SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1811–1827, September 1997

020

WEIGHTED MEANS IN STOCHASTIC APPROXIMATION OF MINIMA*

J. DIPPON† AND J. RENZ‡

Abstract. Weighted averages of Kiefer–Wolfowitz-type procedures, which are driven by larger step lengths than usual, can achieve the optimal rate of convergence. A priori knowledge of a lower bound on the smallest eigenvalue of the Hessian matrix is avoided. The asymptotic mean squared error of the weighted averaging algorithm is the same as would emerge using a Newton-type adaptive algorithm. Several different gradient estimates are considered; one of them leads to a vanishing asymptotic bias. This gradient estimate applied with the weighted averaging algorithm usually yields a better asymptotic mean squared error than applied with the standard algorithm.

Key words. stochastic approximation, acceleration by weighted averaging, weak invariance principle, consistency, Kiefer–Wolfowitz procedure, gradient estimation, optimization

AMS subject classifications. Primary, 60L20; Secondary, 60F05, 60F17, 93E23

PII. S0363012995283789

1. Introduction.

In stochastic approximation the minimizer ϑ of an unknown regression function $f : \mathbb{R}^d \to \mathbb{R}$ can be estimated by running the recursion

$$X_{n+1} = X_n - a_n Y_n, \tag{1.1}$$

where Y_n is a gradient estimate of f at the point X_n and a_n are positive step lengths decreasing to zero. For instance, for $d = 1$ and decreasing span c_n, Kiefer and Wolfowitz [11] used divided differences $Y_n = (Y_{n,1} - Y_{n,2})/(2c_n)$ as approximation of $f'(X_n)$, where $Y_{n,1}$ and $Y_{n,2}$ are error-contaminated observations of $f(X_n + c_n)$ and $f(X_n - c_n)$, respectively. If f is p-times differentiable at ϑ, and if the gradient estimates Y_n are constructed appropriately, one can obtain

$$n^{\frac{\alpha}{2}\left(1-\frac{1}{p}\right)}(X_n - \vartheta) \xrightarrow{\mathcal{D}} N(\mu, K) \quad (n \to \infty)$$

with step lengths $a_n = an^{-\alpha}$ for some $a > 0$ and $\alpha \in (0, 1]$ (see Fabian [8] for $p \geq 3$ odd). Hence, for step lengths $a_n = a/n$, the convergence rate $n^{(1-1/p)/2}$ is obtained. This is the exact minimax order in the problem of estimating the minimizer of f for f belonging to a certain class of p-times differentiable functions (Polyak and Tsybakov [18]).

In this paper we investigate *weighted means*

$$\widetilde{X}_{n,\delta} = \tfrac{1+\delta}{n^{1+\delta}} \sum_{i=1}^{n} i^{\delta} X_i \tag{1.2}$$

of Kiefer–Wolfowitz-type processes (X_n) generated by recursion (1.1) with some gradient estimates Y_n for p-times differentiable regression functions and step lengths converging slower to zero than $1/n$. We obtain

$$n^{\frac{1}{2}\left(1-\frac{1}{p}\right)}\left(\widetilde{X}_{n,\delta} - \vartheta\right) \xrightarrow{\mathcal{D}} N(\widetilde{\mu}, \widetilde{K}) \quad (n \to \infty)$$

*Received by the editors March 27, 1995; accepted for publication (in revised form) July 23, 1996. The research of the second author was supported by a Deutsche Forschungsgemeinschaft grant.
http://www.siam.org/journals/sicon/35-5/28378.html
†Mathematisches Institut A, Universität Stuttgart, 70511 Stuttgart, Germany (dippon@mathematik.uni-stuttgart.de).
‡Landesgirokasse, 70144 Stuttgart, Germany (RiskManagement@t-online.de).

for some weight parameters δ and various types of gradient estimates (Theorems 3.2 and 4.2). The main advantages are the following. First, a priori knowledge of a lower bound on the smallest eigenvalue λ_0 of the Hessian $Hf(\vartheta)$ of f at ϑ is avoided. If, in the standard algorithm with $a_n = a/n$, the constant a is chosen too small, i.e., $a \leq (1 - 1/p)/(2\lambda_0)$, convergence can be very slow. To be safe one might choose a pretty large. But the asymptotic mean squared error (AMSE) produced by the standard algorithm grows approximately linearly in a. These problems do not arise when the averaging algorithm is applied. On the other side, if an asymptotic bias is present, the AMSE of the averaging algorithm cannot be greater than four times the AMSE of the standard algorithm with the optimal, but usually unknown, constant a. In this sense the averaging algorithm can be considered to be more stable than the standard one. Furthermore, the averaging algorithm shows the same limit distribution as the Newton-type adaptive procedure suggested by Fabian [9] (section 5).

The method proposed in this paper is inspired by an idea of Ruppert [21] and Polyak [16], who suggested considering the arithmetic mean of the trajectories of a Robbins–Monro process, which is driven by step lengths slower than $1/n$, too. In this case one obtains the best possible convergence rate and the optimal covariance of the asymptotic distribution in a certain sense [17]. Since then Yin [27], Pechtl [15], Kushner and Yang [13], Györfi and Walk [10], Nazin and Shcherbakov [14], and others have studied this idea.

A further contribution of this paper is a new design to estimate the gradient which leads to a vanishing asymptotic bias $\widetilde{\mu}$ (for $d = 1$ see Renz [19]) regardless of which method (with or without averaging, or with adaptation) is used. Applying the weighted averaging algorithm together with this gradient estimate leads to a second moment of the asymptotic distribution which is minimal within a large class of procedures (relation (5.4)).

Spall [22] introduced another gradient estimate Y_n, the so-called simultaneous gradient perturbation method. It uses only two observations at each step instead of $2d$ observations, as in the standard Kiefer–Wolfowitz method in $\mathbb{R}^d$. This makes it suitable for certain optimization problems in high-dimensional spaces $\mathbb{R}^d$. Taking weighted averages of the process generated with Spall's gradient estimate stabilizes the performance as discussed below (Theorem 4.2 and section 5).

All these central limit theorems require consistency of the stochastic approximation method (Propositions 3.1 and 4.1). To prove the central limit theorems we apply a weak invariance principle stated in Lemma 7.1. Taking weighted averages of the trajectories leads to an accumulation of terms due to the nonlinearity of the regression function. To cope with this effect the assumptions of this lemma are partly stronger than those of a functional central limit theorem for the nonweighted case (see Walk [24]). But fortunately, the additional conditions can be shown to be fulfilled for many stochastic approximation procedures. The assertions of both central limit theorems in this paper can be formulated as invariance principles in the spirit of Lemma 7.1.

As already indicated in Dippon and Renz [4], taking weighted averages of the trajectories works well with the original gradient estimate of Kiefer and Wolfowitz ($p = 3$).

2. Notations. For a d-dimensional Euclidean space the linear space of $d \times d$ matrices is denoted by $\mathcal{L}(\mathbb{R}^d)$. x^* is the transposed vector of $x \in \mathbb{R}^d$, A^* is the adjoint matrix, and $\operatorname{tr} A$ is the trace of $A \in \mathcal{L}(\mathbb{R}^d)$. The tensor product $x \otimes y : \mathbb{R}^d \to \mathbb{R}^d$ is defined by $\langle y, \cdot \rangle x$, where $x, y \in \mathbb{R}^d$ and $\langle \cdot, \cdot \rangle$ is the usual inner product. The space $C([0,1], \mathbb{R}^d)$ of $\mathbb{R}^d$-valued continuous functions on $[0,1]$ is equipped with the maximum

norm. $Hf(\vartheta)$ is the Hessian of a function $f : \mathbb{R}^d \to \mathbb{R}$ at $\vartheta \in \mathbb{R}^d$. For $x \in \mathbb{R}$ we use $\lfloor x \rfloor$ and $\lceil x \rceil$, denoting the integer part of x and the least integer greater than or equal to x, respectively.

Let $(\Omega, \mathcal{A}, P)$ be a probability space. Then a sequence (X_n) of $\mathbb{R}^d$-valued random variables (r.v.'s) is called bounded in probability whenever $\lim_{R\to\infty} \overline{\lim}_n P(\|X_n\| \geq R) = 0$; (X_n) converges to zero almost in L^r or is bounded almost in L^r ($r \in (0,\infty)$) if for each $\varepsilon > 0$ there exists an $\Omega_\varepsilon \in \mathcal{A}$ with $P(\Omega_\varepsilon) \geq 1-\varepsilon$ such that $(\int_{\Omega_\varepsilon} \|X_n\|^r \, dP)^{1/r} = o(1)$ or $= O(1)$, respectively. Convergence almost in L^r implies convergence in probability, but it is weaker than a.s. convergence or convergence in the rth mean.

3. A Kiefer–Wolfowitz procedure with an improved gradient estimate. The Kiefer–Wolfowitz procedure, which finds the minimizer ϑ of a regression function $f : \mathbb{R}^d \to \mathbb{R}$, has been modified by Fabian [6] in such a way that the rate of convergence nearly reaches the rate of a Robbins–Monro procedure if f is assumed to be sufficiently smooth in a neighborhood of ϑ. The method uses multiple observations per step.

We consider here, including the Fabian procedure, a modified Kiefer–Wolfowitz procedure which is given by recursion (1.1). There Y_n is an estimate of the gradient $\nabla f(X_n)$ based on error-contaminated observations of f. It is defined by

$$(3.1)\quad Y_n = c_n^{-1} \sum_{j=1}^{m} v_j \Big(\{ f(X_n + c_n u_j e_i) - V_{n,2j-1}^{(i)} \} - \{ f(X_n - c_n u_j e_i) - V_{n,2j}^{(i)} \} \Big)_{i=1,\dots,d},$$

where the following definitions and relations are used throughout section 3: $m \in \mathbb{N}$, $0 < u_1 < \cdots < u_m \leq 1$, $v_1, \dots, v_m$ are real numbers with $\sum_{j=1}^m v_j u_j^{2i-1} = (1/2)\delta_{1i}$ for all $i = 1, \dots, m$ (as to the existence, compare Fabian [6]), and $c_n = cn^{-\gamma}$ with $c > 0$ and $0 < \gamma < 1/2$. The unit vectors in $\mathbb{R}^d$ are denoted by $e_1, \dots, e_d$.

For future reference, we state the following additional conditions:

(A) ∇f exists on $\mathbb{R}^d$ with $\nabla f(\vartheta) = 0$.

Concerning the local differentiability of f at ϑ we consider two cases. In the first case ($p = 2$) we assume that there exists $\varepsilon > 0$, $\tau \in (0,1]$, K_1 and K_2 such that

(B1a) $Hf(\vartheta)$ exists with $\|\nabla f(x) - Hf(\vartheta)(x-\vartheta)\| \leq K_1 \|x-\vartheta\|^{1+\tau}$ for all $x \in U_\varepsilon(\vartheta)$,
(B1b) $\|\nabla f(x) - \nabla f(y)\| \leq K_2 \|x-y\|$ for all $x, y \in U_\varepsilon(\vartheta)$.

(B1b) holds, for instance, if all second partial derivatives of f exist and are bounded on $U_\varepsilon(\vartheta)$. For the second case ($p \geq 3$), we assume that there exist $\varepsilon > 0$ and L such that

(B2a) derivatives of f up to order $p-1$ exist on $U_\varepsilon(\vartheta)$,
(B2b) the pth derivative of f at ϑ exists,
(B2c) $\|Hf(x) - Hf(y)\| \leq L\|x-y\|$ for all $x, y \in U_\varepsilon(\vartheta)$.

A sufficient condition for (B2c) to hold is that all third partial derivatives of f exist and are bounded on $U_\varepsilon(\vartheta)$.

For brevity, (B1) stands for (B1a) and (B1b), and (B2) for (B2a), (B2b), and (B2c). We use (B) to indicate that either (B1) or (B2) holds.

So far, m has not been specified. The number m must be adapted to the particular value of p given by (B1) or (B2). Fabian [6] considers in this connection the case

(C1) $m := \lfloor p/2 \rfloor = (p-1)/2$ for an odd $p \geq 3$, $\gamma := 1/(2p)$.

We will consider in addition the following case (for $d = 1$ see Renz [19]):

(C2) $m := \lceil p/2 \rceil$ for $p \geq 2$ (p not necessarily odd), $\gamma := 1/(2p)$,

which will result in an unbiased limit distribution, whereas (C1) generally leads to a nonzero bias (Theorem 3.2).

Similarly as above, (C) means that either (C1) or (C2) holds. We note here that the assumptions (B1) and (C1) do not occur together.

The sequence (W_n) of random variables $W_n := \sum_{j=1}^m v_j \left(V_{n,2j-1}^{(i)} - V_{n,2j}^{(i)}\right)_{i=1,\dots,d}$ satisfies

$$\text{(D)} \qquad \underset{n\geq m}{\forall} \quad \|EW_m \otimes W_n\| \leq \varrho_{n-m}(E\|W_m\|^2 E\|W_n\|^2)^{\frac{1}{2}}$$
$$\text{with } \sum_{l=0}^{\infty} \varrho_l < \infty \text{ and } E\|W_n\|^2 = O(1).$$

Regarding assumption (B2b) it is worthwhile to note that this condition is invariant under rotation of coordinates (compare Fabian [8]). As a further comparison with related work (Fabian [8], Spall [23]), we remark that our results, Theorems 3.2 and 4.2, do not assume continuity of the highest-order partial derivatives.

Results about asymptotic normality in stochastic approximation usually rely on local smoothness of the regression function f around ϑ and on the consistency of the procedure. The next proposition shows consistency of the modified procedure. The assumptions imposed on f allow us to decouple the influence of the r.v.'s W_n and to use the weak dependence condition (D).

PROPOSITION 3.1. *Let $a_n = a/n^\alpha$ with $\alpha \in (\max\{1/2 + 1/(2p), 1 - 1/p\}, 1)$ or $a_n = (a \ln n)/n$, $a > 0$. For recursion (1.1) with gradient estimate (3.1), assume that conditions (A) and (D) hold, f is bounded from below and has a Lipschitz continuous derivative with $\nabla f(x) \neq 0$ for all $x \neq \vartheta$, and $\sup\{\|x\| : f(x) \leq \lambda\} < \infty$ for all $\lambda > \inf\{f(x) : x \in \mathbb{R}^d\}$. Then $X_n \to \vartheta$ $(n \to \infty)$ a.s.*

Under condition (C1) a nonweighted analogue of the next theorem can be found in Fabian [8].

THEOREM 3.2. *Let $a_n = (a \ln n)/n$ for $p = 2$ and $a_n = a/n^\alpha$ with $\alpha \in (1/2+1/(2p), 1)$ for $p \geq 3$. For recursion (1.1) with gradient estimate (3.1), assume that conditions (A)–(D) hold, $A := Hf(\vartheta)$ is positive definite, and $X_n \to \vartheta$ a.s. Let $B_n(t) := n^{-1/2}\left\{\sum_{i=1}^{\lfloor nt\rfloor} W_i + (nt - \lfloor nt\rfloor)W_{\lfloor nt\rfloor+1}\right\}$. Suppose the existence of a Brownian motion B with covariance matrix S of $B(1)$ and*

$$B_n \xrightarrow{\mathcal{D}} B \quad \text{in } C([0,1], \mathbb{R}^d) \quad (n \to \infty).$$

Then, for all $\delta > -(p+1)/(2p)$,

$$n^{\frac{1}{2}\left(1-\frac{1}{p}\right)}\left(\widetilde{X}_{n,\delta} - \vartheta\right) \xrightarrow{\mathcal{D}} N\left(\tfrac{2p(1+\delta)}{p+1+2p\delta}\, c^{p-1}A^{-1}b, \tfrac{p(1+\delta)^2}{p+1+2p\delta}\, c^{-2}A^{-1}SA^{-1}\right) \quad (n \to \infty),$$

where $b = -\frac{1}{p!}\left(\sum_{j=1}^m v_j u_j^p (1 + (-1)^{p+1}) \frac{\partial^p}{(\partial x_i)^p} f(\vartheta)\right)_{i=1,\dots,d}$ and $\widetilde{X}_{n,\delta}$ is defined in (1.2). In particular, under condition (C2), $b = 0$.

REMARK 3.3. *The choices $\delta = 0$ and $\delta = -2\gamma = -1/p$ are of special interest. Provided $b \neq 0$, the pair $(\delta, c) = (0, c_0)$ with c_0 as given in (5.1) minimizes the second moment of the limit distribution. However, for fixed $c > 0$, the limit's covariance is minimized by $\delta = -2\gamma = -1/p$. In particular, Theorem 3.2 yields for $n \to \infty$*

$$n^{\frac{1}{2}\left(1-\frac{1}{p}\right)}\left(n^{-1}\sum_{k=1}^n X_k - \vartheta\right) \xrightarrow{\mathcal{D}} N\left(\tfrac{2p}{p+1}\, c^{p-1}A^{-1}b, \tfrac{p}{p+1}\, c^{-2}A^{-1}SA^{-1}\right),$$

$$n^{\frac{1}{2}\left(1-\frac{1}{p}\right)}\left(\tfrac{p-1}{p}\, n^{-\frac{p-1}{p}}\sum_{k=1}^n k^{-\frac{1}{p}} X_k - \vartheta\right) \xrightarrow{\mathcal{D}} N\left(2\, c^{p-1}A^{-1}b,\ \tfrac{p-1}{p}\, c^{-2}A^{-1}SA^{-1}\right).$$

4. A Kiefer–Wolfowitz procedure with simultaneous perturbation gradient approximation. The classical Kiefer–Wolfowitz (finite difference) stochastic approximation method (FDSA) needs $2d$ observations to obtain a finite difference approximation of the gradient belonging to the function $f : \mathbb{R}^d \to \mathbb{R}$ of which the minimizer ϑ is sought. To reduce the number of observations in each step, randomized gradient approximation methods have been considered in the literature. Two examples are random direction stochastic approximation (RDSA), suggested by Kushner and Clark [12], and simultaneous perturbation stochastic approximation (SPSA), suggested by Spall [22]. Both methods are based on only two observations in each iteration. Depending on the dimension d and the third derivatives of the regression function f the AMSE of the SPSA method can be better or worse than that of the FDSA and RDSA methods. At least for second-order polynomials f, the FDSA method needs d times more observations than the SPSA method to achieve the same level of mean squared error asymptotically, when the same span $c_n = cn^{-\gamma}$ is used (Spall [23]).

Before the idea of weighted averages is applied to the SPSA method, we will describe this algorithm in more detail. Again recursion (1.1) is used, but with step lengths $a_n = an^{-\alpha}$ and with the following so-called *simultaneous perturbation gradient estimate* of $\nabla f(X_n)$:

$$Y_n = \frac{1}{2c_n} \begin{pmatrix} (\Delta_n^{(1)})^{-1} \\ \vdots \\ (\Delta_n^{(d)})^{-1} \end{pmatrix} ([f(X_n + c_n\Delta_n) - W_{n,1}] - [f(X_n - c_n\Delta_n) - W_{n,2}]) \tag{4.1}$$

consisting of (artificially generated) random vectors $\Delta_n \in \mathcal{M}(\Omega, \mathbb{R}^d)$, observation errors $W_{n,1}$, $W_{n,2} \in \mathcal{M}(\Omega, \mathbb{R})$, and span c_n.

We consider the following set of conditions.

(E) The components $\Delta_n^{(i)}$ of Δ_n, $i = 1, \dots, d$, for $n \in \mathbb{N}$ fixed, form a set of independent, identically and symmetrically distributed r.v.'s with $|\Delta_n^{(l)}|$ having values between fixed positive numbers $\alpha_0 < \alpha_1$. The r.v. Δ_n is assumed to be independent of $\{X_1, \dots, X_n, \Delta_1, \dots, \Delta_{n-1}\}$. Furthermore, we use $\xi^2 = E|\Delta_n^{(l)}|^2$ and $\rho^2 = E|\Delta_n^{(l)}|^{-2}$. For simplicity, the column vector appearing in (4.1) is denoted by Δ_n^{-1}.

(F) The difference $W_n = W_{n,1} - W_{n,2}$ of the observation errors satisfies $E(W_n \mid \mathcal{F}_n) = 0$ and $\sup_n E\left(W_n^2 \mid \mathcal{G}_n\right) < \infty$ a.s., where $\mathcal{F}_n$ and $\mathcal{G}_n$ denote the σ-fields generated by $\{X_1, \dots, X_n, \Delta_1, \dots, \Delta_n\}$ and $\{X_1, \dots, X_n, \Delta_1, \dots, \Delta_{n-1}\}$, respectively.

(G) $\infty > E(W_n^2 \mid \mathcal{F}_n) \to \sigma^2$ a.s. and $E(W_n^2 1_{[W_n^2 \geq rn]} \mid \mathcal{F}_n) \to 0$ a.s. for every $r > 0$.

(H) (B2) holds for $p = 3$, and $A = Hf(\vartheta)$ is a positive definite matrix.

The proposition below presents conditions for the recursion's consistency. It is related to Blum's result [2] on multivariate Kiefer–Wolfowitz procedures. Under different and less intuitive assumptions and with a different method of proof, Spall [23] asserts consistency as well.

PROPOSITION 4.1. *Let $a_n = a/n^\alpha$ with $\alpha \in (\max\{\gamma + 1/2, 1 - 2\gamma\}, 1]$ and $\gamma > 0$. For recursion* (1.1) *with gradient estimate* (4.1)*, assume that conditions* (A), (E), *and* (F) *hold, and that f is bounded from below and has a Lipschitz continuous gradient.*

(a) *If $\sup\{\|x\| : f(x) \leq \lambda\} < \infty$ for all $\lambda > \inf\{f(x) : x \in \mathbb{R}^d\}$, then $\sup_n \|X_n\| < \infty$ a.s.*

(b) *Assume* $\nabla f(x) \neq 0$ *and* $f(x) > f(\vartheta)$ *for all* $x \neq \vartheta$. *If* $\sup_n \|X_n\| < \infty$ *a.s., then* $X_n \to \vartheta$ $(n \to \infty)$ *a.s.*

A nonweighted analogue of the following theorem is stated in Spall [23].

THEOREM 4.2. *Let* $\alpha \in (2/3, 1)$ *and* $\gamma = 1/6$. *For recursion* (1.1) *with gradient estimate* (4.1), *assume conditions* (A), (E)–(H), *and* $X_n \to \vartheta$ *a.s. Then, for all* $\delta > -2/3$,

$$n^{\frac{1}{3}} \left(\widetilde{X}_{n,\delta} - \vartheta \right) \xrightarrow{\mathcal{D}} N \left(\tfrac{1+\delta}{2/3+\delta} c^2 A^{-1} b, \tfrac{(1+\delta)^2}{4/3+2\delta} c^{-2} A^{-1} S A^{-1} \right) \quad (n \to \infty),$$

where

$$S = \frac{\sigma^2 \rho^2}{4} I, \qquad b = -\frac{1}{6} \xi^2 \left(\frac{\partial^3}{(\partial x_i)^3} f(\vartheta) + 3 \sum_{j=1,\, j \neq i}^{d} \frac{\partial^3}{\partial x_i (\partial x_j)^2} f(\vartheta) \right)_{i=1,\ldots,d},$$

and $\widetilde{X}_{n,\delta}$ *is as defined in* (1.2).

5. Comparison of stochastic approximation procedures with respect to their asymptotic mean squared error and further comments. Based on recursion (1.1) with any of the gradient estimates Y_n discussed in this paper, we consider the following three variants of algorithms:

(i) the basic recursion with $a_n = a/n$,
(ii) an adaptive variant obtained from the basic recursion with $a_n = (a/n) M_n$ and random matrices M_n converging to $M = Hf(\vartheta)^{-1}$ a.s.,
(iii) the basic recursion with a_n converging to zero slower than $1/n$ combined with averaging of the trajectories

and compare the corresponding estimators with regard to their asymptotic behavior. Some of these estimators have been treated in the literature (Fabian [6], [9], Spall [22], [23]). The adaptive procedure has been introduced by Fabian [9] to improve the limit distribution. There the auxiliary sequence M_n is built up from information available up to stage n. For both algorithms (i) and (ii), with any gradient estimate considered in this paper, the limit distribution can be obtained by Theorem 1 in Walk [24] and by the representations derived in the proofs of Theorems 3.2 and 4.2.

Assuming that $A = Hf(\vartheta)$ is a positive definite matrix, the related second moments of the asymptotic distributions turn out to be

$$E(a,c) := \left(2c^{p-1} a \|(2aA - \beta)^{-1} b\| \right)^2 + \frac{a^2}{c^2} \operatorname{tr} \left((2aA - \beta)^{-1} S \right), \quad a > \beta/(2\lambda_0),$$

$$\widehat{E}(a,c) := \left(\frac{2c^{p-1} a}{2a - \beta} \|A^{-1} b\| \right)^2 + \frac{a^2}{c^2 (2a - \beta)} \operatorname{tr} \left(A^{-1} S A^{-1} \right), \quad a > \beta/2,$$

$$\widetilde{E}(\delta,c) := \left(\frac{2c^{p-1}(1+\delta)}{2 - \beta + 2\delta} \|A^{-1} b\| \right)^2 + \frac{(1+\delta)^2}{c^2 (2 - \beta + 2\delta)} \operatorname{tr} \left(A^{-1} S A^{-1} \right), \quad \delta > \beta/2 - 1,$$

respectively, where $\beta = 1 - 1/p$, $\lambda_0 = \min\{\lambda : \lambda \in \operatorname{spec} A\}$, and $c > 0$ (concerning E and $\widehat{E}$, use Theorem 5.8 and Remark 5.9 of [25]). Under appropriate assumptions these quantities are equal to the AMSEs $\lim_n E\|n^{\frac{1}{2}(1-\frac{1}{p})}(X_n - \vartheta)\|^2$ shown by algorithm (i) or (ii), and $\lim_n E\|n^{\frac{1}{2}(1-\frac{1}{p})}(\widetilde{X}_{n,\delta} - \vartheta)\|^2$ shown by algorithm (iii). Apparently it holds that $\widehat{E}(a,c) = \widetilde{E}(a-1,c)$.

If $b \neq 0$, the asymptotic distribution is biased. In this case $\widehat{E}$ and $\widetilde{E}$ are minimized by $(a, c) = (1, c_0)$ and $(\delta, c) = (0, c_0)$, respectively, with

$$c_0 = \left(\frac{(2-\beta)}{4(p-1)} \frac{\operatorname{tr}(A^{-1}SA^{-1})}{\|A^{-1}b\|^2} \right)^{\frac{1}{2p}}, \tag{5.1}$$

which is usually unknown. At the end of section 6 we show that

$$\forall c > 0 \qquad \frac{1}{4} < \left(\frac{p+1}{2p} \right)^2 < \min_{a > \beta/(2\lambda_0)} \frac{E(a,c)}{\widetilde{E}(0,c)} < \sup_{a > \beta/(2\lambda_0)} \frac{E(a,c)}{\widetilde{E}(0,c)} = \infty \tag{5.2}$$

and

$$\frac{1}{4} < \left(\frac{p+1}{2p} \right)^2 < \min_{a > \beta/(2\lambda_0)} \min_{c > 0} \frac{E(a,c)}{\widetilde{E}(0,c_0)} < \sup_{a > \beta/(2\lambda_0)} \min_{c > 0} \frac{E(a,c)}{\widetilde{E}(0,c_0)} = \infty. \tag{5.3}$$

Noticing the last equation of the preceding paragraph, these relations can be rewritten in terms of $\widehat{E}$ instead of $\widetilde{E}$. Thus the AMSE of the adaptive algorithm (ii) and the averaging algorithm (iii) is less than four times the AMSE of the standard algorithm (i) for any admissible a, no matter whether a common c is used or the optimal values of c are chosen. On the opposite side, a bad choice of a (a close to $\beta/(2\lambda_0)$ or a too large) results in an arbitrarily large AMSE of the standard algorithm (i), whereas this difficulty does not arise when the adaptive or averaging method is used. In this sense one may say that the averaging and adaptive algorithms are more stable than the standard algorithm.

In the one-dimensional case the AMSE of the standard algorithm (i) is minimized by $a_0' = 1/A$ and $c_0' = (\frac{2-\beta}{4(p-1)} \frac{S}{b^2})^{1/(2p)}$. Hence, for $d = 1$, the second relation in (5.3) can be sharpened to $E(a_0', c_0')/\widetilde{E}(0, c_0) = 1$.

In section 3 the design $(u_1, \ldots, u_m)$ was fixed. If condition (C1) holds, the gradient estimate (3.1) usually produces an asymptotic bias. In this case Fabian [7] and Erickson, Fabian, and Mařik [5] investigated how the AMSE can be further reduced by the choice of an optimal design.

If the gradient estimate (3.1) is constructed under condition (C2), the bias is vanishing (since $b = 0$). Then, for a fixed positive c, the AMSEs $\widehat{E}$ and $\widetilde{E}$ attain their minimum $c^{-2}(1 - 1/p) \operatorname{tr}(A^{-1}SA^{-1})$ for $a = 1 - 1/p$ and $\delta = -1/p$, respectively. We get

$$\forall c > 0 \qquad 1 \leq \min_{a > \beta/(2\lambda_0)} \frac{E(a,c)}{\widetilde{E}(-1/p, c)} < \sup_{a > \beta/(2\lambda_0)} \frac{E(a,c)}{\widetilde{E}(-1/p, c)} = \infty. \tag{5.4}$$

Assume that a_0 ($> \beta/(2\lambda_0)$) minimizes $E(a, c)$ for a fixed c. Then, only in special cases can $E(a_0, c) = \widetilde{E}(-1/p, c)$ be achieved. In any of the three variants the related AMSE can be made arbitrarily small by choosing c sufficiently large. Hence, with respect to the AMSE criterion, the procedures using the gradient estimate leading to $b = 0$ are superior to those leading to $b \neq 0$, although they need $2d$ more observations per step.

It must be emphasized that in the case $b \neq 0$ the adaptive recursion employing consistent estimators (M_n) of $M = A^{-1}$ instead of some other matrix M is, due to (5.2) and (5.3), a fairly good choice but not the best one with respect to the optimal AMSE. For fixed $c > 0$, a better choice would require consistent estimators M_n of

a matrix M which minimizes $c^{2p-2}\|(MA-\frac{\beta}{2}I)^{-1}Mb\|^2 + c^{-2}\operatorname{tr} Z$ where $\min\{\operatorname{re}\lambda : \lambda \in \operatorname{spec}(MA-\frac{\beta}{2}I)\} > 0$ and Z is the unique solution of $(MA-\frac{\beta}{2}I)Z + Z(MA-\frac{\beta}{2}I)^* = MSM^*$. Hence, averaging applied in the Kiefer–Wolfowitz situation with nonvanishing asymptotic bias does not optimize the AMSE. This is in contrast to the Robbins–Monro situation. Minimizing the expression above in both M and c would lead to an even better AMSE, but we do not pursue this possibility here.

Let $\vartheta_n(f)$ be an estimator of the minimum $\vartheta(f)$ of a p-times differentiable regression function f on $\mathbb{R}$ using n observations. Consider, for fixed $c > 0$,

$$\sup_f P[n^{\frac{p-1}{2p}}|\vartheta_n(f) - \vartheta(f)| > c],$$

where the supremum is taken over all regression functions f satisfying conditions (A) and (B) and some further boundedness conditions. Then, according to results by Chen [3] and by Polyak and Tsybakov [18], this supremum as a function of n has a universal positive lower bound independent of the choice of $\vartheta_n(f)$. This raises the interesting problem of determining which type of algorithm, together with which type of gradient estimate, leads to the smallest supremum above.

The condition $\sup\{\|x\| : f(x) \le \lambda\} < \infty$ for all $\lambda > \inf\{f(x) : x \in \mathbb{R}^d\}$ appearing in Propositions 3.1 and 4.1 is equivalent to $\inf\{f(x) : \|x\| \ge K\} \to \infty$ as $K \to \infty$. In applications this condition can be satisfied by adding the function values of an appropriate increasing and differentiable function to the basic observations taken at x. A possible choice is $x \mapsto \|x - d\frac{x}{\|x\|}\|^2 1_{[\|x\|\ge d]}$ for a fixed d large enough.

Finally, it is worth mentioning that the weighted means $\widetilde{X}_{n,\delta}$ can easily be recursified by

$$\begin{pmatrix} X_{n+1} \\ \widetilde{X}_{n+1,\delta} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{1+\delta}{n+1} & \left(\frac{n}{n+1}\right)^{1+\delta} \end{pmatrix} \begin{pmatrix} X_n \\ \widetilde{X}_{n,\delta} \end{pmatrix} - a_n \begin{pmatrix} Y_n \\ \frac{1+\delta}{n+1} Y_n \end{pmatrix}, \quad \widetilde{X}_{1,\delta} = (1+\delta)X_1.$$

6. Proofs.

Proof of Proposition 3.1. For $x \in \mathbb{R}^d$ and $h > 0$ define

$$g(x,h) := \left(g^{(i)}(x,h)\right)_{i=1,\dots,d} := \sum_{j=1}^m v_j (f(x+hu_je_i) - f(x-hu_je_i))_{i=1,\dots,d}.$$

Then, with $V_n := c^{-1}W_n$ and $H_n := \nabla f(X_n) - c_n^{-1}g(X_n, c_n)$, we obtain

$$X_{n+1} = X_n - a_n(\nabla f(X_n) - n^\gamma V_n - H_n). \tag{6.1}$$

By Lipschitz continuity of ∇f we have $f \in C^1(\mathbb{R}^d)$. This leads, with a Lipschitz constant K, to

$$\begin{aligned} \left|h^{-1}g^{(i)}(x,h) - \tfrac{\partial}{\partial x_i} f(x)\right| &= \left|\sum_{j=1}^m v_j u_j \int_{-1}^1 (\tfrac{\partial}{\partial x_i} f(x + shu_je_i) - \tfrac{\partial}{\partial x_i} f(x))\, ds\right| \\ &\le \sum_{j=1}^m |v_j| u_j \int_{-1}^1 \|\nabla f(x + shu_je_i) - \nabla f(x)\|\, ds \le K \sum_{j=1}^m |v_j| u_j^2 h. \end{aligned}$$

Therefore $\|H_n\| \le \sqrt{d}K \sum_{j=1}^m |v_j| u_j^2 c_n$. Our assumptions yield $\sum a_n^2 n^{1/p} (\log n)^2 < \infty$ and $\sum a_n n^{-1/p} < \infty$. Proposition 4.1 in Dippon and Renz [4] implies the assertion. ∎

Proof of Theorem 3.2. *First step*: *Expansions for* $h^{-1}g(x,h)$. In what follows, we assume $x \in U_{\varepsilon/2}(\vartheta)$ and $h \in (0, \varepsilon/2)$. As a consequence we have $x + she_i$, $\vartheta + t(x - \vartheta) \pm hu_j e_i \in U_\varepsilon(\vartheta)$ for all $t \in [0,1]$ and $s \in [-1,1]$.

First we consider the case of an at least three-times differentiable function f ($p \geq 3$). Using (B2c), we obtain $f \in C^2(U_\varepsilon(\vartheta))$, and therefore, according to Taylor's formula,

$$
\begin{aligned}
(6.2)\quad g^{(i)}(x,h) &= \sum_{j=1}^{m} v_j (f(x + hu_j e_i) - f(x - hu_j e_i)) \\
&= \sum_{j=1}^{m} v_j (f(\vartheta + hu_j e_i) - f(\vartheta - hu_j e_i)) \\
&\quad + \sum_{j=1}^{m} v_j (\nabla f(\vartheta + hu_j e_i) - \nabla f(\vartheta - hu_j e_i))^* (x - \vartheta) \\
&\quad + (x - \vartheta)^* \int_0^1 (1-t) \sum_{j=1}^{m} v_j (Hf(\vartheta + t(x - \vartheta) + hu_j e_i) \\
&\qquad\qquad - Hf(\vartheta + t(x - \vartheta) - hu_j e_i))\, dt\, (x - \vartheta)\,.
\end{aligned}
$$

Let us denote the first term of this sum by $t^{(i)}(h)$ $(i = 1, \ldots, d)$. (B2a) implies

$$
\tfrac{d^l}{(dh)^l} t^{(i)}(h) = \sum_{j=1}^{m} v_j u_j^l \left(\tfrac{\partial^l}{(\partial x_i)^l} f(\vartheta + hu_j e_i) + (-1)^{l+1} \tfrac{\partial^l}{(\partial x_i)^l} f(\vartheta - hu_j e_i) \right)
$$

for $l = 0, \ldots, p-1$. Then $\frac{d^l}{(dh)^l} t^{(i)}(0) = 0$ for all $l = 0, \ldots, p-1$. This is obvious for l even. For l odd with $1 \leq l \leq 2m-1$, this follows from (A) and the choice of the v_k. In the case $m := \lceil p/2 \rceil$, we have $p - 1 \leq 2m - 1$, and in the case $m := \lfloor p/2 \rfloor = (p-1)/2$, p odd, we have $2m - 1 = p - 2$ and $p - 1$ is even. (B2b) implies

$$
\tfrac{d^p}{(dh)^p} t^{(i)}(0) = \sum_{j=1}^{m} v_j u_j^p \left(1 + (-1)^{p+1}\right) \tfrac{\partial^p}{(\partial x_i)^p} f(\vartheta).
$$

In the case $m := \lceil p/2 \rceil$ we obtain $\frac{d^p}{(dh)^p} t^{(i)}(0) = 0$. For p even, this is again obvious, and for p odd, it follows from $2m - 1 = p$ and from the choice of the v_k. Taylor's formula yields

$$
(6.3)\qquad t^{(i)}(h) = \tfrac{h^p}{p!} \left(\tfrac{d^p}{(dh)^p} t^{(i)}(0) + o(1) \right) \quad (h \to 0).
$$

For the discussion of the second term of the sum on the right-hand side (r.h.s.) in (6.2) we define

$$
s^{(i,k)}(h) := \sum_{j=1}^{m} v_j \left(\tfrac{\partial}{\partial x_k} f(\vartheta + hu_j e_i) - \tfrac{\partial}{\partial x_k} f(\vartheta - hu_j e_i) \right) \quad (i,k = 1, \ldots, d).
$$

Using (B2a) we obtain by a consideration analogous to that above

$$
\tfrac{d^l}{(dh)^l}\, s^{(i,k)}(h) = \sum_{j=1}^{m} v_j u_j^l \left(\tfrac{\partial^l}{(\partial x_i)^l} \tfrac{\partial}{\partial x_k} f(\vartheta + hu_j e_i) + (-1)^{l+1} \tfrac{\partial^l}{(\partial x_i)^l} \tfrac{\partial}{\partial x_k} f(\vartheta - hu_j e_i) \right)
$$

for $l = 0, \dots, p-2$, where $s^{(i,k)}(0) = 0$, $\frac{d}{dh} s^{(i,k)}(0) = \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_k} f(\vartheta)$ and $\frac{d^l}{(dh)^l} s^{(i,k)}(0) = 0$ for all $l = 2, \dots, p-2$. (B2b) implies, by reasoning similar to that in the case of $\frac{d^p}{(dh)^p} t^{(i)}(0)$,

$$\tfrac{d^{p-1}}{(dh)^{p-1}} s^{(i,k)}(0) \;=\; \sum_{j=1}^{m} v_j u_j^{p-1} \left(1 + (-1)^p\right) \tfrac{\partial^{p-1}}{(\partial x_i)^{p-1}} \tfrac{\partial}{\partial x_k} f(\vartheta) \;=\; 0.$$

Again, by using Taylor's formula, we obtain

$$s^{(i,k)}(h) = h \tfrac{\partial}{\partial x_i} \tfrac{\partial}{\partial x_k} f(\vartheta) + h^{p-1} o(1) \quad (h \to 0). \tag{6.4}$$

Finally, for every $i = 1, \dots, d$, the expression

$$q^{(i)}(x,h) \;:=\; (x-\vartheta)^* \int_0^1 (1-t) \sum_{j=1}^{m} v_j \left(Hf(\vartheta + t(x-\vartheta) + h u_j e_i) - Hf(\vartheta + t(x-\vartheta) - h u_j e_i)\right) dt$$

can be bounded in the following way by using (B2c):

$$\|q^{(i)}(x,h)\| \;\le\; h \sum_{j=1}^{m} |v_j|\, u_j\, L \|x - \vartheta\|. \tag{6.5}$$

Because of (6.2), (6.3), (6.4), and (6.5) we obtain the following representation:

$$h^{-1} g(x,h) = \left(Hf(\vartheta) + h^{p-2} P(h) + Q(x,h)\right)(x-\vartheta) - \tfrac{h^{p-1}}{c^{p-1}} T(h) \tag{6.6}$$

with matrices $P(h)$, $Q(x,h)$ and a vector $T(h)$ satisfying the relations $\|P(h)\| = o(1)$ $(h \to 0)$, $\|Q(x,h)\| \le \sqrt{d} L \sum_{j=1}^{m} |v_j|\, u_j \, \|x - \vartheta\|$, and $T(h) \to T = c^{p-1} b$ $(h \to 0)$. Notice that Q is a measurable function $(x \in U_{\varepsilon/2}(\vartheta),\, h \in (0, \varepsilon/2))$.

Now we are going to consider the case of a twice differentiable function f $(p = 2)$. (B1a) implies the existence of a measurable matrix-valued function R with

$$\nabla f(x) = \left(Hf(\vartheta) + R(x)\right)(x - \vartheta), \;\; \text{where } \|R(x)\| \le K_1 \|x - \vartheta\|^\tau \text{ for } x \in U_\varepsilon(\vartheta).$$

Because of $p = 2$ we have $m = 1$. Without loss of generality, we may assume that $u_1 = 1$. Then we have $v_1 = 1/2$. By (B1b) we obtain $f \in C^1(U_\varepsilon(\vartheta))$, and therefore

$$\begin{aligned} g^{(i)}(x,h) &= \tfrac{1}{2}(f(x + h e_i) - f(x - h e_i)) = \tfrac{1}{2} h \int_{-1}^{1} \tfrac{\partial}{\partial x_i} f(x + s h e_i)\, ds \\ &= \tfrac{1}{2} h \int_{-1}^{1} (\tfrac{\partial}{\partial x_i} f(x + s h e_i) - \tfrac{\partial}{\partial x_i} f(x))\, ds + h \tfrac{\partial}{\partial x_i} f(x). \end{aligned}$$

Putting the last two relations together gives the following representation:

$$h^{-1} g(x,h) = \left(Hf(\vartheta) + R(x)\right)(x - \vartheta) + s(x,h) \tag{6.7}$$

with vector $s(x,h)$ satisfying $\|s(x,h)\| \le 0.5 \sqrt{d} K_2\, h$. Notice that s is also a measurable function $(x \in U_{\varepsilon/2}(\vartheta),\, h \in (0, \varepsilon/2))$.

While the last representation and Lemma 7.1(b) yield a rate of convergence, we will need a second representation to apply Lemma 7.1(a). We obtain

$$g^{(i)}(x,h) = \tfrac{1}{2}\,h\int_{-1}^{1}\Big(\tfrac{\partial}{\partial x_i}f(x+she_i) - (\nabla\tfrac{\partial}{\partial x_i}f(\vartheta))^*(x-\vartheta+she_i)\Big)\,ds \\ +h\,(\nabla\tfrac{\partial}{\partial x_i}f(\vartheta))^*(x-\vartheta).$$

For $r^{(i)}(x,h) := \frac{1}{2}\int_{-1}^{1}(\frac{\partial}{\partial x_i}f(x+she_i)-(\nabla\frac{\partial}{\partial x_i}f(\vartheta))^*(x-\vartheta+she_i))\,ds$, our assumptions imply

$$|r^{(i)}(x,h)| \\ \le \tfrac{1}{2}\int_{-1}^{1}\|\nabla f(x+she_i) - Hf(\vartheta)(x-\vartheta+she_i)\|\,ds \;\le\; \tfrac{1}{2}\int_{-1}^{1}K_1\,\|x-\vartheta+she_i\|^{1+\tau}\,ds \\ \le 2^{\tau-1}K_1\int_{-1}^{1}\big(\|x-\vartheta\|^{1+\tau}+|s|^{1+\tau}h^{1+\tau}\big)\,ds \;\le\; 2^{\tau}K_1\left(\|x-\vartheta\|^{1+\tau}+\tfrac{h^{1+\tau}}{2+\tau}\right).$$

As a consequence of the last two relations we obtain the following representation:

$$h^{-1}g(x,h) = Hf(\vartheta)(x-\vartheta) + r(x,h), \tag{6.8}$$

where the vector $r(x,h) = (r^{(i)}(x,h))$ is bounded by $\|r(x,h)\| \le 2^{\tau}\sqrt{d}K_1(\|x-\vartheta\|^{1+\tau} + h^{1+\tau}/(2+\tau))$. Again, r is a measurable function ($x \in U_{\varepsilon/2}(\vartheta)$, $h \in (0,\varepsilon/2)$).

Second step: *Rate of convergence for $X_n \to \vartheta$ and proof of asymptotic normality.* Let us define $U_n := X_n - \vartheta$, $V_n := c^{-1}W_n$, and $\Omega(n) := [\|U_n\| < \varepsilon/2 \text{ and } c_n < \varepsilon/2]$.

First we consider the case $p \ge 3$. We define $A_n := (Hf(\vartheta) + c_n^{p-2}P(c_n) + Q(X_n,c_n))1_{\Omega(n)}$ and $T_n := T(c_n)1_{\Omega(n)} - n^{1/2}c^{-1}g(X_n,c_n)1_{\Omega(n)^c}$. In the case $p=2$ we define $A_n := (Hf(\vartheta)+R(X_n))1_{\Omega(n)}$ and $T_n := -n^{1/4}s(X_n,c_n)1_{\Omega(n)} - n^{1/2}c^{-1}g(X_n,c_n)\cdot 1_{\Omega(n)^c}$. Regarding properties (6.1), (6.6), and (6.7), the so-defined quantities fulfill recursion (7.2) and satisfy $A_n \to A = Hf(\vartheta)$ a.s. and $T_n = O(1)$ almost in L^2. Condition (7.12) holds by assumption (D). Therefore, Lemma 7.1(b) yields $U_n = O(a_n^{1/2}n^{1/(2p)})$ almost in L^2.

For $p \ge 3$ the above quantities satisfy $T_n \to T$ a.s. and $\|A_n - A\| \le C_1\,n^{-(p-2)/(2p)} + C_2\,\|U_n\| + C_3 1_{\Omega(n)^c}$. In the case $p=2$ we have to alter the definition of A_n and T_n. Let $A_n := Hf(\vartheta)1_{\Omega(n)}$ and $T_n := -n^{1/4}r(X_n,c_n)1_{\Omega(n)} - n^{1/2}c^{-1}g(X_n,c_n)1_{\Omega(n)^c}$. Due to (6.8) recursion (7.2) holds. Note that $\|n^{1/4}r(X_n,c_n)1_{\Omega(n)}\| \le C_4\,n^{1/4}\|U_n\|^{1+\tau} + C_5\,n^{-\tau/4}$.

In both cases we get $T_n - T = o(1)$ almost in L^1 and $A_n - A = o(1/\sqrt{na_n})$ almost in L^2 , where we have used $2/p \le 1/2 + 1/(2p) < \alpha$ for $p \ge 3$. Thus the assertion follows from Lemma 7.1 (a). □

Proof of Proposition 4.1. We may assume $\vartheta = 0$ and $f(\vartheta) = 0$. Lipschitz continuity of ∇f implies

$$|f(x+h) - f(x-h) - \langle 2h, \nabla f(x)\rangle| = \left|\int_{-1}^{1}\langle h, \nabla f(x+sh) - \nabla f(x)\rangle\,ds\right| \\ \le \|h\|\int_{-1}^{1}K|s|\,\|h\|\,ds \;=\; K\|h\|^2, \tag{6.9}$$

where K is, here and in the following inequalities, a constant that may vary from formula to formula. The last inequality, together with

$$E\left(\tfrac{\Delta_n^{(k)}}{\Delta_n^{(l)}}\tfrac{\partial}{\partial x_k}f(X_n) \mid \mathcal{G}_n\right) = \delta_{kl}\tfrac{\partial}{\partial x_k}f(X_n) \text{ a.s.},$$

proves

$$\|E\,(Y_n \mid \mathcal{G}_n) - \nabla f(X_n)\| \le K c_n \quad \text{a.s.} \tag{6.10}$$

Due to (6.9), (6.10), and assumption (F), we obtain

$$E(\|Y_n\|^2 \mid \mathcal{G}_n) \le K c_n^2 + K\,\|\nabla f(X_n)\|^2 + K\,c_n^{-2}\,E(W_n^2 \mid \mathcal{G}_n) \quad \text{a.s.} \tag{6.11}$$

and

$$\langle \nabla f(X_n), E\,(Y_n \mid \mathcal{G}_n)\rangle \ge \|\nabla f(X_n)\|^2 - K c_n\,\|\nabla f(X_n)\| \quad \text{a.s.} \tag{6.12}$$

Lipschitz continuity of ∇f implies, as above,

$$f(X_{n+1}) \le f(X_n) - a_n\langle \nabla f(X_n), Y_n\rangle + K a_n^2\,\|Y_n\|^2.$$

Taking conditional expectations and using inequalities (6.11) and (6.12), we obtain

$$\begin{aligned}
&E\,(f(X_{n+1}) \mid \mathcal{G}_n)\\
&\le f(X_n) - a_n\left(\|\nabla f(X_n)\|^2 - K c_n\,\|\nabla f(X_n)\|\right)\\
&\qquad\qquad + K a_n^2\,\|\nabla f(X_n)\|^2 + K a_n^2/c_n^2\left(E(W_n^2 \mid \mathcal{G}_n) + 1\right)\\
&\le f(X_n) - a_n/2\,\left(\|\nabla f(X_n)\| - K c_n\right)^2 + K^2/2\,a_n c_n^2 + K a_n^2/c_n^2\left(E(W_n^2 \mid \mathcal{G}_n) + 1\right) \quad \text{a.s.}
\end{aligned}$$

for all n with $K a_n < 1/2$. Let $A_n := a_n/2\,(\|\nabla f(X_n)\| - K c_n)^2$ and $B_n := K^2/2\,a_n c_n^2 + K a_n^2/c_n^2\left(E(W_n^2 \mid \mathcal{G}_n) + 1\right)$. For n large enough

$$E\,(f(X_{n+1}) \mid \mathcal{G}_n) \le f(X_n) - A_n + B_n \quad \text{a.s.},$$

where $A_n \ge 0$, $B_n \ge 0$, and $\sum_{n=1}^{\infty} B_n < \infty$ a.s. On a set Ω_0 of measure 1 we have convergence of $f(X_n)$ and $\sum_{n=1}^{\infty} A_n$ according to a theorem of Robbins and Siegmund [20] for nonnegative almost-supermartingales.

Fix $\omega \in \Omega_0$ and denote $x_n := X_n(\omega)$. Then for almost all n the relation $f(x_n) \le \lambda := \lim f(x_n) + 1$ holds. Since $\{x : f(x) \le \lambda\}$ is bounded, (x_n) is bounded as well.

To prove (b) fix $\omega \in \Omega_0$ with $\sup_n \|x_n\| < \infty$. Select a subsequence $(x_{n'})$ with $\nabla f(x_{n'}) \to 0$. Then there exists a convergent subsequence $(x_{n''})$ of $(x_{n'})$. Since $\nabla f(x_{n''}) \to 0$ and ∇f is continuous, $(x_{n''})$ converges to zero. Hence $f(x_{n''}) \to 0$ and $f(x_n) \to 0$. Choose $\varepsilon > 0$ such that $\|x_n\| < 1/\varepsilon$ for all n. For n sufficiently large we have $f(x_n) < \inf\{f(x) : \varepsilon < \|x\| < 1/\varepsilon\}$. This proves $x_n \to 0$. □

Proof of Theorem 4.2. We will verify the assumptions of Lemma 7.1. For this purpose let us define $U_n := X_n - \vartheta$, $D_n := (2c_n)^{-1}\Delta_n^{-1}(f(X_n + c_n\Delta_n) - f(X_n - c_n\Delta_n))$, $V_{n,1} := (2c)^{-1}\Delta_n^{-1}W_n$, $\Omega(n) := [\|U_n\| < \varepsilon/2$ and $c_n < \varepsilon/(2d^{1/2}\alpha_1)]$, $V_{n,2} := -n^{-1/6}(D_n 1_{\Omega(n)} - E(D_n 1_{\Omega(n)} \mid \mathcal{G}_n))$, $V_n := V_{n,1} + V_{n,2}$, and $T := c^2 b$.

For $x, z \in \mathbb{R}^d$ and $h > 0$ with $x, x \pm hz, \vartheta \pm hz \in U_\varepsilon(\vartheta)$, we obtain by condition (H)

$$\begin{aligned}
&f(x + hz) - f(x - hz)\\
&\quad = f(\vartheta + hz) - f(\vartheta - hz) + \langle \nabla f(\vartheta + hz) - \nabla f(\vartheta - hz), x - \vartheta\rangle\\
&\qquad + (x - \vartheta)^* \int_0^1 (1-t)(Hf(\vartheta + t(x - \vartheta) + hz) - Hf(\vartheta + t(x - \vartheta) - hz))dt\,(x - \vartheta)
\end{aligned}$$

where

$$f(\vartheta + hz) - f(\vartheta - hz) = \frac{2h^3}{6}\sum_{i,j,k} \tfrac{\partial^3}{\partial x_i \partial x_j \partial x_k} f(\vartheta)\, z_i z_j z_k \;+\; o(h^3\|z\|^3),$$

$$\nabla f(\vartheta + hz) - \nabla f(\vartheta - hz) = 2h\,Hf(\vartheta)z \;+\; o(h^2\|z\|^2),$$

and

$$\left\| \int_0^1 (1-t)(Hf(\vartheta + t(x-\vartheta) + hz) - Hf(\vartheta + t(x-\vartheta) - hz))dt \right\| \leq Lh \, \|z\|.$$

This expansion, together with condition (E), leads to the following representation:

$$E\left(D_n 1_{\Omega(n)} \mid \mathcal{G}_n\right) = \left(Hf(\vartheta) + o(c_n) + O(\|U_n\|)\right) U_n 1_{\Omega(n)} - n^{-\frac{1}{3}} \left(T + o(1)\right) 1_{\Omega(n)}.$$

With $A_n := (Hf(\vartheta)+o(c_n)+O(\|U_n\|))1_{\Omega(n)}$, $T_n := (T+o(1))1_{\Omega(n)} - n^{1/3} D_n 1_{\Omega(n)^c}$, and the quantities defined at the beginning of the proof, recursion (1.1) can be rewritten in the form of recursion (7.2).

Let $B_{n,j}(t) := 1/\sqrt{n}\,(\sum_{i=1}^{\lfloor nt \rfloor} V_{i,j} + (nt - \lfloor nt \rfloor) V_{\lfloor nt \rfloor + 1, j})$, $t \in [0,1]$, $j \in \{1,2\}$. To show that $B_{n,1}$ converges in distribution to a Brownian motion B, and that $B_{n,2}$ converges to zero in probability, we apply an invariance principle for martingale difference sequences of Berger [1].

We first consider the case $j = 1$. Since

$$E(V_{n,1} \mid \mathcal{F}_n) = \tfrac{1}{2c} \Delta_n^{-1} E(W_n \mid \mathcal{F}_n) = 0 \text{ a.s.}$$

and $V_{n,1}$ is $\mathcal{F}_{n+1}$-measurable, $(V_{n,1})$ is a martingale difference sequence with respect to $(\mathcal{F}_{n+1})$. Similarly, we get from the assumptions,

$$\begin{aligned}
&\tfrac{1}{n} \sum_{i=1}^n E(V_{i,1} \otimes V_{i,1} \mid \mathcal{F}_i) \\
&\quad = \tfrac{1}{4c^2} \tfrac{1}{n} \sum_{i=1}^n \Delta_i^{-1} \otimes \Delta_i^{-1} E(W_i^2 \mid \mathcal{F}_i) \\
&\quad = \tfrac{\sigma^2}{4c^2} \tfrac{1}{n} \sum_{i=1}^n \Delta_i^{-1} \otimes \Delta_i^{-1} + \tfrac{1}{4c^2} \tfrac{1}{n} \sum_{i=1}^n \Delta_i^{-1} \otimes \Delta_i^{-1} \left(E(W_i^2 \mid \mathcal{F}_i) - \sigma^2\right) \\
&\quad \to \frac{\sigma^2 \rho^2}{4c^2} I \quad (n \to \infty) \text{ a.s.}
\end{aligned}$$

according to Kolmogorov's strong law of large numbers. Further, we obtain

$$\begin{aligned}
\tfrac{1}{n} \sum_{i=1}^n E(\|V_{i,1}\|^2 1_{[\|V_{i,1}\|^2 \geq ri]} \mid \mathcal{F}_i) &\leq \tfrac{1}{n} \sum_{i=1}^n E(\tfrac{d}{4c^2\alpha_0^2} W_i^2 1_{\left[W_i^2 \geq \frac{4c^2\alpha_0^2 r}{d} i\right]} \mid \mathcal{F}_i) \\
&\xrightarrow{P} 0 \quad (n \to \infty)
\end{aligned}$$

since $E(W_i^2 1_{[W_i^2 \geq \tilde{r} i]} \mid \mathcal{F}_i)$ is converging to zero a.s.

To get (7.4) for the sequence $(V_{n,1})$, we check that

$$\sup_n E\left(\|V_{n,1}\|^2 \mid \mathcal{F}_n\right) \leq \tfrac{1}{4c^2} \sup_n \|\Delta_n^{-1}\|^2 \sup_n E\left(W_n^2 \mid \mathcal{F}_n\right) < \infty \quad \text{a.s.},$$

which holds in view of the assumptions.

Likewise, we treat the case $j = 2$. Note that $V_{n,2}$ is $\mathcal{G}_{n+1}$-measurable, and $E(V_{n,2} \mid \mathcal{G}_n) = 0$ a.s. Condition (H) implies

$$\begin{aligned}
\|E\left(V_{n,2} \otimes V_{n,2} \mid \mathcal{G}_n\right)\| &\leq \tfrac{d}{4c^2\alpha_0^2} E\left(|f(X_n + c_n\Delta_n) - f(X_n - c_n\Delta_n)|^2 \; 1_{\Omega(n)} \mid \mathcal{G}_n\right) \\
&\leq (dL\alpha_1/\alpha_0)^2 n^{-2\gamma} \to 0 \quad (n \to \infty) \quad \text{a.s.}
\end{aligned}$$

This implies $\sup_n E(\|V_{n,2}\|^2 \mid \mathcal{G}_n) < \infty$ a.s., and thus validity of (7.4) for the sequence $(V_{n,2})$. Additionally

$$E\left(\|V_{n,2}\|^2 1_{[\|V_{n,2}\|^2 \geq rn]} \mid \mathcal{G}_n\right) \to 0 \quad (n \to \infty) \quad \text{a.s.}$$

Once more, the invariance principle in Berger [1] can be applied to prove the desired result.

Since $X_n \to \vartheta$ a.s., we obtain $A_n \to A$ and $T_n \to T = c^2 b$ a.s. The latter is sufficient for (7.5) and (7.9). Furthermore, the sequence (V_n) fulfills (7.10) and (7.11) with respect to $(\mathcal{G}_{n+1})$. Now Lemma 7.1(b) asserts $U_n = O(n^{(1/6)-(\alpha/2)})$ almost in L^2. To obtain $A_n - A = o(n^{(\alpha-1)/2})$ almost in L^2, one has to choose $\alpha > 2/3$. This completes the proof. □

Proof of relations (5.2)–(5.4). Since $\operatorname{spec}(2aA - \beta) \subset (0, \infty)$ we obtain

$$\begin{aligned} a\|(2aA-\beta)^{-1}b\| &= \tfrac{1}{2} \left\| \left(\tfrac{2a}{\beta}A - I\right)^{-1} \left(\tfrac{2a}{\beta}A - I\right) A^{-1}b + \left(\tfrac{2a}{\beta}A - I\right)^{-1} A^{-1}b \right\| \\ &= \tfrac{1}{2} \left\| \left(I + \left(\tfrac{2a}{\beta}A - I\right)^{-1}\right) A^{-1}b \right\| \\ &\geq \tfrac{1}{2} \min\left\{ \lambda \in \operatorname{spec}\left(I + \left(\tfrac{2a}{\beta}A - I\right)^{-1}\right)\right\} \|A^{-1}b\| \\ &\geq \tfrac{1}{2} \|A^{-1}b\| \end{aligned}$$

and, by Theorem 1 in Wei [26],

$$\operatorname{tr}\left(a^2(2aA-\beta)^{-1}S\right) \geq \operatorname{tr}\left(\beta A^{-1}SA^{-1}\right). \tag{6.13}$$

Noticing that $4\beta/(2-\beta) > 1$ and $(2-\beta)/2 > 1/2$, this yields

$$E(a,c) \geq c^{2p-2}\|A^{-1}b\|^2 + \tfrac{\beta}{c^2}\operatorname{tr}(A^{-1}SA^{-1}) \;>\; \left(\tfrac{2-\beta}{2}\right)^2 \widetilde{E}(0,c) \;>\; \tfrac{1}{4}\widetilde{E}(0,c).$$

The last relation of (5.2) is obvious.

The first two relations of (5.3) follow from (5.2). To prove the last one, we find for a given admissible a that

$$c_0(a) = \left(\frac{\operatorname{tr}((2aA-\beta)^{-1}S)}{4(p-1)\|(2aA-\beta)^{-1}b\|^2}\right)^{\frac{1}{2p}}$$

minimizes $E(a,c)$. Now observe that $E(a, c_0(a)) \to \infty$ as $a \to \infty$ or $a \searrow \beta/(2\lambda_0)$.

The first relation of (5.4) follows from (6.13), and the third one is as shown above. □

7. Appendix: A weak invariance principle for weighted means in stochastic approximation. For the following lemma, which is a consequence of Theorems 3.1 and 4.1 in Dippon and Renz [4], let (a_n) be a sequence decreasing to 0 with

$$na_n \nearrow \infty \quad (n \to \infty) \tag{7.1}$$

and satisfying the relation $a_n - a_{n+1} = o_n a_n^2$ with

$$\sum_{n=1}^{\infty} |o_n - o_{n+1}| < \infty.$$

Note that under (7.1) $o_n \le 1/(na_n) \to 0$ as $n \to \infty$. Examples for sequences having all these properties are (a/n^α) and $(a \log n/n)$ with $\alpha \in (0,1)$, $a > 0$.

LEMMA 7.1. *Let $\gamma \in [0, 1/2)$ and $\delta > -1/2 - \gamma$. For $\mathbb{R}^d$-valued random variables U_n, V_n, T_n and $\mathcal{L}(\mathbb{R}^d)$-valued random variables A_n, assume the following recursion:*

$$U_{n+1} = (I - a_n A_n)\, U_n + a_n n^\gamma \left(V_n + n^{-\frac{1}{2}} T_n \right). \tag{7.2}$$

Suppose that $A \in \mathcal{L}(\mathbb{R}^d)$ satisfies

$$\min\{\operatorname{re} \lambda : \lambda \in \operatorname{spec} A\} > 0.$$

(a) *Let $B_n(t) := n^{-1/2} \left\{ \sum_{i=1}^{\lfloor nt \rfloor} V_i + (nt - \lfloor nt \rfloor) V_{\lfloor nt \rfloor + 1} \right\}$, $t \in [0,1]$, $n \in \mathbb{N}$. Assume the existence of a centered Brownian motion B with covariance matrix S of $B(1)$ and with*

$$B_n \xrightarrow{\mathcal{D}} B \quad \text{in } C([0,1], \mathbb{R}^d) \qquad (n \to \infty), \tag{7.3}$$

$$B_n(1) = O(1) \quad \text{almost in } L^1. \tag{7.4}$$

If there exists $T \in \mathbb{R}^d$ such that

$$T_n - T = o(1) \quad \text{almost in } L^1, \tag{7.5}$$

$$U_n = O\left(n^\gamma \sqrt{a_n}\right) \quad \text{almost in } L^2, \tag{7.6}$$

$$A_n - A = o\left(1/\sqrt{na_n}\right) \quad \text{almost in } L^2, \tag{7.7}$$

then

$$n^{1/2-\gamma}\, t^{-\min\{1,\gamma+\delta\}} \tfrac{1+\delta}{n^{1+\delta}} \left(\sum_{k=1}^{\lfloor nt \rfloor} k^\delta U_k + (nt - \lfloor nt \rfloor)(\lfloor nt \rfloor + 1)^\delta U_{\lfloor nt \rfloor + 1} \right)$$

$$\xrightarrow{\mathcal{D}} G(t) := (1+\delta)\, t^{\max\{0,\gamma+\delta-1\}} A^{-1} \left(\int_{(0,1]} u^{\gamma+\delta}\, dB(tu) + \tfrac{t^{1/2}}{1/2+\gamma+\delta} T \right)$$

in $C([0,1], \mathbb{R}^d)$ for $n \to \infty$, where $G(1)$ is a Gaussian distributed random variable in $\mathbb{R}^d$ with expectation $2(1+\delta)/(1+2\gamma+2\delta) A^{-1} T$ and covariance matrix $(1+\delta)^2/(1+2\gamma+2\delta)$ $A^{-1} S A^{-1*}$.

(b) *Assume*

$$A_n \to A \ \text{ a.s. } (n \to \infty), \tag{7.8}$$

$$T_n = O(1) \ \text{ almost in } L^2, \tag{7.9}$$

and

$$E(V_n \mid \mathcal{F}_{n-1}) = 0 \ \text{ a.s.}, \tag{7.10}$$

$$\sup_n E(\|V_n\|^2 \mid \mathcal{F}_{n-1}) < \infty \ \text{ a.s. or } \ E\|V_n\|^2 = O(1), \tag{7.11}$$

where $(\mathcal{F}_n)$ is a filtration and (V_n) is adapted to $(\mathcal{F}_n)$, or, instead of (7.10) *and* (7.11), *alternatively:*

$$\underset{n \ge m}{\forall} \ \|EV_m \otimes V_n\| \le \varrho_{n-m} (E\|V_m\|^2 E\|V_n\|^2)^{\frac{1}{2}} \tag{7.12}$$

$$\text{with } \sum_{l=0}^{\infty} \varrho_l < \infty \ \text{ and } \ E\|V_n\|^2 = O(1).$$

Then condition (7.6) *holds.*

REMARK 7.2. (a) *For conditions implying* (7.3) *in case of a martingale difference sequence* (V_n)*, see Theorem* 5.1 *in Berger* [1].

(b) *Condition* (7.4) *is implied by* (7.10) *and* (7.11)*, or by* (7.12).

(c) *In applications,* (7.6) *can often be used to show* (7.7). *Usually,* (7.8) *follows from the consistency of the stochastic approximation procedure.*

Acknowledgments. Part of this work was done while the second author was visiting the Department of Statistics and Probability, Michigan State University, East Lansing, and the Department of Statistics, University of Illinois at Urbana-Champaign. He thanks the members of both departments for their hospitality.

The authors wish to thank Professor V. Fabian and the referees for helpful comments.

REFERENCES

[1] E. BERGER, *Asymptotic behaviour of a class of stochastic approximation procedures*, Probab. Theory Related Fields, 71 (1986), pp. 517–552.

[2] J.R. BLUM, *Multidimensional stochastic approximation methods*, Ann. Math. Statist., 25 (1954), pp. 737–744.

[3] H. CHEN, *Lower rate of convergence for locating a maximum of a function*, Ann. Statist., 16 (1988), pp. 1330–1334.

[4] J. DIPPON AND J. RENZ, *Weighted means of processes in stochastic approximation*, Math. Meth. Statist., 5 (1996), pp. 32–60.

[5] R.E. ERICKSON, V. FABIAN, AND J. MAŘIK, *An optimum design for estimating the first derivative*, Ann. Statist., 23 (1995), pp. 1234–1247.

[6] V. FABIAN, *Stochastic approximation of minima with improved asymptotic speed*, Ann. Math. Statist., 38 (1967), pp. 191–200.

[7] V. FABIAN, *On the choice of design in stochastic approximation*, Ann. Math. Statist., 39 (1968), pp. 457–465.

[8] V. FABIAN, *On asymptotic normality in stochastic approximation*, Ann. Math. Statist., 39 (1968), pp. 1327–1332.

[9] V. FABIAN, *Stochastic approximation*, in Optimizing Methods in Statistics, J.S. Rustagi, ed., Academic Press, New York, 1971, pp. 439–470.

[10] L. GYÖRFI AND H. WALK, *On the averaged stochastic approximation for linear regression*, SIAM J. Control Optim., 34 (1996), pp. 31–61.

[11] J. KIEFER AND J. WOLFOWITZ, *Stochastic estimation of the maximum of a regression function*, Ann. Math. Statist., 23 (1952), pp. 462–466.

[12] H.J. KUSHNER AND D.S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New York, 1978.

[13] H.J. KUSHNER AND J. YANG, *Stochastic approximation with averaging the iterates: Optimal asymptotic rate of convergence for general processes*, SIAM J. Control Optim., 31 (1993), pp. 1045–1062.

[14] A.V. NAZIN AND P.S. SHCHERBAKOV, *Method of averaging along trajectories in passive stochastic approximation*, Prob. Inform. Trans., 29 (1993), pp. 328–338.

[15] A. PECHTL, *Arithmetic means and invariance principles in stochastic approximation*, J. Theoret. Probab., 6 (1993), pp. 153–173.

[16] B.T. POLYAK, *New method of stochastic approximation type*, Automat. Remote Control, 51 (1990), pp. 937–946.

[17] B.T. POLYAK AND A.B. JUDITSKY, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855.

[18] B.T. POLYAK AND A.B. TSYBAKOV, *Optimal orders of accuracy for search algorithms of stochastic optimization*, Prob. Inform. Trans., 26 (1990), pp. 126–133.

[19] J. RENZ, *Konvergenzgeschwindigkeit und asymptotische Konfidenzintervalle in der stochastischen Approximation*, Ph.D. thesis, Universität Stuttgart, Germany, 1991.

[20] H. ROBBINS AND D. SIEGMUND, *A convergence theorem for nonnegative almost supermartingales and some applications*, in Optimizing Methods in Statistics, J.S. Rustagi, ed., Academic Press, New York, 1971, pp. 233–257.

[21] D. Ruppert, *Efficient Estimators from a Slowly Converging Robbins-Monro Process*, Tech. Rep. No. 781, School of Oper. Res. and Ind. Engrg., Cornell University, Ithaca, NY, 1988. (See also §2.8 of D. Ruppert, *Stochastic approximation*, in Handbook of Sequential Analysis, B.K. Ghosh and P.K. Sen, eds., Marcel Dekker, New York, 1991, pp. 503–529.)

[22] J.C. Spall, *A stochastic approximation algorithm for large-dimensional systems in the Kiefer-Wolfowitz setting*, in Proc. IEEE Conf. Decision Contr., 1988, pp. 1544–1548.

[23] J.C. Spall, *Multivariate stochastic approximation using a simultaneous perturbation gradient approximation*, IEEE Trans. Automat. Control, 37 (1992), pp. 332–341.

[24] H. Walk, *Limit behaviour of stochastic approximation processes*, Statist. Decisions, 6 (1988), pp. 109–128.

[25] H. Walk, *Foundations of stochastic approximation*, in Stochastic Approximation and Optimization of Random Systems, L. Ljung, G. Pflug, and H. Walk, eds., Birkhäuser, Basel, 1992, pp. 1–51.

[26] C.Z. Wei, *Multivariate adaptive stochastic approximation*, Ann. Statist., 15 (1987), pp. 1115–1130.

[27] G. Yin, *On extensions of Polyak's averaging approach to stochastic approximation*, Stochastics Stochastic Rep., 36 (1991), pp. 245–264.