Bernoulli **26**(3), 2020, 1956–1988
https://doi.org/10.3150/19-BEJ1178

On sampling from a log-concave density using kinetic Langevin diffusions

ARNAK S. DALALYAN[1] and LIONEL RIOU-DURAND[2]

[1]*CREST, ENSAE, Institut Polytechnique de Paris, 5 avenue Le Chatelier, 91129 Palaiseau, France. E-mail: arnak.dalalyan@ensae.fr*
[2]*Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom. E-mail: lionel.riou-durand@warwick.ac.uk*

Langevin diffusion processes and their discretizations are often used for sampling from a target density. The most convenient framework for assessing the quality of such a sampling scheme corresponds to smooth and strongly log-concave densities defined on $\mathbb{R}^p$. The present work focuses on this framework and studies the behavior of the Monte Carlo algorithm based on discretizations of the kinetic Langevin diffusion. We first prove the geometric mixing property of the kinetic Langevin diffusion with a mixing rate that is optimal in terms of its dependence on the condition number. We then use this result for obtaining improved guarantees of sampling using the kinetic Langevin Monte Carlo method, when the quality of sampling is measured by the Wasserstein distance. We also consider the situation where the Hessian of the log-density of the target distribution is Lipschitz-continuous. In this case, we introduce a new discretization of the kinetic Langevin diffusion and prove that this leads to a substantial improvement of the upper bound on the sampling error measured in Wasserstein distance.

Keywords: Hamiltonian Monte Carlo; kinetic Langevin; Langevin algorithm; Markov Chain Monte Carlo; mixing rate

1. Introduction

Markov processes and, more particularly, diffusion processes are often used in order to solve the problem of sampling from a given density π. This problem can be formulated as follows. Assume that we are able to generate an arbitrary number of independent standard Gaussian random variables $\xi_1, \ldots, \xi_K$. For a given precision level $\varepsilon > 0$ and a given metric d on the space of probability measures, the goal is to devise a function F_ε such that the distribution ν_K of the random variable $\vartheta_K = F_\varepsilon(\xi_1, \ldots, \xi_K)$ satisfies $d(\nu_K, \pi) \le \varepsilon$. For solving this task, it is often assumed that we can have access to the evaluations of the probability density function of π as well as its derivatives. Among different functions F_ε having the aforementioned property, the most interesting are those that require the smallest number of computations.

Markov Chain Monte Carlo methods hinge on random variables ϑ_K and associated functions F_ε defined by recursion $\vartheta_k = G_\varepsilon(\vartheta_{k-1}, \xi_k)$, $k = 1, \ldots, K$, where G_ε is some function of two arguments. For a given target distribution π, if one succeeds to design a function G_ε such that the Markov process $\{\vartheta_k; k \in \mathbb{N}\}$ is ergodic with invariant density π then, for large K, the distribution of ϑ_K will be close to π. Therefore, if the evaluation of G_ε involves only simple operations, we get a solution of the task of approximate sampling from π. Of course, it is important to address the problem of the choice of the number of iterations K ensuring that the sampling error

1350-7265 © 2020 ISI/BS

is smaller than ε. However, it is even more important to be able to design functions G_ε, often referred to as the update rule, with desired properties presented above.

Discretization of continuous-time Markov processes is a successful generic method for defining update rules. The idea is to start by specifying a continuous-time Markov process, $\{L_t : t \geq 0\}$, which is provably positive recurrent and has the target π as invariant distribution.[1] The second step is to set-up a suitable time-discretization of the continuous-time process. More precisely, since $\{L_t\}$ is a Markov process, for any step-size $h > 0$, there is a mapping G such that $L_{kh} \overset{\mathscr{D}}{=} G(L_{(k-1)h}, \xi_k)$, $k = 1, \ldots, K$, where ξ_k is a standard Gaussian random variable independent of $L_{(k-1)h}$. This mapping G might not be available in a closed form. Therefore, the last step is to approximate G by a tractable mapping G_ε. Langevin diffusions are a class of continuous-time Markov processes for which the invariant density is available in closed form. For this reason, they are suitable candidates for applying the generic approach of the previous paragraph.

Let m and M be two positive constants such that $m \leq M$. Throughout this work, we will assume that the target distribution π has a density with respect to the Lebesgue measure on $\mathbb{R}^p$, which is of the form $\pi(\boldsymbol{\theta}) = Ce^{-f(\boldsymbol{\theta})}$ for a function f that is m-strongly convex and with a M-Lipschitz gradient. The (highly overdamped) Langevin diffusion having π as invariant distribution is defined as a strong solution to the stochastic differential equation

$$d\boldsymbol{L}_t = -\nabla f(\boldsymbol{L}_t)\, dt + \sqrt{2}\, d\boldsymbol{W}_t, \quad t \geq 0, \tag{1}$$

where $\boldsymbol{W}$ is a p-dimensional standard Brownian motion. The update rule associated to this process, obtained by using the Euler discretization, is given by the equation $G_\varepsilon(\boldsymbol{L}_{(k-1)h}, \boldsymbol{\xi}_k) = -h\nabla f(\boldsymbol{L}_{(k-1)h}) + \sqrt{2h}\,\boldsymbol{\xi}_k$ with $\boldsymbol{\xi}_k \overset{\mathscr{D}}{=} h^{-1/2}(\boldsymbol{W}_{kh} - \boldsymbol{W}_{(k-1)h})$ being a p-dimension standard Gaussian vector. The resulting approximate sampling method is often called the Langevin Monte Carlo (LMC) or Unadjusted Langevin Algorithm (ULA). Its update rule follows from (1) by replacing the function $t \mapsto \nabla f(\boldsymbol{L}_t)$ by its piecewise constant approximation. Therefore, the behavior of the LMC is governed by the following two characteristics of the continuous-time process: the mixing rate and the smoothness of the sample paths. A quantitative bound on the mixing rate allows us to choose a time horizon T such that the distribution of the random vector $\boldsymbol{L}_T$ is within a distance $\varepsilon/2$ of the target distribution, whereas the smoothness of sample paths helps us to design a step-size h so that the distribution of the discretized process at $K = T/h$ is within a distance $\varepsilon/2$ of the distribution of $\boldsymbol{L}_T$. For the LMC, we know that the Langevin diffusion mixes exponentially fast with the precise rate e^{-mt}. In addition, almost all sample paths of $\boldsymbol{L}$ are Hölder continuous of degree α, for every $\alpha < 1/2$. Combining these properties, it has been shown that it suffices $K_\varepsilon = O((p/\varepsilon^2)\log(p/\varepsilon^2))$ iterations for the LMC algorithm to achieve an error smaller than ε (both in total-variation and Wasserstein distances); see (Dalalyan [15]) for the first nonasymptotic result of this type and (Dalalyan and Karagulyan [16], Durmus and Moulines [22,23]) for improved versions of it.

Under the same assumptions on the log-target f, one can consider the kinetic Langevin diffusion, also known as the second-order Langevin process, defined by

$$d\begin{bmatrix}\boldsymbol{V}_t \\ \boldsymbol{L}_t\end{bmatrix} = \begin{bmatrix}-\big(\gamma \boldsymbol{V}_t + u\nabla f(\boldsymbol{L}_t)\big) \\ \boldsymbol{V}_t\end{bmatrix} dt + \sqrt{2\gamma u}\begin{bmatrix}\mathbf{I}_p \\ \mathbf{0}_{p\times p}\end{bmatrix} d\boldsymbol{W}_t, \quad t \geq 0, \tag{2}$$

[1] More generally, one can consider a Markov process having an invariant distribution that is close to π.

where $\gamma > 0$ is the friction coefficient and $u > 0$ is the inverse mass. As proved in (Nelson [32], Theorem 10.1), the highly overdamped Langevin diffusion (1) is obtained as a limit of the rescaled kinetic diffusion $\bar{\boldsymbol{L}}_t = \boldsymbol{L}_{\gamma t}$, where $\boldsymbol{L}$ is defined as in (2) with $u = 1$, when the friction coefficient γ tends to infinity.

The continuous-time Markov process $(\boldsymbol{L}_t, \boldsymbol{V}_t)$ is positive recurrent and its invariant distribution is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{2p}$. The corresponding invariant density is given by

$$p_*(\boldsymbol{\theta}, \boldsymbol{v}) \propto \exp\left\{-f(\boldsymbol{\theta}) - \frac{1}{2u}\|\boldsymbol{v}\|_2^2\right\}, \quad \boldsymbol{\theta} \in \mathbb{R}^p, \boldsymbol{v} \in \mathbb{R}^p.$$

This means that under the invariant distribution, the components $\boldsymbol{L}$ and $\boldsymbol{V}$ are independent, $\boldsymbol{L}$ is distributed according to the target π, whereas $\boldsymbol{V}/\sqrt{u}$ is a standard Gaussian vector. Therefore, one can use this process for solving the problem of sampling from π. As discussed above, the quality of the resulting sampler will depend on two key properties of the process: rate of mixing and smoothness of sample paths. The rate of mixing of kinetic diffusions has been recently studied by Eberle et al. [25] under conditions that are more general than strong convexity of f. In strongly convex case, a more tractable result has been obtained by Cheng et al. [12]. It establishes that for $\gamma = 2$ and $u = 1/M$, the mixing rate in the Wasserstein distance is $e^{-(m/2M)t}$; see Theorem 5 in (Cheng et al. [12]). On the other hand, sample paths of the process $\{\boldsymbol{L}\}$ defined in (2) are smooth of order $1 + \alpha$, for every $\alpha \in [0, 1/2[$. Combining these two properties, (Cheng et al. [12]) prove that a suitable discretization of (2) leads to a sampler that achieves an error smaller than ε in a number of iterations K satisfying $K = O((p/\varepsilon^2)^{1/2}\log(p/\varepsilon))$.

It follows from the discussion of previous paragraphs that the kinetic LMC based on (2) converges faster than the standard LMC based on (1). Furthermore, this improved rate of convergence is mainly due to the higher smoothness of sample paths of the underlying Markov process. The main purpose of the present work is to pursue the investigation of the kinetic Langevin Monte Carlo (KLMC) initiated in (Cheng et al. [12]) by addressing the following questions:

- Q1. What is the rate of mixing of the continuous-time kinetic Langevin diffusion for general values of the parameters u and γ?
- Q2. Is it possible to improve the rate of convergence of the KLMC by optimizing it over the choice of u, γ and the step-size?
- Q3. If the function f happens to have a Lipschitz-continuous Hessian, is it possible to devise a discretization that takes advantage of this additional smoothness and leads to improved rates of convergence?

The rest of the paper is devoted to answering these questions. The rate of mixing for the continuous-time process is discussed in Section 2. In a nutshell, we show that if $\gamma \geq \sqrt{(M+m)u}$, then the rate of mixing is at least of order $e^{-(um/\gamma)t}$. Nonasymptotic guarantees for the KLMC algorithm are stated and discussed in Section 3. They are in the same spirit as those established in (Cheng et al. [12]), but have an improved dependence on the condition number, the ratio of the Lipschitz constant M and the strong convexity constant m. Our result has also improved constants and is much less sensitive to the choice of the initial distribution. These improvements are achieved thanks to a more careful analysis of the discretization error of the

Langevin process. Finally, we present in Section 4 a new discretization, termed second-order KLMC, of the kinetic Langevin diffusion that exploits the knowledge of the Hessian of f. Its error, measured in the Wasserstein distance W_2 is shown to be bounded by ε for a number of iterations that scales as $(p/\varepsilon)^{1/2}$. Thus, we get an improvement of order $(1/\varepsilon)^{1/2}$ over the first-order KLMC algorithm.

2. Mixing rate of the kinetic Langevin diffusion

Let us denote by $\mathbf{P}_t^{\boldsymbol{L}}$ the transition probability at time t of the kinetic diffusion $\boldsymbol{L}$ defined by (2). This means that $\mathbf{P}_t^{\boldsymbol{L}}$ is a Markov kernel given by $\mathbf{P}_t^{\boldsymbol{L}}((\boldsymbol{x},\boldsymbol{v}),B) = \mathbf{P}(\boldsymbol{L}_t \in B | \boldsymbol{V}_0 = \boldsymbol{v}, \boldsymbol{L}_0 = \boldsymbol{x})$, for every $\boldsymbol{v},\boldsymbol{x} \in \mathbb{R}^p$ and any Borel set $B \subset \mathbb{R}^p$. For any probability distribution μ on $\mathbb{R}^p \times \mathbb{R}^p$, we denote $\mu\mathbf{P}_t^{\boldsymbol{L}}$ the (unconditional) distribution of the random variable $\boldsymbol{L}_t$ when the starting distribution of the process $(\boldsymbol{V},\boldsymbol{L})$ is μ (i.e., when $(\boldsymbol{V},\boldsymbol{L}_0) \sim \mu$).

Since the process $(\boldsymbol{V},\boldsymbol{L})$ is ergodic, whatever the initial distribution, for large values of t the distribution of $\boldsymbol{L}_t$ is close to the invariant distribution. We want to quantify how fast does this convergence occur. Furthermore, we are interested in a nonasymptotic result in the Wasserstein–Kantorovich distance W_2, valid for a large set of possible values (γ,u).

A first observation is that, without loss of generality, we can focus our attention to the case $u = 1$. This is made formal in the next lemma.

Lemma 1. *Let $(\boldsymbol{V},\boldsymbol{L})$ be the kinetic Langevin diffusion defined by (2). The modified process $(\bar{\boldsymbol{V}}_t,\bar{\boldsymbol{L}}_t) = (u^{-1/2}\boldsymbol{V}_{t/\sqrt{u}},\boldsymbol{L}_{t/\sqrt{u}})$ is an kinetic Langevin diffusion as well with associated parameters $\bar{\gamma} = \gamma/\sqrt{u}$ and $\bar{u} = 1$.*

The proof of this result is straightforward and, therefore, is omitted. Note that it shows that the parameter u merely represents a time scale (the speed of running over the path of the process $\boldsymbol{L}$). Therefore, in the rest of this paper, we will consider the parameter u to be equal to 1.

Theorem 1. *Assume that the function f is twice differentiable with a Hessian matrix $\nabla^2 f$ satisfying $m\mathbf{I}_p \preceq \nabla^2 f(\boldsymbol{x}) \preceq M\mathbf{I}_p$ for every $\boldsymbol{x} \in \mathbb{R}^p$. Let μ_1, μ_2 and μ_2' be three probability measures on $\mathbb{R}^p$. Let us define the product measures $\mu = \mu_1 \otimes \mu_2$ and $\mu' = \mu_1 \otimes \mu_2'$. For every $\gamma, t > 0$, there exist numbers $\alpha \le \sqrt{2}/\gamma$ and $\beta \ge \{m \wedge (\gamma^2 - M)\}/\gamma$ such that*

$$W_2\big(\mu\mathbf{P}_t^{\boldsymbol{L}},\mu'\mathbf{P}_t^{\boldsymbol{L}}\big) \le \alpha e^{-\beta t} W_2\big(\mu,\mu'\big). \tag{3}$$

More precisely, for every $v \in [0,\gamma/2[$, we have[2]

$$W_2\big(\mu\mathbf{P}_t^{\boldsymbol{L}},\mu'\mathbf{P}_t^{\boldsymbol{L}}\big) \le \frac{\sqrt{2((\gamma-v)^2+v^2)}}{\gamma-2v} \exp\left\{\frac{(v^2-m)\vee(M-(\gamma-v)^2)}{\gamma-2v}t\right\} W_2\big(\mu,\mu'\big). \tag{4}$$

[2]One can observe that (3) can be deduced from (4) by taking $v = 0$.

Table 1. Lower bounds on the rates of contraction of the distribution of the kinetic Langevin diffusion $\boldsymbol{L}_t$ for $u = 1$ and varying γ. The reported values are obtained by optimizing the bound in Theorem 1 with respect to v. In the overdamped case $\gamma^2 \geq 3m + M$, the obtained rates coincide with those that can be directly computed for quadratic functions f and, therefore, are optimal

$\gamma^2 \in$	$]0, M]$	$]M, m+M]$	$[m+M, 3m+M[$	$[3m+M, +\infty[$
Rate of contraction, β	NA	$\frac{\gamma^2-M}{\gamma}$	$\frac{\gamma}{2} - \frac{M-m}{2\sqrt{2(m+M)-\gamma^2}}$	$\frac{\gamma-\sqrt{\gamma^2-4m}}{2}$
Obtained by Theorem 1 with	–	$v = 0$	$v = \frac{\gamma-\sqrt{2(m+M)-\gamma^2}}{2}$	$v = \frac{\gamma-\sqrt{\gamma^2-4m}}{2}$

The proof of this result is postponed to Section 7. Here, we will discuss some consequences of it and present the main ingredient of the proof. First of all, note that this result implies that for $\gamma^2 > 2 \vee M$, the operator $\mathbf{P}_t^{\boldsymbol{L}}$ is a contraction. The rate of this contraction is characterized by the parameter β. If we optimize the exponent in (4) with respect to v, we get the lower bounds on the rates of contraction reported in Table 1.

If we consider the case $\gamma = 2\sqrt{Mu} = 2\sqrt{M}$ previously studied in (Cheng et al. [12]), then the best lower bound on the contraction rate provided by (4) corresponds to $v = \sqrt{M} - \sqrt{M-m}$, and the upper bound of Theorem 1 reads as

$$W_2\big(\mu\mathbf{P}_t^{\boldsymbol{L}}, \mu'\mathbf{P}_t^{\boldsymbol{L}}\big) \leq \left(\frac{2M-m}{M-m}\right)^{1/2} \exp\big\{-(\sqrt{M} - \sqrt{M-m})t\big\} W_2\big(\mu, \mu'\big). \tag{5}$$

One can check that the constant $\sqrt{M} - \sqrt{M-m}$ that we obtain within the exponential is optimal, in the sense that one gets exactly this constant in the case where f is the bivariate quadratic function $f(x_1, x_2) = (m/2)x_1^2 + (M/2)x_2^2$. This constant is slightly better than the one obtained in (Cheng et al. [12], Lemma 8) for the particular choice of the time scale $u = 1/M$. Indeed, if we rewrite the two results in the common time-scale $u = 1$, (Cheng et al. [12], Lemma 8) provides the contraction rate $\beta = m/(2\sqrt{M})$, which is smaller than (but asymptotically equivalent to) $\sqrt{M} - \sqrt{M-m}$.

Another relevant consequence is obtained by instantiating (3) to the case $\gamma \geq \sqrt{M+m}$. This leads to the bound

$$\gamma \geq \sqrt{M+m} \quad \Longrightarrow \quad W_2\big(\mu\mathbf{P}_t^{\boldsymbol{L}}, \mu'\mathbf{P}_t^{\boldsymbol{L}}\big) \leq \sqrt{2}\exp\big\{-(m/\gamma)t\big\} W_2\big(\mu, \mu'\big).$$

An appealing feature of this result is that we can optimize the argument of the exponent with respect to γ for fixed t. The corresponding optimized constant is $m/\sqrt{M+m}$, which improves on the constant obtained in (5) for $\gamma = 2\sqrt{M}$. When M/m becomes large, the improvement factor gets close to 2.

We now describe the main steps of the proof of Theorem 1. The main idea is to consider along with the process $(\boldsymbol{V}, \boldsymbol{L})$, another process $(\boldsymbol{V}', \boldsymbol{L}')$ that satisfies the same SDE (2) as $(\boldsymbol{V}, \boldsymbol{L})$, with the same Brownian motion but with different initial conditions. One easily checks that

$$d\begin{bmatrix}\boldsymbol{V}_t - \boldsymbol{V}'_t \\ \boldsymbol{L}_t - \boldsymbol{L}'_t\end{bmatrix} = \begin{bmatrix}-\big(\gamma\big(\boldsymbol{V}_t - \boldsymbol{V}'_t\big) + \nabla f(\boldsymbol{L}_t) - \nabla f\big(\boldsymbol{L}'_t\big)\big) \\ \boldsymbol{V}_t - \boldsymbol{V}'_t\end{bmatrix} dt, \quad t \geq 0. \tag{6}$$

Using the mean value theorem, we infer that for a suitable symmetric matrix $\mathbf{H}_t$, we have $\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}'_t) = \mathbf{H}_t(\boldsymbol{L}_t - \boldsymbol{L}'_t)$. Furthermore, $\mathbf{H}_t$ being the Hessian of a strongly convex function satisfies $\mathbf{H}_t \succeq m\mathbf{I}_p$. Then (6) can be rewritten as

$$\frac{d}{dt}\begin{bmatrix}\boldsymbol{V}_t - \boldsymbol{V}'_t \\ \boldsymbol{L}_t - \boldsymbol{L}'_t\end{bmatrix} = \begin{bmatrix}-\gamma\mathbf{I}_p & -\mathbf{H}_t \\ \mathbf{I}_p & \mathbf{0}_{p\times p}\end{bmatrix}\begin{bmatrix}\boldsymbol{V}_t - \boldsymbol{V}'_t \\ \boldsymbol{L}_t - \boldsymbol{L}'_t\end{bmatrix}, \quad t \geq 0. \tag{7}$$

In a small neighborhood of any fixed time instance t_0, (7) is close to a linear differential equation with the associated matrix

$$\mathbf{M}(t_0) = \begin{bmatrix}-\gamma\mathbf{I}_p & -\mathbf{H}_{t_0} \\ \mathbf{I}_p & \mathbf{0}_{p\times p}\end{bmatrix}.$$

It is well known that the solution of such a differential equation will tend to zero if and only if the real parts of all the eigenvalues of $\mathbf{M}(t_0)$ are negative. The matrix $\mathbf{M}(t_0)$ is not symmetric; it is in most cases diagonalizable but its eigenvectors generally depend on t_0. To circumvent this difficulty, we determine the transformations diagonalizing the surrogate matrix

$$\mathbf{M} = \begin{bmatrix}-\gamma\mathbf{I}_p & -v^2\mathbf{I}_p \\ \mathbf{I}_p & \mathbf{0}_{p\times p}\end{bmatrix}, \quad \text{for some } v \in [0, \gamma/2[.$$

This yields an invertible matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{M}\mathbf{P}$ is diagonal. We can thus rewrite (7) in the form

$$\frac{d}{dt}\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_t - \boldsymbol{V}'_t \\ \boldsymbol{L}_t - \boldsymbol{L}'_t\end{bmatrix} = \{\mathbf{P}^{-1}\mathbf{M}(t)\mathbf{P}\}\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_t - \boldsymbol{V}'_t \\ \boldsymbol{L}_t - \boldsymbol{L}'_t\end{bmatrix}, \quad t \geq 0.$$

Interestingly, we prove that the quadratic form associated with the matrix $\mathbf{P}^{-1}\mathbf{M}(t)\mathbf{P}$ is negative definite and this provides the desired result. Furthermore, we use this same matrix $\mathbf{P}$ (corresponding to $v = 0$) for analyzing the discretized version of the kinetic Langevin diffusion and proving the main result of the next section.

3. Error bound for the KLMC in Wasserstein distance

Let us start this section by recalling the KLMC algorithm, the sampler derived from a suitable time-discretization of the kinetic diffusion, introduced by Cheng et al. [12]. Let us define the sequence of functions ψ_k by $\psi_0(t) = e^{-\gamma t}$ and $\psi_{k+1}(t) = \int_0^t \psi_k(s)\,ds$. Recall that f is assumed twice differentiable and, without loss of generality, the parameter u is assumed to be equal to one. The discretization involves a step-size $h > 0$ and is defined by the following recursion:

$$\begin{bmatrix}\boldsymbol{v}_{k+1} \\ \boldsymbol{\vartheta}_{k+1}\end{bmatrix} = \begin{bmatrix}\psi_0(h)\boldsymbol{v}_k - \psi_1(h)\nabla f(\boldsymbol{\vartheta}_k) \\ \boldsymbol{\vartheta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)\nabla f(\boldsymbol{\vartheta}_k)\end{bmatrix} + \sqrt{2\gamma}\begin{bmatrix}\boldsymbol{\xi}_{k+1} \\ \boldsymbol{\xi}'_{k+1}\end{bmatrix}, \tag{8}$$

where $(\boldsymbol{\xi}_{k+1}, \boldsymbol{\xi}'_{k+1})$ is a $2p$-dimensional centered Gaussian vector satisfying the following conditions:

- $(\boldsymbol{\xi}_j, \boldsymbol{\xi}'_j)$'s are iid and independent of the initial condition $(\boldsymbol{v}_0, \boldsymbol{\vartheta}_0)$,
- for any fixed j, the random vectors $((\boldsymbol{\xi}_j)_1, (\boldsymbol{\xi}'_j)_1)$, $((\boldsymbol{\xi}_j)_2, (\boldsymbol{\xi}'_j)_2)$, …, $((\boldsymbol{\xi}_j)_p, (\boldsymbol{\xi}'_j)_p)$ are iid with the covariance matrix

$$\mathbf{C} = \int_0^h \big[\psi_0(t)\psi_1(t)\big]^\top \big[\psi_0(t)\psi_1(t)\big]\, dt.$$

This recursion may appear surprising, but one can check that it is obtained by first replacing in (2), on each time interval $t \in [kh, (k+1)h]$, the gradient $\nabla f(\boldsymbol{L}_t)$ by $\nabla f(\boldsymbol{L}_{kh})$, by renaming $(\boldsymbol{V}_{kh}, \boldsymbol{L}_{kh})$ into $(\boldsymbol{v}_k, \boldsymbol{\vartheta}_k)$ and by explicitly solving the obtained linear SDE (which leads to an Ornstein–Uhlenbeck process). To the best of our knowledge, the algorithm (8), that we will refer to as KLMC, has been first proposed by Cheng et al. [12]. The next result characterizes its approximation properties.

Theorem 2. *Assume that the function f is twice differentiable with a Hessian matrix $\nabla^2 f$ satisfying $m\mathbf{I}_p \preceq \nabla^2 f(\boldsymbol{x}) \preceq M\mathbf{I}_p$ for every $\boldsymbol{x} \in \mathbb{R}^p$. In addition, let the initial condition of the KLMC algorithm be drawn from the product distribution $\mu = \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p) \otimes \nu_0$. For every $\gamma \geq \sqrt{m+M}$ and $h \leq m/(4\gamma M)$, the distribution ν_k of the kth iterate $\boldsymbol{\vartheta}_k$ of the KLMC algorithm (8) satisfies*

$$W_2(\nu_k, \pi) \leq \sqrt{2}\left(1 - \frac{0.75mh}{\gamma}\right)^k W_2(\nu_0, \pi) + \frac{Mh\sqrt{2p}}{m}.$$

The proof of this theorem, postponed to Section 8, is inspired by the proof in (Cheng et al. [12]), but with a better control of the discretization error. This allows us to achieve the following improvements as compared to aforementioned paper:

- The second term in the upper bound provided by Theorem 2 scales linearly as a function of the condition number $\varkappa \triangleq M/m$, whereas the corresponding term in (Cheng et al. [12]) scales as $\varkappa^{3/2}$.
- The impact of the initial distribution ν_0 on the overall error of sampling appears only in the first term, which is multiplied by a sequence that has an exponential decay in k. As a consequence, if we denote by K the number of iterations sufficient for the error to be smaller than a prescribed level ε, our result leads to an expression of K in which $W_2(\nu_0, \pi)$ is within a logarithm. Recall that the expression of K in (Cheng et al. [12], Theorem 1) scales linearly in $W_2(\nu_0, \pi)$.
- The numerical constants of Theorem 2 are much smaller than those of the corresponding result in (Cheng et al. [12]).

In order to ease the comparison of our result to (Cheng et al. [12], Theorem 1), let us apply Theorem 2 to

$$h = \frac{m}{4M\sqrt{m+M}} \wedge \frac{0.94\varepsilon}{\varkappa\sqrt{2p}} \tag{9}$$

and $\gamma = \sqrt{m+M}$, which corresponds to the tightest upper bound furnished by our theorem. Note that in (Cheng et al. [12]) it is implicitly assumed that p/ε^2 is large enough so that the

second term in the minimum appearing in (9) is smaller than the first term. From (9), we obtain that[3]

$$K_{\mathrm{KLMC}} \geq \frac{\sqrt{m+M}}{0.75m}\left(\frac{4M\sqrt{m+M}}{m} \vee \frac{\varkappa\sqrt{2p}}{0.94\varepsilon}\right)\log\left(\frac{24W_2(\nu_0,\pi)}{\varepsilon}\right) \tag{10}$$

iterations are sufficient for having $W_2(\nu_K,\pi) \leq \varepsilon$. After some simplifications, we get

$$K_{\mathrm{KLMC}} \geq 3\varkappa^{3/2}\left\{(16\varkappa) \vee \frac{p}{m\varepsilon^2}\right\}^{1/2}\log\left(\frac{24W_2(\nu_0,\pi)}{\varepsilon}\right). \tag{11}$$

Remind that the corresponding result in (Cheng et al. [12]) requires K to satisfy[4]

$$K \geq 52\varkappa^2\left\{\frac{p}{m\varepsilon^2}\right\}^{1/2}\log\left(\frac{24W_2(\nu_0,\pi)}{\varepsilon}\right).$$

Thus, the improvement in terms of the number of iterations we obtain is at least by a factor $17\sqrt{\varkappa}$, whenever $\kappa \leq p/(16m\varepsilon^2)$.

It is also helpful to compare the obtained result (11) to the analogous result for the highly overdamped Langevin diffusion (Durmus and Moulines [23]). Using (Durmus et al. [21], equation (22)), one can check that this is enough to choose an integer

$$K_{\mathrm{LMC}} \geq 2\varkappa\left\{1 \vee \frac{2.18p}{m\varepsilon^2}\right\}\log\left(\frac{24W_2(\nu_0,\pi)}{\varepsilon}\right), \tag{12}$$

such that K_{LMC} iterations of the LMC algorithm are sufficient to arrive at an error bounded by ε. Comparing (11) and (12), we see that the KLMC achieves a smaller error bound than LMC when $p/(m\varepsilon^2)$ is large as compared to the condition number $\varkappa$. This is typically the case when the dimension is high or a high precision approximation is required. The order of preference is reversed when the condition number $\varkappa$ is large as compared to $p/(m\varepsilon^2)$. Such a situation corresponds to settings where the target log-density has a gradient that may increase very fast (M is much larger than p/ε^2). It is important to temper any comparison between the efficiency of the two samplers since there were no lower bounds proven so far for LMC or KLMC. As an important conclusion, we can note that none of these two methods is yet proven to be superior to the other in full generality. The plot in Figure 1 illustrates this fact by showing in gray the regions where the known bounds on LMC are better than those of KLMC.

4. Second-order KLMC and a bound on its error

In this section, we propose another discretization of the kinetic Langevin process, which is applicable when the function f is twice differentiable. We show below that this new discretization

[3]This value of K is obtained by choosing h and K so that the second term in the upper bound of Theorem 2 is equal to $(1-\sqrt{2}/24)\epsilon$ whereas the first term is smaller than $(\sqrt{2}/24)\epsilon$.

[4]This lower bound on K is obtained by replacing $\mathcal{D}^2 \triangleq \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2$ by 0 in (Cheng et al. [12], Theorem 1).

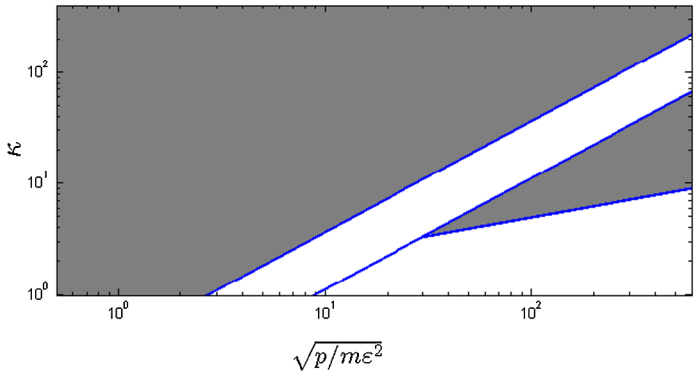

Figure 1. This plot represents in the plane defined by coordinates $(\sqrt{p/m\varepsilon^2}, \varkappa)$ the regions where the best known bound on LMC leads to a smaller error than the bound (11) established for KLMC (in gray). Please note that the axes are in logarithmic scale.

leads to a provably better sampling error bound under the condition that the Hessian matrix of f is Lipschitz-continuous with respect to the spectral norm. At any iteration $k \in \mathbb{N}$, we define $\mathcal{H}_k = \nabla^2 f(\boldsymbol{\vartheta}_k)$ and

$$\begin{bmatrix} \boldsymbol{v}_{k+1} \\ \boldsymbol{\vartheta}_{k+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\boldsymbol{v}_k - \psi_1(h)\nabla f(\boldsymbol{\vartheta}_k) - \varphi_2(h)\mathcal{H}_k\boldsymbol{v}_k \\ \boldsymbol{\vartheta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)\nabla f(\boldsymbol{\vartheta}_k) - \varphi_3(h)\mathcal{H}_k\boldsymbol{v}_k \end{bmatrix} + \sqrt{2\gamma}\begin{bmatrix} \boldsymbol{\xi}_{k+1}^{(1)} - \mathcal{H}_k\boldsymbol{\xi}_{k+1}^{(3)} \\ \boldsymbol{\xi}_{k+1}^{(2)} - \mathcal{H}_k\boldsymbol{\xi}_{k+1}^{(4)} \end{bmatrix}, \tag{13}$$

where

- ψ_0, ψ_1, ψ_2 are defined as in the beginning of the previous section,
- $\varphi_{k+1}(t) = \int_0^t e^{-\gamma(t-s)}\psi_k(s)\,ds$ for every $t > 0$,
- the $4p$ dimensional random vectors $(\boldsymbol{\xi}_{k+1}^{(1)}, \boldsymbol{\xi}_{k+1}^{(2)}, \boldsymbol{\xi}_{k+1}^{(3)}, \boldsymbol{\xi}_{k+1}^{(4)})$ are iid Gaussian with zero mean,
- for any fixed j, the 4-dimensional random vectors $([(\boldsymbol{\xi}_j^{(1)})_1, (\boldsymbol{\xi}_j^{(2)})_1, (\boldsymbol{\xi}_j^{(3)})_1, (\boldsymbol{\xi}_j^{(4)})_1], \ldots, [(\boldsymbol{\xi}_j^{(1)})_p, (\boldsymbol{\xi}_j^{(2)})_p, (\boldsymbol{\xi}_j^{(3)})_p, (\boldsymbol{\xi}_j^{(4)})_p])$ are iid with the covariance matrix

$$\bar{\mathbf{C}} = \int_0^h \big[\psi_0(t); \psi_1(t); \varphi_2(t); \varphi_3(t)\big]^\top \big[\psi_0(t); \psi_1(t); \varphi_2(t); \varphi_3(t)\big]\,dt.$$

This definition is somewhat complicated, but it follows from an application of the second-order Taylor approximation to the drift term of the kinetic Langevin diffusion.[5] At this stage, one can note that if the Hessian $\mathbf{H}_k$ is zero, then the update rule (13) boils down to the update rule of

[5]For more detailed explanations, see Section 9.1.

the KLMC algorithm in (8). Iterating the update rule (13), we get a random variable that will be henceforth called KLMC2 or second-order kinetic Langevin Monte-Carlo algorithm.

Theorem 3. *Assume that, for some constants $m, M, M_2 > 0$, the function f is m-strongly convex, its gradient is M-Lipschitz and its Hessian is M_2-Lipschitz for the spectral norm. In addition, let the initial condition of the second-order KLMC algorithm be drawn from the product distribution $\mu_0 = \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p) \otimes \nu_0$. For every*

$$\gamma \geq \sqrt{m+M} \quad \text{and} \quad h \leq \frac{m}{5\gamma M} \wedge \frac{m}{4\sqrt{5p}M_2},$$

the distribution ν_k of the kth iterate $\boldsymbol{\vartheta}_k$ of the second-order KLMC algorithm (13) satisfies

$$W_2(\nu_k, \pi) \leq \sqrt{2}\left(1 - \frac{mh}{4\gamma}\right)^k W_2(\nu_0, \pi) + \frac{2h^2 M_2 p}{m} + \frac{h^2 M\sqrt{2Mp}}{m} + \frac{8M}{m} h e^{-\frac{m^2}{160 M_2^2 h^2}}.$$

Several important consequences can be drawn from this result. First, the value of the parameter γ minimizing the right-hand side is its smallest possible value $\gamma = \sqrt{m+M}$. Second, one can note that the last term of the obtained upper bound is independent of dimension p and decreases exponentially fast in $1/h$. This term is in most cases negligible with respect to the other terms involved in the upper bound. In particular, we deduce from this result that if the Lipschitz constants M and M_2 are bounded and the strong convexity constant m is bounded away from zero, then the KLMC2 algorithm achieves the precision level ε after K_ε iterations, with K_ε being of order $\sqrt{p/\varepsilon}$, up to a logarithmic factor. Finally, if we neglect the last term in the upper bound of Theorem 3, and choose the parameters h and k so that the other terms are equal to $\varepsilon/\sqrt{4m}$, we get that the number of iteration K_ε to achieve an error $\varepsilon/\sqrt{m}$ scales, up to a logarithmic factor, as $\sqrt{M}/(mh_\varepsilon) = \sqrt{p}\varkappa_2^2 + \sqrt{p/\varepsilon}\varkappa_2^{5/4}$, where $\varkappa_2 = (M_2^{2/3} + Mp^{-1/3})/m$ is a version of the condition number taking into account the Hessian–Lipschitz assumption.

It is interesting to compare this result to the convergence result for the LMCO algorithm established in (Dalalyan and Karagulyan [16]). We can note that the number of iterations that are sufficient for the KLMC2 to achieve the error ε is much smaller than the corresponding number for the LMCO: $\sqrt{p/\varepsilon}$ versus p/ε. In addition, the KLMC2 algorithm does not need to compute matrix exponentials neither to do matrix inversion. The most costly operations are that of computing the products of the $p \times p$ Hessian and the vectors $\boldsymbol{v}_k$, $\boldsymbol{\xi}_{k+1}^3$ and $\boldsymbol{\xi}_{k+1}^3$. In most cases, the complexity of these computations scales linearly in p. In addition, the computational complexity of the Hessian-vector product is provably of the same order as that of evaluating the gradient; see (Griewank [26]).

As a conclusion, to the best of our knowledge, the second-order KLMC algorithm provides the best known upper bound on the 2-Wasserstein mixing time, in terms of its scaling $\sqrt{p/\varepsilon}$ with respect to the dimension p and the precision level ε, for a target density π having a log-density that is concave and Hessian–Lipschitz.

5. Related work

The idea of using the Langevin diffusion (see (Pavliotis [33]) for an introduction to this topic) for approximating a random variable drawn from its invariant distribution is quite old and can be traced back at least to (Roberts and Tweedie [38]). Since then, many papers focused on analyzing the asymptotic behavior of the Langevin-based methods under various assumptions, see (Bou-Rabee and Hairer [4], Douc et al. [20], Lamberton and Pagès [28,29], Pillai et al. [34], Roberts and Rosenthal [36], Roberts and Stramer [37], Stramer and Tweedie [39,40], Xifara et al. [41]) and the references therein. Convergence to the invariant distribution for Langevin processes is studied in (Desvillettes and Villani [17], Dolbeault et al. [19], Helffer and Nier [27]).

The coupling approach for obtaining rates of convergence to equilibrium for kinetic Langevin diffusions was initiated by (Bolley et al. [3]). Nonasymptotic and computable bounds for this convergence have been recently obtained in (Cheng et al. [11,12], Eberle et al. [25]). While (Cheng et al. [12]) considers only the convex case, (Cheng et al. [11], Eberle et al. [25]) deal also with nonconvexity. On the one hand, (Cheng et al. [11]) provide results only for a fixed value of parameters $(\gamma, u) = (2, 1/M)$. On the other hand, if we instantiate results of (Eberle et al. [25]) to the case of convex functions f, convergence to the invariant density is proved under the condition $\gamma^2 \geq 30Mu$. This is to be compared to the conditions of Theorem 1 that establishes exponential convergence as soon as $\gamma^2 > Mu$.

Nonasymptotic bounds on the precision of the Langevin Monte Carlo under strong convexity have been established in (Dalalyan [15]) and then extended and refined in a series of papers (Bernton [2], Brosse et al. [6], Bubeck et al. [7], Cheng and Bartlett [10], Dalalyan [14], Durmus et al. [21], Durmus and Moulines [22,23], Luu et al. [30]). Very recently, it was proved in (Dwivedi et al. [24]) that applying a Metropolis–Hastings correction to the LMC leads to improved dependence on the target precision ϵ of the number of gradient evaluations. The fact that the discretized version of the kinetic Langevin diffusion may outperform its highly overdamped counterpart was observed and quantified in (Cheng et al. [12]).

Previous work has also studied the precision of Langevin algorithms in the case when the gradient evaluations are contaminated by some noise (Baker et al. [1], Chatterji et al. [8], Cheng et al. [12], Dalalyan [14], Dalalyan and Karagulyan [16]) and the relation with stochastic optimization (Dieuleveut et al. [18], Raginsky et al. [35], Xu et al. [42], Zhang et al. [43]). The kinetic Langevin Monte Carlo method has some similarities with the Hamiltonian Monte Carlo, for which convergence to equilibrium has been recently studied by Bou-Rabee and Sanz-Serna [5], Chen, and Vempala [9] and Mangoubi and Smith [31]. There are certainly many other papers related to the present work that are not mentioned in this section. There is a vast literature on this topic and it will be impossible to quote all the papers. We believe that the papers cited here and the references therein provide a good overview of the state of the art.

6. Conclusion

In order to summarize the content of the previous sections, let us return, on by one, to the questions raised in the Introduction. First, concerning the mixing properties of the kinetic Langevin diffusion for general values of u and γ, we have established that as soon as $\gamma^2 > Mu$, the process mixes exponentially fast with a rate at least equal to $\{mu \wedge (\gamma^2 - Mu)\}/\gamma$. Therefore, for fixed

Table 2. The order of magnitude of the known upper bounds on the number of steps of Langevin related algorithms in the strongly convex case

1st-order LMC	1st-order KLMC	2nd-order KLMC
(Durmus and Moulines [23])	(Cheng et al. [12])	Theorem 3
(Dalalyan and Karagulyan [16])	and Theorem 2	
p/ε	$\sqrt{p}/\varepsilon$	$\sqrt{p/\varepsilon}$

values of m, M and u, the nearly fastest rate of mixing is obtained for $\gamma^2 = (m + M)u$ and is equal to $m/\sqrt{m + M}$.

To answer the second question, we have seen that optimization with respect to γ and u leads to improved constants but does not improve the rate. Indeed, the values of γ and u used in (Cheng et al. [12]) (i.e., $\gamma = 2$ and $u = 1/M$, which in view of Lemma 1 are equivalent to $\gamma = 2\sqrt{M}$ and $u = 1$) lead to a bound on the number of iterates sufficient to achieve a precision ε that is of the same order as the optimized one given in (10). Interestingly, our analysis revealed that not only the numerical constants of the result in (Cheng et al. [12]) can be improved, but also the dependence on the condition number $\varkappa = M/m$ can be made better. Indeed, we have managed to replace the factor $\varkappa^2$ by $\varkappa^{3/2}$. Such an improvement might have important consequences in generalizing the results to the case of a convex function which is not strongly convex. This line of research will be explored in a future work. Our bound exhibits also a better dependence on the error of the first step: it is logarithmic in our result while it was linear in (Cheng et al. [12]).

Finally, we have given an affirmative answer to the third question. We have shown that leveraging second-order information may reduce the number of steps of the algorithm by a factor proportional to $1/\sqrt{\varepsilon}$, where ε is the target precision. In order to better situate this improvement in the context of prior work, Table 2 reports the order of magnitude of the number of steps[6] of Langevin related algorithms in the strongly convex case.

7. Proof of the mixing rate

This section is devoted to proofs of the results stated in Section 2. Let $\boldsymbol{L}_0$, $\boldsymbol{L}'_0$ and $\boldsymbol{V}_0$ be three p-dimensional random vectors defined on the same probability space such that

- $\boldsymbol{V}_0$ is independent of $(\boldsymbol{L}_0, \boldsymbol{L}'_0)$,
- $\boldsymbol{V}_0 \sim \mu_1$, whereas $\boldsymbol{L}_0 \sim \mu_2$ and $\boldsymbol{L}'_0 \sim \mu'_2$,
- $W_2^2(\mu_2, \mu'_2) = \mathbf{E}[\|\boldsymbol{L}_0 - \boldsymbol{L}'_0\|_2^2]$.

Let $\boldsymbol{W}$ be a Brownian motion on the same probability space. We define $(\boldsymbol{V}, \boldsymbol{L})$ and $(\boldsymbol{V}', \boldsymbol{L}')$ as kinetic Langevin diffusion processes driven by the same Brownian motion $\boldsymbol{W}$ and satisfying the initial condition $\boldsymbol{V}'_0 = \boldsymbol{V}_0$. From the definition of the Wasserstein distance, it follows that

$$W_2^2\big(\mu \mathbf{P}_t^{\boldsymbol{L}}, \mu' \mathbf{P}_t^{\boldsymbol{L}}\big) \le \mathbf{E}\big[\big\|\boldsymbol{L}_t - \boldsymbol{L}'_t\big\|_2^2\big].$$

[6]To ease the comparison, we consider $\varkappa$ as a fixed constant and do not report the dependence on $\varkappa$ in this table.

In view of this inequality, it suffices to find an appropriate upper bound on the right-hand side of the last display, in order to prove Theorem 1. This upper bound is provided below in Proposition 1.

Proposition 1. *Let $\boldsymbol{V}_0$, $\boldsymbol{L}_0$ and $\boldsymbol{L}'_0$ be random vectors in $\mathbb{R}^p$. Let $(\boldsymbol{V}_t, \boldsymbol{L}_t)$ and $(\boldsymbol{V}'_t, \boldsymbol{L}'_t)$ be kinetic Langevin diffusions driven by the same Brownian motion and starting from $(\boldsymbol{V}_0, \boldsymbol{L}_0)$ and $(\boldsymbol{V}_0, \boldsymbol{L}'_0)$, respectively. Let v be an arbitrary real number from $[0, \gamma/2)$. We have for every $t \geq 0$*

$$\big\|\boldsymbol{L}_t - \boldsymbol{L}'_t\big\|_2 \leq \frac{\sqrt{2((\gamma - v)^2 + v^2)}}{\gamma - 2v} \exp\left\{\frac{(v^2 - m) \vee (M - (\gamma - v)^2)}{\gamma - 2v} t\right\} \big\|\boldsymbol{L}_0 - \boldsymbol{L}'_0\big\|_2.$$

Remark 1. As a consequence, we can see that for $\gamma^2 \geq 2(M + m)$ by setting

$$v = \frac{\gamma - \sqrt{\gamma^2 - 4m}}{2} \geq \frac{m}{\gamma}$$

we arrive at

$$\big\|\boldsymbol{L}_t - \boldsymbol{L}'_t\big\|_2 \leq \left(\frac{2\gamma^2 - 4m}{\gamma^2 - 4m}\right)^{1/2} e^{-vt} \big\|\boldsymbol{L}_0 - \boldsymbol{L}'_0\big\|_2, \quad \forall t \geq 0.$$

Proof. We will use the following short hand notation $y_t \triangleq (\boldsymbol{V}_t + a\boldsymbol{L}_t) - (\boldsymbol{V}'_t + a\boldsymbol{L}'_t)$ and $z_t \triangleq (-\boldsymbol{V}_t - b\boldsymbol{L}_t) + \boldsymbol{V}'_t + b\boldsymbol{L}'_t$, where a and b are two positive numbers such that $a + b = \gamma$ and $a > b$. First, note that using Taylor's theorem with the remainder term in integral form, we get

$$\nabla f(\boldsymbol{L}_t) - \nabla f\big(\boldsymbol{L}'_t\big) = \mathbf{H}_t\big(\boldsymbol{L}_t - \boldsymbol{L}'_t\big)$$

with $\mathbf{H}_t \triangleq \int_0^1 \nabla^2 f(\boldsymbol{L}_t - x(\boldsymbol{L}_t - \boldsymbol{L}'_t))\,dx$. In view of this formula and the fact that $(\boldsymbol{V}, \boldsymbol{L})$ and $(\boldsymbol{V}', \boldsymbol{L}')$ satisfy the SDE (2), we obtain

$$\begin{aligned}
\frac{d}{dt} y_t &= -\gamma\big(\boldsymbol{V}_t - \boldsymbol{V}'_t\big) - \big(\nabla f(\boldsymbol{L}_t) - \nabla f\big(\boldsymbol{L}'_t\big)\big) + a\big(\boldsymbol{V}_t - \boldsymbol{V}'_t\big) \\
&= \frac{(a - \gamma)(by_t + az_t)}{b - a} - \frac{\mathbf{H}_t(y_t + z_t)}{a - b} \\
&= \frac{(b^2\mathbf{I}_p - \mathbf{H}_t)y_t + (ba\mathbf{I}_p - \mathbf{H}_t)z_t}{a - b}.
\end{aligned}$$

In the above inequalities, we have used that $a - \gamma = -b$. Similar computations yield

$$\begin{aligned}
\frac{d}{dt} z_t &= \gamma\big(\boldsymbol{V}_t - \boldsymbol{V}'_t\big) + \big(\nabla f(\boldsymbol{L}_t) - \nabla f\big(\boldsymbol{L}'_t\big)\big) - b\big(\boldsymbol{V}_t - \boldsymbol{V}'_t\big) \\
&= \frac{(\gamma - b)(by_t + az_t)}{b - a} + \frac{\mathbf{H}_t(y_t + z_t)}{a - b} \\
&= \frac{(\mathbf{H}_t - ba\mathbf{I}_p)y_t + (\mathbf{H}_t - a^2\mathbf{I}_p)z_t}{a - b}.
\end{aligned}$$

From these equations, we deduce that

$$\begin{aligned}\frac{d}{dt}\left\|\begin{bmatrix}y_t\\ z_t\end{bmatrix}\right\|_2^2 &= 2y_t^\top\frac{dy_t}{dt}+2z_t^\top\frac{dz_t}{dt}\\ &= \frac{2}{a-b}\big\{y_t^\top\big(b^2\mathbf{I}_p-\mathbf{H}_t\big)y_t+z_t^\top\big(\mathbf{H}_t-a^2\mathbf{I}_p\big)z_t\big\}\\ &\le \frac{2}{a-b}\big\{\big(b^2-m\big)\|y_t\|_2^2+\big(M-a^2\big)\|z_t\|_2^2\big\}\\ &\le \frac{2\{(b^2-m)\vee(M-a^2)\}}{a-b}\left\|\begin{bmatrix}y_t\\ z_t\end{bmatrix}\right\|_2^2.\end{aligned}$$

An application of Gronwall's inequality yields

$$\left\|\begin{bmatrix}y_t\\ z_t\end{bmatrix}\right\|_2\le\exp\left\{\frac{(b^2-m)\vee(M-a^2)}{a-b}t\right\}\left\|\begin{bmatrix}y_0\\ z_0\end{bmatrix}\right\|_2,\quad\forall t\ge 0.$$

Since $\boldsymbol{V}_0=\boldsymbol{V}_0'$ and $\boldsymbol{L}_t-\boldsymbol{L}_t'=(y_t+z_t)/(a-b)$, we get

$$\begin{aligned}\big\|\boldsymbol{L}_t-\boldsymbol{L}_t'\big\|_2&\le\frac{\sqrt{2}}{a-b}\left\|\begin{bmatrix}y_t\\ z_t\end{bmatrix}\right\|_2\\ &\le\frac{\sqrt{2(a^2+b^2)}}{a-b}\exp\left\{\frac{(b^2-m)\vee(M-a^2)}{a-b}t\right\}\big\|\boldsymbol{L}_0-\boldsymbol{L}_0'\big\|_2,\quad\forall t\ge 0,\end{aligned}$$

and the claim of the proposition follows. □

8. Proof of the convergence of the first-order KLMC

This section contains the complete proof of Theorem 2. We first write

$$W_2(\nu_k,\pi)=W_2\big(\nu_k,\mu^*\mathbf{P}_{kh}^{\boldsymbol{L}}\big),$$

where $\mu^*=\mathcal{N}(\mathbf{0}_p,\mathbf{I}_p)\otimes\pi$ and $\mu^*\mathbf{P}_{kh}^{\boldsymbol{L}}$ is the distribution[7] of the kinetic Langevin process $\boldsymbol{L}$ at time instant kh when the initial condition of this process is drawn from μ^*. In order to upper bound the term in the right-hand side of the last display, we introduce the discretized version of the kinetic Langevin diffusion: $(\widetilde{\boldsymbol{V}}_0,\widetilde{\boldsymbol{L}}_0)\sim\mu^*$ and for every $j=0,1,\dots,k$ and for every $t\in\,]jh,(j+1)h]$,

$$\widetilde{\boldsymbol{V}}_t=\widetilde{\boldsymbol{V}}_{jh}e^{-\gamma(t-jh)}-\int_{jh}^t e^{-\gamma(t-s)}\,ds\nabla f(\widetilde{\boldsymbol{L}}_{jh})+\sqrt{2\gamma}\int_{jh}^t e^{-\gamma(t-s)}\,d\boldsymbol{W}_{jh+s},$$

[7]In other words, $\mu^*\mathbf{P}_{kh}^{\boldsymbol{L}}$ is the first marginal of the distribution $\mu^*\mathbf{P}_{kh}^{(\boldsymbol{L},\boldsymbol{V})}$, the last notation being standard in the theory of Markov processes.

$$\widetilde{\boldsymbol{L}}_t = \widetilde{\boldsymbol{L}}_{jh} + \int_{jh}^{t} \widetilde{\boldsymbol{V}}_{jh+s}\, ds.$$

We stress that $\boldsymbol{W}$ in the above formula is the same Brownian motion as the one used for defining the process $(\boldsymbol{V}, \boldsymbol{L})$. Furthermore, we choose $\widetilde{\boldsymbol{V}}_0 = \boldsymbol{V}_0$ and $(\boldsymbol{L}_0, \widetilde{\boldsymbol{L}}_0)$ so that

$$W_2^2(\nu_0, \pi) = \mathbf{E}\big[\|\boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0\|_2^2\big].$$

It is important to note that the process $(\widetilde{\boldsymbol{V}}, \widetilde{\boldsymbol{L}})$ realizes the synchronous coupling between the two sequences $\{(\boldsymbol{v}_j, \boldsymbol{\vartheta}_j);\, j = 0, \dots, k\}$ and $\{(\boldsymbol{V}_{jh}, \boldsymbol{L}_{jh});\, j = 0, \dots, k\}$. Indeed, one easily checks by mathematical induction that $(\widetilde{\boldsymbol{V}}_{jh}, \widetilde{\boldsymbol{L}}_{jh})$ has exactly the same distribution as the vector $(\boldsymbol{v}_j, \boldsymbol{\vartheta}_j)$. Therefore, we have

$$W_2\big(\nu_k, \mu^* \mathbf{P}_{kh}^{\boldsymbol{L}}\big) \le \big(\mathbf{E}\big[\|\widetilde{\boldsymbol{L}}_{kh} - \boldsymbol{L}_{kh}\|_2^2\big]\big)^{1/2} \triangleq \|\widetilde{\boldsymbol{L}}_{kh} - \boldsymbol{L}_{kh}\|_{\mathbb{L}_2}.$$

Let $\mathbf{P}$ be the matrix used in the proof of the contraction in continuous time for $v = 0$, that is,

$$\mathbf{P} = \frac{1}{\gamma}\begin{bmatrix} \mathbf{0}_{p\times p} & -\gamma \mathbf{I}_p \\ \mathbf{I}_p & \mathbf{I}_p \end{bmatrix}, \qquad \mathbf{P}^{-1} = \begin{bmatrix} \mathbf{I}_p & \gamma \mathbf{I}_p \\ -\mathbf{I}_p & \mathbf{0}_{p\times p} \end{bmatrix}.$$

We will now evaluate the sequence

$$A_k \triangleq \left\| \mathbf{P}^{-1} \begin{bmatrix} \widetilde{\boldsymbol{V}}_{kh} - \boldsymbol{V}_{kh} \\ \widetilde{\boldsymbol{L}}_{kh} - \boldsymbol{L}_{kh} \end{bmatrix} \right\|_{\mathbb{L}_2}.$$

The rest of the proof, devoted to upper bounding the last $\mathbb{L}_2$-norm, is done by mathematical induction. On each time interval $[jh, (j+1)h]$, we introduce an auxiliary continuous-time kinetic Langevin process $(\boldsymbol{V}', \boldsymbol{L}')$ such that $(\boldsymbol{V}'_{jh}, \boldsymbol{L}'_{jh}) = (\widetilde{\boldsymbol{V}}_{jh}, \widetilde{\boldsymbol{L}}_{jh})$ and

$$d\begin{bmatrix} \boldsymbol{V}'_t \\ \boldsymbol{L}'_t \end{bmatrix} = \begin{bmatrix} -\big(\gamma \boldsymbol{V}'_t + \nabla f\big(\boldsymbol{L}'_t\big)\big) \\ \boldsymbol{V}'_t \end{bmatrix} dt + \sqrt{2\gamma u} \begin{bmatrix} \mathbf{I}_p \\ \mathbf{0}_{p\times p} \end{bmatrix} d\boldsymbol{W}_t, \quad t \in \big[jh, (j+1)h\big].$$

By the triangle inequality, we have

$$\begin{aligned} A_{j+1} &\le \left\| \mathbf{P}^{-1} \begin{bmatrix} \widetilde{\boldsymbol{V}}_{(j+1)h} - \boldsymbol{V}'_{(j+1)h} \\ \widetilde{\boldsymbol{L}}_{(j+1)h} - \boldsymbol{L}'_{(j+1)h} \end{bmatrix} \right\|_{\mathbb{L}_2} + \left\| \mathbf{P}^{-1} \begin{bmatrix} \boldsymbol{V}'_{(j+1)h} - \boldsymbol{V}_{(j+1)h} \\ \boldsymbol{L}'_{(j+1)h} - \boldsymbol{L}_{(j+1)h} \end{bmatrix} \right\|_{\mathbb{L}_2} \\ &\le \left\| \mathbf{P}^{-1} \begin{bmatrix} \widetilde{\boldsymbol{V}}_{(j+1)h} - \boldsymbol{V}'_{(j+1)h} \\ \widetilde{\boldsymbol{L}}_{(j+1)h} - \boldsymbol{L}'_{(j+1)h} \end{bmatrix} \right\|_{\mathbb{L}_2} + e^{-mh/\gamma} A_j, \end{aligned} \tag{14}$$

where in the last inequality we have used the contraction established in continuous time. For the first norm in the right-hand side of the last display, we use the fact that the considered processes

$(\boldsymbol{V}', \boldsymbol{L}')$ and $(\widetilde{\boldsymbol{V}}, \widetilde{\boldsymbol{L}})$ have the same value at the time instant jh. Therefore,

$$\begin{aligned}
\big\|\widetilde{\boldsymbol{V}}_t - \boldsymbol{V}'_t\big\|_{\mathbb{L}_2} &= \bigg\|\int_{jh}^t e^{-\gamma(t-s)}\big(\nabla f\big(\boldsymbol{L}'_s\big) - \nabla f\big(\boldsymbol{L}'_{jh}\big)\big)\,ds\bigg\|_{\mathbb{L}_2} \\
&\le \int_{jh}^t \big\|\nabla f\big(\boldsymbol{L}'_s\big) - \nabla f\big(\boldsymbol{L}'_{jh}\big)\big\|_{\mathbb{L}_2}\,ds \\
&\le M\int_{jh}^t \big\|\boldsymbol{L}'_s - \boldsymbol{L}'_{jh}\big\|_{\mathbb{L}_2}\,ds \\
&\le M\int_{jh}^t\int_{jh}^s \big\|\boldsymbol{V}'_u\big\|_{\mathbb{L}_2}\,du\,ds \\
&= M\int_{jh}^t (t-u)\big\|\boldsymbol{V}'_u\big\|_{\mathbb{L}_2}\,du \\
&\le M\int_{jh}^t (t-u)\,du \max_{u\in[jh,(j+1)h]}\big\|\boldsymbol{V}'_u\big\|_{\mathbb{L}_2} \\
&= \frac{M(t-jh)^2}{2}\max_{u\in[jh,(j+1)h]}\big\|\boldsymbol{V}'_u\big\|_{\mathbb{L}_2}
\end{aligned}$$

and

$$\begin{aligned}
\big\|\widetilde{\boldsymbol{L}}_{(j+1)h} - \boldsymbol{L}'_{(j+1)h}\big\|_{\mathbb{L}_2} &= \bigg\|\int_{jh}^{(j+1)h}\big(\widetilde{\boldsymbol{V}}_t - \boldsymbol{V}'_t\big)\,dt\bigg\|_{\mathbb{L}_2} \\
&\le \int_{jh}^{(j+1)h}\big\|\widetilde{\boldsymbol{V}}_t - \boldsymbol{V}'_t\big\|_{\mathbb{L}_2}\,dt \\
&\le \frac{M}{2}\int_{jh}^{(j+1)h}(t-jh)^2\,dt \max_{u\in[jh,(j+1)h]}\big\|\boldsymbol{V}'_u\big\|_{\mathbb{L}_2} \\
&\le \frac{Mh^3}{6}\max_{u\in[jh,(j+1)h]}\big\|\boldsymbol{V}'_u\big\|_{\mathbb{L}_2}.
\end{aligned}$$

Lemma 2. *For every $u \in [jh, (j+1)h]$, we have*

$$\big\|\boldsymbol{V}'_u\big\|_{\mathbb{L}_2} \le \sqrt{p} + A_j.$$

Proof. We have

$$\begin{aligned}
\big\|\boldsymbol{V}'_u\big\|_{\mathbb{L}_2} &= \|\boldsymbol{V}_u\|_{\mathbb{L}_2} + \big\|\boldsymbol{V}'_u - \boldsymbol{V}_u\big\|_{\mathbb{L}_2} \\
&= \sqrt{p} + \big\|[\mathbf{I}_p,\ \mathbf{0}_p]\mathbf{P}\mathbf{P}^{-1}\big[\big(\boldsymbol{V}'_u - \boldsymbol{V}_u\big)^\top, \big(\boldsymbol{L}'_u - \boldsymbol{L}_u\big)^\top\big]\big\|_{\mathbb{L}_2} \\
&\le \sqrt{p} + \big\|[\mathbf{I}_p,\ \mathbf{0}_p]\mathbf{P}\big\| \times \big\|\mathbf{P}^{-1}\big[\big(\boldsymbol{V}'_u - \boldsymbol{V}_u\big)^\top, \big(\boldsymbol{L}'_u - \boldsymbol{L}_u\big)^\top\big]\big\|_{\mathbb{L}_2}
\end{aligned}$$

$$\begin{aligned}
&\le \sqrt{p} + \big\|[\mathbf{I}_p,\ \mathbf{0}_p]\mathbf{P}\big\| \times \big\|\mathbf{P}^{-1}\big[\big(\boldsymbol{V}'_{jh} - \boldsymbol{V}_{jh}\big)^\top, \big(\boldsymbol{L}'_{jh} - \boldsymbol{L}_{jh}\big)^\top\big]\big\|_{\mathbb{L}_2} \\
&= \sqrt{p} + \big\|[\mathbf{I}_p,\ \mathbf{0}_p]\mathbf{P}\big\| \times A_j.
\end{aligned}$$

Recall that

$$\mathbf{P} = \frac{1}{\gamma}\begin{bmatrix} \mathbf{0}_{p\times p} & -\gamma\mathbf{I}_p \\ \mathbf{I}_p & \mathbf{I}_p \end{bmatrix},$$

which implies that $\|[\mathbf{I}_p, \mathbf{0}_p]\mathbf{P}\| = 1$. This completes the proof of the lemma. □

From this lemma and previous inequalities, we infer that

$$\begin{aligned}
&\left\|\mathbf{P}^{-1}\begin{bmatrix} \widetilde{\boldsymbol{V}}_{(j+1)h} - \boldsymbol{V}'_{(j+1)h} \\ \widetilde{\boldsymbol{L}}_{(j+1)h} - \boldsymbol{L}'_{(j+1)h} \end{bmatrix}\right\|_{\mathbb{L}_2} \\
&\quad\le \Big\{\big(\big\|\widetilde{\boldsymbol{V}}_{(j+1)h} - \boldsymbol{V}'_{(j+1)h}\big\|_{\mathbb{L}_2} + \gamma\big\|\widetilde{\boldsymbol{L}}_{(j+1)h} - \boldsymbol{L}'_{(j+1)h}\big\|_{\mathbb{L}_2}\big)^2 \\
&\qquad + \big\|\widetilde{\boldsymbol{V}}_{(j+1)h} - \boldsymbol{V}'_{(j+1)h}\big\|^2_{\mathbb{L}_2}\Big\}^{1/2} \\
&\quad\le \left\{\left(1 + \frac{\gamma h}{3}\right)^2 + 1\right\}^{1/2} \frac{Mh^2}{2}(\sqrt{p} + A_j).
\end{aligned}$$

Choosing $h \le 1/(4\gamma)$, we arrive at

$$\left\|\mathbf{P}^{-1}\begin{bmatrix} \widetilde{\boldsymbol{V}}_{(j+1)h} - \boldsymbol{V}'_{(j+1)h} \\ \widetilde{\boldsymbol{L}}_{(j+1)h} - \boldsymbol{L}'_{(j+1)h} \end{bmatrix}\right\|_{\mathbb{L}_2} \le 0.75Mh^2(\sqrt{p} + A_j).$$

Combining this inequality and (14), for every $h \le m/(4\gamma M)$, we get

$$\begin{aligned}
A_{j+1} &\le 0.75Mh^2(\sqrt{p} + A_j) + e^{-hm/\gamma}A_j \\
&= 0.75Mh^2\sqrt{p} + \big(e^{-hm/\gamma} + 0.75Mh^2\big)A_j. \qquad (15)
\end{aligned}$$

Using the inequality $e^{-x} \le 1 - x + \frac{1}{2}x^2$, we can derive from (15) that

$$\begin{aligned}
A_{j+1} &\le 0.75Mh^2\sqrt{p} + \left(1 - \frac{hm}{\gamma} + \frac{h^2m^2}{2\gamma^2} + 0.75Mh^2\right)A_j \\
&\le 0.75Mh^2\sqrt{p} + \left(1 - \frac{0.75mh}{\gamma}\right)A_j.
\end{aligned}$$

Unfolding this recursive inequality, we arrive at

$$A_k \le \frac{Mh\gamma\sqrt{p}}{m} + \left(1 - \frac{0.75mh}{\gamma}\right)^k A_0.$$

Finally, one easily checks that $A_0 = \gamma W_2(\nu_0, \pi)$ and

$$\|\widetilde{\boldsymbol{L}}_{kh} - \boldsymbol{L}_{kh}\|_{\mathbb{L}_2} \le \big\|[\mathbf{0}_{p\times p}\ \mathbf{I}_p]\mathbf{P}\big\| A_k = \gamma^{-1}\sqrt{2}A_k.$$

Putting all these pieces together, we arrive at

$$\begin{aligned} W_2(\nu_k, \pi) &\le \|\widetilde{\boldsymbol{L}}_{kh} - \boldsymbol{L}_{kh}\|_{\mathbb{L}_2} \\ &\le \gamma^{-1}\sqrt{2}A_k \\ &\le \frac{Mh\sqrt{2p}}{m} + \sqrt{2}\left(1 - \frac{0.75mh}{\gamma}\right)^k (A_0/\gamma) \\ &= \frac{Mh\sqrt{2p}}{m} + \sqrt{2}\left(1 - \frac{0.75mh}{\gamma}\right)^k W_2(\nu_0, \pi), \end{aligned}$$

and the claim of Theorem 2 follows.

9. Proofs for the second-order discretization of the kinetic Langevin diffusion

We start this section by providing some explanations on the origin of the KLMC2 algorithm. We turn then to the proof of Theorem 3. In the computations of this section, we make repeated use of the maps ψ_k and φ_k introduced in Section 3 and Section 4. We recall their definitions here, for a fixed value of the friction parameter γ. For every $t > 0$ and every $k \in \mathbb{N}$, we have:

$$\psi_0(t) = e^{-\gamma t}, \qquad \psi_{k+1}(t) = \int_0^t \psi_k(s)\, ds, \qquad \varphi_{k+1}(t) = \int_0^t e^{-\gamma(t-s)}\psi_k(s)\, ds. \tag{16}$$

These short-hand notation are convenient in the following proofs. The notation indicates that for small values of $t > 0$ both $\psi_k(t)$ and $\varphi_k(t)$ behave like k-degrees polynoms.

9.1. Heuristics on the origin of the KLMC2 algorithm

Recall that the kinetic diffusion is given by the equation

$$d\begin{bmatrix} \boldsymbol{V}_t \\ \boldsymbol{L}_t \end{bmatrix} = \begin{bmatrix} -\big(\gamma \boldsymbol{V}_t + \nabla f(\boldsymbol{L}_t)\big) \\ \boldsymbol{V}_t \end{bmatrix} dt + \sqrt{2\gamma}\begin{bmatrix} \mathbf{I}_p \\ \mathbf{0}_{p\times p} \end{bmatrix} d\boldsymbol{W}_t. \tag{17}$$

From (17), by integration by parts, we can deduce that

$$\begin{aligned} e^{\gamma t}\boldsymbol{V}_t &= \boldsymbol{V}_0 + \int_0^t e^{\gamma s}\, d\boldsymbol{V}_s + \gamma \int_0^t e^{\gamma s}\boldsymbol{V}_s\, ds \\ &= \boldsymbol{V}_0 - \int_0^t e^{\gamma s}\nabla f(\boldsymbol{L}_s)\, ds + \sqrt{2\gamma}\int_0^t e^{\gamma s}\, d\boldsymbol{W}_s. \end{aligned}$$

Therefore, we have

$$\begin{aligned} \boldsymbol{V}_t &= e^{-\gamma t}\boldsymbol{V}_0 - \int_0^t e^{-\gamma(t-s)}\nabla f(\boldsymbol{L}_s)\,ds + \sqrt{2\gamma}\int_0^t e^{-\gamma(t-s)}\,d\boldsymbol{W}_s, \\ \boldsymbol{L}_t &= \boldsymbol{L}_0 + \int_0^t \boldsymbol{V}_s\,ds. \end{aligned} \tag{18}$$

If the function f is twice continuously differentiable, then, for small values of s, the value $\nabla f(\boldsymbol{L}_s)$ appearing in (18) can be approximated by an affine function of $\boldsymbol{L}_s$:

$$\begin{aligned} \nabla f(\boldsymbol{L}_s) &\approx \nabla f(\boldsymbol{L}_0) + \nabla^2 f(\boldsymbol{L}_0)(\boldsymbol{L}_s - \boldsymbol{L}_0) \\ &= \nabla f(\boldsymbol{L}_0) + \nabla^2 f(\boldsymbol{L}_0)\int_0^s \boldsymbol{V}_w\,dw \\ &\approx \nabla f(\boldsymbol{L}_0) + \psi_1(s)\nabla^2 f(\boldsymbol{L}_0)\boldsymbol{V}_0 + \sqrt{2\gamma}\nabla^2 f(\boldsymbol{L}_0)\int_0^s \psi_1(s-w)\,d\boldsymbol{W}_w. \end{aligned}$$

From the above approximation, we can infer that

$$\begin{aligned} \int_0^t e^{-\gamma(t-s)}\nabla f(\boldsymbol{L}_s)\,ds &\approx \psi_1(t)\nabla f(\boldsymbol{L}_0) + \varphi_2(t)\nabla^2 f(\boldsymbol{L}_0)\boldsymbol{V}_0 \\ &\quad + \sqrt{2\gamma}\nabla^2 f(\boldsymbol{L}_0)\int_0^t e^{-\gamma(t-s)}\int_0^s \psi_1(s-w)\,d\boldsymbol{W}_w\,ds. \end{aligned} \tag{19}$$

Lemma 3. *For every $\gamma > 0$ and $t > 0$, we have for any $k, j \in \mathbb{N}$,*

$$\varphi_{k+1}(t) = \int_0^t \varphi_k(s)\,ds, \qquad \varphi_{k+j+1}(t) = \int_0^t \psi_k(s)\psi_j(t-s)\,ds.$$

Proof. Fubini's theorem and a change of variables yield

$$\begin{aligned} \int_0^t \varphi_k(s)\,ds &= \int_0^t\int_0^s e^{-\gamma(s-r)}\psi_{k-1}(r)\,dr\,ds \\ &= \int_0^t\int_0^{t-r} e^{-\gamma s}\psi_{k-1}(r)\,ds\,dr \\ &= \int_0^t\int_0^{t-s} e^{-\gamma s}\psi_{k-1}(r)\,dr\,ds \\ &= \int_0^t e^{-\gamma s}\psi_k(t-s)\,ds = \varphi_{k+1}(t). \end{aligned}$$

This is the first claim of the lemma.

The second claim of the lemma is true for $j = 0$ and any $k \in \mathbb{N}$ by definition. By induction, we get

$$\begin{aligned}\int_0^t \psi_k(s)\psi_j(t-s)\,ds &= \int_0^t \psi_k(s)\int_0^{t-s} \psi_{j-1}(r)\,dr\,ds \\ &= \int_0^t \int_0^{t-r} \psi_k(s)\psi_{j-1}(r)\,ds\,dr \\ &= \int_0^t \psi_{k+1}(t-r)\psi_{j-1}(r)\,dr \\ &= \int_0^t \psi_{k+j}(r)\psi_0(t-r)\,dr = \varphi_{k+j+1}(t).\end{aligned}$$

This completes the proof of the lemma. □

Applied to the double integral in (19), Lemma 3 yields

$$\begin{aligned}\int_0^t e^{-\gamma(t-s)}\int_0^s \psi_1(s-w)\,d\boldsymbol{W}_w\,ds &= \int_0^t \int_w^t e^{-\gamma(t-s)}\psi_1(s-w)\,ds\,d\boldsymbol{W}_w \\ &= \int_0^t \int_0^{t-w} e^{-\gamma(t-w-u)}\psi_1(u)\,du\,d\boldsymbol{W}_w \\ &= \int_0^t \varphi_2(t-w)\,d\boldsymbol{W}_w.\end{aligned}$$

Therefore, (19) becomes

$$\begin{aligned}\int_0^t e^{-\gamma(t-s)}\nabla f(\boldsymbol{L}_s)\,ds \approx{}& \psi_1(t)\nabla f(\boldsymbol{L}_0) + \varphi_2(t)\nabla^2 f(\boldsymbol{L}_0)\boldsymbol{V}_0 \\ &+ \sqrt{2\gamma}\nabla^2 f(\boldsymbol{L}_0)\int_0^t \varphi_2(t-w)\,d\boldsymbol{W}_w.\end{aligned}$$

Combining the approximation above and the diffusion equation (18), we get for small values of $t > 0$:

$$\begin{aligned}\boldsymbol{V}_t \approx{}& e^{-\gamma t}\boldsymbol{V}_0 - \psi_1(t)\nabla f(\boldsymbol{L}_0) - \varphi_2(t)\nabla^2 f(\boldsymbol{L}_0)\boldsymbol{V}_0 \\ &- \sqrt{2\gamma}\nabla^2 f(\boldsymbol{L}_0)\int_0^t \varphi_2(t-s)\,d\boldsymbol{W}_s + \sqrt{2\gamma}\int_0^t e^{-\gamma(t-s)}\,d\boldsymbol{W}_s.\end{aligned}$$

This approximation will be used for defining the discretized version of the process $\boldsymbol{V}$. In order to define the discretized version of $\boldsymbol{L}$, we will simply use the plug-in approximation of $\boldsymbol{V}$, and

then integrate. This leads to

$$\begin{aligned}\boldsymbol{L}_t &= \boldsymbol{L}_0 + \int_0^t \boldsymbol{V}_s\,ds \\ &\approx \boldsymbol{L}_0 + \psi_1(t)\boldsymbol{V}_0 - \psi_2(t)\nabla f(\boldsymbol{L}_0) - \varphi_3(t)\nabla^2 f(\boldsymbol{L}_0)\boldsymbol{V}_0 \\ &\quad - \sqrt{2\gamma}\nabla^2 f(\boldsymbol{L}_0)\int_0^t \varphi_3(t-w)\,d\boldsymbol{W}_w + \sqrt{2\gamma}\int_0^t \psi_1(t-w)\,d\boldsymbol{W}_w.\end{aligned}$$

9.2. Proof of Theorem 3

We now turn to the proof of Theorem 3. Similar to the proof of Theorem 2, we first write

$$W_2(\nu_k,\pi) = W_2\big(\nu_k, \mu^*\mathbf{P}_{kh}^{\boldsymbol{L}}\big),$$

where $\mu^* = \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)\otimes\pi$ and $\mu^*\mathbf{P}_{kh}^{\boldsymbol{L}}$ is the distribution of the kinetic Langevin process $\boldsymbol{L}$ at time instant kh when the initial condition of this process is drawn from μ^*. To provide an upper bound on the right-hand side, we introduce the second-order discretization of the kinetic Langevin diffusion: $(\widetilde{\boldsymbol{V}}_0, \widetilde{\boldsymbol{L}}_0)\sim\mu^*$ and for every $j=0,1,\dots,k$ and for every $t\in\,]jh,(j+1)h]$,

$$\begin{aligned}\widetilde{\boldsymbol{V}}_t &= \widetilde{\boldsymbol{V}}_{jh}e^{-\gamma(t-jh)} - \big(\psi_1(t-jh)\nabla f(\widetilde{\boldsymbol{L}}_0) + \varphi_2(t-jh)\nabla^2 f(\widetilde{\boldsymbol{L}}_{jh})\widetilde{\boldsymbol{V}}_{jh}\big) \\ &\quad + \sqrt{2\gamma}\bigg(\int_{jh}^t e^{-\gamma(t-s)}\,d\boldsymbol{W}_{jh+s} - \nabla^2 f(\widetilde{\boldsymbol{L}}_{jh})\int_{jh}^t \varphi_2(t-s)\,d\boldsymbol{W}_{jh+s}\bigg) \qquad (20)\end{aligned}$$

and

$$\begin{aligned}\widetilde{\boldsymbol{L}}_t &= \widetilde{\boldsymbol{L}}_{jh} + \psi_1(t-jh)\widetilde{\boldsymbol{V}}_{jh} - \big(\psi_2(t-jh)\nabla f(\widetilde{\boldsymbol{L}}_{jh}) + \varphi_3(t-jh)\nabla^2 f(\widetilde{\boldsymbol{L}}_{jh})\widetilde{\boldsymbol{V}}_{jh}\big) \\ &\quad + \sqrt{2\gamma}\bigg(\int_{jh}^t \psi_1(t-s)\,d\boldsymbol{W}_{jh+s} - \nabla^2 f(\widetilde{\boldsymbol{L}}_{jh})\int_{jh}^t \varphi_3(t-s)\,d\boldsymbol{W}_{jh+s}\bigg). \qquad (21)\end{aligned}$$

The origin of those processes is explained in Section 9.1. Similar to the proof of Theorem 2, the Brownian motion $\boldsymbol{W}$ in the above formula is the same as the one used for defining the process $(\boldsymbol{V},\boldsymbol{L})$. Furthermore, we also choose $\widetilde{\boldsymbol{V}}_0 = \boldsymbol{V}_0$ and $(\boldsymbol{L}_0, \widetilde{\boldsymbol{L}}_0)$ so that

$$W_2^2(\nu_0,\pi) = \mathbf{E}\big[\|\boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0\|_2^2\big].$$

In what follows, we will use the following matrices to perform a linear transformation of the space $\mathbb{R}^{2p}$:

$$\mathbf{P} = \gamma^{-1}\begin{bmatrix}\mathbf{0}_{p\times p} & -\gamma\mathbf{I}_p \\ \mathbf{I}_p & \mathbf{I}_p\end{bmatrix}, \qquad \mathbf{P}^{-1} = \begin{bmatrix}\mathbf{I}_p & \gamma\mathbf{I}_p \\ -\mathbf{I}_p & \mathbf{0}_{p\times p}\end{bmatrix}. \qquad (22)$$

On each time interval $[jh,(j+1)h]$, we introduce an auxiliary discretized process $(\widehat{\boldsymbol{V}},\widehat{\boldsymbol{L}})$ that evolves according to exactly the same dynamics as $(\widetilde{\boldsymbol{V}},\widetilde{\boldsymbol{L}})$ defined in (20) and (21), but is such that $(\widehat{\boldsymbol{V}}_{jh},\widehat{\boldsymbol{L}}_{jh}) = (\boldsymbol{V}_{jh},\boldsymbol{L}_{jh})$.

The proof of Theorem 3 is divided into three propositions. For the purposes of notation, in both proofs of Propositions 2 and 3 we control the error of one single iteration of the KLMC2 algorithm. It should be understood that this analysis actually holds for every iteration of the KLMC2 algorithm. In the proof of Theorem 3, we combine the results of Propositions 2 and 3 to unfold the induction and complete the proof of the theorem.

Proposition 2. *Assume that, for some constants $m, M, M_2 > 0$, the function f is m-strongly convex, its gradient is M-Lipschitz and its Hessian is M_2-Lipschitz for the spectral norm. If the parameter γ and the step size h of the KLMC2 algorithm are such that*

$$h \le \frac{1}{5\gamma},$$

then, for the $(2p) \times (2p)$ matrix $\mathbf{P}$ defined in (22), and for every $t \in [0, h]$ we have

$$\left\| \mathbf{P}^{-1} \begin{bmatrix} \boldsymbol{V}_t - \widehat{\boldsymbol{V}}_t \\ \boldsymbol{L}_t - \widehat{\boldsymbol{L}}_t \end{bmatrix} \right\|_{\mathbb{L}_2} \le 0.25 \times t^3 \big(M_2 \sqrt{p^2 + 2p} + M^{3/2} \sqrt{p} \big).$$

Proof. From the definition of $\mathbf{P}^{-1}$ introduced in (22), we compute

$$\begin{aligned} \left\| \mathbf{P}^{-1} \begin{bmatrix} \boldsymbol{V}_t - \widehat{\boldsymbol{V}}_t \\ \boldsymbol{L}_t - \widehat{\boldsymbol{L}}_t \end{bmatrix} \right\|_{\mathbb{L}_2} &= \big\{ \big\| \boldsymbol{V}_t - \widehat{\boldsymbol{V}}_t + \gamma (\boldsymbol{L}_t - \widehat{\boldsymbol{L}}_t) \big\|_{\mathbb{L}_2}^2 + \| \boldsymbol{V}_t - \widehat{\boldsymbol{V}}_t \|_{\mathbb{L}_2}^2 \big\}^{1/2} \\ &\le \big\{ \big(\| \boldsymbol{V}_t - \widehat{\boldsymbol{V}}_t \|_{\mathbb{L}_2} + \gamma \| \boldsymbol{L}_t - \widehat{\boldsymbol{L}}_t \|_{\mathbb{L}_2} \big)^2 + \| \boldsymbol{V}_t - \widehat{\boldsymbol{V}}_t \|_{\mathbb{L}_2}^2 \big\}^{1/2}, \end{aligned}$$

where the upper bound follows from Minkowski's inequality. We now give upper bounds for the $\mathbb{L}_2$-norm of processes $\boldsymbol{V} - \widehat{\boldsymbol{V}}$ and $\boldsymbol{L} - \widehat{\boldsymbol{L}}$.

Lemma 4. *For every $t \in [0, h]$, we have*

$$\begin{aligned} \| \widehat{\boldsymbol{V}}_t - \boldsymbol{V}_t \|_{\mathbb{L}_2} &\le \frac{t^3 (M_2 \sqrt{p^2 + 2p} + M^{3/2} \sqrt{p})}{6}, \\ \| \widehat{\boldsymbol{L}}_t - \boldsymbol{L}_t \|_{\mathbb{L}_2} &\le \frac{t^4 (M_2 \sqrt{p^2 + 2p} + M^{3/2} \sqrt{p})}{24}. \end{aligned}$$

Proof. Recall that $\psi_1(t) = \int_0^t e^{-\gamma(t-s)}\, ds$, $\psi_2(t) = \int_0^t s e^{-\gamma(t-s)}\, ds$ and

$$\boldsymbol{V}_t = e^{-\gamma t} \boldsymbol{V}_0 - \int_0^t e^{-\gamma(t-s)} \nabla f(\boldsymbol{L}_s)\, ds + \sqrt{2\gamma} \int_0^t e^{-\gamma(t-s)}\, d\boldsymbol{W}_s.$$

We compute

$$\begin{aligned} \widehat{\boldsymbol{V}}_t - \boldsymbol{V}_t = \int_0^t e^{-\gamma(t-s)} \big(\nabla f(\boldsymbol{L}_s) - \nabla f(\boldsymbol{L}_0) \big)\, ds - \varphi_2(t) \nabla^2 f(\boldsymbol{L}_0) \boldsymbol{V}_0 \\ - \sqrt{2\gamma} \nabla^2 f(\boldsymbol{L}_0) \int_0^t \varphi_2(t-s)\, d\boldsymbol{W}_s. \end{aligned}$$

By Taylor's theorem, we have for every $s > 0$

$$\nabla f(\boldsymbol{L}_s) - \nabla f(\boldsymbol{L}_0) = \bar{\mathbf{H}}_s(\boldsymbol{L}_s - \boldsymbol{L}_0),$$
$$\bar{\mathbf{H}}_s \triangleq \int_0^1 \nabla^2 f\big(\boldsymbol{L}_s + x(\boldsymbol{L}_0 - \boldsymbol{L}_s)\big)\, dx.$$

This yields the following rewriting of the first integral:

$$\begin{aligned}
&\int_0^t e^{-\gamma(t-s)}\big(\nabla f(\boldsymbol{L}_s) - \nabla f(\boldsymbol{L}_0)\big)\, ds \\
&\quad = \underbrace{\int_0^t e^{-\gamma(t-s)}\big(\bar{\mathbf{H}}_s - \nabla^2 f(\boldsymbol{L}_0)\big)(\boldsymbol{L}_s - \boldsymbol{L}_0)\, ds}_{\triangleq \boldsymbol{A}_t} \\
&\qquad + \underbrace{\nabla^2 f(\boldsymbol{L}_0) \int_0^t \int_0^s e^{-\gamma(t-s)} \boldsymbol{V}_r\, dr\, ds}_{\triangleq \boldsymbol{C}_t}.
\end{aligned}$$

Now, we replace $\boldsymbol{V}_r$ by its explicit expression

$$\boldsymbol{V}_r = e^{-\gamma r}\boldsymbol{V}_0 - \int_0^r e^{-\gamma(r-w)}\nabla f(\boldsymbol{L}_w)\, dw + \sqrt{2\gamma}\int_0^r e^{-\gamma(r-w)}\, d\boldsymbol{W}_w.$$

By integrating twice, we compute

$$\begin{aligned}
\boldsymbol{C}_t &= \varphi_2(t)\nabla^2 f(\boldsymbol{L}_0)\boldsymbol{V}_0 + \sqrt{2\gamma}\nabla^2 f(\boldsymbol{L}_0)\int_0^t \varphi_2(t-s)\, d\boldsymbol{W}_s \\
&\quad - \underbrace{\nabla^2 f(\boldsymbol{L}_0)\int_0^t \int_0^s \int_0^r e^{-\gamma(t-s)} e^{-\gamma(r-w)}\nabla f(\boldsymbol{L}_w)\, dw\, dr\, ds}_{\triangleq \boldsymbol{B}_t}.
\end{aligned}$$

Summing the two expressions allows some terms to cancel out leading to

$$\widehat{\boldsymbol{V}}_t - \boldsymbol{V}_t = \boldsymbol{A}_t - \boldsymbol{B}_t,$$

where

$$\begin{aligned}
\boldsymbol{A}_t &= \int_0^t \int_0^1 e^{-\gamma(t-s)}\big(\nabla^2 f\big(\boldsymbol{L}_s + h(\boldsymbol{L}_0 - \boldsymbol{L}_s)\big) - \nabla^2 f(\boldsymbol{L}_0)\big)(\boldsymbol{L}_s - \boldsymbol{L}_0)\, dh\, ds, \\
\boldsymbol{B}_t &= \nabla^2 f(\boldsymbol{L}_0)\int_0^t \int_0^s \int_0^r e^{-\gamma(t-s)} e^{-\gamma(r-w)}\nabla f(\boldsymbol{L}_w)\, dw\, dr\, ds.
\end{aligned}$$

We now control $\mathbb{L}_2$-norm of processes $\boldsymbol{A}_t$ and $\boldsymbol{B}_t$. Bounding $e^{-\gamma(t-s)}$ by one, Minkowski's inequality in its integral version and the Lipschitz assumption on the Hessian yield

$$\begin{aligned}
\|\boldsymbol{A}_t\|_{\mathbb{L}_2} &\le \int_0^t \int_0^1 \mathbf{E}\big[\big\|\big(\nabla^2 f\big(\boldsymbol{L}_s + h(\boldsymbol{L}_0 - \boldsymbol{L}_s)\big) - \nabla^2 f(\boldsymbol{L}_0)\big)(\boldsymbol{L}_s - \boldsymbol{L}_0)\big\|_2^2\big]^{1/2} dh\, ds \\
&\le M_2 \int_0^t \int_0^1 \mathbf{E}\big[(1-h)^2 \|\boldsymbol{L}_s - \boldsymbol{L}_0\|_2^4\big]^{1/2} dh\, ds \\
&= \frac{M_2}{2} \int_0^t \left\{\mathbf{E}\left[\left\|\int_0^s \boldsymbol{V}_r\, dr\right\|_2^4\right]^{1/4}\right\}^2 ds \\
&\le \frac{M_2}{2} \int_0^t \left\{\int_0^s \mathbf{E}\big[\|\boldsymbol{V}_r\|_2^4\big]^{1/4} dr\right\}^2 ds \\
&= \frac{M_2}{2} \int_0^t \left\{\int_0^s \mathbf{E}\big[\|\boldsymbol{V}_0\|_2^4\big]^{1/4} dr\right\}^2 ds \\
&= \frac{M_2 t^3}{6} \mathbf{E}\big[\|\boldsymbol{V}_0\|_2^4\big]^{1/2},
\end{aligned}$$

where we have used the stationarity of the process $\boldsymbol{V}_r$. Since $\boldsymbol{V}_0$ is standard Gaussian, we get $\mathbf{E}[\|\boldsymbol{V}_0\|_2^4] = p^2 + 2p$.

In the same way, Minkowski's inequality in its integral version yields

$$\begin{aligned}
\|\boldsymbol{B}_t\|_{\mathbb{L}_2} &\le \int_0^t \int_0^s \int_0^r \big\|\nabla^2 f(\boldsymbol{L}_0) \nabla f(\boldsymbol{L}_w)\big\|_{\mathbb{L}_2} dw\, dr\, ds \\
&\le \int_0^t \int_0^s \int_0^r M \big\|\nabla f(\boldsymbol{L}_w)\big\|_{\mathbb{L}_2} dw\, dr\, ds \\
&= M \big\|\nabla f(\boldsymbol{L}_0)\big\|_{\mathbb{L}_2} \int_0^t \int_0^s \int_0^r dw\, dr\, ds = \frac{t^3 M}{6} \big\|\nabla f(\boldsymbol{L}_0)\big\|_{\mathbb{L}_2},
\end{aligned}$$

where last equalities follow from the stationarity of $\boldsymbol{L}_w$. Since $\boldsymbol{L}_0 \sim \pi$ (Dalalyan [14], Lemma 2) ensures that $\|\nabla f(\boldsymbol{L}_0)\|_{\mathbb{L}_2} \le \sqrt{Mp}$, and the first claim of the lemma follows.

The bound for process $\boldsymbol{L} - \widehat{\boldsymbol{L}}$ follows from Minkowski's inequality combined with the bound just proven:

$$\begin{aligned}
\|\widehat{\boldsymbol{L}}_t - \boldsymbol{L}_t\|_{\mathbb{L}_2} &\le \int_0^t \|\widehat{\boldsymbol{V}}_s - \boldsymbol{V}_s\|_{\mathbb{L}_2}\, ds \\
&\le \int_0^t \left(\frac{t^3 (M_2 \sqrt{p^2 + 2p} + M^{3/2} \sqrt{p})}{6}\right) ds \\
&= \frac{t^4 (M_2 \sqrt{p^2 + 2p} + M^{3/2} \sqrt{p})}{24}.
\end{aligned}$$

This completes the proof of the lemma. □

The claim of the proposition follows from the assumption $\gamma t \leq 1/5$ and that

$$\sqrt{\left(\frac{1}{6}+\frac{1}{5\times 24}\right)^2+\left(\frac{1}{6}\right)^2} \leq 0.25. \qquad \square$$

The next, perhaps the most important, step of the proof is to assess the distance between the random vectors $(\widehat{\boldsymbol{V}}_t, \widehat{\boldsymbol{L}}_t)$ and $(\widetilde{\boldsymbol{V}}_t, \widetilde{\boldsymbol{L}}_t)$.

Proposition 3. *Assume that, for some constants $m, M, M_2 > 0$, the function f is m-strongly convex, its gradient is M-Lipschitz and its Hessian is M_2-Lipschitz for the spectral norm. If the parameter γ and the step size h of the KLMC2 algorithm satisfy the inequalities*

$$\gamma^2 \geq m + M, \quad h \leq \frac{1}{5\gamma\varkappa},$$

then, for the $(2p)\times(2p)$ matrix $\mathbf{P}$ defined in (22), and for every $a \geq 5p$ and $t \in [0,h]$, it holds

$$\left\|\mathbf{P}^{-1}\begin{bmatrix}\widehat{\boldsymbol{V}}_t-\widetilde{\boldsymbol{V}}_t\\ \widehat{\boldsymbol{L}}_t-\widetilde{\boldsymbol{L}}_t\end{bmatrix}\right\|_{\mathbb{L}_2} \leq \left(1-\frac{mt}{2\gamma}+\frac{M_2\sqrt{a}t^2}{\gamma}\right)\left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_0-\widetilde{\boldsymbol{V}}_0\\ \boldsymbol{L}_0-\widetilde{\boldsymbol{L}}_0\end{bmatrix}\right\|_{\mathbb{L}_2} + \sqrt{2}t^2(M-m)e^{-(a-p)/8}.$$

Proof. *Step 1*: We first introduce some shorthand notation, convenient for writing the processes into their matrix form. Define

$$\mathbf{H}_0 \triangleq \int_0^1 \nabla^2 f\big(\boldsymbol{L}_0 - x(\boldsymbol{L}_0-\widetilde{\boldsymbol{L}}_0)\big)\,dx,$$

such that Taylor's expansion yields

$$\nabla f(\boldsymbol{L}_0) - \nabla f(\widetilde{\boldsymbol{L}}_0) = \mathbf{H}_0(\boldsymbol{L}_0-\widetilde{\boldsymbol{L}}_0).$$

We also introduce the following $2p\times 2p$ matrices:

$$\mathbf{R}_0 \triangleq \begin{bmatrix}\gamma\mathbf{I}_p & \mathbf{H}_0\\ -\mathbf{I}_p & \mathbf{0}_{p\times p}\end{bmatrix}, \qquad \mathbf{E}_0(t) \triangleq \begin{bmatrix}\varphi_2(t)\nabla^2 f(\widetilde{\boldsymbol{L}}_0) & \mathbf{0}_{p\times p}\\ \varphi_3(t)\nabla^2 f(\widetilde{\boldsymbol{L}}_0) & -\psi_2(t)\mathbf{H}_0\end{bmatrix}. \tag{23}$$

From the definitions (20) and (21), we write the process $(\boldsymbol{V}-\widetilde{\boldsymbol{V}}, \boldsymbol{L}-\widetilde{\boldsymbol{L}})$ in matrix form:

$$\begin{bmatrix}\widehat{\boldsymbol{V}}_t-\widetilde{\boldsymbol{V}}_t\\ \widehat{\boldsymbol{L}}_t-\widetilde{\boldsymbol{L}}_t\end{bmatrix} = \big\{\mathbf{I}_{2p}-\psi_1(t)\mathbf{R}_0-\mathbf{E}_0(t)\big\}\begin{bmatrix}\boldsymbol{V}_0-\widetilde{\boldsymbol{V}}_0\\ \boldsymbol{L}_0-\widetilde{\boldsymbol{L}}_0\end{bmatrix} + \begin{bmatrix}\varphi_2(t)\big(\nabla^2 f(\boldsymbol{L}_0)-\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\big)\boldsymbol{V}_0\\ \varphi_3(t)\big(\nabla^2 f(\boldsymbol{L}_0)-\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\big)\boldsymbol{V}_0\end{bmatrix}.$$

We now multiply the process $(\boldsymbol{V}-\widetilde{\boldsymbol{V}}, \boldsymbol{L}-\widetilde{\boldsymbol{L}})$ by $\mathbf{P}^{-1}$ introduced in (22). To this end, we first introduce the matrices

$$\mathbf{Q}_0 \triangleq \mathbf{P}^{-1}\mathbf{R}_0\mathbf{P}, \qquad \mathbf{N}_0(t) \triangleq \mathbf{P}^{-1}\mathbf{E}_0(t)\mathbf{P}.$$

After change of basis, the process rewrites

$$\mathbf{P}^{-1}\begin{bmatrix}\widehat{\boldsymbol{V}}_t-\widetilde{\boldsymbol{V}}_t\\ \widehat{\boldsymbol{L}}_t-\widetilde{\boldsymbol{L}}_t\end{bmatrix}=\big\{\mathbf{I}_{2p}-\psi_1(t)\mathbf{Q}_0-\mathbf{N}_0(t)\big\}\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_0-\widetilde{\boldsymbol{V}}_0\\ \boldsymbol{L}_0-\widetilde{\boldsymbol{L}}_0\end{bmatrix}+\mathbf{P}^{-1}\begin{bmatrix}\varphi_2(t)\big(\nabla^2 f(\boldsymbol{L}_0)-\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\big)\boldsymbol{V}_0\\ \varphi_3(t)\big(\nabla^2 f(\boldsymbol{L}_0)-\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\big)\boldsymbol{V}_0\end{bmatrix},$$

therefore, Minkowski's inequality yields

$$\left\|\mathbf{P}^{-1}\begin{bmatrix}\widehat{\boldsymbol{V}}_t-\widetilde{\boldsymbol{V}}_t\\ \widehat{\boldsymbol{L}}_t-\widetilde{\boldsymbol{L}}_t\end{bmatrix}\right\|_{\mathbb{L}_2}\le\left\|\big\{\mathbf{I}_{2p}-\psi_1(t)\mathbf{Q}_0-\mathbf{N}_0(t)\big\}\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_0-\widetilde{\boldsymbol{V}}_0\\ \boldsymbol{L}_0-\widetilde{\boldsymbol{L}}_0\end{bmatrix}\right\|_{\mathbb{L}_2}+\left\|\mathbf{P}^{-1}\begin{bmatrix}\varphi_2(t)\big(\nabla^2 f(\boldsymbol{L}_0)-\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\big)\boldsymbol{V}_0\\ \varphi_3(t)\big(\nabla^2 f(\boldsymbol{L}_0)-\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\big)\boldsymbol{V}_0\end{bmatrix}\right\|_{\mathbb{L}_2}. \tag{24}$$

In this last expression, the right-hand side of the inequality is a sum of two terms. Step 2 is devoted to the control of the first term while Step 3 focuses on the second one.

Step 2: To control the first term of the right-hand side of (24), our main task boils down to providing an upper bound for spectral norm of $\mathbf{I}_{2p}-\psi_1(t)\mathbf{Q}_0-\mathbf{N}_0(t)$.

The triangle inequality yields

$$\big\|\mathbf{I}_{2p}-\psi_1(t)\mathbf{Q}_0-\mathbf{N}_0(t)\big\|\le\big\|\mathbf{I}_{2p}-\psi_1(t)\mathbf{Q}_0\big\|+\big\|\mathbf{N}_0(t)\big\|.$$

Define $\alpha\triangleq\max(1-M/\gamma^2,3M/\gamma^2-1)$. We first prove that

$$\big\|\mathbf{I}_{2p}-\psi_1(t)\mathbf{Q}_0\big\|\le 1-\psi_1(t)(m/\gamma)+0.5\psi_1(t)^2M\big(\alpha+m^2/\big(M\gamma^2\big)\big).$$

To this end, we control the eigenvalues of

$$\big(\mathbf{I}_{2p}-\psi_1(t)\mathbf{Q}_0\big)\big(\mathbf{I}_{2p}-\psi_1(t)\mathbf{Q}_0\big)^\top=\mathbf{I}_{2p}-2\psi_1(t)\left(\frac{\mathbf{Q}_0+\mathbf{Q}_0^\top}{2}\right)+\psi_1(t)^2\mathbf{Q}_0\mathbf{Q}_0^\top. \tag{25}$$

For the purpose of notation, we introduce the following matrices:

$$\boldsymbol{\Sigma}_0\triangleq\gamma^{-1}\mathbf{H}_0,\qquad \mathbf{S}_0\triangleq\frac{\mathbf{Q}_0+\mathbf{Q}_0^\top}{2}.$$

Recall that $\mathbf{Q}_0=\mathbf{P}^{-1}\mathbf{R}_0\mathbf{P}$ where $\mathbf{P}$ is defined in (22) and $\mathbf{R}_0$ is defined in (23). Direct matrix computations yield

$$\mathbf{Q}_0=\begin{bmatrix}\boldsymbol{\Sigma}_0 & \boldsymbol{\Sigma}_0\\ -\boldsymbol{\Sigma}_0 & \gamma\mathbf{I}_p-\boldsymbol{\Sigma}_0\end{bmatrix},\qquad \mathbf{S}_0=\begin{bmatrix}\boldsymbol{\Sigma}_0 & 0_{p\times p}\\ 0_{p\times p} & \gamma\mathbf{I}_p-\boldsymbol{\Sigma}_0\end{bmatrix},$$

$$\mathbf{Q}_0\mathbf{Q}_0^\top=\begin{bmatrix}2\boldsymbol{\Sigma}_0^2 & \gamma\boldsymbol{\Sigma}_0-2\boldsymbol{\Sigma}_0^2\\ \gamma\boldsymbol{\Sigma}_0-2\boldsymbol{\Sigma}_0^2 & (\gamma\mathbf{I}_p-\boldsymbol{\Sigma}_0)^2+\boldsymbol{\Sigma}_0^2\end{bmatrix}.$$

We also introduce the symmetric matrix

$$\mathbf{U}_0 \triangleq \begin{bmatrix} \boldsymbol{\Sigma}_0^2 & \mathbf{H}_0 - 2\boldsymbol{\Sigma}_0^2 \\ \mathbf{H}_0 - 2\boldsymbol{\Sigma}_0^2 & \boldsymbol{\Sigma}_0^2 \end{bmatrix},$$

such that the following equality holds:

$$\mathbf{Q}_0\mathbf{Q}_0^\top = \mathbf{S}_0^2 + \mathbf{U}_0.$$

Combining this last equality with (25), we regroup the quadratic form as follows:

$$\big(\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0\big)\big(\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0\big)^\top = \big(\mathbf{I}_{2p} - \psi_1(t)\mathbf{S}_0\big)^2 + \psi_1(t)^2\mathbf{U}_0. \tag{26}$$

Lemma 5. *Assume that $\gamma^2 \geq m + M$, then the following holds*:

$$(m/\gamma)\mathbf{I}_{2p} \preceq \mathbf{S}_0 \preceq (\gamma - m/\gamma)\mathbf{I}_{2p}, \qquad \|\mathbf{U}_0\| \leq M\alpha.$$

Proof. The condition $\gamma^2 \geq m + M$ implies that $(m/\gamma)\mathbf{I}_p \preceq \boldsymbol{\Sigma}_0 \preceq (\gamma - m/\gamma)\mathbf{I}_p$. The first claim of the lemma follows directly.

Now, let us compute the eigenvalues of the symmetric matrix $\mathbf{U}_0$. We diagonalize $\mathbf{H}_0$ and note $(\lambda_j^{\mathbf{H}_0})_{j=1,\dots,p}$ its eigenvalues. By solving $\det(\mathbf{U}_0 - \lambda\mathbf{I}_{2p}) = 0$ we get p equations, i.e. for every $j = 1, \dots, p$ we need to solve:

$$\lambda^2 - 2a_j\lambda + a_j^2 - b_j^2 = 0, \qquad a_j = \big(\gamma^{-1}\lambda_j^{\mathbf{H}_0}\big)^2, \qquad b_j = \lambda_j^{\mathbf{H}_0} - 2a_j.$$

The solutions are $\lambda_j = a_j \pm |b_j|$. For every $j = 1, \dots, p$, we get

$$|\lambda_j| \leq \max\big(\lambda_j^{\mathbf{H}_0} - \big(\gamma^{-1}\lambda_j^{\mathbf{H}_0}\big)^2, 3\big(\gamma^{-1}\lambda_j^{\mathbf{H}_0}\big)^2 - \lambda_j^{\mathbf{H}_0}\big).$$

The function $x \mapsto \max(x - (x/\gamma)^2, 3(x/\gamma)^2 - x)$ is increasing on $\mathbb{R}_+$. Since $\lambda_j^{\mathbf{H}_0}$ is upper bounded by M, the second claim of the lemma follows. □

Now, we apply Lemma 5 to control the norm of (26). Since $\mathbf{S}_0$ and $\mathbf{U}_0$ are symmetric, we have

$$\begin{aligned} \big\|\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0\big\|^2 &\leq \big\|\big(\mathbf{I}_{2p} - \psi_1(t)\mathbf{S}_0\big)^2\big\| + \psi_1(t)^2\|\mathbf{U}_0\| \\ &\leq \big(1 - \psi_1(t)(m/\gamma)\big)^2 + \psi_1(t)^2 M\alpha \\ &= 1 - 2\psi_1(t)(m/\gamma) + \psi_1(t)^2 M\big(\alpha + m^2/\big(M\gamma^2\big)\big). \end{aligned}$$

For any $x \leq 1$, it holds that $\sqrt{1-x} \leq (1 - x/2)$. Applying this inequality yields

$$\big\|\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0\big\| \leq 1 - \psi_1(t)(m/\gamma) + 0.5\psi_1(t)^2 M\big(\alpha + m^2/\big(M\gamma^2\big)\big). \tag{27}$$

We now control the spectral norm of $\mathbf{N}_0(t)$. Recall that $\mathbf{N}_0(t) = \mathbf{P}^{-1}\mathbf{E}_0(t)\mathbf{P}$, where $\mathbf{P}$ is defined in (22), and $\mathbf{E}_0(t)$ is defined in (23). Direct matrix computation yields

$$\begin{aligned}
\mathbf{N}_0(t) &= \gamma^{-1}\begin{bmatrix} \mathbf{I}_p & \gamma\mathbf{I}_p \\ -\mathbf{I}_p & \mathbf{0}_{p\times p}\end{bmatrix}\begin{bmatrix} \varphi_2(t)\nabla^2 f(\widetilde{\boldsymbol{L}}_0) & \mathbf{0}_{p\times p} \\ \varphi_3(t)\nabla^2 f(\widetilde{\boldsymbol{L}}_0) & -\psi_2(t)\mathbf{H}_0\end{bmatrix}\begin{bmatrix} \mathbf{0}_{p\times p} & -\gamma\mathbf{I}_p \\ \mathbf{I}_p & \mathbf{I}_p\end{bmatrix} \\
&= \gamma^{-1}\begin{bmatrix} (\varphi_2+\gamma\varphi_3)\nabla^2 f(\widetilde{\boldsymbol{L}}_0) & \mathbf{0} \\ -\varphi_2\nabla^2 f(\widetilde{\boldsymbol{L}}_0) & \mathbf{0}\end{bmatrix}\begin{bmatrix} \mathbf{0} & -\gamma\mathbf{I}_p \\ \mathbf{I}_p & \mathbf{I}_p\end{bmatrix} - \psi_2(t)\gamma^{-1}\begin{bmatrix} \mathbf{0} & \gamma\mathbf{H}_0 \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix} \mathbf{0} & -\gamma\mathbf{I}_p \\ \mathbf{I}_p & \mathbf{I}_p\end{bmatrix} \\
&= \begin{bmatrix} \mathbf{0}_{p\times p} & -\big(\varphi_2(t)+\gamma\varphi_3(t)\big)\nabla^2 f(\widetilde{\boldsymbol{L}}_0) \\ \mathbf{0}_{p\times p} & \varphi_2(t)\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\end{bmatrix} - \psi_2(t)\begin{bmatrix} \mathbf{H}_0 & \mathbf{H}_0 \\ \mathbf{0}_{p\times p} & \mathbf{0}_{p\times p}\end{bmatrix}.
\end{aligned}$$

From definition (16), we know that for every $t > 0$, $\psi_2(t) \le \varphi_2(t) \le t^2/2$. Moreover,

$$\begin{aligned}
\varphi_2(t) + \gamma\varphi_3(t) &= \int_0^t e^{-\gamma(t-s)}\big(\psi_1(s) + \gamma\psi_2(s)\big)\,ds \\
&= \frac{1}{\gamma}\int_0^t e^{-\gamma(t-s)}\big(1 - e^{-\gamma s} + s\gamma - 1 + e^{-\gamma s}\big)\,ds \le t^2/2. \qquad (28)
\end{aligned}$$

Combining these inequalities with the fact that $\nabla^2 f(\widetilde{\boldsymbol{L}}_0)$ and $\mathbf{H}_0$ both have eigenvalues bounded by M, we get

$$\big\|\mathbf{N}_0(t)\big\| \le \sqrt{2}Mt^2. \qquad (29)$$

Summing the two upper bounds (27) and (29) yield

$$\big\|\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0 - \mathbf{N}_0(t)\big\| \le \rho_t \triangleq \left\{1 - \frac{\psi_1(t)m}{\gamma} + \frac{\psi_1(t)^2 M}{2}\left(\alpha + \frac{m^2}{M\gamma^2}\right) + M\sqrt{2}t^2\right\}.$$

From the definition (16), we know that $t - \gamma t^2/2 \le \psi_1(t) \le t$, therefore,

$$\rho_t \le 1 - \frac{mt}{\gamma} + \frac{Mt^2}{2}\underbrace{\left(\alpha + \frac{m^2}{M\gamma^2} + \frac{m}{M} + 2\sqrt{2}\right)}_{\le 2+2\sqrt{2}\le 5}.$$

Finally, we use the condition $t \le 1/(5\gamma\varkappa)$ to bound ρ_t by $1 - mt/(2\gamma)$. This yields

$$\left\|\big\{\mathbf{I}_{2p} - \psi_1(t)\mathbf{Q}_0 - \mathbf{N}_0(t)\big\}\mathbf{P}^{-1}\begin{bmatrix} \boldsymbol{V}_0 - \widetilde{\boldsymbol{V}}_0 \\ \boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0\end{bmatrix}\right\|_{\mathbb{L}_2} \le \left(1 - \frac{mt}{2\gamma}\right)\left\|\mathbf{P}^{-1}\begin{bmatrix} \boldsymbol{V}_0 - \widetilde{\boldsymbol{V}}_0 \\ \boldsymbol{L}_0 - \widetilde{\boldsymbol{L}}_0\end{bmatrix}\right\|_{\mathbb{L}_2}. \qquad (30)$$

Step 3: We now control the second term of the right-hand side of (24). Define

$$\zeta_2(t) \triangleq \sqrt{\big(\varphi_2(t) + \gamma\varphi_3(t)\big)^2 + \varphi_2(t)^2}.$$

From the definition of $\mathbf{P}^{-1}$ introduced in (22), direct computations yield

$$\left\|\mathbf{P}^{-1}\begin{bmatrix}\varphi_2(t)\big(\nabla^2 f(\boldsymbol{L}_0)-\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\big)\boldsymbol{V}_0\\ \varphi_3(t)\big(\nabla^2 f(\boldsymbol{L}_0)-\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\big)\boldsymbol{V}_0\end{bmatrix}\right\|_{\mathbb{L}_2}\le \xi_2(t)\big\|\big(\nabla^2 f(\boldsymbol{L}_0)-\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\big)\boldsymbol{V}_0\big\|_{\mathbb{L}_2}.$$

Moreover, using inequality (28) in the definition of $\zeta_2(t)$, we get the following bound:

$$\zeta_2(t)\le t^2/\sqrt{2}. \tag{31}$$

Our remaining task boils down to control the $\mathbb{L}_2$-norm of $(\nabla^2 f(\boldsymbol{L}_0)-\nabla^2 f(\widetilde{\boldsymbol{L}}_0))\boldsymbol{V}_0$.

By assumption, for every $x\in\mathbb{R}^p$, we have $m\mathbf{I}_p\preccurlyeq\nabla^2 f(x)\preccurlyeq M\mathbf{I}_p$, and the Hessian $\nabla^2 f$ is M_2-Lipschitz, therefore,

$$\big\|\big(\nabla^2 f(\boldsymbol{L}_0)-\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\big)\boldsymbol{V}_0\big\|_2\le \min\big(M-m, M_2\|\boldsymbol{L}_0-\widetilde{\boldsymbol{L}}_0\|_2\big)\|\boldsymbol{V}_0\|_2.$$

Using the inequality $\|\boldsymbol{V}_0\|_2^2\le a+(\|\boldsymbol{V}_0\|_2^2-a)_+$, for every $a>0$, this implies that

$$\begin{aligned}\big\|\big(\nabla^2 f(\boldsymbol{L}_0)-\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\big)\boldsymbol{V}_0\big\|_{\mathbb{L}_2}^2 &\le \mathbf{E}\big[\min\big((M-m)^2, M_2^2\|\boldsymbol{L}_0-\widetilde{\boldsymbol{L}}_0\|_2^2\big)\|\boldsymbol{V}_0\|_2^2\big]\\ &\le M_2^2 a\mathbf{E}\big[\|\boldsymbol{L}_0-\widetilde{\boldsymbol{L}}_0\|_2^2\big]+(M-m)^2\mathbf{E}\big[\big(\|\boldsymbol{V}_0\|_2^2-a\big)_+\big]\\ &\le M_2^2 a\|\boldsymbol{L}_0-\widetilde{\boldsymbol{L}}_0\|_{\mathbb{L}_2}^2+4(M-m)^2 e^{-(a-p)/4},\end{aligned} \tag{32}$$

where the last inequality is valid for every $a\ge 5p$ according to well-known bounds on the χ^2 distribution; see, for instance, (Collier and Dalalyan [13], Lemmas 5–6). Finally, recall that by the definition of $\mathbf{P}^{-1}$ in (22), we have

$$\|\boldsymbol{L}_0-\widetilde{\boldsymbol{L}}_0\|_{\mathbb{L}_2}\le \gamma^{-1}\sqrt{2}\left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_0-\widetilde{\boldsymbol{V}}_0\\ \boldsymbol{L}_0-\widetilde{\boldsymbol{L}}_0\end{bmatrix}\right\|_{\mathbb{L}_2}.$$

Using this last inequality in (32) and taking square roots yields the following bound:

$$\begin{aligned}\big\|\big(\nabla^2 f(\boldsymbol{L}_0)-\nabla^2 f(\widetilde{\boldsymbol{L}}_0)\big)\boldsymbol{V}_0\big\|_{\mathbb{L}_2}\le M_2\gamma^{-1}\sqrt{2a}\left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_0-\widetilde{\boldsymbol{V}}_0\\ \boldsymbol{L}_0-\widetilde{\boldsymbol{L}}_0\end{bmatrix}\right\|_{\mathbb{L}_2}\\ +2(M-m)e^{-(a-p)/8}.\end{aligned} \tag{33}$$

The claim of the proposition follows from combining the bound (30) from Step 2 with the bounds (31) and (33) from Step 3 into (24). □

The last piece of the proof of the theorem is the following proposition. It is obtained by combining the results of Propositions 2 and 3 to unfold the induction, and control the sampling error after k iterates.

Proposition 4. *Assume that, for some constants $m, M, M_2>0$, the function f is m-strongly convex, its gradient is M-Lipschitz and its Hessian is M_2-Lipschitz for the spectral norm. If the*

parameter γ and the step size h of the KLMC2 algorithm satisfy the inequalities

$$\gamma^2 \geq m + M, \qquad h \leq \frac{1}{5\gamma\varkappa} \wedge \frac{m}{4\sqrt{5p}M_2},$$

then

$$\begin{aligned}\left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_{kh}-\widetilde{\boldsymbol{V}}_{kh}\\ \boldsymbol{L}_{kh}-\widetilde{\boldsymbol{L}}_{kh}\end{bmatrix}\right\|_{\mathbb{L}_2} &\leq \left(1-\frac{mh}{4\gamma}\right)^k \left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_0-\widetilde{\boldsymbol{V}}_0\\ \boldsymbol{L}_0-\widetilde{\boldsymbol{L}}_0\end{bmatrix}\right\|_{\mathbb{L}_2}\\ &\quad+\frac{4\sqrt{2}(M-m)}{m}\gamma h e^{-\frac{m^2}{160M_2^2h^2}}\\ &\quad+\gamma h^2\left(\frac{M_2}{m}\sqrt{p^2+2p}+\frac{M^{3/2}}{m}\sqrt{p}\right).\end{aligned}$$

Proof. Minkowski's inequality yields

$$\left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_{kh}-\widetilde{\boldsymbol{V}}_{kh}\\ \boldsymbol{L}_{kh}-\widetilde{\boldsymbol{L}}_{kh}\end{bmatrix}\right\|_{\mathbb{L}_2} \leq \left\|\mathbf{P}^{-1}\begin{bmatrix}\widehat{\boldsymbol{V}}_{kh}-\widetilde{\boldsymbol{V}}_{kh}\\ \widehat{\boldsymbol{L}}_{kh}-\widetilde{\boldsymbol{L}}_{kh}\end{bmatrix}\right\|_{\mathbb{L}_2} + \left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_{kh}-\widehat{\boldsymbol{V}}_{kh}\\ \boldsymbol{L}_{kh}-\widehat{\boldsymbol{L}}_{kh}\end{bmatrix}\right\|_{\mathbb{L}_2}.$$

For $k \geq 0$, define

$$x_k = \left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_{kh}-\widetilde{\boldsymbol{V}}_{kh}\\ \boldsymbol{L}_{kh}-\widetilde{\boldsymbol{L}}_{kh}\end{bmatrix}\right\|_{\mathbb{L}_2}.$$

By Proposition 2 and Proposition 3, we thus have

$$\begin{aligned}x_{k+1} &\leq \left(1-\frac{mh}{2\gamma}+\frac{M_2\sqrt{a}h^2}{\gamma}\right)x_k + \sqrt{2}h^2(M-m)e^{-(a-p)/8}\\ &\quad+0.25h^3\big(M_2\sqrt{p^2+2p}+M^{3/2}\sqrt{p}\big).\end{aligned}$$

Assuming that $\sqrt{a} = m/(4M_2h) \geq \sqrt{5p}$ and unfolding the last recursion, we get

$$\begin{aligned}x_{k+1} &\leq \left(1-\frac{mh}{4\gamma}\right)^{k+1}x_0 + \frac{4\sqrt{2}(M-m)}{m}\gamma h e^{-(a-p)/8}\\ &\quad+\gamma h^2\left(\frac{M_2}{m}\sqrt{p^2+2p}+\frac{M^{3/2}}{m}\sqrt{p}\right).\end{aligned}$$

Easy algebra shows that

$$\frac{a-p}{8} = \frac{a}{10}+\frac{a-5p}{40} \geq \frac{a}{10} = \frac{m^2}{160M_2^2h^2}.$$

This is exactly the claim of the proposition. □

To complete the proof of Theorem 3, we need to do some simple algebra. First of all, using the relations

$$W_2(\nu_k,\pi)\le\gamma^{-1}\sqrt{2}\left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_{kh}-\widetilde{\boldsymbol{V}}_{kh}\\ \boldsymbol{L}_{kh}-\widetilde{\boldsymbol{L}}_{kh}\end{bmatrix}\right\|_{\mathbb{L}_2},\qquad W_2(\nu_0,\pi)=\gamma^{-1}\left\|\mathbf{P}^{-1}\begin{bmatrix}\boldsymbol{V}_{0}-\widetilde{\boldsymbol{V}}_{0}\\ \boldsymbol{L}_{0}-\widetilde{\boldsymbol{L}}_{0}\end{bmatrix}\right\|_{\mathbb{L}_2}$$

as well as the inequality $p^2+2p\le 2p^2$ (since $p\ge 2$), we arrive at

$$\begin{aligned}W_2(\nu_k,\pi)\le\sqrt{2}\left(1-\frac{mh}{4\gamma}\right)^k W_2(\nu_0,\pi)\\ +\frac{8(M-m)}{m}he^{-\frac{m^2}{160M_2^2h^2}}+\sqrt{2}h^2\left(\frac{M_2p}{m}\sqrt{2}+\frac{M^{3/2}}{m}\sqrt{p}\right).\end{aligned}$$

This leads to the claim of the theorem.

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

[1] Baker, J., Fearnhead, P., Fox, E.B. and Nemeth, C. (2019). Control variates for stochastic gradient MCMC. *Stat. Comput.* **29** 599–615. MR3969063 https://doi.org/10.1007/s11222-018-9826-2

[2] Bernton, E. (2018). Langevin Monte Carlo and JKO splitting. In *Proceedings of the* 31*st Conference on Learning Theory* (S. Bubeck, V. Perchet and P. Rigollet, eds.). *Proceedings of Machine Learning Research* **75** 1777–1798.

[3] Bolley, F., Guillin, A. and Malrieu, F. (2010). Trend to equilibrium and particle approximation for a weakly selfconsistent Vlasov–Fokker–Planck equation. *ESAIM Math. Model. Numer. Anal.* **44** 867–884. MR2731396 https://doi.org/10.1051/m2an/2010045

[4] Bou-Rabee, N. and Hairer, M. (2013). Nonasymptotic mixing of the MALA algorithm. *IMA J. Numer. Anal.* **33** 80–110. MR3020951 https://doi.org/10.1093/imanum/drs003

[5] Bou-Rabee, N., Eberle, A. and Zimmer, R. (2018). Coupling and convergence for Hamiltonian Monte Carlo.

[6] Brosse, N., Durmus, A., Moulines, E. and Pereyra, M. (2017). Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo. In *Proceedings of the* 2017 *Conference on Learning Theory* (S. Kale and O. Shamir, eds.). *Proceedings of Machine Learning Research* **65** 319–342. Amsterdam, Netherlands.

[7] Bubeck, S., Eldan, R. and Lehec, J. (2018). Sampling from a log-concave distribution with projected Langevin Monte Carlo. *Discrete Comput. Geom.* **59** 757–783. MR3802303 https://doi.org/10.1007/s00454-018-9992-1

[8] Chatterji, N., Flammarion, N., Ma, Y., Bartlett, P. and Jordan, M. (2018). On the theory of variance reduction for stochastic gradient Monte Carlo. In *Proceedings of the* 35*th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 764–773. Stockholm, Sweden: Stockholmsmassan.

[9] Chen, Z. and Vempala, S.S. (2019). Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions. In *Approximation*, *Randomization*, *and Combinatorial Optimization*. *Algorithms and Techniques*. *LIPIcs*. *Leibniz Int*. *Proc*. *Inform*. **145** Art. No. 64, 12. Wadern: Schloss Dagstuhl. Leibniz-Zent. Inform. MR4012714

[10] Cheng, X. and Bartlett, P. (2018). Convergence of Langevin MCMC in KL-divergence. In *Algorithmic Learning Theory* 2018. *Proc*. *Mach*. *Learn*. *Res*. (*PMLR*) **83** 26. MR3857306

[11] Cheng, X., Chatterji, N.S., Abbasi-Yadkori, Y., Bartlett, P.L. and Jordan, M.I. (2018). Sharp convergence rates for Langevin dynamics in the nonconvex setting. ArXiv e-prints.

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

[13] Collier, O. and Dalalyan, A.S. (2017). Minimax estimation of a p-dimensional linear functional in sparse gaussian models and robust estimation of the mean. Available at arXiv:1712.05495.

[14] Dalalyan, A. (2017). Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Proceedings of the* 2017 *Conference on Learning Theory* (S. Kale and O. Shamir, eds.). *Proceedings of Machine Learning Research* **65** 678–689.

[15] Dalalyan, A.S. (2017). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **79** 651–676. MR3641401 https://doi.org/10.1111/rssb.12183

[16] Dalalyan, A.S. and Karagulyan, A. (2019). User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Process*. *Appl*. **129** 5278–5311. MR4025705 https://doi.org/10.1016/j.spa.2019.02.016

[17] Desvillettes, L. and Villani, C. (2001). On the trend to global equilibrium in spatially inhomogeneous entropy-dissipating systems: The linear Fokker–Planck equation. *Comm*. *Pure Appl*. *Math*. **54** 1–42. MR1787105 https://doi.org/10.1002/1097-0312(200101)54:1<1::AID-CPA1>3.0.CO;2-Q

[18] Dieuleveut, A., Durmus, A. and Bach, F. (2017). Bridging the gap between constant step size stochastic gradient descent and Markov chains. ArXiv e-prints.

[19] Dolbeault, J., Mouhot, C. and Schmeiser, C. (2015). Hypocoercivity for linear kinetic equations conserving mass. *Trans*. *Amer*. *Math*. *Soc*. **367** 3807–3828. MR3324910 https://doi.org/10.1090/S0002-9947-2015-06012-7

[20] Douc, R., Moulines, E. and Rosenthal, J.S. (2004). Quantitative bounds on convergence of time-inhomogeneous Markov chains. *Ann*. *Appl*. *Probab*. **14** 1643–1665. MR2099647 https://doi.org/10.1214/105051604000000620

[21] Durmus, A., Majewski, S. and Miasojedow, B. (2019). Analysis of Langevin Monte Carlo via convex optimization. *J*. *Mach*. *Learn*. *Res*. **20** Paper No. 73, 46. MR3960927

[22] Durmus, A. and Moulines, É. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann*. *Appl*. *Probab*. **27** 1551–1587. MR3678479 https://doi.org/10.1214/16-AAP1238

[23] Durmus, A. and Moulines, É. (2019). High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli* **25** 2854–2882. MR4003567 https://doi.org/10.3150/18-BEJ1073

[24] Dwivedi, R., Chen, Y., Wainwright, M.J. and Yu, B. (2018). Log-concave sampling: Metropolis–Hastings algorithms are fast! In *Proceedings of the* 31*st Conference on Learning Theory*. *Proceedings of Machine Learning Research* **75** 793–797.

[25] Eberle, A., Guillin, A. and Zimmer, R. (2019). Couplings and quantitative contraction rates for Langevin dynamics. *Ann*. *Probab*. **47** 1982–2010. MR3980913 https://doi.org/10.1214/18-AOP1299

[26] Griewank, A. (1993). Some bounds on the complexity of gradients, Jacobians, and Hessians. In *Complexity in Numerical Optimization* 128–162. River Edge, NJ: World Sci. Publ. MR1358844

[27] Helffer, B. and Nier, F. (2005). *Hypoelliptic Estimates and Spectral Theory for Fokker–Planck Operators and Witten Laplacians*. *Lecture Notes in Math*. **1862**. Berlin: Springer. MR2130405 https://doi.org/10.1007/b104762

[28] Lamberton, D. and Pagès, G. (2002). Recursive computation of the invariant distribution of a diffusion. *Bernoulli* **8** 367–405. MR1913112 https://doi.org/10.1142/S0219493703000838

[29] Lamberton, D. and Pagès, G. (2003). Recursive computation of the invariant distribution of a diffusion: The case of a weakly mean reverting drift. *Stoch. Dyn.* **3** 435–451. MR2030742 https://doi.org/10.1142/S0219493703000838

[30] Luu, T.D., Fadili, J. and Chesneau, C. (2017). Sampling from non-smooth distribution through Langevin diffusion. Working paper or Preprint.

[31] Mangoubi, O. and Smith, A. (2019). Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 2: Numerical integrators. In *Proceedings of Machine Learning Research* (K. Chaudhuri and M. Sugiyama, eds.). *Proceedings of Machine Learning Research* **89** 586–595. PMLR.

[32] Nelson, E. (1967). *Dynamical Theories of Brownian Motion*. Princeton, N.J.: Princeton Univ. Press. MR0214150

[33] Pavliotis, G.A. (2014). *Stochastic Processes and Applications*: *Diffusion Processes*, *the Fokker-Planck and Langevin Equations*. *Texts in Applied Mathematics* **60**. New York: Springer. MR3288096 https://doi.org/10.1007/978-1-4939-1323-7

[34] Pillai, N.S., Stuart, A.M. and Thiéry, A.H. (2012). Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *Ann. Appl. Probab.* **22** 2320–2356. MR3024970 https://doi.org/10.1214/11-AAP828

[35] Raginsky, M., Rakhlin, A. and Telgarsky, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *Proceedings of the* 2017 *Conference on Learning Theory* (S. Kale and O. Shamir, eds.). *Proceedings of Machine Learning Research* **65** 1674–1703. Amsterdam, Netherlands.

[36] 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** 255–268. MR1625691 https://doi.org/10.1111/1467-9868.00123

[37] Roberts, G.O. and Stramer, O. (2002). Langevin diffusions and Metropolis–Hastings algorithms *Methodol. Comput. Appl. Probab.*. **4** 337–357. MR2002247 https://doi.org/10.1023/A:1023562417138

[38] Roberts, G.O. and Tweedie, R.L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2** 341–363. MR1440273 https://doi.org/10.2307/3318418

[39] Stramer, O. and Tweedie, R.L. (1999). Langevin-type models. I. Diffusions with given stationary distributions and their discretizations. *Methodol. Comput. Appl. Probab.* **1** 283–306. MR1730651 https://doi.org/10.1023/A:1010086427957

[40] Stramer, O. and Tweedie, R.L. (1999). Langevin-type models. II. Self-targeting candidates for MCMC algorithms. *Methodol. Comput. Appl. Probab.* **1** 307–328. MR1730652 https://doi.org/10.1023/A:1010090512027

[41] Xifara, T., Sherlock, C., Livingstone, S., Byrne, S. and Girolami, M. (2014). Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statist. Probab. Lett.* **91** 14–19. MR3208109 https://doi.org/10.1016/j.spl.2014.04.002

[42] Xu, P., Chen, J., Zou, D. and Gu, Q. (2017). Global convergence of Langevin dynamics based algorithms for nonconvex optimization. ArXiv e-prints.

[43] Zhang, Y., Liang, P. and Charikar, M. (2017). A hitting time analysis of stochastic gradient Langevin dynamics. In *Proceedings of the* 2017 *Conference on Learning Theory* (S. Kale and O. Shamir, eds.). *Proceedings of Machine Learning Research* **65** 1980–2022. Amsterdam, Netherlands.

Received November 2018 and revised October 2019