Proceedings of Machine Learning Research vol 195:1–36, 2023 36th Annual Conference on Learning Theory

Improved Discretization Analysis for Underdamped Langevin Monte Carlo

Matthew Shunshi Zhang MATTHEW.ZHANG@MAIL.UTORONTO.CA
University of Toronto and Vector Institute

Sinho Chewi SCHEWI@MIT.EDU
Massachusetts Institute of Technology

Mufan Bill Li MUFAN.LI@MAIL.UTORONTO.CA
University of Toronto and Vector Institute

Krishnakumar Balasubramanian KBALA@UCDAVIS.EDU
University of California, Davis

Murat A. Erdogdu ERDOGDU@CS.TORONTO.EDU
University of Toronto and Vector Institute

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Underdamped Langevin Monte Carlo (ULMC) is an algorithm used to sample from unnormalized densities by leveraging the momentum of a particle moving in a potential well. We provide a novel analysis of ULMC, motivated by two central questions: (1) *Can we obtain improved sampling guarantees beyond strong log-concavity?* (2) *Can we improve the condition number dependence in sampling?*

For (1), prior results for ULMC only hold under a log-Sobolev inequality together with a restrictive Hessian smoothness condition. Here, we relax these assumptions by removing the Hessian smoothness condition and by considering distributions satisfying a Poincaré inequality. Our analysis achieves the state of art dimension dependence, and is also flexible enough to handle weakly smooth potentials. As a byproduct, we also obtain the first KL divergence guarantees for ULMC without Hessian smoothness under strong log-concavity, which is based on a new result on the log-Sobolev constant along the underdamped Langevin diffusion.

For (2), the recent breakthrough of Cao, Lu, and Wang (2020) established the first accelerated result for the underdamped Langevin diffusion in continuous time via PDE methods. Our discretization analysis translates their result into an algorithmic guarantee, which indeed enjoys better condition number dependence than prior works on ULMC, although we leave open the question of full acceleration in discrete time.

Both (1) and (2) necessitate Rényi discretization bounds, which are more challenging than the typically used Wasserstein coupling arguments. We address this using a flexible discretization analysis based on Girsanov's theorem that easily extends to more general settings.

Keywords: Girsanov's theorem, log-Sobolev inequality, Poincaré inequality, Rényi divergence, underdamped Langevin Monte Carlo

1. Introduction

The problem of sampling from a high-dimensional distribution $\pi \propto \exp(-U)$ on $\mathbb{R}^d$, when the normalizing constant is unknown and only the potential U is given, has increasing relevancy in a number of application domains, including economics, physics, and scientific computing (Johannes and Polson, 2010; Von Toussaint, 2011; Kobyzev et al., 2020). Recent progress on this problem has been driven by a strong connection with the field of optimization, starting from the seminal work of Jordan et al. (1998); see Chewi (2023) for an exposition.

Given the success of momentum-based algorithms for optimization (Nesterov, 1983), it is natural to investigate momentum-based algorithms for sampling. The hope is that such methods can improve the dependence of the convergence estimates on key problem parameters, such as the condition number κ, the dimension d, and the error tolerance ε. One such method is underdamped Langevin Monte Carlo (ULMC), which is a discretization of the underdamped Langevin diffusion (ULD):

$$\begin{aligned} \mathrm{d}x_t &= v_t \,\mathrm{d}t \,, \\ \mathrm{d}v_t &= -\gamma v_t \,\mathrm{d}t - \nabla U(x_t) \,\mathrm{d}t + \sqrt{2\gamma} \,\mathrm{d}B_t \,, \end{aligned} \tag{ULD}$$

where $\{B_t\}_{t \geq 0}$ is the standard d-dimensional Brownian motion. The stationary distribution of ULD is $\mu(x, v) \propto \exp(-U(x) - \|v\|^2/2)$, and in particular, the x-marginal of μ is the desired target distribution π. Therefore, by taking a small step size for the discretization and a large number of iterations, ULMC will yield an approximate sample from π.

We also note that in the limiting case where $\gamma = 0$, ULMC closely resembles the Hamiltonian Monte Carlo algorithm, which is known to achieve better condition number dependence and discretization error in some limited settings (Vishnoi, 2021; Apers et al., 2022; Bou-Rabee and Marsden, 2022; Wang and Wibisono, 2022).

While there is currently no analysis of ULMC that yields acceleration for sampling (i.e., square root dependence on the condition number κ), ULMC is known to improve the dependence on other parameters such as the dimension d and the error tolerance ε (Cheng et al., 2018a,b; Dalalyan and Riou-Durand, 2020), at least for guarantees in the Wasserstein metric. However, compared to the extensive literature on the simpler (overdamped) Langevin Monte Carlo (LMC) algorithm, existing analyses of ULMC are not easily extended to stronger performance metrics such as the KL and Rényi divergences. In turn, this limits the scope of the results for ULMC; see the discussion in Section 1.1.

In light of these shortcomings, in this work, we ask the following two questions:

1. *Can we obtain sampling guarantees beyond the strongly log-concave case via ULMC?*
2. *Can we obtain better condition number dependence for sampling via ULMC?*

1.1. Our Contributions

We address the two questions above by providing a new Girsanov discretization bound for ULMC. Our bound holds in the strong Rényi divergence metric and applies under general assumptions (in particular, it does not require strong log-concavity of the target π, and it allows for weakly smooth potentials). Consequently, it leads to the following new state-of-the-art results for ULMC:

- We obtain an ε^2-guarantee in KL divergence with iteration complexity $\widetilde{\mathcal{O}}(\kappa^{3/2} d^{1/2} \varepsilon^{-1})$ for strongly log-concave and log-smooth distributions, which removes the Lipschitz Hessian assumption of Ma et al. (2021); here, κ is the condition number of the distribution.

- We obtain an ε-guarantee in TV distance with iteration complexity $\widetilde{\mathcal{O}}(C_{\mathsf{LSI}}^{3/2} L^{3/2} d^{1/2} \varepsilon^{-1})$ under a log-Sobolev inequality (LSI) and L-smooth potential, again without assuming a Lipschitz Hessian. This is the state-of-the-art guarantee for this class of distributions with regards to dimension dependence.
- We obtain ε^2-guarantees in the stronger Rényi divergence metric of any order in $[1, 2)$ with iteration complexity $\widetilde{\mathcal{O}}(C_{\mathsf{PI}}^{3/2} L^{3/2}\, d^2 \varepsilon^{-1})$ under a Poincaré inequality and a L-smooth potential, which improves to $\widetilde{\mathcal{O}}(C_{\mathsf{PI}} L d^2 \varepsilon^{-1})$ under log-concavity. These are the first guarantees for ULMC known in these settings, and they substantially improve upon the corresponding results for LMC in these settings (Chewi et al., 2021).
- In the Poincaré case, we also consider weakly smooth potentials (i.e., Hölder continuous gradients with coefficient $s \in (0, 1]$), which more realistically reflect the delicate smoothness properties of distributions satisfying a Poincaré inequality.

We now discuss our results in more detail in the context of the existing literature.

Guarantees under Weaker Assumptions. Prior works, Cheng et al. (2018b); Dalalyan and Riou-Durand (2020); Ganesh and Talwar (2020), require strong log-concavity of the target. Whereas for works which operate under isoperimetric assumptions, we are only aware of Ma et al. (2021), which further assumes a restrictive Lipschitz Hessian condition for the potential. In contrast, we make no such assumption on the Hessian of U, and we obtain results under a log-Sobolev inequality (LSI), or under the even weaker assumption of a Poincaré inequality (PI), for which sampling analysis is known to be challenging (Chewi et al., 2021).

As noted above, our result for sampling from distributions satisfying LSI and smoothness assumptions are state-of-the-art with regards to the dimension dependence ($d^{1/2}$); in contrast, the previous best results had linear dependence on d (Chewi et al., 2021; Chen et al., 2022). Moreover, in the Poincaré case, we can also consider weakly smooth potentials, which have not been previously considered in the context of ULMC.

Guarantees in Stronger Metrics. Key to achieving these results is our discretization analysis in the Rényi divergence metric. Indeed, the continuous-time convergence results for ULD under LSI or PI hold in the KL or Rényi divergence metrics, and translating these guarantees to the ULMC algorithm necessitates studying the discretization in Rényi. This is the main technical challenge, as we can no longer rely on Wasserstein coupling arguments which are standard in the literature (Cheng et al., 2018b; Dalalyan and Riou-Durand, 2020). Two notable exceptions are the Rényi discretization argument of Ganesh and Talwar (2020), which incurs suboptimal dependence on ε, and the KL divergence argument of Ma et al. (2021), which requires stringent smoothness assumptions.

In this work, we provide the first KL divergence guarantee for sampling from strongly log-concave and log-smooth distributions via ULMC without Hessian smoothness, based on a new LSI along the trajectory (discussed further below).

Condition Number Dependence in Sampling. Our work is also motivated by the breakthrough result of Cao et al. (2020), which achieves for the first time an accelerated convergence guarantee for ULD in continuous time. Our discretization bound allows us to convert this result into an algorithmic guarantee which indeed improves the dependence on the condition number κ[1] for ULMC: whereas

1. In the Poincaré case, the condition number is $\kappa := C_{\mathsf{PI}} L$, which is consistent with the definition in the strongly log-concave case.

prior results incurred a dependence of at least $\kappa^{3/2}$, our dependence is linear in κ in the log-concave case. While this falls short of proving full acceleration for sampling (i.e., an improvement to $\kappa^{1/2}$), our result is a significant step in improving the known condition number dependence in sampling.

A New Log-Sobolev Inequality along the ULD Trajectory. Finally, en route to proving the KL divergence guarantee in the strongly log-concave case, we establish a new log-Sobolev inequality along ULD (Proposition 10), which is of independent interest. While such a result was previously known for the overdamped Langevin diffusion, to the best of our knowledge it is new for the underdamped Langevin diffusion.

This result is then applied to our discretization analysis, which is done using Girsanov's Theorem. While such a technique has been seen before in sampling (Ganesh and Talwar, 2020; Chewi et al., 2021), the application in the underdamped case is by no means straightforward. This requires two technical novelties: *(i)* the aforementioned LSI for the iterates along the trajectory, and *(ii)* a sub-Gaussian tail bound along the ULMC iterates, established via a matrix Grönwall inequality.

1.2. More Related Work

Langevin Monte Carlo. The study of non-asymptotic convergence guarantees for the standard LMC algorithm has a long history (Dalalyan and Tsybakov, 2012; Durmus and Moulines, 2017; Dalalyan, 2017). Guarantees in KL divergence under a log-Sobolev inequality were obtained by Vempala and Wibisono (2019), which developed an appealing continuous-time framework for analyzing LMC under functional inequalities. With some difficulty, this result was extended to Rényi divergences by Ganesh and Talwar (2020); Erdogdu et al. (2022). At the same time, a body of literature studied convergence in KL divergence under tail-growth conditions such as dissipativity (Raginsky et al., 2017; Erdogdu and Hosseinzadeh, 2021; Mou et al., 2022), which usually imply functional inequalities.

Most related to the current work, Chewi et al. (2021) extended the continuous-time approach from Vempala and Wibisono (2019) to Rényi divergences, and moreover introduced a novel discretization analysis using Girsanov's theorem, which also holds for weakly smooth potentials. The present work builds upon the Girsanov techniques introduced in Chewi et al. (2021) to study ULMC.

Underdamped Langevin Diffusion. ULMC is a discretization of the underdamped Langevin diffusion (ULD). First studied by Kolmogorov (1934) and Hörmander (1967) in their pioneering works on hypoellipticity, it was quickly understood that establishing quantitative convergence to stationarity is technically challenging, let alone capturing any acceleration phenomenon. The seminal work of Villani (2002, 2009) developed the hypocoercivity approach, providing the first convergence guarantees under functional inequalities; see also (Hérau, 2006; Dolbeault et al., 2009, 2015; Roussel and Stoltz, 2018). We also refer to Bernard et al. (2022) and references therein for a comprehensive discussion of qualitative and quantitative convergence results for ULD.

As mentioned earlier, the most recent breakthrough by Cao et al. (2020) achieved acceleration in continuous time in χ_2-divergence when the target distribution π is log-concave. This work was built on an approach using the dual Sobolev space $\mathcal{H}^{-1}$ (Albritton et al., 2019). However, since this method relies on the duality of the L^2 space and its connections to the Poincaré inequality, it is difficult to extend to L^p spaces or to other functional inequalities.

Other Discretizations. Many alternative discretization schemes have since been proposed in this setting (Shen and Lee, 2019; Foster et al., 2021; Monmarché, 2021; Foster et al., 2022; Johnston

et al., 2023), albeit all of the analyses up to this point were limited to $\mathcal{W}_2$ distance and did not achieve acceleration in terms of the condition number κ. Other works which cover momentum-based methods include Zou et al. (2019); Gao et al. (2022), although their regimes are quite different from our own.

1.3. Organization

The remainder of this paper will be organized as follows. In Section 2, we will review the required definitions and assumptions. In Section 3, we will state our main results and briefly sketch their proofs. In Section 4, we highlight several implications of our theorems through some examples. In Section 5, we briefly sketch the proofs of our main results, before concluding in Section 6 with a discussion of future directions.

2. Background

2.1. Notation

Hereafter, we will use $\|\cdot\|$ to denote the 2-norm on vectors. In general, we will only work with measures that admit densities on $\mathbb{R}^d$, and we will abuse notation slightly to conflate a measure with its density for convenience. The notation $a = \mathcal{O}(b)$ signifies that there exists an absolute constant $C > 0$ such that $a \leq Cb$, and $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors. Similarly we write $a = \Theta(b)$ if there exist constants $c, C > 0$ such that $cb \leq a \leq Cb$, and $\widetilde{\Theta}(\cdot)$ hides logarithmic factors. The stationary measure (in the position coordinate) is $\pi \propto \exp(-U)$, and U will be referred to as the potential. We will use $L^2(\pi)$ to denote test functions f where $\mathbb{E}_\pi f^2 < \infty$, and $\mathcal{H}^1(\pi)$ to denote weakly differentiable $L^2(\pi)$ functions where $\partial_{x_i} f \in L^2(\pi)$. Finally, the notations $\lesssim, \gtrsim, \asymp$ represent $\leq, \geq, =$ up to absolute constants. Further notations are introduced in subsequent sections.

2.2. Definitions and Assumptions

In this subsection, we will define the relevant processes, divergences, and isoperimetric inequalities. Firstly, we define the ULMC algorithm by the following stochastic differential equation (SDE):

$$\begin{aligned} \mathrm{d}x_t &= v_t \,\mathrm{d}t\,, \\ \mathrm{d}v_t &= -\gamma v_t \,\mathrm{d}t + \nabla U(x_{kh})\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t\,, \end{aligned} \tag{ULMC}$$

where $t \in [kh, (k+1)h)$ for some step size $h > 0$. We note this formulation of ULMC can be integrated in closed form (see Appendix A).

Next, we define a few measures of distance between two probability distributions μ and π on $\mathbb{R}^d$. We define the total variation distance as

$$\|\mu - \pi\|_{\mathsf{TV}} \coloneqq \sup |\mu(A) - \pi(A)|\,, \tag{2.1}$$

where the sup is taken over Borel measurable sets $A \subset \mathbb{R}^d$. We further define the KL divergence as

$$\mathsf{KL}(\mu \parallel \pi) \coloneqq \int \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\,\mathrm{d}\pi\,, \tag{2.2}$$

and $\mathsf{KL}(\mu \,\|\, \pi) := +\infty$ if μ is not absolutely continuous with respect to π. Finally, we define the Rényi divergence with order $q > 1$ as

$$\mathcal{R}_q(\mu \,\|\, \pi) := \frac{1}{q-1} \log \int \left| \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right|^q \mathrm{d}\pi \,,$$

and similarly $\mathcal{R}_q(\mu \,\|\, \pi) := +\infty$ if $\mu \not\ll \pi$. The Rényi divergence upper bounds KL for all orders, i.e., $\mathsf{KL}(\mu \,\|\, \pi) \le \mathcal{R}_q(\mu \,\|\, \pi)$ for any order $q > 1$, and $\mathcal{R}_q$ is monotonic in q. In particular, when $q = 2$, we also get χ_2 divergence, i.e., $\chi_2(\mu \,\|\, \pi) = \exp(\mathcal{R}_2(\mu \,\|\, \pi)) - 1$.

Our primary results are provided under the following smoothness conditions.

Definition 1 (Smoothness) *The potential U is (L, s)-**weakly smooth** if U is differentiable and ∇U is s-Hölder continuous satisfying*

$$\|\nabla U(x) - \nabla U(y)\| \le L \|x - y\|^s \,, \tag{2.3}$$

for all $x, y \in \mathbb{R}^d$ and some $L \ge 0$, $s \in (0, 1]$. In the particular case where $s = 1$, we say that the potential is L-smooth, or that ∇U is L-Lipschitz.

We conduct three lines of analysis. The first assumes strong convexity of the potential, i.e.:

Definition 2 (Strong Convexity) *The potential U is m-**strongly convex** for some $m \ge 0$ if for all $x, y \in \mathbb{R}^d$:*

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \ge \frac{m}{2} \|x - y\|^2 \,.$$

In the case $m = 0$ above, we say that U is convex. If a potential function U is (strongly) convex, then we say the distribution $\pi \propto \exp(-U)$ is (strongly) log-concave.

A second, strictly more general assumption is the log-Sobolev inequality.

Definition 3 (Log-Sobolev Inequality) *A measure π satisfies a **log-Sobolev inequality** (LSI) with parameter $C_{\mathsf{LSI}} > 0$ if for all $g \in \mathcal{H}^1(\pi)$:*

$$\mathsf{ent}_\pi(g^2) \le 2 C_{\mathsf{LSI}} \, \mathbb{E}_\pi[\|\nabla g\|^2] \,, \tag{LSI}$$

where $\mathsf{ent}_\pi(g^2) := \mathbb{E}_\pi[g^2 \log(g^2 / \mathbb{E}_\pi[g^2])]$.

An m-strongly convex potential is known to satisfy (LSI) with constant m^{-1} (Bakry et al., 2014). More generally, we can consider the following weaker isoperimetric inequality, which corresponds to a linearization of (LSI).

Definition 4 (Poincaré Inequality) *A measure π satisfies a **Poincaré inequality** with parameter $C_{\mathsf{PI}} > 0$ if for all $g \in \mathcal{H}^1(\pi)$:*

$$\mathsf{var}_\pi(g) \le C_{\mathsf{PI}} \, \mathbb{E}_\pi[\|\nabla g\|^2] \,, \tag{PI}$$

where $\mathsf{var}_\pi(g) = \mathbb{E}_\pi[|g - \mathbb{E}_\pi[g]|^2]$.

Conditions (LSI) and (PI) are standard assumptions made on the stationary distribution in the theory of Markov diffusions as well as sampling (Bakry et al., 2014; Vempala and Wibisono, 2019; Chewi et al., 2021; Chewi, 2023). They are known to be satisfied by a broad class of targets such as log-concave distributions or certain mixture distributions (Chen, 2021; Chen et al., 2021).

We define the condition number for an m-strongly log-concave target with (L, s)-weakly smooth potential as $\kappa := L/m$. In the case where instead of strong convexity, the target only satisfies (LSI) (respectively (PI)), the condition number is instead $\kappa := C_{\mathsf{LSI}} L$ (respectively $\kappa := C_{\mathsf{PI}} L$).

Finally, we collect several mild assumptions to simplify computing the bounds below, which have also appeared in prior work; see in particular the discussion in Chewi et al. (2021, Appendix A).

Assumption 1 *The expectation of the norm (in the position coordinate) is quantitatively bounded by some constant, $\mathbb{E}_\pi[\|\cdot\|] \le \mathfrak{m} = \widetilde{\mathcal{O}}(d)$[2], for some constant $\mathfrak{m} < \infty$. Furthermore, we assume that $\nabla U(0) = 0$ (without loss of generality), and that $U(0) - \min U = \widetilde{\mathcal{O}}(d)$.*

Remark *On an intuitive level, Assumption 1 asks for bounds on the noncentral moment, whereas isoperimetric inequalities only imply bounds on the central moments. For instance, one can construct a sub-Gaussian target centered at some parameter θ with $\|\theta\| \asymp d^2$, which would satisfy (LSI) but not Assumption 1.*

3. Main Theorems

In the sequel, we always take the initial distribution of the momentum ρ_0 to be equal to the stationary distribution $\rho \propto \exp(-\|\cdot\|^2/2)$. Then, under Assumption 1 we can find an initial distribution π_0 for the position which is a centered Gaussian with variance specified in Appendix D, such that π_0 has some appropriately bounded initial divergence (e.g. $\mathsf{KL}, \mathcal{R}_q$) with respect to π. Lastly, we initialize ULMC by sampling from the distribution $\mu_0(x, v) = \pi_0(x) \times \rho_0(v)$, i.e. with x and v independent.

3.1. Convergence in KL and TV

In order to state our results for ULMC in KL and TV, we leverage the following result in continuous-time from Ma et al. (2021), which relies on an entropic hypocoercivity argument, after a time-change of the coordinates (see Appendix B.1 for a proof).

Lemma 5 (Adapted from Ma et al. (2021, Proposition 1)) *Define the Lyapunov functional*

$$\mathcal{F}(\mu' \,\|\, \mu) := \mathsf{KL}(\mu' \,\|\, \mu) + \mathbb{E}_{\mu'}\left[\left\|\mathfrak{M}^{1/2} \nabla \log \frac{\mu'}{\mu}\right\|^2\right], \quad \text{where } \mathfrak{M} = \begin{bmatrix} \frac{1}{4L} & \frac{1}{\sqrt{2L}} \\ \frac{1}{\sqrt{2L}} & 4 \end{bmatrix} \otimes I_d \,. \tag{3.1}$$

For targets π that are L-smooth and satisfy (LSI) with parameter C_{LSI}, let $\gamma = 2\sqrt{2L}$. Then the law μ_t of ULD satisfies

$$\partial_t \mathcal{F}(\mu_t \,\|\, \mu) \le -\frac{1}{10 C_{\mathsf{LSI}} \sqrt{2L}} \mathcal{F}(\mu_t \,\|\, \mu) \,.$$

We now proceed to state our main results more precisely. First, we obtain the following KL divergence guarantee under strong log-concavity and smoothness.

2. This holds for instance when $U(x) = \|x\|^\alpha$ for $1 \le \alpha \le 2$.

Theorem 6 (Convergence in KL under Strong Log-Concavity) *Let the potential U be m-strongly convex and L-smooth, and additionally satisfy Assumption 1. Then, for*

$$h = \widetilde{\Theta}\Big(\frac{\varepsilon m^{1/2}}{L d^{1/2}}\Big) \quad \text{and} \quad \gamma \asymp \sqrt{L},$$

the following holds for $\hat{\mu}_{Nh}$, the law of the N-th iterate of ULMC initialized at a centered Gaussian (with variance specified in Appendix D):

$$\mathsf{KL}(\hat{\mu}_{Nh} \,\|\, \mu) \le \varepsilon^2 \qquad \text{after} \qquad N = \widetilde{\Theta}\Big(\frac{\kappa^{3/2}\, d^{1/2}}{\varepsilon}\Big) \quad \text{iterations}\,.$$

Here, we justify the choice of error tolerance for KL to be ε^2. Based on Pinsker's and Talagrand's transport inequalities, we know KL is on the order of $\mathsf{TV}^2, \mathcal{W}_2^2$. Hence, this allows for a fair comparison of convergence guarantees in terms of KL with TV and $\mathcal{W}_2$. Weakening the strong convexity assumption to (LSI), we obtain a result in TV.

Theorem 7 (Convergence in TV under (LSI)) *Let the potential be L-smooth, satisfy (LSI) with constant C_{LSI}, and satisfy Assumption 1. Then, for*

$$h = \widetilde{\Theta}\Big(\frac{\varepsilon}{C_{\mathsf{LSI}}^{1/2} L d^{1/2}}\Big), \quad \text{and} \quad \gamma \asymp \sqrt{L},$$

the following holds for $\hat{\mu}_{Nh}$, the law of the N-th iterate of ULMC initialized at a centered Gaussian (with variance specified in Appendix D):

$$\|\hat{\mu}_{Nh} - \mu\|_{\mathsf{TV}} \le \varepsilon \qquad \text{after} \qquad N = \widetilde{\Theta}\Big(\frac{C_{\mathsf{LSI}}^{3/2}\, L^{3/2}\, d^{1/2}}{\varepsilon}\Big) \quad \text{iterations}\,.$$

3.2. Convergence in $\mathcal{R}_q$ and Improving the Condition Number κ

To state our convergence results in $\mathcal{R}_q$, we additionally inherit the following technical assumption from Cao et al. (2020).

Assumption 2 *$\mathcal{H}^1(\mu) \hookrightarrow L^2(\mu)$ is a compact embedding. Secondly, assume that U is twice continuously differentiable, and that for all $x \in \mathbb{R}^d$, we have*

$$\|\nabla^2 U(x)\| \le \mathfrak{L}\,(1 + \|\nabla U(x)\|)\,.$$

Remark Hooton (1981, Theorem 3.1) shows the first part of this assumption is always satisfied if the potential has super-linear tail growth, i.e. $U(x) \propto \|x\|^\alpha$ for $\alpha > 1$ and large $\|x\|$. In the case where the tail is strictly linear, we can instead construct an arbitrarily close approximation with super-linear tails; thus, it generically holds for all targets we consider in this work. As also remarked in Cao et al. (2020), the above assumption is required solely due to technical reasons and is likely not a necessary condition.

The second part of the assumption is satisfied under L-smoothness of the gradient with the same constant. In the convex case or the case where $\nabla^2 U$ is lower bounded, the constant $\mathfrak{L}$ does not show up in the bounds. As a result, for weakly smooth potentials in this setting, we can approximate using twice differentiable potentials to obtain a rate estimate.

In the light of the above discussion, we emphasize that this additional assumption largely does not hinder the applicability of our results. Under this assumption, Cao et al. (2020) established the following guarantee on (ULD) in continuous time.

Lemma 8 (Rapid Convergence in L^2; Adapted from Cao et al. (2020, Theorem 1)) *Under Assumption 2, and if π additionally satisfies (PI) with constant C_{PI}, then the following holds for the law μ_t of ULD initialized at μ_0, where $C_0 > 0$ is an absolute constant:*

$$\chi_2(\mu_t \parallel \mu) \le C_0 \exp\bigl(-\mathfrak{q}(\gamma)\, t\bigr)\, \chi_2(\mu_0 \parallel \mu)\,,$$

where the coefficient inside the exponent is

$$\mathfrak{q}(\gamma) \coloneqq \frac{C_{\mathsf{PI}}^{-1}\gamma}{C_0\,(C_{\mathsf{PI}}^{-1} + R^2 + \gamma^2)}, \tag{3.2}$$

and the constant R is

$$R = \begin{cases} 0 & \text{if } U \text{ convex}\,, \\ \sqrt{K} & \text{if } \inf_{x\in\mathbb{R}^d} \nabla^2 U(x) \succeq -K I_d\,, \\ \mathfrak{L}\sqrt{d} & \text{if } \|\nabla^2 U(x)\|_{\mathrm{op}} \le \mathfrak{L}\,(1 + \|\nabla U(x)\|) \text{ for all } x \in \mathbb{R}^d\,. \end{cases}$$

Remark In the strongly log-concave case, Lemma 8 actually yields a better decay of order $\sqrt{m}$ than Lemma 5, which has dependence $m/\sqrt{L}$.

Our final result leverages the above accelerated convergence guarantees of ULD, and establishes the first bound for ULMC in Rényi divergence with an improved condition number dependence.

Theorem 9 (Convergence in $\mathcal{R}_q$ under (PI)) *Let the potential be (L, s)-weakly smooth, satisfy (PI) with constant C_{PI}, and satisfy Assumption 1. Let it also satisfy the additional technical condition Assumption 2. Then, for $\xi \in (0, 1)$*

$$h = \tilde{\Theta}\Bigl(\frac{\gamma^{1/(2s)}\varepsilon^{1/s}\xi^{1/s}\mathfrak{q}(\gamma)^{1/(2s)}}{L^{1/s}d^{1/2}\,(L \vee d)^{1/(2s)}}\Bigr),$$

the following holds for $\hat{\mu}_{Nh}$, the law of the N-th iterate of ULMC initialized at a centered Gaussian (variance specified in Appendix D) for $q = 2 - \xi \in [1, 2)$ and with $\mathfrak{q}$ defined in (3.2):

$$\mathcal{R}_q(\hat{\mu}_{Nh} \parallel \mu) \le \varepsilon^2 \qquad \textit{after} \qquad N = \tilde{\Theta}\Bigl(\frac{L^{1/s}\, d^{1/2}\, (L \vee d)^{1+1/(2s)}}{\gamma^{1/(2s)}\, \varepsilon^{1/s}\, \xi^{1/s}\, \mathfrak{q}(\gamma)^{1+1/(2s)}}\Bigr) \quad \textit{iterations}\,,$$

Remark The optimal choice is to take $\gamma \asymp \sqrt{C_{\mathsf{PI}}^{-1} + R^2}$. If the potential U is convex, then we set $\gamma \asymp \mathfrak{q}(1/\sqrt{C_{\mathsf{PI}}}) \asymp 1/\sqrt{C_{\mathsf{PI}}}$, which is known to be an optimal choice (Cao et al., 2020). As a result, in the convex and smooth case, the iteration complexity has the condition number dependence κ, which improves upon the κ^2 dependence seen in Chewi et al. (2021). The dependence on dimension d and error tolerance ε are also improved.

These results are compared against the known upper bounds in Table 3.2. To summarize our improvements on existing literature, we note that *(i)* our results in the strongly log-concave case are in the "stronger" divergence of $\sqrt{\mathsf{KL}}$ compared to the previous known guarantees in $\mathcal{W}_2$, *(ii)* our results under LSI have better condition number dependence, and remove dependence on the Frobenius Lipschitz constant of the Hessian (which scales like $\mathcal{O}(d)$), *(iii)* the PI regime is a novel result, which to our knowledge has not been seen before in previous works.

Source	Condition	Metric	Complexity
[Dalalyan and Riou-Durand '20]	Strongly Log-Concave	$\mathcal{W}_2$	$\kappa^{3/2}d^{1/2}/\varepsilon$
Theorem 6	Strongly Log-Concave	$\sqrt{\mathsf{KL}}$	$\kappa^{3/2}d^{1/2}/\varepsilon$
[Ma et al. '21]	LSI	$\sqrt{\mathsf{KL}}$	$C_{\mathsf{LSI}}^2 L^2 L_H d^{1/2}/\varepsilon$
Theorem 7	LSI	TV	$C_{\mathsf{LSI}}^{3/2} L^{3/2} d^{1/2}/\varepsilon$
Theorem 8	PI	$\sqrt{\mathcal{R}_{3/2}}$	$C_{\mathsf{PI}} L d^{1/2}/\varepsilon$

Table 1: We compare our guarantees against existing results. The result of [Ma et al. '21] contains dependence on the Hessian Frobenius smoothness constant L_H, which generally scales like $\mathcal{O}(d)$. Our q-Rényi result also holds for $q \in [1, 2)$.

4. Examples

Example 1 We consider the potential $U(x) = \sqrt{1 + \|x\|^2}$, which satisfies (PI) with constant $\mathcal{O}(d)$ (Bobkov, 2003) and is $(1, 1)$-smooth. Assuming the compact embedding condition of Assumption 2, Theorem 9 gives a complexity of $\widetilde{\mathcal{O}}(d^3\xi^{-1}\varepsilon^{-1})$ for ε^2-guarantees in $\mathcal{R}_{2-\xi}$ after optimizing for γ, since in this case the potential is log-concave. In this case, the dimension dependence equates to that of the proximal sampler with rejection sampling (Chen et al., 2022, Corollary 8), which is $\widetilde{\mathcal{O}}(d^3)$; it surpasses Chewi et al. (2021, Theorem 8), which can only obtain $\widetilde{\mathcal{O}}(d^4\varepsilon^{-2})$ for the same guarantees. However, it is important to note that the latter two works obtain these for any order of Rényi divergence and are not limited to order $q = 2 - \xi < 2$, which cannot presently be obtained using our results for ULMC.

Example 2 Consider an m-strongly log-concave and L-log-smooth distribution. Non-trivial examples of this can be found in Bayesian regression (see e.g., Dalalyan (2017, Section 6)); we will examine the first one, where $\pi(x) \propto \exp(-\|x - \boldsymbol{a}\|^2/2) + \exp(-\|x + \boldsymbol{a}\|^2/2)$ for some $\boldsymbol{a} \in \mathbb{R}^d$: $\|\boldsymbol{a}\| = 1/3$. Here, our Theorem 6 gives a complexity of $N = \widetilde{\mathcal{O}}(d^{1/2}\varepsilon^{-1})$ to obtain a ε^2-guarantee for the KL divergence. In contrast, the Hessian is $\nabla^2 U(x) = I_d - 4\boldsymbol{a}\boldsymbol{a}^\top \exp(2x^\mathsf{T}\boldsymbol{a})/(1+\exp(2x^\mathsf{T}\boldsymbol{a}))^2$, which has $L_\mathsf{H} \asymp d$, where L_H is the Lipschitz constant of the Hessian in the Frobenius norm. Consequently, Ma et al. (2021, Theorem 1) is stated as $N = \widetilde{\mathcal{O}}(d^{1/2}L_H m^{-2}\varepsilon^{-1})$, which in this case gives $N = \widetilde{\mathcal{O}}(d^{3/2}\varepsilon^{-1})$ to obtain the same ε^2-accuracy guarantee. This is worse in the dimension-dependence. Finally, it is possible to compare with the discretization bounds achieved in Ganesh and Talwar (2020, Theorem 28), where in combination with our continuous time results (using the same proof technique as Theorem 6) to yield $N = \widetilde{O}(d^{1/2}\varepsilon^{-2})$ iterations, which is suboptimal in the order of ε, but has the same dimension dependence.

Example 3 We can analyze L-smooth distributions satisfying a log-Sobolev inequality with parameter C_{LSI}. One such instance arises when considering any bounded perturbation of a strongly convex potential. In this case, let $U_{\boldsymbol{a}}$ be the potential of the target in Example 2. Then consider a target with modified potential $U_{\boldsymbol{a}} + f$, with $\sup_x |f(x)| \vee \|\nabla f(x)\| \vee \|\nabla^2 f(x)\|_{\mathsf{op}} \le \mathfrak{B}$ for some $\mathfrak{B} < \infty$, and let $\nabla^2 f$ be $\mathcal{O}(d)$-Frobenius Lipschitz. We can bound the log-Sobolev constant of this potential using the Holley–Stroock Lemma (Holley and Stroock, 1987). Let this new potential have condition number κ. We achieve ε-accuracy in TV distance with $N = \widetilde{\mathcal{O}}(\kappa^{3/2}d^{1/2}\varepsilon^{-1})$. For comparison, the previous bound (Ma et al., 2021, Theorem 1) gives $N = \tilde{\mathcal{O}}(\kappa^2 d^{3/2}\varepsilon^{-1})$ to arrive at

the same guarantee in TV, which is worse in the dimension. However, note that the guarantees in Ma et al. (2021, Theorem 1) are in KL, which is stronger than TV. Finally, we note that Ganesh and Talwar (2020) requires strong log-concavity, and hence cannot provide a guarantee in this setting.

Example 4 Consider a $(1, s)$-weakly log-smooth target that is log-concave and satisfies a Poincaré inequality with $C_{\mathsf{PI}} = \mathcal{O}(d)$. Consequently, Theorem 9 yields $N = \tilde{\mathcal{O}}(d^{2+1/s}\xi^{-1/s}\varepsilon^{-1/s})$ to obtain ε^2-guarantees for $\mathcal{R}_{2-\xi}$. Chewi et al. (2021, Theorem 7) yields $N = \tilde{\mathcal{O}}(d^{3+2/s}\varepsilon^{-2/s})$ for the same guarantees, which is worse in both parameters. On the other hand, take the specific case of a distribution with potential $U(x) = \|x\|^\alpha$, which has $C_{\mathsf{PI}} = \mathcal{O}(d^{2/\alpha-1})$ (Bobkov, 2003), is log-convex and $(1, \alpha-1)$-weakly log-smooth. Consequently, Theorem 9 yields $N = \tilde{\mathcal{O}}(d^{\alpha/(\alpha-1)}\xi^{-1/(\alpha-1)}\varepsilon^{-1/(\alpha-1)})$ for ε^2-accuracy guarantees in $\mathcal{R}_{2-\xi}$ divergence. This is worse by a factor of d than the rate estimate obtained in Chewi et al. (2021, Example 9), as they leverage a stronger class of functional inequalities that interpolate between (PI) and (LSI), whereas our analysis cannot capture this improvement. Our convergence guarantee is still better in terms of ε-dependence.

5. Proof Sketches

5.1. Continuous Time Results

For results under both the Poincaré and log-Sobolev inequalities, we leverage the existing results as stated in Cao et al. (2020); Ma et al. (2021), which we present in Lemmas 5 and 8. These allow us to bound $\chi_2(\mu_t \,\|\, \mu)$, $\mathsf{KL}(\mu_t \,\|\, \mu)$ with exponentially decaying quantities.

With the additional assumption of strong convexity, we can obtain a contraction in an alternate system of coordinates $(\phi, \psi) \coloneqq \mathcal{M}(x, v) \coloneqq (x, x + \frac{2}{\gamma} v)$ (see Appendix B). This allows us to consider the distributions of the continuous time iterates and the target in these alternate coordinates $\mu_t^{\mathcal{M}}, \mu^{\mathcal{M}}$ respectively. From this, we obtain the following proposition.

Proposition 10 (Log-Sobolev Inequality Along the Trajectory) *Suppose U is m-strongly convex and L-smooth. Let $\mu_t^{\mathcal{M}}$ now denote the law of the continuous-time underdamped Langevin diffusion with $\gamma = c\sqrt{L}$ for $c \geq \sqrt{2}$ in the (ϕ, ψ) coordinates. Suppose the initial distribution μ_0 has* (LSI) *constant (in the altered coordinates) $C_{\mathsf{LSI}}(\mu_0^{\mathcal{M}})$, then $\{\mu_t^{\mathcal{M}}\}_{t \geq 0}$ satisfies* (LSI) *with constant that can be uniformly upper bounded by*

$$C_{\mathsf{LSI}}(\mu_t^{\mathcal{M}}) \leq \exp\Big(-m\sqrt{\frac{2}{L}}\, t\Big)\, C_{\mathsf{LSI}}(\mu_0^{\mathcal{M}}) + \frac{2}{m}\,.$$

The main idea behind the proof of this proposition is to analyze the discretization (ULMC) of the underdamped Langevin diffusion in the coordinates (ϕ, ψ). Note that this can be written in the following form, for some matrix $\overline{\Sigma} \in \mathbb{R}^{2d \times 2d}$ and function $\bar{F} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$,

$$(\phi_{(k+1)h}, \psi_{(k+1)h}) \overset{\mathsf{d}}{=} \bar{F}(\phi_{kh}, \psi_{kh}) + \mathcal{N}(0, \overline{\Sigma})\,.$$

This is the composition of a deterministic function $\bar{F}$ giving the mean of the next iterate of ULMC started at (ϕ, ψ), followed by addition with a Gaussian distribution giving the variance of the resulting iterate. In particular, we show that for coordinates $(\phi(x, v), \psi(x, v)) \coloneqq (x, x + \frac{2}{\gamma} v)$, we can find an almost sure strict contraction under $\bar{F}$ in the sense that

$$\|\bar{F}\|_{\mathsf{Lip}} \leq 1 - \frac{m}{\sqrt{2L}}\, h + \mathcal{O}(Lh^2)\,,$$

where by abuse of notation $\bar{F} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$, and the seminorm $\|g\|_{\mathrm{Lip}}$ of a function $g : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ refers to the Lipschitz constant of the function.

Since $\bar{F}$ is a contraction for small enough h, each push forward improves the log-Sobolev constant by a multiplicative factor (Vempala and Wibisono, 2019, Lemma 19). At the same time, a Gaussian convolution can only worsen the log-Sobolev constant by an additive constant (Chafai, 2004, Corollary 3.1). Subsequently, the log-Sobolev constant at each iterate forms a (truncated) geometric sum, and therefore can be bounded by the infinite series. This incidentally can be used to bound the log-Sobolev constant of the ULMC iterates. Taking an appropriate limit of $h \to 0$ while keeping $Nh = t$, we arrive at the stated bound in the proposition. Consequently, considering the decomposition of the KL, a simple application of Cauchy–Schwarz tells us that

$$\begin{aligned}\mathsf{KL}(\hat{\mu}_t^{\mathcal{M}} \parallel \mu^{\mathcal{M}}) &= \int \log \frac{\hat{\mu}_t^{\mathcal{M}}}{\mu^{\mathcal{M}}} \, \mathrm{d}\hat{\mu}_t^{\mathcal{M}} = \mathsf{KL}(\hat{\mu}_t^{\mathcal{M}} \parallel \mu_t^{\mathcal{M}}) + \int \log \frac{\mu_t^{\mathcal{M}}}{\mu^{\mathcal{M}}} \, \mathrm{d}\hat{\mu}_t^{\mathcal{M}} \\ &\leq \mathsf{KL}(\hat{\mu}_t^{\mathcal{M}} \parallel \mu_t^{\mathcal{M}}) + \mathsf{KL}(\mu_t^{\mathcal{M}} \parallel \mu^{\mathcal{M}}) + \sqrt{\chi^2(\hat{\mu}_t^{\mathcal{M}} \parallel \mu_t^{\mathcal{M}}) \times \mathsf{var}_{\mu_t^{\mathcal{M}}}\Big(\log \frac{\mu_t^{\mathcal{M}}}{\mu^{\mathcal{M}}}\Big)} \, .\end{aligned}$$

The log-Sobolev inequality for $\mu_t^{\mathcal{M}}$ implies a Poincaré inequality, which allows us to bound the variance term by the Fisher information $\mathsf{FI}(\mu_t^{\mathcal{M}} \parallel \mu^{\mathcal{M}}) = \mathbb{E}_{\mu_t^{\mathcal{M}}} \|\nabla \log(\mu_t^{\mathcal{M}} / \mu^{\mathcal{M}})\|^2$. This can be bounded by the same entropic hypocoercivity argument from Ma et al. (2021) that is used to generate our TV bounds, while the remaining two terms are handled respectively via the discretization analysis and again the entropic hypocoercivity argument.

5.2. Discretization Analysis

The main result we use to control the discretization error can be found below.

Proposition 11 *Let $(\hat{\mu}_t)_{t \geq 0}$ denote the law of (ULMC) and let $(\mu_t)_{t \geq 0}$ denote the law of the continuous-time underdamped Langevin diffusion (ULD), both initialized at some μ_0. Assume that the potential U is (L, s)-weakly smooth. If the step size h satisfies*

$$h = \widetilde{\mathcal{O}}_s\Big(\frac{\gamma^{1/(2s)} \, \varepsilon^{1/s}}{L^{1/s} \, T^{1/(2s)} \, (d + \mathcal{R}_2(\mu_0 \parallel \mu^{(a)}))^{1/2}}\Big) \, , \tag{5.1}$$

where the notation $\widetilde{\mathcal{O}}_s$ hides constants depending on s as well as polylogarithmic factors including $\log N$, and $\mu^{(a)}$ is a modified target distribution (see Appendix C.3 for details), then

$$\mathcal{R}_q(\hat{\mu}_T \parallel \mu_T) \leq \varepsilon^2 \, .$$

Remark The condition on h is dependent on N only through logarithmic factors. Secondly, this is shown under generic assumptions, and can be combined with continuous-time results in $\mathcal{R}_q$ in any setting, such as the log-Sobolev or Latała–Oleszkiewicz inequalities seen in Chewi et al. (2021).

We outline the proof of this result below. Similar to the work of Chewi et al. (2021), we first invoke the data processing inequality, allowing us to bound the Rényi between the time marginal distributions of the iterates with Rényi between the path measures

$$\mathcal{R}_q(\hat{\mu}_T \parallel \mu_T) \leq \mathcal{R}_q(P_T \parallel Q_T) \, ,$$

where P_T, Q_T are probability measures of (ULMC), (ULD) respectively on the space of paths $C([0,T], \mathbb{R}^{2d})$. Subsequently, we invoke Girsanov's theorem, which allows us to exactly bound the pathwise divergence by the difference between the drifts of the two processes:

$$\mathcal{R}_{2q}(P_T \,\|\, Q_T) \lesssim \log \mathbb{E} \exp\Big(\frac{4q^2}{\gamma} \int_0^T \|\nabla U(x_t) - \nabla U(x_{\lfloor t/h \rfloor h})\|^2 \,\mathrm{d}t \Big) .$$

It remains to bound the term inside the expectation. We achieve this by conditioning on the event that $\sup_{t\in[0,T]} \|x_t - x_{\lfloor t/h \rfloor h}\|^2$ is bounded by a vanishing quantity as $h \to 0$, which we must demonstrate occurs with sufficiently high probability. To show this, we begin with a single-step analysis, i.e., we bound the above for $T \leq h$. Compared to LMC, the main gain in this analysis is that the SDEs (ULD) and (ULMC) match exactly in the position coordinate, while the difference between the drifts manifests solely in the momentum. After integration of the momentum, the order of error is better in the position coordinate (the dominant term is $\mathcal{O}(dh^2)$ compared to $\mathcal{O}(dh)$ seen in Chewi et al. (2021, Lemma 24)).

The technique for extending this analysis from a single step to the full time interval follows closely that seen in Chewi et al. (2021). In particular, we obtain a dependence for $\|x_t\|$ on $\|x_{kh}\|$ in the interval $t \in [kh, (k+1)h)$. Controlling the latter is quite complicated when the potential satisfies only a Poincaré inequality, since it is equivalent to showing sub-Gaussian tail bounds on the iterates, while the target itself is not sub-Gaussian in the position coordinate. By comparing against an auxiliary potential, we can show that for our choice of initialization, the iterates remain sub-Gaussian for all iterations up to N (albeit with a growing constant). Finally, this allows us to recover our discretization result in the proposition above.

6. Conclusion

This work provides state-of-the-art convergence guarantees for underdamped Langevin Monte Carlo algorithm in several regimes. Our discretization analysis (Proposition 11) in particular is generic and can be extended to any order of Rényi, under various conditions on the potential (Latała–Oleszkiewicz, weak smoothness, etc.). Consequently, our results serve as a key step towards a complete understanding of the ULMC algorithm. However, limitations of the current continuous-time techniques do not permit us to obtain stronger iteration complexity results. More specifically, it is not understood how to analyze Rényi divergence of order greater than 2, or if hypercontractive decay is possible when the potential satisfies a log-Sobolev inequality. Secondly, our discretization approach via Girsanov is currently suboptimal in the condition number (a fact noted in Chewi et al. (2021)), and thus does not obtain the expected dependence of $\sqrt{\kappa}$ after discretization. An improvement in the proof techniques would be necessary to sharpen this result. We believe the results and techniques developed in this work will be of interest to stimulate future research.

Acknowledgments

We thank Jason M. Altschuler, Alain Durmus, and Aram-Alexandre Pooladian for helpful conversations. MSZ was supported by NSERC PGS-D (award 579155-2023). KB was supported by NSF grant DMS-2053918. SC was supported by the NSF TRIPODS program (award DMS-2022448). MAE was supported by NSERC Grant [2019-06167], the Connaught New Researcher Award, the CIFAR AI Chairs program, and the CIFAR AI Catalyst grant. ML was supported by the Ontario Graduate Scholarship and Vector Institute.

References

Dallas Albritton, Scott Armstrong, Jean-Christophe Mourrat, and Matthew Novack. Variational methods for the kinetic Fokker–Planck equation. *arXiv preprint arXiv:1902.04037*, 2019.

Simon Apers, Sander Gribling, and Dániel Szilágyi. Hamiltonian Monte Carlo for efficient Gaussian sampling: long and random steps. *arXiv preprint arXiv:2209.12771*, 2022.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2014.

Etienne Bernard, Max Fathi, Antoine Levitt, and Gabriel Stoltz. Hypocoercivity with schur complements. *Annales Henri Lebesgue*, 5:523–557, 2022.

Sergey G. Bobkov. Spectral gap and concentration for some spherically symmetric probability measures. In *Geometric aspects of functional analysis*, volume 1807 of *Lecture Notes in Math.*, pages 37–43. Springer, Berlin, 2003.

Nawaf Bou-Rabee and Milo Marsden. Unadjusted Hamiltonian MCMC with stratified Monte Carlo time integration. *arXiv preprint arXiv:2211.11003*, 2022.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

Yu Cao, Jianfeng Lu, and Lihan Wang. On explicit L^2-convergence rate estimate for underdamped Langevin dynamics. *arXiv e-prints*, art. arXiv:1908.04746, 2020.

Djalil Chafai. Entropies, convexity, and functional inequalities: on Φ-entropies and Φ-Sobolev inequalities. *J. Math. Kyoto Univ.*, 44(2):325–363, 2004.

Jagdish Chandra and Paul W. Davis. Linear generalizations of Gronwall's inequality. *Proceedings of the American Mathematical Society*, 60(1):157–160, 1976.

Hong-Bin Chen, Sinho Chewi, and Jonathan Niles-Weed. Dimension-free log-Sobolev inequalities for mixture distributions. *Journal of Functional Analysis*, 281(11):109236, 2021.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2984–3014. PMLR, 02–05 Jul 2022.

Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the KLS conjecture. *Geom. Funct. Anal.*, 31(1):34–61, 2021.

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

Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: a non-asymptotic analysis. In *Conference on Learning Theory*, pages 300–323. PMLR, 2018b.

Sinho Chewi. *Log-concave sampling*. 2023. Book draft available at https://chewisinho.github.io/.

Sinho Chewi, Murat A. Erdogdu, Mufan Bill Li, Ruoqi Shen, and Matthew Zhang. Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev. *arXiv preprint arXiv:2112.12662*, 2021.

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

Arnak S. Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

Arnak S Dalalyan and Alexandre B Tsybakov. Sparse regression learning by aggregation and langevin monte-carlo. *Journal of Computer and System Sciences*, 78(5):1423–1443, 2012.

Jean Dolbeault, Clément Mouhot, and Christian Schmeiser. Hypocoercivity for kinetic equations with linear relaxation terms. *Comptes Rendus Mathematique*, 347(9-10):511–516, 2009.

Jean Dolbeault, Clément Mouhot, and Christian Schmeiser. Hypocoercivity for linear kinetic equations conserving mass. *Transactions of the American Mathematical Society*, 367(6):3807–3828, 2015.

Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. 2017.

Murat A. Erdogdu and Rasa Hosseinzadeh. On the convergence of Langevin Monte Carlo: the interplay between tail growth and smoothness. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1776–1822. PMLR, 2021.

Murat A. Erdogdu, Rasa Hosseinzadeh, and Shunshi Zhang. Convergence of Langevin Monte Carlo in chi-squared and Rényi divergence. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 8151–8175. PMLR, 28–30 Mar 2022.

James Foster, Terry Lyons, and Harald Oberhauser. The shifted ODE method for underdamped Langevin MCMC. *arXiv preprint arXiv:2101.03446*, 2021.

James Foster, Goncalo dos Reis, and Calum Strange. High order splitting methods for SDEs satisfying a commutativity condition. *arXiv preprint arXiv:2210.17543*, 2022.

Arun Ganesh and Kunal Talwar. Faster differentially private samplers via Rényi divergence analysis of discretized Langevin MCMC. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7222–7233. Curran Associates, Inc., 2020.

Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Global convergence of stochastic gradient hamiltonian monte carlo for nonconvex stochastic optimization: Nonasymptotic performance bounds and momentum-based acceleration. *Operations Research*, 70(5):2931–2947, 2022.

Frédéric Hérau. Hypocoercivity and exponential time decay for the linear inhomogeneous relaxation Boltzmann equation. *Asymptotic Analysis*, 46(3-4):349–359, 2006.

Richard Holley and Daniel Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. *J. Statist. Phys.*, 46(5-6):1159–1194, 1987.

James G. Hooton. Compact Sobolev imbeddings on finite measure spaces. *Journal of Mathematical Analysis and Applications*, 83(2):570–581, 1981.

Lars Hörmander. Hypoelliptic second order differential equations. *Acta Mathematica*, 119:147–171, 1967.

Michael Johannes and Nicholas Polson. MCMC methods for continuous-time financial econometrics. In *Handbook of financial econometrics: applications*, pages 1–72. Elsevier, 2010.

Tim Johnston, Iosif Lytras, and Sotirios Sabanis. Kinetic Langevin MCMC Sampling Without Gradient Lipschitz Continuity–the Strongly Convex Case. *arXiv preprint arXiv:2301.08039*, 2023.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Ivan Kobyzev, Simon JD Prince, and Marcus A. Brubaker. Normalizing flows: an introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (11):3964–3979, 2020.

Andrey Kolmogorov. Zufallige bewegungen (zur theorie der Brownschen bewegung). *Annals of Mathematics*, pages 116–117, 1934.

Yi-An Ma, Niladri S. Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L. Bartlett, and Michael I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3): 1942 – 1992, 2021.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

Pierre Monmarché. High-dimensional MCMC with a standard splitting scheme for the underdamped Langevin diffusion. *Electronic Journal of Statistics*, 15(2):4117–4166, 2021.

Wenlong Mou, Nicolas Flammarion, Martin J. Wainwright, and Peter L. Bartlett. Improved bounds for discretization of Langevin diffusions: near-optimal rates without convexity. *Bernoulli*, 28(3): 1577–1601, 2022.

Yurii E. Nesterov. A method of solving a convex programming problem with convergence rate $O(\frac{1}{k^2})$. In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences, 1983.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703. PMLR, 2017.

Julien Roussel and Gabriel Stoltz. Spectral methods for Langevin dynamics and associated error estimates. *ESAIM: Mathematical Modelling and Numerical Analysis*, 52(3):1051–1083, 2018.

Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Cédric Villani. Limites hydrodynamiques de l'équation de Boltzmann. *Astérisque, SMF*, 282: 365–405, 2002.

Cédric Villani. Hypocoercivity. *Mem. Amer. Math. Soc.*, 202(950):iv+141, 2009.

Nisheeth K Vishnoi. An introduction to Hamiltonian Monte Carlo method for sampling. *arXiv preprint arXiv:2108.12107*, 2021.

Udo Von Toussaint. Bayesian inference in physics. *Reviews of Modern Physics*, 83(3):943, 2011.

Jun-Kun Wang and Andre Wibisono. Accelerating Hamiltonian Monte Carlo via Chebyshev integration time. *arXiv preprint arXiv:2207.02189*, 2022.

Difan Zou, Pan Xu, and Quanquan Gu. Stochastic gradient hamiltonian monte carlo methods with recursive variance reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

Appendix A. Explicit Form for the Underdamped Langevin Diffusion

Recall that we evolve (x_t, v_t) for time $t \in [kh, (k+1)h)$ explicitly according to the SDE (ULMC), which we repeat here for convenience:

$$\mathrm{d}x_t \coloneqq v_t \,\mathrm{d}t\,, \tag{A.1}$$

$$\mathrm{d}v_t \coloneqq -\gamma v_t + \nabla U(x_{kh})\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t\,. \tag{A.2}$$

Consequently, since we fix the position x_{kh} in the non-linear term, this permits an explicit solution

$$x_{(k+1)h} = x_{kh} + \gamma^{-1}\left(1-\exp(-\gamma h)\right) v_{kh} - \gamma^{-1}\left(h - \gamma^{-1}\left(1-\exp(-\gamma h)\right)\right)\nabla U(x_{kh}) + W_k^x\,, \tag{A.3}$$

$$v_{(k+1)h} = \exp(-\gamma h)\, v_{kh} - \gamma^{-1}\left(1-\exp(-\gamma h)\right)\nabla U(x_{kh}) + W_k^v\,, \tag{A.4}$$

where $(W_k^x, W_k^v)_{k\in\mathbb{N}}$ is an independent sequence of pairs of variables, where each pair has the joint distribution

$$\begin{bmatrix} W_k^x \\ W_k^v \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \frac{2}{\gamma}\left(h - \frac{2}{\gamma}\left(1-\exp(-\gamma h)\right) + \frac{1}{2\gamma}\left(1-\exp(-2\gamma h)\right)\right) & * \\ \frac{1}{\gamma}\left(1-2\exp(-\gamma h)+\exp(-2\gamma h)\right) & 1-\exp(-2\gamma h) \end{bmatrix}\right),$$

where $*$ is identical to the bottom left entry.

Appendix B. Continuous-Time Results

B.1. Entropic Hypocoercivity

Our proof of Lemma 5 is based on adapting the argument on the decay of a Lyapunov function from Ma et al. (2021) (based on entropic hypocoercivity, see Villani (2009)) and combining it with a time change argument (Dalalyan and Riou-Durand, 2020, Lemma 1). We provide the details below for completeness.

Proof of Lemma 5 First note that variables x_t, v_t with $\gamma = 2\sqrt{2L}$ following (ULMC) can be changed into $(\tilde{x}_t, \tilde{v}_t) = (x_{t\sqrt{\xi}}, \frac{1}{\sqrt{\xi}} v_{t\sqrt{\xi}})$, which satisfies the process given by

$$\begin{aligned} \mathrm{d}\tilde{x}_t &= \xi \tilde{v}_t \,\mathrm{d}t\,, \\ \mathrm{d}\tilde{v}_t &= -\xi\tilde{\gamma}\tilde{v}_t\,\mathrm{d}t - \nabla U(\tilde{x}_t)\,\mathrm{d}t + \sqrt{2\tilde{\gamma}}\,\mathrm{d}B_t\,, \end{aligned}$$

with $\tilde{\gamma} = 2$, $\xi = 2L$, which are the parameters satisfying Ma et al. (2021, Proposition 1). From that Proposition, we know that the Lyapunov functional given by

$$\tilde{\mathcal{F}}(\tilde{\mu}' \parallel \tilde{\mu}) = \mathsf{KL}(\tilde{\mu}' \parallel \tilde{\mu}) + \mathbb{E}_{\tilde{\mu}'}\left[\left\|\mathfrak{N}^{1/2}\nabla \log \frac{\tilde{\mu}'}{\tilde{\mu}}\right\|^2\right], \qquad \text{where } \mathfrak{N} = \frac{1}{L}\begin{bmatrix} 1/4 & 1/2 \\ 1/2 & 2 \end{bmatrix} \otimes I_d\,,$$

decays with $\partial_t \tilde{\mathcal{F}}(\tilde{\mu}_t \parallel \tilde{\mu}) \le -\frac{1}{10 C_{\mathsf{LSI}}} \tilde{\mathcal{F}}(\tilde{\mu}_t \parallel \tilde{\mu})$. Here the LSI constant does not change under our coordinate transform, but now $\tilde{\mu}_t$ represents the joint law of $(\tilde{x}_t, \tilde{v}_t)$, while the stationary measure has the form $\tilde{\mu}(\tilde{x}, \tilde{v}) \propto \pi(\tilde{x}) \times \exp(-\xi\,\|\tilde{v}\|^2/2)$. The statement of our theorem immediately follows by reversing our change of variables, which involves scaling up the gradients of the momenta by $\xi^{1/2}$, while the time is scaled down by $\xi^{1/2}$. ■

B.2. Contraction of ULMC

In this section, we prove a contraction result for ULMC and use this to deduce a log-Sobolev inequality along the trajectory of the underdamped Langevin diffusion. The mean of the next iterate of ULMC started at (x, v) is given by

$$F(x,v) := \Big(x + \frac{1-\exp(-\gamma h)}{\gamma}\, v - \frac{h - \gamma^{-1}\left(1-\exp(-\gamma h)\right)}{\gamma}\nabla U(x), \\ \exp(-\gamma h)\, v - \frac{1-\exp(-\gamma h)}{\gamma}\nabla U(x)\Big)\,.$$

We will use the change of coordinates

$$(\phi, \psi) := \mathcal{M}(x,v) := \Big(x, x + \frac{2}{\gamma}\, v\Big)\,.$$

In these new coordinates, the mean of the next iterate of ULMC started at (ϕ, ψ) is $\bar{F}(\phi, \psi)$, where $\bar{F} = \mathcal{M} \circ F \circ \mathcal{M}^{-1}$. Since $\mathcal{M}^{-1}(\phi, \psi) = (\phi, \frac{\gamma}{2}(\psi - \phi))$, we can explicitly write

$$\bar{F}(\phi, \psi) = \Big(\phi + \frac{1 - \exp(-\gamma h)}{2} (\psi - \phi) - \frac{h - \gamma^{-1}(1 - \exp(-\gamma h))}{\gamma} \nabla U(\phi), \\ \phi + \frac{1 + \exp(-\gamma h)}{2} (\psi - \phi) - \frac{h + \gamma^{-1}(1 - \exp(-\gamma h))}{\gamma} \nabla U(\phi) \Big) .$$

Lemma 12 *Consider the mapping $\bar{F} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$ defined above. Assume that $m I_d \preceq \nabla^2 U \preceq L I_d$. Then, for $h \lesssim 1$ and $\gamma = c\sqrt{L}$ for some $c \geq \sqrt{2}$, $\bar{F}$ is a contraction with parameter*

$$\|\bar{F}\|_{\mathrm{Lip}} \leq 1 - \frac{m}{\sqrt{2L}} h + O(Lh^2) .$$

Proof We compute the partial derivatives

$$\partial_\phi \bar{F}(\phi, \psi)_\phi = \frac{1 + \exp(-\gamma h)}{2} I_d - \frac{h - \gamma^{-1}(1 - \exp(-\gamma h))}{\gamma} \nabla^2 U(\phi) , \\ \partial_\phi \bar{F}(\phi, \psi)_\psi = \frac{1 - \exp(-\gamma h)}{2} I_d - \frac{h + \gamma^{-1}(1 - \exp(-\gamma h))}{\gamma} \nabla^2 U(\phi) , \\ \partial_\psi \bar{F}(\phi, \psi)_\phi = \frac{1 - \exp(-\gamma h)}{2} I_d , \\ \partial_\psi \bar{F}(\phi, \psi)_\psi = \frac{1 + \exp(-\gamma h)}{2} I_d .$$

Let $a \coloneqq \exp(-\gamma h)$ and $b \coloneqq \frac{2}{\gamma}(h + \gamma^{-1}(1 - \exp(-\gamma h)))$. Since

$$\frac{h - \gamma^{-1}(1 - \exp(-\gamma h))}{\gamma} = O(h^2) ,$$

we have

$$\|\nabla \bar{F}(\phi, \psi)\|_{\mathrm{op}} \leq \frac{1}{2} \Big\| \underbrace{\begin{bmatrix} (1+a) I_d & (1-a) I_d - b \nabla^2 U(\phi) \\ (1-a) I_d & (1+a) I_d \end{bmatrix}}_{=:A} \Big\|_{\mathrm{op}} + O(Lh^2) .$$

Then,

$$AA^\mathsf{T} = \begin{bmatrix} (1+a)^2 I_d + ((1-a) I_d - b \nabla^2 U(\phi))^2 & * \\ 2(1-a^2) I_d - (1+a) b \nabla^2 U(\phi) & \{(1-a)^2 + (1+a)^2\} I_d \end{bmatrix} ,$$

where the upper right entry is determined by symmetry. Since $1 - a = \Theta(\gamma h)$ and $b = O(h/\gamma)$, one can simplify this as follows:

$$\Big\| AA^\mathsf{T} - 2 \underbrace{\begin{bmatrix} (1+a^2) I_d & (1-a^2) I_d - b \nabla^2 U(\phi) \\ (1-a^2) I_d - b \nabla^2 U(\phi) & (1+a^2) I_d \end{bmatrix}}_{=:B} \Big\|_{\mathrm{op}} \\ \leq O\big(\frac{L^2 h^2}{\gamma^2} + Lh^2\big) .$$

One can check that the eigenvalues of the matrix B are $1 + a^2 \pm (1 - a^2 - b\lambda)$, where λ ranges over the eigenvalues of $\nabla^2 U(\phi)$. Hence, we can bound

$$\|B\|_{\mathrm{op}} \leq \max\{2a^2 + Lb, 2 - bm\}\,.$$

We note that

$$\begin{aligned} 2a^2 + Lb &= 2\exp(-2\gamma h) + \frac{2L\left(h + \gamma^{-1}\left(1 - \exp(-\gamma h)\right)\right)}{\gamma} \\ &= 2\left\{1 - 2\gamma h + \frac{2Lh}{\gamma} + O(\gamma^2 h^2 + Lh^2)\right\}. \end{aligned}$$

In order for this to be strictly smaller than 2, we must take $\gamma > \sqrt{L}$. We choose $\gamma = c\sqrt{L}$ for $c \geq \sqrt{2}$, in which case

$$\begin{aligned} \|B\|_{\mathrm{op}} &\leq 2\max\Big\{1 - c\sqrt{L}\,h,\ 1 - m\sqrt{\frac{2}{L}}\,h\Big\} + O(Lh^2) \\ &= 2\left(1 - m\sqrt{\frac{2}{L}}\,h\right) + O(Lh^2)\,. \end{aligned}$$

We deduce that

$$\|AA^{\mathsf{T}}\|_{\mathrm{op}} \leq 4\left(1 - m\sqrt{\frac{2}{L}}\,h\right) + O(Lh^2)$$

and therefore

$$\|\nabla \bar{F}(\phi, \psi)\|_{\mathrm{op}} \leq \sqrt{1 - m\sqrt{\frac{2}{L}}\,h} + O(Lh^2) \leq 1 - \frac{m}{\sqrt{2L}}\,h + O(Lh^2)\,.$$

■

The ULMC iterate is

$$(x_{(k+1)h}, v_{(k+1)h}) \overset{\mathsf{d}}{=} F(x_{kh}, v_{kh}) + \mathcal{N}(0, \Sigma)\,,$$

where Σ is the covariance of the Gaussian random vector in the LMC update. In the new coordinates, this iteration can be written

$$(\phi_{(k+1)h}, \psi_{(k+1)h}) \overset{\mathsf{d}}{=} \bar{F}(\phi_{kh}, \psi_{kh}) + \mathcal{N}(0, \mathcal{M}\Sigma\mathcal{M}^{\mathsf{T}})\,.$$

Writing $\mathcal{M}\Sigma\mathcal{M}^{\mathsf{T}} = \bar{\Sigma} \otimes I_d$, we can compute

$$\begin{aligned} \bar{\Sigma}_{1,1} &= \frac{2h}{\gamma} - \frac{3}{\gamma^2} + \frac{4\exp(-\gamma h)}{\gamma^2} - \frac{\exp(-2\gamma h)}{\gamma^2} = O(\gamma h^3)\,, \\ \bar{\Sigma}_{1,2} &= \frac{2h}{\gamma} - \frac{1}{\gamma^2} + \frac{\exp(-2\gamma h)}{\gamma^2} = O(h^2)\,, \\ \bar{\Sigma}_{2,2} &= \frac{2h}{\gamma} + \frac{5}{\gamma^2} - \frac{8\exp(-\gamma h)}{\gamma^2} + \frac{3\exp(-2\gamma h)}{\gamma^2} = \frac{4h}{\gamma^2} + O(h^2)\,. \end{aligned}$$

We conclude that

$$\|\bar{\Sigma}\|_{\mathrm{op}} \leq \frac{4h}{\gamma} + O(h^2)\,.$$

Hence, $C_{\mathsf{LSI}}(\mathcal{N}(0, \mathcal{M}\Sigma\mathcal{M}^{\mathsf{T}})) \leq \frac{4h}{\gamma^2} + O(h^2)$.

Proposition 13 *Let $\hat{\mu}_t^{\mathcal{M}} \coloneqq \mathrm{law}(\phi_t, \psi_t)$. Then, for all $\varepsilon > 0$, for all sufficiently small $h > 0$ (depending on ε), one has*

$$C_{\mathsf{LSI}}(\hat{\mu}_{Nh}^{\mathcal{M}}) \leq \Big(1 - \big(m\sqrt{\frac{2}{L}} - \varepsilon\big)\, h\Big)^N C_{\mathsf{LSI}}(\hat{\mu}_0^{\mathcal{M}}) + \frac{4}{2m - \varepsilon\sqrt{2L}} + O\big(\frac{h\sqrt{L}}{m}\big)\,.$$

Proof The LSI constant evolves according to

$$\begin{aligned} C_{\mathsf{LSI}}(\hat{\mu}_{(k+1)h}^{\mathcal{M}}) &\leq \|\bar{F}\|_{\mathrm{op}}^2\, C_{\mathsf{LSI}}(\hat{\mu}_{kh}^{\mathcal{M}}) + C_{\mathsf{LSI}}\big(\mathcal{N}(0, \mathcal{M}\Sigma\mathcal{M}^{\mathsf{T}})\big) \\ &\leq \Big(1 - m\sqrt{\frac{2}{L}}\, h + O(Lh^2)\Big)\, C_{\mathsf{LSI}}(\hat{\mu}_{kh}^{\mathcal{M}}) + \frac{4h}{\gamma} + O(h^2)\,. \end{aligned}$$

For h sufficiently small, we have

$$C_{\mathsf{LSI}}(\hat{\mu}_{(k+1)h}^{\mathcal{M}}) \leq \Big(1 - \big(m\sqrt{\frac{2}{L}} - \varepsilon\big)\, h\Big)\, C_{\mathsf{LSI}}(\hat{\mu}_{kh}^{\mathcal{M}}) + \frac{4h}{\gamma} + O(h^2)\,.$$

Iterating,

$$C_{\mathsf{LSI}}(\hat{\mu}_{Nh}^{\mathcal{M}}) \leq \Big(1 - \big(m\sqrt{\frac{2}{L}} - \varepsilon\big)\, h\Big)^N C_{\mathsf{LSI}}(\hat{\mu}_0^{\mathcal{M}}) + \frac{4}{2m - \varepsilon\sqrt{2L}} + O\big(\frac{h\sqrt{L}}{m}\big)\,.$$

This completes the proof. ■

Corollary 14 *Let $\mu_t^{\mathcal{M}}$ now denote the law of the continuous-time underdamped Langevin diffusion with $\gamma = c\sqrt{L}$ for $c \geq \sqrt{2}$ in the (ϕ, ψ) coordinates. Then,*

$$C_{\mathsf{LSI}}(\mu_t^{\mathcal{M}}) \leq \exp\Big(-m\sqrt{\frac{2}{L}}\, t\Big)\, C_{\mathsf{LSI}}(\mu_0^{\mathcal{M}}) + \frac{2}{m}\,.$$

Proof In the preceding proposition, let $h \searrow 0$ while $Nh \to t$, and then let $\varepsilon \searrow 0$. ■

Appendix C. Discretization Analysis

We consider the discretization used in Ma et al. (2021), with the following differential form:

$$\begin{aligned} d\hat{x}_t &= \hat{v}_t\, \mathrm{d}t\,, \\ d\hat{v}_t &= -\gamma\hat{v}_t\, \mathrm{d}t - \nabla U(\hat{x}_{kh})\, \mathrm{d}t + \sqrt{2\gamma}\, \mathrm{d}B_t\,, \end{aligned}$$

and we define the variable $\hat{w}_t$ as the tuple $(\hat{x}_t, \hat{v}_t)$, for $t \in [kh, (k+1)h]$.

C.1. Technical Lemmas

Theorem 15 (Girsanov's Theorem, Adapted from Oksendal (2013, Theorem 8.6.8)) *Consider stochastic processes $(x_t)_{t\geq 0}$, $(b_t^P)_{t\geq 0}$, $(b_t^Q)_{t\geq 0}$ adapted to the same filtration, and $\sigma \in \mathbb{R}^{d\times d}$ any constant, possibly degenerate, matrix. Let P_T and Q_T be probability measures on the path space $C([0,T];\mathbb{R}^d)$ such that $(w_t)_{t\geq 0}$ evolves according to*

$$\begin{aligned} \mathrm{d}w_t &= b_t^P\,\mathrm{d}t + \sigma\,\mathrm{d}B_t^P \qquad \text{under } P_T\,,\\ \mathrm{d}w_t &= b_t^Q\,\mathrm{d}t + \sigma\,\mathrm{d}B_t^Q \qquad \text{under } Q_T\,, \end{aligned}$$

where B^P is a P_T-Brownian motion and B^Q is a Q_T-Brownian motion. Furthermore, suppose there exists a process $(u_t)_{t\geq 0}$ such that

$$\sigma\, u_t = b_t^P - b_t^Q\,,$$

and

$$\mathbb{E}^{Q_T}\exp\Big(\frac{1}{2}\int_0^T \|u_s\|^2\,\mathrm{d}s\Big) < \infty\,,$$

Consequently, if we define $\sigma^\dagger$ as the Moore–Penrose pseudo-inverse of σ, then by the previous supposition we have $u_t = \sigma^\dagger\,(b_t^P - b_t^Q)$. Then,

$$\frac{\mathrm{d}P_T}{\mathrm{d}Q_T} = \exp\Big(\int_0^T \langle \sigma^\dagger\,(b_t^P - b_t^Q), \mathrm{d}B_t^Q\rangle - \frac{1}{2}\int_0^T \|\sigma^\dagger\,(b_t^P - b_t^Q)\|^2\,\mathrm{d}t\Big)\,.$$

In fact, we will only need the following corollary.

Corollary 16 *For any event $\mathcal{E}$ and $q \geq 1$,*

$$\mathbb{E}^{Q_T}\big[\big(\frac{\mathrm{d}P_T}{\mathrm{d}Q_T}\big)^q\,\mathbb{1}_{\mathcal{E}}\big] \leq \sqrt{\mathbb{E}\Big[\exp\Big(2q^2\int_0^T \|\sigma^\dagger\,(b_t^P - b_t^Q)\|^2\,\mathrm{d}t\Big)\,\mathbb{1}_{\mathcal{E}}\Big]}\,.$$

Proof Using Cauchy–Schwarz, and then Itô's Lemma, we find

$$\begin{aligned} \mathbb{E}^{Q_T}\big[\big(\frac{\mathrm{d}P_T}{\mathrm{d}Q_T}\big)^q\,\mathbb{1}_{\mathcal{E}}\big] &= \mathbb{E}^{Q_T}\Big[\exp\Big(q\int_0^T \langle \sigma^\dagger\,(b_t^P - b_t^Q), \mathrm{d}B_t^Q\rangle - \frac{q}{2}\int_0^T \|\sigma^\dagger\,(b_t^P - b_t^Q)\|^2\,\mathrm{d}t\Big)\,\mathbb{1}_{\mathcal{E}}\Big]\\ &\leq \sqrt{\mathbb{E}^{Q_T}\Big[\exp\Big((2q^2-q)\int_0^T \|\sigma^\dagger\,(b_t^P - b_t^Q)\|^2\,\mathrm{d}t\Big)\,\mathbb{1}_{\mathcal{E}}\Big]}\\ &\qquad \times \underbrace{\sqrt{\mathbb{E}^{Q_T}\Big[\exp\Big(2q\int_0^T \langle \sigma^\dagger\,(b_t^P - b_t^Q), \mathrm{d}B_t^Q\rangle - 2q^2\int_0^T \|\sigma^\dagger\,(b_t^P - b_t^Q)\|^2\,\mathrm{d}t\Big)\,\mathbb{1}_{\mathcal{E}}\Big]}}_{=1}\\ &\leq \sqrt{\mathbb{E}^{Q_T}\Big[\exp\Big(2q^2\int_0^T \|\sigma^\dagger\,(b_t^P - b_t^Q)\|^2\,\mathrm{d}t\Big)\,\mathbb{1}_{\mathcal{E}}\Big]}\,. \end{aligned}$$

Here, we used the fact that $t \mapsto \exp(\int_0^t \langle u_\tau, \mathrm{d}B_\tau\rangle - \frac{1}{2}\int_0^t \|u_\tau\|^2\,\mathrm{d}\tau)$ is a local martingale. ■

We can identify the following for the process (x_t, v_t):

$$\sigma = \begin{bmatrix} 0 & 0 \\ 0 & \sqrt{2\gamma}\, I_d \end{bmatrix}, \qquad b_t^P = \begin{bmatrix} v_t \\ -\gamma v_t - \nabla U(x_t) \end{bmatrix}, \qquad b_t^Q = \begin{bmatrix} v_t \\ -\gamma v_t - \nabla U(x_{\lfloor t/h \rfloor h}) \end{bmatrix}.$$

In this case, $\|\sigma^\dagger\,(b_t^P - b_t^Q)\| \equiv \frac{1}{\sqrt{2\gamma}}\,\|\nabla U(x_{\lfloor t/h \rfloor h}) - \nabla U(x_t)\|$.

We also adapt the following Lemmas without proof from Chewi et al. (2021).

Lemma 17 (Change of Measure, from Chewi et al. (2021, Lemma 21)) *Let μ, ν be probability measures and let E be any event. Then,*

$$\mu(E) \le \nu(E) + \sqrt{\chi_2(\mu \,\|\, \nu)\, \nu(E)}\,.$$

In particular, if μ and ν are probability measures on $\mathbb{R}^d$ and

$$\nu\{\|\cdot\| \ge R_0 + \eta\} \le C \exp(-c\eta^2) \qquad \text{for all } \eta \ge 0\,,$$

where $C \ge 1$, then

$$\mu\Big\{\|\cdot\| \ge R_0 + \sqrt{\frac{1}{c}\,\mathcal{R}_2(\mu \,\|\, \nu)} + \eta\Big\} \le 2C \exp\big(-\frac{c\eta^2}{2}\big) \qquad \text{for all } \eta \ge 0\,.$$

Lemma 18 *Let $(B_t)_{t\ge 0}$ be a standard Brownian motion in $\mathbb{R}^d$. Then, if $\lambda \ge 0$ and $h \le 1/(4\lambda)$,*

$$\mathbb{E} \exp\big(\lambda \sup_{t\in[0,h]} \|B_t\|^2\big) \le \exp(6dh\lambda)\,.$$

In particular, for all $\eta \ge 0$,

$$\mathbb{P}\big\{\sup_{t\in[0,h]} \|B_t\| \ge \eta\big\} \le 3\exp\big(-\frac{\eta^2}{6dh}\big)\,.$$

Lemma 19 (Ganesh and Talwar (2020, Lemma 14)) *Let $Y > 0$ be a random variable. Assume that for all $0 < \delta < 1/2$ there exists an event $\mathcal{E}_\delta$ with probability at least $1 - \delta$ such that*

$$\mathbb{E}[Y^2 \mid \mathcal{E}_\delta] \le \frac{v}{\delta^\xi}$$

for some $\xi < 1$. Then, $\mathbb{E}\, Y \le 4\sqrt{v}$.

Lemma 20 (Matrix Grönwall Inequality) *Let $x : \mathbb{R}_+ \to \mathbb{R}^d$, and $c \in \mathbb{R}^d$, $A \in \mathbb{R}^{d\times d}$, where A has non-negative entries. Suppose that the following inequality is satisfied componentwise:*

$$x(t) \le c + \int_0^t A x(s)\,\mathrm{d}s\,, \qquad \text{for all } t \ge 0\,. \tag{C.1}$$

Then, the following inequality holds, where $I_d \in \mathbb{R}^{d\times d}$ is the d-dimensional identity matrix:

$$x(t) \le (AA^\dagger\, e^{At} - AA^\dagger + I_d)\, c\,, \tag{C.2}$$

where $A^\dagger$ is the Moore–Penrose pseudo-inverse of A (when A is invertible, this is equivalent to the standard inverse).

Proof This is a special case of Chandra and Davis (1976, Main Theorem). ■

C.2. Movement Bound for ULMC

We next prove a movement bound for the continuous-time Langevin diffusion. The following lemma is a standard fact about the concentration of the norm of a Gaussian vector (see, e.g., Boucheron et al., 2013, Theorem 5.5).

Lemma 21 (Concentration of the Norm) *The following concentration holds: for all $\eta \geq 0$,*

$$\rho(\|\cdot\| \geq \sqrt{d} + \eta) \leq \exp\Big(-\frac{\eta^2}{2}\Big) .$$

Note that $\|v_t - v_0\|$ is of size $\mathcal{O}(\sqrt{dt})$, due to the Brownian motion component of the momentum variable v; this is the same order as the size of the increment of the overdamped Langevin diffusion. However, if we consider the increment in the x-coordinate only, we obtain the following bound.

Lemma 22 *Let $(x_t, v_t)_{t\geq 0}$ denote the continuous-time underdamped Langevin diffusion started at (x_0, v_0), and assume that the gradient ∇U of the potential satisfies $\nabla U(0) = 0$ and is Hölder continuous (satisfies (2.3)). Also, assume that $h \lesssim L^{-1/2} \wedge \gamma^{-1}$ and $0 \leq \lambda \lesssim \frac{1}{\gamma^s d^s h^{3s}}$. Then,*

$$\log \mathbb{E} \exp\big(\lambda \sup_{t\in[0,h]} \|x_t - x_0\|^{2s}\big) \lesssim \big(L^{2s} h^{4s} \left(1 + \|x_0\|^{2s^2}\right) + h^{2s} \|v_0\|^{2s} + \gamma^s d^s h^{3s}\big)\, \lambda .$$

Proof For the interpolant times, we will use Grönwall's matrix inequality (Lemma 20), with the following equation for x:

$$\begin{aligned}
\|x_t - x_0\| &\leq \Big\| \int_0^t v_\tau \,\mathrm{d}\tau \Big\| \leq h \|v_0\| + \Big\| \int_0^t (v_\tau - v_0) \,\mathrm{d}\tau \Big\| \\
&\leq h \|v_0\| + \Big\| \int_0^t \int_0^\tau \gamma v_{\tau'} \,\mathrm{d}\tau' \,\mathrm{d}\tau \Big\| + \Big\| \int_0^t \int_0^\tau \nabla U(x_{\tau'}) \,\mathrm{d}\tau' \,\mathrm{d}\tau \Big\| \\
&\quad + \Big\| \int_0^t \int_0^\tau \sqrt{2\gamma} \,\mathrm{d}B_{\tau'} \,\mathrm{d}\tau \Big\| \\
&\leq h \|v_0\| + \gamma h \Big(h \|v_0\| + \int_0^t \|v_\tau - v_0\| \,\mathrm{d}\tau \Big) + Lh^2 \\
&\quad + Lh \Big(h \|x_0\|^s + \int_0^t \|x_\tau - x_0\| \,\mathrm{d}\tau \Big) + \sqrt{2\gamma}\, h \sup_{t\in[0,h]} \|B_t\| .
\end{aligned}$$

Here we use the Hölder property of ∇U along with $\|x\|^s \leq 1 + \|x\|$. Likewise for v:

$$\begin{aligned}
\|v_t - v_0\| &\leq \Big\| \int_0^t \gamma v_\tau \,\mathrm{d}\tau \Big\| + \Big\| \int_0^t \nabla U(x_\tau) \,\mathrm{d}\tau \Big\| + \Big\| \int_0^t \sqrt{2\gamma} \,\mathrm{d}B_\tau \Big\| \\
&\leq \gamma \Big(h \|v_0\| + \int_0^t \|v_\tau - v_0\| \,\mathrm{d}\tau \Big) + Lh + L \Big(h \|x_0\|^s + \int_0^t \|x_\tau - x_0\| \,\mathrm{d}\tau \Big) \\
&\quad + \sqrt{2\gamma} \sup_{t\in[0,h]} \|B_t\| .
\end{aligned}$$

Consequently, we can use the matrix form of Grönwall's inequality (Lemma 20). While applying that Lemma, let $c = c_1 + c_2$ with c_1, c_2 to be given. First, for c_1:

$$A = \begin{bmatrix} Lh & \gamma h \\ L & \gamma \end{bmatrix}, \qquad c_1 = \begin{bmatrix} Lh^2 \|x_0\|^s + \gamma h^2 \|v_0\| + Lh^2 + \sqrt{2\gamma}\, h \sup_{t\in[0,h]} \|B_t\| \\ Lh \|x_0\|^s + \gamma h \|v_0\| + Lh + \sqrt{2\gamma} \sup_{t\in[0,h]} \|B_t\| \end{bmatrix}.$$

Noting that c_1 lies in the image space of A so that $AA^\dagger c_1 = c_1$, and similarly observing that $\exp(At)\, c_1$ belongs to the image space of A (using the power series representation of the matrix exponential), we obtain for this first component:

$$\begin{aligned} &\sup_{t\in[0,h]} \|x_0 - x_t\| \\ &\quad \le h \exp\big((Lh+\gamma)\, h\big) \Big(\gamma h \|v_0\| + Lh \|x_0\|^s + Lh + \sqrt{2\gamma} \sup_{t\in[0,h]} \|B_t\|\Big) + c_2 \text{ term} \\ &\quad \le 2h \Big(\gamma h \|v_0\| + Lh \|x_0\|^s + Lh + \sqrt{2\gamma} \sup_{t\in[0,h]} \|B_t\|\Big) + c_2 \text{ term}\,, \end{aligned}$$

where in the second line we take $h \lesssim \frac{1}{\sqrt{L}+\gamma}$. Now, taking

$$c_2 = \begin{bmatrix} h \|v_0\| \\ 0 \end{bmatrix},$$

we find the following (where $\mathbf{v}_{(1)}$ denotes the first component of a vector $\mathbf{v}$):

$$((AA^\dagger (e^{Ah} - I_{2d}) + I_{2d})\, c_2)_{(1)} = \frac{Lhe^{(Lh+\gamma)\, h} + \gamma}{Lh + \gamma}\, h \|v_0\|.$$

Finally, for $h \lesssim \frac{1}{\sqrt{L}+\gamma}$, this can be bounded by $2h \|v_0\|$. Using Lemma 18 and plugging this into the expression completes the proof. ∎

C.3. Sub-Gaussianity of the Iterates

Similarly to Chewi et al. (2021), we introduce a modified potential in order to prove sub-Gaussianity of the iterates of ULMC. Firstly, we consider a modified distribution in the x-coordinate, with parameter $a := (\beta, S)$ for some $S, \beta \ge 0$:

$$\pi^{(a)} \propto \exp(-U^{(a)})\,, \qquad U^{(a)}(x) := U(x) + \frac{\beta}{2} (\|x\| - S)_+^2\,. \tag{C.3}$$

The modified potential satisfies the following properties.

Lemma 23 (Properties of the Modified Potential, Chewi et al. (2021, Lemma 23)) *Consider $\pi^{(a)}$ and $U^{(a)}$ defined as in (C.3). Assume that $\nabla U(0) = 0$ and that ∇U satisfies (2.3). Then, the following assertions hold.*

1. *(sub-Gaussian tail bound) Assume that S is chosen so that $\pi(B(0,S)) \ge 1/2$. Then, for all $\eta \ge 0$,*

$$\pi^{(a)}\{\|\cdot\| \ge S + \eta\} \le 2 \exp\big(-\frac{\beta\eta^2}{2}\big)\,.$$

2. *(gradient growth) The gradient $\nabla U^{(a)}$ satisfies*

$$\|\nabla U^{(a)}(x)\| \le L + (\beta + L)\,\|x\|\,.$$

Then, letting $\{(x_t^{(a)}, v_t^{(a)})\}_{t\ge 0}$ be the solution to the underdamped Langevin diffusion with potential $U^{(a)}$ and $\mu^{(a)} := \pi^{(a)} \otimes \rho$, the following lemma holds:

Lemma 24 *Assume that $h \lesssim (\beta + L)^{-1/2} \wedge \gamma^{-1} \wedge d^{-1/2}$, and $\beta \le 1$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{t \le Nh} \|x_t^{(a)}\| - S \lesssim (\beta + L)\, S h^2 + \sqrt{\frac{1}{\beta}\,\mathcal{R}_2(\mu_0^{(a)} \parallel \mu^{(a)})} + \sqrt{\frac{1}{\beta}\log\frac{16N}{\delta}}\,.$$

Proof We can use the change of measure lemma (Lemma 17) together with the sub-Gaussian tail bounds in Lemmas 21, 23 to see that with probability at least $1-\delta$, the following events hold simultaneously:

$$\max_{k\le N} \|x_{kh}^{(a)}\| \le S + \sqrt{\frac{2}{\beta}\,\mathcal{R}_2(\mu_0^{(a)} \parallel \mu^{(a)})} + \sqrt{\frac{4}{\beta}\log\frac{8N}{\delta}}$$
$$\max_{k\le N} \|v_{kh}^{(a)}\| \le \sqrt{d} + \sqrt{2\,\mathcal{R}_2(\mu_0^{(a)} \parallel \mu^{(a)})} + \sqrt{4\log\frac{4N}{\delta}}\,.$$

Here we use a union bound together with the monotonicity of $t \mapsto \mathcal{R}_2(\mu_t^{(a)} \parallel \mu^{(a)})$ in t.

For the interpolant times, we will use Grönwall's matrix inequality, with the following inequality for x:

$$\begin{aligned}
\|x_{kh}^{(a)} - x_{kh+t}^{(a)}\| &\le h\,\|v_{kh}^{(a)}\| + \Big\|\int_0^t\int_0^\tau \gamma v_{kh+\tau'}^{(a)}\,\mathrm{d}\tau'\,\mathrm{d}\tau\Big\| + \Big\|\int_0^t\int_0^\tau \nabla U^{(a)}(x_{kh+\tau'}^{(a)})\,\mathrm{d}\tau'\,\mathrm{d}\tau\Big\| \\
&\quad + \Big\|\int_0^t\int_0^\tau \sqrt{2\gamma}\,\mathrm{d}B_{kh+\tau'}\,\mathrm{d}\tau\Big\| \\
&\le h\,\|v_{kh}^{(a)}\| + \gamma h\left(h\,\|v_{kh}^{(a)}\| + \int_0^t \|v_{kh+\tau}^{(a)} - v_{kh}^{(a)}\|\,\mathrm{d}\tau\right) + Lh^2 \\
&\quad + (\beta+L)\,h\left(h\,\|x_{kh}^{(a)}\| + \int_0^t \|x_{kh+\tau}^{(a)} - x_{kh}^{(a)}\|\,\mathrm{d}\tau\right) \\
&\quad + \sqrt{2\gamma}\,h \sup_{\tau\in[0,h]} \|B_{kh+\tau} - B_{kh}\|\,.
\end{aligned}$$

Likewise,

$$\begin{aligned}
\|v_{kh}^{(a)} - v_{kh+t}^{(a)}\| &\le \Big\|\int_0^t \gamma v_{kh+\tau}^{(a)}\,\mathrm{d}\tau\Big\| + \Big\|\int_0^t \nabla U^{(a)}(x_{kh+\tau}^{(a)})\,\mathrm{d}\tau\Big\| + \Big\|\int_0^t \sqrt{2\gamma}\,\mathrm{d}B_{kh+\tau}\Big\| \\
&\le \gamma\left(h\,\|v_{kh}^{(a)}\| + \int_0^t \|v_{kh+\tau}^{(a)} - v_{kh}^{(a)}\|\,\mathrm{d}\tau\right) + Lh \\
&\quad + (\beta+L)\left(h\,\|x_{kh}^{(a)}\| + \int_0^t \|x_{kh+\tau}^{(a)} - x_{kh}^{(a)}\|\,\mathrm{d}\tau\right) \\
&\quad + \sqrt{2\gamma} \sup_{\tau\in[0,h]} \|B_{kh+\tau} - B_{kh}\|\,.
\end{aligned}$$

Consequently, we can apply the matrix Grönwall inequality analogously to how we did in Lemma 20 with $c = c_1 + c_2$ denoting the following matrices:

$$A = \begin{bmatrix} (\beta+L)\,h & \gamma h \\ (\beta+L) & \gamma \end{bmatrix},$$
$$c_1 = \begin{bmatrix} (\beta+L)\,h^2\,\|x_{kh}^{(a)}\| + \gamma h^2\,\|v_{kh}^{(a)}\| + Lh^2 + \sqrt{2\gamma}\,h \sup_{t\in[0,h]}\|B_{kh+t} - B_{kh}\| \\ (\beta+L)\,h\,\|x_{kh}^{(a)}\| + \gamma h\,\|v_{kh}^{(a)}\| + Lh + \sqrt{2\gamma}\sup_{t\in[0,h]}\|B_{kh+t} - B_{kh}\| \end{bmatrix},$$
$$c_2 = \begin{bmatrix} h\,\|v_{kh}^{(a)}\| \\ 0 \end{bmatrix}.$$

Note that c_1 here is again in the image space of A, so that $(AA^\dagger - I_2)\,c = 0$. Finally, after calculating the matrix exponential we find

$$\begin{aligned}\sup_{t\le h}\|x_{kh}^{(a)} - x_{kh+t}^{(a)}\| &\le h\exp\big((\beta+L)\,h^2 + \gamma h\big)\Big((\beta+L)\,h\,\|x_{kh}^{(a)}\| + \gamma h\,\|v_{kh}^{(a)}\| + Lh \\ &\qquad + \sqrt{2\gamma}\sup_{t\le h}\|B_{kh+t} - B_{kh}\|\Big) \\ &\quad + h\,\frac{(\beta+L)\exp\big((\beta+L)\,h^2 + \gamma h\big)h + \gamma}{(\beta+L)\,h + \gamma}\,\|v_{kh}^{(a)}\| \\ &\le 2h\left((\beta+L)\,h\,\|x_{kh}^{(a)}\| + \|v_{kh}^{(a)}\| + Lh + \sqrt{2\gamma}\sup_{t\le h}\|B_{kh+t} - B_{kh}\|\right),\end{aligned}$$

where in the second line we take $h \lesssim \frac{1}{(\beta+L)^{1/2}} \wedge \frac{1}{\gamma}$. Note that this is also entirely analogous to the calculation in Lemma 22.

Subsequently, we can take a union bound to obtain for any S_1, S_2,

$$\begin{aligned}&\mathbb{P}\Big\{\sup_{t\in[0,Nh]}\|x_t^{(a)}\| \ge \eta\Big\} \\ &\quad\le \mathbb{P}\Big\{\max_{k=0,1,\dots N-1}\|x_{kh}^{(a)}\| \ge S_1\Big\} + \mathbb{P}\Big\{\max_{k=0,1,\dots N-1}\|v_{kh}^{(a)}\| \ge S_2\Big\} \\ &\qquad + \sum_{k=0}^{N-1}\mathbb{P}\Big\{\sup_{t\in[0,h]}\|x_{kh+t}^{(a)} - x_{kh}^{(a)}\| \ge \eta - S_1\Big\} \\ &\quad\le \mathbb{P}\Big\{\max_{k=0,1,\dots N-1}\|x_{kh}^{(a)}\| \ge S_1\Big\} + \mathbb{P}\Big\{\max_{k=0,1,\dots N-1}\|v_{kh}^{(a)}\| \ge S_2\Big\} \\ &\qquad + \sum_{k=0}^{N-1}\mathbb{P}\Big\{\sup_{t\in[0,h]}\sqrt{2\gamma}\,\|B_{kh+t} - B_{kh}\| \ge \frac{\eta - S_1}{2h} - (\beta+L)\,S_1 h - S_2 - Lh\Big\}.\end{aligned}$$

Subsequently, taking respectively $S_1 = S + \sqrt{\frac{2}{\beta}\,\mathcal{R}_2(\mu_0^{(a)} \parallel \mu^{(a)})} + \sqrt{\frac{4}{\beta}\log\frac{8N}{\delta}}$, $S_2 = \sqrt{d} + \sqrt{2\,\mathcal{R}_2(\mu_0^{(a)} \parallel \mu^{(a)})} + \sqrt{4\log\frac{4N}{\delta}}$, we use the Brownian motion tail bound (Lemma 18) to get with probability $1 - 2\delta$:

$$\sup_{t\le Nh}\|x_t^{(a)}\| - S_1 \lesssim (\beta+L)\,S_1 h^2 + S_2 h + Lh^2 + \sqrt{\gamma d h^3 \log\frac{3N}{\delta}}\,.$$

If we assume that $\beta \leq 1$ and $h \lesssim \frac{1}{\sqrt{d}}$, then we can further simplify this bound to yield

$$\sup_{t \leq Nh} \|x_t^{(a)}\| - S \lesssim (\beta + L)\, S h^2 + \sqrt{\frac{1}{\beta}\, \mathcal{R}_2(\mu_0^{(a)} \parallel \mu^{(a)})} + \sqrt{\frac{1}{\beta} \log \frac{8N}{\delta}}\,.$$

This concludes the proof. ■

To transfer this sub-Gaussianity to the original underdamped Langevin process, we consider the following bound on the chi-squared divergence between these two processes.

Proposition 25 *Let $Q_T, Q_T^{(a)}$ represent respectively the laws on the path space of the original and modified diffusions, under the same initialization μ_0. Then, if $\beta \lesssim \frac{\gamma}{T} \wedge L$ and $h \lesssim (\beta + L)^{-1/2} \wedge \gamma^{-1} \wedge d^{-1/2}$, then*

$$\mathcal{R}_2(Q_T \parallel Q_T^{(a)}) \lesssim \frac{\beta^2 L^2 S^2 T h^4}{\gamma} + \frac{\beta T}{\gamma} \left(\mathcal{R}_2(\mu_0 \parallel \mu^{(a)}) + \log N\right).$$

Proof Conditioning on the event in Lemma 24, which we denote by $\mathcal{E}_\delta$ for some $\delta \leq 1/2$, then using Girsanov's theorem (Corollary 16) we get (for some sufficiently small h so that Novikov's condition is satisfied)

$$\begin{aligned}
\log \mathbb{E}\Big[\Big(\frac{\mathrm{d}Q_T}{\mathrm{d}Q_T^{(a)}}\Big)^4 \mathbb{1}_{\mathcal{E}_\delta}\Big] &\leq \frac{1}{2} \log \mathbb{E}\Big[\exp\Big(\frac{16}{\gamma} \int_0^T \|\nabla U(x_t^{(a)}) - \nabla U^{(a)}(x_t^{(a)})\|^2 \,\mathrm{d}t\Big)\, \mathbb{1}_{\mathcal{E}_\delta}\Big] \\
&\leq \frac{1}{2} \log \mathbb{E}\Big[\exp\Big(\frac{16\beta^2}{\gamma} \int_0^T \big(\|x_t^{(a)}\| - S\big)_+^2 \,\mathrm{d}t\Big)\, \mathbb{1}_{\mathcal{E}_\delta}\Big] \\
&\lesssim \frac{\beta^2 T}{\gamma} \Big\{(\beta + L)^2 S^2 h^4 + \frac{1}{\beta}\, \mathcal{R}_2(\mu_0 \parallel \mu^{(a)}) + \frac{1}{\beta} \log \frac{16N}{\delta}\Big\}.
\end{aligned}$$

If we take $\beta \lesssim \gamma/T$ and that $L \geq \beta$, we can use Lemma 19 to get

$$\mathcal{R}_2(Q_T \parallel Q_T^{(a)}) = \log \mathbb{E}\Big[\Big(\frac{\mathrm{d}Q_T}{\mathrm{d}Q_T^{(a)}}\Big)^2\Big] \lesssim \frac{\beta^2 L^2 S^2 T h^4}{\gamma} + \frac{\beta T}{\gamma} \left(\mathcal{R}_2(\mu_0 \parallel \mu^{(a)}) + \log N\right).$$

This concludes the proof. ■

Proposition 26 *Consider the continuous time diffusion $(x_t, v_t)_{t \geq 0}$ initialized at μ_0. For $h \lesssim (\beta + L)^{-1/2} \wedge \gamma^{-1} \wedge d^{-1/2}$, $S \asymp \mathfrak{m}$, and $\beta \asymp \frac{\gamma}{T}$, for $\delta \in (0, 1/2)$, the following holds with probability $1 - \delta$:*

$$\begin{aligned}
\max_{k=0,\ldots,N-1} \|x_{kh}\| &\lesssim \mathfrak{m} + \sqrt{\frac{T}{\gamma} \left(\mathcal{R}_2(\mu_0 \parallel \mu^{(a)}) + \log \frac{N}{\delta}\right)}, \\
\max_{k=0,\ldots,N-1} \|v_{kh}\| &\lesssim \sqrt{d} + \sqrt{\mathcal{R}_2(\mu_0 \parallel \mu^{(a)}) + \log \frac{N}{\delta}}\,.
\end{aligned}$$

Proof Recall from the proof of Lemma 24 that with probability $1-\delta$,

$$\max_{k=0,\dots N-1} \|x_{kh}^{(a)}\| \lesssim S + \sqrt{\frac{1}{\beta}\mathcal{R}_2(\mu_0 \parallel \mu^{(a)})} + \sqrt{\frac{1}{\beta}\log\frac{8N}{\delta}}\,.$$

In particular, this immediately implies that the following holds: for $\eta \geq 0$,

$$\mathbb{P}\Big(\max_{k=0,\dots N-1} \|x_{kh}^{(a)}\| \gtrsim S + \sqrt{\frac{1}{\beta}\mathcal{R}_2(\mu_0 \parallel \mu^{(a)})} + \sqrt{\frac{1}{\beta}\log\frac{8N}{\delta}} + \eta\Big) \lesssim N\exp(-c\beta\eta^2)\,,$$

for a universal constant $c > 0$.

Then, using the change of measure (Lemma 17) together with the bound in Proposition 25, choosing $S \asymp \mathfrak{m}$, we get with probability $1-\delta$

$$\begin{aligned}\max_{k=0,\dots N-1} \|x_{kh}\| &\lesssim S + \sqrt{\frac{1}{\beta}\mathcal{R}_2(\mu_0 \parallel \mu^{(a)})} + \sqrt{\frac{1}{\beta}\mathcal{R}_2(Q_T \parallel Q_T^{(a)})} + \sqrt{\frac{1}{\beta}\log\frac{N}{\delta}} \\ &\lesssim \mathfrak{m} + \sqrt{\frac{1}{\beta}\left(\mathcal{R}_2(\mu_0 \parallel \mu^{(a)}) + \log\frac{N}{\delta}\right) + \frac{\beta L^2 T h^4 \mathfrak{m}^2}{\gamma}}\,.\end{aligned}$$

We choose $\beta \asymp \gamma/T$ so that

$$\max_{k=0,\dots N-1} \|x_{kh}\| \lesssim \mathfrak{m} + \sqrt{\frac{T}{\gamma}\left(\mathcal{R}_2(\mu_0 \parallel \mu^{(a)}) + \log\frac{N}{\delta}\right)}\,.$$

Finally, combining this with a union bound to control $\|v_{kh}\|$ from Lemma 21, we get the Proposition. ■

C.4. Completing the Discretization Proof

We proceed by following the proof of Chewi et al. (2021).

Proof [Proof of Proposition 11] Let $\{x_t\}_{t\geq 0}$ follow the continuous-time process. Let P_T, Q_T denote the measures on the path space corresponding to the interpolated process and the continuous-time diffusion respectively, with both being initialized at $\mu_0 = \pi_0 \otimes \mathcal{N}(0, I_d)$. Then, define

$$G_t \coloneqq \frac{1}{\sqrt{2\gamma}}\int_0^t \langle \nabla U(x_\tau) - \nabla U(x_{\lfloor \tau/h \rfloor h}), \mathrm{d}B_\tau\rangle - \frac{1}{4\gamma}\int_0^t \|\nabla U(x_\tau) - \nabla U(x_{\lfloor \tau/h\rfloor h})\|^2\,\mathrm{d}\tau.$$

From Girsanov's theorem (Theorem 15), we obtain immediately using Itô's formula

$$\begin{aligned}\mathbb{E}_{Q_T}\left[\left(\frac{\mathrm{d}P_T}{\mathrm{d}Q_T}\right)^q\right] - 1 &= \mathbb{E}\exp(qG_T) - 1 \\ &= \frac{q\,(q-1)}{4\gamma}\,\mathbb{E}\int_0^T \exp(qG_t)\,\|\nabla U(x_t) - \nabla U(x_{\lfloor t/h\rfloor h})\|^2\,\mathrm{d}t \\ &\leq \frac{q^2}{4\gamma}\int_0^T \sqrt{\mathbb{E}[\exp(2qG_t)]\,\mathbb{E}[\|\nabla U(x_t) - \nabla U(x_{\lfloor t/h\rfloor h})\|^4]}\,\mathrm{d}t\,.\end{aligned}$$

Bounding these terms individually, we first use Corollary 16 and (2.3) to get

$$\mathbb{E}\exp(2qG_t) \le \sqrt{\mathbb{E}\exp\Big(\frac{4q^2}{\gamma}\int_0^t \|\nabla U(x_r) - \nabla U(x_{\lfloor r/h\rfloor h})\|^2\,\mathrm{d}r\Big)} \le \sqrt{\mathbb{E}\exp\Big(\frac{4L^2q^2}{\gamma}\int_0^t \|x_r - x_{\lfloor r/h\rfloor h}\|^{2s}\,\mathrm{d}r\Big)}\,.$$

Let us now condition on the event

$$\mathcal{E}_{\delta,kh} \coloneqq \Big\{\max_{j=0,1,\dots,k-1}\|x_{kh}\| \le R^x_\delta,\ \max_{j=0,1,\dots,k-1}\|v_{kh}\| \le R^v_\delta\Big\}\,.$$

By Proposition 26, we can have $\mathbb{P}(\mathcal{E}_{\delta,kh}) \ge 1-\delta$ while choosing

$$R^x_\delta \lesssim \mathfrak{m} + \sqrt{\frac{\gamma}{T}\left(\mathcal{R}_2(\mu_0 \parallel \mu^{(a)}) + \log\frac{N}{\delta}\right)}\,,$$
$$R^v_\delta \lesssim \sqrt{d} + \sqrt{\mathcal{R}_2(\mu_0 \parallel \mu^{(a)}) + \log\frac{N}{\delta}}\,.$$

We proceed to bound our desired quantity through some careful steps.

One step error. Consider first the error on a single interval $[0,h]$. If we presume $h \lesssim (\gamma^{1-s}/(L^2d^sq^2))^{1/(1+3s)}$, Lemma 22 implies

$$\log\mathbb{E}\exp\Big(\frac{8L^2q^2}{\gamma}\int_0^h\|x_t-x_0\|^2\,\mathrm{d}t\Big) \le \log\mathbb{E}\exp\Big(\frac{8L^2hq^2}{\gamma}\sup_{t\in[0,h]}\|x_t-x_0\|^2\Big)$$
$$\lesssim \frac{L^{2+2s}h^{1+4s}q^2}{\gamma}\left(1+\|x_0\|^{2s^2}\right) + \frac{L^2h^{1+2s}q^2}{\gamma}\|v_0\|^{2s} + \frac{L^2d^sh^{1+3s}q^2}{\gamma^{1-s}}\,.$$

Iteration. If we let $\{\mathcal{F}_t\}_{t\ge0}$ denote the filtration, then writing $H_t = \int_0^t\|x_r - x_{\lfloor r/h\rfloor h}\|^2\,\mathrm{d}r$, we can condition on $\mathcal{F}_{(N-1)h}$ and iterate our one step bound.

$$\log\mathbb{E}\Big[\exp\Big(\frac{8L^2q^2}{\gamma}H_{Nh}\Big)\mathbb{1}_{\mathcal{E}_{\delta,Nh}}\Big]$$
$$\le \log\mathbb{E}\Big[\exp\Big(\frac{8L^2q^2}{\gamma}H_{(N-1)h} + \mathcal{O}\Big(\frac{L^{2+2s}h^{1+4s}q^2}{\gamma}\left(1+\|x_{(N-1)h}\|^{2s^2}\right) + \frac{L^2h^{1+2s}q^2}{\gamma}\|v_{(N-1)h}\|^{2s} + \frac{L^2d^sh^{1+3s}q^2}{\gamma^{1-s}}\Big)\Big)\mathbb{1}_{\mathcal{E}_{\delta,Nh}}\Big]$$
$$\le \log\mathbb{E}\Big[\exp\Big(\frac{8L^2q^2}{\gamma}H_{(N-1)h}\Big)\mathbb{1}_{\mathcal{E}_{\delta,(N-1)h}}\Big] + \mathcal{O}\Big(\frac{L^{2+2s}h^{1+4s}q^2}{\gamma}(R^x_\delta)^{2s^2} + \frac{L^2h^{1+2s}q^2}{\gamma}(R^v_\delta)^{2s} + \frac{L^2d^sh^{1+3s}q^2}{\gamma^{1-s}}\Big)\,.$$

We now make additional simplifying assumptions to obtain more interpretable bounds: we assume $\gamma/T \le 1$ and $h \lesssim \frac{1}{L}\,(1 \wedge \frac{d^{1/2}}{\mathfrak{m}^s})$. With these assumptions,

$$\begin{aligned}
&\log \mathbb{E}\Big[\exp\Big(\frac{8L^2q^2}{\gamma}\, H_{Nh}\Big)\, \mathbb{1}_{\mathcal{E}_{\delta,Nh}}\Big] \\
&\quad \le \log \mathbb{E}\Big[\exp\Big(\frac{8L^2q^2}{\gamma}\, H_{(N-1)h}\Big)\, \mathbb{1}_{\mathcal{E}_{\delta,(N-1)h}}\Big] \\
&\qquad + \mathcal{O}\Big(\frac{L^2h^{1+2s}q^2}{\gamma}\,\big(d + \mathcal{R}_2(\mu_0 \,\|\, \mu^{(a)}) + \log\frac{N}{\delta}\big)^s\Big)\,.
\end{aligned}$$

Completing this iteration yields

$$\log \mathbb{E}\Big[\exp\Big(\frac{8L^2q^2}{\gamma}\, H_{Nh}\Big)\, \mathbb{1}_{\mathcal{E}_{\delta,Nh}}\Big] \lesssim \frac{L^2Th^{2s}q^2}{\gamma}\,\big(d + \mathcal{R}_2(\mu_0 \,\|\, \mu^{(a)}) + \log\frac{N}{\delta}\big)^s\,.$$

Finally, applying Lemma 19 when

$$h \lesssim_s \frac{\gamma^{1/(2s)}}{L^{1/s}T^{1/(2s)}q^{1/s}} \tag{C.4}$$

(where $\lesssim_s$ hides an s-dependent constant), we find

$$\log \mathbb{E}\Big[\exp\Big(\frac{4L^2q^2}{\gamma}\, H_{Nh}\Big)\Big] \lesssim 1 + \frac{L^2Th^{2s}q^2}{\gamma}\,\big(d + \mathcal{R}_2(\mu_0 \,\|\, \mu^{(a)}) + \log N\big)^s\,.$$

It remains to choose the appropriate step size h which makes this whole quantity $\lesssim 1$. In particular, it suffices to choose

$$h \lesssim \widetilde{\mathcal{O}}_s\Big(\frac{\gamma^{1/(2s)}}{L^{1/s}T^{1/(2s)}q^{1/s}\,(d + \mathcal{R}_2(\mu_0 \,\|\, \mu^{(a)}))^{1/2}}\Big)\,. \tag{C.5}$$

Second term. It remains to bound the other term in our original expression. From Lemma 22, we obtain

$$\mathbb{E}[\exp(\lambda\,\|x_{kh+t} - x_{kh}\|^{2s}) \mid w_{kh}] \lesssim 1\,,$$

so long as λ is chosen to be appropriately small, i.e.,

$$\lambda \asymp \frac{1}{\gamma^s d^s h^{3s}} \wedge \frac{1}{L^{2s}h^{4s}\,(1 + \|x_{kh}\|)^{2s^2}} \wedge \frac{1}{h^{2s}\,\|v_{kh}\|^{2s}}\,.$$

This immediately implies a tail bound: for $\eta \ge 0$,

$$\mathbb{P}\{\|x_{kh+t} - x_{kh}\|^{4s} \ge \eta \mid w_{kh}\} \lesssim \exp(-\lambda\sqrt{\eta})\,.$$

Integrating, we get

$$\begin{aligned}
\sqrt{\mathbb{E}[\|\nabla U(x_t) - \nabla U(x_{kh})\|^4]} &\le L^2\sqrt{\mathbb{E}[\|x_t - x_{kh}\|^{4s}]} \lesssim L^2\sqrt{\mathbb{E}\,\frac{1}{\lambda^2}} \\
&\lesssim L^2\gamma^s d^s h^{3s} + L^{2+2s}h^{4s}\sqrt{1 + \mathbb{E}[\|x_{kh}\|^{4s^2}]} + L^2h^{2s}\sqrt{\mathbb{E}[\|v_{kh}\|^{4s}]}\,.
\end{aligned}$$

We can estimate the expectations by integration of our previous tail bound (Proposition 26):

$$\begin{aligned}\sqrt{\mathbb{E}[\|\nabla U(x_t) - \nabla U(x_{kh})\|^4]} &\lesssim L^2\gamma^s d^s h^{3s} + L^{2+2s}h^{4s}\left(\mathfrak{m} + \frac{T}{\gamma}\left(\mathcal{R}_2(\mu_0 \parallel \mu^{(a)}) + \log N\right)\right)^{2s^2} \\ &\quad + L^2 h^{2s}\left(d + \mathcal{R}_2(\mu_0 \parallel \mu^{(a)}) + \log N\right)^s \\ &\leq \widetilde{\mathcal{O}}\Big(L^2 h^{2s}\left(d + \mathcal{R}_2(\mu_0 \parallel \mu^{(a)})\right)^s\Big),\end{aligned}$$

provided that $h \leq \widetilde{\mathcal{O}}(\frac{1}{L}(\frac{d^{1/2}}{\mathfrak{m}^s} \wedge \mathcal{R}_2(\frac{\gamma\mathcal{R}_2}{T})^{s/2}))$, where $\mathcal{R}_2 = \mathcal{R}_2(\mu_0 \parallel \mu^{(a)})$. In our applications, this condition is not dominant and can be disregarded.

Combining the bounds. Finally, we can combine each of these steps to find that, provided (C.5) for the step size holds,

$$\mathbb{E}_{Q_T}\Big[\Big(\frac{\mathrm{d}P_T}{\mathrm{d}Q_T}\Big)^q\Big] - 1 \leq \widetilde{\mathcal{O}}\Big(\frac{Tq^2}{\gamma}L^2h^{2s}\left(d + \mathcal{R}_2(\mu_0 \parallel \mu^{(a)})\right)^s\Big).$$

Finally, the following step size condition suffices to bound the Rényi divergence by ε^2:

$$h \lesssim \widetilde{\mathcal{O}}_s\Big(\frac{\gamma^{1/(2s)}\varepsilon^{1/s}}{L^{1/s}T^{1/(2s)}q^{1/s}\left(d + \mathcal{R}_2(\mu_0 \parallel \mu^{(a)})\right)^{1/2}}\Big).$$

This completes the proof. ∎

Appendix D. Proof of the Main Results

Firstly, we collect some results on feasible initializations from Chewi et al. (2021). Recall that $\pi^{(a)}$ is the modified distribution introduced in Appendix C.3. Let

$$\pi_0 = \mathcal{N}(0, \varsigma I_d),$$

where $\varsigma = (2L + \beta)^{-1}$ is the variance of the Gaussian, and $\beta \asymp 1/T$ is the parameter appearing in the modified potential. The choice of T will be assumption dependent, and we collect the conditions below under our main assumptions:

$$T = \begin{cases} \widetilde{\Theta}(\frac{L+d}{\mathfrak{q}(\gamma)}) & \pi \text{ satisfies (PI)} \\ \widetilde{\Theta}(\sqrt{L}C_{\mathsf{LSI}}) & \pi \text{ satisfies (LSI), or is strongly log-concave,} \end{cases}$$

where $\mathfrak{q}(\gamma)$ is defined in (3.2).

Lemma 27 (Adapted from Chewi et al. (2021, Appendix A)) *Suppose that π satisfies (PI) and the Hölder continuity condition (2.3), as well as $\nabla U(0) = 0$, $U(0) - \min U \lesssim d$. Then the following two properties hold for $\pi_0 = \mathcal{N}(0, (2L + \beta)^{-1}I_d)$, where β is the parameter appearing in the modified potential:*

$$\begin{aligned}\mathcal{R}_q(\pi_0 \parallel \pi) &\leq \widetilde{\mathcal{O}}(\beta + L + d), \\ \mathcal{R}_q(\pi_0 \parallel \pi^{(a)}) &\leq \widetilde{\mathcal{O}}(\beta + L + d).\end{aligned}$$

Proof Apply either Chewi et al. (2021, Lemma 30) or Chewi et al. (2021, Lemma 31). ■

From our analysis we take $\beta \lesssim L$, and if moreover $L \lesssim d$ then it is reasonable to expect that $\mathcal{R}_q(\pi_0 \,\|\, \pi), \mathcal{R}_q(\pi_0 \,\|\, \pi^{(a)}) \leq \widetilde{\mathcal{O}}(d)$. Let $\mu_0 = \pi_0 \otimes \rho$, so that $\mathcal{R}_q(\mu_0 \,\|\, \mu) = \mathcal{R}_q(\pi_0 \,\|\, \pi)$, and similarly $\mathcal{R}_q(\mu_0 \,\|\, \mu^{(a)}) = \mathcal{R}_q(\pi_0 \,\|\, \pi^{(a)})$.

The following lemma gives a bound on the value of the Fisher information at initialization.

Lemma 28 *Under the conditions of the previous lemma, the initialization $\mu_0 = \pi_0 \otimes \rho$ also satisfies* $\mathsf{FI}(\mu_0 \,\|\, \mu) \lesssim Ld + L^{1-s}d^s$.

Proof Note that as $\nabla U(0) = 0$, $\|\nabla \log \pi(x)\|^2 = \|\nabla U(x)\|^2 \leq L^2 \|x\|^{2s}$. Secondly, π_0 satisfies $\mathbb{E}_{x\sim\pi_0}[\|x\|^2] \lesssim d/L$. Hence,

$$\begin{aligned}\mathsf{FI}(\mu_0 \,\|\, \mu) = \mathbb{E}_{\pi_0}\left[\left\|\nabla \log \frac{\pi_0}{\pi}\right\|^2\right] &\leq \mathbb{E}_{x\sim\pi_0}[\|\nabla U(x) - (2L+\beta)\, x\|^2] \\ &\lesssim L^2\, \mathbb{E}_{x\sim\mu_0}[\|x\|^2 + \|x\|^{2s}] \lesssim Ld + L^{1-s}d^s\,,\end{aligned}$$

where we used Jensen's inequality in the last step. ■

D.1. Poincaré Inequality

Proof [Proof of Theorem 9] The continuous-time result from Lemma 8 states that

$$T \gtrsim \frac{1}{\mathfrak{q}(\gamma)} \log \frac{\chi_2(\mu_0 \,\|\, \mu)}{\varepsilon^2} \implies \chi_2(\mu_T \,\|\, \mu) \leq \varepsilon^2\,.$$

Noting that there exists a feasible initialization such that $\log \chi_2(\mu_0 \,\|\, \mu) \leq \widetilde{\mathcal{O}}(L+d)$, then this is satisfied if we choose $T = \widetilde{\mathcal{O}}(\frac{1}{\mathfrak{q}(\gamma)}\,(L + d + \log \frac{1}{\varepsilon}))$. This also shows that $\mathcal{R}_2(\mu_T \,\|\, \mu) = \log(1 + \chi_2(\mu_T \,\|\, \mu)) \lesssim \varepsilon^2$ for $\varepsilon \lesssim 1$.

Note the following decomposition (weak triangle inequality) for the Rényi divergence (see, e.g., Mironov, 2017, Proposition 11):

$$\mathcal{R}_q(P_1 \,\|\, P_2) \leq \frac{q - 1/\mathfrak{c}}{q-1}\, \mathcal{R}_{\mathfrak{c}q}(P_1 \,\|\, P_3) + \mathcal{R}_{\mathfrak{d}(q-1/\mathfrak{c})}(P_3 \,\|\, P_2),$$

for any valid Hölder conjugate pair $\mathfrak{c}, \mathfrak{d}$, i.e., $\frac{1}{\mathfrak{c}} + \frac{1}{\mathfrak{d}} = 1$, $\mathfrak{c}, \mathfrak{d} > 1$, and any three probability distributions P_1, P_2, P_3.

In our case, we let $q = 2 - \xi$ and $\mathfrak{d}(q - 1/\mathfrak{c}) = 2$, so that after solving for $\mathfrak{c}, \mathfrak{d}$, we get the following for $\xi \leq 1/2$:

$$\mathcal{R}_{2-\xi}(P_1 \,\|\, P_2) \leq 2\mathcal{R}_{2/\xi}(P_1 \,\|\, P_3) + \mathcal{R}_2(P_3 \,\|\, P_2)\,.$$

Consequently, let $P_1 = \hat{\mu}_{Nh}$, $P_2 = \mu$, $P_3 = \mu_{Nh}$, and combining this result with the discretization bound of Proposition 11, we then obtain

$$\mathcal{R}_{2-\xi}(\hat{\mu}_{Nh} \,\|\, \mu) \lesssim \mathcal{R}_{2/\xi}(\hat{\mu}_{Nh} \,\|\, \mu_{Nh}) + \mathcal{R}_2(\mu_{Nh} \,\|\, \mu) \lesssim \varepsilon^2\,,$$

so long as

$$h = \widetilde{\Theta}\Big(\frac{\gamma^{1/(2s)}\varepsilon^{1/s}\xi^{1/s}\mathfrak{q}(\gamma)^{1/(2s)}}{L^{1/s}d^{1/2}\,(L\vee d)^{1/(2s)}}\Big)\,,$$
$$N = \widetilde{\Theta}\Big(\frac{L^{1/s}d^{1/2}\,(L\vee d)^{1+1/(2s)}}{\gamma^{1/(2s)}\varepsilon^{1/s}\xi^{1/s}\mathfrak{q}(\gamma)^{1+1/(2s)}}\Big)\,.$$

This completes the proof. ■

D.2. Log-Sobolev Inequality

D.2.1. KL DIVERGENCE

Proof [Proof of Theorem 6] We provide the following Theorem in the twisted coordinates (ϕ, ψ), which were used in Lemma 10. Consider the decomposition of the KL using Cauchy–Schwarz:

$$\begin{aligned}
\mathsf{KL}(\hat{\mu}_T^{\mathcal{M}} \,\|\, \mu^{\mathcal{M}}) &= \int \log \frac{\hat{\mu}_T^{\mathcal{M}}}{\mu^{\mathcal{M}}}\,\mathrm{d}\hat{\mu}_T^{\mathcal{M}} \\
&= \mathsf{KL}(\hat{\mu}_T^{\mathcal{M}} \,\|\, \mu_T^{\mathcal{M}}) + \int \log \frac{\mu_T^{\mathcal{M}}}{\mu^{\mathcal{M}}}\,\mathrm{d}\hat{\mu}_T^{\mathcal{M}} \\
&= \mathsf{KL}(\hat{\mu}_T^{\mathcal{M}} \,\|\, \mu_T^{\mathcal{M}}) + \mathsf{KL}(\mu_T^{\mathcal{M}} \,\|\, \mu^{\mathcal{M}}) + \int \log \frac{\mu_T^{\mathcal{M}}}{\mu^{\mathcal{M}}}\,\mathrm{d}(\hat{\mu}_T^{\mathcal{M}} - \mu_T^{\mathcal{M}}) \\
&\le \mathsf{KL}(\hat{\mu}_T^{\mathcal{M}} \,\|\, \mu_T^{\mathcal{M}}) + \mathsf{KL}(\mu_T^{\mathcal{M}} \,\|\, \mu^{\mathcal{M}}) + \sqrt{\chi^2(\hat{\mu}_T^{\mathcal{M}} \,\|\, \mu_T^{\mathcal{M}}) \times \mathsf{var}_{\mu_T^{\mathcal{M}}}\Big(\log \frac{\mu_T^{\mathcal{M}}}{\mu^{\mathcal{M}}}\Big)}\,.
\end{aligned}$$

Using the log-Sobolev inequality of the iterates via Lemma 10, we find (through the implication that a log-Sobolev inequality implies a Poincaré inequality with the same constant)

$$\mathsf{var}_{\mu_T^{\mathcal{M}}}\Big(\log \frac{\mu_T^{\mathcal{M}}}{\mu^{\mathcal{M}}}\Big) \le C_{\mathsf{LSI}}(\mu_T^{\mathcal{M}})\,\mathbb{E}_{\mu_T^{\mathcal{M}}}\Big[\Big\|\nabla \log \frac{\mu_T^{\mathcal{M}}}{\mu^{\mathcal{M}}}\Big\|^2\Big]\,,$$

where we substitute $\log \frac{\mu_T^{\mathcal{M}}}{\mu^{\mathcal{M}}}$ for the function in (PI). Here, $C_{\mathsf{LSI}}(\mu_T^{\mathcal{M}}) \lesssim 1/m$ for all $t \ge 0$.

Since $\mu^{\mathcal{M}} = \mathcal{M}_{\#}\mu$, then $\mu^{\mathcal{M}}(\phi, \psi) \propto \mu(\mathcal{M}^{-1}(\phi, \psi))$. Therefore,

$$\nabla \log \mu^{\mathcal{M}} = (\mathcal{M}^{-1})^{\mathsf{T}}\,\nabla \log \mu \circ \mathcal{M}^{-1}\,,$$

and similarly for $\nabla \log \mu_T^{\mathcal{M}}$. This yields the expression

$$\mathbb{E}_{\mu_T^{\mathcal{M}}}\Big[\Big\|\nabla \log \frac{\mu_T^{\mathcal{M}}}{\mu^{\mathcal{M}}}\Big\|^2\Big] = \mathbb{E}_{\mu_T}\Big[\Big\|(\mathcal{M}^{-1})^{\mathsf{T}}\,\nabla \log \frac{\mu_T}{\mu}\Big\|^2\Big]\,.$$

Also, one has

$$\mathcal{M}^{-1}\,(\mathcal{M}^{-1})^{\mathsf{T}} = \begin{bmatrix} 1 & -\gamma/2 \\ -\gamma/2 & \gamma^2/2 \end{bmatrix}.$$

For $c_0 > 0$ and $\mathfrak{M}$ defined in Appendix B.1, we have

$$L\,\mathfrak{M} - c_0\,\mathcal{M}^{-1}\,(\mathcal{M}^{-1})^{\mathsf{T}} = \begin{bmatrix} 1/4 - c_0 & \sqrt{L}\,(1/\sqrt{2} + c_0\sqrt{2}) \\ \sqrt{L}\,(1/\sqrt{2} + c_0\sqrt{2}) & L\,(4 - c_0) \end{bmatrix}.$$

The determinant is $L\,((\frac{1}{4} - c_0)\,(4 - c_0) - (\frac{1}{\sqrt{2}} + c_0\sqrt{2})^2) > 0$ for $c_0 > 0$ sufficiently small. This shows that $\mathcal{M}^{-1}\,(\mathcal{M}^{-1})^{\mathsf{T}} \preceq c_0^{-1} L\,\mathfrak{M}$, and therefore

$$\mathbb{E}_{\mu_T^{\mathcal{M}}}\Big[\Big\|\nabla \log \frac{\mu_T^{\mathcal{M}}}{\mu^{\mathcal{M}}}\Big\|^2\Big] \lesssim L\,\mathsf{FI}_{\mathfrak{M}}(\mu_T \parallel \mu)\,.$$

Here we define

$$\mathsf{FI}_{\mathfrak{M}}(\mu' \parallel \mu) \coloneqq \mathbb{E}_{\mu'}\big[\big\|\mathfrak{M}^{1/2}\,\nabla \log \frac{\mu'}{\mu}\big\|^2\big]$$

The decay of the Fisher information via Lemma 5 allows us to set

$$T \gtrsim C_{\mathsf{LSI}}\sqrt{L}\log\Big(\frac{\kappa}{\varepsilon^2}\,\big(\mathsf{KL}(\mu_0 \parallel \mu) + \mathsf{FI}_{\mathfrak{M}}(\mu_0 \parallel \mu)\big)\Big) \implies \mathsf{var}_{\mu_T^{\mathcal{M}}}\Big(\log \frac{\mu_T^{\mathcal{M}}}{\mu^{\mathcal{M}}}\Big) \lesssim \varepsilon^2\,.$$

The same choice of T also ensures that $\mathsf{KL}(\mu_T^{\mathcal{M}} \parallel \mu^{\mathcal{M}}) \le \varepsilon^2$. From our initialization (Lemma 28), we can naively estimate using that

$$\mathsf{FI}_{\mathfrak{M}}(\mu_0 \parallel \mu) \lesssim \frac{1}{L}\,\mathsf{FI}(\pi_0 \parallel \pi) \lesssim d\,,$$

and $\mathsf{KL}(\mu_0 \parallel \mu) \lesssim d \log \kappa$, so that our condition on T is (with $C_{\mathsf{LSI}} \le m^{-1}$)

$$T \ge \widetilde{\mathcal{O}}\Big(\frac{\sqrt{L}}{m}\log\frac{\kappa d}{\varepsilon^2}\Big)\,.$$

Recall as well that this requires $\gamma \asymp \sqrt{L}$. For the remaining $\chi^2(\hat{\mu}_T \parallel \mu_T)$ and $\mathsf{KL}(\hat{\mu}_T \parallel \mu_T)$ terms, we invoke Proposition 11 with the value of $T = Nh$ specified and desired accuracy ε, and with $q = 2$ and $s = 1$, which consequently yields

$$h = \widetilde{\Theta}\Big(\frac{\varepsilon m^{1/2}}{L d^{1/2}}\Big)\,,$$

with

$$N = \widetilde{\Theta}\Big(\frac{\kappa^{3/2} d^{1/2}}{\varepsilon}\Big)$$

(using $N = T/h$). ∎

D.2.2. TV DISTANCE

Proof [Proof of Theorem 7] Notice first that the TV distance is a proper metric, and therefore satisfies the triangle inequality. Subsequently, by two applications of Pinsker's inequality,

$$\begin{aligned}\|\hat{\mu}_{Nh} - \mu\|_{\mathsf{TV}} &\leq \|\hat{\mu}_{Nh} - \mu_{Nh}\|_{\mathsf{TV}} + \|\mu_{Nh} - \mu\|_{\mathsf{TV}} \\ &\lesssim \sqrt{\mathsf{KL}(\hat{\mu}_{Nh} \parallel \mu_{Nh})} + \sqrt{\mathsf{KL}(\mu_{Nh} \parallel \mu)}\,.\end{aligned}$$

These terms can be bounded separately. Analogous to the proof of the prior theorem, using Lemma 5, it suffices to take

$$T \geq \widetilde{\mathcal{O}}\Big(C_{\mathsf{LSI}}\sqrt{L}\log\frac{d}{\varepsilon^2}\Big)\,,$$

and for the other term, it suffices to use Proposition 11 with any value of q, $\gamma \asymp \sqrt{L}$ which combined with the requirement on T yields:

$$h = \widetilde{\Theta}\Big(\frac{\varepsilon}{C_{\mathsf{LSI}}^{1/2} L d^{1/2}}\Big)\,,$$

with

$$N = \widetilde{\Theta}\Big(\frac{C_{\mathsf{LSI}}^{3/2} L^{3/2} d^{1/2}}{\varepsilon}\Big)\,,$$

(using $N = T/h$). ■