LANGEVIN MONTE CARLO FOR STRONGLY LOG-CONCAVE DISTRIBUTIONS: RANDOMIZED MID-POINT REVISITED

Lu Yu
CREST, ENSAE, IP Paris
lu.yu@ensae.fr

Avetik Karagulyan
KAUST
avetik.karagulyan@kaust.edu.sa

Arnak Dalalyan
CREST, ENSAE, IP Paris
arnak.dalalyan@ensae.fr

ABSTRACT

We revisit the problem of sampling from a target distribution that has a smooth strongly log-concave density everywhere in $\mathbb{R}^p$. In this context, if no additional density information is available, the randomized midpoint discretization for the kinetic Langevin diffusion is known to be the most scalable method in high dimensions with large condition numbers. Our main result is a nonasymptotic and easy to compute upper bound on the W_2-error of this method. To provide a more thorough explanation of our method for establishing the computable upper bound, we conduct an analysis of the midpoint discretization for the vanilla Langevin process. This analysis helps to clarify the underlying principles and provides valuable insights that we use to establish an improved upper bound for the kinetic Langevin process with the midpoint discretization. Furthermore, by applying these techniques we establish new guarantees for the kinetic Langevin process with Euler discretization, which have a better dependence on the condition number than existing upper bounds.

1 INTRODUCTION

The task of sampling from target distributions with smooth, strongly log-concave densities has been a long-standing challenge in various fields such as statistics, machine learning, and computational physics (Andrieu et al., 2003; Krauth, 2006; Andrieu et al., 2010). Over the years, researchers have developed several algorithms to tackle this problem, and one prominent approach are the Langevin algorithms (Rogers & Williams, 2000; Oksendal, 2013; Robert et al., 1999). Langevin algorithms leverage the Langevin equation to design efficient and effective sampling algorithms. These methods generate a Markov chain by iteratively updating the position of a particle based on the Langevin equation. By simulating the particle's motion over time, these algorithms explore the target distribution and eventually converge to samples that approximate the desired distribution (Robert et al., 1999).

The canonical sampling algorithm, Langevin Monte Carlo (LMC) (Roberts & Tweedie, 1996; Dalalyan, 2017; Durmus & Moulines, 2017; Erdogdu & Hosseinzadeh, 2021; Mousavi-Hosseini et al., 2023; Raginsky et al., 2017; Erdogdu et al., 2018; Mou et al., 2022; Erdogdu et al., 2022), is a Markov chain Monte Carlo (MCMC) method that simulates the dynamics of a fictitious particle moving through a potential energy landscape defined by the target distribution. Formally, it is the Euler-Maruyama discretization of an SDE known as the Langevin diffusion. The underlying idea can be traced back to the early 20th century when Paul Langevin introduced a stochastic differential equation (SDE) to describe the motion of a particle in a fluid (Langevin, 1908). This SDE, combines deterministic and random components to model the particle's behavior under the influence of both a deterministic force and random noise.

One popular variant of the Langevin Monte Carlo is based on discretizing the kinetic Langevin diffusion, which introduces a friction term to control the exploration-exploitation trade-off during the sampling process (Einstein, 1905; Von Smoluchowski, 1906). According to Nelson (1967), the Langevin diffusion is the rescaled limit of the kinetic Langevin diffusion. Its ergodicity and mixing-time properties are studied in Eberle et al. (2019); Dalalyan & Riou-Durand (2020). Euler-Maruyama time discretization of this SDE, called kinetic Langevin Monte Carlo (KLMC), is prevalent in the sampling literature (Cheng et al., 2018b; Dalalyan & Riou-Durand, 2020; Shen & Lee, 2019; Ma et al., 2021; Zhang et al., 2023).

The randomized midpoint discretization method, as an alternative to the Euler-Maruyama scheme for KLMC, is proposed by Shen & Lee (2019). They demonstrate the superior performance of this method in terms of both tolerance and condition number dependency. More recently, He et al. (2020) analyze probabilistic properties of the randomized midpoint discretization method for the (kinetic) Langevin diffusion. In this work, we undertake a comprehensive and thorough analysis of the randomized midpoint discretization scheme for the kinetic Langevin diffusion under strongly log-concavity. To achieve this, we introduce a novel proof technique relying on summation by part, which helps to establish improved non-asymptotic and computable upper bounds on the discretization error for this method. Our contributions can be summarized as follows.

- To lay the groundwork for our analysis, we initially delve into the midpoint discretization technique applied to the vanilla Langevin process. In this context, we introduce our novel proof technique, which plays a pivotal role in our study. Notably, in Theorem 1, we provide the convergence guarantees for RLMC in W_2-distance with explicit constants and a transparent reliance on the initialization. These guarantees are competitive with the best available results for LMC, and could be leveraged to derive an improved upper bound specifically tailored for RKLMC.
- We further extend these techniques to RKLMC, and provide the corresponding convergence guarantees in W_2-distance in Theorem 2. Compared to the previous works, our bound **a)** contains small constants and the explicit dependence on the initialization, **b)** does not require the initialization to be at the minimizer of the potential, **c)** and is free from the linear dependence on the sample size, which serves as a crucial step towards the method applied to non-convex potentials.
- Employing the same techniques, we finally examine the convergence behavior of the KLMC algorithm with the Euler-Maruyama discretization. In Theorem 3, we provide an upper bound on the accuracy of this scheme in W_2-distance with improved dependence on the condition number.

We offer a systematic and unified treatment of the variants of LMC, which empowers us to derive enhanced upper bounds for the W_2-error associated with RKLMC, RLMC, and KLMC algorithms. Furthermore, our techniques facilitate the determination of explicit constants and the dependence on initialization, providing us with a clearer basis for choosing the step size and comparing the convergence rates across these methods.

Notation. Denote the p-dimensional Euclidean space by $\mathbb{R}^p$. The letter $\boldsymbol{\theta}$ denotes the deterministic vector and its calligraphic counterpart $\boldsymbol{\vartheta}$ denotes the random vector. We use $\mathbf{I}_p$ and $\mathbf{0}_p$ to denote, respectively, the $p \times p$ identity and zero matrices. Define the relations $\mathbf{A} \preccurlyeq \mathbf{B}$ and $\mathbf{B} \succcurlyeq \mathbf{A}$ for two symmetric $p \times p$ matrices $\mathbf{A}$ and $\mathbf{B}$ to mean that $\mathbf{B} - \mathbf{A}$ is positive semi-definite. The gradient and the Hessian of a function $f : \mathbb{R}^p \to \mathbb{R}$ are denoted by ∇f and $\nabla^2 f$, respectively. Given any pair of measures μ and ν defined on $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$, the Wasserstein-2 distance between μ and ν is defined as

$$\mathsf{W}_2(\mu, \nu) = \Big(\inf_{\varrho \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^p \times \mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 \, \mathrm{d}\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}') \Big)^{1/2},$$

where the infimum is taken over all joint distributions ϱ that have μ and ν as marginals.

2 UNDERSTANDING THE RANDOMIZED MIDPOINT DISCRETIZATION: THE VANILLA LANGEVIN DIFFUSION

The goal is to sample a random vector in $\mathbb{R}^p$ according to a given distribution π of the form

$$\pi(\boldsymbol{\theta}) \propto \exp\{-f(\boldsymbol{\theta})\}, \qquad \boldsymbol{\theta} \in \mathbb{R}^p,$$

with a function $f : \mathbb{R}^p \to \mathbb{R}$, referred to as the potential. Throughout the paper, we assume that the potential function f is M-smooth and m-strongly convex for some constants $0 < m \leqslant M < \infty$.

Assumption 1. *The function* $f : \mathbb{R}^p \to \mathbb{R}$ *is twice differentiable, and its Hessian matrix* $\nabla^2 f$ *satisfies*

$$m\mathbf{I}_p \preccurlyeq \nabla^2 f(\boldsymbol{\theta}) \preccurlyeq M\mathbf{I}_p, \qquad \forall \boldsymbol{\theta} \in \mathbb{R}^p.$$

Let $\boldsymbol{\vartheta}_0$ be a random vector drawn from a distribution ν on $\mathbb{R}^p$ and let $\boldsymbol{W} = (\boldsymbol{W}_t : t \geqslant 0)$ be a p-dimensional Brownian motion independent of $\boldsymbol{\vartheta}_0$. Using the potential f, the random variable $\boldsymbol{\vartheta}_0$ and the process $\boldsymbol{W}$, one can define the stochastic differential equation

$$d\boldsymbol{L}_t^{\mathsf{LD}} = -\nabla f(\boldsymbol{L}_t^{\mathsf{LD}})\,dt + \sqrt{2}\,\mathrm{d}\boldsymbol{W}_t, \qquad t \geqslant 0, \qquad \boldsymbol{L}_0^{\mathsf{LD}} = \boldsymbol{\vartheta}_0. \tag{1}$$

This equation has a unique strong solution, which is a continuous-time Markov process, termed Langevin diffusion. Under some further assumptions on f, such as strong convexity or dissipativity, the Langevin diffusion is ergodic, geometrically mixing and has π as its unique invariant distribution (Bhattacharya, 1978). Furthermore, the mixing properties of this process can be quantified. For instance, if π satisfies the Poincaré inequality with constant C_{P}, then (see e.g. Chewi et al. (2020)) the distribution ν_t^{LD} of $\boldsymbol{L}_t^{\mathsf{LD}}$ satisfies

$$\mathsf{W}_2(\nu_t^{\mathsf{LD}}, \pi) \leqslant e^{-t/C_{\mathsf{P}}}\sqrt{2C_{\mathsf{P}}\chi^2(\nu\|\pi)}, \qquad \forall t \geqslant 0.$$

These results suggest that we can sample from the distribution π by using a suitable discretization of the Langevin diffusion. The Langevin Monte Carlo (LMC) method is based on this idea, combining the aforementioned considerations with the Euler discretization. Specifically, for small values of $h \geqslant 0$ and $\Delta_h \boldsymbol{W}_t = \boldsymbol{W}_{t+h} - \boldsymbol{W}_t$, the following approximation holds

$$\boldsymbol{L}_{t+h}^{\mathsf{LD}} = \boldsymbol{L}_t^{\mathsf{LD}} - \int_0^h \nabla f(\boldsymbol{L}_{t+s}^{\mathsf{LD}})\,\mathrm{d}s + \sqrt{2}\,\Delta_h \boldsymbol{W}_t \approx \boldsymbol{L}_t^{\mathsf{LD}} - h\nabla f(\boldsymbol{L}_t^{\mathsf{LD}}) + \sqrt{2}\,\Delta_h \boldsymbol{W}_t.$$

By repeatedly applying this approximation with a small step-size h, we can construct a Markov chain $(\boldsymbol{\vartheta}_k^{\mathsf{LMC}} : k \in \mathbb{N})$ that converges to the target distribution π as h goes to zero. More precisely, $\boldsymbol{\vartheta}_k^{\mathsf{LMC}} \approx \boldsymbol{L}_{kh}^{\mathsf{LD}}$, for $k \in \mathbb{N}$, is given by

$$\boldsymbol{\vartheta}_{k+1}^{\mathsf{LMC}} = \boldsymbol{\vartheta}_k^{\mathsf{LMC}} - h\nabla f(\boldsymbol{\vartheta}_k^{\mathsf{LMC}}) + \sqrt{2}\,(\boldsymbol{W}_{(k+1)h} - \boldsymbol{W}_{kh}).$$

This method is computationally efficient and has been widely used in statistics and machine learning for sampling from high-dimensional distributions (Gal & Ghahramani, 2016; Izmailov et al., 2020; 2021). To assess the discretization error, consider the case where $\boldsymbol{L}_0^{\mathsf{LD}}$ is drawn from the invariant distribution π and note that

$$\begin{aligned}\boldsymbol{L}_{(k+1)h}^{\mathsf{LD}} - \boldsymbol{\vartheta}_{k+1}^{\mathsf{LMC}} &= \boldsymbol{L}_{kh}^{\mathsf{LD}} - \boldsymbol{\vartheta}_k^{\mathsf{LMC}} - \int_0^h \nabla f(\boldsymbol{L}_{kh+s}^{\mathsf{LD}})\,\mathrm{d}s + h\nabla f(\boldsymbol{\vartheta}_k^{\mathsf{LMC}}) \\ &= \boldsymbol{L}_{kh}^{\mathsf{LD}} - \boldsymbol{\vartheta}_k^{\mathsf{LMC}} - h\big(\nabla f(\boldsymbol{L}_{kh}^{\mathsf{LD}}) - \nabla f(\boldsymbol{\vartheta}_k^{\mathsf{LMC}})\big) - \boldsymbol{\zeta}_k,\end{aligned} \tag{2}$$

where $\boldsymbol{\zeta}_k = \int_0^h \big(\nabla f(\boldsymbol{L}_{kh+s}^{\mathsf{LD}}) - \nabla f(\boldsymbol{L}_{kh}^{\mathsf{LD}})\big)\,\mathrm{d}s$ is a zero-mean random "noise" vector. Previous work on LMC demonstrated that the squared $\mathbb{L}_2$ norm of $\boldsymbol{\zeta}_k$ is of order M^2h^3p, whereas the term $\boldsymbol{L}_{kh}^{\mathsf{LD}} - \boldsymbol{\vartheta}_k^{\mathsf{LMC}} - h\big(\nabla f(\boldsymbol{L}_{kh}^{\mathsf{LD}}) - \nabla f(\boldsymbol{\vartheta}_k^{\mathsf{LMC}})\big)$ satisfies the contraction inequality

$$\big\|\boldsymbol{L}_{kh}^{\mathsf{LD}} - \boldsymbol{\vartheta}_k^{\mathsf{LMC}} - h\big(\nabla f(\boldsymbol{L}_{kh}^{\mathsf{LD}}) - \nabla f(\boldsymbol{\vartheta}_k^{\mathsf{LMC}})\big)\big\|_{\mathbb{L}_2}^2 \leqslant (1-mh)^2\|\boldsymbol{L}_{kh}^{\mathsf{LD}} - \boldsymbol{\vartheta}_k^{\mathsf{LMC}}\|_{\mathbb{L}_2}^2. \tag{3}$$

If we denote by r_k the correlation between $\boldsymbol{\zeta}_k$ and $\boldsymbol{L}_{kh}^{\mathsf{LD}} - \boldsymbol{\vartheta}_k^{\mathsf{LMC}}$, and by Err_k the error $\|\boldsymbol{L}_{kh}^{\mathsf{LD}} - \boldsymbol{\vartheta}_k^{\mathsf{LMC}}\|_{\mathbb{L}_2}$, we infer from equation 2 and equation 3 that

$$\mathrm{Err}_{k+1}^2 \leqslant (1-mh)^2\mathrm{Err}_k^2 + CMhr_k\mathrm{Err}_k\sqrt{hp} + CM^2h^3p,$$

for some universal constant C. If we were able to check that r_k is small enough so that the second term of the right-hand side can be neglected, we would get $\mathrm{Err}_{k+1}^2 \leqslant (1-mh)^2\mathrm{Err}_k^2 + C_1M^2h^3p$, which would eventually lead to $\mathrm{Err}_{k+1}^2 \leqslant (1-mh)^{2k}\mathrm{Err}_1^2 + C_2M^2h^2(p/m)$. This would amount to

$$|r_k| \ll 1 \qquad \Longrightarrow \qquad \mathrm{Err}_{k+1} \leqslant (1-mh)^k\,\mathrm{Err}_1 + C_3Mh\sqrt{p/m}. \tag{4}$$

Unfortunately, without any additional conditions on f, the correlation r_k cannot be shown to be small, and one can only deduce from equation 3 that $\text{Err}_{k+1} \leqslant (1-mh)\text{Err}_k + CMh\sqrt{ph}$, which eventually yields

$$|r_k| \not\ll 1 \qquad \Longrightarrow \qquad \text{Err}_{k+1} \leqslant (1-mh)^k\,\text{Err}_1 + C_4(M/m)\sqrt{ph}. \tag{5}$$

This inequality is established under the standard assumption $Mh \leqslant 1$, which implies that the last term in equation 4 is significantly smaller than equation 5. To get such an error deflation, we need the correlations r_k to be small. While this is not guaranteed for the Euler discretization, we will see that the randomized midpoint method allows us to achieve such a reduction.

Let U be a random variable uniformly distributed in $[0,1]$ and independent of the Brownian motion $\boldsymbol{W}$. The randomized midpoint method exploits the approximation

$$\boldsymbol{L}^{\mathsf{LD}}_{t+h} = \boldsymbol{L}^{\mathsf{LD}}_{t} - \int_0^h \nabla f(\boldsymbol{L}^{\mathsf{LD}}_{t+s})\,\mathrm{d}s + \sqrt{2}\,\Delta_h \boldsymbol{W}_t \approx \boldsymbol{L}^{\mathsf{LD}}_{t} - h\nabla f(\boldsymbol{L}^{\mathsf{LD}}_{t+hU}) + \sqrt{2}\,\Delta_h \boldsymbol{W}_t.$$

The noise counterpart of $\boldsymbol{\zeta}_k$ in this case is $\boldsymbol{\zeta}^{\mathsf{R}}_k = \int_0^h \nabla f(\boldsymbol{L}^{\mathsf{LD}}_{kh+s})\,\mathrm{d}s - \nabla f(\boldsymbol{L}^{\mathsf{LD}}_{kh+Uh})$. It is clearly centered and uncorrelated with all the random vectors independent of U such as $\boldsymbol{L}^{\mathsf{LD}}_{kh}$, $\boldsymbol{\vartheta}^{\mathsf{LMC}}_k$ and the gradient of f evaluated at these points.

The explanation above provides the intuition of the randomized midpoint method, and a hint to why it is preferable to the Euler discretization, but it cannot be taken as a formal definition of the method. The formal definition of the randomized midpoint method for the Langevin Monte Carlo (RLMC) is defined as follows: at each iteration $k = 1, 2, \ldots,$

1. we randomly, and independently of all the variables generated during the previous steps, generate a pair of random vectors $(\boldsymbol{\xi}'_k, \boldsymbol{\xi}''_k)$ and a random variable U_k such that
   - U_k is uniformly distributed in $[0,1]$ and independent of $(\boldsymbol{\xi}'_k, \boldsymbol{\xi}''_k)$,
   - $(\boldsymbol{\xi}'_k, \boldsymbol{\xi}''_k)$ are independent $\mathcal{N}_p(0, \mathbf{I}_p)$.
2. we set $\boldsymbol{\xi}_k = \sqrt{U_k}\,\boldsymbol{\xi}'_k + \sqrt{1-U_k}\boldsymbol{\xi}''_k$ and define the $(k+1)$th iterate $\boldsymbol{\vartheta}^{\mathsf{RLMC}}$ by

$$\boldsymbol{\vartheta}^{\mathsf{RLMC}}_{k+U} = \boldsymbol{\vartheta}^{\mathsf{RLMC}}_{k} - hU_k\nabla f(\boldsymbol{\vartheta}^{\mathsf{RLMC}}_{k}) + \sqrt{2hU_k}\,\boldsymbol{\xi}'_k \tag{6}$$

$$\boldsymbol{\vartheta}^{\mathsf{RLMC}}_{k+1} = \boldsymbol{\vartheta}^{\mathsf{RLMC}}_{k} - h\nabla f(\boldsymbol{\vartheta}^{\mathsf{RLMC}}_{k+U}) + \sqrt{2h}\,\boldsymbol{\xi}_k. \tag{7}$$

With a small step-size h and a large number of iterations n, the distribution of $\boldsymbol{\vartheta}^{\mathsf{RLMC}}_n$ can closely approximate the target distribution π. In a smooth and strongly convex setting, it is even possible to obtain a reliable estimate of the sampling error, as demonstrated in the following theorem (the the proof is included in the supplementary material).

If the step-size h is small and the number of iterations n is large, the distribution of $\boldsymbol{\vartheta}^{\mathsf{RLMC}}_n$ is close to the target π. Interestingly, in the smooth and strongly convex setting it is possible to get a good evaluation of the error of sampling as shown in the next theorem (the proof is deferred to the supplementary material).

Theorem 1. *Assume the function $f:\mathbb{R}^p \to \mathbb{R}$ satisfies Assumption 1. Let h be such that $Mh + \sqrt{\kappa}\,(Mh)^{3/2} \leqslant 1/4$ with $\kappa = M/m$. Then, every $n \geqslant 1$, the distribution ν^{RLMC}_n of $\boldsymbol{\vartheta}^{\mathsf{RLMC}}_n$ satisfies*

$$\mathsf{W}_2(\nu^{\mathsf{RLMC}}_n, \pi) \leqslant 1.11 e^{-mnh/2}\mathsf{W}_2(\nu_0, \pi) + \big(2.4\sqrt{\kappa Mh} + 1.77\big)Mh\sqrt{p/m}. \tag{8}$$

Prior to discussing the relation of the above error estimate to those available in the literature, let us state a consequence of it.

Corollary 1. *Let $\varepsilon \in (0,1)$ be a small number. If we choose $h > 0$ and $n \in \mathbb{N}$ so that*

$$Mh = \frac{\varepsilon}{1.5 + (6.5\kappa\varepsilon)^{1/3}}, \quad \textit{and} \quad n \geqslant \bigg(\frac{3\kappa}{\varepsilon} + \frac{3.8\kappa^{4/3}}{\varepsilon^{2/3}}\bigg)\bigg(\log(20/\varepsilon) + \frac{1}{2}\log\Big(\frac{m}{p}\mathsf{W}_2^2(\nu_0,\pi)\Big)\bigg)$$

then[1] *we have $\mathsf{W}_2(\nu^{\mathsf{RLMC}}_n, \pi) \leqslant \varepsilon\sqrt{p/m}$.*

---

[1] This follows from the fact that $(6\kappa/\varepsilon) + 4.2\kappa^{4/3}/\varepsilon^{2/3} \leqslant 2\kappa/(Mh) = 2/mh$.

Our results can be compared to the best available results for the Langevin Monte Carlo (LMC) under Assumption 1 (Durmus et al., 2019, Eq. 22). We recall that LMC is defined by a recursive relation of the same form as equation 7, with the only difference that $\nabla f(\boldsymbol{\vartheta}_{k+U})$ is replaced by $\nabla f(\boldsymbol{\vartheta}_k)$. The tightest known bound for LMC is given by

$$\mathsf{W}_2(\nu_n^{\mathsf{LMC}}, \pi) \leqslant (1 - mh)^{-n/2}\mathsf{W}_2(\nu_0, \pi) + \sqrt{2Mhp/m},$$

with $Mh \leqslant 1$. By choosing $2Mh = (19/20)^2\varepsilon^2$ and

$$n \geqslant 2.22(\kappa/\varepsilon^2)\big\{\log(20/\varepsilon) + \tfrac{1}{2}\log\big(\tfrac{m}{p}\mathsf{W}_2^2(\nu_0,\pi)\big)\big\},$$

we can ensure that $\mathsf{W}_2(\nu_n^{\mathsf{LMC}}, \pi) \leqslant \varepsilon\sqrt{p/m}$. Therefore, the complexity bound of Corollary 1 derived from our result for RLMC is better than the best-known complexity bound for LMC in the regime of κ of smaller order than[2] ε^{-4}.

To the best of our knowledge, the first results on the error analysis of RLMC have been obtained in (He et al., 2020). They derived an upper bound on the discretization error (the second term on the right-hand side of equation 8) under the assumption that the initial point of the algorithm is the minimizer of the potential function f. Their bound takes the form $C(\sqrt{\kappa Mh}+1)Mh\sqrt{p/m}\times\sqrt{mnh}$, where C is a universal but unspecified constant. Compared to our bound, the one obtained in He et al. (2020) has an additional factor $\sqrt{mnh}$. While this factor may not be very harmful in the case of geometric ergodicity where the number of iterations n is chosen such that nmh goes to infinity at the logarithmic rate $\log(1/\varepsilon)$, removing it can be an important step toward extending these results to potentials that are not strongly convex.

While the proof of this theorem is deferred to the supplementary material, we can outline the main argument that allowed us to remove the factor $\sqrt{nmh}$ from the error bound. To convey the main idea, let us consider three positive sequences a_n, b_n, c_n satisfying, for every $n \in \mathbb{N}$,

$$a_{n+1} \leqslant (1-\alpha)a_n + b_n \tag{9}$$

$$c_{n+1} \leqslant c_n - b_n + \mathsf{C}, \tag{10}$$

with some $\alpha \in (0, 1)$ and $\mathsf{C} > 0$. Using the standard telescoping sums argument, frequently employed for proving the convergence of convex optimization algorithms, one can infer from equation 10 that

$$\sum\nolimits_{k=0}^{n} b_n \leqslant c_0 - c_{n+1} + n\mathsf{C} \leqslant c_0 + n\mathsf{C}. \tag{11}$$

On the other hand, it follows from equation 9 that

$$a_{n+1} \leqslant (1-\alpha)^{n+1}a_0 + \sum\nolimits_{k=0}^{n}(1-\alpha)^{n-k}b_k. \tag{12}$$

Upper bounding $(1-\alpha)^{n-k}$ by one, and using equation 11, we arrive at

$$a_{n+1} \leqslant (1-\alpha)^{n+1}a_0 + c_0 + n\mathsf{C}. \tag{13}$$

This type of argument, used in previous papers on RKLMC (Shen & Lee, 2019), is sub-optimal and leads to the extra factor $\sqrt{nmh}$. A tighter bound can be obtained by replacing the telescoping sum argument by the summation by parts. More precisely, one can check that equation 10 and equation 12 yield

$$\begin{aligned} a_{n+1} &\leqslant (1-\alpha)^{n+1}a_0 + \sum\nolimits_{k=0}^{n}(1-\alpha)^{n-k}(c_k - c_{k+1}) + \mathsf{C}\sum\nolimits_{k=0}^{n}(1-\alpha)^{n-k} \\ &\leqslant (1-\alpha)^{n+1}a_0 + (1-\alpha)^n c_0 + \alpha\sum\nolimits_{k=0}^{n}(1-\alpha)^{n-k}c_k + \frac{\mathsf{C}}{\alpha}. \end{aligned} \tag{14}$$

The upper bound provided by equation 14 has two advantages as compared to equation 13: the term $n\mathsf{C}$ is replaced by C/α, which is generally smaller, and the dependence on the initial value is $(1-\alpha)^n c_0$ instead of c_0. This comes also with a challenge consisting in upper bounding the sum present in the right-hand side of equation 14, which we managed to overcome using the strong convexity (or, more precisely, the Polyak-Lojasiewicz condition). The full details are deferred to the supplementary material.

---

[2] The condition $\kappa = o(\varepsilon^{-4})$ is obtained by simple algebra from the condition $\kappa^{4/3}/\varepsilon^{2/3} = o(\kappa/\varepsilon^2)$.

3 RANDOMIZED MIDPOINT METHOD FOR THE KINETIC LANGEVIN DIFFUSION

The randomized midpoint method, introduced and studied in Shen & Lee (2019), aims at providing a discretization of the kinetic Langevin process that reduces the bias of sampling as compared to more conventional discretizations. Recall that the kinetic Langevin process $\boldsymbol{L}^{\mathsf{KLD}}$ is a solution to a second-order stochastic differential equation that can be informally written as

$$\tfrac{1}{\gamma}\ddot{\boldsymbol{L}}_t^{\mathsf{KLD}} + \dot{\boldsymbol{L}}_t^{\mathsf{KLD}} = -\nabla f(\boldsymbol{L}_t^{\mathsf{KLD}}) + \sqrt{2}\,\dot{\boldsymbol{W}}_t, \tag{15}$$

with initial conditions $\boldsymbol{L}_0^{\mathsf{KLD}} = \boldsymbol{\vartheta}_0$ and $\dot{\boldsymbol{L}}_0^{\mathsf{KLD}} = \boldsymbol{v}_0$. In equation 15, $\gamma > 0$, $\boldsymbol{W}$ is a standard p-dimensional Brownian motion and dots are used to designate derivatives with respect to time $t \geqslant 0$. This can be formalized using Itô's calculus and introducing the velocity field $\boldsymbol{V}^{\mathsf{KLD}}$ so that the joint process $(\boldsymbol{L}^{\mathsf{KLD}}, \boldsymbol{V}^{\mathsf{KLD}})$ satisfies

$$\mathrm{d}\boldsymbol{L}_t^{\mathsf{KLD}} = \boldsymbol{V}_t^{\mathsf{KLD}}\,\mathrm{d}t; \quad \tfrac{1}{\gamma}\mathrm{d}\boldsymbol{V}_t^{\mathsf{KLD}} = -\big(\boldsymbol{V}_t^{\mathsf{KLD}} + \nabla f(\boldsymbol{L}_t^{\mathsf{KLD}})\big)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{W}_t. \tag{16}$$

Similar to the vanilla Langevin diffusion equation 1, the kinetic Langevin diffusion $(\boldsymbol{L}^{\mathsf{KLD}}, \boldsymbol{V}^{\mathsf{KLD}})$ is a Markov process that exhibits ergodic properties when the potential f is strongly convex (see (Eberle et al., 2019) and references therein). The invariant density of this process is given by

$$p_*(\boldsymbol{\theta},\boldsymbol{v}) \propto \exp\{-f(\boldsymbol{\theta}) - \tfrac{1}{2\gamma}\|\boldsymbol{v}\|^2\}, \qquad \text{for all} \quad \boldsymbol{\theta}, \boldsymbol{v} \in \mathbb{R}^p.$$

Note that the marginal of p_* corresponds to $\boldsymbol{\theta}$ coincides with the target density π. However, unlike the vanilla Langevin diffusion, the kinetic Langevin is not reversible. It is interesting to note that the distribution of the process $\boldsymbol{L}^{\mathsf{KLD}}$ approaches that of the vanilla Langevin process as γ approaches infinity (see e.g. (Nelson, 1967)). Therefore, $\boldsymbol{L}^{\mathsf{LD}}$ and $\boldsymbol{L}^{\mathsf{KLD}}$ are often referred to as the overdamped and underdamped Langevin processes, respectively (where increasing the friction parameter γ is characterized as damping).

The kinetic Langevin diffusion $\boldsymbol{L}^{\mathsf{KLD}}$ is particularly attractive for sampling because its distribution ν_t^{KLD} converges to the invariant distribution exponentially fast. This is especially true for strongly convex potentials, as proven in[3] (Dalalyan & Riou-Durand, 2020, Prop. 1), where it is shown that the following inequality holds:

$$\mathsf{W}_2\left(\mathbf{C}\begin{bmatrix}\boldsymbol{V}_t^{\mathsf{KLD}}\\ \boldsymbol{L}_t^{\mathsf{KLD}}\end{bmatrix}, \mathbf{C}\begin{bmatrix}\boldsymbol{v}\\ \boldsymbol{\vartheta}\end{bmatrix}\right) \leqslant e^{-mt}\mathsf{W}_2\left(\mathbf{C}\begin{bmatrix}\boldsymbol{V}_0\\ \boldsymbol{L}_0\end{bmatrix}, \mathbf{C}\begin{bmatrix}\boldsymbol{v}\\ \boldsymbol{\vartheta}\end{bmatrix}\right), \quad \mathbf{C} = \begin{bmatrix}\mathbf{I}_p & \mathbf{0}_p\\ \mathbf{I}_p & \gamma\mathbf{I}_p\end{bmatrix}$$

for every $t \geqslant 0$, provided that $\gamma \geqslant m + M$.

To discretize this continuous-time process and make it applicable to the sampling problem, Shen & Lee (2019) proposed the following procedure: at each iteration $k = 1, 2, \ldots$,

1. randomly, and independently of all the variables generated at the previous steps, generate random vectors $(\boldsymbol{\xi}_k', \boldsymbol{\xi}_k'', \boldsymbol{\xi}_k''')$ and a random variable U_k such that
   - U_k is uniformly distributed in $[0, 1]$,
   - conditionally to $U_k = u$, $(\boldsymbol{\xi}_k', \boldsymbol{\xi}_k'', \boldsymbol{\xi}_k''')$ has the same joint distribution as $\big(\boldsymbol{B}_u - e^{-\gamma h u}\boldsymbol{G}_u, \boldsymbol{B}_1 - e^{-\gamma h}\boldsymbol{G}_1, \gamma e^{-\gamma h}\boldsymbol{G}_1\big)$, where $\boldsymbol{B}$ is a p-dimensional Brownian motion and $\boldsymbol{G}_t = \int_0^t e^{\gamma h s}\,\mathrm{d}\boldsymbol{B}_s$.
2. set $\psi(x) = (1 - e^{-x})/x$ and define the $(k+1)$th iterate of $\boldsymbol{\vartheta}^{\mathsf{RKLMC}}$ by

$$\begin{aligned}
\boldsymbol{\vartheta}_{k+U} &= \boldsymbol{\vartheta}_k + Uh\psi(\gamma Uh)\boldsymbol{v}_k - Uh\big(1 - \psi(\gamma Uh)\big)\nabla f(\boldsymbol{\vartheta}_k) + \sqrt{2h}\,\boldsymbol{\xi}_k'\\
\boldsymbol{\vartheta}_{k+1} &= \boldsymbol{\vartheta}_k + h\psi(\gamma h)\boldsymbol{v}_k - \gamma h^2(1-U)\psi\big(\gamma h(1-U)\big)\nabla f(\boldsymbol{\vartheta}_{k+U}) + \sqrt{2h}\boldsymbol{\xi}_k''\\
\boldsymbol{v}_{k+1} &= e^{-\gamma h}\boldsymbol{v}_k - \gamma h e^{-\gamma h(1-U)}\nabla f(\boldsymbol{\vartheta}_{k+U}) + \sqrt{2h}\,\boldsymbol{\xi}_k'''.
\end{aligned}$$

Although the sequence $(\boldsymbol{v}_k^{\mathsf{RKLMC}}, \boldsymbol{\vartheta}_k^{\mathsf{RKLMC}})$ approximates $(\boldsymbol{V}_{kh}^{\mathsf{KLD}}, \boldsymbol{L}_{kh}^{\mathsf{KLD}})$, it is not immediately apparent. The supplementary material clarifies this point. We state now the main result of this paper, providing a simple upper bound for the error of the RKLMC algorithm.

---

[3] For the sake of the self-containedness of this paper, we reproduce the proof of this inequality in Proposition 1 deferred to the Appendix.

Theorem 2. *Assume the function $f : \mathbb{R}^p \to \mathbb{R}$ satisfies Assumption 1. Choose γ and h so that $\gamma \geqslant 5M$ and $\gamma h \leqslant 0.1\kappa^{-1/6}$, where $\kappa = M/m$. Assume that $\boldsymbol{\vartheta}_0$ is independent of $\boldsymbol{v}_0$ and that $\boldsymbol{v}_0 \sim \mathcal{N}_p(0, \gamma\mathbf{I}_p)$. Then, for any $n \geqslant 1$, the distribution ν_n^{RKLMC} of $\boldsymbol{\vartheta}_n^{\mathsf{RKLMC}}$ satisfies*

$$\mathsf{W}_2(\nu_n^{\mathsf{RKLMC}}, \pi) \leqslant 1.6\varrho^n \mathsf{W}_2(\nu_0, \pi) + 0.1\sqrt{\varrho^n \mathbb{E}[f(\boldsymbol{\vartheta}_0) - f(\boldsymbol{\theta}_*)]/m} + 0.2(\gamma h)^3\sqrt{\kappa p/m} + 10(\gamma h)^{3/2}\sqrt{p/m}\,,$$

where $\varrho = \exp(-mh)$, and $\boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta})$.

This result has several strengths and limitations, which are discussed below, after the corollary providing the number of required iterations to attain a predetermined level of accuracy.

Corollary 2. *Let $\varepsilon \in (0, 1)$ be a small constant. If $\gamma = 5M$, $\boldsymbol{\vartheta}_0 = \boldsymbol{\theta}_*$ and we choose $h > 0$ and $n \in \mathbb{N}$ so that*

$$\gamma h = \frac{\varepsilon^{2/3}}{5 + 0.6(\varepsilon^2\kappa)^{1/6}}, \quad \text{and} \quad n \geqslant \kappa\varepsilon^{-2/3}\big(25 + 3(\varepsilon^2\kappa)^{1/6}\big)\log(20/\varepsilon)\,,$$

then we have $\mathsf{W}_2(\nu_n^{\mathsf{RKLMC}}, \pi) \leqslant \varepsilon\sqrt{p/m}$.

The corollary presented above gives the best-known convergence rate for the number of gradient evaluations required to achieve a prescribed error level in the case of a gradient Lipschitz potential, without any additional assumptions on its structure or smoothness. This rate, $\kappa\varepsilon^{-2/3}(1 + (\varepsilon^2\kappa)^{1/6})$, was first discovered by Shen & Lee (2019) (see also (He et al., 2020)). By employing our proposed proof technique described in Section 2, the result in Theorem 2 gets rid of the factor nmh from the discretization error, which was present in the previous upper bounds of the sampling error. Furthermore, our bound contains only small and explicit constants. Finally, our result does not require the RKLMC algorithm to be initialized at the minimizer of the potential, which is important for extending the method to non-convex potentials.

On the downside, the condition $\gamma \geqslant 5M$ is stronger than the corresponding conditions used in prior work on the KLMC (without randomization). Indeed, these prior results generally require $\gamma \geqslant 2M$. Having a proof of Theorem 2 that reduces the factor 5 in $\gamma \geqslant 5M$ would lead to significant savings in running time.

4 IMPROVED ERROR BOUND FOR THE KINETIC LANGEVIN WITH EULER DISCRETIZATION

The proof techniques presented in the previous section can be used to derive an upper bound on the error of the kinetic Langevin Monte Carlo (KLMC) algorithm. KLMC is a discretized version of KLD equation 16, where the term $\nabla f(\boldsymbol{L}_t)$ is replaced by $\nabla f(\boldsymbol{L}_{kh})$ on each interval $[kh, (k+1)h)$. The resulting error bound, given in the following theorem, exhibits a better dependence on κ than previously established bounds.

Theorem 3. *Let $f : \mathbb{R}^p \to \mathbb{R}$ satisfy $m\mathbf{I}_p \preccurlyeq \nabla^2 f(\boldsymbol{\theta}) \preccurlyeq M\mathbf{I}_p$ for every $\boldsymbol{\theta} \in \mathbb{R}^p$. Choose γ and h so that $\gamma \geqslant 5M$ and $\sqrt{\kappa}\gamma h \leqslant 0.1$, where $\kappa = M/m$. Assume that $\boldsymbol{\vartheta}_0$ is independent of $\boldsymbol{v}_0$ and that $\boldsymbol{v}_0 \sim \mathcal{N}_p(0, \gamma\mathbf{I}_p)$. Then, for any $n \geqslant 1$, the distribution ν_n^{KLMC} of $\boldsymbol{\vartheta}_n^{\mathsf{KLMC}}$ satisfies*

$$\mathsf{W}_2(\nu_n^{\mathsf{KLMC}}, \pi) \leqslant 2\varrho^n \mathsf{W}_2(\nu_0, \pi) + 0.05\sqrt{\varrho^n \mathbb{E}[f(\boldsymbol{\vartheta}_0) - f(\boldsymbol{\theta}_*)]/m} + 0.9\gamma h\sqrt{\kappa p/m}\,,$$

where $\varrho = \exp(-mh)$, and $\boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta})$.

Bounds on the error of KLMC under convexity assumption, or other related conditions, can be found in recent papers (Cheng et al., 2018b; Dalalyan & Riou-Durand, 2020; Monmarché, 2021; Monmarché, 2023). Our result has the advantage of providing an upper bound with the best known dependence on the condition number κ and having relatively small numerical constants, as shown in the next corollary.

Corollary 3. *Let $\varepsilon \in (0, 0.1)$. If $\gamma = 5M$, $\boldsymbol{\vartheta}_0 = \boldsymbol{\theta}_*$ and we choose $h > 0$ and $n \in \mathbb{N}$ so that*

$$\gamma h = \varepsilon\kappa^{-1/2}, \quad \text{and} \quad n \geqslant 5\kappa^{3/2}\varepsilon^{-1}\log(20/\varepsilon)$$

then we have $\mathsf{W}_2(\nu_n^{\mathsf{KLMC}}, \pi) \leqslant \varepsilon\sqrt{p/m}$.

It is worth noting that our error bounds, along with the other bounds mentioned previously under strong convexity, rely on the synchronous coupling between the KLMC and the KLD. However, in the case of the vanilla Langevin, it has been shown in Durmus et al. (2019) that the dependence of the error bound on κ can be improved by considering other couplings (in their case, the coupling is hidden in the analytical arguments). We conjecture that the dependence on κ in the kinetic Langevin Monte Carlo algorithm can also be improved through non-synchronous coupling. Specifically, we conjecture that the number of iterations required to achieve a W_2-error bounded by $\varepsilon\sqrt{p/m}$ should scale as κ/ε rather than $\kappa^{3/2}/\varepsilon$, as obtained in previous work and in Theorem 3.

5 NUMERICAL EXPERIMENTS

In this section, we compare the performance of LMC, KLMC, RLMC, and RKLMC algorithms. We apply the four algorithms to the posterior density of penalized logistic regression, defined by $\pi(\boldsymbol{\vartheta}) \propto \exp(-f(\boldsymbol{\vartheta}))$, with the potential function

$$f(\boldsymbol{\vartheta}) = \frac{\lambda}{2}\,\|\boldsymbol{\vartheta}\|^2 + \frac{1}{n_{\text{data}}}\sum_{i=1}^{n_{\text{data}}} \log(1+\exp(-y_i\boldsymbol{x}_i^\top\boldsymbol{\vartheta}))\,,$$

where $\lambda > 0$ denotes the tuning parameter. The data $\{\boldsymbol{x}_i, y_i\}_{i=1}^m$, composed of binary labels $y_i \in \{-1,1\}$ and features $\boldsymbol{x}_i \in \mathbb{R}^p$ generated from $x_{i,j} \overset{iid}{\sim} \mathcal{N}(0,1), \mathcal{N}(0,5)$, and $\mathcal{N}(0,10)$, corresponding to the plots from left to right, respectively. In our experiments, we have chosen $\lambda = 1/100$, $p = 3$ and $n_{\text{data}} = 100$.

Figure 1 shows the W_2-distance measured along the first dimension between the empirical distributions of the samples from the four algorithms and the target distribution[4], with different choices of h. These numerical results confirm our theoretical results. Indeed, we see that the randomized midpoint versions of LMC and KLMC perform better than their vanilla counterparts when the condition number is not too large (the leftmost plot). This order changes when κ becomes large, as we see in the rightmost plot, where KLMC outperforms the other algorithms.

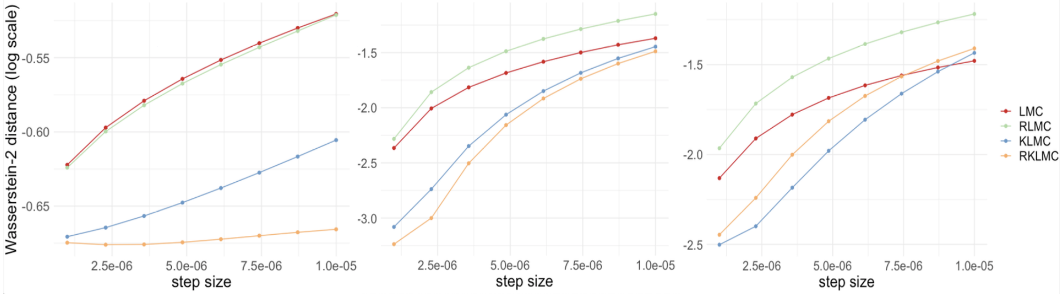

Figure 1: Error of {L,RL,KL,RKL}MC with different choice of step size.

6 DISCUSSION OF ASSUMPTIONS AND LIMITATIONS

The results presented in this paper provide easily computable guarantees for performing sampling with assured accuracy. These guarantees are conservative, implying that the actual sampling error may be smaller than ε even if the upper bounds stated in our theorems are larger than ε. However, these bounds represent the most reliable technique available in the existing literature. The importance of having such guarantees is further emphasized by the lack of reliable practical measures to assess the quality of sampling methods. To better understand the computational complexity implied by our bounds for various Monte Carlo algorithms, we present in Table 1 the number of gradient evaluations required to achieve the accuracy of $\varepsilon\sqrt{p/m}$ for different combinations of (ε,κ).

Strong convexity The assumption of strong convexity is often seen as too restrictive. In our theorems, strong convexity is used for three purposes: (a) to ensure the contraction of the continuous-time

---

[4] Here, we execute the LMC algorithm with a small step size over an extended duration to approximate the true distribution.

Table 1: The number of iterations that are sufficient for the algorithms {L,RL,KL,RKL}MC to achieve an error in W_2 distance bounded by $\varepsilon\sqrt{p/m}$, provided that they are initialized at the minimum of the potential f.

(ε,κ)	$(0.1^1, 10^1)$	$(0.1^1, 10^3)$	$(0.1^1, 10^5)$	$(0.1^1, 10^7)$	$(0.1^1, 10^9)$	$(0.1^1, 10^{11})$
LMC	1.2×10^4	1.2×10^6	1.2×10^8	1.2×10^{10}	1.2×10^{12}	1.2×10^{14}
RLMC	3.6×10^3	1.1×10^6	4.5×10^8	2.0×10^{11}	9.3×10^{13}	4.3×10^{16}
KLMC	8.4×10^3	8.4×10^6	8.4×10^9	8.4×10^{12}	8.4×10^{15}	8.4×10^{18}
RKLMC	1.0×10^4	1.1×10^6	1.1×10^8	1.3×10^{10}	2.2×10^{12}	4.2×10^{14}
(ε,κ)	$(0.1^3, 10^1)$	$(0.1^3, 10^3)$	$(0.1^3, 10^5)$	$(0.1^3, 10^7)$	$(0.1^3, 10^9)$	$(0.1^3, 10^{11})$
LMC	2.2×10^8	2.2×10^{10}	2.2×10^{12}	2.2×10^{14}	2.2×10^{16}	2.2×10^{18}
RLMC	3.8×10^5	6.8×10^7	2.0×10^{10}	8.4×10^{12}	3.8×10^{15}	1.7×10^{18}
KLMC	1.6×10^6	1.6×10^9	1.6×10^{12}	1.6×10^{15}	1.6×10^{18}	1.6×10^{21}
RKLMC	4.5×10^5	4.5×10^7	4.5×10^9	4.5×10^{11}	4.7×10^{13}	5.7×10^{15}
(ε,κ)	$(0.1^5, 10^1)$	$(0.1^5, 10^3)$	$(0.1^5, 10^5)$	$(0.1^5, 10^7)$	$(0.1^5, 10^9)$	$(0.1^5, 10^{11})$
LMC	3.2×10^{12}	3.2×10^{14}	3.2×10^{16}	3.2×10^{18}	3.2×10^{20}	3.2×10^{22}
RLMC	4.6×10^7	5.5×10^9	9.9×10^{11}	3.0×10^{14}	1.2×10^{17}	5.5×10^{19}
KLMC	2.3×10^8	2.3×10^{11}	2.3×10^{14}	2.3×10^{17}	2.3×10^{20}	2.3×10^{23}
RKLMC	1.5×10^7	1.5×10^9	1.5×10^{11}	1.5×10^{13}	1.5×10^{15}	1.5×10^{17}

Langevin dynamics, (b) to relate the potential's values to its gradient through the Polyak-Lojasiewicz condition $\|\nabla f(\boldsymbol{\theta})\|^2 \geqslant 2m(f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}_*))$ (Polyak, 1963; Łojasiewicz, 1963), and (c) to provide the following simple upper bound on the 2-Wasserstein distance $\mathsf{W}_2(\delta_{\boldsymbol{\theta}_*}, \pi) \leqslant \sqrt{p/m}$ (Durmus & Moulines, 2019, Prop. 1). The last two inequalities can be satisfied for many non-convex functions, but the same is not true for the contraction of the Langevin dynamics.

Alternatively, we can assume that the function is only strongly convex outside a ball of radius $R > 0$, whereas within the ball it is smooth but otherwise arbitrary. This approach requires an additional factor of order e^{MR^2} in the number of iterations necessary to achieve a specified error level (Cheng et al., 2018a; Ma et al., 2019). We can also assume that the Markov semi-group has a spectral gap and use this gap in the risk bounds. However, this approach goes against the spirit of our paper, which aims to provide guarantees that are easy to interpret and verify.

Another important point to note is that the results obtained under the assumption of strong convexity can be used as ready-made results in other frameworks as well. For instance, this is applicable to weakly convex potentials or potentials supported on a compact set (Dalalyan et al., 2022; Dwivedi et al., 2018; Brosse et al., 2017).

Smoothness Smoothness of f is a critical assumption for the results obtained in this paper. However, in statistical applications, this assumption may not hold, such as when using a Laplace prior. In such cases, various approaches have been proposed, mainly involving gradient approximation techniques, as explored in the literature (Durmus et al., 2018; Chatterji et al., 2020). Our results open the door for similar extensions of the randomized midpoint method for such scenarios.

It should also be stressed that if the potential is more than twice differentiable with a bounded tensor of higher-order derivatives, then it is possible to design Monte Carlo algorithms that perform better than the LMC and the KLMC (Dalalyan & Karagulyan, 2019; Dalalyan & Riou-Durand, 2020; Ma et al., 2021). The same is true if the function f has some specific structure (Mou et al., 2021).

Functional inequalities Functional inequalities such as the Poincaré and the log-Sobolev inequalities provide a convenient framework for analyzing sampling methods derived from continuous-time Markov processes. This line of research was developed in a series of papers (Chewi et al., 2020; Vempala & Wibisono, 2019; Chewi et al., 2022). Extending the techniques of this paper from strong log-concavity to the framework of distributions satisfying one of the aforementioned functional inequalities is a non-trivial task.

Other distances The Wasserstein-2 distance, utilized in this paper, serves as a natural metric for measuring the error in sampling due to its connection with optimal transport. However, it is worth noting that recent literature on gradient-based sampling has explored other metrics such as total variation distance, KL divergence, and χ^2 divergence (Ma et al., 2021; Vempala & Wibisono, 2019; Durmus et al., 2019; Chewi et al., 2020; Balasubramanian et al., 2022; Zhang et al., 2023). An interesting direction for future research involves establishing error guarantees for the randomized midpoint method with respect to these alternative distances.