Série des Documents de Travail

n° 2017-20
User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient

A.S. DALALYAN [1]
A. KARAGULYAN [2]

[1] Modal'X, Université Paris-Nanterre and CREST, ENSAE
[2] Modal'X, Université Paris-Nanterre and CREST, ENSAE

Submitted to arXiv

User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient

Arnak S. Dalalyan, Avetik Karagulyan

CREST, ENSAE, Université Paris-Saclay

Abstract. In this paper, we revisit the recently established theoretical guarantees for the convergence of the Langevin Monte Carlo algorithm of sampling from a smooth and (strongly) log-concave density. We improve, in terms of constants, the existing results when the accuracy of sampling is measured in the Wasserstein distance and provide further insights on relations between, on the one hand, the Langevin Monte Carlo for sampling and, on the other hand, the gradient descent for optimization. More importantly, we establish non-asymptotic guarantees for the accuracy of a version of the Langevin Monte Carlo algorithm that is based on inaccurate evaluations of the gradient. Finally, we propose a variable-step version of the Langevin Monte Carlo algorithm that has two advantages. First, its step-sizes are independent of the target accuracy and, second, its rate provides a logarithmic improvement over the constant-step Langevin Monte Carlo algorithm.

MSC 2010 subject classifications: Primary 62J05; secondary 62H12.

Key words and phrases: Markov Chain Monte Carlo, Approximate sampling, Rates of convergence, Langevin algorithm, Gradient descent.

1. INTRODUCTION

The problem of sampling a random vector distributed according to a given target distribution is central in many applications. In the present paper, we consider this problem in the case of a target distribution has a smooth and log-concave density and the sampling is performed by the Langevin Monte Carlo algorithm. More precisely, let p be a positive integer and $f : \mathbb{R}^p \to \mathbb{R}$ be a continuously differentiable function satisfying, for some positive constants m and M, the conditions

$$\begin{cases} f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}') - \nabla f(\boldsymbol{\theta}')^\top(\boldsymbol{\theta} - \boldsymbol{\theta}') \ge ({m}/{2})\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2, \\ \|\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}')\|_2 \le M\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \end{cases} \qquad \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p, \tag{1}$$

where ∇f stands for the gradient of f and $\|\cdot\|_2$ is the Euclidean norm. The target distributions considered in this paper are those having a density with respect to the Lebesgue measure on

3 avenue Pierre Larousse, 92245 Malakoff, France.

$\mathbb{R}^p$ given by

$$\pi(\boldsymbol{\theta}) = \frac{e^{-f(\boldsymbol{\theta})}}{\int_{\mathbb{R}^p} e^{-f(\boldsymbol{u})}\, d\boldsymbol{u}}.$$

We say that the density $\pi(\boldsymbol{\theta}) \propto e^{-f(\boldsymbol{\theta})}$ is log-concave (resp. strongly log-concave) if the function f satisfies the first inequality of (1) with $m = 0$ (resp. $m > 0$).

The Langevin Monte Carlo (LMC) algorithm studied throughout this work is the analogue of the gradient descent algorithm for optimization. Starting from an initial point $\boldsymbol{\vartheta}_0 \in \mathbb{R}^p$ that may be deterministic or random, the iterations of the algorithm are defined by the update rule

$$\boldsymbol{\vartheta}_{k+1,\boldsymbol{h}} = \boldsymbol{\vartheta}_{k,\boldsymbol{h}} - h_{k+1}\nabla f(\boldsymbol{\vartheta}_{k,\boldsymbol{h}}) + \sqrt{2h_{k+1}}\, \boldsymbol{\xi}_{k+1}; \qquad k = 0, 1, 2, \ldots \tag{2}$$

where $\boldsymbol{h} = \{h_k\}_{k\in\mathbb{N}}$ is a sequence of positive parameters, referred to as the step-sizes, and $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k, \ldots$ is a sequence of mutually independent, and independent of $\boldsymbol{\vartheta}_0$, centered Gaussian vectors with covariance matrices equal to identity.

When all the h_k's are equal to some value $h > 0$, we will call the sequence in (2) the constant step-size LMC and will denote it by $\boldsymbol{\vartheta}_{k+1,h}$. Under the assumptions imposed on f, when h is small and k is large (so that the product kh is large), the distribution of $\boldsymbol{\vartheta}_{k,h}$ is close in various metrics to the distribution with density $\pi(\boldsymbol{\theta})$, hereafter referred to as the target distribution. An important question is to quantify this closeness. We address this question by establishing user friendly non asymptotic upper bounds on the error of sampling; these kind of bounds are particularly useful for deriving a stopping rule for the LMC algorithm.

In this paper, we measure the error of sampling in the Wasserstein-Monge-Kantorovich distance W_2. For two measures μ and ν defined on $(\mathbb{R}^p, \mathscr{B}(\mathbb{R}^p))$, and for a real number $q \geq 1$, W_q is defined by

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

where the inf is with respect to all joint distributions ϱ having μ and ν as marginal distributions. This distance is perhaps more suitable for quantifying the quality of approximate sampling schemes than other metrics such as the total variation or the Kullback-Leibler divergence. Indeed, on the one hand, bounds on the Wasserstein distance—unlike the bounds on the total-variation distance—directly provide the level of approximating the first and the second order moments. For instance, if μ and ν are two Dirac measures at the points $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, respectively, then the total-variation distance $D_{\rm TV}(\delta_{\boldsymbol{\theta}}, \delta_{\boldsymbol{\theta}'})$ equals one whenever $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$, whereas $W_2(\delta_{\boldsymbol{\theta}}, \delta_{\boldsymbol{\theta}'}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$ is a smoothly increasing function of the Euclidean distance between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. This seems to better correspond to the intuition on the closeness of two distributions.

Throughout this work, for any matrix $\mathbf{M}$, we denote by $\|\mathbf{M}\|$ and $\|\mathbf{M}\|_F$, respectively, the spectral norm and the Frobenius norm of $\mathbf{M}$.

Asymptotic properties of the LMC algorithm, also known as Unadjusted Langevin Algorithm (ULA), and its Metropolis adjusted version, MALA, have been studied in a number of papers (Jarner and Hansen, 2000; Roberts and Rosenthal, 1998; Roberts and Stramer, 2002; Roberts and Tweedie, 1996; Stramer and Tweedie, 1999a,b). These results do not emphasize the effect of the dimension on the computational complexity of the algorithm, which is roughly proportional to the number of iterations. Non asymptotic bounds on the total variation error of the LMC for log-concave and strongly log-concave distributions have been established by Dalalyan (2014). If a warm start is available, the results in Dalalyan (2014) imply that after $O(p/\epsilon^2)$ iterations the LMC algorithm has an error bounded from above by ϵ. Furthermore, if we assume that in addition to (1) the function f has a Lipschitz continuous Hessian, then a modified version of the LMC, the LMC with Ozaki discretization (LMCO), needs $O(p/\epsilon)$ iterations to achieve a precision level ϵ. These results were improved and extended to the Wasserstein distance by (Durmus and Moulines, 2016; Durmus and Moulines, 2017). More precisely, they removed the condition of the warm start and proved that under the Lipschitz continuity assumption on the Hessian of f, it is not necessary to modify the LMC for getting the rate $O(p/\epsilon)$. The last result is closely related to an error bound between a diffusion process and its Euler discretization established by Alfonsi et al. (2014).

On a related note, (Bubeck et al., 2015) studied the convergence of the LMC algorithm with reflection at the boundary of a compact set, which makes it possible to sample from a compactly supported density (see also (Brosse et al., 2017)). Extensions to non-smooth densities were presented in (Durmus et al., 2016; Luu et al., 2017). (Cheng and Bartlett, 2017) obtained guarantees similar to those in (Dalalyan, 2014) when the error is measured by the Kullback-Leibler divergence. Very recently, (Cheng et al., 2017) derived non asymptotic guarantees for the underdamped LMC which turned out to improve on the previously known results. Langevin dynamics was used in (Andrieu et al., 2016; Brosse et al., 2017) in order to approximate normalizing constants of target distributions. Huggins and Zou (2017) established tight bounds in Wasserstein distance between the invariant distributions of two (Langevin) diffusions; the bounds involve mixing rates of the diffusions and the deviation in their drifts.

The goal of the present work is to push further the study of the LMC and its variants both by improving the existing guarantees and by extending them in some directions. Our main contributions can be summarized as follows:

- We get improved and simplified guarantees in Wasserstein distance both for the LMC and the LMCO when the step-size is constant, see Theorem 1 and Theorem 5.
- We propose varying-step LMC which avoids a logarithmic factor in the number of iterations required to achieve a precision level ϵ, see Theorem 2.
- We extend the previous guarantees to the case where accurate evaluations of the gradient are unavailable. Thus, at each iteration of the algorithm, the gradient is computed within an error that has a deterministic and a stochastic component. Theorem 3 deals with functions f satisfying (1), whereas Theorem 4 requires the additional assumption of the smoothness of the Hessian of f.
- We propose a new second-order sampling algorithm termed LMCO'. It has a per-iteration computational cost comparable to that of the LMC and enjoys nearly the same guarantees as the LMCO, when the Hessian of f is Lipschitz continuous, see Theorem 5.
- We provide a detailed discussion of the relations between, on the one hand, the sampling methods and guarantees of their convergence and, on the other hand, optimization methods and guarantees of their convergence (see Section 5).

We have to emphasize right away that theorem 1 is a corrected version of (Dalalyan, 2017, Theorem 1), whereas Theorem 3 extends (Dalalyan, 2017, Theorem 3) to more general noise. In particular, Theorem 3 removes the unbiasedness and independence conditions. Furthermore, thanks to a shrewd use of a recursive inequality, the upper bound in Theorem 3 is tighter than the one in (Dalalyan, 2017, Theorem 3).

As an illustration of the first two bullets mentioned in the above summary of our contributions, let us consider the following example. Assume that $m = 10$, $M = 20$ and we have at our disposal an initial sampling distribution ν_0 satisfying $W_2(\nu_0, \pi) = p + (p/m)$. The main inequalities in Theorem 1 and Theorem 2 imply that after K iterations, we have

$$W_2(\nu_K, \pi) \leq (1 - mh)^K W_2(\nu_0, \pi) + 1.65(M/m)(hp)^{1/2} \tag{3}$$

for the constant step LMC and

$$W_2(\nu_K, \pi) \leq \frac{3.5M\sqrt{p}}{m\sqrt{M + m + (2/3)m(K - K_1)}} \tag{4}$$

for the varying step LMC, where K_1 is an integer the precise value of which is provided in Theorem 2. One can compare these inequalities with the corresponding bound in (Durmus and Moulines, 2016): adapted to the constant-step, it takes the form

$$W_2^2(\nu_K, \pi) \leq 2\Big(1 - \frac{mMh}{m + M}\Big)^K W_2^2(\nu, \pi) + \frac{Mhp}{m}(m+M)\Big(h + \frac{m + M}{2mM}\Big)\Big(2 + \frac{M^2h}{m} + \frac{M^2h^2}{6}\Big). \tag{5}$$

For any $\epsilon > 0$, we can derive from these guarantees the smallest number of iterations, K_ϵ, for which there is a $h > 0$ such that the corresponding upper bound is smaller than ϵ. The logarithms of these values K_ϵ for varying $\epsilon \in \{0.001, 0.005, 0.01\}$ and $p \in \{25, 1000\}$ are plotted in Figure 1. We observe that for all the considered values of ϵ and p, the number of iterations derived from (4) (referred to as Theorem 2) is smaller than those derived from (3) (referred to as Theorem 1) and from (5) (referred to as DM bound). The difference between the varying step LMC and the constant step LMC becomes more important when the target precision level ϵ gets smaller. In average over all values of p, when $\epsilon = 0.001$, the number of iterations derived from (5) is 4.6 times larger than that derived from (4), and almost 3 times larger than the number of iterations derived from (3).

2. IMPROVED GUARANTEES IN THE WASSERSTEIN DISTANCE

The rationale behind the LMC algorithm (2) is simple: the Markov chain $\{\boldsymbol{\vartheta}_{k,h}\}_{k\in\mathbb{N}}$ is the Euler discretization of a continuous-time diffusion process $\{\boldsymbol{L}_t : t \in \mathbb{R}_+\}$, known as Langevin diffusion, that has π as invariant density (Bhattacharya, 1978, Thm. 3.5). The Langevin diffusion is defined by the stochastic differential equation

$$d\boldsymbol{L}_t = -\nabla f(\boldsymbol{L}_t)\,dt + \sqrt{2}\,d\boldsymbol{W}_t, \qquad t \geq 0, \tag{6}$$

where $\{\boldsymbol{W}_t : t \geq 0\}$ is a p-dimensional Brownian motion. When f satisfies condition (1), equation (6) has a unique strong solution which is a Markov process. Let ν_k be the distribution of the k-th iterate of the LMC algorithm, that is $\vartheta_{k,h} \sim \nu_k$. In what follows, we present user-friendly guarantees on the closeness of ν_k and π in the strongly convex and non-strongly convex situations.

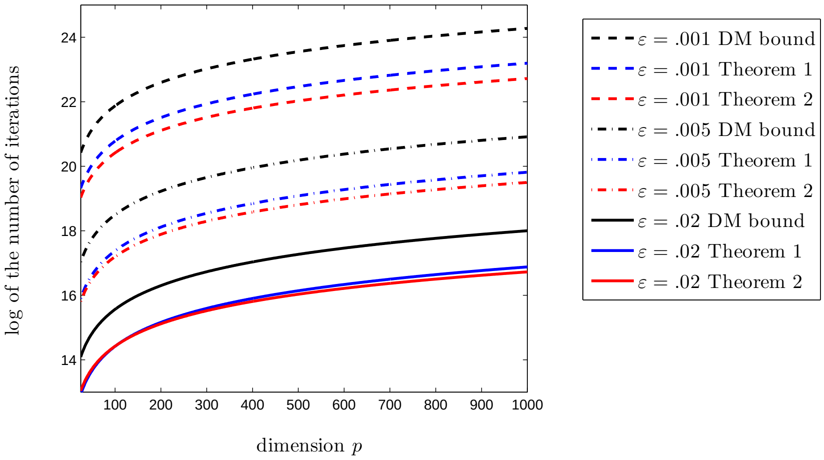

FIG 1. *Plots showing the logarithm of the number of iterations as function of dimension p for several values of ϵ. The plotted values are derived from (3)-(5) using the data $m = 10$, $M = 20$, $W_2(\nu_0, \pi) = p + (p/m)$.*

2.1 Guarantees under strong convexity for the constant step LMC

We start by considering the simpler case where the function f is m-strongly convex, that is it satisfies (1).

THEOREM 1. *Assume that $h \in (0, 2/M)$. The following claims hold:*

(a) *If $h \le 2/(m+M)$ then $W_2(\nu_K, \pi) \le (1 - mh)^K W_2(\nu_0, \pi) + 1.65(M/m)(hp)^{1/2}$.*

(b) *If $h \ge 2/(m+M)$ then $W_2(\nu_K, \pi) \le (Mh - 1)^K W_2(\nu_0, \pi) + \dfrac{1.65Mh}{2 - Mh}(hp)^{1/2}$.*

The proof of this theorem is postponed to Section 7. The factor 1.65 is obtained by upper bounding $7\sqrt{2}/6$.

In practice, a relevant approach to getting an accuracy of at most ϵ is to minimize the upper bound provided by Theorem 1 with respect to h, for a fixed K. Then, one can choose the smallest K for which the obtained upper bound is smaller than ϵ. One useful observation is that the upper bound of case (b) is an increasing function of h. Its minimum is always attained at $h = 2/(m + M)$, which means that one can always look for a step-size in the interval $(0, 2/(m+M)]$ by minimizing the upper bound in (a). This can be done using standard line-search methods such as the bisection algorithm.

Note that if the initial value $\boldsymbol{\vartheta}_0 = \boldsymbol{\theta}_0$ is deterministic then, using the notation $\bar{\boldsymbol{\theta}} = \int_{\mathbb{R}^p} \boldsymbol{\theta}\pi(d\boldsymbol{\theta})$, in view of (Durmus and Moulines, 2016, Theorem 1), we have

$$\begin{aligned} W_2(\nu_0, \pi)^2 &= \int_{\mathbb{R}^p} \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|_2^2 \pi(d\boldsymbol{\theta}) \\ &= \|\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}\|_2^2 + \int_{\mathbb{R}^p} \|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \pi(d\boldsymbol{\theta}) \\ &\le \|\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}\|_2^2 + p/m. \end{aligned} \tag{7}$$

Finally, let us remark that if we choose h and K so that

$$h \leq {}^2\!/\!{}_{(m+M)}, \qquad e^{-mhK} W_2(\nu_0, \pi) \leq \varepsilon/2, \quad 1.65(M/m)(hp)^{1/2} \leq \varepsilon/2, \tag{8}$$

then we have $W_2(\nu_K, \pi) \leq \varepsilon$. In other words, conditions (8) are sufficient for the density of the output of the LMC algorithm with K iterations to be within the precision ε of the target density when the precision is measured using the Wasserstein distance. This readily yields

$$h \leq \frac{m^2\varepsilon^2}{11M^2p} \wedge \frac{2}{m+M} \quad \text{and} \quad hK \geq \frac{1}{m}\log\Big(\frac{2(\|\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}\|_2^2 + p/m)^{1/2}}{\varepsilon}\Big)$$

Assuming m, M and $\|\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}\|_2^2/p$ to be constants, we can deduce from the last display that it suffices $K = C(p/\varepsilon^2)\log(p/\varepsilon^2)$ number of iterations in order to reach the precision level ε. This fact has been first established in (Dalalyan, 2014) for the LMC algorithm with a warm start and the total-variation distance. It was later improved by Durmus and Moulines (2016); Durmus and Moulines (2017), who showed that the same result holds for any starting point and established similar bounds for the Wasserstein distance. Theorem 1 above can be seen as a user-friendly version of the corresponding result established by Durmus and Moulines (2016).

Remark 2.1. *Although the upper bound on $W_2(\nu_0, \pi)$ provided by (7) is relevant for understanding the order of magnitude of $W_2(\nu_0, \pi)$, it has limited applicability since the distance $\|\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}\|$ might be hard to evaluate. An attractive alternative to that bound is the following*[1]*:*

$$\begin{aligned} W_2(\nu_0, \pi)^2 &= \int_{\mathbb{R}^p} \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|_2^2 \pi(d\boldsymbol{\theta}) \\ &\leq \frac{2}{m}\int_{\mathbb{R}^p} \Big(f(\boldsymbol{\theta}_0) - f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})^\top(\boldsymbol{\theta}_0 - \boldsymbol{\theta})\Big)\pi(d\boldsymbol{\theta}) \\ &= \frac{2}{m}\Big(f(\boldsymbol{\theta}_0) - \int_{\mathbb{R}^p} f(\boldsymbol{\theta})\,\pi(d\boldsymbol{\theta}) + p\Big). \end{aligned}$$

If f is lower bounded by some known constant, for instance if $f \geq 0$, the last inequality provides the computable upper bound $W_2(\nu_0, \pi)^2 \leq \frac{2}{m}\big(f(\boldsymbol{\theta}_0) + p\big)$.

2.2 Guarantees under strong convexity for the variable step LMC

The result of previous section provides a guarantee for the constant step LMC. One may wonder if using a variable step sizes $\boldsymbol{h} = \{h_k\}_{k\in\mathbb{N}}$ can improve the convergence. Note that in (Durmus and Moulines, 2016, Theorem 5), guarantees for the variable step LMC are established. However, they do not lead to a clear message on the choice of the step-sizes. The next result fills this gap by showing that an appropriate selection of step-sizes improves on the constant step LMC with an improvement factor logarithmic in p/ϵ^2.

Theorem 2. *Let us consider the LMC algorithm with varying step-size h_{k+1} defined by*

$$h_{k+1} = \frac{2}{M + m + ({}^2\!/\!{}_3)m(k - K_1)_+}, \qquad k = 1, 2, \ldots \tag{9}$$

[1]The second line follows from strong convexity whereas the third line is a consequence of the two identities $\int_{\mathbb{R}^p} \nabla f(\boldsymbol{\theta})\pi(d\boldsymbol{\theta}) = 0$ and $\int_{\mathbb{R}^p} \boldsymbol{\theta}^\top \nabla f(\boldsymbol{\theta})\pi(d\boldsymbol{\theta}) = p$. These identities follow from the fundamental theorem of calculus and the integration by parts formula, respectively.

where K_1 is the smallest non-negative integer satisfying

$$K_1 \geq \frac{\ln\big(W_2(\nu_0,\pi)/\sqrt{p}\big) + \ln(m/M) + (1/2)\ln(M+m)}{\ln(1 + 2m/M-m)}.$$

For every positive integer $k \geq K_1$, we have

$$W_2(\nu_k,\pi) \leq \frac{3.5M\sqrt{p}}{m\sqrt{M+m+(2/3)m(k-K_1)}}. \tag{10}$$

The step size (9) has two important advantages as compared to the constant steps. The first advantage is that it is independent of the target precision level ϵ. The second advantage is that we get rid of the logarithmic terms in the number of iterations required to achieve the precision level ϵ. Indeed, it suffices $K = K_1 + (27M^2/2m^3)(p/\epsilon^2)$ iterations to get the right hand side of (10) smaller than ϵ, where K_1 depends neither on the dimension p nor on the precision level ϵ.

Since the choice of h_{k+1} in (9) might appear mysterious, we provide below a quick explanation of the main computations underpinning this choice. The main step of the proof of upper bounds on $W_2(\nu_k,\pi)$, is the following recursive inequality (see Proposition 2 in Section 7)

$$W_2(\nu_{k+1},\pi) \leq (1-mh_{k+1})W_2(\nu_k,\pi) + 1.65M\sqrt{p}\,h_{k+1}^{3/2}.$$

Using the notation $B_k = \frac{2(m/3)^{3/2}}{1.65M\sqrt{p}}W_2(\nu_k,\pi)$, this inequality can be rewritten as

$$B_{k+1} \leq (1-mh_{k+1})B_k + 2(mh_{k+1}/3)^{3/2}.$$

Minimizing the right hand side with respect to h_{k+1}, we find that the minimum is attained at the stationary point

$$h_{k+1} = \frac{3}{m}B_k^2. \tag{11}$$

With this h_{k+1}, one checks that the sequence B_k satisfies the recursive inequality

$$B_{k+1}^2 \leq B_k^2(1-B_k^2)^2 \leq \frac{B_k^2}{1+B_k^2}.$$

The function $g(x) = x/(1+x)$ being increasing in $(0,\infty)$, we get

$$B_{k+1}^2 \leq \frac{B_k^2}{1+B_k^2} \leq \frac{\frac{B_{k-1}^2}{1+B_{k-1}^2}}{1+\frac{B_{k-1}^2}{1+B_{k-1}^2}} = \frac{B_{k-1}^2}{1+2B_{k-1}^2}.$$

By repetitive application of the same argument, we get

$$B_{k+1}^2 \leq \frac{B_{K_1}^2}{1+(k+1-K_1)B_{K_1}^2}.$$

The integer K_1 was chosen so that $B_{K_1}^2 \leq \frac{2m}{3(M+m)}$, see (23). Inserting this upper bound in the right hand side of the last display, we get

$$B_{k+1}^2 \leq \frac{2m}{3(M+m)+2m(k+1-K_1)}.$$

Finally, replacing in (11) B_k^2 by its upper bound derived from the last display, we get the suggested value for h_{k+1}.

3. GUARANTEES FOR THE INACCURATE GRADIENT VERSION

In some situations, the precise evaluation of the gradient $\nabla f(\boldsymbol{\theta})$ is computationally expensive or practically impossible, but it is possible to obtain noisy evaluations of ∇f at any point. This is the setting considered in the present section. More precisely, we assume that at any point $\boldsymbol{\vartheta}_{k,h} \in \mathbb{R}^p$ of the LMC algorithm, we can observe the value

$$\boldsymbol{Y}_{k,h} = \nabla f(\boldsymbol{\vartheta}_{k,h}) + \boldsymbol{\zeta}_k,$$

where $\{\boldsymbol{\zeta}_k : k = 0, 1, \ldots\}$ is a sequence of random (noise) vectors. The noisy LMC (nLMC) algorithm is defined as

$$\boldsymbol{\vartheta}_{k+1,h} = \boldsymbol{\vartheta}_{k,h} - h\boldsymbol{Y}_{k,h} + \sqrt{2h}\,\boldsymbol{\xi}_{k+1}; \qquad k = 0, 1, 2, \ldots \tag{12}$$

where $h > 0$ and $\boldsymbol{\xi}_{k+1}$ are as in (2). The noise $\{\boldsymbol{\zeta}_k : k = 0, 1, \ldots\}$ is assumed to satisfy the following condition.

Condition N: for some $\delta > 0$ and $\sigma > 0$ and for every $k \in \mathbb{N}$,

- (bounded bias) $\mathbf{E}\big[\big\|\mathbf{E}(\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,h})\big\|_2^2\big] \le \delta^2 p$,
- (bounded variance) $\mathbf{E}[\|\boldsymbol{\zeta}_k - \mathbf{E}(\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,h})\|_2^2] \le \sigma^2 p$,
- (independence of updates) $\boldsymbol{\xi}_{k+1}$ in (12) is independent of $(\boldsymbol{\zeta}_0, \ldots, \boldsymbol{\zeta}_k)$.

The next theorem extends the guarantees of Theorem 1 to the inaccurate-gradient setting and to the nLMC algorithm.

THEOREM 3. *Let $\boldsymbol{\vartheta}_{K,h}$ be the K-th iterate of the nLMC algorithm (12) and ν_K be its distribution. If the function f satisfies condition (1) and $h \le 2/(m+M)$ then*

$$W_2(\nu_K, \pi) \le (1 - mh)^K W_2(\nu_0, \pi) + 1.65(M/m)(hp)^{1/2} + \frac{\delta\sqrt{p}}{m} + \frac{\sigma^2 (hp)^{1/2}}{1.65M + \sigma\sqrt{m}}\,. \tag{13}$$

To the best of our knowledge, this is the first result providing guarantees for sampling from a distribution in the scenario when precise evaluations of the log-density or its gradient are not available. Even in an asymptotic set-up, this problem apparently has not been the object of any investigation. The closely related problem of computing an average value with respect to a distribution, when the gradient of its log-density is known up to an additive noise, has been studied by Nagapetyan et al. (2017); Teh et al. (2016); Vollmer and Zygalakis (2015). Note that these settings are of the same flavor as those of stochastic approximation, an active area of research in optimization and machine learning.

To understand the potential scope of applicability of Theorem 3, let us consider a generic example in which $f(\boldsymbol{\theta})$ is the average of n functions defined through independent random variables $X_1, \ldots, X_n$:

$$f(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, X_i).$$

When the gradient of $\ell(\boldsymbol{\theta}, X_i)$ with respect to parameter $\boldsymbol{\theta}$ is hard to compute, one can replace the evaluation of $\nabla f(\boldsymbol{\vartheta}_{k,h})$ at each step k by that of $Y_k = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\vartheta}_{k,h}, X_{N_k})$, where N_k is a random variable uniformly distributed in $\{1, \ldots, n\}$ and independent of $\boldsymbol{\vartheta}_{k,h}$. Under suitable assumptions, this random vector satisfies the conditions of Theorem 3 with $\delta = 0$ and constant σ^2. Therefore, if we analyze the upper bound provided by (13), we see that the last term, due to the subsampling, is of the same order of magnitude as the second term. Thus, using the subsampled gradient in the LMC algorithm does not cause a significant deterioration of the precision while reducing considerably the computational burden.

Note that Theorem 3 allows to handle situations in which the approximations of the gradient are biased. This bias is controlled by the parameter δ. Such a bias can appear when using deterministic approximations of integrals or differentials. For instance, in statistical models with latent variables, the log-likelihood has often an integral form. Such integrals can be approximated using quadrature rules, yielding a bias term, or Monte Carlo methods, yielding a variance term.

In the preliminary version (Dalalyan, 2017) of this work, we made a mistake by claiming that the stochastic gradient version of the LMC, introduced in (Welling and Teh, 2011) and often referred to as Stochastic Gradient Langevin Dynamics (SGLD), has an error of the same order as the non-stochastic version of it. This claim is wrong, since when $f(\boldsymbol{\theta}) = \sum_{i=1}^n \ell(\boldsymbol{\theta}, X_i)$ with a strongly convex function $\boldsymbol{\theta} \mapsto \ell(\boldsymbol{\theta}, x)$ and iid variables $X_1, \ldots, X_n$, we have m and M proportional to n. Therefore, choosing $Y_k = n\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\vartheta}_{k,h}, X_{N_k})$ as a noisy version of the gradient (where N_k is a uniformly over $\{1, \ldots, n\}$ distributed random variable independent of $\boldsymbol{\vartheta}_{k,h}$), we get $\delta = 0$ but σ^2 proportional to n^2. Therefore, the last term in (13) is of order $(nhp)^{1/2}$ and dominates the other terms. Furthermore, replacing Y_k by $Y_k = \frac{n}{s}\sum_{j=1}^s \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\vartheta}_{k,h}, X_{N_k^j})$ with iid variables $N_k^1, \ldots, N_k^s$ does not help, since then σ^2 is of order n^2/s and the last term in (13) is of order $(nhp/s)^{1/2}$, which is still larger than the term $(M/m)(hp)^{1/2}$. This discussion shows that Theorem 3 does not provide any interesting result when applied to SGLD. For a more in-depth analysis of the SGLD, we refer the reader to (Nagapetyan et al., 2017; Raginsky et al., 2017; Xu et al., 2017).

It is also worth mentioning here that another example of approximate gradient—based on a quadratic approximation of the log-likelihood of the generalized linear model—has been considered in (Huggins and Zou, 2017, Section 5). It corresponds, in terms of condition N, to a situation in which the variance σ^2 vanishes but the bias δ is non-zero.

An important ingredient of the proof of Theorem 3 is the following simple result, which can be useful in other contexts as well.

LEMMA 1. *Let A, B and C be given non-negative numbers such that $A \in (0, 1)$. Assume that the sequence of non-negative numbers $\{x_k\}_{k=0,1,2,\ldots}$ satisfies the recursive inequality*

$$x_{k+1}^2 \leq [(1 - A)x_k + C]^2 + B^2$$

for every integer $k \geq 0$. Then, for all integers $k \geq 0$,

$$x_k \leq (1 - A)^k x_0 + \frac{C}{A} + \frac{B^2}{C + \sqrt{A}\,B}.$$

Thanks to this lemma, the upper bound on the Wasserstein distance provided by (13) is sharper than the one proposed in (Dalalyan, 2017).

4. GUARANTEES UNDER ADDITIONAL SMOOTHNESS ASSUMPTIONS

When the function f has Lipschitz continuous Hessian, one can get improved rates of convergence. This has been noted by (Dalalyan, 2014), who proposed to use a modified version of the LMC algorithm, the LMC with Ozaki discretization, in order to take advantage of the smoothness of the Hessian. Durmus and Moulines (2016) showed that the same rate in terms of p and ϵ can be achieved by the original LMC without any modification.

Condition F: the function f is twice differentiable and for some positive numbers m, M and M_2,

- (strong convexity) $\nabla^2 f(\boldsymbol{\theta}) \succeq m\mathbf{I}_p$, for every $\boldsymbol{\theta} \in \mathbb{R}^p$,
- (bounded second derivative) $\nabla^2 f(\boldsymbol{\theta}) \preceq M\mathbf{I}_p$, for every $\boldsymbol{\theta} \in \mathbb{R}^p$,
- (further smoothness) $\|\nabla^2 f(\boldsymbol{\theta}) - \nabla^2 f(\boldsymbol{\theta}')\| \le M_2\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$, for every $\boldsymbol{\theta} \in \mathbb{R}^p$.

THEOREM 4. *Let $\boldsymbol{\vartheta}_{K,h}$ be the K-th iterate of the nLMC algorithm (12) and ν_K be its distribution. Assume that conditions* **F** *and* **N** *are satisfied. Then, for every $h \le 2/(m+M)$*

$$W_2(\nu_K, \pi) \le (1 - mh)^K W_2(\nu_0, \pi) + \frac{M_2 hp}{2m} + \frac{11Mh\sqrt{Mp}}{5m} + \frac{\delta\sqrt{p}}{m} + \frac{2\sigma^2\sqrt{hp}}{M_2\sqrt{hp} + 2\sigma\sqrt{m}}.$$

In the last inequality, $11/5$ is an upper bound for $0.5 + 2\sqrt{2/3} \approx 2.133$. Theorem 4 is essentially a constant step-size version of (Durmus and Moulines, 2016, Theorem 8), with optimized constants and an extension to the scenario of inaccurate gradient.

When applying the nLMC algorithm to sample from a target density, the user may usually specify four parameters: the step-size h, the number of iterations K, the tolerated precision δ of the deterministic approximation and the precision σ of the stochastic approximation. An attractive feature of Theorem 4 is that the contributions of these four parameters are well separated, especially if we upper bound the last term by $2\sigma^2/M_2$. As a consequence, in order to have an error of order ϵ in Wasserstein distance, we might choose: σ at most of order $\sqrt{\epsilon}$, δ at most of order $m\epsilon/\sqrt{p}$, h of order ϵ/p and K of order $(p/\epsilon)\log(p/\epsilon)$. Akin to Theorem 2, one can use variable step-sizes to avoid the logarithmic factor; we leave these computations to the reader[2].

Under the assumption of Lipschitz continuity of the Hessian of f, one may wonder whether second-order methods that make use of the Hessian in addition to the gradient are able to outperform the standard LMC algorithm. The most relevant candidate algorithms for this are the LMC with Ozaki discretization (LMCO) and a variant of it, LMCO', a slightly modified version of an algorithm introduced in (Dalalyan, 2014). The LMCO is defined as follows: For every $k \ge 0$, we set $\mathbf{H}_k = \nabla^2 f(\boldsymbol{\vartheta}^{\rm LMCO}_{k,h})$, which is an invertible $p \times p$ matrix since f is strongly convex, and define

$$\begin{aligned} &\mathbf{M}_k = \big(\mathbf{I}_p - e^{-h\mathbf{H}_k}\big)\mathbf{H}_k^{-1}, \qquad \boldsymbol{\Sigma}_k = \big(\mathbf{I}_p - e^{-2h\mathbf{H}_k}\big)\mathbf{H}_k^{-1}, \\ &\boldsymbol{\vartheta}^{\rm LMCO}_{k+1,h} = \boldsymbol{\vartheta}^{\rm LMCO}_{k,h} - \mathbf{M}_k \nabla f\big(\boldsymbol{\vartheta}^{\rm LMCO}_{k,h}\big) + \boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\xi}_{k+1}, \end{aligned} \tag{14}$$

where $\{\boldsymbol{\xi}_k : k \in \mathbb{N}\}$ is a sequence of independent random vectors distributed according to the $\mathcal{N}_p(0, \mathbf{I}_p)$ distribution. The LMCO' algorithm is based on approximating the matrix

[2] A bound of that type is established in (Bonis, 2016, Corollary 3).

exponentials by linear functions, more precisely, for $\mathbf{H}'_k = \nabla^2 f(\boldsymbol{\vartheta}^{\rm LMCO'}_{k,h})$,

$$\boldsymbol{\vartheta}^{\rm LMCO'}_{k+1,h} = \boldsymbol{\vartheta}^{\rm LMCO'}_{k,h} - h\Big(\mathbf{I}_p - \frac{1}{2}h\mathbf{H}'_k\Big)\nabla f\big(\boldsymbol{\vartheta}^{\rm LMCO'}_{k,h}\big) + \sqrt{2h}\Big(\mathbf{I}_p - h\mathbf{H}'_k + \frac{1}{3}h^2(\mathbf{H}'_k)^2\Big)^{1/2}\boldsymbol{\xi}_{k+1}. \tag{15}$$

Let us mention right away that the stochastic perturbation present in the last display can be computed in practice without taking the matrix square-root. Indeed, it suffices to generate two independent standard Gaussian vectors $\boldsymbol{\eta}_{k+1}$ and $\boldsymbol{\eta}'_{k+1}$; then the random vector

$$\big(\mathbf{I}_p - (1/2)h\mathbf{H}'_k\big)\boldsymbol{\eta}_{k+1} + (\sqrt{3}/6)\,h\,\mathbf{H}'_k\boldsymbol{\eta}'_{k+1}$$

has exactly the same distribution as the vector $\big(\mathbf{I}_p - h\mathbf{H}'_k + (1/3)h^2(\mathbf{H}'_k)^2\big)^{1/2}\boldsymbol{\xi}_{k+1}$.

In the rest of this section, we provide guarantees for methods LMCO and LMCO'. Note that we consider only the case where the gradient and the Hessian of f are computed exactly, that is without any approximation.

Theorem 5. *Let $\nu_K^{\rm LMCO}$ and $\nu_K^{\rm LMCO'}$ be, respectively, the distributions of the K-th iterate of the LMCO algorithm (14) and the LMCO' algorithm (15) with an initial distribution ν_0. Assume that conditions* **F** *and* **N** *are satisfied. Then, for every $h \le m/M^2$,*

$$W_2(\nu_K^{\rm LMCO}, \pi) \le (1 - 0.25mh)^k W_2(\nu_0, \pi) + \frac{11.5M_2h(p+1)}{m}. \tag{16}$$

If, in addition, $h \le 3m/4M^2$, then

$$W_2(\nu_K^{\rm LMCO'}, \pi) \le (1 - 0.25mh)^k W_2(\nu_0, \pi) + \frac{1.3M^2h^2\sqrt{Mp}}{m} + \frac{7.3M_2h(p+1)}{m}. \tag{17}$$

A very rough consequence of this theorem is that one has similar theoretical guarantees for the LMCO and the LMCO' algorithms, since in most situations the middle term in the right hand side of (17) is smaller than the last term. On the other hand, the per-iteration cost of the modified algorithm LMCO' is significantly smaller than the per-iteration cost of the original LMCO. Indeed, for the LMCO' there is no need to compute matrix exponentials neither to invert matrices, one only needs to perform matrix-vector multiplication for $p \times p$ matrices. Note that for many matrices such a multiplication operation might be very cheap using the fast Fourier transform or other similar techniques. In addition, the computational complexity of the Hessian-vector product is provably of the same order as that of evaluating the gradient, see (Griewank, 1993). Therefore, one iteration of the LMCO' algorithm is not more costly than one iteration of the LMC. At the same time, the error bound (17) for the LMCO' is smaller than the one for the LMC provided by Theorem 4. Indeed, the term $Mh\sqrt{Mp}$ present in the bound of Theorem 4 is generally of larger order than the term $(Mh)^2\sqrt{Mp}$ appearing in (17).

5. RELATION WITH OPTIMIZATION

We have already mentioned that the LMC algorithm is very close to the gradient descent algorithm for computing the minimum $\boldsymbol{\theta}^*$ of the function f. However, when we compare the guarantees of Theorem 1 with those available for the optimization problem, we remark the following striking difference. The approximate computation of $\boldsymbol{\theta}^*$ requires a number of steps of the order of $\log(1/\varepsilon)$ to reach the precision ε, whereas, for reaching the same precision in sampling from π, the LMC algorithm needs a number of iterations proportional to $(p/\varepsilon^2)\log(p/\varepsilon)$. The goal of this section is to explain that this, at first sight very disappointing behavior of the LMC algorithm is, in fact, continuously connected to the exponential convergence of the gradient descent.

The main ingredient for the explanation is that the function $f(\boldsymbol{\theta})$ and the function $f_\tau(\boldsymbol{\theta}) = f(\boldsymbol{\theta})/\tau$ have the same point of minimum $\boldsymbol{\theta}^*$, whatever the real number $\tau > 0$. In addition, if we define the density function $\pi_\tau(\boldsymbol{\theta}) \propto \exp\big(-f_\tau(\boldsymbol{\theta})\big)$, then the average value

$$\bar{\boldsymbol{\theta}}_\tau = \int_{\mathbb{R}^p} \boldsymbol{\theta}\, \pi_\tau(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

tends to the minimum point $\boldsymbol{\theta}^*$ when τ goes to zero. Furthermore, the distribution $\pi_\tau(d\boldsymbol{\theta})$ tends to the Dirac measure at $\boldsymbol{\theta}^*$. Clearly, f_τ satisfies (1) with the constants $m_\tau = m/\tau$ and $M_\tau = M/\tau$. Therefore, on the one hand, we can apply to π_τ claim (a) of Theorem 1, which tells us that if we choose $h = 1/M_\tau = \tau/M$, then

$$W_2(\nu_K, \pi_\tau) \le \Big(1 - \frac{m}{M}\Big)^K W_2(\delta_{\boldsymbol{\theta}_0}, \pi_\tau) + 1.65\Big(\frac{M}{m}\Big)\Big(\frac{p\tau}{M}\Big)^{1/2}. \tag{18}$$

On the other hand, the LMC algorithm with the step-size $h = \tau/M$ applied to f_τ reads as

$$\boldsymbol{\vartheta}_{k+1,h} = \boldsymbol{\vartheta}_{k,h} - \frac{1}{M}\nabla f(\boldsymbol{\vartheta}_{k,h}) + \sqrt{\frac{2\tau}{M}}\,\boldsymbol{\xi}_{k+1}; \qquad k = 0, 1, 2, \ldots \tag{19}$$

When the parameter τ goes to zero, the LMC sequence (19) tends to the gradient descent sequence $\boldsymbol{\theta}_k$. Therefore, the limiting case of (18) corresponding to $\tau \to 0$ writes as

$$\|\boldsymbol{\theta}^{(K)} - \boldsymbol{\theta}^*\|_2 \le \Big(1 - \frac{m}{M}\Big)^K \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2,$$

which is a well-known result in Optimization. This clearly shows that Theorem 1 is a natural extension of the results of convergence from optimization to sampling.

Such an analogy holds true for the Newton method as well. Its counterpart in sampling is the LMCO algorithm. Indeed, one easily checks that if f is replaced by f_τ with τ going to zero, then, for fixed step-size h, the matrix $\boldsymbol{\Sigma}_k$ in (14) tends to zero. This implies that the stochastic perturbation vanishes. On the other hand, the term $\mathbf{M}_{k,\tau}\nabla f_\tau(\boldsymbol{\vartheta}_{k,h}^{\rm LMCO})$ tends to $\{\nabla^2 f(\boldsymbol{\vartheta}_{k,h}^{\rm LMCO})\}^{-1}\nabla f(\boldsymbol{\vartheta}_{k,h}^{\rm LMCO})$, as $\tau \to 0$. Thus, the updates of the Newton algorithm can be seen as the limit case, when τ goes to zero, of the updates of the LMCO.

However, if we replace f by f_τ in the upper bounds stated in Theorem 5 and we let τ go to zero, we do not retrieve the well-known guarantees for the Newtons method. The main reason is that Theorem 5 describes the behavior of the LMCO algorithm in the regime of small step-size h, whereas Newton's method corresponds to (a limit case of) the LMCO with a fixed h. Using arguments similar to those employed in the proof of Theorem 5, one can establish the following result, the proof of which is postponed to Section 7.

Proposition 1. *Let $\nu_K^{\rm LMCO}$ be the distributions of the K-th iterate of the LMCO algorithm (14) with an initial distribution ν_0. Assume that conditions* **F** *and* **N** *are satisfied. Then, for every $h>0$ and $K\in\mathbb{N}$,*

$$W_2(\nu_K^{\rm LMCO},\pi)\le\frac{2m}{M_2}\big(w_K\exp(v_K w_K^{-2^K})\big)^{2^K} \tag{20}$$

with

$$w_K=\frac{M_2W_{2^{K+1}}(\nu_0,\pi)}{2m}+\frac{1}{2}e^{-mh},\quad \textit{and}\quad v_K=\frac{2M_2M^{3/2}\sqrt{2p+2^K}+m^3e^{-mh}}{m^3}.$$

If we replace in the right hand side of (20) the quantities m, M and M_2, respectively, by $m_\tau=m/\tau$, $M_\tau=M/\tau$ and $M_{2,\tau}=M_2/\tau$, and we let τ go to zero, then it is clear that the term v_K vanishes. On the other hand, if ν_0 is the Dirac mass at some point $\boldsymbol{\theta}_0$, then w_K converges to $M_2\|\boldsymbol{\theta}_0-\boldsymbol{\theta}^*\|_2/(2m)$. Therefore, for Newton's algorithm as a limiting case of (20) we get

$$\|\boldsymbol{\theta}_K^{\rm Newton}-\boldsymbol{\theta}^*\|_2\le\frac{2m}{M_2}\bigg(\frac{M_2\|\boldsymbol{\theta}_0-\boldsymbol{\theta}^*\|_2}{2m}\bigg)^{2^K}.$$

The latter provides the so called quadratic rate of convergence, which is a well-known result that can be found in many textbooks; see, for instance, (Chong and Zak, 2013, Theorem 9.1).

There are certainly other interesting connections to uncover between sampling and optimization. One can think of lower bounds for sampling or finding a sampling counterpart of Nesterov acceleration. Some recent advances on the gradient flow (Wibisono et al., 2016) might be useful for achieving these goals.

6. CONCLUSION

We have presented easy-to-use finite-sample guarantees for sampling from a strongly log-concave density using the Langevin Monte-Carlo algorithm with a fixed step-size and extended it to the case where the gradient of the log-density can be evaluated up to some error term. Our results cover both deterministic and random error terms. We have also demonstrated that if the log-density f has a Lipschitz continuous second-order derivative, then one can choose a larger step-size and obtain improved convergence rate.

We have also uncovered some analogies between sampling and optimization. The underlying principle is that an optimization algorithm may be seen as a limit case of a sampling algorithm. Therefore, the results characterizing the convergence of the optimization schemes should have their counterparts for sampling strategies. We have described these analogues for the steepest gradient descent and for the Newton algorithm. However, while in the optimization the relevant characteristics of the problem are the dimension p, the desired accuracy ϵ and the condition number M/m, the problem sampling involves an additional characteristic which is the scale given by the strong-convexity constant m. Indeed, if we increase m by keeping the condition number M/m constant, the number of iterations for the LMC to reach the precision ϵ will decrease. In this respect, we have shown that the LMC with Ozaki discretization, termed LMCO, has a better dependence on the overall scale of f than the original LMC algorithm. However, the weakness of the LMCO is the high computational cost of each iteration. Therefore, we have proposed a new algorithm, LMCO', that improves the LMC in terms of its dependence on the scale and each iteration of LMCO' is computationally much cheaper than each iteration of the LMCO.

Another interesting finding is that, in the case of accurate gradient evaluations (*i.e.*, when there is no error in gradient evaluations), a suitably chosen variable step-size leads to logarithmic improvement in the convergence rate of the LMC algorithm.

Interesting directions for future research are establishing lower bounds in the spirit of those existing in optimization, obtaining user-friendly guarantees for computing the posterior mean or for sampling from a non-smooth density. Some of these problems have already been tackled in several papers mentioned in previous sections, but we believe that the techniques developed in this work might be helpful for revisiting and deepening the existing results.

7. PROOFS

The basis of the proofs of all the theorems stated in previous sections is a recursive inequality that upper bounds the error at the step $k+1$, $W_2(\nu_{k+1},\pi)$, by an expression involving the error of the previous step, $W_2(\nu_k,\pi)$. We will also make repeated use of the Minkowski inequality and its integral version

$$\Big\{\mathbf{E}\Big[\Big(\int_a^b X_t\,dt\Big)^p\Big]\Big\}^{1/p} \le \int_a^b \big\{\mathbf{E}\big[|X_t|^p\big]\big\}^{1/p}\,dt, \qquad \forall p \ge \mathbb{N}^*, \tag{21}$$

where X is a random process almost all paths of which are integrable over the interval $[a,b]$. Furthermore, for any random vector $\boldsymbol{X}$, we define the norm $\|\boldsymbol{X}\|_{L_2} = (\mathbf{E}[\|\boldsymbol{X}\|_2^2])^{1/2}$.

Proposition 2. *Let us introduce the constant* $\varrho_{k+1} = \max(1-mh_{k+1}, Mh_{k+1}-1)$ *(since* $h\in(0,2/M)$*, this value* ϱ *satisfies* $0<\varrho<1$*). If* f *satisfies* (1) *and* $h_{k+1}\le 2/M$*, then*

$$W_2(\nu_{k+1},\pi)^2 \le \big\{\varrho_{k+1}W_2(\nu_k,\pi) + \alpha M(h_{k+1}^3 p)^{1/2} + h_{k+1}\delta\sqrt{p}\big\}^2 + \sigma^2 h_{k+1}^2 p,$$

with $\alpha = 7\sqrt{2}/6 \le 1.65$.

Proof. To simplify notation, and since there is no risk of confusion, we will write h instead of h_{k+1}. Let $\boldsymbol{L}_0$ be a random vector drawn from π such that $W_2(\nu_k,\pi) = \mathbf{E}[\|\boldsymbol{L}_0 - \boldsymbol{\vartheta}_{k,\boldsymbol{h}}\|_2^2]$ and $\mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}},\boldsymbol{L}_0] = \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}]$. Let $\boldsymbol{W}$ be a p-dimensional Brownian Motion independent of $(\boldsymbol{\vartheta}_{k,\boldsymbol{h}},\boldsymbol{L}_0,\boldsymbol{\zeta}_k)$, such that $\boldsymbol{W}_h = \sqrt{h}\,\boldsymbol{\xi}_{k+1}$. We define the stochastic process $\boldsymbol{L}$ so that

$$\boldsymbol{L}_t = \boldsymbol{L}_0 - \int_0^t \nabla f(\boldsymbol{L}_s)\,ds + \sqrt{2}\,\boldsymbol{W}_t, \qquad \forall\, t>0. \tag{22}$$

It is clear that this equation implies that

$$\begin{aligned}\boldsymbol{L}_h &= \boldsymbol{L}_0 - \int_0^h \nabla f(\boldsymbol{L}_s)\,ds + \sqrt{2}\,\boldsymbol{W}_h\\ &= \boldsymbol{L}_0 - \int_0^h \nabla f(\boldsymbol{L}_s)\,ds + \sqrt{2h}\,\boldsymbol{\xi}_{k+1}.\end{aligned}$$

Furthermore, $\{\boldsymbol{L}_t : t \geq 0\}$ is a diffusion process having π as the stationary distribution. Since the initial value $\boldsymbol{L}_0$ is drawn from π, we have $\boldsymbol{L}_t \sim \pi$ for every $t \geq 0$.

Let us denote $\boldsymbol{\Delta}_k = \boldsymbol{L}_0 - \boldsymbol{\vartheta}_{k,\boldsymbol{h}}$ and $\boldsymbol{\Delta}_{k+1} = \boldsymbol{L}_h - \boldsymbol{\vartheta}_{k+1,\boldsymbol{h}}$. We have

$$\begin{aligned}
\boldsymbol{\Delta}_{k+1} &= \boldsymbol{\Delta}_k + h\boldsymbol{Y}_{k,\boldsymbol{h}} - \int_0^h \nabla f(\boldsymbol{L}_t)\,dt \\
&= \boldsymbol{\Delta}_k - h\big(\underbrace{\nabla f(\boldsymbol{\vartheta}_{k,\boldsymbol{h}} + \boldsymbol{\Delta}_k) - \nabla f(\boldsymbol{\vartheta}_{k,\boldsymbol{h}})}_{:=\boldsymbol{U}}\big) + h\boldsymbol{\zeta}_k - \underbrace{\int_0^h \big(\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_0)\big)\,dt}_{:=\boldsymbol{V}}.
\end{aligned}$$

Using the equalities $\mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\Delta}_k, \boldsymbol{U}, \boldsymbol{V}] = \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}, \boldsymbol{L}_0, \boldsymbol{W}] = \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}, \boldsymbol{L}_0] = \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}]$, we get

$$\begin{aligned}
\|\boldsymbol{\Delta}_{k+1}\|_{L_2}^2 &= \big\|\boldsymbol{\Delta}_k - h\boldsymbol{U} - \boldsymbol{V} + h\mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}]\big\|_{L_2}^2 + h^2\big\|\boldsymbol{\zeta}_k - \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}]\big\|_{L_2}^2 \\
&\leq \big\|\boldsymbol{\Delta}_k - h\boldsymbol{U} - \boldsymbol{V} + h\mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}]\big\|_{L_2}^2 + \sigma^2 h^2 p \\
&\leq \big\{\|\boldsymbol{\Delta}_k - h\boldsymbol{U}\|_{L_2} + h\delta\sqrt{p} + \|\boldsymbol{V}\|_{L_2}\big\}^2 + \sigma^2 h^2 p.
\end{aligned}$$

We need now three technical lemmas. The proofs of Lemma 2 and Lemma 3 can be found in (Dalalyan, 2017). Lemma 4 is an improved version of (Dalalyan, 2017, Lemma 3); its proof is postponed to Section 7.6.

LEMMA 2. *It holds that* $\|\boldsymbol{\Delta}_k - h\boldsymbol{U}\|_2 \leq \varrho\|\boldsymbol{\Delta}_k\|_2$.

LEMMA 3. *If the function f is continuously differentiable and the gradient of f is Lipschitz with constant M, then* $\int_{\mathbb{R}^p} \|\nabla f(\boldsymbol{x})\|_2^2\,\pi(\boldsymbol{x})\,d\boldsymbol{x} \leq Mp$.

LEMMA 4. *If the function f has a Lipschitz-continuous gradient with the Lipschitz constant M, $\boldsymbol{L}$ is the Langevin diffusion (22) and $\boldsymbol{V}(a) = \int_a^{a+h} \big(\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_a)\big)\,dt$ for some $a \geq 0$, then*

$$\|\boldsymbol{V}(a)\|_{L_2} \leq \frac{1}{2}\big(h^4 M^3 p\big)^{1/2} + \frac{2}{3}(2h^3 p)^{1/2} M.$$

Using Lemma 2 and Lemma 4, as well as the trivial inequality $W_2(\nu_{k+1}, \pi)^2 \leq \mathbf{E}[\|\boldsymbol{\Delta}_{k+1}\|_2^2]$, we get

$$W_2(\nu_{k+1}, \pi)^2 \leq \big\{\varrho W_2(\nu_k, \pi) + \alpha M(h^3 p)^{1/2} + h\delta\sqrt{p}\big\}^2 + \sigma^2 h^2 p,$$

with $\alpha = 7\sqrt{2}/6 \leq 1.65$. □

7.1 Proof of Theorem 1

Using Proposition 2 with $\sigma = \delta = 0$, we get $W_2(\nu_{k+1}, \pi) \leq \varrho\, W_2(\nu_k, \pi) + \|\boldsymbol{V}\|_{L_2}$ for all $k \in \mathbb{N}$. In view of Lemma 4, this yields

$$W_2(\nu_{k+1}, \pi) \leq \varrho\, W_2(\nu_k, \pi) + \alpha M(h^3 p)^{1/2}.$$

Using this inequality repeatedly for $k+1, k, k-1, \ldots, 1$, we get

$$\begin{aligned} W_2(\nu_{k+1}, \pi) &\le \varrho^{k+1} W_2(\nu_0, \pi) + \alpha M (h^3 p)^{1/2} (1 + \varrho + \ldots + \varrho^k) \\ &\le \varrho^{k+1} W_2(\nu_0, \pi) + \alpha M (h^3 p)^{1/2} (1 - \varrho)^{-1}. \end{aligned}$$

This completes the proof.

7.2 Proof of Theorem 2

Let us denote $\alpha = 7\sqrt{2}/6 \le 1.65$. Theorem 1 implies that using the step-size $h_k = 2/(M + m)$ for $k = 1, \ldots, K_1$, we get

$$\begin{aligned} W_2(\nu_{K_1}, \pi) &\le \Big(1 + \frac{2m}{M-m}\Big)^{-K_1} W_2(\nu_0, \pi) + \frac{\alpha M}{m}\Big(\frac{2p}{m+M}\Big)^{1/2} \\ &\le \frac{3.5M}{m}\Big(\frac{p}{M+m}\Big)^{1/2}. \end{aligned} \tag{23}$$

Starting from this iteration K_1, we use a decreasing step-size

$$h_{k+1} = \frac{2}{M + m + (2/3)m(k - K_1)}.$$

Let us show by induction over k that

$$W_2(\nu_k, \pi) \le \frac{3.5M}{m}\bigg(\frac{p}{M + m + (2/3)m(k-K_1)}\bigg)^{1/2}, \qquad \forall\, k \ge K_1. \tag{24}$$

For $k = K_1$, this inequality is true in view of (23). Assume now that (24) is true for some k. For $k+1$, we have

$$\begin{aligned} W_2(\nu_{k+1}, \pi) &\le (1 - m h_{k+1}) W_2(\nu_k, \pi) + \alpha M \sqrt{p}\, h_{k+1}^{3/2} \\ &\le (1 - m h_{k+1}) \frac{3.5M\sqrt{p}\,(h_{k+1}/2)^{1/2}}{m} + \alpha M \sqrt{p}\, h_{k+1}^{3/2} \\ &\le (1 - \frac{1}{3} m h_{k+1}) \frac{3.5M\sqrt{p}\,(h_{k+1}/2)^{1/2}}{m}. \end{aligned}$$

One can check that

$$\begin{aligned} (1 - \frac{1}{3} m h_{k+1})(h_{k+1}/2)^{1/2} &= \frac{\sqrt{3}\,[m + 3M + 2m(k-K_1)]}{[3m + 3M + 2m(k-K_1)]^{3/2}} \\ &\le \frac{\sqrt{3}\,[m + 3M + 2m(k-K_1)]^{1/2}}{3m + 3M + 2m(k-K_1)} \\ &\le \frac{\sqrt{3}}{[3m + 3M + 2m(k+1-K_1)]^{1/2}}. \end{aligned}$$

This completes the proof of the theorem.

7.3 Proof of Theorem 3

As explained in Section 3, the main new ingredient of the proof is Lemma 1, that has to be combined with Proposition 2. We postpone the proof of Lemma 1 to Section 7.6 and do it in a more general form (see Lemma 7).

In view of Proposition 2, we have

$$W_2(\nu_{k+1},\pi)^2 \le \big\{(1-mh)W_2(\nu_k,\pi) + \alpha M(h^3p)^{1/2} + h\delta\sqrt{p}\big\}^2 + \sigma^2h^2p.$$

We apply now Lemma 1 with $A = mh$, $B = \sigma h\sqrt{p}$ and $C = \alpha M(h^3p)^{1/2} + h\delta\sqrt{p}$, which implies that

$$W_2(\nu_k,\pi) \le (1-mh)^k W_2(\nu_0,\pi) + \frac{\alpha M(hp)^{1/2}+\delta\sqrt{p}}{m} + \frac{\sigma^2 h\sqrt{p}}{\alpha M h^{1/2}+\delta+(mh)^{1/2}\,\sigma}.$$

This completes the proof of the theorem.

7.4 Proof of Theorem 4

Using the same construction and the same definitions as in the proof of Proposition 2, for $\boldsymbol{\Delta}_k = \boldsymbol{L}_0 - \boldsymbol{\vartheta}_{k,\boldsymbol{h}}$, we have

$$\begin{aligned}\boldsymbol{\Delta}_{k+1} &= \boldsymbol{\Delta}_k + h\boldsymbol{Y}_{k,\boldsymbol{h}} - \int_{I_k}\nabla f(\boldsymbol{L}_t)\,dt\\ &= \boldsymbol{\Delta}_k - h\big(\underbrace{\nabla f(\boldsymbol{\vartheta}_{k,\boldsymbol{h}}+\boldsymbol{\Delta}_k) - \nabla f(\boldsymbol{\vartheta}_{k,\boldsymbol{h}})}_{:=\boldsymbol{U}}\big) - \sqrt{2}\underbrace{\int_0^h\int_0^t \nabla^2 f(\boldsymbol{L}_s)d\boldsymbol{W}_s\,dt}_{:=\boldsymbol{S}} + h\boldsymbol{\zeta}_k\\ &\qquad - \underbrace{\int_0^h\Big(\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_0) - \sqrt{2}\int_0^t\nabla^2 f(\boldsymbol{L}_s)d\boldsymbol{W}_s\Big)\,dt}_{:=\bar{\boldsymbol{V}}}\,.\end{aligned}$$

Using the following equalities of conditional expectations $\mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\Delta}_k,\boldsymbol{U},\bar{\boldsymbol{V}}] = \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}},\boldsymbol{L}_0,\boldsymbol{W}] = \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}},\boldsymbol{L}_0] = \mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}]$ and $\mathbf{E}[\boldsymbol{S}_h|\boldsymbol{\vartheta}_{k,\boldsymbol{h}},\boldsymbol{L}_0]=0$, we get

$$\begin{aligned}\|\boldsymbol{\Delta}_{k+1}\|_{L_2}^2 &= \big\|\boldsymbol{\Delta}_k - h\boldsymbol{U} - \bar{\boldsymbol{V}} - \sqrt{2}\boldsymbol{S}_h + h\mathbf{E}[\boldsymbol{\zeta}_k|\boldsymbol{\vartheta}_{k,\boldsymbol{h}}]\big\|_{L_2}^2 + \sigma^2h^2p\\ &\le \big\{\big(\|\boldsymbol{\Delta}_k - h\boldsymbol{U}\|_{L_2}^2 + 2\|\boldsymbol{S}_h\|_{L_2}^2\big)^{1/2} + h\delta\sqrt{p} + \|\bar{\boldsymbol{V}}\|_{L_2}\big\}^2 + \sigma^2h^2p.\end{aligned}$$

In addition, we have

$$\|\boldsymbol{S}_h\|_{L_2}^2 = \bigg\|\int_0^h (h-s)\nabla^2 f(\boldsymbol{L}_s)\,d\boldsymbol{W}_s\bigg\|_{L_2}^2 = \int_0^h (h-s)^2\mathbf{E}[\|\nabla^2 f(\boldsymbol{L}_s)\|_F^2]\,ds \le (1/3)\,M^2h^3p.$$

Setting $x_k = \|\boldsymbol{\Delta}_k\|_{L_2} = W_2(\nu_k,\pi)$ and using Lemma 2, this yields

$$x_{k+1}^2 \le \big\{\big((1-mh)^2x_k^2 + (2/3)\,M^2h^3p\big)^{1/2} + h\delta\sqrt{p} + \|\bar{\boldsymbol{V}}\|_{L_2}\big\}^2 + \sigma^2h^2p.$$

Let us define $A = mh$, $F = (2/3)\,M^2h^3p$, $G = \sigma^2h^2p$ and[3]

$$C = h\delta\sqrt{p} + 0.5M_2h^2p + 0.5M^{3/2}h^2\sqrt{p}.$$

[3] In view of Lemma 6 in Section 7.6, we have $h\delta\sqrt{p} + \|\bar{\boldsymbol{V}}\|_{L_2} \le C$.

Then

$$x_{k+1}^2 \leq \big\{\big((1-A)^2 x_k^2 + F\big)^{1/2} + C\big\}^2 + G.$$

One can deduce from this inequality that $x_{k+1}^2 \leq \big((1-A)x_k + C\big)^2 + F + G + 2C\sqrt{F}$. Therefore, using (41) of Lemma 7 below, we get

$$\begin{aligned} x_k &\leq (1-A)^k x_0 + \frac{C}{A} + \frac{F+G+2C\sqrt{F}}{C + \big(A(F+G+2C\sqrt{F})\big)^{1/2}} \\ &\leq (1-A)^k x_0 + (C/A) + 2(F/A)^{1/2} + \frac{G}{C+\sqrt{AG}}. \end{aligned}$$

Replacing A, C, F and G by their respective expressions, we get the claim of the theorem.

7.5 Proof of Theorem 5

To ease notation, throughout this proof, we will write ν_k and ν_k' instead of ν_k^{LMCO} and $\nu_k^{\text{LMCO}'}$, respectively.

Let $\boldsymbol{D}_0 \sim \nu_k$ and $\boldsymbol{L}_0 \sim \pi$ be two random variables such that $\|\boldsymbol{D}_0 - \boldsymbol{L}_0\|_{L_2}^2 = W_2(\nu_k, \pi)$. Let $\boldsymbol{W}$ be a p-dimensional Brownian motion independent of $(\boldsymbol{D}_0, \boldsymbol{L}_0)$. We define $\boldsymbol{L}$ to be the Langevin diffusion process (22) driven by $\boldsymbol{W}$ and starting at $\boldsymbol{L}_0$, whereas $\boldsymbol{D}$ is the process starting at $\boldsymbol{D}_0$ and satisfying the stochastic differential equation

$$d\boldsymbol{D}_t = -[\nabla f(\boldsymbol{D}_0) + \nabla^2 f(\boldsymbol{D}_0)(\boldsymbol{D}_t - \boldsymbol{D}_0)]\,dt + \sqrt{2}\,d\boldsymbol{W}_t, \quad t \geq 0.$$

This is an Ornstein-Uhlenbeck process. It can be expressed explicitly as a function of $\boldsymbol{D}_0$ and $\boldsymbol{W}$. The corresponding expression implies that $\boldsymbol{D}_h \sim \nu_{k+1}$ and, hence, $W_2(\nu_{k+1}, \pi) \leq \|\boldsymbol{D}_h - \boldsymbol{L}_h\|_{L_2}^2$.

An important ingredient of our proof is the following version of the Gronwall lemma, the proof of which is postponed to Section 7.6.

LEMMA 5. *Let $\boldsymbol{\alpha} : [0,T] \times \Omega \to \mathbb{R}^p$ be a continuous semi-martingale and $\mathbf{H} : [0,T] \times \Omega \to \mathbb{R}^{p\times p}$ be a random process with continuous paths in the space of all symmetric $p \times p$ matrices such that $\mathbf{H}_s\mathbf{H}_t = \mathbf{H}_t\mathbf{H}_s$ for every $s, t \in [0,T]$. If $\boldsymbol{x} : [0,T] \times \Omega \to \mathbb{R}^p$ is a semi-martingale satisfying the identity*

$$\boldsymbol{x}_t = \boldsymbol{\alpha}_t - \int_0^t \mathbf{H}_s \boldsymbol{x}_s\,ds, \qquad \forall t \in [0,T], \tag{25}$$

then, for every $t \in [0,T]$,

$$\boldsymbol{x}_t = \exp\Big\{-\int_0^t \mathbf{H}_s\,ds\Big\}\boldsymbol{\alpha}_0 + \int_0^t \exp\Big\{-\int_s^t \mathbf{H}_u\,du\Big\}d\boldsymbol{\alpha}_s. \tag{26}$$

We denote $\boldsymbol{X}_t = \boldsymbol{L}_t - \boldsymbol{L}_0 - (\boldsymbol{D}_t - \boldsymbol{D}_0)$, where $\boldsymbol{D}_t$ is the random process defined by

$$d\boldsymbol{D}_t = -[\nabla f(\boldsymbol{D}_0) + \nabla^2 f(\boldsymbol{D}_0)(\boldsymbol{D}_t - \boldsymbol{D}_0)]\,dt + \sqrt{2}\,d\boldsymbol{W}_t, \quad \boldsymbol{D}_0 \sim \nu_k, \quad t \in [0,h]$$

and $\boldsymbol{L}_t$ is the Langevin diffusion driven by the same Wiener process $\boldsymbol{W}$ and with initial condition $\boldsymbol{L}_0 \sim \pi$. It is clear that

$$\begin{aligned}
\boldsymbol{X}_t &= -\int_0^t \nabla f(\boldsymbol{L}_s)\,ds + \int_0^t [\nabla f(\boldsymbol{D}_0) + \nabla^2 f(\boldsymbol{D}_0)(\boldsymbol{D}_s - \boldsymbol{D}_0)]\,ds \\
&= -\int_0^t \big\{\nabla f(\boldsymbol{L}_s) - \nabla f(\boldsymbol{D}_0) - \nabla^2 f(\boldsymbol{D}_0)(\boldsymbol{L}_s - \boldsymbol{L}_0)\big\}\,ds - \int_0^t \nabla^2 f(\boldsymbol{D}_0)\boldsymbol{X}_s\,ds.
\end{aligned}$$

Using Lemma 5, we get

$$\begin{aligned}
\boldsymbol{X}_t &= -\int_0^t e^{-s\nabla^2 f(\boldsymbol{D}_0)}\big\{\nabla f(\boldsymbol{L}_s) - \nabla f(\boldsymbol{D}_0) - \nabla^2 f(\boldsymbol{D}_0)(\boldsymbol{L}_s - \boldsymbol{L}_0)\big\}\,ds \\
&= \int_0^t e^{-s\nabla^2 f(\boldsymbol{D}_0)}\,ds[\nabla f(\boldsymbol{D}_0) - \nabla f(\boldsymbol{L}_0)] \\
&\quad - \int_0^t e^{-s\nabla^2 f(\boldsymbol{D}_0)}\big\{\nabla f(\boldsymbol{L}_s) - \nabla f(\boldsymbol{L}_0) - \nabla^2 f(\boldsymbol{L}_0)(\boldsymbol{L}_s - \boldsymbol{L}_0)\big\}\,ds \\
&\quad - \int_0^t e^{-s\nabla^2 f(\boldsymbol{D}_0)}[\nabla^2 f(\boldsymbol{D}_0) - \nabla^2 f(\boldsymbol{L}_0)]\int_0^s \nabla f(\boldsymbol{L}_u)\,du\,ds \\
&\quad + \sqrt{2}\int_0^t e^{-s\nabla^2 f(\boldsymbol{D}_0)}[\nabla^2 f(\boldsymbol{D}_0) - \nabla^2 f(\boldsymbol{L}_0)]\boldsymbol{W}_s\,ds. \qquad (27)
\end{aligned}$$

Let us set $\boldsymbol{\Delta}_t = \boldsymbol{L}_t - \boldsymbol{D}_t$. We have $\boldsymbol{X}_t = \boldsymbol{\Delta}_t - \boldsymbol{\Delta}_0 = A_t - B_t - C_t + S_t$, where A_t, B_t, C_t and S_t stand for the four integrals in (27). We now evaluate these terms separately. For the first one, using the notation $\mathbf{H}_0 = \nabla^2 f(\boldsymbol{D}_0)$ and the identity $\nabla f(\boldsymbol{L}_0) - \nabla f(\boldsymbol{D}_0) = \int_0^1 \nabla^2 f(\boldsymbol{D}_0 + x\boldsymbol{\Delta}_0)\,dx\boldsymbol{\Delta}_0$, we get

$$\begin{aligned}
\|\boldsymbol{\Delta}_0 + A_t\|_2 &\le \|\boldsymbol{\Delta}_0 - t\big(\nabla f(\boldsymbol{L}_0) - \nabla f(\boldsymbol{D}_0)\big)\|_2 + \int_0^t \|\mathbf{I} - e^{-s\mathbf{H}_0}\|\,ds\big\|\nabla f(\boldsymbol{L}_0) - \nabla f(\boldsymbol{D}_0)\big\|_2 \\
&\le (1 - mt + 0.5M^2t^2)\|\boldsymbol{\Delta}_0\|_2. \qquad (28)
\end{aligned}$$

For the term B_t with $t \le h \le m/M^2 \le 1/M$, we can apply (39) to infer that

$$\|B_t\|_{L_2}^2 \le 0.88M_2t^2(p^2 + 2p)^{1/2}. \qquad (29)$$

As for C_t, in view of the inequality $\|\nabla^2 f(\boldsymbol{L}_0) - \nabla^2 f(\boldsymbol{D}_0)\| \le M_2\|\boldsymbol{\Delta}_0\|_2 \wedge M \le \sqrt{MM_2\|\boldsymbol{\Delta}_0\|_2}$, we have

$$\begin{aligned}
\|C_t\|_2 &\le \sqrt{MM_2\|\boldsymbol{\Delta}_0\|_2}\int_0^t\int_0^s \|\nabla f(\boldsymbol{L}_u)\|_2\,du\,ds \\
&\le \mu\|\boldsymbol{\Delta}_0\|_2 + (4\mu)^{-1}MM_2\bigg(\int_0^t (t-u)\|\nabla f(\boldsymbol{L}_u)\|_2\,du\bigg)^2.
\end{aligned}$$

On the other hand, the fact that $\mathbf{E}[\|\nabla f(\boldsymbol{L}_u)\|_2^4] \le M^2(p^2 + 2p)$ yields

$$\bigg(\int_0^t (t-u)(\mathbf{E}[\|\nabla f(\boldsymbol{L}_u)\|_2^4])^{1/4}\,du\bigg)^2 \le \frac{Mt^4(p^2+2p)^{1/2}}{4}. \qquad (30)$$

This implies the inequality

$$\|C_t\|_{L_2} \le \mu W_2(\nu_k, \pi) + (16\mu)^{-1}M^2M_2t^4(p+1). \qquad (31)$$

Finally, using the integration by parts formula for semi-martingales, one can easily write S_t as a stochastic integral with respect to $\boldsymbol{W}$ and derive from that representation the inequality

$$\begin{aligned}\|S_t\|_{L_2}^2 &\le 2\mathbf{E}\bigg[\int_0^t \bigg\|\int_u^t e^{-s\mathbf{H}_0}\,ds\big(\nabla^2 f(\boldsymbol{L}_0)-\nabla^2 f(\boldsymbol{D}_0)\big)\bigg\|_F^2\,du\bigg]\\ &\le 2p\mathbf{E}[(M_2\|\boldsymbol{\Delta}_0\|_2\wedge M)^2]\int_0^t (t-u)^2\,du \le (2/3)M_2Mpt^3\|\boldsymbol{\Delta}_0\|_{L_2}^2. \end{aligned}\tag{32}$$

Putting all these pieces together, taking the expectation, using the Minkowski inequality, the equality $\mathbf{E}[(\boldsymbol{\Delta}_0+A_h)^\top S_h]=0$ and the inequality $\sqrt{a^2+b}\le a+b/(2a)$, we get

$$\begin{aligned}\|\boldsymbol{\Delta}_h\|_{L_2}^2 &= \|\boldsymbol{\Delta}_0+A_h-B_h-C_h+S_h\|_{L_2}^2\\ &\le \big(\|\boldsymbol{\Delta}_0+A_h\|_{L_2}^2+\|S_h\|_{L_2}^2\big)^{1/2}+\|B_h\|_{L_2}^2+\|C_h\|_{L_2}^2\\ &\le \big(1-mh+0.5M^2h^2+\mu\big)\|\boldsymbol{\Delta}_0\|_{L_2}^2+\frac{M_2Mph^3}{3(1-mh+0.5M^2h^2)}\\ &\qquad +0.88M_2h^2(p^2+2p)^{1/2}+\frac{M^2M_2h^4}{16\mu}(p+1). \end{aligned}\tag{33}$$

Let μ be any real number smaller than $0.5h(m-0.5M^2h)$; Eq. (33) and the inequality $p^2+2p\le(p+1)^2$ yield

$$W_2(\nu_{k+1},\pi)\le(1-\mu)W_2(\nu_k,\pi)+\frac{M_2Mph^3}{3(1-2\mu)}+0.88M_2h^2(p+1)+\frac{M^2M_2h^4}{16\mu}(p+1).$$

Since $h\le m/M^2$, we can choose $\mu=0.25mh$ so that $1-2\mu=1-0.5mh\ge 0.5$ and

$$\begin{aligned}W_2(\nu_{k+1},\pi)&\le(1-0.25mh)W_2(\nu_k,\pi)+\frac{2M_2Mph^3}{3}+0.88M_2h^2(p+1)+\frac{M^2M_2h^3}{4m}(p+1)\\ &\le(1-0.25mh)W_2(\nu_k,\pi)+1.8M_2h^2(p+1).\end{aligned}$$

This recursion implies the inequality

$$\begin{aligned}W_2(\nu_k,\pi)&\le(1-0.25mh)^kW_2(\nu_0,\pi)+\frac{1.8M_2h(p+1)}{0.25m}\\ &=(1-0.25mh)^kW_2(\nu_0,\pi)+\frac{7.2M_2h(p+1)}{m}.\end{aligned}$$

This completes the proof of claim (16) of the theorem.

To establish inequality (17), we follow the same steps as in the proof of (16), with a slightly different choice of the process $\boldsymbol{D}$. More precisely, we define $\boldsymbol{D}$ by

$$\boldsymbol{D}_t-\boldsymbol{D}_0=-(t\mathbf{I}_p-0.5t^2\nabla^2 f(\boldsymbol{D}_0))\nabla f(\boldsymbol{D}_0)+\sqrt{2}\int_0^t(\mathbf{I}-(t-u)\nabla^2 f(\boldsymbol{D}_0))\,d\boldsymbol{W}_u.$$

One can check that the conditional distribution of $\boldsymbol{D}_h$ given $\boldsymbol{D}_0=\boldsymbol{x}$ coincides with the conditional distribution of $\boldsymbol{\vartheta}_{k+1,h}^{\mathrm{LMCO}'}$ given $\boldsymbol{\vartheta}_{k,h}^{\mathrm{LMCO}'}=\boldsymbol{x}$. Therefore, if $\boldsymbol{D}_0\sim\nu_k'$, then $\boldsymbol{D}_h\sim\nu_{k+1}'$ and, consequently, $W_2(\nu_{k+1}',\pi)^2\le\mathbf{E}[\|\boldsymbol{D}_h-\boldsymbol{L}_h\|_2^2]$.

To ease notation, we set $\mathbf{H}_0=\nabla^2 f(\boldsymbol{D}_0)$. The process $\boldsymbol{D}$ satisfies the SDE

$$d\boldsymbol{D}_t=-\big[(\mathbf{I}_p-t\nabla^2 f(\boldsymbol{D}_0))\nabla f(\boldsymbol{D}_0)+\sqrt{2}\,\mathbf{H}_0\boldsymbol{W}_t\big]\,dt+\sqrt{2}\,d\boldsymbol{W}_t,$$

which implies that

$$\begin{aligned}d\boldsymbol{D}_t =& - \big[\nabla f(\boldsymbol{D}_0) + \nabla^2 f(\boldsymbol{D}_0)(\boldsymbol{D}_t - \boldsymbol{D}_0)\big]\,dt + \sqrt{2}\,d\boldsymbol{W}_t\\ &- 0.5t^2\mathbf{H}_0^2\nabla f(\boldsymbol{D}_0)\,dt - \sqrt{2}\,\mathbf{H}_0^2\int_0^t (t-u)\,d\boldsymbol{W}_u\,dt.\end{aligned}$$

Proceeding in the same way as for getting (27), we arrive at the decomposition $\boldsymbol{X}_t = \boldsymbol{\Delta}_t - \boldsymbol{\Delta}_0 = A_t - B_t - C_t + S_t - E_t - F_t$, where A_t, B_t, C_t and S_t stand for the four integrals in (27) whereas E_t and F_t are

$$\begin{aligned}E_t &= 0.5\int_0^t e^{-s\mathbf{H}_0}s^2\,ds\,\mathbf{H}_0^2\nabla f(\boldsymbol{D}_0)\\ F_t &= \sqrt{2}\,\mathbf{H}_0^2\int_0^t e^{-s\mathbf{H}_0}\int_0^s (s-u)\,d\boldsymbol{W}_u\,ds.\end{aligned}$$

Using the properties of the stochastic integral, we get

$$\begin{aligned}\mathbf{E}[\|F_h\|_2^2] &= 2\mathbf{E}\Big[\Big\|\mathbf{H}_0^2\int_0^h e^{-s\mathbf{H}_0}\int_0^s (s-u)\,d\boldsymbol{W}_u\,ds\Big\|_2^2\Big]\\ &= 2\mathbf{E}\Big[\Big\|\int_0^h\int_u^h \mathbf{H}_0^2 e^{-s\mathbf{H}_0}(s-u)\,ds\,d\boldsymbol{W}_u\Big\|_2^2\Big]\\ &= 2\int_0^h \Big\|\int_u^h \mathbf{H}_0^2 e^{-s\mathbf{H}_0}(s-u)\,ds\Big\|_F^2\,du\\ &\le 2M^4p\int_0^h\Big(\int_u^h (s-u)\,ds\Big)^2\,du = \frac{M^4h^5p}{10}.\end{aligned} \tag{34}$$

On the other hand,

$$\|E_h\|_2 \le 0.5M^2\int_0^h s^2\,ds\|\nabla f(\boldsymbol{D}_0)\|_2 \le \frac{M^2h^3}{6}\big(\|\nabla f(\boldsymbol{L}_0)\|_2 + M\|\boldsymbol{\Delta}_0\|_2\big),$$

which, in view of Lemma 3, implies that

$$\|E_h\|_{L_2}^2 \le \frac{M^2h^3}{6}\big(\sqrt{Mp} + MW_2(\nu_k', \pi)\big). \tag{35}$$

Proceeding as in (33) and using (30), we get

$$\begin{aligned}\|\boldsymbol{\Delta}_h\|_{L_2} &= \|\boldsymbol{\Delta}_0 + A_h - B_h - C_h + S_h - E_h - F_h\|_{L_2}\\ &\le \|\boldsymbol{\Delta}_0 + A_h + S_h - F_h\|_{L_2} + \|B_h\|_{L_2} + \|C_h\|_{L_2} + \|E_h\|_{L_2}\\ &\le (\|\boldsymbol{\Delta}_0 + A_h\|_{L_2}^2 + \|S_h - F_h\|_{L_2}^2)^{1/2} + \|B_h\|_{L_2} + \|C_h\|_{L_2} + \|E_h\|_{L_2}.\end{aligned} \tag{36}$$

Using the last but one estimate in (32), in conjunction with (34), we get inequalities

$$\|S_h\|_{L_2}^2 \le (2/3)M_2Mh^3pW_2(\nu_k', \pi) \quad\text{and}\quad |\mathbf{E}[S_h^\top F_h]| \le (1/\sqrt{15})M^2M_2h^4pW_2(\nu_k', \pi),$$

which, for $h \le 3m/(4M^2)$, imply that

$$\begin{aligned}\|S_h - F_h\|_{L_2}^2 &\le (2/3)M_2Mh^3pW_2(\nu_k', \pi) + (2/\sqrt{15})M^2M_2h^4pW_2(\nu_k', \pi) + (1/10)M^4h^5p\\ &\le 1.06M_2Mh^3pW_2(\nu_k', \pi) + 0.1M^4h^5p.\end{aligned}$$

Injecting this bound, (28), (29), (31) and (35) in (36), we arrive at

$$\begin{aligned}\|\boldsymbol{\Delta}_h\|_{L_2} &\le \big\{\big[(1-mh+0.5M^2h^2)^2W_2(\nu_k',\pi)^2+1.06M_2Mh^3pW_2(\nu_k',\pi)+0.1M^4h^5p\big\}^{1/2}\\ &\quad+0.88M_2h^2(p+1)+\Big(\mu+\frac{M^3h^3}{6}\Big)W_2(\nu_k',\pi)+\frac{M^2M_2h^4(p+1)}{16\mu}+\frac{M^{5/2}h^3\sqrt{p}}{6}.\end{aligned}$$

In view of the inequality $\sqrt{a^2+b+c}\le\sqrt{a^2+c}+({}^{b}/_{2a})$, the last display leads to

$$\begin{aligned}W_2(\nu_{k+1}',\pi) &\le \big\{\big[(1-mh+0.5M^2h^2)^2W_2(\nu_k',\pi)^2+0.1M^4h^5p\big\}^{1/2}\\ &\quad+\frac{0.53M_2Mh^3p}{1-mh+0.5M^2h^2}+0.88M_2h^2(p+1)+\Big(\mu+\frac{M^3h^3}{6}\Big)W_2(\nu_k',\pi)\\ &\quad+\frac{M^2M_2h^4(p+1)}{16\mu}+\frac{M^{5/2}h^3\sqrt{p}}{6}.\end{aligned}$$

For $h\le 3m/(4M^2)$ and $\mu=0.25mh$, we can use the inequality $1-mh+0.5M^2h^2\ge 17/32$ and simplify the last display as follows:

$$\begin{aligned}W_2(\nu_{k+1}',\pi) &\le \big\{\big[(1-mh+0.5M^2h^2)^2W_2(\nu_k',\pi)^2+0.1M^4h^5p\big\}^{1/2}\\ &\quad+\frac{0.3975M_2h^2(p+1)}{1-mh+0.5M^2h^2}+0.88M_2h^2(p+1)+\Big(\mu+\frac{M^3h^3}{6}\Big)W_2(\nu_k',\pi)\\ &\quad+\frac{3M_2h^2(p+1)}{16}+\frac{M^{5/2}h^3\sqrt{p}}{6}\\ &\le \big\{(1-mh+0.5M^2h^2)^2W_2(\nu_k',\pi)^2+0.1M^4h^5p\big\}^{1/2}\\ &\quad+\Big(0.25mh+\frac{M^3h^3}{6}\Big)W_2(\nu_k',\pi)+1.82M_2h^2(p+1)+\frac{M^{5/2}h^3\sqrt{p}}{6}.\end{aligned}$$

We apply Lemma 9 to the sequence $x_k=W_2(\nu_k',\pi)$ with $A=mh-0.5M^2h^2$ and $D=0.25mh+M^3h^3/6$. For $h\le 3m/(4M^2)$ we have $A-D=0.75mh-0.5M^2h^2-(Mh)^3/6\ge 0.25mh$ and $A+D\le 1.25mh-(3/8)M^2h^2\le 0.727$. This yields

$$\begin{aligned}W_2(\nu_{k+1}',\pi) &\le (1-0.25mh)^kW_2(\nu_0',\pi)+\frac{7.28M_2h(p+1)}{m}+\frac{2M^{5/2}h^2\sqrt{p}}{3m}+\frac{2\sqrt{0.1}\,M^2h^2\sqrt{p}}{\sqrt{1.273m}}\\ &\le (1-0.25mh)^kW_2(\nu_0',\pi)+\frac{7.28M_2h(p+1)}{m}+\frac{1.23M^{5/2}h^2\sqrt{p}}{m}.\end{aligned}$$

This completes the proof of (17) and that of the theorem.

Proof of Proposition 1. Let us denote $\mathbf{M}_k=\int_0^h e^{-s\mathbf{H}_k}\,ds\int_0^1\nabla^2 f(\boldsymbol{D}_{kh}+x\boldsymbol{\Delta}_k)\,dx$. From (27), we have $\boldsymbol{\Delta}_{k+1}=\boldsymbol{\Delta}_k+A_{k,h}+G_{k,h}$ with

$$\begin{aligned}A_{k,h} &= \int_0^h e^{-s\mathbf{H}_k}\,ds\big(\nabla f(\boldsymbol{D}_{kh})-\nabla f(\boldsymbol{L}_{kh})\big)=-\mathbf{M}_k\boldsymbol{\Delta}_k,\\ G_{k,h} &= \int_0^h e^{-s\mathbf{H}_k}\big(\nabla f(\boldsymbol{L}_{kh})-\nabla f(\boldsymbol{L}_s)+\mathbf{H}_k(\boldsymbol{L}_s-\boldsymbol{L}_{kh})\big)\,ds.\end{aligned}$$

Using the fact that

$$\Big\|\int_0^1 \nabla^2 f(\boldsymbol{D}_{kh}+x\boldsymbol{\Delta}_k)\,dx - \mathbf{H}_k\Big\| \le \int_0^1 \big\|\nabla^2 f(\boldsymbol{D}_{kh}+x\boldsymbol{\Delta}_k) - \mathbf{H}_k\big\|\,dx \le \frac{M_2}{2}\,\|\boldsymbol{\Delta}_k\|_2,$$

we get $\|\boldsymbol{\Delta}_k + A_{k,h}\|_2 = \|(\mathbf{I}-\mathbf{M}_k)\boldsymbol{\Delta}_k\|_2 \le \frac{M_2}{2m}\,\|\boldsymbol{\Delta}_k\|_2^2 + e^{-mh}\|\boldsymbol{\Delta}_k\|_2$. This further leads to the recursive inequality

$$\|\boldsymbol{\Delta}_{k+1}\|_2 \le \frac{M_2}{2m}\,\|\boldsymbol{\Delta}_k\|_2^2 + e^{-mh}\|\boldsymbol{\Delta}_k\|_2 + \|G_{k,h}\|_2.$$

In view of the Minkowski inequality, this yields

$$(\mathbf{E}[\|\boldsymbol{\Delta}_{k+1}\|_2^q])^{1/q} \le \frac{M_2}{2m}\,\mathbf{E}[\|\boldsymbol{\Delta}_k\|_2^{2q}]^{1/q} + e^{-mh}\mathbf{E}[\|\boldsymbol{\Delta}_k\|_2^{2q}]^{1/2q} + \mathbf{E}[\|G_{k,h}\|_2^q]^{1/q}. \tag{37}$$

We choose some $K\in\mathbb{N}$ and define the sequence $\{x_0,\dots,x_K\}$ by setting $x_k^{2^{K+1-k}} = \mathbf{E}[\|\boldsymbol{\Delta}_k\|_2^{2^{K+1-k}}]$. Choosing in (37) $q=2^{K-k}$, we get

$$x_{k+1} \le \frac{M_2}{2m}\,x_k^2 + e^{-mh}x_k + \mathbf{E}[\|G_{k,h}\|_2^{2^{K-k}}]^{2^{k-K}},\qquad k=0,1,\dots,K-1.$$

We are in a position to apply Lemma 8 to the sequence $\{x_k\}_{k=0,\dots,K}$. This yields

$$x_K \le \frac{2m}{M_2}\bigg(\frac{M_2x_0}{2m}+\frac{1}{2}e^{-mh}\bigg)^{2^K}\exp\bigg\{2^K\,\frac{M_2\max_k \mathbf{E}[\|G_{k,h}\|_2^{2^K}]^{2^{-K}} + me^{-mh}}{m(\frac{M_2x_0}{2m}+\frac{1}{2}e^{-mh})^{2^K}}\bigg\}, \tag{38}$$

where $\max_k$ is a short notation for $\max_{k=0,1,\dots,K-1}$. It suffices now to upper bound the moments of $\|G_{k,h}\|_2$. We have

$$\begin{aligned}
\mathbf{E}[\|G_{k,h}\|_2^q]^{1/q} &\le M\int_0^h e^{-sm}\big(\mathbf{E}[\|\boldsymbol{L}_{kh+s}-\boldsymbol{L}_{kh}\|_2^q]\big)^{1/q}\,ds\\
&\le M\int_0^h e^{-sm}\Big\{\Big(\mathbf{E}[\|\int_0^s \nabla f(\boldsymbol{L}_{kh+u})\,du\|_2^q]\Big)^{1/q} + \sqrt{2}\big(\mathbf{E}[\|\boldsymbol{W}_s\|_2^q]\big)^{1/q}\Big\}\,ds\\
&\le M\int_0^h e^{-sm}s\,ds\big(\mathbf{E}[\|\nabla f(\boldsymbol{L}_0)\|_2^q]\big)^{1/q} + M\sqrt{2p+q-2}\int_0^s e^{-sm}\sqrt{s}\,ds\\
&\le \frac{M}{m^2}\big(\mathbf{E}[\|\nabla f(\boldsymbol{L}_0)\|_2^q]\big)^{1/q} + \frac{M}{2m^{3/2}}\sqrt{(2p+q-2)\pi}.
\end{aligned}$$

On the other hand, by integration by parts, for every $q\in 2\mathbb{N}$, we have

$$\begin{aligned}
\mathbf{E}[\|\nabla f(\boldsymbol{L}_0)\|_2^q] &= -\int_{\mathbb{R}^p}\|\nabla f(\boldsymbol{x})\|_2^{q-2}\,\nabla f(\boldsymbol{x})^\top d\pi(\boldsymbol{x})\\
&= \sum_{\ell=1}^p\int_{\mathbb{R}^p}\partial_\ell\Big(\|\nabla f(\boldsymbol{x})\|_2^{q-2}\,\partial_\ell f(\boldsymbol{x})\Big)\pi(\boldsymbol{x})\,d\boldsymbol{x}\\
&\le M(p+q-2)\mathbf{E}[\|\nabla f(\boldsymbol{L}_0)\|_2^{q-2}].
\end{aligned}$$

This yields $(\mathbf{E}[\|\nabla f(\boldsymbol{L}_0)\|_2^q])^{1/q} \le \sqrt{M(p+0.5q-1)}$. Combining all these estimates, we arrive at

$$\mathbf{E}[\|G_{k,h}\|_2^q]^{1/q} \le \frac{1.6M^{3/2}\sqrt{2p+q-2}}{m^2}.$$

Combining this inequality with (38) and replacing x_K by $(\mathbf{E}[\|\boldsymbol{\Delta}_K\|_2^2])^{1/2}$, we get

$$(\mathbf{E}[\|\boldsymbol{\Delta}_K\|_2^2])^{1/2} \le \frac{2m}{M_2}\bigg(\frac{M_2x_0}{2m}+\frac{1}{2}e^{-mh}\bigg)^{2^K}\exp\bigg\{2^K\frac{1.6M_2M^{3/2}\sqrt{2p+2^{K-1}-2}+m^3e^{-mh}}{m^3(\frac{M_2x_0}{2m}+\frac{1}{2}e^{-mh})^{2^K}}\bigg\}.$$

This completes the proof of the proposition. □

7.6 Proofs of lemmas

Proofs of Lemma 2 and Lemma 3 can be found in (Dalalyan, 2017). Lemma 4 being an improved version of Lemma 3 from (Dalalyan, 2017), its proof is presented below.

Proof of Lemma 4. Since the process $\boldsymbol{L}$ is stationary, $V(a)$ has the same distribution as $V(0)$. For this reason, it suffices to prove the claim of the lemma for $a = 0$ only. Using the Cauchy-Schwarz inequality and the Lipschitz continuity of f, we get

$$\begin{aligned}\|\boldsymbol{V}(0)\|_{L_2} &= \Big\|\int_0^h \big(\nabla f(\boldsymbol{L}_t)-\nabla f(\boldsymbol{L}_0)\big)\,dt\Big\|_{L_2}\\ &\le \int_0^h \big\|\nabla f(\boldsymbol{L}_t)-\nabla f(\boldsymbol{L}_0)\big\|_{L_2}\,dt\\ &\le M\int_0^h \big\|\boldsymbol{L}_t-\boldsymbol{L}_0\big\|_{L_2}\,dt.\end{aligned}$$

Combining this inequality with the definition of $\boldsymbol{L}_t$, we arrive at

$$\begin{aligned}\|\boldsymbol{V}(0)\|_{L_2} &\le M\int_0^h \Big\|-\int_0^t \nabla f(\boldsymbol{L}_s)\,ds+\sqrt{2}\,\boldsymbol{W}_t\Big\|_{L_2}\,dt\\ &\le M\int_0^h \Big\|\int_0^t \nabla f(\boldsymbol{L}_s)\,ds\Big\|_{L_2}\,dt + M\int_0^h \big\|\sqrt{2}\,\boldsymbol{W}_t\big\|_{L_2}\,dt\\ &\le M\int_0^h\int_0^t \|\nabla f(\boldsymbol{L}_s)\|_{L_2}\,ds\,dt + M\int_0^h \sqrt{2pt}\,dt.\end{aligned}$$

In view of the stationarity of $\boldsymbol{L}_t$, we have $\|\nabla f(\boldsymbol{L}_s)\|_{L_2} = \|\nabla f(\boldsymbol{L}_0)\|_{L_2}$, which leads to

$$\|\boldsymbol{V}(0)\|_{L_2} \le (1/2)Mh^2\big\|\nabla f(\boldsymbol{L}_0)\big\|_{L_2} + (2/3)M\sqrt{2p}\,h^{3/2}.$$

To complete the proof, it suffices to apply Lemma 3. □

Lemma 6. *Let us denote*

$$\begin{aligned}\widetilde{\boldsymbol{V}} &= \int_0^h \big(\nabla f(\boldsymbol{L}_t)-\nabla f(\boldsymbol{L}_0)-\nabla^2 f(\boldsymbol{L}_0)(\boldsymbol{L}_t-\boldsymbol{L}_0)\big)\,dt,\\ \bar{\boldsymbol{V}} &= \int_0^h \Big\{\nabla f(\boldsymbol{L}_t)-\nabla f(\boldsymbol{L}_0)-\sqrt{2}\int_0^t \nabla^2 f(\boldsymbol{L}_s)d\boldsymbol{W}_s\Big\}\,dt,\end{aligned}$$

with f satisfying **Condition F** *and $h \le 1/M$, then*

$$(\mathbf{E}[\|\widetilde{\boldsymbol{V}}\|_2^2])^{1/2} \le 0.877M_2h^2(p^2+2p)^{1/2}, \tag{39}$$

$$\|\bar{\boldsymbol{V}}\|_{L_2} \le (1/2)(M^{3/2}\sqrt{p}+M_2p)h^2. \tag{40}$$

PROOF. We first note that we have

$$\begin{aligned}\|\widetilde{\boldsymbol{V}}\|_2 &\le \int_0^h \Big\| \int_0^1 \big(\nabla^2 f(\boldsymbol{L}_0 + x(\boldsymbol{L}_t - \boldsymbol{L}_0)) - \nabla^2 f(\boldsymbol{L}_0)\big)\, dx (\boldsymbol{L}_t - \boldsymbol{L}_0)\Big\|_2\, dt \\ &\le 0.5 M_2 \int_0^h \|\boldsymbol{L}_t - \boldsymbol{L}_0\|_2^2\, dt.\end{aligned}$$

In view of (21), this implies that $(\mathbf{E}[\|\widetilde{\boldsymbol{V}}\|_2^2])^{1/2} \le 0.5M_2 \int_0^h (\mathbf{E}[\|\boldsymbol{L}_t - \boldsymbol{L}_0\|_2^4])^{1/2}\, dt$. Using the triangle inequality and integration by parts (precise details of the computations are omitted in the interest of saving space), we arrive at

$$\begin{aligned}\mathbf{E}[\|\boldsymbol{L}_t - \boldsymbol{L}_0\|_2^4] &\le \mathbf{E}\Big[\Big\|\int_0^t \nabla f(\boldsymbol{L}_s)\Big\|_2^4\Big] + 12\bigg(\mathbf{E}\Big[\Big\|\int_0^t \nabla f(\boldsymbol{L}_s)\Big\|_2^4\Big]\mathbf{E}[\|\sqrt{2}\boldsymbol{W}_t\|_2^4]\bigg)^{1/2} + 4\mathbf{E}[\|\boldsymbol{W}_t\|_2^4] \\ &\le t^4 M^2 p(2+p) + 12t^3 Mp(2+p) + 4t^2 p(2+p) \\ &= p(2+p)t^2(t^2M^2 + 12tM + 4).\end{aligned}$$

Integrating this inequality, we get

$$\begin{aligned}(\mathbf{E}[\|\widetilde{\boldsymbol{V}}\|_2^2])^{1/2} &\le 0.5M_2(p^2+2p)^{1/2}\int_0^h t(t^2M^2 + 12tM + 4)^{1/2}\, dt \\ &\le \frac{0.5M_2(p^2+2p)^{1/2}}{M^2}\int_0^{Mh} t(t^2 + 12t + 4)^{1/2}\, dt \\ &\le 0.5M_2h^2(p^2+2p)^{1/2} \sup_{x\in(0,2]} \frac{1}{x^2}\int_0^x t(t^2+12t+4)^{1/2}\, dt \\ &= \frac{0.5M_2h^2(p^2+2p)^{1/2}}{4}\int_0^2 t(t^2+12t+4)^{1/2}\, dt \\ &\le 1.16 M_2 h^2 (p^2+2p)^{1/2}.\end{aligned}$$

This completes the proof of (39). To prove (40), we first assume that f is three times continuously differentiable and apply the Ito formula:

$$\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_0) = \int_0^t \nabla^2 f(\boldsymbol{L}_s)\, d\boldsymbol{L}_s + \int_0^t \Delta[\nabla f(\boldsymbol{L}_s)]\, ds.$$

Let us check that $\|\Delta[\nabla f(\boldsymbol{x})]\|_2 = \|\nabla[\Delta f(\boldsymbol{x})]\|_2 \le M_2 p$ for every $\boldsymbol{x} \in \mathbb{R}^p$. Indeed, let us introduce the fanction $g : \mathbb{R}^p \to \mathbb{R}$ defined by $g(\boldsymbol{x}) = \Delta f(\boldsymbol{x}) = \operatorname{tr}[\nabla^2 f(\boldsymbol{x})]$. The third item of condition F implies that $|g(\boldsymbol{x} + t\boldsymbol{u}) - g(\boldsymbol{x})| \le pM_2|t|$ for every $t \in \mathbb{R}$ and every unit vector $\boldsymbol{u} \in \mathbb{R}^p$. Therefore, letting t go to zero, we get $|\boldsymbol{u}^\top \nabla g(\boldsymbol{x})| \le pM_2$ for every unit vector $\boldsymbol{u}$. Choosing $\boldsymbol{u}$ proportional to $\nabla g(\boldsymbol{x})$, we get the inequality $\|\nabla g(\boldsymbol{x})\|_2 = \|\nabla[\Delta f(\boldsymbol{x})]\|_2 \le pM_2$. This leads to

$$\begin{aligned}\|\bar{\boldsymbol{V}}\|_{L_2} &\le \int_0^h \int_0^t \big\|\nabla^2 f(\boldsymbol{L}_s)\nabla f(\boldsymbol{L}_s) - \Delta[\nabla f(\boldsymbol{L}_s)]\big\|_{L^2}\, ds\, dt \\ &\le \int_0^h \int_0^t \big(M\big\|\nabla f(\boldsymbol{L}_s)\big\|_{L^2} + M_2 p\big)\, ds\, dt \\ &= (1/2)(M^{3/2}\sqrt{p} + M_2 p)h^2.\end{aligned}$$

This completes the proof of the lemma in the case of three times continuously differentiable functions f. If f is two-times differentiable with a second-order derivative satisfying the Lipschitz condition, then we can choose an arbitrarily small $\delta > 0$ and apply the previous result to the smoothed function $f_\delta = f * \varphi_\delta$. Here, φ_δ denotes the density of the Gaussian distribution $\mathcal{N}_p(0, \delta^2 \mathbf{I}_p)$ and "$*$" is the convolution operator. The formula $\nabla^2 f_\delta = (\nabla^2 f) * \varphi_\delta$ implies that f_δ satisfies the required smoothness assumptions with the same constants M and M_2 as the function f. Thus, defining $\bar{\boldsymbol{V}}_\delta$ in the same way as $\bar{\boldsymbol{V}}$ with f_δ instead of f, we get

$$\|\bar{\boldsymbol{V}}_\delta\|_{L_2} \le (1/2)(M^{3/2}\sqrt{p} + M_2 p)h^2.$$

On the other hand, setting $g_\delta = f - f_\delta$, we get

$$\begin{aligned}\|\bar{\boldsymbol{V}}_\delta - \bar{\boldsymbol{V}}\|_{L_2} &\le \int_0^h \Big\|\nabla g_\delta(\boldsymbol{L}_t) - \nabla g_\delta(\boldsymbol{L}_0) - \sqrt{2}\int_0^t \nabla^2 g_\delta(\boldsymbol{L}_s) d\boldsymbol{W}_s\Big\|_{L^2} dt \\ &\le \int_0^h \big\|\nabla g_\delta(\boldsymbol{L}_t) - \nabla g_\delta(\boldsymbol{L}_0)\big\|_{L^2} dt + \sqrt{2p}\int_0^h \Big(\int_0^t \mathbf{E}\|\nabla^2 g_\delta(\boldsymbol{L}_s)\|^2 ds\Big)^{1/2} dt.\end{aligned}$$

Using the Lipschitz continuity of ∇f and $\nabla^2 f$, one easily checks that

$$\begin{aligned}\|\nabla g_\delta(\boldsymbol{x})\|_2 &\le \int_{\mathbb{R}^p} \|\nabla f(\boldsymbol{x}-\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|_2 \varphi_\delta(\boldsymbol{y})\, d\boldsymbol{y} \le M\int_{\mathbb{R}^p} \|\boldsymbol{y}\|_2 \varphi_\delta(\boldsymbol{y})\, d\boldsymbol{y} \le M\delta\sqrt{p},\\ \|\nabla^2 g_\delta(\boldsymbol{x})\| &\le \int_{\mathbb{R}^p} \|\nabla^2 f(\boldsymbol{x}-\boldsymbol{y}) - \nabla^2 f(\boldsymbol{x})\| \varphi_\delta(\boldsymbol{y})\, d\boldsymbol{y} \le M_2\int_{\mathbb{R}^p} \|\boldsymbol{y}\|_2 \varphi_\delta(\boldsymbol{y})\, d\boldsymbol{y} \le M_2\delta\sqrt{p}.\end{aligned}$$

This implies that the limit, when δ tends to zero, of $\|\bar{\boldsymbol{V}}_\delta - \bar{\boldsymbol{V}}\|_{L_2}$ is equal to zero. As a consequence,

$$\begin{aligned}\|\bar{\boldsymbol{V}}\|_{L_2} &\le \lim_{\delta\to 0}\big(\|\bar{\boldsymbol{V}}_\delta\|_{L_2} + \|\bar{\boldsymbol{V}}_\delta - \bar{\boldsymbol{V}}\|_{L_2}\big)\\ &\le (1/2)(M^{3/2}\sqrt{p} + M_2 p)h^2 + \lim_{\delta\to 0}\|\bar{\boldsymbol{V}}_\delta - \bar{\boldsymbol{V}}\|_{L_2}\\ &\le (1/2)(M^{3/2}\sqrt{p} + M_2 p)h^2.\end{aligned}$$

This completes the proof of the lemma. □

LEMMA 7. *Let A, B and C be given non-negative numbers such that $A \in (0,1)$. Assume that the sequence of non-negative numbers $\{x_k\}_{k\in\mathbb{N}}$ satisfies the recursive inequality*

$$x_{k+1}^2 \le [(1-A)x_k + C]^2 + B^2$$

for every integer $k \ge 0$. Let us denote

$$\begin{aligned}E &= \frac{(1-A)C + \big\{C^2 + (2A - A^2)B^2\big\}^{1/2}}{2A - A^2} \ge \frac{(1-A)C}{A(2-A)} + \frac{B}{\sqrt{A(2-A)}}\\ D &= \big\{[(1-A)E + C]^2 + B^2\big\}^{1/2} - (1-A)E \le C + \frac{B^2 A}{C + \sqrt{A(2-A)}\,B}\end{aligned}$$

Then

$$x_k \le (1-A)^k x_0 + \frac{D}{A} \le (1-A)^k x_0 + \frac{C}{A} + \frac{B^2}{C + \sqrt{A(2-A)}\,B} \tag{41}$$

for all integers $k \ge 0$.

PROOF. We will repeatedly use the fact that $D = EA$. Let us introduce the sequence y_k defined as follows: $y_0 = x_0 + E$ and

$$y_{k+1} = (1-A)y_k + D, \quad k = 0, 1, 2, \ldots$$

We will first show that $y_k \geq x_k \vee E$ for every $k \geq 0$. This can be done by mathematical induction. For $k = 0$, this claim directly follows from the definition of y_0. Assume that for some k, we have $x_k \leq y_k$ and $y_k \geq E$. Then, for $k+1$, we have

$$\begin{aligned}
x_{k+1} &\leq \big([(1-A)x_k + C]^2 + B^2\big)^{1/2} \\
&\leq \big([(1-A)y_k + C]^2 + B^2\big)^{1/2} \\
&= (1-A)y_k + \big([(1-A)y_k + C]^2 + B^2\big)^{1/2} - (1-A)y_k \\
&\leq (1-A)y_k + \big([(1-A)E + C]^2 + B^2\big)^{1/2} - (1-A)E = y_{k+1}
\end{aligned}$$

and, since $D = EA$, $y_{k+1} = (1-A)y_k + D \geq (1-A)E + EA = E$. Thus, we have checked that the sequence x_k is dominated by the sequence y_k. It remains to establish an upper bound on y_k. This is an easy task since y_k satisfies a first-order linear recurrence relation. We get

$$\begin{aligned}
y_k &= (1-A)^{k-1}y_1 + \sum_{j=0}^{k-2}(1-A)^j D \\
&= (1-A)^{k-1}\Big(x_1 + \frac{D}{A}\Big) + \frac{D}{A}\big(1 - (1-A)^{k-1}\big) \\
&= (1-A)^{k-1}x_1 + \frac{D}{A}.
\end{aligned}$$

This completes the proof of (41). ☐

PROOF OF LEMMA 5. We introduce the process $\boldsymbol{v}_t = -\exp\{\int_0^t \mathbf{H}_u\,du\}\int_0^t \mathbf{H}_s\boldsymbol{x}_s\,ds$. The time derivative of this process satisfies

$$\boldsymbol{v}_t' = -\exp\Big\{\int_0^t \mathbf{H}_u\,du\Big\}\mathbf{H}_t\boldsymbol{\alpha}_t.$$

This implies that $\boldsymbol{v}_t = -\int_0^t \exp\{\int_0^s \mathbf{H}_u\,du\}\mathbf{H}_s\boldsymbol{\alpha}_s\,ds$. Using the definition of $\boldsymbol{v}_t$, we can check that $\int_0^t \mathbf{H}_s\boldsymbol{x}_s\,ds = -\exp\{-\int_0^t \mathbf{H}_u\,du\}\boldsymbol{v}_t = \int_0^t \exp\{-\int_s^t \mathbf{H}_u\,du\}\mathbf{H}_s\boldsymbol{\alpha}_s\,ds$. Substituting this in (25), we get

$$\boldsymbol{x}_t = \boldsymbol{\alpha}_t - \int_0^t \exp\Big\{-\int_s^t \mathbf{H}_u\,du\Big\}\mathbf{H}_s\boldsymbol{\alpha}_s\,ds. \tag{42}$$

On the other hand—using the notation $\mathbf{M}_t = \exp\{\int_0^t \mathbf{H}_u\,du\}$ and the integration by parts formula for semi-martingales—the second integral on the right hand side of (26) can be modified as follows:

$$\begin{aligned}
\int_0^t \exp\Big\{-\int_s^t \mathbf{H}_u\,du\Big\}d\boldsymbol{\alpha}_s &= \mathbf{M}_t^{-1}\int_0^t \mathbf{M}_s d\boldsymbol{\alpha}_s \\
&= \mathbf{M}_t^{-1}\Big(\mathbf{M}_t\boldsymbol{\alpha}_t - \mathbf{M}_0\boldsymbol{\alpha}_0 - \int_0^t d\mathbf{M}_s\,\boldsymbol{\alpha}_s\Big) \\
&= \boldsymbol{\alpha}_t - \exp\Big\{-\int_0^t \mathbf{H}_u\,du\Big\}\boldsymbol{\alpha}_0 - \int_0^t \exp\Big\{-\int_s^t \mathbf{H}_u\,du\Big\}\mathbf{H}_s\boldsymbol{\alpha}_s\,ds.
\end{aligned}$$

Combining this equation with (42), we get the claim of the lemma. ☐

LEMMA 8. *Let A and B be given positive numbers and $\{C_k\}_{k\in\mathbb{N}}$ be a given sequence of real numbers. Assume that the sequence $\{x_k\}_{k\in\mathbb{N}}$ satisfies the recursive inequality*

$$x_{k+1} \le Ax_k^2 + 2Bx_k + C_k, \qquad \forall k \in \mathbb{N}.$$

Then, for all $k \in \mathbb{N}$,

$$x_k \le \frac{1}{A}\big(Ax_0 + B\big)^{2^k} \exp\bigg\{\sum_{j=0}^{k-1} 2^{k-1-j}\,\frac{AC_j + B(1-B)}{(Ax_0+B)^{2^{j+1}}}\bigg\}.$$

PROOF. Let us introduce the sequences $\{y_k\}_{k\in\mathbb{N}}$ and $\{z_k\}_{k\in\mathbb{N}}$ defined by the relations $y_0 = x_0$,

$$\begin{aligned} y_{k+1} &= Ay_k^2 + 2By_k + C_k,\\ z_k &= (Ax_0 + B)^{2^k} \exp\bigg\{\sum_{j=0}^{k-1} 2^{k-1-j}\,\frac{AC_j + B(1-B)}{(Ax_0+B)^{2^{j+1}}}\bigg\}. \end{aligned}$$

Using mathematical induction, one easily shows that inequalities

$$x_k \le y_k \qquad \text{and} \qquad (Ax_0 + B)^{2^k} \le Ay_k + B \le z_k$$

hold for every $k \in \mathbb{N}$. As a consequence, we get

$$x_k \le \frac{Ax_k + B}{A} \le \frac{Ay_k + B}{A} \le \frac{z_k}{A}.$$

This completes the proof of the lemma. $\square$

LEMMA 9. *Let A, B, C, D be positive numbers satisfying $D < A < 1$ and $\{x_k\}_{k\in\mathbb{N}}$ be a sequence of positive numbers satisfying the inequality*

$$x_{k+1} \le \big((1-A)^2 x_k^2 + B^2\big)^{1/2} + C + Dx_k.$$

Then, for every $k \ge 0$, we have

$$x_k \le (1 - A + D)^k x_0 + \frac{C}{A-D} + \frac{B}{\sqrt{(A-D)(2-A-D)}}.$$

PROOF. We start by setting

$$E = \frac{B}{\sqrt{(A-D)(2-A-D)}}, \qquad F = C + (A-D)E$$

and by defining a new sequence $\{y_k\}_{k\in\mathbb{N}}$ by $y_0 = x_0 + E$ and

$$y_{k+1} = (1 - A + D)y_k + F.$$

Our goal is to prove that $y_k \geq x_k \vee E$ for every k. This claim is clearly true for $k = 0$. Let us assume that it is true for the value k and prove its validity for $k + 1$. Since the function $x \mapsto \sqrt{x^2 + a^2} - x$ is decreasing, we have

$$\begin{aligned} x_{k+1} &\leq \sqrt{(1-A)^2 y_k^2 + B^2} + C + D y_k \\ &\leq (1 - A + D) y_k + C + \sqrt{(1-A)^2 y_k^2 + B^2} - (1-A) y_k \\ &\leq (1 - A + D) y_k + C + \sqrt{(1-A)^2 E^2 + B^2} - (1-A) E = y_{k+1}. \end{aligned}$$

On the other hand,

$$\begin{aligned} y_{k+1} &\geq (1 - A + D) y_k + (A - D) E \\ &\geq (1 - A + D) E + (A - D) E = E. \end{aligned}$$

This implies, in particular, that $x_k \leq y_k$ for every $k \in \mathbb{N}$. Since $\{y_k\}$ satisfies a first-order linear recursion, we get $y_k = (1 - A + D)^k y_0 + F(1 - (1 - A + D)^k)/(A - D)$. $\square$

ACKNOWLEDGMENTS

The work of AD was partially supported by the grant Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

REFERENCES

Alfonsi, A., Jourdain, B., and Kohatsu-Higa, A. (2014). Pathwise optimal transport bounds between a one-dimensional diffusion and its euler scheme. *Ann. Appl. Probab.*, 24(3):1049–1080.

Andrieu, C., Ridgway, J., and Whiteley, N. (2016). Sampling normalizing constants in high dimensions using inhomogeneous diffusions. *ArXiv e-prints*.

Bhattacharya, R. N. (1978). Criteria for recurrence and existence of invariant measures for multidimensional diffusions. *Ann. Probab.*, 6(4):541–553.

Bonis, T. (2016). Guarantees in Wasserstein Distance for the Langevin Monte Carlo Algorithm. *ArXiv e-prints*.

Brosse, N., Durmus, A., and Moulines, É. (2017). Normalizing constants of log-concave densities. *ArXiv e-prints*.

Brosse, N., Durmus, A., ric Moulines, and Pereyra, M. (2017). Sampling from a log-concave distribution with compact support with proximal langevin monte carlo. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 319–342, Amsterdam, Netherlands. PMLR.

Bubeck, S., Eldan, R., and Lehec, J. (2015). Sampling from a log-concave distribution with Projected Langevin Monte Carlo. *ArXiv e-prints*.

Cheng, X. and Bartlett, P. (2017). Convergence of Langevin MCMC in KL-divergence. *ArXiv e-prints*.

Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. (2017). Underdamped Langevin MCMC: A non-asymptotic analysis. *ArXiv e-prints*.

Chong, E. and Zak, S. (2013). *An Introduction to Optimization*. Wiley Series in Discrete Mathematics and Optimization. Wiley.

Dalalyan, A. (2017). Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 678–689.

Dalalyan, A. S. (2014). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *ArXiv e-prints*.

Durmus, A. and Moulines, E. (2016). High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *ArXiv e-prints*.

Durmus, A. and Moulines, E. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587.

Durmus, A., Moulines, E., and Pereyra, M. (2016). Sampling from convex non continuously differentiable functions, when Moreau meets Langevin.

Griewank, A. (1993). Some bounds on the complexity of gradients, jacobians, and hessians. In Pardalos, P., editor, *Complexity in Nonlinear Optimization*, pages 128–161. World Scientific publishers.

Huggins, J. and Zou, J. (2017). Quantifying the accuracy of approximate diffusions and Markov chains. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 382–391, Fort Lauderdale, FL, USA. PMLR.

Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.*, 85(2):341–361.

Luu, T. D., Fadili, J., and Chesneau, C. (2017). Sampling from non-smooth distribution through Langevin diffusion. working paper or preprint.

Nagapetyan, T., Duncan, A. B., Hasenclever, L., Vollmer, S. J., Szpruch, L., and Zygalakis, K. (2017). The True Cost of Stochastic Gradient Langevin Dynamics. *ArXiv e-prints*.

Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017). Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703, Amsterdam, Netherlands. PMLR.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(1):255–268.

Roberts, G. O. and Stramer, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.*, 4(4):337–357 (2003).

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

Stramer, O. and Tweedie, R. L. (1999a). Langevin-type models. I. Diffusions with given stationary distributions and their discretizations. *Methodol. Comput. Appl. Probab.*, 1(3):283–306.

Stramer, O. and Tweedie, R. L. (1999b). Langevin-type models. II. Self-targeting candidates for MCMC algorithms. *Methodol. Comput. Appl. Probab.*, 1(3):307–328.

Teh, Y. W., Thiery, A. H., and Vollmer, S. J. (2016). Consistency and fluctuations for stochastic gradient langevin dynamics. *Journal of Machine Learning Research*, 17(7):1–33.

Vollmer, S. J. and Zygalakis, K. C. (2015). (Non-) asymptotic properties of Stochastic Gradient Langevin Dynamics. *ArXiv e-prints*.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 681–688.

Wibisono, A., Wilson, A. C., and Jordan, M. I. (2016). A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358.

Xu, P., Chen, J., and Gu, Q. (2017). Global Convergence of Langevin Dynamics Based Algorithms for Non-convex Optimization. *ArXiv e-prints*.