Numerische Mathematik (2019) 142:279–328
https://doi.org/10.1007/s00211-019-01024-y

Numerische Mathematik

General multilevel adaptations for stochastic approximation algorithms of Robbins–Monro and Polyak–Ruppert type

Steffen Dereich[1] · **Thomas Müller-Gronbach**[2]

Received: 1 May 2017 / Revised: 7 January 2019 / Published online: 6 February 2019

Abstract
In this article we present and analyse new multilevel adaptations of classical stochastic approximation algorithms for the computation of a zero of a function $f\colon D \to \mathbb{R}^d$ defined on a convex domain $D \subset \mathbb{R}^d$, which is given as a parameterised family of expectations. The analysis of the error and the computational cost of our method is based on similar assumptions as used in Giles (Oper Res 56(3):607–617, 2008) for the computation of a single expectation. Additionally, we essentially only require that f satisfies a classical contraction property from stochastic approximation theory. Under these assumptions we establish error bounds in pth mean for our multilevel Robbins–Monro and Polyak–Ruppert schemes that decay in the computational time as fast as the classical error bounds for multilevel Monte Carlo approximations of single expectations known from Giles (Oper Res 56(3):607–617, 2008). Our approach is universal in the sense that having multilevel implementations for a particular application at hand it is straightforward to implement the corresponding stochastic approximation algorithm.

Mathematics Subject Classification Primary 62L20; Secondary 60J10 · 65C05

1 Introduction

Let $D \subset \mathbb{R}^d$ be closed and convex and let U be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a set $\mathcal{U}$ equipped with some σ-field. We study the

✉ Thomas Müller-Gronbach
thomas.mueller-gronbach@uni-passau.de

Steffen Dereich
steffen.dereich@wwu.de

[1] Fachbereich 10: Mathematik und Informatik, Institut für Mathematische Statistik, Westfälische Wilhelms-Universität Münster, Orléans-Ring 10, 48149 Münster, Germany

[2] Fakultät für Informatik und Mathematik, Universität Passau, Innstraße 33, 94032 Passau, Germany

problem of computing zeros of functions $f\colon D \to \mathbb{R}^d$ of the form

$$f(\theta) = \mathbb{E}[F(\theta, U)],$$

where $F\colon D \times \mathcal{U} \to \mathbb{R}^d$ is a product measurable function such that all expectations $\mathbb{E}[F(\theta, U)]$ are well-defined. In this article we focus on the case where the random variables $F(\theta, U)$ cannot be simulated directly so that one has to work with appropriate approximations in numerical simulations. For example, one may think of U being a Brownian motion and of $F(\theta, U)$ being the payoff of an option, where θ is a parameter affecting the payoff and/or the dynamics of the price process. Alternatively, $F(\theta, U)$ might be the value of a PDE at certain positions with U representing random coefficients and θ a parameter of the equation.

In previous years the multilevel paradigm introduced by Heinrich [6] and Giles [5] has proved to be a very efficient tool in the numerical computation of expectations. By Frikha [3] it has recently been shown that the efficiency of the multilevel paradigm prevails when combined with stochastic approximation algorithms. In the present paper we take a different approach than the one introduced by the latter author. Instead of employing a sequence of coupled Robbins–Monro algorithms to construct a multilevel estimate of a zero of f we basically propose a single Robbins–Monro algorithm that uses in the $(n+1)$th step a multilevel estimate of $\mathbb{E}[F(\theta_n, U)]$ with a complexity that is adapted to the actual state θ_n of the system and increases in the number of steps.

Our approach is universal in the sense that having multilevel implementations for a particular application at hand it is straightforward to implement the corresponding stochastic approximation algorithm. Moreover, previous research on multilevel Monte Carlo can be incorporated in a natural way. This is due to the fact that the analysis of the error and the computational cost of our method is based on similar assumptions on the biases, the pth central moments and the simulation cost of the underlying approximations of $F(\theta, U)$ as used in Giles [5], see Assumptions C.1 and C.2 in Sect. 3. Additionally, we require that f satisfies a classical contraction property from stochastic approximation theory: there exist $L > 0$ and a zero θ^* of f such that for all $\theta \in D$,

$$\langle f(\theta), \theta - \theta^* \rangle \leq -L\|\theta - \theta^*\|^2,$$

where $\langle \cdot, \cdot \rangle$ denotes an inner product on $\mathbb{R}^d$. Moreover, f has to satisfy a linear growth condition relative to the zero θ^*, see Assumption A.1 and Remark 2.1 in Sect. 2. Note that the contraction property implies that the zero θ^* is unique. Theorem 3.1 asserts that under these assumptions the maximum pth mean error $\sup_{k \geq n} \mathbb{E}[\|\theta_k - \theta^*\|^p]$ of our properly tuned multilevel Robbins–Monro scheme $(\theta_n)_{n\in\mathbb{N}}$ satisfies the same upper bounds in terms of the computational time needed to compute θ_n as the bounds obtained in Giles [5] for the multilevel computation of a single expectation.

In general, the design of this algorithm requires knowledge on the constant L in the contraction property of f. To bypass this problem without loss of efficiency one may work with a Polyak–Ruppert average of our algorithm. Theorem 3.2 states that under Assumptions C.1 and C.2 on the approximations of $F(\theta, U)$ and Assumption

B.1 on f, which is slightly stronger than condition A.1 a properly tuned multilevel Polyak–Ruppert average $(\bar{\theta}_n)_{n\in\mathbb{N}}$ achieves, for $q < p$, the same upper bounds in the relation of the qth mean error $\mathbb{E}[\|\bar{\theta}_n - \theta^*\|^q]$ and the corresponding computational time as the previously introduced multilevel Robbins–Monro method.

We briefly outline the content of the paper. The multilevel algorithms and the respective complexity theorems are presented in Sect. 3 for the case where $D = \mathbb{R}^d$. General closed convex domains D are covered in Sect. 4. In Sect. 5 we present the construction of our multilevel algorithms in two examples and we illustrate the theoretical error bounds by simulation experiments. We add that Sects. 3 and 4 are self-contained and a reader interested in the multilevel schemes only, can immediately start reading in Sect. 3.

The error analysis of the multilevel stochastic approximation algorithms is based on new estimates of the pth mean error of Robbins–Monro and Polyak–Ruppert algorithms. These results are presented in Sect. 2. As a technical tool we employ a modified Burkholder-Davis-Gundy inequality, which is established in the appendix and might be of interest in itself, see Theorem 5.1.

We add that formally all results of the following sections remain true when replacing $(\mathbb{R}^d, \langle\cdot,\cdot\rangle)$ by an arbitrary separable Hilbert space. However in that case the definition (65) of the computational cost of a multilevel algorithm might not be appropriate in general.

2 New error estimates for stochastic approximation algorithms

Since the pioneering work of Robbins and Monro [17] in 1951 a large body of research has been devoted to the analysis of stochastic approximation algorithms with a strong focus on pathwise and weak convergence properties. In particular, laws of iterated logarithm and central limit theorems have been established that allow to optimise the parameters of the schemes with respect to the almost sure and weak convergence rates and the size of the limiting covariance. See e.g. [4,7,10,11,14–16,18,19] for results and further references as well as the survey articles and monographs [1,2,8,9,12]. Less attention has been paid to an error control in L_p-norm for arbitrary orders $p \geq 2$. We provide such estimates for the Robbins–Monro approximation and the Polyak–Rupert averaging introduced by Ruppert [19] and Polyak [16] under mild conditions on the ingredients of these schemes. These estimates build the basis for the error analysis of the multilevel schemes introduced in Sect. 3.

Throughout this section we fix $p \in [2, \infty)$, a probability space $(\Omega, \mathcal{F}, P)$ equipped with a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$, a scalar product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^d$ with induced norm $\|\cdot\|$. Furthermore, we fix a measurable function $f\colon \mathbb{R}^d \to \mathbb{R}^d$ that has a unique zero $\theta^* \in \mathbb{R}^d$.

We consider an adapted $\mathbb{R}^d$-valued dynamical system $(\theta_n)_{n\in\mathbb{N}_0}$ iteratively defined by

$$\theta_n = \theta_{n-1} + \gamma_n\big(f(\theta_{n-1}) + \varepsilon_n R_n + \sigma_n D_n\big), \tag{1}$$

for $n \in \mathbb{N}$, where $\theta_0 \in \mathbb{R}^d$ is a fixed deterministic starting value,

(I) $(R_n)_{n\in\mathbb{N}}$ is a previsible process, the *remainder/bias*,

(II) $(D_n)_{n\in\mathbb{N}}$ is a sequence of martingale *differences*,
(III) $(\gamma_n)_{n\in\mathbb{N}}$ is a sequence of positive reals tending to zero, and $(\varepsilon_n)_{n\in\mathbb{N}}$ and $(\sigma_n)_{n\in\mathbb{N}}$ are sequences of non-negative real numbers.

2.1 Estimates for the Robbins–Monro algorithm

Our goal is to quantify the speed of convergence of the sequence $(\theta_n)_{n\in\mathbb{N}_0}$ to θ^* in the pth mean sense in terms of the *step-sizes* γ_n, the *bias-levels* ε_n and the *noise-levels* σ_n.

To this end we employ the following set of assumptions in addition to (I)–(III).

A.1 (Assumptions on f and θ^*)
There exist $L, L' \in (0,\infty)$ such that for all $\theta \in \mathbb{R}^d$

(i) $\langle \theta-\theta^*, f(\theta)\rangle \le -L\,\|\theta-\theta^*\|^2$ and
(ii) $\langle \theta-\theta^*, f(\theta)\rangle \le -L'\,\|f(\theta)\|^2$.

A.2 (Assumptions on $(R_n)_{n\in\mathbb{N}}$ and $(D_n)_{n\in\mathbb{N}}$)
It holds

(i) $\sup_{n\in\mathbb{N}} \operatorname{esssup} \|R_n\| < \infty$ and
(ii) $\sup_{n\in\mathbb{N}} \mathbb{E}[\|D_n\|^p] < \infty$.

Remark 2.1 (Discussion of Assumption A.1) We briefly discuss A.1(i) and A.1(ii). Let $\theta \in \mathbb{R}^d$ and $c_1, c_2, c_2', \gamma \in (0,\infty)$, and consider the conditions

$$\langle \theta-\theta^*, f(\theta)\rangle \le -c_1\,\|\theta-\theta^*\|^2, \tag{i}$$

$$\langle \theta-\theta^*, f(\theta)\rangle \le -c_2\,\|f(\theta)\|^2, \tag{ii}$$

$$\|f(\theta)\| \le c_2'\,\|\theta-\theta^*\|, \tag{ii'}$$

$$\|\theta-\theta^*+\gamma f(\theta)\|^2 \le \|\theta-\theta^*\|^2\big(1-\gamma c_1(2-\tfrac{\gamma}{c_2})\big). \tag{$*$}$$

By the Cauchy-Schwartz inequality we have

$$f \text{ satisfies (ii)} \;\Rightarrow\; f \text{ satisfies (ii') for every } c_2' \ge 1/c_2, \tag{2}$$

and the choice $f(\theta)=\theta$ shows that the reverse implication is not valid in general. However, it is easy to check that

$$f \text{ satisfies (i) and (ii')} \;\Rightarrow\; f \text{ satisfies (ii) for any } c_2 \le c_1/(c_2')^2.$$

Thus, in the presence of condition A.1(i), condition A.1(ii) is equivalent to a linear growth condition on the function f relative to the zero θ^*.

Finally, conditions (i) and (ii) jointly imply the contraction property (*), which is crucial for the analysis of the Robbins–Monro scheme. We have

$$f \text{ satisfies (i) and (ii)} \;\Rightarrow\; f \text{ satisfies } (*) \text{ for every } \gamma \le 2c_2. \tag{3}$$

In fact, let $\gamma \le 2c_2$ and use (ii) and then (i) to conclude that

$$\begin{aligned}\|\theta - \theta^* + \gamma f(\theta)\|^2 &= \|\theta - \theta^*\|^2 + 2\gamma\langle\theta - \theta^*, f(\theta)\rangle + \gamma^2\|f(\theta)\|^2\\ &\le \|\theta - \theta^*\|^2 + \langle\theta - \theta^*, f(\theta)\rangle\left(2\gamma - \frac{\gamma^2}{c_2}\right)\\ &\le \|\theta - \theta^*\|^2 - c_1\|\theta - \theta^*\|^2\left(2\gamma - \frac{\gamma^2}{c_2}\right).\end{aligned}$$

We illustrate Assumption A.1 by a multidimensional regression problem.

Example 2.2 (Regression) Let $\langle\cdot,\cdot\rangle$ be the Euclidean scalar product on $\mathbb{R}^d$. Consider a d-dimensional random vector X on $(\Omega, \mathcal{F}, \mathbb{P})$ that satisfies

$$\mathbb{E}[\|X\|^{2p}] < \infty \text{ and } B := \mathbb{E}[XX^\top] \in \mathbb{R}^{d\times d} \text{ is positive definite,} \tag{4}$$

and let $g\colon \mathbb{R}^d \to \mathbb{R}$ be a Borel-measurable function such that

$$\mathbb{E}[|g(X)|^{2p}] < \infty. \tag{5}$$

By (4) and (5), the vector

$$\theta^* = B^{-1}\mathbb{E}[g(X)X]$$

is well defined, and because

$$\langle\theta^*, X\rangle = (\mathbb{E}[g(X)X_1], \ldots, \mathbb{E}[g(X)X_d])B^{-1}X$$

is the orthogonal projection of $g(X)$ onto the linear subspace of $L_2(\Omega, \mathcal{F}, \mathbb{P})$ generated by $X_1, \ldots, X_d$ it follows that θ^* solves the regression problem

$$\mathbb{E}[|\langle\theta^*, X\rangle - g(X)|^2] = \min_{\theta\in\mathbb{R}^d}\mathbb{E}[|\langle\theta, X\rangle - g(X)|^2].$$

Clearly, θ^* is the unique zero of the function $f\colon \mathbb{R}^d \to \mathbb{R}^d$ given by

$$f(\theta) = \mathbb{E}[g(X)X] - B\theta. \tag{6}$$

We show that f satisfies Assumption A.1. Since B is symmetric and positive definite we have

$$\forall\,\theta \in \mathbb{R}^d\colon \langle\theta, B\theta\rangle \ge \max\bigl(\lambda_{\min}\|\theta\|^2, \lambda_{\max}^{-1}\|B\theta\|^2\bigr), \tag{7}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimal and the maximal eigenvalue of B, respectively. This follows easily from the fact that $B = O\Lambda O^\top$, where $O \in \mathbb{R}^{d\times d}$ is orthogonal and $\Lambda \in \mathbb{R}^{d\times d}$ is the diagonal matrix of the eigenvalues of B. Using inequality (7) it is straightforward to see that f satisfies condition A.1(i) for any $L \in (0, \lambda_{\min})$ and the condition A.1(ii) for any $L' \in (0, \lambda_{\max}^{-1})$.

In the following we put for $r \in (0, \infty)$ and $n, k \in \mathbb{N}$ with $n \geq k$,

$$\tau_{k,n}(r) = \prod_{j=k+1}^{n} (1 - \gamma_j r), \quad e_{k,n}(r) = \max_{j=k,\ldots,n} \varepsilon_j \, \tau_{j,n}(r),$$
$$s_{k,n}^2(r) = \sum_{j=k}^{n} \gamma_j^2 \, \sigma_j^2 \, (\tau_{j,n}(r))^2. \tag{8}$$

First we provide pth mean error estimates of the Robbins–Monro scheme (1) in terms of the quantities introduced in (8).

Proposition 2.3 *Assume that (I)–(III) and A.1 and A.2 are satisfied. Then for every $r \in (0, L)$ there exist $n_0 \in \mathbb{N}$ and $\kappa \in (0, \infty)$ such that for all $n \geq k_0 \geq n_0$ we have $\tau_{k_0,n}(r) \in (0, 1)$ and*

$$\mathbb{E}\big[\|\theta_n - \theta^*\|^p\big]^{1/p} \leq \kappa \left(\tau_{k_0,n}(r)\, \mathbb{E}[\|\theta_{k_0} - \theta^*\|^p]^{1/p} + e_{k_0,n}(r) + s_{k_0,n}(r)\right). \tag{9}$$

Proof Without loss of generality we may assume that $\theta^* = 0$.

By Assumption A.2 there exists $\kappa_1 \in (0, \infty)$ such that for all $n \in \mathbb{N}$,

$$\|R_n\| \leq \kappa_1 \text{ a.s.} \tag{10}$$

and

$$\mathbb{E}[\|D_n\|^p] \leq \kappa_1. \tag{11}$$

Note further that (2) in Remark 2.1 implies that the dynamical system (1) satisfies $\|\theta_n\| \leq (1 + \gamma_n/L')\|\theta_{n-1}\| + \gamma_n \varepsilon_n \|R_n\| + \gamma_n \sigma_n \|D_n\|$ for every $n \in \mathbb{N}$. With Assumption A.2 we conclude that $\theta_n \in L_p(\Omega, \mathcal{F}, P)$ for every $n \in \mathbb{N}$.

Let $r \in (0, L)$. Since $\lim_{n\to\infty} \gamma_n = 0$ we may choose $n_0 \in \mathbb{N}$ such that $1 - \gamma_n L > 0$ and $1 - \frac{1}{2}\gamma_n/L' \geq (r + L)/(2L)$ for all $n \geq n_0$. Using (3) in Remark 2.1 we obtain that for all $\theta \in \mathbb{R}^d$ and for all $n \geq n_0$,

$$\|\theta + \gamma_n f(\theta)\|^2 \leq (1 - 2\gamma_n L(1 - \tfrac{1}{2}\gamma_n/L'))\|\theta\|^2 \leq (1 - \gamma_n(r + L)/2)^2 \, \|\theta\|^2. \tag{12}$$

In the following we write $\tau_{k,n}$, $e_{k,n}$ and $s_{k,n}$ in place of $\tau_{k,n}(r)$, $e_{k,n}(r)$ and $s_{k,n}(r)$, respectively. Let $k_0 \geq n_0$ and put

$$\zeta_n = \frac{\theta_n}{\tau_{k_0,n}}, \quad \xi_n = \frac{\theta_{n-1} + \gamma_n \, (f(\theta_{n-1}) + \varepsilon_n \, R_n)}{\tau_{k_0,n}}, \quad M_n = \zeta_{k_0} + \sum_{k=k_0+1}^{n} \frac{\gamma_k \, \sigma_k \, D_k}{\tau_{k_0,k}} \tag{13}$$

for $n \geq k_0$. Then $(\zeta_n)_{n\geq k_0}$ is adapted, $(\xi_n)_{n>k_0}$ is previsible, $(M_n)_{n\geq k_0}$ is a martingale and for all $n > k_0$ we have

$$\zeta_n = \xi_n + \Delta M_n. \tag{14}$$

Below we show that there exists a constant $\kappa_2 \in (0, \infty)$, which only depends on L, r and κ_1 such that a.s. for all $n > k_0$,

$$\|\xi_n\| \leq \|\zeta_{n-1}\| \vee \kappa_2 \frac{\varepsilon_n}{\tau_{k_0,n}} \tag{15}$$

and

$$\mathbb{E}\big[[M]_n^{p/2}\big]^{2/p} \leq \mathbb{E}[\|\theta_{k_0}\|^p]^{2/p} + \kappa_2 \frac{s_{k_0,n}^2}{\tau_{k_0,n}^2}. \tag{16}$$

Observing (14) and (15) we may apply the Burkholder–Davis–Gundy inequality, see Theorem 5.1, to the processes $(\zeta_n)_{n\geq k_0}$, $(\xi_n)_{n>k_0}$ and $(M_n)_{n\geq k_0}$ to obtain for $n \geq k_0$ that

$$\mathbb{E}\left[\max_{k_0\leq k\leq n} \|\zeta_k\|^p\right] \leq \kappa_3 \left(\mathbb{E}\big[[M]_n^{p/2}\big] + \left(\kappa_2 \frac{e_{k_0,n}}{\tau_{k_0,n}}\right)^p\right), \tag{17}$$

where the constant $\kappa_3 > 0$ only depends on p. Using (16) we conclude that

$$\mathbb{E}\big[\|\theta_n\|^p\big] = \tau_{k_0,n}^p\, \mathbb{E}\big[\|\zeta_n\|^p\big] \leq 2^{p/2}\kappa_3 \left(\tau_{k_0,n}^p\, \mathbb{E}[\|\theta_{k_0}\|^p] + \kappa_2^{p/2} s_{k_0,n}^p + \kappa_2^p\, e_{k_0,n}^p\right),$$

which completes the proof of the theorem up to the justification of (15) and (16).

For the proof of (15) we use (10) and (12) to obtain that a.s. for $n > k_0$,

$$\begin{aligned}\|\xi_n\| &\leq \left\| \frac{\theta_{n-1} + \gamma_n f(\theta_{n-1})}{1-\gamma_n r} \frac{1}{\tau_{k_0,n-1}} \right\| + \frac{\gamma_n\, \varepsilon_n}{\tau_{k_0,n}} \|R_n\| \\ &\leq \frac{1-\gamma_n(r+L)/2}{1-\gamma_n r} \|\zeta_{n-1}\| + \kappa_1 \frac{\gamma_n\, \varepsilon_n}{\tau_{k_0,n}} \leq \left(1 - \gamma_n \frac{L-r}{2}\right) \|\zeta_{n-1}\| + \kappa_1 \frac{\gamma_n\, \varepsilon_n}{\tau_{k_0,n}},\end{aligned}$$

where the last inequality follows from the fact that $\frac{1-a}{1-b} \leq 1-a+b$ for $0 \leq b \leq a \leq 1$. Hence, if $\frac{L-r}{2}\|\zeta_{n-1}\| \geq \kappa_1 \varepsilon_n/\tau_{k_0,n}$ then

$$\|\xi_n\| \leq \|\zeta_{n-1}\|,$$

while in the case $\frac{L-r}{2}\|\zeta_{n-1}\| < \kappa_1 \varepsilon_n/\tau_{k_0,n}$,

$$\|\xi_n\| \leq \frac{2\kappa_1}{L-r} \frac{\varepsilon_n}{\tau_{k_0,n}}.$$

Thus (15) holds for any $\kappa_2 \geq 2\kappa_1/(L-r)$.

It remains to show (16). Using (11) we get

$$\begin{aligned}\mathbb{E}\big[[M]_n^{p/2}\big]^{2/p} &= \mathbb{E}\left[\left(\|\theta_{k_0}\|^2 + \sum_{k=k_0+1}^{n} \|\Delta M_k\|^2\right)^{p/2}\right]^{2/p} \\ &\le \mathbb{E}[\|\theta_{k_0}\|^p]^{2/p} + \sum_{k=k_0+1}^{n} \frac{\gamma_k^2\sigma_k^2}{\tau_{k_0,k}^2}\big(\mathbb{E}[\|D_k\|^p]\big)^{2/p} \\ &\le \mathbb{E}[\|\theta_{k_0}\|^p]^{2/p} + \kappa_1^{2/p} \sum_{k=k_0+1}^{n} \frac{\gamma_k^2\,\sigma_k^2}{\tau_{k_0,k}^2} = \mathbb{E}[\|\theta_{k_0}\|^p]^{2/p} + \frac{\kappa_1^{2/p}}{\tau_{k_0,n}^2} s_{k_0,n}^2. \end{aligned} \tag{18}$$

Hence (16) holds for any $\kappa_2 \ge \kappa_1^{2/p}$, which completes the proof. □

Remark 2.4 The proof of the pth mean error estimate (9) in Proposition 2.3 for the times $n \ge k_0 \ge n_0$ makes use of the recursion (1) for n strictly larger than k_0 only. Hence, if $m_0 \in \mathbb{N}_0$ and $(\tilde\theta_n)_{n\ge m_0}$ is the dynamical system given by the recursion (1) with an arbitrary random starting value $\tilde\theta_{m_0} \in L_p(\Omega, \mathcal{F}_{m_0}, P)$ then estimate (9) is valid for $\tilde\theta_n$ in place of θ_n with the same constant κ for all $n \ge k_0 \ge \max(n_0, m_0)$.

The following theorem provides an estimate for the pth mean error of θ_n in terms of the product

$$v_n = \sqrt{\gamma_n}\,\sigma_n.$$

It requires the following additional assumptions on the step-sizes γ_n, the bias-levels ε_n and the noise-levels σ_n.

A.3 (Assumptions on $(\gamma_n)_{n\in\mathbb{N}}$, $(\varepsilon_n)_{n\in\mathbb{N}}$ and $(\sigma_n)_{n\in\mathbb{N}}$)

We have $v_n > 0$ for all $n \in \mathbb{N}$. Furthermore, with L according to A.1(i),

(i) $\limsup_{n\to\infty} \frac{\varepsilon_n}{v_n} < \infty$, and
(ii) $\limsup_{n\to\infty} \frac{1}{\gamma_n} \frac{v_{n-1}-v_n}{v_{n-1}} < L$.

Theorem 2.5 (Robbins–Monro approximation) *Assume that conditions (I)–(III), A.1, A.2 and A.3 are satisfied. Then there exists $\kappa \in (0,\infty)$ such that for all $n \in \mathbb{N}$,*

$$\mathbb{E}\big[\|\theta_n - \theta^*\|^p\big]^{1/p} \le \kappa\, v_n.$$

Proof Below we show that there exist $r \in (0, L)$, $\kappa_1 \in (0,\infty)$ and $n_1 \in \mathbb{N}$ such that for all $n \ge n_1$ we have $\gamma_n < 1/L$ and

$$\tau_{n_1,n}(r) + e_{n_1,n}(r) + s_{n_1,n}(r) \le \kappa_1\, v_n. \tag{19}$$

Then, by choosing $n_0 \in \mathbb{N}$ and $\kappa \in (0,\infty)$ according to Proposition 2.3 and taking $k_0 = \max(n_0, n_1)$ we have for $n \ge k_0$ that $\tau_{k_0,n}(r) = \tau_{n_1,n}(r)/\tau_{n_1,k_0}(r)$, $e_{k_0,n}(r) \le$

$e_{n_1,n}(r)$ and $s_{k_0,n}(r) \le s_{n_1,n}(r)$, and therefore for all $n \ge k_0$

$$\mathbb{E}\big[\|\theta_n - \theta^*\|^p\big]^{1/p} \le \kappa \left(\mathbb{E}[\|\theta_{k_0} - \theta^*\|^p]^{1/p}/\tau_{n_1,k_0}(r) + 1\right)\big(\tau_{n_1,n}(r) + e_{n_1,n}(r) + s_{n_1,n}(r)\big),$$

which finishes the proof of the theorem, up to the justification of (19).

By Assumption A.3 there exist $r_1 \in (0, L)$, $\kappa_2 \in (0, \infty)$ and $n_1 \in \mathbb{N}$ such that for all $n > n_1$,

$$\frac{v_{n-1}}{v_n} \le \frac{1}{1 - \gamma_n r_1} \tag{20}$$

as well as

$$\varepsilon_n \le \kappa_2\, v_n. \tag{21}$$

Take $r \in (r_1, L)$ and assume without loss of generality that $1 - \gamma_n r > 0$ for all $n \ge n_1$. In the following we write $\tau_{k,n}$, $e_{k,n}$ and $s_{k,n}$ in place of $\tau_{k,n}(r)$, $e_{k,n}(r)$ and $s_{k,n}(r)$, respectively. It follows from (20) and $r > r_1$ that the sequence $(v_n/\tau_{n_1,n})_{n \ge n_1}$ is increasing and therefore, for all $n \ge n_1$,

$$\tau_{n_1,n} = \frac{\tau_{n_1,n}}{v_{n_1}} \frac{v_{n_1}}{\tau_{n_1,n_1}} \le \frac{v_n}{v_{n_1}}. \tag{22}$$

Furthermore, observing (21) we also have for all $n \ge n_1$,

$$e_{n_1,n} \le \kappa_2 \max_{j=n_1,\dots,n} v_j\, \tau_{j,n} = \kappa_2\, \tau_{n_1,n} \max_{j=n_1,\dots,n} \frac{v_j}{\tau_{n_1,j}} = \kappa_2\, v_n. \tag{23}$$

Put

$$\varphi(n) = \frac{s^2_{n_1,n}}{v_n^2}$$

for $n \ge n_1$. Observing (20) we obtain that for $n > n_1$,

$$\begin{aligned}\varphi(n) &= \frac{v_{n-1}^2}{v_n^2}(1 - \gamma_n r)^2\, \varphi(n-1) + \gamma_n \le \frac{(1 - \gamma_n r)^2}{(1 - \gamma_n r_1)^2}\, \varphi(n-1) + \gamma_n \\ &= \left(1 - \gamma_n \frac{r - r_1}{1 - \gamma_n r_1}\right)^2 \varphi(n-1) + \gamma_n \le (1 - \gamma_n (r - r_1))\, \varphi(n-1) + \gamma_n.\end{aligned}$$

This entails that

$$\varphi(n) - 1/(r - r_1) \le (1 - \gamma_n (r - r_1))(\varphi(n-1) - 1/(r - r_1)),$$

so that $\varphi(n) \le \varphi(n-1) \vee 1/(r - r_1)$. Hence, by induction, for all $n \ge n_1$,

$$\varphi(n) \le \varphi(n_1) \vee 1/(r - r_1),$$

so that

$$s_{n_1,n} \leq (\gamma_{n_1} \vee 1/(r-r_1))^{1/2}\, v_n. \tag{24}$$

Combining (22) to (24) yields (19). □

As a particular consequence of Theorem 2.5 we obtain error estimates in the case of polynomial step-sizes γ_n and noise-levels σ_n.

Corollary 2.6 (Polynomial step-sizes and noise-levels) *Assume that conditions (I)–(III), A.1 and A.2 are satisfied and choose L according to A.1(i). Take $\gamma_0, \sigma_0 \in (0,\infty)$, $r_1 \in (0,1]$ and $r_2 \in \mathbb{R}$ with*

$$r_1 < 1 \text{ or } \big(r_1 = 1 \text{ and } \gamma_0 > \frac{1+r_2}{2L}\big)$$

and let for all $n \in \mathbb{N}$,

$$\gamma_n = \gamma_0 \frac{1}{n^{r_1}}, \quad \sigma_n^2 = \sigma_0^2 \frac{1}{n^{r_2}}.$$

Assume further that

$$\limsup_{n\to\infty} n^{(r_1+r_2)/2}\, \varepsilon_n < \infty.$$

Then there exists a constant $\kappa \in (0,\infty)$ such that for all $n \in \mathbb{N}$,

$$\mathbb{E}\big[\|\theta_n - \theta^*\|^p\big]^{\frac{1}{p}} \leq \kappa\, n^{-\frac{r_1+r_2}{2}}. \tag{25}$$

Proof We first verify that Assumption A.3 is satisfied. By definition of γ_n and σ_n we have

$$v_n = \sqrt{\gamma_0}\sigma_0 \frac{1}{n^{(r_1+r_2)/2}}.$$

Thus, A.3(i) is satisfied due to the assumption on the sequence $(\varepsilon_n)_{n\in\mathbb{N}}$. Moreover, it is easy to see that

$$\lim_{n\to\infty} \frac{1}{\gamma_n}\Big(1 - \frac{v_n}{v_{n-1}}\Big) = \begin{cases} 0, & \text{if } r_1 < 1, \\ \frac{1+r_2}{2\gamma_0}, & \text{if } r_1 = 1 \end{cases}$$

and therefore A.3(ii) is satisfied as well. Since conditions (I)–(III), A.1 and A.2 are part of the corollary, we may apply Theorem 2.5 to obtain the claimed error estimate. □

Remark 2.7 (Exponential decay of noise-levels) Assumption A.3(ii) may also be satisfied in the case that the noise-levels σ_n have a superpolynomial decay. For instance, if

$$\gamma_n = \frac{a_1}{n^{r_1}}, \quad \sigma_n^2 = \frac{a_2}{n^{r_2}}\, \exp(-a_3 n^{r_3})$$

for all $n \in \mathbb{N}$, where $a_1, a_2, a_3 > 0$, $r_1 > 0$, $r_2 \in \mathbb{R}$ and $r_3 \in (0, 1)$, then

$$\lim_{n\to\infty} \frac{1}{\gamma_n}\left(1 - \frac{v_n}{v_{n-1}}\right) = \begin{cases} 0, & \text{if } r_1 < 1 - r_3, \\ \frac{a_3 r_3}{2a_1}, & \text{if } r_1 = 1 - r_3, \\ \infty, & \text{if } r_1 > 1 - r_3. \end{cases}$$

On the other hand side, if the noise-levels σ_n are decreasing with exponential decay and the step-sizes γ_n are monotonically decreasing then Assumption A.3(ii) is typically not satisfied. In fact, if $\gamma_n \geq \gamma_{n+1}$ for $n \geq n_0$, $\lim_{n\to\infty} \gamma_n = 0$ and $\limsup_{n\to\infty} \sigma_{n+1}/\sigma_n < 1$ then $\lim_{n\to\infty} \gamma_n^{-1}(1 - v_n/v_{n-1}) = \infty$.

The case of an exponential decay of the noise-levels σ_n can be treated by applying Proposition 2.3. Assume that conditions (I)–(III), A.1 and A.2 are satisfied. Assume further that there exist $r \in (0, L)$ and $c \in (0, \infty)$ such that for all $n \in \mathbb{N}$,

(a) $\sigma_n^2 \leq c \exp(-2rn)$ and
(b) $\varepsilon_n \leq c \exp\bigl(-r \sum_{k=1}^{n} \gamma_k\bigr)$.

Then there exists $\kappa \in (0, \infty)$ such that for all $n \in \mathbb{N}$,

$$\mathbb{E}\bigl[\|\theta_n - \theta^*\|^p\bigr]^{1/p} \leq \kappa \exp\left(-r \sum_{k=1}^{n} \gamma_k\right). \tag{26}$$

Proof of (26) Since $\lim_{n\to\infty} \gamma_n = 0$ and $1 - x \leq \exp(-x)$ for all $x \in [0, 1]$ we have $(1 - \gamma_n r) \leq \exp(-r\gamma_n)$ for n sufficiently large. Hence there exists $n_1 \in \mathbb{N}$ such that for all $n \geq j \geq n_1$,

$$\tau_{j,n}(r) \leq \exp\left(-r \sum_{k=j+1}^{n} \gamma_k\right). \tag{27}$$

Using (27) as well as Assumption (b) we get for all $n \geq j \geq n_1$,

$$e_{j,n}(r) \leq (1 + c) \exp\left(-r \sum_{k=1}^{n} \gamma_k\right). \tag{28}$$

Choosing n_1 large enough we may also assume that $\gamma_n \leq 1/2$ for all $n \geq n_1$. Employing (27) and Assumption (a) we then conclude that for all $n \geq j \geq n_1$,

$$\begin{aligned} s_{j,n}^2(r) &= \sum_{k=j}^{n} \gamma_k^2 \sigma_k^2 (\tau_{k,n}(r))^2 \leq \sum_{k=j}^{n} (1 + c) \exp\left(-2rk - 2r \sum_{\ell=k+1}^{n} \gamma_\ell\right) \\ &= \exp\left(-2r \sum_{\ell=j+1}^{n} \gamma_\ell\right) \sum_{k=j}^{n} (1 + c) \exp\left(-2rk + 2r \sum_{\ell=j+1}^{k} \gamma_\ell\right) \end{aligned}$$

$$\leq \exp\left(-2r \sum_{\ell=j+1}^{n} \gamma_\ell\right) \sum_{k=j}^{n} (1+c)\exp(-rk)$$
$$\leq \exp\left(-2r \sum_{\ell=j+1}^{n} \gamma_\ell\right) \frac{(1+c)}{1-\exp(-r)}. \tag{29}$$

Combining (27) to (29) with Proposition 2.3 completes the proof of (26). □

So far we proved error estimates for the single random variables θ_n. In the following theorem we establish error estimates, which allow to control the quality of approximation for the whole sequence $(\theta_n)_{n\geq k_0}$ starting from some time k_0.

To this end we employ Assumption A.4 from below, which is stronger than Assumption A.3.

A.4 (Assumptions on $(\gamma_n)_{n\in\mathbb{N}}$, $(\varepsilon_n)_{n\in\mathbb{N}}$ and $(\sigma_n)_{n\in\mathbb{N}}$)

We have $v_n > 0$ for all $n \in \mathbb{N}$. Furthermore, with L according to A.1(i), there exist $c_1, c_2, \eta_1 \in (0,\infty)$ as well as $\eta_2 \in (0,1]$ such that $\eta_1 > (1-\eta_2)/p$ and

(i) $\limsup_{n\to\infty} \frac{\varepsilon_n}{v_n} < \infty$,

(ii) $\limsup_{n\to\infty} \frac{1}{\gamma_n} \frac{v_{n-1}-v_n}{v_{n-1}} < L$ and $v_n \leq \frac{c_1}{n^{\eta_1}}$ for all but finitely many $n \in \mathbb{N}$,

(iii) $\gamma_n \leq \frac{c_2}{n^{\eta_2}}$ for all but finitely many $n \in \mathbb{N}$.

Theorem 2.8 (Robbins–Monro approximation) *Assume that conditions (I)–(III), A.1, A.2 and A.4 are satisfied and let*

$$\eta^* = \eta_1 - (1-\eta_2)/p.$$

Then for all $\eta \in (0,\eta^*)$ there exists a constant $\kappa \in (0,\infty)$ and $n_0 \in \mathbb{N}$ such that for all $k_0 \geq n_0$

$$\mathbb{E}\left[\sup_{k\geq k_0} k^{p\eta}\,\|\theta_k - \theta^*\|^p\right]^{1/p} \leq \kappa\, k_0^{-(\eta^*-\eta)}. \tag{30}$$

Proof Clearly, we may assume that $\theta^* = 0$. Fix $\eta \in (0,\eta^*)$.

We again use the quantities introduced in (8). Since Assumption A.4 is stronger than Assumption A.3 we see from the proof of Theorem 2.5 that there exist $r \in (0,L)$, $\kappa_1 \in (0,\infty)$ and $n_1 \in \mathbb{N}$ such that for all $n \geq n_1$ we have $\gamma_n < 1/L$ and

$$\tau_{n_1,n}(r) + e_{n_1,n}(r) + s_{n_1,n}(r) \leq \kappa_1 v_n, \tag{31}$$

cf. (19). By A.4(ii) and A.4(iii) we may further assume that for all $n \geq n_1$,

$$v_n \leq c_1/n^{\eta_1} \quad \text{and} \quad \gamma_n < \min(1, 1/(2r)). \tag{32}$$

Fix $k_0 \geq n_1$ and define a strictly increasing sequence $(k_\ell)_{\ell\in\mathbb{N}_0}$ in $\mathbb{N}$ by

$$k_\ell = \min\left\{ m \geq k_0 : \sum_{k=k_0+1}^{m} \gamma_k \geq \ell \right\}.$$

Observing the upper bound for γ_n in (32) it is then easy to see that for all $\ell \in \mathbb{N}$,

$$\sum_{k=k_{\ell-1}+1}^{k_\ell} \gamma_k \leq 2. \tag{33}$$

In the following we write $\tau_{k,n}$, $e_{k,n}$ and $s_{k,n}$ in place of $\tau_{k,n}(r)$, $e_{k,n}(r)$ and $s_{k,n}(r)$, respectively. We estimate the decay of the sequence $(\tau_{k_0,k_\ell})_{\ell\in\mathbb{N}}$. Let $\ell \in \mathbb{N}$. Using (32), the fact that $1 - x \geq \exp(-2x)$ for all $x \in [0, 1/2]$, the estimate (33) and the fact that $1 - x \leq \exp(-x)$ for all $x \in [0, 1]$ we get

$$\tau_{k_0,k_{\ell-1}} = \tau_{k_0,k_\ell} \prod_{k=k_{\ell-1}+1}^{k_\ell} (1-\gamma_k r)^{-1} \leq \tau_{k_0,k_\ell} \prod_{k=k_{\ell-1}+1}^{k_\ell} \exp(2r\gamma_k) \leq \tau_{k_0,k_\ell} \exp(4r) \tag{34}$$

as well as

$$\tau_{k_0,k_\ell} \leq \prod_{k=k_0+1}^{k_\ell} \exp(-r\gamma_k) \leq \exp(-r\ell). \tag{35}$$

Next, we establish a lower bound for the growth of the sequence $(k_\ell)_{\ell\in\mathbb{N}_0}$, namely

$$k_\ell \geq K_\ell \tag{36}$$

for all $\ell \in N_0$, where

$$K_\ell = \begin{cases} \left(\ell(1-\eta_2)/c_2 + k_0^{1-\eta_2}\right)^{\frac{1}{1-\eta_2}}, & \text{if } \eta_2 < 1, \\ k_0 \exp(\ell/c_2), & \text{if } \eta_2 = 1. \end{cases}$$

In fact, by A.4(iii) we get

$$\ell \leq \sum_{k=k_0+1}^{k_\ell} \gamma_k \leq \sum_{k=k_0+1}^{k_\ell} \frac{c_2}{k^{\eta_2}} \leq c_2 \int_{k_0}^{k_\ell} x^{-\eta_2}\, dx = \begin{cases} \frac{c_2}{1-\eta_2}\left(k_\ell^{1-\eta_2} - k_0^{1-\eta_2}\right), & \text{if } \eta_2 < 1, \\ c_2 \ln\left(\frac{k_\ell}{k_0}\right), & \text{if } \eta_2 = 1. \end{cases}$$

which yields (36).

We are ready to establish the claimed estimate in pth mean (30). Similar to the proof of Proposition 2.3 we consider the process $(\zeta_n)_{n\geq n_1}$ and the martingale $(M_n)_{n\geq n_1}$ given by (13), where k_0 is replaced by n_1. As in the proof of Proposition 2.3 we obtain the maximum estimate in pth mean (17) for the process $(\zeta_n)_{n\geq n_1}$ and the estimate in $p/2$th mean (18) for the quadratic variation $([M]_n)_{n\geq n_1}$. Combining these two estimates we

see that for sufficiently large n_1 there exists a constant $\kappa_2 \in (0,\infty)$, such that for every $n \geq n_1$ we have

$$\mathbb{E}\left[\max_{n_1\leq k\leq n} \|\zeta_k\|^p\right] \leq \kappa_2 \left(\mathbb{E}\big[\|\theta_{n_1}\|^p\big] + \frac{s_{n_1,n}^p + e_{n_1,n}^p}{\tau_{n_1,n}^p}\right).$$

Using the latter inequality as well as (34), Theorem 2.5 and (31) we may thus conclude that there exists a constant $\kappa_3 \in (0,\infty)$, which may depend on n_1 but not on k_0 such that for every $\ell \in \mathbb{N}$ we have

$$\begin{aligned}\mathbb{E}\left[\max_{k=k_{\ell-1}+1,\dots,k_\ell} \|\theta_k\|^p\right] &\leq \tau_{n_1,k_{\ell-1}}^p\, \mathbb{E}\left[\max_{k=k_{\ell-1}+1,\dots,k_\ell} \|\zeta_k\|^p\right]\\ &\leq \kappa_2\, \exp(4rp)\, \tau_{n_1,k_\ell}^p \left(\mathbb{E}\big[\|\theta_{n_1}\|^p\big] + \frac{s_{n_1,k_\ell}^p + e_{n_1,k_\ell}^p}{\tau_{n_1,k_\ell}^p}\right)\\ &\leq \kappa_3 \left(\tau_{n_1,k_\ell}^p + s_{n_1,k_\ell}^p + e_{n_1,k_\ell}^p\right) \leq \kappa_3 \kappa_1^p\, v_{k_\ell}^p.\end{aligned}$$

Hence, there exists a constant $\kappa_4 \in (0,\infty)$ that does not depend on k_0 such that

$$\mathbb{E}\left[\sup_{k>k_0} k^{p\eta}\|\theta_k\|^p\right] \leq \sum_{\ell\in\mathbb{N}} \mathbb{E}\left[\max_{k=k_{\ell-1}+1,\dots,k_\ell} k_\ell^{p\eta}\|\theta_k\|^p\right] \leq \kappa_4 \sum_{\ell\in\mathbb{N}} k_\ell^{p\eta}\, v_{k_\ell}^p. \tag{37}$$

Using (32), the fact that $p(\eta_1-\eta) > 1-\eta_2$, due to the choice of η, and the lower bound in (36) we obtain

$$\begin{aligned}\sum_{\ell\in\mathbb{N}} k_\ell^{p\eta}\, v_{k_\ell}^p &\leq c_1^p \sum_{\ell\in\mathbb{N}} k_\ell^{-p(\eta_1-\eta)}\\ &\leq c_1^p \begin{cases} \sum_{\ell\in\mathbb{N}} \big(\ell(1-\eta_2)/c_2 + k_0^{1-\eta_2}\big)^{-\frac{p(\eta_1-\eta)}{1-\eta_2}}, & \text{if } \eta_2 < 1,\\ \sum_{\ell\in\mathbb{N}} \big(k_0 \exp(\ell/c_2)\big)^{-p(\eta_1-\eta)}, & \text{if } \eta_2 = 1,\end{cases}\\ &\leq \kappa_5\, k_0^{-p(\eta_1-\eta)+1-\eta_2}\end{aligned} \tag{38}$$

with a constant $\kappa_5 \in (0,\infty)$ that does not depend on k_0. Combining (37) with (38) yields the claimed maximum estimate in pth mean. □

In analogy to Corollary 2.6 we next treat the particular case of polynomial step-sizes γ_n and noise-levels σ_n.

Corollary 2.9 (Polynomial step-sizes and noise-levels) *Assume that conditions (I)–(III), A.1 and A.2 are satisfied and choose L according to A.1(i). Take $\gamma_0, \sigma_0 \in (0,\infty)$, $r_1 \in (0,1]$ and $r_2 \in (-r_1,\infty)$ with*

(a) $r_1 < 1$ *or* $\big(r_1 = 1$ *and* $\gamma_0 > \frac{1+r_2}{2L}\big)$,
(b) $\frac{r_1+r_2}{2} > \frac{1-r_1}{p}$

and let for all $n \in \mathbb{N}$,

$$\gamma_n = \gamma_0 \frac{1}{n^{r_1}}, \quad \sigma_n^2 = \sigma_0^2 \frac{1}{n^{r_2}}.$$

Assume further that

$$\limsup_{n\to\infty} n^{(r_1+r_2)/2}\, \varepsilon_n < \infty.$$

Then for all $\eta \in (0, \frac{r_1+r_2}{2} - \frac{1-r_1}{p})$ *there exists a constant* $\kappa \in (0,\infty)$ *such that for all* $k_0 \in \mathbb{N}$,

$$\mathbb{E}\Big[\sup_{k\geq k_0} k^{p\eta}\, \|\theta_k - \theta^*\|^p\Big]^{1/p} \leq \kappa\, k_0^{-\left(\frac{r_1+r_2}{2} - \frac{1-r_1}{p} - \eta\right)}. \tag{39}$$

Proof We first verify Assumption A.4. By definition of γ_n and σ_n we have

$$v_n = \sqrt{\gamma_0}\sigma_0 \frac{1}{n^{(r_1+r_2)/2}}.$$

Thus, A.4(i) is satisfied due to the assumption on the sequence $(\varepsilon_n)_{n\in\mathbb{N}}$ and the first part of A.4(ii) is satisfied due to Assumption (a), see the proof of Corollary 2.6. Observing Assumption (b) it is obvious that the second part of A.4(ii) and Assumption A.4(iii) are satisfied for

$$\eta_1 = (r_1 + r_2)/2, \quad \eta_2 = r_1, \quad c_1 = \sqrt{\gamma_0}\sigma_0, \quad c_2 = \gamma_0.$$

Since conditions (I)–(III), A.1 and A.2 are part of the corollary, we may apply Theorem 2.8 to obtain the claimed error estimate. □

2.2 Estimates for the Polyak–Ruppert algorithm

Now we turn to the analysis of Polyak–Ruppert averaging. For $n \in \mathbb{N}$ we let

$$\bar{\theta}_n = \frac{1}{\bar{b}_n} \sum_{k=1}^{n} b_k\, \theta_k, \tag{40}$$

where $(b_k)_{k\in\mathbb{N}}$ is a fixed sequence of strictly positive reals and

$$\bar{b}_n = \sum_{k=1}^{n} b_k.$$

We estimate the speed of convergence of $(\bar{\theta}_n)_{n\in\mathbb{N}}$ to θ^* in pth mean in terms of the sequence $(\bar{v}_n)_{n\in\mathbb{N}}$ given by

$$\bar{v}_n = \frac{v_n}{\sqrt{n\,\gamma_n}} = \frac{\sigma_n}{\sqrt{n}}.$$

To this end we will replace Assumptions A.1, A.2 and A.3 by Assumptions B.1, B.2 and B.3 from below. Note that B.2 coincides with A.2 while B.1 is stronger than A.1 and B.3 is stronger than A.3, see Remark 2.10 below.

B.1 (Assumptions on f and θ^*)
There exist $L, L', L'', \lambda \in (0,\infty)$ and a matrix $H \in \mathbb{R}^{d\times d}$ such that for all $\theta \in \mathbb{R}^d$

(i) $\langle \theta - \theta^*, f(\theta)\rangle \le -L\,\|\theta-\theta^*\|^2$,
(ii) $\langle \theta - \theta^*, f(\theta)\rangle \le -L'\,\|f(\theta)\|^2$ and
(iii) $\|f(\theta) - H(\theta-\theta^*)\| \le L''\,\|\theta-\theta^*\|^{1+\lambda}$.

B.2 (Assumptions on $(R_n)_{n\in\mathbb{N}}$ and $(D_n)_{n\in\mathbb{N}}$)
It holds

(i) $\sup_{n\in\mathbb{N}} \operatorname{esssup} \|R_n\| < \infty$ and
(ii) $\sup_{n\in\mathbb{N}} \mathbb{E}[\|D_n\|^p] < \infty$.

B.3 (Assumptions on $(\gamma_n)_{n\in\mathbb{N}}$, $(\varepsilon_n)_{n\in\mathbb{N}}$, $(\sigma_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$)
We have $\sigma_n > 0$ for all $n \in \mathbb{N}$. The sequence $(\gamma_n)_{n\in\mathbb{N}}$ is decreasing and the sequences $(n\gamma_n)_{n\in\mathbb{N}}$ and $(b_n\sigma_n)_{n\in\mathbb{N}}$ are increasing. Moreover, with L and λ according to B.1 there exist $\nu, c_1, c_2, c_3 \in [0,\infty)$ with $c_2 > (\nu+1)/L$ such that

(i) $\limsup_{n\to\infty} \frac{\varepsilon_n}{\bar v_n} < \infty$,
(ii) $\limsup_{n\to\infty} \frac{1}{\gamma_n}\frac{\bar v_{n-1}-\bar v_n}{\bar v_{n-1}} < L$ and $v_n^{1+\lambda} \le c_1 \bar v_n$ for all but finitely many $n \in \mathbb{N}$,
(iii) $\gamma_n \ge \dfrac{c_2}{n}$ for all but finitely many $n \in \mathbb{N}$,
(iv) $b_m \le c_3\, b_n \Big(\dfrac{m}{n}\Big)^{\nu}$ for all $m \ge n \ge 1$,

i.e., the sequence $(b_n)_{n\in\mathbb{N}}$ has at most polynomial growth.

Remark 2.10 (Discussion of Assumptions B.1 and B.3) We first show that Assumption B.3 implies Assumption A.3. Since $(n\gamma_n)_{n\in\mathbb{N}}$ is increasing we have $\varepsilon_n/v_n = \varepsilon_n/(\sqrt{n\gamma_n}\bar v_n) \le \varepsilon_n/(\sqrt{\gamma_1}\,\bar v_n)$ for every $n \in \mathbb{N}$, which proves that B.3 implies A.3(i). Furthermore,

$$\frac{v_n - v_{n+1}}{v_n} = \frac{\bar v_n - \frac{\sqrt{(n+1)\gamma_{n+1}}}{\sqrt{n\gamma_n}}\bar v_{n+1}}{\bar v_n} \le \frac{\bar v_n - \bar v_{n+1}}{\bar v_n}$$

for every $n \in \mathbb{N}$, which proves that B.3 implies A.3(ii).

We add that, due to the presence of Assumption B.1(ii), it is sufficient to require that f satisfies the inequality in B.1(iii) on some open ball around θ^*. In fact, let $D \subset \mathbb{R}^d$ and $\delta \in (0,\infty)$ be such that $B(\theta^*,\delta) = \{\theta \in \mathbb{R}^d\colon \|\theta-\theta^*\| < \delta\} \subset D$. Let $c_2, c_3, c_3', \lambda \in (0,\infty)$ and $H \in \mathbb{R}^{d\times d}$ and consider the conditions

$$\forall \theta \in D\colon \langle \theta-\theta^*, f(\theta)\rangle \le -c_2\,\|f(\theta)\|^2, \tag{ii}$$

$$\forall \theta \in D\colon \|f(\theta) - H(\theta-\theta^*)\| \le c_3\|\theta-\theta^*\|^{1+\lambda}, \tag{iii}$$

$$\forall \theta \in B(\theta^*,\delta)\colon \|f(\theta) - H(\theta-\theta^*)\| \le c_3'\|\theta-\theta^*\|^{1+\lambda}. \tag{iii'}$$

Then

$$f \text{ satisfies (ii) and (iii')} \Rightarrow \|H\| \le 1/c_2 \text{ and } f \text{ satisfies (iii) for every } c_3 \ge \max\left(c_3', \tfrac{2}{c_2\delta^\lambda}\right), \tag{41}$$

where $\|H\|$ denotes the induced matrix norm of H.

For a proof of (41) we first note that (iii') implies that $H(\theta) = \lim_{0<\varepsilon\to 0} \frac{1}{\varepsilon} f(\theta^* + \varepsilon\theta)$ for every $\theta \in \mathbb{R}^d$. Using (2) we conclude that $\|H\| \le 1/c_2$. For $\theta \in D \setminus B(\theta^*, \delta)$ we have $\|\theta - \theta^*\| \ge \delta$. Observing the latter fact and using (2) again we conclude

$$\|f(\theta) - H(\theta - \theta^*)\| \le \|f(\theta)\| + \|H(\theta - \theta^*)\| \le \tfrac{2}{c_2}\|\theta - \theta^*\| \le \tfrac{2}{c_2\delta^\lambda}\|\theta - \theta^*\|^{1+\lambda}.$$

As an immediate consequence of (41) with the choice $\delta \le 1$ we obtain that if f satisfies B.1(ii),(iii) then f satisfies B.1(iii) for every $\lambda' \in [0, \lambda]$ with L'' replaced by $\max(L'', 2/(L'\delta^{\lambda'}))$.

We proceed with the regression problem of Example 2.2.

Example 2.11 (Regression) Consider the function f given by (6) in Example 2.2, which satisfies B.1(i) and (ii). Clearly, for all $\theta \in \mathbb{R}^d$,

$$f(\theta) = f(\theta) - f(\theta^*) = B(\theta^* - \theta),$$

so that f satisfies B.1(iii) with $H = -B$ and any $L'', \lambda \in (0, \infty)$.

Theorem 2.12 (Polyak–Ruppert approximation) *Assume that conditions (I)–(III) and B.1-B.3 are satisfied. Put* $q = \frac{p}{1+\lambda}$ *with* λ *according to B.1. Then there exists* $\kappa \in (0, \infty)$ *such that the Polyak–Ruppert algorithm* (40) *satisfies for all* $n \in \mathbb{N}$

$$\mathbb{E}\big[\|\bar\theta_n - \theta^*\|^q\big]^{1/q} \le \kappa\, \bar v_n.$$

For the proof of Theorem 2.12 we follow the approach of the classical paper [16] by first comparing the dynamical system $(\theta_n)_{n\ge 0}$ with a linearised version $(y_n)_{n\ge 0}$ given by $y_0 = \theta_0$ and

$$y_n = y_{n-1} + \gamma_n\big(H(y_{n-1} - \theta^*) + \sigma_n D_n\big)$$

for $n \in \mathbb{N}$.

Lemma 2.13 *Assume that conditions (I)–(III) and B.1-B.3 are satisfied. Put* $q = \frac{p}{1+\lambda}$ *with* λ *according to B.1. Then there exists* $\kappa \in (0, \infty)$ *such that for all* $n \in \mathbb{N}$

$$\mathbb{E}\big[\|\theta_n - y_n\|^q\big]^{1/q} \le \kappa\, \bar v_n.$$

Proof Without loss of generality we may assume that $\theta^* = 0$.

Using B.3(i),(ii) we see that there exist $r \in (0, L)$, $n_0 \in \mathbb{N}$ and $\kappa_0 \in (0, \infty)$ such that for all $n \geq n_0$ we have

$$\varepsilon_n \leq \kappa_0 \, \bar{v}_n \tag{42}$$

and

$$\frac{\bar{v}_{n-1}}{\bar{v}_n} \leq \frac{1}{1 - \gamma_n r}. \tag{43}$$

Since $\lim_{n\to\infty} \gamma_n = 0$ we may assume that $\gamma_n \leq 1/(2r)$ for all $n \geq n_0$.

By Remark 2.10 conditions A.1-A.3 are satisfied and we may apply Theorem 2.5 to obtain the existence of $\kappa_1 \in (0, \infty)$ such that for all $n \in \mathbb{N}$,

$$\mathbb{E}[\|\theta_n\|^p]^{1/p} \leq \kappa_1 \, v_n. \tag{44}$$

Furthermore, estimate (12) in the proof of Proposition 2.3 is valid, i.e., there exists $n_1 \in \mathbb{N}$ such that for all $n \geq n_1$ and all $\theta \in \mathbb{R}^d$,

$$\|\theta + \gamma_n f(\theta)\| \leq (1 - \gamma_n(r + L)/2)\|\theta\|. \tag{45}$$

By Assumption B.1(iii) we have

$$H\theta = \lim_{\varepsilon \downarrow 0} \varepsilon^{-1} f(\varepsilon\theta) \tag{46}$$

for every $\theta \in \mathbb{R}^d$. Using (45) we may therefore conclude that for all $n \geq n_1$ and all $\theta \in \mathbb{R}^d$,

$$\|\theta + \gamma_n H\theta\| \leq (1 - \gamma_n(r + L)/2)\|\theta\|. \tag{47}$$

Let $n_2 = \max(n_0, n_1)$. For $n \geq n_2$ we put

$$z_n = \theta_n - y_n \text{ and } \delta_n = \frac{\mathbb{E}\big[\|z_n\|^q\big]^{1/q}}{\bar{v}_n}.$$

Let $n > n_2$. Using (47), Assumptions B.1(iii), B.2(i) and (42) we see that there exists $\kappa_2 \in (0, \infty)$ such that

$$\begin{aligned}
\|z_n\| &= \|z_{n-1} + \gamma_n\big(Hz_{n-1} + f(\theta_{n-1}) - H\theta_{n-1} + \varepsilon_n \, R_n\big)\| \\
&\leq \|z_{n-1} + \gamma_n H z_{n-1}\| + \gamma_n \|f(\theta_{n-1}) - H\theta_{n-1}\| + \gamma_n \, \varepsilon_n \, \|R_n\| \\
&\leq (1 - \gamma_n(r + L)/2)\|z_{n-1}\| + \gamma_n \, L'' \, \|\theta_{n-1}\|^{1+\lambda} + \kappa_2 \, \gamma_n \, \bar{v}_n \quad \text{a.s.},
\end{aligned}$$

and employing (43), (44), B.3(ii) and the fact that $\gamma_n \leq 1/(2r)$ we conclude that

$$\begin{aligned}
\delta_n &\leq (1 - \gamma_n(r + L)/2) \, \frac{\bar{v}_{n-1}}{\bar{v}_n} \, \delta_{n-1} + L'' \, \gamma_n \frac{\bar{v}_{n-1}}{\bar{v}_n} \, \frac{\mathbb{E}\big[\|\theta_{n-1}\|^p\big]^{1/q}}{\bar{v}_{n-1}} + \kappa_2 \gamma_n \\
&\leq \frac{1 - \gamma_n(r + L)/2}{1 - \gamma_n \, r} \, \delta_{n-1} + \frac{L'' \, \kappa_1^{1+\lambda} \, c_1}{1 - \gamma_n \, r} \, \gamma_n + \kappa_2 \gamma_n \\
&\leq (1 - \gamma_n(L - r)/2) \, \delta_{n-1} + \big(2L'' \, \kappa_1^{1+\lambda} \, c_1 + \kappa_2\big) \, \gamma_n.
\end{aligned} \tag{48}$$

Put $\kappa_3 = (L-r)/2 > 0$ and $\kappa_4 = 2L'' \kappa_1^{1+\lambda} c_1 + \kappa_2$. By (48) we have for $n \geq n_2$ that $\delta_n \leq (1 - \kappa_3 \gamma_n)\, \delta_{n-1} + \kappa_4 \gamma_n$ or, equivalently,

$$\frac{\delta_n}{\kappa_4} - \frac{1}{\kappa_3} \leq (1 - \kappa_3 \gamma_n)\Big(\frac{\delta_{n-1}}{\kappa_4} - \frac{1}{\kappa_3}\Big),$$

which yields,

$$\frac{\delta_n}{\kappa_4} - \frac{1}{\kappa_3} \leq \left(\frac{\delta_{n_2}}{\kappa_4} - \frac{1}{\kappa_3}\right) \exp\left(-\kappa_3 \sum_{k=n_2+1}^{n} \gamma_k\right).$$

Since $\sum_{k\in\mathbb{N}} \gamma_k = \infty$, due to Assumption B.3(iii), we conclude that

$$\limsup_{n\to\infty} (\delta_n/\kappa_4 - 1/\kappa_3) \leq 0,$$

which finishes the proof. □

Proof of Theorem 2.12 Without loss of generality we may assume that $\theta^* = 0$.

For all $n \in \mathbb{N}$ we have

$$\bar{\theta}_n = \frac{1}{\bar{b}_n} \sum_{k=1}^{n} b_k(\theta_k - y_k) + \frac{1}{\bar{b}_n} \sum_{k=1}^{n} b_k y_k. \tag{49}$$

We separately analyze the two terms on the right hand side of (49).

By Assumption B.3(iv) it follows that

$$\bar{b}_n \geq \frac{b_n}{c_3} \sum_{k=1}^{n} \left(\frac{k}{n}\right)^{\nu} \geq \kappa_1 n b_n, \tag{50}$$

where $\kappa_1 = (c_3(\nu+1))^{-1}$.

Employing Lemma 2.13 as well as (50) and the fact that $(b_k\sigma_k)_{k\in\mathbb{N}}$ is increasing we see that there exists $\kappa_2 \in (0,\infty)$ such that for all $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{\bar{b}_n}\sum_{k=1}^{n} b_k(\theta_k - y_k)\right\|^q\right]^{1/q} &\leq \frac{\kappa_2}{n\, b_n} \sum_{k=1}^{n} b_k \bar{v}_k = \frac{\kappa_2}{n\, b_n} \sum_{k=1}^{n} \frac{b_k\, \sigma_k}{\sqrt{k}} \\ &\leq \frac{\kappa_2\, \sigma_n}{n} \sum_{k=1}^{n} \frac{1}{\sqrt{k}} \leq 2\kappa_2\, \bar{v}_n, \end{aligned} \tag{51}$$

where we used $\sum_{k=1}^{n} 1/\sqrt{k} \leq 2\sqrt{n}$ in the latter step.

Next, put $b_0 = 0$ and let

$$\Upsilon_{k,n} = \prod_{\ell=k+1}^{n} (I_d + \gamma_\ell H),$$

where $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix, as well as

$$\bar{\Upsilon}_{k,n} = \sum_{m=k}^{n} b_m \Upsilon_{k,m}$$

for $0 \le k \le n$. Then, for all $n \in \mathbb{N}$,

$$y_n = \Upsilon_{0,n}\,\theta_0 + \sum_{k=1}^{n} \gamma_k\,\sigma_k\,\Upsilon_{k,n}\,D_k$$

and

$$\sum_{k=1}^{n} b_k\,y_k = \bar{\Upsilon}_{0,n}\,\theta_0 + \sum_{k=1}^{n} \gamma_k\,\sigma_k\bar{\Upsilon}_{k,n}\,D_k.$$

Using the Burkholder–Davis–Gundy inequality we obtain that there exists a constant $\kappa_3 \in (0,\infty)$ such that for all $n \in \mathbb{N}$,

$$\mathbb{E}\left[\left\|\sum_{k=1}^{n} b_k\,y_k\right\|^q\right]^{1/q} \le \kappa_3 \left(\|\bar{\Upsilon}_{0,n}\|\,\|\theta_0\| + \mathbb{E}\left[\left(\sum_{k=1}^{n}\gamma_k^2\,\sigma_k^2\,\|\bar{\Upsilon}_{k,n}\|^2\|D_k\|^2\right)^{q/2}\right]^{1/q}\right). \tag{52}$$

By Assumption B.2(ii) there exists a constant $\kappa_4 \in (0,\infty)$ such that for all $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{k=1}^{n}\gamma_k^2\,\sigma_k^2\,\|\bar{\Upsilon}_{k,n}\|^2\,\|D_k\|^2\right)^{q/2}\right]^{1/q} &\le \left(\sum_{k=1}^{n}\gamma_k^2\,\sigma_k^2\,\|\bar{\Upsilon}_{k,n}\|^2\,\mathbb{E}\big[\|D_k\|^q\big]^{2/q}\right)^{1/2}\\ &\le \kappa_4\left(\sum_{k=1}^{n}\gamma_k^2\,\sigma_k^2\,\|\bar{\Upsilon}_{k,n}\|^2\right)^{1/2}.\end{aligned} \tag{53}$$

We proceed with estimating the norms $\|\bar{\Upsilon}_{k,n}\|$. Since $c_2 > (\nu+1)/L$ we can fix $r \in ((\nu+1)/c_2, L)$ and proceed as in the proof of Lemma 2.13 to conclude that there exists $n_1 \in \mathbb{N}$ such that for all $n \ge n_1$

$$\|I_d + \gamma_n H\| \le 1 - \gamma_n(r+L)/2,$$

see (47). The latter fact and the assumption that the sequence $(n\gamma_n)_{n\in\mathbb{N}}$ is increasing jointly imply that for $n \ge k \ge n_1 - 1$

$$\|\Upsilon_{k,n}\| \le \prod_{\ell=k+1}^{n}\left(1-\frac{\gamma_\ell\,(r+L)}{2}\right) \le \prod_{\ell=k+1}^{n}\left(1-\frac{\gamma_k\,k\,(r+L)}{2\ell}\right)$$
$$\le \exp\left(-\frac{(r+L)\gamma_k k}{2}\sum_{\ell=k+1}^{n}\frac{1}{\ell}\right) \le \left(\frac{k+1}{n+1}\right)^{(r+L)\gamma_k k/2},$$

where we used that $1-z \le e^{-z}$ for all $z \in \mathbb{R}$. Employing the latter estimate as well as Assumption B.3(iv) we get that for $n \ge k \ge n_1 - 1$

$$\begin{aligned}\|\bar{\Upsilon}_{k,n}\| &\le \sum_{\ell=k}^{n} b_\ell\,\|\Upsilon_{k,\ell}\| \le c_3\,b_k \sum_{\ell=k}^{n}\left(\frac{\ell}{k}\right)^{\nu}\left(\frac{k+1}{\ell+1}\right)^{(r+L)\gamma_k k/2}\\ &\le c_3\,b_k\,2^{\nu}\sum_{\ell=k}^{n}\left(\frac{k+1}{\ell+1}\right)^{(r+L)\gamma_k k/2-\nu}.\end{aligned} \tag{54}$$

Put $\beta_k = (r+L)\gamma_k k/2 - \nu$ and note that by the choice of r and by B.3(iii) one has for k large enough that

$$\beta_k = (L-r)\gamma_k k/2 + r\gamma_k k - \nu > (L-r)\gamma_k k/2 + 1.$$

Choosing n_1 large enough we therefore conclude that there exists $\kappa_5 \in (0,\infty)$ such that for $n \ge k \ge n_1 - 1$,

$$\sum_{\ell=k}^{n}\Big(\frac{k+1}{\ell+1}\Big)^{(r+L)\gamma_k k/2-\nu} \le 1 + (k+1)^{\beta_k}\int_{k+1}^{\infty} t^{-\beta_k}\,dt = 1 + \frac{k+1}{\beta_k - 1} \le \frac{\kappa_5}{\gamma_k}.$$

In combination with (54) we see that there exists $\kappa_6 \in (0,\infty)$ such that for all $n \ge k \ge n_1 - 1$,

$$\|\bar{\Upsilon}_{k,n}\| \le \kappa_6\,\frac{b_k}{\gamma_k}. \tag{55}$$

For $0 \le k < n_1 - 1 \le n$ we have

$$\bar{\Upsilon}_{k,n} = \sum_{\ell=k}^{n_1-2} b_\ell\,\Upsilon_{k,\ell} + \bar{\Upsilon}_{n_1-1,n}\,\Upsilon_{k,n_1-1}$$

and, observing (55), we may thus conclude that for $0 \le k < n_1 - 1 \le n$,

$$\|\bar{\Upsilon}_{k,n}\| \le \left(\max_{0\le j\le\ell\le n_1-1}\|\Upsilon_{j,\ell}\|\right)\left(\sum_{\ell=1}^{n_1-2} b_\ell + \kappa_6\,\frac{b_{n_1-1}}{\gamma_{n_1-1}}\right). \tag{56}$$

Using (55) as well as (56) and the fact that the sequence $(b_n\sigma_n)_{n\in\mathbb{N}}$ is increasing we conclude that there exists $\kappa_7 \in (0,\infty)$ such that for all $n \in \mathbb{N}$,

$$\left(\sum_{k=1}^{n} \gamma_k^2\, \sigma_k^2\, \|\bar{\Upsilon}_{k,n}\|^2\right)^{1/2} \le \kappa_7\, \sqrt{n}\, b_n\, \sigma_n. \tag{57}$$

Combining (52), (53) and (57), employing again that the sequence $(b_n\sigma_n)_{n\in\mathbb{N}}$ is increasing and observing (50) we see that there exists a constant $\kappa_8 \in (0,\infty)$ such that for all $n \in \mathbb{N}$,

$$\mathbb{E}\left[\left\|\frac{1}{\bar{b}_n}\sum_{k=1}^{n} b_k\, y_k\right\|^q\right]^{1/q} \le \frac{\kappa_8}{\bar{b}_n}\,\sqrt{n}\, b_n\, \sigma_n \le \frac{\kappa_8}{\kappa_1}\,\bar{v}_n \tag{58}$$

Combining (51) with (58) completes the proof of the theorem. □

We consider the particular case of polynomial step-sizes γ_n, noise-levels σ_n and weights b_n.

Corollary 2.14 (Polynomial step-sizes, noise-levels and weights) *Assume that conditions (I)–(III), B.1 and B.2 are satisfied and let* $q \in [\frac{p}{1+\lambda}, p)$ *with* λ *according to B.1(iii).*

Take $\gamma_0, \sigma_0, b_0 \in (0,\infty)$, $r_1 \in (0,1)$, $r_2 \in (-r_1,\infty)$ *and* $r_3 \in [r_2/2,\infty)$ *with*

$$\frac{1+r_2}{r_1+r_2} \le \frac{p}{q}$$

and let for all $n \in \mathbb{N}$,

$$\gamma_n = \gamma_0 \frac{1}{n^{r_1}}, \quad \sigma_n^2 = \sigma_0^2 \frac{1}{n^{r_2}}, \quad b_n = b_0 n^{r_3}.$$

Assume further that

$$\limsup_{n\to\infty} n^{(1+r_2)/2}\, \varepsilon_n < \infty$$

Then there exists a constant $\kappa \in (0,\infty)$ *such that for all* $n \in \mathbb{N}$,

$$\mathbb{E}[\|\bar{\theta}_n - \theta^*\|^q]^{1/q} \le \kappa\, n^{-\frac{1}{2}(r_2+1)}.$$

Proof Conditions (I)–(III), B.1 and B.2 are part of the corollary. Further note that by Remark 2.10 condition B.1(iii) remains true when replacing λ by $\lambda' = \frac{p}{q} - 1 \in (0,\lambda]$. We verify that condition B.3 holds as well with λ' in place of λ. Then the corollary is a consequence of Theorem 2.12 and the fact that $\bar{v}_n = \sigma_n/\sqrt{n} = \sigma_0 n^{-\frac{1}{2}(r_2+1)}$.

Since $r_1 \in (0, 1)$ it is clear that $(\gamma_n)_{n\in\mathbb{N}}$ is decreasing and $(n\gamma_n)_{n\in\mathbb{N}}$ is increasing. Furthermore,

$$(b_n\sigma_n)_{n\in\mathbb{N}} = (\sigma_0 b_0 n^{r_3 - r_2/2})_{n\in\mathbb{N}}$$

is increasing since $r_3 \geq r_2/2$. B.3(i) is satisfied due to the assumption on $(\varepsilon_n)_{n\in\mathbb{N}}$. Since $r_1 < 1$ the limes superior in B.3(ii) is zero. Moreover, the second estimate of B.3(ii) holds for an appropriate positive constant c_1 since $v_n^{1+\lambda'} = (\sqrt{\gamma_n}\sigma_n)^{1+\lambda'} = (\gamma_0\sigma_0^2)^{\frac{p}{2q}} n^{-\frac{p}{q}\frac{r_1+r_2}{2}}$ and $\frac{p}{q}\frac{r_1+r_2}{2} \geq \frac{r_2+1}{2}$ by assumption. Condition B.3(iii) is satisfied for any $c_2 \in (0, \infty)$ since $r_1 < 1$, and thus condition B.3(iv) is satisfied with $\nu = r_3$. □

3 Multilevel stochastic approximation

Throughout this section we fix $p \in [2, \infty)$, a probability space $(\Omega, \mathcal{F}, P)$, a scalar product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^d$ with induced norm $\|\cdot\|$, a non-empty set $\mathcal{U}$ equipped with some σ-field, a random variable $U : \Omega \to \mathcal{U}$ and a product-measurable function

$$F : \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}^d,$$

such that $F(\theta, U)$ is integrable for every $\theta \in \mathbb{R}^d$. We consider the function $f : \mathbb{R}^d \to \mathbb{R}^d$ given by

$$f(\theta) = \mathbb{E}[F(\theta, U)] \tag{59}$$

and we assume that f has a unique zero $\theta^* \in \mathbb{R}^d$.

Our goal is to compute θ^* by means of stochastic approximation algorithms based on the multilevel Monte Carlo approach. To this end we suppose that we are given a hierarchical scheme

$$F_1, F_2, \ldots : \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}^d$$

of suitable product-measurable approximations to F, such that $F_k(\theta, U)$ is integrable and $F_k(\theta, U) - F_{k-1}(\theta, U)$ can be simulated for all $\theta \in \mathbb{R}^d$ and $k \in \mathbb{N}$, where $F_0 = 0$.

For example one may think of U being Brownian motion and $F(\theta, U)$ being the payoff of an option in a U-driven financial market model. Then θ may be a parameter that influences the dynamics of the assets (e.g. drift) or has an immediate impact on the payoff (e.g. strike). $F_k(\theta, U)$ would typically be obtained by using a numerical scheme like the Euler-Maruyama scheme or the Milstein scheme to approximate the solution of the SDE. Examples are provided in Sect. 5 below.

To each random vector $F_k(\theta, U) - F_{k-1}(\theta, U)$ we assign a positive number $C_k \in (0, \infty)$, which depends only on the level k and may represent a deterministic worst case upper bound of the average computational cost or average runtime needed to compute a single simulation of $F_k(\theta, U) - F_{k-1}(\theta, U)$. As announced in the introduction we impose assumptions on the approximations F_k and the cost bounds C_k that are similar in spirit to the classical multilevel Monte Carlo setting, see [5].

C.1 (Assumptions on $(F_k)_{k\in\mathbb{N}}$ and $(C_k)_{k\in\mathbb{N}}$)

There exist measurable functions $\Gamma_1, \Gamma_2 \colon \mathbb{R}^d \to (0,\infty)$ and constants $M \in (1,\infty)$ and $K, \alpha, \beta \in (0,\infty)$ with $\alpha \geq \beta$ such that for all $k \in \mathbb{N}$ and all $\theta \in \mathbb{R}^d$

(i) $\mathbb{E}[\|F_k(\theta,U)-F_{k-1}(\theta,U)-\mathbb{E}[F_k(\theta,U)-F_{k-1}(\theta,U)]\|^p]^{1/p} \leq \Gamma_1(\theta)\, M^{-k\beta}$,
(ii) $\|\mathbb{E}[F_k(\theta,U) - F(\theta,U)]\| \leq \Gamma_2(\theta)\, M^{-k\alpha}$, and
(iii) $C_k \leq K M^k$.

We combine the Robbins–Monro algorithm with the classical multilevel approach taken in [5]. The proposed method uses in each Robbins–Monro step a multilevel estimate with a complexity that is adapted to the actual state of the system and increases in time.

The algorithm is specified by the parameters $\Gamma_1, \Gamma_2, M, \alpha, \beta$ from Assumption C.1, an initial vector $\theta_0 \in \mathbb{R}^d$,

(i) a sequence of step-sizes $(\gamma_n)_{n\in\mathbb{N}} \subset (0,\infty)$ tending to zero,
(ii) a sequence of bias-levels $(\varepsilon_n)_{n\in\mathbb{N}} \subset (0,\infty)$, and
(iii) a sequence of noise-levels $(\sigma_n)_{n\in\mathbb{N}} \subset (0,\infty)$.

The maximal level $m_n(\theta)$ and the number of replications $N_{n,k}(\theta)$ on level $k \in \{1,\dots,m_n(\theta)\}$ that are used by the multilevel estimator in the nth Robbins–Monro step depend on $\theta \in \mathbb{R}^d$ and are determined in the following way. We take

$$m_n(\theta) = 1 \vee \left\lceil \frac{1}{\alpha} \log_M \left(\frac{\Gamma_2(\theta)}{\varepsilon_n} \right) \right\rceil \in \mathbb{N}, \tag{60}$$

i.e. $m_n(\theta)$ is the smallest $m \in \mathbb{N}$ such that $\Gamma_2(\theta) M^{-\alpha m} \leq \varepsilon_n$ holds true for the bias bound in Assumption C.1(ii). Furthermore,

$$N_{n,k}(\theta) = \left\lceil \kappa_n(\theta)\, M^{-k(\beta+1/2)} \right\rceil, \tag{61}$$

where

$$\kappa_n(\theta) = \begin{cases} (\Gamma_1(\theta)/\sigma_n)^2\, M^{m_n(\theta)(\frac{1}{2}-\beta)_+}, & \text{if } \beta \neq 1/2, \\ (\Gamma_1(\theta)/\sigma_n)^2\, m_n(\theta), & \text{if } \beta = 1/2. \end{cases} \tag{62}$$

Take a sequence $(U_{n,k,\ell})_{n,k,\ell\in\mathbb{N}}$ of independent copies of U. We use

$$Z_n(\theta) = \sum_{k=1}^{m_n(\theta)} \frac{1}{N_{n,k}(\theta)} \sum_{\ell=1}^{N_{n,k}(\theta)} \big(F_k(\theta, U_{n,k,\ell}) - F_{k-1}(\theta, U_{n,k,\ell}) \big) \tag{63}$$

as a multilevel approximation of $f(\theta)$ in the nth Robbins–Monro step, and we study the sequence of Robbins–Monro approximations $(\theta_n)_{n\in\mathbb{N}_0}$ given by

$$\theta_n = \theta_{n-1} + \gamma_n\, Z_n(\theta_{n-1}). \tag{64}$$

Note that the bias-levels ε_n and the noise-levels σ_n are used to determine the maximal levels $m_n(\theta)$, see (60), and the number of replications $N_{n,k}(\theta)$, see (61). We stress

that for the computation of $m_n(\theta)$ and $N_{n,k}(\theta)$ it is sufficient to know the functions Γ_1 and Γ_2 (or upper bounds of these functions) up to multiplicative constants only, since such constants can be compensated by the choice of the bias-levels ε_n and the noise-levels σ_n. See Sect. 5 for concrete examples. In Theorems 3.1 and 3.2 below the levels ε_n and σ_n are chosen to be of polynomial order and such that $\sigma_n/\varepsilon_n = n^{1/2}$ up to a multiplicative constant.

We measure the computational cost of θ_n by the quantity

$$\text{cost}_n = \mathbb{E}\left[\sum_{j=1}^{n}\sum_{k=1}^{m_j(\theta_{j-1})} N_{j,k}(\theta_{j-1})\, C_k\right]. \tag{65}$$

That means we take the mean computational cost for simulating the random vectors $F_k(\theta, U_{j,k,\ell}) - F_{k-1}(\theta, U_{j,k,\ell})$ for the first n iterations into account and we ignore the cost of the involved arithmetical operations. Note, however, that the number of arithmetical operations needed to compute θ_n is essentially proportional to $\sum_{j=1}^{n}\sum_{k=1}^{m_j(\theta_{j-1})} N_{j,k}(\theta_{j-1})$, and the average of the latter quantity is captured by cost_n under the weak assumption that $\inf_k C_k > 0$.

Note further that the quantity cost_n depends on the parameters $\Gamma_1, \Gamma_2, M, \alpha, \beta, \theta_0$, $(\gamma_k)_{k=1,\dots,n}$, $(\varepsilon_k)_{k=1,\dots,n}$ and $(\sigma_k)_{k=1,\dots,n}$, which determine the algorithm $(\theta_k)_{k\in\mathbb{N}}$ up to time n. For ease of notation we do not explicitly indicate this dependence in the notation cost_n.

To obtain upper bounds of cost_n we need an additional assumption on the functions Γ_1, Γ_2, which implies that both the variance estimate in C.1(i) and the bias estimate in C.1(ii) are at most of polynomial growth in $\theta \in \mathbb{R}^d$ with exponents related to the parameters α, β and p.

C.2 (Assumption on Γ_1, Γ_2)

With $\alpha, \beta, \Gamma_1, \Gamma_2$ according to Assumption C.1 there exists $K_1 \in (0, \infty)$ and

$$\beta_1 \in \begin{cases} [0, \min(\beta, 1/2)], & \text{if } \beta \neq 1/2, \\ [0, 1/2), & \text{if } \beta = 1/2, \end{cases}$$

such that for all $\theta \in \mathbb{R}^d$

$$\Gamma_1(\theta) \leq K_1\,(1 + \|\theta\|)^{\beta_1 p} \text{ and } \Gamma_2(\theta) \leq K_1\,(1 + \|\theta\|)^{\alpha p}. \tag{66}$$

We are now in the position to state the central complexity theorem on the multilevel Robbins–Monro algorithm.

Theorem 3.1 (Multilevel Robbins–Monro approximation) *Suppose that Assumption A.1 is satisfied for the function f given by (59) and that Assumptions C.1 and C.2 are satisfied. Take $L \in (0, \infty)$ according to A.1, take $\Gamma_1, \Gamma_2, M, \alpha, \beta$ according to C.1 and C.2 and let $\theta_0 \in \mathbb{R}^d$.*

Take $r \in (-1, \infty)$, $\gamma_0 \in (\frac{1+r}{2L}, \infty)$, $\sigma_0, \varepsilon_0 \in (0, \infty)$, *and let* $\rho = \frac{1}{2}(1+r)$ *and for all* $n \in \mathbb{N}$,

$$\gamma_n = \gamma_0 \frac{1}{n}, \quad \sigma_n^2 = \sigma_0^2 \frac{1}{n^r}, \quad \varepsilon_n = \varepsilon_0 \frac{1}{n^\rho}.$$

Then for all $\eta \in (0, \rho)$ *there exists* $\kappa \in (0, \infty)$ *such that for all* $n \in \mathbb{N}$,

$$\mathbb{E}\big[\sup_{k \geq n} k^{\eta p} \, \|\theta_k - \theta^*\|^p\big]^{1/p} \leq \kappa \, n^{-(\rho - \eta)}.$$

In particular, for all $\delta \in (0, \infty)$ *we have* $\lim_{n \to \infty} n^{\rho - \delta} \, \|\theta_n - \theta^*\| = 0$ *almost surely.*

If additionally $\alpha > \beta \wedge 1/2$ *and* $r > \frac{\beta \wedge 1/2}{\alpha - \beta \wedge 1/2}$ *then there exists* $\kappa' \in (0, \infty)$ *such that for all* $n \in \mathbb{N}$,

$$\mathrm{cost}_n \leq \kappa' \begin{cases} n^{2\rho}, & \text{if } \beta > 1/2, \\ n^{2\rho} \, (\ln(n+1))^2, & \text{if } \beta = 1/2, \\ n^{2\rho \left(1 + \frac{1-2\beta}{2\alpha}\right)}, & \text{if } \beta < 1/2. \end{cases}$$

The implementation of the multilevel Robbins–Monro approximation from Theorem 3.1 requires the knowledge of a positive lower bound for the parameter L from Assumption A.1. This difficulty is overcome by applying the Polyak–Ruppert averaging methodology. That means we consider the approximations

$$\bar{\theta}_n = \frac{1}{\bar{b}_n} \sum_{k=1}^{n} b_k \, \theta_k, \tag{67}$$

were $(\theta_n)_{n \in \mathbb{N}}$ is the multilevel Robbins–Monro scheme specified by (64), $(b_k)_{k \in \mathbb{N}}$ is a sequence of positive reals and

$$\bar{b}_n = \sum_{k=1}^{n} b_k$$

for $n \in \mathbb{N}$, see Sect. 2.

Note that the cost to compute $\bar{\theta}_n$ differs from the cost to compute θ_n at most by a deterministic factor, which does not depend on n. Therefore we again measure the computational cost for the computation of $\bar{\theta}_n$ by the quantity cost_n given by (65).

We state the second complexity theorem, which concerns Polyak–Ruppert averaging.

Theorem 3.2 (Multilevel Polyak–Ruppert approximation) *Suppose that Assumption B.1 is satisfied for the function* f *given by* (59) *and that Assumptions C.1 and C.2 are satisfied. Take* $\lambda \in (0, \infty)$ *according to B.1, take* $\Gamma_1, \Gamma_2, M, \alpha, \beta$ *according to C.1 and C.2 and let* $\theta_0 \in \mathbb{R}^d$.

Let $q \in [\frac{p}{1+\lambda}, p)$. *Take* $\gamma_0, \sigma_0, \varepsilon_0, b_0 \in (0, \infty)$, $r_1 \in (0, 1)$, $r_2 \in (-r_1, \infty)$ *and* $r_3 \in [r_2/2, \infty)$ *with*

$$\frac{1 + r_2}{r_1 + r_2} \leq \frac{p}{q},$$

and let $\rho = \frac{1}{2}(1 + r_2)$ *and for all* $n \in \mathbb{N}$,

$$\gamma_n = \gamma_0 \frac{1}{n^{r_1}}, \quad \varepsilon_n = \varepsilon_0 \frac{1}{n^{\rho}}, \quad \sigma_n^2 = \sigma_0^2 \frac{1}{n^{r_2}}, \quad b_n = b_0\, n^{r_3}.$$

Then there exists $\kappa \in (0, \infty)$ *such that for all* $n \in \mathbb{N}$

$$\mathbb{E}\big[\|\bar{\theta}_n - \theta^*\|^q\big]^{1/q} \leq \kappa\, n^{-\rho}.$$

If additionally $\alpha > \beta \wedge 1/2$ *and* $r_2 \geq \frac{\beta \wedge 1/2}{\alpha - \beta \wedge 1/2}$ *then there exists* $\kappa' \in (0, \infty)$ *such that for all* $n \in \mathbb{N}$

$$\mathrm{cost}_n \leq \kappa' \begin{cases} n^{2\rho}, & \text{if } \beta > 1/2, \\ n^{2\rho}\,(\ln(n+1))^2, & \text{if } \beta = 1/2, \\ n^{2\rho\left(1 + \frac{1-2\beta}{2\alpha}\right)}, & \text{if } \beta < 1/2. \end{cases}$$

Remark 3.3 Assume the setting of Theorem 3.1 or 3.2 and let $e_n = \mathbb{E}[\|\theta_n - \theta^*\|^p]^{1/p}$ or $e_n = \mathbb{E}[\|\bar{\theta}_n - \theta^*\|^q]^{1/q}$, respectively. Then there exists $\kappa \in (0, \infty)$ such that for every $n \in \mathbb{N}$,

$$\mathrm{cost}_n \leq \kappa \begin{cases} e_n^{-2}, & \text{if } \beta > 1/2, \\ e_n^{-2}(\ln(1 + e_n^{-1}))^2, & \text{if } \beta = 1/2, \\ e_n^{-2 - \frac{1-2\beta}{2\alpha}}, & \text{if } \beta < 1/2. \end{cases} \tag{68}$$

Note that these bounds for the computational cost in terms of the error coincide with the respective bounds for the multilevel computation of a single expectation presented in [5].

Remark 3.4 The multilevel stochastic approximation algorithms analysed in Theorems 3.1 and 3.2 are based on evaluations of the increments $F_k - F_{k-1}$. Consider, more generally, a sequence of measurable mappings $P_k : \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}^d$, $k \in \mathbb{N}$, such that for all $k \in \mathbb{N}$,

$$\mathbb{E}[P_k(\theta, U)] = \mathbb{E}[F_k(\theta, U) - F_{k-1}(\theta, U)]$$

and C_k is a worst case cost bound for simulating $P_k(\theta, U)$. Then Theorems 3.1 and 3.2 are still valid for the algorithm obtained by using P_k as a substitute for the increment $F_k - F_{k-1}$ in (63) if Assumption C.1(i) is satisfied with P_k in place of $F_k - F_{k-1}$.

Remark 3.5 We compare the schemes analysed in Theorems 3.1 and 3.2 with stochastic approximation algorithms that apply classical Monte Carlo.

Assume that Assumptions A.1 and C.1 are satisfied and consider the Robbins–Monro scheme

$$\theta_n^{\mathrm{mc}} = \theta_{n-1}^{\mathrm{mc}} + \gamma_n \, Z_n^{\mathrm{mc}}(\theta_{n-1}^{\mathrm{mc}}),$$

with

$$Z_n^{\mathrm{mc}}(\theta) = \frac{1}{N_n(\theta)} \sum_{\ell=1}^{N_n(\theta)} F_{m_n(\theta)}(\theta, U_{n,\ell}),$$

where $(U_{n,\ell})_{n,\ell\in\mathbb{N}}$ is a sequence of independent copies of U, the level at the nth step $m_n(\theta)$ is given by (60) and

$$N_n(\theta) = \left\lceil \frac{\Gamma_1(\theta)^2}{\sigma_n^2} \right\rceil$$

is the number of replications at the nth step. Note that the parameter β from Assumption C.1(i) has no impact on the construction of θ_n^{mc}.

If the step-sizes γ_n, the bias-levels ε_n and the noise-levels σ_n are chosen as in Theorem 3.1 then the error estimates in Theorem 3.1 are still valid for θ_n^{mc}, in particular, there exists $\kappa \in (0, \infty)$ such that for all $n \in \mathbb{N}$,

$$\mathbb{E}\big[\|\theta_n^{\mathrm{mc}} - \theta^*\|^p\big]^{1/p} \leq \kappa \, n^{-\rho}.$$

Similarly, if additionally Assumptions B.1 and C.1 are satisfied and if the step-sizes γ_n, the bias-levels ε_n, the noise-levels σ_n and the averaging weights b_n are chosen as in Theorem 3.2 then the error estimates in Theorem 3.2 are still valid for the Polyak–Ruppert averaging

$$\bar{\theta}_n^{\mathrm{mc}} = \frac{1}{\bar{b}_n} \sum_{k=1}^{n} \theta_k^{\mathrm{mc}},$$

i.e. there exists $\kappa \in (0, \infty)$ such that for all $n \in \mathbb{N}$,

$$\mathbb{E}\big[\|\bar{\theta}_n^{\mathrm{mc}} - \theta^*\|^q\big]^{1/q} \leq \kappa \, n^{-\rho}.$$

As a measure of the computational cost of θ_n^{mc} and $\bar{\theta}_n^{\mathrm{mc}}$ we use the quantity

$$\mathrm{cost}_n^{\mathrm{mc}} = \mathbb{E}\left[\sum_{j=1}^{n} N_j(\theta_{j-1}^{\mathrm{mc}}) \, C_{m_j(\theta_{j-1}^{\mathrm{mc}})}\right]$$

in accordance with the definition of the computational cost in (65). For the analysis of $\mathrm{cost}_n^{\mathrm{mc}}$ we need, however, to employ a growth condition on the functions Γ_1 and Γ_2 which is stronger than Assumption C.2, namely, that there exist $K > 0$ and $\tilde{\alpha}, \tilde{\beta} \in (0, \infty)$ such that $2\tilde{\beta} + \tilde{\alpha}/\alpha \leq 1$ and for all $\theta \in \mathbb{R}^d$

$$\Gamma_1(\theta) \leq K(1 + \|\theta\|)^{\tilde{\beta} p} \text{ and } \Gamma_2(\theta) \leq K(1 + \|\theta\|)^{\tilde{\alpha} p}. \tag{69}$$

If Assumptions A.1, C.1 and (69) are satisfied then there exists $\kappa \in (0, \infty)$ such that for all $n \in \mathbb{N}$,

$$\mathrm{cost}_n^{\mathrm{mc}} \leq \kappa\, n^{\rho(2+1/\alpha)}.$$

Note that, formally, this cost bound coincides with the cost bounds in Theorems 3.1 and 3.2 for $\beta = 0$.

The proof of these facts can be carried out in the same way as the proof of Theorems 3.1 and 3.2. The resulting upper bound of the computational cost $\mathrm{cost}_n^{\mathrm{mc}}$ in terms of the error $e_n^{\mathrm{mc}} = \mathbb{E}[\|\theta_n^{\mathrm{mc}} - \theta^*\|^p]^{1/p}$ or $e_n^{\mathrm{mc}} = \mathbb{E}[\|\bar{\theta}_n^{\mathrm{mc}} - \theta^*\|^q]^{1/q}$ is given by

$$\mathrm{cost}_n^{\mathrm{mc}} \leq \kappa^2 \big(e_n^{\mathrm{mc}}\big)^{-2-1/\alpha},$$

which coincides with the well-known respective cost bound for the computation of a single expectation and is considerably larger than the cost bound (68) for the multilevel methods in Remark 3.3.

The proofs of Theorems 3.1 and 3.2 are based on the following proposition, which shows that under Assumptions C.1(i),(ii) the scheme (64) can be represented as a Robbins–Monro scheme of the general form (1) studied in Sect. 2. It further provides an estimate of the computational cost (65) based on Assumptions C.1(iii) and C.2 only.

Proposition 3.6 (i) *Suppose that Assumptions C.1(i),(ii) are satisfied. Let $\mathcal{F}_n$ denote the σ-field generated by the variables $U_{m,k,\ell}$ with $m, k, \ell \in \mathbb{N}$ and $m \leq n$, and let $\mathcal{F}_0$ denote the trivial σ-field. The scheme $(\theta_n)_{n\in\mathbb{N}_0}$ given by (64) satisfies*

$$\theta_n = \theta_{n-1} + \gamma_n\big(f(\theta_{n-1}) + \varepsilon_n\, R_n + \sigma_n\, D_n\big)$$

for every $n \in \mathbb{N}$, where $(R_n)_{n\in\mathbb{N}}$ is a previsible process with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ and $(D_n)_{n\in\mathbb{N}}$ is a sequence of martingale differences with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ and $(R_n)_{n\in\mathbb{N}}$ and $(D_n)_{n\in\mathbb{N}}$ satisfy Assumption A.2.

(ii) *Suppose that Assumptions C.1(iii) and C.2 are satisfied. Then there exists a constant $\kappa \in (0, \infty)$ such that for all $n \in \mathbb{N}$ the computational cost (65) of θ_n given by (64) satisfies*

$$\begin{aligned}\mathrm{cost}_n &\le \kappa \max_{k=1,\dots,n-1} \mathbb{E}[(1+\|\theta_k\|)^p] \\ &\begin{cases} \sum_{j=1}^n \big(\varepsilon_j^{-1/\alpha} + \sigma_j^{-2}\, \varepsilon_j^{-\frac{(1-2\beta)_+}{\alpha}}\big), & \text{if } \beta \neq 1/2, \\ \sum_{j=1}^n \big(\varepsilon_j^{-1/\alpha} + \sigma_j^{-2}\, (\log_M(1/\varepsilon_j))^2\big), & \text{if } \beta = 1/2. \end{cases}\end{aligned} \tag{70}$$

Proof We first prove statement (i) of the proposition. Put $f_n(\theta) = \mathbb{E}[F_n(\theta, U)]$ and let

$$P_{n,k,\ell}(\theta) = F_k(\theta, U_{n,k,\ell}) - F_{k-1}(\theta, U_{n,k,\ell})$$

for $n, k, \ell \in \mathbb{N}$ and $\theta \in \mathbb{R}^d$. By Assumptions C.1(i),(ii) we have

$$\mathbb{E}[\|P_{n,k,\ell}(\theta) - \mathbb{E}[P_{n,k,\ell}(\theta)]\|^p]^{1/p} \le \Gamma_1(\theta)\, M^{-k\beta} \quad \text{and} \quad \|f_k(\theta) - f(\theta)\| \le \Gamma_2(\theta)\, M^{-k\alpha} \tag{71}$$

for all $n, k, \ell \in \mathbb{N}$ and $\theta \in \mathbb{R}^d$.

By (71) and the definition (60) of $m_n(\theta)$ we get for all $n \in \mathbb{N}$ and $\theta \in \mathbb{R}^d$ that

$$\|\mathbb{E}[Z_n(\theta)] - f(\theta)\| = \|\mathbb{E}[F_{m_n(\theta)}(\theta, U)] - f(\theta)\| \le \Gamma_2(\theta)\, M^{-m_n(\theta)\alpha} \le \varepsilon_n. \tag{72}$$

Furthermore, by the Burkholder–Davis–Gundy inequality, the triangle inequality on the $L^{p/2}$-space, (71) and the definition (61) of $N_{n,k}(\theta)$ there exists $c_1 \in (0, \infty)$, which only depends on p, such that

$$\begin{aligned}&\mathbb{E}[\|Z_n(\theta) - \mathbb{E}[Z_n(\theta)]\|^p]^{2/p} \\ &= \mathbb{E}\left[\left\|\sum_{k=1}^{m_n(\theta)} \sum_{\ell=1}^{N_{n,k}(\theta)} \frac{1}{N_{n,k}(\theta)} \big(P_{n,k,\ell}(\theta) - \mathbb{E}[P_{n,k,\ell}(\theta)]\big)\right\|^p\right]^{2/p} \\ &\le c_1\, \mathbb{E}\left[\left(\sum_{k=1}^{m_n(\theta)} \sum_{\ell=1}^{N_{n,k}(\theta)} \frac{1}{N_{n,k}(\theta)^2} \big\|P_{n,k,\ell}(\theta) - \mathbb{E}[P_{n,k,\ell}(\theta)]\big\|^2\right)^{p/2}\right]^{2/p} \\ &\le c_1 \sum_{k=1}^{m_n(\theta)} \frac{1}{N_{n,k}(\theta)^2} \sum_{\ell=1}^{N_{n,k}(\theta)} \mathbb{E}\big[\big\|P_{n,k,\ell}(\theta) - \mathbb{E}[P_{n,k,\ell}(\theta)]\big\|^p\big]^{2/p} \\ &\le c_1\, \Gamma_1(\theta)^2 \sum_{k=1}^{m_n(\theta)} \frac{1}{N_{n,k}(\theta)}\, M^{-2\beta k} \le c_1\, \frac{\Gamma_1(\theta)^2}{\kappa_n(\theta)} \sum_{k=1}^{m_n(\theta)} M^{k(1/2-\beta)}.\end{aligned}$$

Recalling the definition of $\kappa_n(\theta)$, see (62), we conclude that there exists $c_2 \in (0, \infty)$ such that for all $n \in \mathbb{N}$ and $\theta \in \mathbb{R}^d$

$$\mathbb{E}[\|Z_n(\theta) - \mathbb{E}[Z_n(\theta)]\|^p]^{2/p} \le c_2\, \sigma_n^2. \tag{73}$$

With

$$R_n := \frac{1}{\varepsilon_n}\,\mathbb{E}[Z_n(\theta_{n-1}) - f(\theta_{n-1})|\mathcal{F}_{n-1}] \text{ and}$$
$$D_n := \frac{1}{\sigma_n}\left(Z_n(\theta_{n-1}) - f(\theta_{n-1}) - \mathbb{E}[Z_n(\theta_{n-1}) - f(\theta_{n-1})|\mathcal{F}_{n-1}]\right)$$

we obtain that $\theta_n = \theta_{n-1} + \gamma_n(f(\theta_{n-1}) + \varepsilon_n R_n + \sigma_n D_n)$. We verify Assumption A.2.

The process $(R_n)_{n\in\mathbb{N}}$ is predictable and using the independence of $(U_{n,k,\ell})_{k,\ell\in\mathbb{N}}$ and $\mathcal{F}_{n-1}$ we conclude with (72) that $\sup_{n\in\mathbb{N}} \|R_n\| \le 1$. By the latter independence it further follows that $(D_n)_{n\in\mathbb{N}}$ is a sequence of martingale differences, which satisfies $\sup_{n\in\mathbb{N}} \mathbb{E}[\|D_n\|^p] \le c_2^{p/2}$ as a consequence of (73). This completes the proof of statement (i).

We turn to the proof of statement (ii). Let $j \in \mathbb{N}$ and $\theta \in \mathbb{R}^d$. Using Assumption C.1(iii), we conclude that there exists $c \in (0,\infty)$, which only depends on K, M and β such that

$$\begin{aligned}\sum_{k=1}^{m_j(\theta)} N_{j,k}(\theta)\, C_k &\le \sum_{k=1}^{m_j(\theta)} \left(1 + \kappa_j(\theta)\, M^{-k(\beta+1/2)}\right) K M^k \\ &\le \frac{K}{1-M^{-1}} M^{m_j(\theta)} + K\kappa_j(\theta) \begin{cases} \frac{M^{m_j(\theta)(1/2-\beta)_+}}{1-M^{-|1/2-\beta|}}, & \text{if } \beta \neq 1/2, \\ m_j(\theta), & \text{if } \beta = 1/2 \end{cases} \\ &\le c\, M^{m_j(\theta)} + c\,\frac{\Gamma_1^2(\theta)}{\sigma_j^2} \begin{cases} M^{m_j(\theta)(1-2\beta)_+}, & \text{if } \beta \neq 1/2, \\ (m_j(\theta))^2, & \text{if } \beta = 1/2. \end{cases}\end{aligned} \tag{74}$$

Furthermore, (60) yields that

$$m_j(\theta) \le \alpha^{-1}(\log_M(\Gamma_2(\theta)) + \log_M(\varepsilon_j^{-1})) + 1 \text{ and } M^{m_j(\theta)} \le M\,\varepsilon_j^{-1/\alpha}(\Gamma_2(\theta))^{1/\alpha}. \tag{75}$$

Combining (74) with (75) and employing Assumption C.2 we see that there exists $c_1 \in (0,\infty)$, which only depends on K, K_1, M, β and α, such that

$$\begin{aligned}&\sum_{k=1}^{m_j(\theta)} N_{j,k}(\theta)\, C_k \le c_1 \varepsilon_j^{-1/\alpha}(1+\|\theta\|)^p \\ &\quad + c_1\sigma_j^{-2}(1+\|\theta\|)^{2\beta_1 p} \begin{cases} \varepsilon_j^{-(1-2\beta)_+/\alpha}(1+\|\theta\|)^{(1-2\beta)_+ p}, & \text{if } \beta \neq 1/2, \\ (\log_M(\varepsilon_j^{-1}(1+\|\theta\|)^{\alpha p}))^2, & \text{if } \beta = 1/2. \end{cases}\end{aligned} \tag{76}$$

Suppose that $\beta \neq 1/2$. Then (76) implies

$$\sum_{k=1}^{m_j(\theta)} N_{j,k}(\theta)\, C_k \le c_1(1+\|\theta\|)^p\left(\varepsilon_j^{-1/\alpha} + \sigma_j^{-2}\varepsilon_j^{-\frac{(1-2\beta)_+}{\alpha}}\right),$$

which finishes the proof for the case $\beta \neq 1/2$. In the case $\beta = 1/2$ we have $\beta_1 < 1/2$ and therefore the existence of a constant $c_2 \in (0, \infty)$, which does not depend on θ, such that

$$(\log_M(1 + \|\theta\|^{\alpha p}))^2 \leq c_2 \, (1 + \|\theta\|)^{p(1-2\beta_1)}.$$

One completes the proof of statement (ii) by combining the latter estimate with (76). □

We turn to the proof of Theorem 3.1.

Proof of Theorem 3.1 The error estimate follows by Corollary 2.9 since Assumption A.1 is part of the theorem and Assumptions (I)–(III) and A.2 are satisfied by Proposition 3.6(i).

It remains to prove the cost estimate. The error estimate implies that $\sup_{n\in\mathbb{N}} \mathbb{E}[(1 + \|\theta_n\|)^p] < \infty$. Employing Proposition 3.6(ii) we thus see that there exists $c_1 \in (0, \infty)$ such that for every $n \in \mathbb{N}$

$$\mathrm{cost}_n \leq c_1 \begin{cases} \sum_{j=1}^n \big(j^{\rho/\alpha} + j^{2\rho(1+(1/2-\beta)_+/\alpha)-1}\big), & \text{if } \beta \neq 1/2, \\ \sum_{j=1}^n \big(j^{\rho/\alpha} + \rho^2 \, j^{2\rho-1} (\log_M(j))^2\big), & \text{if } \beta = 1/2. \end{cases} \tag{77}$$

Hence there exists $c_2 \in (0, \infty)$ such that for every $n \in \mathbb{N}$

$$\mathrm{cost}_n \leq c_2 \begin{cases} n^{\rho/\alpha+1} + n^{2\rho(1+(1/2-\beta)_+/\alpha)}, & \text{if } \beta \neq 1/2, \\ n^{\rho/\alpha+1} + n^{2\rho} (\log_M(n))^2, & \text{if } \beta = 1/2. \end{cases} \tag{78}$$

If $\beta \geq 1/2$ then $\alpha > 1/2$ and $r > \frac{1}{2\alpha-1}$, which implies that $\rho/\alpha < r = 2\rho - 1$ and therefore

$$n^{\rho/\alpha+1} \leq n^{2\rho} = n^{2\rho(1+(1/2-\beta)_+/\alpha)}.$$

If $\beta < 1/2$ then $\alpha > \beta$ and $r > \frac{\beta}{\alpha-\beta}$, which implies that $2\rho\frac{\alpha-\beta}{\alpha} > 1$ and therefore

$$n^{\rho/\alpha+1} \leq n^{\rho/\alpha+2\rho\frac{\alpha-\beta}{\alpha}} = n^{2\rho(1+(1/2-\beta)_+/\alpha)}.$$

This completes the proof. □

We proceed with the proof of Theorem 3.2.

Proof of Theorem 3.2 The error estimate follows with Corollary 2.14 since Assumption B.1 is part of the theorem and Assumptions (I)–(III) and B.2 hold by Proposition 3.6(i). The cost estimate in the theorem is proved in the same way as the cost estimate in Theorem 3.1. One only observes that $\sup_{n\in\mathbb{N}} \mathbb{E}[(1 + \|\theta_n\|)^p] < \infty$ is valid since the assumptions in Corollary 2.14 are stronger than the assumptions in Corollary 2.6. □

4 General convex closed domains

In this section we extend the results of Sects. 2 and 3 to convex domains. In the following D denotes a convex and closed subset of $\mathbb{R}^d$ and $f\colon D \to \mathbb{R}^d$ is a function with a unique zero $\theta^* \in D$. We start with the Robbins–Monro scheme.

Let

$$\mathrm{pr}_D\colon \mathbb{R}^d \to D$$

denote the orthogonal projection on D with respect to the given inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^d$ and define the dynamical system $(\theta_n)_{n\in\mathbb{N}_0}$ by the recursion

$$\theta_n = \mathrm{pr}_D\big(\theta_{n-1} + \gamma_n(f(\theta_{n-1}) + \varepsilon_n R_n + \sigma_n D_n)\big) \tag{79}$$

in place of (1), where $\theta_0 \in D$ is a deterministic starting value in D. Then the following fact follows by a straightforward modification in the proofs of Proposition 2.3 and Theorem 2.8 using the contraction property of pr_D.

Extension 4.1 *Proposition 2.3, Theorem 2.5, Corollary 2.6, Theorem 2.8, Corollary 2.9 and the statement on the system (1) in Remark 2.7 remain valid for the system (79) in place of (1) if $\mathbb{R}^d$ is replaced by D in Assumption A.1.*

Analogously, we extend Theorem 3.1 in Sect. 3 on the multilevel Robbins–Monro approximation to the case where the mappings $F, F_1, F_2, \ldots$ are defined on $D \times \mathcal{U}$ with D being a closed and convex subset of $\mathbb{R}^d$ and

$$f\colon D \to \mathbb{R}^d, \quad \theta \mapsto \mathbb{E}[F(\theta, U)]$$

has a unique zero $\theta^* \in D$. In this case we proceed analogously to Extension 4.1 and employ the projected multilevel Robbins–Monro scheme

$$\theta_n = \mathrm{pr}_D\big(\theta_{n-1} + \gamma_n Z_n(\theta_{n-1})\big) \tag{80}$$

with $\theta_0 \in D$ and Z_n given by (63), in place of the multilevel scheme (64).

Note that if pr_D can be evaluated on $\mathbb{R}^d$ with constant cost then, up to a constant depending on D only, the computational cost of the projected approximation θ_n is still bounded by the quantity cost_n given by (105) since the computation of θ_n requires n evaluations of pr_D and $\mathrm{cost}_n \geq C_1 n$.

Employing Proposition 3.6 as well as Extension 4.1 one easily gets the following result.

Extension 4.2 *Theorem 3.1 remains valid for the scheme (80) in place of (64) if $\mathbb{R}^d$ is replaced by D in Assumptions A.1, C.1 and C.2.*

Next we consider the Polyak–Ruppert scheme. In this case we additionally suppose that D contains an open ball $B(\theta^*, \delta) = \{\theta \in \mathbb{R}^d\colon \|\theta - \theta^*\| < \delta\}$ around the unique zero $\theta^* \in D$ and we extend the function f on $\mathbb{R}^d$: for $c \in (0, \infty)$ define

$$f_c\colon \mathbb{R}^d \to \mathbb{R}^d, \quad x \mapsto -c(x - \mathrm{pr}_D(x)) + f(\mathrm{pr}_D(x)). \tag{81}$$

The following lemma shows that property B.1 is preserved for appropriately chosen $c > 0$.

Lemma 4.3 *Let $\delta > 0$ and suppose that $B(\theta^*, \delta) \subset D$ and that $f\colon D \to \mathbb{R}^d$ satisfies B.1(i) to B.1(iii) on D. Take L, L', L'', λ, H according to B.1, let $c \in (1/(2L'), \infty)$ and put*

$$r_c = 1 - \tfrac{L}{c}\big(2 - \tfrac{1}{cL'}\big) \in [0, 1).$$

Then f_c satisfies B.1(i) to B.1(iii) on $\mathbb{R}^d$ with

$$L_c = c\big(1 - \sqrt{r_c}\big), \quad L_c' = \frac{L_c}{\frac{2}{(L')^2} + 2c^2}, \quad L_c'' = \big(c + \tfrac{1}{L'}\big)\tfrac{1}{\delta^\lambda} + L'' \tag{82}$$

in place of L, L' and L'', respectively.

Proof Using (3) with $c_1 = L$, $c_2 = L'$ and $\gamma = 1/c$ it follows that $r_c \in [0, 1)$. By (3) and the contractivity of the projection pr_D we have for any $\theta \in \mathbb{R}^d$ that

$$\begin{aligned}
\langle \theta - \theta^*, f_c(\theta)\rangle &= \langle \theta - \theta^*, -c\,(\theta - \mathrm{pr}_D(\theta)) + f(\mathrm{pr}_D(\theta))\rangle \\
&= -c\,\|\theta - \theta^*\|^2 + \langle \theta - \theta^*, -c\,(\theta^* - \mathrm{pr}_D(\theta)) + f(\mathrm{pr}_D(\theta))\rangle \\
&\le -c\,\|\theta - \theta^*\|^2 + c\,\|\theta - \theta^*\|\,\|\mathrm{pr}_D(\theta) - \theta^* + \tfrac{1}{c} f(\mathrm{pr}_D(\theta))\| \\
&\le -c\,\|\theta - \theta^*\|^2 + c\,\sqrt{r_c}\,\|\theta - \theta^*\|\,\|\mathrm{pr}_D(\theta) - \theta^*\| \\
&\le -c\,(1 - \sqrt{r_c})\,\|\theta - \theta^*\|^2,
\end{aligned}$$

which shows that f_c satisfies B.1(i) on $\mathbb{R}^d$ with L_c in place of L.

Using the latter estimate, (2) with $c_2' = 1/L'$ and the Lipschitz continuity of pr_D we get for any $\theta \in \mathbb{R}^d$ that

$$\begin{aligned}
\|f_c(\theta)\|^2 + \tfrac{1}{L_c'}\langle \theta - \theta^*, f_c(\theta)\rangle &\le \|-c\,(\theta - \mathrm{pr}_D(\theta)) + f(\mathrm{pr}_D(\theta))\|^2 - \tfrac{L_c}{L_c'}\|\theta - \theta^*\|^2 \\
&\le 2c^2\|\theta - \mathrm{pr}_D(\theta)\|^2 + 2\|f(\mathrm{pr}_D(\theta))\|^2 - \tfrac{L_c}{L_c'}\|\theta - \theta^*\|^2 \\
&\le 2c^2\|\theta - \mathrm{pr}_D(\theta)\|^2 + \tfrac{2}{(L')^2}\|\mathrm{pr}_D(\theta) - \theta^*\|^2 - \tfrac{L_c}{L_c'}\|\theta - \theta^*\|^2 \\
&\le (2c^2 + \tfrac{2}{(L')^2} - \tfrac{L_c}{L_c'})\|\theta - \theta^*\|^2 = 0,
\end{aligned}$$

which shows that f_c satisfies B.1(ii) on $\mathbb{R}^d$ with L_c' in place of L'.

Finally, let $\theta \in \mathbb{R}^d \setminus D$, which implies that $\|\theta - \theta^*\| \ge \delta$. Using the latter fact and the projection property and the contractivity of pr_D we get

$$\begin{aligned}
\|f_c(\theta) - H(\theta - \theta^*)\| = \|&-c(\theta - \mathrm{pr}_D(\theta)) + f(\mathrm{pr}_D(\theta)) - H(\mathrm{pr}_D(\theta) - \theta^*) \\
&- H(\theta - \mathrm{pr}_D(\theta))\|
\end{aligned}$$

$$\begin{aligned}
&\le \|(c\, I_d + H)(\theta - \mathrm{pr}_D(\theta))\| + \|f(\mathrm{pr}_D(\theta)) \\
&\quad - H(\mathrm{pr}_D(\theta) - \theta^*)\| \\
&\le \|c\, I_d + H\|\|\theta - \mathrm{pr}_D(\theta)\| + L''\, \|\mathrm{pr}_D(\theta) - \theta^*\|^{1+\lambda} \\
&\le (c + \|H\|)\|\theta - \theta^*\| + L''\, \|\theta - \theta^*\|^{1+\lambda} \\
&\le \big((c + \|H\|)\tfrac{1}{\delta^\lambda} + L''\big)\|\theta - \theta^*\|^{1+\lambda}.
\end{aligned}$$

Observing that $\|H\| \le 1/L'$, see (41), completes the proof of the lemma.

Replacing f by f_c in (1) we obtain the dynamical system

$$\theta_{c,n} = \theta_{c,n-1} + \gamma_n\big(f_c(\theta_{c,n-1}) + \varepsilon_n\, R_n + \sigma_n\, D_n\big), \tag{83}$$

for $n \in \mathbb{N}$, where $\theta_{c,0} \in D$ is a deterministic starting value in D.

Employing Lemma 4.3 we immediately arrive at the following fact.

Extension 4.4 *Assume that $B(\theta^*, \delta) \subset D$ for some $\delta \in (0, \infty)$. Then Corollary 2.14 remains valid for the modified Polyak–Ruppert algorithm*

$$\bar\theta_{c,n} = \frac{1}{\bar b_n} \sum_{k=1}^{n} b_k \theta_{c,k}, \qquad n \in \mathbb{N}, \tag{84}$$

in place of the scheme (40), if $\mathbb{R}^d$ is replaced by D in Assumption B.1 and $c \in (1/(2L'), \infty)$ with L' according to B.1(ii).

Moreover, Theorem 2.12 remains valid for the scheme (84) as well if, additionally, Assumption B.3 is satisfied with L_c given by (82) in place of L.

Similar to Extension 4.4 we can extend Theorem 3.2 on the multilevel Polyak–Ruppert averaging. To this end we define for $c \in (0, \infty)$ extensions $F_c, F_{c,1}, F_{c,2}, \ldots : \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}^d$ of the mappings $F, F_1, F_2 \ldots : D \times \mathcal{U} \to \mathbb{R}^d$ by taking

$$G_c \colon \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}^d, \quad (\theta, u) \mapsto -c(\theta - \mathrm{pr}_D(\theta)) + G(\mathrm{pr}_D(\theta), u)$$

for $G \in \{F, F_1, F_2, \ldots\}$. Note that $\mathbb{E}[F_c(\theta, U)] = f_c(\theta)$ and $f(\theta^*) = f_c(\theta^*) = 0$ with f_c given by (81).

Clearly, if the mappings $F, F_1, F_2, \ldots$ satisfy C.1(i),(ii) on D then the mappings $F_c, F_{c,1}, F_{c,2}, \ldots$ satisfy C.1(i),(ii) on $\mathbb{R}^d$ with $\Gamma_1 \circ \mathrm{pr}_D$, $\Gamma_2 \circ \mathrm{pr}_D$ in place of Γ_1, Γ_2. Furthermore, if Γ_1, Γ_2 satisfy Assumption C.2 on D then $\Gamma_1 \circ \mathrm{pr}_D$, $\Gamma_2 \circ \mathrm{pr}_D$ satisfy Assumption C.2 on $\mathbb{R}^d$, since we have $\|\mathrm{pr}_D(\theta)\| \le \|\theta\| + \|\mathrm{pr}_D(0)\|$ for every $\theta \in \mathbb{R}^d$.

We thus take

$$\begin{aligned}
Z_{c,n}(\theta) &= \sum_{k=1}^{m_n(\mathrm{pr}_D(\theta))} \frac{1}{N_{n,k}(\mathrm{pr}_D(\theta))} \sum_{\ell=1}^{N_{n,k}(\mathrm{pr}_D(\theta))} \big(F_{c,k}(\theta, U_{n,k,\ell}) - F_{c,k-1}(\theta, U_{n,k,\ell})\big) \\
&= -c(\theta - \mathrm{pr}_D(\theta)) + Z_n(\mathrm{pr}_D(\theta)),
\end{aligned}$$

with m_n, $N_{n,k}$ and Z_n given by (60),(61) and (63), respectively, as a multilevel approximation of $f_c(\theta)$ in the nth Robbins–Monro step, and we use the multilevel scheme

$$\theta_{c,n} = \theta_{c,n-1} + \gamma_n \, Z_{c,n}(\theta_{c,n-1}) \tag{85}$$

for Polyak–Ruppert averaging.

Employing Lemma 4.3 we get the following result.

Extension 4.5 *Assume that $B(\theta^*, \delta) \subset D$ for some $\delta \in (0, \infty)$. Then Theorem 3.2 remains valid for the modified Polyak–Ruppert algorithm*

$$\bar{\theta}_{c,n} = \frac{1}{\bar{b}_n} \sum_{k=1}^{n} b_k \theta_{c,k}, \qquad n \in \mathbb{N}, \tag{86}$$

with $(\theta_{c,n})_{n\in\mathbb{N}}$ given by (85) in place of the scheme (40), if $\mathbb{R}^d$ is replaced by D in Assumptions B.1, C.1, C.2 and $c \in (1/(2L'), \infty)$ with L' according to B.1(ii).

5 Examples and numerical experiments

We illustrate the construction of our multilevel methods and the corresponding theoretical findings by two examples. First we consider the simple one-dimensional problem of computing the volatility in a Black-Scholes model based on the price of a European call. Second we study a regression problem in a 3-dimensional Black-Scholes model. In both cases the explicit solution is known and we provide log–log plots for the errors obtained in our numerical tests which are in line with our theoretical findings.

5.1 Computing a volatility

Fix $T, \mu, s_0, K \in (0, \infty)$ and let U denote a one-dimensional Brownian motion on $[0, T]$. For every $\theta \in (0, \infty)$ we use S^θ to denote the geometric Brownian motion on $[0, T]$ with initial value s_0, trend μ and volatility θ, i.e.

$$\begin{aligned} S_0^\theta &= s_0, \\ dS_t^\theta &= \mu S_t^\theta \, dt + \theta S_t^\theta \, dU_t, \quad t \in [0, T]. \end{aligned} \tag{87}$$

In a Black-Scholes model with fixed interest rate μ the fair price of a European call with maturity T, strike K and underlying geometric Brownian motion with volatility θ is given by

$$pr(\theta) = \mathbb{E}[C(\theta, U)],$$

where

$$C(\theta, U) = \exp(-\mu T)(S_T^\theta - K)_+,$$

and according to the Black-Scholes formula $pr(\theta)$ satisfies

$$pr(\theta) = s_0\,\Phi\Big(\tfrac{\ln(s_0/K)+(\mu+\theta^2/2)T}{\theta\sqrt{T}}\Big) - \exp(-\mu T)K\,\Phi\Big(\tfrac{\ln(s_0/K)+(\mu-\theta^2/2)T}{\theta\sqrt{T}}\Big),$$

where Φ denotes the standard normal distribution function.

Fix $\vartheta_1 < \vartheta_2$ as well as $\theta^* \in [\vartheta_1, \vartheta_2]$. Our computational goal is to approximate θ^* based on the knowledge of ϑ_1, ϑ_2 and the value of the price $pr(\theta^*)$.

Within the framework of Sects. 3 and 4 we take $p \in [2, \infty)$, $d = 1$, $D = [\vartheta_0, \vartheta_1]$ and

$$F(\theta, U) = pr(\theta^*) - C(\theta, U), \quad \theta \in D.$$

Moreover, we approximate $F(\theta, U)$ by employing equidistant Milstein schemes: for $M, k \in \mathbb{N}$ with $M \geq 2$ and $\theta \in D$ we define

$$F_{M,k}(\theta, U) = pr(\theta^*) - \exp(-\mu T)(\widehat{S}^{\theta}_{M^k,T} - K)_+,$$

where $\widehat{S}^{\theta}_{M^k,T}$ denotes the Milstein approximation of S^{θ}_T based on M^k equidistant steps, i.e.

$$\widehat{S}^{\theta}_{M^k,T} = s_0 \prod_{\ell=1}^{M^k} \Big(1 + \mu\tfrac{T}{M^k} + \theta\Delta_\ell U + \tfrac{\theta^2}{2}\big((\Delta_\ell U)^2 - \tfrac{T}{M^k}\big)\Big)$$

with $\Delta_\ell U = U(\ell T/M^k) - U((\ell-1)T/M^k)$.

We briefly check the validity of Assumptions B.1, C.1 and C.2. Clearly, the mapping $f = \mathbb{E}[F(\cdot, U)]\colon D \to \mathbb{R}$ satisfies

$$f(\theta) = pr(\theta^*) - pr(\theta), \quad \theta \in D.$$

Note that pr is two times differentiable with respect to θ on $(0, \infty)$ with

$$\begin{aligned} \tfrac{\partial pr}{\partial\theta}(\theta) &= s_0\sqrt{T}\varphi\Big(\tfrac{\ln(s_0/K)+(\mu+\theta^2/2)T}{\theta\sqrt{T}}\Big), \\ \tfrac{\partial^2 pr}{\partial\theta^2}(\theta) &= \tfrac{(\ln(s_0/K)+(\mu+\theta^2/2)T)(\ln(s_0/K)+(\mu-\theta^2/2)T)}{\theta^3 T}\,\tfrac{\partial pr}{\partial\theta}(\theta), \end{aligned} \tag{88}$$

where φ denotes the density of the standard normal distribution. Let $g(\theta) = \frac{\ln(s_0/K)+(\mu+\theta^2/2)T}{\theta\sqrt{T}}$ and put $u = 2(\ln(s_0/K) + \mu T)/T$ as well as

$$z^* = s_0\sqrt{T}\begin{cases} \min\big(\varphi(g(\vartheta_1)), \varphi(g(\vartheta_2))\big), & \text{if } u \notin (\vartheta_1^2, \vartheta_2^2), \\ \min\big(\varphi(\max(g(\vartheta_1), g(\vartheta_2))), \varphi(\sqrt{u})\big) & \text{if } u \in (\vartheta_1^2, \vartheta_2^2). \end{cases}$$

Using (88) it is then straightforward to verify that f satisfies Assumption B.1 on D with parameters

$$L = s_0\sqrt{T}\min_{\theta\in[\vartheta_1,\vartheta_2]}\varphi(g(\theta)) = z^* \tag{89}$$

and

$$L' = \frac{\sqrt{2\pi}}{s_0\sqrt{T}}, \quad H = -\frac{\partial\, pr}{\partial\theta}(\theta^*), \quad L'' = \max_{\theta\in[\vartheta_1,\vartheta_2]}\Big|\frac{\partial^2 pr}{\partial\theta^2}(\theta)\Big|, \quad \lambda = 1. \tag{90}$$

Furthermore, it is well known that there exists a constant $c(T, \mu, \vartheta_1, \vartheta_2, p) \in (0, \infty)$, which depends only on $T, \mu, \vartheta_1, \vartheta_2, p$, such that

$$\sup_{\theta\in D} \mathbb{E}[|\widehat{S}^{\theta}_{M^k,T} - S^{\theta}_T|^p]^{1/p} \le c(T, \mu, \vartheta_1, \vartheta_2, p) M^{-k}$$

Since $|F_{M,k}(\theta, U) - F(\theta, U)| \le |\widehat{S}^{\theta}_{M^k,T} - S^{\theta}_T|$ we conclude that Assumption C.1 is satisfied on D with parameters

$$\alpha = \beta = 1, \quad \Gamma_1 = \Gamma_2 = c(T, \mu, \vartheta_1, \vartheta_2, p)$$

for some constant $c(T, \mu, \vartheta_1, \vartheta_2, p) \in (1, \infty)$, which depends only on $T, \mu, \vartheta_1, \vartheta_2, p$. Consequently, Assumption C.2 is satisfied on D as well.

First, we consider the projected multilevel Robbins–Monro scheme (80) with step-size γ_n, noise-level σ_n and bias-level ε_n given by

$$\gamma_n = \frac{2}{Ln}, \quad \sigma_n = \frac{c(T, \mu, \vartheta_1, \vartheta_2, p)}{n^{3/2}}, \quad \varepsilon_n = \frac{c(T, \mu, \vartheta_1, \vartheta_2, p)}{n^2}.$$

Note that the constant $c(T, \mu, \vartheta_1, \vartheta_2, p)$ does not need to be known in order to implement the scheme. We have

$$\theta_n = \mathrm{proj}_{[\vartheta_1,\vartheta_2]}\big(\theta_{n-1} + \tfrac{2}{Ln} Z_n(\theta_{n-1})\big),$$

where $\theta_0 \in [\vartheta_1, \vartheta_2]$ and for all $\theta \in D$

$$Z_n(\theta) = \sum_{k=1}^{1\vee\lceil 2\log_M(n)\rceil} \frac{1}{\lceil n^3 M^{-3k/2}\rceil} \sum_{\ell=1}^{\lceil n^3 M^{-3k/2}\rceil} \big(F_{M,k}(\theta, U_{n,k,\ell}) - F_{M,k-1}(\theta, U_{n,k,\ell})\big) \tag{91}$$

with independent copies $U_{n,k,\ell}$ of U. Then by Extension 4.2 of Theorem 3.1, for every $p \in [2, \infty)$ there exists $\kappa \in (0, \infty)$ such that for every $n \in \mathbb{N}$,

$$\mathbb{E}[|\theta_n - \theta^*|^p]^{1/p} \le \kappa\, n^{-2}, \quad \mathrm{cost}_n \le \kappa\, n^4. \tag{92}$$

As a consequence, by the Borel–Cantelli Lemma we obtain that for every $\delta > 0$ we also have with probability one,

$$\sup_{n\in\mathbb{N}} n^{2-\delta}\, |\theta_n - \theta^*| < \infty. \tag{93}$$

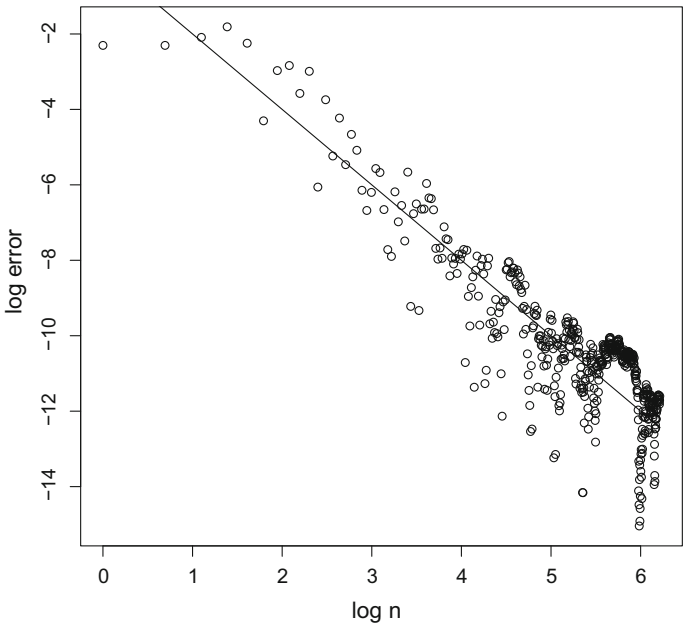

Fig. 1 Multilevel Robbins–Monro: error trajectory for $n = 1, \ldots, 500$

In the following we use the model parameters

$$s_0 = 10, T = 2, \quad \mu = 0.01, \quad K = 11, \quad \vartheta_1 = 0.05, \quad \vartheta_2 = 0.5, \quad \theta^* = 0.2 \tag{94}$$

and we choose

$$M = 4, \quad \theta_0 = 0.1 \tag{95}$$

in the definition of the Robbins–Monro scheme.

Figure 1 shows log–log plots of a simulation of the first 500 steps of the error process $(|\theta_n - \theta^*|)_{n \in \mathbb{N}_0}$ and the curve $x \mapsto x^{-2}$, which illustrates the pathwise behaviour of the error process and indicates the accordance with the theoretical result (93).

Figure 2 shows the log–log plot of Monte Carlo estimates of the root mean squared error of θ_n and the corresponding average computational times for $n = 1, \ldots, 100$ based on $N = 200$ replications. Additionally, the log–log plots of the curves $x \mapsto x^{-2}$ and $x \mapsto e^{-10} x^4$ are drawn to illustrate the accordance with the theoretical bounds in (92).

Next, we consider the multilevel Polyak–Rupert averaging (86) with step-size γ_n, noise-level σ_n, bias-level ε_n, weight b_n and extension parameter c given by

$$\gamma_n = \frac{1}{n^{0.9}}, \quad \sigma_n = \frac{c(T, \mu, \vartheta_1, \vartheta_2, M)}{n^{3/2}}, \quad \varepsilon_n = \frac{c(T, \mu, \vartheta_1, \vartheta_2, M)}{n^2}, \quad b_n = n^2, \quad c = \frac{1}{L'}.$$

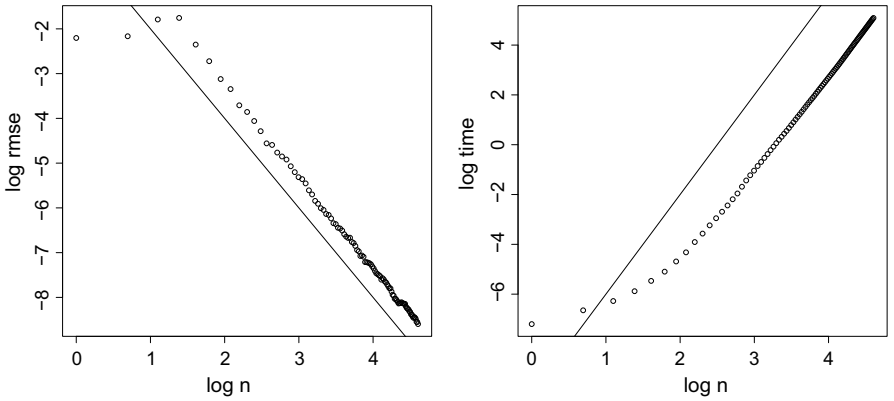

Fig. 2 Multilevel Robbins–Monro: estimated root mean squared error and average computational time for $n = 1, \dots, 100$

Thus

$$\bar{\theta}_{c,n} = \frac{6}{n(n+1)(2n+1)} \sum_{k=1}^{n} k^2 \, \theta_{k,c},$$

where

$$\theta_{n,c} = \theta_{n-1,c} + \frac{1}{n^{0.9}} \Big(-\frac{1}{L'} (\theta_{c,n-1} - \mathrm{proj}_{[\vartheta_1,\vartheta_2]}(\theta_{c,n-1})) + Z_n(\mathrm{proj}_{[\vartheta_1,\vartheta_2]}(\theta_{c,n-1})) \Big),$$

with Z_n given by (91) and a deterministic $\theta_{0,c} \in [\vartheta_1, \vartheta_2]$. By applying Extension 4.5 of Theorem 3.2 with $q = \frac{0.9+3}{1+3} \, p = 0.975 \, p \in [p/2, p)$ we see that for every $p \in [2, \infty)$ there exists a constant $\kappa \in (0, \infty)$ such that for every $n \in \mathbb{N}$,

$$\mathbb{E}[\|\bar{\theta}_{c,n} - \theta^*|^p]^{1/p} \le \kappa n^{-2}, \quad \mathrm{cost}_n \le \kappa n^4. \tag{96}$$

Moreover, for every $\delta > 0$ we have with probability one,

$$\sup_{n \in \mathbb{N}} n^{2-\delta} \, |\bar{\theta}_{c,n} - \theta^*| < \infty. \tag{97}$$

We choose the parameters $s_0, T, \mu, K, \vartheta_1, \vartheta_2, M, \theta_{c,0} = \theta_0$ as in (94) and (95). Figure 3 shows the log–log plot of a trajectory of the error process $(|\bar{\theta}_{c,n} - \theta^*|)_{n \in \mathbb{N}_0}$ until $n = 500$ together with the curve $x \mapsto x^{-2}$ to illustrate the theorectical bound from (97). A comparison with Fig. 1 shows the smoothing effect of the Polyak–Ruppert averaging.

Figure 4 shows the log–log plot of Monte Carlo estimates of the root mean squared error of $\bar{\theta}_{c,n}$ and the corresponding average computational times for $n = 1, \dots, 100$ based on $N = 200$ replications. Additionally, the log–log plots of the curves $x \mapsto x^{-2}$ and $x \mapsto e^{-10} \, x^4$ are drawn to illustrate the accordance with the theoretical bounds in (96).

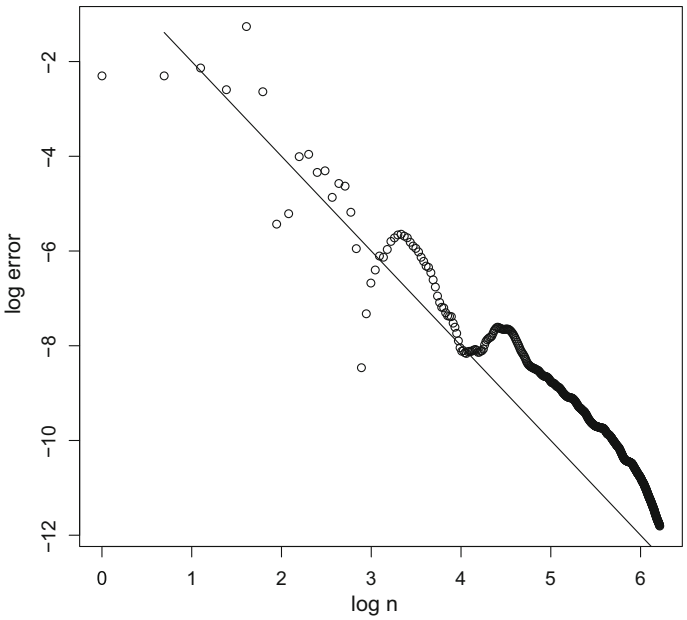

Fig. 3 Multilevel Polyak–Ruppert: error trajectory for $n = 1, \ldots, 500$

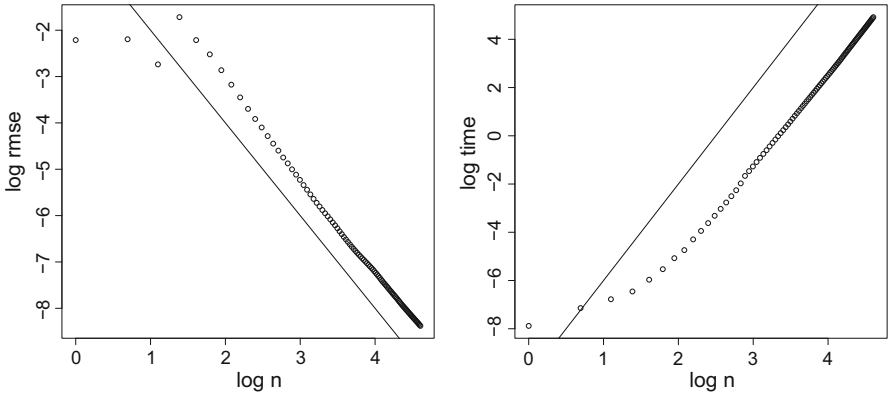

Fig. 4 Multilevel Polyak Ruppert: estimated root mean squared error and average computational time for $n = 1, \ldots, 100$

5.2 A regression problem

Fix $\mu \in \mathbb{R}^d$, $V \in \mathbb{R}^{d\times d}$ with $\operatorname{rank}(V) = d$, $y_0 \in (0, \infty)^d$ and let U be a d-dimensional Brownian motion on $[0, 1]$. We consider a d-dimensional geometric Brownian motion Y given by the SDE

$$\begin{aligned} dY_t &= a(Y_t)\,dt + b(Y_t)\,dU_t, \quad t \in [0, 1], \\ Y_0 &= y_0 \in \mathbb{R}^d, \end{aligned} \tag{98}$$

with drift coefficient

$$a\colon \mathbb{R}^d \to \mathbb{R}^d, \quad y \mapsto \operatorname{diag}(y)\,\mu$$

and diffusion coefficient

$$b\colon \mathbb{R}^d \to \mathbb{R}^{d\times d}, \quad y \mapsto \operatorname{diag}(y)\,V,$$

where $\operatorname{diag}(y_1,\dots,y_d)$ denotes the diagonal matrix in $\mathbb{R}^{d\times d}$ with entries $y_1,\dots,y_d$. Note that (98) has a unique strong solution $Y=(Y^{(1)},\dots,Y^{(d)})$, which is explicitly given by

$$Y_t^{(i)} = y_{0,i}\, e^{(\mu_i - \sum_{k=1}^d V_{i,k}^2/2)t + \sum_{k=1}^d V_{i,k}\,U_{t,k}} \tag{99}$$

for all $t\in[0,1]$ and $i=1,\dots,d$, and Y satisfies

$$\mathbb{E}[\sup_{t\in[0,1]} \|Y_t\|^q] < \infty \tag{100}$$

for every $q\in(0,\infty)$. Moreover, $b(y_0)(b(y_0))^\top = \operatorname{diag}(y_0)\,V\,V^\top\,\operatorname{diag}(y_0)$ is positive definite, which implies that

$$B := \mathbb{E}[Y_1 Y_1^\top] \text{ is positive definite,} \tag{101}$$

see, e.g. [13, Thm. 2.3.1].

We consider the regression problem from Example 2.2 with $X=Y_1$ and $g\colon \mathbb{R}^d \to \mathbb{R}$ being differentiable with a derivative that satisfies a polynomial growth condition. We thus want to compute the unique zero

$$\theta^* = B^{-1}\mathbb{E}[g(Y_1)Y_1]$$

of the function $f(\theta)=\mathbb{E}[g(Y_1)Y_1]-B\theta$, which solves the regression problem

$$\mathbb{E}[|\langle\theta^*, Y_1\rangle - g(Y_1)|^2] = \min_{\theta\in\mathbb{R}^d} \mathbb{E}[|\langle\theta, Y_1\rangle - g(Y_1)|^2].$$

Recall from Examples 2.2 and 2.11 that f satisfies Assumption B.1.

In the framework of Sect. 3 we have $f(\theta)=\mathbb{E}[F(\theta,U)]$ with

$$F(\theta,U) = g(Y_1)Y_1 - Y_1 Y_1^\top \theta$$

for $\theta\in\mathbb{R}^d$. We approximate $F(\theta,U)$ by employing equidistant Milstein schemes. Fix $N, M\in\mathbb{N}$ with $M\geq 2$. For $k\in\mathbb{N}$ and $\theta\in\mathbb{R}^d$ we define

$$F_k(\theta,U) = g(\widehat{Y}_{NM^k,1})\widehat{Y}_{NM^k,1} - \widehat{Y}_{NM^k,1}\widehat{Y}_{NM^k,1}^\top\theta,$$

where $\widehat{Y}_{NM^k,1}$ denotes the Milstein approximation of Y_1 based on NM^k equidistant steps, i.e.

$$\widehat{Y}^{(i)}_{NM^k,1} = y_{0,i} \prod_{\ell=1}^{NM^k} \Big(1 + \mu_i \frac{1}{NM^k} + \sum_{j=1}^{d} V_{i,j}\, \Delta_\ell U^{(j)} + \frac{1}{2}\Big(\Big(\sum_{j=1}^{d} V_{i,j}\, \Delta_\ell U^{(j)}\Big)^2 - \frac{1}{NM^k} \sum_{j=1}^{d} V_{i,j}^2\Big)\Big)$$

with $\Delta_\ell U^{(j)} = U^{(j)}(\ell/(NM^k)) - U^{(j)}((\ell-1)/(NM^k))$.

We next show that Assumptions $C.1$ and $C.2$ are satisfied. Note that the coefficients a and b have bounded partial derivatives of any order. Hence for every $q \in (0,\infty)$ there exists $c \in (0,\infty)$ such that for every $k \in \mathbb{N}$,

$$\mathbb{E}[\|Y_1 - \widehat{Y}_{NM^k,1}\|^q]^{1/q} \le c/M^k.$$

Using the latter estimate, the smoothness properties of g and (75) it is easy to see that for every $p \in [2,\infty)$ there exists $c \in (0,\infty)$ such that for every $k \in \mathbb{N}$ and $\theta \in \mathbb{R}^d$ we have

$$\mathbb{E}[\|F_k(\theta,U) - F(\theta,U)\|^p]^{1/p} \le c(1+\|\theta\|)M^{-k}. \tag{102}$$

It follows that Assumptions C.1(i) and C.1(ii) are satisfied with

$$\alpha = \beta = 1 \text{ and } \Gamma_1(\theta) = \Gamma_2(\theta) = c^*\,(1+\|\theta\|)$$

for some constant $c^* \in (0,\infty)$. Clearly, the computational cost to simulate $F_k(\theta,U)$ is proportional to the number of steps NM^k of the corresponding Milstein approximation $\widehat{Y}_{NM^k,1}$. Hence, Assumption C.1(iii) is satisfied as well. Finally, Assumption C.2 is obviously satisfied with $\beta_1 = 1/2$ and $K_1 = c^*$.

We consider the multilevel Polyak–Ruppert averaging (67) with step-size γ_n, noise-level σ_n, bias-level ε_n and weight b_n chosen according to Theorem 3.2. To this end we take $r_1 \in (0,1)$, $r_2 \in (-r_1,\infty)$ and $r_3 \in [r_2/2,\infty)$ and we use

$$\gamma_n = \frac{1}{n^{r_1}}, \quad \sigma_n = \frac{c^*}{n^{r_2/2}}, \quad \varepsilon_n = \frac{c^*}{n^{(1+r_2)/2}}, \quad b_n = n^{r_3}$$

for every $n \in \mathbb{N}$. Note that the constant c^* does not have to be known in order to implement the scheme. The resulting maximal levels and repetition numbers are

$$m_n(\theta) = 1 \vee \big\lceil \log_M\big(n^{(1+r_2)/2}(1+\|\theta\|)\big)\big\rceil, \quad N_{n,k}(\theta) = \big\lceil (1+\|\theta\|)^2\, n^{r_2}\, M^{-3k/2}\big\rceil \tag{103}$$

for $n \in \mathbb{N}$ and $k = 1,\dots,m_n(\theta)$, and we obtain the Polyak–Ruppert scheme

$$\bar{\theta}_n = \Big(\sum_{k=1}^{n} k^{r_3}\Big)^{-1} \sum_{k=1}^{n} k^{r_3}\theta_k,$$

with

$$\theta_n = \theta_{n-1} + \frac{1}{n^{r_1}} Z_n(\theta_{n-1}),$$

where $\theta_0 \in \mathbb{R}^d$ and $Z_n(\theta)$ is given by (63) for every $\theta \in \mathbb{R}^d$.

Note that the assumptions in Theorem 3.2 are satisfied for any $p \in [2, \infty)$ and any $\lambda \in (0, \infty)$. We may therefore choose $q = \frac{r_1+r_2}{1+r_2}\, p$ in Theorem 3.2 with p arbitrarily large. We thus obtain that for every $q \in (0, \infty)$ there exists a constant $\kappa > 0$ such that for every $n \in \mathbb{N}$,

$$\mathbb{E}\big[\|\bar{\theta}_n - \theta^*\|^q\big]^{1/q} \le \kappa\, n^{-(1+r_2)/2} \quad \text{and} \quad \mathrm{cost}_n \le \kappa\, n^{1+r_2}. \tag{104}$$

In the following we use the model parameters

$$d = 3, \quad y_0 = (1, 1, 1), \quad \mu = (0.1, 0.2, 0), \quad V = \begin{pmatrix} 1 & 0 & -0.2 \\ 0 & -0.7 & 0.1 \\ 0 & 0 & 0.6 \end{pmatrix}$$

and we take

$$g(y) = 0.2 \sum_{i=1}^{3} y_i^2.$$

Put $\Sigma = V V^\top$. Using (99) one obtains

$$\mathbb{E}[g(Y_1) Y_1^{(i)})] = 0.2 \sum_{j=1}^{3} \mathbb{E}\big[(Y_1^{(j)})^2 Y_1^{(i)}\big] = y_{0,i}\, e^{\mu_i} \sum_{j=1}^{3} y_{0,j}^2\, e^{2\mu_j + \Sigma_{j,j} + 2\Sigma_{i,j}}$$

for $i = 1, 2, 3$ and

$$B = \big(y_{0,i}\, y_{0,j}\, e^{\mu_i + \mu_j + \Sigma_{i,j}}\big)_{1 \le i,j \le d},$$

which yields

$$\theta^* = (1.8433086\ldots, 0.1999581\ldots, -0.2275381\ldots).$$

We take $N = 100$ and $M = 2$ as the step-size parameters of the Milstein scheme and choose the parameters r_1, r_2, r_3 of the Polyak–Ruppert scheme as

$$r_1 = 0.9, \quad r_2 = 2, \quad r_3 = 2,$$

similar to the preceding example on computing a volatility. According to (104) we then have

$$\mathbb{E}\big[\|\bar{\theta}_n - \theta^*\|^2\big]^{1/2} \le \kappa\, n^{-3/2} \text{ and } \mathrm{cost}_n \le \kappa\, n^3. \tag{105}$$

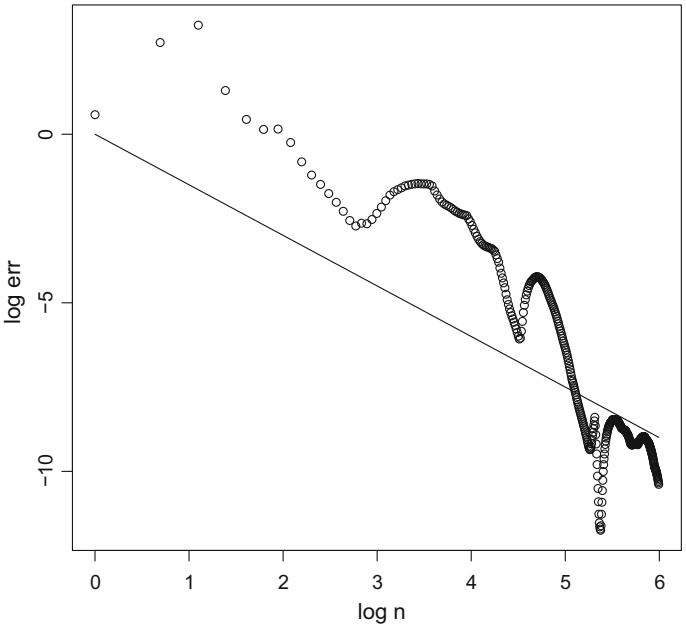

Fig. 5 Multilevel Polyak Ruppert: error trajectory for $n = 1, \ldots, 400$

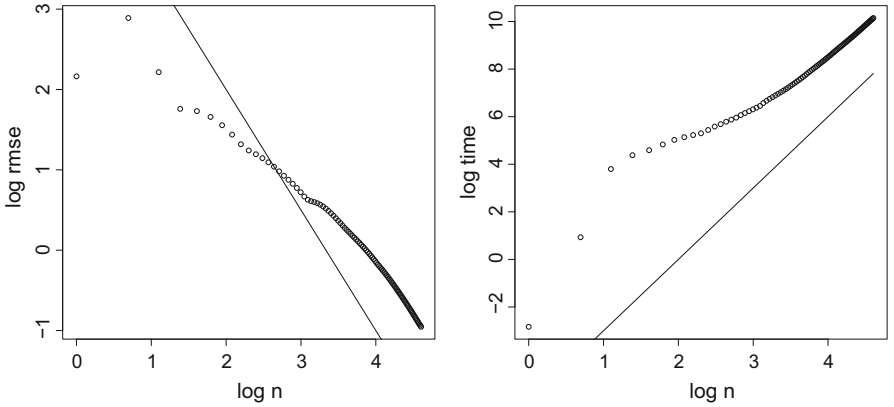

Fig. 6 Multilevel Polyak Ruppert: estimated root mean squared error and average computational time for $n = 1, \ldots, 100$

and for every $\delta > 0$ with probability one the pathwise upper bound $n^{-(3/2-\delta)}$, up to a multiplicative constant, for the error $\|\bar{\theta}_n - \theta^*\|$.

Figure 5 shows the log–log-plot of a trajectory of the error process $(\|\bar{\theta}_n - \theta^*\|)_{n\in\mathbb{N}}$ until $n = 400$. Additionally, the log–log plot of the curve $x \mapsto x^{-3/2}$ is drawn to illustrate the theoretical pathwise bound.

Figure 6 shows the log–log plot of Monte Carlo estimates of the root mean squared error of $\bar{\theta}_n$ and the corresponding average computational times cost_n for

$n = 1, \dots, 100$ based on $N = 200$ replications. Additionally, the log–log plots of the curves $x \mapsto e^5\, x^{-3/2}$ and $x \mapsto e^{-6}\, x^3$ are drawn to illustrate the accordance with the theoretical bounds in (105).

Acknowledgements We thank two anonymous referees for their valuable comments, which improved the presentation of the material.

Appendix

Let $(\Omega, \mathcal{F}, P)$ be a probability space endowed with a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ and let $\|\cdot\|$ denote a Hilbert space norm on $\mathbb{R}^d$.

In this section we provide pth mean estimates for an adapted d-dimensional dynamical system $(\zeta_n)_{n\in\mathbb{N}_0}$ with the property that for each $n \in \mathbb{N}$, ζ_n is a zero-mean perturbation of a previsible proposal ξ_n being comparable in size to ζ_{n-1}. More formally, we assume that there exist a previsible d-dimensional process $(\xi_n)_{n\in\mathbb{N}}$, a d-dimensional martingale $(M_n)_{n\in\mathbb{N}_0}$ with $M_0 = \zeta_0$ and a constant $c \geq 0$ such that for all $n \in \mathbb{N}$

$$\begin{aligned} \zeta_n &= \xi_n + \Delta M_n, \\ \|\xi_n\| &\leq \|\zeta_{n-1}\| \vee c, \end{aligned} \tag{106}$$

where $\Delta M_n = M_n - M_{n-1}$. Note that necessarily $\xi_n = \mathbb{E}[\zeta_n|\mathcal{F}_{n-1}]$.

Theorem 5.1 *Assume that $(\zeta_n)_{n\in\mathbb{N}_0}$ is an adapted d-dimensional process, which satisfies (106), and let $p \in [1, \infty)$. Then there exists a constant $\kappa \in (0, \infty)$, which only depends on p, such that for every $n \in \mathbb{N}_0$,*

$$\mathbb{E}\left[\max_{0\leq k\leq n} \|\zeta_k\|^p\right] \leq \kappa \left(\mathbb{E}\big[[M]_n^{p/2}\big] + c^p\right),$$

where

$$[M]_n = \sum_{k=1}^{n} \|\Delta M_k\|^2 + \|M_0\|^2.$$

Proof Fix $p \in [1, \infty)$.

We first consider the case where $c = 0$. Recall that by the Burkholder-Davis-Gundy inequality there exists a constant $\bar\kappa > 0$ depending only on p such that for every d-dimensional martingale $(M_n)_{n\in\mathbb{N}_0}$,

$$\mathbb{E}\big[\max_{0\leq k\leq n} \|M_k\|^p\big] \leq \bar\kappa\, \mathbb{E}\big[[M]_n^{p/2}\big].$$

We fix a time horizon $T \in \mathbb{N}_0$ and prove the statement of the theorem with $\kappa = \bar\kappa$ by induction: we say that the statement holds up to time $t \in \{0, \dots, T\}$, if for every d-dimensional adapted process $(\zeta_n)_{n\in\mathbb{N}_0}$, for every d-dimensional previsible process $(\xi_n)_{n\in\mathbb{N}}$ and for every d-dimensional martingale $(M_n)_{n\in\mathbb{N}_0}$ with

$$\begin{cases} \zeta_0 = M_0, & \\ \|\xi_n\| \le \|\zeta_{n-1}\|, & \text{if } 1 \le n \le t, \\ \zeta_n = \xi_n + \Delta M_n, & \text{if } 1 \le n \le t, \\ \zeta_n = \zeta_{n-1} + \Delta M_n, & \text{if } n > t, \end{cases} \tag{C_t}$$

one has

$$\mathbb{E}\left[\max_{0\le n\le T} \|\zeta_n\|^p\right] \le \bar{\kappa}\, \mathbb{E}\big[[M]_T^{p/2}\big].$$

Clearly, the statement is satisfied up to time 0 as a consequence of the Burkholder–Davis–Gundy inequality. Next, suppose that the statement is satisfied up to time $t \in \{0, \dots, T-1\}$. Let $(\zeta_n)_{n\in\mathbb{N}_0}$ be a d-dimensional adapted process, $(\xi_n)_{n\in\mathbb{N}}$ be a d-dimensional previsible process and $(M_n)_{n\in\mathbb{N}_0}$ be a d-dimensional martingale satisfying property (C_{t+1}). Consider any $\mathcal{F}_t$-measurable random orthonormal transformation U on $(\mathbb{R}^d, \|\cdot\|)$ and put

$$\zeta_n^U = \begin{cases} \zeta_n, & \text{if } n \le t, \\ \zeta_t + U(M_n - M_t), & \text{if } n > t \end{cases}$$

as well as

$$M_n^U = \begin{cases} M_n, & \text{if } n \le t, \\ M_t + U(M_n - M_t), & \text{if } n > t. \end{cases}$$

Then it is easy to check that $(M_n^U)_{n\in\mathbb{N}_0}$ is a martingale with $[M^U]_n = [M]_n$ for all $n \in \mathbb{N}$. Furthermore, $(\zeta_n^U)_{n\in\mathbb{N}_0}$ is adapted and the triple (ζ^U, ξ, M^U) satisfies property (C_t). Hence, by the induction hypothesis,

$$\mathbb{E}\big[\max_{0\le n\le T} \|\zeta_n^U\|^p\big] \le \bar{\kappa}\, \mathbb{E}\big[[M^U]_T^{p/2}\big] = \bar{\kappa}\, \mathbb{E}\big[[M]_T^{p/2}\big]. \tag{107}$$

Note that for any such random orthonormal transformation U, the norm of the random variable ζ_n^U is the same as the norm of the variable $\bar{\zeta}_n^U$ given by

$$\bar{\zeta}_n^U = \begin{cases} \zeta_n, & \text{if } n \le t, \\ U^*\zeta_t + M_n - M_t, & \text{if } n > t, \end{cases}$$

whence

$$\mathbb{E}\left[\max_{0\le n\le T} \|\bar{\zeta}_n^U\|^p\right] = \mathbb{E}\left[\max_{0\le n\le T} \|\zeta_n^U\|^p\right]. \tag{108}$$

Clearly, we can choose an $\mathcal{F}_t$-measurable random orthonormal transformation U on $(\mathbb{R}^d, \|\cdot\|)$ such that

$$U^*\zeta_t = \frac{\|\zeta_t\|}{\|\xi_{t+1}\|}\xi_{t+1}$$

holds on $\{\xi_{t+1} \neq 0\}$. Let

$$\alpha = \frac{\|\xi_{t+1}\| + \|\zeta_t\|}{2\|\zeta_t\|} \cdot 1_{\{\zeta_t \neq 0\}}.$$

Then α is $\mathcal{F}_t$-measurable and takes values in $[0, 1]$ since $\|\xi_{t+1}\| \leq \|\zeta_t\|$. Moreover, we have $\xi_{t+1} = \alpha U^* \zeta_t + (1-\alpha)(-U)^* \zeta_t$ so that by property (C_{t+1}) of the triple (ζ, ξ, M),

$$\zeta_n = \xi_{t+1} + M_n - M_t = \alpha \bar{\zeta}_n^U + (1-\alpha)\bar{\zeta}_n^{-U}$$

for $n = t+1, \ldots, T$. Note that $\zeta_n = \zeta_n^U = \zeta_n^{-U}$ for $n = 0, \ldots, t$. By convexity of $\|\cdot\|^p$ we thus obtain

$$\begin{aligned}\max_{0\leq n\leq T} \|\bar{\zeta}_n^U\|^p &= \max_{0\leq n\leq T} \|\alpha\bar{\zeta}_n^U + (1-\alpha)\bar{\zeta}_n^{-U}\|^p \\ &\leq \alpha \max_{0\leq n\leq T} \|\bar{\zeta}_n^U\|^p + (1-\alpha) \max_{0\leq n\leq T} \|\bar{\zeta}_n^{-U}\|^p.\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{E}\left[\max_{0\leq n\leq T} \|\zeta_n\|^p | \mathcal{F}_t\right] &\leq \alpha \mathbb{E}\left[\max_{0\leq n\leq T} \|\bar{\zeta}_n^U\|^p | \mathcal{F}_t\right] + (1-\alpha)\mathbb{E}\left[\max_{0\leq n\leq T} \|\bar{\zeta}_n^{-U}\|^p | \mathcal{F}_t\right] \\ &\leq \mathbb{E}\left[\max_{0\leq n\leq T} \|\bar{\zeta}_n^{U'}\|^p | \mathcal{F}_t\right],\end{aligned}$$

where U' is the $\mathcal{F}_t$-measurable random orthonormal transformation given by

$$U'(\omega) = \begin{cases} U(\omega) & \text{if } \omega \in \big\{\mathbb{E}[\max_{0\leq n\leq T} \|\bar{\zeta}_n^U\|^p | \mathcal{F}_t] \geq \mathbb{E}[\max_{0\leq n\leq T} \|\bar{\zeta}_n^{-U}\|^p | \mathcal{F}_t]\big\}, \\ -U(\omega) & \text{otherwise.}\end{cases}$$

Applying (107) and (108) with $U = U'$ finishes the induction step.

Next, we consider the case of $c > 0$. Suppose that ζ, ξ and M are as stated in the theorem. For $n \in \mathbb{N}$ we put

$$\tilde{\xi}_n = (1 - c/\|\xi_n\|)_+ \cdot \xi_n$$

and

$$\tilde{\zeta}_n = \tilde{\xi}_n + \Delta M_n.$$

Furthermore, let $\tilde{\zeta}_0 = \zeta_0 = M_0$. We will show that the triple $(\tilde{\zeta}, \tilde{\xi}, M)$ satisfies (106) with $c = 0$. Clearly, $(\tilde{\zeta}_n)_{n\in\mathbb{N}_0}$ is adapted and $(\tilde{\xi}_n)_{n\in\mathbb{N}}$ is previsible. Moreover, one has for $n \in \mathbb{N}$ on $\{\|\xi_n\| \geq c\}$ that

$$\|\tilde{\xi}_n\| = \|\xi_n\| - c \le \|\zeta_{n-1}\| - c = \|\tilde{\zeta}_{n-1} + \xi_{n-1} - \tilde{\xi}_{n-1}\| - c$$
$$\le \|\tilde{\zeta}_{n-1}\| + \|\xi_{n-1} - \tilde{\xi}_{n-1}\| - c = \|\tilde{\zeta}_{n-1}\|$$

and on $\{\|\xi_n\| < c\}$ that $\|\tilde{\xi}_n\| = 0 \le \|\tilde{\zeta}_{n-1}\|$. We may thus apply Theorem 5.1 with $c = 0$ to obtain that for every $n \in \mathbb{N}$,

$$\mathbb{E}\left[\max_{0\le k\le n} \|\tilde{\zeta}_n\|^p\right] \le \bar{\kappa}\, \mathbb{E}\big[[M]_n^{p/2}\big].$$

Since for every $n \in \mathbb{N}$,

$$\|\zeta_n\|^p = \|\tilde{\zeta}_n + \xi_n - \tilde{\xi}_n\|^p \le 2^p(\|\tilde{\zeta}_n\|^p + c^p),$$

we conclude that

$$\mathbb{E}\left[\max_{0\le k\le n} \|\zeta_n\|^p\right] \le 2^p\big(\bar{\kappa}\, \mathbb{E}\big[[M]_n^{p/2}\big] + c^p\big) \le 2^p(\bar{\kappa} \vee 1) \cdot \big(\mathbb{E}\big[[M]_n^{p/2}\big] + c^p\big),$$

which completes the proof. □

References

1. Benveniste, A., Métivier, M., Priouret, P.: Adaptive Algorithms and Stochastic Approximations. Volume 22 of Applications of Mathematics (New York), vol. 22. Springer, Berlin (1990)
2. Duflo, M.: Algorithmes Stochastiques. Volume 23 of Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 23. Springer, Berlin (1996)
3. Frikha, N.: Multi-level stochastic approximation algorithms. Ann. Appl. Probab. **26**, 933–985 (2016)
4. Gaposhkin, V.F., Krasulina, T.P.: On the law of the iterated logarithm in stochastic approximation processes. Theory Probab. Appl. **19**(4), 844–850 (1974)
5. Giles, M.B.: Multilevel Monte Carlo path simulation. Oper. Res. **56**(3), 607–617 (2008)
6. Heinrich, S.: Multilevel Monte Carlo methods. In: Margenov, S., Waśniewski, J., Yalamov, P. (eds.) Large-Scale Scientific Computing, pp. 58–67. Springer, Berlin (2001)
7. Kushner, H.J., Yang, J.: Stochastic approximation with averaging of the iterates: optimal asymptotic rate of convergence for general processes. SIAM J. Control Optim. **31**(4), 1045–1062 (1993)
8. Kushner, H.J., Yin, G.G.: Stochastic Approximation and Recursive Algorithms and Applications, Volume 35 of Applications of Mathematics (New York). Stochastic Modelling and Applied Probability, 2nd edn. Springer, New York (2003)
9. Lai, T.L.: Stochastic approximation. Ann. Stat. **31**(2), 391–406 (2003). Dedicated to the memory of Herbert E. Robbins
10. Lai, T.L., Robbins, H.: Limit theorems for weighted sums and stochastic approximation processes. Proc. Nat. Acad. Sci. U.S.A. **75**, 1068–1070 (1978)
11. Le Breton, A., Novikov, A.: Some results about averaging in stochastic approximation. Metrika **42**(3–4):153–171 (1995). Second International Conference on Mathematical Statistics (Smolenice Castle, 1994)
12. Ljung, L., Pflug, G., Walk, H.: Stochastic Approximation and Optimization of Random Systems. Volume 17 of DMV Seminar, vol. 17. Birkhäuser Verlag, Basel (1992)
13. Nualart, D.: The Malliavin Calculus and Related Topics. Probability and Its Applications (New York), 2nd edn. Springer, Berlin (2006)
14. Pelletier, M.: On the almost sure asymptotic behaviour of stochastic algorithms. Stoch. Process. Appl. **78**(2), 217–244 (1998)

15. Pelletier, M.: Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. Ann. Appl. Probab. **8**(1), 10–44 (1998)
16. Polyak, B.T.: A new method of stochastic approximation type. Avtomat. i Telemekh. **51**(7), 937–1008 (1998)
17. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**, 400–407 (1951)
18. Ruppert, D.: Almost sure approximations to the Robbins-Monro and Kiefer-Wolfowitz processes with dependent noise. Ann. Probab. **10**, 178–187 (1982)
19. Ruppert, D.: Stochastic Approximation. In: Ghosh, B.K., Sen, P.K. (eds.) Handbook of Sequential Analysis. Volume 118 of Statist. Textbooks Monogr., pp. 503–529. Dekker, New York (1991)