On stochastic gradient Langevin dynamics with dependent data streams

On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case *

N. H. Chau† É. Moulines‡ M. Rásonyi§ S. Sabanis¶ Y. Zhang||

February 3, 2021

Abstract

We consider the problem of sampling from a target distribution, which is *not necessarily logconcave*, in the context of empirical risk minimization and stochastic optimization as presented in [33]. Non-asymptotic analysis results are established in the L^1-Wasserstein distance for the behaviour of Stochastic Gradient Langevin Dynamics (SGLD) algorithms. We allow the estimation of gradients to be performed even in the presence of *dependent* data streams. Our convergence estimates are sharper and *uniform* in the number of iterations, in contrast to those in previous studies.

Keywords: stochastic gradient, Langevin dynamics, convergence guarantees, non-convex optimization, contraction estimates for diffusions

MSC2020 classification: 65C05, 62L10, 93E35

1 Introduction

In this paper, the problem of approximate sampling from a target distribution

$$\pi_\beta(\theta) \propto \exp(-\beta U(\theta))\mathrm{d}\theta \tag{1}$$

is investigated, where $\theta \in \mathbb{R}^d$, $\beta > 0$, and the function $U : \mathbb{R}^d \to \mathbb{R}_+$ is differentiable, ∇U is Lipshitz-continuous, and U satisfies a certain dissipativity condition. If U has a unique minimizer θ^* then sampling from (1) with a large β amounts to finding θ^*.

It is well-known that (1) is the stationary law of the Langevin stochastic differential equation

$$\mathrm{d}L_t = -\nabla U(L_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t, \tag{2}$$

where B is a the standard Brownian motion in $\mathbb{R}^d$ and $\beta > 0$ is the so-called inverse temperature parameter. Euler discretizations of (2) lead to the extensively studied unadjusted Langevin

*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. We thank the Alan Turing Institute, London, UK; the Rényi Institute, Budapest, Hungary and the École Polytechnique, Palaiseau, France for hosting research meetings of the authors.

†Osaka University, Japan.

‡Centre de Mathématiques Appliquées, UMR 7641, Ecole Polytechnique, France

§Alfréd Rényi Institute of Mathematics, 1053 Budapest, Reáltanoda utca 13–15, Hungary
E-mail: rasonyi.miklos@renyi.hu

¶School of Mathematics, The University of Edinburgh and The Alan Turing Institute, UK.

||School of Mathematics, The University of Edinburgh, UK.

algorithm. When only estimates for the gradient ∇U are available, we arrive at the Stochastic Gradient Langevin Dynamics (SGLD) algorithm ((4) below), introduced in [38], which is the focus of our interests in the present article.

Imagine that we wish to tune the parameter θ of some software optimally so as to minimize $U(\theta)$ which is the expectation of a given cost function depending on θ and on an observed random data sequence whose law is unknown (and might slowly change over time). In such a situation our optimization must be data-driven and one may use e.g. fixed gain stochastic gradient algorithms, see [4]. In a nonconvex setting, however, there can be several local minima. By injecting extra noise, SGLD is a powerful tool for solving such problems, see Section 4 below for a more thorough discussion.

For an i.i.d. data sequence, the remarkable study [33] provided theoretical guarantees in the form of non-asymptotic convergence estimates for SGLD in the quadratic Wasserstein distance. The purpose of the present paper is to significantly sharpen these estimates by providing optimal rates in terms of the stepsize, using another metric, for the first time in the literature. We refer to [33] for further details about this method of optimization in the big data context. We stress, however, the applicability of SGLD also in the context of online parameter optimization where dependent data is commonly encountered.

Non-asymptotic convergence rates of Langevin dynamics based algorithms for approximate sampling of log-concave distributions have been intensively studied in recent years, starting with [9, 11]. This was followed by [12, 16, 17, 6, 3] amongst others.

Relaxing log-concavity is a more challenging problem. In [28], the log-concavity assumption is replaced by a "monotonicity at infinity" condition, convergence rates are obtained in L^1- and L^2-Wasserstein distances. In a similar setting, [7] analyzes sampling errors in the L^1-Wasserstein distance for both overdamped and underdamped Langevin MCMC.

Our starting point is [33], where a dissipativity condition is assumed and convergence rates are obtained in the L^2-Wasserstein distance. Moreover, a clear and strong link between sampling via SGLD algorithms and non-convex optimization is highlighted. One can further consult [39, 10] and references therein.

In the present paper, we impose the same dissipativity condition as in [33]. Using the L^1-Wasserstein metric, we obtain sharper estimates and allow for possibly dependent data sequences. The key new idea is comparing the SGLD algorithm to a suitable auxiliary continuous time processes inspired by (2) and then relying on contraction results developed in [18] for (2).

Notations and conventions. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We denote by $\mathbb{E}[X]$ the expectation of a random variable X. For $1 \leq p < \infty$, L^p is used to denote the usual space of p-integrable real-valued random variables. Fix an integer $d \geq 1$. For an $\mathbb{R}^d$-valued random variable X, its law on $\mathcal{B}(\mathbb{R}^d)$ (the Borel sigma-algebra of $\mathbb{R}^d$) is denoted by $\mathcal{L}(X)$. Scalar product is denoted by $\langle \cdot, \cdot \rangle$, with $|\cdot|$ standing for the corresponding norm (where the dimension of the space may vary depending on the context). For $r \in \mathbb{R}_+$, denote by B_r the closed ball centered at 0 with radius r.

For any integer $q \geq 1$, let $\mathcal{P}(\mathbb{R}^q)$ denote the set of probability measures on $\mathcal{B}(\mathbb{R}^q)$. For $\mu \in \mathcal{P}(\mathbb{R}^d)$ and for a non-negative measurable $f : \mathbb{R}^d \to \mathbb{R}$, we denote $\mu(f) := \int_{\mathbb{R}^d} f(\theta)\mu(\mathrm{d}\theta)$.

For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, let $\mathcal{C}(\mu, \nu)$ denote the set of probability measures ζ on $\mathcal{B}(\mathbb{R}^{2d})$ such that its respective marginals are μ, ν. Define, for $p \geq 1$,

$$W_p(\mu, \nu) := \left(\inf_{\zeta \in \mathcal{C}(\mu,\nu)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\theta - \theta'|^p \zeta(\mathrm{d}\theta \mathrm{d}\theta') \right)^{1/p}, \tag{3}$$

which is the L^p-Wasserstein distance associated to the Euclidean distance. We consider below only the cases $p = 1, 2$.

2 Main results

Fix an $\mathbb{R}^d$-valued random variable θ_0, representing the initial value of the procedure we consider. Let $(\mathcal{G}_n)_{n\in\mathbb{N}}$ be a given filtration representing the flow of past information. The notation $\mathcal{G}_\infty$ is self-explanatory. Let $(X_n)_{n\in\mathbb{N}}$ be a $(\mathcal{G}_n)$-adapted process. Let furthermore $(\mathcal{G}_n^+)_{n\in\mathbb{N}}$ be a decreasing sequence of σ-fields which represent the future information at the respective time instants. We assume in the sequel that for each $n \in \mathbb{N}$, the σ-fields $\mathcal{G}_n$ and $\mathcal{G}_n^+$ are independent.

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

$$\theta_0^\lambda := \theta_0, \quad \theta_{n+1}^\lambda := \theta_n^\lambda - \lambda H(\theta_n^\lambda, X_n) + \{2\lambda\beta^{-1}\}^{1/2}\,\xi_{n+1},\ n \in \mathbb{N}, \tag{4}$$

where $H : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ is a measurable function and $(\xi_n)_{n\in\mathbb{N}}$ is an independent sequence of standard d-dimensional Gaussian random variables.

We interpret $(X_n)_{n\in\mathbb{N}}$ as a stream of data and $(\xi_n)_{n\in\mathbb{N}}$ as an artificially generated noise sequence. We assume throughout the paper that θ_0, $\mathcal{G}_\infty$ and $(\xi_n)_{n\in\mathbb{N}}$ are independent.

Let $U : \mathbb{R}^d \to \mathbb{R}_+$ be continuously differentiable with gradient $h := \nabla U$. Let us define the probability

$$\pi_\beta(A) := \frac{\int_A \mathrm{e}^{-\beta U(\theta)}\,\mathrm{d}\theta}{\int_{\mathbb{R}^d} \mathrm{e}^{-\beta U(\theta)}\,\mathrm{d}\theta},\ A \in \mathcal{B}(\mathbb{R}^d).$$

It is implicitly assumed that $\int_{\mathbb{R}^d} \mathrm{e}^{-\beta U(\theta)}\,\mathrm{d}\theta < \infty$ and this is indeed the case under **H** 4 below, as easily seen. Our objective is to (approximately) sample from the distribution π_β using the scheme (4).

We now present our assumptions. First, the moments of the initial condition need to be controlled.

H1. $|\theta_0| \in \bigcap_{p\geq 1} L^p$.

Next, we require joint Lipschitz-continuity of every coordinate function H^i, $i = 1, \ldots, d$.

H2. *There exist positive constants K_1^i, K_2^i, $i = 1, \ldots, d$ such that for all $\theta, \theta' \in \mathbb{R}^d$ and $x, x' \in \mathbb{R}^m$,*

$$|H^i(\theta, x) - H^i(\theta', x')| \leq K_1^i|\theta - \theta'| + K_2^i|x - x'|.$$

We set

$$H^* := |H(0,0)|, \quad K_1 := \sum_{i=1}^d K_1^i, \quad K_2 := \sum_{i=1}^d K_2^i \tag{5}$$

and notice that, clearly,

$$|H(\theta, x) - H(\theta', x')| \leq K_1|\theta - \theta'| + K_2|x - x'|. \tag{6}$$

Remark 2.1. The reader may wonder why we did not assume just (6) directly for some K_1, K_2. The reason is that our estimates in the proof of Lemma 3.16 below lead to constants depending on K_1, K_2 as defined by the sums of the respective Lipschitz-constants for the coordinate mappings.

The data sequence $(X_n)_{n\in\mathbb{N}}$ need not be i.i.d., we require only a mixing property, defined in Section 3.1 below.

H 3. *Let $\mathcal{G}_n$, $n \in \mathbb{N}$ be a given filtration with $\mathcal{G}_0 = \{\emptyset, \Omega\}$. Let $\mathcal{G}_n^+$, $n \in \mathbb{N}$ be a decreasing family of sigma-algebras such that $\mathcal{G}_n$ is independent of $\mathcal{G}_n^+$ for all $n \in \mathbb{N}$. The process $(X_n)_{n\in\mathbb{N}}$ is conditionally L-mixing with respect to $(\mathcal{G}_n, \mathcal{G}_n^+)_{n\in\mathbb{N}}$. It satisfies for each $\theta \in \mathbb{R}^d$ and $n \geq 1$,*

$$\mathbb{E}[H(\theta, X_n)] = h(\theta)\,. \tag{7}$$

If the process $(X_n)_{n\geq 1}$ happens to be strictly stationary then (7) clearly holds. Finally, we present a dissipativity condition on H.

H4. *There exist a, $b > 0$ such that, for all $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$,*

$$\langle H(\theta, x), \theta \rangle \geq a|\theta|^2 - b. \tag{8}$$

When $X_n = c$ for all $n \in \mathbb{N}$ for some $c \in \mathbb{R}^m$ (i.e. when $H(\theta, X_{n+1})$ is replaced by $h(\theta)$ in (4)) then we arrive at the well-known unadjusted Langevin algorithm whose convergence properties have been amply analyzed, see e.g. [11, 17, 16, 7, 28] and the references therein. The case of i.i.d. $(X_n)_{n\in\mathbb{N}}$ has also been investigated in great detail, see e.g. [33, 39, 28].

In the present article, better estimates are obtained for the distance between $\mathcal{L}(\theta_n^\lambda)$ and π_β than those of [33] and [39]. Such rates have already been obtained in [2] for strongly convex U and in [28] for U that satisfies a monotonicity condition outside a compact set. Here we make no convexity assumptions at all. This comes at the price of using the metric W_1 defined in (3) below while [33, 39, 28, 2] use Wasserstein distances with respect to the standard Euclidean metric, see (3) below.

Another novelty of our paper is that, just like in [2], we allow the data sample $(X_n)_{n\in\mathbb{N}}$ to be dependent. As observed data have no reason to be i.i.d., we believe that such a result is fundamental to assure the robustness of online optimization procedures based on the SGLD (4).

Remark 2.2. In this work, the constants appearing are often denoted by C_j for some natural number $j \in \mathbb{N}$. Without further mention, these constants depend on K_1, K_2, a, b, H^*, β, d, and from the process $(X_n)_{n\in\mathbb{N}}$ (such as its moments). Unless otherwise stated, they do not depend on anything else. In case of further dependencies (e.g. in Lemma 3.5 dependence on the order of the moment p appears), we indicate these in parentheses, e.g. $C_6(p)$.

Our main contribution is summarized in the following result. Set

$$\lambda_{\max} = \min\{a/2K_1^2, 1/a\}\,, \tag{9}$$

where K_1 and a are defined in **H**2 and **H**4, respectively.

Theorem 2.3. *Assume **H**1, **H**2, **H**3 and **H**4. Then there are positive constants C_0, C_1, C_2 such that, for every $0 < \lambda \leq \lambda_{\max}$, $\beta > 0$ and $n \in \mathbb{N}$,*

$$W_1(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \leq C_1 \mathrm{e}^{-C_0\lambda n}\mathbb{E}[|\theta_0|^4 + 1] + C_2\sqrt{\lambda}\,, \tag{10}$$

where W_1 is defined in (3)*.*

Remark 2.4. Our assumptions can be somewhat weakened, as seen from a careful reading of the proofs. Indeed, the above theorem remains valid if we assume (instead of conditional L-mixing) only that X_n, $n \in \mathbb{N}$ are L^4-bounded and, for some $\epsilon > 0$, the sequences $M_{2+\epsilon}^n(X), \Gamma_{2+\epsilon}^n(X), n \in \mathbb{N}$ are bounded in L^2 for some $\epsilon > 0$. Furthermore, Assumption 1 can be weakened to $|\theta_0| \in L^6$.

[2, Example 3.4] suggests that the best rate we can hope to get in (10) is $\sqrt{\lambda}$, even in the convex case. The above theorem achieves this rate. We remark that, although the statement of Theorem 2.3 concerns the discrete-time recursive scheme (4), its proof is carried out entirely in a continuous-time setting, in Section 3. It relies on techniques from [2] and [18]. The principal new ideas are the introduction of the auxiliary process $\tilde{Y}_t^\lambda(\mathbf{x})$, $t \in \mathbb{R}_+$ (see (25) below) and reliance on the contractivity of the continuous system dynamics in a suitable semimetric (see Proposition 3.14 below).

2.1 Related work and our contributions

In [33], a non-convex empirical risk minimization problem is considered. The excess risk is decomposed into a sampling error resulting from the application of Stochastic Gradient Langevin Dynamic (SGLD), a generalization error and a suboptimality error. Our aim is to improve the

sampling error in the non-convex setting and provide sharper convergence estimates under more relaxed conditions. To this end, we focus on the comparison of our results with [33, Proposition 3.3].

[33, Assumption (A.5)] is (much) stronger than **H** 1 above. **H** 4 is identical to [33, Assumption (A3)]. [33, Assumption (A.2)] corresponds to Lipschitz-continuity of H in its first variable with a Lipschitz-constant independent from its second variable and $(A.1)$ there means that $H(0,\cdot)$, $u(0,\cdot)$ are bounded where $U(\theta) = \mathbb{E}[u(\theta, X_0)]$ and $H(\cdot,\cdot) = \partial_\theta u(\cdot,\cdot)$. Hence **H** 2 here is neither stronger nor weaker than $(A.2)$ of [33], they are incomparable conditions. In any case, **H** 2 does not seem to be restrictive. Condition $(A.4)$ in [33] is implied by **H**2 and **H**3.

We obtain stronger rates (which we believe to be optimal) than those of [33]. More precisely, we obtain a rate $\lambda^{1/2}$ in (10) for the W_1 distance while [33] only obtains $\lambda^{5/4}n$ (which depends on n) but in the W_2 distance. Furthermore, [33] is applicable only if $(X_n)_{n\in\mathbb{N}}$ is i.i.d. while **H** 3 suffices for the derivation of our results.

Now let us turn to [28]. That paper assumes a strengthening of our dissipativity assumption: they require **H**2 and that there exist $b, a > 0$ such that, for each $\theta, \theta' \in \mathbb{R}^d$ satisfying $|\theta - \theta'| > b$,

$$\langle h(\theta) - h(\theta'), \theta - \theta' \rangle \geq a|\theta - \theta'|^2,\ x \in \mathbb{R}^m. \tag{11}$$

Note, however, that this is stipulated only for h in [28] while we need our dissipativity assumption for $H(\cdot, x)$, for all x, as we allow dependent data streams. Furthermore, Assumption 1.3 in [28] requires that the variance of $H(\theta, X_0)$ is controlled by a power of the step size λ while we do not need such an assumption. The second conclusion of their Theorem 1.4 (with $\alpha = 1$, using their notation α) is the same as that of our Theorem 2.3.

3 Proofs

3.1 Conditional L-mixing

A key mixing assumption is required about X_n, $n \in \mathbb{N}$. In this subsection we present some related concepts and results. The material presented here is from [2].

L-mixing processes and random fields were introduced in [20]. In [4], the closely related concept of *conditional* L-mixing was created.

We assume that the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is equipped with a discrete-time filtration $(\mathcal{R}_n)_{n\in\mathbb{N}}$ as well as with a decreasing sequence of sigma-fields $(\mathcal{R}_n^+)_{n\in\mathbb{N}}$ such that the σ-fields $\mathcal{R}_n$ and $\mathcal{R}_n^+$ are independent for all $n \in \mathbb{N}$. A random process $(U_n)_{n\in\mathbb{N}}$ is called *L^r-bounded* for some $r \geq 1$ if

$$\sup_{n\in\mathbb{N}} \mathbb{E}^{1/r}[|U_n|^r] < \infty.$$

Define, for each $n \in \mathbb{N}$, $i = 1, \ldots, d$,

$$\begin{aligned}
\tilde{M}_r^n(U, i) &:= \sup_{m\in\mathbb{N}} \mathbb{E}^{1/r} \left[|U_{n+m}^i|^r \,\middle|\, \mathcal{F}_n \right], \\
\tilde{\gamma}_r^n(\tau, U, i) &:= \sup_{m\geq\tau} \mathbb{E}^{1/r} \left[|U_{n+m}^i - \mathbb{E}\left[U_{n+m}^i \,\middle|\, \mathcal{F}_{n+m-\tau}^+ \vee \mathcal{F}_n \right]|^r \,\middle|\, \mathcal{F}_n \right],\ \tau \geq 0,
\end{aligned}$$

where U_{n+m}^i refers to the ith coordinate of U_{n+m} in the above expressions. Finally, set

$$\tilde{\Gamma}_r^n(U, i) := \sum_{\tau=0}^{\infty} \tilde{\gamma}_r^n(\tau, U, i),\ M_r^n(U) := \sum_{i=1}^{k} \tilde{M}_r^n(U, i), \text{ and } \Gamma_r^n(U) := \sum_{i=1}^{k} \tilde{\Gamma}_r^n(U, i). \tag{12}$$

Definition 3.1 (Conditional L-mixing)**.** *We say that the random process $(U_n)_{n\in\mathbb{N}}$ is* conditionally L-mixing *with respect to $(\mathcal{R}_n, \mathcal{R}_n^+)_{n\in\mathbb{N}}$ if $(U_n)_{n\in\mathbb{N}}$ is adapted to $(\mathcal{R}_n)_{n\in\mathbb{N}}$ for all $\theta \in \Theta$; for all $r \geq 1$, it is L^r-bounded; and the sequences $(M_r^n(U))_{n\in\mathbb{N}}$, $(\Gamma_r^n(U))_{n\in\mathbb{N}}$ are also L^r-bounded for all $r \geq 1$.*

Conditionally L-mixing encompasses a broad class of stochastic models (i.i.d. with finite moments of all orders, linear processes, functionals of Markov processes, etc.), see in [2, Example 2.1].

It is convenient to extend the L-mixing property to the continuous-time setting. 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 $t \in \mathbb{R}_+$. Consider an $\mathbb{R}^d$-valued continuous-time stochastic process $(W_t)_{t\in\mathbb{R}_+}$ which is progressively measurable (i.e. $W : [0,t] \times \Omega \to \mathbb{R}^d$ is $\mathcal{B}([0,t]) \otimes \mathcal{R}_t$-measurable for all $t \in \mathbb{R}_+$). From now on we assume that $W_t \in L^1$, $t \in \mathbb{R}_+$. We define the quantities[1]

$$\tilde{M}_r^i(\mathbf{W}) := \operatorname{ess.sup}_{t\in\mathbb{R}_+} \mathbb{E}^{1/r}\left[|W_t^i|^r \,\middle|\, \mathcal{R}_0\right],$$
$$\tilde{\gamma}_r^i(\tau,\mathbf{W}) := \operatorname{ess.sup}_{t\geq\tau} \mathbb{E}^{1/r}\left[|W_t^i - \mathbb{E}\left[W_t^i \,\middle|\, \mathcal{R}_{t-\tau}^+ \vee \mathcal{R}_0\right]|^r \,\middle|\, \mathcal{R}_0\right],\ \tau \in \mathbb{R}_+,$$

and set

$$M_r(\mathbf{W}) := \sum_{i=1}^d \tilde{M}_r^i(\mathbf{W}),\ \tilde{\Gamma}_r^i(\mathbf{W}) := \sum_{\tau=0}^\infty \tilde{\gamma}_r^i(\tau,\mathbf{W}),\ \text{and}\ \Gamma_r(\mathbf{W}) := \sum_{i=1}^d \tilde{\Gamma}_r^i(\mathbf{W})$$

where W_t^i refers to the ith coordinate of W_t. We recall [2, Theorem B.5] which is key to further developments.

Theorem 3.2. *Let $(W_t)_{t\in\mathbb{R}_+}$ be L^r-bounded for some $r > 2$ and let $M_r(\mathbf{W}) + \Gamma_r(\mathbf{W}) < \infty$ a.s. Assume $\mathbb{E}[W_t \,|\, \mathcal{R}_0] = 0$ 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*

$$\mathbb{E}^{1/r}\left[\sup_{s\in[0,T]} \left|\int_0^s f_t W_t \,\mathrm{d}t\right|^r \,\middle|\, \mathcal{R}_0\right] \leq C'(r)\left(\int_0^T f_t^2 \,\mathrm{d}t\right)^{1/2} [M_r(\mathbf{W}) + \Gamma_r(\mathbf{W})], \text{a.s.} \tag{13}$$

We can actually take

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

Estimates for $M_r(\mathbf{W}), \Gamma_r(\mathbf{W})$ imply similar estimates for functionals of $\mathbf{W}$.

Lemma 3.3. *Assume **H**2. Then, for each $i \in \mathbb{N}$ and $\theta \in \mathrm{B}_i$, $(H(\theta, W_t))_{t\in\mathbb{R}_+}$ satisfies*

$$M_r(H(\theta,\mathbf{W})) \leq K_1 i + K_2 M_r(\mathbf{W}) + H^*, \tag{14}$$

where H^* is defined in (5) *and*

$$\Gamma_r(H(\theta,\mathbf{W})) \leq 2K_2\Gamma_r(\mathbf{W}). \tag{15}$$

Proof. Identical to the proofs in [2, Lemma 6.4 and Example 2.4], using Lipschitz-continuity of the coordinate functions H^i with the respective constants K_1^i, K_2^i. □

One of the main advantages of the mixing concepts we use is that one can plug in $\mathcal{R}_0$-measurable random variables into θ and still preserve the mixing properties.

Lemma 3.4. *Assume **H**2 and set $i \in \mathbb{N}$. let $(Z_s)_{s\geq 0}$ be a family of B_i-valued random variables satisfying $Z : \mathbb{R}_+ \times \Omega \to \mathbb{R}^d$ is $\mathcal{B}(\mathbb{R}_+) \otimes \mathcal{R}_0$-measurable. Define the process $Y_t = H(Z_t, W_t)$ for $t \in \mathbb{R}_+$. Then*

$$M_p(\mathbf{Y}) \leq K_1 i + K_2 M_r(\mathbf{W}) + H^*,$$

and

$$\Gamma_r(\mathbf{Y}) \leq 2K_2\Gamma_r(\mathbf{W}).$$

Proof. The proof is identical to that of [4, Lemma A.3], noting the Lipschitz continuity. □

[1] 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 it dominates almost surely all the Z_i and it is a.s. dominated by any other random variable with this property. For an existence proof, see e.g. [32, Proposition VI.1.1].

3.2 Further notations and introduction of auxiliary processses

Note that **H**2 implies

$$|h(\theta) - h(\theta')| \leq K_1|\theta - \theta'|,\ \theta, \theta' \in \mathbb{R}^d, \tag{16}$$

and Assumption **H**4 implies

$$\langle h(\theta), \theta \rangle \geq a|\theta|^2 - b,\ \theta \in \mathbb{R}^d. \tag{17}$$

Also, **H**2 implies

$$|H(\theta, x)| \leq K_1|\theta| + K_2|x| + H^*, \tag{18}$$

with the constant H^* defined in (5). Define, for each $p \geq 2$,

$$V_p(\theta) = \mathrm{v}_p(|\theta|)\,, \quad \text{where for } u \in \mathbb{R}_+,\ \mathrm{v}_p(u) := (1 + u^2)^{p/2}. \tag{19}$$

We use V_p as a Lyapunov function which allows to obtain uniform bounds for the moments of various processes. Notice that each V_p is twice continuously differentiable and

$$\lim_{|\theta| \to \infty} \frac{\nabla V_p(\theta)}{V_p(\theta)} = 0. \tag{20}$$

Let $\mathcal{P}_{V_p}(\mathbb{R}^d)$ denote the subset of $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfying $\int_{\mathbb{R}^d} V_p(\theta)\,\mu(\mathrm{d}\theta) < \infty$. The following functional is pivotal in our arguments as it is used to measure the distance between probability measures. We define, for any $p \geq 1$ and $\mu, \nu \in \mathcal{P}_{V_p}(\mathbb{R}^d)$,

$$w_{1,p}(\mu, \nu) := \inf_{\zeta \in \mathcal{C}(\mu,\nu)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \{1 \wedge |\theta - \theta'|\}\{1 + V_p(\theta) + V_p(\theta')\}\zeta(\mathrm{d}\theta \mathrm{d}\theta'), \tag{21}$$

Though $w_{1,p}$ is not a metric, it satisfies

$$W_1(\mu, \nu) \leq w_{1,p}(\mu, \nu), \tag{22}$$

as easily seen, where W_1 is defined in (3). In the sequel we solely consider the case $p = 2$, that is, $w_{1,2}$.

Our estimations are carried out below in a *continuous-time* setting, so we define and discuss a number of auxiliary continuous-time processes below. First, consider $(L_t)_{t \in \mathbb{R}_+}$ defined by the stochastic differential equation (SDE)

$$\mathrm{d}L_t = -h(L_t)\,\mathrm{d}t + \{2\beta^{-1}\}^{1/2}\,\mathrm{d}B_t, \quad L_0 := \theta_0, \tag{23}$$

where $(B_t)_{t \geq 0}$ is standard Brownian motion on $(\Omega, \mathcal{F}, \mathbb{P})$, independent of $\mathcal{G}_\infty \vee \sigma(\theta_0)$ with its natural filtration denoted by $(\mathcal{F}_t)_{t \in \mathbb{R}_+}$ henceforth. The meaning of $\mathcal{F}_\infty$ is clear.

Equation (23) has a unique solution on $\mathbb{R}_+$ adapted to $(\mathcal{F}_t)_{t \in \mathbb{R}_+}$ since h is Lipschitz-continuous by (16). We proceed by defining, for each $\lambda > 0$ convenient time-changed versions of L_t, $t \in \mathbb{R}_+$:

$$L^\lambda_t := L_{\lambda t},\ t \in \mathbb{R}_+.$$

Notice that $\tilde{B}^\lambda_t := B_{\lambda t}/\sqrt{\lambda}$, $t \in \mathbb{R}_+$ is also a Brownian motion and

$$\mathrm{d}L^\lambda_t = -\lambda h(L^\lambda_t)\,\mathrm{d}t + \{2\lambda\beta^{-1}\}^{1/2}\,\mathrm{d}\tilde{B}^\lambda_t,\ L^\lambda_0 = \theta_0. \tag{24}$$

Define $\mathcal{F}^\lambda_t := \mathcal{F}_{\lambda t}$, $\lambda \in \mathbb{R}_+$, $t \in \mathbb{R}_+$, the natural filtration of $(\tilde{B}^\lambda_t)_{t \geq 0}$.

Our recursion (4) is defined in terms of the data sequence X_n, $n \in \mathbb{N}$. However, it is more convenient to *freeze* the values of this sequence and to do the analysis initially with such framework. To this end, for each $\lambda > 0$ and $\mathbf{x} = (x_0, x_1, \ldots) \in (\mathbb{R}^m)^{\mathbb{N}}$, consider the process $(\tilde{Y}^\lambda_t(\mathbf{x}))_{t \in \mathbb{R}_+}$ defined as

$$\mathrm{d}\tilde{Y}^\lambda_t(\mathbf{x}) = -\lambda H(\tilde{Y}^\lambda_t(\mathbf{x}), x_{\lfloor t \rfloor})\,\mathrm{d}t + \{2\lambda\beta^{-1}\}^{1/2}\,\mathrm{d}\tilde{B}^\lambda_t, \tag{25}$$

with initial condition $\tilde{Y}_0^\lambda(\mathbf{x}) = \theta_0$. Due to **H**2, there is a unique solution to (25) which is adapted to $(\mathcal{F}_t^\lambda)_{t\in\mathbb{R}_+}$. This process provides a continuous-time "approximation" for our recursive procedures and plays an important role in the estimations below.

Moreover, for any given $s \geq 0$ and $t \geq s$, consider the following auxiliary process, which follows the same dynamics as (25) but its starting time and value are prescribed:

$$\mathrm{d}\tilde{Y}_{s,t}^\lambda(\mathbf{x},\theta) = -\lambda H(\tilde{Y}_{s,t}^\lambda(\mathbf{x},\theta), x_{\lfloor t\rfloor})\mathrm{d}t + \{2\lambda\beta^{-1}\}^{1/2}\,\mathrm{d}\tilde{B}_t^\lambda, \qquad \text{for } t > s, \tag{26}$$

with initial condition $\tilde{Y}_{s,s}^\lambda(\mathbf{x},\theta) = \theta \in \mathbb{R}^d$. Note that $\tilde{Y}_{s,t}^\lambda(\mathbf{x}, \tilde{Y}_s^\lambda(\mathbf{x})) = \tilde{Y}_t^\lambda(\mathbf{x})$ for all $t > s$ and for all $\mathbf{x} = (x_0, x_1, \dots) \in (\mathbb{R}^m)^{\mathbb{N}}$.

Let us now define the continuously interpolated Euler-Maruyama approximation of $(\tilde{Y}_t^\lambda(\mathbf{x}))_{t\in\mathbb{R}_+}$ via

$$\mathrm{d}Y_t^\lambda(\mathbf{x}) = -\lambda H(Y_{\lfloor t\rfloor}^\lambda(\mathbf{x}), x_{\lfloor t\rfloor})\,\mathrm{d}t + \{2\lambda\beta^{-1}\}^{1/2}\,\mathrm{d}\tilde{B}_t^\lambda, \tag{27}$$

with initial condition $Y_0^\lambda(\mathbf{x}) = \theta_0$. Notice at this point that (27) can be solved by a simple recursion.

Now we explain the relationship of the latter process to θ_n^λ, $n \in \mathbb{N}$, defined in (4). If one considers $(Y_t^\lambda(\mathbf{X}))_{t\in\mathbb{R}_+}$, where $\mathbf{X} = (X_0, X_1, \dots)$ is a random element in $(\mathbb{R}^m)^{\mathbb{N}}$, then for each integer $n \in \mathbb{N}$,

$$\mathcal{L}(Y_n^\lambda(\mathbf{X})) = \mathcal{L}(\theta_n^\lambda), \tag{28}$$

since $\tilde{B}_{n+1}^\lambda - \tilde{B}_n^\lambda$ has standard Gaussian law on $\mathbb{R}^d$, for all $n \in \mathbb{N}$.

3.3 Layout of the proof

In view of the observation (28), the main objective is to bound $W_1(\mathcal{L}(Y_t^\lambda(\mathbf{X})), \pi_\beta)$. This task can be decomposed as follows:

$$W_1(\mathcal{L}(Y_t^\lambda(\mathbf{X})), \pi_\beta) \leq W_1(\mathcal{L}(Y_t^\lambda(\mathbf{X})), \mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{X}))) + W_1(\mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{X})), \mathcal{L}(L_t^\lambda)) + W_1(\mathcal{L}(L_t^\lambda), \pi_\beta)\,, \tag{29}$$

where π_β is defined in (1). Here the last term is controlled below by standard arguments which entail that L_t^λ converges in law to π_β as $t \to \infty$. The drift condition (31) below (which follows from the dissipativity Assumption **H**4 and Lipschitzness of the mean field h, see (16)) ensure the applicability of classical results.

The second term is controlled uniformly in t by a quantity which is proportional to $\sqrt{\lambda}$. To this end, we follow the line of attack used in [2] which consists in estimating, on intervals of length $1/\lambda$, the L^2-distance between $\tilde{Y}_t^\lambda(\mathbf{X})$ and another process that coincides with it at the initial point of the interval but follows the averaged dynamics (24) (see (56) for a precise definition and Lemma 3.17 for details). Here we rely on a maximal inequality for functionals of a conditionally L-mixing process, given as Theorem 3.2 above. We put together estimates on separate intervals and thus obtain a bound on $W_1(\mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{X})), \mathcal{L}(L_t^\lambda))$ in Lemma 3.18, relying on novel results by [18], which give us a contraction rate for the diffusion L_t^λ, $t \geq 0$ in the semimetric $w_{1,2}$, see Proposition 3.14 and, in particular, (60).

Finally, the first term is controlled uniformly in t by a quantity which is also proportional to $\sqrt{\lambda}$, see Corollary 3.23. This is based on Kullback-Leibler distance estimates which go back to [9] but which are somewhat trickier as we need to employ measurable selection to pass from bounds for $W_1(\mathcal{L}(Y_t^\lambda(\mathbf{x})), \mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{x})))$ with fixed $\mathbf{x}$ to ones for $W_1(\mathcal{L}(Y_t^\lambda(\mathbf{X})), \mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{X})))$.

3.4 Moment estimates

Define the following notation for $\lambda > 0$, $\beta > 0$, $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}^m$,

$$\begin{aligned} \bar{\mathrm{L}}_\beta^\lambda V_p(\theta) &:= \lambda\beta^{-1}\Delta V_p(\theta) - \lambda\left\langle h(\theta), \nabla V_p(\theta)\right\rangle\,, \\ \mathrm{L}_{\beta,x}^\lambda V_p(\theta) &:= \lambda\beta^{-1}\Delta V_p(\theta) - \lambda\left\langle H(\theta,x), \nabla V_p(\theta)\right\rangle\,. \end{aligned} \tag{30}$$

Lemma 3.5. *Assume **H4**. For each $p \geq 2$, $\theta \in \mathbb{R}^d$, and $x \in \mathbb{R}^m$,*

$$\bar{\mathrm{L}}^1_\beta V_p(\theta) \leq -C_6(p)V_p(\theta) + C_7(p) \tag{31}$$

$$\mathrm{L}^1_{\beta,x} V_p(\theta) \leq -C_6(p)V_p(\theta) + C_7(p),\ \theta \in \mathbb{R}^d, \tag{32}$$

where $C_6(p) = ap/4$, $C_7(p) = (3/4)ap\mathrm{v}_p(\overline{M}(p))$ with

$$\overline{M}(p) = \sqrt{1/3 + 4b/(3a) + 4d/(3a\beta) + 4(p-2)/(3a\beta)}\,. \tag{33}$$

Proof. By direct calculation,

$$\bar{\mathrm{L}}^1_\beta V_p(\theta) = \beta^{-1}dpV_{p-2}(\theta) + \beta^{-1}p(p-2)(|\theta|^2+1)^{(p-4)/2}|\theta|^2 - pV_{p-2}(\theta)\left\langle h(\theta),\theta\right\rangle\,. \tag{34}$$

By **H**4, see also (17), the third term of (34) is dominated by

$$-pa|\theta|^2(|\theta|^2+1)^{(p-2)/2} + pb(|\theta|^2+1)^{(p-2)/2}. \tag{35}$$

Then, for $|\theta| > \overline{M}(p)$, one observes that $\bar{\mathrm{L}}^1_\beta V_p(\theta) \leq -(ap/4)V_p(\theta)$. As for $|\theta| \leq \overline{M}(p)$, one obtains $\bar{\mathrm{L}}^1_\beta V_p(\theta) \leq (3/4)ap\mathrm{v}_p(\overline{M}(p))$. Eq. (31) follows. The statement (32) follows in an identical way, noting that the constants which appear do not depend on $x \in \mathbb{R}^m$. □

Now, we proceed with the required moment estimates which play a crucial role in the derivation of the main result as given in Theorem 2.3.

Lemma 3.6. *Assume **H1**, **H2** and **H4**. Let $p \geq 2$ and $\tilde{\theta} \in L^{2p-2}$. For any $t > s \geq 0$,*

$$\sup_{\mathbf{x}\in(\mathbb{R}^m)^{\mathbb{N}}} \mathbb{E}[V_p(\tilde{Y}^\lambda_{s,t}(\mathbf{x},\tilde{\theta}))] \leq \mathrm{e}^{-\lambda C_6(p)(t-s)}\mathbb{E}[V_p(\tilde{\theta})] + 3\mathrm{v}_p(\overline{M}(p)) \tag{36}$$

where $\overline{M}(p)$ is defined in (33)*.*

Proof. We note that $2p-2 \geq p$ for $p \geq 2$, hence $\mathbb{E}[V_p(\tilde{\theta})] < \infty$. For any fixed sequence $\mathbf{x} \in (\mathbb{R}^m)^{\mathbb{N}}$ and $t > s \geq 0$, by Itô's formula, one obtains almost surely,

$$\mathrm{d}V_p(\tilde{Y}^\lambda_{s,t}(\mathbf{x},\tilde{\theta})) = \mathrm{L}^\lambda_{\beta,x_{\lfloor t\rfloor}} V_p(\tilde{Y}^\lambda_{s,t}(\mathbf{x},\tilde{\theta}))\mathrm{d}t + \{2\lambda\beta^{-1}\}^{1/2}\left\langle \nabla V_p(\tilde{Y}^\lambda_{s,t}(\mathbf{x},\tilde{\theta})), \mathrm{d}\tilde{B}^\lambda_t\right\rangle,$$

Since $\sup_{0\leq s\leq t}\mathbb{E}[|\nabla V_p(\tilde{Y}^\lambda_{s,t}(\mathbf{x},\tilde{\theta}))|^2] < \infty$ using $\tilde{\theta} \in L^{2p-2}$, the expectation of the stochastic integral vanishes and

$$\mathbb{E}[V_p(\tilde{Y}^\lambda_{s,t}(\mathbf{x},\tilde{\theta}))] = \mathbb{E}[V_p(\tilde{\theta})] + \int_s^t \mathbb{E}\left[\mathrm{L}^\lambda_{\beta,x_{\lfloor u\rfloor}} V_p(\tilde{Y}^\lambda_{s,u}(\mathbf{x},\tilde{\theta}))\right]\mathrm{d}u,$$

Differentiating both sides and using Lemma 3.5 yields that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[V_p(\tilde{Y}^\lambda_{s,t}(\mathbf{x},\tilde{\theta}))] = \mathbb{E}\left[\mathrm{L}^\lambda_{\beta,x_{\lfloor t\rfloor}} V_p(\tilde{Y}^\lambda_{s,t}(\mathbf{x},\tilde{\theta}))\right] \leq -\lambda C_6(p)\mathbb{E}[V_p(\tilde{Y}^\lambda_{s,t}(\mathbf{x},\tilde{\theta}))] + \lambda C_7(p). \tag{37}$$

Hence, by calculating the derivative of $\mathrm{e}^{\lambda C_6(p)(t-s)}\mathbb{E}[V_p(\tilde{Y}^\lambda_{s,t}(\mathbf{x},\tilde{\theta}))]$ and in view of the above relationship (37), one obtains (36). □

Corollary 3.7. *Assume **H1**, **H2** and **H4**. For any integer $p \geq 2$ and $t \in \mathbb{R}_+$,*

$$\sup_{\mathbf{x}\in(\mathbb{R}^m)^{\mathbb{N}}} \mathbb{E}[V_p(\tilde{Y}^\lambda_t(\mathbf{x}))] \leq \mathrm{e}^{-\lambda C_6(p)t}\mathbb{E}[V_p(\theta_0)] + 3\mathrm{v}_p(\overline{M}(p)), \tag{38}$$

where $\overline{M}(p)$ is defined in (33)*.*

Proof. By noting that $\tilde{Y}_t^\lambda(\mathbf{x}) = \tilde{Y}_{0,t}^\lambda(\mathbf{x}, \theta_0)$, one immediately recovers the desired result from Lemma 3.6. □

Corollary 3.8. *Assume **H**1, **H**2 and **H**4. For any integer $p \geq 2$ and $t \in \mathbb{R}_+$,*

$$\mathbb{E}[V_p(\tilde{Y}_t^\lambda(\mathbf{X}))] \leq \mathrm{e}^{-\lambda C_6(p)t}\mathbb{E}[V_p(\theta_0)] + 3\mathrm{v}_p(\overline{M}(p)), \tag{39}$$

where the constant $\overline{M}(p)$ is defined in (33)*.*

Proof. Due to the fact that the dissipativity condition **H** 4 is uniform in x, all estimates are independent of x and therefore the result follows immediately from Corollary 3.7. □

While the moment estimates for $\tilde{Y}^\lambda(\mathbf{x})$ have been rather straightforward, similar bounds for $Y_t^\lambda(\mathbf{x})$ require more substantial calculations, based again on dissipativity, see Assumption **H**4.

Lemma 3.9. *Assume **H**1, **H**2 and **H**4. For any $\lambda < \lambda_{\max}$, as given in* (9)*, $n \in \mathbb{N}$, $t \in (n, n+1]$, $p \in \mathbb{N}^*$, and any sequence $\mathbf{x} \in (\mathbb{R}^m)^{\mathbb{N}}$,*

$$\begin{aligned}\mathbb{E}[|Y_t^\lambda(\mathbf{x})|^{2p}] &\leq (1 - a\lambda(t-n))(1-a\lambda)^n\mathbb{E}|\theta_0|^{2p} \\ &\quad + \lambda a M(p,d)\left\{|x_n|^{2p} + (1 - a\lambda(t-n))\sum\nolimits_{j=1}^n (1-a\lambda)^{j-1}\,|x_{n-j}|^{2p}\right\} + \widehat{M}(p,d),\end{aligned} \tag{40}$$

where the constants $M(p,d)$ and $\widehat{M}(p,d)$ are given by

$$M(p,d) = (2\lambda_{\max} + 4/a)^{p-1}\left[1/a + d\tilde{M}^2(p)\right]c_0^p \tag{41}$$

and

$$\widehat{M}(p,d) \quad = \quad M(p,d)(c_2/c_0)^p \quad + \quad \tilde{M}^2(p)\,(\lambda_{\max} + 2/a)^{p-1}\left(d + (1/\beta)^{p-1}\,(2dp(2p-1))^p\right)$$

with

$$\tilde{M}(p) := 2^p\sqrt{p(2p-1)/(a\beta)}. \tag{42}$$

and c_0 and c_1 are defined by

$$c_0 = 8K_2^2\lambda_{\max},\; c_1 = a^{-1}(c_2 + 2d\beta^{-1}) \quad \textit{and} \quad c_2 = 2b + 8\lambda_{\max}(H^*)^2\,. \tag{43}$$

In particular,

$$\mathbb{E}|Y_{n+1}^\lambda(\mathbf{x})|^2 \leq (1-a\lambda)^{n+1}\mathbb{E}|\theta_0|^2 + \lambda c_0\sum_{j=0}^n(1-a\lambda)^j|x_{n-j}|^2 + c_1\,, \tag{44}$$

Proof. For any $n \in \mathbb{N}$ and $t \in (n, n+1]$, define $\Delta_{n,t}(\mathbf{x}) = Y_n^\lambda(\mathbf{x}) - \lambda H(Y_n^\lambda(\mathbf{x}), x_n)(t-n)$. It is easily seen that for $t \in (n, n+1]$

$$\mathbb{E}\left[\,|Y_t^\lambda(\mathbf{x})|^2 \,\middle|\, Y_n^\lambda(\mathbf{x})\right] = |\Delta_{n,t}(\mathbf{x})|^2 + (2\lambda/\beta)d(t-n).$$

Using **H**2 and **H**4, one obtains for all $\lambda \leq \lambda_{\max}$,

$$\begin{aligned}&|\Delta_{n,t}(\mathbf{x})|^2 = |Y_n^\lambda(\mathbf{x})|^2 - 2\lambda(t-n)\left\langle Y_n^\lambda(\mathbf{x}), H(Y_n^\lambda(\mathbf{x}), x_n)\right\rangle + \lambda^2|H(Y_n^\lambda(\mathbf{x}), x_n)(t-n)|^2 \\ &\leq (1 - 2a\lambda(t-n))|Y_n^\lambda(\mathbf{x})|^2 + 2b\lambda(t-n) + 2\lambda^2(t-n)^2\{K_1^2|Y_n^\lambda(\mathbf{x})|^2 + 4K_2^2|x_n|^2 + 4(H^*)^2\} \\ &\leq (1 - a\lambda(t-n))|Y_n^\lambda(\mathbf{x})|^2 + \lambda(t-n)(c_0|x_n|^2 + c_2).\end{aligned} \tag{45}$$

The desired result (44) follows from an easy induction. For higher moments, the calculation is somewhat more involved. To this end, one calculates, by setting $U_{n,t}^\lambda = \{2\lambda\beta^{-1}\}^{1/2}(\tilde{B}_t^\lambda - \tilde{B}_n^\lambda)$, for $t \in [n, n+1)$,

$$\mathbb{E}[|Y_t^\lambda(\mathbf{x})|^{2p}|Y_n^\lambda(\mathbf{x})] \leq |\Delta_{n,t}(\mathbf{x})|^{2p} + 2p\mathbb{E}\left[|\Delta_{n,t}(\mathbf{x})|^{2p-2}\left\langle \Delta_{n,t}(\mathbf{x}), U_{n,t}^\lambda\right\rangle \middle| Y_n^\lambda(\mathbf{x})\right] + \sum_{k=2}^{2p}\binom{2p}{k}\mathbb{E}\left[|\Delta_{n,t}(\mathbf{x})|^{2p-k}\left|U_{n,t}^\lambda\right|^k \middle| Y_n^\lambda(\mathbf{x})\right],$$

where the last inequality is due to Lemma A.3. The following inequality is used in the subsequent analysis

$$(r+s)^p \leq (1+\epsilon)^{p-1}r^p + (1+\epsilon^{-1})^{p-1}s^p, \tag{46}$$

where $p \geq 2$, $r,\ s \geq 0$ and $\epsilon > 0$. We continue as follows

$$\begin{aligned}
&\mathbb{E}[|Y_t^\lambda(\mathbf{x})|^{2p}|Y_n^\lambda(\mathbf{x})] \\
&\leq |\Delta_{n,t}(\mathbf{x})|^{2p} + \sum_{l=0}^{2(p-1)}\binom{2p}{l+2}\mathbb{E}\left[|\Delta_{n,t}(\mathbf{x})|^{2(p-1)-l}|U_{n,t}^\lambda|^l|U_{n,t}^\lambda|^2 \,\middle|\, Y_n^\lambda(\mathbf{x})\right] \\
&\leq |\Delta_{n,t}(\mathbf{x})|^{2p} + \binom{2p}{2}\sum_{l=0}^{2(p-1)}\binom{2(p-1)}{l}\mathbb{E}\left[|\Delta_{n,t}(\mathbf{x})|^{2(p-1)-l}|U_{n,t}^\lambda|^l|U_{n,t}^\lambda|^2 \,\middle|\, Y_n^\lambda(\mathbf{x})\right] \\
&= |\Delta_{n,t}(\mathbf{x})|^{2p} + p(2p-1)\mathbb{E}\left[\left(|\Delta_{n,t}(\mathbf{x})| + |U_{n,t}^\lambda|\right)^{2p-2}|U_{n,t}^\lambda|^2 \,\middle|\, Y_n^\lambda(\mathbf{x})\right] \\
&= |\Delta_{n,t}(\mathbf{x})|^{2p} + \lambda(t-n)2^{2p-2}p(2p-1)d\beta^{-1}|\Delta_{n,t}(\mathbf{x})|^{2p-2} + 2^{2p-3}p(2p-1)\mathbb{E}\left[|U_{n,t}^\lambda|^{2p}\right]
\end{aligned}$$

which yields, using moment estimates given in [29, Theorem 7.1, Chapter 1], that

$$\mathbb{E}[|Y_t^\lambda(\mathbf{x})|^{2p}|Y_n^\lambda(\mathbf{x})] \leq |\Delta_{n,t}(\mathbf{x})|^{2p} + \lambda(t-n)2^{2p-2}p(2p-1)d\beta^{-1}|\Delta_{n,t}(\mathbf{x})|^{2p-2} + 2^{3p-3}\left(\lambda(t-n)\right)^p\left(p(2p-1)\right)^{p+1}\left\{d\beta^{-1}\right\}^p. \tag{47}$$

Using (43) and the inequalities (45) and (46) with $\epsilon = a\lambda(t-n)/2$, one calculates

$$\begin{aligned}
&|\Delta_{n,t}(\mathbf{x})|^{2p} \leq \{(1-a\lambda(t-n))|Y_n^\lambda(\mathbf{x})|^2 + \lambda(t-n)(c_0|x_n|^2 + c_2)\}^p \\
&\leq (1+\frac{a\lambda(t-n)}{2})^{p-1}(1-a\lambda(t-n))^p|Y_n^\lambda(\mathbf{x})|^{2p} + (1+\frac{2}{a\lambda(t-n)})^{p-1}\lambda^p(t-n)^p(c_0|x_n|^2+c_2)^p \\
&\leq a_{n,t}^{\lambda,p}|Y_n^\lambda(\mathbf{x})|^{2p} + b_{n,t}^{\lambda,p}
\end{aligned} \tag{48}$$

where $a_{n,t}^{\lambda,p} = (1-a\lambda(t-n)/2)^{p-1}(1-a\lambda(t-n))$ and $b_{n,t}^{\lambda,p} = (\lambda(t-n)+2/a)^{p-1}\lambda(t-n)(c_0|x_n|^2+c_2)^p$. Substituting (48) into (47) yields

$$\begin{aligned}
\mathbb{E}[|Y_t^\lambda(\mathbf{x})|^{2p}|Y_n^\lambda(\mathbf{x})] &\leq a_{n,t}^{\lambda,p}|Y_n^\lambda(\mathbf{x})|^{2p} + b_{n,t}^{\lambda,p} + \lambda(t-n)2^{2p-2}p(2p-1)d\beta^{-1} \\
&\times\left[a_{n,t}^{\lambda,p-1}|Y_n^\lambda(\mathbf{x})|^{2(p-1)} + b_{n,t}^{\lambda,p-1}\right] + 2^{3p-3}\left(\lambda(t-n)\right)^p\left(p(2p-1)\right)^{p+1}\left(d\beta^{-1}\right)^p.
\end{aligned} \tag{49}$$

Define $\tilde{M}(p)$ as in (42) and observe that for $|Y_n^\lambda(\mathbf{x})| \geq \sqrt{d}\tilde{M}(p)$

$$\frac{a\lambda(t-n)}{4}|Y_n^\lambda(\mathbf{x})|^{2p} \geq \lambda(t-n)2^{2p}p(2p-1)\frac{d}{4\beta}|Y_n^\lambda(\mathbf{x})|^{2(p-1)}.$$

Consequently, on $\{|Y_n^\lambda(\mathbf{x})| \geq \sqrt{d}\tilde{M}(p)\}$ the inequality (49) yields

$$\begin{aligned}
&\mathbb{E}[|Y_t^\lambda(\mathbf{x})|^{2p}|Y_n^\lambda(\mathbf{x})] \leq (1-a\lambda(t-n)/4)a_{n,t}^{\lambda,p-1}|Y_n^\lambda(\mathbf{x})|^{2p} + b_{n,t}^{\lambda,p} \\
&+ \lambda(t-n)2^{2p-2}p(2p-1)(d/\beta)b_{n,t}^{\lambda,p-1} + \lambda^p(t-n)^p2^{3p-3}(p(2p-1))^{p+1}\left(d/\beta\right)^p \\
&\leq (1-a\lambda(t-n))|Y_n^\lambda(\mathbf{x})|^{2p} + \lambda(t-n)a\left(M(p,d)|x_n|^{2p} + \widehat{M}(p,d)\right),
\end{aligned} \tag{50}$$

where the constants $M(p,d)$ and $\widehat{M}(p,d)$ are defined in (41). Moreover, on $\{|Y_n^\lambda(\mathbf{x})| < \sqrt{d}\tilde{M}(p)\}$ the inequality (49) yields again

$$\mathbb{E}[|Y_t^\lambda(\mathbf{x})|^{2p}|Y_n^\lambda(\mathbf{x})] \;\leq\; (1 - a\lambda(t-n))|Y_n^\lambda(\mathbf{x})|^{2p} + \lambda(t-n)a\left(M(p,d)|x_n|^{2p} + \widehat{M}(p,d)\right) \quad (51)$$

Eq. (40) follows immediately from (51) and (50). □

Remark 3.10. One notes here that $\left(\mathbb{E}[|Y_t^\lambda(\mathbf{x})|^{2p}]\right)^{1/(2p)}$ is of order $\sqrt{d}$, where d denotes the dimension of the problem.

Corollary 3.11. *Assume **H**1, **H**2 and **H**4. For each $0 < \lambda \leq \lambda_{\max}$ and $0 \leq s \leq t$, let $\tilde{Y}_{s,t}^\lambda(\mathbf{x},\theta)$ be the solution of* (26) *with initial condition θ. Then for each $k \geq 1$,*

$$\mathbb{E}[V_4(\tilde{Y}_{(k-1)T,kT}^\lambda(\mathbf{x}, Y_{(k-1)T}^\lambda(\mathbf{x})))] \leq 2\mathrm{e}^{-a}(1-a\lambda)^{(k-1)T}\mathbb{E}|\theta_0|^4 + 2\mathrm{e}^{-a}\left\{1 + a\lambda M(2,d)\sum_{j=0}^{(k-1)T-1}(1-a\lambda)^j\,|x_{(k-1)T-1-j}|^4 + \widehat{M}(2,d)\right\} + 3\mathrm{v}_4(\overline{M}(4)), \quad (52)$$

where the constants $M(2,d)$ and $\widehat{M}(2,d)$ are given by (41) *and* (42) *with $p = 2$.*

Proof. A direct consequence of Lemma 3.6, (40) and the fact that $C_6(4) = a$. □

We now define a continuous-time filtration $(\mathcal{H}_t)_{t\geq 0}$ that encapsulates the information flow of X_n, $n \in \mathbb{N}$ as well as all the "auxiliary" randomness of the Brownian motion B_t, $t \in \mathbb{R}_+$. We also introduce the corresponding decreasing family of σ-algebras $(\mathcal{H}_t^+)_{t\geq 0}$.

$$\mathcal{H}_t := \mathcal{F}_\infty \vee \mathcal{G}_{\lfloor t \rfloor} \quad \text{and} \quad \mathcal{H}_t^+ := \mathcal{G}_{\lfloor t \rfloor}^+, \quad t \in \mathbb{R}_+ \quad (53)$$

where $(\mathcal{G}_n, \mathcal{G}_n^+)_{n\in\mathbb{N}}$ are as in Assumption 3.

We introduce another auxiliary process that play a prominent rôle in the sequel. Let $L_{s,t}^\lambda(\vartheta)$, $t \geq s$ denote the solution of the SDE

$$\mathrm{d}L_{s,t}^\lambda(\vartheta) = -\lambda h(L_{s,t}^\lambda(\vartheta))\,\mathrm{d}t + \{2\lambda\beta^{-1}\}^{1/2}\mathrm{d}\tilde{B}_t^\lambda, \quad (54)$$

with initial condition $L_{s,s}^\lambda(\vartheta) := \vartheta$ for some $\mathcal{H}_s^\lambda$-measurable random variable ϑ. Note that $L_t^\lambda = L_{0,t}^\lambda(\theta_0)$. At this point, we introduce

$$T := \lfloor 1/\lambda \rfloor, \quad (55)$$

which is used for the creation of a suitable set of grid points. Fix $n \in \mathbb{N}$ and define for any $t \in [nT, \infty)$

$$\overline{L}_{nT,t}^\lambda = L_{nT,t}^\lambda(\tilde{Y}_{nT}^\lambda(\mathbf{X})). \quad (56)$$

Note that $\overline{L}_{nT,t}^\lambda$ is $\mathcal{H}_{nT}$-measurable for all $t \geq nT$.

Lemma 3.12. *Assume **H**1, **H**2 and **H**4. For any integers $p \geq 2$, $n \in \mathbb{N}$, $\lambda > 0$ and $t \geq nT$,*

$$\mathbb{E}[V_p(\overline{L}_{nT,t}^\lambda)] \leq \mathrm{e}^{-\lambda C_6(p)t}\mathbb{E}[V_p(\theta_0)] + 6\mathrm{v}_p(\overline{M}(p)), \quad (57)$$

where $\overline{M}(p)$ and V_p are defined in (33) *and* (19)*.*

Proof. By taking into consideration (31) and by arguing as in Lemma 3.6, one obtains $\mathbb{E}[V_p(\overline{L}_{nT,t}^\lambda)] \leq \mathrm{e}^{-\lambda C_6(p)(t-nT)}\mathbb{E}[V_p(\tilde{Y}_{nT}^\lambda(\mathbf{X}))] + 3\mathrm{v}_p(\overline{M}(p))$. Hence, the desired result follows from Corollary 3.8. □

Control of the supremum process of $\overline{L}_{nT,t}^\lambda$ is an essential ingredient in the proof of Lemma 3.17 below.

Corollary 3.13. *Assume **H**1, **H**2 and **H**4. For any integer $p \geq 2$,*

$$\mathbb{E}[\sup_{nT\leq t\leq (n+1)T} V_p(\overline{L}^{\lambda}_{nT,t})] \leq 3\mathrm{e}^{-\lambda C_6(p)nT}\mathbb{E}[V_p(\theta_0)] + C_{12}(p), \tag{58}$$

where T and $\overline{M}(2p)$ are given in (55) *and* (33) *respectively, and*

$$C_{12}(p) := 9(1+(3ap)^{1/2}/2)\,\mathrm{v}_p(\overline{M}(2p))\,. \tag{59}$$

Proof. For any $n \in \mathbb{N}$, $q \geq 2$ and any bounded stopping time $\tau_n \geq nT$ (a.s.), arguing as in Lemma 3.5 results in

$$\begin{aligned}\mathbb{E}\left[V_q(\overline{L}^{\lambda}_{nT,\tau_n})\,\middle|\,\mathcal{H}_{nT}\right] &\leq V_q(\tilde{Y}^{\lambda}_{nT}(\mathbf{X})) + \mathbb{E}\left[\int_{nT}^{\tau_n}\left(-\lambda C_6(q)V_q(\overline{L}^{\lambda}_{nT,s}) + \lambda C_7(q)\right)\,\mathrm{d}s\,\middle|\,\mathcal{H}_{nT}\right]\\ &\leq V_q(\tilde{Y}^{\lambda}_{nT}(\mathbf{X})) + \lambda C_7(q)\mathbb{E}\left[(\tau_n - nT)\,|\,\mathcal{H}_{nT}\right]\,.\end{aligned}$$

Then, according to Lenglart's domination inequality, see [35, Chapter IV, Proposition 4.7], with dominating process

$$A_t := V_q\left(\tilde{Y}^{\lambda}_{nT}(\mathbf{X})\right) + \lambda C_7(q)(t-nT),\ \text{for any } t \geq nT,$$

one obtains, for any $k \in (0,1)$,

$$\mathbb{E}\left[\left(\sup_{nT\leq t\leq (n+1)T} V_q(\overline{L}^{\lambda}_{nT,t})\right)^k\right] \leq \frac{2-k}{1-k}\mathbb{E}[A^k_{(n+1)T}]$$

Thus, using $(a+b)^k \leq a^k + b^k$ for any $a, b \geq 0$ and $k \in (0,1)$, we get

$$\mathbb{E}\left[\left(\sup_{nT\leq t\leq (n+1)T} V_q(\overline{L}^{\lambda}_{nT,t})\right)^k\right] \leq \frac{2-k}{1-k}\left\{\mathbb{E}\left[\left(V_q(\tilde{Y}^{\lambda}_{nT}(\mathbf{X}))\right)^k\right] + C_7^k(q)(\lambda T)^k\right\}\,.$$

Consequently, for $k = 1/2$ and $q = 2p$ and in view of Corollary 3.8, the desired result holds. □

3.5 Contraction estimates

A crucial contraction property is formulated in the next theorem, based on the deep results of [18].

Proposition 3.14. *Let $(L'_t)_{t\in\mathbb{R}_+}$ be the solution of* (23) *with initial condition $L'_0 = \theta'_0$ which is independent of $\mathcal{F}_\infty$ and satisfies $\theta'_0 \in L^2$. Then,*

$$w_{1,2}(\mathcal{L}(L_t), \mathcal{L}(L'_t)) \leq C_9\mathrm{e}^{-C_8 t} w_{1,2}(\mathcal{L}(\theta_0), \mathcal{L}(\theta'_0)),\ t \in \mathbb{R}_+, \tag{60}$$

where the constants C_8 and C_9 are given explicitly in Lemma 3.24 and $w_{1,2}$ is defined in (21)*. Fix a positive integer m. Suppose, for any $t > m$, $\tilde{Y}^{\lambda}_{m,t}(\mathbf{x}, \tilde{\theta})$ and $\tilde{Y}^{\lambda}_{m,t}(\mathbf{x}, \tilde{\theta}')$ are the solutions of* (26) *with initial conditions $\tilde{\theta}$, $\tilde{\theta}' \in L^2$, which are independent of $\mathcal{F}_\infty$. Then, for any $t > m$, we get*

$$w_{1,2}(\mathcal{L}(\tilde{Y}^{\lambda}_{m,t}(\mathbf{x}, \tilde{\theta})), \mathcal{L}(\tilde{Y}^{\lambda}_{m,t}(\mathbf{x}, \tilde{\theta}'))) \leq C_9\mathrm{e}^{-C_8\lambda(t-m)} w_{1,2}(\mathcal{L}(\tilde{\theta}), \mathcal{L}(\tilde{\theta}'))\,. \tag{61}$$

Proof. We first treat L_t, L'_t. [18, Assumption 2.1] holds with κ constant (and equal to K_1) due to **H**2. [18, Assumption 2.5] holds due to (20). [18, Assumption 2.2] holds with $V = V_2$ due to Lemma 3.5 (note that in that paper the diffusion coefficient is assumed to be 1 while in our case it is $\sqrt{2/\beta}$ but this does not affect the validity of the arguments, only the values of the constants). Thus, in view of [18, Corollary 2.3],

$$\mathcal{W}_{\rho_2}(\mathcal{L}(L_t), \mathcal{L}(L'_t)) \leq \mathrm{e}^{-C_8 t}\mathcal{W}_{\rho_2}(\mathcal{L}(\theta_0), \mathcal{L}(\theta'_0)),\ t \in \mathbb{R}_+,$$

where C_8 is given in Lemma 3.24 and the functional $\mathcal{W}_{\rho_2}$ comes from [18] with the choice $V := V_2$, for $\mu, \nu \in \mathcal{P}_{V_2}(\mathbb{R}^d)$

$$\mathcal{W}_{\rho_2}(\mu,\nu) := \inf_{\zeta\in\mathcal{C}(\mu,\nu)} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(|\theta-\theta'|)(1+\epsilon V_2(\theta)+\epsilon V_2(\theta'))\zeta(\mathrm{d}\theta\mathrm{d}\theta'), \tag{62}$$

where f is a concave, bounded and non-decreasing continuous function and ϵ is a positive constant, for more details see [18, Section 5]. Consequently, by using the definition of $\mathcal{W}_{\rho_2}$, one obtains

$$C_{10}w_{1,2}(\mu,\nu) \le \mathcal{W}_{\rho_2}(\mu,\nu) \le C_{11}w_{1,2}(\mu,\nu), \quad \mu,\nu\in\mathcal{P}_{V_2}(\mathbb{R}^d), \tag{63}$$

where C_{10}, C_{11} are calculated in Lemma 3.24 below. Statement (60) follows with $C_9 = C_{11}/C_{10}$.

The same approach is used for $\tilde{Y}^{\lambda}_{m,t}(\mathbf{x},\tilde{\theta})$ and $\tilde{Y}^{\lambda}_{m,t}(\mathbf{x},\tilde{\theta}')$, with the only difference being that we derive first the contraction on an interval of length at most one, since the contribution from the data sequence, through $x_{\lfloor t\rfloor}$, remains constant and thus, the drift coefficient remains autonomous for such an interval. More concretely, [18, Assumption 2.1] holds in this case too with κ constant and equal to K_1 due to **H**2. [18, Assumption 2.2] is true with $V = V_2$ due to Lemma 3.5. Note that the statements in these Assumptions are uniform in x (and thus identical for different values of $x_{\lfloor t\rfloor}$). Finally, [18, Assumption 2.5] is also true due to (20). Thus, the results of [18, Corollary 2.3] apply in this case, too, and one concludes that

$$\begin{aligned}
&\mathcal{W}_{\rho_2}(\mathcal{L}(\tilde{Y}^{\lambda}_{m,t}(\mathbf{x},\tilde{\theta})),\mathcal{L}(\tilde{Y}^{\lambda}_{m,t}(\mathbf{x},\tilde{\theta}')))\\
&\quad= \mathcal{W}_{\rho_2}(\mathcal{L}(\tilde{Y}^{\lambda}_{\lfloor t\rfloor,t}(\mathbf{x},\tilde{Y}^{\lambda}_{m,\lfloor t\rfloor}(\mathbf{x},\tilde{\theta}))),\mathcal{L}(\tilde{Y}^{\lambda}_{\lfloor t\rfloor,t}(\mathbf{x},\tilde{Y}^{\lambda}_{m,\lfloor t\rfloor}(\mathbf{x},\tilde{\theta}'))))\\
&\quad\le \mathrm{e}^{-C_8\lambda(t-\lfloor t\rfloor)}\mathcal{W}_{\rho_2}\left(\mathcal{L}(\tilde{Y}^{\lambda}_{m,\lfloor t\rfloor}(\mathbf{x},\tilde{\theta})),\mathcal{L}(\tilde{Y}^{\lambda}_{m,\lfloor t\rfloor}(\mathbf{x},\tilde{\theta}'))\right)\\
&\quad\le \mathrm{e}^{-C_8\lambda(t-(\lfloor t\rfloor-1))}\mathcal{W}_{\rho_2}(\mathcal{L}(\tilde{Y}^{\lambda}_{m,\lfloor t\rfloor-1}(\mathbf{x},\tilde{\theta})),\mathcal{L}(\tilde{Y}^{\lambda}_{m,\lfloor t\rfloor-1}(\mathbf{x},\tilde{\theta}')))\\
&\quad\le \dots\\
&\quad\le \mathrm{e}^{-C_8\lambda(t-m)}\mathcal{W}_{\rho_2}(\mathcal{L}(\tilde{\theta}),\mathcal{L}(\tilde{\theta}')).
\end{aligned} \tag{64}$$

Observing as above that $\mathcal{W}_{\rho_2}$ is controlled from above and below by multiples of $w_{1,2}$, (64) yields the result. □

3.6 The core lemmas

Our arguments for handling the second term in (29) rest upon Lemmas 3.17 and 3.18 below. As a preparation, we first recall two lemmas: one on regular versions and one on moment estimates that are closely related to the conditional L-mixing property.

Lemma 3.15. *For each $n\in\mathbb{N}$, there exists a measurable function $h:\Omega\times[nT,\infty)\times\mathbb{R}^d\to\mathbb{R}^d$ such that, for each $t\ge nT$ and $\theta\in\mathbb{R}^d$, $h_{nT,t}(\theta)(\omega)$ is a version of $\mathbb{E}\left[H(\theta,X_{\lfloor t\rfloor})\,\middle|\,\mathcal{H}_{nT}\right]$ for almost every $\omega\in\Omega$, $\theta\to h_{nT,t}(\theta)(\omega)$ is continuous.*

Proof. As $h_{nT,t}$, $t\in[k,k+1)$ can be assumed constant for each $k\in\mathbb{N}$, it suffices to prove the existence of a measurable $h_{nT,k}:\Omega\times\mathbb{R}^d\to\mathbb{R}^d$ which is continuous in its second variable, for each fixed k. This follows from [2, Lemma 8.5]. □

Lemma 3.16. *Assume **H**2 and **H**3 and let $p\ge 1$. Then,*

$$\sup_{n\in\mathbb{N}}\mathbb{E}^{1/p}\left[\left(\sum\nolimits_{k=nT}^{\infty}\sup\nolimits_{\theta\in\mathbb{R}^d}\|h_{k,nT}(\theta)-h(\theta)\|\right)^p\right] \le 2K_2\Gamma_p^0(X),$$

where $\Gamma_p^0(X)$ is defined in (12)*.*

Proof. See [2, Lemma 4.9]. □

Now we present the first core lemma.

Lemma 3.17. *Assume **H**1, **H**2 and **H**4 hold. There is C_{13} such that, for each $0 < \lambda \leq \lambda_{\max}$, and for all $t \in [nT, (n+1)T]$,*

$$W_2(\mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{X})), \mathcal{L}(\overline{L}_{nT,t}^\lambda)) \leq C_{13}\lambda^{1/2}[\mathrm{e}^{-an/4}\mathbb{E}^{1/2}[V_2(\theta_0)] + 1].$$

Proof. Fix $t \in [nT, (n+1)T]$. Let us estimate, using **H**2,

$$\begin{aligned}
&\left|\tilde{Y}_t^\lambda(\mathbf{X}) - \overline{L}_{nT,t}^\lambda\right| \leq \lambda \left|\int_{nT}^t \left[H(\tilde{Y}_s^\lambda(\mathbf{X}), X_{\lfloor s \rfloor}) - h(\overline{L}_{nT,s}^\lambda)\right] \mathrm{d}s\right| \\
&\leq \lambda \int_{nT}^t \left|H(\tilde{Y}_s^\lambda(\mathbf{X}), X_{\lfloor s \rfloor}) - H(\overline{L}_{nT,s}^\lambda, X_{\lfloor s \rfloor})\right| \mathrm{d}s + \lambda \left|\int_{nT}^t \left[H(\overline{L}_{nT,s}^\lambda, X_{\lfloor s \rfloor}) - h_{nT,s}(\overline{L}_{nT,s}^\lambda)\right] \mathrm{d}s\right| \\
&+ \lambda \int_{nT}^t \left|h_{nT,s}(\overline{L}_{nT,s}^\lambda) - h(\overline{L}_{nT,s}^\lambda)\right| \mathrm{d}s \\
&\leq \lambda K_1 \int_{nT}^t \left|\tilde{Y}_s^\lambda(\mathbf{X}) - \overline{L}_{nT,s}^\lambda\right| \mathrm{d}s + \lambda A + \lambda B
\end{aligned}$$

where $h_{nT,s}$ is defined in Lemma 3.15 and

$$\begin{aligned}
A &:= \sup_{u \in [nT,(n+1)T]} \left|\int_{nT}^u \left[H(\overline{L}_{nT,s}^\lambda, X_{\lfloor s \rfloor}) - h_{nT,s}(\overline{L}_{nT,s}^\lambda)\right] \mathrm{d}s\right| \\
B &:= \int_{nT}^\infty \sup_{\theta \in \mathbb{R}^d} |h_{nT,s}(\theta) - h(\theta)| \, \mathrm{d}s,
\end{aligned}$$

Now let us apply Grönwall's lemma and take the square of both sides. Using the elementary $(x+y)^2 \leq 2(x^2+y^2)$, $x, y \geq 0$, we arrive at

$$\left|\tilde{Y}_t^\lambda(\mathbf{X}) - \overline{L}_{nT,t}^\lambda\right|^2 \leq 2\lambda^2 \mathrm{e}^{2K_1\lambda T}\{A^2 + B^2\} \tag{65}$$

Introduce for all $i \in \mathbb{N}$ the events

$$F_i^{nT} := \{i \leq \sup_{s \in [nT,(n+1)T]} |\overline{L}_{nT,s}^\lambda| < i+1\},$$

which are $\mathcal{H}_{nT}$-measurable.

We apply below Theorem 3.2 in the following setting. Let $\mathcal{R}_s = \mathcal{H}_{nT+s}$ and $\mathcal{R}_s^+ = \mathcal{H}_{nT+s}^+$ for $s \in \mathbb{R}_+$. Furthermore, let $W_s = W_{s-nT}^{nT,i}$ where we define

$$W_s^{nT,i} := \left(H(\overline{L}_{nT,s}^\lambda, X_{\lfloor s \rfloor}) - h_{nT,s}(\overline{L}_{nT,s}^\lambda)\right) \mathbb{1}_{F_i^{nT}}, \qquad i \in \mathbb{N}, s \geq nT.$$

Clearly, for $s \geq nT$, $\mathbb{E}\left[W_s^{nT,i} \,\middle|\, \mathcal{H}_{nT}\right] = 0$. We now estimate the quantities $M_p(\mathbf{W})$, $\Gamma_p(\mathbf{W})$ appearing in Theorem 3.2.

For each fixed θ, Lemma 3.3 implies that the auxiliary process $\tilde{W}_s^\theta := H(\theta, X_{\lfloor nT+s \rfloor})\mathbb{1}_{F_i^{nT}}$, $s \in \mathbb{R}_+$ satisfies

$$M_p(\tilde{\mathbf{W}}^\theta) \leq K_1 i + K_2 M_p^{nT}(\mathbf{X}) + H^*$$

as well as

$$\Gamma_p(\tilde{\mathbf{W}}^\theta) \leq 2K_2 \Gamma_p^{nT}(\mathbf{X}).$$

Hence Lemma 3.4 guarantees that we can plug in the $\mathcal{H}_{nT}$-measurable process $\overline{L}_{nT,s}^\lambda$ into $\tilde{W}_s^\theta$, getting

$$M_p(\hat{\mathbf{W}}) \leq K_1 i + K_2 M_p^{nT}(\mathbf{X}) + H^*, \ \Gamma_p(\hat{\mathbf{W}}) \leq 2K_2 \Gamma_p^{nT}(\mathbf{X})$$

for the process defined by

$$\hat{W}_s := H(\overline{L}^{\lambda}_{nT,s}, X_{\lfloor nT+s \rfloor}) \mathbb{1}_{F_i^{nT}},\ s \in \mathbb{R}_+.$$

Finally, by [4, Remark A.4] (or after a moment's reflection), we find that

$$M_p(\mathbf{W}) \leq 2[K_1 i + K_2 M_p^{nT}(\mathbf{X}) + H^*],\ \Gamma_p(\mathbf{W}) \leq 2K_2 \Gamma_p^{nT}(\mathbf{X}).$$

Applying Theorem 3.2 with $r := 3$, we obtain

$$\begin{aligned}
&\mathbb{E}^{1/2}\left[\sup_{u\in[nT,(n+1)T]}\left|\int_{nT}^{u}[H(\overline{L}^{\lambda}_{nT,s}, X_{\lfloor s\rfloor}) - h_{nT,s}(\overline{L}^{\lambda}_{nT,s})]\,\mathrm{d}s\right|^2 \mathbb{1}_{F_i^{nT}}\,\middle|\,\mathcal{H}_{nT}\right] \\
&\leq \mathbb{E}^{1/3}\left[\sup_{u\in[nT,(n+1)T]}\left|\int_{nT}^{u}[H(\overline{L}^{\lambda}_{nT,s}, X_{\lfloor s\rfloor}) - h_{nT,s}(\overline{L}^{\lambda}_{nT,s})]\,\mathrm{d}s\right|^3 \mathbb{1}_{F_i^{nT}}\,\middle|\,\mathcal{H}_{nT}\right] \\
&\leq 2C'(3)\sqrt{T}[K_1 i + K_2 M_3^{nT}(\mathbf{X}) + K_2\Gamma_3^{nT}(\mathbf{X}) + H^*]\mathbb{1}_{F_i^{nT}} \\
&\leq 20\sqrt{T}[K_1(1+\sup_{s\in[nT,(n+1)T]}|\overline{L}^{\lambda}_{nT,s}|) + K_2 M_3^{nT}(\mathbf{X}) + K_2\Gamma_3^{nT}(\mathbf{X}) + H^*]\mathbb{1}_{F_i^{nT}},
\end{aligned}$$

noting that the constant $C'(3)$ appearing in Theorem 3.2 satisfies $C'(3) \leq 10$. We can then estimate, noting $C_6(2) = a/2$ (see Lemma 3.5),

$$\begin{aligned}
\mathbb{E}^{1/2}\left[A^2\right] \leq& 20\sqrt{T}[K_1(E^{1/2}[\sup_{s\in[nT,(n+1)T]}|\overline{L}^{\lambda}_{nT,s}|^2] + 1) + K_2 E^{1/2}[(M_3^{nT})^2] \\
&+ K_2 E^{1/2}[(\Gamma_3^{nT})^2] + H^*] \\
\leq& 20\lambda^{-1/2}[K_1(\sqrt{3}(\mathrm{e}^{-\lambda a nT/4}\mathbb{E}^{1/2}[V_2(\theta_0)] + 1 + C_{12}^{1/2}(2)) + K_2 E^{1/2}[(M_3^{nT})^2] \\
&+ K_2 E^{1/2}[(\Gamma_3^{nT})^2] + H^*]
\end{aligned}$$

using Corollary 3.13 with the choice $p = 2$.

Finally, for any $t \in [nT, (n+1)T]$ and $\lambda \in (0, \lambda_{\max}]$, using (65) and Lemma 3.16 we get

$$\begin{aligned}
W_2(\mathcal{L}(\tilde{Y}_t^{\lambda}(\mathbf{X})), \mathcal{L}(\overline{L}^{\lambda}_{nT,t})) \leq& \mathbb{E}^{1/2}\left|\tilde{Y}_t^{\lambda}(\mathbf{X}) - \overline{L}^{\lambda}_{nT,t}\right|^2 \\
\leq& 20\sqrt{2}\mathrm{e}^{K_1}\lambda^{1/2}[K_1(\sqrt{3}(\mathrm{e}^{-\lambda a nT/4}\mathbb{E}^{1/2}[V_2(\theta_0)] + 1 + H^* + C_{12}^{1/2}(2)) \\
&+ K_2 E^{1/2}[(M_3^{nT})^2] + K_2 E^{1/2}[(\Gamma_3^{nT})^2] + 2\lambda_{\max} K_2 \Gamma_2^0(\mathbf{X})]. \qquad (66)
\end{aligned}$$

So we can conclude choosing

$$\begin{aligned}
C_{13} \;=\;& 20\sqrt{2}\mathrm{e}^{K_1}[K_1\sqrt{3} + K_1(1 + C_{12}^{1/2}(2)) + \\
+\;& K_2 E^{1/2}[(M_3^{nT})^2] + K_2 E^{1/2}[(\Gamma_3^{nT})^2] + H^* + 2\lambda_{\max} K_2 \Gamma_2^0(\mathbf{X})].
\end{aligned}$$

□

The second core lemma follows from the first and from Proposition 3.14.

Lemma 3.18. *Assume **H1, H2 and H4**. For each $0 < \lambda \leq \lambda_{\max}$, $n \in \mathbb{N}$ and $t \in [nT, (n+1)T)$,*

$$W_1(\mathcal{L}(\overline{L}^{\lambda}_{nT,t}), \mathcal{L}(L_t^{\lambda})) \leq C_{14}[1 + \mathrm{e}^{-\min\{C_8, a/4\}n/2}\mathbb{E}^{3/4}[V_4(\theta_0)]]\sqrt{\lambda}$$

for a suitable C_{14}, explicitly given in the proof.

Proof. Using telescopic sums, (22) and Proposition 3.14, we get

$$
\begin{aligned}
W_1(\mathcal{L}(\overline{L}^{\lambda}_{nT,t}), \mathcal{L}(L^{\lambda}_t)) \leq & \sum_{k=1}^{n} W_1\left(\mathcal{L}(L^{\lambda}_{kT,t}(\tilde{Y}^{\lambda}_{kT}(\mathbf{X}))), \mathcal{L}(L^{\lambda}_{(k-1)T,t}(\tilde{Y}^{\lambda}_{(k-1)T}(\mathbf{X})))\right) \qquad (67)\\
\leq & \sum_{k=1}^{n} w_{1,2}(\mathcal{L}(L^{\lambda}_{kT,t}(\tilde{Y}^{\lambda}_{kT}(\mathbf{X}))), \mathcal{L}(L^{\lambda}_{kT,t}(L^{\lambda}_{(k-1)T,kT}(\tilde{Y}^{\lambda}_{(k-1)T}(\mathbf{X})))))\\
\leq & C_9 \sum_{k=1}^{n} \exp\left(-C_8(n-k)\right) w_{1,2}(\mathcal{L}(\tilde{Y}^{\lambda}_{kT}(\mathbf{X})), \mathcal{L}(\overline{L}^{\lambda}_{(k-1)T,kT})).
\end{aligned}
$$

Using the definitions (21) and (3) of $w_{1,2}$ and W_2, we get from the Cauchy inequality that

$$
\begin{aligned}
w_{1,2}(\mathcal{L}(\tilde{Y}^{\lambda}_{kT}(\mathbf{X})), \mathcal{L}(\overline{L}^{\lambda}_{(k-1)T,kT})) \leq & W_2(\mathcal{L}(\tilde{Y}^{\lambda}_{kT}(\mathbf{X})), \mathcal{L}(\overline{L}^{\lambda}_{(k-1)T,kT}))\\
& \times [1 + \{\mathbb{E}[V_4(\tilde{Y}^{\lambda}_{kT}(\mathbf{X}))]\}^{1/2} + \{\mathbb{E}[V_4(\overline{L}^{\lambda}_{(k-1)T,kT})]\}^{1/2}].
\end{aligned}
$$

Corollary 3.8, Lemma 3.12 and Lemma 3.17 imply that

$$
\begin{aligned}
w_{1,2}(\mathcal{L}(\tilde{Y}^{\lambda}_{kT}(\mathbf{X})), \mathcal{L}(\overline{L}^{\lambda}_{(k-1)T,kT})) \leq & C_{13}\lambda^{1/2}[\mathrm{e}^{-a(k-1)/4}\mathbb{E}^{1/2}[V_2(\theta_0)] + 1]\\
& \times [1 + 2\mathrm{e}^{-ak/2}\{\mathbb{E}[V_4(\theta_0)]\}^{1/2} + \sqrt{3}\mathrm{v}_2(\overline{M}(4)) + \sqrt{6}\mathrm{v}_2(\overline{M}(4))]
\end{aligned}
$$

since $C_6(4) = a$. For each $y \geq 0$ and $\alpha > 0$, $e^{-\alpha y}(y+1) \leq 1 + 1/\alpha$. In the estimations below we apply this latter observation with $\alpha = \min(C_8, a/4)/2$ and $y = n-1$. Noticing that $\mathbb{E}^{1/2}[V_2(\theta_0)] \leq \mathbb{E}^{1/4}[V_4(\theta_0)]$, we can proceed as

$$
\begin{aligned}
& \sum_{k=1}^{n} \exp\left(-C_8(n-k)\right) w_{1,2}(\mathcal{L}(\tilde{Y}^{\lambda}_{kT}(\mathbf{X})), \mathcal{L}(\overline{L}^{\lambda}_{(k-1)T,kT}))\\
\leq & C_{13}\lambda^{1/2}[2\mathbb{E}^{3/4}[V_4(\theta_0)] + \mathbb{E}^{1/2}[V_4(\theta_0)] + (5\mathrm{v}_2(\overline{M}(4)) + 1)\mathbb{E}^{1/4}[V_4(\theta_0)]]\\
\times & \sum_{k=1}^{n} \exp\left(-\min\{C_8, a/4\}(n-k+k-1)\right)\\
+ & C_{13}\lambda^{1/2}\frac{5\mathrm{v}_2(\overline{M}(4)) + 1}{1-\mathrm{e}^{-C_8}}\\
\leq & C_{13}\lambda^{1/2} n \exp\left(-\min\{C_8, a/4\}(n-1)\right)[(5\mathrm{v}_2(\overline{M}(4)) + 4)\mathbb{E}^{3/4}[V_4(\theta_0)] + 5\mathrm{v}_2(\overline{M}(4)) + 1 + 1]\\
+ & C_{13}\lambda^{1/2}\frac{5\mathrm{v}_2(\overline{M}(4)) + 1}{1-\mathrm{e}^{-C_8}}\\
\leq & C_{13}\lambda^{1/2}\exp\left(-\min\{C_8, a/4\}(n-1)/2\right)\left(1 + \frac{2}{\min\{C_8, a/4\}}\right)\\
\times & [(5\mathrm{v}_2(\overline{M}(4)) + 4)\mathbb{E}^{3/4}[V_4(\theta_0)] + 5\mathrm{v}_2(\overline{M}(4)) + 2]\\
+ & C_{13}\lambda^{1/2}\frac{5\mathrm{v}_2(\overline{M}(4)) + 1}{1-\mathrm{e}^{-C_8}},
\end{aligned}
$$

and we can set

$$
\begin{aligned}
C_{14} = & C_9 C_{13}\left(1 + \frac{2}{\min\{C_8, a/4\}}\right)\mathrm{e}^{\min\{C_8, a/4\}}\left[5\mathrm{v}_2(\overline{M}(4)) + 4\right]\\
+ & C_9 C_{13}\left[\frac{5\mathrm{v}_2(\overline{M}(4)) + 1}{1-\mathrm{e}^{-C_8}} + (5\mathrm{v}_2(\overline{M}(4)) + 2)\left(1 + \frac{2}{\min\{C_8, a/4\}}\right)\right].
\end{aligned}
$$

□

Corollary 3.19. *For each $nT \le t < (n+1)T$,*

$$W_1(\mathcal{L}(L_t^\lambda), \mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{X}))) \le C_{15}[1 + \exp(-\min\{C_8, a/4\}n/2)\mathbb{E}^{3/4}[V_4(\theta_0)]]\sqrt{\lambda},$$

for some C_{15}, explicitly given in the proof.

Proof. Notice that $\mathbb{E}^{1/2}[V_2(\theta_0)] \le \mathbb{E}^{1/4}[V_4(\theta_0)]$. Putting together our previous estimations, we arrive at

$$\begin{aligned}
& W_1(\mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{X})), \mathcal{L}(L_t^\lambda)) \\
\le\ & W_1(\mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{X})), \mathcal{L}(\overline{L}_{nT,t}^\lambda)) + W_1(\mathcal{L}(\overline{L}_{nT,t}^\lambda), \mathcal{L}(L_t^\lambda)) \\
\le\ & \sqrt{\lambda}[\mathrm{e}^{-\min\{C_8, a/4\}n/2}[C_{14}\mathbb{E}^{3/4}[V_4(\theta_0)] + C_{13}\mathbb{E}^{1/4}[V_4(\theta_0)]] + C_{14} + C_{13}] \\
\le\ & \sqrt{\lambda}[\mathrm{e}^{-\min\{C_8, a/4\}n/2}\mathbb{E}^{3/4}[V_4(\theta_0)](C_{14} + C_{13}) + C_{14} + 2C_{13}]
\end{aligned}$$

so we can set $C_{15} := C_{14} + 2C_{13}$. □

3.7 Entropy estimates

We develop in this subsection the estimates that are necessary for coping with the third term in (29). Although the principal ideas are well-known, see e.g. [9, 11, 17], the details require rather tedious technicalities since the estimates depend on the "frozen" data stream $\mathbf{x} = (x_n)_{n\in\mathbb{N}}$.

Lemma 3.20. *Assume **H**1, **H**2 and **H**4 hold. For each $0 < \lambda \le \lambda_{\max}$ (see (9)) and $n \in \mathbb{N}$ we have, for all $t \in (nT, (n+1)T]$ and $\mathbf{x} \in (\mathbb{R}^m)^{\mathbb{N}}$, that*

$$W_1(\mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{x})), \mathcal{L}(Y_t^\lambda(\mathbf{x}))) \le \lambda^{1/2}\mathrm{e}^{-\min(C_8,a)n/2}C_{17}\mathbb{E}[|\theta_0|^4] + \lambda^{1/2}C_{18}(\mathbf{x}, n, \lambda)\ .$$

where C_{17} and $C_{18}(\mathbf{x}, n, \lambda)$ are given by (87) and (88) below, respectively.

Proof. Recall (26) and observe that $\tilde{Y}_t^\lambda(\mathbf{x}) = \tilde{Y}_{0,t}^\lambda(\mathbf{x}, \theta_0)$. Using telescopic sums, we get for $t \in (nT, (n+1)T]$,

$$\begin{aligned}
& W_1(\mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{x})), \mathcal{L}(Y_t^\lambda(\mathbf{x}))) = W_1(\mathcal{L}(\tilde{Y}_{0,t}^\lambda(\mathbf{x}, \theta_0)), \mathcal{L}(Y_t^\lambda(\mathbf{x}))) \\
& \le \sum_{k=1}^{n} W_1(\mathcal{L}(\tilde{Y}_{kT,t}^\lambda(\mathbf{x}, Y_{kT}^\lambda(\mathbf{x}))), \mathcal{L}(\tilde{Y}_{(k-1)T,t}^\lambda(\mathbf{x}, Y_{(k-1)T}^\lambda(\mathbf{x})))) \\
& + W_1(\mathcal{L}(\tilde{Y}_{nT,t}^\lambda(\mathbf{x}, Y_{nT}^\lambda(\mathbf{x}))), \mathcal{L}(Y_t^\lambda(\mathbf{x}))) \\
& \le \sum_{k=1}^{n} w_{1,2}(\mathcal{L}(\tilde{Y}_{kT,t}^\lambda(\mathbf{x}, Y_{kT}^\lambda(\mathbf{x}))), \mathcal{L}(\tilde{Y}_{kT,t}^\lambda(\mathbf{x}, \tilde{Y}_{(k-1)T,kT}^\lambda(\mathbf{x}, Y_{(k-1)T}^\lambda(\mathbf{x})))))) \\
& + w_{1,2}(\mathcal{L}(\tilde{Y}_{nT,t}^\lambda(\mathbf{x}, Y_{nT}^\lambda(\mathbf{x}))), \mathcal{L}(Y_t^\lambda(\mathbf{x}))),
\end{aligned}$$

where the domination of W_1 by $w_{1,2}$ is used, see (22). In view of Proposition 3.14, and in particular inequality (61), one obtains

$$\begin{aligned}
W_1(\mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{x})), \mathcal{L}(Y_t^\lambda(\mathbf{x}))) \le C_9 \sum_{k=1}^{n} \mathrm{e}^{-C_8(n-k)} & w_{1,2}(\mathcal{L}(Y_{kT}^\lambda(\mathbf{x})), \mathcal{L}(\tilde{Y}_{(k-1)T,kT}^\lambda(\mathbf{x}, Y_{(k-1)T}^\lambda(\mathbf{x}))) \\
& + w_{1,2}(\mathcal{L}(\tilde{Y}_{nT,t}^\lambda(\mathbf{x}, Y_{nT}^\lambda(\mathbf{x}))), \mathcal{L}(Y_t^\lambda(\mathbf{x}))). \quad (68)
\end{aligned}$$

At this point, one notes that due to Lemma A.2, for any two probability measures μ and ν on $\mathcal{B}(\mathbb{R}^d)$,

$$w_{1,2}(\mu, \nu) \le \sqrt{2}\left\{1 + [\mu(V_4)]^{1/2} + [\nu(V_4)]^{1/2}\right\}\{\mathrm{KL}(\mu, \nu)\}^{1/2}\ . \quad (69)$$

where $\mathrm{KL}(\mu,\nu)$ denotes the Kullback-Leibler divergence. Thus

$$w_{1,2}(\mathcal{L}(Y^\lambda_{kT}(\mathbf{x})),\mathcal{L}(\tilde{Y}^\lambda_{(k-1)T,kT}(\mathbf{x},Y^\lambda_{(k-1)T}(\mathbf{x})))\le\sqrt{2\lambda}A_k^{1/2}B_k^{1/2}\le\sqrt{\lambda/2}\{A_k+B_k\} \tag{70}$$

where

$$A_k:=\lambda^{-1}\,\mathrm{KL}\left(\mathcal{L}(Y^\lambda_{kT}(\mathbf{x})),\mathcal{L}(\tilde{Y}^\lambda_{(k-1)T,kT}(\mathbf{x},Y^\lambda_{(k-1)T}(\mathbf{x})))\right) \tag{71}$$

$$B_k:=\{1+\mathbb{E}^{1/2}[V_4(Y^\lambda_{kT}(\mathbf{x}))]+\mathbb{E}^{1/2}[V_4(\tilde{Y}^\lambda_{(k-1)T,kT}(\mathbf{x},Y^\lambda_{(k-1)T}(\mathbf{x})))]\}^2 \tag{72}$$

and $1\le k\le n$. For $a<b$, $\mathbf{C}[a,b]$ denotes the Banach space of $\mathbb{R}^d$-valued continuous functions on the interval $[a,b]$. Let $\hat{\mathcal{Q}}_k$ denote the law of the process $\tilde{Y}^\lambda_{(k-1)T,s}(\mathbf{x},Y^\lambda_{(k-1)T}(\mathbf{x}))$, $s\in[(k-1)T,kT]$ on $\mathbf{C}[(k-1)T,kT]$. Similarly, let $\mathcal{Q}_k$ denote the law of $Y^\lambda_s(\mathbf{x})$, $s\in[(k-1)T,kT]$. Lemma A.1 implies that these two probability laws are equivalent. Thus, in view of (96), one then calculates

$$\begin{aligned}
A_k&\le\frac{1}{\lambda}\,\mathrm{KL}(\hat{\mathcal{Q}}_k\|\mathcal{Q}_k)\\
&=\frac{1}{\lambda}\frac{1}{2}\frac{\beta}{2\lambda}\lambda^2\int_{(k-1)T}^{kT}\mathbb{E}|H(Y^\lambda_{\lfloor s\rfloor}(\mathbf{x}),x_{\lfloor s\rfloor})-H(Y^\lambda_s(\mathbf{x}),x_{\lfloor s\rfloor})|^2\,\mathrm{d}s\\
&\le\frac{\beta K_1^2}{4}\int_{(k-1)T}^{kT}\mathbb{E}|Y^\lambda_{\lfloor s\rfloor}(\mathbf{x})-Y^\lambda_s(\mathbf{x})|^2\mathrm{d}s\\
&=\frac{\beta K_1^2}{4}\sum_{j=(k-1)T}^{kT-1}\int_j^{j+1}\mathbb{E}|-\lambda H(Y^\lambda_j(\mathbf{x}),x_j)(s-j)+\sqrt{2\lambda/\beta}(\tilde{B}^\lambda_s-\tilde{B}^\lambda_j)|^2\mathrm{d}s\\
&=\frac{\beta K_1^2}{4}\sum_{j=(k-1)T}^{kT-1}\left\{(1/3)\lambda^2\mathbb{E}|H(Y^\lambda_j(\mathbf{x}),x_j)|^2+d\lambda/\beta\right\}\\
&\le\frac{\beta K_1^2}{4}\sum_{j=(k-1)T}^{kT-1}\left\{\lambda^2\left[(H^*)^2+K_1^2\mathbb{E}|Y^\lambda_j(\mathbf{x})|^2+K_2^2|x_j|^2\right]+d\lambda/\beta\right\}\\
&\le\bar{C}^0(\lambda,\theta_0)(1-a\lambda)^{(k-1)T}+\bar{C}^1_k(\mathbf{x},\lambda)
\end{aligned} \tag{73}$$

where, due to (44), $\bar{C}^0(\lambda,\theta_0)=\lambda\beta K_1^4/(4a)\mathbb{E}|\theta_0|^2$ and

$$\begin{aligned}
\bar{C}^1_k(\mathbf{x},\lambda)=K_1^2\{1+\lambda\beta(H^*)^2+\lambda\beta c_1K_1^2\}/4+(\lambda^2\beta K_1^2K_2^2/4)\sum\nolimits_{j=(k-1)T}^{kT-1}|x_j|^2\\
+(\lambda^3\beta K_1^4c_0/4)\sum\nolimits_{j=(k-1)T}^{kT-1}\sum\nolimits_{l=0}^{j-1}(1-a\lambda)^l|x_{j-1-l}|^2
\end{aligned} \tag{74}$$

where in the case of $k=1$ and $j=0$ the last sum is meant to be 0. Moreover, one calculates the bound for B_k. Using Lemma 3.9 yields that

$$\begin{aligned}
\mathbb{E}[V_4(Y^\lambda_{kT}(\mathbf{x}))]&\le 2+2(1-a\lambda)^{kT}\mathbb{E}|\theta_0|^4+2\lambda aM(2,d)\sum_{j=1}^{kT-1}(1-a\lambda)^{j-1}|x_{kT-1-j}|^4+2\widehat{M}(2,d)\\
&=2(1-a\lambda)^{kT}\mathbb{E}|\theta_0|^4+D_k(\mathbf{x},\lambda),
\end{aligned} \tag{75}$$

where

$$D_k(\mathbf{x},\lambda):=2\lambda aM(2,d)\sum_{j=1}^{kT-1}(1-a\lambda)^{j-1}|x_{kT-1-j}|^4+2\widehat{M}(2,d)+2. \tag{76}$$

Similarly, one obtains, due to Lemma 3.9 and Corollary 3.11,

$$\mathbb{E}[V_4(\tilde{Y}^\lambda_{(k-1)T,kT}(\mathbf{x},Y^\lambda_{(k-1)T}(\mathbf{x})))]\le 2\mathrm{e}^{-a}(1-a\lambda)^{(k-1)T}\mathbb{E}|\theta_0|^4+\mathrm{e}^{-a}D_{k-1}(\mathbf{x},\lambda)+3\mathrm{v}_4(\overline{M}(4)) \tag{77}$$

By observing (70), (73), (75) and (77), it follows that, for $k = 1, \dots, n$,

$$w_{1,2}(\mathcal{L}(Y^{\lambda}_{kT}(\mathbf{x})), \mathcal{L}(\tilde{Y}^{\lambda}_{(k-1)T,kT}(\mathbf{x}, Y^{\lambda}_{(k-1)T}(\mathbf{x})))) \leq \sqrt{\lambda}\left\{(1-a\lambda)^{(k-1)T}\hat{C}^0(\lambda,\theta_0) + \hat{C}^1_k(\mathbf{x},\lambda)\right\}, \quad (78)$$

where $\hat{C}^0(\lambda,\theta_0) = \bar{C}^0(\lambda,\theta_0) + 12\mathbb{E}|\theta_0|^4$ and

$$\hat{C}^1_k(\mathbf{x},\lambda) = \bar{C}^1_k(\mathbf{x},\lambda) + 3 + 3D_k(\mathbf{x},\lambda) + 3D_{k-1}(\mathbf{x},\lambda) + 9\mathrm{v}_4(\overline{M}(4)). \quad (79)$$

In a similar manner as above, see (70), one estimates, for any $t \in (nT, (n+1)T]$,

$$w_{1,2}(\mathcal{L}(\tilde{Y}^{\lambda}_{nT,t}(\mathbf{x}, Y^{\lambda}_{nT}(\mathbf{x}))), \mathcal{L}(Y^{\lambda}_t(\mathbf{x}))) \leq \sqrt{\frac{\lambda}{2}}\{(1-a\lambda)^{nT}\bar{C}^0(\lambda,\theta_0) + \bar{C}^1_{n+1}(\mathbf{x},\lambda) + B\} \quad (80)$$

where

$$B := \{1 + \mathbb{E}^{1/2}[V_4(Y^{\lambda}_t(\mathbf{x})] + \mathbb{E}^{1/2}[V_4(\tilde{Y}^{\lambda}_{nT,t}(\mathbf{x}, Y^{\lambda}_{nT}(\mathbf{x})))]\}^2. \quad (81)$$

Thus, due to Lemmas 3.6 and 3.9 and Corollary 3.11,

$$\mathbb{E}[V_4(\tilde{Y}^{\lambda}_{nT,t}(\mathbf{x}, Y^{\lambda}_{(k-1)T}(\mathbf{x})))] \leq 2(1-a\lambda)^{nT}\mathbb{E}|\theta_0|^4 + D_n(\mathbf{x},\lambda) + 3\mathrm{v}_4(\overline{M}(4)) \quad (82)$$

and, analogously, due to equation (40), for any $t \in (m, m+1] \subset (nT, (n+1)T]$, where m is a positive integer, the following holds

$$\begin{aligned}\mathbb{E}[V_4(Y^{\lambda}_t(\mathbf{x}))] \leq& 2 + 2(1-a\lambda(t-m))(1-a\lambda)^m\mathbb{E}|\theta_0|^4 + 2\widehat{M}(2,d)\\ &+ 2\lambda a M(2,d)\left\{|x_m|^4 + (1-a\lambda(t-m))\sum\nolimits_{j=1}^{m}(1-a\lambda)^{j-1}|x_{m-j}|^4\right\}\\ \leq& 2(1-a\lambda)^{nT}\mathbb{E}|\theta_0|^4 + D_{t,T}(\mathbf{x},\lambda), \end{aligned} \quad (83)$$

where

$$D_{t,T} := 2 + 2\lambda a M(2,d)\left\{|x_m|^4 + \sum\nolimits_{j=1}^{m}(1-a\lambda)^{j-1}|x_{m-j}|^4\right\} + 2\widehat{M}(2,d).$$

Consequently, equations (80), (81), (82) and (83), yield that

$$w_{1,2}(\mathcal{L}(\tilde{Y}^{\lambda}_{nT,t}(\mathbf{x}, Y^{\lambda}_{nT}(\mathbf{x}))), \mathcal{L}(Y^{\lambda}_t(\mathbf{x}))) \leq \sqrt{\lambda}\left\{(1-a\lambda)^{nT}\hat{C}^0(\lambda,\theta_0) + \hat{C}^1_{t,T}(\mathbf{x},\lambda)\right\}, \quad (84)$$

where

$$\hat{C}^1_{t,T}(\mathbf{x},\lambda) = \bar{C}^1_{n+1}(\mathbf{x},\lambda) + 3 + 3D_n(\mathbf{x},\lambda) + 3D_{t,T}(\mathbf{x},\lambda) + 9\mathrm{v}_4(\overline{M}(4)). \quad (85)$$

Finally, equations (68), (78) and (84) yield that

$$\begin{aligned}W_1(\mathcal{L}(\tilde{Y}^{\lambda}_t(\mathbf{x})), \mathcal{L}(Y^{\lambda}_t(\mathbf{x}))) \leq& \sqrt{\lambda}\left(C_9\sum_{k=1}^{n}\mathrm{e}^{-C_8(n-k)}\left[(1-a\lambda)^{(k-1)T}\hat{C}^0(\lambda,\theta_0) + \hat{C}^1_k(\mathbf{x},\lambda)\right]\right)\\ &+\sqrt{\lambda}\left\{(1-a\lambda)^{nT}\hat{C}^0(\lambda,\theta_0) + \hat{C}^1_{t,T}(\mathbf{x},\lambda)\right\}\\ \leq& \sqrt{\lambda}\mathrm{e}^{-\min(C_8,a)n}(n+1)C^{\sharp}\mathbb{E}[|\theta_0|^2(1+|\theta_0|^2)] + \sqrt{\lambda}C^{\flat}(\mathbf{x},n,\lambda),\end{aligned}$$

where

$$C^{\sharp} := (C_9+1)\left(\lambda_{\max}\beta K_1^4/(4a) + 12\right),$$

and

$$C^{\flat}(\mathbf{x},n,\lambda) := C_9\sum_{k=1}^{n}e^{-C_8(n-k)}\hat{C}^1_k(\mathbf{x},\lambda) + \hat{C}^1_{t,T}(\mathbf{x},\lambda). \quad (86)$$

Notice that $\mathbb{E}[|\theta_0|^2] \leq \mathbb{E}[|\theta_0|^4] + 1$. Furthermore, for each $y \geq 0$ and $\alpha > 0$, $e^{-\alpha y}(y+1) \leq 1 + 1/\alpha$. Applying this latter observation with $\alpha = \min(C_8, a)/2$ and $y = n$, it follows that

$$\begin{aligned} & \mathrm{e}^{-\min(C_8,a)n}(n+1)C^\sharp\mathbb{E}[|\theta_0|^2(1+|\theta_0|^2)] + C^\flat(\mathbf{x},n,\lambda) \\ \leq \quad & \mathrm{e}^{-\min(C_8,a)n/2}\left[1 + \frac{2}{\min(C_8,a)}\right] C^\sharp(2\mathbb{E}[|\theta_0|^4]+1) + C^\flat(\mathbf{x},n,\lambda) \end{aligned}$$

so we can set

$$C_{17} := 2\left[1 + \frac{2}{\min(C_8,a)}\right] C^\sharp, \tag{87}$$

and

$$C_{18}(\mathbf{x},n,\lambda) := C^\flat(\mathbf{x},n,\lambda) + \left[1 + \frac{2}{\min(C_8,a)}\right] C^\sharp. \tag{88}$$

□

Recall that $\mathcal{P}(\mathbb{R}^q)$ is the set of probability measures on $\mathcal{B}(\mathbb{R}^q)$ equipped with topology of weak convergence. It is known that $\mathcal{P}(\mathbb{R}^q)$ can be equipped with the structure of a complete separable metric space such that the generated topology coincides with the topology of weak convergence. Let us denote by $\mathsf{S} := (\mathbb{R}^m)^{\mathbb{N}}$ and by $\mathcal{S}$ the Borel σ-algebra associated to the product topology on S.

Lemma 3.21. *Let **H**2 and **H**1 be in force. The mappings $\tilde{\mu} : \mathbf{x} \to \mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{x}))$ and $\mu : \mathbf{x} \to \mathcal{L}(Y_t^\lambda(\mathbf{x}))$ $\mathcal{S}/\mathcal{B}(\mathcal{P}(\mathbb{R}^d))$-measurable for all $0 < \lambda$.*

Proof. Recall that if a sequence $\mathbf{x}^n \in \mathsf{S}$ converges to $\mathbf{x} \in \mathsf{S}$ in the product topology, $n \to \infty$ then $\mathbf{x}_i^n \to \mathbf{x}_i$ for each coordinate $i \in \mathbb{N}$. We show below, by induction on $j \in \mathbb{N}$ that

$$Y_t^\lambda(\mathbf{x}^n) \to Y_t^\lambda(\mathbf{x}) \tag{89}$$

for all $t \in (j, j+1]$ almost surely, $n \to \infty$. Note that (89) is trivial for $t = 0$.

Now notice that

$$Y_t^\lambda(\mathbf{x}^n) = \lambda(t-j)H(Y_j^\lambda(\mathbf{x}^n), \mathbf{x}_j^n) + \sqrt{2\lambda}[\tilde{B}_t^\lambda - \tilde{B}_j^\lambda],$$

so this tends a.s. to $Y_t^\lambda(\mathbf{x})$ as $n \to \infty$, by continuity of $H(\cdot,\cdot)$ and by the induction hypothesis. Since almost sure convergence entails convergence in law, this shows that μ is, in fact, a continuous functional of $\mathbf{x}$.

Now we turn our attention to $\tilde{\mu}$. For each $\mathbf{x} \in \mathsf{S}$, we define a recursive (Picard-type) iteration:

$$D_s^0(\mathbf{x}) := \theta_0,\ 0 \leq s \leq t,\ D_s^{k+1}(\mathbf{x}) := \theta_0 + \lambda \int_0^s H(D_u^k(\mathbf{x}), \mathbf{x}_{\lfloor u \rfloor})\, du + \sqrt{2\lambda}\tilde{B}_s^\lambda,\ k \in \mathbb{N}.$$

Define $\Phi_k(\mathbf{x}) := \mathcal{L}(D_t^k(\mathbf{x}))$, $\mathbf{x} \in \mathsf{S}$, $k \in \mathbb{N}$.

We now establish for each $k \in \mathbb{N}$ that, when $\mathbf{x}^n \to \mathbf{x}$, $n \to \infty$, we have $D_s^k(\mathbf{x}^n) \to D_s^k(\mathbf{x})$ in L^1 (hence also in law). We check by induction on k that

$$\sup_{0 \leq s \leq t} \mathbb{E}|D_s^k(\mathbf{x}^n) - D_s^k(\mathbf{x})| \to 0,$$

which is slightly more (but it is needed for the induction to work). The case $k = 0$ is trivial. Otherwise, using Lipschitz-continuity of $H(\cdot,\cdot)$, for any $s \in [0,T]$,

$$\begin{aligned} & \mathbb{E}|D_s^{k+1}(\mathbf{x}^n) - D_s^{k+1}(\mathbf{x})| \\ & \leq \lambda \int_0^s \mathbb{E}|H(D_u^k(\mathbf{x}^n), \mathbf{x}_{\lfloor u \rfloor}^n) - H(D_u^k(\mathbf{x}), \mathbf{x}_{\lfloor u \rfloor})|\, \mathrm{d}u \end{aligned}$$

$$\leq \lambda \int_0^s \left\{ \mathbb{E}|H(D_u^k(\mathbf{x}^n), \mathbf{x}^n_{\lfloor u \rfloor}) - H(D_u^k(\mathbf{x}), \mathbf{x}^n_{\lfloor u \rfloor})| + \mathbb{E}|H(D_u^k(\mathbf{x}), \mathbf{x}^n_{\lfloor u \rfloor}) - H(D_u^k(\mathbf{x}), \mathbf{x}_{\lfloor u \rfloor})| \right\} \mathrm{d}u$$
$$\leq \lambda \int_0^t \left\{ K_1 \mathbb{E}|D_u^k(\mathbf{x}^n) - D_u^k(\mathbf{x})| + K_2 \max_{0 \leq i \leq \lfloor t \rfloor} |\mathbf{x}_i^n - \mathbf{x}_i^n| \right\} \mathrm{d}u.$$

It follows that

$$\sup_{0 \leq s \leq t} \mathbb{E}|D_s^{k+1}(\mathbf{x}^n) - D_s^{k+1}(\mathbf{x})| \leq \lambda t \left\{ K_1 \sup_{0 \leq s \leq t} \mathbb{E}|D_s^k(\mathbf{x}^n) - D_s^k(\mathbf{x})| + K_2 \max_{0 \leq i \leq \lfloor t \rfloor} |\mathbf{x}_i^n - \mathbf{x}_i^n| \right\},$$

which tends to 0 as $n \to \infty$ by the induction hypothesis and the definition of the convergence in S. We deduce that, for each k, the functional $\Phi_k : \mathcal{R} \to \mathcal{P}$ is continuous on $\mathcal{R}$.

Noting $\theta_0 \in L^2$, it is well-known (see e.g. [1, Theorem 6.2.2]) that $D_t^k(\mathbf{x}) \to \tilde{Y}_t^\lambda(\mathbf{x})$, $k \to \infty$ in L^2. This implies $\Phi_k(\mathbf{x}) \to \mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{x}))$ in law, for each $\mathbf{x} \in \mathcal{R}$, which shows that the functional $\tilde{\mu}$ is measurable, being a pointwise limit of continuous functionals. The proof is complete. □

Lemma 3.22. *Let $(\mathsf{U}, \mathcal{U})$ be a measurable space and let the mappings $\mu : \mathsf{U} \to \mathcal{P}(\mathbb{R}^d)$, $\tilde{\mu} : \mathsf{U} \to \mathcal{P}(\mathbb{R}^d)$ be $\mathcal{U}/\mathcal{B}(\mathcal{P}(\mathbb{R}^d))$-measurable. Let ζ be a probability law on $\mathcal{U}$. If $W_1(\tilde{\mu}(u), \mu(u)) \leq \kappa(u)$ holds for every $u \in \mathsf{U}$ where $\kappa : \mathsf{U} \to [0,1]$ is a measurable function then*

$$W_1\left(\int_{\mathsf{U}} \tilde{\mu}(u)\, \zeta(\mathrm{d}u), \int_{\mathsf{U}} \mu(u)\, \zeta(\mathrm{d}u) \right) \leq \int_{\mathsf{U}} \kappa(u)\, \zeta(\mathrm{d}u).$$

Proof. By [36, Corollary 5.22], there is a measurable choice $u \to \pi(u)$ such that for each u, $\pi(u)$ is a W_1-optimal transference plan between $\mu(u)$ and $\tilde{\mu}(u)$. For any $A \in \mathbb{R}^d$, $\int_{\mathsf{U}} \zeta(\mathrm{d}u)\pi(u)(A \times \mathbb{R}^d) = \int_{\mathbb{R}^d} \zeta(\mathrm{d}u)\mu(u)(A)$ and $\int_{\mathsf{U}} \zeta(\mathrm{d}u)\pi(u)(\mathbb{R}^d \times A) = \int_{\mathbb{R}^d} \zeta(\mathrm{d}u)\tilde{\mu}(u)(A)$. Therefore

$$W_1\left(\int_{\mathsf{U}} \tilde{\mu}(u)\, \zeta(\mathrm{d}u), \int_{\mathsf{U}} \mu(u)\, \zeta(\mathrm{d}u) \right) \leq \int_{\mathsf{U}} \zeta(\mathrm{d}u) \int_{\mathbb{R}^{2d}} \pi(u)(\mathrm{d}x\mathrm{d}y)|x - y|.$$

The proof follows since $\int_{\mathbb{R}^{2d}} \pi(u)(\mathrm{d}x\mathrm{d}y)|x - y| = W_1(\mu(u), \tilde{\mu}(u)) \leq \kappa(u)$. □

Corollary 3.23. *For each $0 < \lambda \leq \lambda_{\max}$ and $t \in (nT, (n+1)T]$, we get*

$$W_1(\mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{X})), \mathcal{L}(Y_t^\lambda(\mathbf{X}))) \leq \lambda^{1/2}[\mathrm{e}^{-\min(C_8, a)n/2} C_{17} E[|\theta_0|^4] + C_{19}],$$

where $C_{19} := \sup_{\lambda \leq \lambda_{\max}} \sup_{n \in \mathbb{N}} E[C_{18}(\mathbf{X}, n, \lambda)] < \infty$.

Proof. Recall first that as X is conditionally L-mixing, $A := 1 + \sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|^4] < \infty$. Fix n such that $n < t \leq n + 1$. Denote by ζ the law of $\mathbf{X}$. Define

$$\tilde{\mu}(\mathbf{x}) := \mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{x})), \quad \mu(\mathbf{x}) := \mathcal{L}(Y_t^\lambda(\mathbf{x})).$$

Lemma 3.21 implies the measurability of these functionals. Let

$$\kappa(\mathbf{x}, t) := \lambda^{1/2}(\mathrm{e}^{-\min(C_8, a)n/2} C_{17} \mathbb{E}[|\theta_0|^4] + C_{18}(\mathbf{x}, n, \lambda)),$$

for each $\mathbf{x} \in \mathcal{R}$, where $C_{18}(\mathbf{x}, n, \lambda)$ is given in (88). Now the statement follows by Lemma 3.22 provided that we show $C_{19} < \infty$. By the definitions of $\hat{C}_k^1(\mathbf{x}, \lambda)$ and $\hat{C}_{t,T}^1(\mathbf{x}, \lambda)$ this boils down to showing that $\sup_{\lambda \leq \lambda_{\max}} \sup_k \mathbb{E}[S_1(\lambda, k) + S_2(\lambda, k)] < \infty$, where

$$S_1(\lambda, k) = \lambda^3 \sum_{j=(k-1)T}^{kT-1} \sum_{l=0}^{j} (1 - a\lambda)^l E|X_{j-l}|^2 + \lambda^2 \sum_{l=(k-1)T}^{kT-1} E|X_l|^2$$
$$S_2(\lambda, k) = \lambda \sum_{j=0}^{kT} (1 - a\lambda)^j\, E|X_{(k-1)T-j}|^4.$$

This is clear since

$$\mathbb{E}[S_1(\lambda,k)] \leq \lambda^3 \frac{A}{a\lambda}\frac{1}{\lambda} + \lambda^2 \frac{A}{\lambda} \leq A\lambda_{\max}\left(1+\frac{1}{a}\right),$$

and

$$\mathbb{E}[S_2(\lambda,k)] \leq \lambda \frac{A}{a\lambda} \leq \frac{A}{a}.$$

□

Lemma 3.24. *The contraction constant in Proposition 3.14 is given by*

$$C_8 = \min\{\bar{\phi}, C_6(p), 4C_7(p)\epsilon C_6(p)\}/2,$$

where the explicit expressions for $C_6(p)$ and $C_7(p)$ can be found in Lemma 3.5 and ϕ is given by

$$\bar{\phi} = \left(\sqrt{4\pi/K_1}\bar{b}\exp\left(\left(\bar{b}\sqrt{K_1}/2 + 2/\sqrt{K_1}\right)^2\right)\right)^{-1}.$$

Furthermore, any ϵ can be chosen which satisfies the following inequality

$$\epsilon \leq 1 \wedge \left(8C_7(p)\sqrt{\pi/K_1}\int_0^{\tilde{b}} \exp\left(\left(s\sqrt{K_1}/2 + 2/\sqrt{K_1}\right)^2\right) \mathrm{d}s\right)^{-1},$$

where $\tilde{b} = \sqrt{2C_7(p)/C_6(p) - 1}$, $\bar{b} = \sqrt{4C_7(p)(1+C_6(p))/C_6(p) - 1}$. The constant C_9 is given as the ratio C_{11}/C_{10}, where C_{11}, C_{10} are given explicitly in the proof below.

Proof. Consider the Lyapunov function $V_p(\theta) = (|\theta|^2+1)^{p/2}$, for any $\theta \in \mathbb{R}^d$ and $p \geq 2$. Notice that $\nabla V_p(\theta) = p\theta(|\theta|^2+1)^{p/2-1}$. As in [18], define a bounded non-decreasing function: $Q(\epsilon) : (0,\infty) \to \mathbb{R}_+$ by

$$Q(\epsilon) = \sup \frac{|\nabla V_p|}{\max\{V_p, 1/\epsilon\}}.$$

For calculating the constants we need an estimate for $Q(\epsilon)$. We consider the following three cases:

1. Consider $\epsilon \in (0, 2^{-p/2})$. For $|\theta| < \sqrt{(1/\epsilon)^{2/p} - 1}$, we have $V_p(\theta) < 1/\epsilon$, and

   $$Q(\epsilon) = \sup_{|\theta| < \sqrt{(1/\epsilon)^{2/p}-1}} \epsilon p|\theta|(|\theta|^2+1)^{p/2-1} = \epsilon^{2/p} p\sqrt{(1/\epsilon)^{2/p} - 1}.$$

   On the other hand, for $|\theta| \geq \sqrt{(1/\epsilon)^{2/p} - 1}$, $V_p(\theta) \geq 1/\epsilon$, and

   $$Q(\epsilon) = \sup \frac{p|\theta|}{|\theta|^2+1} = \epsilon^{2/p} p\sqrt{(1/\epsilon)^{2/p} - 1},$$

   since for $\epsilon \in (0, 2^{-p/2})$, $|\theta| > 1$. Therefore, $Q(\epsilon) = \epsilon^{2/p} p\sqrt{(1/\epsilon)^{2/p} - 1} \leq p/2$ for all $\epsilon \in (0, 2^{-p/2})$.

2. For the second case, consider $\epsilon \in (2^{-p/2}, 1)$. Then, by using the same arguments as above, one obtains for $|\theta| < \sqrt{(1/\epsilon)^{2/p} - 1}$, $Q(\epsilon) = \epsilon^{2/p} p\sqrt{(1/\epsilon)^{2/p} - 1}$, while for $|\theta| \geq \sqrt{(1/\epsilon)^{2/p} - 1}$, $Q(\epsilon) = p/2$. Thus, one obtains $Q(\epsilon) \leq p/2$ for all $\epsilon \in (2^{-p/2}, 1)$.

3. Finally, for $\epsilon \geq 1$, we have $Q(\epsilon) = p/2$, since $V_p(\theta) \geq 1$ for all $\theta \in \mathbb{R}^d$.

In the first two cases above, we used the fact that $p/2 \geq \epsilon^{2/p} p\sqrt{(1/\epsilon)^{2/p}-1}$ for all $\epsilon \in (0,1)$. Indeed, this is true since, squaring both sides, we have

$$1 \geq 4\epsilon^{4/p}((1/\epsilon)^{2/p}-1) \iff 4\epsilon^{4/p} - 4\epsilon^{2/p} + 1 \geq 0 \iff (2\epsilon^{2/p}-1)^2 \geq 0.$$

Combining all the three cases, one obtains $Q(\epsilon) \leq p/2$ for all $\epsilon > 0$.

In [18] a further key qunatity is $R_2 \geq 0$. We note that, by its definition in Section 2 of [18], it satisfies

$$\begin{aligned} & R_2 \leq 2\sup\{|\theta| : V_p(\theta) \leq 4C_7(p)(1+C_6(p))/C_6(p)\} \\ \implies \quad & R_2 \leq \overline{R}_2 := 2\sqrt{(4C_7(p)(1+C_6(p))/C_6(p))^{2/p}-1} \end{aligned}$$

as well as

$$R_2 \geq \underline{R}_2 := \sqrt{(4C_7(p)(1+C_6(p))/C_6(p)-1)^{2/p}-1}.$$

We now check the requirements of [18, Theorem 2.2] for ϵ. It is required that

$$(4C_7(p)\epsilon)^{-1} \geq \int_0^{R_1}\int_0^s \exp\left(\frac{1}{2}\int_r^s u\kappa(u)\,du + 2Q(\epsilon)(s-r)\right)\,dr\,ds,$$

where in our case $\kappa(u) = K_1$ and $Q(\epsilon) \leq p/2$. Hence ϵ is suitable whenever

$$\begin{aligned} (4C_7(p)\epsilon)^{-1} &\geq \int_0^{R_1}\int_0^s \exp\left(\frac{1}{2}\int_r^s K_1 u\,du + p(s-r)\right)\,dr\,ds \\ &= \int_0^{R_1}\int_0^s \exp\left(\frac{K_1}{4}(s^2-r^2) + p(s-r)\right)\,dr\,ds \\ &= \int_0^{R_1} \exp\left(\left(\frac{\sqrt{K_1}}{2}s + \frac{p}{\sqrt{K_1}}\right)^2\right)\int_0^s \exp\left(-\left(\frac{\sqrt{K_1}}{2}r + \frac{p}{\sqrt{K_1}}\right)^2\right)\,dr\,ds, \end{aligned}$$

which implies by setting $v/\sqrt{2} = \sqrt{K_1}r/2 + p/\sqrt{K_1}$, ($dv = \sqrt{K_1/2}dr$)

$$\begin{aligned} (4C_7(p)\epsilon)^{-1} &\geq \sqrt{\frac{2}{K_1}}\int_0^{R_1} \exp\left(\left(\frac{\sqrt{K_1}}{2}s + \frac{p}{\sqrt{K_1}}\right)^2\right)\int_{p\sqrt{2/K_1}}^{\sqrt{K_1/2}s + p\sqrt{2/K_1}} \exp\left(-\frac{v^2}{2}\right)\,dv\,ds \\ &= \sqrt{\frac{4\pi}{K_1}}\int_0^{\tilde{b}} \exp\left(\left(\frac{\sqrt{K_1}}{2}s + \frac{p}{\sqrt{K_1}}\right)^2\right)\left(\Phi\left(\sqrt{K_1/2}s + p\sqrt{2/K_1}\right) - \Phi\left(p\sqrt{2/K_1}\right)\right)\,ds, \end{aligned}$$

where $\tilde{b} = \sqrt{(2C_7(p)/C_6(p))^{2/p}-1} > 0$ and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

The inrements of a cumulative distribution function can be at most one. To ease the calculations of C_{10} and C_{11} below, it is thus enough for ϵ to satisfy the following inequality:

$$\epsilon \leq 1 \wedge \left(8C_7(p)\sqrt{\frac{\pi}{K_1}}\int_0^{\tilde{b}} \exp\left(\left(\frac{\sqrt{K_1}}{2}s + \frac{p}{\sqrt{K_1}}\right)^2\right)\,\mathrm{d}s\right)^{-1}.$$

In [18] a further key quantity is β, which we denote by ϕ in order to avoid a clash of notation. We calculate ϕ using its definition in Theorem 2.2 of [18], noting that $Q(\epsilon) \leq p/2$,

$$\phi^{-1} = \int_0^{R_2}\int_0^s \exp\left(\frac{1}{2}\int_r^s K_1 u\,\mathrm{d}u + 2Q(\epsilon)(s-r)\right)\,\mathrm{d}r\,\mathrm{d}s$$

$$\leq \int_0^{R_2} \int_0^s \exp\left(\frac{1}{2}\int_r^s K_1 u\,\mathrm{d}u + p(s-r)\right)\,\mathrm{d}r\,\mathrm{d}s$$
$$= \sqrt{\frac{4\pi}{K_1}}\int_0^{\bar{b}} \exp\left(\left(\frac{\sqrt{K_1}}{2}s + \frac{p}{\sqrt{K_1}}\right)^2\right)\left(\Phi\left(\sqrt{K_1/2}s + p\sqrt{2/K_1}\right) - \Phi\left(p\sqrt{2/K_1}\right)\right)\,ds,$$

where $\bar{b} = \sqrt{(4C_7(p)(1+C_6(p))/C_6(p))^{2/p} - 1} > 0$. One notices that

$$\phi \geq \bar{\phi} = \left(\sqrt{\frac{4\pi}{K_1}}\bar{b}\exp\left(\left(\frac{\sqrt{K_1}}{2}\bar{b} + \frac{2}{\sqrt{K_1}}\right)^2\right)\right)^{-1}$$

hence Theorem 2.2 of [18] implies that we can choose

$$C_8 = \bar{C}_8 := \min\{\bar{\phi}, C_6(p), 4C_7(p)\epsilon C_6(p)\}/2.$$

As for the calculations of C_{10} and C_{11} in (63), recall the definition of $\mathcal{W}_{\rho_2}$ in (62). Moreover, recall that f, F and R_2 are given in [18, Section 5] and satisfy $\frac{1}{2}F(r) \leq f(r) \leq F(r)$ for $r \leq R_2$ and $f(r) = f(R_2)$ for $r \geq R_2$. In addition, $r\exp\left(-K_1R_2^2/4 - pR_2\right) \leq F(r) \leq r$ for all $r \leq R_2$ and $f(r) \leq R_2$ for all $r > 0$.

With these tools at hand, we take θ, θ' such that $|\theta - \theta'| = r \leq R_2$ and estimate

$$\begin{aligned}
&[1 \wedge |\theta - \theta'|](1 + V_2(\theta) + V_2(\theta')) \\
&\leq \epsilon^{-1}|\theta - \theta'|(\epsilon + \epsilon V_2(\theta) + \epsilon V_2(\theta')) \\
&\leq 2\epsilon^{-1}\exp\left(K_1R_2^2/4 + pR_2\right)\left(\frac{1}{2}F(|\theta - \theta'|)\right)(1 + \epsilon V_2(\theta) + \epsilon V_2(\theta')) \\
&\leq \bar{C}_{10}^{-1} f(|\theta - \theta'|)(1 + \epsilon V_2(\theta) + \epsilon V_2(\theta')),
\end{aligned}$$

where $\bar{C}_{10} = \frac{\epsilon}{2}\exp\left(-K_1\overline{R}_2^2/4 - p\overline{R}_2\right)$. For $r > R_2$ we get

$$\begin{aligned}
f(|\theta - \theta'|)(1 + \epsilon V_2(\theta) + \epsilon V_2(\theta')) &= f(R_2)(1 + \epsilon V_2(\theta) + \epsilon V_2(\theta')) \\
&\geq \tilde{C}_{10}[1 \wedge |\theta - \theta'|](1 + V_2(\theta) + V_2(\theta')),
\end{aligned}$$

where $\tilde{C}_{10} = \frac{\epsilon}{2}\underline{R}_2\exp\left(-K_1\overline{R}_2^2/4 - p\overline{R}_2\right)$. We can thus take $C_{10} = \min\{\bar{C}_{10}, \tilde{C}_{10}\}$.

To calculate C_{11}, one considers, for $|\theta - \theta'| = r \leq R_2$

$$\begin{aligned}
f(|\theta - \theta'|)(1 + \epsilon V_2(\theta) + \epsilon V_2(\theta')) &\leq |\theta - \theta'|(1 + \epsilon V_2(\theta) + \epsilon V_2(\theta')) \\
&\leq C_{11}[1 \wedge |\theta - \theta'|](1 + V_2(\theta) + V_2(\theta')),
\end{aligned}$$

where $C_{11} = 1 + R_2$. In the case where $r > R_2$, we also get

$$\begin{aligned}
f(|\theta - \theta'|)(1 + \epsilon V_2(\theta) + \epsilon V_2(\theta')) &= f(R_2)(1 + \epsilon V_2(\theta) + \epsilon V_2(\theta')) \\
&\leq C_{11}[1 \wedge |\theta - \theta'|](1 + V_2(\theta) + V_2(\theta')),
\end{aligned}$$

hence the choice $C_{11} = 1 + R_2$ is indeed fine. □

3.8 Proof of Main Result

Proof of Theorem 2.3. Trivially, $\mathbb{E}^{3/4}[V_4(\theta_0)] \leq 1 + \mathbb{E}[V_4(\theta_0)]$ and $\mathbb{E}[V_4(\theta_0)] \leq 2 + 2\mathbb{E}[|\theta_0|^4]$. We estimate, for $kT \leq t \leq (k+1)T$,

$$W_1(\mathcal{L}(Y_t^{\lambda}(\mathbf{X}), \pi_\beta)$$

$$
\begin{aligned}
&\leq W_1(\mathcal{L}(Y_t^\lambda(\mathbf{X}), \mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{X})) + W_1(\mathcal{L}(\tilde{Y}_t^\lambda(\mathbf{X})), \mathcal{L}(L_t^\lambda)) + W_1(\mathcal{L}(L_t^\lambda), \pi_\beta) \\
&\leq \lambda^{1/2}\left[\mathrm{e}^{-\min\{C_8, a/4\}k/2}[C_{19}\mathbb{E}[V_4(\theta_0)] + C_{15}\mathbb{E}^{3/4}[V_4(\theta_0)]] + C_{17} + C_{15}\right] + C_9\mathrm{e}^{-C_8\lambda t}w_{1,2}(\theta_0, \pi_\beta) \\
&\leq \mathrm{e}^{-\min\{C_8, a/4\}k/2}\lambda^{1/2}(C_{19} + C_{15})[\mathbb{E}[V_4(\theta_0)] + 1] \\
&+ (C_{17} + C_{15})\sqrt{\lambda} + C_9\mathrm{e}^{-\min\{C_8, a/4\}k/2}[1 + E[V_2(\theta_0)] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(\mathrm{d}\theta)] \\
&\leq \mathrm{e}^{-\min\{C_8, a/4\}k/2}\lambda^{1/2}(C_{19} + C_{15})[2\mathbb{E}[|\theta_0|^4] + 3] \\
&+ (C_{17} + C_{15})\sqrt{\lambda} + C_9\mathrm{e}^{-\min\{C_8, a/4\}k/2}[2 + E[V_4(\theta_0)] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(\mathrm{d}\theta)] \\
&\leq \mathrm{e}^{-\min\{C_8, a/4\}k/2}\lambda_{\max}^{1/2}(C_{19} + C_{15})[2\mathbb{E}[|\theta_0|^4] + 3] \\
&+ (C_{17} + C_{15})\sqrt{\lambda} + C_9\mathrm{e}^{-\min\{C_8, a/4\}k/2}[4 + 2\mathbb{E}[|\theta_0|^4] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(\mathrm{d}\theta)] \\
&\leq \mathrm{e}^{-\min\{C_8, a/4\}k/2}[2\lambda_{\max}^{1/2}(C_{19} + C_{15}) + 2C_9]E[|\theta_0|^4] \\
&+ \mathrm{e}^{-\min\{C_8, a/4\}k/2}[3(C_{19} + C_{15})\lambda_{\max}^{1/2} + 4C_9 + C_9\int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(\mathrm{d}\theta)] \\
&+ (C_{17} + C_{15})\sqrt{\lambda},
\end{aligned}
$$

by Corollary 3.23, Corollary 3.19, Proposition 3.14 and by (22). Noting (28) and $\lfloor n\lambda \rfloor \lfloor 1/\lambda \rfloor \leq n$, this implies, for all $n \in \mathbb{N}$,

$$
W_1(\mathcal{L}(\theta_n^\lambda, \pi_\beta) \leq \mathrm{e}^{-C_0\lfloor n\lambda \rfloor}\bar{C}_1[1 + \mathbb{E}[|\theta_0|^4]] + C_2\sqrt{\lambda},
$$

where $C_0 = \min\{C_8, a/4\}/2$,

$$
\bar{C}_1 = \lambda_{\max}^{1/2}(C_{19} + C_{15}) + 2C_9 + 3\lambda_{\max}^{1/2}(C_{19} + C_{15}) + 4C_9 + C_9\int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(\mathrm{d}\theta)
$$

and $C_2 = C_{17} + C_{15}$. We can thus set $C_1 := e^{C_0}\bar{C}_1$ and conclude. □

Remark 3.25. The proof of Lemma 3.24 shows that C_9 has a rather poor (exponential) dependence on the dimension d, see the definitions of C_{10} and $\overline{R}_2$ therein as well as the definition of $C_7(p)$ in Lemma 3.5. Improvements on the dimension dependence here would require enhancing the coupling arguments of [18, Corollary 2.3] significantly.

4 Applications to non-convex optimization

Example 4.1. Let $Z_n \in \mathbb{R}^m$, $n \in \mathbb{Z}$ be a (strict-sense) stationary sequence. Let us consider the problem of online nonlinear prediction of Z_n as a function of the p previous observations $Z_{n-1}, \ldots, Z_{n-p}$. We use a predictor of the form $\hat{Z}_n(\theta) = f_\theta(Z_{n-1}, \ldots, Z_{n-p})$, where $f_\theta : \mathbb{R}^{p\times m} \to \mathbb{R}^m$, $\theta \in \mathbb{R}^d$ is a parametric family of (non-linear) twice continuously differentiable functions, such as the output of a neural network. We seek to minimize the regularized mean-square error, that is,

$$
U(\theta) = \mathbb{E}[|Z_p - f_\theta(Z_{p-1}, \ldots, Z_0)|^2] + c|\theta|^2 \tag{90}
$$

for some $c > 0$. Here

$$
H^i(\theta, z) = 2\left\langle z^p - f_\theta(z^{p-1}, \ldots, z^0), \frac{\partial f_\theta(z^{p-1}, \ldots, z^0)}{\partial \theta^i}\right\rangle_{\mathbb{R}^m} + 2c\theta^i
$$

for each $i = 1, \ldots, d$. Let Z be conditionally L-mixing. If we assume that f_θ, $\partial_\theta f_\theta$, $\partial_{\theta\theta} f_\theta$ as well as Z_n are all bounded and $z \to f_\theta(z)$ and $z \to \partial_\theta f_\theta(z)$ are Lipschitz, then the assumptions of our

paper hold, as easily checked. The SGLD then provides an algorithm to optimize the prediction procedure.

Let us now apply this framework to online price prediction, a procedure of paramount importance for econometric analysis and algorithmic trading (see e.g. the paper [23] which surveys 27 methods, including several neural network-based approaches).

Denote by $Z_n \in \mathbb{R}^m$ the return vector on m assets at time t. While stationarity of the process Z holds on appropriate time scales (see Subsection 3.1 of [8]), independence badly fails (see Section 5 of [8] and the references therein). Conditional L-mixing holds e.g. when Z is a (possibly non-linear) Lipschitz functional of an infinite moving average processes, see [5]. Another example for conditional L-mixing is rough volatility models, see [19, 5]. Our SGLD algorithm with dependent data can then be used to find the optimizer of (90).

Financial applications provide a rich source of problems where stochastic approximation needs to be used in settings with dependent data: optimal posting of orders, optimal split of orders, etc. Here we do not enter into more details, see [26, 24, 25].

Example 4.2. We sketch a general optimization framework here. It is often the case that a deterministic function we wish to minimize has some representation as the expectation of a functional of a random variable. It is also often clear that the optimizer necessarily lies in some (big) compact set $\mathrm{B}_{R''}$ where R'' can be estimated.

Let $R > 0$ and let $\overline{U} : \mathrm{B_R} \to \mathbb{R}_+$ be a possibly non-convex function. We assume that it admits a stochastic representation $\overline{U}(\theta) = E[\overline{u}(\theta, X)]$, $\theta \in \mathrm{B}_R$ where $\overline{u} : \mathrm{B}_{R'} \times \mathbb{R}^m \to \mathbb{R}_+$ is continuous and continuously differentiable in $\operatorname{int} \mathrm{B}_{R'}$ for some $R' > R$ and $\frac{\partial}{\partial\theta}\overline{u}(\theta, x)$ is jointly Lipschitz-continuous in $(\theta, x) \in \mathrm{B}_R \times \mathbb{R}^m$. X is a $\mathbb{R}^m$-valued random variable which we assume bounded, for simplicity. We assume that $\overline{U}$ has a unique minimizer $\theta^* \in \operatorname{int} \mathrm{B}_{R''}$ with some $R'' < R$. We may always assume $u^* := \overline{U}(\theta^*) \geq 0$ (by adding a suitably large constant).

We propose an approach to find θ^* using SGLD. The case of multiple global minimizers can be handled similarly. Since $\theta^* \in \operatorname{int} \mathrm{B}_{R''}$, one deduces $\inf_{|\theta|=R''} \overline{U}(\theta) \geq u^* + \kappa$ for some $\kappa > 0$. Then, continuity implies that for some $\delta > 1$ which is close enough to 1,

$$\inf_{\lambda\in[1,\delta^2],|\theta|=R''} E\left[\overline{u}(\theta\sqrt{\lambda}, X)\left(1 - \frac{\lambda}{\delta^2}\right)\right] + |\theta\sqrt{\lambda}|^2\frac{\lambda}{\delta^2} \geq u^* + \kappa/2.$$

One reasonably assumes that $R''\delta < R$ and proceeds by defining, for $\lambda \in [1, \delta^2]$ and for all θ with $|\theta| = R''$, $u(\theta\sqrt{\lambda}, x) := \overline{u}(\theta\sqrt{\lambda}, X)\left(1 - \frac{\lambda}{\delta^2}\right) + |\theta\sqrt{\lambda}|^2\frac{\lambda}{\delta^2}$. An alternative way of writing this is

$$u(\theta, x) = \overline{u}(\theta, X)\left(1 - \frac{|\theta|^2}{\delta^2(R'')^2}\right) + |\theta|^2\frac{|\theta|^2}{\delta^2(R'')^2} \tag{91}$$

for θ with $R'' \leq |\theta| \leq R''\delta$. Furthermore, for $\theta \in \operatorname{int} \mathrm{B}_{R''}$ define $u(\theta, x) := \overline{u}(\theta, x)$, $x \in \mathbb{R}^m$ and for $\theta \notin \mathrm{B}_{R''\delta}$ set $u(\theta, x) := |\theta|^2$. Define also $U(\theta) := E[u(\theta, X)]$, $\theta \in \mathbb{R}^d$.

It is claimed that θ^* is also the unique minimizer of $U(\theta)$. To show this, it suffices to demonstrate that $U(\theta) > u^*$ holds for all θ with $|\theta| \geq R''$. This holds for $R'' \leq \theta \leq R''\delta$ by the choice of δ and it is trivial for $|\theta| \geq R''\delta$ since U is monotone in $|\theta|$ over that set.

It is not difficult to see, using (91), that $u(\theta, x)$ is continuously differentiable and $H(\theta, x) := \frac{\partial}{\partial\theta}u(\theta, x)$ is jointly Lipschitz-continuous. The dissipativity condition is trivial for H as it is obvious for $\theta \to |\theta|^2$ and H coincides with the latter function outside a compact set. Let X_n, $n \geq 1$ be a stationary sequence with common law equal to that of X, satisfying Assumption 3. One then implements the SGLD algorithm for β large, λ small. For n large enough, θ_n^λ is a good approximate sample from π_β and hence, by maximality of θ^*, a good estimate for θ^* (where goodness of the estimate is quantified by our main result, Theorem 2.3).

Example 4.3. A deep neural network with weight constraints falls under the scope of the present section. We denote by d_1 the number of hidden layers, let d_2 be the number of nodes in each layer.

The parameter space is $\mathbb{R}^d$, where $d := d_1 \times d_2 \times d_2$. Elements $\mathbf{w} \in \mathbb{R}^d$ are weight matrices where $\mathbf{w} = [w_{k,j,l}]$ and the indices' ranges are $k = 0, \ldots, d_1 - 1$, $j, l = 1, \ldots, d_2$.

The (training) data sequence X_n, $n \geq 1$ is m-dimensional bounded and stationary, satisfying Assumption 3. For simplicity we assume $m = d_2 + 1$.

A generic element of $\mathbb{R}^m$ is denoted by $\mathbf{x} := (x_1, \ldots, x_m)$. We fix an activation function $\alpha : \mathbb{R} \to \mathbb{R}$ that is twice continuously differentiable.

A possible specification is

$$\alpha(u) := \frac{1}{1 + e^{-u}} + 1, \ u \in \mathbb{R},$$

the well-known sigmoid function.

We recursively define a doubly indexed sequence of functions $f_{k,l} : \mathbb{R}_+^d \times \mathbb{R}^m \to \mathbb{R}$ by

$$f_{0,l}(\mathbf{w}, \mathbf{x}) := x_l, \ 1 \leq l \leq d_2, \ (\mathbf{w}, \mathbf{x}) \in \mathbb{R}^{d+m},$$

and

$$f_{k+1,l}(\mathbf{w}, \mathbf{x}) := \alpha \left(\sum_{j=1}^{d_2} w_{k,j,l} f_{k,j}(\mathbf{w}, \mathbf{x}) \right), \ 1 \leq l \leq d_2,$$

for $k := 0, \ldots, d_1 - 1$. We set

$$\overline{u}(\mathbf{w}, \mathbf{x}) := \left(\frac{1}{d_2} \sum_{l=1}^{d_2} f_{d_1,l} - x_{d_2+1} \right)^2$$

and we aim at minimizing $E[F(\mathbf{w}, X_0)]$ in the parameter $\mathbf{w}$.

We imagine that the last coordinate of the X_n are some (noisy) functionals of their first d_2 coordinates and we try to use a neural network, characterized by the weights $\mathbf{w}$, that mimics this functional relationship in the best possible way. This would amount to minimizing $E[\overline{u}(\mathbf{w}, X_0)]$ in $\mathbf{w}$.

It is a standard technique against overfitting to set a maximum for the norm of $\mathbf{w}$ when optimizing. This means maximizing over B_R for some R. By induction, it is easy to show that $f_{k,l}$ are twice continuously differentiable in $\mathbf{w}, \mathbf{x}$ and so is $\overline{u}$. This implies joint Lipschitz-continuity of $\frac{\partial}{\partial \mathbf{w}} \overline{u}$ on the compact set B_R, for each R. It follows that we can apply the optimization procedure outlined above and obtain reassuring theoretical guarantees for the convergence of SGLD iterates for deep neural networks with weight constraints.

Example 4.4. For a given input vector $x \in \mathbb{R}^{m_1}$, an autoencoder aims to learn a cost-effective representation of x. Consider the following neural network:

$$\widehat{x} = \sigma_2(g_2(W_2)\sigma_1(g_1(W_1)x + b_1) + b_2),$$

where $W_1 \in \mathbb{R}^{d_1 \times m_1}, W_2 \in \mathbb{R}^{m_1 \times d_1}$ are the weights, $b_1 \in \mathbb{R}^{d_1}, b_2 \in \mathbb{R}^{m_1}$ are the biases, $\sigma_1 : \mathbb{R}^{d_1} \to \mathbb{R}^{d_1}, \sigma_2 : \mathbb{R}^{m_1} \to \mathbb{R}^{m_1}$ are the elementwise activation functions, and moreover, the (i_1, j_1)-th element of $g_1 : \mathbb{R}^{d_1 \times m_1} \to \mathbb{R}^{d_1 \times m_1}$ is given by $g_1^{(i_1,j_1)}(W_1) = C_1 \tanh(W_1^{(i_1,j_1)}/C_1)$ whereas the (i_2, j_2)-th element of $g_2 : \mathbb{R}^{m_1 \times d_1} \to \mathbb{R}^{m_1 \times d_1}$ is given by $g_2^{(i_2,j_2)}(W_2) = C_2 \tanh(W_2^{(i_2,j_2)}/C_2)$ with $C_1, C_2 > 0$. Here, we assume that the weight parameters are bounded, which is reasonable from both practical and analytical viewpoints (see [34] and references therein). To achieve this, we apply g_1, g_2 to the weight matrices W_1, W_2, then, each element of $g(W_1), g(W_2)$ are bounded by C_1, C_2, respectively. We consider $g(z) = c \tanh(z/c)$, $z \in \mathbb{R}$, $c > 0$ for the weight transformation as it is bounded, while for fixed $\epsilon > 0$, $|g(z) - z| < \epsilon$ for $z \in (-I(c), I(c))$ with $0 < I(c) < c$. Moreover, it is continuously differentiable with bounded first derivative and monotonically increasing on $\mathbb{R}$. Denote by $[W_1], [W_2]$ the vectors of all elements in W_1, W_2 respectively, and denote

by $\theta = ([W_1], [W_2], b_1, b_2) \in \mathbb{R}^d$ with $d = 2d_1m_1 + d_1 + m_1$. We aim to minimize the regularized objective function:

$$\min_\theta U(\theta) = \min_\theta \left(\mathbb{E}[|X - \widehat{X}|^2] + c|\theta|^2 \right),$$

where $c > 0$. Autoencoders can be applied to extract the market implied features, see [14, 21]. A denoising autoencoder (DAE) with masking noise (see Subsection 3.3 of [37]) is considered in [22], which is used to learn the features of the missing yield parameters from the bond yields in a chosen surrogate liquid market, and thus to obtain missing bond yields in illiquid market. For the DAE algorithm, a noisy version $\widetilde{x} \in \mathbb{R}^{m_1}$ of the input vector x is used in the neural network, which is obtained by applying the corruption process $q_D(\widetilde{x}|x)$. The objective function of the DAE algorithm becomes

$$\min_\theta \overline{U}(\theta) = \min_\theta \left(\mathbb{E}[|X - \overline{X}|^2] + c|\theta|^2 \right),$$

where $\overline{x} = \sigma_2(g_2(W_2)\sigma_1(g_1(W_1)\widetilde{x} + b_1) + b_2)$.

In the online setting, the sequence of the (dependent) input vectors $\{x_n\}_{n\in\mathbb{Z}}$, which are bond yields in the liquid market, can be generated using the Nelson– Siegel model (see [31], [13] and [30]). Moreover, the input datapoints are scaled with the maximum yield equal to one, and the corrupted version of each input $\{\widetilde{x}_n\}_{n\in\mathbb{Z}}$ is generated according to the distribution $\widetilde{x}_i \sim q_D(\widetilde{x}_i|x_i)$ for each $i \in \mathbb{Z}$ before feeding into the neural network. The activation functions σ_1, σ_2 are set to be elementwise sigmoid function. Denote by $H : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ the stochastic gradient given by

$$\begin{aligned} H(\theta, z) = (&H_{W_1^{(1,1)}}(\theta, z), \dots, H_{W_1^{(d_1,m_1)}}(\theta, z), H_{W_2^{(1,1)}}(\theta, z), \dots, H_{W_2^{(m_1,d_1)}}(\theta, z), \\ &H_{b_1^{(1)}}(\theta, z), \dots, H_{b_1^{(d_1)}}(\theta, z), H_{b_2^{(1)}}(\theta, z), \dots, H_{b_2^{(m_1)}}(\theta, z)), \end{aligned}$$

where $z = (x, \widetilde{x}) \in \mathbb{R}^m$ with $m = 2m_1$. Then, one obtains, for any $i_1 = 1, \dots, d_1$, $j_1 = 1, \dots, m_1$,

$$\begin{aligned} &H_{W_1^{(i_1,j_1)}}(\theta, z) \\ &= 2cW_1^{(i_1,j_1)} - 2\sum_{k=1}^{m}(x^{(k)} - \overline{x}^{(k)})\partial_{W_1^{(i_1,j_1)}}\sigma_2^{(k)}(g_2^{(k,\cdot)}(W_2)\sigma_1(g_1(W_1)\widetilde{x} + b_1) + b_2^{(k)}) \\ &\quad \times g_2^{(k,i_1)}(W_2)\partial_{W_1^{(i_1,j_1)}}\sigma_1^{(i_1)}(g_1(W_1)^{(i_1,\cdot)}\widetilde{x} + b_1^{(i_1)})\partial_{W_1^{(i_1,j_1)}}g_1^{(i_1,j_1)}(W_1)\widetilde{x}^{(j_1)}. \end{aligned}$$

One can check that all of the assumptions hold for $H_{W_1^{(i_1,j_1)}}(\theta, z)$. Similarly, the assumptions hold for $H_{W_2^{(i_2,j_2)}}(\theta, z), H_{b_1^{(i_3)}}(\theta, z), H_{b_2^{(i_4)}}(\theta, z)$ with $i_2, i_4 = 1, \dots, m_1$, $j_2, i_3 = 1, \dots, d_1$.

A Auxiliary results

We present a simpler version of [27, Theorem 7.19], which is suitable for the purposes of this article.

Lemma A.1. *Let $(\xi_t)_{t\geq 0}$ and $(\eta_t)_{t\geq 0}$ be two diffusion type processes with*

$$\mathrm{d}\xi_t = a_t(\xi)\mathrm{d}t + \sigma\mathrm{d}B_t, \qquad \textit{for } t > 0, \tag{92}$$

and

$$\mathrm{d}\eta_t = b_t(\eta)\mathrm{d}t + \sigma\mathrm{d}B_t \qquad \textit{for } t > 0, \tag{93}$$

where $\xi_0 = \eta_0$ is an $\mathcal{F}_0$ measurable random variable and $\sigma*$ is a positive constant. Suppose also that the nonanticipative functionals $(a_t)_{t\geq 0}$ and $(b_t)_{t\geq 0}$ are such that a unique (continuous) strong solution exist for (92) *and* (93) *respectively. If, for any fixed $T > 0$,*

$$\int_0^T [|a_s(\xi)|^2 + |b_s(\xi)|^2]\mathrm{d}s < \infty \textit{ (a.s.) and } \int_0^T [|a_s(\eta)|^2 + |b_s(\eta)|^2]\mathrm{d}s < \infty \textit{ (a.s.),}$$

then $\mu_\xi^T = \mathcal{L}(\xi_{[0,T]}) \backsim \mu_\eta^T = \mathcal{L}(\eta_{[0,T]})$ *and the densities are given by*

$$\frac{\mathrm{d}\mu_\eta^T}{\mathrm{d}\mu_\xi^T}(\xi) = \exp\left(-\sigma^{-2}\int_0^T \langle a_s(\xi) - b_s(\xi), \mathrm{d}\xi_s\rangle + \frac{1}{2\sigma^2}\int_0^T [|a_s(\xi)|^2 - |b_s(\xi)|^2]\mathrm{d}s\right) \tag{94}$$

and

$$\frac{\mathrm{d}\mu_\xi^T}{\mathrm{d}\mu_\eta^T}(\eta) = \exp\left(\sigma^{-2}\int_0^T \langle a_s(\eta) - b_s(\eta), \mathrm{d}\eta_s\rangle - \frac{1}{2\sigma^2}\int_0^T [|a_s(\eta)|^2 - |b_s(\eta)|^2]\mathrm{d}s\right). \tag{95}$$

Finally, the Kullback-Leibler divergence is given by

$$\mathrm{KL}(\mu_\xi^T, \mu_\eta^T) = \frac{1}{2\sigma^2}\mathbb{E}\left[\int_0^T |a_s(\xi) - b_s(\xi)|^2\mathrm{d}s\right]. \tag{96}$$

Proof. The proof follows from a straightforward extension of [27, Theorem 7.19] to the multidimensional case. The computation of the Kullback-Leibler distance is a direct application of the definition. □

Let $V : \mathbb{R}^d \to [1, \infty)$ be a measurable function. For a measurable function $f : \mathbb{R}^d \to \mathbb{R}$, the V-norm of f is given by $\|f\|_V = \sup_{x\in\mathbb{R}^d} |f(x)|/V(x)$. For ξ and ξ' two probability measures on $\mathbb{R}^d$, the V-total variation distance of ξ and ξ' is given by

$$\|\xi - \xi'\|_V = \sup_{\|f\|_V\le 1}\int_{\mathbb{R}^d} f(\theta)\mathrm{d}\{\xi - \xi'\}(\theta).$$

If $V \equiv 1$, then $\|\cdot\|_V$ is the total variation distance. The V-total variation distance is also characterized in terms of coupling (see [15, Theorem 19.1.7]):

$$\|\xi - \xi'\|_V = \inf_{\zeta\in\mathcal{C}(\xi,\xi')}\iint_{\mathbb{R}^d\times\mathbb{R}^d}\{V(\theta) + V(\theta')\}\mathbb{1}_{\{\theta\neq\theta'\}}\zeta(\mathrm{d}\theta, \mathrm{d}\theta')$$

where $\mathcal{C}(\xi, \xi')$ is the set of coupling of ξ and ξ'. An optimal coupling is given by (see [15, Theorem 19.1.6])

$$\gamma^*(B) = \{1 - \xi\wedge\xi'(\mathbb{R}^d)\}\beta(B) + \int_B \xi\wedge\xi'(\mathrm{d}\theta)\delta_\theta(\mathrm{d}\theta')$$

where $\xi\wedge\xi'$ is the infimum of probability measures ξ and ξ' and β is any coupling of η and η' where

$$\eta = \frac{\xi - \xi\wedge\xi'}{1 - \xi\wedge\xi'(\mathbb{R}^d)} \quad \text{and} \quad \eta' = \frac{\xi - \xi\wedge\xi'}{1 - \xi\wedge\xi'(\mathbb{R}^d)}$$

Lemma A.2. *For any probability measures* ξ *and* ξ' *on* $\mathbb{R}^d$*, and* $p \ge 1$*, we get*

$$w_{1,p}(\xi, \xi') \le \sqrt{2}\left\{1 + [\xi(V_{2p})]^{1/2} + [\xi'(V_{2p})]^{1/2}\right\}\{\mathrm{KL}(\xi, \xi')\}^{1/2}.$$

Proof.

$$\begin{aligned}
w_{1,p}(\xi, \xi') &= \inf_{\zeta\in\mathcal{C}(\xi,\xi')}\iint_{\mathbb{R}^{2d}}(1\wedge|\theta - \theta'|)\{1 + V_p(\theta) + V_p(\theta')\}\zeta(\mathrm{d}\theta\mathrm{d}\theta') \\
&\le \iint_{\mathbb{R}^{2d}}(1\wedge|\theta - \theta'|)\{1 + V_p(\theta) + V_p(\theta')\}\gamma^*(\mathrm{d}\theta\mathrm{d}\theta') \\
&\le \{1 - \xi\wedge\xi'(\mathbb{R}^d)\}\iint_{\mathbb{R}^{2d}}\{1 + V_p(\theta) + V_p(\theta')\}\beta(\mathrm{d}\theta\mathrm{d}\theta') \\
&= \|\xi - \xi'\|_{\mathrm{TV}} + \|\xi - \xi'\|_{V_p}.
\end{aligned}$$

The proof then follows from the weighted Pinsker's inequality; see [17, Lemma 24]. □

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

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

Proof. Note that

$$\sum_{\substack{i+j+k=p \\ \{i\neq p-1\}\cap\{j\neq 1\}}} \frac{p!}{i!j!k!}\|x\|^{2i}\big(2\langle x,y\rangle\big)^j\|y\|^{2k} \quad \leq \sum_{\substack{i+j+k=p \\ \{i\neq p-1\}\cap\{j\neq 1\}}} \frac{p!}{i!j!k!}\|x\|^{2i}\big(2\|x\|\|y\|\big)^j\|y\|^{2k}. \tag{97}$$

Moreover,

$$\begin{aligned}\sum_{k=0}^{2p}\binom{2p}{k}\|x\|^{2p-k}\|y\|^k =&(\|x\|+\|y\|)^{2p} = (\|x\|^2+2\|x\|\|y\|+\|y\|^2)^p\\ &= \sum_{i+j+k=p}\frac{p!}{i!j!k!}\|x\|^{2i}\big(2\|x\|\|y\|\big)^j\|y\|^{2k}.\end{aligned}$$

Consequently,

$$\sum_{\substack{k=0 \\ k\neq 1}}^{2p}\binom{2p}{k}\|x\|^{2p-k}\|y\|^k = \sum_{\substack{i+j+k=p \\ \{i\neq p-1\}\cap\{j\neq 1\}}}\frac{p!}{i!j!k!}\|x\|^{2i}\big(2\|x\|\|y\|\big)^j\|y\|^{2k}. \tag{98}$$

Thus, in view of (97) and (98), the desired result is obtained. □

References

[1] Ludwig Arnold. *Stochastic differential equations: theory and applications*. Wiley-Interscience [John Wiley & Sons], New York-London-Sydney, 1974. Translated from the German.

[2] M Barkhagen, NH Chau, É Moulines, M Rásonyi, S Sabanis, and Y Zhang. On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. *Bernoulli*, 27:1–33, 2021.

[3] N. Brosse, A. Durmus, É. Moulines, and S. Sabanis. The tamed unadjusted Langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663, 2019.

[4] Huy N. Chau, Chaman Kumar, Miklós Rásonyi, and Sotirios Sabanis. On fixed gain recursive estimators with discontinuity in the parameters. *ESAIM Probab. Stat.*, 23:217–244, 2019.

[5] N. H. Chau, Ch. Kumar, M. Rásonyi, and S. Sabanis. On fixed gain recursive estimators with discontinuity in the parameters. *ESAIM: Probability and Statistics*, 23:217–244, 2019.

[6] Xiang Cheng and Peter L Bartlett. Convergence of Langevin MCMC in KL-divergence. *PMLR 83*, (83):186–211, 2018.

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

[8] R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1:223–236, 2001.

[9] A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78(5):1423–1443, 2012.

[10] Arnak S. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *COLT*, 2017.

[11] Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(3):651–676, 2017.

[12] Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 2019.

[13] Francis X Diebold and Canlin Li. Forecasting the term structure of government bond yields. *Journal of econometrics*, 130(2):337–364, 2006.

[14] M.F. Dixon, I. Halperin, and P.A. Bilokon. *Machine Learning in Finance: From Theory to Practice*. Springer International Publishing, 2020.

[15] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Springer, December 2018.

[16] A. Durmus and E. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

[17] Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017.

[18] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Quantitative Harris-type theorems for diffusions and McKean-Vlasov processes. *Trans. Amer. Math. Soc.*, 371(10):7135–7173, 2019.

[19] J. Gatheral, T. Jaisson, and M. Rosenbaum. Volatility is rough. *Quantitative Finance*, 18:933–949, 2018.

[20] László Gerencsér. On a class of mixing processes. *Stochastics Stochastics Rep.*, 26(3):165–191, 1989.

[21] James B Heaton, Nick G Polson, and Jan Hendrik Witte. Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1):3–12, 2017.

[22] Greg Kirczenow, Masoud Hashemi, Ali Fathi, and Matt Davison. Machine learning for yield curve feature extraction: Application to illiquid corporate bonds, 2018.

[23] J. Lago, F. De Ridder, and B. De Schutter. Forecasting spot electricity prices: deep learning approaches and empirical comparison of traditional algorithms. *Applied Energy*, 221:386–405, 2018.

[24] S. Laruelle, Ch.-A. Lehalle, and G. Pagès. Optimal split of orders across liquidity pools: a stochastic algorithm approach. *SIAM Journal of Financial Mathematics*, 2:1042–1076, 2011.

[25] S. Laruelle, Ch.-A. Lehalle, and G. Pagès. Optimal posting price of limit orders: learning by trading. *Mathematics and Financial Economics*, 7:359–403, 2013.

[26] S. Laruelle and G. Pagès. Stochastic approximation with averaging innovation applied to finance. *Monte Carlo Methods and Applications*, 18:1–51, 2012.

[27] Robert S. Liptser and Albert N. Shiryaev. *Statistics of random processes. I*, volume 5 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, expanded edition, 2001. General theory, Translated from the 1974 Russian original by A. B. Aries, Stochastic Modelling and Applied Probability.

[28] Mateusz B Majka, Aleksandar Mijatović, and Lukasz Szpruch. Non-asymptotic bounds for sampling algorithms without log-concavity. *arXiv preprint arXiv:1808.07105*, 2018.

[29] Xuerong Mao. *Stochastic differential equations and their applications*. Horwood Publishing Series in Mathematics & Applications. Horwood Publishing Limited, Chichester, 1997.

[30] Eduardo Mineo, Airlane Pereira Alencar, Marcelo Moura, and Antonio Elias Fabris. Forecasting the term structure of interest rates with dynamic constrained smoothing b-splines. *Journal of Risk and Financial Management*, 13(4), 2020.

[31] Charles R Nelson and Andrew F Siegel. Parsimonious modeling of yield curves. *Journal of business*, pages 473–489, 1987.

[32] Jacques Neveu. *Discrete-parameter martingales*. North-Holland, 1975.

[33] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. *PMLR 65*, (65):1674–1703, 2017.

[34] Nageswara SV Rao and Vladimir Protopopescu. Function estimation by feedforward sigmoidal networks with bounded weights. *Neural Processing Letters*, 7(3):125–131, 1998.

[35] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edition, 1999.

[36] C. Villani. *Optimal transport : old and new*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

[37] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

[38] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

[39] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3122–3133, 2018.