SQRT(D) DIMENSION DEPENDENCE OF LANGEVIN MONTE CARLO

Anonymous authors
Paper under double-blind review

ABSTRACT

This article considers the popular MCMC method of unadjusted Langevin Monte Carlo (LMC) and provides a non-asymptotic analysis of its sampling error in 2-Wasserstein distance. The proof is based on a mean-square analysis framework refined from Li et al. (2019), which works for a large class of sampling algorithms based on discretizations of contractive SDEs. We establish an $\widetilde{\mathcal{O}}\left(\sqrt{d}/\epsilon\right)$ mixing time bound for LMC, without warm start, under the common log-smooth and log-strongly-convex conditions, plus a growth condition on the 3rd-order derivative of the potential of target measures. This bound improves the best previously known $\widetilde{\mathcal{O}}\left(d/\epsilon\right)$ result and is optimal (in terms of order) in both dimension d and accuracy tolerance ϵ for target measures satisfying the aforementioned assumptions. Our theoretical analysis is further validated by numerical experiments.

1 INTRODUCTION

The problem of sampling statistical distributions has attracted considerable attention, not only in the fields of statistics and scientific computing, but also in machine learning (Robert & Casella, 2013; Andrieu et al., 2003; Liu, 2008); for example, how various sampling algorithms scale with the dimension of the target distribution is a popular recent topic in statistical deep learning (see discussions below for references). For samplers that can be viewed as discretizations of SDEs, the idea is to use an ergodic SDE whose equilibrium distribution agrees with the target distribution, and employ an appropriate numerical algorithm that discretizes (the time of) the SDE. The iterates of the numerical algorithm will approximately follow the target distribution when converged, and can be used for various downstream applications such as Bayesian inference and inverse problem (Dashti & Stuart, 2017). One notable example is the Langevin Monte Carlo algorithm (LMC), which corresponds to Euler-Maruyama discretization of the overdamped Langevin equation. Its study dated back to at least the 90s (Roberts et al., 1996) but keeps on leading to important discoveries, for example, on non-asymptotics and dimension dependence, which are relevant to machine learning (e.g., Dalalyan (2017a;b); Cheng et al. (2018a); Durmus et al. (2019); Durmus & Moulines (2019); Vempala & Wibisono (2019); Dalalyan & Riou-Durand (2020); Li et al. (2019); Erdogdu & Hosseinzadeh (2021); Mou et al. (2019); Lehec (2021)). LMC is closely related to SGD too (e.g., Mandt et al. (2017)). Many other examples exist, based on alternative SDEs and/or different discretizations (e.g., Dalalyan & Riou-Durand (2020); Ma et al. (2021); Mou et al. (2021); Li et al. (2020); Roberts & Rosenthal (1998); Chewi et al. (2021); Shen & Lee (2019)).

Quantitatively characterizing the non-asymptotic sampling error of numerical algorithms is usually critical for choosing the appropriate algorithm for a specific downstream application, for providing practical guidance on hyperparameter selection and experiment design, and for designing improved samplers. A powerful tool that dates back to (Jordan et al., 1998) is a paradigm of non-asymptotic error analysis, namely to view sampling as optimization in probability space, and it led to many important recent results (e.g., Liu & Wang (2016); Dalalyan (2017a); Wibisono (2018); Zhang et al. (2018); Frogner & Poggio (2020); Chizat & Bach (2018); Chen et al. (2018); Ma et al. (2021); Erdogdu & Hosseinzadeh (2021)). It works by choosing an objective functional, typically some statistical distances/diverges, and showing that the law of the iterates of sampling algorithms converges in that objective functional. However, the choice of the objective functional often needs to be customized for different sampling algorithms. For example, KL divergence works for LMC (Cheng & Bartlett, 2018), but a carefully hand-crafted cross term needs to be added to KL divergence for analyzing

KLMC (Ma et al., 2021). Even for the same underlying SDE, different discretization schemes exist and lead to different sampling algorithms, and the analyses of them had usually been case by case (e.g., Cheng et al. (2018b); Dalalyan & Riou-Durand (2020); Shen & Lee (2019)). Therefore, it would be a desirable complement to have a unified, general framework to study the non-asymptotic error of SDE-based sampling algorithms. Toward this goal, an alternative approach to analysis has recently started attracting attention, namely to resort to the numerical analysis of SDE integrators (e.g., Milstein & Tretyakov (2013); Kloeden & Platen (1992)) and quantitatively connect the integration error to the sampling error. One remarkable work in this direction is Li et al. (2019), which will be discussed in greater details later on.

The main tool of analysis in this paper will be a strengthened version (in specific aspects that will be clarified soon) of the result in Li et al. (2019). Although this analysis framework is rather general and applicable to a broad family of numerical methods that discretize contractive[1] SDEs, the main innovation focuses on a specific sampling algorithm, namely LMC, which is widely used in practice. Its stochastic gradient version is implemented in common machine learning systems, such as Tensorflow (Abadi et al., 2016), and is the off-the-shelf algorithm for large scale Bayesian inference. With the ever-growing size of parameter space, the non-asymptotic error of LMC is of central theoretical and practical interest, in particular, its dependence on the dimension of the sample space. The best current known upper bound of the mixing time in 2-Wasserstein distance for LMC is $\widetilde{\mathcal{O}}\left(\frac{d}{\epsilon}\right)$ (Durmus & Moulines, 2019). Motivated by a recent result (Chewi et al., 2021) that shows better dimension dependence for a Metropolis-Adjusted improvement of LMC, we will investigate if the current bound for (unadjusted) LMC is tight, and if not, what is the optimal dimension dependence.

Our contribution

The main contribution of this work is an improved $\widetilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\epsilon}\right)$ mixing time upper bound for LMC in 2-Wasserstein distance, under reasonable regularity assumptions. More specifically, we study LMC for sampling from a Gibbs distribution $d\mu \propto \exp\left(-f(\boldsymbol{x})\right) d\boldsymbol{x}$. Under the standard smoothness and strong-convexity assumptions, plus an additional linear growth condition on the third-order derivative of the potential (which also shares connections to popular assumptions in the frontier literature), our bound improves upon the previously best known $\widetilde{\mathcal{O}}\left(\frac{d}{\epsilon}\right)$ result (Durmus & Moulines, 2019) in terms of dimension dependence. For a comparison, note it was known that discretized **kinetic** Langevin dynamics can lead to $\sqrt{d}$ dependence on dimension (Cheng & Bartlett, 2018; Dalalyan & Riou-Durand, 2020) and some believe that it is the introduction of momentum that improves the dimension dependence, but our result shows that discretized overdamped Langevin (no momentum) can also have mixing time scaling like $\sqrt{d}$. In fact, it is important to mention that recently shown was that Metropolis-Adjusted Euler-Maruyama discretization of **overdamped** Langevin (i.e., MALA) has an optimal dimension dependence of $\widetilde{\mathcal{O}}\left(\sqrt{d}\right)$ (Chewi et al., 2021), while what we analyze here is the **unadjusted** version (i.e., LMC), and it has the same dimension dependence (note however that our ϵ dependence is not as good as that for MALA; more discussion in Section 4). We also constructed an example which shows that the mixing time of LMC is at least $\widetilde{\Omega}\left(\frac{\sqrt{d}}{\epsilon}\right)$. Hence, our mixing time bound has the optimal dependence on both d and ϵ, in terms of order, for the family of target measures satisfying those regularity assumptions. Our theoretical analysis is further validated by empirical investigation of numerical examples.

A minor contribution of this work is the error analysis framework that we use. It is based on the classical mean-square analysis (Milstein & Tretyakov, 2013) in numerical SDE literature, however extended from finite time to infinite time. It is a minor contribution because this extension was already pioneered in the milestone work of Li et al. (2019), although we will develop a refined version. Same as in classical mean-square analysis and in Li et al. (2019), the final (sampling in this case) error is only half order lower than the order of local strong integration error (p_2). This will lead to a $\widetilde{\mathcal{O}}\left(C^{\frac{1}{p_2-\frac{1}{2}}}/\epsilon^{\frac{1}{p_2-\frac{1}{2}}}\right)$ mixing time upper bound in 2-Wasserstein distance for the family of algorithms, where C is a constant containing various information of the underlying problem, e.g., the dimension d. Nevertheless, the following two are **new** to this paper: (i) We weakened the requirement

[1] possibly after a coordinate transformation

on local strong and weak errors. More precisely, Li et al. (2019) requires uniform bounds on local errors, but this could be a nontrivial requirement for SDE integrators; the improvement here only requires non-uniform bounds (although establishing the same result consequently needs notably more efforts, these are included in this paper too). (ii) The detailed expressions of our bounds are not the same as those in Li et al. (2019) (even if local errors could be uniformly bounded), and as we are interested in dimension-dependence of LMC, we work out constants and carefully track their dimension-dependences. Bounds and constants in Li et al. (2019) might not be specifically designed for tightly tracking dimension dependences, as the focus of their seminal paper was more on ϵ dependence; consequently, its general error bound only led to a $\tilde{O}(d)$-dependence in mixing time when applied to LMC (see Example 1 in Li et al. (2019)), whereas our result leads to $\tilde{O}(\sqrt{d})$.

2 PRELIMINARIES

Notation Use the symbol $\boldsymbol{x}$ to denote a d-dimensional vector, and the plain symbol x to denote a scalar variable. Use $\|\boldsymbol{x}\|$ to denote the Euclidean norm of vector $\boldsymbol{x}$. A numerical algorithm is denoted by $\mathcal{A}$ and its k-th iterate is denoted by $\bar{\boldsymbol{x}}_k$. We slightly abuse notation by identifying measures with their density function w.r.t. Lebesgue measure. We use the convention $\widetilde{\mathcal{O}}(\cdot) = \mathcal{O}(\cdot)\log^{\mathcal{O}(1)}(\cdot)$, i.e., the $\widetilde{\mathcal{O}}(\cdot)$ notation ignores the dependence on logarithmic factors. We use the notation $\widetilde{\Omega}(\cdot)$ similarly. Denote 2-Wasserstein distance by $W_2(\mu_1,\mu_2) = \left(\inf_{(\boldsymbol{X},\boldsymbol{Y})\sim\Pi(\mu_1,\mu_2)} \mathbb{E}\,\|\boldsymbol{X}-\boldsymbol{Y}\|^2\right)^{\frac{1}{2}}$, where $\Pi(\mu_1,\mu_2)$ is the set of couplings, i.e. all joint measures with X and Y marginals being μ_1 and μ_2. Denote the target distribution by μ and the law of a random variable $\boldsymbol{X}$ by $\mathrm{Law}(\boldsymbol{X})$. Finally, denote the mixing time of an sampling algorithm $\mathcal{A}$ converging to its target distribution μ in 2-Wasserstein distance by $\tau_{\mathrm{mix}}(\epsilon; W_2; \mathcal{A}) = \inf\{k \geq 0 | W_2(\mathrm{Law}(\bar{\boldsymbol{x}}_k),\mu) \leq \epsilon\}$.

SDE for Sampling Consider a general SDE

$$d\boldsymbol{x}_t = \boldsymbol{b}(t,\boldsymbol{x}_t)dt + \boldsymbol{\sigma}(t,\boldsymbol{x}_t)d\boldsymbol{B}_t \tag{1}$$

where $\boldsymbol{b} \in \mathbb{R}^d$ is a drift term, $\boldsymbol{\sigma} \in \mathbb{R}^{d\times l}$ is a diffusion coefficient matrix and $\boldsymbol{B}_t$ is a l-dimensional Wiener process. Under mild condition (Pavliotis, 2014, Theorem 3.1), there exists a unique strong solution $\boldsymbol{x}_t$ to Eq. (1). Some SDEs admit geometric ergodicity, so that their solutions converge exponentially fast to a unique invariant distribution, and examples include the classical overdamped and kinetic Langevin dynamics, but are not limited to those (e.g., Mou et al. (2021); Li et al. (2020)). Such SDE are desired for sampling purposes, because one can set the target distribution to be the invariant distribution by choosing an SDE with an appropriate potential, and then solve the solution $\boldsymbol{x}_t$ of the SDE and push the time t to infinity, so that (approximate) samples of the target distribution can be obtained. Except for a few known cases, however, explicit solutions of Eq. (1) are elusive and we have to resort to numerical schemes to simulate/integrate SDE. Such example schemes include, but are not limited to Euler-Maruyama method, Milstein methods and Runge-Kutta method (e.g., Kloeden & Platen (1992); Milstein & Tretyakov (2013)). With constant stepsize h and at k-th iteration, a typical numerical algorithm takes a previous iterate $\bar{\boldsymbol{x}}_{k-1}$ and outputs a new iterate $\bar{\boldsymbol{x}}_k$ as an approximation of the solution $\boldsymbol{x}_t$ of Eq. (1) at time $t = kh$.

Langevin Monte Carlo Algorithm LMC algorithm is defined by the following update rule

$$\bar{\boldsymbol{x}}_k = \bar{\boldsymbol{x}}_{k-1} - h\nabla f(\bar{\boldsymbol{x}}_{k-1}) + \sqrt{2h}\boldsymbol{\xi}_k, \quad k = 1,2,\cdots \tag{2}$$

where $\{\boldsymbol{\xi}_k\}_{k\in\mathbb{Z}_{>0}}$ are i.i.d. standard d-dimensional Gaussian vectors. LMC corresponds to an Euler-Maruyama discretization of the continuous overdamped Langevin dynamics $d\boldsymbol{x}_t = -\nabla f(\boldsymbol{x}_t)dt + \sqrt{2}d\boldsymbol{B}_t$, which converges to an equilibrium distribution $\mu \sim \exp(-f(\boldsymbol{x}))$.

Dalalyan (2017b) provided a non-asymptotic analysis of LMC. An $\widetilde{\mathcal{O}}\left(\frac{d}{\epsilon^2}\right)$ mixing time bound in W_2 for log-smooth and log-strongly-convex target measures (Dalalyan, 2017a; Cheng et al., 2018a; Durmus et al., 2019) has been established. It was further improved to $\widetilde{\mathcal{O}}\left(\frac{d}{\epsilon}\right)$ under an additional Hessian Lipschitz condition (Durmus & Moulines, 2019). Mixing time bounds of LMC in other statistical distances/divergences have also been studied, including total variation distance (Dalalyan, 2017b; Durmus & Moulines, 2017) and KL divergence (Cheng & Bartlett, 2018).

Classical Mean-Square Analysis A powerful framework for quantifying the *global* discretization error of a numerical algorithm for Eq. (1), i.e., $e_k = \left\{\mathbb{E}\left\|\boldsymbol{x}_{kh} - \bar{\boldsymbol{x}}_k\right\|\right\}^{\frac{1}{2}}$, is mean-square analysis (e.g., Milstein & Tretyakov (2013)). Mean-square analysis studies how *local* integration error propagate and accumulate into global integration error; in particular, if one-step (local) weak error and strong error (both the exact and numerical solutions start from the same initial value $\boldsymbol{x}$) satisfy

$$\begin{aligned} \left\|\mathbb{E}\boldsymbol{x}_h - \mathbb{E}\bar{\boldsymbol{x}}_1\right\| \leq & C_1 \left(1 + \mathbb{E}\left\|\boldsymbol{x}\right\|^2\right)^{\frac{1}{2}} h^{p_1}, \quad \text{(local weak error)} \\ \left(\mathbb{E}\left\|\boldsymbol{x}_h - \bar{\boldsymbol{x}}_1\right\|^2\right)^{\frac{1}{2}} \leq & C_2 \left(1 + \mathbb{E}\left\|\boldsymbol{x}\right\|^2\right)^{\frac{1}{2}} h^{p_2}, \quad \text{(local strong error)} \end{aligned} \tag{3}$$

over a time interval $[0, Kh]$ for some constants $C_1, C_2 > 0$, $p_2 \geq \frac{1}{2}$ and $p_1 \geq p_2 + \frac{1}{2}$, then the global error is bounded by $e_k \leq C\left(1 + \mathbb{E}\left\|\boldsymbol{x}_0\right\|^2\right)^{\frac{1}{2}} h^{p_2 - \frac{1}{2}}$, $k = 1, \cdots, K$ for some constant $C > 0$ dependent on Kh.

Although classical mean-square analysis is only concerned with numerical integration error, sampling error can be also inferred. However, there is a limitation that prevents its direct employment in analyzing sampling algorithms: the global error bound only holds in finite time because the constant C can grow exponentially as K increases, rendering the bound useless when $K \to \infty$.

3 Mean-Square Analysis of Samplers Based on Contractive SDE

We now review and present some improved results on how to use mean-square analysis of **integration** error to quantify **sampling** error. A seminal paper in this direction is Li et al. (2019). What is known / new will be clarified. In all cases, the first step is to lift the finite time limitation when the SDE being discretized has some decaying property so that local integration errors do not amplify with time.

The specific type of decaying property we will work with is contractivity (after coordinate transformation). It is a sufficient condition for the underlying SDE to converge to a statistical distribution.

Definition 3.1. *A stochastic differential equation is contractive if there exists a non-singular constant matrix $A \in \mathbb{R}^{d\times d}$, a constant $\beta > 0$, such that any pair of solutions of the SDE satisfy*

$$\left(\mathbb{E}\left\|A\left(\boldsymbol{x}_t - \boldsymbol{y}_t\right)\right\|^2\right)^{\frac{1}{2}} \leq \left(\mathbb{E}\left\|A\left(\boldsymbol{x} - \boldsymbol{y}\right)\right\|^2\right)^{\frac{1}{2}} \exp(-\beta t), \tag{4}$$

where $\boldsymbol{x}_t, \boldsymbol{y}_t$ are two solutions, driven by the same Brownian motion but evolved respectively from initial conditions $\boldsymbol{x}$ and $\boldsymbol{y}$.

Remark. *As long as $\boldsymbol{b}$ and $\boldsymbol{\sigma}$ in* (1) *are not explicitly dependent on time, it suffices to find an arbitrarily small $t_0 > 0$ and show* (4) *holds for all $t < t_0$.*

Remark. *Overdamped Langevin dynamics, of which LMC is a discretization, is contractive when f is strongly convex and smooth.*

Remark. *Sometimes contraction is not easy to establish directly, but can be shown after an appropriate coordinate transformation, see (Dalalyan & Riou-Durand, 2020, Proposition 1) for such a treatment for kinetic Langevin dynamics. The introduction of A permits such transformations.*

We now use contractivity to remove the finite time limitation. We first need a short time lemma.

Lemma 3.2. *(Milstein & Tretyakov, 2013, Lemma 1.3) Suppose $\boldsymbol{b}$ and $\boldsymbol{\sigma}$ in Eq.*(1) *are Lipschitz continuous. For two solutions $\boldsymbol{x}_t, \boldsymbol{y}_t$ of Eq.* (1) *starting from $\boldsymbol{x}, \boldsymbol{y}$ respectively, denote $\boldsymbol{z} := (\boldsymbol{x}_t - \boldsymbol{x}) - (\boldsymbol{y}_t - \boldsymbol{y})$, then there exist $C_0 > 0$ and $h_0 > 0$ such that*

$$\mathbb{E}\left\|\boldsymbol{z}\right\|^2 \leq C_0 \left\|\boldsymbol{x} - \boldsymbol{y}\right\|^2 t, \quad \forall \boldsymbol{x}, \boldsymbol{y},\ 0 < t \leq h_0. \tag{5}$$

Then we have a sequence of results that connects statistical property with integration property. We will see that a non-asymptotic **sampling** error analysis only requires bounding the orders of local weak and strong **integration** errors (if the continuous dynamics can be shown contractive).

Theorem 3.3. ***(Global Integration Error, Infinite Time Version)*** *Suppose Eq.*(1) *is contractive with rate β and with respect to a non-singular matrix $A \in \mathbb{R}^{d\times d}$, with Lipschitz continuous $\boldsymbol{b}$ and $\boldsymbol{\sigma}$,*

and there is a numerical algorithm $\mathcal{A}$ with step size h simulating the solution $\boldsymbol{x}_t$ of the SDE, whose iterates are denoted by $\bar{\boldsymbol{x}}_k, k = 0, 1, \cdots$. Suppose there exists $0 < h_0 \leq 1, C_1, C_2 > 0, D_1, D_2 \geq 0, p_1 \geq 1, \frac{1}{2} < p_2 \leq p_1 - \frac{1}{2}$ such that for any $0 < h \leq h_0$, the algorithm $\mathcal{A}$ has, respectively, local weak and strong error of order p_1 and p_2, defined as

$$\begin{cases} \left\|\mathbb{E}\left(\boldsymbol{x}_h - \bar{\boldsymbol{x}}_1\right)\right\| \leq \left(C_1 + D_1\sqrt{\mathbb{E}\left\|\boldsymbol{x}\right\|^2}\right) h^{p_1}, \\ \left(\mathbb{E}\left\|\boldsymbol{x}_h - \bar{\boldsymbol{x}}_1\right\|^2\right)^{\frac{1}{2}} \leq \left(C_2^2 + D_2^2\mathbb{E}\left\|\boldsymbol{x}\right\|^2\right)^{\frac{1}{2}} h^{p_2}, \end{cases} \tag{6}$$

where $\boldsymbol{x}_h$ solves Eq.(1) with any initial value $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}_1$ is the result of applying $\mathcal{A}$ to $\boldsymbol{x}$ for one step. If the solution of SDE $\boldsymbol{x}_t$ and algorithm $\mathcal{A}$ both start from $\boldsymbol{x}_0$, then for $0 < h \leq h_1 \triangleq \min\left\{h_0, \frac{1}{4\beta}, \left(\frac{\sqrt{\beta}}{4\sqrt{2}\kappa_A D_2}\right)^{\frac{1}{p_2-\frac{1}{2}}}, \left(\frac{\beta}{8\sqrt{2}\kappa_A^2(D_1+C_0D_2)}\right)^{\frac{1}{p_2-\frac{1}{2}}}\right\}$, the global error $\boldsymbol{e}_k$ is bounded as

$$e_k := \left(\mathbb{E}\|\boldsymbol{x}_{kh} - \bar{\boldsymbol{x}}_k\|^2\right)^{\frac{1}{2}} \leq Ch^{p_2-\frac{1}{2}}, \quad k = 0, 1, 2, \cdots \tag{7}$$

where

$$C = \frac{2}{\sqrt{\beta}}\kappa_A^2\left(\frac{C_1 + C_0C_2 + \sqrt{2}U(D_1 + C_0D_2)}{\sqrt{\beta}} + C_2 + \sqrt{2}D_2U\right), \tag{8}$$

C_0 is from Eq. (5), κ_A is the condition number of matrix A and $U^2 \triangleq 4\mathbb{E}\left\|\boldsymbol{x}_0\right\|^2 + 6\mathbb{E}_\mu\left\|\boldsymbol{x}\right\|^2$.

Remark (what's new)**.** *Thm.3.3 refines the seminal results in Li et al. (2019) in the sense that it only requires non-uniform bounds on the local error* (6)*, whereas Li et al. (2019) requires uniform bounds, i.e., $D_1 = D_2 = 0$ in* (6)*. Therefore, the refinement we present has wider applicability.*

In general, local errors tend to depend on the current step's value, i.e. $D_1 \neq 0, D_2 \neq 0$. Allowing local error bounds to be non-uniform enabled applications such as proving the vanishing bias of mirror Langevin algorithm (Li et al., 2021). For a simpler illustration, consider LMC for 1D standard Gaussian target distribution, then we have

$$\left\|\mathbb{E}\left(\boldsymbol{x}_h - \bar{\boldsymbol{x}}_1\right)\right\| = (e^{-h} - 1 + h)\mathbb{E}\|\boldsymbol{x}\| = (\frac{h^2}{2} + o(h^2))\mathbb{E}\|\boldsymbol{x}\|$$

One can see that the local error does depend on $\boldsymbol{x}$ and is not uniform. Meanwhile, our non-uniform condition still holds because $(\frac{h^2}{2} + o(h^2))\mathbb{E}\|\boldsymbol{x}\| \leq (\frac{h^2}{2} + o(h^2))\sqrt{\mathbb{E}\|\boldsymbol{x}\|^2}$ (and thus $p_1 = 2$). Note if the discretization does converge to a neighborhood of the target distribution, it is possible that $\mathbb{E}\|\boldsymbol{x}\|^2$ and/or $\mathbb{E}\|\boldsymbol{x}\|$ become bounded near the convergence, and in this case the D_1, D_2 parts can be absorbed into C_1 and C_2; however, this 'if' clause is exactly what we'd like to prove.

Nevertheless, we state for rigor that the convention $1/0 = \infty$ is used when $D_1 = D_2 = 0$. Another remark is, even in this case, our bound has a different expression from the seminal results. We will carefully work out, track, and combine dimension-dependence of constants using our bound.

Following Theorem 3.3, we obtain the following non-asymptotic bound of the sampling error in W_2:

Theorem 3.4. ***(Non-Asymptotic Sampling Error Bound: General Case)*** *Under the same assumption and with the same notation of Theorem 3.3, we have*

$$W_2(Law(\bar{\boldsymbol{x}}_k), \mu) \leq e^{-\beta kh}W_2(Law(\boldsymbol{x}_0), \mu) + Ch^{p_2-\frac{1}{2}}, \quad \forall 0 < h \leq h_1.$$

A corollary of Theorem 3.4 is a bound on the mixing time of the sampling algorithm:

Corollary 3.5. ***(Upper Bound of Mixing Time: General Case)*** *Under the same assumption and with the same notation of Theorem 3.3, we have*

$$\tau_{\text{mix}}(\epsilon; W_2; \mathcal{A}) \leq \max\left\{\frac{1}{\beta h_1}, \frac{1}{\beta}\left(\frac{2C}{\epsilon}\right)^{\frac{1}{p_2-\frac{1}{2}}}\right\}\log\frac{2W_2(Law(\boldsymbol{x}_0)\mu)}{\epsilon}$$

In particular, when high accuracy is needed, i.e., $\epsilon < 2Ch_1^{p_2-\frac{1}{2}}$, we have

$$\tau_{\text{mix}}(\epsilon; W_2; \mathcal{A}) \leq \frac{(2C)^{\frac{1}{p_2-\frac{1}{2}}}}{\beta}\frac{1}{\epsilon^{\frac{1}{p_2-\frac{1}{2}}}}\log\frac{2W_2(Law(\boldsymbol{x}_0), \mu)}{\epsilon} = \widetilde{\mathcal{O}}\left(\frac{C^{\frac{1}{p_2-\frac{1}{2}}}}{\beta}\frac{1}{\epsilon^{\frac{1}{p_2-\frac{1}{2}}}}\right) \tag{9}$$

Corollary 3.5 states how mixing time depends on the order of local (strong) error (i.e., p_2) of a numerical algorithm. The larger p_2 is, the shorter the mixing time of the algorithm is, in term of the dependence on accuracy tolerance parameter ϵ. It is important to note that for constant stepsize discretizations that are deterministic on the filtration of the driving Brownian motion and use only its increments, there is a strong order barrier, namely $p_2 \leq 1.5$ (Rüemelin, 1982); however, methods involving multiple stochastic integrals (e.g., Kloeden & Platen (1992); Milstein & Tretyakov (2013)) and randomization (e.g., Shen & Lee (2019)) can yield a larger p_2.

The constant C defined in Eq. (7) typically contains rich information about the underlying SDE, e.g. dimension, Lipschitz constant of drift and noise diffusion, and the initial value $\boldsymbol{x}_0$ of the sampling algorithm. Through C, we can uncover the dependence of mixing time bound on various parameters, such as the dimension d. This will be detailed for Langevin Monte Carlo in the next section.

It is worth clarifying that once Thm.3.3 is proved, establishing Theorem 3.4 and Corollary 3.5 is relatively easy. In fact, analogous results have already been provided in Li et al. (2019), although they also required uniform local errors as consequences of their Thm.1. Nevertheless, we do not claim novelty in Theorem 3.4 and Corollary 3.5 and they are just presented for completeness. Our main refinement is just Thm.3.3 over Thm.1 in Li et al. (2019), and the non-triviality lies in its proof.

4 Non-Asymptotic Analysis of Langevin Monte Carlo Algorithm

We now quantify how LMC samples from Gibbs target distribution $\mu \sim \exp\big(-f(\boldsymbol{x})\big)$ that has a finite second moment, i.e., $\int_{\mathbb{R}^d} \|\boldsymbol{x}\|^2 \, d\mu < \infty$. Assume without loss of generality that the origin is a local minimizer of f, i.e. $\nabla f(\mathbf{0}) = \mathbf{0}$; this is for notational convenience in the analysis and can be realized via a simple coordinate shift, and it is not needed in the practical implementation. In addition, we assume the following two conditions hold:

A 1. ***(Smoothness and Strong Convexity)*** *Assume $f \in \mathcal{C}^2$ and is L-smooth and m-strongly-convex, i.e. there exists $0 < m \leq L$ such that $mI_d \preccurlyeq \nabla^2 f(\boldsymbol{x}) \preccurlyeq LI_d, \quad \forall \boldsymbol{x} \in \mathbb{R}^d$.*

Denote the condition number of f by $\kappa \triangleq \frac{L}{m}$. The smoothness and strong-convexity assumption is the standard assumption in the literature of analyzing LMC algorithm (Dalalyan, 2017a;b; Cheng & Bartlett, 2018; Durmus et al., 2019; Durmus & Moulines, 2019).

A 2. ***(Linear Growth of the 3rd-order Derivative)*** *Assume $f \in \mathcal{C}^3$ and the operator $\nabla(\Delta f)$ grows at most linearly, i.e., there exists a constant $G > 0$ such that $\big\|\nabla(\Delta f(\boldsymbol{x}))\big\| \leq G\left(1 + \|\boldsymbol{x}\|\right)$.*

Remark. *The linear growth (at infinity) condition on $\nabla\Delta f$ is actually not as restrictive as it appears, and in some sense even weaker than some classical condition for the existence of solutions to SDE. For example, a standard condition for ensuring the existence and uniqueness of a global solution to SDE is at most a linear growth (at infinity) of the drift (Pavliotis, 2014, Theorem 3.1). If we consider monomial potentials, i.e., $f(x) = x^p, p \in \mathbb{N}_+$, then the linear growth condition on $\nabla\Delta f$ is met when $p \leq 4$, whereas the classical condition for the existence of solutions holds only when $p \leq 2$.*

Remark. *Another additional assumption, namely Hessian Lipschitz condition, is commonly used in the literature (e.g., Durmus & Moulines (2019); Ma et al. (2021)). It requires the existence of a constant $\tilde{L}$, such that $\|\nabla^2 f(\boldsymbol{y}) - \nabla^2 f(\boldsymbol{x})\| \leq \tilde{L}\|\boldsymbol{y} - \boldsymbol{x}\|$. It can be shown that smoothness and Hessian Lipschitzness imply A2. Meanwhile, examples that satisfy A2 but are not Hessian Lipschitz exist, e.g., $f(x) = x^4$, and thus A2 is not necessarily stronger than Hessian Lipschitzness.*

Remark. *Same as L and m in A1, we implicitly assume the constant G introduced in A2 to be independent of dimension. Meanwhile, it is important to note examples for which G depends on the dimension do exist, and this is also true for other regularity constants including not only L and m but also the Hessian Lipschitz constant $\tilde{L}$. This part of the assumption is a strong one.*

To apply mean-square analysis to study LMC algorithm, we will need to ensure the underlying Langevin dynamics is contractive, which we verify in Section C and D in the appendix. In addition, we work out all required constants to determine the C in Eq. 7 explicitly in the appendix. With all these necessary ingredients, we now invoke Theorem 3.4 and obtain the following result:

Theorem 4.1. ***(Non-Asymptotic Error Bound: LMC)*** *Suppose Assumption 1 and 2 hold. LMC iteration $\bar{\boldsymbol{x}}_{k+1} = \bar{\boldsymbol{x}}_k - h\nabla f(\bar{\boldsymbol{x}}_k) + \sqrt{2h}\xi_k$ satisfies*

$$W_2(\mathit{Law}(\bar{\boldsymbol{x}}_k), \mu) \leq e^{-mkh} W_2(\mathit{Law}(\boldsymbol{x}_0), \mu) + C_{\mathit{LMC}} h, \quad 0 < h \leq \frac{1}{4\kappa L}, k \in \mathbb{N} \tag{10}$$

Table 1: Comparison of mixing time results in 2-Wassertein distance of LMC with L-smooth and m-strongly-convex potential. Constant step size is used and accuracy tolerance ϵ is small enough.

	mixing time	Additional Assumption
(Dalalyan, 2017a, Theorem 1)	$\widetilde{\mathcal{O}}\left(\frac{\kappa^2}{m}\cdot\frac{d}{\epsilon^2}\right)$	N/A
(Cheng & Bartlett, 2018, Theorem 1)	$\widetilde{\mathcal{O}}\left(\frac{\kappa^2}{m}\cdot\frac{d}{\epsilon^2}\right)$	N/A
(Durmus et al., 2019, Corollary 10)	$\widetilde{\mathcal{O}}\left(\frac{\kappa}{m}\cdot\frac{d}{\epsilon^2}\right)$	N/A
(Durmus & Moulines, 2019, Theorem 8)	$\widetilde{\mathcal{O}}\left(\frac{d}{\epsilon}\right)$ [1]	$\left\lVert\nabla^2 f(\boldsymbol{x})-\nabla^2 f(\boldsymbol{y})\right\rVert \le \widetilde{L}\left\lVert\boldsymbol{x}-\boldsymbol{y}\right\rVert$
This work (Theorem 4.2)	$\widetilde{\mathcal{O}}\left(\frac{\kappa^2}{\sqrt{m}}\cdot\frac{\sqrt{d}}{\epsilon}\right)$	Assumption 2 and $G=\mathcal{O}(L^2)$ [2]

where $C_{LMC}=\frac{10(L^2+G)}{m^{\frac{3}{2}}}\sqrt{2d+m\left(\mathbb{E}\left\lVert\boldsymbol{x}_0\right\rVert^2+1\right)}=\mathcal{O}(\sqrt{d})$.

Corollary 3.5 combined with the above result gives the following bound on the mixing time of LMC:

Theorem 4.2. ***(Upper Bound of Mixing Time: LMC)*** *Suppose Assumption 1 and 2 hold. If running LMC from* $\boldsymbol{x}_0$*, we then have*

$$\tau_{\text{mix}}(\epsilon;W_2;\text{LMC})\le\max\{4\kappa^2,\frac{2C_{LMC}}{m}\frac{1}{\epsilon}\}\log\frac{2W_2(Law(\boldsymbol{x}_0),\mu)}{\epsilon}$$

where C_{LMC} *is the same in Theorem 4.1. When high accuracy is needed, i.e.,* $\epsilon\le\frac{C_{LMC}}{2m\kappa^2}$*, we have*

$$\tau_{\text{mix}}(\epsilon;W_2;\text{LMC})\le\frac{2C_{LMC}}{m}\frac{1}{\epsilon}\log\frac{2W_2(Law(\boldsymbol{x}_0),\mu)}{\epsilon}=\widetilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\epsilon}\right).$$

The $\widetilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\epsilon}\right)$ mixing time bound in W_2 distance improves upon the previous ones (Dalalyan, 2017a; Cheng & Bartlett, 2018; Durmus & Moulines, 2019; Durmus et al., 2019) in the dependence of d and/or ϵ. If further assuming $G=\mathcal{O}(L^2)$, we then have $C_{\text{LMC}}=\mathcal{O}(\kappa^2\sqrt{m}\sqrt{d})$ and Thm.4.2 shows the mixing time is $\widetilde{\mathcal{O}}\left(\frac{\kappa^2}{\sqrt{m}}\frac{\sqrt{d}}{\epsilon}\right)$, which also improves the κ dependence in some previous results (Dalalyan, 2017a; Cheng & Bartlett, 2018) in the $m\le 1$ regime. A brief summary is in Table 1.

Remark (more comparison)**.** *The seminal work of Li et al. (2019) provided mean-square analysis (their Thm.1) and obtained a* $\widetilde{\mathcal{O}}\left(\frac{d}{\epsilon}\right)$ *mixing time bound for LMC (their Example 1) under smoothness, strong convexity and Hessian Lipschitz conditions, consistent with that in Durmus & Moulines (2019). By using our version (Thm.3.3) and tracking down constants' dimension-dependence, we are able to tighten it to* $\widetilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\epsilon}\right)$*. Worth clarifying is, dimension-dependence might not be the focus of Li et al. (2019); instead, it considered* ϵ*-dependence and other discretizations, and showed for example that 1.5 SRK discretization has improved mixing time bound of* $\widetilde{\mathcal{O}}\left(\frac{d}{\epsilon^{2/3}}\right)$*. The dimension dependence of this discretization, for example, can possibly be improved by our results too.*

Optimality In fact, the $\widetilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\epsilon}\right)$ mixing time of LMC has the optimal scaling one can expect. This is in terms of dependence on d and ϵ, over the class of all log-smooth and log-strongly-convex target measures. To illustrate this, consider the following Gaussian target distribution whose potential is

$$f(\boldsymbol{x})=\frac{m}{2}\sum_{i=1}^{d}x_i^2+\frac{L}{2}\sum_{i=d+1}^{2d}x_i^2,\quad\text{with } m=1, L\ge 4m. \tag{11}$$

We now establish a lower bound on the mixing time of LMC algorithm for this target measure.

[1] The dependence on κ is not readily available from Theorem 8 in Durmus & Moulines (2019).

[2] The $G=\mathcal{O}(L^2)$ assumption is only for κ, m dependence. Removing it doesn't affect d,ϵ dependence.

Theorem 4.3. ***(Lower Bound of Mixing Time)*** *Suppose we run LMC for the target measure defined in Eq.* (11) *from* $\boldsymbol{x}_0 = \mathbf{1}_{2d}$*, then for any choice of step size* $h > 0$ *within stability limit, we have*

$$\tau_{\text{mix}}(\epsilon; W_2; \text{LMC}) \geq \frac{\sqrt{d}}{8\epsilon} \log \frac{\sqrt{d}}{\epsilon} = \widetilde{\Omega}\left(\frac{\sqrt{d}}{\epsilon}\right).$$

Combining Theorem 4.2 and 4.3, we see that mean-square analysis provides a tight bound for LMC and $\widetilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\epsilon}\right)$ is the optimal scaling of LMC for target measures satisfying Assumptions 1 and 2.

Note that the above optimality results only partially, but *not* completely, close the gap between the upper and lower bounds of LMC over the entire family of log-smooth and log-strongly-convex target measures, because of one limitation of our result — A(ssumption)2 is, despite of its close relation to the Hessian Lipschitz condition frequently used in the literature, still an extra condition. We tend to believe that A2 may not be essential, but rather an artifact of our proof technique. However, at this moment we cannot eliminate the possibility that the best scaling one can get out of Assumption 1 only (no A2) is worse than $\widetilde{\mathcal{O}}\left(\sqrt{d}/\epsilon\right)$. We'd like to further investigate this in future work.

Discussion Besides Li et al. (2019) (see the previous Remark), let's briefly discuss three more important sampling algorithms related to LMC. Two of them are Kinetic Langevin Monte Carlo (KLMC) and Randomized Midpoint Algorithm (RMA), both of which are discretizations of kinetic Langevin dynamics. The other is Metropolis-Adjusted Langevin Algorithm (MALA) which uses the one-step update of LMC as a proposal and then accepts/rejects it with Metropolis-Hastings.

The $\widetilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\epsilon}\right)$ mixing time in 2-Wasserstein distance of KLMC has been established for log-smooth and log-strongly-convex target measures in existing literature (Cheng et al., 2018b; Dalalyan & Riou-Durand, 2020) and that was a milestone. Due to its better dimension dependence over previously best known results of LMC, KLMC is understood to be the analog of Nesterov's accelerated gradient method for sampling (Ma et al., 2021). Our findings show that LMC is able to achieve the same mixing time, although under an additional growth-at-infinity condition. However, this does not say anything about whether/how KLMC accelerates LMC, as the existing KLMC bound may still be not optimal. We also note KLMC has better condition number dependence than our current LMC result, although the κ dependence in our bound may not be tight.

RMA (Shen & Lee, 2019) is based on a brilliant randomized discretization of kinetic Langevin dynamics and shown to have further improved dimension dependence (and other pleasant properties). From the perspective of this work, we think it is because RMA is able to break the strong order barrier due to the randomization, and more investigations based on mean-square analysis should be possible.

For MALA, a recent breakthrough (Chewi et al., 2021) establishes a $\widetilde{\mathcal{O}}\left(\sqrt{d}\right)$ mixing time in W2 distance with warm start, and the dimension dependence is shown to be optimal. We see that without the Metropolis adjustment, LMC (under additional assumptions such as A2) can also achieve the same dimension dependence as MALA. But unlike LMC, MALA only has logarithmic dependence on $\frac{1}{\epsilon}$. With warm-start, is it possible/how to improve the dependence of $\frac{1}{\epsilon}$ for LMC, from polynomial to logarithmic? This question is beyond the scope of this paper but worth further investigation.

5 Numerical Examples

This section numerically verifies our theoretical findings for LMC in Section 4, with a particular focus on the dependence of the discretization error in Theorem 4.1 on dimension d and step size h. To this end, we consider two target measures specified by the following two potentials:

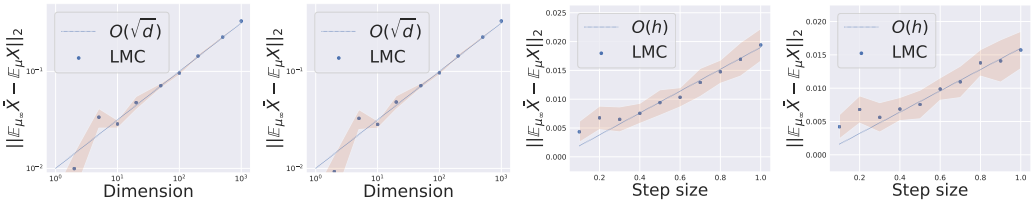
$$f_1(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2 + \log\left(\sum_{i=1}^{d} e^{x_i}\right) \quad \text{and} \quad f_2(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2 - \frac{1}{2d^{\frac{1}{2}}}\sum_{i=1}^{d}\cos\left(d^{\frac{1}{4}}x_i\right). \tag{12}$$

It is not hard to see f_1 is 2-smooth and 1-strongly convex, f_2 is $\frac{3}{2}$-smooth and 1-strongly-convex, and both satisfy Assumption 2. f_2 is also used in (Chewi et al., 2021) to illustrate the optimal dimension dependence of MALA. Explicit expression of 2-Wasserstein distance between non-Gaussian

distributions is typically not available, instead, we use the Euclidean norm of the mean error as a surrogate because $\left\|\mathbb{E}\bar{\boldsymbol{x}}_k - \mathbb{E}_\mu \boldsymbol{x}\right\| \leq W_2(\mathrm{Law}(\bar{\boldsymbol{x}}_k), \mu)$ due to Jensen's inequality. To obtain an accurate estimate of the ground truth, we run 10^8 independent LMC realizations using a tiny step size (h = 0.001), each till a fixed, long enough time, and use the empirical average to approximate $\mathbb{E}_\mu \boldsymbol{x}$.

To study the dimension dependence of sampling error, we fix step size $h = 0.1$, and for each $d \in \{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000\}$, we simulate 10^4 independent Markov chains using LMC algorithm for 100 iterations, which is long enough for the chain to be well-mixed. The mean and the standard deviation of the sampling error corresponding to the last 10 iterates are recorded.

To study step size dependence of sampling error, we fix $d = 10$ and experiment with step size $h \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \times 10^{-1}$. We run LMC till $T = 20$, i.e., $\lceil \frac{T}{h} \rceil$ iterations for each h. The procedure is repeated 10^4 times with different random seeds to obtain independent samples. When the corresponding continuous time $t = kh > 10$, we see from Eq. (10) that LMC is well converged and the sampling error is saturated by the discretization error. Therefore, for each h, we take the last $\lceil \frac{10}{h} \rceil$ iterates and record the mean and standard deviation of their sampling error.



(a) f_1: d dependence (b) f_2: d dependence (c) f_1: h dependence (d) f_2: h dependence

Figure 1: (a) Dependence of the sampling error of LMC on dimension d and step size h for f_1 and f_2. Both axes in Fig.1a & 1b are in log scale. Shaded areas in Fig.1a & 1b represent one std. of the last 10 iterations. Shaded areas in Fig.1c & 1d represent one std. of the last $\lceil \frac{10}{h} \rceil$ iterations.

The experiment results shown in Figure 1 are consistent with our theoretical analysis of the sampling error.Both linear dependence on $\sqrt{d}$ and h can be identified in and supported by the empirical evidence. Note results with smaller h are less accurate because one starts to see the error of empirical approximation due to finite samples. Experiments were conducted on a machine with a 2.20GHz Intel(R) Xeon(R) E5-2630 v4 CPU and an Nvidia GeForce GTX 1080 GPU.

6 Conclusion

Via a refined mean-square analysis of Langevin Monte Carlo algorithm, we obtain an improved and optimal $\widetilde{\mathcal{O}}\left(\sqrt{d}/\epsilon\right)$ bound on its mixing time, which was previously thought to be obtainable only with the addition of momentum. This was under the standard smoothness and strongly-convexity assumption, plus an addition linear growth condition on the third-order derivative of the potential function, similar to Hessian Lipschitz condition already popularized in the literature.

Here are some possible directions worth further investigations. (i) Combine mean-square analysis with stochastic gradient analysis to study SDE-basd stochastic gradient MCMC methods; (ii) Is it still possible to obtain $\sqrt{d}$-dependence without A2, i.e., only under log-smooth and log-strongly-convex conditions? (iii) Applications of mean-square analysis to other SDEs and/or discretizations; (iv) Motivated by Chewi et al. (2021), it would be interesting to know whether the dependence on $\frac{1}{\epsilon}$ can be improved to logarithmic, for example if LMC is initialized at a warm start.

References

Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale

machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283, 2016.

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1):5–43, 2003.

Changyou Chen, Ruiyi Zhang, Wenlin Wang, Bai Li, and Liqun Chen. A unified particle-optimization framework for scalable bayesian sampling. In *The Conference on Uncertainty in Artificial Intelligence*, 2018.

Xiang Cheng and Peter L Bartlett. Convergence of langevin mcmc in kl-divergence. *PMLR 83*, (83): 186–211, 2018.

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

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. *Proceedings of the 31st Conference On Learning Theory, PMLR*, 2018b.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the metropolis-adjusted langevin algorithm. *COLT*, 2021.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pp. 3036–3046, 2018.

Arnak S Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. *Conference on Learning Theory*, pp. 678–689, 2017a.

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

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

Masoumeh Dashti and Andrew M. Stuart. *The Bayesian Approach to Inverse Problems*, pp. 311–428. Springer International Publishing, Cham, 2017. ISBN 978-3-319-12385-1. doi: 10.1007/978-3-319-12385-1_7.

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

Alain Durmus and Eric Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Alain Durmus, Szymon Majewski, and Blazej Miasojedow. Analysis of langevin monte carlo via convex optimization. *Journal of Machine Learning Research*, 20:73–1, 2019.

Murat A Erdogdu and Rasa Hosseinzadeh. On the convergence of langevin monte carlo: The interplay between tail growth and smoothness. *COLT*, 2021.

Charlie Frogner and Tomaso Poggio. Approximate inference with wasserstein gradient flows. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Peter E Kloeden and Eckhard Platen. Numerical solution of stochastic differential equations. Springer, 1992.

Joseph Lehec. The langevin monte carlo algorithm in the non-smooth log-concave case. *arXiv preprint arXiv:2101.10695*, 2021.

Ruilin Li, Hongyuan Zha, and Molei Tao. Hessian-free high-resolution nesterov acceleration for sampling. *arXiv preprint arXiv:2006.09230*, 2020.

Ruilin Li, Molei Tao, Santosh S Vempala, and Andre Wibisono. The mirror langevin algorithm converges with vanishing bias. *arXiv preprint arXiv:2109.12077*, 2021.

Xuechen Li, Denny Wu, Lester Mackey, and Murat A Erdogdu. Stochastic Runge-Kutta accelerates Langevin Monte Carlo and beyond. *NeurIPS*, 2019.

Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in neural information processing systems*, pp. 2378–2386, 2016.

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. doi: 10.3150/20-BEJ1297.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Grigori Noah Milstein and Michael V Tretyakov. *Stochastic numerics for mathematical physics*. Springer Science & Business Media, 2013.

Wenlong Mou, Nicolas Flammarion, Martin J Wainwright, and Peter L Bartlett. Improved bounds for discretization of langevin diffusions: Near-optimal rates without convexity. *arXiv preprint arXiv:1907.11331*, 2019.

Wenlong Mou, Yi-An Ma, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. High-order langevin diffusion yields an accelerated mcmc algorithm. *Journal of Machine Learning Research*, 22(42):1–41, 2021.

Grigorios A Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014.

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1): 255–268, 1998.

Gareth O Roberts, Richard L Tweedie, et al. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

W Rüemelin. Numerical treatment of stochastic differential equations. *SIAM Journal on Numerical Analysis*, 19(3):604–613, 1982.

Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, pp. 2098–2109, 2019.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, pp. 8092–8104, 2019.

Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference On Learning Theory*, pp. 2093–3027, 2018.

Ruiyi Zhang, Changyou Chen, Chunyuan Li, and Lawrence Carin. Policy optimization as wasserstein gradient flows. In *International Conference on Machine Learning*, pp. 5737–5746, 2018.