Adv. Appl. Probab. **55,** 382–406 (2023)
*doi:*10.1017/apr.2022.25

CONVERGENCE OF THE KIEFER–WOLFOWITZ ALGORITHM IN THE PRESENCE OF DISCONTINUITIES

MIKLÓS RÁSONYI* ** AND
KINGA TIKOSI,* *** *Rényi Institute, Budapest, and Mathematical Institute, Warsaw*

Abstract

In this paper we estimate the expected error of a stochastic approximation algorithm where the maximum of a function is found using finite differences of a stochastic representation of that function. An error estimate of the order $n^{-1/5}$ for the nth iteration is achieved using suitable parameters. The novelty with respect to previous studies is that we allow the stochastic representation to be discontinuous and to consist of possibly dependent random variables (satisfying a mixing condition).

Keywords: Stochastic approximation; dependent data; threshold strategies

2020 Mathematics Subject Classification: Primary 93E35
Secondary 91G60

1. Introduction

We are interested in maximizing a function $U : \mathbb{R}^d \to \mathbb{R}$ which is unknown. However, we can observe a sequence $J(\theta, X_n)$, $n \geq 1$, where $J : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ is measurable,

$$\mathbb{E}[J(\theta, X_1)] = U(\theta), \ \theta \in \mathbb{R}^d, \tag{1}$$

and X_n, $n \geq 1$, is an $\mathbb{R}^m$-valued stationary process in the strong sense. The stochastic representations $J(\theta, X_n)$ are often interpreted as noisy measurements of $U(\theta)$. In this paper we focus on applications to mathematical finance, described in Section 6 below, where $J(\theta, X_t)$ are functionals of observed economic variables X_t and θ determines an investor's portfolio strategy. In that context, stochasticity does not come from measurement errors; rather, it is an intrinsic property of the system. Maximizing U serves to find the best investment policy in an online, adaptive manner.

We study a recursive algorithm employing finite differences, as proposed by Kiefer and Wolfowitz in [14]. This is a variant of the Robbins–Monro stochastic gradient method [19] where, instead of the objective function itself, its gradient is assumed to admit a stochastic representation.

The novelty in our work is that we do not assume differentiability, nor even continuity, of $\theta \to J(\theta, \cdot)$, and the sequence X_n may well be dependent as long as it satisfies a mixing condition. The only result in such a setting that we are aware of is in [16], which, however, studies only almost sure convergence, without a convergence rate. Our purpose is not to find

Received 11 August 2020; revision received 25 April 2022.
* Postal address: Reáltanoda utca 13-15, 1053 Budapest, Hungary; ul. Śniadeckich 8, 00-656 Warszawa, Poland.
** Email address: rasonyi@renyi.hu
*** Email address: kinga.tikosi@gmail.com

the weakest possible hypotheses but to arouse keen interest in the given problem that may lead to further, more general results. Our work is also a continuation of [7, 4], where discontinuous stochastic gradient procedures were treated.

The main theorems are stated in Section 2 and proved in Section 3. Section 4 recalls earlier results that we are relying on. A numerical example is provided in Section 5. We explain the significance of our results for algorithmic trading in Section 6.

2. Set-up and results

For real-valued quantities X, Y, the notation $X = O(Y)$ means that there is a constant $C > 0$ such that $|X| \leq CY$. We will always work on a fixed probability space $(\Omega, \mathcal{F}, P)$ equipped with a filtration $\mathcal{F}_n$, $n \in \mathbb{N}$, such that $\mathcal{F}_0 = \{\emptyset, \Omega\}$. A decreasing sequence of sigma-algebras $\mathcal{F}_n^+$, $n \in \mathbb{N}$, is also given, such that, for each $n \in \mathbb{N}$, $\mathcal{F}_n$ and $\mathcal{F}_n^+$ are independent and X_n is adapted to $\mathcal{F}_n$. The notation $\mathbb{E}[X]$ refers to the expectation of a real-valued random variable X, while $\mathbb{E}_k[X]$ is a shorthand notation for $\mathbb{E}[X|\mathcal{F}_k]$, $k \in \mathbb{N}$. $P_k(A)$ refers to the conditional probability $P(A|\mathcal{F}_k)$. We denote by $\mathbb{1}_A$ the indicator of a set A. The notation ω refers to a generic element of Ω. For $r \geq 1$, we refer to the set of random variables with finite rth moments as L^r. Vertical bars $|\cdot|$ denote the Euclidean norm in $\mathbb{R}^k$, where k may vary according to the context.

For $i = 1, \ldots, d$, let $\mathbf{e}_i \in \mathbb{R}^d$ denote the vector in which the ith coordinate is 1 and the other coordinates are 0. For two vectors $v, w \in \mathbb{R}^m$, the relation $v \leq w$ expresses that $v^i \leq w^i$ for all the components $i = 1, \ldots, m$. Let $B_r := \left\{\theta \in \mathbb{R}^d \colon |\theta| \leq r\right\}$ denote the ball of radius r, for $r \geq 0$.

Let the function $U : \mathbb{R}^d \to \mathbb{R}$ have a unique maximum at the point $\theta^* \in \mathbb{R}^d$. Consider the following recursive stochastic approximation scheme for finding θ^*:

$$\theta_{k+1} = \theta_k + \lambda_k H\big(\theta_k, X_{k+1}, c_k\big), \quad \text{for } k \in \mathbb{N}, \tag{2}$$

starting from some initial (deterministic) guess $\theta_0 \in \mathbb{R}^d$, where H is an estimator of the gradient of J, defined as

$$H(\theta, x, c) = \sum_{i=1}^{d} \frac{J(\theta + c\mathbf{e}_i, x) - J(\theta - c\mathbf{e}_i, x)}{2c} \mathbf{e}_i,$$

for all $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}^m$, and $c > 0$.

The sequences $(\lambda_k)_{k\in\mathbb{N}}$ and $(c_k)_{k\in\mathbb{N}}$ appearing in (2) will consist of positive real numbers, which are to be specified later. We will distinguish the cases where λ_k, c_k tend to zero and where they are kept constant, the former being called *decreasing-gain* approximation and the latter *fixed-gain* approximation.

Remark 2.1. Our results below could easily be formulated in a more general setting where $J\big(\theta_k + c_k\mathbf{e}_i, X_{k+1}(i)\big)$ and $J\big(\theta_k - c_k\mathbf{e}_i, X'_{k+1}(i)\big)$, $i = 1, \ldots, d$, are considered with distinct $X_{k+1}(i)$ and $X'_{k+1}(i)$. In the applications that motivate us this is not the case; hence, for reasons of simplicity, we stay in the present setting.

Assumption 2.1. *U is continuously differentiable with unique maximum $\theta^* \in \mathbb{R}^d$. Let $G(\theta) = \nabla U(\theta)$. The function G is assumed Lipschitz-continuous with Lipschitz constant L_G.*

We assume in the sequel that the function J in (1) has a specific form. Note that though J is not continuous, U can nonetheless be continuously differentiable, by the smoothing effect of randomness.

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

Assumption 2.2. *Let the function J be of the following specific form:*

$$J(\theta, x) = l_0(\theta)\mathbb{1}_{A_0(x)} + \sum_{i=1}^{m_s} \mathbb{1}_{A_i(x)} l_i(\theta, x),$$

where $l_i : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ are Lipschitz-continuous (in both variables) for $i = 1, \dots, m_s$, and for some $m_p, m'_p \in \mathbb{N}$,

$$A_i(x) := \left(\cap_{j=1}^{m_p} \left\{ \theta : x \le g_i^j(\theta) \right\} \right) \bigcap \left(\cap_{j=1}^{m'_p} \left\{ \theta : x > h_i^j(\theta) \right\} \right), \ i = 1, \dots, m_s,$$

with Lipschitz-continuous functions $g_i^j, h_i^j : \mathbb{R}^d \to \mathbb{R}^m$. Furthermore, $A_0(x) := \mathbb{R}^d \setminus \cup_{i=1}^{m_s} A_i(x)$ and

$$\cup_{x \in \mathbb{R}^m} \cup_{i=1}^{m_s} A_i(x) \subset B_D$$

for some $D > 0$. The function l_0 is twice continuously differentiable, and there are constants L_1'', L_2'' such that

$$L_1'' I \le \nabla\nabla l_0 \le L_2'' I,$$

where I is the $d \times d$ identity matrix.

Remark 2.2. Assumption 2.2 implies that ∇l_0 grows linearly, and hence l_0 itself is locally Lipschitz with linearly growing Lipschitz coefficient; that is,

$$|l_0(\theta_1) - l_0(\theta_2)| \le L_0(1 + |\theta_1| + |\theta_2|)|\theta_1 - \theta_2|,$$

with some $L_0 > 0$, for all $\theta_1, \theta_2 \in \mathbb{R}^d$.

In plain English, we consider J which is smooth on a finite number of bounded domains (the interior of the constraint sets $A_i(x)$, $i = 1, \dots, m_s$) but may have discontinuities at the boundaries. Furthermore, J (and hence also U) is required to be quadratic 'near infinity' (on $A_0(x)$).

We briefly explain why such a hypothesis is not restrictive for real-life applications. Normally, there is a compact set Q (e.g. a cube or a ball) such that only parameters from Q are relevant, i.e. U is defined only on Q. Assume it has some stochastic representation

$$U(\theta) = \mathbb{E}[J(\theta, X_0)], \qquad \theta \in Q, \tag{3}$$

and a unique maximum $\theta^* \in Q$. Assume that $Q \subset B_D$ for some D. Extend U outside B_D as $U(\theta) = -A|\theta|^2 + B$ for suitable A, B. Extend U and J to $B_D \setminus Q$ as well in such a way that U is continuously differentiable, and $U(\theta) < U(\theta^*)$ for all $\theta \neq \theta^*$ (see Section 4 of [5] for a rigorous construction of this kind). Set $J := U$ outside Q. Then our maximization procedure can be applied to this setting for finding θ^*.

Defining $U = l_0$ to be (essentially) quadratic outside a compact set is one way of solving the problem that such procedures often leave their effective domain Q. Other solutions are resetting (see e.g. [9]) or an analysis of the probability of divergence (see e.g. [2]).

The next assumption postulates that the process X should be bounded and the conditional laws of X_{k+1} should be absolutely continuous with a bounded density.

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

Assumption 2.3. *For each $k \in \mathbb{N}$,*

$$P_k\big(X_{k+1} \in A\big)(\omega) = \int_A p_k(u, \omega)du, \qquad P - \textit{almost surely}, \ A \in \mathcal{B}\big(\mathbb{R}^d\big)$$

for some measurable $p_k : \mathbb{R}^d \times \Omega \to \mathbb{R}_+$, and there is a fixed constant F such that $p_k(u, \omega) \le F$ holds for all k, ω, u. The random variable X_0 satisfies $|X_0| \le K_0$ for some constant K_0.

Note that, by strong stationarity, the process X_k is uniformly bounded under Assumption 2.3.

We will assume a certain mixing property about the process X_n which we recall now. A family of $\mathbb{R}^d$-valued random variables Z_i, $i \in \mathcal{I}$, is called *L^r-bounded* for some $r \ge 1$ if $\sup_{i\in\mathcal{I}} \mathbb{E}|Z_i|^r < \infty$; here $\mathcal{I}$ may be an arbitrary index set.

For a random field $Y_n(\theta)$, $n \in \mathbb{N}$, $\theta \in \mathbb{R}^d$, bounded in L^r for some $r \ge 1$, we define, for all $n \in \mathbb{N}$,

$$M_r^n(Y) = \operatorname*{ess\,sup}_\theta \sup_{k\in\mathbb{N}} \mathbb{E}^{1/r}[|Y_{n+k}(\theta)|^r \,|\mathcal{F}_n],$$

$$\gamma_r^n(\tau, Y) = \operatorname*{ess\,sup}_\theta \sup_{k\ge\tau} \mathbb{E}^{1/r}\left[\left|Y_{n+k}(\theta) - \mathbb{E}\big[Y_{n+k}(\theta)\big|\mathcal{F}^+_{n+k-\tau} \vee \mathcal{F}_n\big]\right|^r |\mathcal{F}_n\right], \qquad \tau \ge 0,$$

$$\Gamma_r^n(Y) = \sum_{\tau=0}^{\infty} \gamma_r^n(\tau, Y).$$

These quantities clearly make sense also for any L^r-bounded stochastic process Y_n, $n \in \mathbb{N}$ (the essential suprema disappear in this case). $M_r^n(Y)$ measures the (conditional) moments of Y, while $\Gamma_r^n(Y)$ describes its dependence structure (like covariance decay). In particular, one can define $M_r^n(X)$, $\Gamma_r^n(X)$. We clearly have $M_r^n(X) \le K_0$ under Assumption 2.3. The quantities $\Gamma_r^n(X)$ will figure in certain estimates later.

Assumption 2.4. *For some $\epsilon > 0$, $\gamma_3^n(\tau, X) = O\big((1+\tau)^{-4-\epsilon}\big)$, where the constant of $O(\cdot)$ is independent of ω, τ, and n. Furthermore,*

$$\mathbb{E}\big[\big|X_{n+k} - \mathbb{E}\big[X_{n+k}|\mathcal{F}_n^+\big]\big|\big] = O\big(k^{-2-\epsilon}\big), \qquad k \ge 1,$$

where the constant of $O(\cdot)$ is independent of n, k.

Both requirements in Assumption 2.4 are about how the effect of the past on the present decreases as we go back farther in time.

Example 2.1. Let ε_n, $n \in \mathbb{N}$, be a bounded sequence of independent and identically distributed (i.i.d.) random variables in $\mathbb{R}^m$ with bounded density w.r.t. the Lebesgue measure, and choose $\mathcal{F}_k := \sigma(\varepsilon_j, \ j \le k)$ and $\mathcal{F}_k^+ := \sigma(\varepsilon_j, \ j \ge k+1)$. Then $X_n := \varepsilon_n$, $n \in \mathbb{N}$, satisfies Assumptions 2.3 and 2.4. A causal infinite moving average process whose coefficients decay sufficiently fast is another pertinent example. Indeed, using the argument of Lemma 4.2 of [4], one can show that $X_n := \sum_{j=0}^{\infty} s_j\varepsilon_{n-j}$, $n \in \mathbb{N}$, satisfies Assumption 2.4, where the ε_i are as above, $s_0 \ne 0$, and $|s_j| \le (1+j)^{-\beta}$ holds for some $\beta > 9/2$. Assumption 2.3 is also clearly satisfied in that model.

Remark 2.3. A random field $Y_n(\theta)$, $n \in \mathbb{N}$, is called uniformly conditionally L-mixing if $Y_n(\theta)$ is adapted to the filtration $\mathcal{F}_n$, $n \in \mathbb{N}$, for all θ, and the sequences $M_r^n(Y)$, $\Gamma_r^n(Y)$, $n \in \mathbb{N}$, are bounded in L^r for each $r \ge 1$. Our Assumption 2.4 thus requires a sort of related mixing property. Conditional L-mixing was introduced in [4], inspired by [8].

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

2.1. Decreasing-gain stochastic approximation

The usual assumption on the sequences $(\lambda_k)_{k=1,2,\dots}$ and $(c_k)_{k=1,2,\dots}$ in the definition of the recursive scheme (2) are the following (see [14]):

$$\begin{aligned} c_k &\to 0, \quad k \to \infty, \\ \sum_{k=1}^{\infty} \lambda_k &= \infty, \\ \sum_{k=1}^{\infty} \lambda_k c_k &< \infty, \\ \sum_{k=1}^{\infty} \lambda_k^2 c_k^{-2} &< \infty. \end{aligned} \tag{4}$$

In the sequel we stick to a more concrete choice which clearly fulfills the conditions in (4) above.

Assumption 2.5. *We fix λ_0, $c_0 > 0$, $\gamma \in (0, 1/3)$, and set*

$$\lambda_k = \lambda_0 \int_k^{k+1} \frac{1}{u}\, du$$

and $c_k = c_0 k^{-\gamma}$, $k \geq 1$. We also assume $c_0 \leq 1$.

Asymptotically λ_k behaves like λ_0/k. However, our choice somewhat simplifies the otherwise already involved theoretical analysis.

The ordinary differential equation (ODE) associated with the problem is

$$\dot{y}_t = \frac{\lambda_0}{t} G(y_t). \tag{5}$$

The idea of using an associated deterministic ODE to study the asymptotic properties of recursive schemes was introduced by Ljung in [17]. The intuition behind this association is that in the long run the noise effects average out and the asymptotic behavior is determined by this 'mean' differential equation. A heuristic connection between the dynamics of the recursive scheme and the ODE can be seen if one looks at the Euler discretization of the latter.

The solution of (5) with initial condition $y_s = \xi$ will be denoted by $y(t, s, \xi)$ for $0 < s \leq t$.

Assumption 2.6. *The ODE (5) fulfills the stability assumption formulated below: there exist $C^* > 0$ and $\alpha > 0$ such that*

$$\left| \frac{\partial y(t, s, \xi)}{\partial \xi} \right| \leq C^* \left(\frac{s}{t} \right)^{\alpha}$$

for all $0 < s < t$.

Our main result comes next.

Theorem 2.1. *Let Assumptions 2.1, 2.2, 2.3, 2.4, 2.5, and 2.6 hold. Then $\mathbb{E}|\theta_n - \theta^*| = O\big(n^{-\chi} + n^{-\alpha}\big)$, $n \geq 1$, where $\chi = \min\big\{\frac{1}{2} - \frac{3}{2}\gamma, \gamma\big\}$ and the constant in $O(\cdot)$ depends only on θ_0.*

To get the best result, set $\gamma = \frac{1}{5}$. In this case the convergence rate is $\chi = \frac{1}{5}$ (provided that $\alpha \geq 1/5$). For Kiefer–Wolfowitz procedures, [20] establishes a convergence rate $n^{-1/3}$ under

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

fairly restrictive conditions (e.g. J is assumed smooth and X is i.i.d.). Our approach is entirely different from that of [20] and relies on the ODE method (see e.g. [15]) in the spirit of [9, 11, 10], where so-called SPSA procedures were analyzed.

Theoretical analysis in the present case is much more involved for two reasons: the discontinuities of J and the state-dependent setting (hardly analyzed in the literature at all). Our results are closest to those of [10], where a rate of $n^{-2/7}$ is obtained for the SPSA algorithm (a close relative of Kiefer–Wolfowitz), with strong smoothness assumptions imposed on J. As already remarked, in the absence of smoothness ours is the first study providing a convergence rate. Further strengthening of our result seems to be difficult and will be the object of further investigations.

We also point to two related papers where stochastic gradient procedures are analyzed: [7] treats the Markovian case while [4] is about possibly non-Markovian settings. In these studies, the gradient of J is assumed to exist, but it may be discontinuous. Some of the ideas of [4] apply in the present, more difficult case where even the continuity of J fails.

2.2. Fixed-gain stochastic approximation

Let us also consider a modified recursive scheme

$$\theta_{k+1} = \theta_k + aH(\theta_k, X_{k+1}, c), \qquad k \in \mathbb{N}, \tag{6}$$

where a and c are fixed (small) positive reals, independent of k. In contrast with the previous scheme (2), which is meant to converge to the maximum of the function, this method is expected to *track* the maximum.

The ODEs associated with the problem are

$$\dot{y}_t = \lambda G(y_t), \tag{7}$$

for each $\lambda > 0$.

Note that, by an exponential time change, one can show that Assumption 2.6 on the ODE (5) implies (7) being exponentially stable, i.e. satisfying

$$\left| \frac{\partial y(t, s, \xi)}{\partial \xi} \right| \le C^* e^{-\alpha\lambda(t-s)}, \qquad 0 < s \le t,$$

for some $\alpha > 0$ (possibly different from the one in (5)).

Theorem 2.2. *Let Assumptions 2.1, 2.2, 2.3, 2.4, and 2.6 hold. Then* $\mathbb{E}|\theta_n - \theta^*| = O\left(\max\left(c^2, \sqrt{\frac{a}{c}}\right) + e^{-a\alpha n}\right)$ *holds for all* $n \ge 1$, *where the constant in* $O(\cdot)$ *depends only on* θ_0.

Note that, similarly to the decreasing-gain setting, this leads to the best choice being $c = a^{\frac{1}{5}}$. We know of no other papers where the fixed-gain case has been treated. In the case of stochastic gradients there are many such studies obtaining a rate of $\sqrt{a}$ for step size a; see e.g. [4] and the references therein.

3. Proofs

The following lemma will play a pivotal role in our estimates: it establishes the *conditional* Lipschitz-continuity of the difference function obtained from J.

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

Lemma 3.1. *Under Assumptions 2.2 and 2.3, there is $C_\flat > 0$ such that, for each $i = 1, \dots, d$ and $c \le 1$,*

$$\mathbb{E}_k\left|J\left(\bar\theta_1 + c\mathbf{e}_i, X_{k+1}\right) - J\left(\bar\theta_1 - c\mathbf{e}_i, X_{k+1}\right) - J\left(\bar\theta_2 + c\mathbf{e}_i, X_{k+1}\right) + J\left(\bar\theta_2 - c\mathbf{e}_i, X_{k+1}\right)\right| \le C_\flat\left[|\bar\theta_1 - \bar\theta_2| + c^2\right]$$

holds for all $k \in \mathbb{N}$ and for all pairs of $\mathcal{F}_k$-measurable $\mathbb{R}^d$-valued random variables $\bar\theta_1$, $\bar\theta_2$.

Proof. We assume that $m_s = 1$, $m_p = 1$, $m_p' = 0$. We will briefly refer to the general case later. We thus assume that $J(\theta, x) = l_1(\theta, x)\mathbb{1}_{\{x \le g(\theta)\}} + l_0(\theta)\mathbb{1}_{A_0(x)}$ with some Lipschitz-continuous g, l_1 with Lipschitz constant L_1 (for both). Let K_1 be an upper bound for l_1 in B_{D+2}.

Consider first the event $A_1 := \left\{\bar\theta_1, \bar\theta_2 \in B_{D+1}\right\}$ and the corresponding indicator $I_1 := \mathbb{1}_{A_1}$. Note that on I_1 we have $\bar\theta_j \pm c\mathbf{e}_i \in B_{D+2}$, $j = 1, 2$. Now estimate

$$\begin{aligned}
&\mathbb{E}_k\left|I_1 l_1\left(\bar\theta_1 + c\mathbf{e}_i, X_{k+1}\right)\mathbb{1}_{\left\{X_{k+1} \le g\left(\bar\theta_1 + c\mathbf{e}_i\right)\right\}} - I_1 l_1\left(\bar\theta_2 + c\mathbf{e}_i, X_{k+1}\right)\mathbb{1}_{\left\{X_{k+1} \le g\left(\bar\theta_2 + c\mathbf{e}_i\right)\right\}}\right| \\
&\le \mathbb{E}_k\left|I_1 l_1\left(\bar\theta_1 + c\mathbf{e}_i, X_{k+1}\right)\mathbb{1}_{\left\{X_{k+1} \le g\left(\bar\theta_1 + c\mathbf{e}_i\right)\right\}} - I_1 l_1\left(\bar\theta_2 + c\mathbf{e}_i, X_{k+1}\right)\mathbb{1}_{\left\{X_{k+1} \le g\left(\bar\theta_1 + c\mathbf{e}_i\right)\right\}}\right| \\
&\quad + \mathbb{E}_k\left|I_1 l_1\left(\bar\theta_2 + c\mathbf{e}_i, X_{k+1}\right)\mathbb{1}_{\left\{X_{k+1} \le g\left(\bar\theta_1 + c\mathbf{e}_i\right)\right\}} - I_1 l_1\left(\bar\theta_2 + c\mathbf{e}_i, X_{k+1}\right)\mathbb{1}_{\left\{X_{k+1} \le g\left(\bar\theta_2 + c\mathbf{e}_i\right)\right\}}\right| \\
&\le L_1\mathbb{E}_k\left|\bar\theta_1 - \bar\theta_2\right| + K_1\sum_{j=1}^{m}\Big[P_k\left(g^j\left(\bar\theta_2 + c\mathbf{e}_i\right) < X_{k+1}^j \le g^j\left(\bar\theta_1 + c\mathbf{e}_i\right)\right) \\
&\qquad\qquad + P_k\left(g^j\left(\bar\theta_1 + c\mathbf{e}_i\right) < X_{k+1}^j \le g^j\left(\bar\theta_2 + c\mathbf{e}_i\right)\right)\Big] \\
&\le L_1\left|\bar\theta_1 - \bar\theta_2\right| + 2mK_1L_1F\left|\bar\theta_1 - \bar\theta_2\right|. \qquad (8)
\end{aligned}$$

In the same way, we also get

$$\mathbb{E}_k\left|I_1 l_1\left(\bar\theta_1 - c\mathbf{e}_i, X_{k+1}\right) - I_1 l_1\left(\bar\theta_2 - c\mathbf{e}_i, X_{k+1}\right)\right| \le L_1\left|\bar\theta_1 - \bar\theta_2\right| + 2mK_1L_1F\left|\bar\theta_1 - \bar\theta_2\right|.$$

As l_0 is clearly Lipschitz on B_{D+2}, we also have

$$\left|I_1 l_0\left(\bar\theta_1 \pm c\mathbf{e}_i, X_{k+1}\right) - I_1 l_0\left(\bar\theta_2 \pm c\mathbf{e}_i, X_{k+1}\right)\right| = O\left(\left|\bar\theta_1 - \bar\theta_2\right|\right).$$

Let L_2'' be an upper bound for the second derivative $\nabla\nabla l_0$; recall Assumption 2.2. Now let A_2 be the event that the line from $\bar\theta_1$ to $\bar\theta_2$ does not intersect B_{D+1} at all; let $I_2 := \mathbb{1}_{A_2}$. It follows in particular that neither $\bar\theta_1 \pm c\mathbf{e}_i$ nor $\bar\theta_2 \pm c\mathbf{e}_i$ fall into B_D. Since $J = l_0$ outside B_D we can write, by the Lagrange mean value theorem,

$$\begin{aligned}
&\mathbb{E}_k I_2\left|J\left(\bar\theta_1 + c\mathbf{e}_i, X_{k+1}\right) - J\left(\bar\theta_2 + c\mathbf{e}_i, X_{k+1}\right) - J\left(\bar\theta_1 - c\mathbf{e}_i, X_{k+1}\right) + J\left(\bar\theta_2 - c\mathbf{e}_i, X_{k+1}\right)\right| \\
&= 2c\mathbb{E}_k I_2\left|\partial_{\theta_i} l_0(\xi_1) - \partial_{\theta_i} l_0(\xi_2)\right| \\
&\le 2c\sup_{u \in \mathbb{R}^d}\left|\nabla\left(\partial_{\theta_i} l_0(u)\right)\right|\mathbb{E}_k|\xi_1 - \xi_2| \\
&\le 2cL_2''\mathbb{E}_k|\xi_1 - \xi_2| \\
&\le 2cL_2''\left[\left|\bar\theta_1 - \bar\theta_2\right| + 2c\right] \le 2L_2''\left|\bar\theta_1 - \bar\theta_2\right| + 4c^2L_2'',
\end{aligned}$$

with some random variables $\xi_j \in \left[\bar\theta_j - c\mathbf{e}_i, \bar\theta_j + c\mathbf{e}_i\right]$, $j = 1, 2$, remembering our assumptions on l_0 and $c \le 1$.

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

Turning to the event $\Omega \setminus (A_1 \cup A_2)$, let us consider the directed straight line from $\bar{\theta}_1(\omega)$ to $\bar{\theta}_2(\omega)$; let its first intersection point with the boundary of B_{D+1} be denoted by $\kappa_1(\omega)$ and its second intersection point by $\kappa_2(\omega)$. In the case where there is only one intersection point, it is denoted by $\kappa_1(\omega)$. Let I_3 be the indicator of the event that there is only one intersection point (κ_1) with B_{D+1} and that $\bar{\theta}_1$ is inside B_{D+1}. The arguments of the previous two cases guarantee that

$$\begin{aligned}
&\mathbb{E}_k I_3 \big| J\big(\bar{\theta}_1 + c\mathbf{e}_i, X_{k+1}\big) - J\big(\bar{\theta}_2 + c\mathbf{e}_i, X_{k+1}\big) - J\big(\bar{\theta}_1 - c\mathbf{e}_i, X_{k+1}\big) + J\big(\bar{\theta}_2 - c\mathbf{e}_i, X_{k+1}\big)\big| \\
&\le \mathbb{E}_k I_3 \big| J\big(\bar{\theta}_1 + c\mathbf{e}_i, X_{k+1}\big) - J\big(\kappa_1 + c\mathbf{e}_i, X_{k+1}\big) - J\big(\bar{\theta}_1 - c\mathbf{e}_i, X_{k+1}\big) + J\big(\kappa_1 - c\mathbf{e}_i, X_{k+1}\big)\big| \\
&\quad + \mathbb{E}_k I_3 \big| J\big(\kappa_1 + c\mathbf{e}_i, X_{k+1}\big) - J\big(\bar{\theta}_2 + c\mathbf{e}_i, X_{k+1}\big) - J\big(\kappa_1 - c\mathbf{e}_i, X_{k+1}\big) + J\big(\bar{\theta}_2 - c\mathbf{e}_i, X_{k+1}\big)\big| \\
&= O\big(|\bar{\theta}_1 - \kappa_1|\big) + O\big(|\kappa_1 - \bar{\theta}_2|\big) + O\big(c^2\big) \\
&= O\big(|\bar{\theta}_1 - \bar{\theta}_2|\big) + O\big(c^2\big).
\end{aligned}$$

Similarly, if I_4 is the indicator of the event where there is one intersection point and $\bar{\theta}_2$ is inside B_{D+1}, then we also get

$$\begin{aligned}
&\mathbb{E}_k I_4 \big| J\big(\bar{\theta}_1 + c\mathbf{e}_i, X_{k+1}\big) - J\big(\bar{\theta}_2 + c\mathbf{e}_i, X_{k+1}\big) - J\big(\bar{\theta}_1 - c\mathbf{e}_i, X_{k+1}\big) + J\big(\bar{\theta}_2 - c\mathbf{e}_i, X_{k+1}\big)\big| \\
&\qquad = O\big(|\bar{\theta}_1 - \bar{\theta}_2| + c^2\big).
\end{aligned}$$

Let I_5 denote the indicator of the case where both $\bar{\theta}_1, \bar{\theta}_2$ are outside B_{D+1} and there are two intersection points κ_1, κ_2. We get, as above,

$$\begin{aligned}
&\mathbb{E}_k I_5 \big| J\big(\bar{\theta}_1 + c\mathbf{e}_i, X_{k+1}\big) - J\big(\bar{\theta}_2 + c\mathbf{e}_i, X_{k+1}\big) - J\big(\bar{\theta}_1 - c\mathbf{e}_i, X_{k+1}\big) + J\big(\bar{\theta}_2 - c\mathbf{e}_i, X_{k+1}\big)\big| \\
&= O\big(|\bar{\theta}_1 - \kappa_1|\big) + O\big(|\kappa_1 - \kappa_2|\big) + O\big(|\kappa_2 - \bar{\theta}_2|\big) + O\big(c^2\big) \\
&= O\big(|\bar{\theta}_1 - \bar{\theta}_2| + c^2\big).
\end{aligned}$$

Finally, in the remaining case (where there is only one intersection point with B_{D+1} though both $\bar{\theta}_1, \bar{\theta}_2$ are outside B_{D+1}), we similarly get an estimate of the order $O\big(|\bar{\theta}_1 - \bar{\theta}_2| + c^2\big)$, and hence we eventually obtain the statement of the lemma.

When $m_p = 0$ and $m'_p = 1$, the same ideas work. When $m_p + m'_p > 1$ we can rely on the elementary observation that

$$\left| \prod_{j=1}^{m_p} \mathbb{1}_{\{X_{k+1} \le g^j(\bar{\theta}_1 + c\mathbf{e}_i)\}} - \prod_{j=1}^{m_p} \mathbb{1}_{\{X_{k+1} \le g^j(\bar{\theta}_2 + c\mathbf{e}_i)\}} \right| \le \sum_{j=1}^{m_p} \left| \mathbb{1}_{\{X_{k+1} \le g^j(\bar{\theta}_1 + c\mathbf{e}_i)\}} - \mathbb{1}_{\{X_{k+1} \le g^j(\bar{\theta}_2 + c\mathbf{e}_i)\}} \right|,$$

and on its counterpart for the h^j. Estimates can be repeated for each summand in the definition of J, so the case $m_s > 1$ follows, too. □

The arguments of the previous lemma, (8) in particular, also give us the following.

Lemma 3.2. *Under Assumptions 2.2 and 2.3, there is $C_\natural > 0$ such that, for each $i = 1, \dots, d$,*

$$\mathbb{E}_k \big| J\big(\bar{\theta} + c\mathbf{e}_i, X_{k+1}\big) - J\big(\bar{\theta} - c\mathbf{e}_i, X_{k+1}\big)\big| \le C_\natural c, \qquad 0 < c \le 1,$$

holds for all $k \in \mathbb{N}$ and for all $\mathcal{F}_k$-measurable B_{D+1}-valued random variables $\bar{\theta}$. □

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

3.1. Moment estimates

In this subsection, we will prove that the first moments of our iteration scheme remain bounded. This will be followed by other moment estimates. We start with a preliminary lemma on deterministic sequences.

Lemma 3.3. *Let $x_k \geq 0$, $k \in \mathbb{N}$, be a sequence; let $\zeta_k > 0$, $k \geq 1$, be another sequence. If they satisfy $\nu\zeta_k < 1$, $k \geq 1$, and*

$$x_k \leq (1 - \nu\zeta_k)x_{k-1} + \underline{c}\zeta_k, \qquad k \geq 1,$$

with some $\underline{c}$, $\nu > 0$, then

$$\sup_{k\in\mathbb{N}} x_k \leq x_0 + \frac{\underline{c}}{\nu}.$$

Proof. Following the argument of Lemma 1 in [6], we notice that

$$x_k \leq \prod_{i=1}^{k} (1 - \nu\zeta_i)x_0 + \underline{c} \sum_{i=1}^{k} \zeta_i \prod_{j=i+1}^{k} (1 - \nu\zeta_j),$$

where an empty product is meant to be 1. We can write

$$\begin{aligned}
&\sum_{i=1}^{k} \zeta_i \prod_{j=i+1}^{k} (1 - \nu\zeta_j) \\
&= \frac{1}{\nu} \sum_{i=1}^{k} \left(\prod_{j=i+1}^{k} (1 - \nu\zeta_j) - \prod_{j=i}^{k} (1 - \nu\zeta_j) \right) \\
&\leq \frac{1}{\nu}.
\end{aligned}$$

This shows the claim. □

Certain calculations are easier to carry out if we consider the continuous-time embedding of the discrete-time processes. Consider the following extension θ_t, $t \in \mathbb{R}_+$, of θ_k, $k \in \mathbb{N}$: let

$$\theta_t := \theta_k + \int_k^t a_u H(u, \theta_k) du$$

for all $k \in \mathbb{N}$ and for all $k \leq t < k + 1$, where $H(u, \theta) = H(\theta, X_{k+1}, c_k)$ for all $k \in \mathbb{N}$, and for all $k \leq u < k + 1$, $c_u = c_k$ and $a_u = \lambda_0 / \max\{u, 1\}$, $u \geq 0$. Extend the filtration to continuous time by $\mathcal{F}_t := \mathcal{F}_{\lceil t \rceil}$, $t \in \mathbb{R}_+$. Now fix $\mu > 1$. We introduce an auxiliary process that will play a crucial role in later estimates. For each $n \geq 1$ and for $\lceil n^\mu \rceil \leq t < \lceil (n+1)^\mu \rceil$, define $\bar{y}_t := y\big(t, \lceil n^\mu \rceil, \theta_{\lceil n^\mu \rceil}\big)$, i.e. the solution of (5) starting at $\lceil n^\mu \rceil$ with initial condition $\bar{y}_{\lceil n^\mu \rceil} = \theta_{\lceil n^\mu \rceil}$.

We introduce the L^1-norm

$$\|Z\|_1 := \mathbb{E}|Z|$$

for each $\mathbb{R}^d$-valued random variable Z.

Lemma 3.4. *Under Assumptions 2.2 and 2.3, we have*

$$\sup_{t\geq 1} \|\bar{y}_t\|_1 + \sup_{t\geq 1} E\|\theta_t\|_1 < \infty.$$

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

Proof. Note that $2c_k H^j(\theta, x, c_k) = l_0\big(\theta + c_k \mathbf{e}_j\big) - l_0\big(\theta - c_k \mathbf{e}_j\big)$, for all $x, j = 1, \ldots, d$, when $\theta \notin B_{D+1}$. Furthermore, the function $l_0\big(\theta + c_k \mathbf{e}_j\big) - l_0\big(\theta - c_k \mathbf{e}_j\big)$ is Lipschitz on B_{D+1}, which together with Lemma 3.2 implies

$$\left\| \frac{l_0\big(\theta + c_k \mathbf{e}_j\big) - l_0\big(\theta - c_k \mathbf{e}_j\big)}{2c_k} - H^j\big(\theta, X_{k+1}, c_k\big) \right\|_1 \leq \bar{C}, \qquad \theta \in \mathbb{R}^d, \tag{9}$$

for a fixed constant $\bar{C}$. Clearly,

$$\begin{aligned} ||\theta_{k+1}||_1 \leq & \left\| \theta_k - \frac{\lambda_k}{2c_k} \sum_{j=1}^{d} \big[l_0\big(\theta_k + c_k \mathbf{e}_j\big) - l_0\big(\theta_k - c_k \mathbf{e}_j\big) \big] \mathbf{e}_j \right\|_1 \\ & + \left\| \frac{\lambda_k}{2c_k} \sum_{j=1}^{d} \big[l_0\big(\theta_k + c_k \mathbf{e}_j\big) - l_0\big(\theta_k - c_k \mathbf{e}_j\big) \big] \mathbf{e}_j - \lambda_k H\big(\theta_k, X_{k+1}, c_k\big) \right\|_1 . \end{aligned}$$

Note that, by Assumption 2.2, l_0 is strongly convex; in particular,

$$\langle \nabla l_0(\theta) - \nabla l_0(0), \theta \rangle \geq A_0 |\theta|^2, \qquad \theta \in \mathbb{R}^d,$$

for all θ, with some $A_0 > 0$. Hence also

$$\langle \nabla l_0(\theta), \theta \rangle \geq A|\theta|^2 - B, \qquad \theta \in \mathbb{R}^d,$$

for suitable $A, B > 0$. But then for all $a > 0$ small enough,

$$|\theta - a \nabla l_0(\theta)| \leq (1 - A'a)|\theta| + aB', \qquad \theta \in \mathbb{R}^d,$$

for suitable $A', B' > 0$. By the mean value theorem,

$$l_0\big(\theta_k + c_k \mathbf{e}_j\big) - l_0\big(\theta_k - c_k \mathbf{e}_j\big) = 2c_k \partial_j l_0(\xi_j)$$

for some random variable $\xi_j \in \big[\theta_k - c_k \mathbf{e}_j, \theta_k + c_k \mathbf{e}_j\big]$. Since ∇l_0 is Lipschitz,

$$\max_j \big\| \nabla l_0(\theta_k) - \nabla l_0(\xi_j) \big\|_1 \leq L'$$

for some $L' > 0$. It then follows easily that, for $k \geq k_0$ large enough so that λ_k is small enough,

$$\begin{aligned} & \left\| \theta_k - \frac{\lambda_k}{2c_k} \sum_{j=1}^{d} \big[l_0\big(\theta_k + c_k \mathbf{e}_j\big) - l_0\big(\theta_k - c_k \mathbf{e}_j\big) \big] \mathbf{e}_j \right\|_1 \\ & \leq \lambda_k dL' + ||\theta_k - \lambda_k \nabla l_0(\theta_k)||_1 \\ & \leq \lambda_k \big(B' + dL'\big) + \big(1 - A'\lambda_k\big) ||\theta_k||_1 \end{aligned}$$

holds. By (9),

$$\left\| \frac{\lambda_k}{2c_k} \sum_{j=1}^{d} \big[l_0\big(\theta_k + c_k \mathbf{e}_j\big) - l_0\big(\theta_k - c_k \mathbf{e}_j\big) \big] \mathbf{e}_j - \lambda_k H\big(\theta_k, X_{k+1}, c_k\big) \right\|_1 \leq \lambda_k d \bar{C}.$$

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

Apply Lemma 3.3 with the choice $x_k := ||\theta_k||_1$, $\underline{c} := d(L' + \bar{C}) + B'$ and $\zeta_k := \lambda_k$, $\nu := A'$ to obtain that $\sup_{k \geq k_0} \|\theta_k\|_1 < \infty$. Then trivially also $\sup_{n \in \mathbb{N}} \|\theta_n\|_1 < \infty$ holds, which easily implies $\sup_{t \geq 1} \|\theta_t\|_1 < \infty$ as well.

Now turning to $\bar{y}_t$ we see that, for $n \geq 1$ and $\lceil n^\mu \rceil \leq t < \lceil (n+1)^\mu \rceil$,

$$\begin{aligned} |\bar{y}_t - \theta^*| &= \left|\bar{y}_t - y\left(t, \lceil n^\mu \rceil, \theta^*\right)\right| \\ &\leq |\theta_{\lceil n^\mu \rceil} - \theta^*| C^*, \end{aligned}$$

finishing the proof. □

Lemma 3.5. *Let Assumptions 2.2 and 2.3 hold. Then there exists $C_l > 0$ such that $\sup_{k \geq 1} |J(\theta, X_k)| \leq C_l\left(1 + |\theta|^2\right)$.*

Proof. Recall that

$$|J(\theta, x)| \leq |l_0(\theta)| + \sum_{i=1}^{m_s} \mathbb{1}_{A_i(x)} |l_i(x, \theta)|,$$

where the functions l_i are bounded on the bounded sets $\cup_{x \in \mathbb{R}^d} A_i(x)$ for $i = 1, \ldots, d$, and l_0 grows quadratically. □

The difficulty of the following lemma consists in handling the discontinuities and the dependence of the sequence X_k at the same time.

Lemma 3.6. *Let Assumptions 2.2 and 2.3 hold. Then for each $R > 0$ the random field $J(\theta, X_n)$, $\theta \in B_R$, $n \in \mathbb{N}$, satisfies*

$$\begin{aligned} M_3^n(J(\theta, X)) &\leq C_l\left(1 + R^2\right), \\ \Gamma_3^n(J(\theta, X)) &\leq L\left(1 + R^2\right), \end{aligned}$$

for some $L > 0$, where C_l is as in Lemma 3.5.

Proof. The first statement is clear from Lemma 3.5. Let $n \geq 0$, $\tau \geq 1$ be fixed. For $k \geq \tau$, define $X_k^+ = \mathbb{E}\left[X_{n+k} | \mathcal{F}_{n+k-\tau}^+ \vee \mathcal{F}_n\right]$. For the sake of simplicity, we assume that $m_s = 1$ in the definition of J, $m_p = 0$, but the same argument would work for several summands, too. We also take the process X unidimensional ($m := 1$), noting that the same arguments easily carry over to a general m.

We now perform an auxiliary estimate. Let $\epsilon_\tau > 0$ be a parameter to be chosen later and let $1 \leq j \leq m_p'$. We will write h below instead of h_1. Define $Z_k = X_{n+k} - X_k^+$ and estimate

$$\begin{aligned} &\mathbb{E}_n \left| \mathbb{1}_{\left\{X_{n+k} > h^j(\theta)\right\}} - \mathbb{1}_{\left\{X_k^+ > h^j\left(\theta\right)\right\}} \right|^3 = \mathbb{E}_n \left| \mathbb{1}_{\left\{X_{n+k} > h^j(\theta)\right\}} - \mathbb{1}_{\left\{X_k^+ > h^j\left(\theta\right)\right\}} \right| \\ &\leq \mathbb{P}_n \left(X_{n+k} \in \left(h^j(\theta) - |Z_k|, h^j(\theta) + |Z_k|\right)\right) \\ &\leq \mathbb{P}_n \left(X_{n+k} \in \left(h^j(\theta) - |Z_k|, h^j(\theta) + |Z_k|\right), |Z_k| \leq \epsilon_\tau \right) + \mathbb{P}_n(|Z_k| \geq \epsilon_\tau) \\ &\leq 2F\epsilon_\tau + \frac{\mathbb{E}_n\left[\left|X_{n+k} - X_k^+\right|^3\right]}{\epsilon_\tau^3}, \end{aligned}$$

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

where the last inequality follows from Assumption 2.3 and the Markov inequality. Now estimate

$$\begin{aligned}
&\mathbb{E}_n^{1/3}\left|\left(\prod_{j=1}^{m_p'} 1_{\{X_{n+k}>h^j(\theta)\}}\right) l_1(X_{n+k},\theta) - \left(\prod_{j=1}^{m_p'} 1_{\{X_k^+>h^j(\theta)\}}\right) l_1\big(X_k^+,\theta\big)\right|^3 \\
&\le \mathbb{E}_n^{1/3}\left|\left(\prod_{j=1}^{m_p'} 1_{\{X_{n+k}>h^j(\theta)\}} - \prod_{j=1}^{m_p'} 1_{\{X_k^+>h^j(\theta)\}}\right) l_1(X_{n+k},\theta)\right|^3 \\
&\quad + \mathbb{E}_n^{1/3}\left|\big(l_1(X_{n+k},\theta) - l_1\big(X_k^+,\theta\big)\big)\prod_{j=1}^{m_p'} 1_{\{X_k^+>h^j(\theta)\}}\right|^3 \\
&\le \mathbb{E}_n^{1/3}\left[\left|\sum_{j=1}^{m_p'}\left|1_{\{X_{n+k}>h^j(\theta)\}} - 1_{\{X_k^+>h^j(\theta)\}}\right|\right|^3 (L_1(|X_{n+k}|+R)+|l_1(0,0)|)^3\right] \\
&\quad + L_1\mathbb{E}_n^{1/3}\big|X_{n+k}-X_k^+\big|^3 \\
&\le C_1(1+R)\left(\epsilon_\tau^{1/3} + \frac{\mathbb{E}_n^{1/3}\Big[\big|X_{n+k}-X_k^+\big|^3\Big]}{\epsilon_\tau}\right)
\end{aligned}$$

for some C_1, where we used the Lipschitz-continuity of the function l_1, as well as the observation that

$$\left|\prod_{j=1}^{m_p'} 1_{\{X_{n+k}>h^j(\theta)\}} - \prod_{j=1}^{m_p'} 1_{\{X_k^+>h^j(\theta)\}}\right| \le \sum_{j=1}^{m_p'}\left|1_{\{X_{n+k}>h^j(\theta)\}} - 1_{\{X_k^+>h^j(\theta)\}}\right|.$$

A similar estimate works for l_0, but we get the upper bound

$$\begin{aligned}
&\mathbb{E}_n^{1/3}\left|\mathbb{1}_{A_0(X_{n+k})}l_0(\theta) - \mathbb{1}_{A_0\left(X_k^+\right)}l_0(\theta)\right|^3 \\
&\le C_1\big(1+R^2\big)\left(\epsilon_\tau^{1/3} + \frac{\mathbb{E}_n^{1/3}\big[|X_{n+k}-X_k^+|^3\big]}{\epsilon_\tau}\right)
\end{aligned}$$

instead. For the second inequality of the present lemma, note first that Lemma 4.1 below implies

$$\begin{aligned}
&\mathbb{E}_n^{1/3}\left[\left|J\big(\theta,X_{n+k}\big) - \mathbb{E}\big[J\big(\theta,X_{n+k}\big)|\mathcal{F}_n\vee\mathcal{F}_{n+k-\tau}^+\big]\right|^3\right] \\
&\le 2\mathbb{E}_n^{1/3}\left[\left|J\big(\theta,X_{n+k}\big) - J\big(\theta,X_k^+\big)\right|^3\right];
\end{aligned}$$

hence it suffices to estimate the latter quantity. From our previous estimates it follows that, for some $C>0$,

$$\mathbb{E}_n^{1/3}\left[\left|J\big(\theta,X_{n+k}\big) - J\big(\theta,X_k^+\big)\right|^3\right] \le C\big(1+R^2\big)\left[\sqrt[3]{\epsilon_\tau} + \frac{\mathbb{E}_n^{1/3}\big|X_{n+k}-X_k^+\big|^3}{\epsilon_\tau}\right]. \tag{10}$$

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

Choose $\epsilon_\tau := (\tau+1)^{-3-\epsilon/2}$. Summing up the right-hand side for $\tau \geq 1$ we see that, by Assumption 2.4, the sum has an upper bound independent of k. The statement follows as the case $\tau = 0$ is easy. □

3.2. Decreasing-gain case

The following lemma contains the core estimates of the present paper.

Lemma 3.7. *Let $n \geq 1$. Let $\lceil n^\mu \rceil \leq t < \lceil (n+1)^\mu \rceil$ for $\mu := 1/\gamma$ and let Assumptions 2.1, 2.2, 2.3, 2.4, 2.5, and 2.6 hold. Then $\mathbb{E}|\theta_t - \bar{y}_t| = O\big(n^{-\beta}\big)$, where $\beta = \min\Big(\frac{1}{2\gamma} - \frac{1}{2}, 2\Big)$.*

Proof. For $\lceil n^\mu \rceil \leq t < \lceil (n+1)^\mu \rceil$,

$$
\begin{aligned}
|\theta_{\lfloor t \rfloor} - \bar{y}_t| &\leq |\bar{y}_{\lfloor t \rfloor} - \bar{y}_t| + |\theta_{\lfloor t \rfloor} - \bar{y}_{\lfloor t \rfloor}| \\
&\leq \int_{\lfloor t \rfloor}^{t} a_u \left| G\big(\bar{y}_u\big) \right| \, du + \left| \int_{\lceil n^\mu \rceil}^{\lfloor t \rfloor} a_u \big(H(u, \theta_{\lfloor u \rfloor}) - G\big(\bar{y}_u\big) \big) \, du \right| \\
&\leq a_{n^\mu} \int_{\lfloor t \rfloor}^{t} \left| G\big(\bar{y}_u\big) \right| \, du \\
&\quad + \left| \int_{\lceil n^\mu \rceil}^{\lfloor t \rfloor} a_u \left(H(u, \theta_{\lfloor u \rfloor}) - H\big(u, \bar{y}_u\big) \right) \, du \right| \\
&\quad + \left| \int_{\lceil n^\mu \rceil}^{\lfloor t \rfloor} a_u \left(H\big(u, \bar{y}_u\big) - \mathbb{E}\big[H\big(u, \bar{y}_u\big) | \mathcal{F}_{\lceil n^\mu \rceil} \big] \right) \, du \right| \\
&\quad + \left| \int_{\lceil n^\mu \rceil}^{\lfloor t \rfloor} a_u \left(\mathbb{E}\big[H\big(u, \bar{y}_u\big) | \mathcal{F}_{\lceil n^\mu \rceil} \big] - G\big(\bar{y}_u\big) \right) \, du \right| \\
&=: \Sigma_0 + \Sigma_1 + \Sigma_2 + \Sigma_3.
\end{aligned}
$$

Estimation of Σ_0. Since G has at most linear growth, Lemma 3.4 guarantees that

$$
\mathbb{E}[\Sigma_0] = O\left(a_{n^\mu} \int_{\lfloor t \rfloor}^{t} (\mathbb{E}|\bar{y}_u| + 1) \, du \right) = O(n^{-\mu}).
$$

Estimation of Σ_1. Recall that, by the tower property for conditional expectations,

$$
\mathbb{E} \left| H(u, \theta_u) - H\big(u, \bar{y}_u\big) \right| = \mathbb{E}\mathbb{E}_k \left| H(u, \theta_u) - H\big(u, \bar{y}_u\big) \right|
$$

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

for all $k \in \mathbb{N}$. Applying this observation to $k = \lfloor u \rfloor$, Lemma 3.1 implies that

$$\begin{aligned}\mathbb{E}[\Sigma_1] &= \mathbb{E}\left| \int_{\lceil n^\mu \rceil}^{\lfloor t \rfloor} a_u \left(H(u, \theta_{\lfloor u \rfloor}) - H\big(u, \bar{y}_u\big) \right) du \right| \\ &\le \int_{\lceil n^\mu \rceil}^{t} a_u \mathbb{E} \left| H(u, \theta_{\lfloor u \rfloor}) - H\big(u, \bar{y}_u\big) \right| du \\ &\le C_\flat \int_{\lceil n^\mu \rceil}^{t} \frac{a_u}{c_u} \mathbb{E} \left| \theta_{\lfloor u \rfloor} - \bar{y}_u \right| du + C_\flat \int_{\lceil n^\mu \rceil}^{t} \frac{a_u}{c_u} c_u^2 \, du. \end{aligned} \tag{11}$$

Henceforth we will write

$$\Sigma_1' := C_\flat \int_{\lceil n^\mu \rceil}^{\lceil (n+1)^\mu \rceil} \frac{a_u}{c_u} c_u^2 \, du.$$

Notice that

$$\mathbb{E}\big[\Sigma_1'\big] = O\big(n^{-\mu\gamma-1}\big) = O\big(n^{-2}\big).$$

Estimation of Σ_2. Notice that $H(u, \bar{\theta}) = \mathbb{E}\big[H(u, \bar{\theta}) | \mathcal{F}_{\lceil n^\mu \rceil}\big]$ for all $\mathcal{F}_{\lceil n^\mu \rceil}$-measurable $\bar{\theta}$ such that, almost surely, $\bar{\theta} \notin B_D$, since $J(\theta, x)$ does not depend on x outside B_D by Assumption 2.2. Thus

$$\Sigma_2 \le \sup_{\lceil n^\mu \rceil \le t < \lceil (n+1)^\mu \rceil} \left| \int_{\lceil n^\mu \rceil}^{t} a_u \mathbb{1}_{\{\bar{y}_u \in B_D\}} \left(H\big(u, \bar{y}_u\big) - \mathbb{E}[H\big(u, \bar{y}_u\big) | \mathcal{F}_{n^\mu}] \right) du \right|.$$

We will use the inequality of Theorem 4.1 below with $r = 3$, with $\mathcal{R}_t := \mathcal{F}_{t+\lceil n^\mu \rceil}$, $t \in \mathbb{R}_+$, $\mathcal{R}_t^+ := \mathcal{F}^+_{t+\lceil n^\mu \rceil}$, with the process defined by

$$W_t = \mathbb{1}_{\{\bar{y}_{t+\lceil n^\mu \rceil} \in B_D\}} c_{t+\lceil n^\mu \rceil} \left(H\big(t, \bar{y}_{t+\lceil n^\mu \rceil}\big) - \mathbb{E}\big[H\big(t, \bar{y}_{t+\lceil n^\mu \rceil}\big) | \mathcal{F}_{\lceil n^\mu \rceil}\big] \right), \ t \ge 0, \tag{12}$$

and with the function $f_t = a_{t+\lceil n^\mu \rceil} / c_{t+\lceil n^\mu \rceil}$. Note that $\{\bar{y}_t \in B_D\} \in \mathcal{F}_{\lceil n^\mu \rceil}$ for all $\lceil n^\mu \rceil \le t < \lceil (n+1)^\mu \rceil$. We get from Lemma 4.2 below and from the cited inequality that

$$\begin{aligned}\mathbb{E}[\Sigma_2] &= \mathbb{E}\big[\mathbb{E}\big[\Sigma_2 | \mathcal{F}_{\lceil n^\mu \rceil}\big]\big] \le \mathbb{E}\big[\mathbb{E}^{1/3}\big[\Sigma_2^3 \big| \mathcal{F}_{\lceil n^\mu \rceil}\big]\big] \\ &\le C'(3) \left(\int_{\lceil n^\mu \rceil}^{\lceil (n+1)^\mu \rceil} \left(\frac{a_u}{c_u} \right)^2 du \right)^{1/2} \mathbb{E}\big[\tilde{M}_3 + \tilde{\Gamma}_3\big] \\ &\le C'(3) \left(\int_{\lceil n^\mu \rceil}^{\lceil (n+1)^\mu \rceil} \left(\frac{a_u}{c_u} \right)^2 du \right)^{1/2} C\big(1 + D^2\big). \end{aligned}$$

We thus get

$$\mathbb{E}[\Sigma_2] = O\left(\int_{\lceil n^\mu \rceil}^{\lceil (n+1)^\mu \rceil} \left(\frac{a_u}{c_u} \right)^2 du \right)^{1/2} = O\Big(n^{\frac{-\mu+2\mu\gamma-1}{2}}\Big).$$

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

Estimation of Σ_3. We have

$$\begin{aligned}\mathbb{E}[\Sigma_3] &\le \mathbb{E}\left[\int_{\lceil n^\mu\rceil}^t a_u \left|\mathbb{E}\big[H\big(u,\bar{y}_u\big)|\mathcal{F}_{\lceil n^\mu\rceil}\big] - G\big(\bar{y}_u\big)\right| du\right] \\ &\le \mathbb{E}\left[\int_{\lceil n^\mu\rceil}^t a_u \sup_{\vartheta\in\mathbb{R}^d}\left|\mathbb{E}\big[H\big(\vartheta, X_{\lfloor u\rfloor+1}, c_{\lfloor u\rfloor}\big)|\mathcal{F}_{\lceil n^\mu\rceil}\big] - \mathbb{E}\big[H\big(\vartheta, X_{\lfloor u\rfloor+1}, c_{\lfloor u\rfloor}\big)\big]\right| du\right] \\ &\quad + \mathbb{E}\left[\int_{\lceil n^\mu\rceil}^t a_u \sup_{\vartheta\in\mathbb{R}^d}\left|\mathbb{E}\big[H\big(\vartheta, X_{\lfloor u\rfloor+1}, c_{\lfloor u\rfloor}\big)\big] - G(\vartheta)\right| du\right]. \end{aligned} \tag{13}$$

To handle the second sum, note that, for each $i = 1, \dots, d$,

$$\mathbb{E}[H^i(\vartheta, X_{k+1}, c_k)] = \frac{U(\vartheta + c_k\mathbf{e}_i) - U(\vartheta - c_k\mathbf{e}_i)}{2c_k} = G^i\big(\xi_k^i\big)$$

for some $\xi_k^i \in [\vartheta - c_k\mathbf{e}_i, \vartheta + c_k\mathbf{e}_i]$. The Lipschitz-continuity of G implies that $|G^i\big(\xi_k^i\big) - G^i(\vartheta)| \le L_G c_k$, so

$$\begin{aligned}\mathbb{E}\left[\int_{\lceil n^\mu\rceil}^t a_u \sup_{\vartheta\in\mathbb{R}^d}\left|\mathbb{E}\big[H\big(\vartheta, X_{\lfloor u\rfloor+1}, c_{\lfloor u\rfloor}\big)\big] - G(\vartheta)\right| du\right] &\le \int_{\lceil n^\mu\rceil}^t a_u dL_G c_{\lceil u\rceil}\, du \\ &= O\left(\int_{\lceil n^\mu\rceil}^{\lceil (n+1)^\mu\rceil} u^{-1-\gamma}\, du\right) = O\big(n^{-2}\big).\end{aligned}$$

Now we turn to the first sum in (13). Define $X_k^+ = \mathbb{E}\big[X_k|\mathcal{F}^+_{\lceil n^\mu\rceil}\big]$, $k \ge \lceil n^\mu\rceil$. First let us estimate

$$\mathbb{E}_{\lceil n^\mu\rceil}\big[\big|H\big(\vartheta, X_{k+1+\lceil n^\mu\rceil}, c_{k+\lceil n^\mu\rceil}\big) - H\big(\vartheta, X^+_{k+1+\lceil n^\mu\rceil}, c_{k+\lceil n^\mu\rceil}\big)\big|\big].$$

Fix $\epsilon_k > 0$ to be chosen later. By an argument similar to that of Lemma 3.6 (using the first instead of the third moment in Markov's inequality) we get that, for some constant C_1,

$$\begin{aligned}&c_{k+\lceil n^\mu\rceil}\mathbb{E}_{\lceil n^\mu\rceil}\big[\big|H\big(\vartheta, X_{\lceil n^\mu\rceil+k+1}, c_{k+\lceil n^\mu\rceil}\big) - H\big(\vartheta, X^+_{\lceil n^\mu\rceil+k+1}, c_{k+\lceil n^\mu\rceil}\big)\big|\big] \\ &\le C_1\left[\epsilon_k + \frac{\mathbb{E}_{\lceil n^\mu\rceil}\big[\big|X_{\lceil n^\mu\rceil+k+1} - X^+_{\lceil n^\mu\rceil+k+1}\big|\big]}{\epsilon_k}\right].\end{aligned}$$

Choose $\epsilon_k = (1+k)^{-1-\epsilon/2}$. Then using Assumption 2.4 we get

$$c_{k+\lceil n^\mu\rceil}\sup_{\vartheta\in\mathbb{R}^d}\mathbb{E}\big[\big|H\big(\vartheta, X_{\lceil n^\mu\rceil+k+1}, c_{\lceil n^\mu\rceil+k}\big) - H\big(\vartheta, X^+_{\lceil n^\mu\rceil+k+1}, c_{k+\lceil n^\mu\rceil}\big)\big|\big|\mathcal{F}_{\lceil n^\mu\rceil}\big] = O\big(k^{-1-\epsilon/2}\big),$$

which also implies

$$c_{k+\lceil n^\mu\rceil}\sup_{\vartheta\in\mathbb{R}^d}\mathbb{E}[|H(\vartheta, X_{\lceil n^\mu\rceil+k+1}, c_{\lceil n^\mu\rceil+k}) - H(\vartheta, X^+_{\lceil n^\mu\rceil+k+1}, c_{\lceil n^\mu\rceil+k})|] = O\big(k^{-1-\epsilon/2}\big).$$

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

Since

$$\mathbb{E}\big[H\big(\vartheta, X_{k+1}^{+}, c_{\lceil n^{\mu}\rceil+k}\big)|\mathcal{F}_{\lceil n^{\mu}\rceil}\big] = \mathbb{E}\big[H\big(\vartheta, X_{k+1}^{+}, c_{k+\lceil n^{\mu}\rceil}\big)\big]$$

for $k \geq \lceil n^{\mu}\rceil$ by independence of $\mathcal{F}_{\lceil n^{\mu}\rceil}$ and $\mathcal{F}_{\lceil n^{\mu}\rceil}^{+}$, we have

$$\begin{aligned}
&\left[\int_{\lceil n^{\mu}\rceil}^{t} a_u \mathbb{E}\sup_{\vartheta\in\mathbb{R}^d}\left|\mathbb{E}\big[H\big(\vartheta, X_{\lfloor u\rfloor+1}, c_{\lfloor u\rfloor}\big)|\mathcal{F}_{\lceil n^{\mu}\rceil}\big] - \mathbb{E}\big[H\big(\vartheta, X_{\lfloor u\rfloor+1}, c_{\lfloor u\rfloor}\big)\big]\right| du\right] \\
&\leq \int_{\lceil n^{\mu}\rceil}^{\infty} \frac{a_u}{c_u} c_u \mathbb{E}\sup_{\vartheta\in\mathbb{R}^d}\mathbb{E}\left[\left|H\big(\vartheta, X_{\lfloor u\rfloor+1}, c_k\big) - H\Big(\vartheta, X_{\lfloor u\rfloor+1}^{+}, c_k\Big)\right| |\mathcal{F}_{\lceil n^{\mu}\rceil}\right] du \\
&\quad + \int_{\lceil n^{\mu}\rceil}^{\infty} \frac{a_u}{c_u} c_u \sup_{\vartheta\in\mathbb{R}^d}\mathbb{E}\left[\left|H\big(\vartheta, X_{\lfloor u\rfloor+1}, c_k\big) - H\Big(\vartheta, X_{\lfloor u\rfloor+1}^{+}, c_k\Big)\right|\right] du \\
&\leq C_2 \frac{a_{\lceil n^{\mu}\rceil}}{c_{\lceil n^{\mu}\rceil}} \sum_{k=1}^{\infty} k^{-1-\epsilon/2}
\end{aligned}$$

with some C_2 so

$$\mathbb{E}[\Sigma_3] = O\big(n^{\mu(\gamma-1)}\big).$$

Combining the estimates we have so far, we get

$$E\big[\Sigma_0 + \Sigma_1' + \Sigma_2 + \Sigma_3\big] = O\Big(n^{-\mu} + n^{-2} + n^{\frac{-\mu+2\mu\gamma-1}{2}} + n^{-2} + n^{\mu(\gamma-1)}\Big). \tag{14}$$

Notice that $\mathbb{E}|\theta_t - \bar{y}_t|$ is always finite; see Lemma 3.4 above. Use Gronwall's lemma and (11) to obtain the inequality

$$\mathbb{E}\big[|\theta_{\lfloor t\rfloor} - \bar{y}_t|\big] \leq E[\Sigma_0 + \Sigma_1' + \Sigma_2 + \Sigma_3] \exp\left(C_3 \int_{n^{\mu}}^{\lceil (n+1)^{\mu}\rceil} \frac{a_u}{c_u}\, du\right)$$

with some constant C_3. From Lemma 3.2 it is also easy to check that $\mathbb{E}|\theta_t - \theta_{\lfloor t\rfloor}| = O(n^{-\mu})$. Note furthermore that the terms $n^{-\mu}$ and $n^{\mu(\gamma-1)}$ are always negligible in (14). These observations lead to

$$\begin{aligned}
&\mathbb{E}|\theta_t - \bar{y}_t| \\
&= O\Big(n^{\frac{-\mu+2\mu\gamma-1}{2}} + n^{-2}\Big) \exp\big(C_4 n^{\mu\gamma-1}\big) \\
&= O\Big(n^{\frac{1}{2}-\frac{1}{2\gamma}} + n^{-2}\Big)
\end{aligned}$$

with some C_4, finishing the proof. □

Proof of Theorem 2.1. Let

$$d_i = \sup_{\lceil i^{\mu}\rceil \leq s < \lceil (i+1)^{\mu}\rceil} \mathbb{E}|\theta_s - \bar{y}_s|, \qquad i = 1, 2, \ldots.$$

By Fatou's lemma, we also have

$$\mathbb{E}\big|\theta_{\lceil (i+1)^{\mu}\rceil} - \bar{y}_{\lceil (i+1)^{\mu}\rceil -}\big| \leq d_i,$$

where $\bar{y}_{s-}$ denotes the left limit of $\bar{y}$ at s.

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

It follows from Lemma 3.7 that $d_i = O(i^{-\beta})$. Combining this with Assumption 2.6 and using telescoping sums we get, for each integer $N \geq 1$,

$$\begin{aligned}\mathbb{E}|y(\lceil N^\mu \rceil, 1, \theta_1) - \theta_{\lceil N^\mu \rceil}| &= \mathbb{E}|y(\lceil N^\mu \rceil, 1, \theta_1) - y(\lceil N^\mu \rceil, \lceil N^\mu \rceil, \theta_{\lceil N^\mu \rceil})| \\ &\leq \sum_{i=2}^{N} \mathbb{E}\left|y(\lceil N^\mu \rceil, \lceil (i-1)^\mu \rceil, \theta_{\lceil (i-1)^\mu \rceil}) - y(\lceil N^\mu \rceil, \lceil i^\mu \rceil, \theta_{\lceil i^\mu \rceil})\right| \\ &\leq \sum_{i=2}^{N} \mathbb{E}\Big|y(\lceil N^\mu \rceil, \lceil i^\mu \rceil, y(\lceil i^\mu \rceil, \lceil (i-1)^\mu \rceil, \theta_{\lceil (i-1)^\mu \rceil})) \\ &\qquad - y(\lceil N^\mu \rceil, \lceil i^\mu \rceil, \theta_{\lceil i^\mu \rceil})\Big| \\ &\leq C^* \sum_{i=2}^{N} \left(\frac{i+1}{N}\right)^{\alpha\mu} d_{i-1} = O\big(N^{-\beta+1}\big),\end{aligned}$$

noting that $y(\lceil i^\mu \rceil, \lceil (i-1)^\mu \rceil, \theta_{\lceil (i-1)^\mu \rceil})$ equals the left limit $\bar{y}_{\lceil i^\mu \rceil -}$. A similar argument provides, for all $t \in (\lceil N^\mu \rceil, \lceil (N+1)^\mu \rceil)$,

$$\mathbb{E}|\theta_t - y(t, 1, \theta_1)| = O\big(N^{-\beta+1}\big).$$

Taking the μth root we obtain

$$\mathbb{E}|\theta_t - y(t, 1, \theta_1)| = O\Big(t^{\frac{-\beta+1}{\mu}}\Big), \qquad t \geq 1.$$

To conclude, note that by the stability Assumption 2.6, $|y(t, 1, \theta_1) - \theta^*| \leq C^*|\theta_1 - \theta^*|t^{-\alpha}$ and that $\mathbb{E}|\theta_1| < \infty$, as easily seen using Lemma 3.2. □

3.3. Fixed-gain stochastic approximation

Define $T = \frac{c}{a}$. For $nT \leq t < (n+1)T$, define $\bar{y}_t = y(t, nT, \theta_{nT})$, i.e. the solution of (5) with the initial condition $y_{nT} = \theta_{nT}$. We use the piecewise linear extension $\overline{\theta}_t$ of θ_t and the piecewise constant extension $H(t, \theta)$ of $H(\theta, X_{k+1}, c)$ as defined in the decreasing-gain setting, but a and c are now constants.

Lemma 3.8. *Let Assumptions 2.1, 2.2, 2.3, 2.4 and 2.6 hold. Then for $t \in [nT, (n+1)T]$ there is $\overline{C} > 0$ such that $\mathbb{E}|\theta_t - \bar{y}_t| \leq \overline{C} \max\left(c^2, \sqrt{\frac{a}{c}}\right)$.*

Proof. Using essentially the same estimates we derived in the decreasing-gain setting, for fixed a and c we get

$$\mathbb{E}[\Sigma_0] \leq C_0 a, \tag{15}$$

$$\mathbb{E}[\Sigma_1] \leq C_1 \left[\frac{a}{c} \sum_{nT}^{t-1} \mathbb{E}|\theta_k - \overline{y_k}| + c^2\right], \tag{16}$$

$$\mathbb{E}[\Sigma_2] \leq C_2 \left(\sum_{nT}^{t-1} \left(\frac{a^2}{c^2}\right)\right)^{1/2} \leq C_2 \left(\frac{c}{a}\frac{a^2}{c^2}\right)^{1/2} = C_2\sqrt{\frac{a}{c}}, \tag{17}$$

$$\mathbb{E}[\Sigma_3] \leq C_3 \left[\frac{a}{c} + \sum_{nT}^{t-1} ac\right] = C_3 Tac + C_3\frac{a}{c} = O\left(c^2 + \frac{a}{c}\right), \tag{18}$$

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

with suitable constants C_0, C_1, C_2, C_3. Combine these estimates and use Gronwall's lemma to get the statement. To choose optimally, set $c^2 = \sqrt{\frac{a}{c}}$, that is, $c = a^{\frac{1}{5}}$. In this case $\mathbb{E}|\theta_t - \overline{y}_t| \le C_4 a^{\frac{2}{5}}$ for some C_4. □

Proof of Theorem 2.2. Let

$$d_i = \sup_{iT \le s < (i+1)T} \mathbb{E}|\theta_s - \overline{y}^i_s|.$$

It follows from Lemma 3.8 that $d_i \le \overline{C} \max\left(c^2, \sqrt{\frac{a}{c}}\right)$. Combining this with Assumption 2.6 and using telescoping sums we get

$$\begin{aligned}
\mathbb{E}|y(NT, 1, \theta_1) - \theta_{NT}| &= \mathbb{E}|y(NT, 1, \theta_1) - y(NT, NT, \theta_{NT})| \\
&\le \sum_{i=2}^{N} \mathbb{E}\left|y(NT, (i-1)T, \theta_{(i-1)T}) - y(NT, iT, \theta_{iT})\right| \\
&\le \sum_{i=2}^{N} \mathbb{E}\left|y(NT, iT, y(iT, (i-1)T, \theta_{(i-1)T})) - y(NT, iT, \theta_{iT})\right| \\
&\le \sum_{i=2}^{N} \left(C^* e^{-a\alpha(NT-iT)}\right) d_{i-1} \le \hat{C} \max\left(c^2, \sqrt{\frac{a}{c}}\right)
\end{aligned}$$

with some $\hat{C}$, since $\sum_{i=2}^{N} e^{-a\alpha(NT-iT)}$ has an upper bound independent of N. We similarly get

$$\sup_{NT \le t < (N+1)T} \mathbb{E}|\theta_t - y(t, 1, \theta_1)| \le \check{C} \max\left(c^2, \sqrt{\frac{a}{c}}\right)$$

with some $\check{C}$. To conclude, note that by the stability Assumption 2.6, $|y(t, 1, \theta_1) - \theta^*| \le C^*|\theta_1 - \theta^*| e^{-a\alpha t}$ and therefore

$$\mathbb{E}|\theta_t - \theta^*| = O\left(\max\left(c^2, \sqrt{\frac{a}{c}}\right) + e^{-a\alpha t}\right).$$ □

4. Auxiliary results

We define continuous-time analogues of the key quantities M and Γ from Assumption 2.4 and establish a pivotal maximal inequality for them.

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

We consider an $\mathbb{R}^d$-valued continuous-time stochastic process $(W_t)_{t \in \mathbb{R}_+}$ 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 $W_t \in L^1$, $t \in \mathbb{R}_+$. Fix $r \ge 1$. We define the quantities

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

and set $\tilde{\Gamma}_r := \sum_{\tau=0}^{\infty} \tilde{\gamma}_r(\tau)$.

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

Now we recall a powerful maximal inequality, Theorem B.3 of [1].

Theorem 4.1. *Let $(W_t)_{t\in\mathbb{R}_+}$ be L^r-bounded for some $r > 2$ and let $\tilde{M}_r + \tilde{\Gamma}_r < \infty$ almost surely. Assume $\mathbb{E}[W_t|\mathcal{R}_0] = 0$ almost surely for $t \in \mathbb{R}_+$. Let $f : [0, T] \to \mathbb{R}$ be $\mathcal{B}([0, T])$-measurable with $\int_0^T f_t^2\, dt < \infty$. Then there is a constant $C'(r)$ such that*

$$\mathbb{E}^{1/r}\left[\sup_{s\in[0,T]}\left|\int_0^s f_t W_t\, dt\right|^r |\mathcal{R}_0\right] \le C'(r)\left(\int_0^T f_t^2\, dt\right)^{1/2}\left[\tilde{M}_r + \tilde{\Gamma}_r\right], \tag{19}$$

almost surely. □

We also recall Lemma A.1 of [4].

Lemma 4.1. *Let $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ be sigma-algebras. Let $X, Y \in \mathbb{R}^d$ be random variables in L^p such that Y is measurable with respect to $\mathcal{H} \vee \mathcal{G}$. Then for any $p \ge 1$,*

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

Lemma 4.2. *Let the process W_t be defined by (12). Taking the filtration $\mathcal{R}_t := \mathcal{F}_{t+\lceil n^\mu\rceil}$ and $\mathcal{R}_t^+ := \mathcal{F}^+_{t+\lceil n^\mu\rceil}$, we get $\tilde{M}_r + \tilde{\Gamma}_r \le C\left(1 + D^2\right)$ for some $C > 0$.*

Proof. Estimations of Lemma 3.6 with $R = D$ and with $\bar{y}_{t+\lceil n^\mu\rceil}$ instead of θ imply the statement. □

5. Numerical experiments

In what follows we present numerical results to check the convergence of the algorithm for a simple discontinuous function J, defined as

$$J(\theta, X) = \begin{cases} (\theta - X)^2 + 1 & \text{if } X \le \theta, \\ (\theta - X)^2 & \text{otherwise,} \end{cases}$$

where X is a square-integrable, absolutely continuous random variable. Clearly, this function is not continuous in the parameter, but its expectation *is* continuous:

$$\begin{aligned} U(\theta) = \mathbb{E}J(X, \theta) &= \int_{-\infty}^{\theta}\left(\left(x - \theta^2\right) + 1\right)f(x)dx + \int_{\theta}^{\infty}(x - \theta)^2 f(x)dx \\ &= \mathbb{E}(X - \theta)^2 + F(\theta) = \mathbb{E}X^2 - 2\theta\mathbb{E}X + \theta^2 + F(\theta), \end{aligned}$$

where $f(\cdot)$ and $F(\cdot)$ are respectively the density function and the distribution function of X. Assuming that F is differentiable, we need to solve

$$\frac{\partial U(\theta)}{\partial\theta} = -2\mathbb{E}X + 2\theta - f(\theta) = 0$$

in order to find $\arg\min \mathbb{E}J(X, \theta)$.

For the numerical examples we will use the recursion

$$\theta_{k+1} = \theta_k + \frac{1}{k + k_0}\frac{J\left(\theta + (k + k_0)^{-1/5}, X_{k+1}\right) - J\left(\theta - (k + k_0)^{-1/5}, X'_{k+1}\right)}{(k + k_0)^{-1/5}}. \tag{20}$$

To compute the expected error, Monte Carlo simulations were used with 10000 sample paths and the number of steps k ranging from 2^8 to 2^{20}. We fit regression on the log–log plot to get the convergence rate only on $\left[2^{13}, 2^{20}\right]$ and set $k_0 = 10000$ to avoid the initial fluctuations of the algorithm.

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

TABLE 1. Convergence speed for different distributions of i.i.d. noise.

	independent X_{k+1}, X'_{k+1}	identical $X_{k+1} = X'_{k+1}$
N(0, 1)	-0.299 ($R^2 = 0.999$)	-0.459 ($R^2 = 0.999$)
U([0, 1])	-0.14 ($R^2 = 0.997$)	-0.14 ($R^2 = 0.997$)
Beta(2, 2)	-0.374 ($R^2 = 0.999$)	-0.393 ($R^2 = 0.999$)

5.1. Independent innovations

In this section we assume that the consecutive 'measurement noises' X_n are i.i.d. We consider three different choices for the distribution of the noise: the normal, uniform, and beta distributions. Note that the normal distribution violates boundedness, and for the uniform distribution the differentiability of F fails; however, convergence is achieved even in these cases. We also distinguish between the case where the observations X_{k+1} and X'_{k+1} are the same and the case where they are independent. Here we refer back to Remark 2.1, where we point out that this choice does not influence our theoretical results, although it may make a visible difference numerically. This phenomenon has already been observed; see [12] for more about the variance reduction technique called *common random numbers* (CRN). The values in Table 1 below represent the slope of the linear regression we fit on the log–log plot of the average absolute error versus the number of steps, together with the R-squared value measuring the goodness of the fit.

The lower limit that we theoretically achieved for the convergence rate in Theorem 2.1 was -0.2; however, the numerical experiments we present show that the practical convergence rate can outperform this.

5.1.1. *Standard normal distribution.* Assume that $X \sim N(0, 1)$. Then the function whose minimum we aim to find is $U_1(\theta) = 1 + \theta^2 + \Phi(\theta)$, where Φ denotes the cumulative distribution function of the standard normal distribution. We get the solution $\theta^* = -\sqrt{W\left(\frac{1}{8\pi}\right)} \approx -0.19569$, where W is the Lambert W function.

Figure 1 illustrates the convergence of two variations of the algorithm (20) for U_1, starting the iteration from $\theta_0 = -0.1$. In Figure (1a) we present the case where X_{k+1} and X'_{k+1} are independent on a log–log plot; we observe a convergence rate of $k^{-0.299}$. Figure (1b) shows the case where $X_{k+1} = X'_{k+1}$, which yields a convergence rate of $k^{-0.459}$.

5.1.2. *Uniform distribution on* [0, 1]. Let $X \sim U([0, 1])$. Then the function whose minimum we aim to find is $U_2(\theta) = 1/3 - \theta + \theta^2 + F_{uni}(\theta)$, where F_{uni} denotes the cumulative distribution function of the Uniform([0, 1]) distribution. We get the solution $\theta^* = 0$.

Figure 2 illustrates the convergence of two variations of the algorithm (20) for U_2, starting the iteration from $\theta_0 = 1$. In Figure (2a) we present the case where X_{k+1} and X'_{k+1} are independent on a log–log plot, while (2b) shows the case where $X_{k+1} = X'_{k+1}$. Both cases yield a convergence rate of $k^{-0.14}$, which is worse than the theoretical rate $k^{-0.2}$.

5.1.3. *Beta(2, 2) distribution.* Let $X \sim Beta(2, 2)$. Then the function whose minimum we aim to find is $U_3(\theta) = 0.3 - \theta + \theta^2 + F_\beta(\theta)$, where F_β denotes the cumulative distribution function of the Beta(2,2) distribution. We get the solution $\theta^* = \frac{2-\sqrt{2.5}}{3} \approx 0.13962$.

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

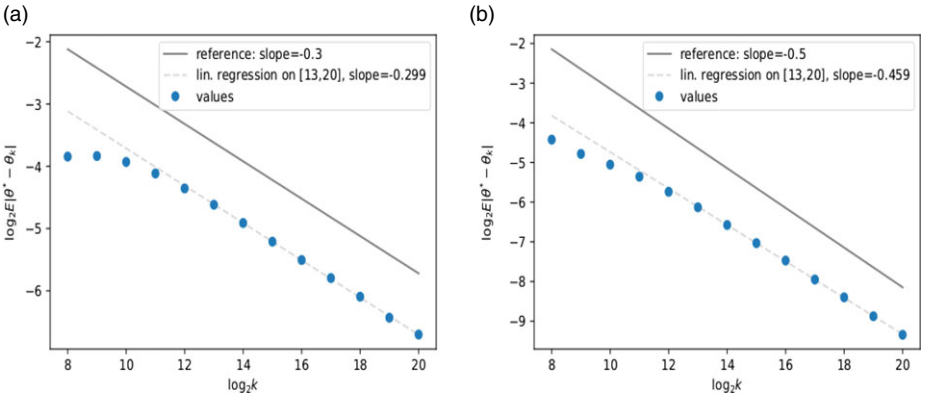

FIGURE 1. Log–log plot of $\mathbb{E}|\theta^* - \theta_k|$ versus number of iterations for i.i.d. standard normal innovations.

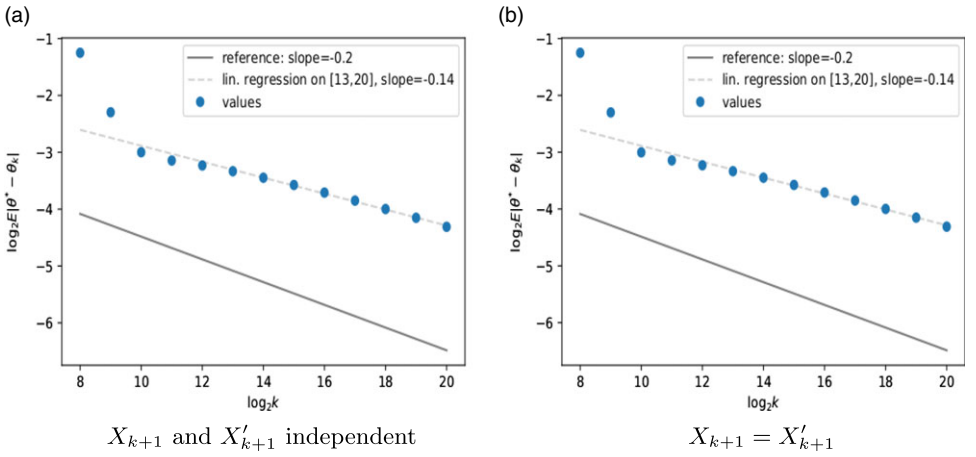

FIGURE 2. Log–log plot of $\mathbb{E}|\theta^* - \theta_k|$ versus number of iterations for i.i.d. uniform innovations.

Figure 3 illustrates the convergence of two variations of the algorithm (20) for U_3, starting the iteration from $\theta_0 = 1$. In Figure (3a) we present the case where X_{k+1} and X'_{k+1} are independent on a log–log plot; we observe a convergence rate of $k^{-0.374}$. Figure (3b) shows the case where $X_{k+1} = X'_{k+1}$, which yields a convergence rate of $k^{-0.393}$.

5.2. AR(1) innovations

For an example with non-i.i.d. X_t, assume that the 'noise' is an AR(1) process defined as

$$Y_{t+1} = \kappa Y_t + \varepsilon_{t+1}, \quad \text{for } t \in \mathbb{Z},$$

where ε_t is standard normal for $t \in \mathbb{Z}$ and $|\kappa| < 1$. Clearly, $Y_t = \sum_{k=0}^{\infty} \kappa^k \varepsilon_{t-k}$, and therefore $Y_t \sim N\left(0, \frac{1}{1-\kappa^2}\right)$. For the sequences X_t and X'_t we have two options: either we take consecutive

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

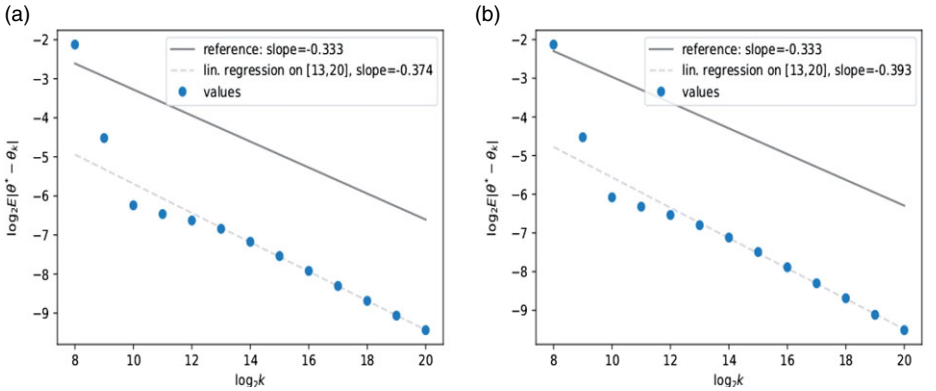

FIGURE 3. Log–log plot of $\mathbb{E}|\theta^* - \theta_k|$ versus number of iterations for i.i.d. beta innovations.

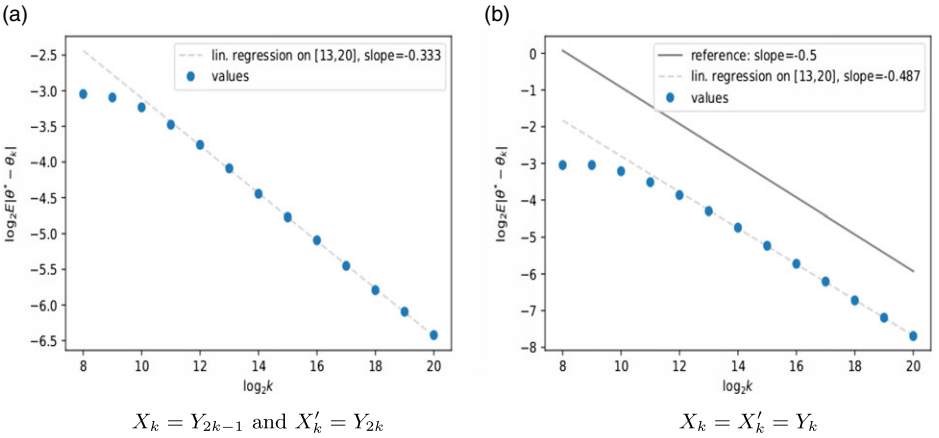

FIGURE 4. Log–log plot of $\mathbb{E}|\theta^* - \theta_k|$ versus number of iterations for AR(1) innovations.

measurements, i.e. $X_k = Y_{2k-1}$ and $X'_k = Y_{2k}$, or we use identical values, i.e. $X_k = X'_k = Y_k$. In both cases,

$$U_4(\theta) = \mathbb{E}J(\theta, X) = \theta^2 + \frac{1}{1-\kappa^2} + \Phi\left(\theta\sqrt{1-\kappa^2}\right).$$

Solving this for $\kappa = 0.75$, we get the optimal value $\theta^* \approx -0.13144$.

Figure 4 and Table 2 illustrate the convergence rate of the algorithm (20) for the function U_4, starting from $\theta_0 = 0$. In Figure (4a) we present the rate in the case where we take consecutive measurements of the AR(1) process ($X_k = Y_{2k-1}$ and $X'_k = Y_{2k}$); the convergence rate of $k^{-0.333}$ is observed. Figure (4b) shows the case where the two measurements are the same $\left(X_k = X'_k = Y_k\right)$, with the rate $k^{-0.487}$.

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

TABLE 2. Convergence rate for AR(1) noise.

	consecutive observations: $X_k = Y_{2k-1}$ and $X'_k = Y_{2k}$	identical: $X_k = X'_k = Y_k$
AR(1)	–0.333 $\left(R^2 = 0.999\right)$	–0.487 $\left(R^2 = 0.999\right)$

6. Application to mathematical finance

The price of a financial asset either follows a trend during a given period of time or just rambles around its 'fair' price value—at least, so it seems to many actual traders. This 'rambling', in more mathematical terms, means that the price is reverting to its long-term average. Such a mean-reversion phenomenon can be exploited by 'buying low, selling high'-type strategies. Discussions on this topic involve plenty of common-sense advice and benevolent concrete suggestions; see e.g. [22–24]. There also exist theoretical studies about optimal trading with such prices; see e.g. [13]. However, a rigorous approach to *adaptive* trading algorithms of this type is lacking.

The results of the present paper provide theoretical convergence guarantees for such algorithms which cannot be deduced from the existing literature on stochastic approximation. The most conspicuous feature of mean-reversion strategies is that they are triggered when the price reaches a certain level. This means that their payoffs are *discontinuous* with respect to the parameters; gradients do not exist, and only finite-difference approximations can be used (the Kiefer–Wolfowitz method). Their convergence in the given discontinuous case cannot be shown based on available results; hence we fill an important and practically relevant gap here.

Below, we describe a trading model in some detail and explain how it fits into the framework used in the previous sections. Let the price of the observed financial asset be described by a real-valued stochastic process S_t, $t \in \mathbb{Z}$, adapted to a given filtration $\mathcal{F}_t$, $t \in \mathbb{Z}$, representing the flow of information. (Alternatively, S_t may be the *increment* of the price at t, which can safely be assumed to follow a stationary process.)

Our algorithm will be based on several dynamically updated estimators which are assumed to be functionals of the trajectories of S_t and possibly of another adapted process F_t describing important economic factors. The estimate for the long-term average of the process is denoted by $A_t(\theta)$ at time t. The upper and lower bandwith processes will be denoted by $B_t^+(\theta)$ and $B_t^-(\theta)$; they are non-negative. All these estimates depend on a parameter θ to be tuned, where θ ranges over a subset Q of $\mathbb{R}^d$.

In practice, $A_t(\theta)$ is some moving average (or exponential moving average) of previous values of S, but it may depend on the other indicators F (market indices, etc.). Here θ determines, for instance, the weights of the moving average estimate. The quantities $B_t^{\pm}(\theta)$ are normally based on standard deviation estimates for S but, again, may be more complex, with θ describing weighting of past information. If we peek from time t back to time $t-p$ with some $p \in \mathbb{N}$, then $A_t(\theta)$, $B_t^{\pm}(\theta)$ are functionals of $(S_{t-p}, F_{t-p}, \dots, S_t, F_t)$.

The price range $\left[A_t - B_t^-, A_t + B_t^+\right]$ is considered to be 'normal' by the algorithm, while quitting that interval suggests 'extremal' behavior that the market should soon correct. For example, reaching the level $A_t - B_t^-$ means that the price is abnormally low for the present circumstances; hence it is worth buying a quantity $b(\theta)$ of the asset, where, again, the parameter θ should be optimally found. When the price returns to $A_{t'}$ at some later time t', the asset will be sold and a profit realized. Similarly, when the price reaches $A_t + B_t^+$, a quantity $s(\theta)$ of the asset is sold (the price being abnormally high), and it will be repurchased once the 'normal' level $A_{t'}$ is reached at some future $t' > t$, with the aim of realizing a profit.

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

The value of the parameter θ will be updated at times tN, $t \in \mathbb{N}$, where $N \geq 1$ is fixed. The (random) profit (or loss) resulting from trading on the interval $[N(t-1), Nt]$ is denoted by $u(\theta, X_t)$ with $X_t = \big(S_{N(t-1)-p}, F_{N(t-1)-p}, \dots, S_{Nt}, F_{Nt}\big)$. We could even write an explicit expression for u based on the description of the trading mechanism in the previous paragraph, but it would be very cumbersome without providing additional insight, so we omit it. We also add that, in many cases, a fee must also be paid at every transaction. Such strategies being 'threshold-type', the function u is generically a *discontinuous* function of θ.

We furthermore argue that one *cannot* smooth out u and make it continuous without losing *essential* features of the problem. At first sight it may look reasonable to approximate the indicator function of the interval $[0, \infty)$ by a function f which is 1 on $[0, \infty)$, 0 on $(-\infty, -\epsilon]$ for some small $\epsilon > 0$, and linear on $(-\epsilon, 0)$, but in this way we get a Lipschitz approximation with a huge Lipschitz constant, hence with a poor convergence rate! This is just to stress that although such simple tricks might work in certain practical situations, they only obscure the real issues in the theoretical analysis (namely, there *is* a discontinuity to be handled).

The algorithm described above is very close to what actual investors do; see [22, 23, 24]. We also mention the related theoretical studies [3, 18], which, however, do not take an adaptive view and calculate optimal strategies for concrete models.

Taking a more realistic, adaptive approach, the investor may seek to maximize $\mathbb{E}u(\theta, X_0)$ by dynamically updating θ at every instant tN, $t \in \mathbb{N}$. Our versions of the Kiefer–Wolfowitz algorithm, presented in the previous sections, are tailor-made for such online optimization, both the decreasing- and the fixed-gain version, depending on the circumstances. Theorems 2.1 and 2.2 provide solid theoretical convergence guarantees for such procedures.

Acknowledgements

Part of this research was performed while the authors were participating in the Simons Semester on Stochastic Modeling and Control at Banach Center, Warsaw, in 2019. During the preparation of this paper the second author attended the PhD school of Central European University, Budapest, and was also affiliated with Eötvös Loránd University, Budapest. We thank all these institutions for their hospitality.

Funding information

Both authors were supported by the Lendület grant LP 2015-6. The second author was also supported by Project No. ED 18-1-2019-0030 (*Application-specific highly reliable IT solutions*), which was implemented with support provided from the National Research, Development and Innovation Fund of Hungary, financed under the Thematic Excellence Programme funding scheme.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process for this article.

References

[1] Barkhagen, M. *et al.* (2021). On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. *Bernoulli* **27**, 1–33.

[2] Benveniste, A., Métivier, M. and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations*. Springer, Berlin, Heidelberg.

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press

[3] CARTEA, Á., JAIMUNGAL, S. AND PENALVA, J. (2015). *Algorithmic and High-Frequency Trading*. Cambridge University Press.
[4] CHAU, N. H., KUMAR, C., RÁSONYI, M. AND SABANIS, S. (2019). On fixed gain recursive estimators with discontinuity in the parameters. *ESAIM Prob. Statist.* **23**, 217–244.
[5] CHAU, N. H. *et al.* (2021). On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. *SIAM J. Math. Data Sci.* **3**, 959–986.
[6] DURMUS, A. AND MOULINES, É. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Prob.* **27**, 1551–1587.
[7] FORT, G, MOULINES, É., SCHRECK, A. AND VIHOLA, M. (2016). Convergence of Markovian stochastic approximation with discontinuous dynamics. *SIAM J. Control Optimization* **54**, 866–893.
[8] GERENCSÉR, L. (1989). On a class of mixing processes. *Stochastics* **26**, 165–191.
[9] GERENCSÉR, L. (1992). Rate of convergence of recursive estimators. *SIAM J. Control Optimization* **30**, 1200–1227.
[10] GERENCSÉR, L. (1998). SPSA with state-dependent noise—a tool for direct adaptive control. In *Proc. 37th IEEE Conference on Decision and Control*, Institute of Electrical and Electronics Engineers, Piscataway, NJ, pp. 3451–3456.
[11] GERENCSÉR, L. (1999). Convergence rate of moments in stochastic approximation with simultaneous perturbation gradient approximation and resetting. *IEEE Trans. Automatic Control* **44**, 894–905.
[12] GLASSERMAN, P. AND YAO, D. D. (1992). Some guidelines and guarantees for common random numbers. *Manag. Sci.* **38**, 884–908.
[13] GUASONI, P., TOLOMEO, A. AND WANG, G. (2019). Should commodity investors follow commodities' prices? *SIAM J. Financial Math.* **10**, 466–490.
[14] KIEFER, J. AND WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* **23**, 462–466.
[15] KUSHNER, H. J. AND CLARK, D. S. (1978). *Stochastic Approximation for Constrained and Unconstrained Systems*. Springer, New York.
[16] LARUELLE, S. AND PAGÈS, G. (2012). Stochastic approximation with averaging innovation applied to finance. *Monte Carlo Meth. Appl.* **18**, 1–51.
[17] L. LJUNG. (1977). Analysis of recursive stochastic algorithms. *IEEE Trans. Automatic Control* **22**, 551–575.
[18] LEUNG, T. AND LI, X. (2015). *Optimal Mean Reversion Trading*. World Scientific, Singapore.
[19] ROBBINS, H. AND MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22**, 400–407.
[20] SACKS, J. (1958). Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Statist.* **29**, 373–405.
[21] SICHEL, H. S., KLEINGELD, W. J. AND ASSIBEY-BONSU, W. (1992). A comparative study of three frequency-distribution models for use in ore valuation. *J. S. Afr. Inst. Mining Metallurgy* **92**, 91–99.
[22] TRADERGAV.COM. What is mean reversion trading strategy. Available at https://tradergav.com/what-is-mean-reversion-trading-strategy.
[23] TRADING STRATEGY GUIDES (2021). Mean reversion trading strategy with a sneaky secret. Available at https://tradingstrategyguides.com/mean-reversion-trading-strategy.
[24] WARRIOR TRADING. Mean reversion trading: is it a profitable strategy? Available at https://www.warriortrading.com/mean-reversion.

https://doi.org/10.1017/apr.2022.25 Published online by Cambridge University Press