Computational Mathematics
Michigan State University
East Lansing, MI, USA
zhishen.huang@colorado.edu

Stephen Becker
Department of Applied Mathematics
University of Colorado Boulder
Boulder, CO, USA
stephen.becker@colorado.edu

***Abstract*—Stochastic gradient Langevin dynamics (SGLD) has gained the attention of optimization researchers due to its global optimization properties. This paper proves an improved convergence property to local minimizers of nonconvex objective functions using SGLD accelerated by variance reductions. Moreover, we prove an ergodicity property of the SGLD scheme, which gives insights on its potential to find global minimizers of nonconvex objectives.**

I. Introduction

In this paper we consider the optimization algorithm *stochastic gradient descent* (SGD) with variance reduction (VR) and Gaussian noise injected at every iteration step. For historical reasons, the particular randomization format of injecting Gaussian noises bears the name *Langevin dynamics* (LD). Thus, the scheme we consider is referred as *stochastic gradient Langevin dynamics with variance reduction* (SGLD-VR). We prove the ergodicity property of SGLD-VR schemes when used as an optimization algorithm, which the normal SGD method without the additional noise does not have. As the ergodicity property implies the non-trivial probability for the LD process to visit the whole space, the set of global minima will also be traversed during the iteration. We also provide convergence results of SGLD-VR to local minima in a similar style to [Xu et al., 2018]. Taken together, the results show that SGLD-VR concentrates around local minima, but is never stuck at a particular point, and thus is useful for global optimization.

We apply the SGLD-VR scheme on the empirical risk minimization (ERM) problem:

$$\text{minimize } f(\boldsymbol{\omega}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{\omega}) \tag{1}$$

which often arises as a sampled version of the stochastic optimization problem: $\min_{\boldsymbol{\omega}} F(\boldsymbol{\omega}) = \int f(\boldsymbol{\omega};\xi)\,\mathrm{d}\xi$, where ξ is the collection of training data and $\boldsymbol{\omega}$ is the parameter for the model, i.e., $F(\boldsymbol{\omega})$ may be the expectation of a loss function with respect to stochastic data ξ, $F(\boldsymbol{\omega}) = \mathbb{E}\,\ell(\boldsymbol{\omega},\xi)$, so then $f_i(\boldsymbol{\omega}) = \ell(\boldsymbol{\omega},\xi_i)$ for i.i.d. realizations $(\xi_i)_{i=1}^n$. In the following text we use $\mathbf{x}$ instead of $\boldsymbol{\omega}$ as the input for the objective f in order to conform to optimization literature conventions. The usual SGD framework for ERM problems is at every gradient step to form a minibatch subsampled from $\{1,\dots,n\}$ and include only the sampled terms (with a reweighting) in the sum, in order to reduce computational complexity. Variance reduction which exploits the finite-sum structure of Eq. (1) accelerates convergence, and LD is is used in the SGD scheme to enable the ergodicity property.

A. Prior art

The Langevin dynamical equation describes the trajectory $X(t)$ of the following stochastic differential equation

$$\mathrm{d}X_t = -\nabla U(X_t)\,\mathrm{d}t + \sigma\,\mathrm{d}B_t, \tag{2}$$

where B_t is a Wiener process. This equation characterizes the continuous motion of a particle subject to fluctuations (due to nonzero temperature) in a potential field U, and in the limit $\sigma \to \infty$ approaches Brownian motion. Using this dynamic as a master equation, through Kramers–Moyal expansion one can derive the Fokker-Planck equation, which gives the spatial distribution of particles at a given time, thus a full characterization of the statistical properties of a particle ensemble [Kramers, 1940].

a) LD and sampling: The connection between Langevin dynamics (LD) and the distribution of particle ensemble reveals the potential of applying LD on sampling. Suppose that the distribution of interest is $\pi(\mathbf{x})$ and that there exists a function U such that $\pi(\mathbf{x}) = \frac{\exp(-U(\mathbf{x}))}{\int \exp(-U(\mathbf{x}))\,\mathrm{d}\mathbf{x}}$, then the LD equation using this U defines a stochastic process with stationary distribution $\pi(\mathbf{x})$. To numerically implement the LD equation for sampling purposes, one needs to discretize the continuous LD equation. A simple version of the discretization is the *unadjusted Langevin algorithm* (ULA),

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta\mathbf{x}_k,\ \Delta\mathbf{x}_k = -\eta_k\nabla U(\mathbf{x}_k) + \rho_0\sqrt{\eta_k}\boldsymbol{\epsilon}_k \tag{3}$$

where $\boldsymbol{\epsilon}_k \sim \mathcal{N}(0,\mathbf{I}_d)$, $\mathbf{x}\in\mathbb{R}^d$ and $\rho_0 = \sigma$. The Gaussian noise term enables the scheme to explore the sample space and the drift term guides the direction of exploration. One common modified scheme is the *Metropolis adjusted Langevin algorithm* (MALA), where upon the suggested update by ULA, there is an additional accept/reject step, with the probability of accepting the the update as $1 \wedge \frac{\pi(\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k+1})}{\pi(\mathbf{x}_{k+1})p(\mathbf{x}_{k+1}|\mathbf{x}_k)}$.

Naturally two central questions related to this sample scheme arise: whether or not the distribution of samples

generated by LD converges, and if so, to π; and what is the mixing time of LD (i.e., how long does it takes for the LD to approximately reach equilibrium hence generating valid samples from the distribution π). The first question motivates the importance of MALA: in terms of the *total variation* (TV) distance, while ULA can fail to converge for either light-tailed or heavy-tailed target distribution, MALA is guaranteed to converge to any continuous target distribution [Meyn and Tweedie, 2009]. Regarding the second question about convergence speed, researchers have investigated the sufficient conditions for ULA and MALA respectively to guarantee exponential (geometric) convergence to target distribution. [Mengersen and Tweedie, 1996] show for distributions over $\mathbb{R}$, the necessary and sufficient condition for MALA to converge to target distribution $\pi(\mathbf{x})$ at geometric speed is that $\pi(\mathbf{x})$ has exponential tails. The sufficiency of this condition is generalized to higher dimension in [Roberts and Tweedie, 1996b]. The seminal work by [Roberts and Tweedie, 1996a] shows that MALA cannot converge at geometric speed to target distributions that are in essence non-localized, or heavy-tailed.

In parallel there have been works to show the convergence of LD for distribution approximation in terms of Wasserstein-2 distance [Dalalyan and Karagulyan, 2017] and KL-divergence respectively [Cheng and Bartlett, 2018].

A particular case of interest for the application of LD on sampling is to find the posterior distribution of parameters $\mathbf{x}$ in the Bayesian setting, where the updates are set as

$$\Delta \mathbf{x}_k = \eta_k \left(\nabla p(\mathbf{x}_k) + \sum_{i=1}^{N} \nabla \log p(\xi_i \mid \mathbf{x}_k) \right) + \sqrt{\eta_k} \boldsymbol{\epsilon}_k \quad (4)$$

where $\boldsymbol{\epsilon}_k \sim \mathcal{N}(0, \mathbf{I})$ and $p(\mathbf{x}, \xi)$ is the joint probability of parameters $\mathbf{x}$ and data ξ. To maximize the likelihood, [Welling and Teh, 2011] suggest to use the format of stochastic gradient descent in the derivative term of (4). [Borkar and Mitter, 1999] show that this minibatch-styled LD will converge to the correct distribution in terms of KL divergence.

b) LD and optimization: The main focus of this paper is on optimization. LD offers an exciting opportunity for global optimization due to the exploring nature of the Brownian motion term. Simulating multiple particles to obtain information about the geometric landscape of the objective function—thus locating a global minimum—is often too computationally expensive to be practical. Notice that when one considers the convergence to a distribution, the exploring nature of LD due to continually injected Gaussian noise of constant variance is the key factor, while for the purpose of optimization, one usually exploits noises with diminishing variance since the goal is to converge to a point.

The technique to achieve this point convergence is *annealing*, which means decreasing the variance of the noise as t grows. Formally, let the objective function be U and we construct the probability distribution $p_T(\mathbf{x}) = \frac{1}{Z} \exp\left(-\frac{U(\mathbf{x})}{T}\right)$, where Z is the normalization factor. The key observation is that as the parameter $T \to 0$, the distribution $p_T(\mathbf{x})$ will concentrate on the global minima. This parameter T is usually referred to as *temperature*, alluding to the alloy annealing process where as temperature decreases, the structure of the metal evolves into the most stable one, hence reaching the state with minimum potential energy. To formulate LD for optimization, one essentially takes the usual LD equation but with the variance term σ now a function of time, $\sigma = \sqrt{T}(t)$:

$$\mathrm{d}\mathbf{x}_t = \nabla U(\mathbf{x}_t)\,\mathrm{d}t + \sqrt{T(t)}\,\mathrm{d}B_t$$

The pioneering work by [Chiang et al., 1987] shows that with the annealing schedule $T(t) \propto (\log t)^{-1}$, then LD will find the global minimum. The work by Chiang et al. does not specify how to simulate the continuous version of Langevin dynamics, thus not providing information on the convergence of discrete approximations (such as the Euler-Maruyama method) for LD. [Gelfand and Mitter, 1991] fill this gap by proving that with an annealing schedule $\eta_k \propto k^{-1}$ and $T_k \propto (k \log\log k)^{-1}$, the discretized LD will converge to the global minima in probability, though the convergence may be slow (and improving slow convergence is the motivation for the variance reduction scheme we analyze). More recently, [Raginsky et al., 2017] use optimal transport formalism to study the empirical risk minimization problem. Their proof uses the Wasserstein-2 distance to evaluate distribution discrepancy and consists of two parts: first they show that the discretization error of LD from continuous LD accumulates linearly with respect to the error tolerance level, and then they show that the continuous LD will converge to the true target distribution exponentially fast.

c) Variance reduction (VR) and LD: In this paper we aim to apply variance reduction techniques in the setting of LD to accelerate the optimization process and to derive an improved time complexity dependence on error tolerance level. We use the term "time complexity" to be proportional to the iteration count.

The main algorithm we consider in this paper is stochastic gradient Langevin dynamics (SGLD) with variance reduction, which consists of two sources of randomness: one from stochastic gradients, the other from Gaussian noise injected at each step. Previous work have investigated the SGLD for optimization to find local minimizers [Chen et al., 2019, Zhang et al., 2017], and reported results for convergence to approximate second order stationary points.

In particular, [Xu et al., 2018] show that with constant-variance Gaussian noise injected at each step, SGLD-VR finds an approximate minimizer with time complexity $\mathcal{O}(\frac{\sqrt{n}}{\varepsilon^2}) \cdot \mathrm{e}^{\mathcal{O}(d)}$, in contrast to the time complexity $\mathcal{O}(\frac{1}{\varepsilon^5}) \cdot \mathrm{e}^{\mathcal{O}(d)}$ for SGLD without variance reduction acceleration, where n is the number of component functions in the ERM; see Figure 1 which lends some experimental evidence that SGLD-VR can outperform regular SGLD. We aim to improve the dependency on ε in the analysis. We also point out that when the variance of the Gaussian noise is set as constant in SGLD, the function value or the point distance between an optimal point and the iterate

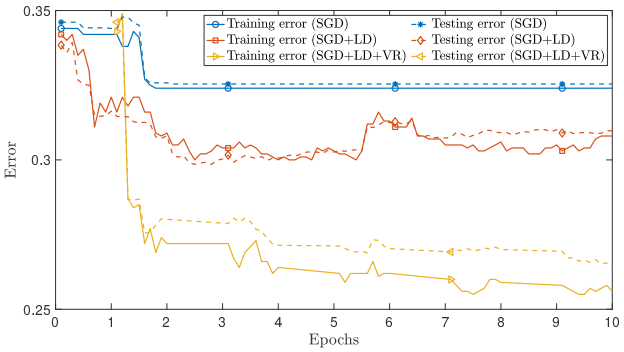

Fig. 1: Example of training a neural net for binary classification with 2 hidden layers, $n = 1000$ training points, sigmoid activation function, $\eta_0 = 10^3$, $\nu = 1$, $\rho_0 = 10^{-2}$, batch size $B_b = 100$, and $B_e = 10$. SGLD-VR converges to a good solution, in terms of both training and testing error, more quickly than either SGLD or regular SGD. The ℓ_2^2 loss was used for training, but the error reported in the figure is the misclassification rate.

can never go to zero, but can at best be bounded by a constant depending on the size of variance.

The specific VR technique we use was originally proposed to reduce the variance of the minibatch gradient estimator in stochastic gradient descent for convex objectives by [Johnson and Zhang, 2013] and separately by [Defazio et al., 2014]. [Reddi et al., 2016, Allen-Zhu and Hazan, 2016] have respectively generalized the application of VR techniques to nonconvex objectives and provided convergence guarantee to first-order stationary points.

In essence, we use the *control variate* technique to construct a new gradient estimator by adding an additional term to the minibatch gradient estimator in SGD (∇_{SGD}) and this term is correlated with ∇_{SGD}, thus reducing the variance of the gradient estimator as a whole. More specifically, consider the classic gradient estimator of function f in (1) at point $\mathbf{x}$: $\nabla_{\mathrm{SGD}} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \nabla f_i(\mathbf{x})$, where $\mathcal{I} \subseteq [n]$. If there is another random variable (r.v.) Y whose expectation is known, then we can construct a new unbiased gradient estimator $\widetilde{\nabla}$ of the function f at point $\mathbf{x}$ as $\widetilde{\nabla} = \nabla_{\mathrm{SGD}} + \alpha(Y - \mathbb{E}\,Y)$, where α is a constant. If $\alpha = -\frac{\mathrm{Cov}(\nabla_{\mathrm{SGD}}, Y)}{\mathrm{Var}[Y]}$, then the variance of the new gradient estimator $\widetilde{\nabla}$ is $\mathrm{Var}[\widetilde{\nabla}] = (1-\kappa^2)\mathrm{Var}[\nabla_{\mathrm{SGD}}] \leq \mathrm{Var}[\nabla_{\mathrm{SGD}}]$, where κ is the Pearson correlation coefficient between ∇_{SGD} and Y. Usually one may not be lucky enough to have access to such a r.v. Y whose covariance with ∇_{SGD} is known, therefore the constant α for the control variate needs to be chosen in a sub-optimal or empirical manner.

Finally we want to mention that [Dubey et al., 2016] introduce the VR technique to Bayesian inference setting where the objective is the posterior and the gradient estimator is constructed with a minibatch of training data. However, the given convergence guarantee is in terms of mean squared error of statistics evaluated based on posterior, instead of posterior distribution of the parameter.

d) Contributions: This paper discusses the convergence properties of SGLD with variance reduction and shows the ergodicity property of the scheme. Our main contributions upon the prior art is the following:

- We provide a better time complexity for the SGLD-VR scheme to converge to a local minimizer than the corresponding result in [Xu et al., 2018].
- We show the ergodicity property of SGLD-VR scheme based on the framework set for SGLD scheme in [Chen et al., 2019].

e) Notation: Bold symbols indicate vectors, for example a vector $\mathbf{x} \in \mathbb{R}^d$, where d stands for the dimension of Euclidean space. We use o to indicate the starting index of a minibatch, $\eta_{a:b}$ as the shorthand for $\sum_{i=a}^{b}\eta_i$, and $[n] = \{1, 2, \ldots, n\}$.

II. Algorithm and Main Results

The main algorithm is given as Algorithm 1.

Algorithm 1 Variance reduced stochastic gradient Langevin dynamics (SGLD-VR)

Require: initial stepsize $\eta_0 > 0$ and variance $\rho_0 > 0$, stepsize decay exponent $\nu \geq 1$, batch size B_b, epoch length B_{e}

Initialize $\mathbf{x}_0 = 0$, $\widetilde{\mathbf{x}}^{(0)} = \mathbf{x}_0$

Define the stepsize and variance sequences:

$$\eta_t = \frac{\eta_0}{t^\nu} \text{ and } \rho_t = \frac{\rho_0}{t^{\nu/2}}. \tag{5}$$

for $s = 0, 1, 2, \cdots, \frac{T}{B_{\mathrm{e}}} - 1$ **do**
  $\widetilde{\mathbf{w}} = \nabla f(\widetilde{\mathbf{x}}^{(s)})$
  for $l = 0, 1, \cdots, B_{\mathrm{e}} - 1$ **do**
    Set index $t = sB_{\mathrm{e}} + l$
    Draw $I_t \subset [n]$ of size $|I_t| = B_b$ ▷ uniformly, with replacement
    Draw $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$
    $\widetilde{\nabla}_t = \frac{1}{B_b}\sum_{i_t\in I_t}\left(\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\widetilde{\mathbf{x}}^{(s)}) + \widetilde{\mathbf{w}}\right)$ ▷ gradient estimator
    update $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t\widetilde{\nabla}_t + \rho_t\boldsymbol{\epsilon}_t$
  end for
  $\widetilde{\mathbf{x}}^{(s)} = \mathbf{x}_{(s+1)B_{\mathrm{e}}}$
end for

A. Convergence to a first-order stationary point

We start stating the main results with computing the time complexity for the SGLD-VR to converge to an $\varepsilon-$first order stationary point. We define $\mathbf{x}^\star$ to be an ε-first order stationary point (FSP) if $\|\nabla f(\mathbf{x}^\star)\| \leq \varepsilon$.

Our first assumption is very standard [Bauschke and Combettes, 2017]:

Assumption 1 (Lipschitz Gradient)**.** *f is continuously differentiable, and there exists a positive constant L such that for all $\mathbf{x}$ and $\mathbf{y}$, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$.*

Theorem 1. *Under Assumption 1, for any $p \in (0, 1)$, then with probability at least $1 - p$, the time complexity for the LD described in Algorithm 1 to converge to an ε-first order stationary point $\mathbf{x}^\star$ is $\mathcal{O}\left(\frac{\Delta_f d}{\varepsilon^2 p}\right)$, where $\Delta_f = f(\mathbf{x}_0) - f(\mathbf{x}^\star)$.*

Method	w. VR	noise magnitude setting	convergence target	Time complexity m
[Raginsky et al., 2017]	no	constant	global min.	$\widetilde{\mathcal{O}}\left(\frac{d+1/\delta_0}{\delta_0\varepsilon^4}\right)$
[Xu et al., 2018]	yes	constant	global min.†	$\widetilde{\mathcal{O}}\left(\frac{\sqrt{n}}{\varepsilon^{5/2}}\right)\exp(\widetilde{\mathcal{O}}(d))$
[Zhang et al., 2017]	no	constant	local min.	$\mathcal{O}\left(\frac{\Delta_f d^4 L^2}{\varepsilon^4}\right)$
This work	yes	diminishing w. poly. speed	local min.*	$\mathcal{O}\left(\frac{\Delta_f}{\varepsilon^2}\right)+\exp\left(\mathcal{O}(\varepsilon d)\right)$

TABLE I: Comparison between convergence results for variants of LD optimization schemes. * indicates convergence target is actually a ε-second-order stationary point, which coincides with a local minimizer when $\varepsilon < \sqrt{q}$ under Assumption 4. † means that with the noise magnitude fixed, the optimized empirical error cannot be arbitrarily close to true minimal empirical error, thus returning an *approximate* global minimizer.

B. Ergodicity

In this work we show that the discretized variance reduced LD (SGLD-VR) has an ergodic property which gives the iteration process the potential of exploring wider space, thus with positive possibility of traversing through the global optimal point. We make the following regularization assumptions.

The following assumption says essentially that f is bounded below by a known value (and without loss of generality, we can assume f is non-negative). This automatically holds for ERM problems when the loss function is non-negative, as is typical.

Assumption 2 (Nonnegative objective)**.** *The objective function is nonnegative.*

The next assumption is more complex and we discuss it in Remark 1.

Assumption 3 (Regularization conditions)**.** *There exist nonnegative constants μ_1, μ_2 and ψ_1, ψ_2 such that for all $\mathbf{x}\in\mathbb{R}^d$,*

$$\|\nabla f(\mathbf{x})\|^2 \geq \mu_1 f(\mathbf{x}) - \psi_1 \tag{6}$$

$$\|\mathbf{x}\|^2 \leq \mu_2 f(\mathbf{x}) + \psi_2 \tag{7}$$

Remark 1. We make the same regularization assumptions as in [Chen et al., 2019]. A similar regularization condition to (6) commonly used in previous literature is the (m,b)-*dissipative condition* [Mattingly et al., 2002, Raginsky et al., 2017, Xu et al., 2018, Zhang et al., 2017], which reads that there exist positive constants m and b such that for all $\mathbf{x}\in\mathbb{R}^d$, $\langle\nabla f(\mathbf{x}),\mathbf{x}\rangle \geq m\|\mathbf{x}\|^2 - b$. [Dong and Tong, 2020] show that the dissipative condition implies (6), which renders the assumption (6) weaker. Another interpretation of (6), in conjunction with Assumption 2, is that this is a slightly weaker version of the Polyak-Łojasiewicz (PL) inequality [Karimi et al., 2016]; choosing ψ_2 to be the minimal value of f gives the PL inequality, but the PL inequality itself is stronger since it implies that any stationary point (i.e., where $\nabla f(x)=0$) is globally optimal. Equation (7) implies that f is supercoercive [Bauschke and Combettes, 2017], and in particular coercive, and thus has bounded level sets.

Consider the function $f(\mathbf{x}) = \sigma(\mathbf{A}\mathbf{x}) + \gamma\|\mathbf{x}\|_2^2 + C$, which describes the connection between one neuron and the layer below it in a feedforward neural network with coefficient matrix $\mathbf{A}$, the activation function σ as $\tanh$ or sigmoid, and a Tikhonov regularization term with magnitude γ and a constant C. This is an example which satisfies Assumptions 1, 2, and 3. Examples with more types of activation functions such as ReLu and more types of regularization terms such as ℓ_1, or extensions to multilayer feedforward networks or convolutional neural networks can also be constructed if they are defined region-wise to cater for near-origin behavior and far-field behavior in the regularization assumption 3 respectively.

In Theorem 2 we show that there is a nonzero probability that the LD iteration will eventually visit any fixed point within a level set of interest.

Theorem 2 (Ergodicity)**.** *Under assumptions 1, 2, and 3, with the same parameter setting as in Lemma 7, for any accuracy $\widetilde{\varepsilon}>0$, failure probability $p>0$, and any point $\mathbf{s}\in\mathbb{R}^d$ which locates in the level set $\{\mathbf{x}: f(\mathbf{x})\leq\mathcal{O}(\widetilde{\varepsilon})\}$, there is a finite time horizon*

$$T = \widetilde{\mathcal{O}}\left(\frac{1+\ln f(\mathbf{x}_0) + \frac{(d\|\mathbf{s}\|+\widetilde{\varepsilon})^d}{\widetilde{\varepsilon}\left((\frac{4}{\sqrt{2\pi}}-1)\mathrm{e}^{-1/2}\widetilde{\varepsilon}\right)^d}}{\widetilde{\varepsilon}p\mu_1\left(\psi_1 + 2\eta_0 L^3\frac{\mu_2 f(\mathbf{x}_0)+2\psi_2}{B_\mathrm{e}} + \frac{\rho_0^2}{\eta_0}Ld\right)}\right) \tag{8}$$

such that

$$\mathbf{Pr}(\|\mathbf{x}_t - \mathbf{s}\| \leq \widetilde{\varepsilon} \text{ for some } t<T) \geq 1-p \tag{9}$$

C. Convergence to an ε-second-order stationary point

An ε-second-order stationary point is a more restrictive type of ε-first order stationary point, and is more likely to be an actual local minimizer.

Definition 1. *Consider a smooth function $f(\mathbf{x})$ with continuous second order derivative. A point $\mathbf{x}$ is an ε-second-order stationary point if*

$$\|\nabla f(\mathbf{x})\| \leq \varepsilon \quad \text{and} \quad \lambda\big(\nabla^2 f(\mathbf{x})\big)_{\min} \geq -\varepsilon^2 \tag{10}$$

where $\lambda(\cdot)_{\min}$ is the smallest eigenvalue.

We make the strict saddle assumption which is common in nonconvex optimization literature [Ge et al., 2015, Lee et al., 2016, Jin et al., 2017, Mokhtari et al., 2018, Lee et al., 2019, Vlatakis-Gkaragkounis et al., 2019, Sun et al., 2019, Liu and Yin, 2019, Li, 2019, Huang and Becker, 2020]; i.e.,

Assumption 4 (Strict saddle)**.** *There exists a constant $q > 0$ such that for all first-order stationary points $\mathbf{x}_{\rm fsp}$, we have*

$$|\lambda(\nabla^2 f(\mathbf{x}_{\rm fsp}))| \geq q > 0.$$

Assumption 5 (Lipschitz Hessian)**.** *f is twice continuously differentiable, and there exists a positive constant L_2 such that for all $\mathbf{x}$ and $\mathbf{y}$, $\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \leq L_2\|\mathbf{x}-\mathbf{y}\|$.*

Theorem 3. *Under Assumptions 1, 4 and 5, setting the stepsize decay parameter $\nu \in [1,2]$ and $\rho_0 = \mathcal{O}(\varepsilon)$, with probability $\mathcal{O}\left(\frac{\varepsilon^{d-1}}{\Gamma(\frac{d-2}{2})L^{d-1}q^{d-1}}\right)$, the time complexity for the LD described in Algorithm 1 to converge to an ε-second order stationary point $\mathbf{x}^\star$ is $\mathcal{O}\left(\frac{\Delta_f}{\varepsilon^2}\right) + \exp\left(\mathcal{O}(\varepsilon d)\right)$, where $\Delta_f = f(\mathbf{x}_0) - f(\mathbf{x}^\star)$.*

III. Proof Sketch

A. Convergence to a first-order stationary point

We first bound the expectation of the square of the gradient norm in a minibatch step of SGLD-VR. To estimate the time needed to converge to a first-order stationary point (FSP), we compute the dependence of the gradient norm bound on the iteration count t. The quantity that plays a central role in the argument is the Lyapunov function, which is essential in constructing the upper bound for gradient norm and connects the argument between successive minibatches.

Lemma 4 (Bound of variance of SVRG gradient estimator [Reddi et al., 2016])**.** *In an epoch, the SVRG gradient estimator satisfies*

$$\mathbb{E}\left[\|\widetilde{\nabla}_t\|^2\right] \leq 2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + 2\frac{L^2}{B_{\rm e}}\mathbb{E}\left[\|\mathbf{x}_t - \widetilde{\mathbf{x}}\|^2\right]. \quad (11)$$

Adapting the framework in [Reddi et al., 2016] for the LD setting, the following lemma bounds the expectation of the gradient norm for the SGLD-VR iteration sequence in a minibatch:

Lemma 5. *Define the weight sequence (c_t) recursively as $c_t = c_{t+1}(1+\beta_t\eta_t + 2\frac{\eta_t^2 L^2}{B_{\rm e}}) + \frac{\eta_t^2 L^3}{B_{\rm e}}$ with $c_{B_{\rm e}} = 0$, and then define the Lyapunov function $R_t = \mathbb{E}\left[f(\mathbf{x}_t) + c_t\|\mathbf{x}_t - \widetilde{\mathbf{x}}\|^2\right]$ for each epoch. Define the normalization sequence $\gamma_t = \eta_t - \frac{c_{t+1}}{\beta_t}\eta_t - \eta_t^2 L - 2c_{t+1}\eta_t^2$ with η_t and $\beta_t > 0$ set to ensure $\gamma_t > 0$. Under Assumption 1, inside an epoch,*

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \leq \frac{R_t - R_{t+1}}{\gamma_t} + \left(\frac{L}{2} + c_{t+1}\right)\frac{d\rho_t^2}{\gamma_t}.$$

Remark 3 in the supplementary material shows that there always exists choices of η_0 and ν (hence η_t via (5)) and (β_t) to ensure $\gamma_t > 0$.

Now we use the bound of the gradient norm within a minibatch to build the norm bound of the SGLD-VR gradient estimator for the whole iteration in the following lemma, with which one can derive the time complexity for the SGLD-VR scheme to converge to a FSP as in Theorem 1:

Lemma 6. *Let $\bar{\gamma} = \min_{0\leq t\leq T-1}\gamma_t$ where γ_t is defined in the previous lemma, and $\nu > 0$. Then under Assumption 1,*

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_a)\|^2\right] \leq \frac{f(\mathbf{x}_0) - f(\mathbf{x}^\star)}{T\bar{\gamma}} + \frac{d}{\bar{\gamma}}\left(\frac{L}{2} + c_0\right)\frac{C_0}{T^\nu}, \quad (12)$$

where $\mathbf{x}_a$ is randomly chosen from the entire iterate sequence and C_0 is a universal constant.

B. Ergodicity

The ergodicity argument is comprised of two parts: recurrence and reachability.

a) Recurrence: The LD term in the optimization scheme, due to its random-walk nature, is the key for the reachability argument. In this section we follow the framework of [Chen et al., 2019] while giving new specific proofs.

We first show that with Langevin dynamics, the iteration process will visit sublevel sets of interest, for instance the collection of compact neighborhoods of all local minimums, infinitely many times. Lemma 7 is the first pillar to establish the ergodicity result. In its proof, we first give a more explicit characterization of function value decrease between two successive SGLD-VR updates than the characterization using the Lyapunov function R_t for the discussion of convergence to first-order stationary points in lemma 5. Next, we construct a supermartingale involving the objective function value and iteration count. Through the introduction of a stopping time sequence which records the time of the iteration visiting targeted sublevel sets, one can establish the expectation of any entry in this stopping time sequence, thus proving the lemma.

Our main lemma is Lemma 7 where we give an explicit upper bound of the expected time of visiting a given level set for the j-th time ($j \geq 1$).

Lemma 7 (Recurrence)**.** *For a fixed $\delta > 0$, let n_0 be the index such that $\eta_{n_0} \leq \delta$, and n_k be the sequence of iteration index $n_{k+1} = \min_s\{s : s > n_k,\, \eta_{n_k:s} \geq \delta\}$.*

Under Assumptions 1, 2 and 3, there exists a constant C_1 such that for constants $\alpha = 1 - 2\exp(-(1-C_1)\mu_1\delta)$, $B = 2(\psi_1 + \frac{2\eta_{n_0}L^3}{B_{\rm e}}\left(\mu_2 f(\mathbf{x}_0) + 2\psi_2\right) + \frac{\rho_0^2 Ld}{2\eta_0})$, $K = \frac{\ln\frac{f(\mathbf{x}_{n_0})}{\delta B}}{(1-C_1)\mu_1\delta}$, the stopping time sequence $\{\tau_k\}$ defined as $\tau_0 = K$ and $\tau_{k+1} = \min\{t : t \geq \tau_k + 1, f(\mathbf{x}_{n_t}) \leq 2\delta B\}$ satisfies

$$\mathbb{E}\left[\tau_j\right] \leq \frac{4}{\alpha} + K + j\left(\frac{1}{2\alpha\delta} + 1\right). \quad (13)$$

Remark 2. As we have assumed that $f \geq 0$ which is common for ERM problems since the loss function is usually nonnegative, it is desirable that $f(\mathbf{x}_{n_t})$ goes to 0 as the iteration proceeds. Thus, the choice of δ for analytical purposes would be $\delta \propto \frac{\widehat{\varepsilon}}{B}$ for some $\widehat{\varepsilon}$-target level one deems appropriate.

Method	Bounded	Grad. Lip.	Hess. Lip.	Regularization	Other assumptions
[Raginsky et al., 2017]	f and $\|\nabla f\|$	yes	no	(m, b)-dissipative	1) stoch. grad. sub-exp. tails 2) init. pt. sub-Gauss. tails
[Xu et al., 2018]	none	yes	no	(m, b)-dissipative	none
[Zhang et al., 2017]	$\|\nabla f\|$ and $\|\nabla^2 f\|$	yes	yes	$(1, 0)$-dissipative	grad. sub-exp. tails
This work	none	yes	yes	Assumption 3	strict saddle

TABLE II: Comparison between assumptions made for variants of LD optimization schemes. The Hessian Lipschitz assumption is used only for claims about second-order convergence.

b) Reachability: As lemma 7 shows, when the expected time for the iterates to revisit a certain level set of interest for j-th time is finite (j is any positive integer), we call such a level set recurrent. We show that when the SGLD iteration sequence starts from a recurrent compact set, there is a positive possibility for the sequence to visit every nearby first-order stationary point.

Lemma 8 is the core lemma to establish the ergodicity result for the LD optimization scheme. The core idea behind its proof is to leverage the exploratory potential of a *radial* Brownian motion process to show that there is a non-trivial probability for the Gaussian noise accumulation in the LD scheme to visit a pre-designated point in space.

Lemma 8 (Ergodicity due to Brownian motion)**.** *Given any sequence $a_k > 0$, let $\mathbf{z}_k = \sum_{i=1}^{k} \rho_0 \sqrt{a_i} \boldsymbol{\epsilon}_i$ where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, I_d)$, for any target vector $\mathbf{z}^\star$ and distance r, there exists a positive function p_1 such that*

$$\mathbf{Pr}(\|\mathbf{z}_n - \mathbf{z}^\star\| \leq r,\ \|\mathbf{z}_k\| \leq \|\mathbf{z}^\star\| + r\ \forall k = 1, \cdots, n) \geq p_1(r, \rho_0, t_n, \mathbf{z}^\star) \quad (14)$$

with $p_1(0, \rho_0, t_n, \mathbf{z}^\star) = p_1(r, 0, t_n, \mathbf{z}^\star) = 0$.

We leverage Lemma 8 to show the reachability of SGLD-VR scheme, which is the second pillar to establish the ergodicity result. The core idea behind the proof of Lemma 9 is to balance the influence on the iterates from gradient descent and Gaussian noise accumulation respectively, and show that the exploratory potential behind the Gaussian noise accumulation will fulfill the desired property of reachability.

Lemma 9 (Reachability)**.** *Assume the same stepsize batch setting as in Lemma 7 and the gradient Lipschitz condition in Assumption 1, with respect to an arbitrary target point $\mathbf{s}$ in the level set $\{\mathbf{x} : f(\mathbf{x}) \leq \varepsilon\}$. there is a constant $C_{21} \propto dL$ and constant C_{22} such that for any $\varepsilon > 0$,*

$$\mathbf{Pr}\left(\|\mathbf{x}_{n_{i+1}} - \mathbf{s}\| \leq \widetilde{\varepsilon}\right) > \frac{1}{2} p_1(\widetilde{\varepsilon}, \rho_0, t_{n_{i+1}}, \mathbf{s}) \quad (15)$$

where $\widetilde{\varepsilon} = \varepsilon + \delta C_{21} + 2\delta\sqrt{C_{22}}$.

Theorem 2 follows by bounding the probability $\mathbf{Pr}(\tau_\star > T)$ for some predesignated T, where we define $\tau_* = \min\{t : t > 0, \|\mathbf{x}_{n_t} - \mathbf{s}\| \leq \widetilde{\varepsilon}\}$. With the marker sequence of iteration index $n_k = \min_s\{s : s > n_{k-1}, \eta_{n_{k-1}:s} \geq \frac{\widetilde{\varepsilon}}{2B}\}$ and the auxiliary stopping time sequence $\tau_0 = K$, $\tau_{t+1} = \min\{t : t \geq \tau_k + 1, f(\mathbf{x}_{n_t}) \leq \widetilde{\varepsilon}\}$, where B and K are objective-specific constants, one writes $\mathbf{Pr}(\tau_* \geq T) = \mathbf{Pr}(\tau_* \geq T, \tau_J > T) + \mathbf{Pr}(\tau_* \geq T, \tau_J < T)$ and bound each term by Lemma 7 and Lemma 9 respectively.

C. Convergence to a second-order stationary point

By far in the literature there are two common ways to argue the convergence to second-order stationary points (SSP)

- show that $f(\mathbf{x}_T) - f(\mathbf{x}_0) < \Delta_f$ with probabilistic guarantee to ensure the continual function value decrease at saddle point ([Jin et al., 2017])
- show that $\|\mathbf{x}_T - \mathbf{x}^\star\|$ decreases in the probabilistic sense as T increases ([Kleinberg et al., 2018]).
  The time complexity of this approach has the exponential dependency on the inverse of the error tolerance. So in this work we resort to the previous approach.

The argument to show sufficient function value decrease from a FSP in [Jin et al., 2017] uses two iterate sequences to demonstrate the continual function value decrease at saddle point. Now that the noise is injected at every iteration, the geometric intuition that the trapping region is thin plus the probabilistic argument should be able to give a similar proof.

In the LD setting we exploit the property of Brownian motion to show the escape from saddle point, i.e. to characterize the perturbed iterate has high probability in the direction of descent,

$$(\mathbf{x}_t - \mathbf{x}_{\text{fsp}})^\intercal \nabla^2 f(\mathbf{x}_{\text{fsp}})(\mathbf{x}_t - \mathbf{x}_{\text{fsp}}) \leq -\zeta$$

The proof contains four steps:

1) We show that $\Delta_i := \sum_{l=n_i}^{n_{i+1}-1} \sqrt{\eta_i}\boldsymbol{\epsilon}_i$ will lead to saddle point escape, i.e. $\Delta_i^\intercal \nabla^2 f(\mathbf{x}_{\text{fsp}})\Delta_i \leq -\zeta$. Specifically, show that Δ_i has projection on the direction of $\lambda_{\min}$ more than ζ with high probability, which exploits the property of Brownian motion and the idea that the trapping region is thin when faced with LD ([Jin et al., 2017]).
2) We show that when $\mathbf{x} \in \mathcal{U}(\mathbf{x}_{\text{fsp}}, r)$ where $\|\Delta_j\| < r$ for $j = n_i, n_i + 1, \cdots, n_{i+1} - 1$, $\|\nabla f(\mathbf{x})\| < \varepsilon$, thus the first order expansion does not contribute to function value change.
3) We show that the update $\mathbf{x}' = \mathbf{x} + \Delta_i$ will lead to function value decrease, thus the SGLD algorithm has to terminate, thus converging to SSP.
4) Compute τ_{SSP} by taking account of the time needed for escaping saddle points and the time needed for achieving sufficient function value decrease.

IV. Conclusion

In this paper we consider the application of the scheme stochastic gradient Langevin dynamics with variance reduction on minimizing nonconvex objectives, prove the probabilistic convergence guarantee to local minimizers, and prove corresponding ergodicity property of the scheme which leads to non-trivial probability for the scheme to visit global minimizers.

Acknowledgments

This material is based upon work supported by the National Science Foundation under grant no. 1819251.

Zhishen Huang thanks Manuel Lladser for helpful discussions on properties of Brownian motion.

The research presented in this paper was performed when ZH was affiliated with Department of Applied Mathematics, University of Colorado Boulder.

References

[Allen-Zhu and Hazan, 2016] Allen-Zhu, Z. and Hazan, E. (2016). Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pages 699–707.

[Bauschke and Combettes, 2017] Bauschke, H. H. and Combettes, P. L. (2017). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer-Verlag, New York, 2 edition.

[Borkar and Mitter, 1999] Borkar, V. and Mitter, S. (1999). A strong approximation theorem for stochastic recursive algorithms. *Journal of Optimization Theory and Applications*, 100:499–513.

[Chen et al., 2019] Chen, X., Du, S. S., and Tong, X. T. (2019). On stationary-point hitting time and ergodicity of stochastic gradient Langevin dynamics. *arXiv e-prints*, page arXiv:1904.13016.

[Cheng and Bartlett, 2018] Cheng, X. and Bartlett, P. (2018). Convergence of Langevin MCMC in KL-divergence. In Janoos, F., Mohri, M., and Sridharan, K., editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 186–211. PMLR.

[Chiang et al., 1987] Chiang, T.-S., Hwang, C.-R., and Sheu, S. J. (1987). Diffusion for global optimization in $\mathbb{R}^n$. *SIAM Journal on Control and Optimization*, 25(3):737–753.

[Dalalyan and Karagulyan, 2017] Dalalyan, A. S. and Karagulyan, A. G. (2017). User-friendly guarantees for the Langevin monte carlo with inaccurate gradient.

[Defazio et al., 2014] Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 1646–1654. Curran Associates, Inc.

[Dong and Tong, 2020] Dong, J. and Tong, X. T. (2020). Replica exchange for non-convex optimization.

[Dubey et al., 2016] Dubey, A., Reddi, S. J., Póczos, B., Smola, A. J., Xing, E. P., and Williamson, S. A. (2016). Variance reduction in stochastic gradient Langevin dynamics. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 1162–1170, Red Hook, NY, USA. Curran Associates Inc.

[Ge et al., 2015] Ge, R., Huang, F., Jin, C., and Yuan, Y. (2015). Escaping from saddle points — online stochastic gradient for tensor decomposition. In Grünwald, P., Hazan, E., and Kale, S., editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 797–842, Paris, France. PMLR.

[Gelfand and Mitter, 1991] Gelfand, S. B. and Mitter, S. K. (1991). Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$. *SIAM Journal on Control and Optimization*, 29(5):999–1018.

[Huang and Becker, 2020] Huang, Z. and Becker, S. (2020). Perturbed proximal descent to escape saddle points for non-convex and non-smooth objective functions. In Oneto, L., Navarin, N., Sperduti, A., and Anguita, D., editors, *Recent Advances in Big Data and Deep Learning*, pages 58–77, Cham. Springer International Publishing.

[Jin et al., 2017] Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. (2017). How to escape saddle points efficiently. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1724–1732, International Convention Centre, Sydney, Australia. PMLR.

[Johnson and Zhang, 2013] Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 315–323. Curran Associates, Inc.

[Karimi et al., 2016] Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer.

[Karlin and Taylor, 1975] Karlin, S. and Taylor, H. M. (1975). Chapter 7 - brownian motion. In Karlin, S. and Taylor, H. M., editors, *A First Course in Stochastic Processes*, pages 340 – 391. Academic Press, Boston, 2nd edition.

[Kleinberg et al., 2018] Kleinberg, B., Li, Y., and Yuan, Y. (2018). An alternative view: When does SGD escape local minima? In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2698–2707, Stockholmsmässan, Stockholm Sweden. PMLR.

[Kramers, 1940] Kramers, H. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284 – 304.

[Lee et al., 2019] Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. (2019). First-order methods almost always avoid strict saddle points. *Math. Program.*, 176(1–2):311–337.

[Lee et al., 2016] Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. (2016). Gradient descent only converges to minimizers. In Feldman, V., Rakhlin, A., and Shamir, O., editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1246–1257, Columbia University, New York, New York, USA. PMLR.

[Li, 2019] Li, Z. (2019). SSRGD: Simple stochastic recursive gradient descent for escaping saddle points. In Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 1523–1533. Curran Associates, Inc.

[Liu and Yin, 2019] Liu, Y. and Yin, W. (2019). An envelope for davis—yin splitting and strict saddle-point avoidance. *J. Optim. Theory Appl.*, 181(2):567–587.

[Mattingly et al., 2002] Mattingly, J., Stuart, A., and Higham, D. (2002). Ergodicity for sdes and approximations: locally lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2):185 – 232.

[Mengersen and Tweedie, 1996] Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the hastings and metropolis algorithms. *The Annals of Statistics*, 24:101–121.

[Meyn and Tweedie, 2009] Meyn, S. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press, USA, 2nd edition.

[Mokhtari et al., 2018] Mokhtari, A., Ozdaglar, A., and Jadbabaie, A. (2018). Escaping saddle points in constrained optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3633–3643, Red Hook, NY, USA. Curran Associates Inc.

[Raginsky et al., 2017] Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703, Amsterdam, Netherlands. PMLR.

[Reddi et al., 2016] Reddi, S. J., Hefny, A., Sra, S., Póczós, B., and Smola, A. (2016). Stochastic variance reduction for nonconvex optimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 314–323. JMLR.org.

[Roberts and Tweedie, 1996a] Roberts, G. O. and Tweedie, R. L. (1996a). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

[Roberts and Tweedie, 1996b] Roberts, G. O. and Tweedie, R. L. (1996b). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110.

[Sun et al., 2019] Sun, T., Li, D., Quan, Z., Jiang, H., Li, S., and Dou, Y. (2019). Heavy-ball algorithms always escape saddle points. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3520–3526. International Joint Conferences on Artificial Intelligence Organization.

[Vlatakis-Gkaragkounis et al., 2019] Vlatakis-Gkaragkounis, E.-V., Flokas, L., and Piliouras, G. (2019). Efficiently avoiding saddle points with zero order methods: No gradients required. In Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 10066–10077. Curran Associates, Inc.

[Welling and Teh, 2011] Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 681–688, Madison, WI, USA. Omnipress.

[Xu et al., 2018] Xu, P., Chen, J., Zou, D., and Gu, Q. (2018). Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3126–3137, Red Hook, NY, USA. Curran Associates Inc.

[Zhang et al., 2017] Zhang, Y., Liang, P., and Charikar, M. (2017). A hitting time analysis of stochastic gradient Langevin dynamics. *arXiv e-prints*, page arXiv:1702.05575.

APPENDIX
SUPPLEMENTARY MATERIAL

A. Proofs of first-order stationary point convergence property

In this section we prove the result for first-order convergence property (theorem 1) as well as the needed lemmas.

Lemma 10 (Repeat of lemma 5)**.** *Define the weight sequence $\{c_t\}$ recursively as $c_t = c_{t+1}(1+\beta_t\eta_t + 2\frac{\eta_t^2 L^2}{B_\mathrm{e}}) + \frac{\eta_t^2 L^3}{B_\mathrm{e}}$ with $c_{B_\mathrm{e}} = 0$, and then define the Lyapunov function $R_t = \mathbb{E}\left[f(\mathbf{x}_t) + c_t\|\mathbf{x}_t - \widetilde{\mathbf{x}}\|^2\right]$ for each epoch. Define the normalization sequence $\gamma_t = \eta_t - \frac{c_{t+1}}{\beta_t}\eta_t - \eta_t^2 L - 2c_{t+1}\eta_t^2$ with η_t and $\beta_t > 0$ set to ensure $\gamma_t > 0$. Under Assumption 1, inside an epoch,*

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \le \frac{R_t - R_{t+1}}{\gamma_t} + \left(\frac{L}{2} + c_{t+1}\right)\frac{d\rho_t^2}{\gamma_t}.$$

Proof. We find upper bounds to the Lyapunov functions R_t in terms of the negative norm of the SVRG gradient estimator, thus proving the lemma. We bound the two terms in the Lyapunov functions respectively. For notational simplicity let $\nabla f(\mathbf{x}_t) = \nabla_t = \mathbb{E}_{I_t}[\widetilde{\nabla}_t]$.

For the first term $f(\mathbf{x}_{t+1})$ in the Lyapunov function, using Prop. 12,

$$\mathbb{E}\left[f(\mathbf{x}_{t+1})\right] = \mathbb{E}\left[f(\mathbf{x}_t) - \eta_t\|\nabla_t\|^2 + \frac{L}{2}\left(\eta_t^2\|\widetilde{\nabla}_t\|^2 + \rho_t^2\|\boldsymbol{\epsilon}_t\|^2\right)\right].$$

For the second term $\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}\|$, as $\langle\nabla_t, \widetilde{\mathbf{x}} - \mathbf{x}_t\rangle \overset{\text{CS}}{\le} \|\nabla_t\|\|\mathbf{x}_t - \widetilde{\mathbf{x}}\| \overset{\text{Young}}{\le} \frac{1}{2\beta_t}\|\nabla_t\|^2 + \frac{\beta_t}{2}\|\mathbf{x}_t - \widetilde{\mathbf{x}}\|^2$,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}\|^2\right] &= \mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t + \mathbf{x}_t - \widetilde{\mathbf{x}}\|^2\right] = \mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}\|^2 + 2\langle\mathbf{x}_{t+1} - \mathbf{x}_t, \mathbf{x}_t - \widetilde{\mathbf{x}}\rangle\right] \\
&= \mathbb{E}\left[\eta_t^2\|\widetilde{\nabla}_t\|^2 + \rho_t^2\|\boldsymbol{\epsilon}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}\|^2 + 2\eta_t\langle\nabla_t, \widetilde{\mathbf{x}} - \mathbf{x}_t\rangle\right] \\
&\le \mathbb{E}\left[\eta_t^2\|\widetilde{\nabla}_t\|^2 + \rho_t^2\|\boldsymbol{\epsilon}_t\|^2 + (1+\eta_t\beta_t)\|\mathbf{x}_t - \widetilde{\mathbf{x}}\|^2 + \frac{\eta_t}{\beta_t}\|\nabla_t\|^2\right]
\end{aligned}$$

Putting these two terms together into R_{t+1}, we have

$$\begin{aligned}
R_{t+1} &\le \mathbb{E}\left[f(\mathbf{x}_t) + \left(\frac{\eta_t c_{t+1}}{\beta_t} - \eta_t\right)\|\nabla_t\|^2 + \left(\frac{L}{2} + c_{t+1}\right)(\eta_t^2\|\widetilde{\nabla}_t\|^2 + \rho_t^2\|\boldsymbol{\epsilon}_t\|^2) + (1+\eta_t\beta_t)c_{t+1}\|\mathbf{x}_t - \widetilde{\mathbf{x}}\|^2\right] \\
&\overset{(11)}{\le} \mathbb{E}\Big[f(\mathbf{x}_t) + \left(\frac{\eta_t c_{t+1}}{\beta_t} - \eta_t + (L + 2c_{t+1})\eta_t^2\right)\|\nabla_t\|^2 + \left(\frac{L}{2} + c_{t+1}\right)\rho_t^2\|\boldsymbol{\epsilon}_t\|^2 \\
&\qquad + \big((1+\eta_t\beta_t)c_{t+1} + (L + 2c_{t+1})\frac{\eta_t^2 L^2}{B_\mathrm{e}}\big)\|\mathbf{x}_t - \widetilde{\mathbf{x}}\|^2\Big] \\
&= \mathbb{E}\left[f(\mathbf{x}_t) - \gamma_t\|\nabla_t\|^2 + \left(\frac{L}{2} + c_{t+1}\right)\rho_t^2\|\boldsymbol{\epsilon}_t\|^2 + c_t\|\mathbf{x}_t - \widetilde{\mathbf{x}}\|^2\right] \\
&= R_t - \mathbb{E}\left[\gamma_t\|\nabla_t\|^2\right] + \mathbb{E}\left[\left(\frac{L}{2} + c_{t+1}\right)\rho_t^2\|\boldsymbol{\epsilon}_t\|^2\right] \\
&= R_t - \mathbb{E}\left[\gamma_t\|\nabla_t\|^2\right] + \left(\frac{L}{2} + c_{t+1}\right)\rho_t^2 d.
\end{aligned}$$

We set $\{\beta_t\}$ and η_0 properly (see the remark below this proof) such that $-\gamma_t = \frac{\eta_t c_{t+1}}{\beta_t} - \eta_t + (L + 2c_{t+1})\eta_t^2 \le 0$ for all $t = 0, 1, \cdots, B_\mathrm{e} - 1$. This can always be achieved as c_t is a decreasing sequence and c_t is negatively related to β_t. Then

$$\mathbb{E}\left[\gamma_t\|\nabla_t\|^2\right] \le -R_{t+1} + R_t + \left(\frac{L}{2} + c_{t+1}\right)\rho_t^2 d. \tag{16}$$

□

Remark 3. We show that the η_0 and $\{\beta_t\}$ sequence setting in the end of the proof of lemma 5 always exists. For now we can assume $\beta_t = \widetilde{\beta}$ is a constant. Then we can define an upper bound sequence for $\{c_t\}$ as

$$\widetilde{c}_t = \widetilde{c}_{t+1}(1 + \widetilde{\beta}\eta_0 + 2\frac{\eta_0^2 L^2}{B_\mathrm{e}}) + \frac{\eta_0^2 L^3}{B_\mathrm{e}}$$

with $\widetilde{c}_{B_\mathrm{e}} = 0$. Then, $c_t \le \widetilde{c}_t$ for $1 \le t \le B_\mathrm{e}$. Consequently, for expression simplicity assuming $q = 1 + \widetilde{\beta}\eta_0 + 2\frac{\eta_0^2 L^2}{B_\mathrm{e}}$ and $D = \frac{\frac{\eta_0^2 L^3}{B_\mathrm{e}}}{\widetilde{\beta}\eta_0 + \frac{2\eta_0^2 L^2}{B_\mathrm{e}}} = \frac{\frac{\eta_0 L^3}{B_\mathrm{e}}}{\widetilde{\beta} + \frac{2\eta_0 L^2}{B_\mathrm{e}}}$, we have

$$\frac{\widetilde{c}_t + D}{\widetilde{c}_{t+1} + D} = q.$$

It follows that $\frac{1}{q^{B_e}}(\widetilde{c}_0 + D) = \widetilde{c}_{B_e} + D = D$, and $\widetilde{c}_0 = (q^{B_e} - 1)D$. We need to set $\widetilde{\beta}$ in a way such that $\gamma_t > 0$ for all $1 \le t \le B_e$. As

$$\gamma_t \ge \left(1 - \frac{\widetilde{c}_0}{\widetilde{\beta}} - \eta_0 L - 2\widetilde{c}_0\eta_0\right)\eta_t \overset{\text{need}}{>} 0,$$

a sufficient condition to assure the second inequality above is

$$\widetilde{c}_0\left(\frac{1}{\widetilde{\beta}} + 2\eta_0\right) + \eta_0 L < 1 \tag{17}$$

Let $\widetilde{\beta}\eta_0$ be small while $\widetilde{\beta} > 1$, then the l.h.s. of (17) is of the order $B_e\eta_0 \frac{\frac{\widetilde{\beta}\eta_0 L^3}{B_e}}{\widetilde{\beta}^2 + 2\frac{\widetilde{\beta}\eta_0 L^2}{B_e}}$, which can ensure (17) to hold.

Lemma 11 (Repeat of lemma 6)**.** *Let $\bar{\gamma} = \min_{0\le t\le T-1}\gamma_t$ where γ_t is defined in the previous lemma, and $\nu > 0$. Then under Assumption 1,*

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_a)\|^2\right] \le \frac{f(\mathbf{x}_0) - f(\mathbf{x}^\star)}{T\bar{\gamma}} + \frac{d}{\bar{\gamma}}\left(\frac{L}{2} + c_0\right)\frac{C_0}{T^\nu}, \tag{18}$$

where $\mathbf{x}_a$ is randomly chosen from the entire iterate sequence and C_0 is a universal constant.

Proof. We set $c_{B_e} = 0$ so that $R_0^{(\alpha)} = f(\mathbf{x}_0^{(\alpha)})$ and $R_{B_e}^{(\alpha)} = f(\mathbf{x}_{B_e}^{(\alpha)})$ for the fixed epoch α. Per line 10 in Algorithm 1, the ending point of the previous epoch is the starting point of the next epoch, i.e., $\mathbf{x}_0^{(\alpha)} = \mathbf{x}_{B_e}^{(\alpha-1)}$. Summing up all the iteration steps in each epoch, we have

$$\sum_{\alpha=0}^{\frac{T}{B_e}-1}\sum_{l=0}^{B_e-1}\mathbb{E}\left[\nabla f(x_l^{(\alpha)})\right] \le \frac{f(\mathbf{x}_0) - f(\mathbf{x}_T)}{\bar{\gamma}} + \frac{\mathbb{E}\left[\|\boldsymbol{\epsilon}\|^2\right]}{\bar{\gamma}}\sum_{t=0}^{T-1}\left(\frac{L}{2} + c_{(t \bmod B_e)+1}\right)\rho_t^2$$

When ρ_t is set as $\mathcal{O}(\frac{1}{t^{\nu/2}})$ where $\nu \ge 1$, as c_t is bounded w.r.t. a fixed epoch, $\sum_{t=0}^{T-1}\rho_t^2 = \mathcal{O}(T^{1-\nu})$. (The $\nu = 1$ case leads to logarithmic growth of summation of ρ_t^2, which does not affect the following result.) Then consider the LHS of the inequality as the average over all iterates, then

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f(\mathbf{x}_a)\|^2\right] &\le \frac{f(\mathbf{x}_0) - f(\mathbf{x}^\star)}{T\bar{\gamma}} + \frac{\mathbb{E}\left[\|\boldsymbol{\epsilon}\|^2\right]}{T\bar{\gamma}}\sum_{t=0}^{T-1}\left(\frac{L}{2} + c_{(t \bmod B_e)+1}\right)\rho_t^2 \\
&\le \frac{f(\mathbf{x}_0) - f(\mathbf{x}^\star)}{T\bar{\gamma}} + \frac{d}{T\bar{\gamma}}\left(\frac{L}{2} + c_0\right)\sum_{t=0}^{T-1}(\frac{L}{2} + c_0)\rho_t^2 \\
&= \frac{f(\mathbf{x}_0) - f(\mathbf{x}^\star)}{T\bar{\gamma}} + \frac{d}{\bar{\gamma}}\left(\frac{L}{2} + c_0\right)\frac{C_0}{T^\nu}.
\end{aligned} \tag{19}$$

□

Proof of Thm. 1. Per (19), we see that the time complexity for the LD to converge to an ε-first order stationary point is $\mathcal{O}\big(\frac{\Delta_f d}{\bar{\gamma}\varepsilon^2}\big)$. Another way to phrase the time complexity is through the hitting time of LD to a first-order stationary point (fsp) τ_{fsp}. To estimate the expected time for the iteration sequence to enter a fsp neighborhood,

$$\begin{aligned}
\mathbf{Pr}(\tau_{\text{fsp}} > T) = \mathbf{Pr}(\|\nabla f(\mathbf{x}_t)\| > \varepsilon,\ \forall t \le T) &\le \mathbf{Pr}\left(\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\| > \varepsilon\right) \\
&\le \frac{\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|\right]}{\varepsilon} \\
&= \frac{\mathbb{E}\left[\|\nabla f(\mathbf{x}_a)\|\right]}{\varepsilon} \le \frac{\sqrt{\mathbb{E}\left[\|\nabla f(\mathbf{x}_a)\|^2\right]}}{\varepsilon},
\end{aligned}$$

where the 2nd inequality is due to Markov's inequality, and the expectation in the final line is taken over choosing a uniformly from $\{1, \ldots, T\}$ in addition to the other random variables, and the final inequality is Jensen's inequality.

Thus, using Lemma 6,

$$\mathbf{Pr}(\tau_{\text{fsp}} > T) \le \frac{1}{\varepsilon}\sqrt{\frac{\Delta_f}{T\bar{\gamma}} + \frac{d}{\bar{\gamma}}\left(\frac{L}{2} + c_0\right)\frac{C_0}{T^\nu}} \overset{\text{let}}{=} p, \tag{20}$$

where p is the failure probability. As $\bar{\gamma}$ is a positive constant independent of d, ε and T, the equation above transforms into $\frac{\Delta_f}{T\bar{\gamma}} + \frac{d}{\bar{\gamma}}\left(\frac{L}{2} + c_0\right)\frac{C_0}{T^\nu} = \varepsilon^2 p$. As $\nu \ge 1$, $T = \mathcal{O}\left(\frac{\Delta_f d}{\bar{\gamma}\varepsilon^2 p}\right)$. □

B. Proofs of ergodicity properties

Let $\mathcal{F}_t$ be the filtration generated by $(\mathbf{x}_0, \dots, \mathbf{x}_t)$ and associated random variables (I_t and $\boldsymbol{\epsilon}_t$) for $(\mathbf{x}_t)$ the sequence from Algo. 1. We write $\mathbb{E}[\cdot \mid \mathcal{F}_t]$ as just $\mathbb{E}[\cdot]$ when the conditioning is clear from context.

We start with a simple proposition that will be used in several of the proofs.

Proposition 12 (variant of the Descent Lemma)**.** *For $\mathbf{x}_t$ generated via Algo. 1, under the Lipschitz assumption 1, then the expectation conditioned on $\mathcal{F}_t$ satisfies*

$$\begin{aligned}\mathbb{E}[f(\mathbf{x}_{t+1})] &\leq \mathbb{E}\left[f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2\right] \\ &= \mathbb{E}\left[f(\mathbf{x}_t) - \eta_t\|\nabla_t\|^2 + \frac{L}{2}\big(\eta_t^2\|\widetilde{\nabla}_t\|^2 + \rho_t^2\|\boldsymbol{\epsilon}_t\|^2\big)\right] \end{aligned} \tag{21}$$

Proof. The first inequality uses the L-smoothness of function f and the second equality uses the SVRG update in algorithm 1 ($\mathbf{x}_{t+1} - \mathbf{x}_t = -\eta_t\widetilde{\nabla}_t + \rho_t\boldsymbol{\epsilon}_t$) and the unbiasedness of the gradient estimator $\widetilde{\nabla}_t$. □

1) Recurrence:

Lemma 13 (Repeat of lemma 7)**.** *For a fixed $\delta > 0$, let n_0 be the index such that $\eta_{n_0} \leq \delta$, and n_k be the sequence of iteration index $n_{k+1} = \min_s\{s : s > n_k,\, \eta_{n_k:s} \geq \delta\}$.*

Under the regularization assumption 3, Lipschitz assumption 1 and nonnegativity assumption 2, there exists a constant C_1 such that for $\alpha = 1 - 2\exp(-(1-C_1)\mu_1\delta)$, $B = 2(\psi_1 + \frac{2\eta_{n_0}L^3}{B_\mathrm{e}}\big(\mu_2 f(\mathbf{x}_0) + 2\psi_2\big) + \frac{\rho_0^2 Ld}{2\eta_0})$ and $K = \frac{\ln\frac{f(\mathbf{x}_{n_0})}{\delta B}}{(1-C_1)\mu_1\delta}$, the stopping time sequence $\{\tau_k\}$, defined as $\tau_0 = K$ and $\tau_{k+1} = \min\{t : t \geq \tau_k + 1, f(\mathbf{x}_{n_t}) \leq 2\delta B\}$ where $(\mathbf{x}_{t'})$ is the sequence generated by Algo. 1, satisfies

$$\mathbb{E}[\tau_j] \leq \frac{4}{\alpha} + K + j\left(\frac{1}{2\alpha\delta} + 1\right). \tag{22}$$

Proof. Conditioned on $\mathcal{F}_t$ and $f(\widetilde{\mathbf{x}}^{(s)}) < f(\mathbf{x}_0)$ for the largest s such that $t \geq sB_\mathrm{e}$, we have

$$\begin{aligned}\mathbb{E}[f(\mathbf{x}_{t+1})] &\overset{(21)}{\leq} f(\mathbf{x}_t) - \eta_t\|\nabla_t\|^2 + \mathbb{E}\,\frac{\eta_t^2 L}{2}\|\widetilde{\nabla}_t\|^2 + \frac{\rho_t^2 L}{2}\mathbb{E}\,\|\boldsymbol{\epsilon}_t\|^2 \\ &\overset{(11)}{\leq} f(\mathbf{x}_t) - \eta_t\|\nabla_t\|^2 + \frac{\eta_t^2 L}{2}\left(2\|\nabla_t\|^2 + 2\frac{L^2}{B_\mathrm{e}}\|\mathbf{x}_t - \widetilde{\mathbf{x}}\|^2\right) + \frac{\rho_t^2 Ld}{2} \\ &= f(\mathbf{x}_t) - (\eta_t - \eta_t^2 L)\|\nabla_t\|^2 + \frac{\eta_t^2 L^3}{B_\mathrm{e}}\|\mathbf{x}_t - \widetilde{\mathbf{x}}\|^2 + \frac{\rho_t^2 Ld}{2} \\ &\overset{(7)}{\leq} f(\mathbf{x}_t) - (\eta_t - \eta_t^2 L)\|\nabla_t\|^2 + \frac{2\eta_t^2 L^3}{B_\mathrm{e}}\big(\mu_2(f(\mathbf{x}_t) + f(\mathbf{x}_0)) + 2\psi_2\big) + \frac{\rho_t^2 Ld}{2} \\ &= (1 + \frac{2\eta_t^2 L^3\mu_2}{B_\mathrm{e}})f(\mathbf{x}_t) - (\eta_t - \eta_t^2 L)\|\nabla_t\|^2 + \frac{2\eta_t^2 L^3}{B_\mathrm{e}}\big(\mu_2 f(\mathbf{x}_0) + 2\psi_2\big) + \frac{\rho_t^2 Ld}{2} \\ &\overset{(6)}{\leq} (1 + \frac{2\eta_t^2 L^3\mu_2}{B_\mathrm{e}})f(\mathbf{x}_t) - (\eta_t - \eta_t^2 L)\big(\mu_1 f(\mathbf{x}_t) - \psi_1\big) + \frac{2\eta_t^2 L^3}{B_\mathrm{e}}\big(\mu_2 f(\mathbf{x}_0) + 2\psi_2\big) + \frac{\rho_t^2 Ld}{2} \\ &= \left(1 - \mu_1\eta_t + \eta_t^2\left(\frac{2L^3\mu_2}{B_\mathrm{e}} + \mu_1 L\right)\right) f(\mathbf{x}_t) + (\eta_t - \eta_t^2 L)\psi_1 + \frac{2\eta_t^2 L^3}{B_\mathrm{e}}\big(\mu_2 f(\mathbf{x}_0) + 2\psi_2\big) + \frac{\rho_t^2 Ld}{2} \\ &\leq \exp\left(-(1-C_1)\mu_1\eta_t\right) f(\mathbf{x}_t) + (\eta_t - \eta_t^2 L)\psi_1 + \frac{2\eta_t^2 L^3}{B_\mathrm{e}}\big(\mu_2 f(\mathbf{x}_0) + 2\psi_2\big) + \frac{\rho_t^2 Ld}{2} \end{aligned} \tag{23}$$

Here C_1 is a positive constant such that $\eta_0(\frac{2L^3\mu_2}{\mu_1 B_\mathrm{e}} + L) < C_1 < 1$ for small enough η_0.

We introduce an index partition to characterize the function value decrease. Let n_0 be the index such that $\eta_{n_0} \leq \delta$, and n_k be the sequence of iteration index $n_{k+1} = \min_s\{s : s > n_k,\, \eta_{n_k:s} \geq \delta\}$. Then $\eta_{n_k:n_{k+1}} \leq 2\delta$.

Before the proof proceeds, we recall the that the stepsize and variance are set (for some $\nu \geq 1$) as

$$\eta_t = \frac{\eta_0}{t^\nu} \text{ and } \rho_t = \frac{\rho_0}{t^{\nu/2}}.$$

Thus $\rho_t = \rho_0 \sqrt{\frac{\eta_t}{\eta_0}}$. Iterating (23) m times, we have

$$\begin{aligned}
\mathbb{E}\, f(\mathbf{x}_{t+m}) &\le \exp\left(-(1-C_1)\mu_1 \eta_{t:t+m-1}\right) f(\mathbf{x}_t) + \\
&\qquad \sum_{i=t}^{t+m-1} \exp\left(-(1-C_1)\mu_1 \eta_{i+1:t+m-1}\right) \eta_i \left(\psi_1 + \frac{2\eta_i L^3}{B_{\mathrm{e}}}\left(\mu_2 f(\mathbf{x}_0) + 2\psi_2\right) + \frac{\rho_0^2 L d}{2\eta_0}\right) \\
&\le \exp\left(-(1-C_1)\mu_1 \eta_{t:t+m-1}\right) f(\mathbf{x}_t) + \sum_{i=t}^{t+m-1} \eta_i \left(\psi_1 + \frac{2\eta_i L^3}{B_{\mathrm{e}}}\left(\mu_2 f(\mathbf{x}_0) + 2\psi_2\right) + \frac{\rho_0^2 L d}{2\eta_0}\right) \\
&\le \exp\left(-(1-C_1)\mu_1 \eta_{t:t+m-1}\right) f(\mathbf{x}_t) + \eta_{t:t+m-1} \left(\psi_1 + \frac{2\eta_t L^3}{B_{\mathrm{e}}}\left(\mu_2 f(\mathbf{x}_0) + 2\psi_2\right) + \frac{\rho_0^2 L d}{2\eta_0}\right) \qquad (24)
\end{aligned}$$

Setting $t = n_{k-1}$ and $m = n_k - n_{k-1}$, inequality (24) takes the form

$$\begin{aligned}
\mathbb{E}\, f(\mathbf{x}_{n_k}) &\le \exp\left(-(1-C_1)\mu_1 \eta_{n_{k-1}:n_k-1}\right) f(\mathbf{x}_{n_{k-1}}) + \eta_{n_{k-1}:n_k-1} \left(\psi_1 + \frac{2\eta_{n_{k-1}} L^3}{B_{\mathrm{e}}}\left(\mu_2 f(\mathbf{x}_0) + 2\psi_2\right) + \frac{\rho_0^2 L d}{2\eta_0}\right) \\
&\le \exp\left(-(1-C_1)\mu_1 \delta\right) f(\mathbf{x}_{n_{k-1}}) + \eta_{n_{k-1}:n_k-1} \left(\psi_1 + \frac{2\eta_{n_{k-1}} L^3}{B_{\mathrm{e}}}\left(\mu_2 f(\mathbf{x}_0) + 2\psi_2\right) + \frac{\rho_0^2 L d}{2\eta_0}\right) \qquad (25) \\
&\le \exp\left(-(1-C_1)k\mu_1 \delta\right) f(\mathbf{x}_{n_0}) + \delta \underbrace{2\left(\psi_1 + \frac{2\eta_{n_0} L^3}{B_{\mathrm{e}}}\left(\mu_2 f(\mathbf{x}_0) + 2\psi_2\right) + \frac{\rho_0^2 L d}{2\eta_0}\right)}_{:=B}. \qquad (26)
\end{aligned}$$

Consider a function value threshold $M := 2\delta B$. From (26), it follows that $\mathbb{E}\, f(\mathbf{x}_{n_k}) \le M$ when

$$k \ge \frac{\ln \frac{f(\mathbf{x}_{n_0})}{\delta B}}{(1-C_1)\mu_1 \delta} := K$$

Now we show that the expected time for the function value to decrease to below this threshold M is upper bounded by a finite number, thus justifying the recurrence of the iteration process to a compact sub-level set. (Note that under the regularization assumption (7), all sub-level sets are compact.) To better exploit the indices partition $\{n_k\}$ of the iteration sequence, define $f(\mathbf{x}_{n_k}) := V_k$, and $\tau = \min\{k : k \ge K, f(\mathbf{x}_{n_k}) \le M\}$. We claim that

$$V_{\tau \wedge k} + \alpha\delta B \cdot (\tau \wedge k)$$

is a supermartingale with $\alpha = 1 - 2\exp\left(-(1-C_1)\mu_1\delta\right)$, i.e.

$$\mathbb{E}\left[V_{\tau\wedge(k+1)} + \alpha\delta B(\tau \wedge (k+1))|V_{\tau\wedge k}\right] \le V_{\tau\wedge k} + \alpha\delta B(\tau \wedge k) \qquad (27)$$

When $\tau \le k$, (27) holds trivially. When $\tau \ge k+1$, then $V_{k+1} > M$. The relation (27) to show in this case takes the form $\alpha\delta B \le V_k - \mathbb{E}\left[V_{k+1}|V_k\right]$. To let this happen, taking inequality (25) into consideration, a sufficient condition is $\mathbb{E}\left[V_{k+1}|V_k\right] \le \exp(-(1-C_1)\mu_1\delta)V_k + \delta B \le V_k - \alpha\delta B$, i.e. $(1+\alpha)\delta B \le (1 - \exp(-(1-C_1)\mu_1\delta))\, V_k$. Considering that $\tau > k+1$ implies $V_k > M$, the previous sufficient condition to show can be further strengthened to $(1+\alpha)\delta B \le (1 - \exp(-(1-C_1)\mu_1\delta)M$, which is catered for per definition of α.

To show that a sub-level set is going to be visited by the iteration sequence for infinitely many times, we introduce the stopping time sequence $\{\tau_k\}$ where $\tau_0 = K$ and $\tau_{k+1} = \min\{t : t \ge \tau_k + 1, f(\mathbf{x}_{n_t}) \le M\}$. Per the same argument as in previous paragraph, $\mathbb{E}\left[V_{\tau_{k+1}} + \alpha\delta B\tau_{k+1}\middle|\tau_k\right] \le V_{\tau_k+1} + \alpha\delta B(\tau_k + 1)$, which gives

$$\alpha\delta B\, \mathbb{E}\left[\tau_{k+1} - \tau_k - 1|\tau_k\right] \le V_{\tau_k+1} - \mathbb{E}\left[V_{\tau_{k+1}}|\tau_k\right]$$

Taking total expectation, and summing over all k from 0 to j with $\tau_0 = K$, we have

$$\alpha\delta B\left(\mathbb{E}\left[\tau_j\right] - K - j\right) \le \sum_{k=0}^{j} \mathbb{E}\left[V_{\tau_k+1} - V_{\tau_{k+1}}\right] \qquad (28)$$

By (23), $\mathbb{E}\left[V_{\tau_k+1}\right] \le \exp\left(-(1-C_1)\mu_1\eta_{\tau_k}\right) V_{\tau_k} + \frac{B}{2} \le V_{\tau_k} + \frac{B}{2}$, thus

$$\alpha\delta B\left(\mathbb{E}\left[\tau_j\right] - K - j\right) \le \mathbb{E}\left[V_K - V_{\tau_j}\right] + j\frac{B}{2} \le 2M + j\frac{B}{2}$$

i.e.

$$\mathbb{E}\left[\tau_j\right] \le \frac{4}{\alpha} + K + j\left(\frac{1}{2\alpha\delta} + 1\right)$$

□

2) Reachability: The following Lemma 14 is stated as a fact, whose proof is straightforward computation, and will be needed for bounding the variance of the variance-reduced gradient estimator later in this section.

Lemma 14 (Variance of subset selection)**.** *Consider a dataset $\{\mathbf{a}_i\}_{i=1}^N$ with mean*

$$\bar{\mathbf{a}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{a}_i.$$

Select b elements uniformly ($1 \le b \le N$) out of this dataset, and denote the index set of these selected elements as $\mathcal{I}$. The subsampled mean, which is a random variable, is

$$\boldsymbol{\xi} = \frac{1}{b}\sum_{i\in\mathcal{I}}\mathbf{a}_i.$$

The variance of $\boldsymbol{\xi}$ is

$$\begin{aligned}\mathbb{E}_{\mathcal{I}}\|\boldsymbol{\xi}-\bar{\mathbf{a}}\|^2 &= \mathbb{E}_{\mathcal{I}}(\boldsymbol{\xi}^2 - 2\langle\boldsymbol{\xi},\bar{\mathbf{a}}\rangle + \bar{\mathbf{a}}^2) = \mathbb{E}_{\mathcal{I}}\boldsymbol{\xi}^2 - \bar{\mathbf{a}}^2 \\ &= \frac{N-b}{N^2 b}\mathrm{Var}[\mathbf{a}] = \frac{N-b}{(N-1)b}\left(\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{a}_i-\bar{\mathbf{a}}\|^2\right) \end{aligned} \tag{29}$$

Lemma 15 will be used to show that inside a stepsize batch, the reachability property will not be hindered by the gradient descent part in the iteration, thus allowing the Gaussian noise terms to give the desired property.

Lemma 15. *Let n be a positive integer, then for any sequence $a_k > 0$ such that there is a constant ν and $\sum_{j=1}^n a_j \le 2\nu$, let $\mathcal{F}_{\mathbf{z}}$ denote the σ-algebra generated by $\mathbf{z}_1, \cdots, \mathbf{z}_n$. Suppose $\boldsymbol{\xi}_k$ is a sequence of random vectors such that*

$$\mathbb{E}\left(\boldsymbol{\xi}_k \,|\, \mathcal{F}_{\mathbf{z}}\right) = 0 \quad \mathbb{E}\left(\|\boldsymbol{\xi}_k\|^2 \,|\, \mathcal{F}_{\mathbf{z}}\right) \le C_2$$

Let $\mathbf{y}_k = \sum_{j=1}^k a_j\boldsymbol{\xi}_j$, then

$$\mathbf{Pr}(\|\mathbf{y}_k\| \le 4\nu\sqrt{C_2}) \ge \frac{1}{2} \tag{30}$$

Proof. In the proof for this lemma, all expectations are conditioned on $\mathcal{F}_{\mathbf{z}}$. With Jensen's inequality, $(\mathbb{E}\,\|\boldsymbol{\xi}_k\|)^2 \le \mathbb{E}\,(\|\boldsymbol{\xi}_k\|^2) \le C_2$. By Markov's inequality,

$$\mathbf{Pr}(\|\mathbf{y}_k\| \ge 4\nu\sqrt{C_2}) \le \frac{\mathbb{E}\,\|\mathbf{y}_k\|}{4\nu\sqrt{C_2}} = \frac{\mathbb{E}\,\|\sum_{j=1}^k a_j\boldsymbol{\xi_j}\|}{4\nu\sqrt{C_2}} \le \frac{\mathbb{E}\sum_{j=1}^k a_j\|\boldsymbol{\xi}_j\|}{4\nu\sqrt{C_2}} \le \frac{1}{2}$$

□

Lemma 16 (Repeat of lemma 8)**.** *Given any sequence $a_k > 0$, let $\mathbf{z}_k = \sum_{i=1}^k \rho_0\sqrt{a_i}\boldsymbol{\epsilon}_i$ where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, I_d)$, for any target vector $\mathbf{z}^\star$ and distance r, there exists a non-negative function p_1 such that*

$$\mathbf{Pr}(\|\mathbf{z}_n - \mathbf{z}^\star\| \le r,\ \|\mathbf{z}_k\| \le \|\mathbf{z}^\star\| + r\ \forall k = 1, \cdots, n) \ge p_1(r, \rho_0, t_n, \mathbf{z}^\star)$$

with $p_1(0, \rho_0, t_n, \mathbf{z}^\star) = p_1(r, 0, t_n, \mathbf{z}^\star) = 0$.

Proof. We first give lower bounds to factors $\mathbf{Pr}(\|\mathbf{z}_n - \mathbf{z}^\star\| \le r)$ and $\mathbf{Pr}(\|\mathbf{z}_k\| \le \|\mathbf{z}\| + r\ \forall k = 1, \cdots, n)$ respectively, and then conclude the proof with $\mathbf{Pr}(\|\mathbf{z}_n - \mathbf{z}^\star\| \le r,\ \|\mathbf{z}_k\| \le \|\mathbf{z}\| + r\ \forall k = 1, \cdots, n) \ge \mathbf{Pr}(\|\mathbf{z}_n - \mathbf{z}^\star\| \le r) \cdot \mathbf{Pr}(\|\mathbf{z}_k\| \le \|\mathbf{z}\| + r\ \forall k = 1, \cdots, n)$.

For $\mathbf{Pr}(\|\mathbf{z}_n - \mathbf{z}^\star\|_2 \le r)$, $\|\mathbf{z}_n - \mathbf{z}^\star\|_2 \le \|\mathbf{z}_n - \mathbf{z}^\star\|_1 = \sum_{\mathrm{dim}=1}^d |(\mathbf{z}_n)_{\mathrm{dim}} - z^\star_{\mathrm{dim}}|$, therefore $\mathbf{Pr}(\|\mathbf{z}_n - \mathbf{z}^\star\|_2 \le r) \ge \mathbf{Pr}(|(\mathbf{z}_n)_{\mathrm{dim}} - z^\star_{\mathrm{dim}}| \le \frac{r}{d}\ \forall \mathrm{dim} \in [d]) = \prod_{\mathrm{dim}=1}^d \mathbf{Pr}(|(\mathbf{z}_n)_{\mathrm{dim}} - z^\star_{\mathrm{dim}}| \le \frac{r}{d})$. Notice that $(\mathbf{z}_n)_{\mathrm{dim}}$ has the distribution of the Brownian motion B_{t_n} where $t_k = \rho_0^2\sum_{i=1}^k a_i$, $k = 1, 2, \cdots, n$. By [Karlin and Taylor, 1975], $\mathbf{Pr}(|(\mathbf{z}_n)_{\mathrm{dim}} - z^\star_{\mathrm{dim}}| \le \frac{r}{d}) = \int_{\max\{z^\star_{\mathrm{dim}} - \frac{r}{d}, 0\}}^{z^\star_{\mathrm{dim}} + \frac{r}{d}} p_{t_n}(z^\star_{\mathrm{dim}}, y)\,\mathrm{d}y$, where $p_t(x,y) = \sqrt{\frac{2}{\pi t}}\exp(-\frac{x^2+y^2}{2t})\cosh(\frac{xy}{t})$. Hence,

$$\mathbf{Pr}(\|\mathbf{z}_n - \mathbf{z}^\star\|_2 \le r) \ge \left(\min_{\mathrm{dim}}\int_{\max\{z^\star_{\mathrm{dim}} - \frac{r}{d}, 0\}}^{z^\star_{\mathrm{dim}} + \frac{r}{d}} p_{t_n}(z^\star_{\mathrm{dim}}, y)\,\mathrm{d}y\right)^d \tag{31}$$

For $\mathbf{Pr}(\|\mathbf{z}_k\| \le \|\mathbf{z}^\star\| + r,\, \forall k \in [n])$, we have the following lower bound:

$$\begin{aligned}
\mathbf{Pr}(\|\mathbf{z}_k\| \le \|\mathbf{z}^\star\| + r,\, \forall k \in [n]) &\ge \mathbf{Pr}\left(\max_k |(\mathbf{z}_k)_{\text{dim}}| \le \frac{\|\mathbf{z}^\star\| + r}{\sqrt{d}},\, \forall \text{dim} \in [d]\right) \\
&= \mathbf{Pr}\left(\max_k |\underbrace{(\mathbf{z}_k)_1}_{\text{1D B.M.}}| \le \frac{\|\mathbf{z}_\star\| + r}{\sqrt{d}}\right)^d \\
&= \left(1 - \mathbf{Pr}\left(\max_k |(\mathbf{z}_k)_1| \ge \frac{\|\mathbf{z}^\star\| + r}{\sqrt{d}}\right)\right)^d
\end{aligned}$$

Now notice $\mathbf{Pr}(\max_k |(\mathbf{z}_k)_1| \ge \frac{\|\mathbf{z}^\star\|+r}{\sqrt{d}}) = \mathbf{Pr}\Big(\max_k(\mathbf{z}_k)_1 > \frac{\|\mathbf{z}^\star\|+r}{\sqrt{d}} \text{ or } \min_k(\mathbf{z}_k)_1 < -\frac{\|\mathbf{z}^\star\|+r}{\sqrt{d}}\Big) \le \mathbf{Pr}\big(\max_k(\mathbf{z}_k)_1 > \frac{\|\mathbf{z}^\star\|+r}{\sqrt{d}}\big) + \mathbf{Pr}\big(\min_k(\mathbf{z}_k)_1 < -\frac{\|\mathbf{z}^\star\|+r}{\sqrt{d}}\big) = 2\mathbf{Pr}\big(\max_k(\mathbf{z}_k)_1 > \frac{\|\mathbf{z}^\star\|+r}{\sqrt{d}}\big)$. Then, by the reflection principle of Brownian motion,

$$\mathbf{Pr}\left(\max_k (\mathbf{z}_k)_1 > \frac{\|\mathbf{z}^\star\| + r}{\sqrt{d}}\right) = 2\mathbf{Pr}\left((\mathbf{z}_n)_1 \ge \frac{\|\mathbf{z}^\star\| + r}{\sqrt{d}}\right)$$

Therefore,

$$\begin{aligned}
\mathbf{Pr}(\|\mathbf{z}_k\| \le \|\mathbf{z}^\star\| + r,\, \forall k \in [n]) &\ge \left(1 - 4\mathbf{Pr}\left(\underbrace{(\mathbf{z}_n)_1}_{\sim\mathcal{N}(0,t_n)} \ge \frac{\|\mathbf{z}^\star\| + r}{\sqrt{d}}\right)\right)^d \\
&= \left(1 - 4\mathbf{Pr}\left(\frac{(\mathbf{z}_n)_1}{\sqrt{t_n}} \ge \frac{\|\mathbf{z}^\star\| + r}{\sqrt{dt_n}}\right)\right)^d \\
&= \left(1 - 2\mathbf{Pr}\left(\frac{(\mathbf{z}_n)_1}{\sqrt{t_n}} \ge \frac{\|\mathbf{z}^\star\| + r}{\sqrt{dt_n}}\right) - 2\mathbf{Pr}\left(\frac{(\mathbf{z}_n)_1}{\sqrt{t_n}} \le -\frac{\|\mathbf{z}^\star\| + r}{\sqrt{dt_n}}\right)\right)^d \\
&= \left(1 - 2\big(1 - \mathbf{Pr}\left(-\frac{\|\mathbf{z}^\star\| + r}{\sqrt{dt_n}} \le \frac{(\mathbf{z}_n)_1}{\sqrt{t_n}} \le \frac{\|\mathbf{z}^\star\| + r}{\sqrt{dt_n}}\right)\big)\right)^d \\
&= \left(2\int_{-\frac{\|\mathbf{z}^\star\|+r}{\sqrt{dt_n}}}^{\frac{\|\mathbf{z}^\star\|+r}{\sqrt{dt_n}}} \frac{\exp(-x^2/2)}{\sqrt{2\pi}}\,\mathrm{d}x - 1\right)^d \\
&\ge \left(4\frac{\|\mathbf{z}^\star\| + r}{\sqrt{2\pi dt_n}} \exp\left(-\frac{1}{2}\frac{(\|\mathbf{z}^\star\| + r)^2}{dt_n}\right) - 1\right)^d
\end{aligned} \tag{32}$$

Let $p_1(r, \rho_0, t_n, \mathbf{z}^\star)$ be the product of two lower bounds (31) and (32) above, recall that $t_n = \rho_0^2 \sum_{i=1}^n a_i$, we define $p_1(r, \rho_0, t_n, \mathbf{z}^\star)$ as the following,

$$p_1(r, \rho_0, t_n, \mathbf{z}^\star) := \left(\min_{\text{dim}} \int_{\max\{z^\star_{\text{dim}} - \frac{r}{d}, 0\}}^{z^\star_{\text{dim}} + \frac{r}{d}} p_{t_n}(z^\star_{\text{dim}}, y)\,\mathrm{d}y\right)^d \cdot \left(4\frac{\|\mathbf{z}^\star\| + r}{\sqrt{2\pi dt_n}} \exp\left(-\frac{1}{2}\frac{(\|\mathbf{z}^\star\| + r)^2}{dt_n}\right) - 1\right)^d \tag{33}$$

where $p_{t_n}(x, y) = \sqrt{\frac{2}{\pi t}} \exp(-\frac{x^2+y^2}{2t_n}) \cosh(\frac{xy}{t_n})$.

To make the dependence of the first factor in p_1 on parameters more explicit, for some $\xi \in (\max\{z^\star_{\text{dim}} - \frac{r}{d}, 0\}, z^\star_{\text{dim}} + \frac{r}{d})$,

$$\begin{aligned}
\int_{\max\{z^\star_{\text{dim}} - \frac{r}{d}, 0\}}^{z^\star_{\text{dim}} + \frac{r}{d}} p_{t_n}(z^\star_{\text{dim}}, y)\,\mathrm{d}y &\ge \sqrt{\frac{2}{\pi t_n}} \exp(-\frac{(z^\star_{\text{dim}})^2 + \xi^2}{2t_n}) \cosh(\frac{z^\star_{\text{dim}}\xi}{t_n}) \frac{r}{d} \\
&= \sqrt{\frac{2}{\pi t_n}} \left(\exp(-\frac{(z^\star_{\text{dim}} - \xi)^2}{2t_n}) + \exp(-\frac{(z^\star_{\text{dim}} + \xi)^2}{2t_n})\right) \frac{r}{2d} \\
&\ge \sqrt{\frac{2}{\pi t_n}} 2\sqrt{\exp(-\frac{(z^\star_{\text{dim}} - \xi)^2}{2t_n} - \frac{(z^\star_{\text{dim}} + \xi)^2}{2t_n})} \frac{r}{2d} \\
&= \sqrt{\frac{2}{\pi t_n}} \exp(-\frac{(z^\star_{\text{dim}})^2 + \xi^2}{2t_n}) \frac{r}{d} \\
&\ge \sqrt{\frac{2}{\pi t_n}} \exp(-\frac{(z^\star_{\text{dim}})^2 + (z^\star_{\text{dim}} + \frac{r}{d})^2}{2t_n}) \frac{r}{d}
\end{aligned}$$

We thus redefine p_1 as

$$p_1(r,\rho_0,t_n,\mathbf{z}^\star) := \left(\min_{\dim}\sqrt{\frac{2}{\pi t_n}}\exp(-\frac{(z^\star_{\dim})^2+(z^\star_{\dim}+\frac{r}{d})^2}{2t_n})\frac{r}{d}\right)^d \cdot \left(4\frac{\|\mathbf{z}^\star\|+r}{\sqrt{2\pi d t_n}}\exp\left(-\frac{1}{2}\frac{(\|\mathbf{z}^\star\|+r)^2}{dt_n}\right)-1\right)^d \tag{34}$$

then we have the lemma 8. □

Lemma 17 (Repeat of lemma 9)**.** *Assume the same stepsize batch setting as in lemma 7 and the gradient Lipschitz condition in assumption 1, with respect to an arbitrary target point* $\mathbf{s}$ *in the level set* $\{\mathbf{x}: f(\mathbf{x}) < \mathcal{O}(\widetilde{\varepsilon})\}$*, there is a constant* $C_{21} \propto dL$ *and a constant* C_{22}*, for any* $\varepsilon > 0$*, there exists a non-negative* $p_1(\widetilde{\varepsilon},\rho_0,t_{n_{i+1}},\mathbf{s})$ *such that*

$$\mathbf{Pr}\left(\|\mathbf{x}_{n_{i+1}}-\mathbf{s}\| \le \widetilde{\varepsilon}\right) > \frac{1}{2}p_1(\widetilde{\varepsilon},\rho_0,t_{n_{i+1}},\mathbf{s}) \tag{35}$$

where $\widetilde{\varepsilon} = \varepsilon + \delta C_{21} + 2\delta\sqrt{C_{22}}$*, and* p_1 *is given in lemma 8.*

Proof. Denote $\mathbf{x}_o = \mathbf{x}_{n_i}$ and $\mathbf{d} = \mathbf{s} - \mathbf{x}_o$. Recall the SGLD scheme in a batch goes as

$$\begin{aligned}
\mathbf{x}_{k+1} &= \mathbf{x}_k - \eta_k\widetilde{\nabla}_k + \rho_k\boldsymbol{\epsilon}_k = \mathbf{x}_o - \sum_{l=0}^{k}\eta_l\widetilde{\nabla}_l + \rho_l\boldsymbol{\epsilon}_l \\
&= \mathbf{x}_o - \sum_{l=0}^{k}\eta_{l+o}\left(\nabla f(\mathbf{x}_{l+o}) + \big(\frac{1}{B}\nabla f_{I_{l+o}}(\mathbf{x}_{l+o}) - \nabla f(\mathbf{x}_l)\big) - \big(\frac{1}{B}\nabla f_{I_{l+o}}(\widetilde{\mathbf{x}}) - \nabla f(\widetilde{\mathbf{x}})\big)\right) + \sum_{l=0}^{k}\rho_{l+o}\boldsymbol{\epsilon}_{l+o} \\
&= \mathbf{x}_o - \sum_{l=0}^{k}\eta_{l+o}\nabla f(\mathbf{x}_{l+o}) - \mathbf{y}_k + \mathbf{z}_k
\end{aligned} \tag{36}$$

where we define $\mathbf{y}_k := \sum_{l=0}^{k}\eta_l\big(\frac{1}{B}\nabla f_{I_l}(\mathbf{x}_l) - \nabla f(\mathbf{x}_l)\big) - \eta_l\big(\frac{1}{B}\nabla f_{I_l}(\widetilde{\mathbf{x}}) - \nabla f(\widetilde{\mathbf{x}})\big)$ and $\mathbf{z}_k := \sum_{l=0}^{k}\rho_0\sqrt{\frac{\eta_l}{\eta_0}}\boldsymbol{\epsilon}_l$. Note that $\widetilde{\mathbf{x}}$ can change as moving from stepsize batch i to $i+1$ may involve different SVRG batch reference points.

Let $m = n_{i+1} - n_i$. Per law of total probability, Denote the event $\mathcal{E}_{i+1} = \{\|\mathbf{z}_{n_{i+1}-1} - \mathbf{d}\| \le \varepsilon,\ \|\mathbf{z}_k\| \le \|\mathbf{d}\| + \varepsilon\ \forall k \in [m] + n_i\}$, then

$$\mathbf{Pr}\left(\|\mathbf{x}_{n_{i+1}}-\mathbf{s}\| \le \widetilde{\varepsilon}\right) \ge \mathbf{Pr}\left(\|\mathbf{x}_{n_{i+1}}-\mathbf{s}\| \le \widetilde{\varepsilon}\,\middle|\,\mathcal{E}_{i+1}\right)\cdot\mathbf{Pr}\left(\mathcal{E}_{i+1}\right) \tag{37}$$

where $\mathbf{Pr}\left(\mathcal{E}_{i+1}\right)$ is lowered bounded by p_1 from lemma 8. What is left in this proof is to bound the first factor in the above equation.

Now we bound the gradient difference terms in (36), thus computing the probability $\mathbf{Pr}\left(\|\mathbf{x}_{n_{i+1}}-\mathbf{s}\| \le \widetilde{\varepsilon}\,\middle|\,\mathcal{E}_{i+1}\right)$.

As the gradient is bounded inside a stepsize batch $\{\eta_{n_i} : \eta_{n_{i+1}-1}\}$, by lemma 14, each summation term in $\mathbf{y}_k$ has the following variance upper bound

$$\frac{n-B}{(n-1)B}\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}_l) - \nabla f(\mathbf{x}_l)\|^2 \le \frac{n-B}{(n-1)B}\frac{1}{n}\max_{l\in[m]+o}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}_l) - \nabla f(\mathbf{x}_l)\|^2 := C_{22}$$

Within the gradient batch, the term $\|\sum_{l=0}^{n_{i+1}-1}\eta_{l+o}\nabla f(\mathbf{x}_{l+o})\| \le \max_{l\in[n_{i+1}-1]}\|\nabla f(\mathbf{x}_{l+o})\|\sum_{l=0}^{n_{i+1}-1}\eta_{l+o} \le \delta\max_{l\in[n_{i+1}-1]}\|\nabla f(\mathbf{x}_{l+o})\|$. After the iteration has proceeded sufficiently, per convergence properties first-order stationary points in lemma 5, the boundedness of gradients in the gradient batch can be controlled by a constant $C_{21} = \mathcal{O}(dL)$. The unbiasedness of SVRG gradient estimator makes lemma 15 applicable to $\mathbf{y}_k$:

$$\begin{aligned}
\mathbf{Pr}(\|\mathbf{x}_{n_{i+1}}-\mathbf{s}\| \le \widetilde{\varepsilon}\,|\,\mathcal{E}_{i+1}) &= \mathbf{Pr}(\|\mathbf{x}_o - \sum_{l=0}^{n_{i+1}-1}\eta_{l+o}\nabla f(\mathbf{x}_{l+o}) - \mathbf{y}_{n_{i+1}-1} + \mathbf{z}_{n_{i+1}-1} - \mathbf{s}\| \le \widetilde{\varepsilon}\,|\,\mathcal{E}_{i+1}) \\
&= \mathbf{Pr}(\| - \sum_{l=0}^{n_{i+1}-1}\eta_{l+o}\nabla f(\mathbf{x}_{l+o}) - \mathbf{y}_{n_{i+1}-1} + \mathbf{z}_{n_{i+1}-1} - \mathbf{d}\| \le \widetilde{\varepsilon}\,|\,\mathcal{E}_{i+1}) \\
&\ge \mathbf{Pr}(\| \sum_{l=0}^{n_{i+1}-1}\eta_{l+o}\nabla f(\mathbf{x}_{l+o})\| \le \delta C_{21}, \|\mathbf{y}_{n_{i+1}-1}\| \le 4\delta C_{22},\,|\,\mathcal{E}_{i+1}) \\
&\ge \mathbf{Pr}(\|\mathbf{y}_{n_{i+1}-1}\| \le 4\delta C_{22},\,|\,\mathcal{E}_{i+1}) \ge \frac{1}{2}
\end{aligned}$$

where the last inequality is due to the fact that the first part of the event is a certain event with the iteration having proceeded sufficiently and the choice of δ to bound the sum of stepsize update in a batch.

□

Now we are ready to prove the ergodicity result for the SGLD-VR scheme with the recurrence and reachability results above.

Theorem 18 (Repeat of theorem 2)**.** *Under regularization condition 3, Lipschitz assumption 1 and nonnegativity assumption 2, with the same parameter setting as in lemma 7, for any $\widetilde{\varepsilon} > 0$, $p > 0$ and a point $\mathbf{s}$ which locates in the level set $\{\mathbf{x} : f(\mathbf{x}) \leq \mathcal{O}(\widetilde{\varepsilon})\}$, there is a*

$$T = \mathcal{O}\left(\frac{1}{p\mu_1\left(4\eta_0 L^3 \frac{\mu_2 f(\mathbf{x}_0)+2\psi_2}{B_{\mathrm{e}}} + \frac{\rho_0^2}{\eta_0} Ld\right)}\left(1 + \ln f(\mathbf{x}_0) + \frac{(d\|\mathbf{s}\| + \widetilde{\varepsilon})^d}{\left((\frac{4}{\sqrt{2\pi}} - 1)\mathrm{e}^{-1/2}\widetilde{\varepsilon}\right)^d}\right)\right) \tag{38}$$

such that

$$\mathbf{Pr}(\|\mathbf{x}_t - \mathbf{s}\| \leq \widetilde{\varepsilon} \textit{ for some } t < T) \geq 1 - p \tag{39}$$

Proof of Thm. 2. Recall the definition of the stopping time sequence: $\tau_0 = K$, $\tau_{t+1} = \min\{t : t \geq \tau_k + 1, f(\mathbf{x}_{n_t}) \leq M = 2\delta B\}$. By defining $\tau_* = \min\{t : t > 0, \|\mathbf{x}_{n_t} - \mathbf{s}\| \leq \widetilde{\varepsilon}\}$ and setting $\delta = \frac{\widetilde{\varepsilon}}{2B}$, we show that $\mathbf{Pr}(\tau_* \geq T) \leq \widetilde{p}$ with a proper choice of T. For any J,

$$\begin{aligned}
\mathbf{Pr}(\tau_* \geq T) &= \mathbf{Pr}(\tau_* \geq T, \tau_J > T) + \mathbf{Pr}(\tau_* \geq T, \tau_J < T) \\
&\leq \mathbf{Pr}(\tau_J > T) + \mathbf{Pr}(\|\mathbf{x}_{n_{\tau_k}+1} - \mathbf{s}\| > \widetilde{\varepsilon}, \tau_J \leq T,\ \forall k \in [J]) \\
&\leq \mathbf{Pr}(\tau_J > T) + \mathbf{Pr}(\|\mathbf{x}_{n_{\tau_k}+1} - \mathbf{s}\| > \widetilde{\varepsilon},\ \forall k \in [J])
\end{aligned}$$

Lemma 7 gives that $\mathbb{E}\,\tau_J \leq \frac{4}{\alpha} + K + J(\frac{1}{2\alpha\delta} + 1)$, thus by Markov inequality

$$\mathbf{Pr}(\tau_J > T) \leq \frac{\mathbb{E}\,[\tau_J]}{T} \leq \frac{\frac{4}{\alpha} + K + J(\frac{1}{2\alpha\delta} + 1)}{T} \tag{40}$$

To ensure the last bound is below the pre-specified threshold $\frac{1}{2}p$, we need to take

$$T = \left\lceil \frac{\frac{4}{\alpha} + K + J(\frac{1}{2\alpha\delta} + 1)}{\frac{p}{2}} \right\rceil + 1 \tag{41}$$

By lemma 9, there is a $p_2 = p_1/2 > 0$ such that

$$\mathbf{Pr}(\|\mathbf{x}_{n_{\tau_k}+1} - \mathbf{s}\| > \widetilde{\varepsilon},\ \forall k \in [J]) = \prod_{k=1}^{J} \mathbf{Pr}(\|\mathbf{x}_{n_{\tau_k}+1} - \mathbf{s}\| > \widetilde{\varepsilon}) \leq (1 - p_2)^J \overset{\text{let}}{\leq} \frac{p}{2} \tag{42}$$

To ensure the upper bound to be less than $\frac{p}{2}$, a sufficient condition is that $J > \dfrac{\ln \frac{p}{2}}{\ln(1 - p_2(\widetilde{\varepsilon}, \rho_0, t_{n_{\tau_J}+1}))} > \dfrac{\ln \frac{2}{p}}{p_2(\widetilde{\varepsilon}, \rho_0, t_{n_{\tau_J}+1})}$.

We consider the dependence of T on the error tolerance $\widetilde{\varepsilon}$, dimension d and initial perturbation parameter ρ_0. Recall $p_2 = \frac{1}{2}p_1$ and we will upper bound it to rid of the dependence on t_n:

$$\begin{aligned}
&p_2(\widetilde{\varepsilon}, \rho_0, t_n, \mathbf{s}) \\
&= \frac{1}{2}\left(\min_{\mathrm{dim}} \sqrt{\frac{2}{\pi t_n}} \exp(-\frac{(s_{\mathrm{dim}})^2 + (s_{\mathrm{dim}} + \frac{\widetilde{\varepsilon}}{d})^2}{2t_n})\frac{\widetilde{\varepsilon}}{d}\right)^d \cdot \left(4\frac{\|\mathbf{s}\| + \widetilde{\varepsilon}}{\sqrt{2\pi d t_n}} \exp\left(-\frac{1}{2}\frac{(\|\mathbf{s}\| + \widetilde{\varepsilon})^2}{d t_n}\right) - 1\right)^d \\
&\leq \frac{1}{2}\left(\min_{\mathrm{dim}} \sqrt{\frac{2}{\pi\left((s_{\mathrm{dim}})^2 + (s_{\mathrm{dim}} + \frac{\widetilde{\varepsilon}}{d})^2\right)}} \exp(-\frac{1}{2})\frac{\widetilde{\varepsilon}}{d}\right)^d \cdot \left(4\frac{1}{\sqrt{2\pi}} - 1\right)^d
\end{aligned}$$

Therefore, a sufficient condition for (42) to hold is

$$J > \left(\ln \frac{2}{p}\right) \max_{\mathrm{dim}} \frac{2}{\left(\sqrt{\frac{2}{\pi\left((s_{\mathrm{dim}})^2 + (s_{\mathrm{dim}} + \frac{\widetilde{\varepsilon}}{d})^2\right)}} \exp(-\frac{1}{2})\frac{\widetilde{\varepsilon}}{d}\right)^d \cdot \left(4\frac{1}{\sqrt{2\pi}} - 1\right)^d} \tag{43}$$

Recall from lemma 7 parameter settings $K = \frac{\ln \frac{f(\mathbf{x}_{n_0})}{\delta B}}{(1-C_1)\mu_1\delta}$, $B = 2(\psi_1 + \frac{2\eta_{n_0}L^3}{B_{\mathrm{e}}}\big(\mu_2 f(\mathbf{x}_0) + 2\psi_2\big) + \frac{\rho_0^2 Ld}{2\eta_0})$ and $\alpha = 1 - 2\exp(-(1-C_1)\mu_1\delta)$. In light of the remark post the lemma 7, δ is to be set as $\delta \propto \widetilde{\varepsilon}B^{-1}$ for the purpose of minimizing the empirical risk. Combining (41) and (43), the total amount of time needed for (9) to hold is

$$\begin{aligned} T &= \left\lceil \frac{\frac{4}{\alpha} + K + J(\frac{1}{2\alpha\delta}+1)}{\frac{p}{2}} \right\rceil + 1 \\ &= \widetilde{\mathcal{O}}\left(\frac{1 + \ln f(\mathbf{x}_0) + \frac{(d\|\mathbf{s}\|+\widetilde{\varepsilon})^d}{\widetilde{\varepsilon}\big((\frac{4}{\sqrt{2\pi}}-1)\mathrm{e}^{-1/2}\widetilde{\varepsilon}\big)^d}}{\widetilde{\varepsilon}p\mu_1\big(\psi_1 + 2\eta_0 L^3 \frac{\mu_2 f(\mathbf{x}_0)+2\psi_2}{B_{\mathrm{e}}} + \frac{\rho_0^2}{\eta_0}Ld\big)} \right) \end{aligned} \tag{44}$$

□

C. Proofs of second-order stationary point convergence property

Proof of Thm. 3.

a) Step 1: Assume the stepsize decay parameter $\nu \in [1, 2]$ for simplicity. We show that $\Delta_i := \sum_{l=n_i}^{n_{i+1}-1} \sqrt{\eta_l}\boldsymbol{\epsilon}_i$ will lead to saddle point escape, i.e. $\Delta_i^\intercal \nabla^2 f(\mathbf{x}_{\mathrm{fsp}})\Delta_i \le -\zeta$. Specifically, we show that the projection of Δ_i on the direction of $\lambda_{\min}$ is than ζ with high probability, which exploits the property of Brownian motion and the idea that the trapping region is thin when faced with LD [Huang and Becker, 2020].

At a fixed first-order stationary point $\mathbf{x}_{\mathrm{fsp}}$, due to the spatial homogeneity of Brownian motion, w.l.o.g. assume that $\mathbf{e}_1$ is the unit eigenvector corresponding to the smallest eigenvalue of $\nabla^2 f(\mathbf{x}_{\mathrm{fsp}})$. To have $\Delta_i^\intercal \nabla^2 f(\mathbf{x}_{\mathrm{fsp}})\Delta_i \le -\zeta$, a sufficient condition is $\lambda_{\min}(\nabla^2 f(\mathbf{x}_{\mathrm{fsp}}))(\Delta_i)_1^2 + L(\|\Delta_i\|^2 - (\Delta_i)_1^2) \le -\zeta$. Assume for now that $\|\Delta_i\|^2 \le r^2$, then this condition can be phrased as $\lambda_{\min}(\nabla^2 f(\mathbf{x}_{\mathrm{fsp}}))(\Delta_i)_1^2 + L(r^2 - (\Delta_i)_1^2) \le -q(\Delta_i)_1^2 + L(r^2 - (\Delta_i)_1^2) \le -\zeta$, i.e. the first component of Δ_i satisfies

$$(\Delta_i)_1^2 \ge \frac{\zeta + Lr^2}{L+q} := Q \tag{45}$$

Now we compute the probability for (45) to fail within the time $T_i := \sum_{l=n_i}^{n_{i+1}-1} \sqrt{\eta_l}$ for a standard 1D Brownian motion. Define $\tau_Q = \min\{t \,|\, \big((\Delta_i)_1(t)\big)^2 \ge Q\}$. Then

$$\mathbf{Pr}(\text{(45) fails to hold within time } T_i) = \mathbf{Pr}(\tau_Q > T_i) \le \frac{\mathbb{E}\,\tau_Q}{T_i} = \frac{Q}{dT_i}.$$

Here we point out that the failure probability for (45) is low due to the large denominator. $T_i = \sum_{l=n_i}^{n_{i+1}-1} \sqrt{\eta_l} \le \sqrt{(n_{i+1}-n_i)\sum_{l=n_i}^{n_{i+1}-1}\eta_l} \approx \sqrt{(n_{i+1}-n_i)\delta}$. Note that $n_{i+1} - n_i = \mathcal{O}\left(n_i \exp(\delta)\right)$, then $n_i = \mathcal{O}(\exp(i\delta))$, thus

$$T_i = \exp(\mathcal{O}(i\delta)) \tag{46}$$

Remark: consider the case $\nu = 1$ as an example for the preceding claim. As $\sum_{l=n_i}^{n_{i+1}-1}\eta_l \approx \delta$ and $n_0 = 1$, $n_i \approx \exp(i\delta)$ and $n_{i+1} = n_i\exp(\delta)$. The corresponding time in continuous domain

$$\begin{aligned} T_i &= \sum_{l=n_i}^{n_{i+1}-1} \sqrt{\eta_l} \approx \int_{n_i}^{n_{i+1}-1} \sqrt{\eta_0}\frac{1}{\sqrt{t}}\,\mathrm{d}t = 2\sqrt{\eta_0}(\sqrt{n_{i+1}-1} - \sqrt{n_i}) \\ &\approx 2\sqrt{\eta_0}(\sqrt{n_{i+1}} - \sqrt{n_i}) = 2\sqrt{\eta_0 n_i}(\sqrt{\exp(\delta)} - 1) = 2\sqrt{\eta_0\exp(i\delta)}(\sqrt{\exp(\delta)} - 1). \end{aligned}$$

b) Step 2: We show that when $\mathbf{x} \in \mathcal{U}(\mathbf{x}_{\mathrm{fsp}}, r)$ where $\|\Delta_j\| < r$ for $j = n_i, n_i+1, \cdots, n_{i+1}-1$, $\|\nabla f(\mathbf{x})\| < \varepsilon$, thus the first order expansion does not contribute to function value change. Due to the Lipschitz gradient, set

$$r = \max\left\{\frac{\varepsilon}{L}, \sqrt{\frac{3}{Lq}}\varepsilon\right\}.$$

While the projection onto $\mathbf{e}_1$ builds up, we compute the probability that the iteration is still constrained within the ε-neighborhood of $\mathbf{x}_{\text{fsp}}$:

$$\begin{aligned}\mathbf{Pr}(\Delta_i^2 - (\Delta_i)_1^2 \leq r^2 - Q \text{ when } t \leq T_i) &= \mathbf{Pr}(\sum_{i=2}^{d} \hat{x}_i^2 \leq r^2 - Q \text{ when } t \leq T_i) \\ &= \int_0^{\sqrt{r^2-Q}} (T_i)^{-\frac{d-1}{2}} \frac{1}{\Gamma(\frac{d-2}{2})} \exp(-\frac{y^2}{2T_i}) y^{d-2}\,\mathrm{d}y \\ &\approx (T_i)^{-\frac{d-1}{2}} \frac{1}{\Gamma(\frac{d-2}{2})} \int_0^{\sqrt{r^2-Q}} y^{d-2}\,\mathrm{d}y \\ &= (T_i)^{-\frac{d-1}{2}} \frac{1}{\Gamma(\frac{d-2}{2})} (r^2-Q)^{\frac{d-1}{2}} := P_i \end{aligned} \tag{47}$$

As T_i increases exponentially w.r.t. index i, P_i decreases accordingly, i.e. , within a stepsize batch, the probability for the iteration to remain bounded within the vicinity of a FSP is decreasing. Hence, the saddle point escape process can be thought of as a binomial trial with decreasing success probability, and the expected time for the iteration process to escape all saddle points is at least proportional to $\Gamma(\frac{d-2}{2})$.

c) Step 3: Show that the update $\mathbf{x}' = \mathbf{x} + \Delta_i$ will lead to function value decrease, thus the SGLD algorithm has to terminate, thus converging to SSP.

Denote the event $\mathcal{A}_i = \{\Delta_i^\intercal \nabla^2 f(\mathbf{x}_{\text{fsp}_i})\Delta_i \leq -\zeta \text{ and } \|\Delta_i\| \leq r\}$. From steps 1 and 2, $\mathbf{Pr}(\mathcal{A}_i) \geq (1 - \frac{Q}{dT_i})P_i$. We show that under the assumption that event $\mathcal{A}_i$ happens, function value decrease is guaranteed. Note that within a minibatch,

$$\begin{aligned}\mathbb{E}\,\|\mathbf{x}_t - \widetilde{\mathbf{x}}\|^2 = \mathbb{E}\left\|\sum_{u=o}^{B_\mathrm{e}-1} \mathbf{x}_{u+1} - \mathbf{x}_u\right\|^2 &= \mathbb{E}\left\|\sum_{u=o}^{B_\mathrm{e}-1} \eta_u\big(\nabla f_{i_u}(\mathbf{x}_{u+1}) - \nabla f_{i_u}(\widetilde{\mathbf{x}}) + \nabla f(\widetilde{\mathbf{x}})\big) - \rho_u \boldsymbol{\epsilon}_u\right\|^2 \\ &\leq 2\mathbb{E}\,\|\sum_{u=o}^{B_\mathrm{e}-1} \eta_u\big(\nabla f_{i_u}(\mathbf{x}_{u+1}) - \nabla f_{i_u}(\widetilde{\mathbf{x}}) + \nabla f(\widetilde{\mathbf{x}})\big)\|^2 + 2\mathbb{E}\,\|\sum_{u=o}^{B_\mathrm{e}-1} \rho_u \boldsymbol{\epsilon}_u\|^2 \\ &= 2\mathbb{E}\,\|\sum_{u=o}^{B_\mathrm{e}-1} \eta_u\big(\nabla f_{i_u}(\mathbf{x}_{u+1}) - \nabla f_{i_u}(\widetilde{\mathbf{x}}) + \nabla f(\widetilde{\mathbf{x}})\big)\|^2 + 2\mathbb{E}\,\|\sum_{u=o}^{B_\mathrm{e}-1} \rho_u \boldsymbol{\epsilon}_u\|^2 \\ &\leq 2\mathbb{E}\sum_{u=o}^{B_\mathrm{e}-1} \eta_u^2 \left(\|\nabla f_{i_u}(\mathbf{x}_{u+1})\|^2 + \|\nabla f(\widetilde{\mathbf{x}}) - \nabla f_{i_u}(\widetilde{\mathbf{x}})\|^2\right) + 2d\sum_{u=o}^{B_\mathrm{e}-1} \rho_u^2 \\ &\overset{(29)}{\leq} 2D_F \sum_{u=o}^{B_\mathrm{e}-1} \eta_u^2 (1 + \frac{2}{B_\mathrm{e}}) + 2d\sum_{u=o}^{B_\mathrm{e}-1} \rho_u^2 \end{aligned} \tag{48}$$

For function value decrease in the descent process, we have

$$\begin{aligned} f(\mathbf{x}_t) - \mathbb{E}\, f(\mathbf{x}_{t+1}) &\geq \mathbb{E}\left[\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1}\rangle - \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2\right] \\ &= \mathbb{E}\left[\langle \nabla f(\mathbf{x}_t), \eta_k \widetilde{\nabla}_k\rangle - \frac{L}{2}\|\eta_k \widetilde{\nabla}_k - \rho_k \boldsymbol{\epsilon}_k\|^2\right] \\ &= \mathbb{E}\left[\eta_t \|\nabla f(\mathbf{x}_t)\|^2 - \frac{L}{2}(\eta_t^2 \|\widetilde{\nabla}_t\|^2 + \rho_t^2 \|\boldsymbol{\epsilon}_t\|^2)\right] \\ &\geq \mathbb{E}\left[\eta_t \|\nabla f(\mathbf{x}_t)\|^2 - \frac{L}{2}\left(\eta_t^2\big(2[\|\nabla f(\mathbf{x}_t)\|^2] + 2\frac{L^2}{B_\mathrm{e}}[\|\mathbf{x}_t - \widetilde{\mathbf{x}}\|^2]\big) + \rho_t^2 \|\boldsymbol{\epsilon}_t\|^2\right)\right] \\ &\overset{(48)}{\geq} (\eta_t - \eta_t^2 L)\|\nabla f(\mathbf{x}_t)\|^2 - \underbrace{\frac{L^3}{B_\mathrm{e}}\eta_t^2 (3D_F \sum_{u=o}^{B_\mathrm{e}-1} \eta_u^2 + 2d\sum_{u=o}^{B_\mathrm{e}-1} \rho_u^2) - \frac{L}{2}\rho_t^2 d}_{\mathcal{R}} = \mathcal{O}\left(\varepsilon^2\right) \end{aligned}$$

Here notice that $\sum_{u=o}^{B_\mathrm{e}-1} \eta_u^2 = \mathcal{O}(\eta_0 \nu^{-1})$ and $\sum_{u=o}^{B_\mathrm{e}-1} \rho_u^2 = \mathcal{O}(\rho_0 \nu^{-1})$, and set $B_\mathrm{e} = \max\{\frac{L^3 D_f d}{\varepsilon^2}, 1\}$ and $\delta = \mathcal{O}(r)$ (which consequently gives the order of η_t^2), then $\mathcal{R} = \mathcal{O}(\varepsilon^2)$.

When a saddle point is encountered, within a minibatch with probability $(1-\frac{Q}{dT_i})P_i$, we have

$$\begin{aligned}
&f(\mathbf{x}_o)-f(\mathbf{x}_o+\Delta_i)=f(\mathbf{x}_o)-f(\mathbf{x}_{\rm fsp})+f(\mathbf{x}_{\rm fsp})-f(\mathbf{x}_{\rm fsp}+\Delta_i)+f(\mathbf{x}_{\rm fsp}+\Delta_i)-f(\mathbf{x}_0+\Delta_i)\\
&=f(\mathbf{x}_{\rm fsp})-\left(f(\mathbf{x}_{\rm fsp})+\frac{1}{2}\Delta_i^\intercal\nabla^2 f(\mathbf{x}_{\rm fsp})\Delta_i+\frac{L_2}{6}\|\Delta_i\|^3\right)+f(\mathbf{x}_o)-f(\mathbf{x}_{\rm fsp})+f(\mathbf{x}_{\rm fsp}+\Delta_i)-f(\mathbf{x}_0+\Delta_i)\\
&\geq f(\mathbf{x}_o)-f(\mathbf{x}_{\rm fsp})+f(\mathbf{x}_{\rm fsp}+\Delta_i)-f(\mathbf{x}_0+\Delta_i)-\frac{1}{2}\Delta_i^\intercal\nabla^2 f(\mathbf{x}_{\rm fsp})\Delta_i+\frac{L_2}{6}\|\Delta_i\|^3\\
&\geq f(\mathbf{x}_o)-f(\mathbf{x}_{\rm fsp})+f(\mathbf{x}_{\rm fsp}+\Delta_i)-f(\mathbf{x}_0+\Delta_i)+\frac{\zeta}{2}-\frac{L_2}{6}r^3\\
&\geq\frac{\zeta}{2}-\frac{L_2}{6}r^3-2Lr^2=\mathcal{O}\left(\varepsilon^2\right)
\end{aligned}$$

Set $\zeta=\frac{5\varepsilon^2}{2L}$, the time complexity to attain sufficient function value decrease before reaching a SSP is $\mathcal{O}\left(\frac{f(\mathbf{x}_0)-f_\star}{\varepsilon^2}\right)$.

d) Step 4: Now we give the description of $\tau_{\rm SSP}$ to finish the proof. In light of setting $\zeta=\frac{5}{2L}\varepsilon^2$, consequently $r^2-Q=\frac{\varepsilon^2}{2Lq(L+q)}$. From (47) together with (46), the probability for constrained perturbation accumulation within the stepsize batch i is given as $P_i=\mathcal{O}\left(\frac{\varepsilon^{d-1}}{\Gamma(\frac{d-2}{2})L^{d-1}q^{d-1}}\right)\cdot\frac{1}{\exp(\mathcal{O}(i\delta d))}$.

Assume the iteration sequence escapes saddle points in each stepsize batch where a saddle point is encountered. Setting the stepsize sum threshold $\delta=\mathcal{O}(\widetilde{\varepsilon}/B)$ to be consistent with the setting of δ in the proof for theorem 2 and lemma 9, with probability $\mathcal{O}\left(\frac{\varepsilon^{d-1}}{\Gamma(\frac{d-2}{2})L^{d-1}q^{d-1}}\right)$, the SGLD converges to a local minimum within time

$$\tau_{\rm SSP}=\underbrace{\mathcal{O}\left(\frac{f(\mathbf{x}_0)-f_\star}{\varepsilon^2}\right)}_{\text{descent}}+\underbrace{\exp\left(\mathcal{O}(\varepsilon d)\right)}_{\text{escaping saddles}},\tag{49}$$

where the first term accounts for the step needed for sufficient function value decrease, and the second term accounts for the time needed to escape saddle points as computed in equation (46).

□