Subsampling Error in Stochastic Gradient Langevin Diffusions

Kexin Jin[*] Princeton University

Chenguang Liu[*] Delft University of Technology

Jonas Latz[†] University of Manchester

Abstract

The Stochastic Gradient Langevin Dynamics (SGLD) are popularly used to approximate Bayesian posterior distributions in statistical learning procedures with large-scale data. As opposed to many usual Markov chain Monte Carlo (MCMC) algorithms, SGLD is not stationary with respect to the posterior distribution; two sources of error appear: The first error is introduced by an Euler–Maruyama discretisation of a Langevin diffusion process, the second error comes from the data subsampling that enables its use in large-scale data settings. In this work, we consider an idealised version of SGLD to analyse the method's pure subsampling error that we then see as a best-case error for diffusion-based subsampling MCMC methods. Indeed, we introduce and study the Stochastic Gradient Langevin Diffusion (SGLDiff), a continuous-time Markov process that follows the Langevin diffusion corresponding to a data subset and switches this data subset after exponential waiting times. There, we show the exponential ergodicity of SLGDiff and that the Wasserstein distance between the posterior and the limiting distribution of SGLDiff is bounded above by a fractional power of the mean waiting time. We bring our results into context with other analyses of SGLD.

1 INTRODUCTION AND MAIN RESULTS

Bayesian machine learning allows the applicant not only to train a model, but also to accurately describe the uncertainty that remains in the model after incorporating the training data. Bayesian approaches are naturally used in conjugate settings, e.g., Gaussian process regression or naive Bayes (Bishop, 2006) or when appropriate approximations are available, e.g., Variational Bayes (Fox and Roberts, 2012). In other situations, none of this is possible and the Bayesian posterior distribution of the trained model needs to be approximated with a Monte Carlo scheme, such as Markov chain Monte Carlo (MCMC) (Neal, 1996). Due to the large amount of available training data and the large computational cost of model/derivative evaluations in, e.g., Bayesian deep learning problems, accurate MCMC techniques (e.g. MALA, Roberts and Tweedie, 1996) are usually inapplicable. Instead, approximate MCMC techniques, such as the Stochastic Gradient Langevin Dynamics (SGLD) Welling and Teh (2011) and its variants are popularly employed. Those methods combine the unadjusted Langevin algorithm (ULA) with *(data) subsampling* as it would be usual in stochastic-gradient-descent-type optimisation algorithms.

In this work, we analyse the error that arises from data subsampling in Langevin-based MCMC algorithms in an idealised dynamical system that we refer to as *Stochastic Gradient Langevin Diffusion* (SGLDiff). Hence, we do not propose a new algorithm, but rather deduct in our continuous-time analysis how well Langevin-based MCMC methods can perform under data subsampling when perfectly sampling the underlying dynamics. Hence, we focus on the intrinsic error that data subsampling has on a Langevin-based MCMC method – independently of how time stepping is used to derive the MCMC sampler from the underlying Langevin dynamics. Interestingly, our best-case error analysis shows for this basic form of SGLDiff a behaviour very similar to the discretised algorithm. Comparisons of discrete-in-time and continuous-in-time dynamical systems are not always conclusive. However, this result may indicate that the Euler–Maruyama discretisation that is implicitly used to obtain SGLD from SGLDiff, is appropriate.

[*]Equal contribution, [†]Corresponding author. Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

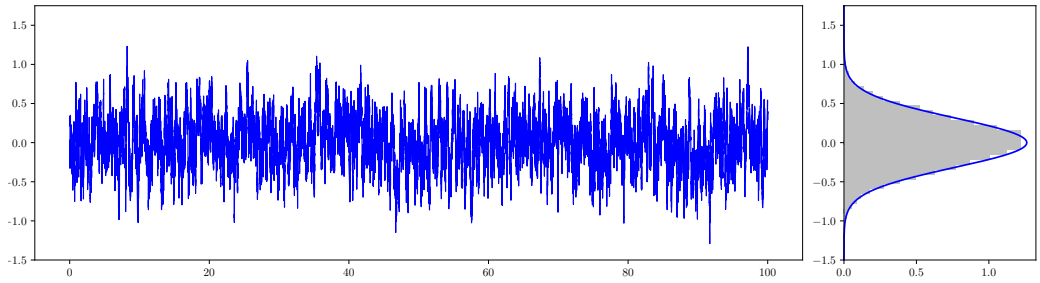

Figure 1.1: Left: Sample path of the diffusion process $\mathrm{d}\zeta_t = -10\zeta_t \mathrm{d}t + \sqrt{2}\mathrm{d}W_t$. Right: Its associated stationary density $\mathrm{N}(0, 0.1)$ and the histogram of the sample path.

2 PROBLEM SETTING

Throughout this work, we aim to approximate a (target) probability distribution μ on a space $X := \mathbb{R}^d$ that we equip with the Euclidean norm $\|\cdot\|$ and its associated Borel-σ-algebra $\mathcal{B}(X)$. We assume that μ is given by

$$\mu(\mathrm{d}\theta) = \frac{1}{Z} \exp\left(-\bar{\Phi}(\theta)\right) \mathrm{d}\theta,$$

where $\bar{\Phi} := \frac{1}{N}\sum_{i=1}^{N} \Phi_i$ is the arithmetic mean of some functions $\Phi_i : X \to \mathbb{R}$ that are bounded below, continuously differentiable, and indexed by $i \in I := \{1, \dots, N\}$, and

$$Z := \int_X \exp\left(-\bar{\Phi}(\theta')\right) \mathrm{d}\theta' \in (0, \infty)$$

is the normalising constant. In a Bayesian learning or inference problem, μ should be thought off as the posterior distribution. In this case, the function Φ_i then refers to the regularised data misfit or the negative log-posterior with respect to the data subset with index $i \in I$. Outside of learning and inference, probability distributions of this form also arise in statistical physics.

We use a Monte Carlo approach to approximate μ, e.g., we generate random samples and then approximate μ by the associated empirical measure. Here, we rely on MCMC techniques that generate a Markov process that is ergodic and stationary with respect to μ, e.g., the sample trajectory can be used to approximate integrals with respect to μ. Under assumptions on $(\Phi_i)_{i\in I}$, an example for such a Markov process is the solution $(\zeta_t)_{t\geq 0}$ of the following *(overdamped) Langevin dynamic*:

$$\mathrm{d}\zeta_t = -\nabla\bar{\Phi}(\zeta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \tag{2.1}$$

where $(W_t)_{t\geq 0}$ is a Brownian motion on X. We show an example where the Langevin diffusion is used to approximate a Gaussian distribution in Figure 1.1. In practice, such a Langevin diffusion is used as an inaccurate MCMC algorithm through an Euler–Maruyama discretisation. Indeed, this is the *unadjusted Langevin algorithm*, where the Markov chain $(\widehat{\zeta}_k)_{k=1}^{\infty}$ is generated by

$$\widehat{\zeta}_{k+1} \leftarrow \widehat{\zeta}_k - \eta\nabla\bar{\Phi}(\widehat{\zeta}_k) + \sqrt{2\eta}\xi_k, \tag{2.2}$$

where $\eta > 0$ is the *learning rate* (or step size) and $(\xi_k)_{k=1}^{\infty}$ is a sequence of indendepent and identically distributed (iid) standard Gaussian random variables on X. ULA approximates $(\zeta_t)_{t\geq 0}$, but does not necessarily converge to μ in its longterm limit.

In practice, N might be very large in which case we may not be able to repeatedly evaluate all N gradients in (2.2). Based on the popular Stochastic Gradient Descent method (Robbins and Monro, 1951) in optimisation , Welling and Teh (2011) have proposed the *Stochastic Gradient Langevin Dynamic*, which is of the form

$$\tilde{\zeta}_{k+1} \leftarrow \tilde{\zeta}_k - \eta\nabla\Phi_{i(k)}(\tilde{\zeta}_k) + \sqrt{2\eta}\xi_k, \tag{2.3}$$

where $i(0), i(1), \dots \sim \mathrm{Unif}(I)$ are iid. This data subsampling that allows us to consider only one gradient at a time introduces again an additional error. In this work, we aim to study this subsampling error, isolatedly from the ULA error. This allows us to obtain a best case error for Langevin-based MCMC methods that are subject to subsampling and is independent from the discretisation. To do so, we will consider the aforementioned *Stochastic Gradient Langevin Diffusion*, a switched diffusion process that is given through the following dynamical system

$$\mathrm{d}\theta_t = -\nabla\Phi_{\boldsymbol{i}(t/\eta)}(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \tag{2.4}$$

where $(\boldsymbol{i}(t))_{t\geq 0}$ is a homogeneous continuous-time Markov process on I that jumps from any state to any other state at rate 1. Here, $\eta > 0$ still has the character of a learning rate. The definition of the SGLDiff is especially motivated by earlier works (Hanu et al., 2023;

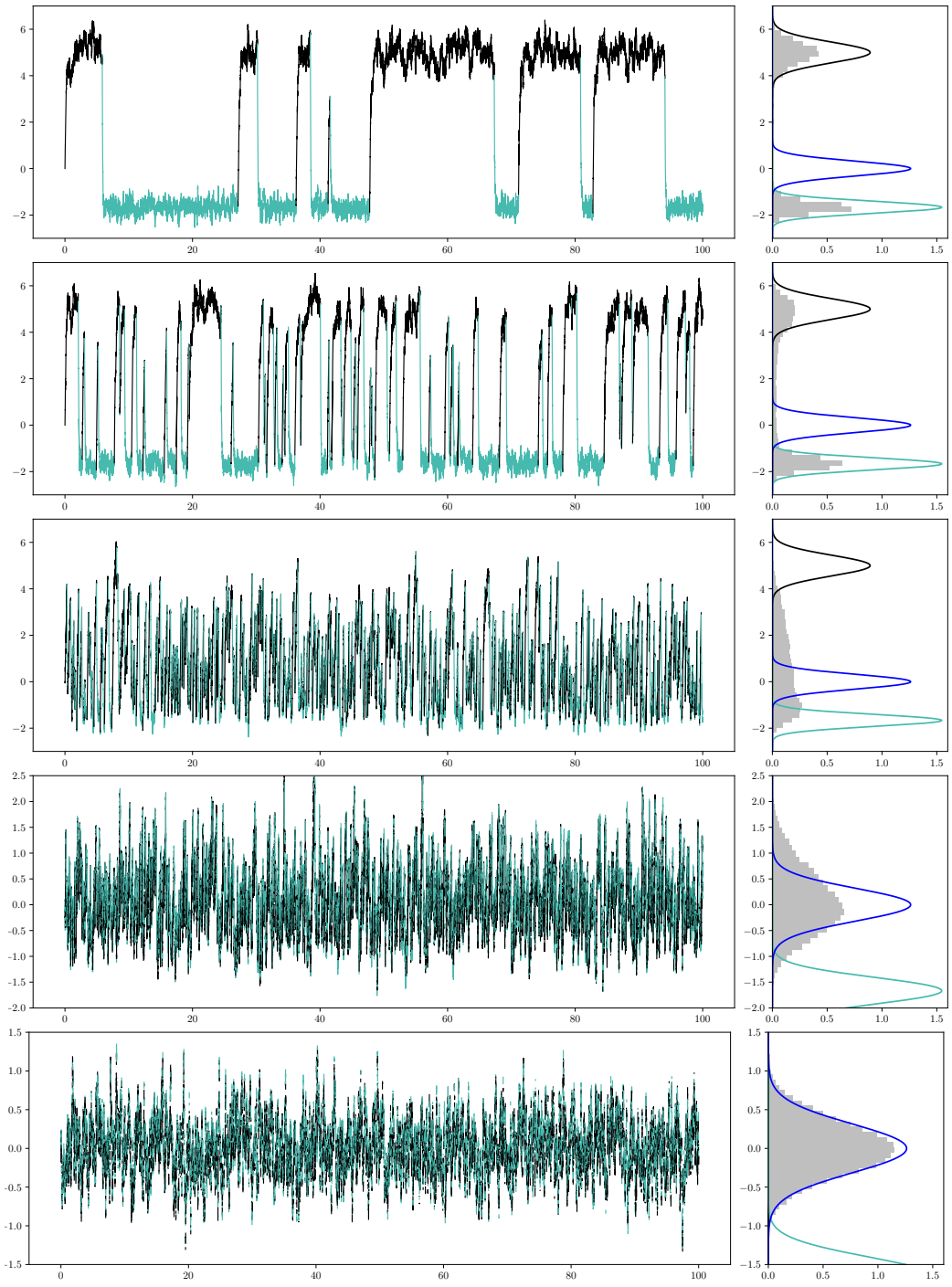

Figure 1.2: Left column: Sample paths of SGLDiffs with $N = 2$ given by $\mathrm{d}\theta_t = a_{\boldsymbol{i}(t)}(b_{\boldsymbol{i}(t)} - \theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$, with $a := (5, 15)$ and $b := (5, -5/3)$, that approximate SDE $\mathrm{d}\zeta_t = -10\zeta_t \mathrm{d}t + \sqrt{2}\mathrm{d}W_t$ and its stationary distribution $\mathrm{N}(0, 0.1)$ given in Figure 1.1, with $\eta = 10^1, 10^0, 10^{-1}, 10^{-2}, 10^{-3}$ (top to bottom). We show the path of $(\theta_t)_{t\geq 0}$ in black whenever $\boldsymbol{i}(t) \equiv 1$ and in teal if $\boldsymbol{i}(t) \equiv 2$. Right column: Stationary densities of subsampled process (e.g., with fixed $\boldsymbol{i}$) in black and teal, respectively, the density of $\mathrm{N}(0, 0.1)$ in blue, and the histogram of the sample path in gray. Note that the scaling of the y-axis changes throughout the plots.

Jin et al., 2022, 2021; Latz, 2021) on continuous-time stochastic gradient descent and different from purely diffusion-based analyses, e.g., such similar to Li et al. (2019). We give examples for sample paths of SGLDiff with different learning rates η in Figure 1.2. There, we especially illustrate that SGLDiff $(\theta_t)_{t\geq 0}$ approximates the Langevin diffusion $(\zeta_t)_{t\geq 0}$, if $\eta \downarrow 0$. Moreover, we can see that SGLDiff also approximates our distribution of interest μ. Throughout this work, we study this approximation of $(\zeta_t)_{t\geq 0}$ using $(\theta_t)_{t\geq 0}$.

2.1 Contributions and outline

We now state the contributions of this work and then give an outline.

From a **learning perspective**, we study the approximation of the Langevin diffusion $(\zeta_t)_{t\geq 0}$ using SGLDiff $(\theta_t)_{t\geq 0}$. Indeed,

- we study convergence and divergence between Langevin dynamic $(\zeta_t)_{t\geq 0}$ and SGLDiff $(\theta_t)_{t\geq 0}$ for small η and large t, respectively,
- we give assumptions under which SGLDiff has a unique stationary distribution μ^η and is ergodic, and we prove an error bound between this distribution μ^η and the target distribution μ, and
- we use the triangle inequality to then also bound the distance between SGLDiff $(\theta_t)_{t\geq 0}$ and target distribution μ, giving us information about bias and convergence at the same time.

From a **probabilistic perspective**, we leverage the key ideas embedded within the ergodic theorem and show the strong convergence between SGLDiff $(\theta_t)_{t\geq 0}$ and Langevin dynamic $(\zeta_t)_{t\geq 0}$ while only having weak convergence between their coefficients $\nabla\Phi_{i(t/\eta)}$ and $\nabla\bar{\Phi}$. We adapt the reflection coupling method in the context of switching diffusion processes and propose an innovative application of this method to address the convergence between the invariant measures μ and μ^η of systems (2.1) and (2.4), respectively.

We formulate the main results of this work in Theorems 2.1–2.3 in Subsection 2.2 and bring them into context with discrete-in-time results in Subsection 2.3. We outline the proofs of Theorem 2.1 and Theorems 2.2–2.3 in Sections 3 and 4 and make them rigorous in Appendices A and B, respectively. We conclude the work in Section 5 and point the reader towards related open problems.

2.2 Main results

We present our main results in this section – starting with two assumptions.

Assumptions 2.1 (Smoothness) *For any $i \in I$, $\Phi_i \in \mathcal{C}^1(X:\mathbb{R})$, i.e., it is continuously differentiable. In addition, $\nabla\Phi_i$ is Lipschitz continuous with Lipschitz constant L, i.e., for any $x, y \in X$,*

$$\|\nabla\Phi_i(x) - \nabla\Phi_i(y)\| \leq L\|x-y\|.$$

Assumptions 2.2 *There exist $K, R > 0$ such that for any $i \in I$ and $\|x-y\| \geq R$,*

$$\langle \nabla\Phi_i(x) - \nabla\Phi_i(y), x-y\rangle \geq K\|x-y\|^2.$$

Assumption 2.1 is usual in the literature (Raginsky et al., 2017; Xu et al., 2018; Zou et al., 2021) and provides existence and uniqueness of the solution to the equation (2.4), see, e.g., Xi (2008) or Yin and Zhu (2010, Chapter 2). Note that the solution is $\mathcal{F}_t^\eta := \sigma(\mathcal{F}_t^B \cup \mathcal{F}_{t/\eta}^I)$-adapted, where $\mathcal{F}_t^B$ is the filtration generated by the Brownian motion $(W_t)_{t\geq 0}$ and $\mathcal{F}_t^I$ is the filtration generated the Markov jump process $(\boldsymbol{i}(t))_{t\geq 0}$. Assumption 2.2 is motivated by Eberle (2011) and Eberle (2016), which allows us to use the reflection coupling method to prove exponential convergence. Intuitively, it states that the Φ_i appear strongly convex if x and y are far from each other for $i \in I$. We remark that while this assumption is weaker than strong convexity, it is stronger than the dissipativeness assumption, which is usually assumed in the discrete-in-time literature for convergence analysis of the non-convex case (e.g. Raginsky et al., 2017; Xu et al., 2018; Zou et al., 2021). See Appendix C where we discuss this connection.

With the above assumptions, we show three convergence results regarding SGLDiff. We begin by showing that the processes $(\zeta_t)_{t\geq 0}$ and $(\theta_t)_{t\geq 0}$ may diverge as $t \to \infty$, but strongly converge at any fixed time t if the learning rate $\eta \downarrow 0$.

Theorem 2.1 *Let $(\theta_t)_{t\geq 0}$ be the solution to (2.4) and $(\zeta_t)_{t\geq 0}$ be the solution to (2.1) with initial value $\theta_0 = \zeta_0$. Under Assumptions 2.1, we have the following inequality*

$$\mathbb{E}[\|\theta_t - \zeta_t\|] \leq C_{\Phi,\theta_0,d} \mathrm{e}^{8(1+L)t}\eta^{\frac{1}{4}},$$

where $C_{\Phi,\theta_0,d} = 8(1+d+\|\theta_0\|^2 + 2\left\|\nabla\bar{\Phi}(0)\right\|^2)^{\frac{1}{2}} C_\Phi^{(1)}$ and $C_\Phi^{(1)} = 1 + L + \sup_{i\in I}\|\nabla\Phi_i(0)\|$.

Here, $C_{\Phi,\theta_0,d}$ depends linearly on $\sqrt{d}$. Next, we study the ergodicity of the SGLDiff (2.4), which we will state in terms of the Wasserstein distance. The Wasserstein distance between two probability measures ν and ν' on $(X, \mathcal{B}(X))$ is given by

$$\mathcal{W}_{\|\cdot\|}(\nu,\nu') = \inf_{\Gamma\in\mathcal{H}(\nu,\nu')} \int_{X\times X} \|y-y'\|\,\Gamma(\mathrm{d}y,\mathrm{d}y'),$$

where $\mathcal{H}(\nu,\nu')$ is the set of coupling between ν and ν', i.e.

$$\mathcal{H}(\nu,\nu') = \{\Gamma \in \Pr(X\times X) : \Gamma(A\times X) = \nu(A), \\ \Gamma(X\times B) = \nu'(B)\ (A,B\in\mathcal{B}(X))\}.$$

We note that Assumptions 2.1 and 2.2 imply that the joint process $(\theta_t, \boldsymbol{i}(t))_{t\geq 0}$ defined in equation (2.4) is Markovian and admits a unique invariant measure $M^\eta(\mathrm{d}\theta, \{i\})$, see for example Yin and Zhu (2010). We denote by $\mu^\eta(\mathrm{d}\theta) := M^\eta(\mathrm{d}\theta, I)$ the $(\theta_t)_{t\geq 0}$-marginal of the stationary distribution M^η and, similarly, the distributions $\nu_t^\eta := \mathbb{P}(\theta_t \in \cdot)$ and $\nu_t := \mathbb{P}(\zeta_t \in \cdot)$ at a fixed time $t > 0$. Finally, we assume in the following that $\boldsymbol{i}(0) \sim \mathrm{Unif}(I)$ and then obtain the following ergodic theorem.

Theorem 2.2 *Under the Assumptions 2.1 and 2.2, we have*

$$\mathcal{W}_{\|\cdot\|}(\nu_t^\eta, \mu^\eta) \leq C\mathrm{e}^{-ct}\mathcal{W}_{\|\cdot\|}(\nu_0, \mu^\eta),$$

where $c = \min\{3L + \frac{2}{R^2}, K\}\mathrm{e}^{-LR^2/2}$ *and* $C = 2\mathrm{e}^{LR^2/2}$.

Theorem 2.2 provides a quantitative way to measure the distance between the distribution of the state of SGLDiff θ_t at time $t > 0$ and its limiting measure, i.e. the exponential convergence between ν_t^η and μ^η. Notice that the constants in the obtained upper bound are independent of the dimension as the reflection coupling reduces the diffusion to a one-dimensional Brownian motion, which will be explained later in the outline of the proof.

In the third convergence result, we study the invariant measures μ and μ^η of $(\zeta_t)_{t\geq 0}$ and $(\theta_t)_{t\geq 0}$. Here, we bound the Wasserstein distance between μ and μ^η and, thus, quantify the asymptotic subsampling error between correct distribution and SGLDiff.

Theorem 2.3 *Under the Assumptions 2.1 and 2.2, the marginal distribution $\mu^\eta(dx)$ converges in $\mathcal{W}_{\|\cdot\|}$ to the target distribution $\mu(dx)$, as $\eta \downarrow 0$. In particular, we have*

$$\mathcal{W}_{\|\cdot\|}(\mu^\eta, \mu) \leq C_{\Phi,d}\eta^{c_\Phi},$$

where $c_\Phi := \frac{c}{32(L+1)+4c}$ *and* $C_{\Phi,d} := C_{\Phi,\theta_0=0,d} + C_d^{(1)}C$, *with* $C_d^{(1)} = \mathcal{O}(\sqrt{d})$.

When t goes to infinity, both, $(\zeta_t)_{t\geq 0}$ and $(\theta_t)_{t\geq 0}$ converge to their invariant measures respectively, and this theorem shows that their invariant measures coincide as the learning rate goes to zero. From Theorem 2.1, we know that the dimension-dependence of the constant $C_{\Phi,d}$ is of order $\mathcal{O}(\sqrt{d})$. Moreover, we have the dimension-independent rate $c_\Phi = 1/4 - \delta$ for some $\delta > 0$. The constant $C_d^{(1)}$ is discussed explicitly in Lemma 4.1.

2.3 Comparison with discrete-in-time Langevin algorithm and related work

There has been an increasing interest in the use of Langevin diffusion-based algorithms for the approximation of Bayesian posterior distributions as these algorithms have demonstrated significant potential for achieving accurate and efficient sampling (Welling and Teh, 2011). The convergence rate has been studied extensively under different log-concavity conditions on the target distribution, see for example Dalalyan (2017a,b); Durmus and Moulines (2016, 2017); Mangoubi and Vishnoi (2019a); as well as in the non-log-concave case, see for example, Balasubramanian et al. (2022); Lee et al. (2018); Ma et al. (2019); Raginsky et al. (2017); Vempala and Wibisono (2019); Xu et al. (2018); Lamperski (2021); Majka et al. (2020); Zhang et al. (2023). In recent years, there has been a growing body of research focused on improving and extending Langevin diffusion-based algorithms for Bayesian sampling. The subsampling-variant of the unadjusted Langevin algorithm, referred to as Stochastic Gradient Langevin Dynamics (SGLD), has proven to be particularly useful for sampling and optimization tasks in which the objective function is nonconvex, noisy, and/or has a large number of parameters. Recall that the Stochastic Gradient Langevin Dynamics updates are defined as in (2.3). The convergence rate of this algorithm and its variants have been studied in for example, Chen et al. (2019); Deng et al. (2020); Gao et al. (2022); Raginsky et al. (2017); Xu et al. (2018); Zhang et al. (2017); Zou et al. (2021). Since then, a significant amount of effort has been put into improving various aspects. For example, SGLD can be combined with variance reduction resulting in a faster convergence rate, such as the Stochastic Variance Reduced Gradient Langevin Dynamics (SVRG-LD), see for instance, Dubey et al. (2016); Huang and Becker (2021); Kinoshita and Suzuki (2022); Xu et al. (2018); Zou et al. (2018b, 2019a, 2021). Another direction of work are higher order MCMC methods, such as Hamiltonian Monte Carlo (see e.g. non-subsampling: Bou-Rabee et al., 2020; Chen et al., 2020; Durmus et al., 2019; Mangoubi and Vishnoi, 2018, 2019b; Neal, 2012, subsampling: Zou and Gu, 2021) and the underdamped Langevin dynamics (see e.g. non-subsampling: Cheng et al., 2018; Eberle, 2016; Eberle et al., 2017, subsampling: Chen et al., 2015, 2017; Gao et al., 2022; Zou et al., 2018a, 2019b; Akyildiz and Sabanis, 2024). For dependent data, see for example Chau et al. (2021).

The convergence rate of the vanilla SGLD in the context of non-convex learning has been studied in several works; see for example, Raginsky et al. (2017); Zou et al. (2021); Li et al. (2023); Majka et al. (2020); Zhang et al. (2023). In particular, under similar conditions as

our Assumption 2.1 and Assumption 2.2, Majka et al. (2020) obtained

$$\mathcal{W}_{\|\cdot\|}(\mathbb{P}(\tilde{\zeta}_k \in \cdot), \mu) \leq e^{R^2}(1 - m\mathrm{e}^{-R^2}\eta)^k + \mathcal{O}\left(e^{R^2}\sqrt{d}\right)\sqrt{\eta},$$

where R is the constant from contractivity-at-infinity condition which is same as the R defined in our Assumption 2.2 and $m > 0$ is independent of d and R. Although R does not depend on the dimension d, there could be a relation between R and d in practice, which makes these constants dimension-dependent, see Bou-Rabee et al. (2020) for details. Even though a direct comparison between our result and this bound may not be possible, it is still noteworthy to observe the analogous error in our continuous scenario, which offers interesting insights. Using the triangle inequality to combine Theorems 2.2 and 2.3, we have

$$\mathcal{W}_{\|\cdot\|}(\nu_t^\eta, \mu) \leq \mathcal{O}(e^{R^2}\sqrt{d})\eta^{c_\Phi} + \mathcal{O}(\mathrm{e}^{R^2})\mathrm{e}^{-\mathcal{O}(\mathrm{e}^{-R^2})t}.$$

The rate in the first term is independent of time, however $c_\Phi \leq 1/4$ indicates slow convergence. The second term decays exponentially in time and it is independent of the dimension d, but depends on R, which, as discussed above, may depend on the dimension d. Hence, our analysis indicates that the ergodic rate does hardly suffer from the Euler-Maruyama discretisation that turns SGLDiff into SGLD; neither does the immediate dimension-dependence of the bias. We do see a similar dependence of R on the bias term in SGLDiff, but also a slower convergence in terms of the learning rate parameter η. Thus, SGLD already achieves an error close to what we obtain in our best-case error analysis obtain.

3 APPROXIMATING THE LANGEVIN DIFFUSION

In this section, we give a sketch of the proof of Theorem 2.1 showing the strong convergence of $\theta_t \to \zeta_t$ for a fixed time $t > 0$, as $\eta \downarrow 0$. The full proof of this theorem and proofs of auxiliary results stated here are deferred to Appendix A. The proof of Theorem 2.1 is inspired by the calculation of the variance for ergodic averages, for example, see Eberle (2023, Chapter 2.2) and Kushner (1984). We notice that $\boldsymbol{i}(\cdot/\eta)$ converges weakly to its invariant measure when $\eta \downarrow 0$. From the ergodic theory for Markov processes, however, we expect that $\int_0^t \nabla\Phi_{\boldsymbol{i}(s/\eta)}(\theta_s)ds = \eta\int_0^{t/\eta}\nabla\Phi_{\boldsymbol{i}(r)}(\theta_{\eta r})dr$ converges to $\int_0^t \nabla\bar{\Phi}(\theta_s)ds$ strongly, which we can then use to prove strong convergence of the full processes. Before sketching the proof of Theorem 2.1, we require some auxiliary results. We start with the following.

Lemma 3.1 *Under Assumption 2.1, for any $t > 0$, we have the following inequality,*

$$\mathbb{E}[\|\zeta_t\|^2] \leq \tilde{c}_{t,\theta_0,d},$$

where $\tilde{c}_{t,\theta_0,d} = \left(\|\theta_0\|^2 + 2\left\|\nabla\bar{\Phi}(0)\right\|^2 + 2td\right)\mathrm{e}^{2(L+1)t}$.

Lemma 3.1 provides the boundedness of $(\zeta_t)_{t\geq 0}$ which will be used repeatedly in the rest of the paper. The following Lemma shows that $(\zeta_t)_{t\geq 0}$ is continuous in time due to the continuity from the drift and the Brownian motion. This continuity allows us to employ a time decomposition later in the proof of Theorem 2.1.

Lemma 3.2 *Under Assumption 2.1, $(\zeta_t)_{t\geq 0}$ is continuous w.r.t time, in the following sense: for $t > s > 0$, we have*

$$\mathbb{E}[\|\zeta_t - \zeta_s\|^2] \leq c_{t,\theta_0,d}\,|t - s|\,,$$

where $c_{t,\theta_0,d} := 2\mathrm{e}^{2(L+1)t}\tilde{c}_{t,\theta_0,d}$.

Notice that the Markov process $(\boldsymbol{i}(t))_{t\geq 0}$ is ergodic, i.e.

$$\frac{1}{T}\int_0^T g_{\boldsymbol{i}(t)}dt \to \frac{1}{N}\sum_{i=1}^N g_i,$$

as $T \to \infty$, for some function $g : I \to X$. The following lemma discusses the precise convergence rate and shows that the time average converges to the space-average with order $\mathcal{O}(1/\sqrt{T})$.

Lemma 3.3 *Let $g : I \to X$ satisfy $\sum_{i=1}^N g_i = 0$. Then*

$$\sup_{\boldsymbol{i}(0)\in I}\mathbb{E}_{\boldsymbol{i}(0)}\left[\left\|\int_0^{\frac{t}{\eta}} g_{\boldsymbol{i}(s)}\mathrm{d}s\right\|^2\right] \leq \frac{2\max_{i=1,\ldots,N}\|g_i\|^2}{N}\frac{t}{\eta}\,.$$

We now have all ingredients to explain how Theorem 2.1 can be proven.

Proof sketch of Theorem 2.1

In the proof of Theorem 2.1, the main idea is to break down the difference of θ_t and ζ_t. First, we examine equations (2.1) and (2.4), we have

$$\begin{aligned}\|\theta_t - \zeta_t\| &= \left\|\int_0^t \nabla\Phi_{\boldsymbol{i}(s/\eta)}(\theta_s) - \nabla\bar{\Phi}(\zeta_s)\mathrm{d}s\right\| \\ &\leq \left\|\int_0^t \nabla\Phi_{\boldsymbol{i}(s/\eta)}(\theta_s) - \nabla\Phi_{\boldsymbol{i}(s/\eta)}(\zeta_s)\mathrm{d}s\right\| \\ &\quad + \left\|\int_0^t \nabla\Phi_{\boldsymbol{i}(s/\eta)}(\zeta_s) - \nabla\bar{\Phi}(\zeta_s)\mathrm{d}s\right\|.\end{aligned}$$

For the term $\nabla\Phi_{\boldsymbol{i}(s/\eta)}(\theta_s) - \nabla\Phi_{\boldsymbol{i}(s/\eta)}(\zeta_s)$, we apply the Lipschitz assumption from Assumption 2.1.

Consequently, $\|\theta_t - \zeta_t\|$ can be bounded by the sum of $\int_0^t \|\theta_s - \zeta_s\| \,\mathrm{d}s$ and $\left\|\int_0^t \nabla\Phi_{\boldsymbol{i}(s/\eta)}(\zeta_s) - \nabla\bar{\Phi}(\zeta_s)\mathrm{d}s\right\|$. Our main goal is to show that the second term is bounded by a constant depending on $t\eta^{1/4}$. To achieve this, we use a discretization technique to estimate the integral (Kushner, 1984 and Jin et al., 2021, proof of Theorem 3). More precisely, one can understand the switching rate η as the discretization time-step and analyze the difference on each time interval of length $\tilde{\eta}$,

$$\int_0^t \nabla\Phi_{\boldsymbol{i}(s/\eta)}(\zeta_s) - \nabla\bar{\Phi}(\zeta_s)\mathrm{d}s = \sum_{j=1}^{1/\tilde{\eta}} \int_{(j-1)t\tilde{\eta}}^{jt\tilde{\eta}} \nabla\Phi_{\boldsymbol{i}(s/\eta)}(\zeta_s) - \nabla\bar{\Phi}(\zeta_s)\mathrm{d}s.$$

Within each time interval $((j-1)t\tilde{\eta}, jt\tilde{\eta}]$ we want to control the variation of $(\zeta_t)_{t\geq 0}$ (using Lemma 3.2), which requires the length $\tilde{\eta}$ to be small. On the other hand, using the ergodicity bound from Lemma 3.3, the fluctuation on each interval has to be large enough so that the overall sum goes to zero. Consequently, we choose $\tilde{\eta} := 1/\lfloor 1/\sqrt{\eta}\rfloor$, where $\lfloor x\rfloor$ is the greatest integer less than or equal to x. This $\tilde{\eta}$ ($\approx\sqrt{\eta}$) optimally satisfies those requirements. Once this bound is established, we apply Grönwall's inequality to obtain the desired result.

4 APPROXIMATING THE DISTRIBUTION OF INTEREST

We now study how well μ^η approximates μ. Again, the full proofs of the main theorems and lemmas are deferred to Appendix B. We begin by showing that $(\theta_t)_{t\geq 0}$ converges exponentially to its stationary measure.

Proof sketch of Theorem 2.2

Before discussing the proof of Theorem 2.2, we recall the exponential contractivity for Markov semi-groups (see e.g. Eberle, 2011, 2023; Latz, 2021). Let $p_t : X \times \mathcal{B}(X) \to [0,1]$ be a homogeneous Markov semi-group and let π be its invariant measure. The exponential contraction in Wasserstein distance induced by some distance d is defined as

$$\mathcal{W}_d(\pi_0 p_t, \pi) \leq \mathrm{e}^{-ct}\mathcal{W}_d(\pi_0, \pi).$$

Now, while the pair $(\theta_t, \boldsymbol{i}(t/\eta))_{t\geq 0}$ is a Markov process, $(\theta_t)_{t\geq 0}$ on its own is not Markovian. Rather than exploring the contractivity of the pair $(\theta_t, \boldsymbol{i}(t/\eta))_{t\geq 0}$, we start the dynamic with $\boldsymbol{i}(0)$ being already distributed according to its invariant measure $\mathrm{Unif}(I)$ and study the contractivity only in $(\theta_t)_{t\geq 0}$. When the potentials $(\Phi_i)_{i\in I}$ are strongly convex, this property is classical and we could use the method in e.g. Latz (2021) to obtain it. More precisely, one can construct a coupled process starting from the invariant measure and run the same dynamic with the same diffusion process. However, in the non-convex case, we do not obtain enough decay solely from the potential hence we need to construct the coupling in a way such that the diffusion term offers extra decay. By selecting an appropriate distance function $F(\cdot)$, it is possible to achieve exponential contractivity even in non-convex potential cases. Here, we choose the distance function to be a supermartingale w.r.t $\mathcal{F}_t^\eta$ and equivalent to the Euclidean distance so that we get exponential decay under this distance and deduce the exponential decay in $\|\cdot\|$. This idea is adapted from the reflection couplings discussed by Eberle (2011, 2016). Intuitively, by diffusing the coupled process along the reflection, we compensate for the lack of decay in the drift.

The classical coupling method fails when the drift term is not strictly contractive. Consider, for instance, a one-dimensional Brownian motion on the torus. In this case, the drift term is 0 and the classical coupling fails: the difference between the two processes is always a constant. In reflection coupling, we couple differently: both processes are driven by the same Brownian motion; however, until they meet, we point the driving processes in opposite directions. Then, the difference between the processes is a one-dimensional Brownian motion and will hit 0 eventually. In our case, the assumption "strongly convex at infinity" allows the processes not to move too far away from one another and to meet eventually. As a result, a fast exponential decay rate can be obtained in the $\mathcal{W}_{\|\cdot\|}$ distance.

Now, we move on to show the error bound between the stationary distribution μ^η and the distribution of interest μ.

Proof sketch of Theorem 2.3

The following lemma shows that μ^η and μ are bounded in terms of their first absolute moments. Recall that μ^η is the marginal distribution of the invariant measure of $(\theta_t)_{t\geq 0}$ and μ is the invariant measure of $(\zeta_t)_{t\geq 0}$.

Lemma 4.1 *Let δ_0 be the Dirac measure concentrated at 0. Under Assumptions 2.1 and 2.2, we have*

$$\mathcal{W}_{\|\cdot\|}(\delta_0, \mu^\eta) \leq C_d^{(1)},$$

where $C_d^{(1)} = \sqrt{C_\Phi K^{-1} d}$ and $C_\Phi = 2(L+K)R^2 + \sup_{i\in I} \frac{\|\nabla\Phi_i(0)\|^2}{K}$.

Specifically, when $N = 1$ – so there is no subsampling – it is easy to conclude that

$$\mathcal{W}_{\|\cdot\|}(\delta_0, \mu) \leq C_d^{(1)}.$$

We first insert ν_t^η and ν_t into the distance between μ^η and μ. Using the triangle inequality, we find that

$$\mathcal{W}_{\|\cdot\|}(\mu^\eta, \mu) \leq \mathcal{W}_{\|\cdot\|}(\mu^\eta, \nu_t^\eta) + \mathcal{W}_{\|\cdot\|}(\nu_t^\eta, \nu_t) + \mathcal{W}_{\|\cdot\|}(\nu_t, \mu).$$

Essentially, the distance between the invariant measures propagates through the distance between their dynamics, $\mathcal{W}_{\|\cdot\|}(\nu_t^\eta, \nu_t)$. Assuming they have the same initial value, this can be controlled using Theorem 2.1 and we obtain an upper bound of order $\eta^{1/4}$. Starting at 0, from Theorem 2.2, we can bound $\mathcal{W}_{\|\cdot\|}(\mu^\eta, \nu_t^\eta)$ and $\mathcal{W}_{\|\cdot\|}(\nu_t, \mu)$ by $\mathcal{W}_{\|\cdot\|}(\delta_0, \mu^\eta) + \mathcal{W}_{\|\cdot\|}(\delta_0, \mu)$ with exponential decay, which are bounded due to Lemma 4.1. Hence the distance between the dynamic and its invariant measure is bounded in both (2.1) and (2.4). Since the left-hand side is independent of t, we choose t freely to obtain an optimal bound. While the contractivity is obtained for each dynamic and their limiting measures, the distance between the dynamics accumulates as t goes to infinity, and the precise rate is given in Theorem 2.1. Hence, we design t as a function of η such that the overall bound goes to 0 as η goes to 0.

5 CONCLUSIONS AND OPEN PROBLEMS

Our analysis has shown that our idealised subsampling MCMC dynamic SGLDiff is able to approximate the distribution of interest μ at high accuracy. We especially learnt that the convergence rate is dimension-independent, only the prefactors depend linearly on the square-root of the dimension of the sample space. However, given that the constant R often depends on the dimension, it is difficult to obtain a fully dimension-independent statement, say, as for preconditioned Crank–Nicolson samplers (Cotter et al., 2013). Overall, we learn that the convergence behaviour of SGLDiff and SGLD are very similar, which might indicate that the Euler–Maruyama discretisation that SGLD is based on is appropriate for the given task.

This result also shows the general usefulness of the SGLDiff as a continuous-time model for subsampling in Langevin-based algorithms and as an analytical tool in their analysis. We list some open questions that could be studied next in this framework below.

Optimisation. SGLD can also be seen as a noisier version of the Stochastic Gradient Descent method (Robbins and Monro, 1951), where additional Gaussian noise is added to the stochastic gradients to further regularize the optimisation problem. In this case, we would probably consider equation (2.4) with an inverse temperature $\beta > 0$, i.e.

$$\mathrm{d}\theta_t = -\nabla \Phi_{\boldsymbol{i}(t/\eta)}(\theta_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}W_t. \tag{5.1}$$

The non-subsampled version of this equation (i.e. setting $\Phi_{\boldsymbol{i}(t/\eta)} = \bar{\Phi}$) has invariant distribution $\mu_\beta(\mathrm{d}\theta) \propto \mathrm{e}^{-\Phi(\theta)/\beta}\mathrm{d}\theta$. With certain assumptions on the potential function $\bar{\Phi}$, μ^β converges to δ_{θ_*} weakly as $\beta \to \infty$, where δ_{θ_*} is the Dirac delta function concentrated in the global minimizer θ_* of $\bar{\Phi}$. Indeed, by rescaling β over time, we can use the associated SDE for global optimisation, see, e.g., Miclo (1992). Going back to subsampling, we may now study the invariant distribution μ_β^η of the process $(\theta_t)_{t\geq 0}$ that solves (5.1). Here, we especially ask, whether $\mu_\beta^\eta \to \delta_{\theta_*}$, if $\beta \uparrow \infty$ and $\eta \downarrow 0$. And thus, whether and how fast this noisier version of Stochastic Gradient Descent can find the global optimiser of $\bar{\Phi}$.

Momentum. Higher-order dynamics have shown to be very successful at optimisation, e.g. ADAM (Kingma and Ba, 2017), and sampling, e.g. the previously mentioned Hamiltonian Monte Carlo. In our work, we can obtain a higher-order dynamic by including a momentum term in equation (2.4) and, thus, obtain an *underdamped Stochastic Gradient Langevin Diffusion*

$$\begin{aligned}\mathrm{d}X_t &= V_t\mathrm{d}t \\ \mathrm{d}V_t &= -\gamma V_t\mathrm{d}t - \nabla\Phi_{\boldsymbol{i}(t/\eta)}(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t,\end{aligned}$$

for which we would study the convergence of the solution $(X_t)_{t\geq 0}$ analogous to that of $(\theta_t)_{t\geq 0}$. The momentum may help to explore complicated energy landscapes in Bayesian deep learning and may reduce the influence of the subsampling. Ideas from Jin et al. (2022) might help the analysis.

Epochs and subsampling without replacement. In practice, the Stochastic Gradient Descent method and the Stochastic Gradient Langevin Dynamics is often employed with full passes through the data set, so-called *epochs*. Here, the index process $\boldsymbol{i}(\cdot)$ actually picks only from subsampled data sets $i \in I$ that were not picked so far and is reset after passing through all of the data. The resulting index process $\boldsymbol{i}(\cdot)$ samples without replacement until the end of an epoch, where it is reset. Then, $\boldsymbol{i}(\cdot)$ is not Markovian per se, but could be lifted into the space $I^{|I|+1}$, where it would be able to track its past in the current epoch and, thus, be Markovian. We would be interested in seeing whether sampling in epochs can improve the convergence of SGLD(iff) and Stochastic Gradient Descent.

Acknowledgements

The authors thank Ö. Deniz Akyildiz and Mateusz Majka for helpful discussions. The third author would like to thank the Isaac Newton Institute for Mathematical

Sciences for support and hospitality during the programme *The mathematical and statistical foundation of future data-driven engineering* when work on this paper was undertaken. This work was supported by: EPSRC Grant Number EP/R014604/1.

References

Akyildiz, Ö. D. and Sabanis, S. (2024). Nonasymptotic analysis of stochastic gradient hamiltonian monte carlo under local conditions for nonconvex optimization.

Balasubramanian, K., Chewi, S., Erdogdu, M. A., Salim, A., and Zhang, S. (2022). Towards a theory of non-log-concave sampling:first-order stationarity guarantees for Langevin Monte Carlo. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2896–2923. PMLR.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Bou-Rabee, N., Eberle, A., and Zimmer, R. (2020). Coupling and convergence for Hamiltonian Monte Carlo. *The Annals of Applied Probability*, 30(3):1209 – 1250.

Chau, N. H., Moulines, E., Rásonyi, M., Sabanis, S., and Zhang, Y. (2021). On stochastic gradient langevin dynamics with dependent data streams: The fully nonconvex case. *SIAM Journal on Mathematics of Data Science*, 3(3):959–986.

Chen, C., Ding, N., and Carin, L. (2015). On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2278–2286.

Chen, C., Wang, W., Zhang, Y., Su, Q., and Carin, L. (2017). A convergence analysis for a class of practical variance-reduction stochastic gradient MCMC. *Science China Information Sciences*, 62.

Chen, Y., Chen, J., Dong, J., Peng, J., and Wang, Z. (2019). Accelerating nonconvex learning via replica exchange Langevin diffusion. *ICLR*.

Chen, Y., Dwivedi, R., Wainwright, M. J., and Yu, B. (2020). Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *J. Mach. Learn. Res.*, 21(1).

Cheng, X., Chatterji, N., Abbasi-Yadkori, Y., Bartlett, P., and Jordan, M. (2018). Sharp convergence rates for Langevin dynamics in the nonconvex setting.

Cotter, S. L., Roberts, G. O., Stuart, A. M., and White, D. (2013). MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster. *Statistical Science*, 28(3):424 – 446.

Dalalyan, A. (2017a). Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 678–689. PMLR.

Dalalyan, A. S. (2017b). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(3):651–676.

Deng, W., Feng, Q., Gao, L., Liang, F., and Lin, G. (2020). Non-convex learning via replica exchange stochastic gradient MCMC. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2474–2483. PMLR.

Dubey, K. A., J. Reddi, S., Williamson, S. A., Poczos, B., Smola, A. J., and Xing, E. P. (2016). Variance reduction in stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Durmus, A. and Moulines, É. (2016). Sampling from a strongly log-concave distribution with the unadjusted Langevin algorithm. *arXiv: Statistics Theory*.

Durmus, A. and Moulines, É. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551 – 1587.

Durmus, A., Moulines, E., and Saksman, E. (2019). On the convergence of Hamiltonian Monte Carlo.

Eberle, A. (2011). Reflection coupling and Wasserstein contractivity without convexity. *Comptes Rendus Mathematique*, 349(19):1101–1104.

Eberle, A. (2016). Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166.

Eberle, A. (2023). Markov processes. *https://uni-bonn.sciebo.de/s/kzTUFff5FrWGAay*.

Eberle, A., Guillin, A., and Zimmer, R. (2017). Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47.

Fox, C. W. and Roberts, S. J. (2012). A tutorial on variational Bayesian inference. *Artificial Intelligence Review*, 38(2):85–95.

Gao, X., Gürbüzbalaban, M., and Zhu, L. (2022). Global convergence of stochastic gradient Hamiltonian Monte Carlo for nonconvex stochastic optimization: Nonasymptotic performance bounds and momentum-based acceleration. *Oper. Res.*, 70(5):2931–2947.

Hanu, M., Latz, J., and Schillings, C. (2023). Subsampling in ensemble Kalman inversion.

Huang, Z. and Becker, S. (2021). Stochastic gradient Langevin dynamics with variance reduction. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Jin, K., Latz, J., Liu, C., and Scagliotti, A. (2022). Losing momentum in continuous-time stochastic optimisation.

Jin, K., Latz, J., Liu, C., and Schönlieb, C.-B. (2021). A continuous-time stochastic gradient descent method for continuous data.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Kinoshita, Y. and Suzuki, T. (2022). Improved convergence rate of stochastic gradient Langevin dynamics with variance reduction and its application to optimization. In *Advances in Neural Information Processing Systems*.

Kushner, H. J. (1984). *Approximation and weak convergence methods for random processes with applications to stochastic systems theory*, volume 6. MIT press.

Lamperski, A. (2021). Projected stochastic gradient langevin algorithms for constrained sampling and non-convex learning. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2891–2937. PMLR.

Latz, J. (2021). Analysis of stochastic gradient descent in continuous time. *Statistics and Computing*, 31.

Lee, H., Risteski, A., and Ge, R. (2018). Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering Langevin Monte Carlo. In *Advances in Neural Information Processing Systems*, volume 31.

Li, L., Liu, J.-G., and Wang, Y. (2023). Geometric ergodicity of sgld via reflection coupling.

Li, Q., Tai, C., and E, W. (2019). Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47.

Ma, Y.-A., Chen, Y., Jin, C., Flammarion, N., and Jordan, M. (2019). Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116:201820003.

Majka, M. B., Mijatović, A., and Szpruch, L. (2020). Nonasymptotic bounds for sampling algorithms without log-concavity. *The Annals of Applied Probability*, 30(4):1534–1581.

Mangoubi, O. and Vishnoi, N. K. (2018). Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. In *Neural Information Processing Systems*.

Mangoubi, O. and Vishnoi, N. K. (2019a). Nonconvex sampling with the Metropolis-adjusted Langevin algorithm. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2259–2293. PMLR.

Mangoubi, O. and Vishnoi, N. K. (2019b). Nonconvex sampling with the Metropolis-adjusted Langevin algorithm. In *COLT*.

Miclo, L. (1992). Recuit simulé sur $\mathbb{R}^n$. Étude de l'évolution de l'énergie libre. *Annales de l'I.H.P. Probabilités et statistiques*, 28(2):235 – 266.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer, New York, NY.

Neal, R. M. (2012). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113 – 162.

Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703. PMLR.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407.

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

Vempala, S. and Wibisono, A. (2019). Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, volume 32.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. ICML'11, page 681–688, Madison, WI, USA.

Xi, F. (2008). On the stability of jump-diffusions with Markovian switching. *Journal of Mathematical Analysis and Applications*, 341(1):588–600.

Xu, P., Chen, J., Zou, D., and Gu, Q. (2018). Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 31.

Yin, G. G. and Zhu, C. (2010). *Hybrid Switching Diffusions: Properties and Applications*. Springer New York, New York, NY.

Zhang, Y., Akyildiz, Ö. D., Damoulas, T., and Sabanis, S. (2023). Nonasymptotic estimates for stochastic gradient langevin dynamics under local conditions in nonconvex optimization. *Applied Mathematics & Optimization*, 87(2).

Zhang, Y., Liang, P., and Charikar, M. (2017). A hitting time analysis of stochastic gradient Langevin dynamics. In *Annual Conference Computational Learning Theory*.

Zou, D. and Gu, Q. (2021). On the convergence of Hamiltonian Monte Carlo with stochastic gradients. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 13012–13022. PMLR.

Zou, D., Xu, P., and Gu, Q. (2018a). Stochastic variance-reduced Hamilton Monte Carlo methods. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 6028–6037. PMLR.

Zou, D., Xu, P., and Gu, Q. (2018b). Subsampled stochastic variance-reduced gradient Langevin dynamics. In *International Conference on Uncertainty in Artificial Intelligence*.

Zou, D., Xu, P., and Gu, Q. (2019a). Sampling from non-log-concave distributions via variance-reduced gradient Langevin dynamics. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2936–2945. PMLR.

Zou, D., Xu, P., and Gu, Q. (2019b). Stochastic gradient Hamiltonian Monte Carlo methods with recursive variance reduction. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zou, D., Xu, P., and Gu, Q. (2021). Faster convergence of stochastic gradient Langevin dynamics for non-log-concave sampling. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1152–1162. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Subsampling Error in Stochastic Gradient Langevin Diffusions: Supplementary Materials

In this supplementary material, we organize our content into three appendices. We provide the proofs for all Theorems and Lemmas from the main paper. Appendix A contains the proof of Theorem 2.1 from the main text. We prove Theorem 2.1 by first proving Lemma 3.1, Lemma 3.2, and Lemma 3.3. Appendix B has the proof of Theorem 2.2 and Theorem 2.3. Appendix C discusses the difference between the dissipativeness assumption and Assumption 2.2.

A Proof of Theorem 2.1

We prove Theorem 2.1 starting with showing Lemma 3.1.

A.1 Proof of Lemma 3.1

The following lemma shows the boundedness of $(\zeta_t)_{t\geq 0}$ which will be used repeatedly in the rest of the paper.

Lemma 3.1 *Under Assumption 2.1, for any $t > 0$, we have the following inequality,*

$$\mathbb{E}[\|\zeta_t\|^2] \leq \tilde{c}_{t,\theta_0,d},$$

where $\tilde{c}_{t,\theta_0,d} = \Big(\|\theta_0\|^2 + 2t \left\|\nabla\bar{\Phi}(0)\right\|^2 + 2td\Big) e^{2(L+1)t}$.

Proof. By Itô's formula, we have

$$\begin{aligned}
\frac{\|\zeta_t\|^2}{2} &= \|\theta_0\|^2 - \int_0^t \left\langle \zeta_s, \nabla\bar{\Phi}(\zeta_s)\right\rangle \mathrm{d}t + \sqrt{2}\int_0^t \langle \zeta_s, \mathrm{d}B_s\rangle + td \\
&= \|\theta_0\|^2 - \int_0^t \left\langle \zeta_s, \nabla\bar{\Phi}(\zeta_s) - \nabla\bar{\Phi}(0)\right\rangle \mathrm{d}s \\
&\quad - \int_0^t \left\langle \zeta_s, \nabla\bar{\Phi}(0)\right\rangle \mathrm{d}s + \sqrt{2}\int_0^t \langle \zeta_s, \mathrm{d}B_s\rangle + td \\
&\leq \|\theta_0\|^2 + L\int_0^t \|\zeta_s\|^2 \,\mathrm{d}s + \left\|\nabla\bar{\Phi}(0)\right\| \int_0^t \|\zeta_s\| \,\mathrm{d}s + \sqrt{2}\int_0^t \langle \zeta_s, \mathrm{d}B_s\rangle + td \\
&\leq \|\theta_0\|^2 + (L+1)\int_0^t \|\zeta_s\|^2 \,\mathrm{d}s + t\left\|\nabla\bar{\Phi}(0)\right\|^2 + \sqrt{2}\int_0^t \langle \zeta_s, \mathrm{d}B_s\rangle + td.
\end{aligned}$$

Taking expectation of both sides, we have

$$\frac{\mathbb{E}[\|\zeta_t\|^2]}{2} \leq \frac{\|\theta_0\|^2}{2} + (L+1)\int_0^t \mathbb{E}[\|\zeta_s\|^2]\mathrm{d}s + t\left\|\nabla\bar{\Phi}(0)\right\|^2 + td.$$

By using Grönwall's inequality, we obtain the bound

$$\mathbb{E}[\|\zeta_t\|^2] \leq \Big(\|\theta_0\|^2 + 2t\left\|\nabla\bar{\Phi}(0)\right\|^2 + td\Big) e^{2(L+1)t},$$

which completes the proof. □

A.2 Proof of Lemma 3.2

Lemma 3.2 *Under Assumption 2.1, $(\theta_t)_{t\geq 0}$ is continuous w.r.t time, in the following sense: for $t > s > 0$, we have*

$$\mathbb{E}[\|\zeta_t - \zeta_s\|^2] \leq c_{t,\theta_0,d}\,|t-s|\,,$$

where $c_{t,\theta_0,d} := 2\mathrm{e}^{2(L+1)t}\tilde{c}_{t,\theta_0,d}$.

Proof. From equation (2.1), we get

$$\|\zeta_t - \zeta_s\| \leq \underbrace{\int_s^t \left\|\nabla\bar{\Phi}(\zeta_r)\right\| \mathrm{d}r}_{(m2.1)} + \underbrace{\sqrt{2}\,\|B_t - B_s\|}_{(m2.2)}.$$

The second term can be bounded by the variance of increments of Brownian motions,

$$\mathbb{E}\Big[(m2.2)^2\Big] = 2\,|t-s|\,.$$

Consider the first term,

$$\begin{aligned}(m2.1) = \int_s^t \left\|\nabla\bar{\Phi}(\zeta_r)\right\| \mathrm{d}r &= \int_s^t (\left\|\nabla\bar{\Phi}(\zeta_r) - \nabla\bar{\Phi}(0)\right\| + \left\|\nabla\bar{\Phi}(0)\right\|)\mathrm{d}r \\ &\leq L\int_s^t \|\zeta_r\|\,\mathrm{d}r + \left\|\nabla\bar{\Phi}(0)\right\|\,|t-s|\,.\end{aligned}$$

By Lemma 3.1, we conclude

$$\mathbb{E}[|(m2.1)|^2] \leq 2(\tilde{c}_{t,\theta_0,d} + \left\|\nabla\bar{\Phi}(0)\right\|^2)t\,|t-s|\,,$$

which yields

$$\mathbb{E}[\|\zeta_t - \zeta_s\|^2] \leq 2\mathbb{E}[|(m2.1)|^2] + 2\mathbb{E}[|(m2.2)|^2] \leq c_{t,\theta_0,d}\,|t-s|$$

for some constant $c_{t,\theta_0,d}$. □

A.3 Proof of Lemma 3.3

Lemma 3.3 *Let $g : I \to X$ satisfy $\sum_{i=1}^N g_i = 0$. Then*

$$\sup_{\boldsymbol{i}(0)\in I} \mathbb{E}_{\boldsymbol{i}(0)}\Big[\,\Big\|\int_0^{\frac{t}{\eta}} g_{\boldsymbol{i}(s)}\mathrm{d}s\Big\|^2\,\Big] \leq \frac{2\max_{i=1,\ldots,N}\|g_i\|^2}{N}\frac{t}{\eta}\,.$$

Proof. We rewrite the square integral and use the Markov property of $(\boldsymbol{i}(t))_{t\geq 0}$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{i}(0)}\Big[\Big\|\int_0^{\frac{t}{\eta}} g_{\boldsymbol{i}(s)}\mathrm{d}s\Big\|^2\Big] =& \mathbb{E}_{\boldsymbol{i}(0)}\Big[\int_0^{\frac{t}{\eta}}\int_0^{\frac{t}{\eta}} \big\langle g_{\boldsymbol{i}(s)}, g_{\boldsymbol{i}(r)}\big\rangle \,\mathrm{d}s\mathrm{d}r\Big]\\
=& 2\mathbb{E}_{\boldsymbol{i}(0)}\Big[\int_0^{\frac{t}{\eta}}\int_0^{\frac{t}{\eta}} \big\langle g_{\boldsymbol{i}(s)}, g_{\boldsymbol{i}(r)}\big\rangle \mathbf{1}_{r\leq s}\mathrm{d}s\mathrm{d}r\Big] \ \textit{(since } s,r \textit{ are symmetric)}\\
=& 2\mathbb{E}_{\boldsymbol{i}(0)}\Big[\int_0^{\frac{t}{\eta}}\int_r^{\frac{t}{\eta}} \big\langle g_{\boldsymbol{i}(s)}, g_{\boldsymbol{i}(r)}\big\rangle \,\mathrm{d}s\mathrm{d}r\Big]\\
=& 2\mathbb{E}_{\boldsymbol{i}(0)}\Big[\int_0^{\frac{t}{\eta}}\int_r^{\frac{t}{\eta}} \mathbb{E}[\big\langle g_{\boldsymbol{i}(s)}, g_{\boldsymbol{i}(r)}\big\rangle |\mathcal{F}_r]\mathrm{d}s\mathrm{d}r\Big]\\
=& 2\mathbb{E}_{\boldsymbol{i}(0)}\Big[\int_0^{\frac{t}{\eta}}\int_r^{\frac{t}{\eta}} \mathbb{E}_{j=\boldsymbol{i}(r)}[\big\langle g_{\boldsymbol{i}(s-r)}, g_j\big\rangle]\mathrm{d}s\mathrm{d}r\Big] \ \textit{(by Markov property)}\\
=& 2\mathbb{E}_{\boldsymbol{i}(0)}\Big[\int_0^{\frac{t}{\eta}}\int_r^{\frac{t}{\eta}} \frac{1-e^{-N(s-r)}}{N}\underbrace{\Big\langle\sum_{i=1}^N g_i, g_{\boldsymbol{i}(r)}\Big\rangle}_{=0} + e^{-N(s-r)}\big\|g_{\boldsymbol{i}(r)}\big\|^2 \mathrm{d}s\mathrm{d}r\Big]\\
& \Big(\frac{1-e^{-N(s-r)}}{N} \textit{is the probability switching from } j \textit{ to any other state in } (s-r,s].\Big)\\
=& 2\int_0^{\frac{t}{\eta}}\int_r^{\frac{t}{\eta}} e^{-N(s-r)}\mathbb{E}_{\boldsymbol{i}(0)}[\big\|g_{\boldsymbol{i}(r)}\big\|^2]\mathrm{d}s\mathrm{d}r\\
\leq& 2\max_{i=1,\dots,N}\|g_i\|^2\int_0^{\frac{t}{\eta}}\int_0^{\frac{t}{\eta}-r} e^{-Nm}\mathrm{d}m\mathrm{d}r\\
\leq& \frac{2\max_{i=1,\dots,N}\|g_i\|^2}{N}\frac{t}{\eta}.
\end{aligned}$$

□

A.4 Proof of Theorem 2.1

Theorem 2.1 *Let $(\theta_t)_{t\geq 0}$ be the solution to (2.4) and $(\zeta_t)_{t\geq 0}$ be the solution to (2.1) with initial value $\theta_0=\zeta_0$. Under Assumption 2.1, we have the following inequality*

$$\mathbb{E}[\|\theta_t-\zeta_t\|]\leq C_{\Phi,\theta_0,d}\mathrm{e}^{8(1+L)t}\eta^{\frac{1}{4}},$$

where $C_{\Phi,\theta_0,d}=8(1+d+\|\theta_0\|^2+2\big\|\nabla\bar{\Phi}(0)\big\|^2)^{\frac{1}{2}}C_\Phi^{(1)}$ and $C_\Phi^{(1)}=1+L+\sup_{i\in I}\|\nabla\Phi_i(0)\|$.

Proof. We decompose $\|\theta_t-\zeta_t\|$ into two terms using equations (2.1) and (2.4),

$$\begin{aligned}
\|\theta_t-\zeta_t\| =& \Big\|\int_0^t \nabla\Phi_{\boldsymbol{i}(s/\eta)}(\theta_s)-\nabla\bar{\Phi}(\zeta_s)\mathrm{d}s\Big\|\\
\leq& \Big\|\int_0^t \nabla\Phi_{\boldsymbol{i}(s/\eta)}(\theta_s)-\nabla\Phi_{\boldsymbol{i}(s/\eta)}(\zeta_s)\mathrm{d}s\Big\| + \Big\|\int_0^t \nabla\Phi_{\boldsymbol{i}(s/\eta)}(\zeta_s)-\nabla\bar{\Phi}(\zeta_s)\mathrm{d}s\Big\|\\
\leq& L\int_0^t\|\theta_s-\zeta_s\|\,\mathrm{d}s + \Big\|\int_0^t \nabla\Phi_{\boldsymbol{i}(s/\eta)}(\zeta_s)-\nabla\bar{\Phi}(\zeta_s)\mathrm{d}s\Big\|. \qquad (A.1)
\end{aligned}$$

We claim that $\mathbb{E}[\big\|\int_0^t \nabla\Phi_{\boldsymbol{i}(s/\eta)}(\zeta_s)-\nabla\bar{\Phi}(\zeta_s)\mathrm{d}s\big\|]$ can be bounded by $C_{t,\theta_0,d}\sqrt{\eta}$ for some $C_{t,\theta_0,d}>0$. Let $\tilde{\eta}:=1/\lfloor 1/\sqrt{\eta}\rfloor$, where $\lfloor x\rfloor$ is the greatest integer less than or equal to x. Then we have the following decomposition

$$\int_0^t\Big(\nabla\Phi_{\boldsymbol{i}(s/\eta)}(\zeta_s)-\nabla\bar{\Phi}(\zeta_s)\Big)\mathrm{d}s=\sum_{i=1}^{1/\tilde{\eta}}\int_{(i-1)t\tilde{\eta}}^{it\tilde{\eta}} G(\boldsymbol{i}(s/\eta),\zeta_s)\mathrm{d}s,$$

where $G(i,x) = \nabla\Phi_i(x) - \nabla\bar{\Phi}(x)$. For fixed i, $G(i,x)$ is Lipschitz continuous with constant L. Hence,

$$\begin{aligned}\left\|\int_{(i-1)t\tilde{\eta}}^{it\tilde{\eta}}(G(\boldsymbol{i}(s/\eta),\zeta_s)\mathrm{d}s\right\| \leq& \left\|\int_{(i-1)t\tilde{\eta}}^{it\tilde{\eta}}(G(\boldsymbol{i}(s/\eta),\zeta_s) - G(\boldsymbol{i}(s/\eta),\zeta_{(i-1)t\tilde{\eta}}))\mathrm{d}s\right\| \\ &+ \left\|\int_{(i-1)t\tilde{\eta}}^{it\tilde{\eta}}(G(\boldsymbol{i}(s/\eta),\zeta_{(i-1)t\tilde{\eta}})\mathrm{d}s\right\| \\ \leq& L\underbrace{\int_{(i-1)t\tilde{\eta}}^{it\tilde{\eta}}\left\|\zeta_s - \zeta_{(i-1)t\tilde{\eta}}\right\|\mathrm{d}s}_{(p2.1)} + \underbrace{\left\|\int_{(i-1)t\tilde{\eta}}^{it\tilde{\eta}}(G(\boldsymbol{i}(s/\eta),\zeta_{(i-1)t\tilde{\eta}})\mathrm{d}s\right\|}_{(p2.2)}.\end{aligned}$$

By Lemma 3.2, we bound the first term as

$$\mathbb{E}[(p2.1)] \leq L(c_{t,\theta_0,d})^{\frac{1}{2}}(t\tilde{\eta})^{\frac{3}{2}}.$$

We first study the second term whilst conditioning on $\mathcal{F}^{\eta}_{(i-1)t\tilde{\eta}}$,

$$\begin{aligned}\mathbb{E}\Big[(p2.2)\Big|\mathcal{F}^{\eta}_{(i-1)t\tilde{\eta}}\Big] &= \mathbb{E}_{i^\eta((i-1)t\tilde{\eta}),x=\zeta_{(i-1)t\tilde{\eta}}}\Big[\left\|\int_0^{t\tilde{\eta}}(G(\boldsymbol{i}(s/\eta),x)\mathrm{d}s\right\|\Big] \\ &\leq \Big[\mathbb{E}_{i^\eta((i-1)t\tilde{\eta}),x=\zeta_{(i-1)t\tilde{\eta}}}\left\|\int_0^{t\tilde{\eta}}(G(\boldsymbol{i}(s/\eta),x)\mathrm{d}s\right\|^2\Big]^{\frac{1}{2}} \\ &\overset{r=s/\eta}{=} \Big[\mathbb{E}_{i^\eta((i-1)t\tilde{\eta}),x=\zeta_{(i-1)t\tilde{\eta}}}\eta^2\left\|\int_0^{t\tilde{\eta}\eta^{-1}}(G(\boldsymbol{i}(r),x)\mathrm{d}r\right\|^2\Big]^{\frac{1}{2}} \\ &\underbrace{\leq}_{\text{Lemma 3.3}} \frac{2\max_{j=1,\ldots,N}\left\|G(j,\zeta_{(i-1)t\tilde{\eta}})\right\|}{\sqrt{N}}\sqrt{t\eta\tilde{\eta}} \\ &= \frac{2\max_{j=1,\ldots,N}\left\|(\nabla\Phi_j - \nabla\bar{\Phi})(\zeta_{(i-1)t\tilde{\eta}})\right\|}{\sqrt{N}}\sqrt{t\eta\tilde{\eta}} \\ &\leq C^{(1)}_{\Phi}(1 + \left\|\zeta_{(i-1)t\tilde{\eta}}\right\|)\sqrt{t\eta\tilde{\eta}},\end{aligned}$$

where $C^{(1)}_{\Phi} = 2(1 + L + \sup_{i\in I}\|\nabla\Phi_i(0)\|)$. By Lemma 3.1, this implies

$$\mathbb{E}[(p2.2)] \leq C^{(1)}_{\Phi}(c_{t,\theta_0,d})^{\frac{1}{2}}\tilde{\eta}^{\frac{3}{2}}.$$

Hence,

$$\mathbb{E}[\|\theta_t - \zeta_t\|] \leq L\int_0^t \mathbb{E}[\|\theta_s - \zeta_s\|]\mathrm{d}s + C^{(1)}_{\Phi}(c_{t,\theta_0,d})^{\frac{1}{2}}(1+\sqrt{t})\tilde{\eta}^{\frac{1}{2}}.$$

Using Grönwall's inequality yields

$$\mathbb{E}[\|\theta_t - \zeta_t\|] \leq C^{(1)}_{\Phi}(c_{t,\theta_0,d})^{\frac{1}{2}}(1+\sqrt{t})\tilde{\eta}^{\frac{1}{2}}e^{Lt}.$$

Recall that $(c_{t,\theta_0,d})^{\frac{1}{2}}(1+\sqrt{t}) = 2(1+\sqrt{t})(1 + 2td + \|\theta_0\|^2 + 2\left\|\nabla\bar{\Phi}(0)\right\|^2)^{\frac{1}{2}}e^{4(L+1)t}$. Therefore,

$$\mathbb{E}[\|\theta_t - \zeta_t\|] \leq C_{\Phi,\theta_0,d}e^{8(L+1)t}\eta^{\frac{1}{4}},$$

where $C_{\Phi,\theta_0,d} = 8(1 + d + \|\theta_0\|^2 + 2\left\|\nabla\bar{\Phi}(0)\right\|^2)^{\frac{1}{2}}(1 + L + \sup_{i\in I}\|\nabla\Phi_i(0)\|)$. □

We remark that the factor in t can be improved with an additional assumption but the rate in η remains unchanged. In fact, with Assumption 2.2, the bound can be improved to $(\mathrm{e}^{8(1+L)t}\eta^{1/4}) \wedge (R + \eta^{1/4})$.

B Proof of Theorem 2.2 and Theorem 2.3

B.1 Proof of Theorem 2.2

Theorem 2.2 *Under Assumptions 2.1 and 2.2, we have*

$$\mathcal{W}_{\|\cdot\|}(\nu_t^\eta, \mu^\eta) \leq C\mathrm{e}^{-ct}\mathcal{W}_{\|\cdot\|}(\nu_0, \mu^\eta),$$

where $c = \min\{\big(3L + \frac{2}{R^2}\big), K\}\mathrm{e}^{-LR^2/2}$ *and* $C = 2\mathrm{e}^{LR^2/2}$.

Proof. We adapt the reflection coupling method introduced in Eberle (2011, 2016). Let $(\theta_t)_{t\geq 0}$ be the solution to equation (2.4) with $\theta_0 \sim \nu$. In the coupling approach, we construct another solution $(\tilde{\theta}_t)_{t\geq 0}$ of the same SDE on the same probability space with the same index process $(\boldsymbol{i}(t/\eta))_{t\geq 0}$ and with a different initial law in θ denoted as $\tilde{\theta}_0 \sim \mu^\eta$, i.e.

$$\begin{cases} d\theta_t &= -\nabla\Phi_{\boldsymbol{i}(t/\eta)}(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \\ d\tilde{\theta}_t &= -\nabla\Phi_{\boldsymbol{i}(t/\eta)}(\tilde{\theta}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\tilde{B}_t \\ i(t=0) &= i_0 \\ \tilde{\theta}(t=0) &= \tilde{\theta}_0 \sim \mu^\eta,\ \theta(t=0) = \theta_0 \sim \nu \end{cases} \tag{B.1}$$

where

$$\tilde{B}_t = \int_0^t (I_d - 2e_s e_s^T \mathbf{1}_{\theta_s \neq \tilde{\theta}_s}) dB_s, \quad e_s = (\theta_s - \tilde{\theta}_s)/\left\|\theta_s - \tilde{\theta}_s\right\|,$$

and I_d is the identity matrix of dimension d. It is not hard to verify $I_d - 2e_s e_s^T$ is an orthogonal matrix, which implies that $\tilde{B}_t$ is a d-dimensional Brownian motion.

Let $T = \inf\{t \geq 0 : \theta_t = \tilde{\theta}_t\}$ and $r_t = \left\|\theta_t - \tilde{\theta}_t\right\|$, then for $t < T$, the difference between θ_t and $\tilde{\theta}_t$ satisfies

$$d(\theta_t - \tilde{\theta}_t) = -(\nabla\Phi_{\boldsymbol{i}(t/\eta)}(\theta_t) - \nabla\Phi_{\boldsymbol{i}(t/\eta)}(\tilde{\theta}_t))\mathrm{d}t + 2\sqrt{2}e_t \mathrm{d}B_t^1, \tag{B.2}$$

where $B_t^1 := \int_0^t e_s \cdot dB_s$, which is a one-dimensional Brownian motion. Hence, for $F \in C^2(\mathbb{R})$, by Itô's formula, we have, for $t < T$,

$$dF(r_t) = \Big[-\Big\langle e_t, \nabla\Phi_{\boldsymbol{i}(t/\eta)}(\theta_t) - \nabla\Phi_{\boldsymbol{i}(t/\eta)}(\tilde{\theta}_t)\Big\rangle F'(r_t) + 4F''(r_t)\Big]dt + 2\sqrt{2}F'(r_t)dB_t^1.$$

We choose $F(r) = \int_0^r e^{-\frac{L\min\{s,R\}^2}{2}}(1 - \frac{1}{2R}\min\{s, R\})ds$. Note that F' is non-increasing. Hence, $e^{-\frac{LR^2}{2}}r/2 \leq F(r) \leq r$. Next, we are going to verify that for some constant $c > 0$,

$$(L\mathbf{1}_{r\leq R} - K\mathbf{1}_{r>R})rF'(r) + 4F''(r) \leq -cF(r). \tag{B.3}$$

When $r > R$, since $F''(r) \leq 0$ and $F'(r) = e^{-\frac{LR^2}{2}}$, (B.3) holds with constant $c \leq Ke^{-\frac{LR^2}{2}}$. For $r \leq R$, we have $F'(r) = e^{-\frac{Lr^2}{2}}(1 - \frac{r}{2R})$ and $F''(r) = e^{-\frac{Lr^2}{2}}(-\frac{2Lr}{2} + \frac{Lr^2}{2R} - \frac{1}{2R})$. Hence, for $r \leq R$, the left side of (B.3) is

$$\begin{aligned} L\mathbf{1}_{r\leq R}rF'(r) + 4F''(r) =& e^{-\frac{Lr^2}{2}}r\big(L - \frac{Lr}{2R} - 4L + \frac{Lr}{2R} - \frac{2}{rR}\big) \\ \leq& -e^{-\frac{Lr^2}{2}}r\big(3L + \frac{2}{rR}\big) \leq -\big(3L + \frac{2}{R^2}\big)e^{-\frac{LR^2}{2}}F(r). \end{aligned}$$

Setting $c = \min\{\big(3L + \frac{2}{R^2}\big), K\}e^{-\frac{LR^2}{2}}$ yields inequality (B.3). By Assumptions 2.1 and 2.2, since $r_t = 0$ for $t \geq T$, we know $e^{ct}F(r_t)$ is a supermartingale w.r.t $\mathcal{F}_t^\eta$. Therefore,

$$\mathbb{E}[F(r_t)] \leq e^{-ct}\mathbb{E}[F(r_0)].$$

Recall that $e^{-\frac{LR^2}{2}}r/2 \leq F(r) \leq r$, we get

$$\mathcal{W}_{\|.\|}(\nu_t^\eta, \tilde{\nu}_t^\eta) \leq Ce^{-ct}\mathcal{W}_{\|.\|}(\nu_0, \mu^\eta)$$

for $C = 2e^{\frac{LR^2}{2}}$. Since μ^η is invariant in time, we have $\tilde{\nu}_t^\eta = \nu_0^\eta = \mu^\eta(\cdot, I)$,which completes the proof. □

B.2 Proof of Lemma 4.1

Lemma 4.1 *Let δ_0 be the Dirac delta function at 0. Under Assumptions 2.1 and 2.2, we have*

$$\mathcal{W}_{\|.\|}(\delta_0, \mu^\eta) \leq C_d^{(1)},$$

where $C_d^{(1)} = \sqrt{C_\Phi K^{-1} d}$ and $C_\Phi = 2(L+K)R^2 + \sup_{i\in I} \frac{\|\nabla \Phi_i(0)\|^2}{K}$.

Specifically, when $N = 1$, it is easy to conclude that

$$\mathcal{W}_{\|.\|}(\delta_0, \mu) \leq C_d^{(1)}.$$

Proof. Let ν_t^η be the distribution of θ_t with $(\theta_0, I_0) = (0, i_0)$ and $i_0 \sim \text{Unif}(I)$, we have

$$\mathcal{W}_{\|.\|}(\delta_0, \mu^\eta) \leq \mathcal{W}_{\|.\|}(\delta_0, \nu_t^\eta) + \mathcal{W}_{\|.\|}(\nu_t^\eta, \mu^\eta).$$

From Theorem 2.2, we can bound the second term via

$$\mathcal{W}_{\|.\|}(\nu_t^\eta, \mu^\eta) \leq Ce^{-ct}\mathcal{W}_{\|.\|}(\delta_0, \mu^\eta).$$

For the first term, by Itô's formula, we have

$$\begin{aligned}
\mathrm{d}\left\|\theta_t\right\|^2 =& -2\left\langle \theta_t, \nabla\Phi_{\boldsymbol{i}(t/\eta)}(\theta_t)\right\rangle \mathrm{d}t + 2\sqrt{2}\left\langle \theta_t, \mathrm{d}B_t\right\rangle + 2d\mathrm{d}t \\
=& -2\left\langle \theta_t, \nabla\Phi_{\boldsymbol{i}(t/\eta)}(\theta_t) - \nabla\Phi_{\boldsymbol{i}(t/\eta)}(0)\right\rangle \mathrm{d}t \\
& -2\left\langle \theta_t, \nabla\Phi_{\boldsymbol{i}(t/\eta)}(0)\right\rangle dt + 2\sqrt{2}\left\langle \theta_t, \mathrm{d}B_t\right\rangle + 2d\mathrm{d}t.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&-2\left\langle \theta_t, \nabla\Phi_{\boldsymbol{i}(t/\eta)}(\theta_t) - \nabla\Phi_{\boldsymbol{i}(t/\eta)}(0)\right\rangle - 2\left\langle \theta_t, \nabla\Phi_{\boldsymbol{i}(t/\eta)}(0)\right\rangle \\
\leq& 2L\left\|\theta_t\right\|^2 \mathbf{1}_{\|\theta_t\|\leq R} - 2K\left\|\theta_t\right\|^2 \mathbf{1}_{\|\theta_t\|>R} + K\left\|\theta_t\right\|^2 + \frac{\left\|\nabla\Phi_{\boldsymbol{i}(t/\eta)}(0)\right\|^2}{K} \\
\leq& 2(L+K)\left\|\theta_t\right\|^2 \mathbf{1}_{\|\theta_t\|\leq R} - K\left\|\theta_t\right\|^2 + \frac{\left\|\nabla\Phi_{\boldsymbol{i}(t/\eta)}(0)\right\|^2}{K} \\
\leq& 2(L+K)R^2 - K\left\|\theta_t\right\|^2 + \frac{\left\|\nabla\Phi_{\boldsymbol{i}(t/\eta)}(0)\right\|^2}{K} \\
\leq& C_\Phi - K\left\|\theta_t\right\|^2,
\end{aligned}$$

where $C_\Phi = 2(L+K)R^2 + \sup_{i\in I} \frac{\|\nabla \Phi_i(0)\|^2}{K}$.

Since we set $\theta_0 = 0$, we have

$$e^{Kt}\mathbb{E}[\|\theta_t\|^2] \leq C_\Phi d\int_0^t e^{Ks}ds,$$

which implies $\mathcal{W}_{\|.\|}(\delta_0, \nu_t^\eta) \leq \sqrt{C_\Phi K^{-1} d}$. Therefore,

$$\mathcal{W}_{\|.\|}(\delta_0, \mu^\eta) \leq \sqrt{C_\Phi K^{-1} d} + Ce^{-ct}\mathcal{W}_{\|.\|}(\delta_0, \mu^\eta).$$

The second term goes to 0 as $t \to \infty$, which yields the proof. □

B.3 Proof of Theorem 2.3

Theorem 2.3 *Under the Assumptions 2.1 and 2.2, the marginal distribution $\mu^\eta(dx)$ converges weakly to the stationary measure of $(\zeta_t)_{t\geq 0}$. In particular, we have*

$$\mathcal{W}_{\|\cdot\|}(\mu^\eta, \mu) \leq C_{\Phi,d}\eta^{c_\Phi},$$

where $c_\Phi := \frac{c}{32(L+1)+4c}$ and $C_{\Phi,d} := C_{\Phi,\theta_0=0,d} + C_d^{(1)}C$, with $C_d^{(1)} = \mathcal{O}(\sqrt{d})$.

Proof. We first bound $\mathcal{W}_{\|.\|}(\mu^\eta, \mu)$ by

$$\mathcal{W}_{\|.\|}(\mu^\eta, \mu) \leq \underbrace{\mathcal{W}_{\|.\|}(\mu^\eta, \nu_t^\eta)}_{(w1.1)} + \underbrace{\mathcal{W}_{\|.\|}(\nu_t^\eta, \nu_t)}_{(w1.2)} + \underbrace{\mathcal{W}_{\|.\|}(\nu_t, \mu)}_{(w1.3)},$$

where $\nu_0^\eta = \nu_0 = \delta_0$. From Theorem 2.2, we have

$$(w1.1) + (w1.3) \leq Ce^{-ct}\Big(\mathcal{W}_{\|.\|}(\delta_0, \mu^\eta) + \mathcal{W}_{\|.\|}(\delta_0, \mu)\Big).$$

From Lemma 4.1, we conclude that $(w1.1) + (w1.3) \leq C^{(1)}Ce^{-ct}$. For the middle term, by using Theorem 2.1 with initial value $\theta_0 = 0$, we get

$$(w1.2) \leq \mathbb{E}[\|\theta_t - \zeta_t\|] \leq C_{\Phi,0,d}e^{8(L+1)t}\eta^{\frac{1}{4}}.$$

We set $t = -\frac{1}{32(L+1)+4c}\log\eta$. Hence,

$$\mathcal{W}_{\|.\|}(\mu^\eta, \mu) \leq (w1.1) + (w1.2) + (w1.3) \leq C_{\Phi,d}\eta^{c_\Phi},$$

where $c_\Phi = \frac{c}{32(L+1)+4c}$ and $C_{\Phi,d} = C_{\Phi,0,d} + C^{(1)}C$. □

C Dissipativeness is weaker than Assumption 2.2

To bring Assumption 2.2 into the context of other analyses of SGLD algorithms, we remark that the dissipativeness assumption assumed in the non-convex analysis of SGLD-type algorithms (see e.g. Raginsky et al., 2017; Xu et al., 2018; Zou et al., 2021) is weaker than Assumption 2.2. Recall the dissipativeness assumption as the following.

Definition C.1 (Dissipativeness) *A function $f(\cdot)$ is (m, b)-dissipative if for some $m > 0$ and $b > 0$,*

$$\langle x, \nabla f(x)\rangle \geq m\|x\|^2 - b, \qquad \forall x \in \mathbb{R}^d.$$

Intuitively, dissipativeness means that the function $f(\cdot)$ grows like a quadratic function outside of a ball. The following lemma shows that Assumption 2.2 implies dissipativeness. The converse implication, however, is incorrect: after proving the lemma, we give an example of a function satisfying the dissipativeness condition, but not Assumption 2.2.

Lemma C.2 *Assume $\{\Phi_i\}_{i\in I}$ satisfy Assumption 2.2 with (R, K). Then there exists a constant $b \geq 0$, such that $\{\Phi_i\}_{i\in I}$ is $(K/2, b)$-dissipative, i.e. for any $i \in I$,*

$$\langle x, \nabla\Phi_i(x)\rangle \geq \frac{K}{2}\|x\|^2 - b.$$

Proof. When $\|x\| \leq R$,

$$\begin{aligned}\langle x, \nabla\Phi_i(x)\rangle &\geq -\|x\|\,\|\nabla\Phi_i(x)\| \\ &\geq \frac{K}{2}\|x\|^2 - \frac{K}{2}R^2 - \|x\|\,\|\nabla\Phi_i(x)\| \\ &\geq \frac{K}{2}\|x\|^2 - \frac{K}{2}R^2 - R\sup_{i\in I}\sup_{\|x\|\leq R}\|\nabla\Phi_i(x)\|.\end{aligned}$$

For $\|x\| \geq R$, by choosing $y = 0$ in Assumption 2.2, we have

$$\langle x, \nabla\Phi_i(x) - \nabla\Phi_i(0)\rangle \geq K\|x\|^2,$$

which implies

$$\langle x, \nabla\Phi_i(x)\rangle \geq K\|x\|^2 + \langle x, \nabla\Phi_i(0)\rangle.$$

By ε-Young's inequality,

$$\langle x, \nabla\Phi_i(0)\rangle \geq -\|x\| \|\nabla\Phi_i(0)\| \geq -\frac{K}{2}\|x\|^2 - \frac{1}{2K}\sup_{i\in I}\|\nabla\Phi_i(0)\|^2 .$$

Hence, for $\|x\| \geq R$, we have for any $i \in I$,

$$\begin{aligned}\langle x, \nabla\Phi_i(x)\rangle \geq & K\|x\|^2 - \frac{K}{2}\|x\|^2 - \frac{1}{2K}\sup_{i\in I}\|\nabla\Phi_i(0)\|^2 \\ \geq & \frac{K}{2}\|x\|^2 - \frac{1}{2K}\sup_{i\in I}\|\nabla\Phi_i(0)\|^2 .\end{aligned}$$

Set $b = \max\{\frac{K}{2}R^2 + R\sup_{i\in I}\sup_{\|x\|\leq R}\|\nabla\Phi_i(x)\|, \frac{1}{2K}\sup_{i\in I}\|\nabla\Phi_i(0)\|^2\}$, we get

$$\langle x, \nabla\Phi_i(x)\rangle \geq \frac{K}{2}\|x\|^2 - b.$$

□

Example C.3 *We let $X := \mathbb{R}$. We give Φ through its derivative $\Phi'(x)$. The latter is the odd function defined in the following way, for $0 \leq x \leq 2$, $\Phi'(x) = x$. In the case $x \geq 2$, there exist $n \geq 1$, such that $2^n \leq x < 2^{n+1}$, we define:*

$$\Phi'(x) = \begin{cases} 2^n, & \text{if } 2^n \leq x \leq 2^n + \log(n); \\ \frac{2^n}{2^n - \log(n)}(x - 2^n - \log(n)) + 2^n, & \text{if } 2^n + \log(n) < x < 2^{n+1}. \end{cases} \tag{C.1}$$

We can verify that $x/2 \leq \Phi'(x) \leq x$ for $x \geq 0$, hence we have $x\Phi'(x) \geq x^2/2$ and Φ satisfies dissipativeness with $(m, b) = (1/2, 0)$. However, for any $n \in \mathbb{N}$ and $x, y \in [2^n, 2^n + \log(n)]$, we have $\Phi'(x) - \Phi'(y) = 0$. Therefore, Φ does not satisfy Assumption 2.2.