On recursive estimation schemes with stationary data streams *

M. Barkhagen N. H. Chau É. Moulines M. Rásonyi
S. Sabanis Y. Zhang

June 5, 2018

Abstract

1 Introduction

We are concerned with sampling from the distribution π defined by

$$\pi(A) := \int_A e^{-U(x)}\, dx / \int_{\mathbb{R}^d} e^{-U(x)}\, dx, \ A \in \mathcal{B}(\mathbb{R}^d),$$

where $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel sigma-field and $U : \mathbb{R}^d \to \mathbb{R}_+$ is continuously differentiable. We assume, for simplicity, that $U(0) = 0$. The sampling algorithms will be based on the derivative $h := \nabla U$ and on its noisy observations. We only treat the case of unbiased observations but a bias could also easily be incorporated.

2 Main results

We are working on a probability space $(\Omega, \mathcal{F}, P)$. Expectation of a random variable X will be denoted by EX. For any $m \geq 1$, for any $\mathbb{R}^m$-valued random variable X and for any $1 \leq p < \infty$, let us set $\|X\|_p := \sqrt[p]{\mathbb{E}[|X|^p]}$. We denote by L^p the set of X with $\|X\|_p < \infty$. The indicator function of a set A will be denoted by 1_A.

Let θ_0 be an $\mathbb{R}^d$-valued random variable, representing the initial value of the procedures we consider. Let $H : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ be a measurable function, let X_t, $t \in \mathbb{Z}$ be an $\mathbb{R}^m$-valued (strict sense) stationary process. Let ξ_t, $t \in \mathbb{N}$ be an independent sequence of standard Gaussian random variables. We assume that θ_0, $(X_t)_{t\in\mathbb{Z}}$, $(\xi_t)_{t\in\mathbb{N}}$ are independent.

For each $\lambda > 0$, define the $\mathbb{R}^d$-valued random process θ_t^λ, $t \in \mathbb{N}$ by recursion:

$$\theta_0^\lambda := \theta_0, \quad \theta_{t+1}^\lambda := \theta_t^\lambda - \lambda H(\theta_t^\lambda, X_{t+1}) + \sqrt{2\lambda}\xi_{t+1}. \tag{1}$$

*All the authors were supported by The Alan Turing Institute, London under the EPSRC grant EP/N510129/1. 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.

We assume that $h(\theta) = E[H(\theta, X_0)]$, $\theta \in \mathbb{R}^d$ (in particular, the expectations are finite), and go on defining the "averaged" version of (1),

$$\overline{\theta}_0^\lambda := \theta_0, \quad \overline{\theta}_{t+1}^\lambda := \overline{\theta}_t^\lambda - \lambda h(\overline{\theta}_t^\lambda) + \sqrt{2\lambda}\xi_{t+1}. \tag{2}$$

Assumption 2.1. *There is $L > 0$ such that*

$$|H(\theta_1, x_1) - H(\theta_2, x_2)| \leq L[|\theta_1 - \theta_2| + |x_1 - x_2|].$$

Furthermore, $X_0 \in L^2$.

Under Assumption 2.1, h is also Lipschitz-continuous. Next, monotonicity conditions are required for H. Scalar product in $\mathbb{R}^d$ is denoted by $\langle \cdot, \cdot \rangle$.

Assumption 2.2. *There is a function $a : \mathbb{R}^m \to \mathbb{R}_+$ such that, for all $\theta_1, \theta_2 \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$,*

$$\langle \theta_1 - \theta_2, H(\theta_1, x) - H(\theta_2, x) \rangle \geq a(x)[|\theta_1 - \theta_2|^2 + |H(\theta_1, x) - H(\theta_2, x)|^2] \tag{3}$$

and $Ea(X_0) > 0$.

This assumption clearly holds if, for all x, $\theta \to H(\theta, x)$ is the derivative of a strongly convex function. IN THIS FIRST DRAFT WE ASSUME $a(\cdot)$ TO BE CONSTANT. THIS CAN BE RELAXED MODULO SOME EXTRA WORK.

For technical purposes it is convenient to assume a specific structure for the probability space and the filtrations we consider.

Assumption 2.3. *Let $\mathcal{X}$ be a Polish space. We assume that $\Omega = (\mathcal{X} \times \mathbb{R}^d)^{\mathbb{Z}}$, $\mathcal{F}$ consists of the Borel sets of Ω and $P = \otimes_{i \in \mathbb{Z}} \psi$ where $\psi = \nu \otimes \varpi$ with ν a fixed probability measure on $\mathcal{X}$ and ϖ standard Gaussian on $\mathbb{R}^d$. The coordinate mappings from Ω to $\mathcal{X}$ will be denoted by $\Lambda_i = (\varepsilon_i, \xi_i)$, $i \in \mathbb{Z}$, where ξ_i, $i \geq 1$ are the random variables figuring in (1) and (2) and $X_t = g(\varepsilon_t, \varepsilon_{t-1}, \ldots)$ with a fixed measurable function $g : \mathcal{X}^{-\mathbb{N}} \to \mathbb{R}^m$. We furthermore set $\mathcal{G}_n := \sigma(\varepsilon_i,\ i \leq n)$, as well as $\mathcal{G}_n^+ := \sigma(\varepsilon_i,\ i > n)$, for each $n \in \mathbb{N}$. Define also $\mathcal{F}_n := \mathcal{G}_n \vee \sigma(\xi_i,\ i \in \mathbb{N})$ and $\mathcal{F}_n^+ := \mathcal{G}_n^+$.*

Our aim is to estimate $\|\theta_t^\lambda - \overline{\theta}_t^\lambda\|_2$, uniformly in t.

Example 2.4. Let $H(\theta, x) := \theta + x$ and let X_n, $n \in \mathbb{N}$ be an independent sequence of standard Gaussian random variables, independent of ξ_n, $n \in \mathbb{N}$. Take $\theta_0 := 0$. It is straightforward to check that

$$\overline{\theta}_t^\lambda - \theta_t^\lambda = \sum_{j=0}^{t-1} (1 - \lambda)^j \lambda X_{t-j}$$

which clearly has variance

$$\sum_{j=0}^{t-1} (1 - \lambda)^{2j} \lambda^2 = \frac{\lambda(1 - (1 - \lambda)^{2t})}{2 - \lambda}.$$

It follows that

$$\sup_{t \in \mathbb{N}} ||\overline{\theta}_t^\lambda - \theta_t^\lambda||_2 = \sqrt{\frac{\lambda}{2 - \lambda}}.$$

This shows that the best estimate we may hope to get is of the order $\sqrt{\lambda}$. Our Theorem 2.5 below achieves this bound modulo a logarithmic factor (which does not seem to matter in practice).

Theorem 2.5. *Let X be conditionally L-mixing of order $(2,4)$ with respect to $(\mathcal{G}_t, \mathcal{G}_t^+)$. Let Assumptions 2.1, 2.2 and 2.3 hold. Then there exists $C^\circ > 0$ such that*

$$||\theta_t^\lambda - \overline{\theta}_t^\lambda||_2 \le C^\circ \sqrt{\lambda} |\ln(\lambda)|^{3/2}, \ t \in \mathbb{N}. \tag{4}$$

The next corollary relates our findings in Theorem 2.5 to the problem of sampling from π. Let W_2 denote the Wasserstein metric of order 2, see e.g. [8] for more information about this distance.

Corollary 2.6. *For each $\kappa > 0$, there exist constants $c_1, c_2 > 0$ such that, for each $\epsilon > 0$ one has*

$$W_2(\mathrm{Law}(\theta_t^\lambda), \pi) \le \epsilon$$

whenever

$$\lambda \le c_1 \epsilon^{2+\kappa} \ \textit{and} \ t \ge \frac{c_2}{\epsilon^{2+\kappa}} \ln(1/\epsilon). \tag{5}$$

Remark 2.7. Corollary 2.6 significantly improves on some of the results in [7] in certain cases, compare also to [9]. In [7] the monotonicity assumption (3) is not imposed, only a dissipativity condition is required and a more general recursive scheme is investigated. However, the input sequence X_t, $t \in \mathbb{N}$ is assumed i.i.d. In that setting, Theorem of [7] applies to (1) (with the choice $\delta = 0$, $\beta = 1$, d fixed, see also the last paragraph of Subsection 1.1 of [7]), and we get that

$$W_2(\mathrm{Law}(\theta_t^\lambda), \pi) \le \epsilon$$

holds whenever $\lambda \le c_3(\epsilon/\ln(1/\epsilon))^4$ and $t \ge \frac{c_4}{\epsilon^4} \ln^5(1/\epsilon)$ with some $c_3, c_4 > 0$. Our results provide the sharper estimates (5). The main purpose of the present note is to provide results for the case where X_t, $t \in \mathbb{N}$ has dependencies, but (5) is new even in the particular case where X_t, $t \in \mathbb{N}$ are i.i.d.

3 Conditional L-mixing

L-mixing processes and random fields were introduced in [5]. They proved to be useful in tackling difficult problems of system identification. In [1] the related concept of *conditional* L-mixing was introduced in order to treat fixed gain recursive estimators with discontinous updating functions. Although the function H is assumed continuous in the present article, it seems that the right setting for the analysis of (1) is provided by conditionally L-mixing random fields, which we will define below.

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

Fix an integer $d \ge 1$ and let $D \subset \mathbb{R}^d$ be a set of parameters. A measurable function $X : \mathbb{N} \times D \times \Omega \to \mathbb{R}^m$ is called a random field. We will drop dependence on $\omega \in \Omega$ and use the notation $X_t(\theta)$, $t \in \mathbb{N}$, $\theta \in D$. A random process X_t, $t \in \mathbb{N}$ corresponds to a random field where D is a singleton. A random field is L^r-*bounded* for some $r \ge 1$ if

$$\sup_{t \in \mathbb{N}} \sup_{\theta \in D} ||X_t(\theta)||_r < \infty.$$

Now we define conditional L-mixing. Recall that, for any family Z_i, $i \in I$ of real-valued random variables, $\mathrm{ess.\,sup}_{i \in I} Z_i$ denotes a random variable that is an almost sure upper bound for each Z_i and it is a.s. smaller than or equal to any other such bound, see e.g. Proposition VI.1.1. of [6].

Let $X_t(\theta)$, $t \in \mathbb{N}$, $\theta \in D$ be a random field bounded in L^r. Define, for each $n \in \mathbb{N}$,

$$
\begin{aligned}
M_r^n(X) &:= \operatorname*{ess\,sup}_{\theta \in D} \sup_{t \in \mathbb{N}} E^{1/r}[|X_{n+t}(\theta)|^r \big| \mathcal{H}_n], \\
\gamma_r^n(\tau, X) &:= \operatorname*{ess\,sup}_{\theta \in D} \sup_{t \geq \tau} E^{1/r}[|X_{n+t}(\theta) - E[X_{n+t}(\theta)|\mathcal{H}_{n+t-\tau}^+ \vee \mathcal{H}_n]|^r \big| \mathcal{H}_n], \ \tau \geq 1, \\
\Gamma_r^n(X) &:= \sum_{\tau=1}^{\infty} \gamma_r^n(\tau, X).
\end{aligned}
$$

When necessary, we will also use the notations $M_r^n(X,D)$, $\gamma_r^n(\tau,X,D)$, $\Gamma_r^n(X,D)$ to signal dependence of these quantities on the domain D which may vary.

For some $r, p \geq 1$, we call $X_t(\theta)$, $t \in \mathbb{N}$, $\theta \in D$ *uniformly conditionally L-mixing of order* (r,p) (UCLM-(r,p)) with respect to $(\mathcal{H}_t, \mathcal{H}_t^+)$ if it is L^r-bounded; $X_t(\theta)$, $t \in \mathbb{N}$ is adapted to $\mathcal{F}_t$, $t \in \mathbb{N}$ for all $\theta \in D$ and the sequences $M_r^n(X)$, $\Gamma_r^n(X)$, $n \in \mathbb{N}$ are bounded in L^p. In the case of stochastic processes (when D is a singleton) the terminology "conditionally L-mixing process of order (r,p)" will be used.

The following maximal inequality is pivotal for our arguments.

Theorem 3.1. *Let Assumption 2.3 be in force. Fix $r > 2$, $n \in \mathbb{N}$. Let W_t, $t \in \mathbb{N}$ be a conditionally L-mixing process of order $(r,1)$ w.r.t. $(\mathcal{H}_t, \mathcal{H}_t^+)$, satisfying $E[W_t|\mathcal{H}_n] = 0$ a.s. for all $t \geq n$. Let $m > n$ and let b_t, $n < t \leq m$ be deterministic numbers. Then we have*

$$
E^{1/r}\left[\sup_{n<t\leq m}\left|\sum_{s=n+1}^{t} b_s W_s\right|^r \big|\mathcal{H}_n\right] \leq C_r \left(\sum_{s=n+1}^{m} b_s^2\right)^{1/2} \sqrt{M_r^n(W)\Gamma_r^n(W)}, \quad (6)
$$

almost surely, where C_r is a deterministic constant depending only on r but independent of n, m.

For each $R \geq 0$ we denote $B(R) := \{x \in \mathbb{R}^d : |x| \leq R\}$, the closed ball of radius R around the origin.

Lemma 3.2. *Let X_t, $t \in \mathbb{N}$ be UCLM-(r,p). Let Assumption 2.1 hold true. Then, for each $j \in \mathbb{N}$, the random field $H(\theta, X_t)$, $t \in \mathbb{N}$, $\theta \in B(j)$ is UCLM-(r,p).*

Proof. Notice that

$$
\begin{aligned}
|H(\theta,x)| \leq |H(\theta,x) - H(\theta,0)| + |H(\theta,0) - H(0,0)| + |H(0,0)| \quad &\leq \\
L|x| + L|\theta| + |H(0,0)|.&
\end{aligned}
$$

Let $\theta \in B(j)$. Then, for $k \geq n$,

$$
E[|H(\theta, X_k)|^r|\mathcal{H}_n] \leq C(r)[E[|X_k|^r|\mathcal{H}_n] + j^r + 1]
$$

for some $C(r) > 0$ hence

$$
M_r^n(H(\theta,X), B(j)) \leq C^{1/r}(r)[M_r^n(X) + j + 1].
$$

We also have

$$
\begin{aligned}
E^{1/r}[|H(\theta,X_k)-E[H(\theta,X_k)|\mathcal{H}_n\vee\mathcal{H}^+_{n-\tau}]|^r|\mathcal{H}_n] &\leq \\
2E^{1/r}[|H(\theta,X_k)-H(\theta,E[X_k|\mathcal{H}_n\vee\mathcal{H}^+_{n-\tau}])|^r|\mathcal{H}_n] &\leq \\
2LE^{1/r}[|X_k-E[X_k|\mathcal{H}_n\vee\mathcal{H}^+_{n-\tau}]|^r|\mathcal{H}_n],&
\end{aligned}
$$

using Lemma 3.3, which implies

$$\Gamma^n_r(H(\theta,X),B(j))\leq 2L\Gamma^n_r(X).$$

□

Lemma 3.3. *Let $\mathcal{G},\mathcal{H}\subset\mathcal{F}$ be sigma-algebras. Let X,Y be random variables in L^p such that Y is measurable with respect to $\mathcal{H}\vee\mathcal{G}$. Then for any $p\geq 1$,*

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

If Y is $\mathcal{H}$-measurable then

$$\|X-E[X|\mathcal{H}]\|_p\leq 2\|X-Y\|_p. \tag{7}$$

Proof. See Lemma 6.2 of [1]. □

4 Ergodic properties of recursive schemes

One of the key observations is the following contraction property of the Markov chain $\overline{\theta}_t$, $t\in\mathbb{N}$.

Lemma 4.1. *Let $\theta_0,\theta_0'\in L^2$ be random variables and ξ standard Gaussian, independent of $\sigma(\theta_0,\theta_0')$. Then there is $\rho>0$ such that, defining*

$$\theta_1:=\theta_0-\lambda h(\theta_0)+\sqrt{2\lambda}\xi,\ \ \theta_1':=\theta_0'-\lambda h(\theta_0')+\sqrt{2\lambda}\xi,$$

we have

$$E[(\theta_1-\theta_1')^2]\leq e^{-\rho\lambda}E[(\theta_0-\theta_0')^2].$$

Proof. See Proposition 3 of [2]. □

Let $V(x):=\exp(U(x)/2)$.

Lemma 4.2. *There exists $\tilde{c}>0$ such that $U(x)/2\geq\tilde{c}x^2$ holds for all $x\in\mathbb{R}^d$.*

Let us fix $\tilde{c}$ as in Lemma 4.2 and define $\tilde{V}(x):=\exp(\tilde{c}x^2)$, $x\in\mathbb{R}^d$.

Lemma 4.3. *Let χ_n, $n\geq 1$ be such that*

$$\sup_{n\geq 1}E\tilde{V}(\chi_n)<\infty.$$

Denote $\zeta_n:=\sup_{1\leq k\leq n}|\chi_k|$. Then, for all $p>0$, and for all $n\geq 1$,

$$E|\zeta_n|^p\leq\tilde{C}(p)\ln^{p/2}(n+1).$$

holds for some $\tilde{C}(p)>0$.

Proof. Define the *convex* function

$$f(x) := e^{\tilde{c}|x|^{2/p}},\ |x| \geq \left(\frac{p}{2\tilde{c}}\right)^{p/2},\ f(x) := e^{p/2},\ |x| < \left(\frac{p}{2\tilde{c}}\right)^{p/2}.$$

Denote $M := \sup_{n\in\mathbb{N}} E\tilde{V}(\xi_n)$. Jensen's inequality and trivial considerations show that

$$f(E|\zeta_n|^p) \leq Ef(|\zeta_n|^p) \leq Ee^{\tilde{c}|\zeta_n|^2} + e^{p/2} \quad \leq$$
$$e^{p/2} + \sum_{j=1}^{n} Ee^{\tilde{c}|\chi_j|^2} \leq nM + e^{p/2}.$$

This implies also

$$e^{\tilde{c}E^{2/p}|\zeta_n|^p} \leq nM + e^{p/2},$$

which leads to

$$E|\zeta_n|^p \leq \tilde{C}(p)(\ln(n+1))^{p/2},$$

for some $\tilde{C}(p) > 0$, as stated. □

Lemma 4.4. *Under Assumption 2.2,* $\sup_n EV(\theta_n) < \infty$.

Proof. Assumption 2.2 implies that V is a Lyapunov function for this stochastic system (see Proposition 8 of [3]) and the statement easily follows from this. □

Lemma 4.5. *Let Assumption 2.2 hold. We also have* $\sup_n EV(\overline{z}_n) < \infty$ *and* $\sup_n EV(\overline{\theta}_n) < \infty$. *A fortiori,* $\sup_n E\tilde{V}(\overline{z}_n) < \infty$

Lemma 4.6. *There is* $C^\flat > 0$ *such that*

$$\sup_n \||H(\overline{\theta}_n, X_{n+1})| + |h(\overline{z}_n|\|_2 \leq C^\flat.$$

Proof. This is quite trivial from Assumption 2.1 and Lemma 4.5, details later. □

Clearly, since X is conditionally L-mixing of order $(2,4)$ with respect to $(\mathcal{G}_t, \mathcal{G}_t^+)$, it remains conditionally L-mixing of order $(2,4)$ with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$, too.

For each $\theta \in \mathbb{R}^d$, $0 \leq s \leq t$, we recursively define

$$z(s,s,\theta) := \theta, \quad z(t+1,s,\theta) := z(t,s,\theta) - \lambda h(z(t,s,\theta)) + \sqrt{2\lambda}\xi_{t+1}.$$

We then set, for each $n \in \mathbb{N}$ and for each $nT \leq t < (n+1)T$, $\overline{z}_t := z(t, nT, \theta_{nT})$. Note that $\overline{z}_t$ is then defined for all $t \in \mathbb{N}$ and that $\overline{\theta}_t = z(t, 0, \theta_0)$.

Lemma 4.7. *There is a random variable* Ξ *such that, for all* $\theta \in \mathbb{R}^d$ *and for all* $n \in \mathbb{N}$,

$$\sum_{k=nT+1}^{\infty} |h_{k,nT}(\theta) - h(\theta)| \leq \Xi$$

and $E[\Xi^2] < \infty$.

Proof. Notice that, since $E[X_k|\mathcal{F}^+_{nT}]$ is independent of $\mathcal{F}_{nT}$,

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

This implies that

$$\begin{aligned}
|h_{k,nT}(\theta) - h(\theta)| &\leq \\
\left|E[H(\theta, X_k)|\mathcal{F}_{nT}] - E[H(\theta, E[X_k|\mathcal{F}^+_{nT}])|\mathcal{F}_{nT}]\right| &+ \\
\left|E[H(\theta, E[X_k|\mathcal{F}^+_{nT}])] - E[H(\theta, X_k)]\right| &\leq \\
LE[|X_k - E[X_k|\mathcal{F}^+_{nT}]||\mathcal{F}_{nT}] + LE[|X_k - E[X_k|\mathcal{F}^+_{nT}]|] &\leq \\
L[\gamma_1^{nT}(X, k-nT) + E\gamma_1^{nT}(X, k-nT)].
\end{aligned}$$

Hence

$$\sum_{k=nT+1}^{\infty} |h_{k,nT}(\theta) - h(\theta)| \leq L[\Gamma_1^{nT}(X) + E\Gamma_1^{nT}(X)].$$

Since X is conditionally L-mixing of order $(2,4)$, it is also conditionally L-mixing of order $(1,2)$ so $E[(\Gamma_1^{nT}(X))^2] < \infty$, This implies the statement. □

Proof of Theorem 2.5. Let $T := \lfloor 1/\lambda \rfloor$. Fix $n \in \mathbb{N}$ and let $nT \leq t < (n+1)T$ be arbitrary. Let us define the (random) functions

$$h_{t,nT}(\theta) := E[H(\theta, X_t)|\mathcal{F}_{nT}],\ \theta \in \mathbb{R}^d.$$

It is tedious but standard to show that there exists a jointly measurable version of

$$(\omega, \theta) \to h_{k,nT}(\theta, \omega),\ (\omega, \theta) \in \Omega \times \mathbb{R}^d.$$

Estimate,

$$\begin{aligned}
|\theta_t - \overline{z}_t| \leq \lambda \left| \sum_{k=nT+1}^{t} (H(\theta_k, X_k) - h(\overline{z}_k)) \right| &\leq \\
\lambda \sum_{k=nT+1}^{t} |H(\theta_k, X_k) - H(\overline{z}_k, X_k)| &+ \\
\lambda \left| \sum_{k=nT+1}^{t} (H(\overline{z}_k, X_k) - h_{k,nT}(\overline{z}_k)) \right| &+ \\
\lambda \sum_{k=nT+1}^{t} |h_{k,nT}(\overline{z}_k) - h(\overline{z}_k)| &\leq \\
\lambda L \sum_{k=nT+1}^{t} |\theta_k - \overline{z}_k| &+ \\
\lambda \max_{nT+1 \leq m < (n+1)T} \left| \sum_{k=nT+1}^{m} (H(\overline{z}_k, X_k) - h_{k,nT}(\overline{z}_k)) \right| &+ \\
\lambda \sum_{k=nT+1}^{\infty} |h_{k,nT}(\overline{z}_k) - h(\overline{z}_k)|,
\end{aligned}$$

by Assumption 2.1. Gronwall's lemma and taking squares lead to

$$|\theta_t - \overline{z}_t|^2 \le$$

$$2\lambda^2 e^{2LT\lambda} \left[\max_{nT+1\le m<(n+1)T} \left| \sum_{k=nT+1}^{m} (H(\overline{z}_k, X_k) - h_{k,nT}(\overline{z}_k)) \right|^2 + \left(\sum_{k=nT+1}^{\infty} |h_{k,nT}(\overline{z}_k) - h(\overline{z}_k)| \right)^2 \right],$$

noting also $(x+y)^2 \le 2(x^2+y^2)$, $x, y \in \mathbb{R}$. Let N be the random variable $N := \sup_{nT+1\le i<(n+1)T} |\overline{z}_i|$. Now, recalling the definition of T and taking $\mathcal{F}_{nT}$-conditional expectations, we can write

$$E\left[|\theta_t - \overline{z}_t|^2 | \mathcal{F}_{nT}\right] \le$$

$$2\lambda^2 e^{2L} \left[\sum_{j=1}^{\infty} 1_{\{j-1\le N<j\}} E\left[\max_{nT+1\le m<(n+1)T} \left| \sum_{k=nT+1}^{m} (H(\overline{z}_k, X_k) - h_{k,nT}(\overline{z}_k)) \right|^2 |\mathcal{F}_{nT} \right] + E\left[\left(\sum_{k=nT+1}^{\infty} |h_{k,nT}(\overline{z}_k) - h(\overline{z}_k)| \right)^2 |\mathcal{F}_{nT} \right] \right].$$

Using the $\mathcal{F}_{nT}$-measurability of $\overline{z}_k$, $nT \le k < (n+1)T$, Lemma 4.3 and taking expectations, we can continue our estimations as

$$E|\theta_t - \overline{z}_t|^2 \le$$

$$2\lambda^2 e^{2L} \sum_{j=1}^{\infty} E[1_{\{j-1\le N<j\}} T \Gamma_2^{nT}(H(\theta, X), B(j)) M_2^{nT}(H(\theta, X), B(j))] +$$

$$2\lambda^2 e^{2L} E\left[\Xi^2\right],$$

see Lemma 4.7. By the Cauchy inequality and the trivial $\{j-1 \le N\} = \{j \le N+1\}$, an application of Theorem 3.1 gives

$$\sum_{j=1}^{\infty} E[1_{\{j-1\le N<j\}} \Gamma_2^{nT}(H(\theta, X), B(j)) M_2^{nT}(H(\theta, X), B(j))] \le$$

$$\sum_{j=1}^{\infty} P^{1/2}(N+1 \ge j) E^{1/2}[(\Gamma_2^{nT}(H(\theta, X), B(j)))^2 (M_2^{nT}(H(\theta, X), B(j)))^2] \le$$

$$\sum_{j=1}^{\infty} \sqrt{\frac{E(N+1)^6}{j^6}} 2L E^{1/2}[(\Gamma_2^{nT}(X))^2 C(2)[M_2^{nT}(X) + j + 1]^2] \le$$

$$\sum_{j=1}^{\infty} \sqrt{\frac{E(N+1)^6}{j^6}} 2L\sqrt{C(2)} E^{1/4}[(\Gamma_2^{nT}(X))^4] E^{1/4}[M_2^{nT}(X) + j + 1]^4] \le$$

$$\check{C}' \sum_{j=1}^{\infty} \frac{\ln^3(T)}{j^2} \le \check{C} \ln^3(T),$$

for suitable $\check{C}, \check{C}' > 0$, noting Lemma 3.2 and the fact that X was assumed to be conditionally L-mixing of order $(2, 4)$. We conclude that

$$E^{1/2}|\theta_t - \overline{z}_t|^2 \le C^{\sharp} \lambda \sqrt{T} |\ln(T)|^{3/2} \le C^{\star} \sqrt{\lambda} |\ln(\lambda)|^{3/2},$$

with some $C^\sharp, C^\star > 0$, for all $t \in \mathbb{N}$.

Now we turn to estimating $|\overline{z}_t - \overline{\theta}_t|$. By Lemmata 4.1 and 4.6,

$$
\begin{aligned}
||\overline{z}_t - \overline{\theta}_t||_2 &\leq \\
\sum_{k=1}^{n} ||z(t, kT, \theta_{kT}) - z(t, (k-1)T, \theta_{(k-1)T})||_2 &= \\
\sum_{k=1}^{n} ||z(t, kT, \theta_{kT}) - z(t, kT, z(kT, (k-1)T, \theta_{(k-1)T}))||_2 &\leq \\
\sum_{k=1}^{n} e^{-\lambda\rho(t-kT)/2} \left[||\overline{\theta}_{kT-1} - \overline{z}_{kT-1}||_2 + \lambda ||H(\overline{\theta}_{kT-1}, X_{kT}) - h(\overline{z}_{kT-1})||_2 \right] &\leq \\
\frac{C^\dagger}{1 - e^{-\rho/2}} \left[1 + C^\flat \right] \sqrt{\lambda} |\ln(\lambda)|^{3/2},
\end{aligned}
$$

for some $C^\dagger > 0$. This completes the proof of the theorem since

$$
|\theta_t - \overline{\theta}_t| \leq |\theta_t - \overline{z}_t| + |\overline{z}_t - \overline{\theta}_t|.
$$

□

5 Examples

6 The bright future

I think that we can significantly generalize the above results using another approach, based on [4]. We will use the metric

$$
w(\mu, \nu) := \inf_{\zeta \in \mathcal{C}(\mu,\nu)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} [1 \wedge |x - y|](1 + V(x) + V(y)) \zeta(dx, dy).
$$

Consider the diffusion process L_t, $t \in \mathbb{R}_+$ defined by

$$
dL_t = -h(L_t)\, dt + dB_t,\ L_0 := \theta_0.
$$

We will also need, for each $\lambda > 0$,

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

Let us introduce the Itô process

$$
dY_t^\lambda = -H(Y_t^\lambda, X_{\lfloor t \rfloor})\ dt + dB_t,\ Y_t^\lambda := \theta_0.
$$

A new approach would consist of the following steps:

1. We only assume dissipativity and Lipschitz-continuity of $h(\theta), H(\theta, x)$ (the latter uniformly in x). This guarantees that L_t is contractive in the metric w, by [4] and by e.g. [3].

2. The arguments presented above can equally well be performed in continuous time and provide $w(\mathrm{Law}(L_t^\lambda), \mathrm{Law}(Y_t^\lambda)) \leq C\sqrt{\lambda}|\ln(\lambda)|^{3/2}$, for each t. (OK, strictly speaking the above argument contains L^2-estimates of the type $E|X - Y|^2$ but these could be ameliorated to get estimates of the form $E|X - Y|(1 + V(X) + V(Y))$ since V is a Lyapunov function.)

3. Now the arguments of [3] (which go back to Dalalyan) could be used to establish the bound

$$||\mathrm{Law}(Y_t^\lambda) - \mathrm{Law}(\theta_t^\lambda)||_V \leq C\sqrt{\lambda}$$

on the weighted total variation norm There is also X_t intervening here but I think that the same arguments (using Kullback-Leibler divergence) should work.

4. As $w(\cdot,\cdot) \leq C||\cdot - \cdot||_V$, this leads to a rate of convergence much better than that of Raginsky, not in W_2, but in w. AND WITHOUT CONVEXITY OF ANY SORT! Also, h does not need to be the derivative of something for 2. and 3. to work.

With Huy we could do 2., I think, while the other part of the team could do 3. The machine learning community will tremble :).

References

[1] N. H. Chau, Ch. Kumar, M. Rásonyi and S. Sabanis. On fixed gain recursive estimators with discontinuity in the parameters. arXiv:1609.05166v3, 2017.

[2] A. Durmus and É. Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. arXiv:1605.01559, 2018.

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

[4] A. Eberle, A. Guillin and R. Zimmer. Quantitative Harris-type theorems for diffusions and McKean-Vlasov processes. *Preprint.*, 2017. arXiv:1606.0612v2

[5] L. Gerencsér. On a class of mixing processes. *Stochastics*, 26:165–191, 1989.

[6] J.Neveu. *Discrete-parameter martingales.* North-Holland, 1975.

[7] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. *Proceedings of Machine Learning Research*, 65:1674–1703, 2017. arXiv:1702.03849

[8] C. Villani. *Optimal transport. Old an new.* Springer, 2009.

[9] P. Xu, J. Chen, D. Zhou and Q. Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. arXiv:1707.06618, 2018.