Stochastic Runge-Kutta Accelerates Langevin Monte Carlo and Beyond

Xuechen Li[1, 2], Denny Wu[1, 2], Lester Mackey[3], Murat A. Erdogdu[1, 2]
University of Toronto[1], Vector Institute[2], Microsoft Research[3]
{lxuechen, dennywu, erdogdu}@cs.toronto.edu, lmackey@microsoft.com

Abstract

Sampling with Markov chain Monte Carlo methods often amounts to discretizing some continuous-time dynamics with numerical integration. In this paper, we establish the convergence rate of sampling algorithms obtained by discretizing smooth Itô diffusions exhibiting fast Wasserstein-2 contraction, based on local deviation properties of the integration scheme. In particular, we study a sampling algorithm constructed by discretizing the overdamped Langevin diffusion with the method of stochastic Runge-Kutta. For strongly convex potentials that are smooth up to a certain order, its iterates converge to the target distribution in 2-Wasserstein distance in $\tilde{\mathcal{O}}(d\epsilon^{-2/3})$ iterations. This improves upon the best-known rate for strongly log-concave sampling based on the overdamped Langevin equation using only the gradient oracle without adjustment. In addition, we extend our analysis of stochastic Runge-Kutta methods to uniformly dissipative diffusions with possibly non-convex potentials and show they achieve better rates compared to the Euler-Maruyama scheme in terms of the dependence on tolerance ϵ. Numerical studies show that these algorithms lead to better stability and lower asymptotic errors.

1 Introduction

Sampling from a probability distribution is a fundamental problem that arises in machine learning, statistics, and optimization. In many situations, the goal is to obtain samples from a target distribution given only the unnormalized density [2, 27, 40]. A prominent approach to this problem is the method of Markov chain Monte Carlo (MCMC), where an ergodic Markov chain is simulated so that iterates converge exactly or approximately to the distribution of interest [43, 2].

MCMC samplers based on numerically integrating continuous-time dynamics have proven very useful due to their ability to accommodate a stochastic gradient oracle [65]. Moreover, when used as optimizations algorithms, these methods can deliver strong theoretical guarantees in non-convex settings [50]. A popular example in this regime is the unadjusted Langevin Monte Carlo (LMC) algorithm [51]. Fast mixing of LMC is inherited from exponential Wasserstein decay of the Langevin diffusion, and numerical integration using the Euler-Maruyama scheme with a sufficiently small step size ensures the Markov chain tracks the diffusion. Asymptotic guarantees of this algorithm are well-studied [51, 26, 42], and non-asymptotic analyses specifying explicit constants in convergence bounds were recently conducted [14, 11, 18, 7, 20, 9].

To the best of our knowledge, the best known rate of LMC in 2-Wasserstein distance is due to Durmus and Moulines [18] – $\tilde{\mathcal{O}}(d\epsilon^{-1})$ iterations are required to reach ϵ accuracy to d-dimensional target distributions with strongly convex potentials under the additional Lipschitz Hessian assumption, where $\tilde{\mathcal{O}}$ hides insubstantial poly-logarithmic factors. Due to its simplicity and well-understood theoretical properties, LMC and its derivatives have found numerous applications in statistics and machine learning [65, 15]. However, from the numerical integration point of view, the Euler-Maruyama

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

scheme is usually less preferred for many problems due to its inferior stability compared to implicit schemes [1] and large integration error compared to high-order schemes [46].

In this paper, we study the convergence rate of MCMC samplers devised from discretizing Itô diffusions with exponential Wasserstein-2 contraction. Our result provides a general framework for establishing convergence rates of existing numerical schemes in the SDE literature when used as sampling algorithms. In particular, we establish *non-asymptotic* convergence bounds for sampling with *stochastic Runge-Kutta* (SRK) methods. For strongly convex potentials, iterates of a variant of SRK applied to the overdamped Langevin diffusion has a convergence rate of $\tilde{\mathcal{O}}(d\epsilon^{-2/3})$. Similar to LMC, the algorithm only queries the gradient oracle of the potential during each update and improves upon the best known rate of $\tilde{\mathcal{O}}(d\epsilon^{-1})$ for strongly log-concave sampling based on the overdamped Langevin diffusion without Metropolis adjustment, under the mild extra assumption that the potential is smooth up to the third order. In addition, we extend our analysis to *uniformly dissipative* diffusions, which enables sampling from non-convex potentials by choosing a non-constant diffusion coefficient. We study a different variant of SRK and obtain the convergence rate of $\tilde{\mathcal{O}}(d^{3/4}m^2\epsilon^{-1})$ for general Itô diffusions, where m is the dimensionality of the Brownian motion. This improves upon the convergence rate of $\tilde{\mathcal{O}}(d\epsilon^{-2})$ for the Euler-Maruyama scheme in terms of the tolerance ϵ, while potentially trading off dimension dependence.

Our contributions can be summarized as follows:

- We provide a broadly applicable theorem for establishing convergence rates of sampling algorithms based on discretizing Itô diffusions exhibiting exponential Wasserstein-2 contraction to the target invariant measure. The convergence rate is explicitly expressed in terms of the contraction rate of the diffusion and local properties of the numerical scheme, both of which can be easily derived.
- We show for strongly convex potentials, a variant of SRK applied to the overdamped Langevin diffusion achieves the improved convergence rate of $\tilde{\mathcal{O}}(d\epsilon^{-2/3})$ by accessing only the gradient oracle, under mild additional smoothness conditions on the potential.
- We establish the convergence rate of a different variant of SRK applied to uniformly dissipative diffusions. By choosing an appropriate diffusion coefficient, we show the corresponding algorithm can sample from certain non-convex potentials and achieves the rate of $\tilde{\mathcal{O}}(d^{3/4}m^2\epsilon^{-1})$.
- We provide examples and numerical studies of sampling from both convex and non-convex potentials with SRK methods and show they lead to better stability and lower asymptotic errors.

1.1 Additional Related Work

High-Order Schemes. Numerically solving SDEs has been a research area for decades [46, 32]. We refer the reader to [3] for a review and to [32] for technical foundations. Chen et al. [5] studied the convergence of smooth functions evaluated at iterates of sampling algorithms obtained by discretizing the Langevin diffusion with high-order numerical schemes. Their focus was on convergence rates of function evaluations under a stochastic gradient oracle using asymptotic arguments. This convergence assessment pertains to analyzing numerical schemes in the weak sense. By contrast, we establish non-asymptotic convergence bounds in the 2-Wasserstein metric, which covers a broader class of functions by the Kantorovich duality [28, 62], and our techniques are based on the mean-square convergence analysis of numerical schemes. Notably, a key ingredient in the proofs by Chen et al. [5], i.e. moment bounds in the guise of a Lyapunov function argument, is assumed without justification, whereas we derive this formally and obtain convergence bounds with explicit dimension dependent constants. Durmus et al. [19] considered convergence of function evaluations of schemes obtained using Richardson-Romberg extrapolation. Sabanis and Zhang [53] introduced a numerical scheme that queries the gradient of the Laplacian based on an integrator that accommodates superlinear drifts [54]. In particular, for potentials with a Lipschitz gradient, they obtained the convergence rate of $\tilde{\mathcal{O}}(d^{4/3}\epsilon^{-2/3})$. In optimization, high-order ordinary differential equation (ODE) integration schemes were introduced to discretize a second-order ODE and achieved acceleration [68].

Non-Convex Learning. The convergence analyses of sampling using the overdamped and underdamped Langevin diffusion were extended to the non-convex setting [9, 39]. For the Langevin diffusion, the most common assumption on the potential is strong convexity outside a ball of finite radius, in addition to Lipschitz smoothness and twice differentiability [9, 38, 39]. More generally, Vempala and Wibisono [61] showed that convergence in the KL divergence of LMC can be derived assuming a log-Sobolev inequality of the target measure with a positive log-Sobolev constant holds. For general Itô diffusions, the notion of *distant dissipativity* [30, 22, 23] is used to study convergence

Table 1: Convergence rates in W_2 for algorithms sampling from strongly convex potentials by discretizing the overdamped Langevin diffusion. "Oracle" refers to highest derivative used in the update. "Smoothness" refers to Lipschitz conditions. Note that faster algorithms exist by discretizing high-order Langevin equations [13, 8, 9, 47, 56] or applying Metropolis adjustment [21, 6].

Method	Convergence Rate	Oracle	Smoothness
Euler-Maruyama [18]	$\tilde{\mathcal{O}}(d\epsilon^{-2})$	1st order	gradient
Euler-Maruyama [18]	$\tilde{\mathcal{O}}(d\epsilon^{-1})$	1st order	gradient & Hessian
Ozaki's [11] [1]	$\tilde{\mathcal{O}}(d\epsilon^{-1})$	2nd order	gradient & Hessian
Tamed Order 1.5 [53] [2]	$\tilde{\mathcal{O}}(d^{4/3}\epsilon^{-2/3})$	3rd order	1st to 3rd derivatives
Stochastic Runge-Kutta (this work)	$\tilde{\mathcal{O}}(d\epsilon^{-2/3})$	1st order	1st to 3rd derivatives

to target measures with non-convex potentials in the 1-Wasserstein distance. Different from these works, our non-convex convergence analysis, due to conducted in W_2, requires the slightly stronger uniform dissipativity condition [30]. In optimization, non-asymptotic results for stochastic gradient Langevin dynamics and its variants have been established for non-convex objectives [50, 67, 24, 69].

Notation. We denote the p-norm of a real vector $x \in \mathbb{R}^d$ by $\|x\|_p$. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we denote its ith derivative by $\nabla^i f(x)$ and its Laplacian by $\Delta f = \sum_{i=1}^d \partial^2 f_i(x)/\partial x_i^2$. For a vector-valued function $g : \mathbb{R}^d \to \mathbb{R}^m$, we denote its vector Laplacian by $\vec{\Delta}(g)$, i.e. $\vec{\Delta}(g)_i = \Delta(g_i)$. For a tensor $T \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_m}$, we define its operator norm recursively as $\|T\|_{\mathrm{op}} = \sup_{\|u\|_2 \le 1} \|T[u]\|_{\mathrm{op}}$, where $T[u]$ denotes the tensor-vector product. For f sufficiently differentiable, we denote the Lipschitz and polynomial coefficients of its ith order derivative as

$$\mu_0(f) = \sup_{x \in \mathbb{R}^d} \|f(x)\|_{\mathrm{op}}, \quad \mu_i(f) = \sup_{x,y \in \mathbb{R}^d, x \neq y} \frac{\left\|\nabla^{i-1} f(x) - \nabla^{i-1} f(y)\right\|_{\mathrm{op}}}{\|x-y\|_2}, \text{ and } \pi_{i,n}(f) = \sup_{x \in \mathbb{R}^d} \frac{\left\|\nabla^{i-1} f(x)\right\|_{\mathrm{op}}^n}{1+\|x\|_2^n},$$

with the exception in Theorem 3, where $\pi_{1,n}(\sigma)$ is used for a sublinear growth condition. We denote Lipschitz and growth coefficients under the Frobenius norm $\|\cdot\|_{\mathrm{F}}$ as $\mu_1^{\mathrm{F}}(\cdot)$ and $\pi_{1,n}^{\mathrm{F}}(\cdot)$, respectively.

Coupling and Wasserstein Distance. We denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field of $\mathbb{R}^d$. Given probability measures ν and ν' on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we define a coupling (or transference plan) ζ between ν and ν' as a probability measure on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ such that $\zeta(A \times \mathbb{R}^d) = \nu(A)$ and $\zeta(\mathbb{R}^d \times A) = \nu'(A)$ for all $A \in \mathcal{B}(\mathbb{R}^d)$. Let couplings$(\nu, \nu')$ denote the set of all such couplings. We define the 2-Wasserstein distance between a pair of probability measures ν and ν' as

$$W_2(\nu, \nu') = \inf_{\zeta \in \mathrm{couplings}(\nu, \nu')} \left(\int \|x-y\|_2^2 \, \mathrm{d}\zeta(\nu, \nu') \right)^{1/2}.$$

2 Sampling with Discretized Diffusions

We study the problem of sampling from a target distribution $p(x)$ with the help of a candidate Itô diffusion [37, 44] given as the solution to the following stochastic differential equation (SDE):

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}B_t, \quad \text{with} \quad X_0 = x_0, \tag{1}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d \times m}$ are termed as the drift and diffusion coefficients, respectively. Here, $\{B_t\}_{t \ge 0}$ is an m-dimensional Brownian motion adapted to the filtration $\{\mathcal{F}_t\}_{t \ge 0}$, whose ith dimension we denote by $\{B_t^{(i)}\}_{t \ge 0}$. A candidate diffusion should be chosen so that (i) its invariant measure is the target distribution $p(x)$ and (ii) it exhibits fast mixing properties. Under mild conditions, one can design a diffusion with the target invariant measure by choosing the drift coefficient as (see e.g. [37, Thm. 2])

$$b(x) = \frac{1}{2p(x)} \langle \nabla, p(x) w(x) \rangle, \quad \text{where} \quad w(x) = \sigma(x)\sigma(x)^\top + c(x), \tag{2}$$

[1] We obtain a rate in W_2 from the discretization analysis in KL [11] via standard techniques [50, 61].

[2] Sabanis and Zhang [53] use the Frobenius norm for matrices and the Euclidean norm of Frobenius norms for 3-tensors. For a fair comparison, we convert their Lipschitz constants to be based on the operator norm.

$c(x) \in \mathbb{R}^{d\times d}$ is any skew-symmetric matrix and $\langle \nabla, \cdot\rangle$ is the divergence operator for a matrix-valued function, i.e. $\langle \nabla, w(x)\rangle_i = \sum_{j=1}^d \partial w_{i,j}(x)/\partial x_j$ for $w : \mathbb{R}^d \to \mathbb{R}^{d\times d}$. To guarantee that this diffusion has fast convergence properties, we will require certain dissipativity conditions to be introduced later. For example, if the target is the Gibbs measure of a strongly convex potential $f : \mathbb{R}^d \to \mathbb{R}$, i.e., $p(x) \propto \exp(-f(x))$, a popular candidate diffusion is the (overdamped) Langevin diffusion which is the solution to the following SDE:

$$\mathrm{d}X_t = -\nabla f(X_t)\ \mathrm{d}t + \sqrt{2}\ \mathrm{d}B_t, \quad \text{with} \quad X_0 = x_0. \tag{3}$$

It is straightforward to verify (2) for the above diffusion which implies that the target $p(x)$ is its invariant measure. Moreover, strong convexity of f implies uniform dissipativity and ensures that the diffusion achieves fast convergence.

2.1 Numerical Schemes and the Itô-Taylor Expansion

In practice, the Itô diffusion (1) (similarly (3)) cannot be simulated in continuous time and is instead approximated by a discrete-time numerical integration scheme. Owing to its simplicity, a common choice is the Euler-Maruyama (EM) scheme [32], which relies on the following update rule,

$$\tilde{X}_{k+1} = \tilde{X}_k + h\, b(\tilde{X}_k) + \sqrt{h}\, \sigma(\tilde{X}_k)\xi_{k+1}, \quad k = 0, 1, \ldots, \tag{4}$$

where h is the step size and $\xi_{k+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ is independent of $\tilde{X}_k$ for all $k \in \mathbb{N}$. The above iteration defines a Markov chain and due to discretization error, its invariant measure $\tilde{p}(x)$ is different from the target distribution $p(x)$; yet, for a sufficiently small step size, the difference between $\tilde{p}(x)$ and $p(x)$ can be characterized (see e.g. [42, Thm. 7.3]).

Analogous to ODE solvers, numerical schemes such as the EM scheme and SRK schemes are derived based on approximating the continuous-time dynamics locally. Similar to the standard Taylor expansion, Itô's lemma induces a stochastic version of the Taylor expansion of a smooth function evaluated at a stochastic process at time t. This is known as the Itô-Taylor (or Wagner-Platen) expansion [46], and one can also interpret the expansion as recursively applying Itô's lemma to terms in the integral form of an SDE. Specifically, for $g : \mathbb{R}^d \to \mathbb{R}^d$, we define the operators:

$$L(g)(x) = \nabla g(x) \cdot b(x) + \tfrac{1}{2}\textstyle\sum_{i=1}^m \nabla^2 g(x)[\sigma_i(x), \sigma_i(x)], \quad \Lambda_j(g)(x) = \nabla g(x) \cdot \sigma_j(x), \tag{5}$$

where $\sigma_i(x)$ denotes the ith column of $\sigma(x)$. Then, applying Itô's lemma to the integral form of the SDE (1) with the starting point X_0 yields the following expansion around X_0 [32, 46]:

$$\begin{aligned} X_t = &\overbrace{\underbrace{X_0 + t\, b(X_0) + \sigma(X_0)B_t}_{\text{Euler-Maruyama update}} + \textstyle\sum_{i,j=1}^m \int_0^t\int_0^s \Lambda_j(\sigma_i)(X_u)\, \mathrm{d}B_u^{(j)}\, \mathrm{d}B_u^{(i)}}^{\text{mean-square order 1.0 stochastic Runge-Kutta update}} + \textstyle\int_0^t\int_0^s L(b)(X_u)\, \mathrm{d}u\, \mathrm{d}s \\ &+ \textstyle\sum_{i=1}^m \int_0^t\int_0^s L(\sigma_i)(X_u)\ \mathrm{d}u\ \mathrm{d}B_s^{(i)} \ + \ \sum_{i=1}^m \int_0^t\int_0^s \Lambda_i(b)(X_u)\ \mathrm{d}B_u^{(i)}\ \mathrm{d}s. \end{aligned} \tag{6}$$

The expansion justifies the update rule of the EM scheme, since the discretization is nothing more than taking the first three terms on the right hand side of (6). Similarly, a mean-square order 1.0 SRK scheme for general Itô diffusions – introduced in Section 4.2 – approximates the first four terms. In principle, one may recursively apply Itô's lemma to terms in the expansion to obtain a more fine-grained approximation. However, the appearance of non-Gaussian terms in the guise of iterated Brownian integrals presents a challenge for simulation. Nevertheless, it is clear that the above SRK scheme will be a more accurate local approximation than the EM scheme, due to accounting more terms in the expansion. As a result, the local deviation between the continuous-time process and Markov chain will be smaller. We characterize this property of a numerical scheme as follows.

Definition 2.1 (Uniform Local Deviation Orders)**.** *Let $\{\tilde{X}_k\}_{k\in\mathbb{N}}$ denote the discretization of an Itô diffusion $\{X_t\}_{t\geq 0}$ based on a numerical integration scheme with constant step size h, and its governing Brownian motion $\{B_t\}_{t\geq 0}$ be adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$. Suppose $\{X_s^{(k)}\}_{s\geq 0}$ is another instance of the same diffusion starting from $\tilde{X}_{k-1}$ at $s = 0$ and governed by the Brownian motion $\{B_{s+h(k-1)}\}_{s\geq 0}$. Then, the numerical integration scheme has local deviation $D_h^{(k)} = \tilde{X}_k - X_h^{(k)}$ with uniform orders (p_1, p_2) if*

$$\mathcal{E}_k^{(1)} = \mathbb{E}\left[\mathbb{E}\big[\|D_h^{(k)}\|_2^2 | \mathcal{F}_{t_{k-1}}\big]\right] \leq \lambda_1 h^{2p_1}, \quad \mathcal{E}_k^{(2)} = \mathbb{E}\left[\big\|\mathbb{E}\big[D_h^{(k)} | \mathcal{F}_{t_{k-1}}\big]\big\|_2^2\right] \leq \lambda_2 h^{2p_2}, \tag{7}$$

for all $k \in \mathbb{N}_+$ and $0 \leq h < C_h$, where constants $0 < \lambda_1, \lambda_2, C_h < \infty$. We say that $\mathcal{E}_k^{(1)}$ and $\mathcal{E}_k^{(2)}$ are the local mean-square deviation and the local mean deviation at iteration k, respectively.

In the SDE literature, local deviation orders are defined to derive the mean-square order (or strong order) of numerical schemes [46], where the mean-square order is defined as the maximum half-integer p such that $\mathbb{E}[\|X_{t_k} - \tilde{X}_k\|_2^2] \leq Ch^{2p}$ for a constant C independent of step size h and all $k \in \mathbb{N}$ where $t_k < T$. Here, $\{X_t\}_{t\geq 0}$ is the continuous-time process, $\tilde{X}_k(k = 0, 1, \dots)$ is the Markov chain with the same Brownian motion as the continuous-time process, and $T < \infty$ is the terminal time. The key difference between our definition of *uniform* local deviation orders and local deviation orders in the SDE literature is we require the extra step of ensuring the expectations of $\varepsilon_k^{(1)}$ and $\varepsilon_k^{(2)}$ are bounded across all iterations, instead of merely requiring the two deviation variables to be bounded by a function of the previous iterate.

3 Convergence Rates of Numerical Schemes for Sampling

We present a user-friendly and broadly applicable theorem that establishes the convergence rate of a diffusion-based sampling algorithm. We develop our explicit bounds in the 2-Wasserstein distance based on two crucial steps. We first verify that the candidate diffusion exhibits exponential Wasserstein-2 contraction and thereafter compute the uniform local deviation orders of the scheme.

Definition 3.1 (Wasserstein-2 rate)**.** *A diffusion X_t has Wasserstein-2 (W_2) rate $r : \mathbb{R}_{\geq 0} \to \mathbb{R}$ if for two instances of the diffusion X_t initiated respectively from x and y, we have*

$$W_2(\delta_x P_t, \delta_y P_t) \leq r(t) \|x - y\|_2, \quad \textit{for all } x, y \in \mathbb{R}^d, t \geq 0,$$

where $\delta_x P_t$ denotes the distribution of the diffusion X_t starting from x. Moreover, if $r(t) = e^{-\alpha t}$ for some $\alpha > 0$, then we say the diffusion has exponential W_2-contraction.

The above condition guarantees fast mixing of the sampling algorithm. For Itô diffusions, uniform dissipativity suffices to ensure exponential W_2-contraction $r(t) = e^{-\alpha t}$ [24, Prop. 3.3].

Definition 3.2 (Uniform Dissipativity)**.** *A diffusion defined by (1) is α-uniformly dissipative if*

$$\langle b(x) - b(y), x - y\rangle + \tfrac{1}{2}\|\sigma(x) - \sigma(y)\|_{\mathrm{F}}^2 \leq -\alpha \|x - y\|_2^2, \quad \textit{for all } x, y \in \mathbb{R}^d.$$

For Itô diffusions with a constant diffusion coefficient, uniform dissipativity is equivalent to one-sided Lipschitz continuity of the drift with coefficient -2α. In particular, for the overdamped Langevin diffusion (3), this reduces to strong convexity of the potential. Moreover, for this special case, exponential W_2-contraction of the diffusion and strong convexity of the potential are equivalent [4]. We will ultimately verify uniform dissipativity for the candidate diffusions, but we first use W_2-contraction to derive the convergence rate of a diffusion-based sampling algorithm.

Theorem 1 (W_2-rate of a numerical scheme)**.** *For a diffusion with invariant measure ν^*, exponentially contracting W_2-rate $r(t) = e^{-\alpha t}$, and Lipschitz drift and diffusion coefficients, suppose its discretization based on a numerical integration scheme has uniform local deviation orders (p_1, p_2) where $p_1 \geq 1/2$ and $p_2 \geq p_1 + 1/2$. Let ν_k be the measure associated with the Markov chain obtained from the discretization after k steps starting from the dirac measure $\nu_0 = \delta_{x_0}$. Then, for constant step size h satisfying*

$$h < 1 \wedge C_h \wedge \frac{1}{2\alpha} \wedge \frac{1}{8\mu_1(b)^2 + 8\mu_1^{\mathrm{F}}(\sigma)^2},$$

where C_h is the step size constraint for obtaining the uniform local deviation orders, we have

$$W_2(\nu_k, \nu^*) \leq \left(1 - \frac{\alpha h}{2}\right)^k W_2(\nu_0, \nu^*) + \left(\frac{8\,(16\mu_1(b)\lambda_1 + \lambda_2)}{\alpha^2} + \frac{2\lambda_1}{\alpha}\right)^{1/2} h^{p_1 - 1/2}. \tag{8}$$

Moreover, if $p_1 > 1/2$ and the step size additionally satisfies

$$h < \left(\frac{2}{\epsilon}\sqrt{\frac{64(16\lambda_1\mu_1(b) + \lambda_2)}{\alpha^2} + \frac{2\lambda_1}{\alpha}}\right)^{-1/(p_1 - 1/2)},$$

then $W_2(\nu_k, \nu^*)$ converges in $\tilde{\mathcal{O}}(\epsilon^{-1/(p_1-1/2)})$ iterations within a sufficiently small positive error ϵ.

Theorem 1 directly translates mean-square order results in the SDE literature to convergence rates of sampling algorithms in W_2. The proof deferred to Appendix A follows from an inductive argument over the local deviation at each step (see e.g. [46]), and the convergence is provided by the exponential W_2-contraction of the diffusion. To invoke the theorem and obtain convergence rates of a sampling algorithm, it suffices to (i) show that the candidate diffusion is uniformly dissipative and (ii) derive the local deviation orders for the underlying discretization. Below, we demonstrate this on both the overdamped Langevin and general Itô diffusions when the EM scheme is used for discretization, as well as the underdamped Langevin diffusion when a linearization is used for discretization [8]. For these schemes, local deviation orders are either well-known or straightforward to derive. Thus, convergence rates for corresponding sampling algorithms can be easily obtained using Theorem 1.

Example 1. Consider sampling from a target distribution whose potential is strongly convex using the overdamped Langevin diffusion (3) discretized by the EM scheme. The scheme has local deviation of orders $(1.5, 2.0)$ for Itô diffusions with constant diffusion coefficients and drift coefficients that are sufficiently smooth [3] (see e.g. [46, Sec. 1.5.4]). Since the potential is strongly convex, the Langevin diffusion is uniformly dissipative and achieves exponential W_2-contraction [18, Prop. 1]. Elementary algebra shows that Markov chain moments are bounded [24, Lem. A.2]. Therefore, Theorem 1 implies that the rate of the sampling is $\tilde{\mathcal{O}}(d\epsilon^{-1})$, where the dimension dependence can be extracted from the explicit bound. This recovers the result by Durmus and Moulines [18, Thm. 8].

Example 2. If a general Itô diffusion (1) with Lipschitz smooth drift and diffusion coefficients is used for the sampling task, local deviation orders of the EM scheme reduce to $(1.0, 1.5)$ due to the approximation of the diffusion term [46] – this term is exact for Langevin diffusion. If we further have uniform dissipativity, it can be shown that Markov chain moments are bounded [24, Lem. A.2]. Hence, Theorem 1 concludes that the convergence rate is $\tilde{\mathcal{O}}(d\epsilon^{-2})$. We note that for the diffusion coefficient, we use the Frobenius norm for the Lipschitz and growth constants which potentially hides dimension dependence factors. The dimension dependence worsens if one were to convert all bounds to be based on the operator norm using the pessimistic inequality $\|\sigma(x)\|_{\mathrm{F}} \le (d^{1/2} + m^{1/2}) \|\sigma(x)\|_{\mathrm{op}}$. Appendix D provides a convergence bound with explicit constants.

Example 3. Consider sampling from a target distribution whose potential is strongly convex using the underdamped Langevin diffusion:

$$\mathrm{d}X_t = V_t \,\mathrm{d}t, \quad \mathrm{d}V_t = -\gamma V_t \,\mathrm{d}t - u \nabla f(X_t) \,\mathrm{d}t + \sqrt{2\gamma u}\,\mathrm{d}B_t.$$

Cheng et al. [8] show that the continuous-time process $\{(X_t, X_t + V_t)\}_{t\ge 0}$ exhibits exponential W_2-contraction when the coefficients γ and u are appropriately chosen [8, Thm. 5]. Moreover, the scheme devised by linearizing the degenerate SDE has uniform local deviation orders $(1.5, 2.0)$ [4] [8, Thm. 9]. Theorem 1 implies that the convergence rate is $\mathcal{O}(d^{1/2}\epsilon^{-1})$, where the dimension dependence is extracted from explicit bounds. This recovers the result by Cheng et al. [8, Thm. 1].

While computing the local deviation orders of a numerical scheme for a single step is often straightforward, it is not immediately clear how one might verify them uniformly for each iteration. This requires a uniform bound on moments of the Markov chain defined by the numerical scheme. As our second principal contribution, we explicitly bound the Markov chain moments of SRK schemes which, combined with Theorem 1, leads to improved rates by only accessing the first-order oracle.

4 Sampling with Stochastic Runge-Kutta and Improved Rates

We show that convergence rates of sampling can be significantly improved if an Itô diffusion with exponential W_2-contraction is discretized using SRK methods. Compared to the EM scheme, SRK schemes we consider query the same order oracle and improve on the deviation orders.

Theorem 1 hints that one may expect the convergence rate of sampling to improve as more terms of the Itô-Taylor expansion are incorporated in the numerical integration scheme. However, in practice, a challenge for simulation is the appearance of non-Gaussian terms in the form of iterated Itô integrals. Fortunately, since the overdamped Langevin diffusion has a constant diffusion coefficient, efficient SRK methods can still be applied to accelerate convergence.

[3] In fact, it suffices to ensure the drift is three-times differentiable with Lipschitz gradient and Hessian.

[4] Cheng et al. [8] derive the uniform local mean-square deviation order. Jensen's inequality implies that the local mean deviation is of the same uniform order. This entails uniform local deviation orders are $(2.0, 2.0)$ and hence also $(1.5, 2.0)$ when step size constraint $C_h \le 1$; note $p_2 \ge p_1 + 1/2$ is required to invoke Theorem 1.

4.1 Sampling from Strongly Convex Potentials with the Langevin Diffusion

We provide a non-asymptotic analysis for integrating the overdamped Langevin diffusion based on a mean-square order 1.5 SRK scheme for SDEs with constant diffusion coefficients [46]. We refer to the sampling algorithm as SRK-LD. Specifically, given a sample from the previous iteration $\tilde{X}_k$,

$$\tilde{H}_1 = \tilde{X}_k + \sqrt{2h}\left[\left(\frac{1}{2} + \frac{1}{\sqrt{6}}\right)\xi_{k+1} + \frac{1}{\sqrt{12}}\eta_{k+1}\right],$$
$$\tilde{H}_2 = \tilde{X}_k - h\nabla f(\tilde{X}_k) + \sqrt{2h}\left[\left(\frac{1}{2} - \frac{1}{\sqrt{6}}\right)\xi_{k+1} + \frac{1}{\sqrt{12}}\eta_{k+1}\right],$$
$$\tilde{X}_{k+1} = \tilde{X}_k - \frac{h}{2}\left(\nabla f(\tilde{H}_1) + \nabla f(\tilde{H}_2)\right) + \sqrt{2h}\xi_{k+1}, \tag{9}$$

where h is the step size and $\xi_{k+1}, \eta_{k+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ are independent of $\tilde{X}_k$ for all $k \in \mathbb{N}$. We refer the reader to [46, Sec. 1.5] for a detailed derivation of the scheme and other background information.

Theorem 2 (SRK-LD). *Let ν^* be the target distribution with a strongly convex potential that is four-times differentiable with Lipschitz continuous first three derivatives. Let ν_k be the distribution of the kth Markov chain iterate defined by (9) starting from the dirac measure $\nu_0 = \delta_{x_0}$. Then, for a sufficiently small step size, 1.5 SRK scheme has uniform local deviation orders $(2.0, 2.5)$, and $W_2(\nu_k, \nu^*)$ converges within ϵ error in $\tilde{\mathcal{O}}(d\epsilon^{-2/3})$ iterations.*

The proof of this theorem is given in Appendix B where we provide explicit constants. The basic idea of the proof is to match up the terms in the Itô-Taylor expansion to terms in the Taylor expansion of the discretization scheme. However, extreme care is needed to ensure a tight dimension dependence.

***Remark*.** For large-scale Bayesian inference, computing the full gradient of the potential can be costly. Fortunately, SRK-LD can be easily adapted to use an unbiased stochastic oracle, provided queries of the latter have a variance not overly large. We provide an informal discussion in Appendix E.

We emphasize that the 1.5 SRK scheme (9) only queries the gradient of the potential and improves the best available W_2-rate of LMC in the same setting from $\tilde{\mathcal{O}}(d\epsilon^{-1})$ to $\tilde{\mathcal{O}}(d\epsilon^{-2/3})$, with merely two extra gradient evaluations per iteration. Remarkably, the dimension dependence stays the same.

4.2 Sampling from Non-Convex Potentials with Itô Diffusions

For the Langevin diffusion, the conclusions of Theorem 1 only apply to distributions with strongly convex potentials, as exponential W_2-contraction of the Langevin diffusion is equivalent to strong convexity of the potential. This shortcoming can be addressed using a non-constant diffusion coefficient which allows us to sample from non-convex potentials using uniformly dissipative candidate diffusions. Below, we use a mean-square order 1.0 SRK scheme for general diffusions [52] and achieve an improved convergence rate compared to sampling with the EM scheme.

We refer to the sampling algorithm as SRK-ID, which has the following update rule:

$$\tilde{H}_1^{(i)} = \tilde{X}_k + \textstyle\sum_{j=1}^m \sigma_l(\tilde{X}_k)\dfrac{I_{(j,i)}}{\sqrt{h}}, \qquad \tilde{H}_2^{(i)} = \tilde{X}_k - \textstyle\sum_{j=1}^m \sigma_l(\tilde{X}_k)\dfrac{I_{(j,i)}}{\sqrt{h}},$$
$$\tilde{X}_{k+1} = \tilde{X}_k + hb(\tilde{X}_k) + \textstyle\sum_{i=1}^m \sigma_i(\tilde{X}_k) I_{(i)} + \dfrac{\sqrt{h}}{2}\textstyle\sum_{i=1}^m\left(\sigma_i(\tilde{H}_1^{(i)}) - \sigma_i(\tilde{H}_2^{(i)})\right), \tag{10}$$

where $I_{(i)} = \int_{t_k}^{t_{k+1}} \mathrm{d}B_s^{(i)}$, $I_{(j,i)} = \int_{t_k}^{t_{k+1}}\int_{t_k}^{s} \mathrm{d}B_u^{(j)}\, \mathrm{d}B_s^{(i)}$. We note that schemes of higher order exist for general diffusions, but they typically require advanced approximations of iterated Itô integrals of the form $\int_0^{t_0}\cdots\int_0^{t_{n-1}} \mathrm{d}B_{t_n}^{(k_n)}\cdots \mathrm{d}B_{t_1}^{(k_1)}$.

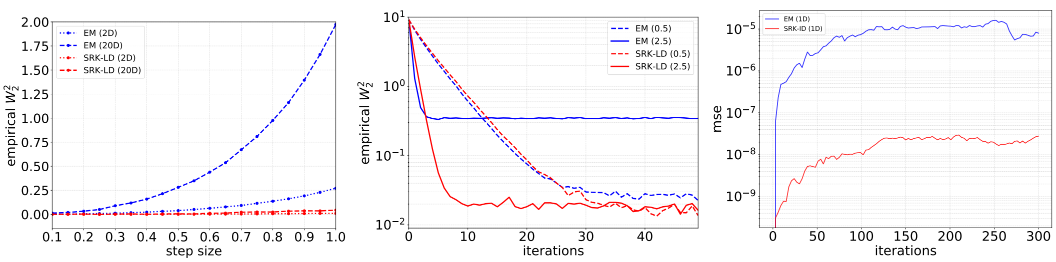
Theorem 3 (SRK-ID). *For a uniformly dissipative diffusion with invariant measure ν_*, Lipschitz drift and diffusion coefficients that have Lipschitz gradients, assume that the diffusion coefficient further satisfies the sublinear growth condition $\|\sigma(x)\|_{\mathrm{op}} \le \pi_{1,1}(\sigma)\left(1 + \|x\|_2^{1/2}\right)$ for all $x \in \mathbb{R}^d$. Let ν_k be the distribution of the kth Markov chain iterate defined by (10) starting from the dirac measure $\nu_0 = \delta_{x_0}$. Then for a sufficiently small step size, iterates of the 1.0 SRK scheme have uniform local deviation orders $(1.5, 2.0)$, and $W_2(\nu_k, \nu^*)$ converges within ϵ error in $\tilde{\mathcal{O}}(d^{3/4}m^2\epsilon^{-1})$ iterations.*

The proof is given in Appendix C where we present explicit constants. We note that the dimension dependence in this case is only better than that of EM due to the extra growth condition on the diffusion. The extra m-dependence comes from the $2m$ evaluations of the diffusion coefficient at $\tilde{H}_1^{(i)}$ and $\tilde{H}_2^{(i)}$ $(i = 1, \ldots, m)$. In the above theorem, we use the Frobenius norm for the Lipschitz and growth constants for the diffusion coefficient which potentially hides dimension dependence. One may convert all bounds to be based on the operator norm with our constants given in the Appendix.

In practice, accurately simulating both the iterated Itô integrals $I_{(j,i)}$ and the Brownian motion increments $I_{(i)}$ simultaneously is difficult. We comment on two possible approximations based on truncating an infinite series in Appendix H.2.

5 Examples and Numerical Studies

We provide examples of our theory and numerical studies showing SRK methods achieve lower asymptotic errors, are stable under large step sizes, and hence converge faster to a prescribed tolerance. We sample from strongly convex potentials with SRK-LD and non-convex potentials with SRK-ID. Since our theory is in W_2, we compare with EM on W_2 and mean squared error (MSE) between iterates of the Markov chain and the target. We do not compare to schemes that require computing derivatives of the drift and diffusion coefficients. Since directly computing W_2 is infeasible, we estimate it using samples instead. However, sample-based estimators have a bias of order $\Omega(n^{-1/d})$ [64], so we perform a heuristic correction whose description is in Appendix G.



(a) Gaussian mixture (b) Bayesian logistic regression (c) non-convex potential

Figure 1: (a) Estimated asymptotic error against step size. (b) Estimated error against number of iterations. (c) MSE against number of iterations. Legends of (a) and (c) denote "scheme (dimensionality)". Legend of (b) denotes "scheme (step size)".

5.1 Strongly Convex Potentials

Gaussian Mixture. We consider sampling from a multivariate Gaussian mixture with density

$$\pi(\theta) \propto \exp\left(-\tfrac{1}{2}\|\theta - a\|_2^2\right) + \exp\left(-\tfrac{1}{2}\|\theta + a\|_2^2\right), \quad \theta \in \mathbb{R}^d,$$

where $a \in \mathbb{R}^d$ is a parameter that measures the separation of two modes. The potential is strongly convex when $\|a\|_2 < 1$ and has Lipschitz gradient and Hessian [11]. Moreover, one can verify that its third derivative is also Lipschitz.

Bayesian Logistic Regression. We consider Bayesian logistic regression (BLR) [11]. Given data samples $\mathrm{X} = \{x_i\}_{i=1}^n \in \mathbb{R}^{n\times d}$, $\mathrm{Y} = \{y_i\}_{i=1}^n \in \mathbb{R}^n$, and parameter $\theta \in \mathbb{R}^d$, logistic regression models the Bernoulli conditional distribution with probability $\Pr(y_i = 1|x_i) = 1/(1 + \exp(-\theta^\top x_i))$. We place a Gaussian prior on θ with mean zero and covariance proportional to Σ_{X}^{-1}, where $\Sigma_{\mathrm{X}} = \mathrm{X}^\top\mathrm{X}/n$ is the sample covariance matrix. We sample from the posterior density

$$\pi(\theta) \propto \exp(-f(\theta)) = \exp\left(\mathrm{Y}^\top\mathrm{X}\theta - \textstyle\sum_{i=1}^n \log(1 + \exp(-\theta^\top x_i)) - \frac{\alpha}{2}\|\Sigma_{\mathrm{X}}^{1/2}\theta\|_2^2\right).$$

The potential is strongly convex and has Lipschitz gradient and Hessian [11]. One can also verify that it has a Lipschitz third derivative.

To obtain the potential, we generate data from the model with the parameter $\theta_* = \mathbf{1}_d$ following [11, 21]. To obtain each x_i, we sample a vector whose components are independently drawn from the

Rademacher distribution and normalize it by the Frobenius norm of the sample matrix X times $d^{-1/2}$. Note that our normalization scheme is different from that adopted in [11, 21], where each x_i is normalized by its Euclidean norm. We sample the corresponding y_i from the model and fix the regularizer $\alpha = 0.3d/\pi^2$.

To characterize the true posterior, we sample 50k particles driven by EM with a step size of 0.001 until convergence. We subsample from these particles 5k examples to represent the true posterior each time we intend to estimate squared W_2. We monitor the kernel Stein discrepancy [5] (KSD) [29, 10, 36] using the inverse multiquadratic kernel [29] with hyperparameters $\beta = -1/2$ and $c = 1$ to measure the distance between the 100k particles and the true posterior. We confirm that these particles faithfully approximate the true posterior with the squared KSD being less than 0.002 in all settings.

When sampling from a Gaussian mixture and the posterior of BLR, we observe that SRK-LD leads to a consistent improvement in the asymptotic error compared to the EM scheme when the same step size is used. In particular, Figure 1 (a) plots the estimated asymptotic error in squared W_2 of different step sizes for 2D and 20D Gaussian mixture problems and shows that SRK-LD is surprisingly stable for exceptionally large step sizes. Figure 1 (b) plots the estimated error in squared W_2 as the number of iterations increases for 2D BLR. We include additional results on problems in 2D and 20D with error estimates in squared W_2 and the energy distance [58] along with a wall time analysis in Appendix H.

5.2 Non-Convex Potentials

We consider sampling from the non-convex potential

$$f(x) = \left(\beta + \|x\|_2^2\right)^{1/2} + \gamma \log\left(\beta + \|x\|_2^2\right), \quad x \in \mathbb{R}^d,$$

where $\beta, \gamma > 0$ are scalar parameters of the distribution. The corresponding density is a simplified abstraction for the posterior distribution of Student's t regression with a pseudo-Huber prior [30]. One can verify that when $\beta + \|x\|_2^2 < 1$ and $(4\gamma + 1)\,\|x\|_2^2 < (2\gamma + 1)\sqrt{\beta + \|x\|_2^2}$, the Hessian has a negative eigenvalue. The candidate diffusion, where the drift coefficient is given by (2) and diffusion coefficient $\sigma(x) = g(x)^{1/2} I_d$ with $g(x) = \left(\beta + \|x\|_2^2\right)^{1/2}$, is uniformly dissipative if $\frac{1}{2} - |\gamma - \frac{1}{2}| \frac{2}{\beta^{1/2}} - \frac{d}{8\beta^{1/2}} > 0$. Indeed, one can verify that $\mu_1(g) \le 1$, $\mu_2(g) \le \frac{2}{\beta^{1/2}}$, and $\mu_1(\sigma) \le \frac{1}{2\beta^{1/4}}$. Therefore,

$$\begin{aligned}\langle b(x) - b(y), x - y\rangle + \tfrac{1}{2}\|\sigma(x) - \sigma(y)\|_{\mathrm{F}}^2 &\le -\left(\tfrac{1}{2} - |\gamma - \tfrac{1}{2}|\mu_2(g) - \tfrac{d}{2}\mu_1(\sigma)^2\right)\|x - y\|_2^2,\\ &\le -\left(\tfrac{1}{2} - |\gamma - \tfrac{1}{2}|\tfrac{2}{\beta^{1/2}} - \tfrac{d}{8\beta^{1/2}}\right)\|x - y\|_2^2.\end{aligned}$$

Moreover, b and σ have Lipschitz first two derivatives, and the latter satisfies the sublinear growth condition in Theorem 3.

To study the behavior of SRK-ID, we simulate using both SRK-ID and EM. For both schemes, we simulate with a step size of 10^{-3} initiated from the same 50k particles approximating the stationary distribution obtained by simulating EM with a step size of 10^{-6} until convergence. We compute the MSE between the continuous-time process and the Markov chain with the same Brownian motion for 300 iterations when we observe the MSE curve plateaus. We approximate the continuous-time process by simulating using the EM scheme with a step size of 10^{-6} similar to the setting in [52]. To obtain final results, we average across ten independent runs. We note that the MSE upper bounds W_2 due to the latter being an infimum over all couplings. Hence, the MSE value serves as an indication of the convergence performance in W_2.

Figure 1 (c) shows that for $\beta = 0.33$, $\gamma = 0.5$ and $d = 1$, when simulating from a good approximation to the target distribution with the same step size, the MSE of SRK-ID remains small, whereas the MSE of EM converges to a larger value. However, this improvement diminishes as the dimensionality of the sampling problem increases. We report additional results with other parameter settings in Appendix H.2.2. Notably, we did not observe significant differences in the estimated squared W_2 values. We suspect this is due to the discrepancy being dominated by the bias of our estimator.

Acknowledgments

MAE is partially funded by NSERC [2019-06167] and CIFAR AI Chairs program at the Vector Institute.

[5] Unfortunately, there appear to be two definitions for KSD and the energy distance in the literature, differing in whether a square root is taken or not. We adopt the version with the square root taken.

References

[1] David F Anderson and Jonathan C Mattingly. A weak trapezoidal method for a class of stochastic differential equations. *arXiv preprint arXiv:0906.3475*, 2009.

[2] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

[3] Kevin Burrage, PM Burrage, and Tianhai Tian. Numerical methods for strong solutions of stochastic differential equations: an overview. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 460(2041):373–402, 2004.

[4] Simone Calogero. Exponential convergence to equilibrium for kinetic Fokker-Planck equations. *Communications in Partial Differential Equations*, 37(8):1357–1390, 2012.

[5] Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Advances in Neural Information Processing Systems*, pages 2278–2286, 2015.

[6] Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *arXiv preprint arXiv:1905.12247*, 2019.

[7] Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. *arXiv preprint arXiv:1705.09048*, 2017.

[8] Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017.

[9] Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

[10] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. JMLR: Workshop and Conference Proceedings, 2016.

[11] Arnak S Dalalyan. 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, 2017.

[12] Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 2019.

[13] Arnak S Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *arXiv preprint arXiv:1807.09382*, 2018.

[14] Arnak S Dalalyan and Alexandre B Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, 78(5):1423–1443, 2012.

[15] Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D Skeel, and Hartmut Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in neural information processing systems*, pages 3203–3211, 2014.

[16] V Dobrić and Joseph E Yukich. Asymptotics for transportation cost in high dimensions. *Journal of Theoretical Probability*, 8(1):97–118, 1995.

[17] RM Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[18] Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

[19] Alain Durmus, Umut Simsekli, Eric Moulines, Roland Badeau, and Gaël Richard. Stochastic gradient Richardson-Romberg Markov chain Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 2047–2055, 2016.

[20] Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

[21] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! *arXiv preprint arXiv:1801.02309*, 2018.

[22] Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166(3-4):851–886, 2016.

[23] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *arXiv preprint arXiv:1703.01617*, 2017.

[24] Murat A Erdogdu, Lester Mackey, and Ohad Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, pages 9694–9703, 2018.

[25] R'emi Flamary and Nicolas Courty. POT python optimal transport library, 2017. URL `https://github.com/rflamary/POT`.

[26] Saul B Gelfand and Sanjoy K Mitter. Recursive stochastic algorithms for global optimization in rˆd. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

[27] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[28] Ivan Gentil, Christian Léonard, and Luigia Ripani. About the analogy between optimal transport and minimal entropy. *arXiv preprint arXiv:1510.08230*, 2015.

[29] Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1292–1301. JMLR. org, 2017.

[30] Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, and Lester Mackey. Measuring sample quality with diffusions. *arXiv preprint arXiv:1611.06972*, 2016.

[31] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[32] Peter E Kloeden and Eckhard Platen. *Numerical solution of stochastic differential equations*, volume 23. Springer Science & Business Media, 2013.

[33] Peter E Kloeden, Eckhard Platen, and IW Wright. The approximation of multiple stochastic integrals. *Stochastic analysis and applications*, 10(4):431–441, 1992.

[34] Dmitriy F Kuznetsov. Explicit one-step strong numerical methods of order 2.5 for Itô stochastic differential equations, based on the unified Taylor-Itô and Taylor-Stratonovich expansions. *arXiv preprint arXiv:1802.04844*, 2018.

[35] Adrien Laurent and Gilles Vilmart. Exotic aromatic b-series for the study of long time integrators for a class of ergodic SDEs. *arXiv preprint arXiv:1707.02877*, 2017.

[36] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pages 276–284, 2016.

[37] Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pages 2917–2925, 2015.

[38] Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I Jordan. Sampling can be faster than optimization. *arXiv preprint arXiv:1811.08413*, 2018.

[39] Yi-An Ma, Niladri Chatterji, Xiang Cheng, Nicolas Flammarion, Peter Bartlett, and Michael I Jordan. Is there an analog of Nesterov acceleration for MCMC? *arXiv preprint arXiv:1902.00996*, 2019.

[40] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[41] Xuerong Mao. *Stochastic differential equations and applications*. Elsevier, 2007.

[42] Jonathan C Mattingly, Andrew M Stuart, and Desmond J Higham. Ergodicity for SDEs and approximations: locally lipschitz vector fields and degenerate noise. *Stochastic processes and their applications*, 101(2):185–232, 2002.

[43] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[44] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[45] GN Milstein and Michael V Tretyakov. Quasi-symplectic methods for Langevin-type equations. *IMA journal of numerical analysis*, 23(4):593–626, 2003.

[46] Grigori Noah Milstein and Michael V Tretyakov. *Stochastic numerics for mathematical physics*. Springer Science & Business Media, 2013.

[47] Wenlong Mou, Yi-An Ma, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. High-order Langevin diffusion yields an accelerated MCMC algorithm. *arXiv preprint arXiv:1908.10859*, 2019.

[48] Bernt Øksendal. Stochastic differential equations. In *Stochastic differential equations*, pages 65–84. Springer, 2003.

[49] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[50] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

[51] Gareth O Roberts, Richard L Tweedie, et al. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[52] Andreas Rößler. Runge–kutta methods for the strong approximation of solutions of stochastic differential equations. *SIAM Journal on Numerical Analysis*, 48(3):922–952, 2010.

[53] Sotirios Sabanis and Ying Zhang. Higher order Langevin Monte Carlo algorithm. *arXiv preprint arXiv:1808.00728*, 2018.

[54] Sotirios Sabanis and Ying Zhang. On explicit order 1.5 approximations with varying coefficients: the case of super-linear diffusion coefficients. *Journal of Complexity*, 50:84–115, 2019.

[55] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, et al. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41 (5):2263–2291, 2013.

[56] Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. *arXiv preprint arXiv:1909.05503*, 2019.

[57] Marvin K Simon. *Probability distributions involving Gaussian random variables: A handbook for engineers and scientists*. Springer Science & Business Media, 2007.

[58] Gábor J Székely. E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05):1–18, 2003.

[59] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

[60] Veeravalli S Varadarajan. On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):23–26, 1958.

[61] Santosh S Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Log-sobolev suffices. *arXiv preprint arXiv:1903.08568*, 2019.

[62] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[63] Gilles Vilmart. Postprocessed integrators for the high order integration of ergodic SDEs. *SIAM Journal on Scientific Computing*, 37(1):A201–A220, 2015.

[64] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

[65] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[66] Magnus Wiktorsson et al. Joint characteristic function and simultaneous simulation of iterated Itô integrals for multiple independent Brownian motions. *The Annals of Applied Probability*, 11 (2):470–487, 2001.

[67] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3122–3133, 2018.

[68] Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, and Ali Jadbabaie. Direct Runge-Kutta discretization achieves acceleration. In *Advances in Neural Information Processing Systems*, pages 3904–3913, 2018.

[69] Difan Zou, Pan Xu, and Quanquan Gu. Sampling from non-log-concave distributions via variance-reduced gradient Langevin dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2936–2945, 2019.

A Proof of Theorem 1

Proof. Let $\{X_t\}_{t\geq 0}$ denote the continuous-time process defined by the SDE (1) initiated from the target stationary distribution, driven by the Brownian motion $\{B_t\}_{t\geq 0}$. Since the continuous-time transition kernel preserves the stationary distribution, the marginal distribution of $\{X_t\}_{t\geq 0}$ remains to be the stationary distribution for all $t \geq 0$.

We denote by t_k $(k = 0, 1, \dots)$ the timestamps of the Markov chain obtained by discretizing the continuous-time process with a numerical integration scheme and assume the Markov chain has a constant step size h that satisfies the conditions in the theorem statement. We denote by $\tilde{X}_k$ the kth iterate of the Markov chain. In the following, we derive a recursion for the quantity

$$A_k = \mathbb{E}\left[\left\|X_{t_k} - \tilde{X}_k\right\|_2^2\right]^{1/2}.$$

Fix $k \in \mathbb{N}$. We define the process $\{\bar{X}_t\}_{t\geq 0}$ such that it is the Markov chain until t_k, starting from which it follows the continuous-time process defined by the SDE (1). We let $\{\bar{X}_t\}_{t\geq 0}$ and the Markov chain $\tilde{X}_k$ $(k = 0, 1, \dots)$ share the same Brownian motion $\{\bar{B}_t\}_{t\geq 0}$. Suppose $\{\mathcal{F}_t\}_{t\geq 0}$ is a filtration to which both $\{B_t\}_{t\geq 0}$ and $\{\bar{B}_t\}_{t\geq 0}$ are adapted. Conditional on $\mathcal{F}_{t_k}$, let $X_{t_{k+1}}$ and $\bar{X}_{t_{k+1}}$ be coupled such that

$$\mathbb{E}\left[\left\|X_{t_{k+1}} - \bar{X}_{t_{k+1}}\right\|_2^2 |\mathcal{F}_{t_k}\right] \leq e^{-2\alpha h}\left\|X_{t_k} - \bar{X}_{t_k}\right\|_2^2. \tag{11}$$

This we can achieve due to exponential W_2-contraction. We define the process $\{Z_s\}_{s\geq t_k}$ as follows

$$Z_s = \left(X_s - \bar{X}_s\right) - \left(X_{t_k} - \bar{X}_{t_k}\right).$$

Note $\int_{t_k}^{t_k+t} \sigma(X_s)\,\mathrm{d}B_s - \int_{t_k}^{t_k+t} \sigma(\bar{X}_s)\,\mathrm{d}\bar{B}_s$ is a Martingale w.r.t. $\{\mathcal{F}_{t_k+t}\}_{t\geq 0}$, since it is adapted and the two component Itô integrals are Martingales w.r.t. the considered filtration. By Fubini's theorem, we switch the order of integrals and obtain

$$\mathbb{E}\left[Z_{t_{k+1}}|\mathcal{F}_{t_k}\right] = \int_{t_k}^{t_{k+1}} \mathbb{E}\left[b(X_s) - b(\bar{X}_s)|\mathcal{F}_{t_k}\right]\,\mathrm{d}s.$$

By Jensen's inequality,

$$\begin{aligned}\left\|\mathbb{E}\left[Z_{t_{k+1}}|\mathcal{F}_{t_k}\right]\right\|_2^2 \leq& h\int_{t_k}^{t_{k+1}} \mathbb{E}\left[\left\|b(X_s) - b(\bar{X}_s)\right\|_2^2 |\mathcal{F}_{t_k}\right]\,\mathrm{d}s\\ \leq& \mu_1(b)^2 h\int_{t_k}^{t_{k+1}} \mathbb{E}\left[\left\|X_s - \bar{X}_s\right\|_2^2 |\mathcal{F}_{t_k}\right]\,\mathrm{d}s.\end{aligned} \tag{12}$$

For $s \in [t_k, t_k + h]$, by Young's inequality, Jensen's inequality, and Itô isometry,

$$\begin{aligned}&\mathbb{E}\left[\left\|X_s - \bar{X}_s\right\|_2^2 |\mathcal{F}_{t_k}\right]\\ =&\mathbb{E}\left[\left\|X_{t_k} - \bar{X}_{t_k} + \int_{t_k}^{s}\left(b(X_u) - b(\bar{X}_u)\right)\,\mathrm{d}u + \int_{t_k}^{s}\left(\sigma(X_u) - \sigma(\bar{X}_u)\right)\,\mathrm{d}B_u\right\|_2^2 |\mathcal{F}_{t_k}\right]\\ \leq& 4\left\|X_{t_k} - \bar{X}_{t_k}\right\|_2^2 + 4(s - t_k)\int_{t_k}^{s} \mathbb{E}\left[\left\|b(X_u) - b(\bar{X}_u)\right\|_2^2 |\mathcal{F}_{t_k}\right]\,\mathrm{d}u\\ &+ 4\int_{t_k}^{s} \mathbb{E}\left[\left\|\sigma(X_u) - \sigma(\bar{X}_u)\right\|_{\mathrm{F}}^2 |\mathcal{F}_{t_k}\right]\,\mathrm{d}u\\ \leq& 4\left\|X_{t_k} - \bar{X}_{t_k}\right\|_2^2 + 4(s - t_k)\mu_1(b)^2\int_{t_k}^{s} \mathbb{E}\left[\left\|X_u - \bar{X}_u\right\|_2^2 |\mathcal{F}_{t_k}\right]\,\mathrm{d}u\\ &+ 4\mu_1^{\mathrm{F}}(\sigma)^2\int_{t_k}^{s} \mathbb{E}\left[\left\|X_u - \bar{X}_u\right\|_2^2 |\mathcal{F}_{t_k}\right] u\\ \leq& 4\left\|X_{t_k} - \bar{X}_{t_k}\right\|_2^2 + 4\left(\mu_1(b)^2 + \mu_1^{\mathrm{F}}(\sigma)^2\right)\int_{t_k}^{s} \mathbb{E}\left[\left\|X_u - \bar{X}_u\right\|_2^2 |\mathcal{F}_{t_k}\right]\,\mathrm{d}u.\end{aligned}$$

By the integral form of Grönwall's inequality for continuous functions,

$$\mathbb{E}\left[\left\|X_s-\bar{X}_s\right\|_2^2|\mathcal{F}_{t_k}\right] \leq 4\exp\left(4\left(\mu_1(b)^2+\mu_1^{\mathrm{F}}(\sigma)^2\right)(s-t_k)\right)\left\|X_{t_k}-\bar{X}_{t_k}\right\|_2^2.$$

Plugging this result into (12), by $h<1/\left(8\mu_1(b)^2+8\mu_1^{\mathrm{F}}(\sigma)^2\right)$,

$$\begin{aligned}\left\|\mathbb{E}\left[Z_{t_{k+1}}|\mathcal{F}_{t_k}\right]\right\|_2^2 \leq&\frac{\mu_1(b)^2h}{\mu_1(b)^2+\mu_1^{\mathrm{F}}(\sigma)^2}\left[\exp\left(4\left(\mu_1(b)^2+\mu_1^{\mathrm{F}}(\sigma)^2\right)h\right)-1\right]\left\|X_{t_k}-\bar{X}_{t_k}\right\|_2^2\\ \leq&\frac{8\mu_1(b)^2h^2}{\mu_1(b)^2+\mu_1^{\mathrm{F}}(\sigma)^2}\left(\mu_1(b)^2+\mu_1^{\mathrm{F}}(\sigma)^2\right)\left\|X_{t_k}-\bar{X}_{t_k}\right\|_2^2\\ \leq&8\mu_1(b)^2h^2\left\|X_{t_k}-\bar{X}_{t_k}\right\|_2^2.\end{aligned} \tag{13}$$

By direct expansion,

$$\mathbb{E}\left[\left\|X_{t_{k+1}}-\bar{X}_{t_{k+1}}\right\|_2^2|\mathcal{F}_{t_k}\right]=\left\|X_{t_k}-\bar{X}_{t_k}\right\|_2^2+\mathbb{E}\left[\left\|Z_{t_{k+1}}\right\|_2^2|\mathcal{F}_{t_k}\right]+2\left\langle X_{t_k}-\bar{X}_{t_k},\mathbb{E}\left[Z_{t_{k+1}}|\mathcal{F}_{t_k}\right]\right\rangle. \tag{14}$$

Combining (11) (13) and (14), by the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}\left[\left\|Z_{t_{k+1}}\right\|_2^2|\mathcal{F}_{t_k}\right] \leq&\left(e^{-2\alpha h}-1\right)\left\|X_{t_k}-\bar{X}_{t_k}\right\|_2^2-2\left\langle X_{t_k}-\bar{X}_{t_k},\mathbb{E}\left[Z_{t_{k+1}}|\mathcal{F}_{t_k}\right]\right\rangle\\ \leq&2\left\|X_{t_k}-\bar{X}_{t_k}\right\|_2\left\|\mathbb{E}\left[Z_{t_{k+1}}|\mathcal{F}_{t_k}\right]\right\|_2\\ \leq&8\mu_1(b)h\left\|X_{t_k}-\bar{X}_{t_k}\right\|_2^2\\ =&8\mu_1(b)h\left\|X_{t_k}-\tilde{X}_k\right\|_2^2.\end{aligned}$$

Hence,

$$\mathbb{E}\left[\left\|Z_{t_{k+1}}\right\|_2^2\right]=\mathbb{E}\left[\mathbb{E}\left[\left\|Z_{t_{k+1}}\right\|_2^2|\mathcal{F}_{t_k}\right]\right]\leq 8\mu_1(b)h\mathbb{E}\left[\left\|X_{t_k}-\tilde{X}_k\right\|_2^2\right]=8\mu_1(b)hA_k^2.$$

Let $\lambda_3=8\lambda_1^{1/2}\mu_1(b)^{1/2}+2\lambda_2^{1/2}$. Then, by the Cauchy–Schwarz inequality, we obtain a recursion

$$\begin{aligned}A_{k+1}^2=&\mathbb{E}\left[\left\|X_{t_{k+1}}-\tilde{X}_{k+1}\right\|_2^2\right]\\ =&\mathbb{E}\left[\left\|X_{t_{k+1}}-\bar{X}_{t_{k+1}}+\bar{X}_{t_{k+1}}-\tilde{X}_{k+1}\right\|_2^2\right]\\ =&\mathbb{E}\left[\left\|X_{t_{k+1}}-\bar{X}_{t_{k+1}}\right\|_2^2+\left\|\bar{X}_{t_{k+1}}-\tilde{X}_{k+1}\right\|_2^2+2\left\langle X_{t_{k+1}}-\bar{X}_{t_{k+1}},\bar{X}_{t_{k+1}}-\tilde{X}_{k+1}\right\rangle\right]\\ =&\mathbb{E}\left[\mathbb{E}\left[\left\|X_{t_{k+1}}-\bar{X}_{t_{k+1}}\right\|_2^2|\mathcal{F}_{t_k}\right]\right]+\mathbb{E}\left[\mathbb{E}\left[\left\|\bar{X}_{t_{k+1}}-\tilde{X}_{k+1}\right\|_2^2|\mathcal{F}_{t_k}\right]\right]\\ &+2\mathbb{E}\left[\mathbb{E}\left[\left\langle X_{t_{k+1}}-\bar{X}_{t_{k+1}},\bar{X}_{t_{k+1}}-\tilde{X}_{k+1}\right\rangle|\mathcal{F}_{t_k}\right]\right]\\ =&\mathbb{E}\left[\mathbb{E}\left[\left\|X_{t_{k+1}}-\bar{X}_{t_{k+1}}\right\|_2^2|\mathcal{F}_{t_k}\right]\right]+\mathbb{E}\left[\mathbb{E}\left[\left\|\bar{X}_{t_{k+1}}-\tilde{X}_{k+1}\right\|_2^2|\mathcal{F}_{t_k}\right]\right]\\ &+2\mathbb{E}\left[\left\langle X_{t_k}-\bar{X}_{t_k},\mathbb{E}\left[\bar{X}_{t_{k+1}}-\tilde{X}_{k+1}|\mathcal{F}_{t_k}\right]\right\rangle\right]\\ &+2\mathbb{E}\left[\left\langle Z_{t_{k+1}},\bar{X}_{t_{k+1}}-\tilde{X}_{k+1}\right\rangle\right]\\ \leq&\mathbb{E}\left[\mathbb{E}\left[\left\|X_{t_{k+1}}-\bar{X}_{t_{k+1}}\right\|_2^2|\mathcal{F}_{t_k}\right]\right]+\mathbb{E}\left[\mathbb{E}\left[\left\|\bar{X}_{t_{k+1}}-\tilde{X}_{k+1}\right\|_2^2|\mathcal{F}_{t_k}\right]\right]\\ &+2\mathbb{E}\left[\left\|X_{t_k}-\bar{X}_{t_k}\right\|_2^2\right]^{1/2}\mathbb{E}\left[\left\|\mathbb{E}\left[\bar{X}_{t_{k+1}}-\tilde{X}_{k+1}|\mathcal{F}_{t_k}\right]\right\|_2^2\right]^{1/2}\\ &+2\mathbb{E}\left[\left\|Z_{t_{k+1}}\right\|_2^2\right]^{1/2}\mathbb{E}\left[\left\|\bar{X}_{t_{k+1}}-\tilde{X}_{k+1}\right\|_2^2\right]^{1/2}\end{aligned}$$

$$
\begin{aligned}
&\leq e^{-2\alpha h} A_k^2 + \lambda_1 h^{2p_1} + 2\lambda_2^{1/2} h^{p_2} A_k + 8\lambda_1^{1/2} \mu_1(b)^{1/2} h^{p_1+1/2} A_k \\
&\leq (1-\alpha h)\, A_k^2 + \lambda_3 h^{p_1+1/2} A_k + \lambda_1 h^{2p_1} \\
&\leq (1-\alpha h)\, A_k^2 + \frac{\alpha h}{2} A_k^2 + \frac{8}{\alpha} \lambda_3^2 h^{2p_1} + \lambda_1 h^{2p_1} \\
&\leq (1-\alpha h/2)\, A_k^2 + \left(8\lambda_3^2/\alpha + \lambda_1\right) h^{2p_1},
\end{aligned} \tag{15}
$$

where the third to last inequality follows from $e^{-2\alpha h} < 1 - \alpha h$ when $\alpha h < 1/2$, and the second to last inequality follows from the elementary relation below with the choice of $\kappa = \alpha/2$

$$
A_k h^{1/2} \cdot \lambda_3 h^{p_1} \leq \kappa A_k^2 h + \frac{4}{\kappa} \lambda_3^2 h^{2p_1}.
$$

Let $\eta = 1 - \alpha h/2 \leq e^{-\alpha h/2} \leq 1$. By unrolling the recursion,

$$
\begin{aligned}
A_k^2 &\leq (1-\alpha h/2)\, A_{k-1}^2 + \left(8\lambda_3^2/\alpha + \lambda_1\right) h^{2p_1} \\
&\leq \eta^k A_0^2 + \left(1 + \eta + \cdots + \eta^{k-1}\right) \left(8\lambda_3^2/\alpha + \lambda_1\right) h^{2p_1} \\
&\leq \eta^k A_0^2 + \left(8\lambda_3^2/\alpha + \lambda_1\right) h^{2p_1}/(1-\eta) \\
&= \eta^k A_0^2 + (16\lambda_3^2/\alpha^2 + 2\lambda_1/\alpha) h^{2p_1-1}.
\end{aligned}
$$

Let ν_k and ν^* be the measures associated with the kth iterate of the Markov chain and the target distribution, respectively. Since W_2 is defined as an infimum over all couplings,

$$
W_2(\nu_k, \nu^*) \leq A_k \leq e^{-\alpha h k/4} A_0 + (16\lambda_3^2/\alpha^2 + 2\lambda_1/\alpha)^{1/2} h^{p_1-1/2}.
$$

To ensure W_2 is less than some small positive tolerance ϵ, we need only ensure the two terms in the above inequality are each less than $\epsilon/2$. Some simple calculations show that it suffices that

$$
\begin{aligned}
h &< \left(\frac{2}{\epsilon}\sqrt{\frac{64(16\lambda_1\mu_1(b)+\lambda_2)}{\alpha^2} + \frac{2\lambda_1}{\alpha}}\right)^{-1/(p_1-1/2)} \wedge \frac{1}{2\alpha} \wedge \frac{1}{8\mu_1(b)^2 + 8\mu_1^{\mathrm{F}}(\sigma)^2}, \\
k &> \left[\left(\frac{2}{\epsilon}\sqrt{\frac{64(16\lambda_1\mu_1(b)+\lambda_2)}{\alpha^2} + \frac{2\lambda_1}{\alpha}}\right)^{1/(p_1-1/2)} \vee 2\alpha \vee \left(8\mu_1(b)^2 + 8\mu_1^{\mathrm{F}}(\sigma)^2\right)\right] \frac{4}{\alpha} \log\left(\frac{2A_0}{\epsilon}\right).
\end{aligned} \tag{16}
$$

Note that for small enough positive tolerance ϵ, when the step size satisfies (16), it suffices that

$$
k = \left\lceil \left(\frac{2}{\epsilon}\sqrt{\frac{64(16\lambda_1\mu_1(b)+\lambda_2)}{\alpha^2} + \frac{2\lambda_1}{\alpha}}\right)^{1/(p_1-1/2)} \frac{4}{\alpha} \log\left(\frac{2A_0}{\epsilon}\right) \right\rceil = \tilde{\mathcal{O}}(\epsilon^{-1/(p_1-1/2)}).
$$

□

B Proof of Theorem 2

B.1 Moment Bounds

Verifying the order conditions in Theorem 1 for SRK-LD requires bounding the second, fourth, and sixth moments of the Markov chain. In principle, one may employ an exponential moment bound argument using a Lyapunov function. However, in this case, the tightness of the final convergence bound may depend on the selection of the Lyapunov function, and reasoning about the dimension dependence can become less obvious. Here, we directly bound all the even moments by expanding the expression. Intuitively, one expects the $2n$th moments of the Markov chain iterates to be $\mathcal{O}(d^n)$. The following proofs assume Lipschitz smoothness of the potential to a certain order and dissipativity.

Definition B.1 (Dissipativity)**.** *For constants $\alpha, \beta > 0$, the diffusion satisfies the following*

$$
\langle \nabla f(x), x \rangle \geq \frac{\alpha}{2} \|x\|_2^2 - \beta, \quad \forall x \in \mathbb{R}^d.
$$

For the Langevin diffusion, dissipativity directly follows from strong convexity of the potential [24]. Here, α can be chosen as the strong convexity parameter, provided β is an appropriate constant of order $\mathcal{O}(d)$.

Additionally, we assume the discretization has a constant step size h and the timestamp of the kth iterate is t_k as per the proof of Theorem 1. To simplify notation, we define the following

$$
\begin{aligned}
\tilde{\nabla} f =& \frac{1}{2}\left(\nabla f(\tilde{H}_1) + \nabla f(\tilde{H}_2)\right), \\
v_1 =& \sqrt{2}\left(\frac{1}{2} + \frac{1}{\sqrt{6}}\right)\xi_{k+1}\sqrt{h}, \\
v_1' =& \sqrt{2}\left(\frac{1}{2} - \frac{1}{\sqrt{6}}\right)\xi_{k+1}\sqrt{h}, \\
v_2 =& \frac{1}{\sqrt{6}}\eta_{k+1}\sqrt{h},
\end{aligned}
$$

where $\xi_{k+1}, \eta_{k+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ independent of $\tilde{X}_k$ for all $k \in \mathbb{N}$. We rewrite $\tilde{H}_1$ and $\tilde{H}_2$ as

$$
\begin{aligned}
\tilde{H}_1 &= \tilde{X}_k + \Delta\tilde{H}_1 = \tilde{X}_k + v_1 + v_2, \\
\tilde{H}_2 &= \tilde{X}_k + \Delta\tilde{H}_2 = \tilde{X}_k + v_1' + v_2 - \nabla f(\tilde{X}_k)h.
\end{aligned}
$$

B.1.1 Second Moment Bound

Lemma 4. *If the second moment of the initial iterate is finite, then the second moments of Markov chain iterates defined in (9) are uniformly bounded by a constant of order $\mathcal{O}(d)$, i.e.*

$$
\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^2\right] \le \mathcal{U}_2, \quad \textit{for all } k \in \mathbb{N},
$$

where $\mathcal{U}_2 = \mathbb{E}\left[\left\|\tilde{X}_0\right\|_2^2\right] + N_6$, and constants N_1 to N_6 are given in the proof, if the step size

$$
h < 1 \wedge \frac{2d}{\pi_{2,2}(f)} \wedge \frac{2\pi_{2,1}(f)}{\pi_{2,2}(f)} \wedge \frac{\alpha}{4\mu_2(f)\pi_{2,2}(f)} \wedge \frac{3\alpha}{2N_1 + 4}.
$$

Proof. By direct computation,

$$
\begin{aligned}
\left\|\tilde{X}_{k+1}\right\|_2^2 =& \left\|\tilde{X}_k - \left(\nabla f(\tilde{H}_1) + \nabla f(\tilde{H}_2)\right)\frac{h}{2} + 2^{1/2}\xi_{k+1}h^{1/2}\right\|_2^2 \\
=& \left\|\tilde{X}_k\right\|_2^2 + \left\|\nabla f(\tilde{H}_1) + \nabla f(\tilde{H}_2)\right\|_2^2 \frac{h^2}{4} + 2\left\|\xi_{k+1}\right\|_2^2 h \\
&- \left\langle \tilde{X}_k, \nabla f(\tilde{H}_1) + \nabla f(\tilde{H}_2)\right\rangle h \\
&+ 2^{3/2}\left\langle \tilde{X}_k, \xi_{k+1}\right\rangle h^{1/2} \\
&- 2^{1/2}\left\langle \nabla f(\tilde{H}_1) + \nabla f(\tilde{H}_2), \xi_{k+1}\right\rangle h^{3/2}.
\end{aligned}
$$

In the following, we bound each term in the expansion separately and obtain a recursion. To achieve this, we first upper bound the second moments of $\tilde{H}_1$ and $\tilde{H}_2$ for $h < 2d \wedge 2\pi_{2,1}(f)/\pi_{2,2}(f)$,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\tilde{H}_1\right\|_2^2 |\mathcal{F}_{t_k}\right] =& \left\|\tilde{X}_k\right\|_2^2 + \mathbb{E}\left[\left\|v_1\right\|_2^2 |\mathcal{F}_{t_k}\right] + \mathbb{E}\left[\left\|v_2\right\|_2^2 |\mathcal{F}_{t_k}\right] \le \left\|\tilde{X}_k\right\|_2^2 + 3dh, \\
\mathbb{E}\left[\left\|\tilde{H}_2\right\|_2^2 |\mathcal{F}_{t_k}\right] =& \left\|\tilde{X}_k\right\|_2^2 + \left\|\nabla f(\tilde{X}_k)\right\|_2^2 h^2 + \mathbb{E}\left[\left\|v_1'\right\|_2^2 |\mathcal{F}_{t_k}\right] + \mathbb{E}\left[\left\|v_2\right\|_2^2 |\mathcal{F}_{t_k}\right] \\
&+ 2\left\langle \tilde{X}_k, \nabla f(\tilde{X}_k)\right\rangle h \\
\le& \left\|\tilde{X}_k\right\|_2^2 + \pi_{2,2}(f)\left(1 + \left\|\tilde{X}_k\right\|_2^2\right)h^2 + dh + 2\pi_{2,1}(f)\left\|\tilde{X}_k\right\|_2^2 h \\
\le& \left\|\tilde{X}_k\right\|_2^2 + 4\pi_{2,1}(f)h\left\|\tilde{X}_k\right\|_2^2 + 3dh.
\end{aligned}
$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla f(\tilde{H}_1)+\nabla f(\tilde{H}_2)\right\|_2^2|\mathcal{F}_{t_k}\right] \leq& 2\mathbb{E}\left[\left\|\nabla f(\tilde{H}_1)\right\|_2^2+\left\|\nabla f(\tilde{H}_2)\right\|_2^2|\mathcal{F}_{t_k}\right]\\
\leq& 2\pi_{2,2}(f)\mathbb{E}\left[2+\left\|\tilde{H}_1\right\|_2^2+\left\|\tilde{H}_2\right\|_2^2|\mathcal{F}_{t_k}\right]\\
=& N_1\left\|\tilde{X}_k\right\|_2^2+N_2,
\end{aligned}$$

where $N_1=2\pi_{2,2}(f)\left(2+4\pi_{2,1}(f)\right)$ and $N_2=2\pi_{2,2}(f)\left(6d+2\right)$.

Additionally, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
-\mathbb{E}\left[\left\langle\nabla f(\tilde{H}_1),\xi_{k+1}\right\rangle|\mathcal{F}_{t_k}\right] \leq& \mathbb{E}\left[\left\|\nabla f(\tilde{H}_1)\right\|_2\|\xi_{k+1}\|_2|\mathcal{F}_{t_k}\right]\\
\leq& \mathbb{E}\left[\left\|\nabla f(\tilde{H}_1)\right\|_2^2|\mathcal{F}_{t_k}\right]^{1/2}\mathbb{E}\left[\|\xi_{k+1}\|_2^2\right]^{1/2}\\
\leq& \sqrt{d\pi_{2,2}(f)}\left(1+\mathbb{E}\left[\left\|\tilde{H}_1\right\|_2^2|\mathcal{F}_{t_k}\right]^{1/2}\right)\\
\leq& \sqrt{d\pi_{2,2}(f)}\left(1+\left\|\tilde{X}_k\right\|_2+\sqrt{3dh}\right).
\end{aligned}\tag{17}$$

Similarly,

$$\begin{aligned}
-\mathbb{E}\left[\left\langle\nabla f(\tilde{H}_2),\xi_{k+1}\right\rangle|\mathcal{F}_{t_k}\right] \leq& \mathbb{E}\left[\left\|\nabla f(\tilde{H}_2)\right\|_2\|\xi_{k+1}\|_2|\mathcal{F}_{t_k}\right]\\
\leq& \mathbb{E}\left[\left\|\nabla f(\tilde{H}_2)\right\|_2^2|\mathcal{F}_{t_k}\right]^{1/2}\mathbb{E}\left[\|\xi_{k+1}\|_2^2|\mathcal{F}_{t_k}\right]^{1/2}\\
\leq& \sqrt{d\pi_{2,2}(f)}\left(1+\mathbb{E}\left[\left\|\tilde{H}_2\right\|_2^2|\mathcal{F}_{t_k}\right]^{1/2}\right)\\
\leq& \sqrt{d\pi_{2,2}(f)}\left(1+\left\|\tilde{X}_k\right\|_2+2\sqrt{\pi_{2,1}(f)h}\left\|\tilde{X}_k\right\|_2+\sqrt{3dh}\right).
\end{aligned}\tag{18}$$

Combining (17) and (18), we obtain the following using AM–GM,

$$\begin{aligned}
-2^{1/2}\mathbb{E}\left[\left\langle\nabla f(\tilde{H}_1)+\nabla f(\tilde{H}_2),\xi_{k+1}\right\rangle|\mathcal{F}_{t_k}\right]h^{3/2} \leq& N_3\left\|\tilde{X}_k\right\|_2h^{3/2}+N_4\\
\leq& \frac{1}{2}\left\|\tilde{X}_k\right\|_2^2h^2+\frac{N_3^2}{2}h+N_4h^{3/2}.
\end{aligned}$$

where $N_3=2\sqrt{2d\pi_{2,2}(f)}\left(1+\sqrt{\pi_{2,1}(f)}\right)$ and $N_4=2\sqrt{2d\pi_{2,2}(f)}\left(1+\sqrt{3d}\right)$.

Now, we lower bound the second moments of $\tilde{H}_1$ and $\tilde{H}_2$ by dissipativity,

$$\begin{aligned}
\mathbb{E}\left[\left\|\tilde{H}_1\right\|_2^2|\mathcal{F}_{t_k}\right] =& \mathbb{E}\left[\left\|\tilde{X}_k+v_1+v_2\right\|_2^2|\mathcal{F}_{t_k}\right]\\
=& \left\|\tilde{X}_k\right\|_2^2+\mathbb{E}\left[\|v_1\|_2^2|\mathcal{F}_{t_k}\right]+\mathbb{E}\left[\|v_2\|_2^2|\mathcal{F}_{t_k}\right]\geq\left\|\tilde{X}_k\right\|_2^2,
\end{aligned}\tag{19}$$

$$\begin{aligned}
\mathbb{E}\left[\left\|\tilde{H}_2\right\|_2^2|\mathcal{F}_{t_k}\right] =& \mathbb{E}\left[\left\|\tilde{X}_k-\nabla f(\tilde{X}_k)h+v_1'+v_2\right\|_2^2|\mathcal{F}_{t_k}\right]\\
=& \left\|\tilde{X}_k\right\|_2^2+\left\|\nabla f(\tilde{X}_k)\right\|_2^2h^2+\mathbb{E}\left[\|v_1'\|_2^2|\mathcal{F}_{t_k}\right]+\mathbb{E}\left[\|v_2\|_2^2|\mathcal{F}_{t_k}\right]\\
&+2\left\langle\tilde{X}_k,\nabla f(\tilde{X}_k)\right\rangle h\\
\geq& \left\|\tilde{X}_k\right\|_2^2+2\left(\frac{\alpha}{2}\left\|\tilde{X}_k\right\|_2^2-\beta\right)h
\end{aligned}$$

$$\geq \left\| \tilde{X}_k \right\|_2^2 - 2\beta h.$$

Additionally, by Stein's lemma for multivariate Gaussians,

$$\begin{aligned}
\mathbb{E}\left[\left\langle \nabla f(\tilde{H}_1), v_1 \right\rangle | \mathcal{F}_{t_k} \right] =& 2h \left(\frac{1}{2} + \frac{1}{\sqrt{6}} \right)^2 \mathbb{E}\left[\Delta(f)(\tilde{H}_1) | \mathcal{F}_{t_k} \right] \leq 2d\mu_3(f)h, \\
\mathbb{E}\left[\left\langle \nabla f(\tilde{H}_1), v_2 \right\rangle | \mathcal{F}_{t_k} \right] =& \frac{1}{6} h \mathbb{E}\left[\Delta(f)(\tilde{H}_1) | \mathcal{F}_{t_k} \right] \leq \frac{1}{6} d\mu_3(f)h, \\
\mathbb{E}\left[\left\langle \nabla f(\tilde{H}_2), v_1' \right\rangle | \mathcal{F}_{t_k} \right] =& 2h \left(\frac{1}{2} - \frac{1}{\sqrt{6}} \right)^2 \mathbb{E}\left[\Delta(f)(\tilde{H}_2) | \mathcal{F}_{t_k} \right] \leq d\mu_3(f)h, \\
\mathbb{E}\left[\left\langle \nabla f(\tilde{H}_2), v_2 \right\rangle | \mathcal{F}_{t_k} \right] =& \frac{1}{6} h \mathbb{E}\left[\Delta(f)(\tilde{H}_2) | \mathcal{F}_{t_k} \right] \leq d\mu_3(f)h.
\end{aligned}$$

Therefore, by dissipativity and the lower bound (19),

$$\begin{aligned}
-\mathbb{E}\left[\left\langle \nabla f(\tilde{H}_1), \tilde{X}_k \right\rangle | \mathcal{F}_{t_k} \right] =& -\mathbb{E}\left[\left\langle \nabla f(\tilde{H}_1), \tilde{H}_1 \right\rangle | \mathcal{F}_{t_k} \right] + \mathbb{E}\left[\left\langle \nabla f(\tilde{H}_1), v_1 + v_2 \right\rangle | \mathcal{F}_{t_k} \right] \\
\leq& -\frac{\alpha}{2} \mathbb{E}\left[\left\| \tilde{H}_1 \right\|_2^2 | \mathcal{F}_{t_k} \right] + \beta + \mathbb{E}\left[\left\langle \nabla f(\tilde{H}_1), v_1 + v_2 \right\rangle | \mathcal{F}_{t_k} \right] \\
\leq& -\frac{\alpha}{2} \left\| \tilde{X}_k \right\|_2^2 + \beta + 3d\mu_3(f)h. \qquad (20)
\end{aligned}$$

To bound the expectation of $-\langle \nabla f(\tilde{H}_2), \tilde{X}_k \rangle$, we first bound the second moment of $\Delta \tilde{H}_2$,

$$\begin{aligned}
\mathbb{E}\left[\left\| \Delta \tilde{H}_2 \right\|_2^2 | \mathcal{F}_{t_k} \right] =& \mathbb{E}\left[\left\| -\nabla f(\tilde{X}_k)h + v_1' + v_2 \right\|_2^2 | \mathcal{F}_{t_k} \right] \\
=& \left\| \nabla f(\tilde{X}_k) \right\|_2^2 h^2 + \mathbb{E}\left[\|v_1'\|_2^2 | \mathcal{F}_{t_k} \right] + \mathbb{E}\left[\|v_2\|_2^2 | \mathcal{F}_{t_k} \right] \\
\leq& \pi_{2,2}(f) \left(1 + \left\| \tilde{X}_k \right\|_2^2 \right) h^2 + dh. \qquad (21)
\end{aligned}$$

Notice the second equality above also implies

$$\left\| \nabla f(\tilde{X}_k) \right\|_2 h \leq \mathbb{E}\left[\left\| \Delta \tilde{H}_2 \right\|_2^2 | \mathcal{F}_{t_k} \right]^{1/2}. \qquad (22)$$

By Taylor's Theorem with the remainder in integral form,

$$\nabla f(\tilde{H}_2) = \nabla f(\tilde{X}_k) + R(t_{k+1}) = \nabla f(\tilde{X}_k) + \int_0^1 \nabla^2 f\left(\tilde{X}_k + \tau \Delta \tilde{H}_2 \right) \Delta \tilde{H}_2 \, \mathrm{d}\tau.$$

Since ∇f is Lipschitz, $\nabla^2 f$ is bounded, and

$$\|R(t_{k+1})\|_2 \leq \int_0^1 \left\| \nabla^2 f\left(\tilde{X}_k + \tau \Delta \tilde{H}_2 \right) \right\|_{\mathrm{op}} \left\| \Delta \tilde{H}_2 \right\|_2 \mathrm{d}\tau \leq \mu_2(f) \left\| \Delta \tilde{H}_2 \right\|_2.$$

By (21) and (22),

$$\begin{aligned}
-\mathbb{E}\left[\left\langle \nabla f(\tilde{H}_2), \nabla f(\tilde{X}_k) \right\rangle | \mathcal{F}_{t_k} \right] =& -\left\| \nabla f(\tilde{X}_k) \right\|_2^2 - \left\langle \mathbb{E}\left[R(t_{k+1}) | \mathcal{F}_{t_k} \right], \nabla f(\tilde{X}_k) \right\rangle \\
\leq& \left\| \mathbb{E}\left[R(t_{k+1}) | \mathcal{F}_{t_k} \right] \right\|_2 \left\| \nabla f(\tilde{X}_k) \right\|_2 \\
\leq& \mathbb{E}\left[\|R(t_{k+1})\|_2 | \mathcal{F}_{t_k} \right] \left\| \nabla f(\tilde{X}_k) \right\|_2 \\
\leq& \mu_2(f) \mathbb{E}\left[\left\| \Delta \tilde{H}_2 \right\|_2 | \mathcal{F}_{t_k} \right] \left\| \nabla f(\tilde{X}_k) \right\|_2 \\
\leq& \mu_2(f) \mathbb{E}\left[\left\| \Delta \tilde{H}_2 \right\|_2^2 | \mathcal{F}_{t_k} \right]^{1/2} \left\| \nabla f(\tilde{X}_k) \right\|_2 \\
\leq& \mu_2(f) \mathbb{E}\left[\left\| \Delta \tilde{H}_2 \right\|_2^2 | \mathcal{F}_{t_k} \right] h^{-1}
\end{aligned}$$

$$\leq \mu_2(f)\pi_{2,2}(f)\left(1+\left\|\tilde{X}_k\right\|_2^2\right)h+d.$$

Therefore, for $h < 1 \wedge \alpha/(4\mu_2(f)\pi_{2,2}(f))$,

$$\begin{aligned}
&-\mathbb{E}\left[\left\langle\nabla f(\tilde{H}_2), \tilde{X}_k\right\rangle|\mathcal{F}_{t_k}\right]\\
=&-\mathbb{E}\Big[\left\langle\nabla f(\tilde{H}_2), \tilde{H}_2\right\rangle+\left\langle\nabla f(\tilde{H}_2), \nabla f(\tilde{X}_k)\right\rangle h-\left\langle\nabla f(\tilde{H}_2), v_1'+v_2\right\rangle|\mathcal{F}_{t_k}\Big]\\
\leq&-\frac{\alpha}{2}\mathbb{E}\left[\left\|\tilde{H}_2\right\|_2^2|\mathcal{F}_{t_k}\right]+\beta-\mathbb{E}\left[\left\langle\nabla f(\tilde{H}_2), \nabla f(\tilde{X}_k)\right\rangle|\mathcal{F}_{t_k}\right]h+\mathbb{E}\left[\left\langle\nabla f(\tilde{H}_2), v_1'+v_2\right\rangle|\mathcal{F}_{t_k}\right]\\
\leq&-\frac{\alpha}{2}\left\|\tilde{X}_k\right\|_2^2+\alpha\beta h+\beta+\mu_2(f)\pi_{2,2}(f)\left(1+\left\|\tilde{X}_k\right\|_2^2\right)h^2+dh+2d\mu_3(f)h\\
\leq&-\frac{\alpha}{4}\left\|\tilde{X}_k\right\|_2^2+\left(\alpha\beta+\mu_2(f)\pi_{2,2}(f)+d+2d\mu_3(f)\right)h+\beta.
\end{aligned}\tag{23}$$

Combining (20) and (23), we have

$$-\mathbb{E}\left[\left\langle\nabla f(\tilde{H}_1)+\nabla f(\tilde{H}_2), \tilde{X}_k\right\rangle|\mathcal{F}_{t_k}\right]\leq-\frac{3}{4}\alpha\left\|\tilde{X}_k\right\|_2^2+N_5,\tag{24}$$

where $N_5=(\alpha\beta+\mu_2(f)\pi_{2,2}(f)+d+5d\mu_3(f))+2\beta$.

Putting things together, for $h < 3\alpha/(2N_1+4)$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|\tilde{X}_{k+1}\right\|_2^2|\mathcal{F}_{t_k}\right]=&\left\|\tilde{X}_k\right\|_2^2+\mathbb{E}\left[\left\|\nabla f(\tilde{H}_1)+\nabla f(\tilde{H}_2)\right\|_2^2|\mathcal{F}_{t_k}\right]\frac{h^2}{4}+2dh\\
&-\mathbb{E}\left[\left\langle\tilde{X}_k, \nabla f(\tilde{H}_1)+\nabla f(\tilde{H}_2)\right\rangle|\mathcal{F}_{t_k}\right]h\\
&-2^{1/2}\mathbb{E}\left[\left\langle\nabla f(\tilde{H}_1)+\nabla f(\tilde{H}_2), \xi_{k+1}\right\rangle|\mathcal{F}_{t_k}\right]h^{3/2}\\
\leq&\left\|\tilde{X}_k\right\|_2^2+\frac{N_1}{4}\left\|\tilde{X}_k\right\|_2^2h^2+\frac{N_2}{4}h^2+2dh\\
&-\frac{3}{4}\alpha h\left\|\tilde{X}_k\right\|_2^2+N_5h\\
&+\frac{1}{2}\left\|\tilde{X}_k\right\|_2^2h^2+\frac{N_3^2}{2}h+N_4h^{3/2}\\
\leq&\left(1-\frac{3}{4}\alpha h+\frac{N_1+2}{4}h^2\right)\left\|\tilde{X}_k\right\|_2^2\\
&+N_2h^2/4+2dh+N_5h+N_3^2h/2+N_4h^{3/2}\\
\leq&\left(1-\frac{3}{8}\alpha h\right)\left\|\tilde{X}_k\right\|_2^2+N_2h^2/4+2dh+N_5h+N_3^2h/2+N_4h^{3/2},
\end{aligned}$$

For $h < 1$, by unrolling the recursion, we obtain the following

$$\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^2\right]\leq\mathbb{E}\left[\left\|\tilde{X}_0\right\|_2^2\right]+N_6,\quad\text{for all } k\in\mathbb{N},$$

where

$$N_6=\frac{1}{3\alpha}\left(2N_2+16d+8N_5+4N_3^2+8N_4\right)=\mathcal{O}(d).$$

□

B.1.2 $2n$th Moment Bound

Lemma 5. *For $n\in\mathbb{N}_+$, if the $2n$th moment of the initial iterate is finite, then the $2n$th moments of Markov chain iterates defined in (9) are uniformly bounded by a constant of order $\mathcal{O}(d^n)$, i.e.*

$$\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^{2n}\right]\leq\mathcal{U}_{2n},\quad\textit{for all } k\in\mathbb{N},$$

where

$$\mathcal{U}_{2n} = \mathbb{E}\left[\left\|\tilde{X}_0\right\|_2^{2n}\right] + \frac{8}{3\alpha n}\left(N_{7,n} + N_{12,n}\right),$$

and constants $N_{7,n}$ to $N_{12,n}$ are given in the proof, if the step size

$$h < 1 \wedge \frac{2d}{\pi_{2,2}(f)} \wedge \frac{2\pi_{2,1}(f)}{\pi_{2,2}(f)} \wedge \frac{\alpha}{4\mu_2(f)\pi_{2,2}(f)} \wedge \frac{3\alpha}{2N_1+4} \wedge \min\left\{\left(\frac{3\alpha l}{8N_{11,l}}\right)^2 : l = 2, \ldots, n\right\}.$$

Proof. Our proof is by induction. The base case is given in Lemma 4. For the inductive case, we prove that the $2n$th moment is uniformly bounded by a constant of order $\mathcal{O}(d^n)$, assuming the $2(n\text{-}1)$th moment is uniformly bounded by a constant of order $\mathcal{O}(d^{n-1})$.

By the multinomial theorem,

$$\begin{aligned}
\mathbb{E}\left[\left\|\tilde{X}_{k+1}\right\|_2^{2n}\right] =& \mathbb{E}\left[\left\|X_k - \tilde{\nabla} f h + 2^{1/2}\xi_{k+1}h^{1/2}\right\|_2^{2n}\right] \\
=& \mathbb{E}\Big[\Big(\|X_k\|_2^2 + \left\|\tilde{\nabla} f\right\|_2^2 h^2 + 2\|\xi_{k+1}\|_2^2 h \\
& - 2\left\langle \tilde{X}_k, \tilde{\nabla} f\right\rangle h + 2^{3/2}\left\langle \tilde{X}_k, \xi_{k+1}\right\rangle h^{1/2} - 2^{3/2}\left\langle \tilde{\nabla} f, \xi_{k+1}\right\rangle h^{3/2}\Big)^n\Big] \\
=& \mathbb{E}\Bigg[\sum_{k_1+\cdots+k_6=n} (-1)^{k_4+k_6}\binom{n}{k_1 \ \ldots \ k_6} 2^{k_3+k_4+\frac{3k_5}{2}+\frac{3k_6}{2}} h^{2k_2+k_3+k_4+\frac{k_5}{2}+\frac{3k_6}{2}} \\
& \left\|\tilde{X}_k\right\|_2^{2k_1}\left\|\tilde{\nabla} f\right\|_2^{2k_2}\|\xi_{k+1}\|_2^{2k_3}\left\langle \tilde{X}_k, \tilde{\nabla} f\right\rangle^{k_4}\left\langle \tilde{X}_k, \xi_{k+1}\right\rangle^{k_5}\left\langle \tilde{\nabla} f, \xi_{k+1}\right\rangle^{k_6}\Bigg] \\
=& \mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^{2n} + Ah + Bh^{3/2}\right],
\end{aligned}$$

where

$$A = 2n\left\|\tilde{X}_k\right\|_2^{2(n-1)}\|\xi_{k+1}\|_2^2 - 2n\left\|\tilde{X}_k\right\|_2^{2(n-1)}\left\langle \tilde{X}_k, \tilde{\nabla} f\right\rangle + 4n(n-1)\left\|\tilde{X}_k\right\|_2^{2(n-2)}\left\langle \tilde{X}_k, \xi_{k+1}\right\rangle^2,$$

$$B \le \sum_{\substack{k_1+\cdots+k_6=n \\ 2k_2+k_3+k_4+\frac{k_5}{2}+\frac{3k_6}{2}>1}} 2^{\frac{3n}{2}}\binom{n}{k_1 \ \ldots \ k_6}\left\|\tilde{X}_k\right\|_2^{2k_1+k_4+k_5}\left\|\tilde{\nabla} f\right\|_2^{2k_2+k_4+k_6}\|\xi_{k+1}\|_2^{2k_3+k_5+k_6}.$$

Now, we bound the expectation of A using (24),

$$\begin{aligned}
\mathbb{E}\left[A|\mathcal{F}_{t_k}\right] \le& 2dn\left\|\tilde{X}_k\right\|_2^{2(n-1)} + 2n\left\|\tilde{X}_k\right\|_2^{2(n-1)}\left(-\frac{3}{8}\alpha\left\|\tilde{X}_k\right\|_2^2 + \frac{N_5}{2}\right) + 4dn(n-1)\left\|\tilde{X}_k\right\|_2^{2(n-1)} \\
\le& -\frac{3}{4}\alpha n\left\|\tilde{X}_k\right\|_2^{2n} + \left(2dn + nN_5 + 4dn(n-1)\right)\left\|\tilde{X}_k\right\|_2^{2(n-1)}.
\end{aligned}$$

Moreover, by the inductive hypothesis,

$$\mathbb{E}\left[A\right] = \mathbb{E}\left[\mathbb{E}\left[A|\mathcal{F}_{t_k}\right]\right] \le -\frac{3}{4}\alpha n\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^{2n}\right] + N_{7,n}, \tag{25}$$

where $N_{7,n} = \left(2dn + nN_5 + 4dn(n-1)\right)\mathcal{U}_{2(n-1)} = \mathcal{O}(d^n)$.

Next, we bound the expectation of B. By the Cauchy–Schwarz inequality,

$$\begin{aligned}
\mathbb{E}\left[B|\mathcal{F}_{t_k}\right] =& \sum_{\substack{k_1+\cdots+k_6=n \\ 2k_2+k_3+k_4+\frac{k_5}{2}+\frac{3k_6}{2}>1}} 2^{\frac{3n}{2}}\binom{n}{k_1 \ \ldots \ k_6}\left\|\tilde{X}_k\right\|_2^{2k_1+k_4+k_5}\mathbb{E}\left[\left\|\tilde{\nabla} f\right\|_2^{2k_2+k_4+k_6}\|\xi_{k+1}\|_2^{2k_3+k_5+k_6}|\mathcal{F}_{t_k}\right] \\
\le& \sum_{\substack{k_1+\cdots+k_6=n \\ 2k_2+k_3+k_4+\frac{k_5}{2}+\frac{3k_6}{2}>1}} 2^{\frac{3n}{2}}\binom{n}{k_1 \ \ldots \ k_6}\left\|\tilde{X}_k\right\|_2^{2k_1+k_4+k_5}
\end{aligned}$$

$$\times \mathbb{E}\left[\left\|\tilde{\nabla} f\right\|_2^{4k_2+2k_4+2k_6}|\mathcal{F}_{t_k}\right]^{1/2} \mathbb{E}\left[\|\xi_{k+1}\|_2^{4k_3+2k_5+2k_6}|\mathcal{F}_{t_k}\right]^{1/2}.$$

Let $\chi(d)^2$ be a chi-squared random variable with d degrees of freedom. Recall its nth moment has a closed form solution and is of order $\mathcal{O}(d^n)$ [57]. Now, we bound the $2p$th moments of $\tilde{H}_1$ and $\tilde{H}_2$ for positive integer p. To achieve this, we first expand the expressions,

$$\begin{aligned}
\left\|\tilde{H}_1\right\|_2^{2p} &= \left\|\tilde{X}_k + v_1 + v_2\right\|_2^{2p} \\
&= \left(\left\|\tilde{X}_k\right\|_2^2 + \|v_1\|_2^2 + \|v_2\|_2^2 + 2\left\langle \tilde{X}_k, v_1\right\rangle + 2\left\langle \tilde{X}_k, v_2\right\rangle + 2\left\langle v_1, v_2\right\rangle\right)^p \\
&\leq \sum_{j_1+\cdots+j_6=p} 2^{j_4+j_5+j_6}\binom{p}{j_1\ \ldots\ j_6}\left\|\tilde{X}_k\right\|_2^{2j_1+j_4+j_5}\|v_1\|_2^{2j_2+j_4+j_6}\|v_2\|_2^{2j_3+j_5+j_6} \\
&\leq \sum_{j_1+\cdots+j_6=p} 2^{j_2+\frac{3}{2}j_4+j_5+\frac{3}{2}j_6} h^{j_2+j_3+\frac{j_4}{2}+\frac{j_5}{2}+j_6}\binom{p}{j_1\ \ldots\ j_6}\left\|\tilde{X}_k\right\|_2^{2j_1+j_4+j_5} \\
&\qquad \times \|\xi_{k+1}\|_2^{2j_2+j_4+j_6}\|\eta_{k+1}\|_2^{2j_3+j_5+j_6} \\
&\leq \sum_{j_1+\cdots+j_6=p} 2^{3p}\binom{p}{j_1\ \ldots\ j_6}\left\|\tilde{X}_k\right\|_2^{2j_1+j_4+j_5}\|\xi_{k+1}\|_2^{2j_2+j_4+j_6}\|\eta_{k+1}\|_2^{2j_3+j_5+j_6} \\
&\leq \sum_{j_1+\cdots+j_6=p} 2^{3p}\binom{p}{j_1\ \ldots\ j_6}\left(\tfrac{2j_1+j_4+j_5}{2p}\|\tilde{X}_k\|_2^{2p} + \tfrac{2j_2+j_4+j_6}{2p}\|\xi_{k+1}\|_2^{2p} + \tfrac{2j_3+j_5+j_6}{2p}\|\eta_{k+1}\|_2^{2p}\right) \\
&\leq 2^{4p}3^p\left(\left\|\tilde{X}_k\right\|_2^{2p} + \|\xi_{k+1}\|_2^{2p} + \|\eta_{k+1}\|_2^{2p}\right),
\end{aligned}$$

where the second to last inequality follows from Young's inequality for products with three variables. Therefore,

$$\mathbb{E}\left[\left\|\tilde{H}_1\right\|_2^{2p}|\mathcal{F}_{t_k}\right] \leq 2^{4p}3^p\left\|\tilde{X}_k\right\|_2^{2p} + 2^{4p+1}3^p\mathbb{E}\left[\chi(d)^{2p}\right]. \tag{26}$$

Similarly,

$$\begin{aligned}
\left\|\tilde{H}_2\right\|_2^{2p} &= \left\|\tilde{X}_k - \nabla f(\tilde{X}_k)h + v_1' + v_2\right\|_2^{2p} \\
&\leq \Big(\left\|\tilde{X}_k\right\|_2^2 + \left\|\nabla f(\tilde{X}_k)\right\|_2^2 h^2 + \|v_1' + v_2\|_2^2 \\
&\qquad - 2\left\langle \tilde{X}_k, \nabla f(\tilde{X}_k)\right\rangle h + 2\left\langle \tilde{X}_k, v_1' + v_2\right\rangle - 2\left\langle \nabla f(\tilde{X}_k), v_1' + v_2\right\rangle\Big)^p \\
&\leq \sum_{j_1+\cdots+j_6=p} 2^{j_4+j_5+j_6}\binom{p}{j_1\ \ldots\ j_6}\left\|\tilde{X}_k\right\|_2^{2j_1+j_4+j_5}\left\|\nabla f(\tilde{X}_k)\right\|_2^{2j_2+j_4+j_6}\|v_1' + v_2\|_2^{2j_3+j_5+j_6} \\
&\leq 2^{4p}3^p\left(\left\|\tilde{X}_k\right\|_2^{2p} + \left\|\nabla f(\tilde{X}_k)\right\|_2^{2p} + \|\xi_{k+1}\|_2^{2p} + \|\eta_{k+1}\|_2^{2p}\right).
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\left\|\tilde{H}_2\right\|_2^{2p}|\mathcal{F}_{t_k}\right] \leq 2^{4p}3^p\left(1 + \pi_{2,2p}(f)\right)\left\|\tilde{X}_k\right\|_2^{2p} + 2^{4p+1}3^p\left(\pi_{2,2p}(f) + \mathbb{E}\left[\chi(d)^{2p}\right]\right). \tag{27}$$

Thus, combining (26) and (27),

$$\begin{aligned}
\mathbb{E}\left[\left\|\tilde{\nabla} f\right\|_2^{2p}|\mathcal{F}_{t_k}\right] &\leq \frac{1}{2}\mathbb{E}\left[\left\|\nabla f(\tilde{H}_1)\right\|_2^{2p} + \left\|\nabla f(\tilde{H}_2)\right\|_2^{2p}|\mathcal{F}_{t_k}\right] \\
&\leq \frac{1}{2}\pi_{2,2p}(f)\mathbb{E}\left[2 + \left\|\tilde{H}_1\right\|_2^{2p} + \left\|\tilde{H}_2\right\|_2^{2p}|\mathcal{F}_{t_k}\right] \\
&\leq N_{8,n}(p)^2\left\|\tilde{X}_k\right\|_2^{2p} + N_{9,n}(p)^2,
\end{aligned}$$

where the p-dependent constants are

$$N_{8,n}(p) = 2^{2p}3^{\frac{p}{2}}\left(\pi_{2,2p}(f)\left(1+\frac{1}{2}\pi_{2,2p}(f)\right)\right)^{\frac{1}{2}},$$
$$N_{9,n}(p) = \left(\pi_{2,2p}(f)\left(2^{4p+1}3^p\mathbb{E}\left[\chi(d)^{2p}\right]+2^{4p}3^p\pi_{2,2p}(f)+1\right)\right)^{\frac{1}{2}} = \mathcal{O}(d^{\frac{p}{2}}).$$

Since $N_{8,n}(p)$ does not depend on the dimension, let

$$N_{8,n} = \max\{N_{8,n}(2k_2+k_4+k_6) : k_1,\ldots,k_6\in\mathbb{N},\ k_1+\cdots+k_6=n,\ 2k_2+k_3+k_4+\tfrac{k_5}{2}+\tfrac{3k_6}{2}>1\}.$$

The bound on B reduces to

$$\begin{aligned}\mathbb{E}\left[B|\mathcal{F}_{t_k}\right] \leq& \sum_{\substack{k_1+\cdots+k_6=n\\ 2k_2+k_3+k_4+\frac{k_5}{2}+\frac{3k_6}{2}>1}} 2^{\frac{3n}{2}}\binom{n}{k_1\ \ldots\ k_6}\left\|\tilde{X}_k\right\|_2^{2k_1+k_4+k_5}\mathbb{E}\left[\chi(d)^{4k_3+2k_5+2k_6}\right]^{1/2}\\ &\times\left(N_{8,n}\left\|\tilde{X}_k\right\|_2^{2k_2+k_4+k_6}+N_{9,n}(2k_2+k_4+k_6)\right)\\ \leq& B_1+B_2,\end{aligned}$$

where

$$B_1 = \sum_{\substack{k_1+\cdots+k_6=n\\ 2k_2+k_3+k_4+\frac{k_5}{2}+\frac{3k_6}{2}>1}} 2^{\frac{3n}{2}}\binom{n}{k_1\ \ldots\ k_6}\mathbb{E}\left[\chi(d)^{4k_3+2k_5+2k_6}\right]^{1/2}N_{8,n}\left\|\tilde{X}_k\right\|_2^{2k_1+2k_2+2k_4+k_5+k_6},$$
$$B_2 = \sum_{\substack{k_1+\cdots+k_6=n\\ 2k_2+k_3+k_4+\frac{k_5}{2}+\frac{3k_6}{2}>1}} 2^{\frac{3n}{2}}\binom{n}{k_1\ \ldots\ k_6}\mathbb{E}\left[\chi(d)^{4k_3+2k_5+2k_6}\right]^{1/2}N_{9,n}(2k_2+k_4+k_6)\left\|\tilde{X}_k\right\|_2^{2k_1+k_4+k_5}.$$

In the following, we bound the expectations of B_1 and B_2 separately. By Young's inequality for products and the function $x\mapsto x^{1/(2k_3+k_5+k_6)}$ being concave on the positive domain,

$$\begin{aligned}&\mathbb{E}\left[\chi(d)^{4k_3+2k_5+2k_6}\right]^{1/2}N_{8,n}\left\|\tilde{X}_k\right\|_2^{2k_1+2k_2+2k_4+k_5+k_6}\\ &\leq N_{8,n}\left(\tfrac{2k_3+k_5+k_6}{2n}\mathbb{E}\left[\chi(d)^{4k_3+2k_5+2k_6}\right]^{\frac{2n}{4k_3+2k_5+2k_6}}+\tfrac{2k_1+2k_2+2k_4+k_5+k_6}{2n}\left\|\tilde{X}_k\right\|_2^{2n}\right)\\ &\leq N_{8,n}\left(\mathbb{E}\left[\chi(d)^2\right]^n+\left\|\tilde{X}_k\right\|_2^{2n}\right).\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}\left[B_1|\mathcal{F}_{t_k}\right] \leq& \sum_{k_1+\cdots+k_6=n} 2^{\frac{3n}{2}}\binom{n}{k_1\ \ldots\ k_6}N_{8,n}\left(\mathbb{E}\left[\chi(d)^2\right]^n+\left\|\tilde{X}_k\right\|_2^{2n}\right)\\ =&2^{\frac{3n}{2}}6^nN_{8,n}\left(d^n+\left\|\tilde{X}_k\right\|_2^{2n}\right).\end{aligned}\tag{28}$$

Similarly,

$$\begin{aligned}&\mathbb{E}\left[\chi(d)^{4k_3+2k_5+2k_6}\right]^{\frac{1}{2}}N_{9,n}(2k_2+k_4+k_6)\left\|\tilde{X}_k\right\|_2^{2k_1+k_4+k_5}\\ &\leq\left(\mathbb{E}\left[\chi(d)^{4k_3+2k_5+2k_6}\right]^{\frac{1}{2}}N_{9,n}(2k_2+k_4+k_6)\right)^{\frac{2n}{2k_2+2k_3+k_4+k_5+2k_6}}+\left\|\tilde{X}_k\right\|_2^{2n}\\ &\leq N_{10,n}+\left\|\tilde{X}_k\right\|_2^{2n},\end{aligned}$$

where

$$N_{10,n} = \max\Big\{\left(\mathbb{E}[\chi(d)^{4k_3+2k_5+2k_6}]^{\frac{1}{2}}N_{9,n}(2k_2+k_4+k_6)\right)^{\frac{2n}{2k_2+2k_3+k_4+k_5+2k_6}} : k_1,\ldots,k_6\in\mathbb{N},\ k_1+\cdots+k_6=n,\ 2k_2+k_3+k_4+\tfrac{k_5}{2}+\tfrac{3k_6}{2}>1\Big\} = \mathcal{O}(d^n).$$

Hence,

$$\begin{aligned}\mathbb{E}\left[B_2|\mathcal{F}_{t_k}\right] \leq& \sum_{k_1+\cdots+k_6=n} 2^{\frac{3n}{2}} \binom{n}{k_1 \ \ldots \ k_6} \left(N_{10,n} + \left\|\tilde{X}_k\right\|_2^{2n}\right) \\ \leq& 2^{\frac{3n}{2}} 6^n \left(N_{10,n} + \left\|\tilde{X}_k\right\|_2^{2n}\right). \end{aligned} \tag{29}$$

Therefore, combining (28) and (29),

$$\mathbb{E}\left[B\right] = \mathbb{E}\left[\mathbb{E}\left[B_1 + B_2|\mathcal{F}_{t_k}\right]\right] \leq N_{11,n}\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^{2n}\right] + N_{12,n}, \tag{30}$$

where

$$\begin{aligned} N_{11,n} =& 2^{\frac{3n}{2}} 6^n \left(1 + N_{8,n}\right), \\ N_{12,n} =& 2^{\frac{3n}{2}} 6^n \left(N_{8,n} d^n + N_{10,n}\right) = \mathcal{O}(d^n). \end{aligned}$$

Thus, when $h < (3n\alpha/8N_{11,n})^2$, by (25) and (30),

$$\begin{aligned} \mathbb{E}\left[\left\|\tilde{X}_{k+1}\right\|^{2n}\right] \leq& \left(1 - \frac{3}{4}\alpha n h + N_{11,n} h^{3/2}\right) \mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^{2n}\right] + N_{7,n} h + N_{12,n} h^{3/2} \\ \leq& \left(1 - \frac{3}{8}\alpha n h\right) \mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^{2n}\right] + N_{7,n} h + N_{12,n} h^{3/2}, \end{aligned}$$

Hence,

$$\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^{2n}\right] \leq \mathbb{E}\left[\left\|\tilde{X}_0\right\|_2^{2n}\right] + \frac{8}{3\alpha n}\left(N_{7,n} + N_{12,n}\right).$$

□

B.2 Local Deviation Orders

We first provide two lemmas on bounding the second and fourth moments of the change in the continuous-time process. These will be used later when we verify the order conditions.

Lemma 6. *Suppose X_t is the continuous-time process defined by (3) initiated from some iterate of the Markov chain X_0 defined by (9), then the second moment of X_t is uniformly bounded by a constant of order $\mathcal{O}(d)$, i.e.*

$$\mathbb{E}\left[\|X_t\|_2^2\right] \leq \mathcal{U}_2', \quad \textit{for all } t \geq 0,$$

where $\mathcal{U}_2' = \mathcal{U}_2 + 2(\beta + d)/\alpha$.

Proof. By Itô's lemma and dissipativity,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\|X_t\|_2^2\right] = -2\mathbb{E}\left[\langle \nabla f(X_t), X_t\rangle\right] + 2d \leq -\alpha\mathbb{E}\left[\|X_t\|_2^2\right] + 2(\beta + d).$$

Moreover, by Grönwall's inequality,

$$\mathbb{E}\left[\|X_t\|_2^2\right] \leq e^{-\alpha t}\mathbb{E}\left[\|X_0\|_2^2\right] + 2(\beta + d)/\alpha \leq \mathcal{U}_2 + 2(\beta + d)/\alpha = \mathcal{U}_2'.$$

□

Lemma 7 (Second Moment of Change)**.** *Suppose X_t is the continuous-time process defined by (3) initiated from some iterate of the Markov chain X_0 defined by (9), then*

$$\mathbb{E}\left[\|X_t - X_0\|_2^2\right] \leq C_0 t = \mathcal{O}(dt), \quad \textit{for all } 0 \leq t \leq 1,$$

where $C_0 = 2\pi_{2,2}(f)\left(1 + \mathcal{U}_2'\right) + 4d$.

Proof. By Young's inequality,

$$\begin{aligned}\mathbb{E}\left[\|X_t - X_0\|_2^2\right] &= \mathbb{E}\left[\left\|-\int_0^t \nabla f(X_s)\ \mathrm{d}s + \sqrt{2}B_t\right\|_2^2\right] \\ &\leq 2\mathbb{E}\left[\left\|\int_0^t \nabla f(X_s)\ \mathrm{d}s\right\|_2^2 + 2\|B_t\|_2^2\right] \\ &\leq 2t\int_0^t \mathbb{E}\left[\|\nabla f(X_s)\|_2^2\right]\ \mathrm{d}s + 4\mathbb{E}\left[\|B_t\|_2^2\right] \\ &\leq 2\pi_{2,2}(f)t\int_0^t \mathbb{E}\left[1 + \|X_s\|_2^2\right]\ \mathrm{d}s + 4dt \\ &\leq 2\pi_{2,2}(f)\left(1 + \mathcal{U}_2'\right)t + 4dt.\end{aligned}$$

□

Lemma 8. *Suppose X_t is the continuous-time process defined by (3) initiated from some iterate of the Markov chain X_0 defined by (9), then the fourth moment of X_t is uniformly bounded by a constant of order $\mathcal{O}(d^2)$, i.e.*

$$\mathbb{E}\left[\|X_t\|_2^4\right] \leq \mathcal{U}_4', \quad \textit{for all } t \geq 0,$$

where $\mathcal{U}_4' = \mathcal{U}_4 + (2\beta + 6)\mathcal{U}_2'/\alpha$.

Proof. By Itô's lemma, dissipativity, and Lemma 6,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\|X_t\|_2^4\right] &= -4\mathbb{E}\left[\|X_t\|_2^2\langle \nabla f(X_t), X_t\rangle\right] + 12\mathbb{E}\left[\|X_t\|_2^2\right] \\ &\leq -2\alpha\mathbb{E}\left[\|X_t\|_2^4\right] + (4\beta + 12)\mathbb{E}\left[\|X_t\|_2^2\right] \\ &\leq -2\alpha\mathbb{E}\left[\|X_t\|_2^4\right] + (4\beta + 12)\mathcal{U}_2'.\end{aligned}$$

Moreover, by Grönwall's inequality,

$$\begin{aligned}\mathbb{E}\left[\|X_t\|_2^4\right] &\leq e^{-2\alpha t}\mathbb{E}\left[\|X_0\|_2^4\right] + (2\beta + 6)\mathcal{U}_2'/\alpha \\ &\leq \mathcal{U}_4 + (2\beta + 6)\mathcal{U}_2'/\alpha = \mathcal{U}_4'.\end{aligned}$$

□

Lemma 9 (Fourth Moment of Change)**.** *Suppose X_t is the continuous-time process defined by (3) initiated from some iterate of the Markov chain X_0 defined by (9), then*

$$\mathbb{E}\left[\|X_t - X_0\|_2^4\right] \leq C_1 t^2 = \mathcal{O}(d^2t^2), \quad \textit{for all } 0 \leq t \leq 1,$$

where $C_1 = 8\pi_{2,4}(f)\left(1 + \mathcal{U}_4'\right) + 32d(d + 2)$.

Proof. By Young's inequality,

$$\begin{aligned}\mathbb{E}\left[\|X_t - X_0\|_2^4\right] &= \mathbb{E}\left[\left\|-\int_0^t \nabla f(X_s)\ \mathrm{d}s + \sqrt{2}B_t\right\|_2^4\right] \\ &= \mathbb{E}\left[\left(\left\|-\int_0^t \nabla f(X_s)\ \mathrm{d}s + \sqrt{2}B_t\right\|_2^2\right)^2\right] \\ &\leq \mathbb{E}\left[\left(2\left\|\int_0^t \nabla f(X_s)\ \mathrm{d}s\right\|_2^2 + 4\|B_t\|_2^2\right)^2\right]\end{aligned}$$

$$
\begin{aligned}
&\le \mathbb{E}\left[\left(2t\int_0^t \|\nabla f(X_s)\|_2^2 \,\mathrm{d}s + 4\|B_t\|_2^2\right)^2\right]\\
&\le \mathbb{E}\left[8t^2\left(\int_0^t \|\nabla f(X_s)\|_2^2 \,\mathrm{d}s\right)^2 + 32\|B_t\|_2^4\right]\\
&\le 8t^3\int_0^t \mathbb{E}\left[\|\nabla f(X_s)\|_2^4\right] \mathrm{d}s + 32\mathbb{E}\left[\|B_t\|_2^4\right]\\
&\le 8\pi_{2,4}(f)t^3\int_0^t \mathbb{E}\left[1+\|X_s\|_2^4\right] \mathrm{d}s + 32d(d+2)t^2\\
&\le 8\pi_{2,4}(f)\left(1+\mathcal{U}_4'\right)t^2 + 32d(d+2)t^2.
\end{aligned}
$$

□

B.2.1 Local Mean-Square Deviation

Lemma 10. *Suppose X_t and $\tilde{X}_t$ are the continuous-time process defined by (3) and Markov chain defined by (9) for time $t \ge 0$, respectively. If X_t and $\tilde{X}_t$ are initiated from the same iterate of the Markov chain X_0 and share the same Brownian motion, then*

$$
\mathbb{E}\left[\left\|X_t - \tilde{X}_t\right\|_2^2\right] \le C_2 t^4 = \mathcal{O}(d^2t^4), \quad \textit{for all } 0 \le t \le 1,
$$

where

$$
\begin{aligned}
C_2 =& 8C_1^{1/2}(1+\mathcal{U}_4')^{1/2}\left(\mu_2(f)^2\pi_{3,4}(f)^{1/2} + \mu_3(f)^2\pi_{2,4}(f)^{1/2}\right)\\
&+ \left(8\pi_{2,4}(f)\left(1+\mathcal{U}_4\right) + 116d^2 + 90d + 8C_0\right)\mu_3(f)^2.
\end{aligned}
$$

Proof. Since the two processes share the same Brownian motion,

$$
X_t - \tilde{X}_t = -\int_0^t \nabla f(X_s)\,\mathrm{d}s + \frac{t}{2}\left(\nabla f(\tilde{H}_1) + \nabla f(\tilde{H}_2)\right). \tag{31}
$$

By Itô's lemma,

$$
\begin{aligned}
\nabla f(X_s) =& \nabla f(X_0) - \int_0^s \left(\nabla^2 f(X_u)\nabla f(X_u) - \vec{\Delta}\left(\nabla f\right)(X_u)\right)\mathrm{d}u + \sqrt{2}\int_0^s \nabla^2 f(X_u)\,\mathrm{d}B_u\\
=& \nabla f(X_0) - \nabla^2 f(X_0)\nabla f(X_0)s + \sqrt{2}\nabla^2 f(X_0)B_s + R(s),
\end{aligned}
$$

where the remainder is

$$
\begin{aligned}
R(s) =& \underbrace{\int_0^s \left(-\nabla^2 f(X_u)\nabla f(X_u) + \nabla^2 f(X_0)\nabla f(X_0)\right)\mathrm{d}u}_{R_1(s)} + \underbrace{\int_0^s \vec{\Delta}\left(\nabla f\right)(X_u)\,\mathrm{d}u}_{R_2(s)}\\
&+ \underbrace{\sqrt{2}\int_0^s \left(\nabla^2 f(X_u) - \nabla^2 f(X_0)\right)\mathrm{d}B_u}_{R_3(s)}.
\end{aligned}
$$

We bound the second moment of $R(s)$ by bounding those of $R_1(s)$, $R_2(s)$, and $R_3(s)$ separately. For $R_1(s)$, by the Cauchy–Schwarz inequality,

$$
\begin{aligned}
\mathbb{E}\left[\|R_1(s)\|_2^2\right] =& \mathbb{E}\left[\left\|\int_0^s \left(\nabla^2 f(X_u)\nabla f(X_u) - \nabla^2 f(X_0)\nabla f(X_0)\right)\mathrm{d}u\right\|_2^2\right]\\
=& 2\mathbb{E}\left[\left\|\int_0^s \left(\nabla^2 f(X_u)\nabla f(X_u) - \nabla^2 f(X_0)\nabla f(X_u)\right)\mathrm{d}u\right\|_2^2\right]\\
&+ 2\mathbb{E}\left[\left\|\int_0^s \left(\nabla^2 f(X_0)\nabla f(X_u) - \nabla^2 f(X_0)\nabla f(X_0)\right)\mathrm{d}u\right\|_2^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq 2s\int_0^s \mathbb{E}\left[\left\|\nabla^2 f(X_u)\nabla f(X_u)-\nabla^2 f(X_0)\nabla f(X_u)\right\|_2^2\right]\,\mathrm{d}u\\
&\quad+2s\int_0^s \mathbb{E}\left[\left\|\nabla^2 f(X_0)\nabla f(X_u)-\nabla^2 f(X_0)\nabla f(X_0)\right\|_2^2\right]\,\mathrm{d}u\\
&\leq 2s\int_0^s \mathbb{E}\left[\left\|\nabla^2 f(X_u)-\nabla^2 f(X_0)\right\|_{\mathrm{op}}^2\left\|\nabla f(X_u)\right\|_2^2\right]\,\mathrm{d}u\\
&\quad+2s\int_0^s \mathbb{E}\left[\left\|\nabla^2 f(X_0)\right\|_{\mathrm{op}}^2\left\|\nabla f(X_u)-\nabla f(X_0)\right\|_2^2\right]\,\mathrm{d}u\\
&\leq 2\mu_3(f)^2 s\int_0^s \mathbb{E}\left[\left\|X_u-X_0\right\|_2^2\left\|\nabla f(X_u)\right\|_2^2\right]\,\mathrm{d}u\\
&\quad+2\mu_2(f)^2 s\int_0^s \mathbb{E}\left[\left\|\nabla^2 f(X_0)\right\|_{\mathrm{op}}^2\left\|X_u-X_0\right\|_2^2\right]\,\mathrm{d}u\\
&\leq 2\mu_3(f)^2 s\int_0^s \mathbb{E}\left[\left\|X_u-X_0\right\|_2^4\right]^{1/2}\mathbb{E}\left[\left\|\nabla f(X_u)\right\|_2^4\right]^{1/2}\,\mathrm{d}u\\
&\quad+2\mu_2(f)^2 s\int_0^s \mathbb{E}\left[\left\|\nabla^2 f(X_0)\right\|_{\mathrm{op}}^4\right]^{1/2}\mathbb{E}\left[\left\|X_u-X_0\right\|_2^4\right]^{1/2}\,\mathrm{d}u\\
&\leq 2\mu_3(f)^2\pi_{2,4}(f)^{1/2}C_1^{1/2}\left(1+\mathcal{U}_4'\right)^{1/2}\int_0^s u\,\mathrm{d}u\\
&\quad+2\mu_2(f)^2\pi_{3,4}(f)^{1/2}C_1^{1/2}\left(1+\mathcal{U}_4'\right)^{1/2}s\int_0^s u\,\mathrm{d}u\\
&\leq C_1^{1/2}\left(1+\mathcal{U}_4'\right)^{1/2}\left(\mu_2(f)^2\pi_{3,4}(f)^{1/2}+\mu_3(f)^2\pi_{2,4}(f)^{1/2}\right)s^3. \qquad (32)
\end{aligned}
$$

For $R_2(s)$, by Lemma 34,

$$
\begin{aligned}
\mathbb{E}\left[\left\|R_2(s)\right\|_2^2\right] &= \mathbb{E}\left[\left\|\int_0^s \vec{\Delta}\left(\nabla f\right)(X_u)\,\mathrm{d}u\right\|_2^2\right]\\
&\leq s\int_0^s \mathbb{E}\left[\left\|\vec{\Delta}\left(\nabla f\right)(X_u)\right\|_2^2\right]\,\mathrm{d}u\\
&\leq \mu_3(f)^2 d^2 s^2. \qquad (33)
\end{aligned}
$$

For $R_3(s)$, by Itô isometry,

$$
\begin{aligned}
\mathbb{E}\left[\left\|R_3(s)\right\|_2^2\right] &= 2\mathbb{E}\left[\left\|\int_0^s \left(\nabla^2 f(X_u)-\nabla^2 f(X_0)\right)\,\mathrm{d}B_u\right\|_2^2\right]\\
&= 2\mathbb{E}\left[\int_0^s \left\|\nabla^2 f(X_u)-\nabla^2 f(X_0)\right\|_2^2\,\mathrm{d}u\right]\\
&\leq 2\mu_3(f)^2\int_0^s \mathbb{E}\left[\left\|X_u-X_0\right\|_2^2\right]\,\mathrm{d}u\\
&\leq 2\mu_3(f)^2 C_0\int_0^s u\,\mathrm{d}u\\
&\leq \mu_3(f)^2 C_0 s^2. \qquad (34)
\end{aligned}
$$

Thus, combining (32), (33), and (34),

$$
\begin{aligned}
\mathbb{E}\left[\left\|R(s)\right\|_2^2\right] &\leq 4\mathbb{E}\left[\left\|R_1(s)\right\|_2^2\right]+4\mathbb{E}\left[\left\|R_2(s)\right\|_2^2\right]+4\mathbb{E}\left[\left\|R_3(s)\right\|_2^2\right]\\
&\leq 4C_1^{1/2}(1+\mathcal{U}_4')^{1/2}\left(\mu_2(f)^2\pi_{3,4}(f)^{1/2}+\mu_3(f)^2\pi_{2,4}(f)^{1/2}\right)s^2\\
&\quad+4\mu_3(f)^2\left(d^2+C_0\right)s^2.
\end{aligned}
$$

Next, we characterize the terms in the Markov chain update. By Taylor's theorem,

$$
\nabla f(\tilde{H}_1)=\nabla f(X_0)+\nabla^2 f(X_0)\Delta\tilde{H}_1+\rho_1(t),
$$

$$\nabla f(\tilde{H}_2) = \nabla f(X_0) + \nabla^2 f(X_0)\Delta\tilde{H}_2 + \rho_2(t),$$

where

$$\begin{aligned}
\rho_1(t) &= \int_0^1 (1-\tau)\nabla^3 f(X_0 + \tau\Delta\tilde{H}_1)[\Delta\tilde{H}_1,\, \Delta\tilde{H}_1]\ \mathrm{d}\tau,\\
\rho_2(t) &= \int_0^1 (1-\tau)\nabla^3 f(X_0 + \tau\Delta\tilde{H}_2)[\Delta\tilde{H}_2,\, \Delta\tilde{H}_2]\ \mathrm{d}\tau,\\
\Delta\tilde{H}_1 &= \sqrt{2}\left(\frac{1}{t}\Psi(t) + \frac{1}{\sqrt{6}}B_t\right),\\
\Delta\tilde{H}_2 &= -\nabla f(X_0)t + \sqrt{2}\left(\frac{1}{t}\Psi(t) - \frac{1}{\sqrt{6}}B_t\right),\\
\Psi(t) &= \int_0^t B_s\ \mathrm{d}s.
\end{aligned}$$

We bound the fourth moments of $\Delta\tilde{H}_1$ and $\Delta\tilde{H}_2$,

$$\begin{aligned}
\mathbb{E}\left[\left\|\Delta\tilde{H}_1\right\|_2^4\right] =& \mathbb{E}\left[\left\|\sqrt{2}\left(\frac{1}{t}\Psi(t) + \frac{1}{\sqrt{6}}B_t\right)\right\|_2^4\right]\\
\le& \frac{32}{t^4}\mathbb{E}\left[\|\Psi(t)\|_2^4\right] + \frac{8}{9}\mathbb{E}\left[\|B_t\|_2^4\right]\\
=& \frac{32}{t^4}\sum_{i=1}^d \mathbb{E}\left[\Psi_i(t)^4\right] + \frac{32}{t^4}\sum_{i,j=1,i\neq j}^d \mathbb{E}\left[\Psi_i(t)^2\right]\mathbb{E}\left[\Psi_j(t)^2\right] + \frac{8}{9}d(d+2)t^2\\
\le& \frac{32}{t^4}\frac{dt^6}{3} + \frac{32}{t^4}\frac{d(d-1)t^6}{9} + \frac{8d(d+2)t^2}{9}\\
=& \left(\frac{32d}{3} + \frac{32d(d-1)}{9} + \frac{8d(d+2)}{9}\right)t^2\\
\le& 2d(6d+5)t^2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}\left[\left\|\Delta\tilde{H}_2\right\|_2^4\right] =& \mathbb{E}\left[\left\|-\nabla f(X_0)t + \sqrt{2}\left(\frac{1}{t}\Psi(t) - \frac{1}{\sqrt{6}}B_t\right)\right\|_2^4\right]\\
\le& 8\mathbb{E}\left[\|\nabla f(X_0)\|_2^4\right]t^4 + 8\mathbb{E}\left[\left\|\sqrt{2}\left(\frac{1}{t}\Psi(t) - \frac{1}{\sqrt{6}}B_t\right)\right\|_2^4\right]\\
\le& 8\pi_{2,4}(f)\mathbb{E}\left[1 + \|X_0\|_2^4\right]t^4 + 16d(6d+5)t^2\\
\le& 8\pi_{2,4}(f)\left(1 + \mathcal{U}_4\right)t^4 + 16d(6d+5)t^2\\
\le& 8\left(\pi_{2,4}(f)\left(1 + \mathcal{U}_4\right) + 2d(6d+5)\right)t^2.
\end{aligned}$$

Using the above information, we bound the second moments of $\rho_1(t)$ and $\rho_2(t)$,

$$\begin{aligned}
\mathbb{E}\left[\|\rho_1(t)\|_2^2\right] =& \mathbb{E}\left[\left\|\int_0^1 (1-\tau)\nabla^3 f(X_0 + \tau\Delta\tilde{H}_1)[\Delta\tilde{H}_1,\, \Delta\tilde{H}_1]\ \mathrm{d}\tau\right\|_2^2\right]\\
\le& \int_0^1 \mathbb{E}\left[\left\|\nabla^3 f(X_0 + \tau\Delta\tilde{H}_1)[\Delta\tilde{H}_1,\, \Delta\tilde{H}_1]\right\|_2^2\right]\mathrm{d}\tau\\
\le& \int_0^1 \mathbb{E}\left[\left\|\nabla^3 f(X_0 + \tau\Delta\tilde{H}_1)\right\|_{\mathrm{op}}^2\left\|\Delta\tilde{H}_1\right\|_2^4\right]\mathrm{d}\tau\\
\le& \mu_3(f)^2\int_0^1 \mathbb{E}\left[\left\|\Delta\tilde{H}_1\right\|_2^4\right]\mathrm{d}\tau\\
\le& 2d(6d+5)\mu_3(f)^2 t^2.
\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}\left[\|\rho_2(t)\|_2^2\right] \le&\mu_3(f)^2\int_0^1 \mathbb{E}\left[\left\|\Delta\tilde{H}_2\right\|_2^4\right]\,\mathrm{d}\tau\\ \le&8\left(\pi_{2,4}(f)\left(1+\mathcal{U}_4\right)+2d(6d+5)\right)\mu_3(f)^2t^2.\end{aligned}$$

Plugging these results into (31),

$$X_t-\tilde{X}_t=-\int_0^t R(s)\,\mathrm{d}s-\frac{t}{2}\left(\rho_1(t)+\rho_2(t)\right).$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\left\|X_t-\tilde{X}_t\right\|_2^2\right]=&\mathbb{E}\left[\left\|-\int_0^t R(s)\,\mathrm{d}s-\frac{t}{2}\left(\rho_1(t)+\rho_2(t)\right)\right\|_2^2\right]\\ \le&4t\int_0^t\mathbb{E}\left[\|R(s)\|_2^2\right]\,\mathrm{d}s+t^2\mathbb{E}\left[\|\rho_1(t)\|_2^2\right]+t^2\mathbb{E}\left[\|\rho_2(t)\|_2^2\right]\\ \le&8C_1^{1/2}(1+\mathcal{U}_4')^{1/2}\left(\mu_2(f)^2\pi_{3,4}(f)^{1/2}+\mu_3(f)^2\pi_{2,4}(f)^{1/2}\right)t^4\\ &+\left(8\pi_{2,4}(f)\left(1+\mathcal{U}_4\right)+116d^2+90d+8C_0\right)\mu_3(f)^2t^4\\ \le&C_2t^4.\end{aligned}$$

□

B.2.2 Local Mean Deviation

Lemma 11. *Suppose X_t and $\tilde{X}_t$ are the continuous-time process defined by (3) and Markov chain defined by (9) for time $t\ge 0$, respectively. If X_t and $\tilde{X}_t$ are initiated from the same iterate of the Markov chain X_0 and share the same Brownian motion, then*

$$\mathbb{E}\left[\left\|\mathbb{E}\left[X_t-\tilde{X}_t|\mathcal{F}_0\right]\right\|_2^2\right]\le C_3t^5=\mathcal{O}(d^3t^5),\quad \textit{for all } 0\le t\le 1,$$

where

$$\begin{aligned}C_3=&4\left(C_1^{1/2}\left(1+\mathcal{U}_4'\right)^{1/2}\left(\mu_2(f)^2\pi_{3,4}(f)^{1/2}+\mu_3(f)^2\pi_{2,4}(f)^{1/2}\right)+C_0d\mu_4(f)^2\right)\\ &+\frac{1}{4}\mu_3(f)^2\pi_{2,4}(f)\left(1+\mathcal{U}_4\right)+8\mu_4(f)^2\left(\pi_{2,6}(f)\left(1+\mathcal{U}_6\right)+73(d+4)^3\right).\end{aligned}$$

Proof. The proof is similiar to that of Lemma 10 with slight variations on truncating the expansions. Recall since the two processes share the same Brownian motion,

$$X_t-\tilde{X}_t=-\int_0^t\nabla f(X_s)\,\mathrm{d}s+\frac{t}{2}\left(\nabla f(\tilde{H}_1)+\nabla f(\tilde{H}_2)\right).$$

By Itô's lemma,

$$\begin{aligned}\nabla f(X_s)=&\nabla f(X_0)-\int_0^s\left(\nabla^2 f(X_u)\nabla f(X_u)-\vec{\Delta}\left(\nabla f\right)(X_u)\right)\,\mathrm{d}u+\sqrt{2}\int_0^s\nabla^2 f(X_u)\,\mathrm{d}B_u\\ =&\nabla f(X_0)-\nabla^2 f(X_0)\nabla f(X_0)s+\sqrt{2}\nabla^2 f(X_0)B_s+\vec{\Delta}(\nabla f)(X_0)s+\bar{R}(s),\end{aligned}$$

where the remainder is

$$\begin{aligned}\bar{R}(s)=&\underbrace{\int_0^s\left(-\nabla^2 f(X_u)\nabla f(X_u)+\nabla^2 f(X_0)\nabla f(X_0)\right)\,\mathrm{d}u}_{\bar{R}_1(s)}\\ &+\underbrace{\int_0^s\left(\vec{\Delta}\left(\nabla f\right)(X_u)-\vec{\Delta}\left(\nabla f\right)(X_0)\right)\,\mathrm{d}u}_{\bar{R}_2(s)}\end{aligned}$$

$$+\underbrace{\sqrt{2}\int_0^s \left(\nabla^2 f(X_u)-\nabla^2 f(X_0)\right)\,\mathrm{d}B_u}_{\bar{R}_3(s)}.$$

By Taylor's theorem with the remainder in integral form,

$$\nabla f(\tilde{H}_1)=\nabla f(X_0)+\nabla^2 f(X_0)\Delta\tilde{H}_1+\frac{1}{2}\nabla^3 f(X_0)[\Delta\tilde{H}_1,\,\Delta\tilde{H}_1]+\bar{\rho}_1(t),$$
$$\nabla f(\tilde{H}_2)=\nabla f(X_0)+\nabla^2 f(X_0)\Delta\tilde{H}_2+\frac{1}{2}\nabla^3 f(X_0)[\Delta\tilde{H}_2,\,\Delta\tilde{H}_2]+\bar{\rho}_2(t),$$

where

$$\bar{\rho}_1(t)=\frac{1}{2}\int_0^1(1-\tau)^2\nabla^4 f(X_0+\tau\Delta\tilde{H}_1)[\Delta\tilde{H}_1,\,\Delta\tilde{H}_1,\,\Delta\tilde{H}_1]\,\mathrm{d}\tau,$$
$$\bar{\rho}_2(t)=\frac{1}{2}\int_0^1(1-\tau)^2\nabla^4 f(X_0+\tau\Delta\tilde{H}_2)[\Delta\tilde{H}_2,\,\Delta\tilde{H}_2,\,\Delta\tilde{H}_2]\,\mathrm{d}\tau.$$

Now, we show the following equality in a component-wise manner,

$$\begin{aligned}\frac{t^2}{2}\mathbb{E}\left[\vec{\Delta}\left(\nabla f\right)(X_0)\right]+\frac{t^3}{4}\mathbb{E}\left[\nabla^3 f(X_0)[\nabla f(X_0),\,\nabla f(X_0)]\right]=\\ \frac{t}{4}\mathbb{E}\left[\nabla^3 f(X_0)[\Delta\tilde{H}_1,\,\Delta\tilde{H}_1]\right]+\frac{t}{4}\mathbb{E}\left[\nabla^3 f(X_0)[\Delta\tilde{H}_2,\,\Delta\tilde{H}_2]\right].\end{aligned}\tag{35}$$

To see this, recall that odd moments of the Brownian motion is zero. So, for each $\partial_i f$,

$$\begin{aligned}\mathbb{E}\left[\left\langle\Delta\tilde{H}_1,\nabla^2(\partial_i f)(X_0)\Delta\tilde{H}_1\right\rangle\right]=&\mathbb{E}\left[\mathbb{E}\left[\mathrm{Tr}\left((\Delta\tilde{H}_1)^\top\Delta\tilde{H}_1\nabla^2(\partial_i f)(X_0)\right)|\mathcal{F}_0\right]\right]\\ =&\mathbb{E}\left[\mathrm{Tr}\left(\mathbb{E}\left[(\Delta\tilde{H}_1)^\top\Delta\tilde{H}_1|\mathcal{F}_0\right]\nabla^2(\partial_i f)(X_0)\right)\right]\\ =&2t\left(\frac{1}{2}+\frac{1}{\sqrt{6}}\right)\mathbb{E}\left[\Delta(\partial_i f)(X_0)\right].\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}\left[\left\langle\Delta\tilde{H}_2,\nabla^2(\partial_i f)(X_0)\Delta\tilde{H}_2\right\rangle\right]=&\mathbb{E}\left[\mathbb{E}\left[\mathrm{Tr}\left((\Delta\tilde{H}_2)^\top\Delta\tilde{H}_2\nabla^2(\partial_i f)(X_0)\right)|\mathcal{F}_0\right]\right]\\ =&\mathbb{E}\left[\mathrm{Tr}\left(\mathbb{E}\left[(\Delta\tilde{H}_2)^\top\Delta\tilde{H}_2|\mathcal{F}_0\right]\nabla^2\partial_i f(X_0)\right)\right]\\ =&2t\left(\frac{1}{2}-\frac{1}{\sqrt{6}}\right)\mathbb{E}\left[\Delta(\partial_i f)(X_0)\right]\\ &+t^2\mathbb{E}\left[\left\langle\nabla f(X_0),\nabla^2(\partial_i f)(X_0)\nabla f(X_0)\right\rangle\right].\end{aligned}$$

Adding the previous two equations together, we obtain the desired equality (35).

Next, we bound the second moments of $\bar{R}_1(s)$ and $\bar{R}_2(s)$. For $\bar{R}_1(s)$, recall from the proof of Lemma 10,

$$\mathbb{E}\left[\left\|\bar{R}_1(s)\right\|_2^2\right]=\mathbb{E}\left[\left\|R_1(s)\right\|_2^2\right]\le C_1^{1/2}\left(1+\mathcal{U}_4'\right)^{1/2}\left(\mu_2(f)^2\pi_{3,4}(f)^{1/2}+\mu_3(f)^2\pi_{2,4}(f)^{1/2}\right)s^3.$$

Additionally for $\bar{R}_2(s)$,

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{R}_2(s)\right\|_2^2\right]=&\mathbb{E}\left[\left\|\int_0^s\left(\vec{\Delta}\left(\nabla f\right)(X_u)-\vec{\Delta}\left(\nabla f\right)(X_0)\right)\mathrm{d}u\right\|_2^2\right]\\ \le&s\int_0^s\mathbb{E}\left[\left\|\vec{\Delta}\left(\nabla f\right)(X_u)-\vec{\Delta}\left(\nabla f\right)(X_0)\right\|_2^2\right]\mathrm{d}u\\ \le&d^2\mu_4(f)^2 s\int_0^s\mathbb{E}\left[\left\|X_u-X_0\right\|_2^2\right]\mathrm{d}u\\ \le&C_0 d^2\mu_4(f)^2 s\int_0^s u\,\mathrm{d}u\end{aligned}$$

$$\le C_0 d^2 \mu_4(f)^2 \frac{s^3}{2}.$$

Since $\bar{R}_3(s)$ is a Martingale,

$$\begin{aligned}
\left\| \mathbb{E}\left[\int_0^t \bar{R}(s)\ \mathrm{d}s|\mathcal{F}_0\right]\right\|_2^2 =& \left\| \mathbb{E}\left[\int_0^t \bar{R}_1(s)\ \mathrm{d}s|\mathcal{F}_0\right] + \mathbb{E}\left[\int_0^t \bar{R}_2(s)\ \mathrm{d}s|\mathcal{F}_0\right]\right\|_2^2 \\
\le& 2\left\| \mathbb{E}\left[\int_0^t \bar{R}_1(s)\ \mathrm{d}s|\mathcal{F}_0\right]\right\|_2^2 + 2\left\| \mathbb{E}\left[\int_0^t \bar{R}_2(s)\ \mathrm{d}s|\mathcal{F}_0\right]\right\|_2^2 \\
\le& 2t\int_0^t \mathbb{E}\left[\left\|\bar{R}_1(s)\right\|_2^2 + \left\|\bar{R}_2(s)\right\|_2^2 |\mathcal{F}_0\right]\ \mathrm{d}s.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\left\| \mathbb{E}\left[\int_0^t \bar{R}(s)\ \mathrm{d}s|\mathcal{F}_0\right]\right\|_2^2\right] \le& 2t\int_0^t \mathbb{E}\left[\left\|\bar{R}_1(s)\right\|_2^2 + \left\|\bar{R}_2(s)\right\|_2^2\right]\ \mathrm{d}s \\
\le& C_1^{1/2}\left(1+\mathcal{U}_4'\right)^{1/2}\left(\mu_2(f)^2\pi_{3,4}(f)^{1/2} + \mu_3(f)^2\pi_{2,4}(f)^{1/2}\right)t^5 \\
&+ C_0 d\mu_4(f)^2 t^5.
\end{aligned}$$

Next, we bound the sixth moments of $\Delta\tilde{H}_1$ and $\Delta\tilde{H}_2$. Note for two random vectors a and b, by Young inequality and Lemma 31, we have

$$\mathbb{E}\left[\|a+b\|_2^6\right] \le \mathbb{E}\left[\left(2\|a\|_2^2 + 2\|b\|_2^2\right)^3\right] \le 32\mathbb{E}\left[\|a\|_2^6 + \|b\|_2^6\right].$$

To simplify notation, we define

$$\begin{aligned}
v_1 &= \sqrt{2}\left(\frac{1}{2}+\frac{1}{\sqrt{6}}\right)\xi\sqrt{t}, \quad v_1' = \sqrt{2}\left(\frac{1}{2}-\frac{1}{\sqrt{6}}\right)\xi\sqrt{t}, \\
v_2 &= \frac{1}{\sqrt{6}}\eta\sqrt{t} \quad \text{where} \quad \xi,\eta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d),
\end{aligned}$$

We bound the sixth moments of v_1, v_1' and v_2 using $1/2 + 1/\sqrt{6} < 1$, $1/2 - 1/\sqrt{6} < 1/2$ and the closed form moments of a chi-squared random variable with d degrees of freedom $\chi(d)^2$ [57],

$$\begin{aligned}
\mathbb{E}\left[\|v_1\|_2^6\right] \le& 8\mathbb{E}\left[\|\xi\|_2^6\right]t^3 = 8\mathbb{E}\left[\chi(d)^6\right]t^3 = 8d(d+2)(d+4)t^3 < 8(d+4)^3t^3, \\
\mathbb{E}\left[\|v_1'\|_2^6\right] \le& \mathbb{E}\left[\|\xi\|_2^6\right]t^3 = \mathbb{E}\left[\chi(d)^6\right]t^3 = d(d+2)(d+4)t^3 < (d+4)^3t^3, \\
\mathbb{E}\left[\|v_2\|_2^6\right] =& \frac{1}{216}\mathbb{E}\left[\|\eta\|_2^6\right]t^3 = \frac{1}{216}\mathbb{E}\left[\chi(d)^6\right]t^3 = \frac{1}{216}d(d+2)(d+4)t^3 < \frac{1}{216}(d+4)^3t^3.
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{E}\left[\left\|\Delta\tilde{H}_1\right\|_2^6\right] =& \mathbb{E}\left[\|v_1+v_2\|_2^6\right] \le 32\mathbb{E}\left[\|v_1\|_2^6 + \|v_2\|_2^6\right] \le 288(d+4)^3t^3, \\
\mathbb{E}\left[\left\|\Delta\tilde{H}_2\right\|_2^6\right] =& \mathbb{E}\left[\|-\nabla f(X_0)t + v_1' + v_2\|_2^6\right] \\
\le& 32\mathbb{E}\left[\|\nabla f(X_0)t\|_2^6\right]t^6 + 32\mathbb{E}\left[\|v_1'+v_2\|_2^6\right] \\
\le& 32\pi_{2,6}(f)\left(1+\mathbb{E}\left[\|X_0\|_2^6\right]\right)t^6 + 1024\mathbb{E}\left[\|v_1'\|_2^6 + \|v_2\|_2^6\right] \\
\le& 32\pi_{2,6}(f)\left(1+\mathcal{U}_6\right)t^3 + 2048(d+4)^3t^3 \\
\le& 32\left(\pi_{2,6}(f)\left(1+\mathcal{U}_6\right) + 64(d+4)^3\right)t^3.
\end{aligned}$$

Now, we bound the second moments of $\bar{\rho}_1(t)$ and $\bar{\rho}_2(t)$ using the derived sixth-moment bounds,

$$\mathbb{E}\left[\|\bar{\rho}_1(t)\|_2^2\right] = \mathbb{E}\left[\left\|\frac{1}{2}\int_0^1 (1-\tau)^2\nabla^4 f(X_0+\tau\Delta\tilde{H}_1)[\Delta\tilde{H}_1,\, \Delta\tilde{H}_1,\, \Delta\tilde{H}_1]\right\|_2^2\right]$$

$$
\begin{aligned}
&\leq \frac{1}{4} \sup_{z\in\mathbb{R}^d} \left\|\nabla^4 f(z)\right\|_{\mathrm{op}}^2 \mathbb{E}\left[\left\|\Delta\tilde{H}_1\right\|_2^6\right] \\
&\leq 72\mu_4(f)^2(d+4)^3 t^3.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\mathbb{E}\left[\|\bar{\rho}_2(t)\|_2^2\right] =& \mathbb{E}\left[\left\|\frac{1}{2}\int_0^1 (1-\tau)^2\nabla^4 f(X_0+\tau\Delta\tilde{H}_2)[\Delta\tilde{H}_2,\, \Delta\tilde{H}_2,\, \Delta\tilde{H}_2]\right\|_2^2\right] \\
\leq& \frac{1}{4} \sup_{z\in\mathbb{R}^d} \left\|\nabla^4 f(z)\right\|_{\mathrm{op}}^2 \mathbb{E}\left[\left\|\Delta\tilde{H}_2\right\|_2^6\right] \\
\leq& 8\mu_4(f)^2\left(\pi_{2,6}(f)\left(1+\mathcal{U}_6\right)+64(d+4)^3\right)t^3.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbb{E}\left[X_t-\tilde{X}_t|\mathcal{F}_0\right]\right\|_2^2\right] \\
=&\mathbb{E}\left[\left\|\mathbb{E}[-\int_0^t \bar{R}(s)\ \mathrm{d}s + \frac{t^3}{4}\nabla^3 f(X_0)[\nabla f(X_0),\, \nabla f(X_0)] + \frac{t}{2}\bar{\rho}_1(t) + \frac{t}{2}\bar{\rho}_2(t)|\mathcal{F}_0]\right\|^2\right] \\
\leq& 4\mathbb{E}\left[\left\|\mathbb{E}\left[\int_0^t \bar{R}(s)\ \mathrm{d}s|\mathcal{F}_0\right]\right\|_2^2\right] + \frac{t^6}{4}\mathbb{E}\left[\left\|\nabla^3 f(X_0)[\nabla f(X_0),\, \nabla f(X_0)]\right\|_2^2\right] \\
&+t^2\mathbb{E}\left[\|\bar{\rho}_1(t)\|_2^2+\|\bar{\rho}_2(t)\|_2^2\right] \\
\leq& 4\left(C_1^{1/2}\left(1+\mathcal{U}_4'\right)^{1/2}\left(\mu_2(f)^2\pi_{3,4}(f)^{1/2}+\mu_3(f)^2\pi_{2,4}(f)^{1/2}\right)+C_0 d\mu_4(f)^2\right)t^5 \\
&+\frac{1}{4}\mu_3(f)^2\mathbb{E}\left[\|\nabla f(X_0)\|_2^4\right]t^6 \\
&+72\mu_4(f)^2(d+4)^3t^5+8\mu_4(f)^2\left(\pi_{2,6}\left(1+\mathcal{U}_6\right)+64(d+4)^3\right)t^5 \\
\leq& 4\left(C_1^{1/2}\left(1+\mathcal{U}_4'\right)^{1/2}\left(\mu_2(f)^2\pi_{3,4}(f)^{1/2}+\mu_3(f)^2\pi_{2,4}(f)^{1/2}\right)+C_0 d\mu_4(f)^2\right)t^5 \\
&+\frac{1}{4}\mu_3(f)^2\pi_{2,4}(f)\left(1+\mathcal{U}_4\right)t^5 \\
&+8\mu_4(f)^2\left(\pi_{2,6}(f)\left(1+\mathcal{U}_6\right)+73(d+4)^3\right)t^5 \\
\leq& C_3 t^5.
\end{aligned}
$$

□

B.3 Invoking Theorem 1

Now, we invoke Theorem 1 with our derived constants. We obtain that if the constant step size

$$
h < 1 \wedge C_h \wedge \frac{1}{2\alpha} \wedge \frac{1}{8\mu_1(b)^2+8\mu_1^{\mathrm{F}}(\sigma)^2},
$$

where

$$
C_h = \frac{2d}{\pi_{2,2}(f)} \wedge \frac{2\pi_{2,1}(f)}{\pi_{2,2}(f)} \wedge \frac{\alpha}{4\mu_2(f)\pi_{2,2}(f)} \wedge \frac{3\alpha}{2N_1+2N_2+4} \wedge \min\left\{\left(\frac{3\alpha l}{8N_{11,l}}\right)^2 : l=2,3\right\},
$$

and the smoothness conditions on the strongly convex potential in Theorem 2 holds, then the uniform local deviation bounds (7) hold with $\lambda_1 = C_2$ and $\lambda_2 = C_3$, and consequently the bound (8) holds. This concludes that to converge to a sufficiently small positive tolerance ϵ, $\tilde{\mathcal{O}}(d\epsilon^{-2/3})$ iterations are required, since C_2 is of order $\mathcal{O}(d^2)$, and C_3 is of order $\mathcal{O}(d^3)$.

C Proof of Theorem 3

C.1 Moment Bounds

Verifying the order conditions in Theorem 1 for SRK-ID requires bounding the second and fourth moments of the Markov chain.

The following proofs only assume Lipschitz smoothness of the drift coefficient b and diffusion coefficient σ to a certain order and a generalized notion of dissipativity for Itô diffusions.

Definition C.1 (Dissipativity). *For constants $\alpha, \beta > 0$, the diffusion satisfies the following*

$$-2 \langle b(x), x \rangle - \|\sigma(x)\|_{\mathrm{F}}^2 \geq \alpha \|x\|_2^2 - \beta, \quad \textit{for all } x \in \mathbb{R}^d.$$

For general Itô diffusions, dissipativity directly follows from uniform dissipativity, where β is an appropriate constant of order $\mathcal{O}(d)$. Additionally, we assume the discretization has a constant step size h and the timestamp of the kth iterate is t_k as per the proof of Theorem 1. To simplify notation, we rewrite the update as

$$\tilde{X}_{k+1} = \tilde{X}_k + b(\tilde{X}_k)h + \sigma(\tilde{X}_k)\xi_{k+1}h^{1/2} + \tilde{Y}_{k+1}, \quad \xi_{k+1} \sim \mathcal{N}(0, I_d),$$

where

$$\tilde{Y}_{k+1}^{(i)} = \left(\sigma_i(\tilde{H}_1^{(i)}) - \sigma_i(\tilde{H}_2^{(i)})\right) h^{1/2}, \quad \tilde{Y}_{k+1} = \frac{1}{2}\sum_{i=1}^{m} \tilde{Y}_{k+1}^{(i)}.$$

Note that ξ_{k+1} and $\tilde{Y}_{k+1}$ are not independent, since we model $I_{(\cdot)} = (I_{(1)}, \dots, I_{(m)})^\top$ as $\xi_{k+1}h^{1/2}$. Moreover, we define the following notation

$$I_{(\cdot,i)} = (I_{(1,i)}, \dots, I_{(m,i)})^\top, \quad \Delta\tilde{H}^{(i)} = \sigma(\tilde{X}_k)I_{(\cdot,i)}h^{-1/2}, \quad i = 1, \dots, m.$$

Hence, the variables $\tilde{H}_1^{(i)}$ and $\tilde{H}_2^{(i)}$ can be written as

$$\tilde{H}_1^{(i)} = \tilde{X}_k + \Delta\tilde{H}^{(i)}, \quad \tilde{H}_2^{(i)} = \tilde{X}_k - \Delta\tilde{H}^{(i)}.$$

We first bound the second moments of $\tilde{Y}_k$, using the following moment inequality.

Theorem 12 ([41, Sec. 1.7, Thm. 7.1]). *Let $p \geq 2$. If $\{G_s\}_{s\geq 0}$ is a $d\times m$ matrix-valued process, and $\{B_t\}_{t\geq 0}$ is a d-dimensional Brownian motion, both of which are adapted to the filtration $\{\mathcal{F}_s\}_{s\geq 0}$ such that for some fixed $t > 0$, the following relation holds*

$$\mathbb{E}\left[\int_0^t \|G_s\|_{\mathrm{F}}^p \,\mathrm{d}s\right] < \infty.$$

Then,

$$\mathbb{E}\left[\left\|\int_0^t G_s \,\mathrm{d}B_s\right\|_2^p\right] \leq \left(\frac{p(p-1)}{2}\right)^{p/2} t^{(p-2)/2}\mathbb{E}\left[\int_0^t \|G_s\|_{\mathrm{F}}^p \,\mathrm{d}s\right].$$

In particular, equality holds when $p = 2$.

The above theorem can be proved directly using Itô's lemma and Itô isometry, with the help of Hölder's inequality. The theorem can also be seen as a natural consequence of the Burkholder-Davis-Gundy Inequality [41].

Corollary 13. *Let even integer $p \geq 2$. Then, the following relation holds*

$$\mathbb{E}\left[\left\|\Delta\tilde{H}^{(i)}\right\|_2^p |\mathcal{F}_{t_k}\right] \leq \left(\frac{p(p-1)}{2}\right)^p \pi_{1,p}^{\mathrm{F}}(\sigma)\left(1 + \left\|\tilde{X}_k\right\|_2^{p/2}\right) h^{p/2}.$$

Proof. It is clear that the integrability condition in Theorem 12 holds for the inner and outer integrals of $\Delta\tilde{H}^{(i)}$. Hence, by repeatedly applying the theorem,

$$\begin{aligned}
\mathbb{E}\left[\left\|\Delta\tilde{H}^{(i)}\right\|_2^p |\mathcal{F}_{t_k}\right] =& \mathbb{E}\left[\left\|\sigma(\tilde{X}_k)I_{(\cdot,i)}\right\|_2^p |\mathcal{F}_{t_k}\right] h^{-p/2} \\
=& \mathbb{E}\left[\left\|\int_{t_k}^{t_{k+1}}\int_{t_k}^{s} \sigma(\tilde{X}_k) \,\mathrm{d}B_u \,\mathrm{d}B_s^{(i)}\right\|_2^p |\mathcal{F}_{t_k}\right] h^{-p/2} \\
\leq& \left(\frac{p(p-1)}{2}\right)^{p/2} h^{-1}\int_{t_k}^{t_{k+1}} \mathbb{E}\left[\left\|\int_{t_k}^{s} \sigma(\tilde{X}_k) \,\mathrm{d}B_u\right\|_2^p |\mathcal{F}_{t_k}\right] \mathrm{d}s
\end{aligned}$$

$$\leq \left(\frac{p(p-1)}{2}\right)^p h^{-1} \int_{t_k}^{t_{k+1}} s^{(p-2)/2} \int_{t_k}^{s} \mathbb{E}\left[\left\|\sigma(\tilde{X}_k)\right\|_{\mathrm{F}}^p |\mathcal{F}_{t_k}\right] \mathrm{d}u \ \mathrm{d}s$$
$$\leq \left(\frac{p(p-1)}{2}\right)^p \pi_{1,p}^{\mathrm{F}}(\sigma) \left(1 + \left\|\tilde{X}_k\right\|_2^{p/2}\right) h^{p/2}.$$

□

Lemma 14 (Second Moment Bounds for $\tilde{Y}_k$)**.** *The following relation holds*

$$\mathbb{E}\left[\left\|\tilde{Y}_{k+1}\right\|_2^2 |\mathcal{F}_{t_k}\right] \leq 2^2 3^4 m^2 \mu_2(\sigma)^2 \pi_{1,4}^{\mathrm{F}}(\sigma) \left(1 + \left\|\tilde{X}_k\right\|_2^2\right) h^3.$$

Proof. By Taylor's Theorem with the remainder in integral form,

$$\begin{aligned}
\left\|\tilde{Y}_{k+1}^{(i)}\right\|_2 &= \left\|\sigma_i(\tilde{X}_k + \Delta\tilde{H}^{(i)}) - \sigma_i(\tilde{X}_k - \Delta\tilde{H}^{(i)})\right\|_2 h^{1/2} \\
&= \left\|\int_0^1 \left(\nabla\sigma_i(\tilde{X}_k + \tau\Delta\tilde{H}^{(i)}) - \nabla\sigma_i(\tilde{X}_k - \tau\Delta\tilde{H}^{(i)})\right) \Delta\tilde{H}^{(i)} \ \mathrm{d}\tau\right\|_2 h^{1/2} \\
&\leq h^{1/2} \int_0^1 \left\|\nabla\sigma_i(\tilde{X}_k + \tau\Delta\tilde{H}^{(i)}) - \nabla\sigma_i(\tilde{X}_k - \tau\Delta\tilde{H}^{(i)})\right\|_{\mathrm{op}} \left\|\Delta\tilde{H}^{(i)}\right\|_2 \mathrm{d}\tau \\
&\leq \mu_2(\sigma) h^{1/2} \left\|\Delta\tilde{H}^{(i)}\right\|_2^2 \int_0^1 2\tau \ \mathrm{d}\tau \\
&\leq \mu_2(\sigma) h^{1/2} \left\|\Delta\tilde{H}^{(i)}\right\|_2^2.
\end{aligned} \tag{36}$$

By (36) and Corollary 13,

$$\mathbb{E}\left[\left\|\tilde{Y}_{k+1}^{(i)}\right\|_2^2 |\mathcal{F}_{t_k}\right] \leq \mu_2(\sigma)^2 \mathbb{E}\left[\left\|\Delta\tilde{H}^{(i)}\right\|_2^4 |\mathcal{F}_{t_k}\right] h \leq 6^4 \mu_2(\sigma)^2 \pi_{1,4}^{\mathrm{F}}(\sigma) \left(1 + \left\|\tilde{X}_k\right\|_2^2\right) h^3$$

Therefore,

$$\mathbb{E}\left[\left\|\tilde{Y}_{k+1}\right\|_2^2 |\mathcal{F}_{t_k}\right] \leq \frac{m}{4} \sum_{i=1}^m \mathbb{E}\left[\left\|\tilde{Y}_{k+1}^{(i)}\right\|_2^2 |\mathcal{F}_{t_k}\right] \leq 2^2 3^4 m^2 \mu_2(\sigma)^2 \pi_{1,4}^{\mathrm{F}}(\sigma) \left(1 + \left\|\tilde{X}_k\right\|_2^2\right) h^3.$$

□

To prove the following moment bound lemmas for SRK-ID, we recall a standard quadratic moment bound result whose proof we omit and provide a reference of.

Lemma 15 ([24, Lemma F.1])**.** *Let even integer $p \geq 2$ and $f : \mathbb{R}^d \to \mathbb{R}^{d\times m}$ be Lipschitz. For $\xi \sim \mathcal{N}(0, I_m)$ independent from the d-dimensional random vector X, the following relation holds*

$$\mathbb{E}\left[\|f(X)\xi\|_2^p\right] \leq (p-1)!!\mathbb{E}\left[\|f(X)\|_{\mathrm{F}}^p\right].$$

C.1.1 Second Moment Bound

Lemma 16. *If the second moment of the initial iterate is finite, then the second moments of Markov chain iterates defined in* (10) *are uniformly bounded, i.e.*

$$\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^2\right] \leq \mathcal{V}_2, \quad \textit{for all } k \in \mathbb{N}$$

where

$$\mathcal{V}_2 = \mathbb{E}\left[\left\|\tilde{X}_0\right\|_2^2\right] + M_2,$$

and constants M_1 and M_2 are given in the proof, if the constant step size

$$h < 1 \wedge \frac{1}{m^2} \wedge \frac{\alpha^2}{4M_1^2}.$$

Proof. By direct computation,

$$\begin{aligned}\left\|\tilde{X}_{k+1}\right\|_2^2 = &\left\|\tilde{X}_k\right\|_2^2 + \left\|b(\tilde{X}_k)\right\|_2^2 h^2 + \left\|\sigma(\tilde{X}_k)\xi_{k+1}\right\|_2^2 h + \left\|\tilde{Y}_{k+1}\right\|_2^2 \\ &+ 2\left\langle \tilde{X}_k, b(\tilde{X}_k)\right\rangle h + 2\left\langle \tilde{X}_k, \sigma(\tilde{X}_k)\xi_{k+1}\right\rangle h^{1/2} + 2\left\langle \tilde{X}_k, \tilde{Y}_{k+1}\right\rangle \\ &+ 2\left\langle b(\tilde{X}_k), \sigma(\tilde{X}_k)\xi_{k+1}\right\rangle h^{3/2} + 2\left\langle b(\tilde{X}_k), \tilde{Y}_{k+1}\right\rangle h \\ &+ 2\left\langle \sigma(\tilde{X}_k)\xi_{k+1}, \tilde{Y}_{k+1}\right\rangle h^{1/2}.\end{aligned}$$

By Lemma 15 and dissipativity,

$$\begin{aligned}\mathbb{E}\left[2\left\langle \tilde{X}_k, b(\tilde{X}_k)\right\rangle h + \left\|\sigma(\tilde{X}_k)\xi_{k+1}\right\|_2^2 h|\mathcal{F}_{t_k}\right] =& 2\left\langle \tilde{X}_k, b(\tilde{X}_k)\right\rangle h + \left\|\sigma(\tilde{X}_k)\right\|_{\mathrm{F}}^2 h \\ \leq& -\alpha\left\|\tilde{X}_k\right\|_2^2 h + \beta h.\end{aligned}$$

We bound the remaining terms by direct computation. By linear growth,

$$\left\|b(\tilde{X}_k)\right\|_2^2 h^2 \leq \pi_{1,2}(b)\left(1 + \left\|\tilde{X}_k\right\|_2^2\right) h^2.$$

By Lemma 14, for $h < 1 \wedge 1/m^2$,

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{Y}_{k+1}\right\|_2^2 |\mathcal{F}_{t_k}\right] \leq& 2^2 3^4 m^2 \mu_2(\sigma)^2 \pi_{1,4}^{\mathrm{F}}(\sigma)\left(1 + \left\|\tilde{X}_k\right\|_2^2\right) h^3 \\ \leq& 2^2 3^4 m \mu_2(\sigma)^2 \pi_{1,4}^{\mathrm{F}}(\sigma)\left(1 + \left\|\tilde{X}_k\right\|_2^2\right) h^{3/2}.\end{aligned}$$

By Lemma 14,

$$\begin{aligned}\mathbb{E}\left[\left\langle \tilde{X}_k, \tilde{Y}_{k+1}\right\rangle |\mathcal{F}_{t_k}\right] \leq& \left\|\tilde{X}_k\right\|_2 \mathbb{E}\left[\left\|\tilde{Y}_{k+1}\right\|_2 |\mathcal{F}_{t_k}\right] \\ \leq& \left\|\tilde{X}_k\right\|_2 \mathbb{E}\left[\left\|\tilde{Y}_{k+1}\right\|_2^2 |\mathcal{F}_{t_k}\right]^{1/2} \\ \leq& 2^2 3^2 m \mu_2(\sigma) \pi_{1,4}^{\mathrm{F}}(\sigma)^{1/2}\left(1 + \left\|\tilde{X}_k\right\|_2^2\right) h^{3/2}.\end{aligned}$$

Similarly, by Lemma 14,

$$\begin{aligned}\mathbb{E}\left[\left\langle b(\tilde{X}_k), \tilde{Y}_{k+1}\right\rangle |\mathcal{F}_{t_k}\right] \leq& \left\|b(\tilde{X}_k)\right\|_2 \mathbb{E}\left[\left\|\tilde{Y}_{k+1}\right\|_2 |\mathcal{F}_{t_k}\right] \\ \leq& \left\|b(\tilde{X}_k)\right\|_2 \mathbb{E}\left[\left\|\tilde{Y}_{k+1}\right\|_2^2 |\mathcal{F}_{t_k}\right]^{1/2} \\ \leq& 2^2 3^2 m \mu_2(\sigma) \pi_{1,4}^{\mathrm{F}}(\sigma)^{1/2} \pi_{1,1}(b)\left(1 + \left\|\tilde{X}_k\right\|_2^2\right) h^{3/2}.\end{aligned}$$

By Lemma 14 and Lemma 15,

$$\begin{aligned}\mathbb{E}\left[\left\langle \sigma(\tilde{X}_k)\xi_{k+1}, \tilde{Y}_{k+1}\right\rangle |\mathcal{F}_{t_k}\right] \leq& \mathbb{E}\left[\left\|\sigma(\tilde{X}_k)\xi_{k+1}\right\|_2 \left\|\tilde{Y}_{k+1}\right\|_2 |\mathcal{F}_{t_k}\right] \\ \leq& \mathbb{E}\left[\left\|\sigma(\tilde{X}_k)\xi_{k+1}\right\|_2^2 |\mathcal{F}_{t_k}\right]^{1/2} \mathbb{E}\left[\left\|\tilde{Y}_{k+1}\right\|_2^2 |\mathcal{F}_{t_k}\right]^{1/2} \\ \leq& \mathbb{E}\left[\left\|\sigma(\tilde{X}_k)\right\|_{\mathrm{F}}^2 |\mathcal{F}_{t_k}\right]^{1/2} \mathbb{E}\left[\left\|\tilde{Y}_{k+1}\right\|_2^2 |\mathcal{F}_{t_k}\right]^{1/2} \\ \leq& 2^2 3^2 m \mu_2(\sigma) \pi_{1,4}^{\mathrm{F}}(\sigma)^{1/2} \pi_{1,2}^{\mathrm{F}}(\sigma)^{1/2}\left(1 + \left\|\tilde{X}_k\right\|_2^2\right) h^{3/2}.\end{aligned}$$

Putting things together, for $h < 1 \wedge \alpha^2/(4M_1^2)$,

$$\mathbb{E}\left[\left\|\tilde{X}_{k+1}\right\|_2^2 | \mathcal{F}_{t_k}\right] \leq \left(1 - \alpha h + M_1 h^{3/2}\right)\left\|\tilde{X}_k\right\|_2^2 + \beta h + M_1 h^{3/2}$$
$$\leq (1 - \alpha h/2)\left\|\tilde{X}_k\right\|_2^2 + \beta h + M_1 h^{3/2},$$

where

$$M_1 = \pi_{1,2}(b) + 2^3 3^2 m \mu_2(\sigma)\pi_{1,4}^{\mathrm{F}}(\sigma)^{1/2}\left(1 + \mu_2(\sigma)\pi_{1,4}^{\mathrm{F}}(\sigma)^{1/2} + \pi_{1,1}(b) + \pi_{1,2}^{\mathrm{F}}(\sigma)^{1/2}\right).$$

Unrolling the recursion gives the following for $h < 1 \wedge 1/m^2$

$$\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^2\right] \leq \mathbb{E}\left[\left\|\tilde{X}_0\right\|_2^2\right] + 2\left(\beta + M_1 h^{1/2}\right)/\alpha$$
$$\leq \mathbb{E}\left[\left\|\tilde{X}_0\right\|_2^2\right] + M_2, \quad \text{for all } k \in \mathbb{N},$$

where

$$M_2 = 2\left(\beta + \pi_{1,2}(b)\pi_{1,2}(b) + 2^3 3^2 \mu_2(\sigma)\pi_{1,4}^{\mathrm{F}}(\sigma)^{1/2}\left(1 + \mu_2(\sigma)\pi_{1,4}^{\mathrm{F}}(\sigma)^{1/2} + \pi_{1,1}(b) + \pi_{1,2}^{\mathrm{F}}(\sigma)^{1/2}\right)\right)/\alpha.$$

□

C.1.2 $2n$th Moment Bound

Before bounding the $2n$th moments, we first generalize Lemma 14 to arbitrary even moments.

Lemma 17. *Let even integer $p \geq 2$ and $\tilde{Z}_{k+1} = \tilde{Y}_{k+1} h^{-3/2}$. Then, the following relation holds*

$$\mathbb{E}\left[\left\|\tilde{Z}_{k+1}\right\|_2^p | \mathcal{F}_{t_k}\right] \leq m^p \mu_2(\sigma)^p \left(\frac{2p(2p-1)}{2}\right)^{2p} \pi_{1,2p}^{\mathrm{F}}(\sigma)\left(1 + \left\|\tilde{X}_k\right\|_2^p\right).$$

Proof. For $i \in \{1, 2, \ldots, m\}$, by (36),

$$\left\|\tilde{Z}_{k+1}^{(i)}\right\|_2 = \tilde{Y}_{k+1}^{(i)} h^{-3/2} \leq \mu_2(\sigma) h^{-1}\left\|\Delta \tilde{H}^{(i)}\right\|_2^2.$$

Hence, by Corollary 13,

$$\mathbb{E}\left[\left\|\tilde{Z}_{k+1}^{(i)}\right\|_2^p | \mathcal{F}_{t_k}\right] \leq \mu_2(\sigma)^p h^{-p} \mathbb{E}\left[\left\|\Delta \tilde{H}^{(i)}\right\|_2^{2p} | \mathcal{F}_{t_k}\right]$$
$$\leq \mu_2(\sigma)^p \left(\frac{2p(2p-1)}{2}\right)^{2p} \pi_{1,2p}^{\mathrm{F}}(\sigma)\left(1 + \left\|\tilde{X}_k\right\|_2^p\right).$$

The remaining follows easily from Lemma 31. □

Lemma 18. *For $n \in \mathbb{N}_+$, if the $2n$th moment of the initial iterate is finite, then the $2n$th moments of Markov chain iterates defined in (10) are uniformly bounded, i.e.*

$$\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^{2n}\right] \leq \mathcal{V}_{2n}, \quad \textit{for all } k \in \mathbb{N}$$

where

$$\mathcal{V}_{2n} = \mathbb{E}\left[\left\|\tilde{X}_0\right\|_2^{2n}\right] + \frac{2}{n\alpha}\left(\beta \mathcal{V}_{2(n-1)} + 2^{23n-1} 10^n n^{8n} \pi_{1,2n}(b)\pi_{1,8n}^{\mathrm{F}}(\sigma)^{1/2}\mu_2(\sigma)^{2n}\right),$$

if the step size

$$h < 1 \wedge \frac{1}{m^2} \wedge \frac{\alpha^2}{4M_1^2} \wedge \min\left\{\left(\frac{\alpha l}{2M_{3,l}}\right)^2 : l = 2, \ldots, n\right\}.$$

Proof. Our proof is by induction. The base case is given in Lemma 16. For the inductive case, we prove that the $2n$th moment is uniformly bounded by a constant, assuming the $2(n\text{-}1)$th moment is uniformly bounded by a constant.

By the multinomial theorem,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\tilde{X}_{k+1}\right\|_2^2 |\mathcal{F}_{t_k}\right] =&\mathbb{E}\Bigg[\bigg(\left\|\tilde{X}_k\right\|_2^2 + \left\|b(\tilde{X}_k)\right\|_2^2 h^2 + \left\|\sigma(\tilde{X}_k)\xi_{k+1}\right\|_2^2 h + \left\|\tilde{Y}_{k+1}\right\|_2^2 \\
&\quad + 2\left\langle \tilde{X}_k, b(\tilde{X}_k)\right\rangle h + 2\left\langle \tilde{X}_k, \sigma(\tilde{X}_k)\xi_{k+1}\right\rangle h^{1/2} + 2\left\langle \tilde{X}_k, \tilde{Y}_{k+1}\right\rangle \\
&\quad + 2\left\langle b(\tilde{X}_k), \sigma(\tilde{X}_k)\xi_{k+1}\right\rangle h^{3/2} + 2\left\langle b(\tilde{X}_k), \tilde{Y}_{k+1}\right\rangle h \\
&\quad + 2\left\langle \sigma(\tilde{X}_k)\xi_{k+1}, \tilde{Y}_{k+1}\right\rangle h^{1/2}\bigg)^n |\mathcal{F}_{t_k}\Bigg] \\
=&\mathbb{E}\Bigg[\bigg(\left\|\tilde{X}_k\right\|_2^2 + \left\|b(\tilde{X}_k)\right\|_2^2 h^2 + \left\|\sigma(\tilde{X}_k)\xi_{k+1}\right\|_2^2 h + \left\|\tilde{Z}_{k+1}\right\|_2^2 h^3 \\
&\quad + 2\left\langle \tilde{X}_k, b(\tilde{X}_k)\right\rangle h + 2\left\langle \tilde{X}_k, \sigma(\tilde{X}_k)\xi_{k+1}\right\rangle h^{1/2} + 2\left\langle \tilde{X}_k, \tilde{Z}_{k+1}\right\rangle h^{3/2} \\
&\quad + 2\left\langle b(\tilde{X}_k), \sigma(\tilde{X}_k)\xi_{k+1}\right\rangle h^{3/2} + 2\left\langle b(\tilde{X}_k), \tilde{Z}_{k+1}\right\rangle h^{5/2} \\
&\quad + 2\left\langle \sigma(\tilde{X}_k)\xi_{k+1}, \tilde{Z}_{k+1}\right\rangle h^2\bigg)^n |\mathcal{F}_{t_k}\Bigg] \\
=&\left\|\tilde{X}_k\right\|_2^{2n} + \mathbb{E}\left[A|\mathcal{F}_{t_k}\right] h + \mathbb{E}\left[B|\mathcal{F}_{t_k}\right] h^{3/2},
\end{aligned}
$$

where by the Cauchy–Schwarz inequality,

$$
\begin{aligned}
A =& n\|\tilde{X}_k\|_2^{2(n-1)}\left(2\langle\tilde{X}_k, b(\tilde{X}_k)\rangle + \|\sigma(\tilde{X}_k)\xi_{k+1}\|_2^2\right) + 2n(n-1)\|\tilde{X}_k\|_2^{2(n-2)}\langle\tilde{X}_k, \sigma(\tilde{X}_k)\xi_{k+1}\rangle, \\
B \le& \sum_{(k_1,\dots,k_{10})\in J} 2^n \binom{n}{k_1 \;\dots\; k_{10}} \left\|\tilde{X}_k\right\|_2^{p_1} \left\|b(\tilde{X}_k)\right\|_2^{p_2} \left\|\sigma(\tilde{X}_k)\xi_{k+1}\right\|_2^{p_3} \left\|\tilde{Z}_{k+1}\right\|_2^{p_4},
\end{aligned}
$$

the indicator set

$$
\begin{aligned}
J =&\Big\{(k_1,\dots,k_{10}) \in \mathbb{N}^{10} : k_1 + \cdots + k_{10} = n, \\
&\qquad 2k_2 + k_3 + 3k_4 + k_5 + \frac{k_6}{2} + \frac{3k_7}{2} + \frac{3k_8}{2} + \frac{5k_9}{2} + 2k_{10} > 1\Big\},
\end{aligned}
$$

and with slight abuse of notation, we hide the explicit dependence on $k_1, \dots, k_{10}$ for the exponents

$$
\begin{aligned}
p_1 =& 2k_1 + k_5 + k_6 + k_7, \\
p_2 =& 2k_2 + k_5 + k_8 + k_9, \\
p_3 =& 2k_3 + k_6 + k_8 + k_{10}, \\
p_4 =& 2k_4 + k_7 + k_9 + k_{10}.
\end{aligned}
$$

By dissipativity,

$$
\mathbb{E}\left[A|\mathcal{F}_{t_k}\right] \le -n\alpha \left\|\tilde{X}_k\right\|_2^{2n} + n\beta \left\|\tilde{X}_k\right\|_2^{2(n-1)}. \tag{37}
$$

Note that $p_1 + p_2 + p_3 + p_4 = 2n$. Since $h < 1 \wedge 1/m^2$, we may cancel out the m factor in some of the terms. One can verify that the only remaining term that is m-dependent is

$$
\left\langle \tilde{X}_k, \tilde{Z}_{k+1}\right\rangle = \mathcal{O}(mh^{3/2}).
$$

Using this information, Lemma 17, Lemma 15, the Cauchy–Schwarz inequality, and $p_3 + p_4 \le 2n$,

$$
\mathbb{E}\left[B|\mathcal{F}_{t_k}\right]
$$

$$
\begin{aligned}
&\le \sum_{(k_1,\ldots,k_{10})\in J} 2^n \binom{n}{k_1 \ \ldots \ k_{10}} \left\|\tilde{X}_k\right\|_2^{p_1} \left\|b(\tilde{X}_k)\right\|_2^{p_2} \mathbb{E}\left[\left\|\sigma(\tilde{X}_k)\xi_{k+1}\right\|_2^{p_3} \left\|\tilde{Z}_{k+1}m^{-1}\right\|_2^{p_4} |\mathcal{F}_{t_k}\right] m \\
&\le \sum_{(k_1,\ldots,k_{10})\in J} 2^n \binom{n}{k_1 \ \ldots \ k_{10}} \left\|\tilde{X}_k\right\|_2^{p_1} \left\|b(\tilde{X}_k)\right\|_2^{p_2} \mathbb{E}\left[\left\|\sigma(\tilde{X}_k)\xi_{k+1}\right\|_2^{2p_3} |\mathcal{F}_{t_k}\right]^{1/2} \mathbb{E}\left[\left\|\tilde{Z}_{k+1}m^{-1}\right\|_2^{2p_4} |\mathcal{F}_{t_k}\right]^{1/2} m \\
&\le \sum_{(k_1,\ldots,k_{10})\in J} 2^n \binom{n}{k_1 \ \ldots \ k_{10}} \left\|\tilde{X}_k\right\|_2^{p_1} \pi_{1,p_2}(b) \left(1+\left\|\tilde{X}_k\right\|_2^{p_2}\right) ((2p_3-1)!!)^{1/2} \pi_{1,p_3}^{\mathrm{F}}(\sigma) \left(1+\left\|\tilde{X}_k\right\|_2^{p_3}\right) \\
&\qquad \times \mu_2(\sigma)^{p_4} \left(8p_4^2\right)^{2p_4} \pi_{1,4p_4}^{\mathrm{F}}(\sigma)^{1/2} \left(1+\left\|\tilde{X}_k\right\|_2^{p_4}\right) m \\
&\le 2^n \left(1+\left\|\tilde{X}_k\right\|_2\right)^{2n} \sum_{(k_1,\ldots,k_{10})\in J} \pi_{1,p_2}(b) ((2p_3-1)!!)^{1/2} \pi_{1,p_3}^{\mathrm{F}}(\sigma) \mu_2(\sigma)^{p_4} \left(8p_4^2\right)^{2p_4} \pi_{1,4p_4}^{\mathrm{F}}(\sigma)^{1/2} \binom{n}{k_1 \ \ldots \ k_{10}} m \\
&\le 2^n \left(1+\left\|\tilde{X}_k\right\|_2\right)^{2n} \pi_{1,2n}(b) \pi_{1,2n}^{\mathrm{F}}(\sigma) \mu_2(\sigma)^{2n} \pi_{1,4n}^{\mathrm{F}}(\sigma)^{1/2} 2^{20n} n^{8n} \sum_{\substack{k_1,\ldots,k_{10}\in\mathbb{N} \\ k_1+\cdots+k_{10}=n}} \binom{n}{k_1 \ \ldots \ k_{10}} m \\
&\le 2^{23n-1} 10^n n^{8n} \pi_{1,2n}(b) \pi_{1,8n}^{\mathrm{F}}(\sigma)^{1/2} \mu_2(\sigma)^{2n} m \left(1+\left\|\tilde{X}_k\right\|_2^{2n}\right). \qquad (38)
\end{aligned}
$$

By the inductive hypothesis, (37) and (38), and $h < 1 \wedge n^2\alpha^2/(4M_{3,n}^2)$, we obtain the recursion

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[\left\|\tilde{X}_{k+1}\right\|_2^{2n} |\mathcal{F}_{t_k}\right]\right] &\le \left(1-n\alpha h + M_{3,n}h^{3/2}\right)\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^{2n}\right] + n\beta h \mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^{2(n-1)}\right] + M_3 h^{3/2} \\
&\le \left(1-n\alpha h + M_{3,n}h^{3/2}\right)\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^{2n}\right] + n\beta \mathcal{V}_{2(n-1)} h + M_{3,n} h^{3/2} \\
&\le (1-n\alpha h/2)\,\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^{2n}\right] + n\beta \mathcal{V}_{2(n-1)} h + M_{3,n} h^{3/2},
\end{aligned}
$$

where the constant $M_{3,n} = 2^{23n-1}10^n n^{8n}\pi_{1,2n}(b)\pi_{1,8n}^{\mathrm{F}}(\sigma)^{1/2}\mu_2(\sigma)^{2n}m$.

For $h < 1 \wedge 1/m^2$, by unrolling the recursion, we obtain

$$
\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^{2n}\right] \le \mathbb{E}\left[\left\|\tilde{X}_0\right\|_2^{2n}\right] + \frac{2}{n\alpha}\left(n\beta\mathcal{V}_{2(n-1)} + M_{3,n}h^{1/2}\right) \le \mathcal{V}_{2n}, \quad \text{for all } k \in \mathbb{N},
$$

where

$$
\mathcal{V}_{2n} = \mathbb{E}\left[\left\|\tilde{X}_0\right\|_2^{2n}\right] + \frac{2}{n\alpha}\left(\beta\mathcal{V}_{2(n-1)} + 2^{23n-1}10^n n^{8n}\pi_{1,2n}(b)\pi_{1,8n}^{\mathrm{F}}(\sigma)^{1/2}\mu_2(\sigma)^{2n}\right).
$$

□

C.2 Local Deviation Orders

In this section, we verify the local deviation orders for SRK-ID. The proofs are again by matching up terms in the Itô-Taylor expansion of the continuous-time process to terms in the Taylor expansion of the numerical integration scheme. Extra care needs to be taken for a tight dimension dependence.

Lemma 19. *Suppose X_t is the continuous-time process defined by (1) initiated from some iterate of the Markov chain X_0 defined by (10), then the second moment of X_t is uniformly bounded, i.e.*

$$
\mathbb{E}\left[\|X_t\|_2^2\right] \le \mathcal{V}_2', \quad \textit{for all } t \ge 0.
$$

where $\mathcal{V}_2' = \mathcal{V}_2 + \beta/\alpha$.

Proof. By Itô's lemma and dissipativity,

$$
\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\|X_t\|_2^2\right] = \mathbb{E}\left[2\langle X_t, b(X_t)\rangle + \|\sigma(X_t)\|_{\mathrm{F}}^2\right] \le -\alpha\mathbb{E}\left[\|X_t\|_2^2\right] + \beta.
$$

Moreover, by Grönwall's inequality,

$$
\mathbb{E}\left[\|X_t\|_2^2\right] \le e^{-\alpha t}\mathbb{E}\left[\|X_0\|_2^2\right] + \beta/\alpha \le \mathcal{V}_2 + \beta/\alpha = \mathcal{V}_2'.
$$

□

Lemma 20 (Second Moment of Change)**.** *Suppose X_t is the continuous-time process defined by (1) initiated from some iterate of the Markov chain X_0 defined by (10), then*

$$\mathbb{E}\left[\|X_t - X_0\|_2^2\right] \le D_0 t, \quad \textit{for all } 0 \le t \le 1,$$

where $D_0 = 2\left(\pi_{1,2}(b) + \pi_{1,2}^{\mathrm{F}}(\sigma)\right)(1 + \mathcal{V}_2')$.

Proof. By Itô isometry,

$$\begin{aligned}
\mathbb{E}\left[\|X_t - X_0\|_2^2\right] &= \mathbb{E}\left[\left\|\int_0^t b(X_s)\,\mathrm{d}s + \int_0^t \sigma(X_s)\,\mathrm{d}B_s\right\|_2^2\right] \\
&\le 2\mathbb{E}\left[\left\|\int_0^t b(X_s)\,\mathrm{d}s\right\|_2^2 + \left\|\int_0^t \sigma(X_s)\,\mathrm{d}B_s\right\|_2^2\right] \\
&\le 2t\int_0^t \mathbb{E}\left[\|b(X_s)\|_2^2\right]\mathrm{d}s + 2\int_0^t \mathbb{E}\left[\|\sigma(X_s)\|_{\mathrm{F}}^2\right]\mathrm{d}s \\
&\le 2\pi_{1,2}(b)t\int_0^t \mathbb{E}\left[1 + \|X_s\|_2^2\right]\mathrm{d}s + 2\pi_{1,2}^{\mathrm{F}}(\sigma)\int_0^t \mathbb{E}\left[1 + \|X_s\|_2^2\right]\mathrm{d}s \\
&\le 2\left(\pi_{1,2}(b) + \pi_{1,2}^{\mathrm{F}}(\sigma)\right)(1 + \mathcal{V}_2')\,t.
\end{aligned}$$

□

To bound the fourth moment of change in continuous-time, we use the following lemma.

Lemma 21 ([24, adapted from Lemma A.1])**.** *Assuming $\{X_t\}_{t\ge 0}$ is the solution to the SDE (1), under the condition that the drift coefficient b and diffusion coefficient σ are Lipschitz. If σ has satisfies the following sublinear growth condition*

$$\|\sigma(x)\|_{\mathrm{F}}^l \le \pi_{1,l}^{\mathrm{F}}(\sigma)\left(1 + \|x\|^{l/2}\right), \quad \textit{for all } x \in \mathbb{R}^d, l = 1, 2, \ldots,$$

and the diffusion is dissipative, then for $n \ge 2$, we have the following relation

$$\mathcal{A}\|x\|_2^n \le -\frac{\alpha n}{4}\|x\|_2^n + \beta_n,$$

where the (infinitesimal) generator $\mathcal{A}$ is defined as

$$\mathcal{A}f(x) = \lim_{t\downarrow 0}\frac{\mathbb{E}\left[f(X_t)|X_0 = x\right] - f(x)}{t},$$

and the constant $\beta_n = \mathcal{O}(d^{\frac{n}{2}})$.

Proof. By definition of the generator and dissipativity,

$$\begin{aligned}
\mathcal{A}\|x\|_2^n =& n\|x\|_2^{n-2}\langle x, b(x)\rangle + \frac{n}{2}\|x\|_2^{n-2}\|\sigma(x)\|_{\mathrm{F}}^2 + \frac{n(n-2)}{2}\|x\|_2^{n-4}\left\langle \mathrm{vec}(xx^\top), \mathrm{vec}(\sigma\sigma^\top(x))\right\rangle \\
\le& -\frac{\alpha n}{2}\|x\|_2^n + \frac{\beta n}{2}\|x\|_2^{n-2} + \frac{n(n-2)}{2}\|x\|_2^{n-2}\pi_{1,2}^{\mathrm{F}}(\sigma)\left(1 + \|x\|_2\right) \\
=& -\frac{\alpha n}{2}\|x\|_2^n + \frac{n(n-2)}{2}\pi_{1,2}^{\mathrm{F}}(\sigma)\|x\|_2^{n-1} + \left(\frac{\beta n}{2} + \frac{n(n-2)}{2}\pi_{1,2}^{\mathrm{F}}(\sigma)\right)\|x\|_2^{n-2}.
\end{aligned}$$

By Young's inequality,

$$\begin{aligned}
\frac{n(n-2)}{2}\pi_{1,2}^{\mathrm{F}}(\sigma)\|x\|_2^{n-1} =& \frac{n(n-2)}{2}\pi_{1,2}^{\mathrm{F}}(\sigma)\left(\frac{8}{\alpha n}\right)^{\frac{n-1}{n}} \cdot \|x\|_2^{n-1}\left(\frac{\alpha n}{8}\right)^{\frac{n-1}{n}} \\
\le& \frac{1}{n}\left(\frac{n(n-2)}{2}\right)^n \pi_{1,2}^{\mathrm{F}}(\sigma)^n\left(\frac{8}{\alpha n}\right)^{n-1} + \frac{n-1}{n}\frac{\alpha n}{8}\|x\|_2^n \\
=& \frac{(n-2)^n}{2^{2n-3}\alpha^{n-1}}\pi_{1,2}^{\mathrm{F}}(\sigma)^n + \frac{\alpha(n-1)}{8}\|x\|_2^n.
\end{aligned}$$

Similarly,

$$\left(\frac{\beta n}{2}+\frac{n(n-2)}{2}\pi_{1,2}^{\mathrm{F}}(\sigma)\right)\|x\|_2^{n-2}=\left(\frac{\beta n}{2}+\frac{n(n-2)}{2}\pi_{1,2}^{\mathrm{F}}(\sigma)\right)\left(\frac{8}{\alpha n}\right)^{\frac{n-2}{n}}\cdot\|x\|_2^{n-2}\left(\frac{\alpha n}{8}\right)^{\frac{n-2}{n}}$$
$$\le\frac{2}{n}\left(\frac{\beta n}{2}+\frac{n(n-2)}{2}\pi_{1,2}^{\mathrm{F}}(\sigma)\right)^{\frac{n}{2}}\left(\frac{\alpha n}{8}\right)^{\frac{n-2}{2}}+\frac{\alpha(n-2)}{8}\|x\|_2^n.$$

We define the following shorthand notation

$$\beta_n^{(1)}=\frac{(n-2)^n}{2^{2n-3}\alpha^{n-1}}\pi_{1,2}^{\mathrm{F}}(\sigma)^n=\mathcal{O}(d^{\frac{n}{2}}),$$
$$\beta_n^{(2)}=\frac{2}{n}\left(\frac{\beta n}{2}+\frac{n(n-2)}{2}\pi_{1,2}^{\mathrm{F}}(\sigma)\right)^{\frac{n}{2}}\left(\frac{\alpha n}{8}\right)^{\frac{n-2}{2}}=\mathcal{O}(d^{\frac{n}{2}}).$$

Putting things together, we obtain the following bound

$$\mathcal{A}\|x\|_2^n\le-\frac{\alpha n}{2}\|x\|_2^n+\frac{\alpha(n-1)}{8}\|x\|_2^n+\frac{\alpha(n-2)}{8}\|x\|_2^n+\beta_n^{(1)}+\beta_n^{(2)}$$
$$\le-\frac{\alpha n}{4}\|x\|_2^n+\beta_n,$$

where $\beta_n=\beta_n^{(1)}+\beta_n^{(2)}=\mathcal{O}(d^{\frac{n}{2}})$. □

Lemma 22. *Suppose X_t is the continuous-time process defined by (1) initiated from some iterate of the Markov chain X_0 defined by (10), then the fourth moment of X_t is uniformly bounded, i.e.*

$$\mathbb{E}\left[\|X_t\|_2^4\right]\le\mathcal{V}_4',\quad \textit{for all } t\ge 0,$$

where $\mathcal{V}_4'=\mathcal{V}_4+\beta_4/\alpha$.

Proof. By Dynkin's formula [48] applied to the function $(t,x)\mapsto e^{\alpha t}\|x\|_2^4$ and Lemma 21,

$$e^{\alpha t}\mathbb{E}\left[\|X_t\|_2^4|\mathcal{F}_0\right]=\|X_0\|_2^4+\int_0^t\mathbb{E}\left[\alpha e^{\alpha s}\|X_s\|_2^4+e^{\alpha s}\mathcal{A}\|X_s\|_2^4|\mathcal{F}_0\right]\mathrm{d}s$$
$$\le\|X_0\|_2^4+\int_0^t\mathbb{E}\left[\alpha e^{\alpha s}\|X_s\|_2^4-\alpha e^{\alpha s}\|X_s\|_2^4+e^{\alpha s}\beta_4|\mathcal{F}_0\right]\mathrm{d}s$$
$$=\|X_0\|_2^4+\frac{e^{\alpha t}-1}{\alpha}\beta_4.$$

Hence,

$$\mathbb{E}\left[\|X_t\|_2^4\right]=\mathbb{E}\left[\mathbb{E}\left[\|X_t\|_2^4|\mathcal{F}_0\right]\right]\le e^{-\alpha t}\mathbb{E}\left[\|X_0\|_2^4\right]+\beta_4/\alpha\le\mathcal{V}_4+\beta_4/\alpha=\mathcal{V}_4'.$$

□

Lemma 23 (Fourth Moment of Change)**.** *Suppose X_t is the continuous-time process defined by (1) from some iterate of the Markov chain X_0 defined by (10), then*

$$\mathbb{E}\left[\|X_t-X_0\|_2^4\right]\le D_1t^2,\quad \textit{for all } 0\le t\le 1,$$

where $D_1=8\left(\pi_{1,4}(b)+36\pi_{1,4}^{\mathrm{F}}(\sigma)\right)(1+\mathcal{V}_4')$.

Proof. By Theorem 12,

$$\mathbb{E}\left[\|X_t-X_0\|_2^4\right]=\mathbb{E}\left[\left\|\int_0^t b(X_s)\,\mathrm{d}s+\int_0^t\sigma(X_s)\,\mathrm{d}B_s\right\|_2^4\right]$$
$$\le 8\mathbb{E}\left[\left\|\int_0^t b(X_s)\,\mathrm{d}s\right\|_2^4+\left\|\int_0^t\sigma(X_s)\,\mathrm{d}B_s\right\|_2^4\right]$$
$$\le 8t^3\int_0^t\mathbb{E}\left[\|b(X_s)\|_2^4\right]\mathrm{d}s+288t\mathbb{E}\left[\int_0^t\|\sigma(X_s)\|_{\mathrm{F}}^4\,\mathrm{d}s\right]$$
$$\le 8\left(\pi_{1,4}(b)+36\pi_{1,4}^{\mathrm{F}}(\sigma)\right)(1+\mathcal{V}_4')\,t^2.$$

□

C.2.1 Local Mean-Square Deviation

Lemma 24. *Suppose X_t and $\tilde{X}_t$ are the continuous-time process defined by (1) and Markov chain defined by (10) for time $t \geq 0$, respectively. If X_t and $\tilde{X}_t$ are initiated from the same iterate of the Markov chain X_0, and they share the same Brownian motion, then*

$$\mathbb{E}\left[\left\|X_t - \tilde{X}_t\right\|_2^2\right] \leq D_3 t^3, \quad \textit{for all } 0 \leq t \leq 1,$$

where

$$\begin{aligned} D_3 = &\Big(16 D_0 \mu_1(b)^2 + \frac{16}{3}\mu_2(\sigma)^2 \pi_{1,4}^{1/2} D_1^{1/2}(1 + \mathcal{V}_2'^{1/2}) m^2 + \frac{16}{3}\mu_1(\sigma)^4 m^2 D_0 \\ &+ 16\mu_1(\sigma)^2 \pi_{1,2}(b)^2 (1 + \mathcal{V}_2') m + 4m^3 \mu_2(\sigma)^2 \pi_{1,4}^{\mathrm{F}}(\sigma)(1 + \mathcal{V}_2') \\ &+ 2^7 3^4 m^2 \mu_2(\sigma)^2 \pi_{1,4}^{\mathrm{F}}(\sigma)\left(1 + \mathcal{V}_2\right)\Big). \end{aligned}$$

Proof. Recall the operators L and Λ_i ($i = 1, \ldots, m$) defined in (5). By Itô's lemma,

$$\begin{aligned} X_t - X_0 = &\int_0^t b(X_s)\ \mathrm{d}s + \sigma(X_0) B_t \\ &+ \sum_{i=1}^m \sum_{l=1}^m \int_0^t \int_0^s \Lambda_l(\sigma_i)(X_u)\ \mathrm{d}B_u^{(l)}\ \mathrm{d}B_s^{(i)} + \sum_{i=1}^m \int_0^t \int_0^s L(\sigma_i)(X_u)\ \mathrm{d}u\ \mathrm{d}B_s^{(i)} \\ = &\int_0^t b(X_s)\ \mathrm{d}s + \sigma(X_0) B_t + \sum_{i=1}^m \sum_{l=1}^m \int_0^t \int_0^s \nabla\sigma_i(X_u)\sigma_l(X_u)\ \mathrm{d}B_u^{(l)}\ \mathrm{d}B_s^{(i)} + S(t), \end{aligned}$$

where

$$S(t) = \underbrace{\sum_{i=1}^m \int_0^t \int_0^s \nabla\sigma_i(X_u) b(X_u)\ \mathrm{d}u\ \mathrm{d}B_s^{(i)}}_{S_1(t)} + \underbrace{\frac{1}{2}\sum_{i=1}^m \sum_{l=1}^m \int_0^t \int_0^s \nabla^2\sigma_i(X_u)[\sigma_l(X_u),\, \sigma_l(X_u)]\ \mathrm{d}u\ \mathrm{d}B_s^{(i)}}_{S_2(t)}.$$

By Taylor's theorem with the remainder in integral form,

$$\begin{aligned} \sigma_i(\tilde{H}_1^{(i)}) &= \sigma_i(X_0) + \nabla\sigma_i(X_0)\Delta\tilde{H}^{(i)} + \phi_1^{(i)}(t), \\ \sigma_i(\tilde{H}_2^{(i)}) &= \sigma_i(X_0) - \nabla\sigma_i(X_0)\Delta\tilde{H}^{(i)} + \phi_2^{(i)}(t), \end{aligned}$$

where

$$\begin{aligned} \phi_1^{(i)}(t) &= \int_0^1 (1-\tau)\nabla^2\sigma_i(X_0 + \tau\Delta\tilde{H}^{(i)})[\Delta\tilde{H}^{(i)},\, \Delta\tilde{H}^{(i)}]\ \mathrm{d}\tau, \\ \phi_2^{(i)}(t) &= \int_0^1 (1-\tau)\nabla^2\sigma_i(X_0 - \tau\Delta\tilde{H}^{(i)})[\Delta\tilde{H}^{(i)},\, \Delta\tilde{H}^{(i)}]\ \mathrm{d}\tau, \\ \Delta\tilde{H}^{(i)} &= \sum_{l=1}^m \sigma_l(X_0)\frac{I_{(l,i)}}{\sqrt{t}}. \end{aligned}$$

Hence,

$$\begin{aligned} X_t - \tilde{X}_t = &\int_0^t (b(X_s) - b(X_0))\ \mathrm{d}s \\ &+ \sum_{i=1}^m \sum_{l=1}^m \int_0^t \int_0^s (\nabla\sigma_i(X_u)\sigma_l(X_u) - \nabla\sigma_i(X_0)\sigma_l(X_0))\ \mathrm{d}B_u^{(l)}\ \mathrm{d}B_s^{(i)} \\ &+ S(t) - \frac{1}{2}\sum_{i=1}^m \left(\phi_1^{(i)}(t) - \phi_2^{(i)}(t)\right)\sqrt{t}. \end{aligned}$$

Since b is $\mu_1(b)$-Lipschitz,

$$\begin{aligned}\mathbb{E}\left[\left\|\int_0^t (b(X_s)-b(X_0))\ \mathrm{d}s\right\|_2^2\right] \leq& \mu_1(b)^2 t\int_0^t \mathbb{E}\left[\|X_s-X_0\|_2^2\right]\ \mathrm{d}s\\ \leq& \mu_1(b)^2 t\int_0^t D_0 s\ \mathrm{d}s\\ \leq& \frac{1}{2}D_0\mu_1(b)^2t^3.\end{aligned}$$

We define the following,

$$A(t) = A_1(t)+A_2(t) = \sum_{i=1}^m\sum_{l=1}^m\int_0^t\int_0^s \left(\nabla\sigma_i(X_u)\sigma_l(X_u)-\nabla\sigma_i(X_0)\sigma_l(X_0)\right)\ \mathrm{d}B_u^{(l)}\ \mathrm{d}B_s^{(i)},$$

where

$$\begin{aligned}A_1(t) =& \sum_{i=1}^m\sum_{l=1}^m\int_0^t\int_0^s \left(\nabla\sigma_i(X_u)\sigma_l(X_u)-\nabla\sigma_i(X_0)\sigma_l(X_u)\right)\ \mathrm{d}B_u^{(l)}\ \mathrm{d}B_s^{(i)},\\ A_2(t) =& \sum_{i=1}^m\sum_{l=1}^m\int_0^t\int_0^s \left(\nabla\sigma_i(X_0)\sigma_l(X_u)-\nabla\sigma_i(X_0)\sigma_l(X_0)\right)\ \mathrm{d}B_u^{(l)}\ \mathrm{d}B_s^{(i)}.\end{aligned}$$

By Itô isometry and the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}\left[\|A_1(t)\|_2^2\right] =& \sum_{i=1}^m\sum_{l=1}^m\int_0^t\int_0^s \mathbb{E}\left[\|\nabla\sigma_i(X_u)\sigma_l(X_u)-\nabla\sigma_i(X_0)\sigma_l(X_u)\|_2^2\right]\ \mathrm{d}u\ \mathrm{d}s\\ \leq& \sum_{i=1}^m\sum_{l=1}^m\int_0^t\int_0^s \mathbb{E}\left[\|\nabla\sigma_i(X_u)-\nabla\sigma_i(X_0)\|_{\mathrm{op}}^2\,\|\sigma_l(X_u)\|_2^2\right]\ \mathrm{d}u\ \mathrm{d}s\\ \leq& \sum_{i=1}^m\sum_{l=1}^m\int_0^t\int_0^s \mathbb{E}\left[\|\nabla\sigma_i(X_u)-\nabla\sigma_i(X_0)\|_{\mathrm{op}}^4\right]^{1/2}\mathbb{E}\left[\|\sigma_l(X_u)\|_2^4\right]^{1/2}\ \mathrm{d}u\ \mathrm{d}s\\ \leq& \mu_2(\sigma)^2\pi_{1,4}(\sigma)^{1/2}m^2\int_0^t\int_0^s \mathbb{E}\left[\|X_u-X_0\|_2^4\right]^{1/2}\mathbb{E}\left[1+\|X_u\|_2^2\right]^{1/2}\ \mathrm{d}u\ \mathrm{d}s\\ \leq& \mu_2(\sigma)^2\pi_{1,4}(\sigma)^{1/2}D_1^{1/2}\left(1+\mathcal{V}_2'^{1/2}\right)m^2\frac{t^3}{6}. \end{aligned}\tag{39}$$

Similarly,

$$\begin{aligned}\mathbb{E}\left[\|A_2(t)\|_2^2\right] =& \sum_{i=1}^m\sum_{l=1}^m\int_0^t\int_0^s \mathbb{E}\left[\|\nabla\sigma_i(X_0)\sigma_l(X_u)-\nabla\sigma_i(X_0)\sigma_l(X_0)\|_{\mathrm{op}}^2\right]\ \mathrm{d}u\ \mathrm{d}s\\ \leq& \sum_{i=1}^m\sum_{l=1}^m\int_0^t\int_0^s \mathbb{E}\left[\|\nabla\sigma_i(X_0)\|_{\mathrm{op}}^2\,\|\sigma_l(X_u)-\sigma_l(X_0)\|_2^2\right]\ \mathrm{d}u\ \mathrm{d}s\\ \leq& \mu_1(\sigma)^2\sum_{i=1}^m\sum_{l=1}^m\int_0^t\int_0^s \mathbb{E}\left[\|\sigma_l(X_u)-\sigma_l(X_0)\|_2^2\right]\ \mathrm{d}u\ \mathrm{d}s\\ \leq& \mu_1(\sigma)^4m^2\int_0^t\int_0^s \mathbb{E}\left[\|X_u-X_0\|_2^2\right]\ \mathrm{d}u\ \mathrm{d}s\\ \leq& \frac{1}{6}\mu_1(\sigma)^4m^2D_0t^3. \end{aligned}\tag{40}$$

By Itô isometry,

$$\mathbb{E}\left[\|S_1(t)\|_2^2\right] = \sum_{i=1}^m\int_0^t s\int_0^s \mathbb{E}\left[\|\nabla\sigma_i(X_u)b(X_u)\|_2^2\right]\ \mathrm{d}u\ \mathrm{d}s$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{m} \int_0^t s \int_0^s \mathbb{E}\left[\|\nabla \sigma_i(X_u)\|_{\text{op}}^2 \|b(X_u)\|_2^2\right] \text{d}u \ \text{d}s \\
&\leq \mu_1(\sigma)^2 \pi_{1,2}(b)^2 \sum_{i=1}^{m} \int_0^t s \int_0^s \mathbb{E}\left[1 + \|X_u\|_2^2\right] \text{d}u \ \text{d}s \\
&= \frac{1}{2}\mu_1(\sigma)^2 \pi_{1,2}(b)^2 \left(1 + \mathcal{V}_2'\right) m t^3.
\end{aligned} \tag{41}
$$

Similarly,

$$
\begin{aligned}
\mathbb{E}\left[\|S_2(t)\|_2^2\right] &= \frac{1}{4} \sum_{i=1}^{m} \int_0^t \mathbb{E}\left[\left\|\int_0^s \sum_{l=1}^{m} \nabla^2 \sigma_i(X_u)[\sigma_l(X_u),\, \sigma_l(X_u)] \ \text{d}u\right\|_2^2 \text{d}s\right] \\
&\leq \frac{1}{4} m \sum_{i=1}^{m} \sum_{l=1}^{m} \int_0^t s \int_0^s \mathbb{E}\left[\left\|\nabla^2 \sigma_i(X_u)[\sigma_l(X_u),\, \sigma_l(X_u)]\right\|_2^2\right] \text{d}u \ \text{d}s \\
&\leq \frac{1}{4} m \sum_{i=1}^{m} \sum_{l=1}^{m} \int_0^t s \int_0^s \mathbb{E}\left[\left\|\nabla^2 \sigma_i(X_u)\right\|_{\text{op}}^2 \|\sigma_l(X_u)\|_2^4\right] \text{d}u \ \text{d}s \\
&\leq \frac{1}{4} \sigma_2(\sigma)^2 \pi_{1,4}(\sigma) m^3 \int_0^t s \int_0^s \mathbb{E}\left[1 + \|X_u\|_2^2\right] \text{d}u \ \text{d}s \\
&\leq \frac{1}{8} \sigma_2(\sigma)^2 \pi_{1,4}(\sigma) m^3 \left(1 + \mathcal{V}_2'\right) t^3.
\end{aligned} \tag{42}
$$

By Corollary 13,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\Delta \tilde{H}^{(i)}\right\|_2^4\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\|\Delta \tilde{H}^{(i)}\right\|_2^4 | \mathcal{F}_{t_k}\right]\right] \\
&\leq 6^4 \pi_{1,4}^{\text{F}}(\sigma) \mathbb{E}\left[1 + \left\|\tilde{X}_k\right\|_2^2\right] t^2 \\
&\leq 6^4 \pi_{1,4}^{\text{F}}(\sigma) \left(1 + \mathcal{V}_2\right) t^2.
\end{aligned}
$$

Now, we bound the second moments of $\phi_1^{(i)}(t)$ and $\phi_2^{(i)}(t)$,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\phi_1^{(i)}(t)\right\|_2^2\right] &= \mathbb{E}\left[\left\|\int_0^1 (1-\tau) \nabla^2 \sigma_i(X_0 + \tau \Delta \tilde{H}^{(i)})[\Delta \tilde{H}^{(i)},\, \Delta \tilde{H}^{(i)}] \ \text{d}\tau\right\|_2^2\right] \\
&\leq \mathbb{E}\left[\left\|\nabla^2 \sigma_i(X_0 + \tau \Delta \tilde{H}^{(i)})\right\|_{\text{op}}^2 \left\|\Delta \tilde{H}^{(i)}\right\|_2^4\right] \\
&\leq 6^4 \mu_2(\sigma)^2 \pi_{1,4}^{\text{F}}(\sigma)(1 + \mathcal{V}_2) t^2.
\end{aligned} \tag{43}
$$

Similarly,

$$
\mathbb{E}\left[\left\|\phi_2^{(i)}(t)\right\|_2^2\right] \leq 6^4 \mu_2(\sigma)^2 \pi_{1,4}^{\text{F}}(\sigma)(1 + \mathcal{V}_2) t^2. \tag{44}
$$

Hence, by (43) and (44),

$$
\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{2} \sum_{i=1}^{m} \left(\phi_1^{(i)}(t) - \phi_2^{(i)}(t)\right) \sqrt{t}\right\|_2^2\right] &\leq \frac{m}{4} t \sum_{i=1}^{m} \mathbb{E}\left[\left\|\phi_1^{(i)}(t) - \phi_2^{(i)}(t)\right\|_2^2\right] \\
&\leq 2^2 3^4 m^2 \mu_2(\sigma)^2 \pi_{1,4}^{\text{F}}(\sigma)(1 + \mathcal{V}_2) t^3.
\end{aligned} \tag{45}
$$

Combining (39), (40), (41), (42), and (45),

$$
\begin{aligned}
\mathbb{E}\left[\left\|X_t - \tilde{X}_t\right\|_2^2\right] \leq & 32\mathbb{E}\left[\left\|\int_0^t \left(b(X_s) - b(X_0)\right) \text{d}s\right\|_2^2\right] \\
&+ 32\mathbb{E}\left[\|A_1(t)\|_2^2 + \|A_2(t)\|_2^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ 32\mathbb{E}\left[\|S_1(t)\|_2^2 + \|S_2(t)\|_2^2\right] \\
&+ 32\mathbb{E}\left[\left\|\frac{1}{2}\sum_{i=1}^{m}\left(\phi_1^{(i)}(t) - \phi_2^{(i)}(t)\right)\sqrt{t}\right\|_2^2\right] \\
\leq&\Big(16D_0\mu_1(b)^2 + \frac{16}{3}\mu_2(\sigma)^2\pi_{1,4}^{1/2}D_1^{1/2}(1+\mathcal{V}_2'^{1/2})m^2 + \frac{16}{3}\mu_1(\sigma)^4 m^2 D_0 \\
&+ 16\mu_1(\sigma)^2\pi_{1,2}(b)^2(1+\mathcal{V}_2')m + 4m^3\mu_2(\sigma)^2\pi_{1,4}^{\mathrm{F}}(\sigma)(1+\mathcal{V}_2') \\
&+ 2^7 3^4 m^2\mu_2(\sigma)^2\pi_{1,4}^{\mathrm{F}}(\sigma)\left(1+\mathcal{V}_2\right)\Big)t^3.
\end{aligned}
$$

□

C.2.2 Local Mean Deviation

Lemma 25. *Suppose X_t and $\tilde{X}_t$ are the continuous-time process defined by (1) and Markov chain defined by (10) for time $t \geq 0$, respectively. If X_t and $\tilde{X}_t$ are initiated from the same iterate of the Markov chain X_0, and they share the same Brownian motion, then*

$$
\mathbb{E}\left[\left\|\mathbb{E}\left[X_t - \tilde{X}_t|\mathcal{F}_0\right]\right\|_2^2\right] \leq D_4 t^4, \quad \textit{for all } 0 \leq t \leq 1,
$$

where

$$
\begin{aligned}
D_4 =&\Big(\frac{4}{3}\mu_1(b)^2\pi_{1,2}(b)\left(1+\mathcal{V}_2'\right) + \frac{1}{3}\mu_2(b)^2\pi_{1,4}(\sigma)\left(1+\mathcal{V}_2'\right)m^2 \\
&+ 2^4 3^5 5^6\mu_3(\sigma)^2\pi_{1,6}^{\mathrm{F}}(\sigma)\left(1+\mathcal{V}_4^{3/4}\right)\Big).
\end{aligned}
$$

Proof. Recall the operators L and Λ_i ($i = 1, \ldots, m$) defined in (5). By Itô's lemma,

$$
\begin{aligned}
X_t - X_0 =& b(X_0)t + \sum_{i=1}^{m}\int_0^t \sigma_i(X_s)\,\mathrm{d}B_s^{(i)} \\
&+ \sum_{i=1}^{m}\int_0^t\int_0^s \Lambda_i(b)(X_u)\,\mathrm{d}B_u^{(i)}\,\mathrm{d}s + \int_0^t\int_0^s L(b)(X_u)\,\mathrm{d}u\,\mathrm{d}s \\
=& b(X_0)t + \sum_{i=1}^{m}\int_0^t \sigma_i(X_s)\,\mathrm{d}B_s^{(i)} + \sum_{i=1}^{m}\int_0^t\int_0^s \nabla b(X_u)\sigma_i(X_u)\,\mathrm{d}B_u^{(i)}\,\mathrm{d}s + \bar{S}(t),
\end{aligned}
$$

where

$$
\bar{S}(t) = \underbrace{\int_0^t\int_0^s \nabla b(X_u)b(X_u)\,\mathrm{d}u\,\mathrm{d}s}_{\bar{S}_1(t)} + \underbrace{\frac{1}{2}\sum_{i=1}^{m}\int_0^t\int_0^s \nabla^2 b(X_u)[\sigma_i(X_u),\,\sigma_i(X_u)]\,\mathrm{d}u\,\mathrm{d}s}_{\bar{S}_2(t)}.
$$

Now, we bound the second moments of $\bar{S}_1(t)$ and $\bar{S}_2(t)$,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\bar{S}_1(t)\right\|_2^2\right] =& \mathbb{E}\left[\left\|\int_0^t\int_0^s \nabla b(X_u)b(X_u)\,\mathrm{d}u\,\mathrm{d}s\right\|_2^2\right] \\
\leq& t\int_0^t s\int_0^s \mathbb{E}\left[\|\nabla b(X_u)b(X_u)\|_2^2\right]\mathrm{d}u\,\mathrm{d}s \\
\leq& t\int_0^t s\int_0^s \mathbb{E}\left[\|\nabla b(X_u)\|_{\mathrm{op}}^2\,\|b(X_u)\|_2^2\right]\mathrm{d}u\,\mathrm{d}s \\
\leq& \mu_1(b)^2\pi_{1,2}(b)t\int_0^t s\int_0^s \mathbb{E}\left[1+\|X_u\|_2^2\right]\mathrm{d}u\,\mathrm{d}s \\
\leq& \frac{1}{3}\mu_1(b)^2\pi_{1,2}(b)\left(1+\mathcal{V}_2'\right)t^4.
\end{aligned} \tag{46}
$$

Similarly,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\bar{S}_2(t)\right\|_2^2\right] =&\mathbb{E}\left[\left\|\frac{1}{2}\sum_{i=1}^m \int_0^t \int_0^s \nabla^2 b(X_u)[\sigma_i(X_u),\, \sigma_i(X_u)]\ \mathrm{d}u\ \mathrm{d}s\right\|_2^2\right] \\
\leq& \frac{m}{4}\sum_{i=1}^m t\int_0^t s\int_0^s \mathbb{E}\left[\left\|\nabla^2 b(X_u)[\sigma_i(X_u),\, \sigma_i(X_u)]\right\|_2^2\right]\ \mathrm{d}u\ \mathrm{d}s \\
\leq& \frac{m}{4}\sum_{i=1}^m t\int_0^t s\int_0^s \mathbb{E}\left[\left\|\nabla^2 b(X_u)\right\|_{\mathrm{op}}^2 \left\|\sigma_i(X_u)\right\|_2^4\right]\ \mathrm{d}u\ \mathrm{d}s \\
\leq& \frac{m}{4}\mu_2(b)^2\sum_{i=1}^m t\int_0^t s\int_0^s \mathbb{E}\left[\left\|\sigma_i(X_u)\right\|_2^4\right]\ \mathrm{d}u\ \mathrm{d}s \\
\leq& \frac{m^2}{4}\mu_2(b)^2\pi_{1,4}(\sigma) t\int_0^t s\int_0^s \mathbb{E}\left[1+\left\|X_u\right\|_2^2\right]\ \mathrm{d}u\ \mathrm{d}s \\
\leq& \frac{1}{12}\mu_2(b)^2\pi_{1,4}(\sigma)\left(1+\mathcal{V}_2'\right)m^2t^4. \qquad (47)
\end{aligned}
$$

By Corollary 13,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\Delta\tilde{H}^{(i)}\right\|_2^6\right] =&\mathbb{E}\left[\mathbb{E}\left[\left\|\Delta\tilde{H}^{(i)}\right\|_2^6|\mathcal{F}_{t_k}\right]\right] \\
\leq& 3^6 5^6 \pi_{1,6}^{\mathrm{F}}(\sigma)\mathbb{E}\left[1+\left\|\tilde{X}_k\right\|_2^3\right]t^3 \\
\leq& 3^6 5^6 \pi_{1,6}^{\mathrm{F}}(\sigma)\left(1+\mathcal{V}_4^{3/4}\right)t^3.
\end{aligned}
$$

Now, we bound the second moment of the difference between $\phi_1^{(i)}(t)$ and $\phi_2^{(i)}(t)$,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\phi_1^{(i)}(t)-\phi_2^{(i)}(t)\right\|_2^2\right] \leq& \mathbb{E}\left[\int_0^1 \left\|\nabla^2\sigma_i(X_0+\tau\Delta\tilde{H}^{(i)})-\nabla^2\sigma_i(X_0-\tau\Delta\tilde{H}^{(i)})\right\|_{\mathrm{op}}^2 \left\|\Delta\tilde{H}^{(i)}\right\|_2^4\ \mathrm{d}\tau\right] \\
\leq& \mu_3(\sigma)^2\int_0^1 \mathbb{E}\left[\left\|2\tau\Delta\tilde{H}^{(i)}\right\|_2^2\left\|\Delta\tilde{H}^{(i)}\right\|_2^4\right]\ \mathrm{d}\tau \\
\leq& \frac{4}{3}\mu_3(\sigma)^2\mathbb{E}\left[\left\|\Delta\tilde{H}^{(i)}\right\|_2^6\right] \\
\leq& 2^2 3^5 5^6 \mu_3(\sigma)^2\pi_{1,6}^{\mathrm{F}}(\sigma)\left(1+\mathcal{V}_4^{3/4}\right)t^3. \qquad (48)
\end{aligned}
$$

Hence, combining (46), (47), and (48),

$$
\begin{aligned}
\mathbb{E}\left[\left\|\mathbb{E}\left[X_t-\tilde{X}_t|\mathcal{F}_0\right]\right\|_2^2\right] =&\mathbb{E}\left[\left\|\mathbb{E}\left[\bar{S}(t)|\mathcal{F}_0\right]-\mathbb{E}\left[\frac{1}{2}\sum_{i=1}^m\left(\phi_1^{(i)}(t)-\phi_2^{(i)}(t)\right)\sqrt{t}|\mathcal{F}_0\right]\right\|_2^2\right] \\
\leq& 4\mathbb{E}\left[\left\|\bar{S}_1(t)\right\|_2^2+\left\|\bar{S}_2(t)\right\|_2^2\right]+4\mathbb{E}\left[\left\|\frac{1}{2}\sum_{i=1}^m\left(\phi_1^{(i)}(t)-\phi_2^{(i)}(t)\right)\sqrt{t}\right\|_2^2\right] \\
\leq& \Big(\frac{4}{3}\mu_1(b)^2\pi_{1,2}(b)\left(1+\mathcal{V}_2'\right)+\frac{1}{3}\mu_2(b)^2\pi_{1,4}(\sigma)\left(1+\mathcal{V}_2'\right)m^2 \\
&+2^4 3^5 5^6\mu_3(\sigma)^2\pi_{1,6}^{\mathrm{F}}(\sigma)\left(1+\mathcal{V}_4^{3/4}\right)\Big)t^4.
\end{aligned}
$$

□

C.3 Invoking Theorem 1

Now, we invoke Theorem 1 with our derived constants. We obtain that if the constant step size

$$
h < 1\wedge C_h \wedge \frac{1}{2\alpha}\wedge\frac{1}{8\mu_1(b)^2+8\mu_1^{\mathrm{F}}(\sigma)^2},
$$

where

$$C_h = \frac{1}{m^2} \wedge \frac{\alpha^2}{4M_1^2} \wedge \frac{\alpha^2}{M_{3,2}^2},$$

and the smoothness conditions in Theorem 3 of the drift and diffusion coefficients are satisfied for a uniformly dissipative diffusion, then the uniform local deviation bounds (7) hold with $\lambda_1 = D_3$ and $\lambda_2 = D_4$, and consequently the bound (8) holds. This concludes that to converge to a sufficiently small positive tolerance ϵ, $\tilde{\mathcal{O}}(d^{3/4}m^2\epsilon^{-1})$ iterations are required, since D_3 is of order $\mathcal{O}(d^{3/2}m^3)$, and D_4 is of order $\mathcal{O}(d^{3/2}m^2)$. Note that the dimension dependence worsens if one were to further convert the Frobenius norm dependent constants to be based on the operator norm.

D Convergence Rate for Example 2

D.1 Moment Bound

Verifying the order conditions in Theorem 1 of the EM scheme for uniformly dissipative diffusions requires bounding the second moments of the Markov chain. Recall, dissipativity (Definition C.1) follows from uniform dissipativity of the Itô diffusion.

Lemma 26. *If the second moment of the initial iterate is finite, then the second moments of Markov chain iterates defined in* (4) *are uniformly bounded, i.e.*

$$\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^2\right] \le \mathcal{W}_2, \quad \textit{for all } k \in \mathbb{N},$$

where $\mathcal{W}_2 = \mathbb{E}\left[\left\|\tilde{X}_0\right\|_2^2\right] + 2(\pi_{1,2}(b)+\beta)/\alpha$, *if the constant step size* $h < 1 \wedge \alpha/(2\pi_{1,2}(b))$.

Proof. By direct computation,

$$\begin{aligned}\left\|\tilde{X}_{k+1}\right\|_2^2 = &\left\|\tilde{X}_k\right\|_2^2 + \left\|b(\tilde{X}_k)\right\|_2^2 h^2 + \left\|\sigma(\tilde{X}_k)\xi_{k+1}\right\|_2^2 h \\ &+ 2\left\langle \tilde{X}_k, b(\tilde{X}_k)\right\rangle h + 2\left\langle \tilde{X}_k, \sigma(\tilde{X}_k)\xi_{k+1}\right\rangle h^{1/2} \\ &+ 2\left\langle b(\tilde{X}_k), \sigma(\tilde{X}_k)\xi_{k+1}\right\rangle h^{3/2}.\end{aligned}$$

Recall by Lemma 15 and dissipativity,

$$\mathbb{E}\left[2\left\langle \tilde{X}_k, b(\tilde{X}_k)\right\rangle h + \left\|\sigma(\tilde{X}_k)\xi_{k+1}\right\|_2^2 h|\mathcal{F}_{t_k}\right] \le -\alpha\left\|\tilde{X}_k\right\|_2^2 h + \beta h.$$

By odd moments of Gaussian variables being zero and the step size condition,

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{X}_{k+1}\right\|_2^2|\mathcal{F}_{t_k}\right] \le &(1-\alpha h)\left\|\tilde{X}_k\right\|_2^2 + \left\|b(\tilde{X}_k)\right\|_2^2 h^2 + \beta h \\ \le &(1-\alpha h + \pi_{1,2}(b)h^2)\left\|\tilde{X}_k\right\|_2^2 + \pi_{1,2}(b)h^2 + \beta h \\ \le &(1-\alpha h/2)\left\|\tilde{X}_k\right\|_2^2 + \pi_{1,2}(b)h^2 + \beta h.\end{aligned}$$

By unrolling the recursion,

$$\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^2\right] \le \mathbb{E}\left[\left\|\tilde{X}_0\right\|_2^2\right] + 2(\pi_{1,2}(b)+\beta)/\alpha, \quad \text{for all } k \in \mathbb{N}.$$

□

D.2 Local Deviation Orders

Before verifying the local deviation orders, we first state two auxiliary lemmas. We omit the proofs, since they are almost identical to that of Lemma 6 and Lemma 7, respectively.

Lemma 27. *Suppose X_t is the continuous-time process defined by (1) initiated from some iterate of the Markov chain X_0 defined by (4), then the second moment of X_t is uniformly bounded, i.e.*

$$\mathbb{E}\left[\|X_t\|_2^2\right] \leq \mathcal{W}_2 + \beta/\alpha = \mathcal{W}_2', \quad \textit{for all } t \geq 0.$$

Lemma 28. *Suppose X_t is the continuous-time process defined by (1) initiated from some iterate of the Markov chain X_0 defined by (4), then*

$$\mathbb{E}\left[\|X_t - X_0\|_2^2\right] \leq E_0 t, \quad \textit{for all } t \geq 0,$$

where $E_0 = 2\left(\pi_{1,2}(b) + \pi_{1,2}^{\mathrm{F}}(\sigma)\right)(1 + \mathcal{W}_2')$.

D.2.1 Local Mean-Square Deviation

Lemma 29. *Suppose X_t and $\tilde{X}_t$ are the continuous-time process defined by (1) and Markov chain defined by (4) for time $t \geq 0$, respectively. If X_t and $\tilde{X}_t$ are initiated from the same iterate of the Markov chain X_0 and share the same Brownian motion, then*

$$\mathbb{E}\left[\left\|X_t - \tilde{X}_t\right\|_2^2\right] \leq E_1 t^2, \quad \textit{for all } 0 \leq t \leq 1,$$

where $E_1 = \left(\mu_1(b)^2 + \mu_1^{\mathrm{F}}(\sigma)^2\right) E_0$.

Proof. By Itô isometry and Lipschitz of the drift and diffusion coefficients,

$$\begin{aligned}
\mathbb{E}\left[\left\|X_t - \tilde{X}_t\right\|_2^2\right] \leq& 2\mathbb{E}\left[\left\|\int_0^t (b(X_s) - b(X_0))\ \mathrm{d}s\right\|_2^2\right] + 2\mathbb{E}\left[\left\|\int_0^t (\sigma(X_s) - \sigma(X_0))\ \mathrm{d}B_s\right\|_2^2\right] \\
\leq& 2t\mathbb{E}\left[\int_0^t \|b(X_s) - b(X_0)\|_2^2\ \mathrm{d}s\right] + 2\mathbb{E}\left[\int_0^t \|\sigma(X_s) - \sigma(X_0)\|_{\mathrm{F}}^2\ \mathrm{d}s\right] \\
\leq& 2\left(\mu_1(b)^2 t + \mu_1^{\mathrm{F}}(\sigma)^2\right) \int_0^t \mathbb{E}\left[\|X_s - X_0\|_2^2\right]\ \mathrm{d}s \\
\leq& \left(\mu_1(b)^2 + \mu_1^{\mathrm{F}}(\sigma)^2\right) E_0 t^2.
\end{aligned}$$

□

D.2.2 Local Mean Deviation

Lemma 30. *Suppose X_t and $\tilde{X}_t$ are the continuous-time process defined by (1) and Markov chain defined by(4) for time $t \geq 0$, respectively. If X_t and $\tilde{X}_t$ are initiated from the same iterate of the Markov chain X_0 and share the same Brownian motion, then*

$$\mathbb{E}\left[\left\|\mathbb{E}\left[X_t - \tilde{X}_t|\mathcal{F}_0\right]\right\|_2^2\right] \leq E_2 t^3, \quad \textit{for all } 0 \leq t \leq 1,$$

where $E_2 = \mu_1(b)E_0/2$.

Proof. By Itô's lemma,

$$\begin{aligned}
X_t - X_0 =& \int_0^t b(X_s)ds + \sigma(X_0)B_t \\
&+ \sum_{i=1}^m \sum_{l=1}^m \int_0^t \int_0^s \Lambda_l(\sigma_i)(X_u)\ \mathrm{d}B_u^{(l)}\ \mathrm{d}B_s^{(i)} + \sum_{i=1}^m \int_0^t \int_0^s L(\sigma_i)(X_u)\ \mathrm{d}u\ \mathrm{d}B_s^{(i)}.
\end{aligned}$$

Since the last two terms in the above inequality are Martingales,

$$\mathbb{E}\left[X_t - X_0|\mathcal{F}_0\right] = \mathbb{E}\left[\int_0^t (b(X_s) - b(X_0))\ \mathrm{d}s|\mathcal{F}_0\right].$$

Hence, by Jensen's inequality,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\mathbb{E}\left[X_t - \tilde{X}_t|\mathcal{F}_0\right]\right\|_2^2\right] =& \mathbb{E}\left[\left\|\mathbb{E}\left[\int_0^t \left(b(X_s) - b(X_0)\right)\,\mathrm{d}s|\mathcal{F}_0\right]\right\|_2^2\right] \\
\leq& \mathbb{E}\left[\left\|\int_0^t \left(b(X_s) - b(X_0)\right)\,\mathrm{d}s\right\|_2^2\right] \\
\leq& \mu_1(b) t \int_0^t \mathbb{E}\left[\|X_s - X_0\|_2^2\right]\,\mathrm{d}s \\
\leq& \mu_1(b) E_0 t^3/2.
\end{aligned}
$$

□

D.3 Invoking Theorem 1

Now, we invoke Theorem 1 with our derived constants. We obtain that if the constant step size

$$
h < 1 \wedge \frac{\alpha}{2\pi_{1,2}(b)} \wedge \frac{1}{2\alpha} \wedge \frac{1}{8\mu_1(b)^2 + 8\mu_1^{\mathrm{F}}(\sigma)^2},
$$

and the smoothness conditions of the drift and diffusion coefficients are satisfied for a uniformly dissipative diffusion, then the uniform local deviation bounds (7) hold with $\lambda_1 = E_1$ and $\lambda_2 = E_2$, and consequently the bound (8) holds. This concludes that for a sufficiently small positive tolerance ϵ, $\tilde{\mathcal{O}}(d\epsilon^{-2})$ iterations are required, since both E_1 and E_2 are of order $\mathcal{O}(d)$. If one were to convert the Frobenius norm dependent constants to be based on the operator norm, then E_1 is of order $\mathcal{O}(d(d+m)^2)$, and E_2 is of order $\mathcal{O}(d(d+m))$. This yields the convergence rate of $\tilde{\mathcal{O}}(d(d+m)^2\epsilon^{-2})$.

E Convergence of SRK-LD Under an Unbiased Stochastic Oracle

We provide an informal analysis on the scenario where the oracle is stochastic. We denote the new interpolated values under the stochastic oracle as $\hat{H}_1$ and $\hat{H}_2$, and the new iterate value as $\hat{X}_k$. We assume (i) the stochastic oracle is unbiased, i.e. $\mathbb{E}[\hat{\nabla} f(x)] = f(x)$ for all $x \in \mathbb{R}^d$, (ii) the stochastic oracle has finite variance at the Markov chain iterates and "interpolated" values, i.e. $\mathbb{E}[\|\hat{\nabla} f(Y) - \nabla f(Y)\|_2^2] \leq \sigma^2 d$, for some finite σ, where Y may be $\hat{X}_k$, $\hat{H}_1$, or $\hat{H}_2$[6], and (iii) the randomness in the stochastic oracle is independent of that of the Brownian motion.

Fix iteration index $k \in \mathbb{N}$, let $\tilde{D}_h^{(k)}$ and $\hat{D}_h^{(k)}$ denote the local deviations under the exact and stochastic oracles, respectively. Then, assuming the step size is chosen sufficiently small such that the Markov chain moments are bounded,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\hat{D}_h^{(k)}\right\|_2^2\right] \leq& 2\mathbb{E}\left[\left\|\tilde{D}_h^{(k)}\right\|_2^2\right] + 2\mathbb{E}\left[\left\|\tilde{D}_h^{(k)} - \hat{D}_h^{(k)}\right\|_2^2\right] \\
\leq& 2\mathbb{E}\left[\left\|\tilde{D}_h^{(k)}\right\|_2^2\right] + 4\mathbb{E}\left[\left\|\hat{\nabla} f(\hat{H}_1) - \nabla f(\tilde{H}_1)\right\|_2^2\right] + 4\mathbb{E}\left[\left\|\hat{\nabla} f(\hat{H}_2) - \nabla f(\tilde{H}_2)\right\|_2^2\right] \\
\leq& 2\mathbb{E}\left[\left\|\tilde{D}_h^{(k)}\right\|_2^2\right] + 4\sigma^2 d + 4\mathbb{E}\left[\left\|\hat{\nabla} f(\hat{H}_2) - \nabla f(\hat{H}_2) + \nabla f(\hat{H}_2) - \nabla f(\tilde{H}_2)\right\|_2^2\right] \\
\leq& \mathcal{O}(h^4 + \sigma^2).
\end{aligned}
$$

Similarly, one can derive the new local mean deviation,

$$
\mathbb{E}\left[\left\|\mathbb{E}\left[\hat{D}_h^{(k)}|\mathcal{F}_{t_{k-1}}\right]\right\|_2^2\right] \leq \mathbb{E}\left[\left\|\mathbb{E}\left[\tilde{D}_h^{(k)}|\mathcal{F}_{t_{k-1}}\right] + \mathbb{E}\left[\hat{D}_h^{(k)} - \tilde{D}_h^{(k)}|\mathcal{F}_{t_{k-1}}\right]\right\|_2^2\right]
$$

[6]There is slight ambiguity in terms of which iteration's interpolated values should $\tilde{H}_1$ and $\tilde{H}_2$ correspond to. For notational simplicity, we have avoided using a subscript or superscript for the iteration index k, and almost always make $\tilde{H}_1$ and $\tilde{H}_2$ appear along with the original iterate $\tilde{X}_k$.

$$
\begin{aligned}
&\leq \mathbb{E}\left[\left\|\mathbb{E}\left[\tilde{D}_h^{(k)}|\mathcal{F}_{t_{k-1}}\right]\right\|_2^2\right] + \mathbb{E}\left[\left\|\hat{D}_h^{(k)} - \tilde{D}_h^{(k)}\right\|_2^2\right] \\
&= \mathcal{O}(h^5 + \sigma^2).
\end{aligned}
$$

One can replace the corresponding terms in (15) and obtain a recursion. Note however, to ensure unrolling the recursion gives a convergence bound, one need that $\sigma^2 < \mathcal{O}(\alpha h)$.

F Auxiliary Lemmas

We list standard results used to develop our theorems and include their proofs for completeness.

Lemma 31. *For $x_1, \ldots, x_m \in \mathbb{R}$ and $m, n \in \mathbb{N}_+$, we have*

$$
\left(\sum_{i=1}^m x_i\right)^n \leq m^{n-1} \sum_{i=1}^m x_i^n.
$$

Proof. Recall the function $f(x) = x^n$ is convex for $n \in \mathbb{N}_+$. Hence,

$$
\left(\frac{\sum_{i=1}^m x_i}{m}\right)^n \leq \frac{\sum_{i=1}^m x_i^n}{m}.
$$

Multiplying both sides of the inequality by m^n completes the proof. □

Lemma 32. *For the d-dimensional Brownian motion $\{B_t\}_{t\geq 0}$,*

$$
Z_t = \int_0^t \int_0^s \mathrm{d}B_u \ \mathrm{d}s \sim \mathcal{N}\left(0, t^3 I_d/3\right).
$$

Proof. We consider the case where $d = 1$. The multi-dimensional case follows naturally, since we assume different dimensions of the Brownian motion vector are independent. Let $t_k = \delta k$, we define

$$
S_m = \sum_{k=0}^{m-1} B_{t_k}(t_{k+1} - t_k) = \sum_{k=1}^{m-1} \left(B_{t_{k+1}} - B_{t_k}\right)(t_k - t).
$$

Since S_m is a sum of Gaussian random variables, it is also Gaussian. By linearity of expectation and independence of Brownian motion increments,

$$
\begin{aligned}
&\mathbb{E}\left[S_m\right] = 0, \\
&\mathbb{E}\left[S_m^2\right] = \sum_{k=1}^{m-1} (t_k - t)^2 \, \mathbb{E}\left[\left(B_{t_{k+1}} - B_{t_k}\right)^2\right] \to \int_0^t (s-t)^2 \ \mathrm{d}s = t^3/3 \quad \text{as} \quad m \to \infty.
\end{aligned}
$$

Since $S_m \overset{\text{a.s.}}{\to} Z_t$ as $m \to \infty$ by the strong law of large numbers, we conclude that $Z_t \sim \mathcal{N}\left(0, t^3/3\right)$. □

Lemma 33. *For $n \in \mathbb{N}$ and the d-dimensional Brownian motion $\{B_t\}_{t\geq 0}$,*

$$
\mathbb{E}\left[\|B_t\|_2^{2n}\right] = t^n d(d+2)\cdots(d+2n-2).
$$

Proof. Note $\|B_t\|_2^2$ may be expressed as the sum of squared Gaussian random variables, i.e.

$$
\|B_t\|_2^2 = t\sum_{i=1}^d \xi_i^2, \quad \text{where} \quad \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).
$$

Observe that this is also a multiple of the chi-squared random variable with d degrees of freedom $\chi(d)^2$. Its nth moment has the following closed form [57],

$$
\mathbb{E}\left[\chi(d)^{2n}\right] = 2^n \frac{\Gamma\left(n + \frac{d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} = d(d+2)\cdots(d+2n-2).
$$

Thus,

$$
\mathbb{E}\left[\|B_t\|_2^{2n}\right] = t^n \mathbb{E}\left[\chi(d)^{2n}\right] = t^n d(d+2)\cdots(d+2n-2).
$$

□

Lemma 34. *For $f : \mathbb{R}^d \to \mathbb{R}$ which is C^3, suppose its Hessian is μ_3-Lipschitz under the operator norm and Euclidean norm, i.e.*

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_{\mathrm{op}} \leq \mu_3 \left\|x - y\right\|_2, \quad \textit{for all } x, y \in \mathbb{R}^d.$$

Then, the vector Laplacian of its gradient is bounded, i.e.

$$\left\|\vec{\Delta}(\nabla f)(x)\right\|_2 \leq d\mu_3, \quad \textit{for all } x \in \mathbb{R}^d.$$

Proof. See proof of Lemma 6 in [12]. □

Lemma 35. *For $f : \mathbb{R}^d \to \mathbb{R}$ which is C^4, suppose its third derivative is μ_4-Lipschitz under the operator norm and Euclidean norm, i.e.*

$$\left\|\nabla^3 f(x) - \nabla^3 f(y)\right\|_{\mathrm{op}} \leq \mu_4 \left\|x - y\right\|_2, \quad \textit{for all } x, y \in \mathbb{R}^d.$$

Then, the vector Laplacian of its gradient is $d\mu_4$-Lipschitz, i.e.

$$\left\|\vec{\Delta}(\nabla f)(x) - \vec{\Delta}(\nabla f)(y)\right\|_2 \leq d\mu_4 \left\|x - y\right\|_2.$$

Proof. Let $g(x) = \Delta(f)(x)$. Since $f \in C^4$, we may switch the order of partial derivatives,

$$\left\|\vec{\Delta}(\nabla f)(x) - \vec{\Delta}(\nabla f)(y)\right\|_2 = \left\|\nabla g(x) - \nabla g(y)\right\|_2.$$

By Taylor's theorem with the remainder in integral form,

$$\begin{aligned}
\left\|\nabla g(x) - \nabla g(y)\right\|_2 &= \left\|\int_0^1 \nabla^2 g\left(y + \tau(x-y)\right)(x-y)\,\mathrm{d}\tau\right\|_2 \\
&\leq \int_0^1 \left\|\nabla^2 g\left(y + \tau(x-y)\right)\right\|_{\mathrm{op}} \left\|x-y\right\|_2 \,\mathrm{d}\tau \\
&\leq \sup_{z \in \mathbb{R}^d} \left\|\nabla^2 g(z)\right\|_{\mathrm{op}} \left\|x-y\right\|_2.
\end{aligned}$$

Note that $\nabla^2 g(x)$ can be written as a sum of d matrices, each being a sub-tensor of $\nabla^4 f(x)$, due to the the trace operator, i.e.

$$\nabla^2 g(x) = \sum_{i=1}^d G_i(x), \quad \text{where} \quad G_i(x)_{jk} = \partial_{iijk} f(x).$$

Since the operator norm of $\nabla^4 f(x)$ upper bounds the operator norm of each of its sub-tensor,

$$\left\|\nabla^2 g(x)\right\|_{\mathrm{op}} \leq \sum_{i=1}^d \left\|G_i(x)\right\|_{\mathrm{op}} \leq d\left\|\nabla^4 f(x)\right\|_{\mathrm{op}}$$

Recall the third derivative is μ_4-Lipschitz, we obtain

$$\left\|\nabla g(x) - \nabla g(y)\right\|_2 \leq d\mu_3 \left\|x-y\right\|_2.$$

□

G Estimating the Wasserstein Distance

For a Borel measure μ defined on a compact and separable topological space $\mathcal{X}$, a sample-based empirical measure μ_n may asymptotically serve as a proxy to μ in the W_p sense for $p \in [1, \infty)$, i.e.

$$W_p(\mu, \hat{\mu}_n) \xrightarrow{\mu\text{-a.s.}} 0.$$

This is a consequence of the Wasserstein distance metrizing weak convergence [62] and that the empirical measure converges weakly to μ almost surely [60].

However, in the finite-sample setting, this distance is typically non-negligible and worsens as the dimensionality increases. Specifically, generalizing previous results based on the 1-Wasserstein distance [17, 16], Weed and Bach [64] showed that for $p \in [1, \infty)$,

$$W_p(\mu, \hat{\mu}_n) \gtrsim n^{-1/t},$$

where t is less than the lower Wasserstein dimension $d_*(\mu)$. This presents a severe challenge in estimating the 2-Wasserstein distance between probability measures using samples.

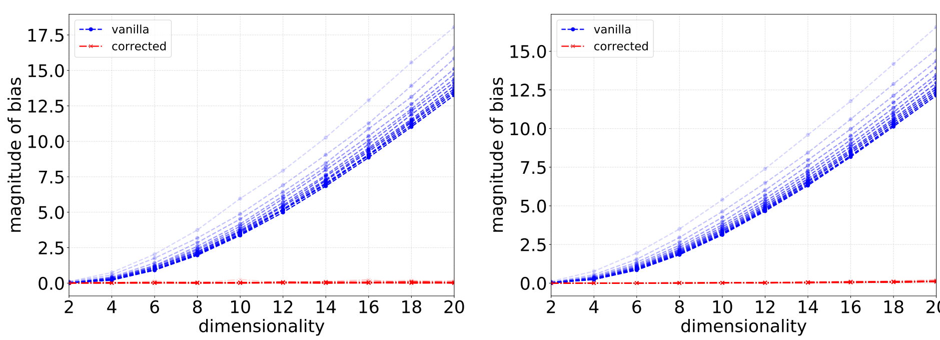
To better detect convergence, we zero center a simple sample-based estimator by subtracting the null responses and obtain the following new estimator:

$$\tilde{W}_2^2(\mu, \nu) = \frac{1}{2}\left(W_2^2(\hat{\mu}_n, \hat{\nu}_n) + W_2^2(\hat{\mu}'_n, \hat{\nu}'_n) - W_2^2(\hat{\mu}_n, \hat{\mu}'_n) - W_2^2(\hat{\nu}_n, \hat{\nu}'_n)\right),$$

where $\hat{\nu}_n$ and $\hat{\nu}'_n$ are based on two independent samples of size n from μ, and similarly for $\hat{\nu}_n$ and $\hat{\nu}'_n$ from ν. This estimator is inspired by the contruction of distances in the maximum mean discrepancy family [31] and the Sinkhorn divergence [49]. Note that the 2-Wasserstein distance between finite samples can be computed conveniently with existing packages [25] that solves a linear program. Although the new estimator is not guaranteed to be unbiased across all settings, it is unbiased when the two distributions are the same.

Since our correction is based on a heuristic, the new estimator is still biased. To empirically characterize the effectiveness of the correction, we compute the discrepancy between the squared 2-Wasserstein distance for two continuous densities and the finite-sample estimate obtained from i.i.d. samples. When μ and ν are Gaussians with means $m_1, m_2 \in \mathbb{R}^d$ and covariance matrices $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$, we have the following convenient closed-form

$$W_2^2(\mu, \nu) = \|m_1 - m_2\|_2^2 + \operatorname{Tr}\left(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right).$$



(a) different in mean

(b) different in covariance

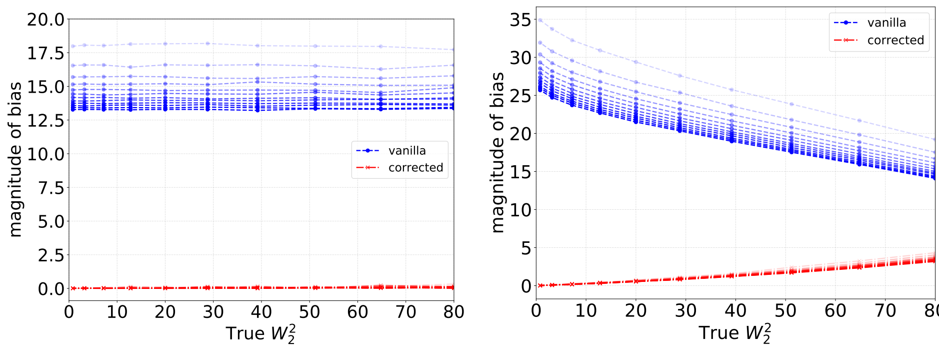
Figure 2: Absolute value between $W_2^2(\mu, \nu)$ and the sample averages of estimators $\hat{W}_2^2$ (vanilla) and $\tilde{W}_2^2$ (corrected) for Gaussian μ and ν. Darker curves correspond to larger number of samples used to compute the empirical estimate (ranging from 100 to 1000). (a) $m_1 = 0, m_2 = \mathbf{1}_d, \Sigma_1 = \Sigma_2 = I_d$. (b) $m_1 = m_2 = 0, \Sigma_1 = I_d, \Sigma_2 = I_d/2 + \mathbf{1}_d\mathbf{1}_d^\top/5$.

We compare the vanilla estimate $\hat{W}_2^2(\mu, \nu, n)$ and the corrected estimate $\tilde{W}_2^2(\mu, \nu, n)$ by their magnitude of deviation from the true value $W_2^2(\mu, \nu)$:

$$\left|W_2^2(\mu, \nu) - \mathbb{E}[\hat{W}_2^2(\mu, \nu, n)]\right|, \quad \left|W_2^2(\mu, \nu) - \mathbb{E}[\tilde{W}_2^2(\mu, \nu, n)]\right|,$$

where the expectations are approximated via averaging 100 independent draws. Figure 2 reports the deviation across different sample sizes and dimensionalities, where μ and ν differ only in either mean or covariance. While the corrected estimator is not unbiased, it is relatively more accurate.

In addition, Figure 3 demonstrates that our bias-corrected estimator becomes more accurate as the two distributions are closer. This indicates that our proposed estimator may provide a more reliable estimate of the 2-Wasserstein distance when the sampling algorithm is close to convergence.



(a) different in mean (b) different in covariance

Figure 3: Absolute value between $W_2^2(\mu,\nu)$ and the sample averages of estimators $\hat{W}_2^2$ (vanilla) and $\tilde{W}_2^2$ (corrected) for Gaussian μ and ν. Darker curves correspond to larger number of samples used to compute the empirical estimate (ranging from 100 to 1000). We fix $d = 20$ and interpolate the mean and the covariance matrix, i.e. $m = \alpha m_1 + (1-\alpha)m_2, \Sigma = \alpha\Sigma_1 + (1-\alpha)\Sigma_2, \alpha \in [0,1]$. (a) $m_1 = 0, m_2 = 2\mathbf{1}_d, \Sigma_1 = \Sigma_2 = I_d$. (b) $m_1 = m_2 = 0, \Sigma_1 = 2I_d, \Sigma_2 = I_d/2 + \mathbf{1}_d\mathbf{1}_d^\top/5$.

H Additional Numerical Studies

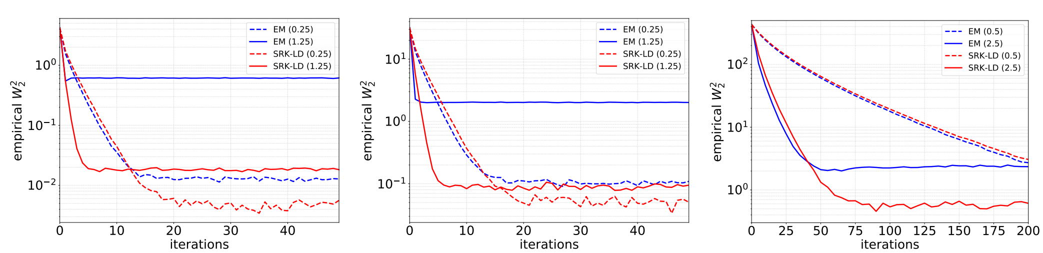
In this section, we include additional numerical studies complementing Section 5.

H.1 Strongly Convex Potentials

We first include additional plots of error estimates in W_2 and the energy distance for sampling from a Gaussian mixture and the posterior of BLR. The results indicate that the reduction in asymptotic error is consistent across problems with varying dimensionalities that we consider. In the end, we conduct a wall time analysis and show that SRK-LD is competitive in practice.

H.1.1 Additional Results

Figure 4 shows the estimated W_2 error as the number of iterations increase for the 2D and 20D Gaussian mixture and BLR problems with the parameter settings described in Section 5. We observe consistent improvement in the asymptotic error across different settings in which we experimented.



(a) Gaussian Mixture (2D) (b) Gaussian Mixture (20D) (c) BLR (20D)

Figure 4: Error in W_2^2 for strongly log-concave sampling. Legend denotes“scheme (step size)”.

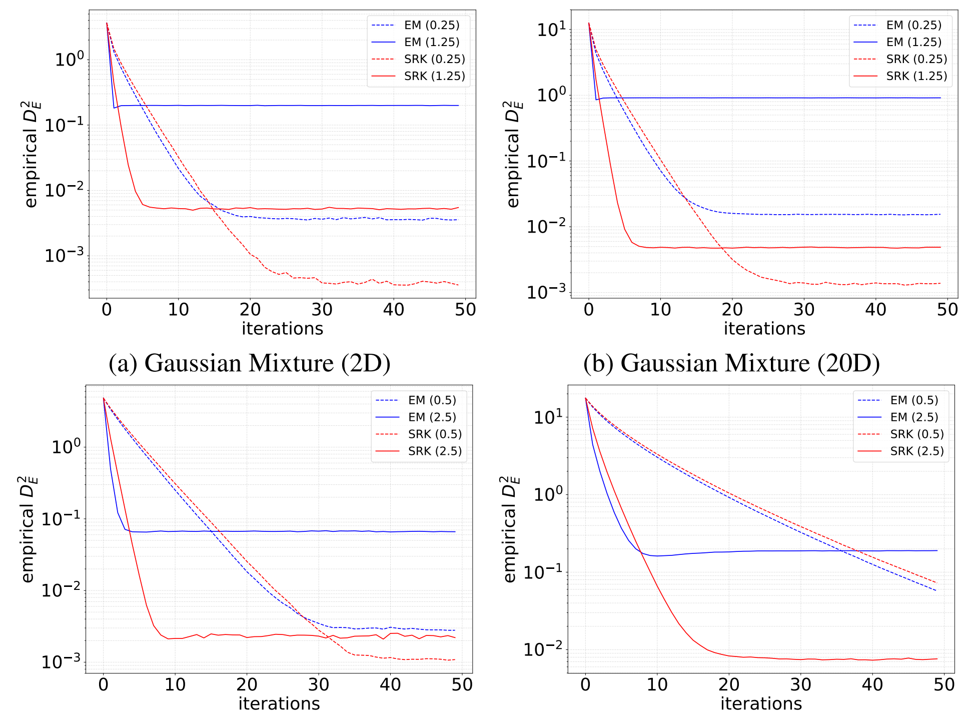
In addition to reporting the estimated squared W_2 values, we also evaluate the two schemes by estimating the energy distance [58, 59] under the Euclidean norm. For probability measures μ and ν on $\mathbb{R}^d$ with finite first moments, this distance is defined to be the square root of

$$D_E(\mu,\nu)^2 = 2\mathbb{E}\left[\|Y - Z\|_2\right] - \mathbb{E}\left[\|Y - Y'\|_2\right] - \mathbb{E}\left[\|Z - Z'\|_2\right], \tag{49}$$

where $Y, Y' \overset{\text{i.i.d.}}{\sim} \mu$ and $Z, Z' \overset{\text{i.i.d.}}{\sim} \nu$. The moment condition is required to ensure that the expectations in (49) is finite. This holds in our settings due to derived moment bounds. Since exactly computing the energy distance is intractable, we estimate the quantity using the following (biased) V-statistic [55]

$$\hat{D}_E(\mu,\nu)^2 = \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\|Y_i - Z_j\|_2 - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|Y_i - Y_j\|_2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|Z_i - Z_j\|_2\,,$$

where $Y_i \overset{\text{i.i.d.}}{\sim} \mu$ for $i = 1,\dots,m$ and $Z_j \overset{\text{i.i.d.}}{\sim} \nu$ for $j = 1,\dots,n$. Figure 5 shows the estimated energy distance as the number of iteration increases on a semi-log scale. We use 5k samples each for the Markov chain and the target distribution to compute the V-statistic, where the target distribution is approximated following the same procedure as described in Section 5.1. These plots show that SRK-LD achieves lower asymptotic errors compared to the EM scheme, where the error is measured in the energy distance. This is consistent with the case where the error is estimated in W_2^2.



(a) Gaussian Mixture (2D) (b) Gaussian Mixture (20D)

(c) BLR (2D) (d) BLR (20D)

Figure 5: Error in D_E^2 for strongly log-concave sampling. Legend denotes“scheme (step size)”.

H.1.2 Asymptotic Error vs Dimensionality and Step Size

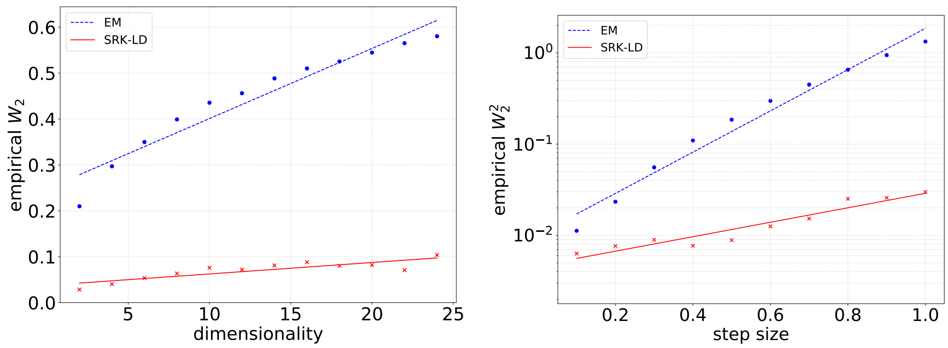
Figure 6 (a) and (b) respectively show the asymptotic error against dimensionality and step size for Gaussian mixture sampling. We perform least squares regression in both plots. Plot (a) shows results when a step size of 0.5 is used. Plot (b) is on semi-log scale, where the quantities are estimated for a 10D problem.

H.1.3 Wall Time

Figure 7 shows the wall time against the estimated W_2^2 of SRK-LD compared to the EM scheme for a 20D Gaussian mixture sampling problem. On a 6-core CPU with 2 threads per core, we observe that SRK-LD is roughly $\times$ 2.5 times as costly as EM per iteration. However, since SRK-LD is more stable for large step sizes, we may choose a step size much larger for SRK-LD compared to EM, in which case its iterates converge to a lower error within less time.

H.2 Non-Convex Potentials

We first discuss how we approximate the iterated Itô integrals, after which we include additional numerical studies varying the dimensionality of the sampling problem.



(a) dimensionality (b) step size

Figure 6: Asymptotic error vs dimensionality and step size.

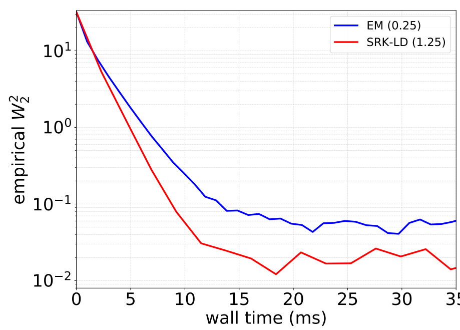


Figure 7: Wall time for sampling from a 20D Gaussian mixture.

H.2.1 Approximating Iterated Itô Integrals

Simulating both the iterated Itô integrals $I_{(l,i)}$ and the Brownian motion increments $I_{(i)}$ exactly is difficult. We adopt the Kloeden-Platen-Wright approximation, which has an MSE of order h^2/n, where n is the number of terms in the truncation [33]. The infinite series can be written as follows:

$$I_{(l,i)} = \frac{I_{(l)} I_{(i)} - h\delta_{li}}{2} + A_{(l,i)},$$
$$A_{(l,i)} = \frac{h}{2\pi} \sum_{k=1}^{\infty} \frac{1}{k} \left(\xi_{l,k} \left(\eta_{i,k} + \sqrt{2/h} \Delta B_h^{(i)} \right) - \xi_{i,k} \left(\eta_{l,k} + \sqrt{2/h} \Delta B_h^{(l)} \right) \right),$$

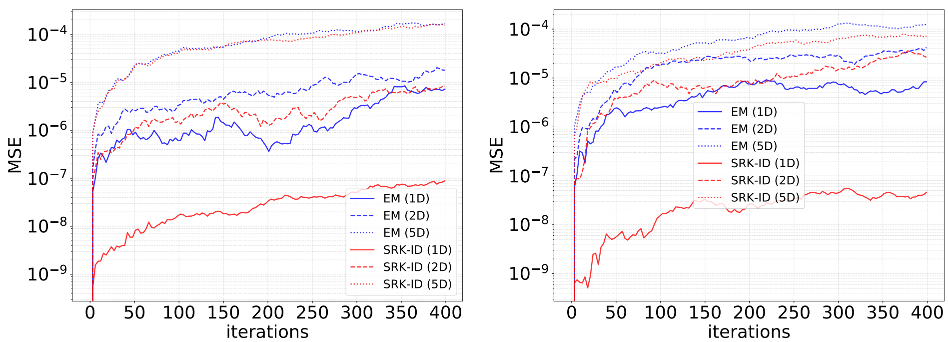
where $\xi_{l,k}, \xi_{i,k}, \eta_{i,k}, \eta_{l,k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. $A_{(l,i)}$ is known as the *Lévy area* and is notoriously hard to simulate [66].

For SDE simulation, in order for the scheme to obtain the same strong convergence order under the approximation, the MSE in the approximation of the Itô integrals must be negligible compared to the local mean-square deviation of the numerical integration scheme. For our experiments, we use $n = 3000$, following the rule of thumb that $n \propto h^{-1}$ [33]. Although simulating the extra terms can become costly, the computation may be vectorized, branched off from the main update, and parallelized on an additional thread, since it does not require any information of the current iterate.

Wiktorsson et al. [66] proposed to add a correction term to the truncated series, which results in an approximation that has an MSE of order h^2/n^2. In this case, $n \propto h^{-1/2}$ terms are effectively required. We note that analyzing and comparing between different Lévy area approximations is beyond the scope of this paper.

H.2.2 Additional Results

Figure 8 shows the MSE of simulations starting from a faithful approximation to the target. We adopt the same simulation settings as described in Section 5.2. We observe diminishing gains as the dimensionality increases across all settings with differing β and γ parameters in which we experimented. These empirical findings corroborate our theoretical results. Note that the corresponding diffusion in all settings are still uniformly dissipativity, yet the potential may become convex when β is large. Nevertheless, the potential is never strongly convex when β is positive due to the linear growth term.



(a) $\beta = 2, \gamma = 0.5$ (b) $\beta = 3, \gamma = 0.5$

Figure 8: MSE for non-convex sampling.