Stochastic Runge-Kutta Accelerates Langevin Monte Carlo and Beyond

Xuechen Li\textsuperscript{1,2}, Denny Wu\textsuperscript{1,2}, Lester Mackey\textsuperscript{3}, Murat A. Erdogdu\textsuperscript{1,2}

University of Toronto\textsuperscript{1}, Vector Institute\textsuperscript{2}, Microsoft Research\textsuperscript{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 \(O(d\epsilon^{-2/3})\) 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 \(\epsilon\). 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 \([2, 25, 38]\). 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 \([41, 2]\).

MCMC samplers based on numerically integrating continuous-time dynamics have proven very useful due to their ability to accomodate a stochastic gradient oracle \([61]\). Moreover, when used as optimizations algorithms, these methods can deliver strong theoretical guarantees in non-convex settings \([47]\). A popular example in this regime is the unadjusted Langevin Monte Carlo (LMC) algorithm \([48]\). 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 \([48, 24, 40]\), and non-asymptotic analyses specifying explicit constants in convergence bounds were recently conducted \([13, 10, 17, 6, 18, 8]\).

To the best of our knowledge, the best known rate of LMC in 2-Wasserstein distance is due to Durmus and Moulines \([17]\) – \(O(d\epsilon^{-1})\) iterations are required to reach \(\epsilon\) accuracy to \(d\)-dimensional target distributions with strongly convex potentials under the additional Lipschitz Hessian assumption, where \(O\) 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 \([61, 14]\). However, from the numerical integration point of view, the Euler-Maruyama
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 [44].

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 $O(d\epsilon^{-2/3})$. Similar to LMC, the algorithm only queries the gradient oracle of the potential during each update and improves upon the best known rate of $O(d\epsilon^{-1})$ for strongly log-concave sampling 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 rate of $\tilde{O}(d^{3/4}m^2\epsilon^{-1})$ for general Itô diffusions, where $m$ is the dimensionality of the Brownian motion. This improves upon the rate of $O(d\epsilon^{-2})$ for the Euler-Maruyama scheme [22] in terms of the tolerance $\epsilon$, 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 $\tilde{O}(d\epsilon^{-2/3})$ 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 non-convex potentials and achieves the convergence rate of $\tilde{O}(d^{3/4}m^2\epsilon^{-1})$.
- 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 [44, 30]. We refer the reader to [3] for a review and to [30] 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 [26, 58], 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. Sabanis and Zhang [50] introduced a numerical scheme that queries the gradient of the Laplacian based on an integrator that accommodates superlinear drifts [51]. In particular, for potentials with a Lipschitz gradient, they obtained the convergence rate of $\tilde{O}(d^{1/3}\epsilon^{-2/3})$. In optimization, high-order ordinary differential equation (ODE) integration schemes were introduced to discretize a second order ODE and achieved acceleration [64]. However, directly applying ODE schemes to SDEs usually produces unsatisfactory results, due to unaccounted terms from the Itô-Taylor expansion [30].

Non-Convex Learning. The convergence analyses of sampling using the overdamped and under-damped Langevin diffusion were extended to the non-convex setting [8, 37]. 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 [8, 36, 37]. More generally, Vempala and Wibisono [57] 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.
We study the problem of sampling from a target distribution $p$. We define the drift coefficient as (see e.g. [35, Thm. 2])

\[ \mu_0(f) = \sup_{x \in \mathbb{R}^d} \|f(x)\|_{\text{op}}, \quad \mu_i(f) = \sup_{x, y \in \mathbb{R}^d, x \neq y} \frac{\|\nabla^{i-1} f(x) - \nabla^{i-1} f(y)\|_2}{\|x - y\|_2^2}, \quad \text{and} \quad \pi_{i,n}(f) = \sup_{x \in \mathbb{R}^d} \frac{\|\nabla^{i-1} f(x)\|_2^n}{1 + \|x\|_2^n}, \]

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 $\|\cdot\|_F$ as $\mu^F_i(\cdot)$ and $\pi^F_{i,n}(\cdot)$, respectively.

### Coupling and Wasserstein Distance

We denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel $\sigma$-field of $\mathbb{R}^d$. Given probability measures $\nu$ and $\nu'$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we define a coupling (or transference plan) $\zeta$ between $\nu$ and $\nu'$ as a probability measure on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d) \times \mathcal{B}(\mathbb{R}^d))$ 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 $\nu$ and $\nu'$ as

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

### 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 [35, 42] given as the solution to the following stochastic differential equation (SDE):

\[ dX_t = b(X_t) \, dt + \sigma(X_t) \, dB_t, \quad \text{with} \quad X_0 = x_0, \]

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, $\{B_t\}_{t \geq 0}$ is an $m$-dimensional Brownian motion adapted to the filtration $\{\mathcal{F}_t\}_{t \geq 0}$. 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. [35, Thm. 2])

\[ b(x) = \frac{1}{2p(x)} (\nabla p(x) w(x)), \quad \text{where} \quad w(x) = \sigma(x)' \sigma(x) + c(x), \]

1 We obtain a rate in $W_2$ from the discretization analysis in KL [10] via standard techniques [47, 57].

2 Sabanis and Zhang [50] use the Frobenius norm for matrices and the Euclidean norm of Frobenius norms for 3-tensors. For a fair comparison, they convert their Lipschitz constants to be based on the operator norm.
Analogous to ODE solvers, numerical schemes such as the EM scheme and SRK schemes are where the expansion justifies the update rule of the EM scheme, since the discretization is nothing more than the local mean-square deviation and the local mean deviation at iteration \( k \). The instance of the same diffusion starting from \( \tilde{X}_k \) will be a more accurate local approximation than the EM scheme, due to accounting more accurately for Brownian integrals presents a challenge for simulation. Nevertheless, it is clear that the above SRK expansion, Itô’s lemma induces a stochastic version of the Taylor expansion of a smooth function \( p \) and \( \tilde{X}_k \) is the solution to the following SDE:
\[
\mathrm{d}X_t = -\nabla f(X_t) \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}B_t, \quad \text{with} \quad X_0 = x_0.
\] (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 [30], which relies on the following update rule,
\[
\tilde{X}_{k+1} = \tilde{X}_k + h \, b(\tilde{X}_k) + \sqrt{h} \, \sigma(\tilde{X}_k) \xi_{k+1}, \quad k = 0, 1, \ldots,
\] (4)
where \( h \) is the step size and \( \xi_{k+1} \mathrm{i.i.d.} \mathcal{N}(0, I_d) \) 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 \( \bar{p}(\tilde{x}) \) is different from the target distribution \( p(x) \); yet, for a sufficiently small step size, the difference between \( \bar{p}(\tilde{x}) \) and \( p(x) \) can be characterized (see e.g. [40, 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 [44], 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:
\[
L(g)(x) = \nabla g(x) \cdot b(x) + \frac{1}{2} \sum_{i,j=1}^m \nabla^2 g(x)[\sigma_i(x), \sigma_j(x)], \quad \Lambda_j(g)(x) = \nabla g(x) \cdot \sigma_j(x),
\] (5)
where \( \sigma_i(x) \) denotes the \( i \)th 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 \( X_0 \)

- mean-square order 1.0 stochastic Runge-Kutta update
  
  \[
  X_t = X_0 + h \sum_{i=1}^m \int_0^t \Lambda_i(g(X_u)) \, \mathrm{d}B_u^{(i)} + \int_0^t L(\sigma(X_u)) \, \mathrm{d}u + \int_0^t \sigma(X_u) \, \mathrm{d}B_u,
  \]

- Euler-Maruyama update
  
  \[
  \begin{aligned}
  X_t &= X_0 + h \sum_{i=1}^m \int_0^t \Lambda_i(g(X_u)) \, \mathrm{d}B_u^{(i)} + \sum_{i=1}^m \int_0^t \Lambda_i(\sigma(X_u)) \, \mathrm{d}u + \sum_{i=1}^m \int_0^t \sigma_i(\sigma(X_u)) \, \mathrm{d}B_u^{(i)} \, \mathrm{d}s.
  \end{aligned}
  \] (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 \( \{F_t\}_{t \geq 0} \). Suppose \( \{X_k^{(h)}\}_{k \geq 0} \) is another instance of the same diffusion starting from \( \tilde{X}_{k-1} \) at \( s = 0 \) and governed by the Brownian motion \( \{B_{h,k-1}(s-k)\}_{s \geq 0} \). Then, the numerical integration scheme has local deviation \( D_h^{(k)} = \tilde{X}_k - X_k^{(h)} \) with uniform orders \( (p_1, p_2) \) if
\[
\mathcal{E}_k^{(1)} = \mathbb{E} \left[ \|D_h^{(k)}\|_2^2 | F_{k-1} \right] \leq \lambda_1 h^{2p_1}, \quad \mathcal{E}_k^{(2)} = \mathbb{E} \left[ \|D_h^{(k)}| F_{k-1} \|_2^2 \right] \leq \lambda_2 h^{2p_2},
\] (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 [44], where the mean-square order is defined as the maximum positive half-integer \( p \) such that there exists a constant \( C \) independent of step size \( h \) with the following \( \mathbb{E}[\|X_{t_k} - \hat{X}_k\|^2] \leq Ch^{2p} \), for all \( k \in \mathbb{N} \) such that \( t_k < T \), where \( \{X_t\}_{t \geq 0} \) is the continuous-time process, \( \hat{X}_k \) 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 that we require the extra step of ensuring the expectations of \( e_k \) and \( e_k \) 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 Wasserstein-2 contraction and thereafter compute the uniform local deviation orders of the scheme.

**Theorem 1** (Wasserstein-2 rate). A diffusion \( X_t \), for \( t \in \mathbb{R}^d \rightarrow \mathbb{R} \), for two instances of the diffusion \( X_t \) initiated respectively from \( x \) and \( y \), the following holds

\[
W_2(\delta_{\sigma} P_t, \delta_{\sigma} P_t) \leq r(t) \|x - y\|_2, \quad \text{for all } x, y \in \mathbb{R}^d, t \geq 0,
\]

where \( \delta_{\sigma} P_t \) 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 \( r(t) = e^{-\alpha t} \) [22, Proposition 3.3].

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

\[
\langle b(x) - b(y), x - y \rangle + \frac{1}{2} \|\sigma(x) - \sigma(y)\|_F^2 \leq -\alpha \|x - y\|_2^2, \quad \text{for all } x, y \in \mathbb{R}^d.
\]

For Itô diffusions with a constant diffusion coefficient, uniform dissipativity is equivalent to one-sided Lipschitz continuity of the drift with coefficient \(-2\alpha\). 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** (Wasserstein-2-rate of a numerical scheme). For a diffusion with invariant measure \( \nu^* \), 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 \( \nu_k \) 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

\[
h < 1 \land C_h \land \frac{1}{2\alpha} \land \frac{1}{8\mu_1(b)^2 + 8\mu_1^p(\sigma)^2},
\]

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

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

Moreover, if 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 \( \tilde{O}(e^{-1/(p_1 - 1/2)}) \) iterations within a sufficiently small positive error \( \epsilon \).

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. [44]), and the convergence is provided by the exponential contracting $W_2$-rate 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 Langevin and general Itô diffusions when the EM scheme is used for discretization. For this scheme, local deviation orders are well-known (and straightforward to derive); thus, the convergence rates for the 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 EM scheme has local deviation orders of (1.5, 2.0) for Itô diffusions with constant diffusion coefficients and drift coefficients that are three-times differentiable with Lipschitz gradient and Hessian [44, 30]. Since the potential is strongly convex, the Langevin diffusion is uniformly dissipative and achieves exponential $W_2$-contraction. Elementary algebra shows that Markov chain moments are bounded [40, 22]. Therefore, Theorem 1 implies that the rate of the sampling is $O(d\varepsilon^{-1})$, where the dimension dependence can be extracted from the explicit bound. This recovers the result by Durmus and Moulines [17, Theorem 8].

**Example 2.** If a general Itô diffusion (1) with 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 [44] – this term is exact for Langevin diffusion. If we further have uniform dissipativity, it can be shown that Markov chain moments are bounded (e.g., Lemma A.2 of [22]). Hence, Theorem 1 concludes that the convergence rate is $O(d\varepsilon^{-2})$. We note that for the diffusion coefficient, we use the Frobenius norm for the Lipschitz and growth constants which potentially hides dimension dependence. The dimension dependence worsens if one were to convert all bounds to be based on the operator norm using the pessimistic inequality $\|\sigma(x)\|_F \leq (d^{1/2} + m^{1/2}) \|\sigma(x)\|_{op}$. Appendix D provides a convergence bound with explicit constants.

While verifying 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.

### 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 [44]. We refer to the sampling algorithm as SRK-LD. Specifically, given a sample from the previous iteration $\tilde{X}_k$,

$$
\tilde{H}_1 = \tilde{X}_k + \sqrt{2h} \left[ \frac{1}{2} + \frac{1}{\sqrt{6}} \right] \xi_{k+1} + \frac{1}{\sqrt{12}} \eta_{k+1},
$$

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

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

where $h$ is the step size and $\xi_{k+1}, \eta_{k+1} \sim_{i.i.d.} N(0, I_d)$ are independent of $\tilde{X}_k$ for all $k \in \mathbb{N}$.
Theorem 2 (SRK-LD). Let $\nu^*$ be the target distribution with a strongly convex potential that is four-times differentiable with Lipschitz continuous first three derivatives. Let $\nu_k$ be the distribution of the $k$th 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 $\epsilon$ error in $\mathcal{O}(d\epsilon^{-2/3})$ 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.

We emphasize that 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 $\mathcal{O}(d\epsilon^{-1})$ to $\mathcal{O}(d\epsilon^{-2/3})$, 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 [49] 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.

$$\tilde{H}_1^{(i)} = \tilde{X}_k + \sum_{j=1}^m \sigma_j(\tilde{X}_k) I_{(j,i)}^{(i)} = X_k - \sum_{j=1}^m \sigma_j(\tilde{X}_k) I_{(j,i)}^{(i)}, \quad \tilde{H}_2^{(i)} = X_k + \sum_{j=1}^m \sigma_j(\tilde{X}_k) I_{(j,i)}^{(i)}.$$  

where $I_{(i)} = \int_t^{t+h} dB_s^{(i)}$, $I_{(j,i)} = \int_t^{t+h} dB_s^{(j)} dB_s^{(i)}$, and $\{B_t^{(i)}\}_{t \geq 0}$ denotes the $i$th dimension of $\{B_t\}_{t \geq 0}$. We note that schemes of higher order exist for general diffusions, but they typically require advanced approximations of iterated Itô integrals of the form $\int_0^t \cdots \int_0^{t-h} dB_s^{(m)} dB_s^{(i)}$.

Theorem 3 (SRK-ID). For a uniformly dissipative diffusion with invariant measure $\nu_\omega$, Lipschitz drift and diffusion coefficients that have Lipschitz gradients, assume that the diffusion coefficient further satisfies the sublinear growth condition $\|\sigma(x)\|_\infty \leq \pi(1)(1+\|x\|_2^2)$ for all $x \in \mathbb{R}^d$. Let $\nu_k$ be the distribution of the $k$th 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 $\epsilon$ error in $\mathcal{O}(d^{3/4}m^2\epsilon^{-1})$ 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, accurate simulation of both the iterated Itô integrals $I_{(j,i)}^{(i)}$ and the Brownian motion increments $I_{(j,i)}$ is difficult. We comment on two possible approximations based on truncating an infinite series in Appendix G.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 $\Omega(n^{-1/d})$ [60], so we perform a heuristic correction whose description is in Appendix F.
We consider sampling from the following non-convex potential
\[ f(x) = (\beta + \|x\|^2_2)^{1/2} + \gamma \log(\beta + \|x\|^2_2), \quad x \in \mathbb{R}^d, \]
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 [28]. One
\footnote{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.}

### 5.1 Strongly Convex Potentials

**Gaussian Mixture.** We consider sampling from a multivariate Gaussian mixture with density
\[ \pi(\theta) \propto \exp\left(-\frac{1}{2}\|\theta - a\|^2_2\right) + \exp\left(-\frac{1}{2}\|\theta + a\|^2_2\right), \quad \theta \in \mathbb{R}^d, \]
where \( a \in \mathbb{R}^d \) is a parameter that measures the separation of two modes. The potential is strongly convex when \( \|a\|^2_2 < 1 \) and has Lipschitz gradient and Hessian [10]. Moreover, one can verify that its third derivative is also Lipschitz.

**Bayesian Logistic Regression.** We consider Bayesian logistic regression (BLR) [10]. Given data samples \( X = \{x_i\}_{i=1}^n \in \mathbb{R}^{n \times d}, Y = \{y_i\}_{i=1}^n \in \mathbb{R}^n \), 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 \( \theta \) with mean zero and covariance proportional to \( \Sigma_X^{-1} \), where \( \Sigma_X = X^\top X/n \) is the sample covariance matrix. We sample from the posterior density
\[ \pi(\theta) \propto \exp(-f(\theta)) = \exp\left(Y^\top X \theta - \sum_i^n \log(1 + \exp(-\theta^\top x_i)) - \frac{\alpha}{2} \|\Sigma_X^{1/2} \theta\|^2_2\right). \]
The potential is strongly convex and has Lipschitz gradient and Hessian [10]. One can also verify that it has a Lipschitz third derivative. 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 \(^3\) (KSD) [27, 9, 34] using the inverse multiquadratic kernel [27] 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. We include other details of the setup in Appendix G.1.1.

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 [54] along with a wall time analysis in Appendix G.

### 5.2 Non-Convex Potentials

We consider sampling from the following non-convex potential
\[ f(x) = (\beta + \|x\|^2_2)^{1/2} + \gamma \log(\beta + \|x\|^2_2), \quad x \in \mathbb{R}^d, \]
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 [28]. One
\footnote{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.}
can verify that when $\beta + \|x\|^2 < 1$ and $(4\gamma + 1)\|x\|^2 < (2\gamma + 1)\sqrt{\beta + \|x\|^2}$, the Hessian has a negative eigenvalue. The candidate diffusion, where the drift coefficient is given by (2) and diffusion coefficient $\sigma(x) = g(x)^{1/2}I_4$ with $g(x) = (\beta + \|x\|^2)^{1/2}$, is uniformly dissipative if $\frac{1}{2} - |\gamma - \frac{1}{2}|\frac{2}{\mu_1} - \frac{d}{8\beta^2} > 0$. Indeed, one can verify that $\mu_1(g) \leq 1$, $\mu_2(g) \leq \frac{2}{\beta^2}$, and $\mu_1(\sigma) \leq \frac{1}{\beta^2}$. Therefore,

$$\langle b(x) - b(y), x - y \rangle + \frac{1}{2} \|\sigma(x) - \sigma(y)\|^2 \leq -\left(\frac{1}{2} - |\gamma - \frac{1}{2}|\mu_2(\sigma) - \frac{d}{4}\mu_1(\sigma)^2\right)\|x - y\|^2,$$

$$\leq -\left(\frac{1}{2} - |\gamma - \frac{1}{2}|\frac{2}{\beta^2} - \frac{d}{8\beta^2}\right)\|x - y\|^2.$$

Moreover, $b$ and $\sigma$ 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 [49]. 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 G.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.

6 Discussions

We established convergence rates of samplings algorithms obtained by discretizing Itô diffusions with exponential $W_2$-contraction based on local properties of numerical schemes. The user-friendly conditions promote one to derive rates based on the uniform orders of the local deviation. In addition, we showed that discretizing diffusions with SRK schemes leads to improved rates in $W_2$ for both convex and non-convex potentials.

Despite focusing on SRK methods, Theorem 1 can be used to obtain convergence rates for other classes of schemes. For the underdamped Langevin diffusion, quasi-symplectic schemes that rely on Runge-Kutta-type update can achieve mean-square order 2.0 and beyond [43]. For general Itô diffusions, there exist schemes of mean-square order 1.5 and beyond, using the Fourier-Legendre series to approximate the Lévy area [32].

A direction of interest is to relax the $W_2$-contraction condition on the diffusion to $W_1$-contraction or $W_1$-decay. This would enable us to leverage results based on distant dissipativity, and consequently allow us to sample from a wider class of non-convex potentials [21]. Orthogonally, for the overdamped Langevin diffusion, the $W_2$-contraction condition may be relaxed to a log-Sobolev inequality condition on the target measure, if the discretization analysis is adapted to be based on the KL divergence [57, 47]. This would also broaden the class of non-convex potentials from which we can sample with theoretical guarantees.

Parallel to studying sampling from a mean-square convergence aspect, works in numerical analysis have established convergence results in the weak sense for SRK schemes applied to ergodic SDEs using aromatic trees and B-series [33, 59]. However, moment bounds in these works are proven by generic arguments [44, e.g. Lemma 2.2.2], and reasoning about the rate’s dimension dependence becomes less obvious. Refined non-asymptotic convergence bounds would provide more insight for these algorithms’ performance on practical problems.

Lastly, the convergence results in $W_2$ for SRK-LD can be augmented to yield a generalization bound for optimization when the excess risk is characterized using the Gibbs distribution [47].
Acknowledgements

We thank Mufan Li for insightful discussions, and Jimmy Ba, David Duvenaud and Taiji Suzuki for helpful comments on an early draft of this work. MAE is partially funded by CIFAR AI Chairs program at the Vector Institute.

References


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 \( t_k \) \((k = 0, 1, \ldots)\) 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 \( X_k \) the \( k \)th iterate of the Markov chain. In the following, we derive a recursion for the quantity

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

Fix \( k \in \mathbb{N} \). We define the process \( \{\bar{X}_k\}_{t \geq t_k} \) 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 \( \{\bar{X}_1\}_{t \geq 0} \) and the Markov chain \( \bar{X}_k \) \((k = 0, 1, \ldots)\) share the same Brownian motion \( \{B_t\}_{t \geq 0} \). 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 \( X_{t_k} \) and \( \bar{X}_{t_k} \) be coupled such that

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

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

\[
Z_s = (X_s - \bar{X}_s) - (X_{t_k} - \bar{X}_{t_k}).
\]

Note \( \int_{t_k}^{t_{k+1}} \sigma(X_s) \, dB_s - \int_{t_k}^{t_{k+1}} \sigma(\bar{X}_s) \, \bar{B}_s \) is a Martingale w.r.t. \( \{\mathcal{F}_t\}_{t \geq 0} \), 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

\[
\mathbb{E} \left[ Z_{t_{k+1}} \mid \mathcal{F}_{t_k} \right] = \int_{t_k}^{t_{k+1}} \mathbb{E} \left[ b(X_u) - b(\bar{X}_u) \mid \mathcal{F}_{t_k} \right] \, du.
\]

By Jensen’s inequality,

\[
\left\| \mathbb{E} \left[ Z_{t_{k+1}} \mid \mathcal{F}_{t_k} \right] \right\|_2 \leq h \int_{t_k}^{t_{k+1}} \mathbb{E} \left[ \left\| b(X_u) - b(\bar{X}_u) \right\|_2^2 \mid \mathcal{F}_{t_k} \right] \, du \leq \mu_1(b) h \int_{t_k}^{t_{k+1}} \mathbb{E} \left[ \left\| X_u - \bar{X}_u \right\|_2^2 \mid \mathcal{F}_{t_k} \right] \, du.
\]

For \( s \in [t_k, t_{k+1}] \), by Young’s inequality, Jensen’s inequality, and Itô isometry,

\[
\mathbb{E} \left[ \left\| X_s - \bar{X}_s \right\|_2^2 \mid \mathcal{F}_{t_k} \right] = \mathbb{E} \left[ \left\| X_{t_k} - \bar{X}_{t_k} + \int_{t_k}^{s} \left( b(X_u) - b(\bar{X}_u) \right) \, du + \int_{t_k}^{s} \left( \sigma(X_u) - \sigma(\bar{X}_u) \right) \, dB_u \right\|_2^2 \mid \mathcal{F}_{t_k} \right]
\]

\[
\leq 4 \left\| X_{t_k} - \bar{X}_{t_k} \right\|_2^2 + 4(s - t_k) \int_{t_k}^{s} \mathbb{E} \left[ \left\| b(X_u) - b(\bar{X}_u) \right\|_2^2 \mid \mathcal{F}_{t_k} \right] \, du
\]

\[
+ 4 \int_{t_k}^{s} \mathbb{E} \left[ \left\| \sigma(X_u) - \sigma(\bar{X}_u) \right\|_2^2 \mid \mathcal{F}_{t_k} \right] \, du
\]

\[
\leq 4 \left\| X_{t_k} - \bar{X}_{t_k} \right\|_2^2 + 4(s - t_k) \mu_1(b)^2 \int_{t_k}^{s} \mathbb{E} \left[ \left\| X_u - \bar{X}_u \right\|_2^2 \mid \mathcal{F}_{t_k} \right] \, du
\]

\[
+ 4 \mu_1^2(b) \int_{t_k}^{s} \mathbb{E} \left[ \left\| X_u - \bar{X}_u \right\|_2^2 \mid \mathcal{F}_{t_k} \right] \, du
\]

\[
\leq 4 \left\| X_{t_k} - \bar{X}_{t_k} \right\|_2^2 + 4 \left( \mu_1(b)^2 + \mu_1^2(b) \right) \int_{t_k}^{s} \mathbb{E} \left[ \left\| X_u - \bar{X}_u \right\|_2^2 \mid \mathcal{F}_{t_k} \right] \, du.
\]
By the integral form of Grönwall’s inequality for continuous functions,

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

Plugging this result into (12), by \( h < 1 / (8 \mu_1 (b)^2 + 8 \mu_1^P (\sigma)^2) \),

\[
\begin{align*}
\left\| \mathbb{E} \left[ Z_{t_{k+1}} \mid \mathcal{F}_{t_k} \right] \right\|_2^2 & \leq \frac{\mu_1 (b)^2 h}{\mathbb{E} \left[ \mathbb{E} \left[ Z_{t_{k+1}} \mid \mathcal{F}_{t_k} \right] \right]_2^2} \left[ \exp \left( 4 (\mu_1 (b)^2 + \mu_1^P (\sigma)^2) h \right) - 1 \right] \left\| X_{t_k} - \tilde{X}_{t_k} \right\|_2^2 \\
& \leq \frac{8 \mu_1 (b)^2 h^2}{\mu_1 (b)^2 + \mu_1^P (\sigma)^2} \left( \mu_1 (b)^2 + \mu_1^P (\sigma)^2 \right) \left\| X_{t_k} - \tilde{X}_{t_k} \right\|_2^2 \\
& \leq 8 \mu_1 (b)^2 h \left\| X_{t_k} - \tilde{X}_{t_k} \right\|_2^2 . \quad (13)
\end{align*}
\]

By direct expansion,

\[
\begin{align*}
\mathbb{E} \left[ \left\| X_{t_{k+1}} - X_{t_{k+1}} \right\|_2 \mid \mathcal{F}_{t_k} \right] &= \left\| X_{t_k} - \tilde{X}_{t_k} \right\|_2^2 + \mathbb{E} \left[ \left\| Z_{t_{k+1}} \right\|_2^2 \mid \mathcal{F}_{t_k} \right] + 2 \left\langle X_{t_k} - \tilde{X}_{t_k}, \mathbb{E} \left[ Z_{t_{k+1}} \mid \mathcal{F}_{t_k} \right] \right\rangle \\
& \quad + 2 \left\langle \mathbb{E} \left[ Z_{t_{k+1}} \mid \mathcal{F}_{t_k} \right], X_{t_k} - \tilde{X}_{t_k} \right\rangle \\
& \quad + 2 \left\langle \mathbb{E} \left[ Z_{t_{k+1}} \mid \mathcal{F}_{t_k} \right], \mathbb{E} \left[ Z_{t_{k+1}} \mid \mathcal{F}_{t_k} \right] \right\rangle . \quad (14)
\end{align*}
\]

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

\[
\begin{align*}
\mathbb{E} \left[ \left\| Z_{t_{k+1}} \right\|_2^2 \mid \mathcal{F}_{t_k} \right] & \leq \mathbb{E} \left[ \mathbb{E} \left[ \left\| Z_{t_{k+1}} \right\|_2^2 \mid \mathcal{F}_{t_k} \right] \right] \\
& \leq 8 \mu_1 (b) h \left\| X_{t_k} - \tilde{X}_{t_k} \right\|_2^2 \\
& \leq 8 \mu_1 (b) h \left\| X_{t_k} - \tilde{X}_{t_k} \right\|_2^2 . \quad (15)
\end{align*}
\]

Hence,

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

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

\[
A_{k+1}^2 = \mathbb{E} \left[ \left\| X_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2^2 \right] \\
= \mathbb{E} \left[ \left\| X_{t_{k+1}} - X_{t_{k+1}} + X_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2^2 \right] \\
= \mathbb{E} \left[ \left\| X_{t_{k+1}} - X_{t_{k+1}} \right\|_2^2 + \left\| X_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2^2 + 2 \left\langle X_{t_{k+1}} - X_{t_{k+1}} - \tilde{X}_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\rangle \right] \\
= \mathbb{E} \left[ \mathbb{E} \left[ \left\| X_{t_{k+1}} - X_{t_{k+1}} \right\|_2^2 \mid \mathcal{F}_{t_k} \right] \right] + \mathbb{E} \left[ \left\| X_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2^2 \mid \mathcal{F}_{t_k} \right] \\
+ 2 \mathbb{E} \left[ \left\langle X_{t_{k+1}} - X_{t_{k+1}}, X_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\rangle \mid \mathcal{F}_{t_k} \right] \\
= \mathbb{E} \left[ \left\| X_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2^2 \left\| \mathcal{F}_{t_k} \right\| + \mathbb{E} \left[ \left\| \tilde{X}_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2^2 \mid \mathcal{F}_{t_k} \right] \\
+ 2 \mathbb{E} \left[ \left\langle X_{t_{k+1}} - \tilde{X}_{t_{k+1}}, X_{t_{k+1}} - \tilde{X}_{t_{k+1}} \mid \mathcal{F}_{t_k} \right\rangle \right] \\
+ 2 \mathbb{E} \left[ \left\langle Z_{t_{k+1}}, \tilde{X}_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\rangle \mid \mathcal{F}_{t_k} \right] \\
\leq \mathbb{E} \left[ \left\| X_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2^2 \left\| \mathcal{F}_{t_k} \right\| + \mathbb{E} \left[ \left\| \tilde{X}_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2^2 \mid \mathcal{F}_{t_k} \right] \\
+ 2 \mathbb{E} \left[ \left\| X_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2 \right]^{1/2} \mathbb{E} \left[ \left\| \left\| X_{t_{k+1}} - \tilde{X}_{t_{k+1}} \mid \mathcal{F}_{t_k} \right\|_2 \right]^{1/2} \\
+ 2 \mathbb{E} \left[ \left\| X_{t_{k+1}} \right\|_2 \right]^{1/2} \mathbb{E} \left[ \left\| \tilde{X}_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2 \right]^{1/2} \\
+ 2 \mathbb{E} \left[ \left\| Z_{t_{k+1}} \right\|_2 \right]^{1/2} \mathbb{E} \left[ \left\| \tilde{X}_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2 \right]^{1/2} \right] \\
= \mathbb{E} \left[ \left\| X_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2^2 \right] + \mathbb{E} \left[ \left\| \tilde{X}_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2^2 \right] \\
+ 2 \mathbb{E} \left[ \left\| X_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2 \right]^{1/2} \mathbb{E} \left[ \left\| \left\| X_{t_{k+1}} - \tilde{X}_{t_{k+1}} \mid \mathcal{F}_{t_k} \right\|_2 \right]^{1/2} \\
+ 2 \mathbb{E} \left[ \left\| X_{t_{k+1}} \right\|_2 \right]^{1/2} \mathbb{E} \left[ \left\| \tilde{X}_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2 \right]^{1/2} \\
+ 2 \mathbb{E} \left[ \left\| Z_{t_{k+1}} \right\|_2 \right]^{1/2} \mathbb{E} \left[ \left\| \tilde{X}_{t_{k+1}} - \tilde{X}_{t_{k+1}} \right\|_2 \right]^{1/2} .
\]
Verifying the order conditions in Theorem 1 for SRK-LD requires bounding the second, fourth, and where the third to last inequality follows from \(e^{-2ah} < 1 - \alpha h\) when \(\alpha h < 1/2\), and the second to last inequality follows from the elementary relation below with the choice of \(\kappa = \alpha/2\)

\[
A_k h^{1/2} \cdot \lambda_3 h^{p_1} \leq \kappa A_k^2 h + \frac{4}{\kappa} \lambda_2^2 h^{2p_1}.
\]

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

\[
A_k^2 \leq (1 - \alpha h/2) A_{k-1}^2 + (8\lambda_2^2/\alpha + \lambda_1) h^{2p_1}
\]

\[
\leq \eta A_0^2 + (1 + \eta + \cdots + \eta^{k-1}) (8\lambda_2^2/\alpha + \lambda_1) h^{2p_1}
\]

\[
\leq \eta^k A_0^2 + (8\lambda_2^2/\alpha + \lambda_1) h^{2p_1}/(1 - \eta)
\]

\[
= \eta^k A_0^2 + (16\lambda_2^2/\alpha^2 + 2\lambda_1/\alpha) h^{2p_1-1}.
\]

Let \(\nu_k\) and \(\nu^*\) be the measures associated with the \(k\)th iterate of the Markov chain and the target distribution, respectively. Since \(W_2\) is defined as an infimum over all couplings,

\[
W_2(\nu_k, \nu^*) \leq A_k \leq e^{-\alpha h k/4} A_0 + (16\lambda_2^2/\alpha^2 + 2\lambda_1/\alpha) h^{1/2} p_1 - 1/2
\]

To ensure \(W_2\) is less than some small positive tolerance \(\epsilon\), 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

\[
h < \left(\frac{2}{\epsilon} \sqrt{\frac{64(16\lambda_1 \mu_1(b) + \lambda_2)}{\alpha^2} + \frac{2\lambda_1}{\alpha}} \right)^{-1/(p_1-1/2)} \land \frac{1}{2\alpha} \land \frac{1}{8\mu_1(b)^2 + 8\mu_1^2(\sigma)^2}.
\]

\[
k > \left[\frac{2}{\epsilon} \sqrt{\frac{64(16\lambda_1 \mu_1(b) + \lambda_2)}{\alpha^2} + \frac{2\lambda_1}{\alpha}} \right]^{1/(p_1-1/2)} \lor 2\alpha \lor \left(8\mu_1(b)^2 + 8\mu_1^2(\sigma)^2\right) \frac{4}{\alpha} \log \left(\frac{2A_0}{\epsilon}\right).
\]

Note that for small enough positive tolerance \(\epsilon\), when the step size satisfies (15), it suffices that

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

\[
\square
\]

### 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 2

\[th\] 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 [22]. Here, \(\alpha\) can be chosen as the strong convexity parameter, provided \(\beta\) 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 $k$th iterate is $t_k$ as per the proof of Theorem 1. To simplify notation, we define the following
\[
\begin{align*}
\hat{\nabla} f &= \frac{1}{2} \left( \nabla f(\hat{H}_1) + \nabla f(\hat{H}_2) \right), \\
v_1 &= \sqrt{2} \left( \frac{1}{2} + \frac{1}{\sqrt{6}} \right) \xi_{k+1} \sqrt{\eta}, \\
v_1' &= \sqrt{2} \left( \frac{1}{2} - \frac{1}{\sqrt{6}} \right) \xi_{k+1} \sqrt{\eta}, \\
v_2 &= \frac{1}{\sqrt{6}} \eta_{k+1} \sqrt{\eta},
\end{align*}
\]
where $\xi_{k+1}, \eta_{k+1} \overset{i.i.d.}{\sim} N(0, I_d)$ independent of $\hat{X}_k$ for all $k \in \mathbb{N}$. We rewrite $\hat{H}_1$ and $\hat{H}_2$ as
\[
\begin{align*}
\hat{H}_1 &= \hat{X}_k + \Delta \hat{H}_1 = \hat{X}_k + v_1 + v_2, \\
\hat{H}_2 &= \hat{X}_k + \Delta \hat{H}_2 = \hat{X}_k + v_1' + v_2 - \nabla f(\hat{X}_k) h.
\end{align*}
\]

### 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.
\[
\mathbb{E} \left[ \left\| \hat{X}_k \right\|_2^2 \right] \leq \mathcal{U}_2, \quad \text{for all } k \in \mathbb{N},
\]
where $\mathcal{U}_2 = \mathbb{E} \left[ \left\| \hat{X}_0 \right\|_2^2 \right] + N_0$, 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,
\[
\begin{align*}
\left\| \hat{X}_{k+1} \right\|_2^2 &= \left\| \hat{X}_k - \left( \nabla f(\hat{H}_1) + \nabla f(\hat{H}_2) \right) \frac{h}{2} + 2^{1/2} \xi_{k+1} h^{1/2} \right\|_2^2 \\
&= \left\| \hat{X}_k \right\|_2^2 + \left\| \nabla f(\hat{H}_1) + \nabla f(\hat{H}_2) \right\|_2^2 h^2 + 2 \left\| \xi_{k+1} \right\|_2^2 h \\
&\quad - 2^{1/2} \left\langle \hat{X}_k, \nabla f(\hat{H}_1) + \nabla f(\hat{H}_2) \right\rangle h \\
&\quad + 2^{3/2} \left\langle \hat{X}_k, \xi_{k+1} \right\rangle h^{1/2} \\
&\quad - 2^{1/2} \left\langle \nabla f(\hat{H}_1) + \nabla f(\hat{H}_2), \xi_{k+1} \right\rangle h^{3/2}.
\end{align*}
\]
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 $\hat{H}_1$ and $\hat{H}_2$ for $h < 2d \wedge 2\pi_{2,1}(f) / \pi_{2,2}(f)$,
\[
\begin{align*}
\mathbb{E} \left[ \left\| \hat{H}_1 \right\|_2^2 \left| \mathcal{F}_{t_k} \right\| \right] &= \left\| \hat{X}_k \right\|_2^2 + \mathbb{E} \left[ \left\| v_1 \right\|_2^2 \left| \mathcal{F}_{t_k} \right\| \right] + \mathbb{E} \left[ \left\| v_2 \right\|_2^2 \left| \mathcal{F}_{t_k} \right\| \right] \leq \left\| \hat{X}_k \right\|_2^2 + 3dh, \\
\mathbb{E} \left[ \left\| \hat{H}_2 \right\|_2^2 \left| \mathcal{F}_{t_k} \right\| \right] &= \left\| \hat{X}_k \right\|_2^2 + \left\| \nabla f(\hat{X}_k) \right\|_2^2 h^2 + \mathbb{E} \left[ \left\| v_1' \right\|_2^2 \left| \mathcal{F}_{t_k} \right\| \right] + \mathbb{E} \left[ \left\| v_2 \right\|_2^2 \left| \mathcal{F}_{t_k} \right\| \right] \\
&\quad + 2 \left\langle \hat{X}_k, \nabla f(\hat{X}_k) \right\rangle h \\
&\quad \leq \left\| \hat{X}_k \right\|_2^2 + \pi_{2,2}(f) \left( 1 + \left\| \hat{X}_k \right\|_2^2 \right) h^2 + dh + 2\pi_{2,1}(f) \left\| \hat{X}_k \right\|_2^2 h \\
&\quad \leq \left\| \hat{X}_k \right\|_2^2 + 4\pi_{2,1}(f) h \left\| \hat{X}_k \right\|_2^2 + 3dh.
\end{align*}
\]
Thus,

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

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

Additionally, by the Cauchy-Schwarz inequality,

\[
-\mathbb{E} \left[ \left\langle \nabla f(\tilde{H}_1), \xi_{k+1} \right\rangle \right] \leq \mathbb{E} \left[ \left\| \nabla f(\tilde{H}_1) \right\|_{\mathcal{F}_{t_k}} \left\| \xi_{k+1} \right\|_{\mathcal{F}_{t_k}} \right] \\
\leq \mathbb{E} \left[ \left\| \nabla f(\tilde{H}_1) \right\|^2_{\mathcal{F}_{t_k}} \right]^{1/2} \mathbb{E} \left[ \left\| \xi_{k+1} \right\|^2_{\mathcal{F}_{t_k}} \right]^{1/2} \\
\leq \sqrt{d\pi_{2,2}(f)} \left( 1 + \mathbb{E} \left[ \left\| \tilde{H}_1 \right\|^2_{\mathcal{F}_{t_k}} \right]^{1/2} \right) \\
\leq \sqrt{d\pi_{2,2}(f)} \left( 1 + \left\| \tilde{X}_k \right\|^2_{\mathcal{F}_{t_k}} + \sqrt{3dh} \right). \tag{16}
\]

Similarly,

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

Combining (16) and (17), we obtain the following using AM–GM,

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

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,

\[
\mathbb{E} \left[ \left\| \tilde{H}_1 \right\|^2_{\mathcal{F}_{t_k}} \right] = \mathbb{E} \left[ \left\| \tilde{X}_k + v_1 + v_2 \right\|^2_{\mathcal{F}_{t_k}} \right] \\
= \left\| \tilde{X}_k \right\|^2_{\mathcal{F}_{t_k}} + \mathbb{E} \left[ \left\| v_1 \right\|^2_{\mathcal{F}_{t_k}} \right] + \mathbb{E} \left[ \left\| v_2 \right\|^2_{\mathcal{F}_{t_k}} \right] \geq \left\| \tilde{X}_k \right\|^2_{\mathcal{F}_{t_k}}, \tag{18}
\]

\[
\mathbb{E} \left[ \left\| \tilde{H}_2 \right\|^2_{\mathcal{F}_{t_k}} \right] = \mathbb{E} \left[ \left\| \tilde{X}_k - \nabla f(\tilde{X}_k)h + v_1' + v_2' \right\|^2_{\mathcal{F}_{t_k}} \right] \\
= \left\| \tilde{X}_k \right\|^2_{\mathcal{F}_{t_k}} + \left\| \nabla f(\tilde{X}_k) \right\|^2_{\mathcal{F}_{t_k}} h^2 + \mathbb{E} \left[ \left\| v_1' \right\|^2_{\mathcal{F}_{t_k}} \right] + \mathbb{E} \left[ \left\| v_2' \right\|^2_{\mathcal{F}_{t_k}} \right] \\
+ 2 \left\langle \tilde{X}_k, \nabla f(\tilde{X}_k) \right\rangle h \\
\geq \left\| \tilde{X}_k \right\|^2_{\mathcal{F}_{t_k}} + 2 \left( \frac{\alpha}{2} \left\| \tilde{X}_k \right\|^2_{\mathcal{F}_{t_k}} - \beta \right) h
\]
Additionally, by Stein’s lemma for multivariate Gaussians,

\[ E \left[ \langle \nabla f(\tilde{H}_1), v_1 \rangle \big| \mathcal{F}_{t_k} \right] = 2h \left( \frac{1}{2} + \frac{1}{\sqrt{6}} \right)^2 \leq 2d\mu_3(f)h, \]

\[ E \left[ \langle \nabla f(\tilde{H}_1), v_2 \rangle \big| \mathcal{F}_{t_k} \right] = \frac{1}{6} h E \left[ \Delta(f)(\tilde{H}_1) \big| \mathcal{F}_{t_k} \right] \leq \frac{1}{6} d\mu_3(f)h, \]

\[ E \left[ \langle \nabla f(\tilde{H}_2), v'_1 \rangle \big| \mathcal{F}_{t_k} \right] = 2h \left( \frac{1}{2} - \frac{1}{\sqrt{6}} \right)^2 \leq d\mu_3(f)h, \]

\[ E \left[ \langle \nabla f(\tilde{H}_2), v_2 \rangle \big| \mathcal{F}_{t_k} \right] = \frac{1}{6} h E \left[ \Delta(f)(\tilde{H}_2) \big| \mathcal{F}_{t_k} \right] \leq d\mu_3(f)h. \]

Therefore, by dissipativity and the lower bound (18),

\[ -E \left[ \langle \nabla f(\tilde{H}_1), \tilde{X}_k \rangle \big| \mathcal{F}_{t_k} \right] = -E \left[ \langle \nabla f(\tilde{H}_1), \tilde{H}_1 \rangle \big| \mathcal{F}_{t_k} \right] + E \left[ \langle \nabla f(\tilde{H}_1), v_1 + v_2 \rangle \big| \mathcal{F}_{t_k} \right] \]

\[ \leq -\frac{\alpha}{2} E \left[ \| \tilde{H}_1 \|^2 \big| \mathcal{F}_{t_k} \right] + \beta + E \left[ \langle \nabla f(\tilde{H}_1), v_1 + v_2 \rangle \big| \mathcal{F}_{t_k} \right] \]

\[ \leq -\frac{\alpha}{2} \| \tilde{X}_k \|^2 + \beta + 3d\mu_3(f)h. \tag{19} \]

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

\[ E \left[ \| \Delta \tilde{H}_2 \|^2 \big| \mathcal{F}_{t_k} \right] = E \left[ \| -\nabla f(\tilde{X}_k)h + v'_1 + v_2 \|^2 \big| \mathcal{F}_{t_k} \right] \]

\[ = \| \nabla f(\tilde{X}_k) \|^2 h^2 + E \left[ \| v'_1 \|^2 \big| \mathcal{F}_{t_k} \right] + E \left[ \| v_2 \| \big| \mathcal{F}_{t_k} \right] \]

\[ \leq \pi_2(f) \left( 1 + \| \tilde{X}_k \|^2 \right) h^2 + dh. \tag{20} \]

Notice the second equality above also implies

\[ \| \nabla f(\tilde{X}_k) \|^2 h \leq E \left[ \| \Delta \tilde{H}_2 \|^2 \big| \mathcal{F}_{t_k} \right]^{1/2}. \tag{21} \]

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

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

Since \(\nabla f\) is Lipschitz, \(\nabla^2 f\) is bounded, and

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

By (20) and (21),

\[ -E \left[ \langle \nabla f(\tilde{H}_2), \nabla f(\tilde{X}_k) \rangle \big| \mathcal{F}_{t_k} \right] \leq \| R(t_{k+1}) \|_{\mathcal{F}_{t_k}} \| \nabla f(\tilde{X}_k) \|^2 \]

\[ \leq \| R(t_{k+1}) \|_{\mathcal{F}_{t_k}} \| \nabla f(\tilde{X}_k) \|^2 \]

\[ \leq \mu_2(f) \| \Delta \tilde{H}_2 \|^2 |F_{t_k}| \| \nabla f(\tilde{X}_k) \|^2 \]

\[ \leq \mu_2(f) \| \Delta \tilde{H}_2 \|^2 |F_{t_k}| \| \nabla f(\tilde{X}_k) \|^2 \]

\[ \leq \mu_2(f) E \left[ \| \Delta \tilde{H}_2 \|^2 \big| \mathcal{F}_{t_k} \right] \| \nabla f(\tilde{X}_k) \|^2 \]

\[ \leq \mu_2(f) E \left[ \| \Delta \tilde{H}_2 \|^2 \big| \mathcal{F}_{t_k} \right] h^{-1} \]
\[ \leq \mu_2(f) \pi_{2,2}(f) \left( 1 + \| \tilde{X}_k \|_2^2 \right) h + d. \]

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

\[ -\mathbb{E} \left[ \langle \nabla f(\tilde{H}_2), \tilde{X}_k \rangle \mid \mathcal{F}_{t_k} \right] \]
\[ = -\mathbb{E} \left[ \langle \nabla f(\tilde{H}_2), \tilde{H}_2 \rangle + \langle \nabla f(\tilde{H}_2), \nabla f(\tilde{X}_k) \rangle h - \langle \nabla f(\tilde{H}_2), v_1 + v_2 \rangle \mid \mathcal{F}_{t_k} \right] \]
\[ \leq -\frac{\alpha}{2} \mathbb{E} \left[ \| \tilde{H}_2 \|_2^2 \mid \mathcal{F}_{t_k} \right] + \beta - \mathbb{E} \left[ \langle \nabla f(\tilde{H}_2), \nabla f(\tilde{X}_k) \rangle \mid \mathcal{F}_{t_k} \right] h + \mathbb{E} \left[ \langle \nabla f(\tilde{H}_2), v_1 + v_2 \rangle \mid \mathcal{F}_{t_k} \right] \]
\[ \leq -\frac{\alpha}{2} \| \tilde{X}_k \|_2^2 + \alpha \beta h + \beta + \mu_2(f)\pi_{2,2}(f) \left( 1 + \| \tilde{X}_k \|_2^2 \right) h^2 + dh + 2d\mu_3(f)h \]
\[ \leq -\frac{\alpha}{4} \| \tilde{X}_k \|_2^2 + (\alpha \beta + \mu_2(f)\pi_{2,2}(f)) h + \beta. \quad (22) \]

Combining (19) and (22), we have

\[ -\mathbb{E} \left[ \langle \nabla f(\tilde{H}_2), \nabla f(\tilde{H}_2), \tilde{X}_k \rangle \mid \mathcal{F}_{t_k} \right] \leq -\frac{3}{4} \alpha \| \tilde{X}_k \|_2^2 + N_{5}, \quad (23) \]

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

\[ \mathbb{E} \left[ \| \tilde{X}_{k+1} \|_2^2 \mid \mathcal{F}_{t_k} \right] = \| \tilde{X}_k \|_2^2 + \mathbb{E} \left[ \| \nabla f(\tilde{H}_1) + \nabla f(\tilde{H}_2) \|_2^2 \mid \mathcal{F}_{t_k} \right] \frac{h^2}{4} + 2dh \]
\[ - \mathbb{E} \left[ \langle \tilde{X}_k, \nabla f(\tilde{H}_1) + \nabla f(\tilde{H}_2) \rangle \mid \mathcal{F}_{t_k} \right] h \]
\[ - 2^{1/2} \mathbb{E} \left[ \langle \nabla f(\tilde{H}_1) + \nabla f(\tilde{H}_2), \tilde{X}_{k+1} \rangle \mid \mathcal{F}_{t_k} \right] h^{3/2} \]
\[ \leq \| \tilde{X}_k \|_2^2 + \frac{N_1}{4} \| \tilde{X}_k \|_2^2 h^2 + \frac{N_2}{4} h^2 + 2dh \]
\[ - \frac{3}{4} \alpha h \| \tilde{X}_k \|_2^2 + N_5 h \]
\[ + \frac{1}{2} \| \tilde{X}_k \|_2^2 h^2 + \frac{N_2}{2} h + N_4 h^{3/2} \]
\[ \leq \left( 1 - \frac{3}{4} \alpha h + \frac{N_1 + 2}{4} h^2 \right) \| \tilde{X}_k \|_2^2 \]
\[ + N_2 h^2/4 + 2dh + N_5 h + N_3 h/2 + N_4 h^{3/2} \]
\[ \leq \left( 1 - \frac{3}{8} \alpha h \right) \| \tilde{X}_k \|_2^2 + 2N_2 h^2/4 + 2dh + N_5 h + N_3 h/2 + N_4 h^{3/2}, \]

For \( h < 1 \), by unrolling the recursion, we obtain the following

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

where

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

\[ \square \]

**B.1.2 2nthMoment Bound**

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

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

20
where
\[ U_{2n} = \mathbb{E} \left[ \| \tilde{X}_0 \|_{12}^{2n} \right] + \frac{8}{3\alpha n} (N_{7,n} + N_{12,n}), \]
and constants \( N_{7,n} \) to \( N_{12,n} \) are given in the proof, if the step size
\[
 h < 1 \wedge \frac{2d}{\pi_2(f)} \wedge \frac{2\pi_2(f)}{\pi_2(f)} \wedge \frac{\alpha}{4\mu_2(f)\pi_2(f)} \wedge \frac{3\alpha}{2N_1 + 4} \wedge \min \left\{ \left( \frac{3\alpha l}{8N_{11,l}} \right)^2 : l = 2, \ldots, n \right\}.
\]

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

By the multinomial theorem,
\[
\mathbb{E} \left[ \| \tilde{X}_{k+1} \|_{12}^{2n} \right] = \mathbb{E} \left[ \left\| X_k - \nabla f h + 2^{1/2} \xi_{k+1} h^{1/2} \right\|_{12}^{2n} \right]
\]
\[
= \mathbb{E} \left[ \left( \| X_k \|_2^2 + \| \nabla f \|_2^2 h^2 + 2 \| \xi_{k+1} \|_2^2 h \right. \right.
\]
\[
\left. \left. - 2 \left\langle X_k, \nabla f \right\rangle h + 2^{3/2} \left( \tilde{X}_k, \xi_{k+1} \right) h^{1/2} - 2^{3/2} \left( \tilde{X}_k, \xi_{k+1} \right) h^{3/2} \right)^n \right]
\]
\[
= \mathbb{E} \left[ \sum_{k_1 + \cdots + k_6 = n} (-1)^{k_4 + k_6} \binom{n}{k_1 \ldots k_6} 2^{k_3 + k_5 + 3k_6} \frac{\| \tilde{X}_k \|_2^{2k_1 + k_4 + k_5} \| \tilde{\nabla f} \|_2^{2k_2 + k_4 + k_6} \| \xi_{k+1} \|_2^{2k_3 + k_5 + k_6} \right]
\]
\[
= \mathbb{E} \left[ \| \tilde{X}_k \|_{12}^{2n} + Ah + Bh^{3/2} \right],
\]
where
\[ A = 2n \| \tilde{X}_k \|_{12}^{2(n-1)} \| \xi_{k+1} \|_2^2 - 2n \| \tilde{X}_k \|_{12}^{2(n-1)} \left\langle \tilde{X}_k, \nabla f \right\rangle + 4n(n - 1) \left\| \tilde{X}_k \right\|_{12}^{2(n-2)} \left\langle \tilde{X}_k, \xi_{k+1} \right\rangle \right]^2,
\]
\[ B \leq \sum_{k_1 + \cdots + k_6 = n} 2^{3n} \binom{n}{k_1 \ldots k_6} \| \tilde{X}_k \|_{12}^{2k_1 + k_4 + k_5} \| \tilde{\nabla f} \|_{12}^{2k_2 + k_4 + k_6} \| \xi_{k+1} \|_{12}^{2k_3 + k_5 + k_6}.
\]

Now, we bound the expectation of \( A \) using (23),
\[
\mathbb{E} [A | F_{t_k}] \leq 2dn \left\| \tilde{X}_k \right\|_{12}^{2(n-1)} + 2n \left\| \tilde{X}_k \right\|_{12}^{2(n-1)} \left\langle \tilde{X}_k, \nabla f \right\rangle + 4dn(n - 1) \left\| \tilde{X}_k \right\|_{12}^{2(n-1)}
\]
\[
\leq - \frac{3}{4} \alpha n \left\| \tilde{X}_k \right\|_{12}^{2n} + (2dn + nN_5 + 4dn(n - 1)) \left\| \tilde{X}_k \right\|_{12}^{2(n-1)}.
\]
Moreover, by the inductive hypothesis,
\[
\mathbb{E} [A] = \mathbb{E} \mathbb{E} [A | F_{t_k}] \leq - \frac{3}{4} \alpha n \mathbb{E} \left[ \left\| \tilde{X}_k \right\|_{12}^{2n} \right] + N_{7,n}, \tag{24}
\]
where \( N_{7,n} = (2dn + nN_5 + 4dn(n - 1)) U_{2(n-1)} = O(d^n) \).

Next, we bound the expectation of \( B \). By the Cauchy–Schwarz inequality,
\[
\mathbb{E} [B | F_{t_k}] = \sum_{k_1 + \cdots + k_6 = n} 2^{3n} \binom{n}{k_1 \ldots k_6} \| \tilde{X}_k \|_{12}^{2k_1 + k_4 + k_5} \mathbb{E} \left[ \| \tilde{\nabla f} \|_{12}^{2k_2 + k_4 + k_6} \| \xi_{k+1} \|_{12}^{2k_3 + k_5 + k_6} | F_{t_k} \right]
\]
\[
\leq \sum_{k_1 + \cdots + k_6 = n} 2^{3n} \binom{n}{k_1 \ldots k_6} \| \tilde{X}_k \|_{12}^{2k_1 + k_4 + k_5}.
\]

21
where the second to last inequality follows from Young’s inequality for products with three variables. Therefore, for positive integer $H$

$$\|H_1\|_2^{2p} = \left(\|\tilde{X}_k\|^2 + (v_1 + v_2)^2_2 \right)$$

$$\leq \sum_{j_1 + \cdots + j_6 \leq p} 2^{j_1 + j_2 + j_5} \|\tilde{X}_k\|^2_2 \left(\|v_1\|^2 + (v_2)^2_2 \right)$$

$$\leq \sum_{j_1 + \cdots + j_6 \leq p} 2^{j_1 + j_2 + j_5} \|\tilde{X}_k\|^2_2 \left(\|v_1\|^2 + (v_2)^2_2 \right)$$

$$\leq \sum_{j_1 + \cdots + j_6 \leq p} 2^{j_1 + j_2 + j_5} \|\tilde{X}_k\|^2_2 \left(\|v_1\|^2 + (v_2)^2_2 \right)$$

$$\leq 2^{4p} \|\tilde{X}_k\|^2_2 + \|\xi_{k+1}\|^2_2 + \|\eta_{k+1}\|^2_2,$$

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

$$\mathbb{E} \left[ \|H_1\|_2^{2p} | \mathcal{F}_{t_k} \right] \leq 2^{4p} \|\tilde{X}_k\|^2_2 + 2^{4p+1} \|\tilde{X}_k\|^2_2 \mathbb{E} [\chi(d)^{2p}]. \quad (25)$$

Similarly,

$$\|\tilde{H}_2\|^2_2 = \left(\|\tilde{X}_k\|^2 + (v_1 + v_2)^2_2 \right)$$

$$\leq \left(\|\tilde{X}_k\|^2 + (v_1 + v_2)^2_2 \right)$$

$$\leq \left(\|\tilde{X}_k\|^2 + (v_1 + v_2)^2_2 \right)$$

$$\leq \left(\|\tilde{X}_k\|^2 + (v_1 + v_2)^2_2 \right)$$

$$\leq \left(\|\tilde{X}_k\|^2 + (v_1 + v_2)^2_2 \right)$$

$$\leq \left(\|\tilde{X}_k\|^2 + (v_1 + v_2)^2_2 \right)$$

$$\leq \left(\|\tilde{X}_k\|^2 + (v_1 + v_2)^2_2 \right)$$

$$\leq \left(\|\tilde{X}_k\|^2 + (v_1 + v_2)^2_2 \right)$$

Therefore,

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

Thus, combining (25) and (26),

$$\mathbb{E} \left[ \|\nabla f\|_2^{2p} | \mathcal{F}_{t_k} \right] \leq \frac{1}{2} \mathbb{E} \left[ \|\nabla f(\tilde{H}_1)\|_2^{2p} + \|\nabla f(\tilde{H}_2)\|_2^{2p} | \mathcal{F}_{t_k} \right]$$

$$\leq \frac{1}{2} \pi_{2,2p}(f) \mathbb{E} \left[ \left(\tilde{H}_1\right)^{2p} + \left(\tilde{H}_2\right)^{2p} | \mathcal{F}_{t_k} \right]$$

$$\leq N_{8,n}(p)^2 \|\tilde{X}_k\|^2_2 + N_{9,n}(p)^2,$$
where the \( p \)-dependent constants are

\[
N_{8,n}(p) = 2^{2p\frac{3}{2}} \left( \pi_{2,2p}(f) \left( 1 + \frac{1}{2} \pi_{2,2p}(f) \right) \right)^{\frac{1}{2}},
\]

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

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

\[
N_{8,n} = \max\{ N_{8,n}(2k_2+k_4+k_6) : k_1, \ldots, k_6 \in \mathbb{N}, k_1 + \cdots + k_6 = n, 2k_2 + k_4 + k_6 + \frac{k_4}{2} + \frac{3k_4}{2} > 1 \}.
\]

The bound on \( B \) reduces to

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

where

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

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

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,

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

Hence,

\[
\mathbb{E} \left[ B_1 \mid \mathcal{F}_k \right] \leq \sum_{k_1 + \cdots + k_6 = n} \frac{n}{2k_2 + k_4 + k_6 + \frac{k_4}{2} + \frac{3k_4}{2} > 1} \frac{2^{4p}}{2} \left( k_1 \cdots k_6 \right) N_{8,n} \left( \mathbb{E} \left[ \chi(d)^2 \right]^n + \left\| \tilde{X}_k \right\|_2^{2n} \right) = 2^{\frac{4p}{3}} 6^n N_{8,n} (d^n + \left\| \tilde{X}_k \right\|_2^{2n}).
\]

Similarly,

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

where

\[
N_{10,n} = \max\{ \left( \mathbb{E} \left[ \chi(d)^{4k_3+2k_5+2k_6} \right]^{1/2} N_{9,n}(2k_2+k_4+k_6) \right) \left\| \tilde{X}_k \right\|_2^{2n} : k_1, \ldots, k_6 \in \mathbb{N}, k_1 + \cdots + k_6 = n, 2k_2 + k_4 + k_6 + \frac{k_4}{2} + \frac{3k_4}{2} > 1 \} = O(d^n).
\]
Hence,

\[ E[B_2|F_{t_k}] \leq \sum_{k_1 \cdots k_6=n} 2^{2n} \left( \prod_{k_1}^{n} \left( N_{10,n} + \| \tilde{X}_k \|_2^{2n} \right) \right) \]

\leq 2^{2n} 6^n \left( N_{10,n} + \| \tilde{X}_k \|_2^{2n} \right) . \tag{28}

Therefore, combining (27) and (28),

\[ E[B] = E[E[B_1 + B_2|F_{t_k}]] \leq N_{11,n} E \left[ \| \tilde{X}_k \|_2^{2n} \right] + N_{12,n}, \tag{29} \]

where

\[ N_{11,n} = 2^{2n} 6^n (1 + N_{8,n}), \]

\[ N_{12,n} = 2^{2n} 6^n (N_{8,n} d^n + N_{10,n}) = O(d^n). \]

Thus, when \( h < (3\alpha/8N_{11,n})^2 \), by (24) and (29),

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

Hence,

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

\[ \square \]

### 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 \( O(d) \), i.e.

\[ E \left[ \| X_t \|_2^2 \right] \leq U'_t, \quad \text{for all } t \geq 0, \]

where \( U'_t = U_2 + 2(\beta + d)/\alpha \).

**Proof.** By Itô’s lemma and dissipativity,

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

Moreover, by Grönwall’s inequality,

\[ E \left[ \| X_t \|_2^2 \right] \leq e^{-\alpha t} E \left[ \| X_0 \|_2^2 \right] + 2(\beta + d)/\alpha \leq U_2 + 2(\beta + d)/\alpha = U'_t. \]

\[ \square \]

**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

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

where \( C_0 = 2\pi_{2,2}(f) (1 + U'_2) + 4d \).
Proof. By Young’s inequality,
\[
\mathbb{E} \left[ \|X_t - X_0\|_2^2 \right] = \mathbb{E} \left[ \left\| - \int_0^t \nabla f(X_s) \ ds + \sqrt{2} B_t \right\|_2^2 \right] \\
\leq 2\mathbb{E} \left[ \left\| \int_0^t \nabla f(X_s) \ ds \right\|_2^2 + 2 \|B_t\|_2^2 \right] \\
\leq 2t \int_0^t \mathbb{E} \left[ \|\nabla f(X_s)\|_2^2 \right] \ ds + 4\mathbb{E} \left[ \|B_t\|_2^2 \right] \\
\leq 2\pi_{2,2}(f) t \int_0^t \mathbb{E} \left[ 1 + \|X_s\|_2^2 \right] \ ds + 4dt \\
\leq 2\pi_{2,2}(f) (1 + U'_t) t + 4dt.
\]

\[\square\]

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 \(O(d^2)\), i.e.
\[
\mathbb{E} \left[ \|X_t\|_2^4 \right] \leq U'_t, \quad \text{for all } t \geq 0,
\]
where \(U'_t = U_t + (2\beta + 6)U'_2/\alpha\).

Proof. By Itô’s lemma, dissipativity, and Lemma 6,
\[
\frac{d}{dt} \mathbb{E} \left[ \|X_t\|_2^2 \right] = -4\mathbb{E} \left[ \|X_t\|_2^2 \langle \nabla f(X_t), X_t \rangle \right] + 12\mathbb{E} \left[ \|X_t\|_2^4 \right] \\
\leq -2\alpha \mathbb{E} \left[ \|X_t\|_2^4 \right] + (4\beta + 12) \mathbb{E} \left[ \|X_t\|_2^2 \right] \\
\leq -2\alpha \mathbb{E} \left[ \|X_t\|_2^4 \right] + (4\beta + 12) U'_t.
\]

Moreover, by Grönwall’s inequality,
\[
\mathbb{E} \left[ \|X_t\|_2^4 \right] \leq e^{-2\alpha t} \mathbb{E} \left[ \|X_0\|_2^4 \right] + (2\beta + 6)U'_2/\alpha \\
\leq U_t + (2\beta + 6)U'_2/\alpha = U'_t.
\]

\[\square\]

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
\[
\mathbb{E} \left[ \|X_t - X_0\|_2^2 \right] \leq C_1 t^2 = O(d^2 t^2), \quad \text{for all } 0 \leq t \leq 1,
\]
where \(C_1 = 8\pi_{2,2}(f) (1 + U'_t) + 32d(d + 2)\).

Proof. By Young’s inequality,
\[
\mathbb{E} \left[ \|X_t - X_0\|_2^4 \right] = \mathbb{E} \left[ \left\| - \int_0^t \nabla f(X_s) \ ds + \sqrt{2} B_t \right\|_2^4 \right] \\
= \mathbb{E} \left[ \left\| - \int_0^t \nabla f(X_s) \ ds + \sqrt{2} B_t \right\|_2^2 \right]^2 \\
\leq \mathbb{E} \left[ \left( 2 \left\| \int_0^t \nabla f(X_s) \ ds \right\|_2^2 + 4 \|B_t\|_2^2 \right)^2 \right]
\]

25
We bound the second moment of Markov chain $X$ where the remainder is

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

By Itô’s lemma,

where

\[ C_2 = 8C_1^{1/2}(1 + U_3^{1/2}) \left( \mu_2(f)^2 \pi_{3,4}(f)^{1/2} + \mu_3(f)^2 \pi_{2,4}(f)^{1/2} \right) + (8\pi_{2,4}(f)(1 + U_4) + 116d^2 + 90d + 8C_0) \mu_3(f). \]

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

By Itô’s lemma,

where the remainder is

\[ R(s) = \sum_{t=1}^s \left( -\nabla^2 f(X_u) \nabla f(X_u) + \nabla^2 f(X_0) \nabla f(X_0) \right) \, du + \int_0^s \nabla^2 f(X_u) \, dB_u \]

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,

\[ \mathbb{E} \left[ \left\| R_1(s) \right\|_2^2 \right] = \mathbb{E} \left[ \left\| \int_0^s \nabla^2 f(X_u) \nabla f(X_u) - \nabla^2 f(X_0) \nabla f(X_0) \, du \right\|_2^2 \right] \]

\[ = 2\mathbb{E} \left[ \left\| \int_0^s \nabla^2 f(X_u) \nabla f(X_u) - \nabla^2 f(X_0) \nabla f(X_0) \, du \right\|_2^2 \right] + 2\mathbb{E} \left[ \left\| \int_0^s \nabla^2 f(X_0) \nabla f(X_u) - \nabla^2 f(X_0) \nabla f(X_0) \, du \right\|_2^2 \right] \]
Thus, combining (31), (32), and (33),

\[ \leq 2s \int_0^s E \left[ \| \nabla^2 f(X_u) \nabla f(X_u) - \nabla^2 f(X_0) \nabla f(X_0) \|_2 \right] \, du \]

\[ + 2s \int_0^s E \left[ \| \nabla^2 f(X_u) \nabla f(X_u) - \nabla^2 f(X_0) \nabla f(X_0) \|_2 \right] \, du \]

\[ \leq 2s \int_0^s E \left[ \| \nabla^2 f(X_u) - \nabla^2 f(X_0) \| \| \nabla f(X_u) \|_2 \right] \, du \]

\[ + 2s \int_0^s E \left[ \| \nabla^2 f(X_0) \|_2^2 \| \nabla f(X_u) - \nabla f(X_0) \|_2 \right] \, du \]

\[ \leq 2\mu_3(f)^2 s \int_0^s E \left[ \| X_u - X_0 \|_2 \| \nabla f(X_u) \|_2 \right] \, du \]

\[ + 2\mu_2(f)^2 s \int_0^s E \left[ \| \nabla^2 f(X_0) \|_2^2 \| X_u - X_0 \|_2 \right] \, du \]

\[ \leq 2\mu_3(f)^2 \| \nabla^2 f(X_0) \|_2 \int_0^s E \left[ \| X_u - X_0 \|_2 \right] \, du \]

\[ + 2\mu_2(f)^2 \| \nabla^2 f(X_0) \|_2 \int_0^s \| X_u - X_0 \|_2 \, du \]

\[ \leq 2\mu_3(f)^2 \| \nabla^2 f(X_0) \|_2 \left( 1 + U_4 \right)^{1/2} \int_0^s \| X_u - X_0 \|_2 \, du \]

\[ + 2\mu_2(f)^2 \| \nabla^2 f(X_0) \|_2 \left( 1 + U_4 \right)^{1/2} \int_0^s \| X_u - X_0 \|_2 \, du \]

\[ \leq C_1^{1/2} \left( 1 + U_4 \right)^{1/2} \left( \mu_2(f)^2 \| \nabla^2 f(X_0) \|_2 \right) \left( 1 + \mu_2(f)^2 \| \nabla^2 f(X_0) \|_2 \right) \, s^3. \]  \( \tag{31} \)

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

\[ E \left[ \| R_2(s) \|_2^2 \right] = E \left[ \left\| \int_0^s \Delta \nabla f(X_u) \, du \right\|_2^2 \right] \]

\[ \leq 2s \int_0^s E \left[ \| \nabla \nabla f(X_u) \|_2^2 \right] \, du \]

\[ \leq \mu_3(f)^2 s^2 \leq 2. \]  \( \tag{32} \)

For \( R_3(s) \), by Itô isometry,

\[ E \left[ \| R_3(s) \|_2^2 \right] = E \left[ \left\| \int_0^s \left( \nabla^2 f(X_u) - \nabla^2 f(X_0) \right) \, dB_u \right\|_2^2 \right] \]

\[ = 2E \left[ \left\| \int_0^s \left( \nabla^2 f(X_u) - \nabla^2 f(X_0) \right) \, du \right\|_2^2 \right] \]

\[ \leq 2\mu_3(f)^2 \int_0^s E \left[ \| X_u - X_0 \|_2^2 \right] \, du \]

\[ \leq 2\mu_3(f)^2 C_0 \int_0^s \| X_u - X_0 \|_2^2 \, du \]

\[ \leq 2\mu_3(f)^2 C_0 s^2. \]  \( \tag{33} \)

Thus, combining (31), (32), and (33),

\[ E \left[ \| R(s) \|_2^2 \right] \leq 4E \left[ \| R_1(s) \|_2^2 \right] + 4E \left[ \| R_2(s) \|_2^2 \right] + 4E \left[ \| R_3(s) \|_2^2 \right] \]

\[ \leq 4C_1^{1/2} \left( 1 + U_4 \right)^{1/2} \left( \mu_2(f)^2 \| \nabla^2 f(X_0) \|_2 \right) \left( 1 + \mu_2(f)^2 \| \nabla^2 f(X_0) \|_2 \right) \, s^2 \]

\[ + 4\mu_3(f)^2 (d^2 + C_0) \, s^2. \]

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

\[ \rho_1(t) = \int_0^1 (1 - \tau) \nabla^3 f(X_0 + \tau \Delta \tilde{H}_1) [\Delta \tilde{H}_1, \Delta \tilde{H}_1] \, d\tau, \]

\[ \rho_2(t) = \int_0^1 (1 - \tau) \nabla^3 f(X_0 + \tau \Delta \tilde{H}_2) [\Delta \tilde{H}_2, \Delta \tilde{H}_2] \, d\tau, \]

\[ \Delta \tilde{H}_1 = \sqrt{2} \left( \frac{1}{t} \Psi(t) + \frac{1}{\sqrt{6}} B_t \right), \]

\[ \Delta \tilde{H}_2 = \nabla f(X_0) t + \sqrt{2} \left( \frac{1}{t} \Psi(t) - \frac{1}{\sqrt{6}} B_t \right), \]

\[ \Psi(t) = \int_0^t B_s \, ds. \]

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

\[ \mathbb{E} \left[ \left\| \Delta \tilde{H}_1 \right\|_2^4 \right] = \mathbb{E} \left[ \left\| \sqrt{2} \left( \frac{1}{t} \Psi(t) + \frac{1}{\sqrt{6}} B_t \right) \right\|_2^4 \right] \]

\[ \leq \frac{32}{t^4} \mathbb{E} \left[ \left\| \Psi(t) \right\|_2^4 \right] + \frac{8}{9} \mathbb{E} \left[ \left\| B_t \right\|_2^4 \right] \]

\[ = \frac{32}{t^4} \sum_{i=1}^d \mathbb{E} [\Psi_i(t)^4] + \frac{32}{t^4} \sum_{i,j=1}^d \sum_{i \neq j}^d \mathbb{E} [\Psi_i(t)^2] \mathbb{E} [\Psi_j(t)^2] + \frac{8}{9} d(d + 2) t^2 \]

\[ \leq \frac{32 dt^6}{3} + \frac{32 d(d - 1) t^6}{9} + \frac{8d(d - 1) t^6}{9} \]

\[ = \left( \frac{32d}{3} + \frac{32 d(d - 1)}{9} + \frac{8d(d - 1)}{9} \right) t^2 \]

\[ \leq 2d(d + 5) t^2. \]

Similarly,

\[ \mathbb{E} \left[ \left\| \Delta \tilde{H}_2 \right\|_2^4 \right] = \mathbb{E} \left[ \left\| -\nabla f(X_0) t + \sqrt{2} \left( \frac{1}{t} \Psi(t) - \frac{1}{\sqrt{6}} B_t \right) \right\|_2^4 \right] \]

\[ \leq 8 \mathbb{E} \left[ \left\| \nabla f(X_0) \right\|_2^4 \right] t^4 + 8 \mathbb{E} \left[ \left\| \sqrt{2} \left( \frac{1}{t} \Psi(t) - \frac{1}{\sqrt{6}} B_t \right) \right\|_2^4 \right] \]

\[ \leq 8 \pi_{2,4}(f) \mathbb{E} \left[ 1 + \left\| X_0 \right\|_2^4 \right] t^4 + 16d(d + 5) t^2 \]

\[ \leq 8 \pi_{2,4}(f) (1 + \mathcal{U}_4) t^4 + 16d(d + 5) t^2 \]

\[ \leq 8 (\pi_{2,4}(f) (1 + \mathcal{U}_4) + 2d(6d + 5)) t^2. \]

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

\[ \mathbb{E} \left[ \left\| \rho_1(t) \right\|_2^2 \right] = \mathbb{E} \left[ \left\| \int_0^1 (1 - \tau) \nabla^3 f(X_0 + \tau \Delta \tilde{H}_1) [\Delta \tilde{H}_1, \Delta \tilde{H}_1] \, d\tau \right\|_2^2 \right] \]

\[ \leq \int_0^1 \mathbb{E} \left[ \left\| \nabla^3 f(X_0 + \tau \Delta \tilde{H}_1) \right\|_2^4 \right] \mathbb{E} \left[ \left\| \Delta \tilde{H}_1 \right\|_2^4 \right] \, d\tau \]

\[ \leq \int_0^1 \mathbb{E} \left[ \left\| \nabla^3 f(X_0 + \tau \Delta \tilde{H}_1) \right\|_2^4 \right] \mathbb{E} \left[ \left\| \Delta \tilde{H}_1 \right\|_2^4 \right] \, d\tau \]

\[ \leq \mu_3(f)^2 \int_0^1 \mathbb{E} \left[ \left\| \Delta \tilde{H}_1 \right\|_2^4 \right] \, d\tau \]

\[ \leq 2d(6d + 5) \mu_3(f)^2 t^2. \]
Similarly,
\[
E \left[ \| \rho_2(t) \|^2 \right] \leq \mu_3(f)^2 \int_0^t E \left[ \| \Delta H_2 \|^2 \right] \, dt \\
\leq 8 (\pi_{2,4}(f) (1 + \mathcal{U}_4) + 2d(6d + 5)) \mu_3(f)^2 t^2.
\]

Plugging these results into (30),
\[
X_t - \bar{X}_t = - \int_0^t R(s) \, ds - \frac{t}{2} (\rho_1(t) + \rho_2(t)).
\]

Thus,
\[
E \left[ \| X_t - \bar{X}_t \|^2 \right] = E \left[ \left\| \int_0^t R(s) \, ds - \frac{t}{2} (\rho_1(t) + \rho_2(t)) \right\|^2 \right] \\
\leq 4t \int_0^t E \left[ \| R(s) \|^2 \right] \, ds + t^2 E \left[ \| \rho_1(t) \|^2 \right] + t^2 E \left[ \| \rho_2(t) \|^2 \right] \\
\leq 8C_1 t^{1/2} (1 + \mathcal{U}_4)^{1/2} \left( \mu_2(f)^2 \pi_{3,4}(f)^{1/2} + \mu_3(f)^2 \pi_{2,4}(f)^{1/2} \right) t^4 \\
+ (8 \pi_{2,4}(f) (1 + \mathcal{U}_4) + 116d^2 + 90d + 8C_0) \mu_3(f)^2 t^4 \\
\leq C_2 t^4.
\]

\[\square\]

B.2.2 Local Mean Deviation

**Lemma 11.** Suppose \( X_t \) and \( \bar{X}_t \) are the continuous-time process defined by (3) and Markov chain defined by (9) for time \( t \geq 0 \), respectively. If \( X_t \) and \( \bar{X}_t \) are initiated from the same iterate of the Markov chain \( X_0 \) and share the same Brownian motion, then
\[
E \left[ \left\| E \left[ X_t - \bar{X}_t | \mathcal{F}_t \right] \right\|^2 \right] \leq C_3 t^5 = \mathcal{O}(d^3 t^5), \quad \text{for all } 0 \leq t \leq 1,
\]

where
\[
C_3 = 4 \left( C_1^{1/2} (1 + \mathcal{U}_4)^{1/2} \left( \mu_2(f)^2 \pi_{3,4}(f)^{1/2} + \mu_3(f)^2 \pi_{2,4}(f)^{1/2} \right) + C_0 \mu_4(f)^2 \right) \\
+ \frac{1}{4} \mu_3(f)^2 \pi_{2,4}(f) (1 + \mathcal{U}_4) + 8 \mu_4(f)^2 \left( \pi_{2,6}(f) (1 + \mathcal{U}_6) + 73(d + 4)^3 \right).
\]

**Proof.** The proof is similar to that of Lemma 10 with slight variations on truncating the expansions. Recall since the two processes share the same Brownian motion,
\[
X_t - \bar{X}_t = - \int_0^t \nabla f(X_u) \, du - \frac{t}{2} \left( \nabla f(H_1) + \nabla f(H_2) \right).
\]

By Itô’s lemma,
\[
\nabla f(X_u) = \nabla f(X_0) - \int_0^u \left( \nabla^2 f(X_u) \nabla f(X_u) - \Delta (\nabla f)(X_u) \right) \, du + \sqrt{2} \int_0^u \nabla f(X_u) \, dB_u \\
= \nabla f(X_0) - \nabla^2 f(X_0) \nabla f(X_0) s + \sqrt{2} \nabla f(X_0) B_u + \Delta (\nabla f)(X_0)s + \bar{R}(s),
\]

where the remainder is
\[
\bar{R}(s) = \int_0^s \left( \nabla f(X_u) \nabla f(X_u) + \nabla^2 f(X_0) \nabla f(X_0) \right) \, du \\
+ \int_0^s \left( \Delta (\nabla f)(X_u) - \Delta (\nabla f)(X_0) \right) \, du
\]

29
Additionally, for 

\[ + \sqrt{2} \int_0^s (\nabla^2 f(X_u) - \nabla^2 f(X_0)) \, dB_u. \]

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

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

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

where

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

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

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

\[
\frac{t^2}{2} \mathbb{E} \left[ \frac{1}{2} \nabla f(X_0) \right] + \frac{t^3}{4} \mathbb{E} \left[ \nabla^4 f(X_0) [\nabla f(X_0), \nabla f(X_0)] \right] = \frac{t}{4} \mathbb{E} \left[ \nabla^3 f(X_0) [\Delta \tilde{H}_1, \Delta \tilde{H}_1] \right] + \frac{t}{4} \mathbb{E} \left[ \nabla^3 f(X_0) [\Delta \tilde{H}_2, \Delta \tilde{H}_2] \right]. \quad (34)
\]

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

\[
\mathbb{E} \left[ \langle \Delta \tilde{H}_1, \nabla^2 (\partial_i f)(X_0) \Delta \tilde{H}_1 \rangle \right] = \mathbb{E} \left[ \mathbb{E} \left[ \mathbb{E} \left[ (\Delta \tilde{H}_1)^T \Delta \tilde{H}_1 \nabla^2 (\partial_i f)(X_0) \right] \right] \right] = 2t \left( \frac{1}{2} + \frac{1}{\sqrt{6}} \right) \mathbb{E} [\Delta (\partial_i f)(X_0)].
\]

Similarly,

\[
\mathbb{E} \left[ \langle \Delta \tilde{H}_2, \nabla^2 (\partial_i f)(X_0) \Delta \tilde{H}_2 \rangle \right] = \mathbb{E} \left[ \mathbb{E} \left[ \mathbb{E} \left[ (\Delta \tilde{H}_2)^T \Delta \tilde{H}_2 \nabla^2 (\partial_i f)(X_0) \right] \right] \right] = 2t \left( \frac{1}{2} - \frac{1}{\sqrt{6}} \right) \mathbb{E} [\Delta (\partial_i f)(X_0)] + t^2 \mathbb{E} [\langle \nabla f(X_0), \nabla^2 (\partial_i f)(X_0) \nabla f(X_0) \rangle].
\]

Adding the previous two equations together, we obtain the desired equality (34).

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,

\[
\mathbb{E} \left[ \| \bar{R}_1(s) \|^2 \right] = \mathbb{E} \left[ \| R_1(s) \|^2 \right] \leq C_1^{1/2} (1 + \mu_4)^{1/2} \left( \mu_2(f)^2 \pi_3(f)^{1/2} + \mu_3(f)^2 \pi_2(f)^{1/2} \right) s^3.
\]

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

\[
\mathbb{E} \left[ \| \bar{R}_2(s) \|^2 \right] = \mathbb{E} \left[ \left\| \int_0^s (\bar{A}(\nabla f)(X_u) - \bar{A}(\nabla f)(X_0)) \, du \right\|^2 \right] \leq s \int_0^s \mathbb{E} \left[ \| \bar{A}(\nabla f)(X_u) - \bar{A}(\nabla f)(X_0) \|^2 \right] \, du \leq C_0 d^2 \mu_4(f)^2 s \int_0^s \| X_u - X_0 \|^2 \, du \leq C_0 d^2 \mu_4(f)^2 s \int_0^s u \, du
\]
\[ \leq C_0 d^2 \mu_4(f)^2 \frac{S^3}{2}. \]

Since \( \tilde{R}_3(s) \) is a Martingale,
\[
\left\| \mathbb{E} \left[ \int_0^t \tilde{R}(s) \, ds \mid \mathcal{F}_0 \right] \right\|_2^2 = \left\| \mathbb{E} \left[ \int_0^t \tilde{R}_1(s) \, ds \mid \mathcal{F}_0 \right] + \mathbb{E} \left[ \int_0^t \tilde{R}_2(s) \, ds \mid \mathcal{F}_0 \right] \right\|_2^2 \\
\leq 2 \left\| \mathbb{E} \left[ \int_0^t \tilde{R}_1(s) \, ds \mid \mathcal{F}_0 \right] \right\|_2^2 + 2 \left\| \mathbb{E} \left[ \int_0^t \tilde{R}_2(s) \, ds \mid \mathcal{F}_0 \right] \right\|_2^2 \\
\leq 2t \int_0^t \mathbb{E} \left[ \| \tilde{R}_1(s) \|_2^2 + \| \tilde{R}_2(s) \|_2^2 \mid \mathcal{F}_0 \right] \, ds.
\]

Therefore,
\[
\mathbb{E} \left[ \left\| \mathbb{E} \left[ \int_0^t \tilde{R}(s) \, ds \mid \mathcal{F}_0 \right] \right\|_2^2 \right] \leq 2t \int_0^t \mathbb{E} \left[ \| \tilde{R}_1(s) \|_2^2 + \| \tilde{R}_2(s) \|_2^2 \right] \, ds \\
\leq C_1 \frac{1}{2} (1 + U_1)^{1/2} \left( \mu_2(f)^2 \pi_{3,4}(f)^{1/2} + \mu_3(f)^2 \pi_{2,4}(f)^{1/2} \right) t^5 \\
+ C_0 d \mu_4(f)^2 t^5.
\]

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

To simplify notation, we define
\[
v_1 = \sqrt{2} \left( \frac{1}{2} + \frac{1}{\sqrt{6}} \right) \xi \sqrt{t}, \quad v'_1 = \sqrt{2} \left( \frac{1}{2} - \frac{1}{\sqrt{6}} \right) \xi \sqrt{t},
\]
\[
v_2 = \frac{1}{\sqrt{6}} \eta \sqrt{t} \quad \text{where} \quad \xi, \eta \text{ i.i.d. } N(0, I_d).
\]

We bound the sixth moments of \( v_1, v'_1 \) and \( v_2 \) using \( 1/2 + 1/\sqrt{3} < 1, 1/2 - 1/\sqrt{3} < 1/2 \) and the closed form moments of a chi-squared random variable with \( d \) degrees of freedom \( \chi(d)^2 \) [53],
\[
\mathbb{E} \left[ \| v_1 \|_2^6 \right] \leq 8 \mathbb{E} \left[ \| \xi \|_2^6 \right] t^3 = 8 \mathbb{E} \left[ \chi(d)^6 \right] t^3 = 8d(d+2)(d+4)t^3 < 8(d+4)^3 t^3,
\]
\[
\mathbb{E} \left[ \| v'_1 \|_2^6 \right] \leq 8 \mathbb{E} \left[ \| \xi \|_2^6 \right] t^3 = 8 \mathbb{E} \left[ \chi(d)^6 \right] t^3 = 8d(d+2)(d+4)t^3 < (d+4)^3 t^3,
\]
\[
\mathbb{E} \left[ \| v_2 \|_2^6 \right] = \frac{1}{216} \mathbb{E} \left[ \| \eta \|_2^6 \right] t^3 = \frac{1}{216} \mathbb{E} \left[ \chi(d)^6 \right] t^3 = \frac{1}{216}d(d+2)(d+4)t^3 < \frac{1}{216}(d+4)^3 t^3.
\]

Then,
\[
\mathbb{E} \left[ \| \Delta \tilde{H}_1 \|_2^6 \right] = \mathbb{E} \left[ \| v_1 + v_2 \|_2^6 \right] \leq 32 \mathbb{E} \left[ \| v_1 \|_2^6 + \| v_2 \|_2^6 \right] \leq 288(d+4)^3 t^3,
\]
\[
\mathbb{E} \left[ \| \Delta \tilde{H}_2 \|_2^6 \right] = \mathbb{E} \left[ \| -\nabla f(X_0) t + v'_1 + v_2 \|_2^6 \right] \\
\leq 32 \mathbb{E} \left[ \| \nabla f(X_0) t \|_2^6 \right] t^6 + 32 \mathbb{E} \left[ \| v'_1 + v_2 \|_2^6 \right] \\
\leq 32 \pi_{2,6}(f) \left( 1 + \mathbb{E} \left[ \| X_0 \|_2^6 \right] \right) t^6 + 1024 \mathbb{E} \left[ \| v'_1 \|_2^6 + \| v_2 \|_2^6 \right] \\
\leq 32 \pi_{2,6}(f) \left( 1 + U_6 \right) t^4 + 2048(d+4)^3 t^3 \\
\leq 32 \left( \pi_{2,6}(f) \left( 1 + U_6 \right) + 64(d+4)^3 \right) t^3.
\]

Now, we bound the second moments of \( \tilde{\rho}_1(t) \) and \( \tilde{\rho}_2(t) \) using the derived sixth-moment bounds,
\[
\mathbb{E} \left[ \| \tilde{\rho}_1(t) \|_2^2 \right] = \mathbb{E} \left[ \left\| \frac{1}{2} \int_0^1 (1-\tau)^2 \nabla^4 f(X_0 + \tau \Delta \tilde{H}_1)(\Delta \tilde{H}_1, \Delta \tilde{H}_1, \Delta \tilde{H}_1) \right\|_2^2 \right]
\]

31
Verifying the order conditions in Theorem 1 for SRK-ID requires bounding the second and fourth moments of the Markov chain.

Thus,\[
\mathbb{E} \left[ \left\| \tilde{\rho}_2(t) \right\|_2^2 \right] = \mathbb{E} \left[ \left\| \frac{1}{2} \int_0^t (1 - \tau)^2 \nabla^4 f(X_0 + \tau \Delta \tilde{H}_2) \Delta \tilde{H}_2, \Delta \tilde{H}_2 \right\|_2^2 \right] \leq \frac{1}{4} \sup_{z \in \mathbb{R}^d} \left\| \nabla^4 f(z) \right\|_{op}^2 \mathbb{E} \left[ \left\| \Delta \tilde{H}_2 \right\|_2^6 \right] \leq 72 \mu_4(f)^2 (d + 4)^3 t^3.
\]

Similarly,\[
\mathbb{E} \left[ \left\| \tilde{\rho}_2(t) \right\|_2^2 \right] = \mathbb{E} \left[ \left\| \frac{1}{2} \int_0^t (1 - \tau)^2 \nabla^4 f(X_0 + \tau \Delta \tilde{H}_2) \Delta \tilde{H}_2, \Delta \tilde{H}_2 \right\|_2^2 \right] \leq \frac{1}{4} \sup_{z \in \mathbb{R}^d} \left\| \nabla^4 f(z) \right\|_{op}^2 \mathbb{E} \left[ \left\| \Delta \tilde{H}_2 \right\|_2^6 \right] \leq 8 \mu_4(f)^2 \left( \pi_{2,6}(f) (1 + \mathcal{U}_0) + 64(d + 4)^3 \right) t^5.
\]

Thus,\[
\mathbb{E} \left[ \mathbb{E} \left[ X_t - \tilde{X}_t | \mathcal{F}_0 \right] \right] \leq 4 \left( C_1^{1/2} (1 + \mathcal{U}_1)^{1/2} \left( \mu_2(f)^2 \pi_{3,4}(f)^{1/2} + \mu_3(f)^2 \pi_{2,4}(f)^{1/2} \right) + C_0 \mu_4(f)^2 \right) t^5 \]
\[
+ \frac{1}{4} \mu_3(f)^2 \mathbb{E} \left[ \left\| \nabla f(X_0) \right\|_2^4 \right] t^6 \]
\[
+ 72 \mu_4(f)^2 (d + 4)^3 t^5 + 8 \mu_4(f)^2 \left( \pi_{2,6}(f) (1 + \mathcal{U}_0) + 64(d + 4)^3 \right) t^5 \leq 4 \left( C_1^{1/2} (1 + \mathcal{U}_1)^{1/2} \left( \mu_2(f)^2 \pi_{3,4}(f)^{1/2} + \mu_3(f)^2 \pi_{2,4}(f)^{1/2} \right) + C_0 \mu_4(f)^2 \right) t^5 \]
\[
+ \frac{1}{4} \mu_3(f)^2 \pi_{2,4}(f) (1 + \mathcal{U}_1) t^5 \]
\[
+ 8 \mu_4(f)^2 \left( \pi_{2,6}(f) (1 + \mathcal{U}_0) + 73(d + 4)^3 \right) t^5 \leq C_3 t^5.
\]

\[\square\]

B.3 Invoking Theorem 1

Now, we invoke Theorem 1 with our derived constants. We obtain that if the constant step size \( h < 1 \) and \( C_h \leq \frac{1}{2} \alpha \), then
\[
\mathbb{E} \left[ X_t - \tilde{X}_t | \mathcal{F}_0 \right] \leq \frac{1}{8 \mu_1(b)^2 + 8 \mu_4(f)^2},
\]

where \( C_h = \frac{2d}{\pi_{2,2}(f)^2} \) 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 \( \epsilon \), \( \mathcal{O}(de^{-2/3}) \) 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 \( \sigma \) 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
\[
-2 \langle b(x), x \rangle - \|\sigma(x)\|^2 \geq \alpha \|x\|^2 - \beta, \quad \text{for all } x \in \mathbb{R}^d.
\]
For general Itô diffusions, dissipativity directly follows from uniform dissipativity, where \( \beta \) is an appropriate constant of order \( O(d) \). Additionally, we assume the discretization has a constant step size \( h \) and the timestamp of the \( k \)th iterate is \( t_k \) as per the proof of Theorem 1. To simplify notation, we rewrite the update as
\[
X_{k+1} = X_k + b(X_k)h + \sigma(X_k)\xi_{k+1}h^{1/2} + \bar{Y}_{k+1}, \quad \xi_{k+1} \sim \mathcal{N}(0, I_d),
\]
where
\[
\bar{Y}^{(i)}_{k+1} = \left( \sigma_i(\tilde{H}^{(i)}_1) - \sigma_i(\tilde{H}^{(i)}_2) \right) h^{1/2}, \quad \tilde{Y}_{k+1} = \frac{1}{2} \sum_{i=1}^m \bar{Y}^{(i)}_{k+1}.
\]
Note that \( \xi_{k+1} \) and \( \tilde{Y}_{k+1} \) are not independent, since we model \( I_{(i)} = (I_{(i,1)}, \ldots, I_{(i,m)})^\top \) as \( \xi_{k+1}h^{1/2} \). Moreover, we define the following notation
\[
I_{(i,j)} = (I_{(i,1)}, \ldots, I_{(m,i)})^\top, \quad \Delta \tilde{H}^{(i)} = \sigma(X_k)I_{(i,i)}h^{-1/2}, \quad i = 1, \ldots, m.
\]
Hence, the variables \( \tilde{H}^{(i)}_1 \) and \( \tilde{H}^{(i)}_2 \) can be written as
\[
\tilde{H}^{(i)}_1 = X_k + \Delta \tilde{H}^{(i)}, \quad \tilde{H}^{(i)}_2 = X_k - \Delta \tilde{H}^{(i)}
\]
We first bound the second moments of \( \tilde{Y}_{k} \), using the following moment inequality.

**Theorem 12 ([39, 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
\[
\mathbb{E} \left[ \int_0^t \|G_s\|^p \|d_s\| \right] < \infty.
\]
Then,
\[
\mathbb{E} \left[ \left\| \int_0^t G_s \, dB_s \right\|^p \right] \leq \left( \frac{p(p - 1)}{2} \right)^{p/2} t^{1(p^{-2})/2} \mathbb{E} \left[ \int_0^t \|G_s\|^p \|d_s\| \right].
\]
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 [39].

**Corollary 13.** Let even integer \( p \geq 2 \). Then, the following relation holds
\[
\mathbb{E} \left[ \left\| \Delta \tilde{H}^{(i)} \right\|^p_2 |\mathcal{F}_{t_k}\right] \leq \left( \frac{p(p - 1)}{2} \right)^{p/2} \pi_{1,p}^1(\sigma) \left( 1 + \left\| \tilde{X}_k \right\|^{p/2}_2 \right) h^{p/2}.
\]

**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,
\[
\mathbb{E} \left[ \left\| \Delta \tilde{H}^{(i)} \right\|^p_2 |\mathcal{F}_{t_k}\right] = \mathbb{E} \left[ \left\| \sigma(X_k)I_{(i,i)} \right\|^p_2 |\mathcal{F}_{t_k}\right] h^{-p/2}
\]
\[
= \mathbb{E} \left[ \left\| \int_{t_k}^{t_{k+1}} \int_{t_k}^s \sigma(X_k) \, dB_u \, dB^{(i)}_s \right\|^p_2 |\mathcal{F}_{t_k}\right] h^{-p/2}
\]
\[
\leq \left( \frac{p(p - 1)}{2} \right)^{p/2} h^{-1} \int_{t_k}^{t_{k+1}} \mathbb{E} \left[ \left\| \int_{t_k}^s \sigma(X_k) \, dB_u \right\|^p_2 |\mathcal{F}_{t_k}\right] ds
\]
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 14 (Second Moment Bounds for \( \tilde{Y}_k \)).** The following relation holds
\[
\mathbb{E} \left[ \left\| \tilde{Y}_{k+1} \right\|_{2 \mathcal{F}_{t_k}}^2 \right] \leq 2^2 3^4 m^2 \mu_2(\sigma)^2 \pi_{1,4}^F(\sigma) \left( 1 + \left\| \tilde{X}_k \right\|_{2}^2 \right) h^3.
\]

**Proof.** By Taylor’s Theorem with the remainder in integral form,
\[
\left\| \tilde{Y}_{k+1}^{(i)} \right\|_{2} = \left\| \sigma_i(\tilde{X}_k + \Delta \tilde{H}^{(i)}) - \sigma_i(\tilde{X}_k - \Delta \tilde{H}^{(i)}) \right\|_{2} h^{1/2}
\]
\[
= \left\| \int_0^{\frac{1}{h}} \left( \nabla \sigma_i(\tilde{X}_k + \tau \Delta \tilde{H}^{(i)}) - \nabla \sigma_i(\tilde{X}_k - \tau \Delta \tilde{H}^{(i)}) \right) \Delta \tilde{H}^{(i)} \, d\tau \right\|_{2} h^{1/2}
\]
\[
\leq h^{1/2} \int_0^{\frac{1}{h}} \left\| \nabla \sigma_i(\tilde{X}_k + \tau \Delta \tilde{H}^{(i)}) - \nabla \sigma_i(\tilde{X}_k - \tau \Delta \tilde{H}^{(i)}) \right\|_{op} \left\| \Delta \tilde{H}^{(i)} \right\|_{2} \, d\tau
\]
\[
\leq \mu_2(\sigma) h^{1/2} \left\| \Delta \tilde{H}^{(i)} \right\|_{2}^2 \int_0^{1/h} 2\tau \, d\tau
\]
\[
\leq \mu_2(\sigma) h^{1/2} \left\| \Delta \tilde{H}^{(i)} \right\|_{2}^2 .
\]

(35)

By (35) and Corollary 13,
\[
\mathbb{E} \left[ \left\| \tilde{Y}_{k+1} \right\|_{2 \mathcal{F}_{t_k}}^2 \right] \leq \mu_2(\sigma)^2 \mathbb{E} \left[ \left\| \Delta \tilde{H}^{(i)} \right\|_{2}^4 \right] h \leq 6^4 \mu_2(\sigma)^2 \pi_{1,4}^F(\sigma) \left( 1 + \left\| \tilde{X}_k \right\|_{2}^2 \right) h^3.
\]

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

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

**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.
\[
\mathbb{E} \left[ \left\| \tilde{X}_k \right\|_{2}^2 \right] \leq \mathcal{V}_2 , \quad \text{for all } k \in \mathbb{N}
\]
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
\[
h < 1 \land \frac{1}{m^2} \land \frac{\sigma^2}{4M_1^2} .
\]

34
Proof. By direct computation,

$$
\|\tilde{X}_{k+1}\|_2^2 = \|\tilde{X}_k\|_2^2 + h^2 \left( \|b(\tilde{X}_k)\|_2^2 + \|\sigma(\tilde{X}_k)\tilde{\xi}_{k+1}\|_2^2 \right) + 2 \left( \tilde{X}_k, b(\tilde{X}_k) \right) h + 2 \left( \tilde{X}_k, \sigma(\tilde{X}_k)\tilde{\xi}_{k+1} \right) h^{1/2} + 2 \left( \tilde{X}_k, \tilde{Y}_{k+1} \right) h \\
+ 2 \left( b(\tilde{X}_k), \sigma(\tilde{X}_k)\tilde{\xi}_{k+1} \right) h^{3/2} + 2 \left( b(\tilde{X}_k), \tilde{Y}_{k+1} \right) h + 2 \left( \sigma(\tilde{X}_k)\tilde{\xi}_{k+1}, \tilde{Y}_{k+1} \right) h^{1/2}.
$$

By Lemma 15 and dissipativity,

$$
\mathbb{E} \left[ 2 \left( \tilde{X}_k, b(\tilde{X}_k) \right) h + \|\sigma(\tilde{X}_k)\tilde{\xi}_{k+1}\|_2^2 h | F_{s_k} \right] = 2 \left( \tilde{X}_k, b(\tilde{X}_k) \right) h + \|\sigma(\tilde{X}_k)\|_F^2 h \\
\leq - \alpha \left( \|\tilde{X}_k\|_2^2 h + \beta h.\right)
$$

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

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

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

$$
\mathbb{E} \left[ \|\tilde{Y}_{k+1}\|_2^2 | F_{s_k} \right] \leq 2^23^4 m^2 \mu_2(\sigma)^2 \pi_{1,4}(\sigma) \left( 1 + \|\tilde{X}_k\|_2^2 \right) h^3 \\
\leq 2^23^4 m \mu_2(\sigma)^2 \pi_{1,4}(\sigma) \left( 1 + \|\tilde{X}_k\|_2^2 \right) h^{3/2}.
$$

By Lemma 14,

$$
\mathbb{E} \left[ \langle \tilde{X}_k, \tilde{Y}_{k+1} \rangle | F_{s_k} \right] \leq \|\tilde{X}_k\|_2 \mathbb{E} \left[ \|\tilde{Y}_{k+1}\|_2 | F_{s_k} \right] \\
\leq \|\tilde{X}_k\|_2 \left( \|\tilde{Y}_{k+1}\|_2 \right)^{1/2} \\
\leq 2^23^2 \mu_2(\sigma)^2 \pi_{1,4}(\sigma)^{1/2} \left( 1 + \|\tilde{X}_k\|_2^2 \right) h^{3/2}.
$$

Similarly, by Lemma 14,

$$
\mathbb{E} \left[ \langle b(\tilde{X}_k), \tilde{Y}_{k+1} \rangle | F_{s_k} \right] \leq \|b(\tilde{X}_k)\|_2 \mathbb{E} \left[ \|\tilde{Y}_{k+1}\|_2 | F_{s_k} \right] \\
\leq \|b(\tilde{X}_k)\|_2 \left( \|\tilde{Y}_{k+1}\|_2 \right)^{1/2} \\
\leq 2^23^2 \mu_2(\sigma)^2 \pi_{1,4}(\sigma)^{1/2} \pi_{1,1}(b) \left( 1 + \|\tilde{X}_k\|_2^2 \right) h^{3/2}.
$$

By Lemma 14 and Lemma 15,

$$
\mathbb{E} \left[ \langle \sigma(\tilde{X}_k)\tilde{\xi}_{k+1}, \tilde{Y}_{k+1} \rangle | F_{s_k} \right] \leq \mathbb{E} \left[ \|\sigma(\tilde{X}_k)\tilde{\xi}_{k+1}\|_2 \|\tilde{Y}_{k+1}\|_2 | F_{s_k} \right] \\
\leq \mathbb{E} \left[ \|\sigma(\tilde{X}_k)\tilde{\xi}_{k+1}\|_2 | F_{s_k} \right] \mathbb{E} \left[ \|\tilde{Y}_{k+1}\|_2 | F_{s_k} \right]^{1/2} \\
\leq \mathbb{E} \left[ \|\sigma(\tilde{X}_k)\|_F | F_{s_k} \right] \mathbb{E} \left[ \|\tilde{Y}_{k+1}\|_2 | F_{s_k} \right]^{1/2} \\
\leq 2^23^2 \mu_2(\sigma)^2 \pi_{1,4}(\sigma)^{1/2} \pi_{1,2}(\sigma)^{1/2} \left( 1 + \|\tilde{X}_k\|_2^2 \right) h^{3/2}.
$$
Putting things together, for \( h < 1 \land \alpha^2/(4M_1^2) \),

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

where

\[
M_1 = \pi_{1,2}(b) + 2^3 \mathbb{E} \mu_2(\sigma) \pi_{1,4}^F(\sigma)^{1/2} \left( 1 + \mu_2(\sigma) \pi_{1,4}^F(\sigma)^{1/2} + \pi_{1,1}(b) + \pi_{1,2}^F(\sigma)^{1/2} \right).
\]

Unrolling the recursion gives the following for \( h < 1 \land 1/m^2 \)

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

where

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

### C.1.2 2nth Moment Bound

Before bounding the 2nth moment, we first generalize Lemma 14 to arbitrary even moments.

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

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

**Proof.** For \( i \in \{1, 2, \ldots, m\} \), by (35),

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

Hence, by Corollary 13,

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

The remaining follows easily from Lemma 31. \( \square \)

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

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

where

\[
\mathcal{V}_{2n} = E \left[ \left\| \bar{X}_0 \right\|_2^{2n} \right] + \frac{2}{n\alpha} \left( \beta \mathcal{V}_{2(n-1)} + 2^{3n-1} \alpha^2 n \mathbb{E} \pi_{1,2n}^F(b) \pi_{1,8n}^F(\sigma)^{1/2} \mu_2(\sigma)^{2n} \right).
\]

If the step size

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

36
Proof. Our proof is by induction. The base case is given in Lemma 16. For the inductive case, we prove that the 2nth moment is uniformly bounded by a constant, assuming the 2(n-1)th moment is uniformly bounded by a constant.

By the multinomial theorem,

\[
E \left[ \left\| \tilde{X}_{k+1} \right\|^2_{2^n} | \mathcal{F}_{t_k} \right] = E \left[ \left( \left\| \tilde{X}_k \right\|^2_{2^n} + \left\| b(\tilde{X}_k) \right\|^2_{2^n} h^2 + \left\| \sigma(\tilde{X}_k)\xi_{k+1} \right\|^2_{2^n} h + \left\| \tilde{Y}_{k+1} \right\|^2_{2^n} \right. \right.
\]
\[
+ 2 \left( \tilde{X}_k, b(\tilde{X}_k) \right) h + 2 \left( \tilde{X}_k, \sigma(\tilde{X}_k)\xi_{k+1} \right) h^{1/2} + 2 \left( \tilde{X}_k, \tilde{Y}_{k+1} \right) h
\]
\[
+ 2 \left( b(\tilde{X}_k), \sigma(\tilde{X}_k)\xi_{k+1} \right) h^{3/2} + 2 \left( b(\tilde{X}_k), \tilde{Y}_{k+1} \right) h
\]
\[
+ 2 \left( \sigma(\tilde{X}_k)\xi_{k+1}, \tilde{Y}_{k+1} \right) h^{1/2} \left. \right] | \mathcal{F}_{t_k} \right]
\]
\[
= \left\| \tilde{X}_{k+1} \right\|^2_{2^n} + E [A|\mathcal{F}_{t_k}] h + E [B|\mathcal{F}_{t_k}] h^{3/2},
\]
where by the Cauchy–Schwarz inequality,

\[
A = n \left\| \tilde{X}_k \right\|^2_{2^{(n-1)}} \left( \left\| \sigma(\tilde{X}_k)\xi_{k+1} \right\|^2_{2^{(n-1)}} + 2n(n-1) \right) \left\| \tilde{X}_k \right\|^2_{2^{(n-2)}} \left\langle \tilde{X}_k, \sigma(\tilde{X}_k)\xi_{k+1} \right\rangle,
\]

\[
B \leq \sum_{(k_1, \ldots, k_{10}) \in J} 2^n \left( n \right) \left\| \tilde{X}_k \right\|_{2^{(n-1)}}^{p_1} \left\| b(\tilde{X}_k) \right\|_{2^{(n-1)}}^{p_2} \left\| \sigma(\tilde{X}_k)\xi_{k+1} \right\|_{2^{(n-1)}}^{p_3} \left\| \tilde{Z}_{k+1} \right\|_{2^{(n-1)}}^{p_4},
\]
the indicator set

\[
J = \left\{ (k_1, \ldots, k_{10}) \in \mathbb{N}^{10} : k_1 + \cdots + k_{10} = n, \right. \]
\[
2k_2 + 3k_3 + 3k_4 + k_5 + \frac{k_6}{2} + \frac{3k_7}{2} + \frac{3k_8}{2} + \frac{5k_9}{2} + 2k_{10} > 1 \right\},
\]
and with slight abuse of notation, we hide the explicit dependence on \(k_1, \ldots, k_{10}\) for the exponents

\[
p_1 = 2k_1 + k_5 + k_6 + k_7,
\]
\[
p_2 = 2k_2 + k_5 + k_8 + k_9,
\]
\[
p_3 = 2k_3 + k_6 + k_8 + k_{10},
\]
\[
p_4 = 2k_4 + k_7 + k_9 + k_{10}.
\]

By dissipativity,

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

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

\[
\left\langle \tilde{X}_k, \tilde{Z}_{k+1} \right\rangle = O(mh^{3/2}).
\]
Using this information, Lemma 17, Lemma 15, the Cauchy–Schwarz inequality, and \(p_3 + p_4 \leq 2n\),

\[
E [B|\mathcal{F}_{t_k}] = \frac{1}{2^n} \left\| \tilde{X}_k \right\|_{2^n}^{2n} + n\beta \left\| \tilde{X}_k \right\|_{2^{(n-1)}}^{2(n-1)}.
\]
\[
\leq \sum_{(k_1, \ldots, k_{10}) \in J} 2^n \|X_{k_1}\|_2^{p_1} E \left[ \|\sigma(X_{k_1})\|_2^{p_3} \left\| \tilde{Z}_{k+1} m^{-1} \right\|_2^{2p_4} F_{k_1} \right] \mbox{m}
\]

\[
\leq \sum_{(k_1, \ldots, k_{10}) \in J} 2^n \|X_{k_1}\|_2^{p_1} E \left[ \|\sigma(X_{k_1})\|_2^{2p_3} F_{k_1} \right]^{1/2} \left\| \tilde{Z}_{k+1} m^{-1} \right\|_2^{2p_4} F_{k_1}^{1/2} \mbox{m}
\]

\[
\leq \sum_{(k_1, \ldots, k_{10}) \in J} 2^n \|X_{k_1}\|_2^{p_1} \pi_{1,p_2} \left( 1 + \|\tilde{X}_k\|_2^{p_2} \right) (2p_3 - 1)! \pi_{1,p_3} \left( 1 + \|\tilde{X}_k\|_2^{p_3} \right) \times \mu_2(\sigma)^{p_4} (8\sigma^2)^{p_4} \pi_{1,4p_4} (\sigma)^{1/2} \left( 1 + \|\tilde{X}_k\|_2^{p_4} \right) m
\]

\[
\leq 2^n \left( 1 + \|\tilde{X}_k\|_2^{p_2} \right) \sum_{(k_1, \ldots, k_{10}) \in J} \pi_{1,p_2} (2p_3 - 1)! \pi_{1,p_3} \mu_2(\sigma)^{p_4} (8\sigma^2)^{p_4} \pi_{1,4p_4} (\sigma)^{1/2} \left( k_1 \ldots k_{10} \right) m
\]

\[
\leq 2^n \left( 1 + \|\tilde{X}_k\|_2^{p_2} \right) \pi_{1,2n} (2p_2 - 1)! \pi_{1,4n} (2n)^{1/2} \sum_{k_1, \ldots, k_{10} \in \mathbb{N}} \pi_{1,8n} (\sigma)^{1/2} \mu_2(\sigma)^{2n} m \left( 1 + \|\tilde{X}_k\|_2^{2n} \right).
\]

By the inductive hypothesis, (36) and (37), and \( h < 1 \land n^2 \alpha^2 / (4M_{3,n}^2) \), we obtain the recursion

\[
E \left[ \left\| X_{k+1} \right\|_2^{2n} \left| F_{k_1} \right. \right] \leq \left( 1 - \alpha h + M_{3,n} h^{3/2} \right) E \left[ \left\| X_k \right\|_2^{2n} \right] + n \beta h E \left[ \left\| X_k \right\|_2^{2(n-1)} \right] + M_5 h^{3/2}
\]

\[
\leq \left( 1 - \alpha h + M_{3,n} h^{3/2} \right) E \left[ \left\| X_k \right\|_2^{2n} \right] + n \beta V_2(n-1) h + M_{3,n} h^{3/2}
\]

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

For \( h < 1 \land 1/m^2 \), by unrolling the recursion, we obtain

\[
E \left[ \left\| X_k \right\|_2^{2n} \right] \leq E \left[ \left\| X_0 \right\|_2^{2n} \right] + \frac{2}{n \alpha} \left( n \beta V_2(n-1) + M_{3,n} h^{1/2} \right) \leq V_{2n}, \quad \text{for all } k \in \mathbb{N},
\]

where

\[
V_{2n} = E \left[ \left\| X_0 \right\|_2^{2n} \right] + \frac{2}{n \alpha} \left( \beta V_2(n-1) + 2^{23n-1} 10^n 8^n \pi_{1,2n} (2n)^{1/2} \mu_2(\sigma)^{2n} \right).
\]

\[\square\]

**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.

\[
E \left[ \|X_t\|_2^2 \right] \leq V_2, \quad \text{for all } t \geq 0.
\]

where \( V_2 = V_2 + \beta / \alpha \).

**Proof.** By Itô’s lemma and dissipativity,

\[
\frac{d}{dt} E \left[ \|X_t\|_2^2 \right] = E \left[ 2 \langle X_t, b(X_t) \rangle + \|\sigma(X_t)\|_F^2 \right] \leq -\alpha E \left[ \|X_t\|_2^2 \right] + \beta.
\]

Moreover, by Grönwall’s inequality,

\[
E \left[ \|X_t\|_2^2 \right] \leq e^{-\alpha t} E \left[ \|X_0\|_2^2 \right] + \beta / \alpha \leq V_2 + \beta / \alpha = V_2.
\]

\[\square\]
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
\[ E \left[ \| X_t - X_0 \|_2^2 \right] \leq D_0 t, \quad \text{for all } 0 \leq t \leq 1, \]
where $D_0 = 2 \left( \pi_{1,2}(b) + \pi_{1,2}^F(\sigma) \right) (1 + V_2')$.

Proof. By Itô isometry,
\[
E \left[ \| X_t - X_0 \|_2^2 \right] = E \left[ \left\| \int_0^t b(X_s) \, ds + \int_0^t \sigma(X_s) \, dB_s \right\|^2_2 \right]
\leq 2E \left[ \left\| \int_0^t b(X_s) \, ds \right\|_2^2 + \left\| \int_0^t \sigma(X_s) \, dB_s \right\|_2^2 \right]
\leq 2t \int_0^t E \left[ \| b(X_s) \|_2^2 \right] \, ds + 2 \int_0^t E \left[ \| \sigma(X_s) \|_F^2 \right] \, ds
\leq 2\pi_{1,2}(b) t \int_0^t E \left[ 1 + \| X_s \|_2^2 \right] \, ds + 2\pi_{1,2}^F(\sigma) \int_0^t E \left[ 1 + \| X_s \|_2^2 \right] \, ds
\leq 2 \left( \pi_{1,2}(b) + \pi_{1,2}^F(\sigma) \right) (1 + V_2') t.
\]

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

Lemma 21 ([22, adapted from Lemma A.1]). Assuming $\{X_t\}_{t \geq 0}$ is the solution to the SDE (1), under the condition that the drift coefficient $b$ and diffusion coefficient $\sigma$ are Lipschitz. If $\sigma$ satisfies the following sublinear growth condition
\[
\| \sigma(x) \|_F^2 \leq \pi_{1,2}^F(\sigma) \left( 1 + \| x \|^{(l/2)} \right), \quad \text{for all } x \in \mathbb{R}^d, l = 1, 2, \ldots,
\]
and the diffusion is dissipative, then for $n \geq 2$, we have the following relation
\[
A \| x \|_2^n \leq -\frac{\alpha n}{4} \| x \|_2^n + \beta_n,
\]
where the (infinitesimal) generator $A$ is defined as
\[
Af(x) = \lim_{t \to 0} \frac{E \left[ f(X_t) | X_0 = x \right] - f(x)}{t},
\]
and the constant $\beta_n = \mathcal{O}(d^{2})$.

Proof. By definition of the generator and dissipativity,
\[
A \| x \|_2^n \leq n \| x \|_2^{n-2} \langle x, b(x) \rangle + \frac{n}{2} \| x \|_2^{n-2} \| \sigma(x) \|_F^2 + \frac{n(n-2)}{2} \| x \|_2^{n-4} \langle \text{vec}(x^\top), \text{vec}(\sigma \sigma^\top(x)) \rangle
\leq -\frac{\alpha n}{2} \| x \|_2^n + \frac{\beta n}{2} \| x \|_2^{n-2} + \frac{n(n-2)}{2} \| x \|_2^{n-2} \pi_{1,2}^F(\sigma) (1 + \| x \|_2)
= -\frac{\alpha n}{2} \| x \|_2^n + \frac{n(n-2)}{2} \pi_{1,2}^F(\sigma) \| x \|_2^{n-1} + \left( \frac{\beta n}{2} + \frac{n(n-2)}{2} \pi_{1,2}^F(\sigma) \right) \| x \|_2^{n-2}.
\]
By Young's inequality,
\[
\frac{n(n-2)}{2} \pi_{1,2}^F(\sigma) \| x \|_2^{n-1} = \frac{n(n-2)}{2} \pi_{1,2}^F(\sigma) \left( \frac{8}{\alpha n} \right)^{\frac{n-1}{n}} \| x \|_2^{n-1} \left( \frac{8}{\alpha n} \right)^{\frac{n-1}{n}}
\leq \frac{\alpha n}{8} \| x \|_2^{n-1} \left( \frac{8}{\alpha n} \right)^{\frac{n-1}{n}} + \frac{n-1}{\alpha n} \| x \|_2^n
= \frac{(n-2)^n}{2^{2n-3} \alpha^{n-1}} \pi_{1,2}^F(\sigma)^n + \frac{\alpha(n-1)}{8} \| x \|_2^n.
\]

39
We define the following shorthand notation
\[ D \] where \[ X \]

Lemma 23

Similarly,
\[
\left( \frac{\beta n}{2} + \frac{n(n-2)}{2} \pi^{F}_{1,2}(\sigma) \right) \| x \|^2 = \left( \frac{\beta n}{2} + \frac{n(n-2)}{2} \pi^{F}_{1,2}(\sigma) \right) \left( \frac{8}{\alpha n} \right)^{n/2} \| x \|^2 \beta n \| x \|^2 + \alpha(n-2) \| x \|^2.
\]

Putting things together, we obtain the following bound
\[
\mathcal{A} \| x \|^2 \leq - \frac{\alpha n}{2} \| x \|^2 + \frac{\alpha(n-1)}{8} \| x \|^2 + \frac{\alpha(n-2)}{8} \| x \|^2 + \beta_n^{(1)} + \beta_n^{(2)}
\]

where \[ \beta_n = \beta_n^{(1)} + \beta_n^{(2)} = \mathcal{O}(d^2). \]

\[ \| X \|^2 \leq \mathcal{V}_4, \quad \text{for all } t \geq 0,
\]

where \[ \mathcal{V}_4 = \mathcal{V}_4 + \beta_4/\alpha. \]

\[ \mathcal{E} \left[ \| X \|^4 \right] \leq \mathcal{V}_4 \]

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.

\[ \mathcal{E} \left[ \| X \|^4 \right] \leq \mathcal{V}_4, \quad \text{for all } t \geq 0,
\]

where \( \mathcal{V}_4 = \mathcal{V}_4 + \beta_4/\alpha. \)

\[ \mathcal{E} \left[ ||X||^4 \right] \leq \mathcal{V}_4 \]

Proof.

By Dynkin's formula [45] applied to the function \( (t, x) \mapsto e^{\alpha t} \| x \|^4 \) and Lemma 21,

\[
e^{\alpha t} \mathcal{E} \left[ ||X||^4 | \mathcal{F}_0 \right] = \| X_0 \|^2 + \int_0^t \mathcal{E} \left[ ||X||^4 + e^{\alpha s} A \| X \|^2 | \mathcal{F}_0 \right] ds
\]

Hence,
\[ \mathcal{E} \left[ ||X||^2 \right] = \mathcal{E} \left[ \| X_0 \|^2 | \mathcal{F}_0 \right] \leq \mathcal{E} \left[ ||X||^2 + \beta_4/\alpha \right] \leq \mathcal{V}_4 + \beta_4/\alpha = \mathcal{V}_4'.
\]

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

\[ \mathcal{E} \left[ ||X_t - X_0||^4 \right] \leq D_1 t^2, \quad \text{for all } 0 \leq t \leq 1,
\]

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

Proof.

By Theorem 12,
\[
\mathcal{E} \left[ ||X_t - X_0||^4 \right] = \mathcal{E} \left[ \| \int_0^t b(X_s) \|_2 + \int_0^t \sigma(X_s) dB_s \|_2^4 \right]
\]

\[ \leq 8 \mathcal{E} \left[ \| \int_0^t b(X_s) \|_2 + \int_0^t \sigma(X_s) dB_s \|_2^4 \right]
\]

\[ \leq 8 t^3 \int_0^t \mathcal{E} \left[ ||b(X_s)||_2^2 \right] ds + 288 t \mathcal{E} \left[ \| \sigma(X_s) \|^2 \right] ds
\]

\[ \leq 8 \left( \pi_{1,4}(b) + 36 \pi_{1,4}(\sigma) \right) (1 + \mathcal{V}_4') t^2.
\]
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

$$
\mathbb{E} \left[ \| X_t - \tilde{X}_t \|^2 \right] \leq D_4 t^3, \quad \text{for all } 0 \leq t \leq 1,
$$

where

$$
D_4 = \left( 16 D_0 \mu_1(b)^2 + \frac{16}{3} \mu_2(\sigma)^2 \pi_{1,4}^{(1/2)} D_1^{1/2} (1 + \mathbb{V}_2^{1/2}) m^2 + \frac{16}{3} \mu_1(\sigma)^4 m^2 D_0 \right.

+ 16 \mu_1(\sigma)^2 \pi_{1,2}(b)^2 (1 + \mathbb{V}_2') m + 4 m^3 \mu_2(\sigma)^2 \pi_{1,4}^F(\sigma) (1 + \mathbb{V}_2')

+ 2^3 4^3 m^2 \mu_2(\sigma)^2 \pi_{1,4}^F(\sigma) (1 + \mathbb{V}_2').
$$

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

$$
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^s \int_0^t \Lambda_i(\sigma_i)(X_u) \, dB^{(i)}_u \, dB^{(i)}_s

+ \sum_{i=1}^m \int_0^t \int_0^s L(\sigma_i)(X_u) \, du \, dB^{(i)}_s

= \int_0^t b(X_s) \, ds + \sigma(X_0) B_t + \sum_{i=1}^m \sum_{l=1}^m \int_0^s \nabla \sigma_i(X_u) \sigma_i(X_u) \, dB^{(i)}_u \, dB^{(i)}_s + S(t),
$$

where

$$
S(t) = \sum_{i=1}^m \int_0^t \int_0^s \nabla \sigma_i(X_u) b(X_u) \, du \, dB^{(i)}_s + \frac{1}{2} \sum_{i=1}^m \sum_{l=1}^m \int_0^s \nabla \sigma_i(X_u) \sigma_i(X_u) \, dB^{(i)}_u \, dB^{(i)}_s.
$$

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

$$
\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),
$$

where

$$
\phi_1^{(i)}(t) = \int_0^1 (1 - \tau) \nabla^2 \sigma_i(X_0 + \tau \Delta \tilde{H}^{(i)}) \Delta \tilde{H}^{(i)} \, 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)} \, d\tau

\Delta \tilde{H}^{(i)} = \sum_{l=1}^m \sigma_i(X_0) \frac{I_{(t,i)}}{\sqrt{t}}.
$$

Hence,

$$
X_t - \tilde{X}_t = \int_0^t (b(X_s) - b(X_0)) \, ds

+ \sum_{i=1}^m \sum_{l=1}^m \int_0^s (\nabla \sigma_i(X_u) \sigma_i(X_u) - \nabla \sigma_i(X_0) \sigma_i(X_0)) \, dB^{(i)}_u \, dB^{(i)}_s

+ S(t) - \frac{1}{2} \sum_{i=1}^m \left( \phi_1^{(i)}(t) - \phi_2^{(i)}(t) \right) \sqrt{t}.
$$
Since $b$ is $\mu_1(b)$-Lipschitz,
\[
\mathbb{E}
\left[
\left|
\left|
\int_0^t (b(X_s) - b(X_0)) \, ds
\right|
\right|_2^2
\right] \leq \mu_1(b)^2 t \int_0^t \mathbb{E}
\left[
\left|
\left|
 X_s - X_0
\right|
\right|_2^2
\right] \, ds
\leq \mu_1(b)^2 t \int_0^t D_0 \, ds
\leq \frac{1}{2} D_0 \mu_1(b)^2 t^3.
\]

We define the following,
\[
A(t) = A_1(t) + A_2(t) = \sum_{i=1}^m \sum_{l=1}^m \int_0^t \int_0^s \left(\nabla \sigma_i(X_u) \sigma_l(X_u) - \nabla \sigma_i(X_0) \sigma_l(X_0)\right) \, dB_u^{(i)} \, dB_s^{(i)},
\]
where
\[
A_1(t) = \sum_{i=1}^m \sum_{l=1}^m \int_0^t \int_0^s \left(\nabla \sigma_i(X_u) \sigma_l(X_u) - \nabla \sigma_i(X_0) \sigma_l(X_0)\right) \, dB_u^{(i)} \, dB_s^{(i)},
\]
\[
A_2(t) = \sum_{i=1}^m \sum_{l=1}^m \int_0^t \int_0^s \left(\nabla \sigma_i(X_0) \sigma_l(X_u) - \nabla \sigma_i(X_0) \sigma_l(X_0)\right) \, dB_u^{(i)} \, dB_s^{(i)}.
\]

By Itô isometry and the Cauchy-Schwarz inequality,
\[
\mathbb{E}
\left[
\left|
\left|
 A_1(t)
\right|
\right|_2^2
\right] = \sum_{i=1}^m \sum_{l=1}^m \int_0^t \int_0^s \mathbb{E}
\left[
\left|
\nabla \sigma_i(X_u) \sigma_l(X_u) - \nabla \sigma_i(X_0) \sigma_l(X_0)\right|
\right]_2^2 \, du \, ds
\leq \sum_{i=1}^m \sum_{l=1}^m \int_0^t \int_0^s \mathbb{E}
\left[
\nabla \sigma_i(X_u) - \nabla \sigma_i(X_0)\right]_2^2 \, du \, ds
\leq \sum_{i=1}^m \sum_{l=1}^m \int_0^t \int_0^s \mathbb{E}
\left[
\nabla \sigma_i(X_u) - \nabla \sigma_i(X_0)\right]_2^2 \, du \, ds
\leq \mu_2(\sigma)^2 \pi_{1,4}(\sigma)^{1/2} m^2 \int_0^t \int_0^s \mathbb{E}
\left[
\left|
 X_u - X_0\right|_2^4 \right]^{1/2} \, du \, ds
\leq \mu_2(\sigma)^2 \pi_{1,4}(\sigma)^{1/2} D_1^{1/2} \left(1 + V_2^1/2\right) m^2 t^3.
\]

Similarly,
\[
\mathbb{E}
\left[
\left|
\left|
 A_2(t)
\right|
\right|_2^2
\right] = \sum_{i=1}^m \sum_{l=1}^m \int_0^t \int_0^s \mathbb{E}
\left[
\left|
\nabla \sigma_i(X_0) \sigma_l(X_u) - \nabla \sigma_i(X_0) \sigma_l(X_0)\right|
\right]_2^2 \, du \, ds
\leq \mu_1(\sigma)^2 \sum_{i=1}^m \sum_{l=1}^m \int_0^t \int_0^s \mathbb{E}
\left[
\left|
\nabla \sigma_i(X_u) - \nabla \sigma_l(X_u)\right|_2^2
\right] \, du \, ds
\leq \mu_1(\sigma)^4 m^2 \int_0^t \int_0^s \mathbb{E}
\left[
\left|
 X_u - X_0\right|_2^2
\right] \, du \, ds
\leq \frac{1}{6} \mu_1(\sigma)^4 m^2 D_0 t^3.
\]

By Itô isometry,
\[
\mathbb{E}
\left[
\left|
\left|
 S_1(t)
\right|
\right|_2^2
\right] = \sum_{i=1}^m \int_0^t s \, \int_0^s \mathbb{E}
\left[
\left|
\nabla \sigma_i(X_u) b(X_u)\right|
\right|_2^2 \, du \, ds
\]
Similarly, Combining (38), (39), (40), (41), and (44),

\begin{align*}
\text{Hence, by (42) and (43),}
\end{align*}

\begin{align*}
- \| X_1 \|_2 = 1 & \leq \mu_1(\sigma)^2 \pi_{1,2}^2 (b)^2 \sum_{i=1}^m t \int_0^t s \int_0^s \mathbb{E} \left[ 1 + \| X_u \|_2^2 \right] du ds \\
& = \frac{1}{2} \mu_1(\sigma)^2 \pi_{1,2}^2 (1 + \mathcal{V}_2') t^3. \\
\text{(40)}
\end{align*}

Similarly,

\begin{align*}
\mathbb{E} \left[ \| S_2(t) \|_2^2 \right] = & \frac{1}{4} \sum_{i=1}^m t \int_0^t s \int_0^s \mathbb{E} \left[ \left\| \int_0^t s \int_0^s \nabla^2 \sigma_i(X_u)[\sigma_i(X_u), \sigma_i(X_u)] du \right\|_2^2 \right] ds \\
& \leq \frac{1}{4} m \sum_{i=1}^m \sum_{i=1}^m t \int_0^t s \int_0^s \mathbb{E} \left[ \left\| \nabla^2 \sigma_i(X_u)[\sigma_i(X_u), \sigma_i(X_u)] \right\|_2^2 \right] du ds \\
& \leq \frac{1}{4} \sigma_2(\sigma)^2 \pi_{1,4}(\sigma) m^3 t \int_0^t s \int_0^s \mathbb{E} \left[ 1 + \| X_u \|_2^2 \right] du ds \\
& \leq \frac{1}{8} \sigma_2(\sigma)^2 \pi_{1,4}(\sigma) m^3 (1 + \mathcal{V}_2') t^3. \\
\text{(41)}
\end{align*}

By Corollary 13,

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

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

\begin{align*}
\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 i \Delta \tilde{H}^{(i)}[\Delta \tilde{H}^{(i)}, \Delta \tilde{H}^{(i)}]) du \right\|_2^2 \right] \\
& \leq \mathbb{E} \left[ \nabla^2 \sigma_i(X_0 + \tau \Delta \tilde{H}^{(i)}) \right\|_2^2 \left\| \Delta \tilde{H}^{(i)} \right\| \right]_2^2 \\
& \leq 6^4 \mu_2(\sigma)^2 \pi_1^4(\sigma) (1 + \mathcal{V}_2) t^2. \\
\text{(42)}
\end{align*}

Similarly,

\begin{align*}
\mathbb{E} \left[ \left\| \phi_2^{(i)}(t) \right\|_2^2 \right] \leq & 6^4 \mu_2(\sigma)^2 \pi_{1,4}^4(\sigma) (1 + \mathcal{V}_2) t^2. \\
\text{(43)}
\end{align*}

Hence, by (42) and (43),

\begin{align*}
\mathbb{E} \left[ \left\| \frac{1}{2} \sum_{i=1}^m (\phi_1^{(i)}(t) - \phi_2^{(i)}(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 2^4 3^4 m^2 \mu_2(\sigma)^2 \pi_{1,4}^4(\sigma) (1 + \mathcal{V}_2) t^3. \\
\text{(44)}
\end{align*}

Combining (38), (39), (40), (41), and (44),

\begin{align*}
\mathbb{E} \left[ \left\| X_t - \tilde{X}_t \right\|_2^2 \right] \leq & 32 \mathbb{E} \left[ \left\| \int_0^t (b(X_s) - b(X_0)) ds \right\|_2^2 \right] \\
& + 32 \mathbb{E} \left[ \| A_1(t) \|_2^2 + \| A_2(t) \|_2^2 \right]
\end{align*}
where 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

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

where

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

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

$$X_t - X_0 = b(X_0)t + \sum_{i=1}^m \int_0^t \sigma_i(X_u) \, dB_u^{(i)} + \sum_{i=1}^m \int_0^t \int_0^s A_i(b(X_u)) \, dB_u^{(i)} ds + \int_0^t \int_0^s L(b(X_u)) \, du \, ds$$

$$= b(X_0)t + \sum_{i=1}^m \int_0^t \sigma_i(X_u) \, dB_u^{(i)} + \sum_{i=1}^m \int_0^t \int_0^s \nabla b(X_u) \sigma_i(X_u) \, dB_u^{(i)} ds + \tilde{S}(t),$$

where

$$\tilde{S}(t) = \int_0^t \int_0^s \nabla b(X_u) \sigma_i(X_u) \, du \, ds + \frac{1}{2} \sum_{i=1}^m \int_0^t \int_0^s \nabla^2 b(X_u)[\sigma_i(X_u), \sigma_i(X_u)] \, du \, ds.$$

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

$$E \left[ \left\| \tilde{S}_1(t) \right\|_2^2 \right] = E \left[ \left\| \int_0^t \int_0^s \nabla b(X_u) b(X_u) \, du \, ds \right\|_2^2 \right]$$

$$\leq t \int_0^t \int_0^s E \left[ \left\| \nabla b(X_u) b(X_u) \right\|_2^2 \right] \, du \, ds$$

$$\leq t \int_0^t \int_0^s E \left[ \left\| \nabla b(X_u) \right\|_{op}^2 \left\| b(X_u) \right\|_2^2 \right] \, du \, ds$$

$$\leq \mu_1(b)^2 \pi_{1,2}(b) t \int_0^t \int_0^s E \left[ 1 + \left\| b(X_u) \right\|_2^2 \right] \, du \, ds$$

$$\leq \frac{1}{3} \mu_1(b)^2 \pi_{1,2}(b) (1 + \mathcal{V}_2') t^4.$$

\[\square\]

### 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

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

where

$$D_4 = \left( \frac{4}{3} \mu_1(b)^2 \pi_{1,2}(b) (1 + \mathcal{V}_2') + \frac{1}{3} \mu_2(b)^2 \pi_{1,4}(\sigma) (1 + \mathcal{V}_2') m^2 + 2^4 3^5 \mu_3(\sigma)^2 \pi_{1,6}(\sigma) \left( 1 + \mathcal{V}_4^{3/4} \right) \right).$$
Similarly,
\[
\mathbb{E} \left[ \left\| S_2(t) \right\|_2^2 \right] = \mathbb{E} \left[ \left\| \frac{1}{2} \sum_{i=1}^{m} \int_0^t \int_0^s \nabla^2 b(X_u) \sigma_i(X_u), \sigma_i(X_u) \right\|_2^2 \right] \\
\leq \frac{m}{4} \sum_{i=1}^{m} \int_0^t \int_0^s \mathbb{E} \left[ \left\| \nabla^2 b(X_u) \sigma_i(X_u), \sigma_i(X_u) \right\|_2^2 \right] \, ds \\
\leq \frac{m}{4} \sum_{i=1}^{m} \int_0^t \int_0^s \mathbb{E} \left[ \left\| \nabla^2 b(X_u) \right\|_{\text{op}}^2 \left\| \sigma_i(X_u) \right\|_2 \right] \, ds \\
\leq \frac{m^2}{4} \mu_2(b)^2 \sum_{i=1}^{m} \int_0^t \int_0^s \mathbb{E} \left[ \left\| \sigma_i(X_u) \right\|_2 \right] \, ds \\
\leq \frac{m^2}{4} \mu_2(b)^2 \pi_{1,4}(\sigma) t \int_0^t \int_0^s \mathbb{E} \left[ 1 + \left\| X_u \right\|_2^2 \right] \, ds \\
\leq \frac{1}{12} \mu_2(b)^2 \pi_{1,4}(\sigma) \left( 1 + \mathcal{V}_2 \right) m^2 t^4. \tag{46}
\]

By Corollary 13,
\[
\mathbb{E} \left[ \left\| \Delta \tilde{H}^{(i)} \right\|_2^6 \right] = \mathbb{E} \left[ \left\| \Delta \tilde{H}^{(i)} \right\|_2^6 \mid \mathcal{F}_k \right] \\
\leq 3^{6} 5^{6} \pi_{1,6}(\sigma) \mathbb{E} \left[ \left\| \tilde{X}_k \right\|_2^3 \right] t^3 \\
\leq 3^{6} 5^{6} \pi_{1,6}(\sigma) \left( 1 + \mathcal{V}_4^{3/4} \right) t^3.
\]

Now, we bound the second moment of the difference between \( \phi_1^{(i)}(t) \) and \( \phi_2^{(i)}(t) \),
\[
\mathbb{E} \left[ \left\| \phi_1^{(i)}(t) - \phi_2^{(i)}(t) \right\|_2^2 \right] \leq \mathbb{E} \left[ \int_0^1 \left\| \nabla^2 \sigma_i(X_0 + \tau \Delta \tilde{H}^{(i)}) - \nabla^2 \sigma_i(X_0 - \tau \Delta \tilde{H}^{(i)}) \right\|_2 \left\| \Delta \tilde{H}^{(i)} \right\|_2^4 \, d\tau \right] \\
\leq \mu_3(\sigma)^2 \int_0^1 \mathbb{E} \left[ \left\| 2\tau \Delta \tilde{H}^{(i)} \right\|_2^2 \left\| \Delta \tilde{H}^{(i)} \right\|_2^4 \right] \, d\tau \\
\leq \frac{4}{3} \mu_3(\sigma)^2 \mathbb{E} \left[ \left\| \Delta \tilde{H}^{(i)} \right\|_2^6 \right] \\
\leq 2^{2} 3^{5} 5^{6} \mu_3(\sigma)^2 \pi_{1,6}(\sigma) \left( 1 + \mathcal{V}_4^{3/4} \right) t^3. \tag{47}
\]

Hence, combining (45), (46), and (47),
\[
\mathbb{E} \left[ \left\| X_t - \tilde{X}_t \mid \mathcal{F}_0 \right\|_2^2 \right] = \mathbb{E} \left[ \mathbb{E} \left[ S(t) \mid \mathcal{F}_0 \right] - \mathbb{E} \left[ \frac{1}{2} \sum_{i=1}^{m} (\phi_1^{(i)}(t) - \phi_2^{(i)}(t)) \sqrt{t} \mid \mathcal{F}_0 \right] \right] \right|_2^2 \\
\leq 4 \mathbb{E} \left[ \left\| S_1(t) \right\|_2^2 \right] + 4 \mathbb{E} \left[ \left\| S_2(t) \right\|_2^2 \right] + 4 \mathbb{E} \left[ \left\| \frac{1}{2} \sum_{i=1}^{m} (\phi_1^{(i)}(t) - \phi_2^{(i)}(t)) \sqrt{t} \right\|_2^2 \right] \\
\leq \left( \frac{4}{3} \mu_1(b)^2 \pi_{1,2}(b) \left( 1 + \mathcal{V}_2 \right) + \frac{1}{3} \mu_2(b)^2 \pi_{1,4}(\sigma) \left( 1 + \mathcal{V}_2 \right) m^2 \right. \\
\left. + 2^{4} 3^{5} 5^{6} \mu_3(\sigma)^2 \pi_{1,6}(\sigma) \left( 1 + \mathcal{V}_4^{3/4} \right) \right) t^4.
\]

C.3 Invoking Theorem 1

Now, we invoke Theorem 1 with our derived constants. We obtain that if the constant step size
\[
h < 1 \land C_h \frac{1}{2\alpha} \land \frac{1}{8 \mu_1(b)^2 + 8 \mu_1(\sigma)^2},
\]

45
where
\[ C_h = \frac{1}{m^2} \land \frac{\alpha^2}{4M_1^2} \land \frac{\alpha^2}{M_3^2}. \]
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 \( \epsilon \), \( \tilde{O}(d^{3/4}m^2\epsilon^{-1}) \) iterations are required, since \( D_3 \) is of order \( \tilde{O}(d^{3/2}m^3) \), and \( D_4 \) is of order \( \tilde{O}(d^{3/2}m^2) \). Note that the dimension dependence worsens if one were to further convert the Frobenius norm dependent constants to be based on the operator norm.

\[ D \quad \text{Convergence Rate for Example 2} \]

\[ D.1 \quad \text{Moment Bound} \]

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.

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

where \( \mathcal{W}_2 = \mathbb{E} \left[ \left\| X_0 \right\|_2^2 \right] + 2(\pi_{1,2}(b) + \beta)/\alpha, \) if the constant step size \( h < 1 \land \alpha/(2\pi_{1,2}(b)). \)

**Proof.** By direct computation,

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

Recall by Lemma 15 and dissipativity,

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

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

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

By unrolling the recursion,

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

\[ D.2 \quad \text{Local Deviation Orders} \]

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.

$$
\mathbb{E} \left[ \|X_t\|^2 \right] \leq \mathcal{W}_2 + \beta/\alpha = \mathcal{W}_2^*, \quad \text{for all } t \geq 0.
$$

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

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

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

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

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

where $E_1 = (\mu_1(b)^2 + \mu_1^F(\sigma)^2) E_0$.

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

$$
\mathbb{E} \left[ \|X_t - \tilde{X}_t\|^2 \right] \leq 2 \mathbb{E} \left[ \left\| \int_0^t (b(X_s) - b(X_0)) \, ds \right\|^2 \right] + 2 \mathbb{E} \left[ \left\| \int_0^t (\sigma(X_s) - \sigma(X_0)) \, dB_s \right\|^2 \right]
\leq 2 t \mathbb{E} \left[ \int_0^t \|b(X_s) - b(X_0)\|^2 \, ds \right] + 2 \mathbb{E} \left[ \int_0^t \|\sigma(X_s) - \sigma(X_0)\|^2 \, dF_s \right]
\leq 2 \left( \mu_1(b)^2 + \mu_1^F(\sigma)^2 \right) \int_0^t \mathbb{E} \left[ \|X_s - X_0\|^2 \right] \, ds
\leq \left( \mu_1(b)^2 + \mu_1^F(\sigma)^2 \right) E_0 t^2.
$$

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

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

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

Proof. By Itô’s lemma,

$$
X_t - X_0 = \int_0^t b(X_s) \, ds + \sigma(X_0) B_t
+ \sum_{i=1}^m \sum_{l=1}^m \int_0^t \int_0^s \Lambda_i(\sigma_l)(X_u) \, dB_u^{(l)} \, dB_s^{(l)}
+ \sum_{i=1}^m \int_0^t \int_0^s \Lambda_i(\sigma_l)(X_u) \, dW_u \, dB_s^{(l)}.
$$

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

$$
\mathbb{E} \left[ X_t - X_0 | \mathcal{F}_0 \right] = \mathbb{E} \left[ \int_0^t (b(X_s) - b(X_0)) \, ds | \mathcal{F}_0 \right].
$$
Hence, by Jensen’s inequality,
\[
\mathbb{E}\left[\left\|\mathbb{E}\left[X_t - \tilde{X}_t | \mathcal{F}_0\right]\right\|_2^2\right] = \mathbb{E}\left[\left\|\mathbb{E}\left[\int_0^t (b(X_s) - b(X_0)) \, ds | \mathcal{F}_0\right]\right\|_2^2\right]
\leq \mathbb{E}\left[\left\|\int_0^t (b(X_s) - b(X_0)) \, ds\right\|_2^2\right]
\leq \mu_1(b) t \int_0^t \mathbb{E}\left[\|X_s - X_0\|_2^2\right] \, ds
\leq \mu_1(b) E_0 t^3 / 2.
\]

\[\square\]

D.3 Invoking Theorem 1

Now, we invoke Theorem 1 with our derived constants. We obtain that if the constant step size
\[
h < 1 \wedge \frac{\alpha}{2 \pi_{1,2}(b)} \wedge \frac{1}{2 \alpha} \wedge \frac{1}{8 \mu_1(b)^2 + 8 \mu_1^2(\sigma)^2},
\]
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 \(\epsilon\), \(\tilde{O}(d \epsilon^{-2})\) iterations are required, since both \(E_1\) and \(E_2\) are of order \(O(d)\). If one were to convert the Frobenius norm dependent constants to be based on the operator norm, then \(E_1\) is of order \(O(d(d + m)^2)\), and \(E_2\) is of order \(O(d(d + m))\). This yields the convergence rate of \(\tilde{O}(d(d + m)^2 \epsilon^{-2})\).

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

Proof. Recall the function \(f(x) = x^n\) is convex for \(n \in \mathbb{N}_+\). Hence,
\[
\left(\frac{\sum_{i=1}^{m} x_i^m}{m}\right)^n \leq \frac{\sum_{i=1}^{m} x_i^n}{m}.
\]
Multiplying both sides of the inequality by \(m^n\) completes the proof. \(\square\)

Lemma 32. For the \(d\)-dimensional Brownian motion \(\{B_t\}_{t \geq 0}\),
\[
Z_t = \int_0^t \int_0^s dB_u \, ds \sim \mathcal{N} \left(0, t^3 I_d / 3\right).
\]

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
\[
S_m = \sum_{k=0}^{m-1} B_{t_k} (t_{k+1} - t_k) = \sum_{k=1}^{m-1} (B_{t_{k+1}} - B_{t_k}) (t_k - t).
\]
Since \(S_m\) is a sum of Gaussian random variables, it is also Gaussian. By linearity of expectation and independence of Brownian motion increments,
\[
\mathbb{E}[S_m] = 0,
\]
\[
\mathbb{E}[S_m^2] = \mathbb{E}\left[\sum_{k=1}^{m-1} (B_{t_{k+1}} - B_{t_k}) (t_k - t)^2\right] = \mathbb{E}\left[\sum_{k=1}^{m-1} (B_{t_{k+1}} - B_{t_k})^2 (t_k - t)\right] = \sum_{k=1}^{m-1} \mathbb{E}[B_{t_{k+1}}^2] (t_k - t) = \sum_{k=1}^{m-1} t_k (t_k - t) = \frac{1}{2} \int_0^m t \, dt = \frac{m^2}{2}.
\]

Therefore, \(Z_t\) is normally distributed with mean 0 and variance \(t^3 / 3\).
Then, the vector Laplacian of its gradient is bounded, i.e.
\[ \|B_t\|^2_2 \leq t^n \|x - y\|^2 \|Df(x)\|_{op} \]
for all \( x, y \in \mathbb{R}^d \).

Then, the vector Laplacian of its gradient is bounded, i.e.
\[ \|\nabla^2 f(x) - \nabla^2 f(y)\|_{op} \leq \mu_3 \|x - y\|_2, \quad \text{for all } x, y \in \mathbb{R}^d. \]

Then, the vector Laplacian of its gradient is bounded, i.e.
\[ \|\Delta(\nabla f)(x)\| \leq \mu_3, \quad \text{for all } x \in \mathbb{R}^d. \]


Lemma 35. For \( f : \mathbb{R}^d \to \mathbb{R} \) which is \( C^4 \), suppose its third derivative is \( \mu_4 \)-Lipschitz under the operator norm and Euclidean norm, i.e.
\[ \|\nabla^3 f(x) - \nabla^3 f(y)\|_{op} \leq \mu_4 \|x - y\|_2, \quad \text{for all } x, y \in \mathbb{R}^d. \]

Then, the vector Laplacian of its gradient is \( d\mu_4 \)-Lipschitz, i.e.
\[ \|\Delta(\nabla f)(x) - \Delta(\nabla f)(y)\|_2 \leq d\mu_4 \|x - y\|_2. \]

Proof. Let \( g(x) = \Delta f(x) \). Since \( f \in C^4 \), we may switch the order of partial derivatives,
\[ \|\Delta(\nabla f)(x) - \Delta(\nabla f)(y)\|_2 = \|\nabla g(x) - \nabla g(y)\|_2. \]

By Taylor’s theorem with the remainder in integral form,
\[ \|\nabla g(x) - \nabla g(y)\|_2 = \left\| \int_0^1 \nabla^2 g(y + \tau(x - y)) \left| x - y \right| \, d\tau \right\|_2 \leq \int_0^1 \|\nabla^2 g(y + \tau(x - y))\|_{op} \|x - y\|_2 \, d\tau \leq \sup_{z \in \mathbb{R}^d} \|\nabla^2 g(z)\|_{op} \|x - y\|_2. \]
We compare the vanilla estimate where the expectations are approximated via averaging 100 independent draws. Figure 2 reports the 
\[ \Sigma \]
where
\[ t \]
To better detect convergence, we zero center a simple sample-based estimator by subtracting the null deviation across different sample sizes and dimensionalities, where
\[ \mu \]
This is a consequence of the Wasserstein distance metrizing weak convergence [58] and that the family [29] and the Sinkhorn divergence [46]. Note that 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.

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.

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

Since the operator norm of \( \nabla^4 f(x) \) upper bounds the operator norm of each of its sub-tensor,
\[ \|\nabla^2 g(x)\|_{op} \leq \sum_{i=1}^{d} \|G_i(x)\|_{op} \leq d \|\nabla^4 f(x)\|_{op} . \]

Recall the third derivative is \( \mu \)-Lipschitz, we obtain
\[ \|\nabla g(x) - \nabla g(y)\|_2 \leq d \mu_3 \|x - y\|_2 . \]

\( \square \)

F Estimating the Wasserstein Distance

For a Borel measure \( \mu \) defined on a compact and separable topological space \( \mathcal{X} \), a sample-based empirical measure \( \mu_n \) may asymptotically serve as a proxy to \( \mu \) in the \( W_p \) sense for \( p \in [1, \infty) \), i.e.
\[ W_p(\mu, \mu_n) \xrightarrow{\text{a.s.}} 0. \]

This is a consequence of the Wasserstein distance metrizing weak convergence [58] and that the empirical measure converges weakly to \( \mu \) almost surely [56].

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 [16, 15], Weed and Bach [60] showed that for \( p \in [1, \infty) \),
\[ W_p(\mu, \mu_n) \gtrsim n^{-1/t} , \]
where \( t \) is less than the lower Wasserstein dimension \( d_s(\mu) \). 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:
\[ \tilde{W}_2^2(\mu, \nu) = \frac{1}{2} \left( W_2^2(\mu_n, \nu_n) + W_2^2(\nu_n, \mu_n) - W_2^2(\mu_n, \nu_n) - W_2^2(\nu_n, \nu_n) \right) , \]
where \( \nu_n \) and \( \nu_n' \) are based on two independent samples of size \( n \) from \( \mu \), and similarly for \( \nu_n \) and \( \nu_n' \) from \( \nu \). This estimator is inspired by the construction of distances in the maximum mean discrepancy family [29] and the Sinkhorn divergence [46]. Note that the 2-Wasserstein distance between finite samples can be computed conveniently with existing packages [23] 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 \( \mu \) and \( \nu \) 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
\[ W_2^2(\mu, \nu) = \|m_1 - m_2\|^2 + \text{Tr} \left( \Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2} \right) . \]

We compare the vanilla estimate \( \tilde{W}_2^2(\mu, \nu) \) and the corrected estimate \( \tilde{W}_2^2(\mu, \nu) \) by their magnitude of deviation from the true value \( W_2^2(\mu, \nu) \):
\[ |W_2^2(\mu, \nu) - \mathbb{E}[\tilde{W}_2^2(\mu, \nu, n)]| , \quad |W_2^2(\mu, \nu) - \mathbb{E}[\tilde{W}_2^2(\mu, \nu, n)]| , \]
where the expectations are approximated via averaging 100 independent draws. Figure 2 reports the deviation across different sample sizes and dimensionalities, where \( \mu \) and \( \nu \) 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.
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 $\mu$ and $\nu$. Darker curves correspond to larger number of samples used to compute the empirical estimate (ranging from 100 to 1000). (a) $m_1 = 0, m_2 = 1, \Sigma_1 = \Sigma_2 = I_d$. (b) $m_1 = m_2 = 0, \Sigma_1 = I_d, \Sigma_2 = I_d/2 + 1_d1_d^T / 5$.

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 $\mu$ and $\nu$. 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) $m_1 = 0, m_2 = 21, \Sigma_1 = \Sigma_2 = I_d$. (b) $m_1 = m_2 = 0, \Sigma_1 = 2I_d, \Sigma_2 = I_d/2 + 1_d1_d^T / 5$.

G Additional Numerical Studies

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

G.1 Strongly Convex Potentials

We first include the detailed setup for BLR in Section 5. We then 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.

G.1.1 Setup for BLR

To obtain the potential, we generate data from the model with the parameter $\theta_* = 1_d$ following [10, 19]. 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 [10, 19], where each $x_i$...
is normalized by its Euclidean norm. We sample the corresponding \( y_i \) from the model and fix the regularizer \( \alpha = 0.3d/\pi^2 \).

### G.1.2 Additional Results

Figure 4 shows the estimated \( W_2^2 \) 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.

![Error in \( W_2^2 \) for strongly log-concave sampling](image)

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 [54, 55] under the Euclidean norm. For probability measures \( \mu \) and \( \nu \) on \( \mathbb{R}^d \) with finite first moments, this distance is defined to be the square root of

\[
D_E(\mu, \nu)^2 = 2E[||Y - Z||_2] - E[||Y - Y'||_2] - E[||Z - Z'||_2],
\]

where \( Y, Y' \sim \mu \) and \( Z, Z' \sim \nu \). The moment condition is required to ensure that the expectations in (48) 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 [52]

\[
\hat{D}_E(\mu, \nu)^2 = \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} ||Y_i - Z_j||_2 + \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} ||Y_i - Y'_j||_2 + \frac{1}{n^2} \sum_{i=1}^{m} \sum_{j=1}^{n} ||Z_i - Z'_j||_2,
\]

where \( Y_i \sim \mu \) for \( i = 1, \ldots, m \) and \( Z_j \sim \nu \) for \( j = 1, \ldots, n \). 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 \( W_2^2 \).

### G.1.3 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.

### G.1.4 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.

### G.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.
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 \[31\]. The infinite series can be written as follows:

$$ I_{(l,i)} = \frac{I_{(l)}I_{(i)}}{2} - h\delta_{l,i} + A_{(l,i)}, $$

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

where $\xi_{l,k}, \xi_{l,k}, \eta_{l,k}, \eta_{l,k} \overset{i.i.d.}{\sim} N(0,1)$. $A_{(l,i)}$ is known as the Lévy area and is notoriously hard to simulate \[62\].

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 \( n \propto h^{-1} \) [31]. 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. [62] 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.

### G.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 dimen-
sionality increases across all settings with differing \( \beta \) and \( \gamma \) 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 \( \beta \) is large.
Nevertheless, the potential is never strongly convex when \( \beta \) is positive due to the linear growth term.

Figure 8: Wall time for sampling from a 20D Gaussian mixture.