Proceedings of Machine Learning Research vol 178:1–28, 2022 35th Annual Conference on Learning Theory

Towards a Theory of Non-Log-Concave Sampling: First-Order Stationarity Guarantees for Langevin Monte Carlo

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

Sinho Chewi SCHEWI@MIT.EDU
Massachusetts Institute of Technology

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

Adil Salim ADILSALIM@MICROSOFT.COM
Microsoft Research

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

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

For the task of sampling from a density $\pi \propto \exp(-V)$ on $\mathbb{R}^d$, where V is possibly non-convex but L-gradient Lipschitz, we prove that averaged Langevin Monte Carlo outputs a sample with ε-relative Fisher information after $O(L^2 d^2/\varepsilon^2)$ iterations. This is the sampling analogue of complexity bounds for finding an ε-approximate first-order stationary points in non-convex optimization and therefore constitutes a first step towards the general theory of non-log-concave sampling. We discuss numerous extensions and applications of our result; in particular, it yields a new state-of-the-art guarantee for sampling from distributions which satisfy a Poincaré inequality.

Keywords: Fisher information, Langevin Monte Carlo, non-log-concave sampling, Poincaré inequality

1. Introduction

Consider the canonical task of sampling from a density $\pi \propto \exp(-V)$ on $\mathbb{R}^d$, given query access to the gradients of V. In the case where V is strongly convex and smooth, this task is well-studied, with a number of works giving precise and non-asymptotic complexity bounds which scale polynomially in the problem parameters. In contrast, there are comparatively few works which study the case when V is non-convex. In this work, we take a first step towards developing a general theory of non-log-concave sampling by formulating the sampling analogue of *stationary point analysis*, which has been highly successful in the non-convex optimization (Nesterov et al., 2018).

Classically, the *Langevin diffusion*, the solution to the stochastic differential equation

$$\mathrm{d}z_t = -\nabla V(z_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \tag{1}$$

has π as its unique stationary distribution and converges to it as $t \to \infty$ under mild conditions. Here, $(B_t)_{t\geq 0}$ is a standard d-dimensional Brownian motion. Discretizing this stochastic process with step size $h > 0$ yields the standard *Langevin Monte Carlo* (LMC) algorithm

$$x_{(k+1)h} \coloneqq x_{kh} - h\,\nabla V(x_{kh}) + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,. \tag{LMC}$$

Several extensions of LMC have been considered in the literature. For instance, a stochastic gradient can be used as an estimate of the "full" gradient $\nabla V(x_{kh})$ at each iteration.

Although LMC and its extensions are ostensibly sampling algorithms, they find applications in optimization. Indeed, LMC and its extensions can be viewed as a variant of (stochastic) gradient descent in which Gaussian noise is explicitly injected in the (stochastic) gradient in each iteration. As explored, for example, in Raginsky et al. (2017) and Jin et al. (2021), the presence of noise allows the iteration to escape local minima and allows for establishing global non-asymptotic convergence guarantees on well-behaved yet non-convex objectives.

Perhaps surprisingly, the connection between optimization and sampling also goes in the other direction: the theory of optimization can be used to understand the performance of sampling algorithms. On a superficial level, this is anticipated because the Langevin diffusion (1) is simply a standard gradient flow to which a Brownian noise has been added. However, there is a much deeper connection, due to Jordan et al. (1998), which interprets the Langevin diffusion as an *exact* gradient flow in the space of probability measures equipped with the geometry of optimal transport, where the objective functional is the Kullback–Leibler (KL) divergence $\mathsf{KL}(\cdot \parallel \pi)$. This perspective has spurred researchers to provide novel optimization-inspired analyses of sampling (Bernton, 2018; Wibisono, 2018; Durmus et al., 2019).

For example, the Wasserstein gradient of $\mathsf{KL}(\cdot \parallel \pi)$ at μ is $\nabla \ln(\mu/\pi)$, and the calculation rules for gradient flows imply that if π_t denotes the law of the Langevin diffusion (1) at time t, then $\partial_t \mathsf{KL}(\pi_t \parallel \pi) = -\mathbb{E}_{\pi_t}[\|\nabla \ln(\pi_t/\pi)\|^2]$ (see Ambrosio et al., 2008; Villani, 2009; Santambrogio, 2015). As this quantity is important in what follows, we explicitly write $\mathsf{FI}(\mu \parallel \pi) := \mathbb{E}_\mu[\|\nabla \ln(\mu/\pi)\|^2]$ for the (relative) *Fisher information* of μ w.r.t. π. If V is convex (resp. strongly convex), then it turns out that the objective functional $\mathsf{KL}(\cdot \parallel \pi)$ is convex (resp. strongly convex) in the Wasserstein geometry, which in turn implies that $\mathsf{KL}(\pi_t \parallel \pi)$ decays to zero at the rate $O(1/t)$ (resp. exponentially fast).

In the case when V is non-convex, however, less is known. Of course, just like non-convex optimization, it is in general impossible to obtain polynomial sampling guarantees for non-log-concave distributions. Recently, Vempala and Wibisono (2019); Chewi et al. (2021a); Ma et al. (2021) study tractable cases of non-log-concave sampling in which the target π satisfies a *functional inequality*, such as the *log-Sobolev inequality* (LSI). Indeed, if LSI holds, then $\mathsf{FI}(\mu \parallel \pi) \gtrsim \mathsf{KL}(\mu \parallel \pi)$ for all μ. In light of the Wasserstein calculus described above, this is the analogue of the *gradient domination* condition (or *Polyak–Łojasiewicz inequality*) in non-convex optimization: $\|\nabla V(x)\|^2 \gtrsim V(x) - \min V$ (Lojasiewicz, 1963; Polyak, 1963; Karimi et al., 2016). Furthermore, Durmus and Moulines (2017); Cheng et al. (2018); Li et al. (2019); Majka et al. (2020); Erdogdu and Hosseinzadeh (2021); He et al. (2022) study tractable classes of non-log-concave sampling based on certain tail-growth conditions. However, the assumptions made in all the above works are very far from capturing the breadth of non-log-concave sampling.

Instead, in general non-convex optimization, the standard approach is to prove convergence to a *stationary point* of the objective function, or from a more quantitative perspective, to determine the complexity of obtaining a point x satisfying $\|\nabla V(x)\|^2 \le \varepsilon$. This complexity is typically $O(1/\varepsilon)$ (Nesterov et al., 2018). Following this paradigm, we propose to use the Fisher information as the sampling analogue of the squared norm of the gradient. Our main result (Theorem 4) establishes that under the sole assumption that ∇V is L-gradient Lipschitz, an averaged version of the LMC algorithm (LMC) outputs a sample whose law μ satisfies $\mathsf{FI}(\mu \parallel \pi) \le \varepsilon$ after $O(L^2 d^2/\varepsilon^2)$ iterations. Intuitively, the Fisher information captures the rapid local mixing of the Langevin diffusion near modes of the distribution π, while ignoring the metastability effects which occur between the modes (Bovier et al., 2002, 2004, 2005). We give an illustrative example in Section 2 which expands upon this intuition.

1.1. Paper organization and contributions

The rest of the paper is organized as follows. In Section 2, we provide intuitions on Fisher information guarantees in sampling. In Section 3, we formally define the Fisher information, and in Section 4, we state our main result in Theorem 4. In Section 5, we consider applications of our main result:

- We show the weak convergence of averaged LMC with decaying step size (Section 5.1).
- We provide new sampling guarantees in total variation distance under Poincaré inequality (Section 5.2). These guarantees are competitive with very recent results by Chewi et al. (2021a) (in fact, our dimension dependence is substantially better).
- We show an accelerated convergence result for LMC when the Hessian of the potential is also Lipschitz and when the potential satisfies a polynomial tail growth condition (Section 5.3).

In Section 6, we consider extensions of LMC involving stochastic gradients.

- First, we consider the general case where the stochastic gradients admit a bounded bias and a bounded variance (Section 6.1).
- As a corollary of this general result, we obtain convergence guarantees for LMC in the case where V is only *weakly* smooth, i.e., ∇V is Hölder continuous (Section 6.2). We employ the Gaussian smoothing technique to obtain this corollary. It implies new sampling guarantees in total variation distance under a Poincaré inequality and weak smoothness.
- We obtain convergence guarantees for LMC in the case where V is a finite sum and the stochastic gradients are defined from mini-batches. In this case, the stochastic gradients have zero bias but unbounded variance. We employ the variance reduction technique (Section 6.3).

Finally, we conclude with open directions in Section 7.

2. Interpretation of approximate first-order stationarity in sampling

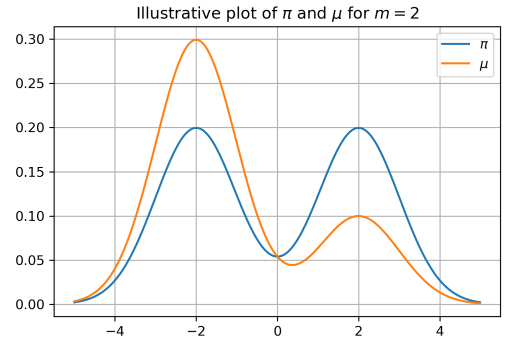
Intriguingly, unlike the situation in non-convex optimization, in sampling there are no "spurious stationary points": if μ and π have positive and smooth densities and $\mathsf{FI}(\mu \,\|\, \pi) = 0$, then $\mu = \pi$. However, for $\varepsilon > 0$, it may be unclear what the guarantee $\mathsf{FI}(\mu \,\|\, \pi) \le \varepsilon$ entails. In this section, we give an example illustrating what conclusions may be drawn from a bound on the Fisher information, which helps to better interpret our result in the next sections.

Consider a mixture of two Gaussians in one dimension as the target distribution:

$$\pi = \frac{1}{2} \underbrace{\mathsf{normal}(-m, 1)}_{\pi_-} + \frac{1}{2} \underbrace{\mathsf{normal}(+m, 1)}_{\pi_+},$$

where $m \gg 0$. Also, consider a mixture of two Gaussians with different weights:

$$\mu := \frac{3}{4}\, \pi_- + \frac{1}{4}\, \pi_+ \,.$$



An illustrative plot of π and μ is provided for the sake of easier visualization. In the appendix, we will prove the following.

Proposition 1 *Let π and μ be as defined above. For all $m \geq 1/80$, it holds that*

$$\|\mu - \pi\|_{\mathrm{TV}} \geq \frac{1}{800} > 0\,.$$

On the other hand,

$$\mathsf{FI}(\mu \,\|\, \pi) \leq 4m^2 \exp\bigl(-\frac{m^2}{2}\bigr) \to 0 \qquad \text{as } m \to \infty\,.$$

In the next section, we will show that averaged LMC can drive the Fisher information to zero at a polynomial rate. For large m, the measure μ has small Fisher information with respect to π, so μ serves as a model for the kind of distribution that averaged LMC can reach. We can draw a few conclusions:

1. Although the Fisher information $\mathsf{FI}(\mu \,\|\, \pi)$ is very small, the total variation distance remains bounded away from zero. This shows that a Fisher information guarantee does *not* ensure fast convergence of averaged LMC in other metrics without further assumptions (anyway, polynomial guarantees for non-log-concave sampling in other metrics are impossible in general).

2. Here, μ *locally* captures the correct shape of π at the two modes. On the other hand, μ has different mixing weights than π, which means that μ is *globally* different from π. Since $\mathsf{FI}(\mu \,\|\, \pi)$ is small for this example, it shows that the Fisher information is not sensitive to the latter effect. Hence, our Fisher information guarantee for averaged LMC captures the fact that the algorithm rapidly gets the local structure of π correct.

3. After a few steps of LMC started at the distribution $\frac{3}{4}\, \delta_{-m} + \frac{1}{4}\, \delta_{+m}$, the algorithm arrives at a measure which closely resembles μ, rather than the true stationary measure π. Indeed, the iterates of LMC do not need to jump from one mode to another to approximate μ. This jumping takes an exponentially long time and is the main barrier to the mixing of LMC, but it is necessary for LMC to learn the global mixing weights—this is known as the *metastability* phenomenon (Bovier et al., 2002, 2004, 2005). Our analysis provides a convenient way to quantify this effect.

Remark 2 *In the context of Bayesian inference, the choice of relative Fisher information metric between the prior and the exact posterior distribution has been proposed by Walker (2016); Holmes and Walker (2017); Shao et al. (2019), as a measure of robustness of the overall inferential procedure. In this regard, our results provide a computational angle to this paradigm: in practice we rarely have access to the exact posterior distribution. Our results algorithmically quantify the distance (in relative Fisher information) between the posterior distribution obtained after a certain number of iterations of LMC and the exact posterior.*

3. Preliminaries

Throughout the paper, we assume that the potential $V : \mathbb{R}^d \to \mathbb{R}$ is a smooth (i.e., twice continuously differentiable) function such that $\int \exp(-V) < \infty$. The target distribution $\pi \propto \exp(-V)$ is therefore well-defined.

For a probability measure μ with a smooth density, we can define the *Fisher information* of μ relative to π via $\mathsf{FI}(\mu \parallel \pi) := \mathbb{E}_\mu[\|\nabla \ln(\mu/\pi)\|^2]$. To extend this definition to other probability measures, we recall from Markov semigroup theory (see Bakry et al., 2014) that we associate with the Langevin diffusion (1) a *Dirichlet energy* $f \mapsto \mathscr{E}(f)$ which maps a subspace $\operatorname{dom}\mathscr{E} \subseteq L^2(\pi)$ to $\mathbb{R}_+$. If f is smooth and compactly supported, then $f \in \operatorname{dom}\mathscr{E}$ and the Dirichlet energy has the explicit expression $\mathscr{E}(f) = \mathbb{E}_\pi[\|\nabla f\|^2]$. The Fisher information is defined from the Dirichlet energy as follows. For an arbitrary probability measure μ, set

$$\mathsf{FI}(\mu \parallel \pi) := \begin{cases} 4\,\mathscr{E}(\sqrt{f})\,, & \text{if } f := \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \text{ exists and } \sqrt{f} \in \operatorname{dom}\mathscr{E}\,, \\ +\infty\,, & \text{otherwise}\,. \end{cases}$$

In particular, if $f = \frac{\mathrm{d}\mu}{\mathrm{d}\pi}$ is positive and smooth, one can check that

$$\mathsf{FI}(\mu \parallel \pi) = \int \|\nabla \ln(f)\|^2 \,\mathrm{d}\mu\,, \qquad \text{or} \qquad \mathsf{FI}(\mu \parallel \pi) = \int \frac{\|\nabla f\|^2}{f}\,\mathrm{d}\pi\,.$$

Using the convexity of $(a,b) \mapsto \|a\|^2/b$ on $\mathbb{R}^d \times \mathbb{R}_+$, the latter formula implies that the Fisher information $\mu \mapsto \mathsf{FI}(\mu \parallel \pi)$ is convex in the classical sense on the space of probability measures. Besides, the Fisher information is also lower semicontinuous in its first argument with respect to the weak topology of measures (see e.g. Wu, 2000, Appendix B).

4. Main result

Recall that the LMC algorithm is given by

$$x_{(k+1)h} := x_{kh} - h\,\nabla V(x_{kh}) + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,.$$

Our main result is stated for the following continuous interpolation of LMC:

$$x_t := x_{kh} - (t-kh)\,\nabla V(x_{kh}) + \sqrt{2}\,(B_t - B_{kh}) \qquad \text{for } t \in [kh,(k+1)h]\,. \tag{2}$$

We write μ_t for the law of x_t.

Assumption 3 *The gradient of V is L-Lipschitz continuous: $\|\nabla V(x_1) - \nabla V(x_2)\| \le L\,\|x_1 - x_2\|$, for all $x_1, x_2 \in \mathbb{R}^d$ and for some $L > 0$.*

Theorem 4 *Let $(\mu_t)_{t\geq 0}$ denote the law of the interpolation (2) of LMC, and let the potential V satisfy Assumption 3. Then, for any step size $h \in (0, \frac{1}{6L})$, it holds that*

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \leq \frac{2\,\mathsf{KL}(\mu_0 \,\|\, \pi)}{Nh} + 8L^2 dh\,.$$

In particular, if $\mathsf{KL}(\mu_0 \,\|\, \pi) \leq K_0$ and we choose $h = \sqrt{K_0}/(2L\sqrt{dN})$, then for $N \geq 9K_0/d$,

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \leq \frac{8L\sqrt{dK_0}}{\sqrt{N}}\,.$$

By the convexity of the Fisher information, it follows that the averaged distribution $\bar\mu_{Nh} := (Nh)^{-1}\int \mu_t\,\mathrm{d}t$ satisfies $\mathsf{FI}(\bar\mu_{Nh} \,\|\, \pi) \leq 8L\sqrt{dK_0/N}$ as well. Also, it is possible to output a sample from $\bar\mu_{Nh}$, as follows:

1. Pick a time $t \in [0, Nh]$ uniformly at random.
2. Let k be the largest integer such that $kh \leq t$, and let x_{kh} be the iterate of LMC at time kh. Then, perform a partial LMC update for time $t - kh$, i.e. set

$$x_t := x_{kh} - (t - kh)\,\nabla V(x_{kh}) + \sqrt{2}\,(B_t - B_{kh})\,.$$

Then, x_t is a sample from $\bar\mu_{Nh}$. Note that it is possible to sample the Brownian increments exactly as long as one can draw standard Gaussian vectors.

Remark 5 *Since we can usually take K_0 to be of order d, see, e.g., Vempala and Wibisono (2019, Lemma 1) or Chewi et al. (2021a, Appendix A), in order for averaged LMC to reach ε accuracy in terms of the Fisher information w.r.t. the target, the iteration complexity is $O(L^2 d^2/\varepsilon^2)$.*

5. Applications

5.1. Asymptotic convergence of averaged LMC with vanishing step size

Our main result immediately implies asymptotic convergence of averaged LMC with decreasing step size under very general conditions. Let $(h_k)_{k=1}^\infty$ be a sequence of positive step sizes such that

$$\sum_{k=1}^\infty h_k = \infty \quad \text{and} \quad \sum_{k=1}^\infty h_k^2 < \infty\,. \tag{3}$$

Write $\tau_n := \sum_{k=1}^n h_k$, and denote by $\bar\mu_{\tau_n} := \tau_n^{-1}\int_0^{\tau_n} \mu_t\,\mathrm{d}t$, where μ_t is the law of x_t defined by

$$x_t = x_{\tau_{n-1}} - (t - \tau_{n-1})\,\nabla V(x_{\tau_{n-1}}) + \sqrt{2}\,(B_t - B_{\tau_{n-1}})\,, \qquad t \in [\tau_{n-1}, \tau_n]\,.$$

Then, we have the following convergence result.

Theorem 6 *Let $(\mu_t)_{t\geq 0}$ denote the law of the interpolation (2) of LMC, and let the potential V satisfy Assumption 3. Suppose that LMC is initialized at a measure μ_0 with $\mathsf{KL}(\mu_0 \,\|\, \pi) < \infty$ and that the step size sequence $(h_k)_{k=1}^\infty$ satisfy $h_k \in (0, \frac{1}{6L})$ for every k, as well as the conditions in (3). Then, $\bar\mu_{\tau_n} \to \pi$ weakly.*

While it might be possible to prove the weak convergence of LMC using other techniques, for example, the ordinary differential equation method from the stochastic approximation literature (Kushner and Yin, 2003) or general results on the analysis of Markov chains (Bakry et al., 2014; Douc et al., 2018), we emphasize that Theorem 6 follows immediately from our main result in Theorem 4 and the property that $\mathsf{FI}(\mu \parallel \pi) = 0$ implies $\mu = \pi$. To the best of our knowledge, explicit results available in the literature on the weak convergence of LMC (e.g., Lamberton and Pages, 2002; Pagès and Panloup, 2012) require Lyapunov-type conditions. In comparison, Theorem 6 holds just under the Lipschitz gradient assumption on the potential V.

5.2. New sampling guarantees under a Poincaré inequality

In this section, we show that if we additionally assume that π satisfies a Poincaré inequality, then we obtain sampling guarantees in total variation distance as a corollary of our main theorem. Surprisingly, the rates we obtain in this manner are competitive with (and arguably better than) the state-of-the-art results for LMC, for these classes of target distributions. To present our result, we recall the following transportation inequality.

Lemma 7 (Guillin et al. (2009, Theorem 3.1)) *Suppose that π satisfies a Poincaré inequality: for all smooth compactly supported functions $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\operatorname{var}_\pi f \le C_{\mathsf{PI}}\, \mathbb{E}_\pi[\|\nabla f\|^2]\,. \tag{PI}$$

Then, for all probability measures μ,

$$\|\mu - \pi\|_{\mathrm{TV}}^2 \le 4 C_{\mathsf{PI}}\, \mathsf{FI}(\mu \parallel \pi)\,.$$

When combined with Theorem 4, we immediately obtain the following corollary.

Corollary 8 *Let $(\mu_t)_{t\ge 0}$ denote the law of the interpolation (2) of LMC, and let the potential V satisfy Assumption 3. If $\mathsf{KL}(\mu_0 \parallel \pi) \le K_0$ and we choose $h = \sqrt{K}_0/(2L\sqrt{dN})$, then for $N \ge 9K_0/d$ and $\bar\mu_{Nh} \coloneqq (Nh)^{-1}\int_0^{Nh} \mu_t \,\mathrm{d}t$,*

$$\|\bar\mu_{Nh} - \pi\|_{\mathrm{TV}}^2 \le \frac{32 C_{\mathsf{PI}} L \sqrt{dK_0}}{\sqrt{N}}\,.$$

Remark 9 *If $K_0 = O(d)$, it implies an iteration complexity of $O(C_{\mathsf{PI}}^2 L^2 d^2/\varepsilon^2)$ to output a sample whose squared total variation distance to π is at most ε. We are aware of only one other work which provides sampling guarantees for smooth potentials satisfying a Poincaré inequality: the recent result of Chewi et al. (2021a, Theorem 7) yields an iteration complexity of $\widetilde{O}(C_{\mathsf{PI}}^2 L^2 d^3/\varepsilon)$ for LMC (without averaging). Our result has worse dependence on the inverse accuracy, but better dependence on the dimension.*

Using *Gaussian smoothing* we can extend Corollary 8 to the case when ∇V is only Hölder continuous rather than Lipschitz continuous. This requires extending the Theorem 4 to accommodate stochastic gradients, and hence it is deferred to Section 6.2.

5.3. Hessian smoothness

While our main results were obtained under Lipschitz smoothness of the gradient of the potential, prior analyses of Langevin algorithms (Dalalyan and Karagulyan, 2019; Mou et al., 2022) suggest that convergence rates are accelerated under a smoothness assumption on the Hessian.

Assumption 10 *The Hessian of V is M-Lipschitz: $\|\nabla^2 V(x_1) - \nabla^2 V(x_2)\|_{\mathrm{op}} \le M\,\|x_1 - x_2\|$, for all $x_1, x_2 \in \mathbb{R}^d$ and for some $M > 0$.*

Additionally, we require an upper bound on the order of growth of the function.

Assumption 11 *There exist parameters $\gamma \in [0, 2]$, $0 \le \xi \le \gamma/2$, and constants $a, b, m > 0$ such that for all $x \in \mathbb{R}^d$,*

$$\langle x, \nabla V(x)\rangle \ge a\,\|x\|^\gamma - b \quad \text{and} \quad \|\nabla V(x)\| \le m\,(1 + \|x\|^\xi)\,. \tag{4}$$

Note that assuming $\gamma > 2$ would contradict Lipschitz smoothness of the gradient. The final condition allows for any polynomial tail for the potential, thus covers a significantly more general setting than the dissipativity assumption appearing in Raginsky et al. (2017); Erdogdu et al. (2018) ($\gamma = 2$) and the growth considered in Chewi et al. (2021a); Erdogdu and Hosseinzadeh (2021) ($\gamma \ge 1$). In the special case where $\xi = \gamma = 0$, it is equivalent to the gradient $\|\nabla V\|$ being uniformly bounded, i.e., V itself is Lipschitz. The growth condition is used to establish new moment bounds for the iterates of LMC (Proposition 25), which are key for discretization analysis.

In the following theorem, we assume for simplicity that $a = 1$ (which can be achieved by rescaling the potential).

Theorem 12 *Let $(\mu_t)_{t\ge 0}$ denote the law of the interpolation (2) of LMC, and let the potential V satisfy Assumptions 3, 10, and 11. Assume $a = 1$ and that the initialization is chosen with $\mathbb{E}[\|x_0\|^4] \le \sigma^2 d^2$ for some $\sigma \ge 3$. Define the parameter $\kappa := 1 \vee L \vee M^{2/3} \vee (M^{1/3} m^{2/3})$. If the step size is chosen to satisfy $0 < h \lesssim \frac{1}{L} \wedge \frac{1}{m^2} \wedge 1$, then*

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \lesssim \frac{\mathsf{KL}(\mu_0 \,\|\, \pi)}{Nh} + \kappa^3 d^2 h^2 + \kappa^6\,(b + \sigma d)^3\,N h^5\,.$$

If $\mathsf{KL}(\mu_0 \,\|\, \pi) \le K_0$, and $h \asymp \frac{K_0^{1/3}}{\kappa\,(b+\sigma d)^{2/3}\,N^{1/3}}$ while $N \gtrsim \frac{K_0\,(L^3 \vee m^6)}{\kappa^3 (b+\sigma d)^2}$, the following bound holds:

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \lesssim \Big((b + \sigma d)^{2/3} K_0^{2/3} + \frac{K_0^{5/3}}{(b + \sigma d)^{1/3}}\Big)\,\frac{\kappa}{N^{2/3}}\,.$$

Note that the bound is independent of the growth exponents γ, ξ found in Assumption 11.

Remark 13 *When $b, K_0 = O(d)$, the iteration complexity implied by this result is $O(d^2/\varepsilon^{3/2})$, which should be compared to the complexity of $O(d^2/\varepsilon^2)$ in Theorem 4. As in Corollary 8, we can combine this result with Lemma 7 to obtain the complexity $O(d^2/\varepsilon^{3/2})$ in squared total variation distance under the additional assumption of a Poincaré inequality on the target.*

6. Extension to stochastic gradients

6.1. General result

We proceed to prove a more general result in which the gradient term in (LMC) is replaced by a stochastic gradient. More precisely, we use a stochastic estimate $G(x_{kh}, \zeta_k)$ of the gradient $\nabla V(x_{kh})$, where the random variables $(\zeta_k)_{k\in\mathbb{N}}$ representing the external randomness are i.i.d. and independent of all other random variables. Thus, we obtain *stochastic gradient Langevin Monte Carlo* (SG-LMC):

$$x_{(k+1)h} \coloneqq x_{kh} - h\, G(x_{kh}, \zeta_k) + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,. \tag{SG-LMC}$$

Assumption 14 (Regularity of the stochastic gradient) *Let $\mathbb{E}_\zeta\, G(y,\zeta) = \nabla\hat{V}(y)$ for some function $\hat{V} : \mathbb{R}^d \to \mathbb{R}$. The stochastic gradient $G(x,\zeta) \in \mathbb{R}^d$ satisfies:*

- *Smoothness of the expected stochastic gradient: $\nabla\hat{V}$ is $\hat{L}$-Lipschitz.*
- *Bias bound: $\|\nabla\hat{V}(x) - \nabla V(x)\|^2 \le \mathsf{bias}^2$ for all $x \in \mathbb{R}^d$.*
- *Variance bound: $\mathbb{E}_\zeta[\|G(x,\zeta) - \nabla\hat{V}(x)\|^2] \le \mathsf{var}$ for all $x \in \mathbb{R}^d$.*

We present the following theorem regarding the convergence:

Theorem 15 *Let $(\mu_t)_{t\ge 0}$ denote the law of the interpolation of (SG-LMC). Assume that the stochastic oracle satisfies Assumption 14. Then, for all $h \in (0, \frac{1}{14\hat{L}})$, we have*

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \le \frac{2\,\mathsf{KL}(\mu_0 \,\|\, \pi)}{Nh} + 16\hat{L}^2 dh + 8\,(\mathsf{bias}^2 + \mathsf{var})\,.$$

In particular, if $\mathsf{KL}(\mu_0 \,\|\, \pi) \le K_0$ and we choose $h = \sqrt{K_0}/(\hat{L}\sqrt{8dN})$, then for $N \ge 25K_0/d$,

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \le \frac{16\hat{L}\sqrt{2dK_0}}{\sqrt{N}} + 8\,(\mathsf{bias}^2 + \mathsf{var})\,.$$

This generic result allows us to use *biased* stochastic gradients. In particular, it can be applied to LMC with Gaussian smoothing.

6.2. Extension to non-smooth potentials via Gaussian smoothing

In this section, we extend our main theorem (Theorem 4) to the case when ∇V is assumed to be Hölder continuous.

Assumption 16 *The gradient of V is Hölder continuous of exponent $s \in (0,1]$:*

$$\|\nabla V(x_1) - \nabla V(x_2)\| \le L\,\|x_1 - x_2\|^s$$

for all $x_1, x_2 \in \mathbb{R}^d$ and for some $L > 0$.

We consider the Gaussian smoothing LMC algorithm analyzed in Chatterji et al. (2020):

$$x_{(k+1)h} = x_{kh} - h\,\nabla V(x_{kh} + \eta\zeta_k) + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,, \tag{5}$$

where $\eta > 0$ is a smoothing parameter and $(\zeta_k)_{k\in\mathbb{N}}$ are i.i.d. standard Gaussian random variables on $\mathbb{R}^d$ independent from x_0 and $(B_t)_{t\geq 0}$. We see that this iteration is a special case of (SG-LMC) with stochastic gradient given by $G(x_{kh}, \zeta_k) = \nabla V(x_{kh} + \eta\zeta_k)$. The expected stochastic gradient is $\mathbb{E}\, G(x,\zeta) = \nabla \hat V(x)$, where $\hat V(x) = \mathbb{E}\, V(x+\eta\zeta)$ and $\zeta \sim \mathsf{normal}(0, I_d)$.

From Chatterji et al. (2020, Lemma 2.2 and Lemma 3.1), $\nabla \hat V$ satisfies the first and third conditions of Assumption 14 with

$$\hat L \leq \frac{L d^{(1-s)/2}}{\eta^{1-s}}\,, \qquad \mathsf{var} \leq 4L^2 d^s \eta^{2s}\,.$$

To control the bias, we extend the result of Nesterov and Spokoiny (2017).

Lemma 17 *The Gaussian smoothed potential $\hat V$ with smoothing parameter μ satisfies the second condition of Assumption 14 with*

$$\mathsf{bias}^2 \lesssim L^2 d^{2+s} \eta^{2s}\,.$$

From the lemma, we see that the bias dominates: $\mathsf{bias}^2 \gtrsim \mathsf{var}$. We obtain the following corollary.

Corollary 18 *Let $(\mu_t)_{t\geq 0}$ denote the law of the interpolation (2) of Gaussian smoothed LMC (5), and let the potential V satisfy Assumption 16. If $\mathsf{KL}(\mu_0 \,\|\, \pi) \leq K_0$, we choose the smoothing to be $\eta \asymp \varepsilon^{1/(2s)}/(L^{1/s} d^{(2+s)/(2s)})$ (where the $\asymp$ hides an absolute constant), and we choose the step size h as in Theorem 15, then the averaged law $\bar\mu_{Nh} \leq (Nh)^{-1}\int_0^{Nh} \mu_t \,\mathrm{d}t$ satisfies $\mathsf{FI}(\bar\mu_{Nh} \,\|\, \pi) \leq \varepsilon$, provided that the number of iterations is*

$$N \gtrsim \frac{K_0 L^{2/s} d^{(2+s-2s^2)/s}}{\varepsilon^{(1+s)/s}}\,.$$

Proof Apply Theorem 15. ■

When $K_0 = O(d)$, the iteration complexity is $O(L^{2/s} d^{2\,(1+s-s^2)/s}/\varepsilon^{(1+s)/s})$. As in Section 5.2, this result can be combined with a Poincaré inequality to yield a convergence result in total variation distance. However, there is a better approach in this case. It turns out that to reduce the variance of the stochastic gradients in Gaussian smoothing, it is advantageous to consider mini-batching: for $B \in \mathbb{N}^+$, we consider

$$x_{(k+1)h} = x_{kh} - \frac{h}{B}\sum_{\ell=1}^{B} \nabla V(x_{kh} + \eta\zeta_{k,\ell}) + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,, \tag{6}$$

where $(\zeta_{k,\ell})_{k,\ell\in\mathbb{N}}$ is a family of i.i.d. standard Gaussians on $\mathbb{R}^d$ independent of x_0 and $(B_t)_{t\geq 0}$.

Corollary 19 *Let $(\mu_t)_{t\geq 0}$ denote the law of the interpolation of the Gaussian smoothed LMC with mini-batching (6), and let the potential V satisfy Assumption 16. Assume moreover that π satisfies the Poincaré inequality (PI) with constant C_{PI}. If $\mathsf{KL}(\mu_0 \,\|\, \pi) \leq K_0$, we choose the smoothing*

η appropriately (see (15)), and we choose the step size h as in Theorem 15, then the averaged law $\bar{\mu}_{Nh} \le (Nh)^{-1} \int_0^{Nh} \mu_t \, \mathrm{d}t$ satisfies $\|\bar{\mu}_{Nh} - \pi\|_{\mathrm{TV}}^2 \le \varepsilon$ (for $0 < \varepsilon \le 1$), with ***total gradient complexity*** *at most*

$$B \times N \lesssim \begin{cases} \dfrac{C_{\mathsf{PI}}^{(1+s)/s} K_0 L^{2/s} d^{3-2s}}{\varepsilon^{(1+s)/s}}\,, & \text{if } s \ge \dfrac{1}{2} \text{ with } B = 1\,, \\ \dfrac{C_{\mathsf{PI}}^{3} K_0 L^{6/(1+s)} d^{3-2s}}{\varepsilon^{(5-s)/(1+s)}}\,, & \text{if } s \le \dfrac{1}{2} \text{ with } B \asymp \dfrac{C_{\mathsf{PI}} L^{2/(1+s)}}{\varepsilon^{(1-s)/(1+s)}}\,. \end{cases}$$

Compared with the result of Chewi et al. (2021a, Theorem 7) which has iteration complexity $\widetilde{O}(C_{\mathsf{PI}}^{(1+s)/s} L^{2/s} d^{(1+2s)/s} / \varepsilon^{1/s})$, we see that our dependence on every problem parameter is better except for the dependence on the inverse accuracy, for which we obtain a better rate only for $s \le 2 - \sqrt{3} \approx 0.27$. In particular, our complexity does not blow up as $s \searrow 0$, so we can set $s = 0$ and get an iteration complexity of $O(C_{\mathsf{PI}}^2 L^6 d^4 / \varepsilon^5)$ for sampling from *Lipschitz potentials satisfying a Poincaré inequality*. To the best of our knowledge, this is the first guarantee for this setting.

6.3. Finite sum setting

Finally, we consider the case $V = \frac{1}{n} \sum_{i=1}^n f_i$ is a finite sum involving a large number n of terms, as it is often the case in machine learning. In the big data regime, mini-batch stochastic gradient-based LMC is preferred to vanilla LMC due to reduced per-iteration costs (Brosse et al., 2018; Chatterji et al., 2018) and is well-studied (Xu et al., 2018; Zou et al., 2021). However, stochastic gradients obtained by randomly selecting a mini-batch of data do not have a bounded variance in general. Therefore, the generic Theorem 15 is not applicable to mini-batching in general.

We consider LMC with a variance-reduced stochastic gradient given by the PAGE estimator (Li et al., 2021). Indeed, in non-convex optimization, the PAGE estimator has been used to reduce the variance in SGD and led to a simple and optimal stochastic non-convex optimization algorithm. We consider a Variance Reduced LMC algorithm:

$$x_{(k+1)h} \coloneqq x_{kh} - h\, g_{kh} + \sqrt{2}\,(B_{(k+1)h} - B_{kh})\,, \tag{VR-LMC}$$

where g_{kh} is defined by

$$g_{(k+1)h} \coloneqq \begin{cases} \nabla V(x_{(k+1)h})\,, & \text{with probability } p\,, \\ g_{kh} + \nabla f_i(x_{(k+1)h}) - \nabla f_i(x_{kh})\,, & \text{with probability } 1-p\,, \end{cases} \tag{7}$$

where $i \sim \mathsf{uniform}([1, \dots, n])$ and $p \in (0, 1]$. Let us describe how $g_{(k+1)h}$ is obtained from g_{kh}. Denote $\mathcal{F}_k = \sigma(g_0, \dots, g_{kh}, x_0, \dots, x_{(k+1)h})$. To obtain $g_{(k+1)h}$, one first samples $B \sim \mathsf{Bernoulli}(p)$, independent of $\mathcal{F}_k$. If $B = 1$, then $g_{(k+1)h} = \nabla V(x_{(k+1)h})$ and if $B = 0$ then one samples a uniform random variable i, independent of $\mathcal{F}_k$ and independent of B, and one sets $g_{(k+1)h} = g_{kh} + \nabla f_i(x_{(k+1)h}) - \nabla f_i(x_{kh})$.

Assuming that g_0 is an unbiased estimate of $\nabla V(x_0)$, then $\mathbb{E}(g_{(k+1)h}) = \mathbb{E}(\nabla V(x_{(k+1)h}))$ by induction. Therefore the PAGE estimator has zero bias. However, its variance is not uniformly bounded in general and Theorem 15 is not applicable to (VR-LMC). Nevertheless, we obtain a result under the following assumption.

Assumption 20 *The potential V is a finite sum $V = n^{-1} \sum_{i=1}^n f_i$ and the gradient of f_i is L-Lipschitz continuous for every $i \in [n]$.*

If a further assumption is made on the growth of V at infinity, bounds in terms of the Wasserstein distance can be obtained for variants of Langevin algorithm with variance reduction to sample from a non-log-concave target distribution (Zou et al., 2019). The next theorem provides Fisher information guarantee under Assumption 20 only.

Theorem 21 *Let $(\mu_t)_{t\geq 0}$ denote the law of the interpolation of (VR-LMC) and let the potential V satisfy Assumption 20. Then, for all $h \in (0, \frac{\sqrt{p}}{5L})$, we have*

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \leq \frac{2C}{Nh} + \frac{18L^2 dh}{p}\,,$$

where

$$C := \mathsf{KL}(\mu_0 \,\|\, \pi) + \frac{3h}{p}\,\mathbb{E}[\|g_0 - \nabla V(x_0)\|^2]\,.$$

In particular, if $h = \frac{\sqrt{pC}}{3L\sqrt{Nd}}$ and $N \geq \frac{2C}{d}$, then

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \leq 12L\sqrt{\frac{Cd}{Np}}\,.$$

Remark 22 *We now elaborate on the* ***total number of individual gradient evaluations*** *based on Theorem 21. Since $\mathsf{KL}(\mu_0 \,\|\, \pi) = O(d)$, if we assume that a full gradient is computed at the first step, then $N \asymp L^2 d^2/(p\varepsilon^2)$ iterations suffice to achieve ε accuracy: $\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \leq \varepsilon$. At each iteration, the algorithm computes $pn + 1 - p = O(pn)$ new gradients in average. Therefore, ε accuracy is achieved after*

$$O(pnN) = O\Big(\frac{L^2 d^2 n}{\varepsilon^2}\Big) \tag{8}$$

gradient computations. A direct application of Theorem 4 would also give a similar order of gradient computations. However, the per-iteration complexity of (VR-LMC) is better than that of (1), making it easier to apply to problems with large n in practice. For instance, taking $p = 1/(n-1)$, the amortized number of gradient computations per iteration is constant (independent of n) equal to 2, which allows for using mini-batch versions of LMC without making the variance boundedness assumption required by Theorem 15.

7. Conclusion and open questions

In this work, we have initiated the study of non-log-concave sampling by proving that, under the sole assumption that the potential has a Lipschitz gradient, averaged LMC drives the Fisher information w.r.t. the target to zero after polynomially many iterations. We have argued that this is the natural sampling analogue of finding approximate first-order stationary points in non-convex optimization.

Although our focus was to work under the minimal assumption of smoothness, surprisingly our analysis yielded new results for sampling from targets satisfying a Poincaré inequality, and moreover our results attain state-of-the-art dimension dependence for these settings for LMC.

We believe there are many intriguing directions for future work, and we list a few to conclude.

1. (lower bounds) We ask whether one can prove lower bounds on the complexity of outputting a sample whose Fisher information w.r.t. the target is ε. Since the setting of this work is fully non-convex, it may be easier to produce lower bound constructions than the strongly log-concave case, in which the theory of lower bounds is nascent (see Chewi et al., 2021b).

2. (improved results and further extensions) Although we have provided results under Hessian smoothness and via variance reduction, our investigation is still preliminary and we believe that these results can be strengthened. Additionally, there are other important extensions to consider; for instance, is there an analogue of second-order stationarity in sampling?

3. (Poincaré case) The iteration complexity we obtained for smooth potentials which satisfy a Poincaré inequality (focusing only on dimension and accuracy) is $O(d^2/\varepsilon^2)$, whereas Chewi et al. (2021a) obtained $\widetilde{O}(d^3/\varepsilon)$. Is it possible to achieve $\widetilde{O}(d^2/\varepsilon)$ with a variant of LMC? If so, is averaging necessary?

Acknowledgments

We thank Mufan (Bill) Li and Ruoqi Shen for helpful conversations. KB was supported by a seed grant from Center for Data Science and Artificial Intelligence Research, UC Davis and NSF Grant DMS-2053918. SC was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. MAE was supported by NSERC Grant [2019-06167], Connaught New Researcher Award, CIFAR AI Chairs program, and CIFAR AI Catalyst grant. AS was supported by a Simons–Berkeley Research Fellowship. This work was done while several of the authors were visiting the Simons Institute for the Theory of Computing.

References

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

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.

Espen Bernton. Langevin Monte Carlo and JKO splitting. In *Conference on Learning Theory (COLT)*, pages 1777–1798, 2018.

Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. Metastability and low lying spectra in reversible Markov chains. *Communications in Mathematical Physics*, 228(2): 219–255, 2002.

Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes I: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424, 2004.

Anton Bovier, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes II: Precise asymptotics for small eigenvalues. *Journal of the European Mathematical Society*, 7(1):69–99, 2005.

Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of stochastic gradient Langevin dynamics. *Advances in Neural Information Processing Systems*, 31, 2018.

Niladri Chatterji, Nicolas Flammarion, Yian Ma, Peter Bartlett, and Michael Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. In *International Conference on Machine Learning*, pages 764–773. PMLR, 2018.

Niladri Chatterji, Jelena Diakonikolas, Michael I Jordan, and Peter Bartlett. Langevin Monte Carlo without smoothness. In *International Conference on Artificial Intelligence and Statistics*, pages 1716–1726. PMLR, 2020.

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

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

Sinho Chewi, Patrik Gerber, Chen Lu, Thibaut Le Gouic, and Philippe Rigollet. The query complexity of sampling from strongly log-concave distributions in one dimension. *arXiv e-prints*, art. arXiv:2105.14163, 2021b.

Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians. *arXiv e-prints*, art. arXiv:1810.08693, 2020.

Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*. Springer, 2018.

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

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

Murat A Erdogdu and Rasa Hosseinzadeh. On the convergence of Langevin Monte Carlo: the interplay between tail growth and smoothness. In *Conference on Learning Theory*, pages 1776–1822. PMLR, 2021.

Murat A Erdogdu, Lester Mackey, and Ohad Shamir. Global non-convex optimization with discretized diffusions. *Advances in Neural Information Processing Systems*, 31, 2018.

Arnaud Guillin, Christian Léonard, Liming Wu, and Nian Yao. Transportation-information inequalities for Markov processes. *Probab. Theory Related Fields*, 144(3-4):669–695, 2009.

Ye He, Krishnakumar Balasubramanian, and Murat A. Erdogdu. Heavy-tailed sampling via transformed unadjusted Langevin algorithm. *arXiv preprint arXiv:2201.08349*, 2022.

Chris C. Holmes and Stephen G. Walker. Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503, 2017.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Lojasiewicz condition. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, page 795–811, 2016.

Harold. J. Kushner and George G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer-Verlag, New York, second edition, 2003.

Damien Lamberton and Gilles Pages. Recursive computation of the invariant distribution of a diffusion. *Bernoulli*, pages 367–405, 2002.

Xuechen Li, Yi Wu, Lester Mackey, and Murat A Erdogdu. Stochastic Runge–Kutta accelerates Langevin Monte Carlo and beyond. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pages 6286–6295. PMLR, 2021.

Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

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.

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

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 Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Gilles Pagès and Fabien Panloup. Ergodic approximation of the distribution of a stationary diffusion: rate of convergence. *The Annals of Applied Probability*, 22(3):1059–1100, 2012.

Boris Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

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

Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.

Stephane Shao, Pierre E Jacob, Jie Ding, and Vahid Tarokh. Bayesian model comparison with the Hyvärinen score: Computation and consistency. *Journal of the American Statistical Association*, 2019.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems 32*, pages 8094–8106. 2019.

Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.

Stephen G Walker. Bayesian information in an experiment and the Fisher information distance. *Statistics & Probability Letters*, 112:5–9, 2016.

Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR, 2018.

Liming Wu. Uniformly integrable operators and large deviations for Markov processes. *J. Funct. Anal.*, 172(2):301–376, 2000.

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Difan Zou, Pan Xu, and Quanquan Gu. Sampling from non-log-concave distributions via variance-reduced gradient Langevin dynamics. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2936–2945. PMLR, 16–18 Apr 2019.

Difan Zou, Pan Xu, and Quanquan Gu. Faster convergence of stochastic gradient Langevin dynamics for non-log-concave sampling. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1152–1162. PMLR, 27–30 Jul 2021.

Appendix A. Proof for the illustrative example

Proof [Proof of Proposition 1] The total variation distance is

$$\|\mu - \pi\|_{\mathrm{TV}} = \frac{1}{2}\int |\mu - \pi| = \frac{1}{8}\int |\pi_+ - \pi_-| = \frac{1}{4}\,\|\pi_+ - \pi_-\|_{\mathrm{TV}}\,.$$

Since $\pi_- = \mathsf{normal}(-m, 1)$ and $\pi_+ = \mathsf{normal}(m, 1)$, the lower bound on $\|\mu - \pi\|_{\mathrm{TV}}$ follows from Devroye et al. (2020, Theorem 1.3).

Next, we have

$$\begin{aligned}\nabla \ln \frac{\mu}{\pi} &= \frac{1}{\mu}\,\bigl(\frac{3}{4}\,\nabla\pi_- + \frac{1}{4}\,\nabla\pi_+\bigr) - \frac{1}{\pi}\,\bigl(\frac{1}{2}\,\nabla\pi_- + \frac{1}{2}\,\nabla\pi_+\bigr)\\ &= \frac{1}{\mu\pi}\,\bigl[\pi\,\bigl(\frac{3}{4}\,\nabla\pi_- + \frac{1}{4}\,\nabla\pi_+\bigr) + \mu\,\bigl(\frac{1}{2}\,\nabla\pi_- + \frac{1}{2}\,\nabla\pi_+\bigr)\bigr]\,.\end{aligned}$$

Writing $s_\mp \coloneqq \nabla \ln \pi_\mp$, some algebra reveals that

$$\nabla \ln \frac{\mu}{\pi} = \frac{1}{4\mu\pi}\,(\pi_+\,\nabla\pi_- - \pi_-\,\nabla\pi_+) = \frac{\pi_-\pi_+}{4\mu\pi}\,(s_- - s_+) = -\frac{\pi_-\pi_+}{2\mu\pi}\,m\,.$$

Therefore,

$$\begin{aligned}\mathsf{FI}(\mu \,\|\, \pi) &= \frac{m^2}{4}\int \frac{\pi_-^2\pi_+^2}{\mu^2\pi^2}\,\mathrm{d}\mu = \frac{m^2}{4}\int \frac{\pi_-^2\pi_+^2}{\mu\pi^2} = \frac{m^2}{4}\int \frac{\pi_-^2\pi_+^2}{\bigl(\frac{3}{4}\,\pi_- + \frac{1}{4}\,\pi_+\bigr)\,\bigl(\frac{1}{2}\,\pi_- + \frac{1}{2}\,\pi_+\bigr)^2}\\ &\le 4m^2\int \frac{\pi_-^2\pi_+^2}{(\pi_- + \pi_+)^3} \le 4m^2\,\Bigl[\int_{\mathbb{R}_-}\frac{\pi_+^2}{\pi_-} + \int_{\mathbb{R}_+}\frac{\pi_-^2}{\pi_+}\Bigr]\,.\end{aligned}$$

Writing $Z \coloneqq (2\pi)^{d/2}$ for the normalizing constant,

$$\begin{aligned}\int_{\mathbb{R}_+}\frac{\pi_-^2}{\pi_+} &= \frac{1}{Z}\int_0^\infty \exp\bigl(-|x+m|^2 + \frac{1}{2}\,|x-m|^2\bigr)\,\mathrm{d}x = \frac{\exp(4m^2)}{Z}\int_0^\infty \exp\bigl(-\frac{1}{2}\,|x+3m|^2\bigr)\,\mathrm{d}x\\ &= \exp(4m^2)\,\mathbb{P}\{\xi \ge 3m\}\end{aligned}$$

where ξ is a standard Gaussian random variable. Using a standard Gaussian tail bound,

$$\mathbb{P}\{\xi \ge 3m\} \le \frac{1}{2}\exp\bigl(-\frac{9m^2}{2}\bigr)\,.$$

A symmetric argument holds for the other integral, and hence

$$\mathsf{FI}(\mu \,\|\, \pi) \le 4m^2\exp\bigl(-\frac{m^2}{2}\bigr)$$

which completes the proof. ∎

Appendix B. Proof of the main theorem

Our proof follows the interpolation argument of Vempala and Wibisono (2019) which proceeds by obtaining a differential inequality for the KL divergence along an interpolation of the algorithm. With an eye towards extensions of the main result, we prove a more general version of the inequality.

Lemma 23 *Consider the stochastic process defined by*

$$x_t \coloneqq x_0 - hg_0 + \sqrt{2}\, B_t\,, \qquad \text{for } t \ge 0\,,$$

where $(B_t)_{t\ge 0}$ is a standard Brownian motion in $\mathbb{R}^d$ which is independent of (x_0, g_0). Then, writing μ_t for the law of x_t,

$$\begin{aligned}\partial_t \operatorname{\mathsf{KL}}(\mu_t \,\|\, \pi) &\le -\frac{3}{4}\operatorname{\mathsf{FI}}(\mu_t \,\|\, \pi) + \mathbb{E}\big[\|\nabla V(x_t) - \mathbb{E}[g_0 \mid x_t]\|^2\big] \\ &\le -\frac{3}{4}\operatorname{\mathsf{FI}}(\mu_t \,\|\, \pi) + \mathbb{E}[\|\nabla V(x_t) - g_0\|^2]\,.\end{aligned}$$

Proof Let $\mathcal{F}_0$ denote the σ-algebra generated by (x_0, g_0), and let $\mu_{t|\mathcal{F}_0}$ denote the conditional law of x_t given $\mathcal{F}_0$. Then, $t \mapsto \mu_{t|\mathcal{F}_0}$ evolves according to the Fokker–Planck equation

$$\partial_t \mu_{t|\mathcal{F}_0}(x) = \Delta \mu_{t|\mathcal{F}_0}(x) + \operatorname{div}_x\big(\mu_{t|\mathcal{F}_0}(x)\, g_0\big)\,.$$

If $\mathbb{P}_0$ denotes the restriction of the probability measure $\mathbb{P}$ on the underlying probability space, then taking the expectation w.r.t. $\mathbb{P}_0$ yields

$$\partial_t \mu_t(x) = \Delta \mu_t(x) + \operatorname{div}_x \mathbb{E}[\mu_{t|\mathcal{F}_0}(x)\, g_0]\,.$$

The second term is

$$\begin{aligned}\mathbb{E}[\mu_{t|\mathcal{F}_0}(x) g_0] &= \int \mu_{t|\mathcal{F}_0}(x \mid \omega)\, g_0(\omega)\, \mathbb{P}_0(\mathrm{d}\omega) = \mu_t(x) \int g_0(\omega)\, \mu_{\mathcal{F}_0|t}(\mathrm{d}\omega \mid x) \\ &= \mu_t(x)\, \mathbb{E}[g_0 \mid x_t = x]\,.\end{aligned}$$

From this, the time derivative of the KL divergence is

$$\begin{aligned}\partial_t \operatorname{\mathsf{KL}}(\mu_t \,\|\, \pi) &= \int \big(\ln \frac{\mu_t}{\pi}\big) \operatorname{div}\big(\mu_t\, (\nabla \ln \mu_t + \mathbb{E}[g_0 \mid x_t = \cdot])\big) \\ &= -\int \big\langle \nabla \ln \frac{\mu_t}{\pi}, \nabla \ln \mu_t + \mathbb{E}[g_0 \mid x_t = \cdot]\big\rangle\, \mathrm{d}\mu_t \\ &= -\operatorname{\mathsf{FI}}(\mu_t \,\|\, \pi) + \int \big\langle \nabla \ln \frac{\mu_t}{\pi}, \nabla V - \mathbb{E}[g_0 \mid x_t = \cdot]\big\rangle\, \mathrm{d}\mu_t\,.\end{aligned}$$

Applying Young's inequality,

$$\int \big\langle \nabla \ln \frac{\mu_t}{\pi}, \nabla V - \mathbb{E}[g_0 \mid x_t = \cdot]\big\rangle\, \mathrm{d}\mu_t \le \frac{1}{4}\operatorname{\mathsf{FI}}(\mu_t \,\|\, \pi) + \mathbb{E}\big[\|\nabla V(x_t) - \mathbb{E}[g_0 \mid x_t]\|^2\big]$$

which completes the proof. ■

We also use the following lemma, which is taken from Chewi et al. (2021a). For the reader's convenience, the proof is reproduced here.

Lemma 24 (Chewi et al. (2021a, Lemma 16)) *Assume that ∇V is L-Lipschitz. For any probability measure μ, it holds that*

$$\mathbb{E}_\mu[\|\nabla V\|^2] \le \mathsf{FI}(\mu \,\|\, \pi) + 2dL\,.$$

Proof Let $\mathscr{L}$ denote the infinitesimal generator of the Langevin diffusion (1), i.e.

$$\mathscr{L} f \coloneqq \langle \nabla V, \nabla f \rangle - \Delta f\,.$$

Observe that $\mathscr{L} V = \|\nabla V\|^2 - \Delta V$. Applying integration by parts,

$$\begin{aligned}
\mathbb{E}_\mu[\|\nabla V\|^2] &= \mathbb{E}_\mu \mathscr{L} V + \mathbb{E}_\mu \Delta V \le \int \mathscr{L} V \,\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\,\mathrm{d}\pi + dL = \int \langle \nabla V, \nabla \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\rangle \,\mathrm{d}\pi + dL \\
&= 2\int \langle \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}}\,\nabla V, \nabla \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}}\rangle \,\mathrm{d}\pi + dL \le \frac{1}{2}\,\mathbb{E}_\mu[\|\nabla V\|^2] + 2\,\mathbb{E}_\pi\big[\|\nabla \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\pi}}\|^2\big] + dL\,.
\end{aligned}$$

Rearrange this inequality to obtain the desired result. ■

We now prove our main result.

Proof [Proof of Theorem 4] Let $(x_t)_{t\ge 0}$ denote the interpolation of LMC (defined in (2)). For $t \in [kh, (k+1)h]$, Lemma 23 yields

$$\partial_t \,\mathsf{KL}(\mu_t \,\|\, \pi) \le -\frac{3}{4}\,\mathsf{FI}(\mu_t \,\|\, \pi) + \mathbb{E}[\|\nabla V(x_t) - \nabla V(x_{kh})\|^2]$$

and the error term is

$$\begin{aligned}
\mathbb{E}[\|\nabla V(x_t) - \nabla V(x_{kh})\|^2] &\le L^2\,\mathbb{E}[\|x_t - x_{kh}\|^2] \\
&\le 2L^2\,(t-kh)^2\,\mathbb{E}[\|\nabla V(x_{kh})\|^2] + 4L^2\,\mathbb{E}[\|B_t - B_{kh}\|^2]\,.
\end{aligned}$$

Next, since ∇V is Lipschitz,

$$\begin{aligned}
\|\nabla V(x_{kh})\| &\le \|\nabla V(x_t)\| + L\,\|x_t - x_{kh}\| \\
&\le \|\nabla V(x_t)\| + Lh\,\|\nabla V(x_{kh})\| + \sqrt{2}L\,\|B_t - B_{kh}\|\,.
\end{aligned}$$

and for $h \le 1/(3L)$ we can rearrange this to yield

$$\|\nabla V(x_{kh})\| \le \frac{3}{2}\,\|\nabla V(x_t)\| + \frac{3L}{\sqrt{2}}\,\|B_t - B_{kh}\|\,.$$

Plugging this in,

$$\|\nabla V(x_t) - \nabla V(x_{kh})\|^2 \le 9L^2\,(t-kh)^2\,\|\nabla V(x_t)\|^2 + 6L^2\,\|B_t - B_{kh}\|^2\,. \tag{9}$$

For the expectation of the first term, we can use Lemma 24 to bound

$$\mathbb{E}_{\mu_t}[\|\nabla V\|^2] \le \mathsf{FI}(\mu_t \,\|\, \pi) + 2Ld\,.$$

Hence, for $h \le 1/(6L)$,

$$\partial_t \mathsf{KL}(\mu_t \,\|\, \pi) \le -\big(\frac{3}{4} - 9L^2h^2\big)\,\mathsf{FI}(\mu_t \,\|\, \pi) + 18L^3 d\,(t-kh)^2 + 6L^2 d\,(t-kh)$$
$$\le -\frac{1}{2}\,\mathsf{FI}(\mu_t \,\|\, \pi) + 18L^3 d\,(t-kh)^2 + 6L^2 d\,(t-kh)\,. \tag{10}$$

Integrating, we obtain

$$\mathsf{KL}(\mu_{(k+1)h} \,\|\, \pi) - \mathsf{KL}(\mu_{kh} \,\|\, \pi) \le -\frac{1}{2}\int_{kh}^{(k+1)h} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t + 6L^3dh^3 + 3L^2dh^2$$
$$\le -\frac{1}{2}\int_{kh}^{(k+1)h} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t + 4L^2dh^2\,. \tag{11}$$

Now by summing, we have

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \le \frac{2\,\mathsf{KL}(\mu_0 \,\|\, \pi)}{Nh} + 8L^2dh\,.$$

This concludes the proof. ■

Appendix C. Proofs for the extensions and applications

C.1. Asymptotic convergence of averaged LMC

Proof [Proof of Theorem 6] The one-step recursion (11) in the proof of Theorem 4 yields

$$\mathsf{KL}(\mu_{\tau_n} \,\|\, \pi) - \mathsf{KL}(\mu_{\tau_{n-1}} \,\|\, \pi) \le -\frac{1}{2}\int_{\tau_{n-1}}^{\tau_n} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t + 4L^2dh_n^2\,.$$

Iterating the above bound, we obtain

$$\mathsf{KL}(\mu_{\tau_n} \,\|\, \pi) \le \mathsf{KL}(\mu_0 \,\|\, \pi) - \frac{1}{2}\int_0^{\tau_n} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t + 4L^2d\sum_{k=1}^n h_k^2\,.$$

Rearranging the terms, dividing by τ_n, and using the convexity of the Fisher information,

$$\mathsf{FI}(\bar\mu_{\tau_n} \,\|\, \pi) \le \frac{1}{\tau_n}\int_0^{\tau_n} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \le \frac{2\,\mathsf{KL}(\mu_0 \,\|\, \pi)}{\tau_n} + \frac{8L^2d}{\tau_n}\,S\,, \tag{12}$$

where $S \coloneqq \sum_{k=1}^\infty h_k^2 < \infty$. On the other hand, if $t \in [\tau_n, \tau_{n+1}]$, integrating (10) between τ_n and t shows that

$$\mathsf{KL}(\mu_t \,\|\, \pi) \le \mathsf{KL}(\mu_{\tau_n} \,\|\, \pi) + 4L^2d\,(t-\tau_n)^2 \le \mathsf{KL}(\mu_0 \,\|\, \pi) + 8L^2dS < \infty\,,$$

so that $\{\mathsf{KL}(\mu_t \,\|\, \pi) \mid t \ge 0\}$ is bounded. By convexity of the KL divergence, it also implies that $\{\mathsf{KL}(\bar\mu_{\tau_n} \,\|\, \pi) \mid n \in \mathbb{N}\}$ is uniformly bounded. Recalling that the sublevel sets of $\mathsf{KL}(\cdot \,\|\, \pi)$ are weakly compact we obtain that $(\bar\mu_{\tau_n})_{n\in\mathbb{N}}$ is tight. To show that $\bar\mu_{\tau_n} \to \pi$ weakly, it suffices to show that every cluster point of $(\bar\mu_{\tau_n})_{n\in\mathbb{N}}$ is equal to π. Consider a subsequence of $(\bar\mu_{\tau_n})_{n\in\mathbb{N}}$ converging to some cluster point $\bar\mu$.

Taking $n\to\infty$ in (12) and noting that $\tau_n\to\infty$ by our assumptions, we have $\mathsf{FI}(\bar\mu_{\tau_n}\parallel\pi)\to 0$, therefore this is still true along the subsequence. Using the weak lower semicontinuity of the Fisher information along the subsequence, $\mathsf{FI}(\bar\mu\parallel\pi)=0$. This means that for $f:=\frac{\mathrm{d}\bar\mu}{\mathrm{d}\pi}$, we have $\sqrt{f}\in\operatorname{dom}\mathscr{E}$ and $\mathscr{E}(\sqrt{f})=0$. Since ∇V is Lipschitz, then π has a continuous and strictly positive density on $\mathbb{R}^d$, so $\mathscr{E}(\sqrt{f})=0$ implies that f is a constant π-a.e., and hence $\bar\mu=\pi$. ■

C.2. Hessian smoothness

We first control the moments of LMC under Assumption 11.

Proposition 25 *Assume that the growth conditions in (4) are satisfied for $\gamma>0$, $0<\xi\le\gamma/2$, and $h\le\frac{a}{4m^2}\wedge 1$. Then, for the LMC iterates $(x_{kh})_{k\in\mathbb{N}}$, we have*

$$\mathbb{E}[\|x_{kh}\|^2]\le\mathbb{E}[\|x_0\|^2]+3\,(a+b+d)\,kh\,, \tag{13}$$

$$\mathbb{E}[\|x_{kh}\|^4]\le\mathbb{E}[\|x_0\|^4]+6\,\Big(\frac{3\,(a+b+d)}{1\wedge a}\Big)^{\frac{2+\gamma}{\gamma}\vee 2}\,kh\,. \tag{14}$$

Proof As before, we denote the interpolation diffusion with $\{x_t\}_{t\ge0}$ and the corresponding filtration with $\{\mathcal{F}_t\}_{t\ge0}$. Using Itô's formula conditioned on $\mathcal{F}_{kh}$, we obtain

$$\begin{aligned}\partial_t\,\mathbb{E}[\|x_t\|^2\mid\mathcal{F}_{kh}]&=-2\,\mathbb{E}[\langle x_t,\nabla V(x_{kh})\rangle\mid\mathcal{F}_{kh}]+2d\\&=-2\,\langle x_{kh},\nabla V(x_{kh})\rangle+2\,(t-kh)\,\|\nabla V(x_{kh})\|^2+2d\\&\le-2a\,(1+\|x_{kh}\|^\gamma)+2\,(a+b+d)+4\,(t-kh)\,m^2\,(1+\|x_{kh}\|^{2\xi})\\&\le 3a+2b+2d\end{aligned}$$

where we used $h\le a/(4m^2)$. Integrating this from kh to $(k+1)h$ and iterating yields the second moment bound in (25).

Similarly for the fourth moment, we write

$$\begin{aligned}\partial_t\,\mathbb{E}[\|x_t\|^4\mid\mathcal{F}_{kh}]&=-4\,\mathbb{E}[\|x_t\|^2\,\langle\nabla V(x_{kh}),x_t\rangle\mid\mathcal{F}_{kh}]+(4d+2)\,\mathbb{E}[\|x_t\|^2\mid\mathcal{F}_{kh}]\\&=4\,\mathbb{E}[\|x_t\|^2\mid\mathcal{F}_{kh}]\,\{-\langle\nabla V(x_{kh}),x_{kh}\rangle+(t-kh)\,\|\nabla V(x_{kh})\|^2+d+1/2\}\\&\qquad-16\,(t-kh)\,\langle x_{kh}-(t-kh)\,\nabla V(x_{kh}),\nabla V(x_{kh})\rangle\end{aligned}$$

where in the last step, we use Gaussian integration by parts

$$\begin{aligned}&-4\,\mathbb{E}[\|x_t\|^2\,\langle\nabla V(x_{kh}),\sqrt{2}\,(B_t-B_{kh})\rangle\mid\mathcal{F}_{kh}]\\&\quad=-16\,(t-kh)\,\langle x_{kh}-(t-kh)\,\nabla V(x_{kh}),\nabla V(x_{kh})\rangle\,.\end{aligned}$$

Therefore, we can use the growth condition and write

$$\begin{aligned}\partial_t\,\mathbb{E}[\|x_t\|^4\mid\mathcal{F}_{kh}]&\le4\,\{\mathbb{E}[\|x_t\|^2\mid\mathcal{F}_{kh}]+4\,(t-kh)\}\\&\qquad\times\{-\langle\nabla V(x_{kh}),x_{kh}\rangle+(t-kh)\,\|\nabla V(x_{kh})\|^2+d+1/2\}\\&\le4\,\{\mathbb{E}[\|x_t\|^2\mid\mathcal{F}_{kh}]+4\,(t-kh)\}\\&\qquad\times\{-a\,\|x_{kh}\|^\gamma+b+2\,(t-kh)\,m^2\,(1+\|x_{kh}\|^{2\xi})+d+1/2\}\,.\end{aligned}$$

Next, recalling that $h \le \min\{1, a/(4m^2)\}$ and using Assumption 11,

$$\partial_t \mathbb{E}[\|x_t\|^4 \mid \mathcal{F}_{kh}] \le 4\,\{\mathbb{E}[\|x_t\|^2 \mid \mathcal{F}_{kh}] + 4\,(t-kh)\} \times \{-\frac{a}{2}\,\|x_{kh}\|^\gamma + \frac{a}{2} + b + d + \frac{1}{2}\}\,.$$

Define $C := a + b + d$. We split into two cases. If $\|x_{kh}\| \ge ((a+2b+2d+1)/a)^{1/\gamma}$, then the time derivative is *negative*. Otherwise, if $\|x_{kh}\| \le ((a+2b+2d+1)/a)^{1/\gamma} \le (3C/a)^{1/\gamma}$, then recalling our second moment bound,

$$\mathbb{E}[\|x_t\|^2 \mid \mathcal{F}_{kh}] + 4\,(t-kh) \le \|x_{kh}\|^2 + (t-kh)\,(3a+2b+2d+4) \le \|x_{kh}\|^2 + 6C$$

and therefore

$$\partial_t \mathbb{E}[\|x_t\|^4 \mid \mathcal{F}_{kh}] \le 3\,\Bigl(\bigl(\frac{3C}{a}\bigr)^{2/\gamma} + 6C\Bigr)\,C \le 6\,\bigl(\frac{3C}{1\wedge a}\bigr)^{\frac{2+\gamma}{\gamma}\vee 2}\,.$$

This concludes the proof. ■

Proof [Proof of Theorem 12] Under Hessian smoothness, we can achieve tighter control on the discretization error via the fourth moment. To do this, we introduce the following lemma, which is derived from an intermediate result in the work of Mou et al. (2022).

Lemma 26 *Under Assumption 10, the following bound holds for the discretization error.*

$$\begin{aligned}
&\mathbb{E}\bigl[\|\nabla V(x_t) - \mathbb{E}[\nabla V(x_{kh}) \mid x_t]\|^2\bigr] \\
&\quad\le 4L^2\,(t-kh)^2\,\mathbb{E}[\|\nabla \ln \mu_{kh}(x_{kh})\|^2] + 12L^4\,(t-kh)^3\,d + 4L^2h^2\,\mathbb{E}[\|\nabla V(x_{kh})\|^2] \\
&\qquad + 4\,(t-kh)^4\,M^2\,\mathbb{E}[\|\nabla V(x_{kh})\|^4] + 48\,(t-kh)^2\,M^2d^2\,.
\end{aligned}$$

Proof This result follows from Mou et al. (2022), by combining the proof of their Lemma 3 (before substitution of $\|\nabla V(x_{kh})\|^4$), their Lemma 4, and the result for the term I_2 in their Lemma 5 with the bound on I_3 in the proof of their Lemma 5 before substituting for $\|\nabla V(x_{kh})\|^2$. ■

We invoke the following Lemma, also from Mou et al. (2022).

Lemma 27 (Mou et al. (2022, Lemma 7)) *For $h \le \frac{1}{2L}$ and all $t \in [kh, (k+1)h]$,*

$$\mathbb{E}[\|\nabla \ln \mu_{kh}(x_{kh})\|^2] \le 8\,\mathbb{E}[\|\nabla \ln \mu_t(x_t)\|^2] + 32M^2d^2h^2\,.$$

Consequently, we first analyze the first term in Lemma 26 for $t \in [kh, (k+1)h]$:

$$\begin{aligned}
\mathbb{E}[\|\nabla \ln \mu_{kh}(x_{kh})\|^2] &\lesssim \mathbb{E}[\|\nabla \ln \mu_t(x_t)\|^2] + M^2d^2h^2 \\
&= \mathbb{E}\bigl[\bigl\|\nabla \ln \frac{\mu_t}{\pi}(x_t) + \nabla V(x_t)\bigr\|^2\bigr] + M^2d^2h^2 \\
&\lesssim \mathbb{E}\bigl[\bigl\|\nabla \ln \frac{\mu_t}{\pi}(x_t)\bigr\|^2\bigr] + \mathbb{E}[\|\nabla V(x_t)\|^2] + M^2d^2h^2 \\
&\lesssim \mathsf{FI}(\mu_t \,\|\, \pi) + Ld + M^2d^2h^2\,,
\end{aligned}$$

where we applied Lemma 24. Similarly, we bound the term

$$\begin{aligned}
\mathbb{E}[\|\nabla V(x_{kh})\|^2] &\lesssim \mathbb{E}[\|\nabla V(x_t)\|^2] + \mathbb{E}[\|\nabla V(x_{kh}) - \nabla V(x_t)\|^2] \\
&\lesssim (1 + L^2\,(t-kh)^2)\,\mathbb{E}[\|\nabla V(x_t)\|^2] + L^2\,(t-kh)\,d \\
&\le \mathsf{FI}(\mu_t \,\|\, \pi) + Ld\,,
\end{aligned}$$

where we used (9), $h \lesssim 1/L$, and Lemma 24.

The primary term of concern is the expected fourth power of the gradient, $\mathbb{E}[\|\nabla V(x_{kh})\|^4]$. For large orders of growth $\xi > 1/2$, we can directly use the fourth moment bound found in Proposition 25, which has a worst case order of d^3. However, when $\xi \le 1/2$, the term $\mathbb{E}[\|\nabla V(x_{kh})\|^4]$ will only grow as the second moment $\mathbb{E}[\|x_{kh}\|^2]$, and consequently the order of this term is d. In both cases, this term is no longer dominant.

Case $\xi > 1/2$: Using our growth assumption, we get using Assumption 11 for $\xi > 1/2$

$$\begin{aligned}\mathbb{E}[\|\nabla V(x_{kh})\|^4] &\lesssim m^4\,(1 + \|x_{kh}\|^4)\\ &\lesssim m^4 \left(1 + \mathbb{E}[\|x_0\|^4] + \Big(\frac{3\,(a+b+d)}{1\wedge a}\Big)^{\frac{2+\gamma}{\gamma}\vee 2} kh\right)\\ &\lesssim m^4\,\big(1 + \mathbb{E}[\|x_0\|^4] + (b+d)^3\,kh\big)\,,\end{aligned}$$

where the last line follows as $\xi > 1/2$ implies $\gamma > 1$.

Case $\xi \le 1/2$: In this case, when we use Assumption 11 for $\xi \le 1/2$

$$\begin{aligned}\mathbb{E}[\|\nabla V(x_{kh})\|^4] &\lesssim m^4\,(1 + \|x_{kh}\|^2)\\ &\lesssim m^4\,\big(1 + \mathbb{E}[\|x_0\|^2] + (b+d)\,kh\big)\,.\end{aligned}$$

As we shall see, it will suffice for simplicity in both cases to use the worst case bound for all $k \le N$,

$$\mathbb{E}[\|\nabla V(x_{kh})\|^4] \lesssim m^4\,(b+\sigma d)^3\,Nh\,.$$

Substituting all of these terms into Lemma 26, we get for $t \in [kh, (k+1)h]$ and $h \lesssim \frac{1}{L}$

$$\begin{aligned}\mathbb{E}\big[\|\nabla V(x_t) - \mathbb{E}[\nabla V(x_{kh}) \mid x_t]\|^2\big] \lesssim L^2h^2\,\mathsf{FI}(\mu_t \,\|\, \pi) + L^3d^2h^2 + L^2M^2d^2h^4\\ + M^2m^4\,(b+\sigma d)^3\,Nh^5 + M^2d^2h^2\,.\end{aligned}$$

Finally, from the differential inequality of Lemma 23, we get

$$\partial_t\,\mathsf{KL}(\mu_t \,\|\, \pi) \le -\frac{3}{4}\,\mathsf{FI}(\mu_t \,\|\, \pi) + \mathbb{E}\big[\|\nabla V(x_t) - \mathbb{E}[\nabla V(x_{kh}) \mid x_t]\|^2\big]$$

and so for $h \lesssim \frac{1}{L}$,

$$\mathsf{FI}(\mu_t \,\|\, \pi) \lesssim -\partial_t\,\mathsf{KL}(\mu_t \,\|\, \pi) + (1 \vee L^3 \vee M^2)\,d^2h^2 + M^2m^4\,(b+\sigma d)^3\,Nh^5\,.$$

Finally, we integrate and average over the time horizon to get

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \lesssim \frac{\mathsf{KL}(\mu_0 \,\|\, \pi)}{Nh} + (1 \vee L^3 \vee M^2)\,d^2h^2 + M^2m^4\,(b+\sigma d)^3\,Nh^5\,.$$

Consequently, if we define $\kappa = 1 \vee L \vee M^{2/3} \vee (M^{1/3}m^{2/3})$, then if $h \asymp \frac{K_0^{1/3}}{\kappa\,(b+\sigma d)^{2/3}\,N^{1/3}}$, we get

$$\frac{1}{Nh}\int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t \lesssim \left((b+\sigma d)^{2/3}K_0^{2/3} + \frac{K_0^{5/3}}{(b+\sigma d)^{1/3}}\right)\frac{\kappa}{N^{2/3}}\,.$$

This completes the proof. ■

C.3. Stochastic gradient setting

Proof [Proof of Theorem 15] Using Lemma 23, we have

$$\partial_t \operatorname{KL}(\mu_t \,\|\, \pi) \le -\frac{3}{4} \operatorname{FI}(\mu_t \,\|\, \pi) + \mathbb{E}[\|\nabla V(x_t) - G(x_{kh}, \zeta_k)\|^2]\,.$$

The error term can be bounded via

$$\begin{aligned}
\mathbb{E}[\|\nabla V(x_t) - G(x_{kh}, \zeta_k)\|^2] &\le 3\,\mathbb{E}[\|\nabla V(x_t) - \nabla \hat V(x_t)\|^2] + 3\,\mathbb{E}[\|\nabla \hat V(x_t) - \nabla \hat V(x_{kh})\|^2] \\
&\quad + 3\,\mathbb{E}[\|\nabla \hat V(x_{kh}) - G(x_{kh}, \zeta_k)\|^2] \\
&\le 3\,\mathsf{bias}^2 + 3\,\mathsf{var} + 3\hat L^2\,\mathbb{E}[\|x_t - x_{kh}\|^2]\,.
\end{aligned}$$

Next, we have

$$\begin{aligned}
\mathbb{E}[\|x_t - x_{kh}\|^2] &= (t-kh)^2\,\mathbb{E}[\|G(x_{kh}, \zeta_k)\|^2] + 2\,\mathbb{E}[\|B_t - B_{kh}\|^2] \\
&\le 2\,\mathsf{var}\,(t-kh)^2 + 2\,(t-kh)^2\,\mathbb{E}[\|\nabla \hat V(x_{kh})\|^2] + 2d\,(t-kh)\,.
\end{aligned}$$

Using smoothness of $\hat V$,

$$\mathbb{E}[\|\nabla \hat V(x_{kh})\|^2] \le 2\,\mathbb{E}[\|\nabla \hat V(x_t)\|^2] + 2\hat L^2\,\mathbb{E}[\|x_t - x_{kh}\|^2]\,.$$

Substitute this into the previous inequality. If $h \le 1/(\sqrt{8}\hat L)$, we can rearrange to obtain

$$\mathbb{E}[\|x_t - x_{kh}\|^2] \le 4\,\mathsf{var}\,(t-kh)^2 + 8\,(t-kh)^2\,\mathbb{E}[\|\nabla \hat V(x_t)\|^2] + 4d\,(t-kh)\,.$$

Next, to bound $\mathbb{E}[\|\nabla \hat V(x_t)\|^2]$, we generalize the proof of Lemma 24. Introduce the generator $\mathcal{L}$ of the Langevin diffusion. Since $\mathcal{L}\hat V = -\Delta \hat V + \langle \nabla V, \nabla \hat V\rangle$, we can write

$$\begin{aligned}
\mathbb{E}_{\mu_t}[\|\nabla \hat V\|^2] &= \mathbb{E}_{\mu_t}[\mathcal{L}\hat V + \Delta \hat V + \langle \nabla \hat V, \nabla \hat V - \nabla V\rangle] \\
&\le \mathbb{E}_{\mu_t}\,\mathcal{L}\hat V + \hat L d + \sqrt{\mathbb{E}_{\mu_t}[\|\nabla \hat V\|^2]\,\mathbb{E}_{\mu_t}[\|\nabla \hat V - \nabla V\|^2]} \\
&\le \mathbb{E}_{\mu_t}\,\mathcal{L}\hat V + \hat L d + \sqrt{\mathsf{bias}^2\,\mathbb{E}_{\mu_t}[\|\nabla \hat V\|^2]}\,.
\end{aligned}$$

For the first term, we can use an integration by parts argument as in the proof of Lemma 24:

$$\mathbb{E}_{\mu_t}\,\mathcal{L}\hat V = \mathbb{E}_{\mu_t}\big\langle \nabla \hat V, \nabla \ln \frac{\mu_t}{\pi}\big\rangle \le \sqrt{\mathbb{E}_{\mu_t}[\|\nabla \hat V\|^2]\operatorname{FI}(\mu_t \,\|\, \pi)}\,.$$

Applying Young's inequality,

$$\mathbb{E}_{\mu_t}[\|\nabla \hat V\|^2] \le \frac{1}{4}\,\mathbb{E}_{\mu_t}[\|\nabla \hat V\|^2] + \operatorname{FI}(\mu_t \,\|\, \pi) + \hat L d + \frac{1}{4}\,\mathbb{E}_{\mu_t}[\|\nabla \hat V\|^2] + \mathsf{bias}^2$$

which is rearranged to yield

$$\mathbb{E}_{\mu_t}[\|\nabla \hat V\|^2] \le 2\operatorname{FI}(\mu_t \,\|\, \pi) + 2\hat L d + 2\,\mathsf{bias}^2\,.$$

Therefore,

$$\begin{aligned}\mathbb{E}[\|\nabla V(x_t) - G(x_{kh}, \zeta_k)\|^2] &\le 3\,\mathsf{bias}^2 + 3\,\mathsf{var} + 12\hat{L}^2\,\mathsf{var}\,(t-kh)^2 \\ &\quad + 48\hat{L}^2\,\{\mathsf{FI}(\mu_t \,\|\, \pi) + \hat{L}d + \mathsf{bias}^2\}\,(t-kh)^2 + 12\hat{L}^2 d\,(t-kh)\end{aligned}$$

and for $h \le 1/(\sqrt{192}\hat{L})$ we can absorb the Fisher information term into the differential inequality for the KL divergence:

$$\begin{aligned}\partial_t \mathsf{KL}(\mu_t \,\|\, \pi) &\le -\frac{1}{2}\,\mathsf{FI}(\mu_t \,\|\, \pi) + 3\,\mathsf{bias}^2 + 3\,\mathsf{var} + 12\hat{L}^2\,\mathsf{var}\,(t-kh)^2 \\ &\quad + 48\hat{L}^2\,(\hat{L}d + \mathsf{bias}^2)\,(t-kh)^2 + 12\hat{L}^2 d\,(t-kh)\,.\end{aligned}$$

Integrating,

$$\begin{aligned}\mathsf{KL}(\mu_{(k+1)h} \,\|\, \pi) - \mathsf{KL}(\mu_{kh} \,\|\, \pi) &\le -\frac{1}{2}\int_{kh}^{(k+1)h} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t + 3h\,\mathsf{bias}^2 + 3h\,\mathsf{var} + 4\hat{L}^2 h^3\,\mathsf{var} \\ &\quad + 16\hat{L}^2 h^3\,(\hat{L}d + \mathsf{bias}^2) + 6\hat{L}^2 d h^2 \\ &\le -\frac{1}{2}\int_{kh}^{(k+1)h} \mathsf{FI}(\mu_t \,\|\, \pi)\,\mathrm{d}t + 4h\,\mathsf{bias}^2 + 4h\,\mathsf{var} + 8\hat{L}^2 d h^2\,.\end{aligned}$$

The proof is concluded in the same way as Theorem 4. ■

C.4. Gaussian smoothing

Proof [Proof of Lemma 17] From a Gaussian integration by parts argument (see Nesterov and Spokoiny, 2017), writing γ for the standard Gaussian measure on $\mathbb{R}^d$,

$$\begin{aligned}\|\nabla\hat{V}(x) - \nabla V(x)\| &= \Big\|\int \Big(\frac{V(x+\eta\zeta) - V(x)}{\eta} - \langle \nabla V(x), \zeta\rangle\Big)\,\zeta\,\gamma(\mathrm{d}\zeta)\Big\| \\ &\le \frac{1}{\eta}\int |V(x+\eta\zeta) - V(x) - \eta\,\langle\nabla V(x), \zeta\rangle|\,\|\zeta\|\,\gamma(\mathrm{d}\zeta) \\ &= \int \Big|\Big\langle \int_0^1 \{\nabla V(x+t\eta\zeta) - \nabla V(x)\}\,\mathrm{d}t, \zeta\Big\rangle\Big|\,\|\zeta\|\,\gamma(\mathrm{d}\zeta) \\ &\le \int \Big(\int_0^1 \|\nabla V(x+t\eta\zeta) - \nabla V(x)\|\,\mathrm{d}t\Big)\,\|\zeta\|^2\,\gamma(\mathrm{d}\zeta) \\ &\le L\eta^s \int \|\zeta\|^{2+s}\,\gamma(\mathrm{d}\zeta) \asymp L d^{(2+s)/2}\eta^s\,.\end{aligned}$$

The last inequality follows from standard bounds on the Gaussian moments. ■

Proof [Proof of Corollary 19] We proceed via the following steps.

1. Control of the bias. Let $\hat{\pi} \propto \exp(-\hat{V})$ and assume that the potential V is normalized so that $\int \exp(-V) = 1$. From Chatterji et al. (2020, Lemma 2.2), we know that $\sup|\hat{V} - V| \le L d^{(1+s)/2}\eta^{1+s}$. Then,

$$\frac{\hat{\pi}}{\pi} = \frac{\exp(V - \hat{V})}{\int \exp(-\hat{V})} \le \frac{\exp(V - \hat{V})}{\exp(-\sup|V - \hat{V}|)\int \exp(-V)} \le \exp\big(2\sup|V - \hat{V}|\big)\,.$$

For η small, we deduce from Pinsker's inequality that

$$\|\hat\pi - \pi\|_{\mathrm{TV}}^2 \lesssim \mathsf{KL}(\hat\pi \,\|\, \pi) \le \ln \sup \frac{\hat\pi}{\pi} \le 2 \sup |V - \hat V| \lesssim L d^{(1+s)/2} \eta^{1+s}\,.$$

Hence, provided $\eta \lesssim \varepsilon^{1/(1+s)}/(L^{1/(1+s)} d^{1/2})$, we can ensure that $\|\hat\pi - \pi\|_{\mathrm{TV}}^2 \le \frac{\varepsilon}{4}$.

2. Convergence to the smoothed potential. We next apply Theorem 15 with the target distribution $\hat\pi$. Due to the mini-batching of the stochastic gradients,

$$\hat L \lesssim \frac{L d^{(1-s)/2}}{\eta^{1-s}}\,, \qquad \mathsf{var} \lesssim \frac{L^2 d^s \eta^{2s}}{B}\,.$$

Since we are viewing the smoothed potential $\hat\pi$ as the target, then $\mathsf{bias}^2 = 0$. Therefore, Theorem 15 implies that $\mathsf{FI}(\bar\mu_{Nh} \,\|\, \hat\pi) \le \delta$ after N iterations, provided that $\eta \lesssim B^{1/(2s)} \delta^{1/(2s)}/(L^{1/s} d^{1/2})$ and

$$N \gtrsim \frac{K_0 L^2 d^{2-s}}{\delta^2 \eta^{2\,(1-s)}}\,.$$

3. The smoothed potential satisfies a Poincaré inequality. From the first step, our choice of η entails that $\hat\pi$ is a bounded perturbation of π, and hence $\hat\pi$ satisfies a Poincaré inequality with constant $\lesssim C_{\mathsf{PI}}$ (Bakry et al., 2014, Proposition 4.2.7). Applying Lemma 7, we obtain

$$\|\bar\mu_{Nh} - \hat\pi\|_{\mathrm{TV}}^2 \lesssim C_{\mathsf{PI}}\, \mathsf{FI}(\bar\mu_{Nh} \,\|\, \hat\pi)\,.$$

Setting $\delta \asymp \varepsilon / C_{\mathsf{PI}}$, we see that provided $\eta \lesssim B^{1/(2s)} \varepsilon^{1/(2s)}/(C_{\mathsf{PI}}^{1/(2s)} L^{1/s} d^{1/2})$ and

$$N \gtrsim \frac{C_{\mathsf{PI}}^2 K_0 L^2 d^{2-s}}{\varepsilon^2 \eta^{2\,(1-s)}}\,,$$

we obtain $\|\bar\mu_{Nh} - \hat\pi\|_{\mathrm{TV}}^2 \le \frac{\varepsilon}{4}$.

4. Conclusion of the proof. Putting the steps together,

$$\|\bar\mu_{Nh} - \pi\|_{\mathrm{TV}}^2 \le 2\, \|\bar\mu_{Nh} - \hat\pi\|_{\mathrm{TV}}^2 + 2\, \|\hat\pi - \pi\|_{\mathrm{TV}}^2 \le \varepsilon\,.$$

To fulfill the conditions on η, we take

$$\eta \asymp \frac{1}{d^{1/2}} \min\Bigl\{ \frac{\varepsilon^{1/(1+s)}}{L^{1/(1+s)}}, \frac{B^{1/(2s)} \varepsilon^{1/(2s)}}{C_{\mathsf{PI}}^{1/(2s)} L^{1/s}} \Bigr\}\,. \tag{15}$$

The gradient complexity is

$$BN \asymp \frac{C_{\mathsf{PI}}^2 K_0 L^2 d^{3-2s}}{\varepsilon^2} \times B \times \max\Bigl\{ \frac{L^{1/(1+s)}}{\varepsilon^{1/(1+s)}}, \frac{C_{\mathsf{PI}}^{1/(2s)} L^{1/s}}{B^{1/(2s)} \varepsilon^{1/(2s)}} \Bigr\}^{2\,(1-s)}\,.$$

Now we optimize over B. If $s \ge 1/2$, then we set $B = 1$, with complexity

$$BN \asymp \frac{C_{\mathsf{PI}}^{(1+s)/s} K_0 L^{2/s} d^{3-2s}}{\varepsilon^{(1+s)/s}}\,.$$

Otherwise, if $s \le 1/2$, we set $B \asymp C_{\mathsf{PI}} L^{2/(1+s)}/\varepsilon^{(1-s)/(1+s)}$, with complexity

$$BN \asymp \frac{C_{\mathsf{PI}}^3 K_0 L^{6/(1+s)} d^{3-2s}}{\varepsilon^{(5-s)/(1+s)}}\,.$$

This completes the proof. ∎

C.5. Finite sum setting

Proof [Proof of Theorem 21] Let $(x_t)_{t\ge 0}$ denote the interpolation of (VR-LMC). Using Lemma 23, for $t \in [kh, (k+1)h]$, we have

$$\begin{aligned}
\partial_t \mathsf{KL}(\mu_t \,\|\, \pi) &\le -\frac{3}{4}\,\mathsf{FI}(\mu_t \,\|\, \pi) + \mathbb{E}[\|\nabla V(x_t) - g_k\|^2] \\
&\le -\frac{3}{4}\,\mathsf{FI}(\mu_t \,\|\, \pi) + 2\,\mathbb{E}[\|\nabla V(x_t) - \nabla V(x_{kh})\|^2] + 2\,\mathbb{E}[\|\nabla V(x_{kh}) - g_k\|^2]\,. \qquad (16)
\end{aligned}$$

The second term in (16) can be further bounded as

$$\begin{aligned}
\mathbb{E}[\|\nabla V(x_t) - \nabla V(x_{kh})\|^2] &\le L^2\,\mathbb{E}[\|x_t - x_{kh}\|^2] = L^2\,(t-kh)^2\,\mathbb{E}[\|g_k\|^2] + 2L^2 d\,(t-kh) \\
&\le L^2 h^2\,\mathbb{E}[\|g_k\|^2] + 2L^2 dh = L^2\,\mathbb{E}[\|x_{(k+1)h} - x_{kh}\|^2]\,. \qquad (17)
\end{aligned}$$

Furthermore, write $\sigma_k^2 = \mathbb{E}[\|g_k - \nabla V(x_{kh})\|^2]$ for the variance term. The third term in (16) can be bounded as

$$\begin{aligned}
\sigma_{k+1}^2 &= (1-p)\,\mathbb{E}[\|g_k - \nabla V(x_{(k+1)h}) + \nabla f_i(x_{(k+1)h}) - \nabla f_i(x_{kh})\|^2] \\
&= (1-p)\,\mathbb{E}[\|g_k - \nabla V(x_{kh}) + \underbrace{(\nabla f_i(x_{(k+1)h}) - \nabla f_i(x_{kh}))}_{:=a_i} \\
&\qquad\qquad - \underbrace{(\nabla V(x_{(k+1)h}) - \nabla V(x_{kh}))}_{:=\bar a = \frac{1}{n}\sum_{\ell=1}^n a_\ell}\|^2] \\
&= (1-p)\,\mathbb{E}[\|g_k - \nabla V(x_{kh})\|^2] + (1-p)\,\frac{1}{n}\sum_{\ell=1}^n \mathbb{E}[\|a_\ell - \bar a\|^2] \\
&\le (1-p)\,\mathbb{E}[\|g_k - \nabla V(x_{kh})\|^2] + (1-p)\,\frac{1}{n}\sum_{\ell=1}^n \mathbb{E}[\|a_\ell\|^2] \\
&\le (1-p)\,\sigma_k^2 + (1-p)\,L^2\,\mathbb{E}[\|x_{(k+1)h} - x_{kh}\|^2]\,.
\end{aligned}$$

In the third equality, we conditioned w.r.t. $\mathcal{F}_k$ and used that i is independent of $\mathcal{F}_k$. Therefore, we obtain the inequality

$$\sigma_k^2 \le \frac{1-p}{p}\,L^2\,\mathbb{E}[\|x_{(k+1)h} - x_{kh}\|^2] - \frac{1}{p}\,(\sigma_{k+1}^2 - \sigma_k^2)\,. \qquad (18)$$

Plugging (18) and (17) into (16) we obtain

$$\partial_t \mathsf{KL}(\mu_t \,\|\, \pi) \le -\frac{3}{4}\,\mathsf{FI}(\mu_t \,\|\, \pi) + \frac{2L^2}{p}\,\mathbb{E}[\|x_{(k+1)h} - x_{kh}\|^2] - \frac{2}{p}\,(\sigma_k^2 - \sigma_{k+1}^2)\,. \qquad (19)$$

Now, we bound the term $\mathbb{E}[\|x_{(k+1)h} - x_{kh}\|^2]$ appearing in (19) as

$$\begin{aligned}
\mathbb{E}[\|x_{(k+1)h} - x_{kh}\|^2] &= h^2\,\mathbb{E}[\|g_k\|^2] + 2hd \\
&\le h^2\,\mathbb{E}[\|\nabla V(x_{kh})\|^2] + h^2\sigma_k^2 + 2hd \\
&\le 2h^2\,\mathbb{E}[\|\nabla V(x_t)\|^2] + 2h^2\,\mathbb{E}[\|\nabla V(x_t) - \nabla V(x_{kh})\|^2] + h^2\sigma_k^2 + 2hd \\
&\le 2h^2\,\mathbb{E}[\|\nabla V(x_t)\|^2] + 2L^2h^2\,\mathbb{E}[\|x_{(k+1)h} - x_{kh}\|^2] + h^2\sigma_k^2 + 2hd\,,
\end{aligned}$$

where we used (17). Further using (18), we obtain

$$\mathbb{E}[\|x_{(k+1)h} - x_{kh}\|^2] \le 2h^2\, \mathbb{E}[\|\nabla V(x_t)\|^2] - \frac{h^2}{p}\,(\sigma_{k+1}^2 - \sigma_k^2) + h^2 L^2\, \frac{1+p}{p}\, \mathbb{E}[\|x_{(k+1)h} - x_{kh}\|^2] + 2hd\,.$$

Assuming $h^2 L^2 \le p/24$, we have

$$\frac{11}{12}\, \mathbb{E}[\|x_{(k+1)h} - x_{kh}\|^2] \le 2h^2\, \mathbb{E}[\|\nabla V(x_t)\|^2] - \frac{h^2}{p}\,(\sigma_{k+1}^2 - \sigma_k^2) + 2hd\,.$$

Using Lemma 24, we obtain

$$\begin{aligned} 2\, \mathbb{E}[\|x_{(k+1)h} - x_{kh}\|^2] &\le 6h^2\, \mathsf{FI}(\mu_t \,\|\, \pi) - \frac{3h^2}{p}\,(\sigma_{k+1}^2 - \sigma_k^2) + 6hd + 12Lh^2 d \\ &\le 6h^2\, \mathsf{FI}(\mu_t \,\|\, \pi) - \frac{3h^2}{p}\,(\sigma_{k+1}^2 - \sigma_k^2) + 9hd\,. \end{aligned} \tag{20}$$

Plugging (20) into (19), we obtain

$$\begin{aligned} \partial_t \mathsf{KL}(\mu_t \,\|\, \pi) &\le \big(-\frac{3}{4} + \frac{6L^2h^2}{p}\big)\, \mathsf{FI}(\mu_t \,\|\, \pi) + \frac{9L^2hd}{p} - \frac{2}{p}\,\big(1 + \frac{3L^2h^2}{2p}\big)\,(\sigma_{k+1}^2 - \sigma_k^2) \\ &\le -\frac{1}{2}\, \mathsf{FI}(\mu_t \,\|\, \pi) + \frac{9L^2hd}{p} - \frac{2}{p}\,\big(1 + \frac{3L^2h^2}{2p}\big)\,(\sigma_{k+1}^2 - \sigma_k^2)\,, \end{aligned} \tag{21}$$

where we used $L^2h^2 \le p/24$. Integrating between kh and $(k+1)h$,

$$\mathscr{L}_{k+1} - \mathscr{L}_k \le -\frac{1}{2} \int_{kh}^{(k+1)h} \mathsf{FI}(\mu_t \,\|\, \pi)\, \mathrm{d}t + \frac{9L^2hd}{p}\,,$$

where $\mathscr{L}_k \coloneqq \mathsf{KL}(\mu_{kh} \,\|\, \pi) + \frac{2h}{p}\,(1 + \frac{3L^2h^2}{2p})\,\sigma_k^2 \ge 0$. Iterating, and using $\mathscr{L}_k \ge 0$,

$$\frac{1}{Nh} \int_0^{Nh} \mathsf{FI}(\mu_t \,\|\, \pi)\, \mathrm{d}t \le \frac{2\mathscr{L}_0}{Nh} + \frac{18L^2hd}{p}\,.$$

Since $h^2L^2 < p/24$, we have

$$\mathscr{L}_0 = \mathsf{KL}(\mu_0 \,\|\, \pi) + \frac{2h}{p}\,\big(1 + \frac{3L^2h^2}{2p}\big)\,\sigma_0^2 \le \mathsf{KL}(\mu_0 \,\|\, \pi) + \frac{3h}{p}\,\sigma_0^2 = C\,,$$

thereby completing the first claim. By setting $h = \frac{\sqrt{pC}}{3L\sqrt{Nd}}$, we obtain the second. ■