Submitted to Bernoulli
arXiv: arXiv:1907.11331

Improved Bounds for Discretization of Langevin Diffusions: Near-Optimal Rates without Convexity

WENLONG MOU[1,*] NICOLAS FLAMMARION[3,**]
MARTIN J. WAINWRIGHT[1,2,†] and PETER L. BARTLETT[1,2,‡]

[1]*Department of EECS, UC Berkeley, Cory Hall, Berkeley, CA, 94720, USA,*
E-mail: [*]wmou@berkeley.edu; [†]wainwrig@berkeley.edu; [‡]peter@berkeley.edu

[2]*Department of Statistics, UC Berkeley, Evans Hall, Berkeley, CA, 94720, USA*

[3]*School of Computer and Communication Sciences, EPFL, INJ 336, Station 14, CH-1015 Lausanne, Switzerland, E-mail:* [**]nicolas.flammarion@epfl.ch

Discretizations of the Langevin diffusion have been proven very useful for developing and analyzing algorithms for sampling and stochastic optimization. We present an improved non-asymptotic analysis of the Euler-Maruyama discretization of the Langevin diffusion. Our analysis does not require global contractivity, and yields polynomial dependence on the time horizon. Compared to existing approaches, we make an additional smoothness assumption, and improve the existing rate in discretization step size from $O(\eta)$ to $O(\eta^2)$ in terms of the KL divergence. This result matches the correct order for numerical SDEs, without suffering from exponential time dependence. When applied to MCMC, this result simultaneously improves on the analyses of a range of sampling algorithms that are based on Dalalyan's approach.

Keywords: Langevin diffusion, Markov chain Monte Carlo, Euler-Maruyama discretization, KL divergence, Non-asymptotic bound.

1. Introduction

In recent years, the machine learning and statistics communities have witnessed a surge of interest in the Langevin diffusion process, and its connections to stochastic algorithms for sampling and optimization. The Langevin diffusion in $\mathbb{R}^d$ is defined via the Itô stochastic differential equation (SDE)

$$dX_t = b(X_t)dt + dB_t, \tag{1.1}$$

where B_t is a standard d-dimensional Brownian motion, and the function $b : \mathbb{R}^d \to \mathbb{R}^d$ is known as the drift term. For a drift term of the form $b(x) = -\frac{1}{2}\nabla U(x)$ for some differentiable function $U : \mathbb{R}^d \to \mathbb{R}$, the Langevin process (1.1) has stationary distribution with density $\gamma(x) \propto e^{-U(x)}$; moreover, under mild growth conditions on U, the diffusion converges to this stationary distribution as $t \to \infty$. See the book [44] for more

background on these facts, which underlie the development of sampling algorithms based on discretizations of the Langevin diffusion. Diffusive processes of this nature also play an important role in understanding stochastic optimization; in this context, the Gaussian noise helps escaping shallow local minima and saddle points in finite time, making it especially useful for non-convex optimization. From a theoretical point of view, the continuous-time process is attractive to analyze, being amenable to a range of tools coming from stochastic calculus and Brownian motion theory [46]. However, in practice, an algorithm can only run in discrete time, so that the understanding of discretized versions of the Langevin diffusion is very important.

The discretization of SDEs is a central topic in the field of scientific computation, with a wide variety of schemes proposed and studied (e.g., see the standard sources [33, 29, 43] and references therein). The most commonly used method is the Euler-Maruyama discretization: parameterized by a step size $\eta > 0$, it is defined by the recursion

$$\hat{X}_{(k+1)\eta} = \hat{X}_{k\eta} + \eta b(\hat{X}_{k\eta}) + \sqrt{\eta}\xi_k, \quad \text{for } k = 0, 1, 2, \ldots. \tag{1.2}$$

Here the sequence $\{\xi_k\}_{k=0}^{+\infty}$ consists of i.i.d. d-dimensional standard Gaussian random vectors.

The behavior of the recursion (1.2), along with higher-order variants thereof, has been intensively studied for decades, under different sets of assumptions. The classical papers [50, 51] provide a guarantee of weak error expansion, giving a first-order error bound on the forward Euler method, albeit one that exhibits exponential dependence on the time horizon T. Under mild assumptions on the potential function, the process (1.2) can be shown to be geometrically ergodic as $T \to +\infty$ (e.g., [32, 41]). The Metropolis-adjusted variant of the process (1.2) has also been studied, with various asymptotic and non-asymptotic guarantees known [8, 7, 47, 22, 24].

As mentioned previously, the Euler-Murayama scheme is known to have first-order accuracy under appropriate smoothness conditions. In particular, if we measure distance using the Wasserstein distance $\mathcal{W}_p$ for some $p \geq 1$, then the distance between the original Langevin diffusion and the discretized version decays as $O(\eta)$ as η decays to zero; here the dependence on the dimension d and time horizon T is subsumed within the order notation [see, e.g., 1]. When the underlying dependence on the time horizon T is explicitly calculated, it can grow exponentially, due to the underlying Grönwall inequality. If the potential U is both suitably smooth and strongly convex, then the scaling with the step size η remains first-order, and the bound becomes independent of time T [16, 21]. These bounds, in conjunction with the coupling method, have been used to provide non-asymptotic and explicit bounds on the mixing time of the unadjusted Langevin algorithm (ULA) for sampling from strongly-log-concave densities. Moreover, this bound aligns well with the classical theory of discretization for ordinary differential equations (ODEs), where finite-time discretization error may suffer from bad dependence on T, and either contraction assumptions or symplectic structures are needed in order to control long-time behavior [31].

On the surface, it might seem that SDEs pose greater numerical challenges than ODEs; however, the presence of randomness actually has been shown to help in the

long-term behavior of discretization. Most closely related to the current paper, the seminal work of Dalalyan [15] showed that the pathwise Kullback-Leibler (KL) divergence between the original Langevin diffusion (1.1) and the Euler-Maruyma discretization (1.2) is bounded as $O(\eta T)$ with only smoothness conditions. This result enables comparison of the discretization with the original diffusion over long time intervals, even without contraction. The discretization techniques of Dalalyan [15] are further developed by Durmus and Moulines [20]; both papers serve as a foundation for a number of recent papers on sampling and non-convex statistical learning, including a line of recent work (e.g., [45, 53, 36, 37]).

On the other hand, this $O(\eta)$ bound on the KL error is likely to be loose in general. Under suitable smoothness conditions, standard transportation inequalities [5] guarantee that such a KL bound can be translated into an $O(\sqrt{\eta})$-bound in Wasserstein (and TV) distance. Yet, as mentioned in the previous paragraph, the Wasserstein (and TV) rate should scale as $O(\eta)$ under an appropriate smoothness assumption. This latter result either requires assuming contraction or leads to exponential time dependence, raising naturally the question: can we achieve best of both worlds? That is, is it possible to prove $O(\eta T)$ bounds in the Wasserstein and TV distances without requiring convexity or other contractivity conditions?

Our contributions: In this paper, we answer the preceding question in the affirmative: more precisely, we close the gap between the correct rate for the Euler-Maruyama method and the linear dependence on time horizon, without any contractivity assumptions. Furthermore, as opposed to prior works where the long-term stability is asymptotic or built upon the ergodicity of the process, our discretization error bound is *non-asymptotic* and *explicit*, and does not require the mixing of the underlying process. As long as the drift term satisfies certain first and second-order smoothness, as well as a certain type of distant growth condition, we show the KL divergence between marginal distributions of the processes (1.1) and equation (1.2), at any time T, is bounded as $O(\eta^2 d^2 T)$. Note that this bound is non-asymptotic, with polynomial dependence on all the smoothness parameters, and linear dependence on T. As a corollary of this improved discretization bound, we give improved bounds for using the unadjusted Langevin algorithm (ULA) for sampling from a distribution satisfying a log-Sobolev inequality. In addition, our improved discretization bound improves a number of previous results on non-convex optimization and inference, all of which are based on the discretized Langevin diffusion.

In the proof of our main theorem, we introduce a number of new techniques. A central challenge is how to study the evolution of time marginals of the interpolation of discrete-time Euler algorithm. In order to do so, we derive a Fokker-Planck equation for the interpolated process, where the drift term is the backward conditional expectation of b at the previous step, conditioned on the current value of x. The difference between this new drift term for the interpolated process and b itself can be much smaller than the difference between b at two time points. Indeed, taking the conditional expectation cancels out the bulk of the noise terms, assuming the density from the previous step is smooth enough. We capture the smoothness of density at the previous step by its Fisher information, and develop a recursive bound using the convolution inequality in order

to control the Fisher information along the path. Combining this regularity estimate with suitable tail bounds leads to our main result. We suspect that our analysis of this interpolated process may be of independent interest.

Related work: Recent years have witnessed a flurry of activity in statistics and machine learning on the Langevin diffusion and related stochastic processes. A standard application is sampling from a density of the form $\gamma(x) \propto e^{-U(x)}$ based on a first-order oracle that returns the pair $(U(x), \nabla U(x))$ for any query point x. In the log-concave case, algorithms for sampling under this model are relatively well-understood, with various methods for discretization and variants of Langevin diffusion proposed in order to refine the dependence on dimension, accuracy level and condition number [15, 21, 14, 35, 39, 22, 17].

When the potential function U is non-convex, the analysis of continuous-time convergence and the discretization error analysis both become much more involved. When the potential satisfies a logarithmic Sobolev inequalities, continuous-time convergence rates can be established [see e.g. 40], and these guarantees have been leveraged for sampling algorithms [4, 56, 37]. Coupling-based results for the Wasserstein distance $\mathcal{W}_1$ have also been shown for variants of Langevin diffusion [13, 6], and [38] show a bound under a stronger $\mathcal{W}_2$ distance under similar assumptions. For the non-convex case, these approaches typically require (strong) convexity outside a ball, which excludes many important probability distributions. See Section A for more discussion.

Beyond sampling, the global convergence nature of Langevin diffusion has been used in non-convex optimization, since the stationary distribution is concentrated around global minima. Langevin-based optimization algorithms have been studied under log-Sobolev inequalities [45], bounds on the Stein factor [25]; in addition, accelerated methods have been studied [11]. The dynamics of Langevin algorithms have also been studied without convergence to stationarity, including exiting times [53], hitting times [57], exploration of a basin-of-attraction [34], and statistical inference using the path [36].

Much of this work in the non-convex setting is based on the discretization methods first introduced by Dalalyan [15], and subsequently refined and generalized by Durmus and Moulines [20]. This approach, which leads to an $O(\sqrt{\eta})$ bound in total variation distance, is based on computing pathwise KL divergence using Girsanov's theorem. Notably, this approach does not lead to the optimal bound for a first-order scheme. In particular, when a second-order smoothness condition is assumed on the drift term $b(\cdot)$, their bounds in terms of TV distance are loose compared to the $O(\eta)$ bound given here. Cheng and Bartlett [12] used this same approach to prove an $O(\eta)$ bound in KL divergence, as opposed to the sharper $O(\eta^2)$ bound that we establish in this paper. Recently, De Bortoli and Durmus [19] used this Girsanov-based technique to establish asymptotic discretization error guarantees. They study a very general class of drift terms, one which does not require global Lipschitzness and allow arbitrary speed of growth. We believe that our method might be fruitfully combined with their analysis, thereby replacing the Girsanov-based argument so as to obtain non-asymptotic guarantees in their more general setting.

[1] We only listed time horizon dependence for methods that guarantee discretization error between

Paper	Contractive operator	[1]Time T	[2] Step size η	Requires mixing	Additional assumptions
[15],[20]	No	$O(\sqrt{T})$	$O(\sqrt{\eta})^*$	No	None
[1]	No	$O(e^{cT})$	$O(\eta)$	No	Second-order smooth drift
[16],[21]	Yes	-	$O(\eta)$	Yes	Second-order smooth drift
[12],[37]	No	-	$O(\sqrt{\eta})^*$	Yes	strong convexity outside a ball
[13],[6][3]	No	-	$O(\eta)$	Yes	strong convexity outside a ball
This paper	No	$O(\sqrt{T})$	$O(\eta)^*$	No	Second-order smooth drift

Table 1. Comparison between non-asymptotic error bounds on discretization of Langevin diffusion and MCMC sampling algorithms.

Finally, in a concurrent and independent line of work, Fang and Giles [26] also studied a multi-level sampling algorithm without imposing a contraction condition, and obtained bounds for the mean-squared error; however, their results do not give explicit dependence on problem parameters. Since the proofs involve bounding the moments of Radon-Nikodym derivative, their results may be exponential in dimension, as opposed to the polynomial-dependence given here.

Notation: We let $\|x\|_2$ denote the Euclidean norm of a vector $x \in \mathbb{R}^d$. For a matrix M we let $\|M\|_{\mathrm{op}}$ denote its spectral norm. For a function $b : \mathbb{R}^d \to \mathbb{R}^d$, we let $\nabla b(x) \in \mathbb{R}^{d\times d}$ denote its Jacobian evaluated at x. We use $\mathcal{L}(X)$ to denote the law of random variable X. When the variable of the integrand is not explicitly written, integrals are taken with respect to the Lebesgue measure: in particular, for an integrable function $g : \mathbb{R}^d \to \mathbb{R}$, we use $\int g$ as a shorthand for $\int_{\mathbb{R}^d} g(x)dx$. For a continuously differentiable probability density function p (with respect to the Lebesgue measure), we use $\mathcal{I}(p)$ to denote the (scalar) Fisher information for p with respect to the Lebesgue measure, i.e.,

$$\mathcal{I}(p) := \int p(x) \|\nabla \log p(x)\|_2^2 \, dx.$$

For real numbers x, y, we use $x \wedge y$ to denote $\min(x, y)$ and use $x \vee y$ to denote $\max(x, y)$.

2. Main Results and their Implications

We now turn to our main results, beginning with our assumptions and statement of our main theorems. We then develop and discuss a number of corollaries of our results.

continuous-time and discrete-time for any time. If the proof requires mixing and does not bound the difference between the finite-time distributions, we mark it as "-".

[2]All distances are measured in $\mathcal{W}_1$. If the original bound was shown for the KL divergence, we have transformed it into the $\mathcal{W}_1$-distance using transportation inequalities. We mark with * if the original bound was shown in the KL divergence.

[3]For Hamiltonian Monte-Carlo, which is based on discretization of ODE, instead of SDE.

2.1. Statement of main results

Our main results involve three conditions on the drift term b, and one on the initialization:

***Assumption* 2.1** (Lipschitz drift term)**.** *There is a finite constant L_1 such that*

$$\|b(x) - b(y)\|_2 \le L_1 \|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^d. \tag{2.1}$$

***Assumption* 2.2** (Smooth drift term)**.** *There is a finite constant L_2 such that*

$$\|\nabla b(x) - \nabla b(y)\|_{op} \le L_2 \|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^d. \tag{2.2}$$

***Assumption* 2.3** (Distant dissipativity)**.** *There exist strictly positive constants μ, β such that*

$$\langle b(x), x\rangle \le -\mu \|x\|_2^2 + \beta \quad \text{for all } x \in \mathbb{R}^d. \tag{2.3}$$

***Assumption* 2.4** (Initial distribution)**.** *The initializations X_0 and $\hat{X}_0$, for the processes (1.1) and (1.2) respectively, are drawn from a density π_0 such that*

$$\left(\mathbb{E}_{\pi_0} \|X\|_2^4\right)^{1/4} \le \sigma_0 \sqrt{d} \qquad \text{for some finite } \sigma_0. \tag{2.4}$$

Note that we do *not* impose any contractivity assumption on the drift term b. Rather, we use the notion of distant dissipativity, which is substantially weaker; moreover, even this assumption is relaxed in Theorem 2. A spectrum of growth conditions on the drift term (or tail conditions for the stationary distribution for sampling problems) have been employed in literature, including (strong) convexity [21, 18], strong convexity outside a ball [37, 23], distant dissipativity [45, 27]; such growth conditions are also closely related to functional inequalities that guarantee rapid mixing, such as the log-Sobolev inequality, Poincaré inequality, and Talagrand's inequality. In Section A of the supplementary material, we carry out a comprehensive discussion of these conditions used in literature.

We note that the initialization condition (2.4) is rather standard, and clearly satisfied, for instance, by the standard Gaussian density with $\sigma_0 = \sqrt{3}$.

With these definitions, the main result of this paper is the following:

Theorem 1. *Consider the Langevin diffusion (1.1) under Assumptions 2.1—2.4. Then there are universal constants (c_0, c_1) such that for any $\eta \in (0, \frac{1}{2L_1})$ and all times $T > 0$, the KL error of the Euler-Maruyama discretization (1.2) is bounded as*

$$D_{KL}(\hat{\pi}_T \| \pi_T) \le c_0 \eta^2 L_1^2 T \left\{ \frac{d \log \eta^{-1} T}{T} \wedge \mathcal{I}(\pi_0) + L_1 d + A_0^2 + L_1^2 \left(\sigma_0^2 d + \tfrac{\beta + d}{\mu} \right) \right\} + c_0 \eta^2 T L_2^2 d^2 + c_1 \eta^4 L_2^2 \left\{ A_0^4 + L_1^4 \left(\sigma_0^4 d^2 + \tfrac{(\beta+d)^2}{\mu^2} + d^2 \right) \right\}, \tag{2.5}$$

where $A_0 := \|b(0)\|_2$.

It is important to note that this bound is *non-asymptotic*, holding for all times, and provides *explicit* dependence on all the problem parameters. The bound allows for any initial distribution π_0 satisfying the moment condition (2.4). When π_0 has bounded Fisher information, then Theorem 1 gives an $O(\eta^2)$ dependency on the stepsize, matching the optimal numerical order for Euler schemes. Note that this bounded Fisher information condition is often satisfied; for instance, in sampling for Monte Carlo estimation, one might simply take a standard Gaussian initialization. On the other hand, even if the initial distribution lacks a density, Theorem 1 still gives an $O(\eta^2 \log \eta^{-1})$ bound in terms of the KL divergence, and moreover, this scaling is near-optimal.

It is worthwhile highlighting certain aspects of the dependency on the triple (η, T, d). In order to do so, we consider the following *canonical setup* of problem parameters:

- The smoothness parameters L_1, L_2 are of constant order, independent of problem dimension.
- The quantity $A_0 = \|b(0)\|_2$ is of order $O(\sqrt{d})$.
- The dissipativity condition 2.3 is satisfied by a pair (μ, β) such that $\mu = \Omega(1)$ and $\beta = O(d)$.
- The initial distribution is taken as standard Gaussian $\mathcal{N}(0, I_d)$.

Note that the first and third requirements are just scaling conditions, and can be satsified by rescaling the coordinates. The second condition—namely, the bound on $\|b(0)\|_2$—can be satisfied by finding an approximate solution to the nonlinear equation $b(x) = 0$, and then translating the coordinate system so as to satisfy this condition.

Under the canonical setup, if we track only the dependence on (η, T, d), the result (2.5) can be summarized as a bound of the form $D_{\mathrm{KL}}(\hat{\pi}_T \| \pi_T) \lesssim \eta^2 d^2 T$. This result should be compared to the $O(\eta d T)$ bound from past work [15] that imposes only Assumption 2.2. It is also worth noticing that the term $\eta^2 d^2 L_2^2 T$ only comes with the third order derivative bound, which coincides with the Wasserstein distance result, based on a coupling proof, as obtained in the papers [21, 16]. However, these works require a contractivity assumption, and do not study separately the discretization error of the discrete process. On the other hand, the bounds from the paper [20] do not require the contractivity condition. To compare with their result, Theorem 1 can be transformed into an $O(\eta d \sqrt{T})$ bound in terms of total variation distance using Pinsker's inequality. In comparison, results in the paper [20] provide a bound that scales as $O(\sqrt{\eta d T})$. Our bound improves upon this result for step size $\eta = O(1/d)$, which is required to make the discretization error small.

Note that Assumption 2.3 can be substantially relaxed when the drift function corresponds to the negative gradient of some function f. Essentially, we only require the function f to be non-negative, along with the smoothness assumptions. In such case, we have the following discretization error bound:

Theorem 2. *Consider the original Langevin diffusion (1.1) under Assumptions 2.1, 2.2, and 2.4, and suppose that $b = -\nabla f$ for some non-negative function f. Then for any stepsize $\eta \in (0, \frac{1}{2L_1})$ and time $T > 0$, the KL error of the Euler-Maruyama discretiza-*

tion (1.2) *is bounded as*

$$D_{KL}(\hat{\pi}_T\|\pi_T) \le c_0\eta^2 L_1^2 T\left(\frac{d\log\eta^{-1}T}{T} \wedge \mathcal{I}(\pi_0) + L_1 d + \frac{\mathbb{E}_{\pi_0}[f(X)]}{T}\right) + c_0\eta^2 T L_2^2 d^2 + c_1\eta^4 L_2^2\left\{A_0^4 + L_1^4\left(\mathbb{E}_{\pi_0}[f(X)^2] + L_1^2T^2\sigma_0^4 d^2 + L_1^2T^4 d^2\right)\right\}, \quad (2.6)$$

where $A_0 := \|b(0)\|_2$.

Once again tracking only the dependence on the triple (η, T, d), the bound (2.6) can be stated succinctly as $D_{\mathrm{KL}}(\hat{\pi}_T\|\pi_T) \lesssim \eta^2 T(1+\eta^2T^3)d^2$, where $\lesssim$ denotes inequality that holds up to quantities independent of (η, T, d). Relative to Theorem 1, this bound has weaker dependency on the time horizon T, but it holds for any non-negative potential function without any growth conditions. We note that neither Theorem 1 nor Theorem 2 depend on the mixing time of the underlying process, which can be exponentially large for multi-modal problems [9]. Indeed, the condition $f \ge 0$ does not even guarantee the existence of a stationary distribution, and Theorem 2 holds true even when the underlying process is not ergodic at all. This bound broadens the class of SDEs for which the long-term structure-preservation is guaranteed.

It is also worthwhile comparing the conditions and bounds in Theorem 2 to those of Durmus and Moulines [20]. When transformed into a bound in TV distance using Pinsker's inequality, Theorem 2 leads to a bound of the order $O(\eta d\sqrt{T(1+\eta^2T^3)})$, while the bound in the paper [20] is of the order $O(\sqrt{\eta dT})$. Our bound leads to an improved numerical order, and is better than their result for a stepsize $\eta \in (0, \frac{1}{T+d})$; note that such a choice is required to make the overall discretization error be bounded as $O(1)$. On the other hand, the bound in Theorem 2 does not require any condition on mixing properties of the underlying Langevin diffusion, and gives explicit dependency on all the problem parameters. This result can be combined with any mixing time bound in TV distance to get the mixing time upper bound in Corollary 1. In comparison, Proposition 2 in the paper [20] assumes exponential ergodicity of the Langevin diffusion. Furthermore, their bound is expressed in terms of $A(\eta, x) = \sup_{k\ge0}\mathbb{E}\|\nabla f(\hat{X}_{k\eta})\|_2^2$, which can also grow with the time horizon T; in contrast, our bound provides explicit upper bound on such quantities when the potential function is uniformly bounded from below.

2.2. Results on mixing and ergodicity

We now discuss some implications of our main results for the analysis of MCMC algorithms. We first present mixing time bounds under both strong assumptions in Theorem 1 and weak assumptions in Theorem 2. Then, an MSE bound is established for the empirical averages of the discretized trajectory.

When the problem of sampling from a target distribution $\gamma(x) \propto e^{-U(x)}$ is considered, the above bounds applied to the drift term $b(x) = -\frac{1}{2}\nabla U(x)$ yield bounds in TV distance, more precisely via the convergence of the Fokker-Planck equation and the

Pinsker inequality [15]. Instead, in this paper, so as to obtain a sharper result, we directly combine the proofs of Theorem 1 with the analysis of Cheng et al. [12]. A notable feature of this strategy is that it completely decouples analyses of the discretization error and of the convergence of the continuous-time diffusion process. The convergence of the continuous-time process is guaranteed when the target distribution satisfies a log-Sobolev inequality [52, 40].

We now state a guarantee for the Euler scheme (1.2), also known as Unadjusted Langevin Algorithm (ULA). Note that compared to Theorem 1, the initialization requirement (2.4) is relaxed in such case, and the bound only requires a moment assumption on the initial distribution.

Theorem 3. *Consider a density of the form $\gamma(x) \propto \exp(-U(x))$ such that:*

(a) The gradient ∇U satisfies Assumptions 2.1—2.3.
(b) The distribution defined by γ satisfies a log-Sobolev inequality with constant $\rho > 0$.

Then for the Gaussian initialization $\mathcal{N}(0, L_1^{-1} I_d)$ and any stepsize parameter $\eta \in \left(0, \frac{1}{4L_1}\right)$, the KL divergence is bounded as

$$D_{KL}(\hat{\pi}_t \| \gamma) \lesssim e^{-\rho t/8} D_{KL}(\hat{\pi}_0 \| \gamma) + \frac{1}{\rho}\left(\eta^2 L_1^4 R_0^2 + \eta^2 L_1^3 d + \eta^4 L_2^2 L_1^4 R_0^4 + \eta^2 L_2^2 d^2\right), \quad (2.7)$$

where $R_0 := \sigma_0 \sqrt{d} + \sqrt{\frac{d+\beta}{\mu}} + \frac{A_0}{L_1}$.

See Section D.1 of the supplementary material for the proof.

The KL bound (2.7) has a number of consequences for mixing times. Given an error tolerance $\varepsilon > 0$ and a distance function dist, we define the associated *mixing time*

$$N(\varepsilon, \text{dist}) := \arg \min_{k=1,2,\ldots} \{\text{dist}(\hat{\pi}_{k\eta}, \gamma) \leq \epsilon\}. \quad (2.8)$$

Under the canonical setup (smoothness parameters independent of the dimension) and the initial distribution $\mathcal{N}(0, I_d)$, for any $\varepsilon > 0$, we can show that the unadjusted Langevin algorithm (1.2) with drift $b = -\frac{1}{2}\nabla U$ and step size $\eta = \frac{\sqrt{\varepsilon\rho}}{d}(\log \frac{1}{\rho})^{-1}$ has mixing times bounded as:

$$\begin{cases} N(\varepsilon, D_{KL}) = \tilde{O}\left(d\, \varepsilon^{-\frac{1}{2}} \rho^{-\frac{3}{2}}\right) & \text{in KL-divergence,} \\ N(\varepsilon, TV) = \tilde{O}\left(d\, \varepsilon^{-1} \rho^{-\frac{3}{2}}\right) & \text{in TV distance,} \\ N(\varepsilon, \mathcal{W}_2) = \tilde{O}\left(d\, \varepsilon^{-1} \rho^{-\frac{5}{2}}\right) & \text{in } \mathcal{W}_2 \text{ distance,} \\ N(\varepsilon, \mathcal{W}_1) = \tilde{O}\left(d^{\frac{3}{2}}\, \varepsilon^{-1} \rho^{-\frac{3}{2}}\right) & \text{in } \mathcal{W}_1 \text{ distance.} \end{cases}$$

See Section D.1 of the supplementary material for the steps leading from the KL bound (2.7) to these mixing time bounds.

It is worth noting that the set of distributions satisfying a log-Sobolev inequality [28] includes strongly log-concave distributions [3] as well as perturbations thereof [30]. For

example, it includes distributions that are strongly log-concave outside of a bounded region, but non-log-concave inside of it, as analyzed in some recent work [37]. Under the additional smoothness Assumption 2.2, we obtain an improved mixing time of $O(d/\varepsilon)$ in total variation distance, compared to the $O(d/\varepsilon^2)$ bound in the paper [37]. On the other hand, we obtain the same mixing time in $\mathcal{W}_2$ distance as the papers [20, 16] but under weaker assumptions on the target distribution—namely, those that satisfy a log-Sobolev inequality as opposed to being strongly log-concave. See Section A in the supplementary file for more discussion regarding the relationship between tail assumptions and the isoperimetric inequalities.

Similar to Theorem 3, when the potential function U does not satisfy the distant dissipativity assumption, but is uniformly bounded from below, and when the continuous-time dynamics is geometrically ergodic, a non-asymptotic mixing time upper bound on the forward Euler scheme can be then derived by directly combining the continuous-time mixing result with Theorem 2 with the triangle inequality.

Corollary 1. *Given any $\varepsilon > 0$ and initial distribution $\pi_0 = \mathcal{N}(0, I_d)$, suppose that there exists $\tau(\varepsilon) > 0$ such that the continuous-time diffusion (1.1) satisfies the bound*

$$d_{TV}(\pi_{\tau(\varepsilon)}, \gamma) < \varepsilon/2.$$

Assume moreover that the potential function U satisfies the smoothness Assumptions 2.1 and 2.2 with L_1, L_2 of constant order, and $\|\nabla U(0)\|_2 = O(\sqrt{d})$, $U(0) = O(d)$, and that $U(\theta) \geq 0$ for all $\theta \in \mathbb{R}^d$. Then with the step size choice $\eta = \frac{c\varepsilon}{d\sqrt{\tau(\varepsilon)}} \wedge \frac{c\sqrt{\varepsilon}}{\tau(\varepsilon)\sqrt{d}}$ for a constant $c > 0$ depending on smoothness parameters but independent of dimension and ε, we have:

$$N(\varepsilon, TV) = O\left(\frac{d\,\tau^{3/2}(\varepsilon)}{\varepsilon} + \frac{\sqrt{d}\,\tau^2(\varepsilon)}{\sqrt{\varepsilon}}\right).$$

Note that Corollary 1 only requires a total variation distance upper bound on the continuous-time dynamics, and directly translates into a mixing time result for the Euler discretization. To have a better understanding of this corollary, we consider the following three examples:

- For $\gamma \propto e^{-U}$ satisfying the log-Sobolev inequality with constant λ and smoothness conditions (2.1) and (2.2), we have:

$$d_{TV}(\pi_t, \gamma) \leq \sqrt{D_{\mathrm{KL}}(\pi_t \| \gamma)} \leq \sqrt{D_{\mathrm{KL}}(\pi_0 \| \gamma) e^{-\lambda t}} \leq \exp\left(O(\log d) - \frac{\lambda t}{2}\right),$$

  which leads to a bound $\tau(\varepsilon) \leq \frac{1}{\lambda} \log \frac{d}{\varepsilon}$. Plugging into Corollary 1, we obtain:

$$N(\varepsilon, TV) = O\left(\frac{\sqrt{d}}{\lambda^{3/2}\varepsilon}\left(\sqrt{d} + \lambda^{-1/2}\right) \log^2 \frac{d}{\varepsilon}\right).$$

- For $\gamma \propto e^{-U}$ satisfying the Poincaré inequality with constant λ and smoothness assumptions (2.1), (2.2), we have:

$$d_{TV}(\pi_t, \gamma) \leq \sqrt{\chi^2(\pi_t||\gamma)} \leq \sqrt{\chi^2(\pi_0||\gamma)e^{-\lambda t}} \leq \exp\left(O(d) - \frac{\lambda t}{2}\right),$$

  which leads to a bound $\tau(\varepsilon) \leq \frac{d}{\lambda} \log \frac{1}{\varepsilon}$. Plugging into Corollary 1, we obtain:

$$N(\varepsilon, TV) = O\left(\frac{d^{5/2}}{\lambda^2 \varepsilon} \log^2 \frac{1}{\varepsilon}\right).$$

- For heavy-tailed potential functions satisfying the Veretennikov condition:

$$\langle \nabla U(x), \tfrac{x}{\|x\|_2} \rangle \geq r \quad \text{for all } x \text{ with } \|x\|_2 \geq M_0 \text{ for some } r > d/2,$$

  the Langevin process has mixing time bounded as $\tau(\varepsilon) = O(\varepsilon^{-\frac{1}{r-d/2}})$ in TV distance [55]. When used in conjunction with Corollary 1, this leads to the bound:

$$N(\varepsilon, TV) = O\left(\varepsilon^{-\frac{4}{2r-d}-\frac{1}{2}} + \varepsilon^{-\frac{3}{2r-d}-1}\right).$$

  Here we have suppressed the problem-dependent pre-factors within the $O(\cdot)$-notation so as to focus on the exponent in $(1/\varepsilon)$ in the convergence rate.

There are also non-asymptotic results for the mixing time of ULA under the Poincaré inequality or related conditions. For instance, Vempala and Wibisono [54] give Rényi divergence upper bounds under the Poincaré inequality. The expression of their bound depends on the Rényi divergence between the stationary distribution of the discrete-time process and the target density. In comparison, our bound gives an explicit dependence on all problem parameters. Dalalyan et al. [18] prove upper bounds on the mixing time in the non-strongly log-concave setting. Log-concavity is known to imply the Poincaré inequality [2]. However, our bounds are expressed in terms of the Poincaré constant, and the problem-dependent constants are not directly comparable. In terms of dependency on the target accuracy ε, Corollary 1 implies an $\tilde{O}(\varepsilon^{-1})$ mixing time bound in TV distance, which improves upon the $O(\varepsilon^{-4})$ bound given in the paper [18].

MCMC is widely used to compute the expectation of a function under the target distribution. In many applications, one can take the empirical average of the value of a given test function h over the trajectory of the Euler scheme:

$$\hat{h}_N := \frac{1}{N} \sum_{k=N_0+1}^{N_0+N} h(\hat{X}_{k\eta}), \tag{2.9}$$

where N is the number of samples taken, and N_0 is a burn-in time. Although the empirical measure supported on N points inevitably has large Wasserstein distance to the target measure, a small error between $\hat{h}_N$ and the expectation $\mathbb{E}_\gamma h(X)$ under the target measure can be obtained. In the following, we show a corollary to Theorem 1 that guarantees the MSE bound for bounded and Lipschitz test functions.

Corollary 2. *Consider a density $\gamma \propto \exp(-U(x))$ such that ∇U satisfies Assumption 2.1, 2.2 and 2.3. Suppose furthermore that the density γ satisfies the log-Sobolev inequality with constant ρ. For a given number of iterations $N > 0$ and burn-in time N_0, define $\hat{h}_N$ as in equation (2.9). Under the canonical setup, we have:*

- *If h is a bounded function, taking step size $\eta \asymp \left(\frac{\log Nd/\rho}{Nd^2}\right)^{\frac{1}{3}}$ and burn-in time $N_0 \asymp \frac{1}{\eta\rho}\log\frac{Nd}{\rho}$, we obtain:*

$$\mathbb{E}\left(\hat{h}_N - \mathbb{E}_\gamma h(X)\right)^2 \leq C\frac{\|h\|_\infty^2}{\rho}\left(\frac{d}{N}\log\frac{Nd}{\rho}\right)^{\frac{2}{3}}, \text{ and}$$

- *If h is a Lipschitz function, taking step size $\eta \asymp \left(\frac{\rho}{Nd}\log\frac{d}{\eta\rho}\right)^{1/3}$ and burn-in time $N_0 \asymp \frac{1}{\eta\rho}\log\frac{Nd}{\rho}$, we obtain:*

$$\mathbb{E}\left(\hat{h}_N - \mathbb{E}_\gamma h(X)\right)^2 \leq C\|h\|_{Lip}^2\left(\frac{d^2}{N\rho^2}\log\frac{Nd}{\rho}\right)^{\frac{2}{3}},$$

where C is a problem-dependent constant independent of (N, η, d, ρ).

Corollary 2 matches the $O(N^{-2/3})$ MSE bound from the paper [21], a rate which is standard in the MCMC literature. However, the result given here holds under milder assumptions, allowing for non-log-concave and multi-modal distributions. We also present the explicit dependency of the MSE bound on problem-dependent parameters (d, ρ). It is also worthwhile comparing to the classical MSE bound [42] obtained using Poisson equation methods. In comparison, our bound requires fewer smoothness assumptions on both the test function h and the potential function U. Notably, our result applies to discontinuous test functions h such as indicator functions, which play an important role in the construction of Bayesian credible sets.

2.3. Overview of proof

In this section, we provide a high-level overview of the three main steps that comprise the proof of Theorem 1; the subsequent Sections 3, 4, and 5 are devoted (respectively) to the details of these three steps.

Step 1: Our first step is to construct a continuous-time interpolation $\{\hat{X}_t\}_{t\geq 0}$ of the discrete-time process $\{\hat{X}_{k\eta}\}_{k=0}^{\infty}$, and prove that its density $\hat{\pi}_t$ satisfies an analogue of the Fokker-Plank equation (see Lemma 1). The elliptic operator of this equation is time-dependent, with a drift term $\hat{b}_t = \mathbb{E}(b(\hat{X}_{k\eta})|\hat{X}_t = x)$ given by the backward conditional expectation of the original drift term b. By direct calculation, the time derivative of the KL divergence between the interpolated and the original Langevin diffusion $D_{\mathrm{KL}}(\hat{\pi}_t\|\pi_t)$

is controlled by the mean squared difference between the drift terms of the Fokker-Planck equations for the original and the interpolated processes, namely the quantity

$$\int \hat{\pi}_t(x)\|b(x) - \hat{b}_t(x)\|_2^2 dx. \tag{2.10}$$

See Lemma 2 for details.

Step 2: Our next step is to control the mean-squared error term (2.10). Compared to the MSE bound obtained from the Girsanov theorem in past work [15], our bound has an additional backward conditional expectation inside the norm. Directly pulling this latter outside the norm by convexity inevitably entails a KL bound $O(\eta)$ due to fluctuations of the Brownian motion. However, taking the backward expectation cancels out most of the noises, as long as the density function of the iterate at each step is smooth enough. This geometric intuition is explained precisely in Section 4.1, and concretely implemented in Section 4.2. The main conclusion from Steps 1 and 2 is summarized as follows:

Proposition 1. *Under Assumptions 2.1 and 2.2, for any $t \in [k\eta, (k+1)\eta]$, we have*

$$\begin{aligned}\frac{d}{dt} D_{KL}(\hat{\pi}_t\|\pi_t) \le 4L_1^2(t-k\eta)^2\mathcal{I}(\hat{\pi}_{k\eta}) + 16L_1^2(t-k\eta)^2\mathbb{E}\left\|b(\hat{X}_{k\eta})\right\|_2^2 + 12L_1^4(t-k\eta)^3 d \\ + 16(t-k\eta)^4 L_2^2\mathbb{E}\left\|b(\hat{X}_{k\eta})\right\|_2^4 + 48(t-k\eta)^2 L_2^2 d^2, \end{aligned} \tag{2.11a}$$

and the moments of drift terms can be further bounded as

$$\mathbb{E}\left\|b(\hat{X}_{k\eta})\right\|_2^2 \le 2A_0^2 + 2L_1^2\mathbb{E}\left\|\hat{X}_{k\eta}\right\|_2^2, \quad \text{and} \quad \mathbb{E}\left\|b(\hat{X}_{k\eta})\right\|_2^4 \le 4A_0^4 + 4L_1^4\mathbb{E}\left\|\hat{X}_{k\eta}\right\|_2^4. \tag{2.11b}$$

See Section 4 for the proof of this claim.

Step 3: The third step is to bound the Fisher information term $\mathcal{I}(\hat{\pi}_{k\eta})$. In order to obtain bound the Fisher information for the density at the discrete time steps, we iteratively apply Stam's convolution inequality for Fisher information [49]. The update rule (1.2) can be viewed as a deterministic update combined with convolution with a Gaussian kernel. The Fisher information can increase under the first step and will decrease under the second step. We can then prove the recursive inequality characterizing the effect of both steps, leading to the following proposition:

Proposition 2. *Under Assumptions 2.1, 2.2 and 2.4, for $T = N\eta$ and $N \in \mathbb{N}_+$, there exists a universal constant $c > 0$, such that the following bound holds true:*

$$\frac{1}{N}\sum_{k=1}^{N}\mathcal{I}(\hat{\pi}_{k\eta}) \le c \cdot \left[\min\left(\frac{d\log T/\eta}{T}, \mathcal{I}(\pi_0)\right) + \frac{L_2^2 d^2}{L_1^2} + L_1 d\right].$$

See Section 5.1 for the proof of this claim.

Note that the bound deals with two cases: when the initial Fisher information is small, the upper bound is independent of the step size η; when the initial Fisher information can be unbounded, we still have a bound with logarithmic dependency on the stepsize.

It remains to bound the moments of $\hat{X}_t$ along the path. In Proposition 1, the second and fourth moment of $\hat{X}_t$ are used. With different assumptions on the drift term, different moments bounds can be established, leading to Theorem 1 and Theorem 2, respectively.

- Under distant dissipativity (Assumption 2.3), the p-th moment of this process can be bounded from above, for arbitrary value of $p > 1$. (see Lemma 8). The proof is based on the Burkholder-Davis-Gundy inequality for continuous martingales. Collecting these results yields equation (2.5), which completes our sketch of the proof of Theorem 1.
- Without Assumption 2.3, if the drift term takes the form $b = -\nabla f$ for some non-negative function f, then the second and fourth moments can still be bounded (see Lemma 9 and 10). Collecting these results yields equation (2.6), and completes our sketch of the proof of Theorem 2.

3. Interpolation, KL Bounds and Fokker-Planck

As in the analysis of Dalalyan [15], the first step of the proof is to construct a continuous-time interpolation for the discrete-time algorithm (1.2). In particular, we define a stochastic process over the interval $t \in [k\eta, (k+1)\eta]$ via

$$\hat{X}_t := \hat{X}_{k\eta} + \int_{k\eta}^{t} b(\hat{X}_{k\eta})ds + \int_{k\eta}^{t} d\hat{B}_s. \tag{3.1}$$

Let $\{\hat{\mathcal{F}}_t \mid t \geq 0\}$ be the natural filtration associated with the Brownian motion $\{\hat{B}_t \mid t \geq 0\}$. Conditionally on $\hat{\mathcal{F}}_{k\eta}$, the process $\{(\hat{X}_t|\hat{\mathcal{F}}_{k\eta}) \mid t \in [k\eta, (k+1)\eta]\}$ is a Brownian motion with constant drift $b(\hat{X}_{k\eta})$ and starting at $\hat{X}_{k\eta}$. This interpolation has been used in past work [15, 12]. In their work, the KL divergence between the law of processes $\{X_t \mid t \in [0,T]\}$ and $\{\hat{X}_t \mid t \in [0,T]\}$ is controlled, via a use of the Girsanov theorem, by bounding Radon-Nikodym derivatives. This approach requires controlling the quantity $\mathbb{E}\|b(\hat{X}_t) - b(\hat{X}_{k\eta})\|_2^2$ for $t \in [k\eta, (k+1)\eta]$. It should be noted that it scales as $O(\eta)$, due to the scale of oscillations in Brownian motion.

In our approach, we overcome this difficulty by only considering the KL divergence of the one-time marginal laws $D_{\mathrm{KL}}(\mathcal{L}(\hat{X}_T)\|\mathcal{L}(X_T))$. Let us denote the densities of X_t and $\hat{X}_t$ with respect to Lebesgue measure in $\mathbb{R}^d$ by π_t and $\hat{\pi}_t$, respectively. It is well-known that when b is Lipschitz, then the density π_t satisfies the Fokker-Planck equation

$$\frac{\partial \pi_t}{\partial t} = -\nabla \cdot (\pi_t b) + \frac{1}{2}\Delta \pi_t, \tag{3.2}$$

where Δ denotes the Laplacian operator. On the other hand, the interpolated process $\hat{X}_{k\eta}$ is not Markovian, and so does not have a semigroup generator. For this reason, it is difficult to directly control the KL divergence between it and the original Langevin diffusion. In the following lemma, we construct a different partial differential equation that is satisfied by $\hat{\pi}_t$.

Lemma 1. *The density $\hat{\pi}_t$ of the process $\hat{X}_t$ defined in equation (3.1) satisfies the PDE*

$$\frac{\partial \hat{\pi}_t}{\partial t} = -\nabla \cdot \left(\hat{\pi}_t \hat{b}_t\right) + \tfrac{1}{2}\Delta \hat{\pi}_t \qquad \text{over the interval } t \in [k\eta, (k+1)\eta], \tag{3.3}$$

where $\hat{b}_t(x) := \mathbb{E}\left(b(\hat{X}_{k\eta}) \middle| \hat{X}_t = x\right)$ is a time-varying drift term.

See Section 3.1 for the proof of this lemma.

The key observation is that, conditioned on the σ-field $\hat{\mathcal{F}}_{k\eta} = \sigma(\hat{X}_t : 0 \le t \le k\eta)$, the process $\left\{(\hat{X}_t \mid \hat{\mathcal{F}}_{k\eta}) \mid t \in [\eta, (k+1)\eta]\right\}$ is a Brownian motion with constant drift, whose conditional density $\hat{\pi}_t \mid \hat{\mathcal{F}}_{k\eta}$ satisfies a Fokker-Planck equation. Taking the expectation on both sides, and interchanging the integral with the derivatives, we obtain the Fokker-Planck equation for the density $\hat{\pi}_t$ unconditionally.

In Lemma 1, we have a Fokker-Planck equation with time-varying coefficients; it is satisfied by the one-time marginal densities of the continuous-time interpolation for the process (1.2). This representation provides convenient tool for bounding the time derivative of KL divergence, a task to which we turn in the next section.

3.1. Proof of Lemma 1

We first consider the conditional distribution of $(\hat{X}_t : k\eta \le t \le (k+1)\eta)$, conditioned on $\hat{\mathcal{F}}_{k\eta}$. At time $t = k\eta$, it starts with an atomic mass (viewed as Dirac δ-function at point $\hat{X}_{k\eta}$, which is a member of the tempered distribution space $\mathcal{S}'$ [see, e.g., 48]. Its derivatives and Hessian are well-defined as well.) For $t > k\eta$, this conditional density follows the Fokker-Planck equation for a Brownian motion with constant drift:

$$\frac{\partial\left(\hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}\right)}{\partial t} = -\nabla \cdot \left(\hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}} b(\hat{X}_{k\eta})\right) + \frac{1}{2}\Delta \hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}, \tag{3.4}$$

where the partial derivatives are in terms of the dummy variable x. Next we take expectations of both sides of (3.4). By interchanging derivative and integration, we obtain the following identities (with rigorous justification of these steps provided below):

$$\mathbb{E}\left(\frac{\partial \hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}}{\partial t}(x)\right) = \frac{\partial \hat{\pi}_t}{\partial t}(x) \tag{3.5a}$$

$$\mathbb{E}\left(\nabla\left(\hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}(x) b(\hat{X}_{k\eta})\right)\right) = \nabla \cdot \left(\hat{\pi}_t(x)\mathbb{E}\left(b(\hat{X}_{k\eta})\middle|\hat{X}_t = x\right)\right) \tag{3.5b}$$

$$\mathbb{E}\left(\Delta \hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}\right) = \Delta \hat{\pi}_t. \tag{3.5c}$$

Proof of equation (3.5a): We show:

$$\mathbb{E}\left(\frac{\partial\hat{\pi}_t\,|_{\hat{\mathcal{F}}_{k\eta}}}{\partial t}(x)\right)=\int_{\mathbb{R}^d}\hat{\pi}_{k\eta}(y)\frac{\partial\hat{\pi}_t\,|_{\hat{\mathcal{F}}_{k\eta}}}{\partial t}(x|y)dy\overset{(i)}{=}\frac{\partial}{\partial t}\int_{\mathbb{R}^d}\hat{\pi}_{k\eta}(y)\hat{\pi}_t\,|_{\hat{\mathcal{F}}_{k\eta}}(x|y)dy=\frac{\partial\hat{\pi}_t}{\partial t}(x),$$

Applying Lemma 1 in Section E of the supplementary material, we can show that the density $\hat{\pi}_{k\eta}$ has a tail decaying as $Ce^{-r\|y\|^2}$. We then note that $\frac{\partial\hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}}{\partial t}(x|y)$ is equal to the semigroup generator of the conditional Brownian motion with constant drift, which also decays exponentially with y, in a small neighborhood of t, for fixed x. So the quantity $\hat{\pi}_{k\eta}(y)\frac{\partial\hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}}{\partial t}(x|y)$ has a dominating function of the form of $C(1+\|y\|)e^{-r\|y\|^2}$ in a small neighborhood of t. Combining with the dominated convergence theorem justifies step (i).

Proof of equation (3.5b): We have:

$$\begin{aligned}\mathbb{E}\left(\nabla\left(\hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}(x)b(\hat{X}_{k\eta})\right)\right)&=\int_{\mathbb{R}^d}\hat{\pi}_{k\eta}(y)\nabla_x\cdot\left(\hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}(x|y)b(y)\right)dy\\&\overset{(i)}{=}\nabla_x\cdot\int_{\mathbb{R}^d}\hat{\pi}_{k\eta}(y)\hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}(x|y)b(y)dy\\&\overset{(ii)}{=}\nabla\cdot\left(\hat{\pi}_t(x)\mathbb{E}\left(b(\hat{X}_{k\eta})\big|\hat{X}_t=x\right)\right).\end{aligned}$$

In order to justify step (i), we first note that, according to Assumption 2.1, both of the functions $y\mapsto b(y)$ and $y\mapsto\nabla_x\log\hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}(x|y)$ grow at most linearly in y, for fixed t. By the rapid decay of the tail of $\hat{\pi}_t$ shown in Lemma 1, and the decay of the tail of $\hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}(x|y)$ obtained by elementary results on the Gaussian density, we have a dominating function of the form of $C(1+\|y\|^2)e^{-r\|y\|^2}$. This justifies step (i) by the dominated convergence theorem. Step (ii) simply follows from the Bayes rule.

Proof of equation(3.5c): We similarly have:

$$\mathbb{E}\left(\Delta\hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}(x)\right)=\int_{\mathbb{R}^d}\Delta_x\left(\hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}(x|y)\right)\hat{\pi}_{k\eta}(y)dy.$$

Note that $\Delta p(x)=(\Delta\log p+\|\nabla\log p\|^2)p$ for any density function p. Since $\log\hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}(x|y)$ is a quadratic function in the variable x, its gradient is linear (it also grows at most linearly with $\|y\|$), and its Laplacian is constant. Therefore, we have a dominating function of form $C(1+\|y\|^2)e^{-r\|y\|^2}$ for the integrand, which guarantees the interchange between the integral and the Laplacian operator. This leads to $\mathbb{E}\left(\Delta\hat{\pi}_t|_{\hat{\mathcal{F}}_{k\eta}}(x)\right)=\Delta\hat{\pi}_t(x)$.

Combining these identities yields

$$\frac{\partial\hat{\pi}_t}{\partial t}(x)=-\nabla\cdot\left(\hat{\pi}_t(x)\hat{b}_t(x)\right)+\frac{1}{2}\Delta\hat{\pi}_t,\quad t\in[k\eta,(k+1)\eta],$$

where $\hat{b}_t(x)=\mathbb{E}\left(b(\hat{X}_{k\eta})\big|\hat{X}_t=x\right)$ for $t\in[k\eta,(k+1)\eta]$.

4. Controlling the KL divergence: Proof of Proposition 1

We now turn to the proof of Proposition 1, which involves bounding the derivative $\frac{d}{dt}D_{\mathrm{KL}}(\hat{\pi}_t\|\pi_t))$. We first compute the derivative using the Fokker-Planck equation established in Lemma 1, and then upper bound it by a regularity estimate of the density $\hat{\pi}_{k\eta}$ and moment bounds on $\hat{X}_{k\eta}$. The key geometric intuition underlying our argument is the following: if the drift b is second-order smooth and the initial distribution at each step is also smooth, most of the Gaussian noise is cancelled out, and only higher-order terms remain. This intuition is fleshed out in Section 4.1.

In the following lemma, we give an explicit upper bound on the KL divergence between the one-time marginal distributions of the interpolated process and the original diffusion, based on Fokker-Planck equations derived above.

Lemma 2. *For any pair of densities π and $\hat{\pi}$ satisfying the Fokker-Planck equations (3.2) and (3.3), respectively, we have*

$$\frac{d}{dt}D_{KL}(\hat{\pi}_t\|\pi_t) \leq \frac{1}{2}\int_{\mathbb{R}^d}\hat{\pi}_t(x)\|\hat{b}_t(x)-b(x)\|_2^2dx. \tag{4.1}$$

See Section B.1 of the supplementary material for the proof of this claim.

It is worth noting the key difference between our approach and the method of Dalalyan [15], which is based on the Girsanov theorem. His analysis controls the KL divergence via the quantity $\int_0^T\mathbb{E}\|b(\hat{X}_{k\eta})-b(\hat{X}_t)\|_2^2dt$, a term which scales as $O(\eta)$ even for the simple case of the Ornstein-Uhlenbeck process. Observe that the Brownian motion contributes to an $O(\eta)$ oscillation in $\|\hat{X}_{k\eta}-\hat{X}_t\|_2^2$, dominating other lower-order terms. By contrast, we control the KL divergence using the quantity $\int_0^T\mathbb{E}\|\hat{b}_t(\hat{X}_t)-b(\hat{X}_t)\|_2^2dt$. Observe that $\hat{b}_t$ is exactly the backward conditional expectation of $b(\hat{X}_{k\eta})$ conditioned on the value of $\hat{X}_t$. Having the conditional expectation inside the norm (rather than outside) leads to cancellation of the lower-order oscillations.

In the remainder of this section, we focus on bounding the integral on the right-hand side of equation (4.1). Since the difference between $\hat{X}_{k\eta}$ and $\hat{X}_t$ comes mostly from an isotropic noise, we may expect it to mostly cancel out. In order to exploit this intuition, we use the third-order smoothness condition (see Assumption 2.2) so as to perform the Taylor expansion

$$\hat{b}_t(x)-b(x)=\mathbb{E}\left(b(\hat{X}_{k\eta})-b(\hat{X}_t)\big|\hat{X}_t=x\right)=\nabla b(x)\mathbb{E}\left(\hat{X}_{k\eta}-\hat{X}_t\big|\hat{X}_t=x\right)+\hat{r}_t(x), \tag{4.2}$$

where the remainder takes the form

$$\hat{r}_t(x):=\mathbb{E}\left(\int_0^1 s\nabla^2 b((1-s)\hat{X}_t+s\hat{X}_{k\eta})[\hat{X}_{k\eta}-\hat{X}_t,\hat{X}_{k\eta}-\hat{X}_t]ds\Big|\hat{X}_t=x\right).$$

The reminder is relatively easy to control, since it contains a $\|\hat{X}_{k\eta} - \hat{X}_t\|_2^2$ factor, which is already of order $O(\eta)$. We summarize as follows:

Lemma 3. *Under Assumptions 2.1 and 2.2, we have*

$$\mathbb{E}\|\hat{r}_t(\hat{X}_t)\|_2^2 \le 8(t-k\eta)^4 L_2^2 \left(A_0^4 + L_1^4 \mathbb{E}\|\hat{X}_{k\eta}\|_2^4\right) + 24(t-k\eta)^2 L_2^2 d^2$$

for any $t \in [k\eta, (k+1)\eta)$.

See Section B.2 of the supplementary material for the proof of this claim.

It remains to control the first order term. From Assumption 2.1, the Jacobian norm $\|\nabla b(x)\|_{\mathrm{op}}$ is at most L_1; accordingly, we only need to control the norm of the vector $\mathbb{E}\left(\hat{X}_{k\eta} - \hat{X}_t \middle| \hat{X}_t = x\right)$. It corresponds to the difference between the best prediction about the past of the path and the current point, given the current information. Herein lies the main technical challenge in the proof of Proposition 1, apart from the construction of the Fokker-Planck equation for the interpolated process. Before entering the technical details, let us first provide some geometric intuition for the argument.

4.1. Geometric intuition

Suppose that we were dealing with the conditional expectation of the continuous process $\hat{X}_t$, conditioned on $\hat{X}_{k\eta}$; in this case, the Gaussian noise would completely cancel out (see equation (3.1)). However, we are indeed reasoning backward, and $\hat{X}_t$ itself depends on the the Gaussian noise $\int_{k\eta}^t d\hat{B}_s$ added to this process. It is unclear whether the cancellation occurs when computing $\mathbb{E}\left(\hat{X}_{k\eta} \middle| \hat{X}_t\right) - \hat{X}_t$. In fact, it occurs only under particular situations, which turn out to be typical for the discretized process.

Due to the dependence between $\hat{X}_t$ and Gaussian noise, we cannot expect cancellation to occur in general. Figure 1(a) illustrates an extremal case, where the initial distribution at time $k\eta$ is an atomic mass. When we condition on the value at $\hat{X}_t$ as well, the process behaves like a Brownian bridge. Consequently, it makes no difference whether the conditional expectation is inside or outside the norm: in either case, there is a term of the form $\|\hat{X}_{k\eta} - \hat{X}_t\|_2$, which scales as $O(\sqrt{\eta})$.

On the other hand, as illustrated in Figure 1(b), if the initial distribution is uniform over some region, the initial point is almost equally likely to be from anywhere around $\hat{X}_t$, up to the drift term, and most of the noise gets cancelled out. In general, if the initial distribution is smooth, locally it looks almost uniform, and similar phenomena should also hold true. Thus we expect $\mathbb{E}(\hat{X}_{k\eta}|\hat{X}_t) - \hat{X}_t$ to be decomposed into terms coming from the drift and terms coming from the smoothness of the initial distribution.

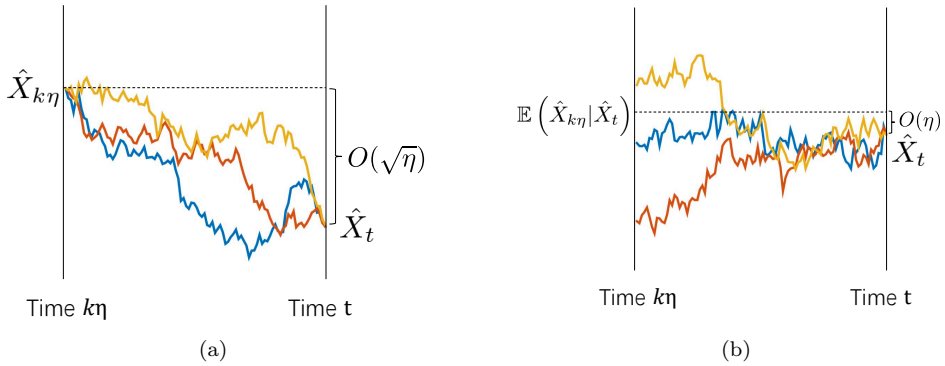

Figure 1. Different cases of the backward conditional expectation $\mathbb{E}(\hat{X}_{k\eta} \mid \hat{X}_t)$ for $t \in [k\eta, (k+1)\eta)$. In order to generate the plots, we take the drift term $b = 0$, and simulate the Brownian motion path within time period $[k\eta, t]$ conditionally on the end point $\hat{X}_t$. (a) In one extreme case, if $\hat{X}_{k\eta}$ is fixed, the difference $|\mathbb{E}(\hat{X}_{k\eta}|\hat{X}_t) - \hat{X}_t|$ is of order $O(\sqrt{\eta})$ with high probability. (b) In another extreme case, if $\hat{X}_{k\eta} \sim \mathcal{N}(0,1)$, which is smooth enough, we can compute the backward conditional expectation in closed form, which leads to $|\mathbb{E}(\hat{X}_{k\eta} \mid \hat{X}_t) - \hat{X}_t| = O(\eta)$ with high probability.

4.2. Upper bound via integration by parts

With this intuition in hand, we now turn to the proof itself. In order to leverage the smoothness of the initial distribution, we use integration by parts to move the derivatives onto the density of $\hat{X}_{k\eta}$. From Bayes' formula, we have

$$\mathbb{E}\left(\hat{X}_{k\eta} - \hat{X}_t \middle| \hat{X}_t = x\right) = \int (y-x) p(\hat{X}_{k\eta} = y | \hat{X}_t = x) dy = \int (y-x) \frac{\hat{\pi}_{k\eta}(y) p(\hat{X}_t = x | \hat{X}_{k\eta} = y)}{\hat{\pi}_t(x)} dy. \tag{4.3}$$

Since the conditional density $p(\hat{X}_t = x \mid \hat{X}_{k\eta} = y)$ is a Gaussian centered at $y-(t-k\eta)b(y)$ with fixed covariance, the gradient with respect to y is the density itself times a linear factor $x-y+(t-k\eta)b(y)$, with an additional factor depending on the Jacobian of b. This elementary fact motivates a decomposition whose goal is to express $\mathbb{E}\big(\hat{X}_{k\eta} - \hat{X}_t \big| \hat{X}_t = x\big)$ using the conditional expectation of $\nabla \log \hat{\pi}_{k\eta}$ and some other terms which are easy to control. More precisely, in order to expose a gradient of the Gaussian density, we decompose the difference $y-x$ into three parts, namely $y-x = a_1(x,y) - a_2(x,y) - a_3(x,y)$, where

$$\begin{aligned} a_1(x,y) &:= (I + (t-k\eta)\nabla b(y))(y - x + (t-k\eta)b(y))\,, \\ a_2(x,y) &:= (t-k\eta)\nabla b(y)(y - x + (t-k\eta)b(y)), \quad \text{and} \\ a_3(x,y) &:= (t-k\eta)b(y). \end{aligned}$$

We define the conditional expectations $I_i(x) := \mathbb{E}\big(a_i(\hat{X}_{k\eta}, \hat{X}_t)\big|\hat{X}_t = x\big)$ for $i = 1, 2, 3$ and control the three terms separately.

Let us denote by φ the d-dimensional standard Gaussian density. The first term I_1 can directly be expressed in terms of the gradient of φ:

$$\begin{aligned} I_1(x) &= \int (I + (t-k\eta)\nabla b(y))(y - x + (t-k\eta)b(y))\varphi\left(\frac{x - y - (t-k\eta)b(y)}{\sqrt{t-k\eta}}\right)\frac{\hat{\pi}_{k\eta}(y)}{\hat{\pi}_t(x)}dy \\ &= (t-k\eta)\int \nabla_y\varphi\left(\frac{x - y - (t-k\eta)b(y)}{\sqrt{t-k\eta}}\right)\frac{\hat{\pi}_{k\eta}(y)}{\hat{\pi}_t(x)}dy, \end{aligned}$$

where we used the chain rule and $\nabla\varphi(y) = -y\varphi(y)$. Thus, applying integration by parts, we write I_1 in a revised form.

Lemma 4. *For all $t \in [k\eta, (k+1)\eta]$, we have $I_1(x) = -(t-k\eta)\mathbb{E}\left(\nabla \log \hat{\pi}_{k\eta}(\hat{X}_{k\eta})\big|\hat{X}_t = x\right)$, and consequently,*

$$\mathbb{E}\|I_1(\hat{X}_t)\|_2^2 \le (t-k\eta)^2 \int \hat{\pi}_{k\eta}(x)\|\nabla \log \hat{\pi}_{k\eta}(x)\|_2^2 dx.$$

See Section 4.3 for the proof of this lemma.

It is clear from Lemma 4 that a regularity estimates on the moments of $\nabla \log \hat{\pi}_{k\eta}(\hat{X}_{k\eta})$ gives an $O(\eta^2)$ estimates on the squared integral. Such a bound with reasonable dimension dependence is nontrivial to obtain, and we postpone this argument to Section 5.

On the other hand, the remaining two terms are relatively easy to control, as summarized in the following:

Lemma 5. *Under Assumption 2.1, the following bounds hold for all $t \in [k\eta, (k+1)\eta]$:*

$$\mathbb{E}\|I_2(\hat{X}_t)\|_2^2 \le 3(t-k\eta)^3 L_1^2 d, \quad \text{and} \tag{4.4a}$$

$$\mathbb{E}\|I_3(\hat{X}_t)\|_2^2 \le (t-k\eta)^2\mathbb{E}\left\|b(\hat{X}_{k\eta})\right\|_2^2 \le 2(t-k\eta)^2(A_0^2 + L_1^2\mathbb{E}\|\hat{X}_{k\eta}\|_2^2). \tag{4.4b}$$

See Section 4.4 for the proof of this lemma.

Combining the Taylor expansion (4.2) with the bounds from Lemma 4 and 5 yields the bound claimed in Proposition 1.

4.3. Proof of Lemma 4

In this section, we prove Lemma 4; it controls the dominant term I_1 in the decomposition (4.3) of $\mathbb{E}\left(\hat{X}_{k\eta} - \hat{X}_t\big|\hat{X}_t = x\right)$. Recall the definition

$$I_1(x) = (t-k\eta)\int \nabla_y\varphi\left(\frac{x - y - (t-k\eta)b(y)}{\sqrt{t-k\eta}}\right)\frac{\hat{\pi}_{k\eta}(y)}{\hat{\pi}_t(x)}dy,$$

where φ is the d-dimensional standard Gaussian density. We first note the tail of the Gaussian density is trivial, and the tail of $\hat{\pi}_{k\eta}$ is controlled by the results in Section E of the supplementary material. Therefore, we may apply integration by parts so as to obtain

$$\begin{aligned}
I_1(x) =& \int (I + (t-k\eta)\nabla b(y))(y - x + (t-k\eta)b(y))(2\pi(t-k\eta))^{-\frac{d}{2}} \\
&\qquad \cdot \exp\left(-\frac{1}{2(t-k\eta)}\|x - y - (t-k\eta)b(y)\|_2^2\right)\frac{\hat{\pi}_{k\eta}(y)}{\hat{\pi}_t(x)}dy \\
=& \int (t-k\eta)\nabla_y \exp\left(-\frac{1}{2(t-k\eta)}\|x - y - (t-k\eta)b(y)\|_2^2\right)\frac{\hat{\pi}_{k\eta}(y)}{\hat{\pi}_t(x)}dy \\
=& -(t-k\eta)\int \exp\left(-\frac{1}{2(t-k\eta)}\|x - y - (t-k\eta)b(y)\|_2^2\right)\frac{\nabla_y \hat{\pi}_{k\eta}(y)}{\hat{\pi}_t(x)}dy \\
=& -(t-k\eta)\int \nabla_y \log \hat{\pi}_{k\eta}(y) p(\hat{X}_t = x|\hat{X}_{k\eta} = y)\frac{\hat{\pi}_{k\eta}(y)}{\hat{\pi}_t(x)}dy \\
=& -(t-k\eta)\mathbb{E}\left(\nabla \log \hat{\pi}_{k\eta}(\hat{X}_{k\eta})\big|\hat{X}_t = x\right).
\end{aligned}$$

Applying the Cauchy-Schwartz inequality yields

$$\begin{aligned}
\mathbb{E}\|I_1(\hat{X}_t)\|_2^2 =& (t-k\eta)^2\mathbb{E}\left\|\mathbb{E}\left(\nabla \log \hat{\pi}_{k\eta}(\hat{X}_{k\eta})\big|\hat{X}_t\right)\right\|_2^2 \\
\leq& (t-k\eta)^2\mathbb{E}\|\nabla \log \hat{\pi}_{k\eta}(\hat{X}_{k\eta})\|_2^2 = (t-k\eta)^2\int \hat{\pi}_{k\eta}\|\nabla \log \hat{\pi}_{k\eta}\|_2^2,
\end{aligned}$$

which concludes the proof.

4.4. Proof of Lemma 5

In this section, we prove Lemma 5; it provides bounds on the remaining two terms I_2 and I_3 of the decomposition (4.3) of $\mathbb{E}\left(\hat{X}_{k\eta} - \hat{X}_t\big|\hat{X}_t = x\right)$. We split our proof into two parts, corresponding to the two bounds.

Proof of the bound (4.4a): We directly bound the Jacobian matrix using Assumption 2.1. In particular, we have

$$\begin{aligned}
\frac{\|I_2(x)\|_2}{t-k\eta} =& \left\|\int \nabla b(y)(y - x + (t-k\eta)b(y))(2\pi(t-k\eta))^{-\frac{d}{2}}\exp\left(-\frac{\|x - y - (t-k\eta)b(y)\|_2^2}{2(t-k\eta)}\right)\frac{\hat{\pi}_{k\eta}(y)}{\hat{\pi}_t(x)}dy\right\|_2 \\
\leq& L_1\int \|(y - x + (t-k\eta)b(y)\|_2\frac{\hat{\pi}_{k\eta}(y)}{\hat{\pi}_t(x)}p(\hat{X}_t = x\big|\hat{X}_{k\eta} = y)dy \\
=& L_1\mathbb{E}\left(\|\hat{X}_{k\eta} + (t-k\eta)b(\hat{X}_{k\eta}) - \hat{X}_t\|_2\Big|\hat{X}_t = x\right) \\
=& L_1\mathbb{E}\left(\|\int_{k\eta}^t dB_s\|_2\Big|\hat{X}_t = x\right).
\end{aligned}$$

Plugging into the squared integral yields

$$\mathbb{E}\|I_2(\hat{X}_t)\|_2^2 \le (t-k\eta)^2 L_1^2 \mathbb{E}\left(\mathbb{E}\left(\|\int_{k\eta}^t dB_s\| \Big| \hat{X}_t\right)\right)^2 \le (t-k\eta)^2 L_1^2 \mathbb{E}\|\int_{k\eta}^t dB_s\|_2^2 \le 3(t-k\eta)^3 L_1^2 d.$$

Proof of the bound (4.4b)**:** The size of norm of I_3 is determined largely by $b(\hat{X}_{k\eta})$, which can be controlled using Assumption 2.1:

$$\mathbb{E}\|I_3(\hat{X}_t)\|_2^2 = (t-k\eta)^2 \mathbb{E}\|\mathbb{E}(b(\hat{X}_{k\eta})|\hat{X}_t)\|_2^2 \le (t-k\eta)^2 \mathbb{E}\|b(\hat{X}_{k\eta})\|_2^2 \le 2(t-k\eta)^2(A_0^2 + L_1^2 \mathbb{E}\|\hat{X}_{k\eta}\|_2^2).$$

5. Bounds on the Fisher information and moments

In the previous section, we established upper bounds on the time derivative of the KL divergence between the Langevin diffusion and its Euler discretization; these bounds involve the Fisher information of $\hat{\pi}_{k\eta}$ and the moments of $\hat{X}_{k\eta}$. In order to show that the above estimate is $O(\eta^2)$, we now derive upper bounds on the Fisher information and the moments that are independent of the step size.

Bounding the discretization error essentially relies on a bound on the Fisher information $\mathcal{I}(\hat{\pi}_{k\eta})$, and control of the higher order moments of the process $\{\hat{X}_{k\eta}\}_{k=0}^{\infty}$. In this section, we provide non-asymptotic bounds for both quantities. The regularity estimate is based on a discrete-time argument via Stam's convolution inequality [49]. This proof technique yields bounds with polynomial dependence on the dimension, in sharp contrast to results from classical PDE regularity theory that exhibit exponential dependence. The moment estimate comes from a standard martingale argument, but with explicit dependence on all the parameters.

5.1. Bounding the Fisher information: Proof of Proposition 2

We now turn to the proof of Proposition 2, which gives control of the Fisher information term that appears in the bound from Proposition 1. Our proof is based on Stam's convolution inequality for Fisher information [49]. This inequality guarantees that for any pair of suitably regular probability density functions p, q on $\mathbb{R}^d$, the Fisher information satisifes the inequality

$$\frac{1}{\mathcal{I}(p*q)} \ge \frac{1}{\mathcal{I}(p)} + \frac{1}{\mathcal{I}(q)}, \tag{5.1}$$

where $p*q$ denotes the convolution of p and q. The discrete-time update (1.2) can be seen as a combination of applying the *deterministic mapping* $\phi_\eta(x) := x + \eta b(x)$ with a convolution step with Gaussian kernel. We exploit the inequality (5.1) so as to bound the Fisher information $\mathcal{I}(\hat{\pi}_{(k+1)\eta})$ in terms of $\mathcal{I}(\hat{\pi}_{k\eta})$. In order to do so, we bound the Fisher information for the intermediate density after the first step.

Lemma 6. *For some stepsize $\eta \in (0, \frac{1}{8L_1})$, let $p_k(\cdot)$ be the density of the random variable $Z_k = \phi_\eta(\hat{X}_{k\eta})$ obtained by applying the deterministic mapping ϕ_η. Then under Assumption 2.1, we have the bound*

$$\int p_k(z)\|\nabla_z \log p_k(z)\|_2^2 dz \le (1+4\eta L_1)\int \hat{\pi}_{k\eta}(x)\|\nabla_x \log \hat{\pi}_{k\eta}(x)\|_2^2 dx + 16\frac{\eta L_2^2 d^2}{L_1}.$$

See Section C.1 of the supplementary material for the proof of this lemma.

Let q_η denote the d-dimensional Gaussian distribution $\mathcal{N}(0, \eta I_d)$. Clearly we have the identity $\mathcal{I}(q_\eta) = \frac{d}{\eta}$. By the update rule (1.2), we have that $\hat{\pi}_{(k+1)\eta} = p_k * q_\eta$, for the density p_k defined in Lemma 6. Invoking the convolution inequality (5.1), for $\eta < \frac{1}{8L_1}$, we have the bound

$$\frac{1}{\mathcal{I}(\hat{\pi}_{(k+1)\eta})} \ge \frac{1}{\mathcal{I}(p_k)} + \frac{1}{\mathcal{I}(q_\eta)} \ge \frac{1}{(1+4\eta L_1)\mathcal{I}(\hat{\pi}_{k\eta}) + 16\eta L_2^2 d^2/L_1} + \frac{\eta}{d}. \tag{5.2}$$

Applying equation (5.1) to the initial distribution yields the bound $\mathcal{I}(\hat{\pi}_\eta) \le \frac{d}{\eta}$.

Now we solve the recursion (5.2). First, we note that if $\mathcal{I}(\hat{\pi}_{k\eta}) > \frac{16L_2^2 d^2}{L_1^2}$, then the recursion becomes

$$\frac{1}{\mathcal{I}(\hat{\pi}_{(k+1)\eta})} \ge \frac{1}{(1+5\eta L_1)\mathcal{I}(\hat{\pi}_{k\eta})} + \frac{\eta}{d} \ge \frac{1-5\eta L_1}{\mathcal{I}(\hat{\pi}_{k\eta})} + \frac{\eta}{d}.$$

If $\mathcal{I}(\hat{\pi}_{k\eta}) < \frac{16L_2^2 d^2}{L_1^2}$, it can be easily seen from equation (5.2) that $\mathcal{I}(\hat{\pi}_{(k+1)\eta}) < \frac{32L_2^2 d^2}{L_1^2}$. Consequently, equation (5.2) implies that

$$\frac{1}{\mathcal{I}(\hat{\pi}_{(k+1)\eta})} \ge \min\left(\frac{L_1^2}{32L_2^2 d^2}, \frac{1-5\eta L_1}{\mathcal{I}(\hat{\pi}_{k\eta})} + \frac{\eta}{d}\right).$$

The solution to the recursion of the form above is given by the following lemma:

Lemma 7. *Given positive constants $\lambda_1, \lambda_2 > 0$ and $\gamma \in (0,1)$, if a non-negative sequence $(u_k)_{k\ge 0}$ satisfies $u_0 \ge 0$ and*

$$u_{k+1} \ge \min\left(\lambda_1, (1-\gamma)u_k + \gamma\lambda_2\right), \quad \text{for } k = 0, 1, 2, \cdots \tag{5.3a}$$

Then we have the following two bounds for all $k \ge 0$:

$$u_k \ge \min\left(\frac{\lambda_2\gamma k}{4}, \frac{\lambda_1}{2}, \frac{\lambda_2}{2}\right), \quad \text{and} \quad u_k \ge \min\left(u_0, \frac{\lambda_1}{2}, \frac{\lambda_2}{2}\right). \tag{5.3b}$$

See Section C.2 of the supplementary material for the proof.

Applying Lemma 7 with $u_k := \frac{1}{\mathcal{I}(\hat{\pi}_{k\eta})}$, $\lambda_1 := \frac{L_1^2}{32L_2^2d^2}$, $\lambda_2 := 5L_1d$ and $\gamma := 5\eta L_1$, we have the bound

$$\mathcal{I}(\hat{\pi}_{k\eta}) \le \max\left(\frac{4d}{\eta k} \wedge \mathcal{I}(\pi_0), 64L_2^2d^2/L_1^2, 10L_1d\right), \quad \text{for } k = 1, 2, \dots \tag{5.4}$$

Summing over $k = 1, 2, \dots, N$, we arrive at the bound

$$\frac{1}{N}\sum_{k=1}^{N}\mathcal{I}(\hat{\pi}_{k\eta}) \le c\left(\frac{d\log N}{\eta N}\mathcal{I}(\pi_0) + \frac{L_2^2d^2}{L_1^2} + L_1d\right),$$

which proves the desired result.

5.2. Moment estimates under dissipativity conditions

In this section, we bound the moments of the process $\hat{X}_{k\eta}$ along the path of the discretized Langevin diffusion. In particular, leveraging Assumption 2.3 yields the following:

Lemma 8. *Suppose that Assumption 2.3 holds for the drift term $b(\cdot)$. Then there is a universal constant $C > 0$ such that the interpolated process (3.1) satisfies*

$$\sup_{t\ge 0}\left(\mathbb{E}\|\hat{X}_t\|_2^p\right)^{\frac{1}{p}} \le C\left(\mathbb{E}\|\hat{X}_0\|_2^p\right)^{\frac{1}{p}} + C\sqrt{\frac{p+\beta+d}{\mu}} \qquad \text{for all } p \ge 1. \tag{5.5}$$

In particular, when the initialization condition (2.4) is also satisfied, we have the bound $\sup_{t\ge 0}\left(\mathbb{E}\|\hat{X}_t\|_2^4\right)^{\frac{1}{4}} \le C\{\sigma_0\sqrt{d} + \sqrt{\frac{\beta+d}{\mu}}\}$.

The proof of this lemma is based on martingale L^p estimates and the Burkholder-Davis-Gundy inequality [10]. See Section C.3 of the supplementary material for the details. It is worth noting the bounds in this lemma depend polynomially on the parameters (μ, β) in Assumption 2.3.

Without Assumption 2.3 and control on the directions of the drift at a far distance, the moment of the iterates can exponentially blow up. A simple counterexample is given by the potential function $U(x) = -\|x\|_2^2$ and associated drift term $b(x) = x$. With these choices, it can be seen that $\|\hat{X}_t\|_2 \gtrsim e^T$. On the other hand, it is possible to significantly weaken Assumption 2.3 when the potential function is non-negative, as we discuss in the following section.

5.3. Moment estimates without dissipativity conditions

Note that Lemma 8 requires the distant dissipativity condition from Assumption 2.3. When the underlying potential function is non-negative, this condition can be relaxed,

albeit with a slightly worse dependence on the time horizon T. In this section, we assume $b = -\nabla f$ for some function f that is non-negative over $\mathbb{R}^d$, which we refer to as the case of a *non-negative potential*. In this special case, we have the following auxiliary results:

Lemma 9. *If Assumptions 2.1 and 2.4 hold for a non-negative potential, then there is a universal constant $C > 0$ such that any stepsize $\eta \in (0, \frac{1}{2L_1})$, we have*

$$\sup_{0 \le k \le T/\eta} \left(\mathbb{E} \left\| \hat{X}_{k\eta} \right\|_2^4 \right) \le C \cdot \left(\mathbb{E} f(\hat{X}_0)^2 + L_1^2 T^2 \sigma_0^4 d^2 + L_1^2 T^4 d^2 \right).$$

See Section C.4 of the supplementary material for the proof of this claim.

We also need the following bounds on the moments of the gradient ∇f:

Lemma 10. *If Assumptions 2.1 and 2.4 hold for a non-negative potential, then there is a universal constant $C > 0$ such that for any stepsize $\eta \in (0, \frac{1}{2L_1})$, we have*

$$\mathbb{E} \left[\eta \sum_{k=0}^{T/\eta} \left\| \nabla f(\hat{X}_{k\eta}) \right\|_2^2 \right] \le 2\mathbb{E} f(\hat{X}_0) + L_1 T d, \quad \text{and}$$

$$\mathbb{E} \left(\eta \sum_{k=0}^{T/\eta} \left\| \nabla f(\hat{X}_{k\eta}) \right\|_2^2 \right)^2 \le 12\mathbb{E} f(\hat{X}_0)^2 + 24\mathbb{E} f(\hat{X}_0) + 12 T L_1 d + 9 L_1^2 T^2 d^2.$$

See Section C.5 of the supplementary material for the proof of this claim.

We caution that the two lemmas have inter-dependence: the proof of Lemma 9 relies on Lemma 10. The main idea in the proof of the two lemmas is straightforward: when the gradient ∇f has large norm, the dynamics of the Langevin algorithm force down the value of f. Since f is non-negative, the average mean-squared norm can be bounded using the initial value of f. The combinatorial techniques used in the proof of Lemma 10 are only able to control moments up to order four, which is all that is needed in the proof of Theorem 2. The on-average gradient norm bound in turn makes it possible to establish bounds on the iterates themselves, by invoking Grönwall's inequality.

Substituting the fourth moment bound from Lemma 9 into Proposition 1 yields the claim of Theorem 2.

6. Discussion

We have presented an improved non-asymptotic analysis of the Euler-Maruyama discretization of the Langevin diffusion. We have shown that as long as the drift term satisfies a second-order smoothness condition, then the KL divergence between the Langevin

diffusion and its discretization is bounded as $O(\eta^2 d^2 T)$. Importantly, this analysis yields a tight $O(\eta)$ rate for the error in the Euler-Maruyama scheme, measured under either Wasserstein or TV distances, without assuming global contractivity. This allows continuous-time results to be directly translated into discrete time results with tight rates. Thus, it serves as a convenient tool for the future study of Langevin algorithms for sampling, optimization, and statistical inference.

We should emphasize that our results apply only to the Langevin diffusion. Considering the discretization of more general diffusions, either with location-varying covariance or second-order structure, such as the underdamped Langevin dynamics [14], is a promising direction for further research. Finally, we note that a class of high-order discretization schemes exist for SDE, including the Talay-Tubaro expansion [51], and the Ozaki discretization using Hessian information [15]. For the discrete-time process defined by those schemes, the backward conditional expectation in our analysis admits a higher-order expansion. It is also an interesting direction of research to extend our analysis and obtain improved bounds for high-order schemes in the non-convex setting.

Acknowledgements

This work was partially supported by Office of Naval Research Grant ONR-N00014-18-1-2640 to MJW, and National Science Foundation Grants IIS-1619362 to PLB and CCF-1909365 and DMS-2023505 to MJW and PLB. We thank Xiang Cheng, Valentin De Bortolo, Yi-An Ma, and Andre Wibisono for helpful discussions, as well as the referees and associate editor for their comments and suggestions that helped to improve the paper.

Supplementary Material

Supplementary material to "Improved Bounds for Discretization of Langevin Diffusions: Near-Optimal Rates without Convexity"
(doi: COMPLETED BY THE TYPESETTER; .pdf). The supplementary material consists of the following sections: some additional discussion on the tail assumptions used in literature; proofs of the technical results (Lemma 2 and 3) in Section 4; proofs of the technical results (Lemma 6, 7, 8, 9, and 10) in Section 5; proofs of the mixing time results (Theorem 3, Corollary 2 and Corollary 1); and coarse tail bounds that facilitate the integration-by-parts arguments used throughout the paper.

References

[1] A. Alfonsi, B. Jourdain, and A. Kohatsu-Higa. Optimal transport bounds between the time-marginals of a multidimensional diffusion and its Euler scheme. *Electron. J. Probab.*, 20:31 pp., 2015.

[2] D. Bakry, F. Barthe, P. Cattiaux, and A. Guillin. A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13:60–66, 2008.

[3] D. Bakry and M. Émery. Diffusions hypercontractives. In *Séminaire de probabilités*, volume 1123 of *Lecture Notes in Math.*, pages 177–206. Springer, 1985.

[4] E. Bernton. Langevin Monte Carlo and JKO splitting. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1777–1798. PMLR, 2018.

[5] F. Bolley and C. Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la Faculté des Sciences de Toulouse. Mathématiques. Série 6*, 14(3):331–352, 2005.

[6] N. Bou-Rabee, A. Eberle, and R. Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *arXiv preprint arXiv:1805.00452*, 2018.

[7] N. Bou-Rabee and M. Hairer. Non-asymptotic mixing of the MALA algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110, 2013.

[8] N. Bou-Rabee and E. Vanden-Eijnden. Pathwise accuracy and ergodicity of metropolized integrators for SDEs. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 63(5):655–696, 2010.

[9] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes I: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424, 2004.

[10] D. L. Burkholder, B. J. Davis, and R. F. Gundy. Integral inequalities for convex functions of operators on martingales. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. II: Probability theory*, pages 223–240. Univ. California Press, 1972.

[11] Y. Chen, J. Chen, J. Dong, J. Peng, and Z. Wang. Accelerating nonconvex learning via replica exchange Langevin diffusion. In *International Conference on Learning Representations*, 2019.

[12] X. Cheng and P. Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 186–211. PMLR, 2018.

[13] X. Cheng, N. S Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

[14] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323. PMLR, 2018.

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

[16] A. Dalalyan and A. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *arXiv preprint arXiv:1710.00095*, 2017.

[17] A. Dalalyan and L. Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *arXiv preprint arXiv:1807.09382*, 2018.
[18] A. S. Dalalyan, L. Riou-Durand, and A. Karagulyan. Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets. *arXiv preprint arXiv:1906.08530*, 2019.
[19] V. De Bortoli and A. Durmus. Convergence of diffusions and their discretizations: from continuous to discrete processes and back. *arXiv preprint arXiv:1904.09808*, 2019.
[20] A. Durmus and É. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.
[21] A. Durmus and É. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.
[22] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 793–797. PMLR, 2018.
[23] A. Eberle. Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166(3-4):851–886, 2016.
[24] A. Eberle and M. B. Majka. Quantitative contraction rates for Markov chains on general state spaces. *Electronic Journal of Probability*, 24, 2019.
[25] M. A Erdogdu, L. Mackey, and O. Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, pages 9694–9703, 2018.
[26] W. Fang and M. B. Giles. Multilevel Monte Carlo method for ergodic SDEs without contractivity. *Journal of Mathematical Analysis and Applications*, 2019.
[27] J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928, 2019.
[28] L. Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.
[29] D. J. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review*, 43(3):525–546, 2001.
[30] R. Holley and D. Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46(5-6):1159–1194, 1987.
[31] A. Iserles. *A First Course in the Numerical Analysis of Differential Equations*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, second edition, 2009.
[32] R. Khasminskii. *Stochastic Stability of Differential Equations*, volume 66. Springer Science & Business Media, 2011.
[33] E. Kloeden, P. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, 1992.
[34] H. Lee, A. Risteski, and R. Ge. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering Langevin Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 7858–7867, 2018.
[35] Y. T. Lee, Z. Song, and S. S. Vempala. Algorithmic theory of ODEs and sampling

from well-conditioned log-concave densities. *arXiv preprint arXiv:1812.06243*, 2018.
[36] T. Liang and W. Su. Statistical inference for the population landscape via moment adjusted stochastic gradients. *arXiv preprint arXiv:1712.07519*, 2017.
[37] Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I Jordan. Sampling can be faster than optimization. *arXiv preprint arXiv:1811.08413*, 2018.
[38] M. B. Majka, A. Mijatović, and L. Szpruch. Non-asymptotic bounds for sampling algorithms without log-concavity. *arXiv preprint arXiv:1808.07105*, 2018.
[39] O. Mangoubi and N. Vishnoi. Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems 31*, pages 6030–6040. Curran Associates, Inc., 2018.
[40] P. A. Markowich and C. Villani. On the trend to equilibrium for the Fokker-Planck equation: an interplay between physics and functional analysis. *Matemática Contemporânea*, 19:1–29, 2000. VI Workshop on Partial Differential Equations, Part II (Rio de Janeiro, 1999).
[41] J. C. Mattingly, A. J. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2):185–232, 2002.
[42] J. C. Mattingly, A. M. Stuart, and M. V. Tretyakov. Convergence of numerical time-averaging and stationary measures via Poisson equations. *SIAM Journal on Numerical Analysis*, 48(2):552–577, 2010.
[43] G. N. Milstein and M. V. Tretyakov. *Stochastic Numerics for Mathematical Physics*. Springer Science & Business Media, 2013.
[44] G. A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin equations*. Springer-Verlag, first edition, 2014.
[45] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703. PMLR, 2017.
[46] D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion*, volume 293. Springer-Verlag, third edition, 1999.
[47] G. Roberts and R. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.
[48] W. Rudin. *Functional Analysis*. McGraw-Hill Inc., second edition, 1991.
[49] A. J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2(2):101–112, 1959.
[50] D. Talay. Simulation and numerical analysis of stochastic differential systems: a review. Research Report RR-1313, INRIA, 1990.
[51] D. Talay and L. Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Analysis and Applications*, 8(4):483–509, 1990.
[52] G. Toscani. Entropy production and the rate of convergence to equilibrium for the Fokker-Planck equation. *Quarterly of Applied Mathematics*, 57(3):521–541, 1999.
[53] B. Tzen, T. Liang, and M. Raginsky. Local optimality and generalization guarantees for the Langevin algorithm via empirical metastability. In *Proceedings of the*

31st Conference on Learning Theory, volume 75 of *Proceedings of Machine Learning Research*, pages 857–875. PMLR, 2018.

[54] S. Vempala and A. Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, pages 8094–8106, 2019.

[55] A. Y. Veretennikov. On polynomial mixing bounds for stochastic differential equations. *Stochastic Processes and their Applications*, 70(1):115–127, 1997.

[56] A. Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 2093–3027. PMLR, 2018.

[57] Y. Zhang, P. Liang, and M. Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1980–2022. PMLR, 2017.