Bernoulli **27**(1), 2021, 1–33
https://doi.org/10.3150/19-BEJ1187

On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case

M. BARKHAGEN[1,*], N.H. CHAU[2], É. MOULINES[3,4], M. RÁSONYI[5], S. SABANIS[1,6,†] and Y. ZHANG[1,‡]

[1]*School of Mathematics, The University of Edinburgh, UK.*
E-mail: [*]*n.barkhagen@ed.ac.uk;* [†]*s.sabanis@ed.ac.uk;* [‡]*ying.zhang@ed.ac.uk*
[2]*Center for Mathematical Modeling and Data Science, Osaka University, Japan.*
E-mail: chau@sigmath.es.osaka-u.ac.jp
[3]*Centre de Mathématiques Appliquées, UMR 7641, École Polytechnique, France.*
E-mail: eric.moulines@polytechnique.edu
[4]*HSE University, Russian Federation*
[5]*Alfred Renyi Institue of Mathematics, Hungary. E-mail: rasonyi.miklos@renyi.hu*
[6]*The Alan Turing Institute, UK*

We study the problem of sampling from a probability distribution π on $\mathbb{R}^d$ which has a density w.r.t. the Lebesgue measure known up to a normalization factor $x \mapsto \mathrm{e}^{-U(x)}/\int_{\mathbb{R}^d} \mathrm{e}^{-U(y)}\,\mathrm{d}y$. We analyze a sampling method based on the Euler discretization of the Langevin stochastic differential equations under the assumptions that the potential U is continuously differentiable, ∇U is Lipschitz, and U is strongly concave. We focus on the case where the gradient of the log-density cannot be directly computed but unbiased estimates of the gradient from possibly dependent observations are available. This setting can be seen as a combination of a stochastic approximation (here stochastic gradient) type algorithms with discretized Langevin dynamics. We obtain an upper bound of the Wasserstein-2 distance between the law of the iterates of this algorithm and the target distribution π with constants depending explicitly on the Lipschitz and strong convexity constants of the potential and the dimension of the space. Finally, under weaker assumptions on U and its gradient but in the presence of independent observations, we obtain analogous results in Wasserstein-2 distance.

Keywords: L-mixing; Langevin diffusion; Monte Carlo methods; stochastic approximation

1. Introduction

Sampling target distributions is an important topic in statistics and applied probability. In this paper, we are concerned with sampling from a distribution π defined by

$$\pi(A) := \int_A \mathrm{e}^{-U(\theta)}\,\mathrm{d}\theta \Big/ \int_{\mathbb{R}^d} \mathrm{e}^{-U(\theta)}\,\mathrm{d}\theta, \quad A \in \mathcal{B}\big(\mathbb{R}^d\big),$$

where $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel sets of $\mathbb{R}^d$ and $U : \mathbb{R}^d \to \mathbb{R}_+$ is continuously differentiable.

One of the sampling schemes considered in this paper is the *unadjusted* Langevin algorithm (a.k.a. Langevin Monte Carlo). The idea is to construct a Markov chain which is the Euler discretization of a continuous-time diffusion process that has an invariant distribution π.

We work on a fixed probability space $(\Omega, \mathcal{F}, P)$ throughout the paper. We consider the so-called overdamped Langevin stochastic differential equation (SDE)

$$\mathrm{d}\theta_t = -h(\theta_t)\,dt + \sqrt{2}\,\mathrm{d}B_t, \tag{1}$$

with a (possibly random) initial condition θ_0, where $h := \nabla U$ and $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. It is well known that, under appropriate conditions, the Markov semigroup associated with the Langevin diffusion (1) is reversible with respect to π, and the rate of convergence to π is geometric in the total variation norm (see [21,27], Theorem 1.2, and [2], Theorem 1.6). The Euler–Maruyama discretization scheme for SDE (1), which is referred to as the unadjusted Langevin algorithm (ULA), is given by

$$\overline{\theta}_0^{\lambda} := \theta_0, \qquad \overline{\theta}_{n+1}^{\lambda} := \overline{\theta}_n^{\lambda} - \lambda h\big(\overline{\theta}_n^{\lambda}\big) + \sqrt{2\lambda}\xi_{n+1}, \tag{2}$$

where $(\xi_n)_{n\in\mathbb{N}}$ is a sequence of independent, standard d-dimensional Gaussian random variables, $\lambda > 0$ is the step size and θ_0 is an $\mathbb{R}^d$-valued random variable denoting the initial values of both (2) and (1). Under appropriate assumptions on the step size λ and the potential U, the homogeneous Markov chain $(\overline{\theta}_n^{\lambda})_{n\in\mathbb{N}}$ converges to a distribution π_λ which differs from π but, for small λ, it is close to π in an appropriate sense; see [7,8,10], and Section 4.1.

We now adopt a framework where the exact gradient h is unknown, however one can observe at each iteration an unbiased estimator. Let $H : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ be a measurable function and $X := (X_n)_{n\in\mathbb{N}}$ an $\mathbb{R}^m$-valued process adapted to some given filtration $\mathcal{G}_n$, $n \in \mathbb{N}$ satisfying

$$h(\theta) = \mathbb{E}\big[H(\theta, X_n)\big], \quad \theta \in \mathbb{R}^d, n \geq 1, \tag{3}$$

where the existence of the expectation is implicitly assumed. Note that (3) holds, in particular, when $(X_n)_{n\geq 1}$ is a strictly stationary process. Denoting by μ the (common) distribution of X_n, $n \geq 1$, we may write

$$h(\theta) = \int H(\theta, x)\mu(\mathrm{d}x), \tag{4}$$

in this case. We also assume henceforth that θ_0, $\mathcal{G}_\infty$, $(\xi_n)_{n\in\mathbb{N}}$ are independent.

For each $\lambda > 0$, define an $\mathbb{R}^d$-valued random process $(\theta_n^{\lambda})_{n\in\mathbb{N}}$ by recursion:

$$\theta_0^{\lambda} := \theta_0, \qquad \theta_{n+1}^{\lambda} := \theta_n^{\lambda} - \lambda H\big(\theta_n^{\lambda}, X_{n+1}\big) + \sqrt{2\lambda}\xi_{n+1}. \tag{5}$$

Such a sampling scheme is often called a stochastic gradient Langevin dynamics (SGLD) algorithm; see [8,30] and [28]. Data sequences $(X_n)_{n\in\mathbb{N}}$ are in general not i.d.d., not even Markovian. They may exhibit strong non-Markovian features as it is observed in various stochastic phenomena. Stochastic approximation for dependent data sequences (gradient and Kiefer–Wolfowitz procedures) has been successfully used in financial applications, see [18,32] and the references therein. With these examples in mind, in the present paper we seek theoretical guarantees for the convergence of the closely related SGLD procedure to ensure its validity for non-independent data sets, too.

The only instance we know of that provides results in such a setting is Theorem 4 of [8]. The main condition of that result (Condition N in [8]) requires estimates on the conditional bias and variance of the updating function with respect to the previous iterate of the recursion (5), see Section 3.3 for extensive discussions. In concrete examples, it seems very difficult to determine the order of these quantities. We follow a different path. Intuitively, if the signal X_n is "sufficiently ergodic" then one should be able to estimate the sampling error, without checking conditions on the conditional bias/variance of specific objects. We will assume a certain mixing condition, *conditional L-mixing* for the data sequence $(X_n)_{n\in\mathbb{N}}$; see Section 2 below for technical details. Theorem 3.5 is obtained which guarantees an (essentially) optimal estimate in terms of the stepsize. Our approach involves several new ideas which serve as a basis for further developments in the case of non-convex U, see [5].

The goal of this work is to establish an upper bound on the Wasserstein distance between the target distribution π and its approximations $(\mathrm{Law}(\theta_n^{\lambda}))_{n\in\mathbb{N}}$ generated by the SGLD algorithm (5). This goal

is achieved while the rate of convergence is improved with respect to the findings in [24], see also [6,31] and [8]. We stress that we prove the validity of sampling procedures driven by SGLD (5) within a framework where $(X_n)_{n\in\mathbb{N}}$ are not assumed i.d.d. and hence θ_n^λ fails to be Markovian and related techniques are not applicable. Algorithms for variance reduction of SGLD have been suggested by [3,31], however, we do not see for the moment how these could be treated by our methods here.

The paper is organized as follows. Section 2 recalls the theoretical concept of conditional L-mixing which is required below for the process $(X_n)_{n\in\mathbb{N}}$. This notion accommodates a large class of (possibly non-Markovian) processes. In Section 3, assumptions and main results are presented in the case where the process $(X_n)_{n\in\mathbb{N}}$ is conditionally L-mixing (Section 3.1) and i.d.d. (Section 3.2), respectively. In Section 3.3, we discuss the contributions of our work with respect to existing results reported in the literature. In Section 4.1 and Section 4.2, the properties of (1), (2), and (5) are analyzed. The proofs of the main theorems are provided in Sections 4 and 5, while certain auxiliary results are presented in Sections A and B.

Notations and conventions. Scalar product in $\mathbb{R}^d$ is denoted by $\langle\cdot,\cdot\rangle$. We use $\|\cdot\|$ to denote the Euclidean norm (where the dimension of the space may vary). $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel σ-field of $\mathbb{R}^d$. For each $x_0\in\mathbb{R}^d$ and $R\geq 0$, we denote $\mathrm{B}(x_0,R):=\{x\in\mathbb{R}^d : \|x-x_0\|\leq R\}$, the closed ball of radius R centered at x_0. For two sigma algebras $\mathcal{F}_1$, $\mathcal{F}_2$, we define $\mathcal{F}_1\vee\mathcal{F}_2:=\sigma(\mathcal{F}_1\cup\mathcal{F}_2)$. The expectation of a random variable X is denoted by $\mathbb{E}[X]$. For any $m\geq 1$, for any $\mathbb{R}^m$-valued random variable X and for any $1\leq p<\infty$, we set $\|X\|_p:=\mathbb{E}^{1/p}[\|X\|^p]$. We denote by L^p the set of X with $\|X\|_p<\infty$. The indicator function of a set A is denoted by $\mathbb{1}_A$. The Wasserstein distance of order $p\geq 1$ between two probability measures μ and ν on $\mathcal{B}(\mathbb{R}^d)$ is defined by

$$W_p(\mu,\nu)=\left(\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathcal{X}}\|x-y\|^p\,\mathrm{d}\pi(x,y)\right)^{1/p}, \tag{6}$$

where $\Pi(\mu,\nu)$ is the set of couplings of (μ,ν), see, for example, [29].

2. Conditional L-mixing

L-mixing processes and random fields were introduced in [12]. They proved to be useful in tackling difficult problems of system identification, see, for example, [13–16,25]. In [4], in the context of stochastic gradient methods, the related concept of *conditional* L-mixing was introduced. We now recall its definition below.

We consider the probability space $(\Omega,\mathcal{F},P)$, equipped with a discrete-time filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ as well as with a decreasing sequence of sigma-fields $(\mathcal{F}_n^+)_{n\in\mathbb{N}}$ such that $\mathcal{F}_n$ is independent of $\mathcal{F}_n^+$, for all $n\in\mathbb{N}$.

For a family $(Z_i)_{i\in I}$ of real-valued random variables (where the index set I may have arbitrary cardinality), there exists one and (up to a.s. equality) only one random variable $g=\operatorname{ess\,sup}_{i\in I} Z_i$ such that:

(i) $g\geq Z_i$, a.s. for all $i\in I$,
(ii) if g' is a random variable, $g'\geq Z_i$, a.s. for all $i\in I$ then $g'\geq g$ P-a.s.,

see, for example, [23], Proposition VI.1.1.

Fix an integer $d\geq 1$ and let $D\subset\mathbb{R}^d$ be a set of parameters. A measurable function $U:\mathbb{N}\times D\times\Omega\to\mathbb{R}^k$ is called a random field. We drop dependence on $\omega\in\Omega$ in the notation henceforth and write

$(U_n(\theta))_{n\in\mathbb{N},\theta\in D}$. A random process $(U_n)_{n\in\mathbb{N}}$ corresponds to a random field where D is a singleton. A random field is L^r*-bounded* for some $r \geq 1$ if

$$\sup_{n\in\mathbb{N}}\sup_{\theta\in D}\big\|U_n(\theta)\big\|_r < \infty.$$

Let $U_n(\theta) \in L^1$, $n \in \mathbb{N}$, $\theta \in D$ and U^i_{n+m} is the i-th coordinate of U_{n+m}. Define, for each $n \in \mathbb{N}$, $i = 1, \dots, k$, and $\tau \in \mathbb{N}$

$$\tilde{M}^n_r(U,i) := \operatorname{ess\,sup}_{\theta\in D}\sup_{m\in\mathbb{N}}\mathbb{E}^{1/r}\big[\big|U^i_{n+m}(\theta)\big|^r|\mathcal{F}_n\big], \tag{7}$$

$$\tilde{\gamma}^n_r(\tau,U,i) := \operatorname{ess\,sup}_{\theta\in D}\sup_{m\geq\tau}\mathbb{E}^{1/r}\big[\big|U^i_{n+m}(\theta) - \mathbb{E}\big[U^i_{n+m}(\theta)|\mathcal{F}^+_{n+m-\tau}\vee\mathcal{F}_n\big]\big|^r|\mathcal{F}_n\big], \tag{8}$$

and set

$$\begin{aligned}\tilde{\Gamma}^n_r(U,i) &:= \sum_{\tau=0}^{\infty}\tilde{\gamma}^n_r(\tau,U,i), \qquad M^n_r(U) := \sum_{i=1}^{k}\tilde{M}^n_r(U,i), \quad \text{and}\\ \Gamma^n_r(U) &:= \sum_{i=1}^{k}\tilde{\Gamma}^n_r(U,i).\end{aligned} \tag{9}$$

When necessary, the notations $M^n_r(U,D)$, $\gamma^n_r(\tau,U,D)$ and $\Gamma^n_r(U,D)$ are used to emphasize dependence of these quantities on the domain D which may vary.

Definition 2.1 (Conditional L-mixing). Let $r, s \geq 1$. We say that the random field $(U_n(\theta))_{n\in\mathbb{N},\theta\in D}$ is *uniformly conditionally L-mixing* (UCLM) of order (r,s) with respect to $(\mathcal{F}_n,\mathcal{F}^+_n)_{n\in\mathbb{N}}$ if $(U_n(\theta))_{n\in\mathbb{N}}$ is adapted to $(\mathcal{F}_n)_{n\in\mathbb{N}}$ for any $\theta \in D$; it is L^r-bounded; and the sequences $(M^n_r(U))_{n\in\mathbb{N}}$, $(\Gamma^n_r(U))_{n\in\mathbb{N}}$ are L^s-bounded. When this holds for all $r, s \geq 1$ then we call the random field simply "uniformly L-mixing". In the case of stochastic processes (when D is a singleton) the terminology "conditionally L-mixing process (of order (r,s))" is used.

Remark 2.2. The definition of conditional L-mixing in [4] is slightly different from the definition above but they are clearly equivalent.

Although we do not use the concept of L-mixing in the present paper it is worth noting that the definition of a uniformly L-mixing process follows naturally from the above definition if one sets $d = 1$, $n = 0$ and $\mathcal{F}_n$ is replaced by the trivial σ-algebra in the definitions of $M^n_r(U)$, $\gamma^n_r(\tau,U)$ and $\Gamma^n_r(U)$. Then, one obtains deterministic $M_r(U)$, $\gamma_r(\tau,U)$, $\Gamma_r(U)$ and the required condition for these quantities becomes $M_r(U) + \Gamma_r(U) < \infty$. For more details, one can consult [4] and [12].

Let $(U_n)_{n\in\mathbb{N}}$ be a conditionally L-mixing process. For later use, we also introduce the quantities for $r, s \geq 1$,

$$\mathcal{M}_r(U) := \sup_{n\in\mathbb{N}}\mathbb{E}\big[\|U_n\|^r\big], \qquad \mathcal{C}_{r,s}(U) := \sup_{n\in\mathbb{N}}\mathbb{E}\big[\big\{\Gamma^n_r(U)\big\}^s\big]. \tag{10}$$

The interpretation of $\mathcal{M}_r(U)$ is straightforward while $\mathcal{C}_{r,s}(U)$ serves as a certain measure of dependence for the process U.

Example 2.3. Let $(X_n)_{n\in\mathbb{N}}$ be i.d.d. random variables ($d = 1$) and set $\mathcal{F}_n := \sigma(X_k, k \leq n)$, $\mathcal{F}^+_n := \sigma(X_k, k > n)$, $n \in \mathbb{N}$. If $\mathbb{E}[|X_0|^r] < \infty$ for any $r \geq 1$, then $(X_n)_{n\in\mathbb{N}}$ is conditionally L-mixing with

respect to $(\mathcal{F}_n, \mathcal{F}_n^+)_{n\in\mathbb{N}}$. Moreover,

$$\mathcal{M}_r(X) = \mathbb{E}\big[|X_0|^r\big], \qquad \mathcal{C}_{r,s}(X) = \mathbb{E}^{s/r}\big[\big|X_0 - \mathbb{E}[X_0]\big|^r\big]r, \quad s \geq 1. \tag{11}$$

Example 2.4. Let us consider, for example, a functional of a linear process $U := \{U_n(\theta)\}_{n\in\mathbb{N}}$, such that

$$U_n(\theta) := G(\theta, X_n), \qquad X_n := \sum_{k=0}^{\infty} a_k \varepsilon_{n-k}, \tag{12}$$

with scalars $(a_k)_{k\in\mathbb{N}}$, some sequence $(\varepsilon_k)_{k\in\mathbb{Z}}$ of i.d.d. $\mathbb{R}$-valued random variables satisfying $\|\varepsilon_0\|_p < \infty$ for all $p \geq 1$ and $G : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ a function satisfying

$$\big|G(\theta, x) - G\big(\theta', x'\big)\big| \leq L_1\big|\theta - \theta'\big| + L_2\big|x - x'\big|.$$

Let $\mathcal{G}_n = \sigma(\varepsilon_j, j \leq n)$, and $\mathcal{G}_n^+ = \sigma(\varepsilon_j, j > n)$ for $n \in \mathbb{N}$. If we further assume that $|a_k| \leq c(1 + k)^{-\beta}$, $k \in \mathbb{N}$ for some $c > 0$, $\beta > 3/2$ then the argument of [4], Lemma 4.2, shows that $(X_n)_{n\in\mathbb{N}}$ is a conditionally L-mixing process with respect to $(\mathcal{G}_n, \mathcal{G}_n^+)_{n\in\mathbb{N}}$. Applying Lemma 4.7 below with $\vartheta = 0$ shows that for all $j \in \mathbb{N}$, $M_r^n(U, \mathrm{B}(0, j))) \leq L_1 j + L_2 M_r^n(X) + |G(0, 0)|$ and $\Gamma_r^n(U, \mathrm{B}(0, j)) \leq 2L_2\Gamma_r^n(X)$.

Remark 2.5. If $(X_n)_{n\in\mathbb{N}}$ is a conditionally L-mixing process with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)_{n\in\mathbb{N}}$ then so is $(F(X_n))_{n\in\mathbb{N}}$ for any Lipschitz-continuous function F, see [4], Remark 2.3. Finally, we know from [11], Example 7.1, that a broad class of functionals of geometrically ergodic Markov chains have the L-mixing property. It is possible to show, along the same lines, the conditional L-mixing property of these functionals, too.

3. Assumptions and main results

3.1. Dependent data

Assumption 3.1. Let $\mathcal{G}_0 := \{\varnothing, \Omega\}$. The process $(X_n)_{n\in\mathbb{N}}$ is conditionally L-mixing with respect to $(\mathcal{G}_n, \mathcal{G}_n^+)_{n\in\mathbb{N}}$, where $(\mathcal{G}_n^+)_{n\in\mathbb{N}}$ is some decreasing sequence of sigma-fields with $\mathcal{G}_n$ independent of $\mathcal{G}_n^+$ for all $n \in \mathbb{N}$. Furthermore, let $\|\theta_0\|_p < \infty$ for all $p \geq 1$.

For $(x, \theta) \in \mathbb{R}^m \times \mathbb{R}^d$, we denote $H(x, \theta) = [H^1(x, \theta), \ldots, H^d(x, \theta)]^T$.

Assumption 3.2. There exist constants $L_1^i, L_2^i > 0$, $i \in \{1, \ldots, d\}$ such that for all $\theta, \theta' \in \mathbb{R}^d$ and $x, x' \in \mathbb{R}^m$, $|H^i(\theta, x) - H^i(\theta', x')| \leq L_1^i\|\theta - \theta'\| + L_2^i\|x - x'\|$.

We set

$$L_1 = \sum_{i=1}^{d} L_1^i \quad \text{and} \quad L_2 = \sum_{i=1}^{d} L_2^i. \tag{13}$$

Note that, under Assumption 3.2, for any $(x, \theta) \in \mathbb{R}^m \times \mathbb{R}^d$ we get

$$\big\|H(x, \theta) - H\big(x, \theta'\big)\big\| \leq L_1\big\|\theta - \theta'\big\| + L_2\big\|x - x'\big\|.$$

Assumption 3.1 implies, in particular, that $\|X_0\| \in L^r$, for any $r \geq 1$, thus, under Assumption 3.1 and 3.2, $h(\theta) := \mathbb{E}[H(\theta, X_0)]$, $\theta \in \mathbb{R}^d$, is indeed well-defined.

Assumption 3.3. There is a constant $a > 0$ such that for all $\theta, \theta' \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$,

$$\big\langle \theta - \theta', H(\theta, x) - H\big(\theta', x\big)\big\rangle \geq a\big\|\theta - \theta'\big\|^2. \tag{14}$$

Two important properties immediately follow from Assumptions 3.2 and 3.3.

(B1) For all $\theta, \theta' \in \mathbb{R}^d$, $\|h(\theta) - h(\theta')\| \leq L_1\|\theta - \theta'\|$.
(B2) There exists a constant $a > 0$ such that, for all $\theta, \theta' \in \mathbb{R}^d$, $\langle \theta - \theta', h(\theta) - h(\theta')\rangle \geq a\|\theta - \theta'\|^2$.

[22], Theorem 2.1.12, shows that, under these assumptions, for all $\theta, \theta' \in \mathbb{R}^d$,

$$\big\langle \theta - \theta', h(\theta) - h\big(\theta'\big)\big\rangle \geq \tilde{a}\big\|\theta - \theta'\big\|^2 + \frac{1}{a + L_1}\big\|h(\theta) - h\big(\theta'\big)\big\|^2, \tag{15}$$

where we have set

$$\tilde{a} = \frac{aL_1}{a + L_1}. \tag{16}$$

Our aim initially is to estimate $\|\theta_n^\lambda - \overline{\theta}_n^\lambda\|_2$, uniformly in $n \in \mathbb{N}$. To begin with, an example is presented where explicit calculations are possible.

Example 3.4. Let $d := 1$, $H(\theta, x) := \theta + x$, $(X_n)_{n\in\mathbb{Z}}$ be a sequence of satisfying (12) with $(\epsilon_j)_{j\in\mathbb{Z}}$ an independent sequence of standard Gaussian random variables independent of $(\xi_n)_{n\in\mathbb{N}}$; and $|a_k| \leq c(1 + k)^{-\beta}$, $k \in \mathbb{N}$ for some $\beta > 3/2$ and

$$0 < m := \inf_{\mu\in[-\pi,\pi]}\left|\sum_{k=0}^{\infty} a_k \mathrm{e}^{-\mathrm{i}\mu k}\right| \leq \sup_{\mu\in[-\pi,\pi]}\left|\sum_{k=0}^{\infty} a_k \mathrm{e}^{-\mathrm{i}\mu k}\right| \leq M < \infty. \tag{17}$$

We observe that the function H satisfies Assumptions 3.2 and 3.3. Take $\theta_0 := 0$. It is straightforward to check that, for any $\lambda \in (0, 1)$,

$$\overline{\theta}_n^\lambda - \theta_n^\lambda = \sum_{j=0}^{n-1}(1 - \lambda)^j \lambda X_{n-j},$$

which clearly has variance

$$\mathbb{E}\big[\big(\overline{\theta}_n^\lambda - \theta_n^\lambda\big)^2\big] = \frac{\lambda^2}{2\pi}\int_{-\pi}^{\pi}\left|\sum_{k=0}^{\infty} a_k \mathrm{e}^{-\mathrm{i}k\mu}\right|^2\left|\sum_{k=0}^{n-1}(1 - \lambda)^k \mathrm{e}^{-\mathrm{i}k\mu}\right|^2 \mathrm{d}\mu$$

It follows that, using (17) and the Parseval–Plancherel theorem

$$m\sqrt{\frac{\lambda\{1 - (1 - \lambda)^{2n}\}}{2 - \lambda}} \leq \big\|\overline{\theta}_n^\lambda - \theta_n^\lambda\big\|_2 \leq M\sqrt{\frac{\lambda\{1 - (1 - \lambda)^{2n}\}}{2 - \lambda}}.$$

This shows that the best estimate we may hope to obtain for $\sup_{n\in\mathbb{N}}\|\overline{\theta}_n^\lambda - \theta_n^\lambda\|_2$ is of the order $\sqrt{\lambda}$. Theorem 3.5 below achieves this bound asymptotically as $p \to \infty$.

Our main results may be stated as follows.

Theorem 3.5. *Let Assumptions 3.1, 3.2 and 3.3 hold. For every even number $p \geq 4$ and $\lambda < \bar{\lambda}$, where*

$$\bar{\lambda} := \frac{2}{a + L_1}, \tag{18}$$

there exists $C_0(p) > 0$ such that

$$\left\| \theta_n^\lambda - \overline{\theta}_n^\lambda \right\|_2 \leq C_0(p) \lambda^{\frac{1}{2} - \frac{1}{p}}, \quad n \in \mathbb{N} \tag{19}$$

holds for a constant $C_0(p)$ that is explicitly given in the proof. It depends only on a, L_1, L_2, d, p and on the process $(X_n)_{n\in\mathbb{N}}$ through the quantities defined in (10).

Proof. The proof of this theorem is postponed to Section 4.3. □

The next result relates our findings in Theorems 3.5 to the problem of sampling from the probability law π.

Theorem 3.6. *Let Assumptions 3.1, 3.2 and 3.3 hold and let $\bar{\lambda}$ be given by (18). For each $\kappa > 0$, there exist constants $c_1(\kappa), c_2(\kappa) > 0$ such that, for each $0 < \epsilon \leq \mathrm{e}^{-1}$ one has*

$$W_2\big(\mathrm{Law}\big(\theta_n^\lambda\big), \pi\big) \leq \epsilon$$

whenever $\lambda < \bar{\lambda}$ satisfies

$$\lambda = c_1(\kappa)\epsilon^{2+\kappa} \quad \textit{and} \quad n \geq \frac{c_2(\kappa)}{\epsilon^{2+\kappa}} \ln(1/\epsilon), \tag{20}$$

where $c_1(\kappa)$, $c_2(\kappa)$ (given explicitly in the proof) depend only on κ, d, a, L_1, L_2 and on the process $(X_n)_{n\in\mathbb{N}}$ through the quantities defined in (10).

Proof. The proof of this theorem is postponed to Section 4.4. □

3.2. Independent data

When the data sequences $(X_n)_{n\in\mathbb{Z}}$ are i.d.d., then the full rate is recovered under more relaxed conditions for the unbiased estimator of the gradient of U. More concretely, one may assume the following assumption.

Assumption 3.7. There exist positive constants L_1, L_2 and ρ such that, for all $x, x' \in \mathbb{R}^m$ and $\theta, \theta' \in \mathbb{R}^d$,

$$\begin{aligned}
&\left\| H(\theta, x) - H\big(\theta', x\big) \right\| \leq L_1 \big(1 + \|x\|\big)^\rho \left\| \theta - \theta' \right\|, \\
&\left\| H(\theta, x) - H\big(\theta, x'\big) \right\| \leq L_2 \big(1 + \|x\| + \left\| x' \right\|\big)^\rho \big(1 + \|\theta\|\big) \left\| x - x' \right\|.
\end{aligned}$$

Assumption 3.8. The process $(X_n)_{n\in\mathbb{N}}$ is i.d.d.with $\|X_0\|_{2(\rho+1)}$ and $\|\theta_0\|_2$ being finite.

Assumption 3.9. There exists a mapping $A : \mathbb{R}^m \to \mathbb{R}^{d\times d}$ such that

$$\langle y, A(x)y\rangle \geq 0, \quad \text{for any } x, y \in \mathbb{R}^d \text{ (positive semidefinite)}$$

and, for all $\theta, \theta' \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$,

$$\langle \theta - \theta', H(\theta, x) - H(\theta', x)\rangle \geq \langle \theta - \theta', A(x)(\theta - \theta')\rangle$$

with the smallest eigenvalue of the matrix $\mathbb{E}[A(X_0)]$ being a positive real number which is denoted by a.

It is clear then that properties (B1) and (B2) are still valid for the gradient h of U, with the only difference that the Lipschitz constant in (B1) is given by $L_1\mathbb{E}[(1 + \|X_0\|)^\rho]$. This allows us to obtain the following result.

Theorem 3.10. *Let Assumptions 3.7, 3.8 and 3.9 hold and let $\bar{\lambda}$ be given by (18). There exist constants $c_1, c_2 > 0$ such that, for each $0 < \epsilon \leq 1/2$,*

$$W_2(\mathrm{Law}(\theta_n^\lambda), \pi) \leq \epsilon.$$

whenever $\lambda \leq \min(a/2L_1^2\mathbb{E}[(1 + \|X_0\|)^{2\rho}], 1/a)$ satisfies

$$\lambda \leq c_1\epsilon^2 \quad \text{and} \quad n \geq \frac{c_2}{\epsilon^2}\ln(1/\epsilon), \tag{21}$$

where c_1, c_2 (given explicitly in the proof) depend only on d, a, $\mathbb{E}[\|X_0\|^{2\rho+2}]$, L_1 and L_2. If $\rho = 0$ in Assumption 3.7, then the above results are true for $\lambda \leq 1/2\min(L_1^{-1}, \bar{\lambda})$.

Proof. The proof of this Theorem is postponed to Section 5. □

3.3. Discussion

Rate of convergence. Theorem 3.6 significantly improves on some of the results in [24] in certain cases, compare also to [31]. In [24] the monotonicity assumption (14) is not imposed, only a dissipativity condition is required and a more general recursive scheme is investigated. However, the input sequence $(X_n)_{n\in\mathbb{N}}$ is assumed i.d.d. In that setting, [24], Theorem 2.1, applies to (5) (with the choice $\delta = 0$, $\beta = 1$ and d fixed, see also the last paragraph of Section 1.1 of [24]), which implies that

$$W_2(\mathrm{Law}(\theta_n^\lambda), \pi) \leq \epsilon$$

holds whenever $\lambda \leq c_3(\epsilon/\ln(1/\epsilon))^4$ and $n \geq \frac{c_4}{\epsilon^4}\ln^5(1/\epsilon)$ with some $c_3, c_4 > 0$. For the case of i.d.d. $(X_n)_{n\in\mathbb{N}}$ see also the very recent [19]. Our results provide the sharper estimates (20) in a setting where $(X_n)_{n\in\mathbb{N}}$ may have dependencies.

Comparison with [8]. One notes, further, that a noisy Langevin Monte Carlo algorithm (nLMC) with inaccurate drift is proposed in [8], where the drift is assumed to be a linear combination of the original gradient and of random noise represented by a dependent sequence of random vectors with non-zero means. Thus, a particular form of dependency is included in this approach. A convergence result, [8], Theorem 4, in Wasserstein-2 distance between nLMC and the target distribution π is provided, which

is in agreement with our findings, that is, rate of convergence equal to $1/2$ is given when the bias term is eliminated.

In [8], Condition N, two quantities enter into play: the upper bound L^2-norm of the conditional bias, $\mathbb{E}[\|\mathbb{E}[H(\theta_k^\lambda, X_{k+1})|\theta_k^\lambda] - h(\theta_k^\lambda)\|^2]$ and the variance $\mathbb{E}[\|H(\theta_k^\lambda, X_{k+1}) - \mathbb{E}[H(\theta_k^\lambda, X_{k+1})|\theta_k^\lambda]\|^2]$. We stress that, when the process $(X_k)_{k\in\mathbb{N}}$ is actually dependent, θ_k^λ and X_{k+1} are dependent and therefore $\mathbb{E}[H(\theta_k^\lambda, X_{k+1})|\theta_k^\lambda] \neq h(\theta_k^\lambda)$ in general. With the exception of a few very simple cases, a precise computation of conditional bias $\mathbb{E}[H(\theta_k^\lambda, X_{k+1})|\theta_k^\lambda] - h(\theta_k^\lambda)$ (or of a tight upper bound for the L^2 norm of this quantity) is out of reach. Using (3) and Assumption 3.2, we get that, for all $k \in \mathbb{N}$,

$$\left\|\mathbb{E}\left[H\left(\theta_k^\lambda, X_{k+1}\right)|\theta_k^\lambda\right] - h\left(\theta_k^\lambda\right)\right\|^2 \leq L_2^2 \int \mathbb{E}\left[\|X_k - x\|^2|\theta_k^\lambda\right]\mu(\mathrm{d}x),$$

where μ denotes the common law of the X_k. This implies that $\mathbb{E}[\|\mathbb{E}[H(\theta_k^\lambda, X_{k+1})|\theta_k^\lambda] - h(\theta_k^\lambda)\|^2] \leq \delta^2 d$ with

$$\delta^2 \leq 2d^{-1}L_2^2\left\{\mathcal{M}_2(X) + \int \|x\|^2\mu(\mathrm{d}x)\right\}.$$

Similarly, using again Assumption 3.2, we get

$$\begin{aligned}
&\mathbb{E}\left[\left\|H\left(\theta_k^\lambda, X_{k+1}\right) - \mathbb{E}\left[H\left(\theta_k^\lambda, X_{k+1}\right)|\theta_k^\lambda\right]\right\|^2\right] \\
&\quad\leq 2\mathbb{E}\left[\left\|H\left(\theta_k^\lambda, X_{k+1}\right) - H\left(\theta_k^\lambda, 0\right)\right\|^2\right] + 2\mathbb{E}\left[\left\|\mathbb{E}\left[H\left(\theta_k^\lambda, X_{k+1}\right) - H\left(\theta_k^\lambda, 0\right)|\theta_k^\lambda\right]\right\|^2\right] \\
&\quad\leq 4L_2^2\mathcal{M}_2(X) =: \sigma^2 d.
\end{aligned}$$

Our assumptions therefore imply [8], Condition N, but the conclusions that we reach in Theorems 3.5 and 3.6 are sharper (note that the bias term in [8], Theorem 4, does not vanish as $\lambda \downarrow 0^+$).

Choice of step size. It is pointed out in [27] that the ergodicity property of (2) is sensitive to the step size λ. Moreover, [20], Lemma 6.3, gives an example in which the Euler–Maruyama discretization is transient. As pointed out in [20], under discretization, the minorization condition is insensitive with appropriate sampling rate while the Lyapunov condition may be lost. An invariant measure exists if the two conditions hold simultaneously, see [20], Theorem 7.3, and also [27], Theorem 3.2, for similar discussions. In this work, an approach similar to [7] is chosen, in that strong convexity of U is assumed together with Lipschitzness of its gradient and, thus, the ergodicity of (2) is obtained.

4. Proof of main results: Dependent data

4.1. The Langevin SDE and its discretization: The strongly convex case

Before proceeding to the demonstrations of the main results, we recall here some recent results on the diffusion of Langevin and its discretization for strongly convex potentials. All the results presented here are classic and can be found in either [9,10] or [8].

By [22], Theorem 2.1.8, U has a unique minimum at some point $\theta^* \in \mathbb{R}^d$. Note that due to the Lipschitz condition (B1), the SDE (1) has a unique strong solution. It is a well-known result that the Langevin SDE (1) admits a unique invariant measure π. By [17], Theorem 4.20, one constructs the associated strongly Markovian semigroup $(P_t)_{t\geq 0}$ given for all $t \geq 0$, $x \in \mathbb{R}^d$ and $\mathrm{A} \in \mathcal{B}(\mathbb{R}^d)$ by $P_t(x, \mathrm{A}) = P(\theta_t \in \mathrm{A}|\theta_0 = x)$.

The following lemma from [10] with adapted statement provides the explicit bound of the second moment of the Langevin diffusion, which allows the analysis of the Wasserstein-2 distance between π and the aforementioned sampling algorithms.

Lemma 4.1 (Proposition 1 in [10]). *Let Assumptions 3.2 and 3.3 hold and thus* (B1), (B2) *are thereby implied.*

(i) *For all $t \geq 0$ and $\vartheta \in \mathbb{R}^d$,*

$$\int_{\mathbb{R}^d} \|\vartheta - \theta^*\|^2 P_t(\theta, \mathrm{d}\vartheta) \leq \|\theta - \theta^*\|^2 \mathrm{e}^{-2at} + (d/a)\big(1 - \mathrm{e}^{-2at}\big).$$

(ii) *The stationary distribution π satisfies*

$$\int_{\mathbb{R}^d} \|\vartheta - \theta^*\|^2 \pi(\mathrm{d}\vartheta) \leq d/a.$$

For a fixed step size $\lambda \in (0, 1]$, consider the Markov kernel R_λ given for all $A \in \mathcal{B}(\mathbb{R}^d)$ and $\theta \in \mathbb{R}^d$ by

$$R_\lambda(\theta, A) = \int_A (4\pi\lambda)^{-d/2} \exp\big(-(4\lambda)^{-1} \|\vartheta - \theta + \lambda h(\theta)\|^2\big)\, \mathrm{d}\vartheta. \tag{22}$$

The discrete-time Langevin recursion (2) defines a time-homogeneous Markov chain, and for any $n \geq 1$, and for any bounded (or non-negative) Borel function $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}\big[f\big(\overline{\theta}_n^\lambda\big) | \overline{\theta}_{n-1}^\lambda\big] = R_\lambda f\big(\overline{\theta}_{n-1}^\lambda\big) = \int_{\mathbb{R}^d} f(\vartheta) R_\lambda\big(\overline{\theta}_{n-1}^\lambda, \mathrm{d}\vartheta\big).$$

Lemma 4.2 below is also a result from [10] and along with Lemma 4.1 are presented here for completeness by using the notation of this article. In particular, Lemma 4.2 states that R_λ admits a unique stationary distribution π_λ, which may differ from π.

Lemma 4.2. *Let Assumption 3.3 hold and thus* (B2) *is thereby implied. Then, for all $\lambda < \bar{\lambda}$, where $\bar{\lambda}$ is defined in (18), the following hold*:

(i) *For all $\theta \in \mathbb{R}^d$, $n \geq 1$,*

$$\int_{\mathbb{R}^d} \|\vartheta - \theta^*\|^2 R_\lambda^n(\theta, \mathrm{d}\vartheta) \leq (1 - 2\tilde{a}\lambda)^n \|\theta - \theta^*\|^2 + (d/\tilde{a})\big(1 - (1 - 2\tilde{a}\lambda)^n\big).$$

(ii) *The Markov kernel R_λ has a unique stationary distribution π_λ which satisfies*

$$\int_{\mathbb{R}^d} \|\theta - \theta^*\|^2 \pi_\lambda(\mathrm{d}\theta) \leq d/\tilde{a}.$$

where $\tilde{a}$ is defined in (16).

(iii) *For all $\theta \in \mathbb{R}^d$, $n \geq 1$,*

$$W_2\big(\delta_\theta R_\lambda^n, \pi_\lambda\big) \leq \mathrm{e}^{-\tilde{a}\lambda n} \sqrt{2}\big(\|\theta - \theta^*\|^2 + d/\tilde{a}\big)^{1/2}.$$

(iv) *For all $n \in \mathbb{N}$ and square-integrable $\mathbb{R}^d$-valued random variables η_1, η_2 with $\sigma(\eta_1, \eta_2)$ independent of ξ_k, $k \in \mathbb{N}$*

$$\mathbb{E}\big[\big\|\overline{\theta}_n^{\lambda}(1) - \overline{\theta}_n^{\lambda}(2)\big\|^2\big] \le \mathrm{e}^{-2\tilde{a}\lambda n}\mathbb{E}\big[\|\eta_1 - \eta_2\|^2\big],$$

where $\overline{\theta}_n^{\lambda}(i)$, $i = 1, 2$ denote the solutions of the recursion (2) with the respective initial conditions $\theta_0 = \eta_i$, $i = 1, 2$.

Proof. For the first three statements, see [10], Propositions 2 and 3. For iv, see the proof of [10], Proposition 3. □

Note that by Lemma 4.2, a Foster–Lyapunov type drift condition is satisfied with $V_1(\theta) := \|\theta - \theta^*\|^2$, which yields that $\sup_{n\ge 0}\|\overline{\theta}_n^{\lambda}\|_2 < \infty$. This allows the analysis of the convergence between the recursive scheme (2) and the stationary distribution π in Wasserstein-2 distance (see Theorem 4.11 below). However, in order to obtain the rate of convergence between (2) and the SGLD scheme (5), the finiteness of higher moments is required. In the following lemma, one obtains the drift condition with $V_p(\theta) := \|\theta - \theta^*\|^{2p}$, $p \in \mathbb{N} \setminus \{0\}$.

Lemma 4.3. *Let Assumptions 3.1, 3.2 and 3.3 hold. For any integer $p \ge 1$, let $V_p(\theta) := \|\theta - \theta^*\|^{2p}$. Then, the process $\overline{\theta}^{\lambda}$ satisfies, for any $n \in \mathbb{N}$ and $\lambda < \bar{\lambda}$, where $\bar{\lambda}$ is defined in (18),*

$$\mathbb{E}\big[V_p\big(\overline{\theta}_{n+1}^{\lambda}\big)|\overline{\theta}_n^{\lambda}\big] \le \rho_\lambda V_p\big(\overline{\theta}_n^{\lambda}\big) + \lambda C'(p), \tag{23}$$

where $\rho_\lambda = 1 - \tilde{a}\lambda \in (0, 1)$ and

$$C'(p) := d^p(2p-1)^p p^p 2^{p(2p-1)}\tilde{a}^{1-p} + (2p-1)p2^{3p-2}2^{2p}d^p p^{\frac{3}{2}p}. \tag{24}$$

Moreover,

$$\sup_{\lambda<\bar{\lambda}}\sup_n \mathbb{E}\big[V_p\big(\overline{\theta}_n^{\lambda}\big)\big] \le \mathbb{E}\big[V_p(\theta_0)\big] + C'(p)/\tilde{a}. \tag{25}$$

and $C'(p)^{1/2p} \le c'(p)$ holds with

$$c'(p) = p\sqrt{d}\big(2^{p+1/2}\tilde{a}^{\frac{1}{2p}-\frac{1}{2}} + 24\big). \tag{26}$$

Proof. Recall equation (2) and define

$$\Delta_n := \overline{\theta}_n^{\lambda} - \theta^* - \lambda\big(h\big(\overline{\theta}_n^{\lambda}\big) - h\big(\theta^*\big)\big), \quad \text{for every } n \ge 0.$$

Then, one calculates

$$\begin{aligned}
\mathbb{E}\big[\big\|\overline{\theta}_{n+1}^{\lambda} - \theta^*\big\|^{2p}|\overline{\theta}_n^{\lambda}\big] &= \mathbb{E}\big[\|\Delta_n + \sqrt{2\lambda}\xi_{n+1}\|^{2p}|\overline{\theta}_n^{\lambda}\big] \\
&= \mathbb{E}\big[\big(\|\Delta_n\|^2 + 2\langle\Delta_n, \sqrt{2\lambda}\xi_{n+1}\rangle + \|\sqrt{2\lambda}\xi_{n+1}\|^2\big)^p|\overline{\theta}_n^{\lambda}\big] \\
&\le \mathbb{E}\Bigg[\sum_{\substack{i+j+k=p \\ \{i\le p-1\}\cap\{j\ne 1\}}} \frac{p!}{i!\,j!\,k!}\|\Delta_n\|^{2i}\big(2\langle\Delta_n, \sqrt{2\lambda}\xi_{n+1}\rangle\big)^j\|\sqrt{2\lambda}\xi_{n+1}\|^{2k}\Big|\overline{\theta}_n^{\lambda}\Bigg] \\
&\quad + \mathbb{E}\big[2p\|\Delta_n\|^{2(p-1)}\langle\Delta_n, \sqrt{2\lambda}\xi_{n+1}\rangle|\overline{\theta}_n^{\lambda}\big]
\end{aligned}$$

where the last term is clearly zero. Thus, due to Lemma A.3,

$$
\begin{aligned}
&\mathbb{E}\big[\big\|\overline{\theta}_{n+1}^{\lambda}-\theta^*\big\|^{2p}|\overline{\theta}_n^{\lambda}\big]\\
&\quad\le \mathbb{E}\Bigg[\sum_{\substack{k=0\\k\neq 1}}^{2p}\binom{2p}{k}\|\Delta_n\|^{2p-k}\|\sqrt{2\lambda}\xi_{n+1}\|^k\Big|\overline{\theta}_n^{\lambda}\Bigg]\\
&\quad\le \|\Delta_n\|^{2p}+\mathbb{E}\Bigg[\sum_{k=2}^{2p}\binom{2p}{k}\|\Delta_n\|^{2p-k}\|\sqrt{2\lambda}\xi_{n+1}\|^k\Big|\overline{\theta}_n^{\lambda}\Bigg]\\
&\quad= \|\Delta_n\|^{2p}+\mathbb{E}\Bigg[\Bigg(\sum_{k=2}^{2p}\binom{2p}{k}\|\Delta_n\|^{2p-k}\|\sqrt{2\lambda}\xi_{n+1}\|^{k-2}\Bigg)\|\sqrt{2\lambda}\xi_{n+1}\|^2\Big|\overline{\theta}_n^{\lambda}\Bigg]\\
&\quad= \|\Delta_n\|^{2p}+\mathbb{E}\Bigg[\Bigg(\sum_{l=0}^{2(p-1)}\binom{2p}{l+2}\|\Delta_n\|^{2(p-1)-l}\|\sqrt{2\lambda}\xi_{n+1}\|^l\Bigg)\|\sqrt{2\lambda}\xi_{n+1}\|^2\Big|\overline{\theta}_n^{\lambda}\Bigg]\\
&\quad\le \|\Delta_n\|^{2p}\\
&\qquad+\mathbb{E}\Bigg[\binom{2p}{2}\Bigg(\sum_{l=0}^{2(p-1)}\binom{2(p-1)}{l}\|\Delta_n\|^{2(p-1)-l}\|\sqrt{2\lambda}\xi_{n+1}\|^l\Bigg)\|\sqrt{2\lambda}\xi_{n+1}\|^2\Big|\overline{\theta}_n^{\lambda}\Bigg]\\
&\quad= \|\Delta_n\|^{2p}+(2p-1)p\mathbb{E}\big[\big(\|\Delta_n\|+\|\sqrt{2\lambda}\xi_{n+1}\|\big)^{2(p-1)}\|\sqrt{2\lambda}\xi_{n+1}\|^2|\overline{\theta}_n^{\lambda}\big]\\
&\quad\le \|\Delta_n\|^{2p}+(2p-1)p2^{2(p-1)}\|\Delta_n\|^{2(p-1)}\mathbb{E}\big[\|\sqrt{2\lambda}\xi_{n+1}\|^2\big]\\
&\qquad+(2p-1)p2^{2(p-1)}\mathbb{E}\big[\|\sqrt{2\lambda}\xi_1\|^{2p}\big].
\end{aligned}
\tag{27}
$$

Moreover, one recalls that for $\lambda<2/(a+L_1)$

$$
\|\Delta_n\|^2\le(1-2\tilde{a}\lambda)\big\|\overline{\theta}_n^{\lambda}-\theta^*\big\|^2.
$$

Consequently

$$
\begin{aligned}
&\mathbb{E}\big[\big\|\overline{\theta}_{n+1}^{\lambda}-\theta^*\big\|^{2p}|\overline{\theta}_n^{\lambda}\big]\\
&\quad\le(1-2\tilde{a}\lambda)^p\big\|\overline{\theta}_n^{\lambda}-\theta^*\big\|^{2p}+(2p-1)p2^{2p-1}\lambda d(1-2\tilde{a}\lambda)^{p-1}\big\|\overline{\theta}_n^{\lambda}-\theta^*\big\|^{2(p-1)}\\
&\qquad+(2p-1)p2^{2(p-1)}\mathbb{E}\big[\|\sqrt{2\lambda}\xi_1\|^{2p}\big]\\
&\quad\le(1-\tilde{a}\lambda)(1-2\tilde{a}\lambda)^{p-1}\big\|\overline{\theta}_n^{\lambda}-\theta^*\big\|^{2p}-\tilde{a}\lambda(1-2\tilde{a}\lambda)^{p-1}\big\|\overline{\theta}_n^{\lambda}-\theta^*\big\|^{2p}\\
&\qquad+(2p-1)p2^{2p-1}\lambda d(1-2\tilde{a}\lambda)^{p-1}\big\|\overline{\theta}_n^{\lambda}-\theta^*\big\|^{2(p-1)}\\
&\qquad+(2p-1)p2^{2(p-1)}\mathbb{E}\big[\|\sqrt{2\lambda}\xi_1\|^{2p}\big].
\end{aligned}
\tag{28}
$$

As a result, for $\|\overline{\theta}_n^{\lambda}-\theta^*\|\ge\overline{M}$, where $\overline{M}=\sqrt{d(2p-1)p2^{2p-1}/\tilde{a}}$, one obtains

$$
\mathbb{E}\big[\big\|\overline{\theta}_{n+1}^{\lambda}-\theta^*\big\|^{2p}|\overline{\theta}_n^{\lambda}\big]\le(1-\tilde{a}\lambda)\big\|\overline{\theta}_n^{\lambda}-\theta^*\big\|^{2p}+\lambda(2p-1)p2^{3p-2}\mathbb{E}\big[\|\xi_1\|^{2p}\big],
$$

whereas, for $\|\overline{\theta}_n^\lambda - \theta^*\| \le \overline{M}$ one obtains

$$\mathbb{E}\big[\big\|\overline{\theta}_{n+1}^\lambda - \theta^*\big\|^{2p} | \overline{\theta}_n^\lambda\big] \le (1-\tilde{a}\lambda)\big\|\overline{\theta}_n^\lambda - \theta^*\big\|^{2p} + \lambda d^p (2p-1)^p p^p 2^{p(2p-1)} \tilde{a}^{1-p} + \lambda(2p-1)p2^{3p-2}\mathbb{E}\big[\|\xi_1\|^{2p}\big]$$

which yields (23). Consequently, by Lemma A.4 below,

$$\mathbb{E}\big[\big\|\overline{\theta}_{n+1}^\lambda - \theta^*\big\|^{2p} | \overline{\theta}_n^\lambda\big] \le (1-\tilde{a}\lambda)^{2p}\big\|\overline{\theta}_0 - \theta^*\big\|^{2p} + \frac{C'(p)}{\tilde{a}}.$$

Thus, one obtains the desired result regarding the uniform bounds. The estimate $C'(p)^{1/2p} \le c'(p)$ follows, noting the trivial inequalities: $p^{1/p} \le 2$, $p \in \mathbb{N}\setminus\{0\}$; $(x+y)^{1/2p} \le x^{1/2p} + y^{1/2p}$, $x, y \ge 0$. □

4.2. Analysis for the SGLD scheme

One notes initially that the process in (2) is Markovian while the one in (5) is not. However, uniform bounds are obtained in Lemma 4.4, below, for the $2p$-th moment of the SGLD scheme (5), for any $p \ge 1$. This result complements the findings of Lemma 4.3 and is used in the proof of Theorem 3.5, which examines the convergence between the two sampling algorithms, ULA (2) and SGLD (5), in Wasserstein-2 distance.

The following inequalities, derived from Assumptions 3.2 and 3.3, are often used:

$$\begin{aligned} &\big\|H(\theta,x)\big\| \le L_1\big\|\theta-\theta^*\big\| + L_2\|x\| + H^*, \quad H^* = \sum_{i=1}^d \big|H^i\big(\theta^*,0\big)\big|, \\ &\big\langle \theta-\theta^*, H(\theta,x)\big\rangle \ge a\big\|\theta-\theta^*\big\|^2 + \big\langle \theta-\theta^*, H\big(\theta^*,x\big)\big\rangle. \end{aligned} \tag{29}$$

Lemma 4.4. *Let Assumptions 3.1, 3.2 and 3.3 hold. Let $V_p(\theta) = \|\theta-\theta^*\|^{2p}$ for some integer $p \ge 1$. The process θ^λ satisfies, for any $n \in \mathbb{N}$ and $\lambda < \bar{\lambda}$, where $\bar{\lambda}$ is defined in (18),*

$$\mathbb{E}\big[V_p\big(\theta_n^\lambda\big)\big] \le (\rho_\lambda)^n \mathbb{E}\big[V_p\big(\theta_0^\lambda\big)\big] + \lambda C''(p), \tag{30}$$

where $\rho_\lambda = 1-\tilde{a}\lambda \in (0,1)$ and

$$\begin{aligned} C''(p) := {} & \big(2^{2p}dp(2p-1)\big)^p (2/\tilde{a})^{p-1} + 2^{5p-4}p(2p-1)2^{2p}d^p p^{\frac{3}{2}p} \\ & + 2^{2p-1}\big\{(2p)^{2p}(2/\tilde{a})^{2p-1} + \big(2^{2p-1}p(2p-1)\big)^p (2/\tilde{a})^{p-1} \\ & + 2^{4p-4}p(2p-1)\big\}\big\{2^{2p-1}L_1^{2p}\big\|\theta^*\big\|^{2p} + 2^{2p-1}L_2^{2p}\mathcal{M}_{2p}(X) + \big\{H^*\big\}^{2p}\big\}. \end{aligned}$$

As a result,

$$\sup_{\lambda<\bar{\lambda}}\sup_n \mathbb{E}\big[V_p\big(\theta_n^\lambda\big)\big] \le \mathbb{E}\big[V_p(\theta_0)\big] + \frac{C''(p)}{\tilde{a}}. \tag{31}$$

It follows also that $C''(p)^{1/2p} \le c''(p)$ where

$$c''(p) := p\sqrt{d}\big(2^{p+1/2}\tilde{a}^{\frac{1}{2p}-\frac{1}{2}} + 48\big)$$

$$+2\big\{4p/\tilde{a}^{1-1/2p}+2^p p\sqrt{2}(2/\tilde{a})^{1/2-1/(2p)}$$
$$+12\big\}\big\{2L_1\|\theta^*\|+2L_2\mathcal{M}_{2p}^{1/2p}(X)+H^*\big\}. \tag{32}$$

Proof. For each $n\in\mathbb{N}$, denote by $\Delta_n=\theta_n^\lambda-\theta^*-\lambda(H(\theta_n^\lambda,X_{n+1})-H(\theta^*,X_{n+1}))$. By direct calculations, one obtains,

$$\begin{aligned}
&\mathbb{E}\big[\|\theta_{n+1}^\lambda-\theta^*\|^{2p}|\theta_n^\lambda\big]\\
&=\mathbb{E}\big[\|\Delta_n+\sqrt{2\lambda}\xi_{n+1}-\lambda H\big(\theta^*,X_{n+1}\big)\|^{2p}|\theta_n^\lambda\big]\\
&=\mathbb{E}\big[\big(\|\Delta_n\|^2+\|\sqrt{2\lambda}\xi_{n+1}-\lambda H\big(\theta^*,X_{n+1}\big)\|^2\\
&\quad+2\big\langle\Delta_n,\sqrt{2\lambda}\xi_{n+1}-\lambda H\big(\theta^*,X_{n+1}\big)\big\rangle\big)^p|\theta_n^\lambda\big]\\
&=\mathbb{E}\Bigg[\sum_{k_1+k_2+k_3=p}\frac{p!}{k_1!k_2!k_3!}\|\Delta_n\|^{2k_1}\|\sqrt{2\lambda}\xi_{n+1}-\lambda H\big(\theta^*,X_{n+1}\big)\|^{2k_2}\\
&\quad\times\big(2\big\langle\Delta_n,\sqrt{2\lambda}\xi_{n+1}-\lambda H\big(\theta^*,X_{n+1}\big)\big\rangle\big)^{k_3}\Big|\theta_n^\lambda\Bigg]\\
&\le\mathbb{E}\big[\|\Delta_n\|^{2p}|\theta_n^\lambda\big]+2p\mathbb{E}\big[\|\Delta_n\|^{2p-2}\big\langle\Delta_n,\sqrt{2\lambda}\xi_{n+1}-\lambda H\big(\theta^*,X_{n+1}\big)\big\rangle|\theta_n^\lambda\big]\\
&\quad+\sum_{k=2}^{2p}\binom{2p}{k}\mathbb{E}\big[\|\Delta_n\|^{2p-k}\|\sqrt{2\lambda}\xi_{n+1}-\lambda H\big(\theta^*,X_{n+1}\big)\|^k|\theta_n^\lambda\big],
\end{aligned}$$

where the last inequality holds due to Lemma A.3, and further calculations yield

$$\begin{aligned}
&\mathbb{E}\big[\|\theta_{n+1}^\lambda-\theta^*\|^{2p}|\theta_n^\lambda\big]\\
&\le\mathbb{E}\big[\|\Delta_n\|^{2p}|\theta_n^\lambda\big]+2p\lambda\mathbb{E}\big[\|\Delta_n\|^{2p-1}\|H\big(\theta^*,X_{n+1}\big)\||\theta_n^\lambda\big]\\
&\quad+\sum_{k=2}^{2p}\binom{2p}{k}\mathbb{E}\big[\|\Delta_n\|^{2p-k}\|\sqrt{2\lambda}\xi_{n+1}-\lambda H\big(\theta^*,X_{n+1}\big)\|^k|\theta_n^\lambda\big]\\
&\le\Big(1+\frac{\tilde{a}\lambda}{2}\Big)\mathbb{E}\big[\|\Delta_n\|^{2p}|\theta_n^\lambda\big]+\lambda(2p)^{2p}\Big(\frac{2}{\tilde{a}}\Big)^{2p-1}\mathbb{E}\big[\|H\big(\theta^*,X_{n+1}\big)\|^{2p}|\theta_n^\lambda\big]\\
&\quad+2^{2p-3}p(2p-1)\mathbb{E}\big[\|\Delta_n\|^{2p-2}\|\sqrt{2\lambda}\xi_{n+1}-\lambda H\big(\theta^*,X_{n+1}\big)\|^2|\theta_n^\lambda\big]\\
&\quad+2^{2p-3}p(2p-1)\mathbb{E}\big[\|\sqrt{2\lambda}\xi_{n+1}-\lambda H\big(\theta^*,X_{n+1}\big)\|^{2p}|\theta_n^\lambda\big]\\
&\le(1+\tilde{a}\lambda)\mathbb{E}\big[\|\Delta_n\|^{2p}|\theta_n^\lambda\big]+\lambda(2p)^{2p}\Big(\frac{2}{\tilde{a}}\Big)^{2p-1}\mathbb{E}\big[\|H\big(\theta^*,X_{n+1}\big)\|^{2p}|\theta_n^\lambda\big]\\
&\quad+\lambda\big(2^{2p-1}p(2p-1)\big)^p\Big(\frac{2}{\tilde{a}}\Big)^{p-1}\mathbb{E}\big[\|H\big(\theta^*,X_{n+1}\big)\|^{2p}|\theta_n^\lambda\big]\\
&\quad+\lambda\big(2^{2p}dp(2p-1)\big)^p\Big(\frac{2}{\tilde{a}}\Big)^{p-1}+\lambda 2^{5p-4}p(2p-1)\mathbb{E}\big[\|\xi_{n+1}\|^{2p}\big]
\end{aligned}$$

$$+\lambda 2^{4p-4}p(2p-1)\mathbb{E}\big[\big\|H\big(\theta^*, X_{n+1}\big)\big\|^{2p}|\theta_n^\lambda\big], \tag{33}$$

where the second inequality follows the same argument as in the proof of Lemma 4.3. Moreover, for $\lambda < 2/(a + L_1)$,

$$\begin{aligned}\mathbb{E}\big[\|\Delta_n\|^{2p}|\theta_n^\lambda\big] &= \mathbb{E}\big[\big(\big\|\theta_n^\lambda - \theta^*\big\|^2 - 2\lambda\big\langle\theta_n^\lambda - \theta^*, H\big(\theta_n^\lambda, X_{n+1}\big) - H\big(\theta^*, X_{n+1}\big)\big\rangle \\ &\quad + \lambda^2\big\|H\big(\theta_n^\lambda, X_{n+1}\big) - H\big(\theta^*, X_{n+1}\big)\big\|^2\big)^p|\theta_n^\lambda\big] \\ &\le (1 - 2\tilde{a}\lambda)^p\big\|\theta_n^\lambda - \theta^*\big\|^{2p}.\end{aligned}$$

Then, substituting the above estimate into (33) yields

$$\mathbb{E}\big[\big\|\theta_{n+1}^\lambda - \theta^*\big\|^{2p}|\theta_n^\lambda\big] \le (1 - \tilde{a}\lambda)\big\|\theta_n^\lambda - \theta^*\big\|^{2p} + \lambda\mathbb{E}\big[g(X_{n+1})|\theta_n^\lambda\big],$$

where

$$\begin{aligned}g(X_{n+1}) &= \big(2^{2p}dp(2p-1)\big)^p(2/\tilde{a})^{p-1} + 2^{5p-4}p(2p-1)\mathbb{E}\big[\|\xi_{n+1}\|^{2p}\big] \\ &\quad + 2^{2p-1}\big\{(2p)^{2p}(2/\tilde{a})^{2p-1} + \big(2^{2p-1}p(2p-1)\big)^p(2/\tilde{a})^{p-1} \\ &\quad + 2^{4p-4}p(2p-1)\big\}\big\{\big(L_1\big\|\theta^*\big\| + L_2\|X_{n+1}\|\big)^{2p} + \big\|H\big(\theta^*, 0\big)\big\|^{2p}\big\}.\end{aligned}$$

Using the trivial $(x + y)^{2p} \le 2^{2p-1}(x^{2p} + y^{2p})$, $x, y \ge 0$ and Lemma A.4, we have

$$E\big[g(X_{n+1})\big] \le C''(p).$$

Finally, denote by $\rho_\lambda = 1 - \tilde{a}\lambda \in (0, 1)$, then by induction, one obtains

$$\mathbb{E}\big[\big\|\theta_{n+1}^\lambda - \theta^*\big\|^{2p}\big] \le (\rho_\lambda)^{n+1}\mathbb{E}\big[\big\|\theta_0 - \theta^*\big\|^{2p}\big] + \frac{C''(p)}{\tilde{a}},$$

which implies $\sup_{\lambda<\bar{\lambda}}\sup_n \mathbb{E}[\|\theta_{n+1}^\lambda - \theta^*\|^{2p}] \le \mathbb{E}[\|\theta_0 - \theta^*\|^{2p}] + C''(p)/\tilde{a}$. It is easy to check $C''(p)^{1/2p} \le c''(p)$, too. □

Uniform L^2 bounds for the process in (5) are obtained in [24] under dissipativity condition on ∇U and the L^2 error of the stochastic gradient, that is, $\mathbb{E}[\|H(\theta, X_n) - h(\theta)\|^2]$, see their Assumptions (A.3), (A.4). In that paper a large size mini-batch could be used to reduce the variance of the estimator, which requires more computational costs. We could also incorporate mini-batches in our algorithm but this is not pursued here. For stability, the variance of the estimator has to be controlled, see [28].

4.3. Proof of Theorem 3.5

We now sketch a roadmap for the proof of Theorem 3.5. The time axis is cut into intervals of size T. An auxiliary process $\overline{z}^\lambda$ is introduced which equals θ^λ at the points nT, $n \in \mathbb{N}$ but it follows the averaged dynamics on $[nT, (n + 1)T)$, see (2).

Using the conditional L-mixing property, one obtains estimates for the L^2-distance between $\overline{z}^\lambda$ and θ^λ. If $\overline{z}^\lambda$ were uniformly bounded, these would be of the order $\sqrt{\lambda}$. However, $\overline{z}^\lambda$ is unbounded hence its maximal process needs to be controlled which leads to the somewhat weaker rate $\lambda^{\frac{1}{2}-\varepsilon}$, for $\varepsilon > 0$ arbitrarily small.

Next, estimates for the difference between $\overline{z}^\lambda$ and $\overline{\theta}^\lambda$ are derived using the contraction property of the dynamics of $\overline{\theta}^\lambda$, see Lemma 4.2. It follows that this is of the same order as $\overline{z}^\lambda - \theta^\lambda$. These observations then allows us to conclude.

We proceed now with the rigorous arguments. Let

$$\mathcal{H}_n := \mathcal{G}_n \vee \sigma(\xi_j, j \in \mathbb{N}), \qquad \mathcal{H}_n^+ := \mathcal{G}_n^+, \quad n \in \mathbb{N}.$$

Observe first that, since $(X_n)_{n\in\mathbb{N}}$ is conditionally L-mixing with respect to $(\mathcal{G}_n, \mathcal{G}_n^+)_{n\in\mathbb{N}}$, it is conditionally L-mixing with respect to $(\mathcal{H}_n, \mathcal{H}_n^+)_{n\in\mathbb{N}}$, too, and the corresponding quantities $(M, \Gamma, \mathcal{C}, \mathcal{M})$ remain the same.

For each $\theta \in \mathbb{R}^d$, $0 \le i \le j$, one recursively defines

$$z^\lambda(i,i,\theta) := \theta, \qquad z^\lambda(j+1,i,\theta) := z^\lambda(j,i,\theta) - \lambda h\big(z^\lambda(j,i,\theta)\big) + \sqrt{2\lambda}\xi_{j+1}.$$

Let $T := \lfloor 1/\lambda \rfloor$, then for each $n \in \mathbb{N}$ and for each $nT \le k < (n+1)T$, one defines

$$\overline{z}_k^\lambda := z^\lambda\big(k, nT, \theta_{nT}^\lambda\big).$$

Consequently, $\overline{z}_k^\lambda$ is defined for all $k \in \mathbb{N}$; $\overline{z}_{nT}^\lambda = \theta_{nT}^\lambda$ for $n \in \mathbb{N}$ and $\overline{\theta}_k^\lambda = z^\lambda(k,0,\theta_0)$. Next, some simple but essential moment estimates are derived.

Lemma 4.5. *Let $q \ge 1$ be an integer. Then, for all $\lambda < \bar{\lambda}$, where $\bar{\lambda}$ is defined in (18),*

$$\sup_{k\in\mathbb{N}} \big\|\overline{z}_k^\lambda - \theta^*\big\|_{2q} \le \underline{C}(q)$$

holds for

$$\underline{C}(q) := \big\|\theta_0 - \theta^*\big\|_{2q} + \frac{c'(q) + c''(q)}{\tilde{a}^{1/(2q)}}, \tag{34}$$

where $c'(q)$, $c''(q)$ are as in Lemmata 4.3 and 4.4.

Proof. Define $V_q(\theta) := \|\theta - \theta^*\|^{2q}$, $\theta \in \mathbb{R}^d$. Let $k \in \mathbb{N}$ be arbitrary and let $n \in \mathbb{N}$ be such that $nT \le k < (n+1)T$. Note that (25) and (31) imply

$$\begin{aligned}\sup_{nT\le k<(n+1)T} \big\|\overline{z}_k^\lambda - \theta^*\big\|_{2q} &\le \left[\mathbb{E}\big[V_q\big(\theta_{nT}^\lambda\big)\big] + \frac{C'(q)}{\tilde{a}}\right]^{1/(2q)} \\ &\le \big\|\theta_0 - \theta^*\big\|_{2q} + \frac{C'(q)^{1/(2q)} + C''(q)^{1/(2q)}}{\tilde{a}^{1/(2q)}}.\end{aligned}$$

□

Lemma 4.6. *For all $\lambda < \bar{\lambda}$, where $\bar{\lambda}$ is defined in (18), it holds that*

$$\sup_{n\in\mathbb{N}}\big[\big\|H\big(\theta_n^\lambda, X_{n+1}\big)\big\|_2 + \big\|h\big(\overline{z}_n^\lambda\big)\big\|_2\big] \le C^\flat,$$

where

$$C^\flat = L_1\left[\big\|\theta_0 - \theta^*\big\|_2 + \frac{C''(1)}{\tilde{a}}\right] + 2L_2\mathcal{M}_2^{1/2}(X) + 2H^* + \underline{C}(1)L_1. \tag{35}$$

Proof. The first inequality of (29) implies

$$\big\|H\big(\theta_n^\lambda, X_{n+1}\big)\big\|_2 \le L_1\big\|\theta_n^\lambda - \theta^*\big\|_2 + L_2\|X_n\|_2 + H^*.$$

Combining this with Lemma 4.4 (applied with $p = 1$) shows that

$$\sup_{\lambda<\bar\lambda}\sup_n\big\|H\big(\theta_n^\lambda, X_{n+1}\big)\big\|_2 \le L_1\left[\mathbb{E}^{1/2}\big[V_1(\theta_0)\big] + \frac{C''(1)^{1/2}}{\tilde a^{1/2}}\right] + L_2\mathcal{M}_2^{1/2}(X) + H^*.$$

A similar argument can be applied to $h(\overline{z}_n^\lambda)$, in view of (11),

$$\begin{aligned}\big\|h\big(\overline{z}_n^\lambda\big)\big\|_2 &\le L_1\big\|\overline{z}_n^\lambda - \theta^*\big\|_2 + L_2\mathcal{M}_2^{1/2}(X) + H^*\\ &\le \underline{C}(1)L_1 + L_2\mathcal{M}_2^{1/2}(X) + H^*,\end{aligned}$$

where $\underline{C}(1)$ is given by (34). □

Lemma 4.7. *For each $j \in \mathbb{N}$, the random field $H(\theta, X_n)$, $n \in \mathbb{N}$, $\theta \in \mathrm{B}(\theta^*, j)$ satisfies*

$$M_r^n\big(H(\theta, X), \mathrm{B}\big(\theta^*, j\big)\big) \le L_1 j + L_2 M_r^n(X) + H^*, \tag{36}$$

$$\Gamma_r^n\big(H(\theta, X), \mathrm{B}\big(\theta^*, j\big)\big) \le 2L_2\Gamma_r^n(X). \tag{37}$$

Proof. Let $\theta \in \mathrm{B}(\theta^*, j)$. The Minkowski's inequality imply for $k \ge n$ and $i \in \{1, \dots, m\}$,

$$\mathbb{E}^{1/r}\big[\big|H^i(\theta, X_k)\big|^r|\mathcal{H}_n\big] \le L_1^i j + L_2^i\mathbb{E}^{1/r}\big[\|X_k\|^r|\mathcal{H}_n\big] + \big|H^i\big(\theta^*, 0\big)\big|.$$

Hence, using $\|X_k\| \le \sum_{j=1}^m |X_k^j|$ and the Minkowski's inequality again, we obtain

$$M_r^n\big(H(\theta, X), \mathrm{B}\big(\theta^*, j\big), i\big) \le L_1^i j + L_2^i M_r^n(X) + \big|H^i\big(\theta^*, 0\big)\big|.$$

Summing the above relation over the indices $i \in \{1, \dots, m\}$ we get (36). One also notes that, due to Lemma A.2,

$$\begin{aligned}&\mathbb{E}^{1/r}\big[\big|H^i(\theta, X_k) - \mathbb{E}\big[H^i(\theta, X_k)|\mathcal{H}_n \vee \mathcal{H}_{n-\tau}^+\big]\big|^r|\mathcal{H}_n\big]\\ &\quad\le 2\mathbb{E}^{1/r}\big[\big|H^i(\theta, X_k) - H^i\big(\theta, \mathbb{E}\big[X_k|\mathcal{H}_n \vee \mathcal{H}_{n-\tau}^+\big]\big)\big|^r|\mathcal{H}_n\big]\\ &\quad\le 2L_2^i\mathbb{E}^{1/r}\big[\big\|X_k - \mathbb{E}\big[X_k|\mathcal{H}_n \vee \mathcal{H}_{n-\tau}^+\big]\big\|^r|\mathcal{H}_n\big] \le 2L_2^i\sum_{j=1}^m\gamma_r^n(X, \tau, j),\end{aligned}$$

which implies (37). □

We shall also need the following measure-theoretical lemma.

Lemma 4.8. *Let $k \ge nT$ be an integer. There exists a version $h_{k,nT} : \Omega \times \mathbb{R}^d \to \mathbb{R}^d$ of $\mathbb{E}[H(\theta, X_k)|\mathcal{H}_{nT}]$, $\theta \in \mathbb{R}^d$ which is jointly measurable.*

Proof. For a fixed $\theta \in \mathbb{R}^d$, the conditional expectation $\mathbb{E}[H(\theta, X_k)|\mathcal{H}_{nT}]$, $\theta \in \mathbb{R}^d$ is a $\mathcal{H}_{nT}$-measurable random variable. We will construct a function $h_{k,nT} : \Omega \times \mathbb{R}^d \to \mathbb{R}^d$ that is measurable in its second variable and, for all $\theta \in \mathbb{R}^d$, $h_{k,nT}$ is a version of $\mathbb{E}[H(\theta, X_k)|\mathcal{H}_{nT}]$. The case $k = nT$ is

trivial. Let $k > nT$. It is enough to construct $h_{k,nT}(\theta)$, $\theta \in \mathrm{B}(\theta^*, N)$ for each $N \in \mathbb{N}$. Consider $\mathbb{B}(N) := \mathbf{C}(\mathrm{B}(\theta^*, N); \mathbb{R}^d)$, the usual Banach space of continuous, $\mathbb{R}^d$-valued functions defined on $\mathrm{B}(\theta^*, N)$, equipped with the maximum norm. The function

$$\omega \in \Omega \to G_N(\omega) := \big(H\big(\theta, X_k(\omega)\big)_{\theta \in \mathrm{B}(\theta^*, N)}\big), \quad \omega \in \Omega,$$

is a $\mathbb{B}(N)$-valued random variable and, by (29),

$$\sup_{\theta \in \mathrm{B}(\theta^*, N)} \big\| H(\theta, X_k) \big\| \le L_1 N + L_2 \|X_k\| + H^*,$$

which clearly has finite expectation as the process X_n, $n \in \mathbb{N}$ is conditionally L-mixing. Moreover, [23], Proposition V.2.5, implies the existence of a $\mathbb{B}(N)$-valued random variable $\mathfrak{G}_N$ such that, for each $\mathbf{b}$ in the dual space $\mathbb{B}'(N)$ of $\mathbb{B}(N)$,

$$\mathbb{E}\big[\mathbf{b}(G_N)|\mathcal{H}_{nT}\big] = \mathbf{b}(\mathfrak{G}_N).$$

This implies, in particular, that for all $\theta \in \mathrm{B}(\theta^*, N)$, $\mathbb{E}[H(\theta, X_k)|\mathcal{H}_{nT}] = \mathfrak{G}_N(\theta)$. We may thus set $h_{k,nT}(\omega, \theta) := \mathfrak{G}_N(\omega, \theta)$. Since $(\omega, \theta) \to \mathfrak{G}_N(\omega, \theta)$ is measurable in its first variable and continuous in the second, it is, in particular, jointly measurable, see, for example, [1], Lemma 4.50. □

Lemma 4.9. *Assume 3.1 and 3.1 and let* $p \ge 1$.

$$\sup_{n \in \mathbb{N}} \mathbb{E}^{1/p}\left[\left(\sum_{k=nT}^{\infty} \sup_{\theta \in \mathbb{R}^d} \big\| h_{k,nT}(\theta) - h(\theta) \big\|\right)^p\right] \le 2L_2 \mathcal{C}_{p,1}(X),$$

where $\mathcal{C}_{p,1}(X)$ *is defined in* (10).

Proof. Let $k \ge nT$. Notice that, since X_k and $\mathcal{G}^+_{nT}$ are independent of $\sigma(\xi_j, j \in \mathbb{N})$, $\mathbb{E}[X_k|\mathcal{H}^+_{nT}] = \mathbb{E}[X_k|\mathcal{G}^+_{nT}]$, $\mathbb{P}$-a.s. Since $\mathcal{G}^+_{nT}$ and $\mathcal{G}_{nT}$ are independent, we get that, for all $k \ge nT$, $\mathbb{P}$-a.s.,

$$\mathbb{E}\big[H\big(\theta, \mathbb{E}\big[X_k|\mathcal{G}^+_{nT}\big]\big)|\mathcal{H}_{nT}\big] = \mathbb{E}\big[H\big(\theta, \mathbb{E}\big[X_k|\mathcal{G}^+_{nT}\big]\big)|\mathcal{G}_{nT}\big] = \mathbb{E}\big[H\big(\theta, \mathbb{E}\big[X_k|\mathcal{G}^+_{nT}\big]\big)\big].$$

This implies that, for all $k \ge nT$,

$$\begin{aligned}
\big\| h_{k,nT}(\theta) - h(\theta) \big\| &\le \big\| \mathbb{E}\big[H(\theta, X_k)|\mathcal{G}_{nT}\big] - \mathbb{E}\big[H\big(\theta, \mathbb{E}\big[X_k|\mathcal{G}^+_{nT}\big]\big)|\mathcal{G}_{nT}\big] \big\| \\
&\quad + \big\| \mathbb{E}\big[H\big(\theta, \mathbb{E}\big[X_k|\mathcal{G}^+_{nT}\big]\big)\big] - \mathbb{E}\big[H(\theta, X_k)\big] \big\| \\
&\le L_2 \mathbb{E}\big[\big\| X_k - \mathbb{E}\big[X_k|\mathcal{G}^+_{nT}\big] \big\| |\mathcal{G}_{nT}\big] + L_2 \mathbb{E}\big[\big\| X_k - \mathbb{E}\big[X_k|\mathcal{G}^+_{nT}\big] \big\|\big].
\end{aligned}$$

Using the Minkowski inequality, we get

$$\begin{aligned}
&\mathbb{E}^{1/p}\Big[\sup_{\theta \in \mathbb{R}^d} \big\| h_{k,nT}(\theta) - h(\theta) \big\|^p\Big] \\
&\le L_2 \mathbb{E}^{1/p}\big[\big\| X_k - \mathbb{E}\big[X_k|\mathcal{G}^+_{nT}\big] \big\|^p\big] + L_2 \mathbb{E}\big[\big\| X_k - \mathbb{E}\big[X_k|\mathcal{G}^+_{nT}\big] \big\|\big] \\
&\le 2L_2 \sum_{i=1}^{m} \gamma^0_p(X, k - nT, i),
\end{aligned}$$

noting that $\mathcal{G}_0$ is the trivial sigma algebra. This concludes the proof since $\mathbb{E}[\Gamma^0_p(X)] \le \mathcal{C}_{p,1}(X)$. □

Proof of Theorem 3.5. Fix $n \in \mathbb{N}$ and let $nT \le k < (n+1)T$ be an arbitrary integer. By the triangle inequality, the difference of θ^λ and $\overline{\theta}^\lambda$ is decomposed into two parts

$$\|\theta_k^\lambda - \overline{\theta}_k^\lambda\| \le \|\theta_k^\lambda - \overline{z}_k^\lambda\| + \|\overline{z}_k^\lambda - \overline{\theta}_k^\lambda\|. \tag{38}$$

Let $h_{k,nT}$ be the functional constructed in Lemma 4.8. Then, one estimates

$$\begin{aligned}
\|\theta_k^\lambda - \overline{z}_k^\lambda\| &\le \lambda \left\| \sum_{i=nT}^{k-1} \big(H(\theta_i^\lambda, X_i) - h(\overline{z}_i^\lambda)\big) \right\| \\
&\le \lambda \sum_{i=nT}^{k-1} \|H(\theta_i^\lambda, X_i) - H(\overline{z}_i^\lambda, X_i)\| + \lambda \left\| \sum_{i=nT}^{k-1} \big(H(\overline{z}_i^\lambda, X_i) - h_{i,nT}(\overline{z}_i^\lambda)\big) \right\| \\
&\quad + \lambda \sum_{i=nT}^{k-1} \|h_{i,nT}(\overline{z}_i^\lambda) - h(\overline{z}_i^\lambda)\| \\
&\le \lambda L_1 \sum_{i=nT}^{k-1} \|\theta_i^\lambda - \overline{z}_i^\lambda\| + \lambda \max_{nT \le m < (n+1)T} \left\| \sum_{i=nT}^{m} \big(H(\overline{z}_i^\lambda, X_i) - h_{i,nT}(\overline{z}_i^\lambda)\big) \right\| \\
&\quad + \lambda \sum_{i=nT}^{\infty} \|h_{i,nT}(\overline{z}_i^\lambda) - h(\overline{z}_i^\lambda)\|
\end{aligned}$$

due to Assumption 3.2. Thanks to Lemmas 4.4, 4.5, 4.6, and 4.9 all the terms on the RHS of the previous inequality are almost surely finite. A discrete-time version of Grönwall's lemma and taking squares lead to

$$\begin{aligned}
\|\theta_k^\lambda - \overline{z}_k^\lambda\|^2 \le 2\lambda^2 \mathrm{e}^{2L_1 T\lambda} \Bigg[&\max_{nT \le m < (n+1)T} \left\| \sum_{i=nT}^{m} \big(H(\overline{z}_i^\lambda, X_i) - h_{i,nT}(\overline{z}_i^\lambda)\big) \right\|^2 \\
&+ \left(\sum_{i=nT}^{\infty} \|h_{i,nT}(\overline{z}_i^\lambda) - h(\overline{z}_i^\lambda)\| \right)^2 \Bigg],
\end{aligned}$$

noting also $(x+y)^2 \le 2(x^2+y^2)$, $x, y \in \mathbb{R}$. Let us define the $\mathcal{H}_{nT}$-measurable random variable

$$N_{nT} := \max_{nT \le k < (n+1)T} \|\overline{z}_k^\lambda - \theta^*\|.$$

Now, by recalling the definition of T and taking $\mathcal{H}_{nT}$-conditional expectations, one obtains

$$\begin{aligned}
\mathbb{E}^{1/2}\big[\|\theta_k^\lambda - \overline{z}_k^\lambda\|^2 | \mathcal{H}_{nT}\big] \le \sqrt{2}\lambda \mathrm{e}^{L_1} &\sum_{j=1}^{\infty} \mathbb{1}_{\{j-1 \le N_{nT} < j\}} \\
&\times \mathbb{E}^{1/2}\left[\max_{nT \le m < (n+1)T} \left\| \sum_{i=nT}^{m} \big(H(\overline{z}_i^\lambda, X_i) - h_{i,nT}(\overline{z}_i^\lambda)\big) \right\|^2 \Big| \mathcal{H}_{nT} \right] \\
&+ \sqrt{2}\lambda \mathrm{e}^{L_1} \sup_{\theta \in \mathbb{R}^d} \sum_{i=nT}^{\infty} \|h_{i,nT}(\theta) - h(\theta)\|.
\end{aligned}$$

Define for $n \in \mathbb{N}$,

$$\tilde{Z}^{\lambda}_{n,k}(j) := \begin{cases} H\big(\overline{z}^{\lambda}_k, X_k\big)\mathbb{1}_{\{\|\overline{z}^{\lambda}_k - \theta^*\| \le j\}}, & nT \le k < (n+1)T, \\ 0, & \text{otherwise.} \end{cases} \tag{39}$$

Recalling the $\mathcal{H}_{nT}$-measurability of $\overline{z}^{\lambda}_k$, $nT \le k < (n+1)T$, and arguing like in Lemma 4.7, one obtains

$$\begin{aligned} M^{nT}_r\big(\tilde{Z}^{\lambda}_n(j)\big) &\le L_1 j + L_2 M^{nT}_r(X) + H^*, \\ \Gamma^{nT}_r\big(\tilde{Z}^{\lambda}_n(j)\big) &\le 2L_2 \Gamma^{nT}_r(X). \end{aligned} \tag{40}$$

With these notation, for each $j \in \mathbb{N}$, the process defined by

$$Z^{\lambda}_{n,k}(j) := \big(H\big(\overline{z}^{\lambda}_k, X_k\big) - h_{k,nT}\big(\overline{z}^{\lambda}_k\big)\big)\mathbb{1}_{\{\|\overline{z}^{\lambda}_k - \theta^*\| \le j\}} = \tilde{Z}^{\lambda}_{n,k}(j) - \mathbb{E}\big[\tilde{Z}^{\lambda}_{n,k}(j)|\mathcal{H}_{nT}\big], \tag{41}$$

for $nT \le k < (n+1)T$, $n \in \mathbb{N}$ satisfies

$$\begin{aligned} M^{nT}_r\big(Z^{\lambda}_n(j)\big) &\le 2\big[L_1 j + L_2 M^{nT}_r(X) + H^*\big], \\ \Gamma^{nT}_r\big(Z^{\lambda}_n(j)\big) &\le 2L_2 \Gamma^{nT}_r(X). \end{aligned} \tag{42}$$

Notice that $Z^{\lambda}_{n,nT}(j) = 0$ hence the maximum can be taken over $nT < m < (n+1)T$ instead of $nT \le m < (n+1)T$. One then applies Theorem B.4 with the choice $n = nT$, $r = 3$, $b_i \equiv 1$, $X_k := Z^{\lambda}_{n,k}(j)$ to obtain

$$\begin{aligned} &\mathbb{1}_{\{N_{nT} \le j\}} \mathbb{E}^{1/2}\left[\max_{nT < m < (n+1)T} \left\| \sum_{i=nT+1}^{m} \big(H\big(\overline{z}^{\lambda}_i, X_i\big) - h_{i,nT}\big(\overline{z}^{\lambda}_i\big)\big)\right\|^2 \Big| \mathcal{H}_{nT}\right] \\ &\quad \le \mathbb{1}_{\{N_{nT} \le j\}} \mathbb{E}^{1/3}\left[\max_{nT < m < (n+1)T} \left\| \sum_{i=nT+1}^{m} \big(H\big(\overline{z}^{\lambda}_i, X_i\big) - h_{i,nT}\big(\overline{z}^{\lambda}_i\big)\big)\right\|^3 \Big| \mathcal{H}_{nT}\right] \\ &\quad \le 10 \mathbb{1}_{\{N_{nT} \le j\}} \sqrt{T}\big[\Gamma^{nT}_3\big(Z^{\lambda}_n(j)\big) + M^{nT}_3\big(Z^{\lambda}_n(j)\big)\big], \end{aligned} \tag{43}$$

noting that $C'(3) \le 10$ holds for the constant $C'(3)$ appearing in Theorem B.4.

Now we turn to estimating N_{nT}. Let $q > 1$ be an arbitrary integer. Let us apply Lemma A.1 with the choice $r := 2$ and $p := 2q$ to obtain

$$\mathbb{E}\big[N^2_{nT}\big] \le T^{2/(2q)} \sup_{nT \le k < (n+1)T} \mathbb{E}^{2/(2q)}\big[\big\|\overline{z}^{\lambda}_k - \theta^*\big\|^{2q}\big], \tag{44}$$

which implies, by Lemma 4.5,

$$\mathbb{E}\big[(N_{nT}+1)^2\big] \le 2\big[1 + T^{2/(2q)} \underline{C}^2(q)\big]. \tag{45}$$

By the Cauchy–Schwarz inequality, (42) and (45) we can perform the auxiliary estimate

$$\begin{aligned} &\sum_{j=1}^{\infty} \mathbb{E}\big[\mathbb{1}_{\{j-1 \le N_{nT} < j\}}\big[\Gamma^{nT}_3\big(Z^{\lambda}_n(j)\big) + M^{nT}_3\big(Z^{\lambda}_n(j)\big)\big]^2\big] \\ &\quad \le 8 \sum_{j=1}^{\infty} \mathbb{E}[\mathbb{1}_{\{j-1 \le N_{nT} < j\}}\big[L^2_2\big(\Gamma^{nT}_3(X)\big)^2 + \big[L_1 j + L_2 M^{nT}_3(X) + H^*\big]^2\big] \end{aligned}$$

$$\begin{aligned}
&\leq 8L_2^2\mathbb{E}\big[\big(\Gamma_3^{nT}(X)\big)^2\big] + 24[\mathbb{E}\big[L_1^2(N_{nT}+1)^2 + L_2^2\big(M_3^{nT}(X)\big)^2 + \big(H^*\big)^2\big] \\
&\leq 8L_2^2\mathcal{C}_{3,2} + 24\big[L_2^2\mathcal{M}_3^{2/3} + \big(H^*\big)^2\big] + 48L_1^2\big[1+T^{2/2q}\big]\underline{C}^2(q) \\
&\leq 96T^{2/p}\big[L_1^2\underline{C}^2(q) + L_2^2\mathcal{C}_{3,2} + L_2^2\mathcal{M}_3^{2/3} + \big(H^*\big)^2\big], \qquad (46)
\end{aligned}$$

using the notation introduced for conditionally L-mixing processes in (10) and the trivial $T \geq 1$ (in the last inequality). We define

$$C^\sharp(p) := 96\big[L_1^2\underline{C}^2(p/2) + L_2^2\mathcal{C}_{3,2} + L_2^2\mathcal{M}_3^{2/3} + \big(H^*\big)^2\big] + 4L_2^2\mathcal{C}_{2,1}^2.$$

Notice that $(C^\sharp)^{1/2} \leq C^\star$, where the latter constant is given by

$$C^\star(p) := 10\big[L_1\underline{C}(p/2) + L_2\mathcal{C}_{3,2}^{1/2} + L_2\mathcal{M}_3^{1/3} + H^*\big] + 2L_2\mathcal{C}_{2,1}. \qquad (47)$$

We conclude from (43), (46) and (47) that

$$\mathbb{E}^{1/2}\big\|\theta_k^\lambda - \overline{z}_k^\lambda\big\|^2 \leq 15e^{L_1}C^\star(p)\big[\lambda\sqrt{T}T^{1/p} + \lambda\big] \leq 30\mathrm{e}^{L_1}C^\star(p)\lambda^{\frac{1}{2}-\frac{1}{p}},$$

for all $k \in \mathbb{N}$, noting also that $\sqrt{2} \leq 3/2$.

Now we turn to estimating $\|\overline{z}_k^\lambda - \overline{\theta}_k^\lambda\|$ for $nT \leq k < (n+1)T$. We compute

$$\begin{aligned}
\big\|\overline{z}_k^\lambda - \overline{\theta}_k^\lambda\big\|_2 &\leq \sum_{i=1}^n \big\|z^\lambda\big(k, iT, \theta_{iT}^\lambda\big) - z^\lambda\big(k, (i-1)T, \theta_{(i-1)T}^\lambda\big)\big\|_2 \\
&= \sum_{i=1}^n \big\|z^\lambda\big(k, iT, \theta_{iT}^\lambda\big) - z^\lambda\big(k, iT, z^\lambda\big(iT, (i-1)T, \theta_{(i-1)T}^\lambda\big)\big)\big\|_2.
\end{aligned}$$

By Lemma 4.2-iv, we estimate

$$\begin{aligned}
&\big\|z^\lambda\big(k, iT, \theta_{iT}^\lambda\big) - z^\lambda\big(k, iT, z^\lambda\big(iT, (i-1)T, \theta_{(i-1)T}^\lambda\big)\big)\big\|_2 \\
&\leq (1-2\tilde{a}\lambda)^{k-iT}\big\|\theta_{iT}^\lambda - z^\lambda\big(iT, (i-1)T, \theta_{(i-1)T}^\lambda\big)\big\|_2 \\
&\leq (1-2\tilde{a}\lambda)^{k-iT}\big\|\theta_{iT-1}^\lambda - \lambda H\big(\theta_{iT-1}^\lambda, X_{iT}\big) - \overline{z}_{iT-1}^\lambda + \lambda h\big(\overline{z}_{iT-1}^\lambda\big)\big\|_2 \\
&\leq (1-2\tilde{a}\lambda)^{k-iT}\big[\big\|\theta_{iT-1}^\lambda - \overline{z}_{iT-1}^\lambda\big\|_2 + \lambda\big\|H\big(\theta_{iT-1}^\lambda, X_{iT}\big) - h\big(\overline{z}_{iT-1}^\lambda\big)\big\|_2\big].
\end{aligned}$$

Using Lemma 4.6, the estimation continues as follows

$$\begin{aligned}
\big\|\overline{z}_k^\lambda - \overline{\theta}_k^\lambda\big\|_2 &\leq \sum_{i=1}^n \mathrm{e}^{-2\tilde{a}\lambda(k-iT)}\big[\big\|\theta_{iT-1}^\lambda - \overline{z}_{iT-1}^\lambda\big\|_2 + \lambda\big\|H\big(\theta_{iT-1}^\lambda, X_{iT}\big) - h\big(\overline{z}_{iT-1}^\lambda\big)\big\|_2\big] \\
&\leq \sum_{i=1}^n \mathrm{e}^{-2a\lambda(n-i)T}\big[30\mathrm{e}^{L_1}C^\star(p)\lambda^{\frac{1}{2}-\frac{1}{p}} + C^\flat\lambda\big] \\
&\leq \frac{30\mathrm{e}^{L_1}C^\star(p) + C^\flat}{1-\mathrm{e}^{-2\tilde{a}\lambda T}}\lambda^{\frac{1}{2}-\frac{1}{p}} \leq \frac{30\mathrm{e}^{L_1}C^\star(p) + C^\flat}{1-\mathrm{e}^{-\tilde{a}}}\lambda^{\frac{1}{2}-\frac{1}{p}}.
\end{aligned}$$

The proof is completed by setting

$$C_0(p) := \frac{30\mathrm{e}^{L_1} C^\star(p) + C^\flat}{1 - \mathrm{e}^{-\tilde{a}}} + C^\star(p) \tag{48}$$

and noting (38). □

Remark 4.10. We track the dependence of the constant $C_0(p)$ (appearing in Theorem 3.5) on the dimension d. Notice that Lemmata 4.3 4.4 provide $c'(p)$ and $c''(p)$, both of which of the order $\sqrt{d}$. This order is inherited by $\underline{C}(q)$ in Lemma 4.5 and thus results in $d^{1/2}$ in $C^\star(p)$ and $C^\flat$, see (47) and (35). We finally get that $C_0(p)$ is of the order $d^{1/2}$.

4.4. Proof of Theorem 3.6

To prove Theorem 3.6, another convergence result is needed, which is the rate of convergence to stationarity of the recursive scheme (2) in Wasserstein-2 distance. Note that with Lemma 4.1 and 4.2, the convergence in Wasserstein-2 distance can be considered. The following is the adapted statement in [10], Corollary 7, using the notation of this article.

Theorem 4.11 ([10], Corollary 7). *Let Assumptions 3.1, 3.2, 3.3 hold and let $\lambda < \bar{\lambda}$ where $\bar{\lambda}$ is defined in (18). Then, the Markov chain $(\overline{\theta}^\lambda_n)_{n\in\mathbb{N}}$ admits an invariant measure π_λ such that, for all $n \in \mathbb{N}$;*

$$W_2\big(\mathrm{Law}\big(\overline{\theta}^\lambda_n\big), \pi_\lambda\big) \le \hat{c}\mathrm{e}^{-a\lambda n}, \quad n \in \mathbb{N},$$

where $\hat{c}$ is coming from (iii) Lemma 4.2:

$$\hat{c} := \sqrt{2}\big(\|\theta_0 - \theta\|^2 + d/\tilde{a}\big)^{1/2}.$$

Furthermore,

$$W_2(\pi, \pi_\lambda) \le c\sqrt{\lambda},$$

where

$$c = \left(L_1^2\tilde{a}^{-1}\big(2\lambda + \tilde{a}^{-1}\big)\left(d + \frac{1}{12}\lambda^2 L_1^2 d + \frac{1}{2}L_1^2\lambda d/a\right)\right)^{1/2}$$

with $\tilde{a}$ defined in (15).

Note that for the Langevin SDE (1), the Euler and Milstein schemes coincide, which implies that the optimal rate of convergence for scheme (2) is 1 instead of 1/2. The bound provided in Theorem 4.11 can thus be improved under an additional smoothness assumption for the drift coefficient of (1). However, as our main focus is the behaviour of the SGLD algorithm (5) and, in view of Example 3.4, it is known that its optimal rate of convergence is 1/2, any improvement on the behaviour of scheme (2) does not change this fact.

Proof of Theorem 3.6. Take p large enough so that $\kappa > 2/(p-1)$ and thus $1/p \le \kappa/(\kappa+2)$ holds. Denote by $\tilde{C} = \max\{C_0(p), \hat{c}, c\}$. Theorems 3.5 and 4.11 imply that

$$W_2\big(\mathrm{Law}\big(\theta^\lambda_n\big), \pi\big) \le W_2\big(\mathrm{Law}\big(\theta^\lambda_n\big), \mathrm{Law}\big(\overline{\theta}^\lambda_n\big)\big) + W_2\big(\mathrm{Law}\big(\overline{\theta}^\lambda_n\big), \pi_\lambda\big) + W_2(\pi_\lambda, \pi)$$

$$\leq \tilde{C}\big[\lambda^{\frac{1}{2}-\frac{3}{2p}} + \mathrm{e}^{-a\lambda n} + \lambda^{\frac{1}{2}}\big]$$
$$\leq 2\tilde{C}\big[\lambda^{\frac{1}{2+\kappa}} + \mathrm{e}^{-a\lambda n}\big].$$

For $0 < \epsilon < \mathrm{e}^{-1}$, choosing $\lambda := \epsilon^{2+\kappa}/(4\tilde{C})^{2+\kappa}$, $2\tilde{C}\lambda^{\frac{1}{2+\kappa}} \leq \epsilon/2$ holds. Now it remains to choose n large enough to have $\tilde{C}\mathrm{e}^{-a\lambda n} \leq \epsilon/2$ or, equivalently, $a\lambda n \geq \ln(2\tilde{C}/\epsilon)$. Noting the choice of λ and $\ln(1/\epsilon) \geq 1$, this is possible if

$$n \geq \frac{c_2(\kappa)}{\epsilon^{2+\kappa}} \ln(1/\epsilon),$$

where $c_2(\kappa) = \frac{(4\tilde{C})^{2+\kappa}}{a}(1 + \ln(2\tilde{C}))$. □

5. Proof of main results: Independent data

For the case of independent data, it is enough to obtain the second moment of the SGLD scheme (5) before considering the convergence in Wasserstein-2 distance. The following lemma provides an upper bound for the second moment of the scheme (5) with explicit constants.

Lemma 5.1. *Let Assumptions 3.7, 3.8 and 3.9 hold. Let*

$$\lambda_0 := \min\big(a/2L_1^2\mathbb{E}\big[\big(1 + \|X_0\|\big)^{2\rho}\big], 1/a\big). \tag{49}$$

For $\lambda \leq \lambda_0$, the function $V_1(\theta) := \|\theta - \theta^*\|^2$ satisfies

$$\mathbb{E}\big[V_1\big(\theta_n^\lambda\big)|\theta_{n-1}^\lambda\big] \leq (1 - a\lambda)V_1\big(\theta_{n-1}^\lambda\big) + \lambda C,$$

where

$$C := 4L_2^2\big(1 + \big\|\theta^*\big\|\big)^2\mathbb{E}\big[\big(1 + \|X_0\|\big)^{2\rho+2}\big] + 4\big\{H^*\big\}^2 + 2d.$$

As a result, $\sup_{\lambda\leq\lambda_0}\sup_{n\in\mathbb{N}}\mathbb{E}[V_1(\theta_n^\lambda)] < \infty$. Moreover, if $\rho = 0$ in Assumption 3.7, then the above result is true for $\lambda \leq \min(1/2L_1, 1/(a + L_1))$.

Proof. By using the SGLD scheme (5), one calculates

$$\begin{aligned}
\big\|\theta_{n+1}^\lambda - \theta^*\big\|^2 &= \big\|\theta_n^\lambda - \theta^*\big\|^2 + 2\big\langle\theta_n^\lambda - \theta^*, -\lambda H\big(\theta_n^\lambda, X_{n+1}\big) + \sqrt{2\lambda}\xi_{n+1}\big\rangle \\
&\quad + \big\|-\lambda H\big(\theta_n^\lambda, X_{n+1}\big) + \sqrt{2\lambda}\xi_{n+1}\big\|^2 \\
&= \big\|\theta_n^\lambda - \theta^*\big\|^2 - 2\lambda\big\langle\theta_n^\lambda - \theta^*, H\big(\theta_n^\lambda, X_{n+1}\big) - H\big(\theta^*, X_{n+1}\big)\big\rangle \\
&\quad + 2\big\langle\theta_n^\lambda - \theta^*, \sqrt{2\lambda}\xi_{n+1}\big\rangle - 2\lambda\big\langle\theta_n^\lambda - \theta^*, H\big(\theta^*, X_{n+1}\big)\big\rangle \\
&\quad + \lambda^2\big\|H\big(\theta_n^\lambda, X_{n+1}\big)\big\|^2 - 2\lambda\big\langle H\big(\theta_n^\lambda, X_{n+1}\big), \sqrt{2\lambda}\xi_{n+1}\big\rangle + 2\lambda\|\xi_{n+1}\|^2
\end{aligned}$$

and thus

$$\begin{aligned}
&\mathbb{E}\big[\big\|\theta_{n+1}^\lambda - \theta^*\big\|^2|\theta_n^\lambda\big] \\
&\quad \leq \big\|\theta_n^\lambda - \theta^*\big\|^2 - 2\lambda\mathbb{E}\big[\big\langle\theta_n^\lambda - \theta^*, A(X_{n+1})\big(\theta_n^\lambda - \theta^*\big)\big\rangle|\theta_n^\lambda\big] - 2\lambda\big\langle\theta_n^\lambda - \theta^*, h\big(\theta^*\big)\big\rangle
\end{aligned}$$

$$
\begin{aligned}
&+\lambda^2\mathbb{E}\big[\big\|H\big(\theta_n^\lambda, X_{n+1}\big)\big\|^2|\theta_n^\lambda\big]+2\lambda d\\
&\le \big\|\theta_n^\lambda-\theta^*\big\|^2-2\lambda a\big\|\theta_n^\lambda-\theta^*\big\|^2+2\lambda^2\mathbb{E}\big[\big\|H\big(\theta_n^\lambda, X_{n+1}\big)-H\big(\theta^*, X_{n+1}\big)\big\|^2|\theta_n^\lambda\big]\\
&+2\lambda^2\mathbb{E}\big[\big\|H\big(\theta^*, X_{n+1}\big)\big\|^2\big]+2\lambda d. \qquad (50)
\end{aligned}
$$

Hence, for $\lambda \le \min(a/2L_1^2\mathbb{E}[(1+\|X_0\|)^{2\rho}], 1/a)$

$$
\begin{aligned}
\mathbb{E}\big[\big\|\theta_{n+1}^\lambda-\theta^*\big\|^2|\theta_n^\lambda\big] &\le (1-\lambda a)\big\|\theta_n^\lambda-\theta^*\big\|^2+4\lambda^2L_2^2\big(1+\big\|\theta^*\big\|\big)^2\mathbb{E}\big[\big(1+\|X_0\|\big)^{2\rho+2}\big]\\
&\quad+4\lambda^2\big\{H^*\big\}^2+2\lambda d\\
\Rightarrow\quad \mathbb{E}\big(\big\|\theta_{n+1}^\lambda-\theta^*\big\|^2|\theta_n^\lambda\big) &\le (1-\lambda a)\big\|\theta_n^\lambda-\theta^*\big\|^2+\lambda C,
\end{aligned}
$$

where $C = 4L_2^2(1+\|\theta^*\|)^2\mathbb{E}[(1+\|X_0\|)^{2\rho+2}]+4\{H^*\}^2+2d$. Consequently, for any $n \ge 1$,

$$
\mathbb{E}\big[\big\|\theta_n^\lambda-\theta^*\big\|^2\big] \le (1-\lambda a)^n\mathbb{E}\big[\big\|\theta_0-\theta^*\big\|^2\big]+\frac{C}{a}<\infty.
$$

Crucially, one observes here that if $\rho = 0$ in Assumption 3.7, then H is co-coercive with the following property, for every $x \in \mathbb{R}^m$ and all $\theta, \theta^* \in \mathbb{R}^d$

$$
\big\langle\theta-\theta', H(\theta,x)-H\big(\theta',x\big)\big\rangle \ge \frac{1}{L_1}\big\|H(\theta,x)-H\big(\theta',x\big)\big\|^2. \qquad (51)
$$

It follows that, in view of (51), one rewrites (50) as follows

$$
\begin{aligned}
&\mathbb{E}\big[\big\|\theta_{n+1}^\lambda-\theta^*\big\|^2|\theta_n^\lambda\big]\\
&\le \big\|\theta_n^\lambda-\theta^*\big\|^2-\lambda\mathbb{E}\big[\big\langle\theta_n^\lambda-\theta^*, A(X_{n+1})\big(\theta_n^\lambda-\theta^*\big)\big\rangle|\theta_n^\lambda\big]\\
&\quad-\frac{\lambda}{L_1}\big\|H\big(\theta_n^\lambda, X_{n+1}\big)-H\big(\theta^*, X_{n+1}\big)\big\|^2+2\lambda\big\langle\theta_n^\lambda-\theta^*, h\big(\theta^*\big)\big\rangle\\
&\quad+\lambda^2\mathbb{E}\big[\big\|H\big(\theta_n^\lambda, X_{n+1}\big)\big\|^2|\theta_n^\lambda\big]+2\lambda d\\
&\le \big\|\theta_n^\lambda-\theta^*\big\|^2-\lambda a\big\|\theta_n^\lambda-\theta^*\big\|^2+\left(2\lambda^2-\frac{\lambda}{L_1}\right)\mathbb{E}\big[\big\|H\big(\theta_n^\lambda, X_{n+1}\big)-H\big(\theta^*, X_{n+1}\big)\big\|^2\big]\\
&\quad+2\lambda^2\mathbb{E}\big[\big\|H\big(\theta^*, X_{n+1}\big)\big\|^2\big]+2\lambda d,
\end{aligned}
$$

which yields, for $\lambda \le 1/2L_1$

$$
\begin{aligned}
\mathbb{E}\big[\big\|\theta_{n+1}^\lambda-\theta^*\big\|^2|\theta_n^\lambda\big] &\le (1-\lambda a)\big\|\theta_n^\lambda-\theta^*\big\|^2+4\lambda^2L_2^2\big(1+\big\|\theta^*\big\|\big)^2\mathbb{E}\big[\big(1+\|X_0\|\big)^2\big]\\
&\quad+4\lambda^2\big\{H^*\big\}^2+2\lambda d\\
\Rightarrow\quad \mathbb{E}\big(\big\|\theta_{n+1}^\lambda-\theta^*\big\|^2|\theta_n^\lambda\big) &\le (1-\lambda a)\big\|\theta_n^\lambda-\theta^*\big\|^2+\lambda C,
\end{aligned}
$$

where $C = 4L_2^2(1+\|\theta^*\|)^2\mathbb{E}[(1+\|X_0\|)^2]+4\{H^*\}^2+2d$. □

Proof of Theorem 3.10. One notes that (B1) is still valid, with the only difference that the Lipschitz constant in (B1) is given by $L_1\mathbb{E}[(1+\|X_0\|)^\rho]$, and (B2) holds with a. Consequently, Theorem 4.11

is still true. The main steps of the proof of Theorem 3.5 need to be reformulated for the i.d.d. case. Initially, one notes that the following result holds due to Lemma 5.1

$$\sup_{\lambda\in(0,\lambda_0)}\sup_{n\geq 0}\mathbb{E}\big[\big\|\theta_n^\lambda\big\|^2\big]<c_0,$$

where $c_0=2\mathbb{E}\|\theta_0-\theta^*\|^2+2C/a+2\|\theta^*\|^2$, and C is given explicitly in Lemma 5.1. Then, using synchronous coupling for the schemes (2) and (5), one obtains

$$\begin{aligned}
&\big\|\theta_{n+1}^\lambda-\bar\theta_{n+1}^\lambda\big\|^2\\
&\quad=\big\|\theta_n^\lambda-\bar\theta_n^\lambda-\lambda\big(H\big(\theta_n^\lambda,X_{n+1}\big)-h\big(\bar\theta_n^\lambda\big)\big)\big\|^2\\
&\quad\leq\big\|\theta_n^\lambda-\bar\theta_n^\lambda\big\|^2-2\lambda\big\langle\theta_n^\lambda-\bar\theta_n^\lambda,H\big(\theta_n^\lambda,X_{n+1}\big)-h\big(\bar\theta_n^\lambda\big)\big\rangle+\lambda^2\big\|H\big(\theta_n^\lambda,X_{n+1}\big)-h\big(\bar\theta_n^\lambda\big)\big\|^2\\
&\quad\leq\big\|\theta_n^\lambda-\bar\theta_n^\lambda\big\|^2-2\lambda\big\langle\theta_n^\lambda-\bar\theta_n^\lambda,h\big(\theta_n^\lambda\big)-h\big(\bar\theta_n^\lambda\big)\big\rangle-2\lambda\big\langle\theta_n^\lambda-\bar\theta_n^\lambda,H\big(\theta_n^\lambda,X_{n+1}\big)-h\big(\theta_n^\lambda\big)\big\rangle\\
&\qquad+2\lambda^2\big\|H\big(\theta_n^\lambda,X_{n+1}\big)-h\big(\theta_n^\lambda\big)\big\|^2+2\lambda^2\big\|h\big(\theta_n^\lambda\big)-h\big(\bar\theta_n^\lambda\big)\big\|^2.
\end{aligned}$$

Taking expectations on both sides and using (15) yields

$$\begin{aligned}
&\mathbb{E}\big[\big\|\theta_{n+1}^\lambda-\bar\theta_{n+1}^\lambda\big\|^2|\theta_n^\lambda,\bar\theta_n^\lambda\big]\\
&\quad\leq\big\|\theta_n^\lambda-\bar\theta_n^\lambda\big\|^2-2\lambda\tilde a\big\|\theta_n^\lambda-\bar\theta_n^\lambda\big\|^2-\frac{2\lambda}{a+L_1}\big\|h\big(\theta_n^\lambda\big)-h\big(\bar\theta_n^\lambda\big)\big\|^2\\
&\qquad+2\lambda^2\mathbb{E}\big[\big\|H\big(\theta_n^\lambda,X_{n+1}\big)-h\big(\theta_n^\lambda\big)\big\|^2|\theta_n^\lambda,\bar\theta_n^\lambda\big]+2\lambda^2\big\|h\big(\theta_n^\lambda\big)-h\big(\bar\theta_n^\lambda\big)\big\|^2,
\end{aligned}$$

where $\tilde a$ is defined in (16). Hence, for $\lambda\leq 1/(a+L_1)$,

$$\begin{aligned}
\mathbb{E}\big[\big\|\theta_{n+1}^\lambda-\bar\theta_{n+1}^\lambda\big\|^2|\theta_n^\lambda,\bar\theta_n^\lambda\big]&\leq(1-\lambda\tilde a)\big\|\theta_n^\lambda-\bar\theta_n^\lambda\big\|^2\\
&\quad+2\lambda^2\mathbb{E}\big[\big\|H\big(\theta_n^\lambda,X_{n+1}\big)-\mathbb{E}\big[H\big(\theta_n^\lambda,X_{n+1}\big)\big]|\theta_n^\lambda,\bar\theta_n^\lambda\big]\big\|^2|\theta_n^\lambda,\bar\theta_n^\lambda\big].
\end{aligned}$$

Thus, due to Lemma A.2,

$$\begin{aligned}
&\mathbb{E}\big[\big\|\theta_{n+1}^\lambda-\bar\theta_{n+1}^\lambda\big\|^2|\theta_n^\lambda,\bar\theta_n^\lambda\big]\\
&\quad\leq(1-\lambda\tilde a)\big\|\theta_n^\lambda-\bar\theta_n^\lambda\big\|^2+8\lambda^2\mathbb{E}\big[\big\|H\big(\theta_n^\lambda,X_{n+1}\big)-H\big(\theta_n^\lambda,\mathbb{E}\big[X_{n+1}|\theta_n^\lambda,\bar\theta_n^\lambda\big]\big)\big\|^2|\theta_n^\lambda,\bar\theta_n^\lambda\big]\\
&\quad\leq(1-\lambda\tilde a)\big\|\theta_n^\lambda-\bar\theta_n^\lambda\big\|^2+8\lambda^2L_2^2\big(1+\big\|\theta_n^\lambda\big\|\big)^2\operatorname{Var}_{\mathcal W}(X_0)
\end{aligned}$$

which implies that

$$\mathbb{E}\big[\big\|\theta_{n+1}^\lambda-\bar\theta_{n+1}^\lambda\big\|^2\big]\leq 8\lambda L_2^2\Big(1+\sup_{n\geq 0}\mathbb{E}\big[\big\|\theta_n^\lambda\big\|^2\big]\Big)]\operatorname{Var}_{\mathcal W}(X_0)\frac{1}{\tilde a},$$

where

$$\operatorname{Var}_{\mathcal W}(X_0):=\mathbb{E}\big[\big(1+\|X_0\|+\big\|\mathbb{E}[X_0]\big\|\big)^{2\rho}\big\|X_0-\mathbb{E}[X_0]\big\|^2\big].$$

Denote by $\bar{c} = \sqrt{8L_2^2(1+c_0)]\operatorname{Var}_{\mathcal{W}}(X_0)\frac{1}{\bar{a}}}$, one obtains $W_2(\operatorname{Law}(\theta_n^\lambda), \operatorname{Law}(\overline{\theta}_n^\lambda)) \le \bar{c}\lambda^{1/2}$. Then, together with Theorem 4.11, the following result can be obtained

$$\begin{aligned} W_2\big(\operatorname{Law}\big(\theta_n^\lambda\big), \pi\big) &\le W_2\big(\operatorname{Law}\big(\theta_n^\lambda\big), \operatorname{Law}\big(\overline{\theta}_n^\lambda\big)\big) + W_2\big(\operatorname{Law}\big(\overline{\theta}_n^\lambda\big), \pi_\lambda\big) + W_2(\pi_\lambda, \pi) \\ &\le \bar{C}\big[\lambda^{\frac{1}{2}} + \mathrm{e}^{-a\lambda n}\big], \end{aligned}$$

where $\bar{C} = \max\{\bar{c}, c_1, c\}$. For any $0 < \epsilon < 1/2$, by letting $\bar{C}\lambda^{\frac{1}{2}} < \epsilon/2$, and $\bar{C}\mathrm{e}^{-a\lambda n} \le \epsilon/2$, one obtains $\lambda < c_1\epsilon^2$ and $n > c_2\epsilon^{-2}\ln(1/\epsilon)$ with $c_1 = (4\bar{C})^{-1}$, $c_2 = (ac_1)^{-1}(\ln(2\bar{C}) + 1)$. □

Appendix A: Technical results

Lemma A.1. *Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of random variables such that for some $p > 0$, $M = \sup_{i\in\mathbb{N}} \mathbb{E}[\|X_i\|^p] < \infty$. Then for $0 < r < p$, $\mathbb{E}[\sup_{1\le i\le j} \|X_i\|^r] \le j^{r/p}M^{r/p}$.*

Proof. One has

$$\mathbb{E}^{p/r}\Big[\sup_{1\le i\le j}\|X_i\|^r\Big] \le \mathbb{E}\Big[\sup_{1\le i\le j}\|X_i\|^p\Big] \le \mathbb{E}\left[\sum_{i=1}^{j}\|X_i\|^p\right] \le jM,$$

by Jensen's inequality. □

Lemma A.2. *Let $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ be sigma-algebras. Let $p \ge 1$. Let X, Y be $\mathbb{R}$-valued random variables in L^p such that Y is measurable with respect to $\mathcal{H} \vee \mathcal{G}$. Then*

$$\mathbb{E}^{1/p}\big[\big\|X - \mathbb{E}[X|\mathcal{H}\vee\mathcal{G}]\big\|^p|\mathcal{G}\big] \le 2\mathbb{E}^{1/p}\big[\|X - Y\|^p|\mathcal{G}\big].$$

Proof. See [4], Lemma 6.1. □

Lemma A.3. *Let $x, y \in \mathbb{R}^d$, then*

$$\sum_{\substack{i+j+k=p\\ \{i\neq p-1\}\cap\{j\neq 1\}}} \frac{p!}{i!j!k!}\|x\|^{2i}\big(2\langle x,y\rangle\big)^j\|y\|^{2k} \le \sum_{\substack{k=0\\ k\neq 1}}^{2p}\binom{2p}{k}\|x\|^{2p-k}\|y\|^k.$$

Proof. Note that

$$\begin{aligned} &\sum_{\substack{i+j+k=p\\ \{i\neq p-1\}\cap\{j\neq 1\}}} \frac{p!}{i!j!k!}\|x\|^{2i}\big(2\langle x,y\rangle\big)^j\|y\|^{2k} \\ &\le \sum_{\substack{i+j+k=p\\ \{i\neq p-1\}\cap\{j\neq 1\}}} \frac{p!}{i!j!k!}\|x\|^{2i}\big(2\|x\|\|y\|\big)^j\|y\|^{2k}. \end{aligned} \tag{A.1}$$

Moreover,

$$\sum_{k=0}^{2p}\binom{2p}{k}\|x\|^{2p-k}\|y\|^k = \big(\|x\|+\|y\|\big)^{2p} = \big(\|x\|^2 + 2\|x\|\|y\| + \|y\|^2\big)^p$$

$$= \sum_{i+j+k=p} \frac{p!}{i!j!k!} \|x\|^{2i} \big(2\|x\|\|y\|\big)^j \|y\|^{2k}.$$

Consequently,

$$\sum_{\substack{k=0\\k\neq 1}}^{2p} \binom{2p}{k} \|x\|^{2p-k}\|y\|^k = \sum_{\substack{i+j+k=p\\ \{i\neq p-1\}\cap\{j\neq 1\}}} \frac{p!}{i!j!k!} \|x\|^{2i} \big(2\|x\|\|y\|\big)^j \|y\|^{2k}. \tag{A.2}$$

Thus, in view of (A.1) and (A.2), the desired result is obtained. □

Lemma A.4. *For each integer $r \geq 1$, $\mathbb{E}[\|\xi_1\|^{2r}] \leq 2^{2r} d^r r^{3r/2}$.*

Proof. Let $\zeta_1, \ldots, \zeta_d$ denote the coordinates of ξ_1. It is well known that $\mathbb{E}[\zeta_1^{2r}] = 2^r \Gamma([2r+1]/2)/\sqrt{\pi}$. Clearly,

$$\begin{aligned}\|\xi_1\|_{2r} &\leq \left(\sum_{i=1}^d \mathbb{E}^{1/r}\big[\zeta_i^{2r}\big]\right)^{1/2} = \big(2d\Gamma^{1/r}\big([2r+1]/2\big)\pi^{-1/(2r)}\big)^{1/2} \\ &\leq \sqrt{2d}\Gamma^{1/2r}(r+1)\pi^{-1/4r} \leq \sqrt{2d}\big(\sqrt{2\pi}r^{r+1/2}\mathrm{e}^{-r}\mathrm{e}^{1/(12r)}\big)^{1/2r}\pi^{-1/4r},\end{aligned}$$

where an estimate for the gamma function from [26] is used in the last inequality. Continuing in a somewhat rough way, one obtains

$$\|\xi_1\|_{2r} \leq 2\sqrt{d}r^{1/2+(1/4r)}\mathrm{e}^{-1/2}\mathrm{e}^{1/2} \leq 2\sqrt{d}r^{3/4}.$$ □

Appendix B: Proof of a pivotal inequality

In this section we prove the analogues of two moment inequalities from [12] for conditional L-mixing processes. One of these has already been shown in [4] but only under specific assumptions on the filtration. Our proofs (which mostly take place in continuous time) follow closely the arguments of [12]. There are, however, a number of small modifications that need to be pointed out.

We consider a continuous-time filtration $(\mathcal{R}_t)_{t\in\mathbb{R}_+}$ as well as a decreasing family of sigma-fields $(\mathcal{R}_t^+)_{t\in\mathbb{R}_+}$. We assume that $\mathcal{R}_t$ is independent of $\mathcal{R}_t^+$, for all $t \in \mathbb{R}_+$.

We consider an $\mathbb{R}^d$-valued continuous-time stochastic process $(X_t)_{t\in\mathbb{R}_+}$ which is progressively measurable (i.e., $X : [0,t] \times \Omega \to \mathbb{R}^d$ is $\mathcal{B}([0,t]) \otimes \mathcal{R}_t$-measurable for all $t \in \mathbb{R}_+$).

From now on we assume that $X_t \in L^1$, $t \in \mathbb{R}_+$. We define the quantities

$$\begin{aligned}\tilde{M}_r^i &:= \operatorname*{ess\,sup}_{t\in\mathbb{R}_+} \mathbb{E}^{1/r}\big[\big|X_t^i\big|^r|\mathcal{R}_0\big], \\ \tilde{\gamma}_r^i(\tau) &:= \operatorname*{ess\,sup}_{t\geq\tau} \mathbb{E}^{1/r}\big[\big|X_t^i - \mathbb{E}\big[X_t^i|\mathcal{R}_{t-\tau}^+ \vee \mathcal{R}_0\big]\big|^r|\mathcal{R}_0\big], \quad \tau \in \mathbb{R}_+,\end{aligned}$$

and set $M_r := \sum_{i=1}^d \tilde{M}_r^i$, $\tilde{\Gamma}_r^i := \sum_{\tau=0}^\infty \gamma_r^i(\tau)$ and $\Gamma_r := \sum_{i=1}^d \tilde{\Gamma}_r^i$ where X_t^i refers to the ith coordinate of X_t.

Remark B.1. If $d = 1$, $\mathcal{R}_0$ is trivial and $\mathcal{R}_t^+$, $t \in \mathbb{R}_+$ is right-continuous then we get back to the setting of [12]. It is shown in Lemma 9.1 of [12] that the (non-random) function $\tau \to \gamma_r(\tau)$, $\tau \in \mathbb{R}_+$ is measurable hence $\overline{\Gamma}_r := \int_0^\infty \gamma_r(\tau)\,\mathrm{d}\tau$ can be defined. [12], Theorems 1.1 and 5.1, formulate inequalities in terms of $\overline{\Gamma}_r$ instead of Γ_r.

We could attempt to define $\overline{\Gamma}_r$ for general $\mathcal{R}_0$ as a random variable but it requires further assumptions and tedious arguments which we do not pursue here. We stay with Γ_r which is easier to handle and it suffices for our purposes.

Theorem B.2. *Let $(X_t)_{t\in\mathbb{R}_+}$ be L^r-bounded for some $r \geq 2$ and let $M_r + \Gamma_r < \infty$ a.s. Assume $\mathbb{E}[X_t|\mathcal{R}_0] = 0$ a.s. for $t \in \mathbb{R}_+$. Let $f : [0, T] \to \mathbb{R}$ be $\mathcal{B}([0, T])$-measurable with $\int_0^T f_t^2\,\mathrm{d}t < \infty$. Then there is a constant $C(r)$ such that*

$$\mathbb{E}^{1/r}\left[\left|\int_0^T f_t X_t\,\mathrm{d}t\right|^r \Big|\mathcal{R}_0\right] \leq C(r)\left(\int_0^T f_t^2\,\mathrm{d}t\right)^{1/2}[M_r + \Gamma_r], \tag{B.1}$$

almost surely. We can actually take $C(r) = \sqrt{r-1}$.

Theorem B.3. *Let the conditions of Theorem B.2 hold for some $r > 2$. Then there is a constant $C'(r)$ such that*

$$\mathbb{E}^{1/r}\left[\sup_{s\in[0,T]}\left|\int_0^s f_t X_t\,\mathrm{d}t\right|^r \Big|\mathcal{R}_0\right] \leq C'(r)\left(\int_0^T f_t^2\,\mathrm{d}t\right)^{1/2}[M_r + \Gamma_r], \tag{B.2}$$

almost surely. We can actually take

$$C'(r) = \frac{\sqrt{r-1}}{2^{1/2} - 2^{1/r}}.$$

Note that the supremum in (B.2) can be taken along rationals hence it defines a random variable. We now state the corresponding results for conditionally L-mixing processes.

Theorem B.4. *Let $(X_n)_{n\in\mathbb{N}}$ be conditionally L-mixing of order $(r, 1)$ for some $r \geq 2$. Let b_i, $1 \leq i \leq m$ be real numbers. Then for each $n \in \mathbb{N}$*

$$\mathbb{E}\left[\left|\sum_{i=1}^m b_i X_{n+i}\right|^r \Big|\mathcal{F}_n\right] \leq C(r)\left(\sum_{i=1}^m b_i^2\right)^{1/2}\left[M_r^n(X) + \Gamma_r^n(X)\right],$$

almost surely. If $r > 2$, then also

$$\mathbb{E}\left[\left|\max_{1\leq k\leq m}\sum_{i=1}^k b_i X_{n+i}\right|^r \Big|\mathcal{F}_n\right] \leq C'(r)\left(\sum_{i=1}^m b_i^2\right)^{1/2}\left[M_r^n(X) + \Gamma_r^n(X)\right] \tag{B.3}$$

holds.

We are proceeding to the proofs of the above results. Since $\mathbb{E}[X_t|\mathcal{R}_{t-\tau_1}^+ \vee \mathcal{R}_0]$ is $\mathcal{R}_{t-\tau_2}^+ \vee \mathcal{R}_0$-measurable for $t \geq \tau_2 \geq \tau_1$, we obtain from Lemma A.2 with the choice $X = X_t$, $Y = \mathbb{E}[X_t|\mathcal{R}_{t-\tau_1}^+ \vee \mathcal{R}_0]$, $\mathcal{H} = \mathcal{R}_{t-\tau_2}^+$, $\mathcal{G} = \mathcal{R}_0$ that

$$\gamma_r(\tau_2) \leq 2\gamma_r(\tau_1). \tag{B.4}$$

We need a measure-theoretical lemma about real-valued random variables Y and Z.

Lemma B.5. *Let $r > 1$, $1/r + 1/q = 1$ and let $Y \in L^r$ be $\mathcal{R}_0 \vee \mathcal{R}_s^+$-measurable for some $s \geq 0$. Then for all $\mathcal{R}_s$-measurable $Z \in L^q$,*

$$\mathbb{E}[YZ|\mathcal{R}_0] = \mathbb{E}[Y|\mathcal{R}_0]\mathbb{E}[Z|\mathcal{R}_0].$$

Proof. Let $A \in \mathcal{R}_0$ be arbitrary. We assume $Y = \mathbb{1}_B \mathbb{1}_C$ with $B \in \mathcal{R}_0$, $C \in \mathcal{R}_s^+$ and $Z = \mathbb{1}_D$ with $D \in \mathcal{R}_s$. Then we find that, by independence of $\mathcal{R}_s$ from $\mathcal{R}_s^+$ and by $\mathcal{R}_0 \subset \mathcal{R}_s$,

$$\begin{aligned}\mathbb{E}[\mathbb{1}_A YZ] &= \mathbb{P}(C)\mathbb{P}(A \cap B \cap D) = \mathbb{P}(C)\mathbb{E}\big[\mathbb{1}_{A\cap B}\mathbb{E}[\mathbb{1}_D|\mathcal{R}_0]\big] \\ &= \mathbb{E}\big[\mathbb{1}_A \mathbb{1}_B \mathbb{P}(C)\mathbb{E}[\mathbb{1}_D|\mathcal{R}_0]\big] = \mathbb{E}\big[\mathbb{1}_A \mathbb{E}[Y|\mathcal{R}_0]\mathbb{E}[Z|\mathcal{R}_0]\big],\end{aligned}$$

which proves the statement for this Y and Z. Now, by standard arguments, one can extend these to $Y = \mathbb{1}_G$ for all $G \in \mathcal{R}_0 \vee \mathcal{R}_s^+$. We thus obtain the result for step functions Y, Z; then for bounded measurable functions and finally we arrive at the general statement. □

Now we formulate, in the present setting, the analogue of [12], Lemma 2.3.

Lemma B.6. *Let the assumptions of Theorem B.2 be in force. Let $d = 1$ and $1/r + 1/q = 1$. We have, for all $0 \leq s \leq t$,*

$$\big|\mathbb{E}[X_t \eta|\mathcal{R}_0]\big| \leq \gamma_r(t-s)\mathbb{E}^{1/q}\big[|\eta|^q|\mathcal{R}_0\big]$$

for each $\eta \in L^q$ which is $\mathcal{R}_s$-measurable.

Proof. Using Lemma B.5,

$$\mathbb{E}[X_t\eta|\mathcal{R}_0] = \mathbb{E}\big[\mathbb{E}\big[X_t|\mathcal{R}_s^+ \vee \mathcal{R}_0\big]|\mathcal{R}_0\big]\mathbb{E}[\eta|\mathcal{R}_0] + \mathbb{E}\big[\big(X_t - \mathbb{E}\big[X_t|\mathcal{R}_s^+ \vee \mathcal{R}_0\big]\big)\eta|\mathcal{R}_0\big].$$

Note that $\mathbb{E}[\mathbb{E}[X_t|\mathcal{R}_s^+ \vee \mathcal{R}_0]|\mathcal{R}_0] = \mathbb{E}[X_t|\mathcal{R}_0] = 0$. The conditional Hölder inequality implies that

$$\big|\mathbb{E}\big[\big(X_t - \mathbb{E}\big[X_t|\mathcal{R}_s^+ \vee \mathcal{R}_0\big]\big)\eta|\mathcal{R}_0\big]\big| \leq \gamma_r(t-s)\mathbb{E}^{1/q}\big[|\eta|^q|\mathcal{R}_0\big],$$

showing the statement. □

Proof of Theorem B.2. First let $d := 1$. For $t \in [0,T]$, define $I_t := \int_0^t f_s X_s \,\mathrm{d}s$ and $g_t := \mathbb{E}[|I_t|^r|\mathcal{R}_0]$. Following verbatim the arguments in the proof of [12], Theorem 1.1, we arrive at

$$|I_T|^r = \int_0^T \int_0^t r(r-1) f_t X_t f_s X_s |I_s|^{r-2} \,\mathrm{d}s \,\mathrm{d}t.$$

Hence, using Lemma B.6, $\lfloor t - s \rfloor \leq t - s$ and (B.4),

$$\begin{aligned}g_T &\leq \int_0^T \int_0^t r(r-1)\big|f_t f_s \mathbb{E}\big[X_t X_s |I_s|^{r-2}|\mathcal{R}_0\big]\big| \,\mathrm{d}s \,\mathrm{d}t \\ &\leq \int_0^T \int_0^t r(r-1)|f_t f_s| 2\gamma_r\big(\lfloor t-s \rfloor\big) M_r g_s^{1-2/r} \,\mathrm{d}s \,\mathrm{d}t \\ &= \int_0^T g_s^{1-2/r} r(r-1)|f_s| \int_s^T |f_t| 2\gamma_r\big(\lfloor t-s \rfloor\big) M_r \,\mathrm{d}t \,\mathrm{d}s\end{aligned}$$

almost surely, whereupon Lemma 2.5 of [12] implies

$$g_T^{1/r} \le \left(\frac{1}{r/2} \int_0^T r(r-1)|f_s| \int_s^T |f_t| 2\gamma_r\big(\lfloor t-s \rfloor\big) M_r \, \mathrm{d}t \, \mathrm{d}s \right)^{1/2}$$

almost surely. The Cauchy inequality leads to

$$g_T^{1/r} \le 2\sqrt{r-1} M_r^{1/2} \left(\int_0^T f_s^2 \, \mathrm{d}s \right)^{1/4} \left(\int_0^T \left(\int_s^T |f_t| \gamma_r\big(\lfloor t-s \rfloor\big) \, \mathrm{d}t \right)^2 \mathrm{d}s \right)^{1/4}.$$

Moreover, by the Minkowski inequality for the Hilbert space $L^2([0,T], \mathcal{B}([0,T]), \mathrm{Leb})$,

$$\begin{aligned}
&\left(\int_0^T \left(\int_s^T |f_t| \gamma_r\big(\lfloor t-s \rfloor\big) \, \mathrm{d}t \right)^2 \mathrm{d}s \right)^{1/2} \\
&\quad = \left(\int_0^T \left(\sum_{k=0}^\infty \gamma_r(k) \int_0^1 |f_{s+k+u}| \mathbb{1}_{\{s+k+u \le T\}} \, \mathrm{d}u \right)^2 \mathrm{d}s \right)^{1/2} \\
&\quad \le \sum_{k=0}^\infty \gamma_r(k) \left(\int_0^T \left(\int_0^1 |f_{s+k+u}| \mathbb{1}_{\{s+k+u \le T\}} \, \mathrm{d}u \right)^2 \mathrm{d}s \right)^{1/2} \\
&\quad \le \sum_{k=0}^\infty \gamma_r(k) \left(\int_0^T \int_0^1 f_{s+k+u}^2 \mathbb{1}_{\{s+k+u \le T\}} \, \mathrm{d}u \, \mathrm{d}s \right)^{1/2} \\
&\quad = \sum_{k=0}^\infty \gamma_r(k) \left(\int_0^1 \int_0^T f_{s+k+u}^2 \mathbb{1}_{\{s+k+u \le T\}} \, \mathrm{d}s \, \mathrm{d}u \right)^{1/2} \\
&\quad \le \sum_{k=0}^\infty \gamma_r(k) \left(\int_0^1 \int_{\min\{k+u,T\}}^T f_t^2 \, \mathrm{d}t \, \mathrm{d}u \right)^{1/2} \\
&\quad \le \sum_{k=0}^\infty \gamma_r(k) \left(\int_0^T f_t^2 \, \mathrm{d}t \right)^{1/2}.
\end{aligned}$$

Thus, we finally arrive at

$$g_T^{1/r} \le 2\sqrt{r-1} M_r^{1/2} \left(\int_0^T f_s^2 \, \mathrm{d}s \right)^{1/4} \left(\int_0^T f_t^2 \, \mathrm{d}t \right)^{1/4} \Gamma_r^{1/2},$$

which allows to conclude since $\sqrt{\Gamma_r M_r} \le [\Gamma_r + M_r]/2$. Now let d be arbitrary. Applying the one-dimensional result componentwise gives the result, noting the the Minkowski inequality and the definitions of M_r, Γ_r as sums of M_r^i, Γ_r^i, respectively. □

Proof of Theorem B.3. Again, let $d := 1$. Let $\mathcal{I} := \{(a,b) : 0 \le a < b \le T, \int_a^b f_s^2 \, \mathrm{d}s > 0\}$ and define, for $(a,b) \in \mathcal{I}$,

$$K_{a,b} := \frac{\sup_{t \in [a,b]} |\int_a^t f_s X_s \, \mathrm{d}s|^r}{\int_a^b f_s^2 \, \mathrm{d}s}$$

which is a random variable since the supremum can be taken along the rational numbers. Set $M_{a,b} := \mathbb{E}^{1/r}[K_{a,b}|\mathcal{R}_0]$. Define, furthermore

$$M_T^* := \operatorname{ess\,sup}_{(a,b)\in\mathcal{I}} M_{a,b}.$$

Noting Theorem B.2 and following verbatim the arguments in the proof of Theorem 5.1 in [12] we arrive at

$$M_T^* \le \frac{\sqrt{r-1}[M_r + \Gamma_r]}{\sqrt{2}} + \frac{2^{1/r}}{\sqrt{2}} M_T^*$$

almost surely, which implies

$$M_T^* \le \frac{\sqrt{r-1}[M_r + \Gamma_r]}{2^{1/2} - 2^{1/r}},$$

showing the statement. The case $d > 1$ follows by a componentwise application of the one-dimensional result. □

Proof of Theorem B.4. Fix $n \in \mathbb{N}$. We define the continuous-time process $\tilde{X}_0 := X_n$,

$$\tilde{X}_t := X_{n+k+1} \quad \text{for } k < t \le k+1, k \in \mathbb{N}.$$

Set $\mathcal{R}_t := \mathcal{F}_{n+\lceil t \rceil}$ and $\mathcal{R}_t^+ := \mathcal{F}_{n+\lceil t \rceil}^+$ for $t \in \mathbb{R}_+$. Notice that, for $\tau \in \mathbb{N}$, $\gamma_r(\tau)$ calculated for $(\tilde{X}_t, \mathcal{R}_t, \mathcal{R}_t^+)_{t\in\mathbb{R}_+}$ coincides with $\gamma_r^n(\tau, X)$ as defined in (8) and (9) for $(X_n, \mathcal{F}_n, \mathcal{F}_n^+)_{n\in\mathbb{N}}$. Similarly, M_r calculated for $\tilde{X}$ coincides with $M_r^n(X)$. Let $T := m$, define $f_t := b_i$, $i - 1 < t \le i$, $i = 1, \dots, m$ and $f_0 = 0$. Clearly,

$$\int_0^T f_t \tilde{X}_t \, \mathrm{d}t = \sum_{i=1}^m b_i X_{n+i}$$

An application of Theorems B.2 and B.3 to $\tilde{X}$ yield the result. □

Acknowledgements

All the authors were supported by The Alan Turing Institute, London under the EPSRC grant EP/N510129/1. N.H.C. and M.R. also enjoyed the support of the NKFIH (National Research, Development and Innovation Office, Hungary) grant KH 126505 and the "Lendület" grant LP 2015-6 of the Hungarian Academy of Sciences. Y.Z. was supported by The Maxwell Institute Graduate School in Analysis and its Applications, a Centre for Doctoral Training funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016508/01), the Scottish Funding Council, Heriot-Watt University and the University of Edinburgh.

The article was prepared within the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project -5-100-.

References

[1] Aliprantis, C.D. and Border, K.C. (2006). *Infinite Dimensional Analysis*: *A Hitchhiker's Guide*, 3rd ed. Berlin: Springer. MR2378491

[2] Ambrosio, L., Savaré, G. and Zambotti, L. (2009). Existence and stability for Fokker–Planck equations with log-concave reference measure. *Probab. Theory Related Fields* **145** 517–564. MR2529438 https://doi.org/10.1007/s00440-008-0177-3
[3] Chatterji, N., Flammarion, N., Ma, Y., Bartlett, P. and Jordan, M. (2018). On the theory of variance reduction for stochastic gradient Monte Carlo. Preprint, arXiv:1802.05431.
[4] Chau, H.N., Kumar, C., Rásonyi, M. and Sabanis, S. (2019). On fixed gain recursive estimators with discontinuity in the parameters. *ESAIM Probab. Stat.* **23** 217–244. MR3945579 https://doi.org/10.1051/ps/2018019
[5] Chau, H.N., Moulines, É., Rásonyi, M., Sabanis, S. and Zhang, Y. (2019). On stochastic gradient Langevin dynamics with dependent data streams: The fully non-convex case. Preprint, arXiv:1905.13142.
[6] 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.
[7] 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
[8] 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
[9] 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
[10] 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
[11] Gerencsér, B. and Rásonyi, M. (2019). On the ergodic properties of certain Markov chains in random environments. arXiv:1807.03568v3.
[12] Gerencsér, L. (1989). On a class of mixing processes. *Stoch. Stoch. Rep.* **26** 165–191. MR1018543 https://doi.org/10.1080/17442508908833555
[13] Gerencsér, L. (1992). AR(∞) estimation and nonparametric stochastic complexity. *IEEE Trans. Inf. Theory* **38** 1768–1778. MR1187818 https://doi.org/10.1109/18.165449
[14] Gerencsér, L. (1993). Strong approximation of the recursive prediction error estimator of the parameters of an ARMA process. *Systems Control Lett.* **21** 347–351. MR1241415 https://doi.org/10.1016/0167-6911(93)90078-K
[15] Gerencsér, L. (1994). On Rissanen's predictive stochastic complexity for stationary ARMA processes. *J. Statist. Plann. Inference* **41** 303–325. MR1309616 https://doi.org/10.1016/0378-3758(94)90026-4
[16] Gerencsér, L. (2006). A representation theorem for the error of recursive estimators. *SIAM J. Control Optim.* **44** 2123–2188. MR2248178 https://doi.org/10.1137/S0363012991217421
[17] Karatzas, I. and Shreve, S.E. (1991). *Brownian Motion and Stochastic Calculus*, 2nd ed. *Graduate Texts in Mathematics* **113**. New York: Springer. MR1121940 https://doi.org/10.1007/978-1-4612-0949-2
[18] Laruelle, S. (2011). Analyse d'Algorithmes Stochastiques Appliqués à la Finance. PhD Thesis, Université Paris VI.
[19] Majka, M.B., Mijatović, A. and Szpruch, L. (2018). Non-asymptotic bounds for sampling algorithms without log-concavity. arXiv:1808.07105.
[20] Mattingly, J.C., Stuart, A.M. and Higham, D.J. (2002). Ergodicity for SDEs and approximations: Locally Lipschitz vector fields and degenerate noise. *Stochastic Process. Appl.* **101** 185–232. MR1931266 https://doi.org/10.1016/S0304-4149(02)00150-3
[21] Meyn, S.P. and Tweedie, R.L. (1993). Stability of Markovian processes. III. Foster–Lyapunov criteria for continuous-time processes. *Adv. in Appl. Probab.* **25** 518–548. MR1234295 https://doi.org/10.2307/1427522
[22] Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization*: *A Basic Course*. *Applied Optimization* **87**. Boston, MA: Kluwer Academic. MR2142598 https://doi.org/10.1007/978-1-4419-8853-9
[23] Neveu, J. (1975). *Discrete-Parameter Martingales*, Revised ed. *North-Holland Mathematical Library* **10**. Amsterdam-Oxford: North-Holland; New York: American Elsevier Publishing Co., Inc. Translated from the French by T.P. Speed. MR0402915
[24] Raginsky, M., Rakhlin, A. and Telgarsky, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. *Proc. Mach. Learn. Res.* **65** 1674–1703.

[25] Rásonyi, M. (2010). On the statistical analysis of quantized Gaussian AR(1) processes. *Internat. J. Adapt. Control Signal Process.* **24** 490–507. MR2666031 https://doi.org/10.1002/acs.1145
[26] Robbins, H. (1955). A remark on Stirling's formula. *Amer. Math. Monthly* **62** 26–29. MR0069328 https://doi.org/10.2307/2308012
[27] 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
[28] Teh, Y.W., Thiery, A.H. and Vollmer, S.J. (2016). Consistency and fluctuations for stochastic gradient Langevin dynamics. *J. Mach. Learn. Res.* **17** Paper No. 7, 33. MR3482927
[29] Villani, C. (2009). *Optimal Transport. Old an New*. Springer.
[30] Welling, M. and Teh, Y.W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the* 28*th International Conference on Machine Learning* 681–688.
[31] Xu, P., Chen, J., Zhou, D. and Gu, Q. (2018). Global convergence of Langevin dynamics-based algorithms for nonconvex optimization. arXiv:1707.06618.
[32] Zhuang, C. (2008). Stochastic approximation methods and applications in financial optimization problems. PhD Thesis, University of Georgia, Athens, Georgia.

Received March 2019 and revised September 2019