The Annals of Statistics
2003, Vol. 31, No. 4, 1260–1281

ACCELERATED RANDOMIZED STOCHASTIC OPTIMIZATION

BY JÜRGEN DIPPON

Universität Stuttgart

We propose a general class of randomized gradient estimates to be employed in a recursive search for the minimum of an unknown multivariate regression function. Here only two observations per iteration step are used. Special cases include random direction stochastic approximation (Kushner and Clark), simultaneous perturbation stochastic approximation (Spall) and a special kernel based stochastic approximation method (Polyak and Tsybakov). If the unknown regression is p-smooth ($p \geq 2$) at the point of minimum, these methods achieve the optimal rate of convergence $O(n^{-(p-1)/(2p)})$. For both the classical stochastic approximation scheme (Kiefer and Wolfowitz) and the averaging scheme (Ruppert and Polyak) the related asymptotic distributions are computed.

1. Introduction. Consider two random variables X and Z with values in $\mathbb{R}^d$ and $\mathbb{R}$, respectively, that have unknown common distribution $P_{X,Z}$. Assume that the regression function $E(Z \mid X = \cdot) : \mathbb{R}^d \to \mathbb{R}$ exists and has a unique minimizer $\vartheta \in \mathbb{R}^d$:

$$\vartheta = \operatorname*{argmin}_{x \in \mathbb{R}^d} E(Z \mid X = x).$$

In this article we discuss a general method to estimate ϑ recursively. For this purpose it is assumed that the statistician can take a sample Z_x distributed according to $P_{Z|X=x}$ for any given $x \in \mathbb{R}^d$. This sample can be viewed as a noisy observation of $f(x) := E(Z \mid X = x)$ corrupted by some random error W_x:

$$Z_x = f(x) - W_x.$$

In 1952, Kiefer and Wolfowitz [10] solved this problem for $d = 1$. Subsequently Blum [2] treated the multivariate case by running the recursion

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

with Y_n taken to be an estimate of $\nabla f(X_n)$, the gradient of f at X_n. In contrast to the gradient method in numerical analysis, the step lengths $a_n > 0$ must converge to zero to obtain consistency. In this approach, Y_n was chosen to be an estimator of a d-dimensional two-sided difference quotient

$$Y_n = \frac{1}{2c_n}\{[f(X_n + c_n e_i) - W_{n,i,1}] - [f(X_n - c_n e_i) - W_{n,i,2}]\}_{i \in \{1,\ldots,d\}}$$

Received September 2000; revised July 2002.
AMS 2000 subject classification. 62L20.
Key words and phrases. Stochastic approximation, stochastic optimization, gradient estimation, randomization, asymptotic normality, optimal rates of convergence.

with positive step lengths c_n converging to zero and canonical basis $\{e_1, \ldots, e_d\}$ of $\mathbb{R}^d$. Observe that this gradient estimate requires $2d$ observations (in square brackets) of f at design points $X_n \pm c_n e_i$. Respective observation errors are denoted by $W_{n,i,j}$.

If f is three times differentiable, the iterates of the algorithm above tend to ϑ with rate $O_P(n^{-1/3})$, provided $a_n = a/n$ with the choice of a large enough $a > 0$. For regression functions f possessing derivatives of higher order p ($p \geq 3$ odd), Fabian [6] gave a modified gradient estimate based on $2d\lfloor p/2 \rfloor = d(p-1)$ observations which attains rate $O_P(n^{-(p-1)/(2p)})$.

More generally, Chen [3] and Polyak and Tsybakov [14] showed that for the class of p-times differentiable regression functions ($p \geq 2$), the optimal rate in a minimax sense is $O(n^{-(p-1)/(2p)})$.

To deal with high dimensional problems, Kushner and Clark [11], Polyak and Tsybakov [14], Spall [17] and others suggested randomized gradient estimators which need only two observations per step. The iterates of these methods still converge with rate $O_P(n^{-1/3})$, but at the expense of a possibly higher asymptotic squared error.

We show that these randomized gradient estimators can be considered as special cases of a *randomized kernel gradient estimate*

$$Y_n = \frac{1}{2c_n} K(\Delta_n)\{[f(X_n + c_n\Delta_n) - W_{n,1}] - [f(X_n - c_n\Delta_n) - W_{n,2}]\} \tag{2}$$

of ∇f at X_n. The kernel function $K : \mathbb{R}^d \to \mathbb{R}^d$ and (artificially generated) independently and identically distributed $\mathbb{R}^d$-valued random vectors Δ_n are chosen such that $E(K(\Delta_n) \otimes \Delta_n) = I_d$ to ensure that, conditionally on X_n, Y_n is an unbiased estimator of the gradient of f at X_n. It will be shown that there are many possibilities to design the gradient estimate such that the iterates attain the optimal rate of convergence. This approach unifies the investigation of different randomized gradient estimators and indicates how to find a whole bunch of new ones.

Proposition 1 formulates sufficient conditions for consistency. To be able to compare these estimators in terms of their asymptotics, we compute their weak limit distributions. This is done in a general frame in Section 4 for two important types of algorithms. In Theorem 2 the traditional scheme (1) is studied, but there is a crucial regularity assumption concerning the gain parameter a in $a_n \sim a/n$ which is connected with the unknown Hessian $Hf(\vartheta)$ of f at ϑ. Motivated by the Polyak–Ruppert idea for the Robbins–Monro stochastic approximation (see [13] and [16]), Dippon and Renz [5] suggested taking weighted averages of the iterates generated by Kiefer–Wolfowitz type algorithms. The iterates themselves are obtained with larger step lengths $a_n = an^{-\alpha}$ ($\alpha < 1$), but without any regularity condition on the Hessian $Hf(\vartheta)$. This approach is adopted in Theorem 3. Both theorems show how to obtain estimators with unbiased limit distribution.

In Section 5 it is shown how to obtain methods discussed in the previous literature as special cases of the kernel function approach which include the random direction method of Kushner and Clark [11] and the simultaneous perturbation method of Spall [18] (Section 5.1). These methods can be generalized in a natural way to higher order methods (Section 5.2). A special kernel method proposed by Polyak and Tsybakov [14] fits in this framework as well. In the case of random vectors Δ_n consisting of a support with finitely many directions only, the randomized kernel gradient estimator can be replaced by an average of estimated directional derivatives. This approach includes Fabian's higher order stochastic method [7] (Section 5.4).

Instead of using two observations per step it is even possible to work with only one observation (cf. [4], [14] and [19]). However, I will not pursue this idea here.

Although randomized and/or higher order methods can outperform the standard methods in terms of asymptotic behavior, it is expected that, for a small or moderate number n of iteration steps, the standard methods may be superior. Hence, for practical applications, one might use a two-stage scheme which switches from a standard method to a randomized or higher order method after a moderate number of iteration steps. For a discussion of nonasymptotic properties of randomized methods in comparison to their nonrandomized counterparts, see page 318ff in [12].

Two competing methods with the same rate of convergence can be compared in terms of their asymptotic mean squared error. I will not delve into these issues, but the asymptotic distributions obtained herein provide a basis for a discussion in this direction (compare pages 252–254 in [11], Section 5 in [5] and pages 337–338 in [18]).

2. Notation. In the Euclidean space $\mathbb{R}^d$ the unit vectors are denoted by $e_1, \dots, e_d$. The ith coordinate of a vector x is indicated by x_i, but if $x, x_1, x_2, \dots$ is a sequence of vectors (possibly without first element x), the ith coordinates are denoted by $x^{(i)}, x_1^{(i)}, x_2^{(i)}, \dots$, respectively. I_d is the identity matrix and $\langle \cdot, \cdot \rangle$ is the usual inner product. The tensor $x \otimes y : \mathbb{R}^d \to \mathbb{R}^d$ of two vectors $x, y \in \mathbb{R}^d$ is the linear mapping defined by $\langle y, \cdot \rangle x$, $Hf(\vartheta)$ is the Hessian of a function $f : \mathbb{R}^d \to \mathbb{R}$ at $\vartheta \in \mathbb{R}^d$ and $U_\varepsilon(\vartheta)$ is the open ball around ϑ with radius ε. Consider the multiindex $m = (m_1, \dots, m_d) \in \mathbb{N}_0^d$. Its length $|m|$ is just the sum $m_1 + \cdots + m_d$, and $m!$ is given by $m_1! \cdots m_d!$. The mth power of $x \in \mathbb{R}^d$ is declared by $x^m = x_1^{m_1} \times \cdots \times x_d^{m_d}$ under the convention $0^0 := 1$. The differential operator D^m is defined by

$$\frac{\partial^{m_1}}{(\partial x_1)^{m_1}} \cdots \frac{\partial^{m_d}}{(\partial x_d)^{m_d}}.$$

For $r \in \mathbb{R}$ we use $\lfloor r \rfloor$ and $\lceil r \rceil$ to denote the integer part of r and the least integer greater than or equal to r, respectively. For a logical expression A, the

value of the indicator function $\mathbb{1}_A$ equals 1 if A is true and 0 otherwise. If B is a set, then $\mathbb{1}_B(\omega)$ is a short form of $\mathbb{1}_{[\omega\in B]}$. The space $C([0,1],\mathbb{R}^d)$ of $\mathbb{R}^d$-valued continuous functions on $[0,1]$ is equipped with the maximum norm and $(\Omega, \mathcal{A}, P)$ is the underlying probability space. Let (X_n) be a sequence of $\mathbb{R}^d$-valued random variables. We write $X_n = O_P(r_n)$ if r_n is increasing to infinity and (X_n/r_n) is bounded in probability, that is, $\lim_{R\to\infty} \operatorname{limsup}_n P(\|X_n/r_n\| \geq R) = 0$. The sequence (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 a $\Omega_\varepsilon \in \mathcal{A}$ 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. Two sequences (a_n) and (b_n) are called asymptotically equivalent, $a_n \sim b_n$, if $\lim_{n\to\infty}(a_n/b_n) = 1$. A set $\{a,\dots,b\}$ of increasing integers is considered to be empty whenever $a > b$.

3. Consistency. Consider the following set of conditions.

(A) $f:\mathbb{R}^d \to \mathbb{R}$ is bounded from below and has a Lipschitz continuous gradient.

(B) $\Delta, \Delta_1, \Delta_2, \dots$ are independent and identically distributed random variables with values in $\mathbb{R}^d$ and finite second moments. The random variable Δ_n is assumed to be independent of $\{X_1,\dots,X_n,\Delta_1,\dots,\Delta_{n-1}\}$.

(C) The kernel $K:\mathbb{R}^d \to \mathbb{R}^d$ is a measurable function with $E(K(\Delta)\otimes\Delta) = I_d$ and $E(\|\Delta\|^4\|K(\Delta)\|^2) < \infty$.

(D) The difference $W_n = (W_{n,1} - W_{n,2})/2$ of the observation errors (divided by 2) satisfies $E(K(\Delta_n)W_n \mid \mathcal{G}_n) = 0$ and $\sup_n E(\|K(\Delta_n)\|^2 W_n^2 \mid \mathcal{G}_n) < \infty$ a.s., where the σ-field $\mathcal{G}_n$ is generated by $\{X_1,\dots,X_n,\Delta_1,\dots,\Delta_{n-1}\}$.

The consistency result below is an extension of Proposition 4.1 in [5]. It is related to Blum's result [2] on multivariate Kiefer–Wolfowitz procedures. Using different techniques in the proof, the conditions on f can be relaxed.

PROPOSITION 1. *Choose sequences* (a_n) *and* (c_n) *satisfying* $a_n \geq 0$, $c_n > 0$, $a_n \to 0$, $\sum_{n=1}^\infty a_n = \infty$, $\sum_{n=1}^\infty a_n c_n^2 < \infty$ *and* $\sum_{n=1}^\infty a_n^2/c_n^2 < \infty$. *For recursion* (1) *with gradient estimate* (2) *assume that Conditions* (A)–(D) *hold.*

(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.*

If there are constants $a, c > 0$, $\alpha \in (\max\{\gamma + 1/2, 1-2\gamma\}, 1]$ and $\gamma \in (0, 1/2)$, then sequences (a_n) and (c_n) with $a_n \sim a/n^\alpha$ or $a_n \sim (a\ln n)/n$ and $c_n \sim cn^{-\gamma}$ fulfill the assumptions of the proposition.

4. Asymptotic normality. We will use the following assumptions:

($\widetilde{\text{A}}$) $f:\mathbb{R}^d \to \mathbb{R}$ is a measurable function and $\nabla f(\vartheta)$ exists and equals 0. For a given real number $p \geq 2$ assume an $\varepsilon > 0$ such that all derivatives of f up to order $\lfloor p \rfloor - 1$ exist on $U_\varepsilon(\vartheta)$, all $\lfloor p \rfloor$th derivatives of f at ϑ exist, f is *p-smooth at ϑ* in the sense that

$$\left\| \nabla f(x) - \sum_{|m| \leq \lfloor p \rfloor - 1} \frac{1}{m!} D^m \nabla f(\vartheta)(x-\vartheta)^m \right\| = o\big(\|x-\vartheta\|^{p-1}\big)$$

and, if $p \in [3, \infty)$, the Hessian of f is locally Lipschitz around ϑ, that is,

$$\underset{x,y \in U_\varepsilon(\vartheta)}{\forall} \ \underset{L<\infty}{\exists} \qquad \|Hf(x) - Hf(y)\| \leq L\|x-y\|.$$

($\widetilde{\text{B}}$) Condition (B) holds. Additionally, the random variables Δ_i are bounded a.s.

($\widetilde{\text{C}}$) Condition (C) holds. Additionally, if $p \geq 3$, for every multiindex $m \in \mathbb{N}_0^d$ with odd length $|m| \in \{3, \dots, \lfloor p \rfloor\}$, $E(\Delta^m K(\Delta)) = 0$ holds.

Set $V_n := K(\Delta_n)W_n$ and define linearly interpolated stochastic processes B_n by $B_n(t) := 1/\sqrt{n}\{\sum_{i=1}^{\lfloor nt \rfloor} V_i + (nt - \lfloor nt \rfloor)V_{\lfloor nt \rfloor + 1}\}$, $t \in [0,1]$.

($\widetilde{\text{D}}$) Condition (D) holds. Additionally, there exist a Brownian motion B, some $\sigma > 0$ with $B_n \to_{\mathcal{D}} B$ in $C([0,1], \mathbb{R}^d)$ as $n \to \infty$ and covariance matrix $S := \sigma^2 E(K(\Delta) \otimes K(\Delta))$ of $B(1)$.

($\widetilde{\text{E}}$) $B_n(1) = O(1)$ almost in $L^1(P)$ (as defined at the end of Section 2).

Furthermore, we have to impose a restriction on the eigenvalues of the Hessian of $f(\vartheta)$:

$$\lambda_0 := \inf\{\operatorname{re}\lambda : \lambda \text{ eigenvalue of } Hf(\vartheta)\}. \tag{3}$$

[The eigenvalues of $Hf(\vartheta)$ need not necessarily be real under the conditions stated on f.]

THEOREM 2. *Assume Conditions* ($\widetilde{\text{A}}$)–($\widetilde{\text{D}}$) *subject to some* $p \geq 2$. *Choose* $c_n = cn^{-\gamma}$ *with* $\gamma = \frac{1}{2p}$ *and choose* $a_n = a/n$ *such that* $\lambda_0 > \frac{1}{2a}(1-2\gamma)$. *If* $X_n \to \vartheta$ *a.s., then*

$$n^{(1-1/p)/2}(X_n - \vartheta) \xrightarrow{\mathcal{D}} N(0, \Sigma)$$

with $\Sigma = \frac{1}{2}(aHf(\vartheta) - \frac{1}{2}(1-2\gamma)I_d)^{-1}(a^2/c^2)S$.

THEOREM 3. *Assume Conditions* ($\widetilde{\text{A}}$)–($\widetilde{\text{E}}$) *subject to some* $p > 2$. *Suppose* $\lambda_0 > 0$. *Choose* $c_n = cn^{-\gamma}$ *with* $\gamma = \frac{1}{2p}$ *and* $a_n = a/n^\alpha$ *with* $\alpha \in (\max\{\frac{1}{2} + \gamma, 4\gamma\}, 1)$ *or* $a_n = a \ln n/n$. *If* $X_n \to \vartheta$ *a.s., then for all* $\delta > -(\frac{1}{2} + \gamma)$,

$$\widetilde{X}_{n,\delta} := \frac{1+\delta}{n^{1+\delta}} \sum_{i=1}^{n} i^\delta X_i$$

achieves

$$n^{(1-1/p)/2}(\widetilde{X}_{n,\delta} - \vartheta) \xrightarrow{\mathcal{D}} N(0, \widetilde{\Sigma})$$

with $\widetilde{\Sigma} = c^{-2}((1+\delta)^2/(1+2\gamma+2\delta))A^{-1}SA^{-1}$ *and* $A = Hf(\vartheta)$.

Consider some $p \geq 3$ whose integer part $\lfloor p \rfloor$ is odd. If Condition ($\widetilde{\mathrm{C}}$) is relaxed to

($\widetilde{\mathrm{C}}'$) Condition (C) holds. For every multiindex $m \in \mathbb{N}_0^d$ with odd length $|m| \in \{3, \ldots, \lfloor p \rfloor - 2\}$, $E(\Delta^m K(\Delta)) = 0$ holds,

which coincides with the condition usually imposed in the literature for various designs (Δ, K) (see Section 5), and if the quantity

$$T := -c^{\lfloor p \rfloor - 1} \sum_{|m| = \lfloor p \rfloor} \frac{1}{m!} D^m f(\vartheta) E\big(\Delta^m K(\Delta)\big)$$

does not vanish [which cannot happen under Condition ($\widetilde{\mathrm{C}}$)], then the asymptotic distribution will no longer be unbiased. If additionally $p > \lfloor p \rfloor$, the optimal rate of convergence will not be attained. These deficiencies occur for several traditional methods such as those discussed in Section 5.1.

COROLLARY 4. *Suppose* $p \geq 3$ *with* $\lfloor p \rfloor$ *odd. Assume Conditions* ($\widetilde{\mathrm{A}}$)–($\widetilde{\mathrm{D}}$) *with* ($\widetilde{\mathrm{C}}$) *replaced by* ($\widetilde{\mathrm{C}}'$). *Choose* $c_n = cn^{-\gamma}$ *with* $\gamma = \frac{1}{2\lfloor p \rfloor}$ *and* $a_n = a/n$ *such that* $\lambda_0 > \frac{1}{2a}(1-2\gamma)$. *If* $X_n \to \vartheta$ *a.s., then*

$$n^{(1-1/\lfloor p \rfloor)/2}(X_n - \vartheta) \xrightarrow{\mathcal{D}} N(\mu, \Sigma)$$

with $\mu = (aHf(\vartheta) - \frac{1}{2}(1-2\gamma)I_d)aT$ *and* $\Sigma = \frac{1}{2}(aHf(\vartheta) - \frac{1}{2}(1-2\gamma)I_d)^{-1} \times (a^2/c^2)S$.

COROLLARY 5. *Suppose* $p \geq 3$ *with* $\lfloor p \rfloor$ *odd. Assume Conditions* ($\widetilde{\mathrm{A}}$)–($\widetilde{\mathrm{E}}$) *with* ($\widetilde{\mathrm{C}}$) *replaced by* ($\widetilde{\mathrm{C}}'$) *and* $\lambda_0 > 0$. *Choose* $c_n = cn^{-\gamma}$ *with* $\gamma = \frac{1}{2\lfloor p \rfloor}$ *and* $a_n = a/n^\alpha$ *with* $\alpha \in (\max\{\frac{1}{2} + \gamma, 4\gamma - \frac{p - \lfloor p \rfloor}{\lfloor p \rfloor}\}, 1)$ *or* $a_n = a \ln n/n$. *If* $X_n \to \vartheta$ *a.s., then for all* $\delta > -(\frac{1}{2} + \gamma)$,

$$\widetilde{X}_{n,\delta} := \frac{1+\delta}{n^{1+\delta}} \sum_{i=1}^{n} i^\delta X_i$$

achieves

$$n^{(1-1/\lfloor p \rfloor)/2}(\widetilde{X}_{n,\delta} - \vartheta) \xrightarrow{\mathcal{D}} N(\widetilde{\mu}, \widetilde{\Sigma})$$

with $\widetilde{\mu} = (2(1+\delta)/(1+2\gamma+2\delta))A^{-1}T$, $\widetilde{\Sigma} = c^{-2}((1+\delta)^2/(1+2\gamma+2\delta))A^{-1}SA^{-1}$ *and* $A = Hf(\vartheta)$.

REMARK 6. (a) Observe that the spectral condition on the Hessian in Theorem 3 is much weaker than in Theorem 2. However, in Theorem 3 the choice $p = 2$ is excluded due to increased demand on the smoothness on f at ϑ necessary in the averaging scheme. In Theorem 2 and Corollary 4 step lengths a_n converging more slowly to zero than a/n are possible as well, but then the rate of convergence of $(X_n - \vartheta)$ will be slower [7].

(b) Our notion of p-smoothness as given in Condition $(\widetilde{\mathrm{A}})$ is slightly stronger than the Hölder condition of order p as given in [14].

(c) Conditions $(\widetilde{\mathrm{D}})$ and $(\widetilde{\mathrm{E}})$ are implied by the following two conditions:

$(\widehat{\mathrm{D}})$ Conditions (B), (C) and (D) hold, $E(W_n \mid \mathcal{F}_n) = 0$ a.s., $E(W_n^2 \mid \mathcal{F}_n) \to \sigma^2$ a.s., $\sup_n E(|W_n|^{2+\varepsilon} \mid \mathcal{F}_n) < \infty$ a.s. and $E\|K(\Delta)\|^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$, where the σ-field $\mathcal{F}_n$ is generated by $\{X_1, \ldots, X_n, \Delta_1, \ldots, \Delta_n\}$.

$(\widehat{\mathrm{E}})$ Assume $\frac{1}{n}\sum_{i=1}^n E(\|K(\Delta_i)\|^2 W_i^2) = O(1)$ [which holds, e.g., if the kernel K is bounded, or if Δ_i and W_i are independent, or if $E(K(\Delta)^4)$ and $\sup_i E(W_i^4)$ are bounded].

(d) If in the recursions considered in Theorem 2 and Corollary 4 the gradient estimates Y_n are premultiplied by a matrix M [such that $MHf(\vartheta)$ instead of $Hf(\vartheta)$ satisfies (3)], the trace of the resulting covariance matrix of the limit distribution attains its minimum for $M = (Hf(\vartheta))^{-1}$ (which exists by assumption). In this case the optimal choice for a is $a = (p-1)/p$. Since in practice $Hf(\vartheta)$ is usually unknown, certain adaptive methods seek to estimate $(Hf(\vartheta))^{-1}$ consistently by a sequence of random matrices M_n built up from information gained up to the nth loop of the iteration. For Kiefer–Wolfowitz type stochastic approximation methods this was suggested by Fabian [8]. Under additional assumptions, results similar to Theorem 2 and Corollary 4 can be achieved for adaptive variants. In these cases the covariance matrix of the limit distribution will coincide with those of Theorem 3 (or Corollary 5, respectively) provided the weighting exponent $\delta = -1/p$ is chosen. Now consider the case of a nonvanishing bias of the asymptotic distribution as in Corollaries 4 and 5. Construction of a procedure which minimizes the second moment of the asymptotic distribution is a difficult problem (see [5] for a further discussion of this issue).

(e) The sequences (a_n) and (c_n) in Theorems 2 and 3 and Corollaries 4 and 5 may be chosen more generally. To keep the formulation of the results and their proofs lucent, we abstain from treating generalizations in this direction. For tools to prove such extensions, see [5] and [20].

5. Examples. In the following examples we discuss several choices for the pair (Δ, K), check the related Conditions (C), $(\widetilde{\mathrm{C}})$ and $(\widetilde{\mathrm{C}}')$, and compute in each case the ingredients of the quantities S and T.

5.1. *Methods for p-smooth regression functions* ($p \in [2,3]$). In this subsection we assume that the regression function f is p-smooth in the sense of Condition ($\widetilde{\mathrm{A}}$) with some $p \in [2,3]$.

5.1.1. *Random finite difference stochastic approximation.* In the classical Kiefer–Wolfowitz stochastic approximation scheme, one takes two-sided differences along each direction of the coordinate axes. A randomized version of this employs the two-sided difference along the direction of a randomly chosen coordinate axis. This means

$$\Delta \sim U(\{e_1, \dots, e_d\}) \quad \text{and} \quad K(\Delta) = d\Delta. \tag{4}$$

Then $E(K(\Delta) \otimes \Delta) = I_d$ and $E(K(\Delta) \otimes K(\Delta)) = d I_d$. If $m \in \mathbb{N}_0^d$ with $|m| = 3$, then $E(\Delta^m K(\Delta)) = e_i$ or $= 0$, whenever there is an i with $m_i = 3$ or not, respectively. Hence $S = d\sigma^2 I_d$ and, if $p = 3$,

$$T = -\frac{c^2}{6}\left(\frac{\partial^3}{(\partial x_i)^3} f(\vartheta)\right)_{i=1,\dots,d}.$$

In comparison, for the classical multivariate Kiefer–Wolfowitz procedure we have the same T, but a different $S = \sigma^2 I_d$.

5.1.2. *Random direction stochastic approximation.* Kushner and Clark ([11], page 58ff) described a method which estimates the gradient of f at X_n by estimating the directional derivative along a randomly chosen direction of the unit sphere S^d. This condition can be relaxed as is done in [12], page 315ff, by choosing a distribution F_{RD} which is concentrated on S^d, has identically distributed projections on coordinate axes and is symmetrically distributed with respect to reflection about each axis.

Consider the design

$$\Delta \sim F_{\mathrm{RD}} \quad \text{and} \quad K(\Delta) = d\Delta. \tag{5}$$

We mention three examples for distributions of Δ: the uniform distribution on the sphere S^d, the uniform distribution on the finite set $\{x \in S^d : |x^{(i)}|^2 = 1/d \text{ for each } i \in \{1, \dots, d\}\}$ and the uniform distribution on the finite set $\{\pm e_i : i \in \{1, \dots, d\}\}$. In [12], the methods corresponding to the first two choices were called the *spherical method* and the *Bernoulli method*, respectively. The third choice is just the random finite difference method as given in display (4).

We have $E(K(\Delta) \otimes \Delta) = d E(\Delta \otimes \Delta) = I_d$ and $E(K(\Delta) \otimes K(\Delta)) = d I_d$. Set $\tau_1 := E(\Delta^{(1)})^4$ and $\tau_2 := E(\Delta^{(1)}\Delta^{(2)})^2 = (1 - d\tau_1)/(d(d-1))$. Observe that for $m \in \mathbb{N}_0^d$ with $|m| = 3$,

$$E\big(\Delta^m K(\Delta)\big) = \begin{cases} d\tau_1 e_i, & \text{if there is an } i \text{ with } m_i = 3, \\ d\tau_2 e_i, & \text{if there are } i \neq j \text{ with } m_i = 1 \text{ and } m_j = 2, \\ 0, & \text{otherwise.} \end{cases}$$

Then $S = d\sigma^2 I_d$ and, if $p = 3$,

$$T = -\frac{c^2 d}{6}\left(\tau_1 \frac{\partial^3}{(\partial x_i)^3} f(\vartheta) + 3\tau_2 \sum_{j=1, j\neq i}^{d} \frac{\partial^3}{\partial x_i (\partial x_j)^2} f(\vartheta)\right)_{i=1,\dots,d}.$$

5.1.3. *Simultaneous perturbation stochastic approximation.* Spall [17, 18] introduced another scheme to estimate the gradient with two observations per step. The difference of two observations taken at $X_n + c_n\Delta_n$ and $X_n - c_n\Delta_n$ with a random direction Δ_n is simultaneously used for each component of the gradient estimate.

Choose a distribution F_{SP} on $\mathbb{R}^d$ which is the d-fold tensor product of a symmetrical distribution concentrated on $\mathbb{R} \setminus \{0\}$. A possible but simple choice for F_{SP} is the uniform distribution concentrated on the vertices of the cube $[-1, 1]^d$. For $\delta = (\delta^{(1)}, \dots, \delta^{(d)})$, define $\delta^{-1} = (1/\delta^{(1)}, \dots, 1/\delta^{(d)})$. Now consider the design given by

$$\Delta \sim F_{\mathrm{SP}} \quad \text{and} \quad K(\Delta) = \Delta^{-1}.$$

Suppose that both $\xi^2 := E((\Delta^{(i)})^2)$ and $\rho^2 := E((\Delta^{(i)})^{-2})$ are finite. We find $E(K(\Delta) \otimes \Delta) = E(\Delta^{-1} \otimes \Delta) = I_d$ and $E(K(\Delta) \otimes K(\Delta)) = E(\Delta^{-1} \otimes \Delta^{-1}) = \rho^2 I_d$.

Let $m \in \mathbb{N}_0^d$ with $|m| = 3$. If Δ has a finite third moment, then $E(\Delta^m K(\Delta)) = \xi^2 e_i$ or $= 0$ whenever there are $i \neq j$ with $m_i = 1$ and $m_j = 2$ or not, respectively. Hence $S = \sigma^2\rho^2 I_d$ and, if $p = 3$,

$$T = -\frac{c^2\xi^2}{6}\left(\frac{\partial^3}{(\partial x_i)^3} f(\vartheta) + 3 \sum_{j=1, j\neq i}^{d} \frac{\partial^3}{\partial x_i (\partial x_j)^2} f(\vartheta)\right)_{i=1,\dots,d}.$$

5.2. *Higher order methods for p-smooth regression functions* ($p \geq 2$). We modify the designs of Section 5.1 to obtain a convergence rate faster than $n^{-1/3}$ whenever $p > 3$ while still requiring only two observations per step. For $p \geq 2$, take $q = \lceil p/2 \rceil$. If $p \geq 3$ is an odd number, the choice $q = \lfloor p/2 \rfloor$ is possible as well, but generally this will lead to a biased asymptotic distribution. Choose numbers $0 < u_1 < \cdots < u_q \leq 1$ and compute $(v_1, \dots, v_q)^t = M^{-1} e_1$, where M is the Vandermonde matrix $(u_j^{2l-1})_{l,\, j \in \{1,\dots,q\}}$. Then $\sum_{j=1}^{q} u_j^{2l-1} v_j = \mathbb{1}_{[l=1]}$ for $l = \{1, \dots, q\}$.

5.2.1. *Higher order random finite difference stochastic approximation.* For regression functions with p-times differentiability at ϑ ($p \geq 3$ odd), Fabian [7] described a method that required $2d\lfloor p/2 \rfloor$ observations per step. In [5] this idea was generalized to p-smooth functions ($p \in \{2, 3, \dots\}$) using $2d\lfloor p/2 \rfloor$ or $2d\lceil p/2 \rceil$ observations per step. Now we give a randomized version which uses two observations per step only and assumes a p-smooth regression function ($p \geq 2$).

Consider the design

$$\Delta = \Phi \cdot \Psi \qquad \text{with independent } \Phi \sim U(\{u_j : 1 \le j \le q\}) \text{ and } \Psi \sim U(\{e_i : 1 \le i \le d\}),$$

$$K(\Delta) = qdv_j\Psi \qquad \text{whenever } \Delta = u_j\Psi.$$

It turns out that

$$E\big(K(\Delta)\otimes\Delta\big) = \frac{1}{qd}\sum_{i=1}^{d}\sum_{j=1}^{q} qdu_jv_j(e_i\otimes e_i) = I_d,$$

$$S = \sigma^2 E\big(K(\Delta)\otimes K(\Delta)\big) = \sigma^2\frac{1}{qd}\sum_{i=1}^{d}\sum_{j=1}^{q}(qd)^2v_j^2(e_i\otimes e_i) = \sigma^2 qd\sum_{j=1}^{q}v_j^2 I_d,$$

since $\sum_{i=1}^{d} e_i\otimes e_i = I_d$.

Let $m\in\mathbb{N}_0^d$. Assume that Δ has the realization u_je_i. Then $\Delta^m = u_j^{|m|}$ whenever $m^{(i)} = |m|$ and $\Delta^m = 0$ otherwise. Hence

$$E\big(\Delta^m K(\Delta)\big) = \begin{cases} \sum_{j=1}^{q} u_j^{|m|}v_je_i, & \text{if there is an } i \text{ with } m^{(i)} = |m|,\\ 0, & \text{otherwise.}\end{cases}$$

If $\lfloor p\rfloor$ is odd and $|m|\in\{3,5,\ldots,\lfloor p\rfloor-2\}$ or if $\lfloor p\rfloor$ is even and $|m|\in\{3,5,\ldots,\lfloor p\rfloor-1\}$, then, due to $\sum_{j=1}^{q} u_j^{|m|}v_j = \mathbb{1}_{[|m|=1]}$, we have $E(\Delta^m K(\Delta)) = 0$. Now consider $\lfloor p\rfloor$ odd and $|m| = \lfloor p\rfloor$. If $q = \lceil p/2\rceil$, $E(\Delta^m K(\Delta)) = 0$ holds as well; thus Condition $(\widetilde{\mathrm{C}})$ is satisfied. However, if we choose $q = \lfloor p/2\rfloor$, we have

$$E\big(\Delta^m K(\Delta)\big) = \begin{cases} \left(\sum_{j=1}^{q} u_j^{\lfloor p\rfloor}v_j\right)e_i, & \text{if there is an } i \text{ with } m^{(i)} = |m|,\\ 0, & \text{otherwise.}\end{cases}$$

This leads to

$$T = -c^{\lfloor p\rfloor-1}\sum_{i=1}^{d}\frac{1}{\lfloor p\rfloor!}\frac{\partial^{\lfloor p\rfloor}}{\partial\vartheta_i^{\lfloor p\rfloor}}f(\vartheta)\left(\sum_{j=1}^{q}u_j^{\lfloor p\rfloor}v_j\right)e_i.$$

Apparently, instead of Condition $(\widetilde{\mathrm{C}})$, the weaker Condition $(\widetilde{\mathrm{C}}')$ is satisfied.

5.2.2. *Higher order random direction stochastic approximation.* Using the notation of Section 5.1.2 and

$$\Delta = \Phi \cdot \Psi \qquad \text{with independent r.v.'s } \Phi \sim U(\{u_j : 1 \le j \le q\}) \text{ and } \Psi \sim F_{\mathrm{RD}},$$

$$K(\Delta) = qdv_j\Psi \qquad \text{whenever } \Delta = u_j\Psi,$$

we obtain

$$E\big(K(\Delta)\otimes\Delta\big)=\frac{1}{q}\sum_{j=1}^{q}qu_jv_jdE(\Psi\otimes\Psi)=I_d,$$

$$E\big(K(\Delta)\otimes K(\Delta)\big)=\frac{1}{q}\sum_{j=1}^{q}q^2d^2v_j^2E(\Psi\otimes\Psi)=qd\sum_{j=1}^{q}v_j^2I_d$$

and

$$E\big(\Delta^mK(\Delta)\big)=\frac{1}{q}\sum_{j=1}^{q}qu_j^{|m|}v_jE(\Psi^m\Psi)=\left(\sum_{j=1}^{q}u_j^{|m|}v_j\right)E(\Psi^m\Psi),$$

which equals zero for every multiindex $m\in\mathbb{N}_0^d$ with odd length $|m|\in\{3,\dots,\lfloor p\rfloor-1\}$. As in Section 5.2.1, for $q=\lceil p/2\rceil$ Condition $(\widetilde{\mathrm{C}})$ is satisfied.

For $p\geq 3$ with $\lfloor p\rfloor$ odd and $q=\lfloor p/2\rfloor$ one obtains

$$T=-c^{\lfloor p\rfloor-1}\sum_{|m|=\lfloor p\rfloor}\frac{1}{m!}D^mf(\vartheta)E\left(\sum_{j=1}^{q}u_j^{|m|}v_j\right)E(\Psi^m\Psi),$$

which is to be used in connection with Condition $(\widetilde{\mathrm{C}}')$.

5.2.3. *Higher order simultaneous perturbation finite difference stochastic approximation.* With the notation of Section 5.1.3, choose

$$\Delta=\Phi\cdot\Psi \qquad \text{with independent r.v.'s } \Phi\sim U(\{u_j:1\leq j\leq q\}) \text{ and } \Psi\sim F_{\mathrm{SP}},$$

$$K(\Delta)=qv_j\Psi^{-1} \qquad \text{whenever } \Delta=u_j\Psi$$

and assume finite moments of Ψ up to order $\lfloor p\rfloor$. Then we obtain

$$E\big(K(\Delta)\otimes\Delta\big)=\frac{1}{q}\sum_{j=1}^{q}qu_jv_jE(\Psi\otimes\Psi^{-1})=I_d,$$

$$E\big(K(\Delta)\otimes K(\Delta)\big)=\frac{1}{q}\sum_{j=1}^{q}q^2v_j^2E(\Psi\otimes\Psi^{-1})=q\rho^2\sum_{j=1}^{q}v_j^2I_d$$

and

$$E\big(\Delta^mK(\Delta)\big)=\frac{1}{q}\sum_{j=1}^{q}qu_j^{|m|}v_jE(\Psi^m\Psi^{-1})=\left(\sum_{j=1}^{q}u_j^{|m|}v_j\right)E(\Psi^m\Psi^{-1}).$$

Again, for the choice $q=\lceil p/2\rceil$ Condition $(\widetilde{\mathrm{C}})$ is satisfied. For $p\geq 3$ with $\lfloor p\rfloor$ odd

and $q = \lfloor p/2 \rfloor$, Condition $(\widetilde{C}')$ applies with

$$T = -c^{\lfloor p \rfloor - 1} \sum_{|m| = \lfloor p \rfloor} \frac{1}{m!} D^m f(\vartheta) E\left(\sum_{j=1}^{q} u_j^{|m|} v_j \right) E(\Psi^m \Psi^{-1}).$$

5.3. *The kernel method of Polyak and Tsybakov.* An interesting design investigated by Polyak and Tsybakov [14] is

$$\Delta \sim U([-0.5, 0.5]^d) \quad \text{and} \quad K(\Delta) = \big(K_1(\Delta), \dots, K_d(\Delta)\big),$$

where the functions $K_i : \mathbb{R}^d \to \mathbb{R}$ are chosen such that

$$K_j(\delta) = k_0(\delta^{(j)}) \prod_{m=1, m \neq j}^{d} \widetilde{k}(\delta^{(m)}), \qquad \delta \in \mathbb{R}^d,$$

with measurable and bounded functions $k_0, \widetilde{k} : \mathbb{R} \to \mathbb{R}$, the supports of which are contained in $[-0.5, 0.5]$, and

$$\underset{i \in \{0, \dots, \lfloor p \rfloor\}}{\forall} \qquad \int u^i k_0(u)\, du = \mathbb{1}_{[i=1]}$$

and

$$\underset{i \in \{0, \dots, \lfloor p \rfloor - 1\}}{\forall} \qquad \int u^i \widetilde{k}(u)\, du = \mathbb{1}_{[i=0]}.$$

Polyak and Tsybakov described a method for constructing the kernels k_0 and $\widetilde{k}$ by using orthogonal Legendre polynomials on $[-0.5, 0.5]$.

We observe that

$$\begin{aligned}
&E\big(\Delta^{(i)} K_j(\Delta)\big) \\
&\quad = E\left(\Delta^{(i)} k_0(\Delta^{(j)}) \prod_{m=1, m \neq j}^{d} \widetilde{k}(\Delta^{(m)}) \right) \\
&\quad = \begin{cases}
E\big(\Delta^{(i)} \widetilde{k}(\Delta^{(i)})\big) E\big(k_0(\Delta^{(j)})\big) \displaystyle\prod_{m=1, m \neq j, m \neq i}^{d} E\big(\widetilde{k}(\Delta^{(m)})\big), & \text{if } i \neq j, \\
E\big(\Delta^{(i)} k_0(\Delta^{(i)})\big) \displaystyle\prod_{m=1, m \neq i}^{d} E\big(\widetilde{k}(\Delta^{(m)})\big), & \text{if } i = j
\end{cases} \\
&\quad = \mathbb{1}_{[i=j]}
\end{aligned}$$

and thus $E(K(\Delta) \otimes \Delta) = I_d$. Furthermore,

$$E\big(\Delta^m K_j(\Delta)\big) = E\big((\Delta^{(j)})^{m_j} k_0(\Delta^{(j)})\big) \prod_{i=1, i \neq j}^{d} E\big((\Delta^{(i)})^{m_i} \widetilde{k}(\Delta^{(i)})\big) = 0$$

whenever $2 \le |m| \le \lfloor p \rfloor$. Hence Condition $(\widetilde{\mathrm{C}})$ is fulfilled. With

$$
\begin{aligned}
&E\big(K_i(\Delta)K_j(\Delta)\big)\\
&\quad= E\left(k_0(\Delta^{(i)})k_0(\Delta^{(j)})\prod_{l=1,l\neq i}^{d}\widetilde{k}(\Delta^{(l)})\prod_{q=1,q\neq j}^{d}\widetilde{k}(\Delta^{(q)})\right)\\
&\quad= \begin{cases} E\big(k_0(\Delta^{(i)})\widetilde{k}(\Delta^{(i)})\big)E\big(k_0(\Delta^{(j)})\widetilde{k}(\Delta^{(j)})\big)\displaystyle\prod_{l=1,l\neq i,l\neq j}^{d}E\big(\widetilde{k}(\Delta^{(l)})^2\big), & \text{if } i\neq j,\\ E\big(k_0(\Delta^{(i)})^2\big)\displaystyle\prod_{l=1,l\neq i}^{d}E\big(\widetilde{k}(\Delta^{(l)})^2\big), & \text{if } i=j,\end{cases}\\
&\quad= \begin{cases} \left(\int k_0\widetilde{k}\right)^2\left(\int\widetilde{k}^2\right)^{d-2}, & \text{if } i\neq j,\\ \left(\int k_0^2\right)\left(\int\widetilde{k}^2\right)^{d-1}, & \text{if } i=j,\end{cases}
\end{aligned}
$$

we obtain the components of $S=\sigma^2E(K(\Delta)\otimes K(\Delta))$.

5.4. *Methods that use more than two observations per step.* Let us consider any of the kernel functions K together with the related distributions of the simulated random variables Φ and Ψ as described in Sections 5.2.1–5.2.3. There, at step n, we took one simulated realization of $\Phi_n\sim U(\{u_j:1\le j\le q\})$, one simulated realization of Ψ_n, and two noisy observations of f at X_n (in square brackets), that is,

$$Y_n=\frac{1}{c_n}K(\Delta_n)\big\{[f(X_n+c_n\Delta_n)-W_{n,1}]-[f(X_n+c_n\Delta_n)-W_{n,1}]\big\},$$

where $\Delta_n=\Phi_n\cdot\Psi_n$. If the distributions of Δ, Φ or Ψ have a support with finitely many points, the use of randomization in Sections 5.1 and 5.2 can be (partly or totally) avoided by taking, in each step of the iteration, the average of $2q$ or $2q\kappa$ observations of function values of the unknown regression function f.

5.4.1. $2q$ *observations per iteration step.* Consider the case of a finitely valued random variable Φ. For observation errors $W_{n,i,l}$, assume that, in lieu of (W_n), the sequence defined by $\overline{W}_n=\frac{1}{q}\sum_{l=1}^{q}(W_{n,1,l}-W_{n,2,l})$, $n\in\mathbb{N}$, satisfies Condition $(\widetilde{\mathrm{D}})$ with σ^2 replaced by σ^2/q. Then the conditional expectation of $E(Y_n\mid X_n,\Psi_n)$ can be estimated unbiasedly by

$$
\begin{aligned}
\overline{Y}_n=\frac{1}{q}\sum_{l=1}^{q}\frac{1}{c_n}K(u_l\Psi_n)\big\{&[f(X_n+c_nu_l\Psi_n)-W_{n,1,l}]\\
&-[f(X_n+c_nu_l\Psi_n)-W_{n,2,l}]\big\},
\end{aligned}
$$

which uses $2q$ observations (in square brackets) of the regression function. Now the assertions of Theorems 2 and 3 and Corollaries 4 and 5 remain valid if we replace Y_n by $\overline{Y}_n$ and S by $(1/q)S$, where S and T are as computed in Section 5.2.

The special case of equidistant $u_1, \dots, u_q$ together with the simultaneous perturbation scheme in Section 5.2.3 was recently suggested by Gerencsér [9].

5.4.2. *$2q\kappa$ observations per iteration step.* If both simulated random variables Φ and Ψ attain finitely many values $u_1, \dots, u_q$ and $\psi_1, \dots, \psi_\kappa$ only, an unbiased estimator of the conditional expectation of $E(Y_n \mid X_n)$ is given by

$$\widehat{Y}_n = \frac{1}{q\kappa} \sum_{j=1}^{q} \sum_{k=1}^{\kappa} \frac{1}{c_n} K(u_j \psi_k) \{ [f(X_n + c_n u_j \psi_k) - W_{n,1,j,k}] - [f(X_n - c_n u_j \psi_k) - W_{n,2,j,k}] \},$$

which uses $2q\kappa$ observations at each step but no randomization. Here we assume that the observation errors $W_{n,i,j,k}$ obey the property that the sequence defined by $\widehat{W}_n = \frac{1}{q\kappa} \sum_{j=1}^{q} \sum_{k=1}^{\kappa} (W_{n,1,j,k} - W_{n,2,j,k})$, $n \in \mathbb{N}$, fulfills Condition $(\widetilde{\mathrm{D}})$ with (W_n) and σ^2 replaced by $\widehat{W}_n$ and $\sigma^2/(q\kappa)$, respectively. Then the assertions of Theorems 2 and 3 and Corollaries 4 and 5 remain valid if we replace Y_n by $\widehat{Y}_n$ and S by $1/(q\kappa)S$, where S and T are as computed in Section 5.2.

If we use the kernel function given in Section 5.2.1, we end up with a generalization of [5] to general p-smooth regression functions with $p \geq 2$, which itself extends Fabian's method [6, 7] from odd numbers $p \geq 3$ (and $q = \lfloor p/2 \rfloor$) to the case of natural numbers $p \geq 2$ (and $q = \lfloor p/2 \rfloor$ or $q = \lceil p/2 \rceil$, possessing unbiased limit distributions in the latter case). In the very special setting $p = 3$ and $q = 1$, this is just the original Kiefer–Wolfowitz method [10] that yields convergence rate $n^{-1/3}$.

6. Proofs.

PROOF OF PROPOSITION 1. Without loss of generality we may assume $\vartheta = 0$ and $f(\vartheta) = 0$. According to Condition (A), ∇f is Lipschitz continuous. Hence

$$\begin{aligned} &\big| f(x+h) - f(x-h) - \langle 2h, \nabla f(x) \rangle \big| \\ &\quad = \left| \int_{-1}^{1} \langle h, \nabla f(x+th) - \nabla f(x) \rangle \, dt \right| \\ &\quad \leq \|h\| \int_{-1}^{1} L|t| \|h\| \, dt = L\|h\|^2, \end{aligned} \tag{6}$$

where L here and in the following inequalities is a constant which may vary from inequality to inequality. Applying (6) and respecting Conditions (B) and (C), we

arrive at

$$
\begin{aligned}
&\|E(Y_n \mid \mathcal{G}_n) - \nabla f(X_n)\| \\
&\quad \le \Big\| E\Big(\frac{1}{2c_n} K(\Delta_n)\big\{ f(X_n + c_n\Delta_n) - f(X_n - c_n\Delta_n) \\
&\qquad\qquad - \langle 2c_n\Delta_n, \nabla f(X_n)\rangle \big\} \mid \mathcal{G}_n \Big) \Big\| \\
&\qquad + \big\| E\big(K(\Delta_n)\langle \Delta_n, \nabla f(X_n)\rangle - \nabla f(X_n) \mid \mathcal{G}_n\big) \big\| \\
&\quad \le c_n^{-1} L E\big((\|c_n\Delta_n\|^2 \|K(\Delta_n)\|) \mid \mathcal{G}_n \big) \\
&\qquad + \big\| E\big(K(\Delta_n)\otimes\Delta_n \mid \mathcal{G}_n\big)\nabla f(X_n) - \nabla f(X_n) \big\| \\
&\quad \le Lc_n \qquad \text{a.s.}
\end{aligned} \tag{7}
$$

Hence $|\langle E(Y_n \mid \mathcal{G}_n) - \nabla f(X_n), \nabla f(X_n)\rangle| \le Lc_n \|\nabla f(X_n)\|$ a.s. and

$$
\langle \nabla f(X_n), E(Y_n \mid \mathcal{G}_n)\rangle \ge \|\nabla f(X_n)\|^2 - Lc_n\|\nabla f(X_n)\| \qquad \text{a.s.} \tag{8}
$$

Referring to (6) and Conditions (C) and (D), we obtain

$$
\begin{aligned}
&E(\|Y_n\|^2 \mid \mathcal{G}_n) \\
&\quad \le E\Big(\Big\| Y_n - \frac{1}{2c_n} K(\Delta_n)\big(f(X_n + c_n\Delta_n) - f(X_n - c_n\Delta_n)\big) \Big\|^2 \mid \mathcal{G}_n \Big) \\
&\qquad + E\Big(\Big\| \frac{1}{2c_n} K(\Delta_n)\big\{ f(X_n + c_n\Delta_n) - f(X_n - c_n\Delta_n) \\
&\qquad\qquad - \langle 2c_n\Delta_n, \nabla f(X_n)\rangle \big\} \Big\|^2 \mid \mathcal{G}_n \Big) \\
&\qquad + E\big(\|K(\Delta_n)\langle \Delta_n, \nabla f(X_n)\rangle\|^2 \mid \mathcal{G}_n \big) \\
&\quad \le Lc_n^{-2} E\big(\|K(\Delta_n)\|^2 W_n^2 \mid \mathcal{G}_n \big) + Lc_n^2 + L\|\nabla f(X_n)\|^2 \qquad \text{a.s.}
\end{aligned} \tag{9}
$$

Lipschitz continuity of ∇f implies, as in (6),

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

Taking conditional expectations and using inequalities (8) and (9), we obtain

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

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

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

where $A_n \geq 0$, $B_n \geq 0$ 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 [15] for nonnegative almost supermartingales.

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

To prove (b), fix $\omega \in \Omega_0$ with $\sup_n \|X_n(\omega)\| < \infty$ and set $x_n := X_n(\omega)$ again. 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 $x_n \to 0$. $\square$

PROOF OF THEOREM 3 AND COROLLARY 5. We apply Lemma 7.1 in [5] to show asymptotic normality.

Step 1. Expansion of $D_n := (1/2c_n)K(\Delta_n)(f(X_n + c_n\Delta_n) - f(X_n - c_n\Delta_n))$. First let us consider the case $p \geq 3$. For x, $\delta \in \mathbb{R}^d$ and $h > 0$ with $x \pm h\delta$, $\vartheta \pm h\delta \in U_\varepsilon(\vartheta)$, Taylor's formula yields

$$\frac{K(\delta)}{2}\{f(x + h\delta) - f(x - h\delta)\} = t(h, \delta) + s(h, \delta) + q(h, \delta)$$

with

$$t(h, \delta) = \frac{K(\delta)}{2}\big(f(\vartheta + h\delta) - f(\vartheta - h\delta)\big),$$

$$s(h, \delta) = \frac{K(\delta)}{2}\langle \nabla f(\vartheta + h\delta) - \nabla f(\vartheta - h\delta), x - \vartheta \rangle,$$

$$q(h, \delta) = \frac{K(\delta)}{2}\left\langle x - \vartheta, \int_0^1 (1 - t)G(t, h, \delta)\, dt\,(x - \vartheta)\right\rangle,$$

where $G(t, h, \delta) := Hf(\vartheta + t(x - \vartheta) + h\delta) - Hf(\vartheta + t(x - \vartheta) - h\delta)$, since f is at least twice differentiable.

Due to Assumption $(\widetilde{\mathrm{A}})$ we have for $\vartheta \pm h\delta \in U_\varepsilon(\vartheta)$,

$$t(h, \delta) = \bar{t}(h, \delta) + K(\delta)o(h^p \|\delta\|^p)$$

with

$$\bar{t}(h, \delta) = \frac{K(\delta)}{2} \sum_{|m| \leq \lfloor p \rfloor} \frac{1}{m!} D^m f(\vartheta)\delta^m h^m \big(1 - (-1)^m\big).$$

Observe that, in the last sum, terms with $|m| = 1$ vanish, that $E(\Delta^m K(\Delta)) = 0$ for $|m|$ odd with $|m| \in \{3, \ldots, \lfloor p \rfloor - 1\}$ and that $1 - (-1)^m = 0$ for $|m|$ even. Hence

$$\begin{aligned} E\bar{t}(h, \Delta) &= \sum_{|m| \leq \lfloor p \rfloor} \frac{1}{m!} D^m f(\vartheta) E\big(\Delta^m K(\Delta)\big) h^m \frac{1 - (-1)^m}{2} \\ &= \sum_{|m| = \lfloor p \rfloor} \frac{1}{m!} D^m f(\vartheta) E\big(\Delta^m K(\Delta)\big) h^m \frac{1 - (-1)^m}{2} =: h^{\lfloor p \rfloor} \widehat{T} \end{aligned} \tag{10}$$

and, since Δ is bounded,

$$Et(h,\Delta) = h^{\lfloor p\rfloor}\widehat{T} + o(h^p).$$

Notice that $\widehat{T} = 0$ under the assumptions of Theorem 3.

Similarly, we have

$$s(x,h,\delta) = \overline{s}(x,h,\delta) + K(\delta)o(h^{p-1}\|\delta\|^{p-1}\|x-\vartheta\|)$$

with

$$\overline{s}(x,h,\delta) = K(\delta)\left\langle \sum_{0\le|m|\le\lfloor p\rfloor-1} \frac{1}{m!} D^m\nabla f(\vartheta)(h\delta)^m \frac{1-(-1)^m}{2}, x-\vartheta \right\rangle.$$

In this sum the term for $|m| = 0$ vanishes. Furthermore, in $E\overline{s}(x,h,\Delta)$ all summands vanish except those with $|m| = 1$, since for odd $|m|$ with $3 \le |m| \le \lfloor p\rfloor - 1$ the term $E(\Delta^m K(\Delta))$ equals zero and for $|m|$ even with $2 \le |m| \le \lfloor p\rfloor - 1$ the term $(1-(-1)^m)$ equals zero too. Hence

$$\begin{aligned} E\overline{s}(x,h,\Delta) &= E\left(K(\Delta)\left\langle \left(\sum_{j=1}^{d} \frac{\partial^2}{\partial x_i\,\partial x_j} f(\vartheta)h\Delta^{(j)}\right)_{i\in\{1,\dots,d\}}, x-\vartheta\right\rangle\right) \\ &= h\sum_{i=1}^{d}\sum_{j=1}^{d} \big(Hf(\vartheta)\big)_{i,j}(x-\vartheta)^{(i)} E\big(\Delta^{(j)}K(\Delta)\big) \\ &= hHf(\vartheta)(x-\vartheta) \end{aligned} \tag{11}$$

and

$$Es(x,h,\Delta) = h\big(Hf(\vartheta) + o(h^{p-2})\big)(x-\vartheta).$$

Due to Lipschitz continuity of the Hessian of f, we obtain for $\vartheta \pm h\delta, x \pm h\delta \in U_\varepsilon(\vartheta)$,

$$\begin{aligned} q(x,h,\delta) &= K(\delta)O\left(\|x-\vartheta\|^2 \int_0^1 (1-t)Lh\|\delta\|\,dt\right) \\ &= K(\delta)O(h\|\delta\|\|x-\vartheta\|^2) \end{aligned}$$

and thus

$$Eq(x,h,\Delta) = O(h\|x-\vartheta\|^2).$$

Set $\overline{D}_n := (1/c_n)\overline{t}(c_n,\Delta_n) + (1/c_n)\overline{s}(X_n,c_n,\Delta_n)$ and $\Omega_n := [\|X_n - \vartheta\| < \varepsilon/2] \in \mathcal{G}_n$. Since Δ is bounded, there is an n_0 such that $\|c_n\Delta_n\| < \varepsilon/2$ on Ω for any $n \ge n_0$. Now observe that

$$\begin{aligned} &(D_n - \overline{D}_n)\mathbb{1}_{\Omega_n} \\ &\quad = K(\Delta_n)\big(o(c_n^{p-1}) + o(c_n^{p-2})\|X_n-\vartheta\| + O(\|X_n-\vartheta\|^2)\big)\mathbb{1}_{\Omega_n} \end{aligned} \tag{12}$$

and, since $X_n \to \vartheta$ a.s., with probability 1,

$$\big(\overline{D}_n - E(\overline{D}_n \mid \mathcal{G}_n)\big)\mathbb{1}_{\Omega_n^c} + (D_n - \overline{D}_n)\mathbb{1}_{\Omega_n^c} = 0 \tag{13}$$

for n sufficiently large.

Then

$$\begin{aligned} D_n &= E(\overline{D}_n \mid \mathcal{G}_n) + \big(\overline{D}_n\mathbb{1}_{\Omega_n} - E\big(\overline{D}_n\mathbb{1}_{\Omega_n} \mid \mathcal{G}_n\big)\big) + \big(\overline{D}_n - E(\overline{D}_n \mid \mathcal{G}_n)\big)\mathbb{1}_{\Omega_n^c} \\ &\quad + \big((D_n - \overline{D}_n)\mathbb{1}_{\Omega_n} - E\big((D_n - \overline{D}_n)\mathbb{1}_{\Omega_n} \mid \mathcal{G}_n\big)\big) \\ &\quad + E\big((D_n - \overline{D}_n)\mathbb{1}_{\Omega_n} \mid \mathcal{G}_n\big) \\ &\quad + (D_n - \overline{D}_n)\mathbb{1}_{\Omega_n^c} \qquad (14) \\ &= \big(Hf(\vartheta) + o(c_n^{p-2}) + O(\|X_n - \vartheta\|)\big)(X_n - \vartheta) + c_n^{\lfloor p\rfloor - 1}\widehat{T} + o(c_n^{p-1}) \\ &\quad + \big(\overline{D}_n\mathbb{1}_{\Omega_n} - E\big(\overline{D}_n\mathbb{1}_{\Omega_n} \mid \mathcal{G}_n\big)\big) \\ &\quad + \big((D_n - \overline{D}_n)\mathbb{1}_{\Omega_n} - E\big((D_n - \overline{D}_n)\mathbb{1}_{\Omega_n} \mid \mathcal{G}_n\big)\big). \end{aligned}$$

Now we consider the case $p \in [2, 3)$. Due to Assumption $(\widetilde{\mathrm{A}})$, for $x \in U_\varepsilon(\vartheta)$, $\nabla f(x) = (Hf(\vartheta) + R(x))(x - \vartheta)$ holds with a matrix-valued remainder term R satisfying $\|R(x)\| = o(\|x - \vartheta\|^{p-2})$. Hence, for $x \pm h\delta \in U_\varepsilon(\vartheta)$ the following representation is valid:

$$\begin{aligned} &\frac{K(\delta)}{2}\{f(x + h\delta) - f(x - h\delta)\} \\ &\quad = \frac{K(\delta)}{2}\int_{-1}^{1}\langle \nabla f(x + th\delta), h\delta\rangle\, dt \\ &\quad = \frac{K(\delta)}{2}\bigg\{\int_{-1}^{1}\langle Hf(\vartheta)(x + th\delta - \vartheta), h\delta\rangle\, dt \\ &\qquad\qquad + \int_{-1}^{1}\langle \nabla f(x + th\delta) - Hf(\vartheta)(x + th\delta - \vartheta), h\delta\rangle\, dt\bigg\} \\ &\quad = \frac{K(\delta)}{2}\bigg\{\langle Hf(\vartheta)(x - \vartheta), 2h\delta\rangle + \int_{-1}^{1}\langle R(x + th\delta)(x + th\delta - \vartheta), h\delta\rangle\, dt\bigg\} \\ &\quad = \frac{K(\delta)}{2}\bigg\{\langle Hf(\vartheta)(x - \vartheta), 2h\delta\rangle \\ &\qquad\qquad + \int_{-1}^{1} o(\|x + th\delta - \vartheta\|^{p-2})\|x + th\delta - \vartheta\|\,\|h\delta\|\, dt\bigg\} \\ &\quad = \frac{K(\delta)}{2}\{\langle Hf(\vartheta)(x - \vartheta), 2h\delta\rangle + o(\|x - \vartheta\|^{p-1} + \|h\delta\|^{p-1})\|h\delta\|\}. \end{aligned}$$

Set $\overline{D}_n := K(\Delta_n)\langle Hf(\vartheta)(X_n - \vartheta), \Delta_n\rangle$ and observe

$$E(\overline{D}_n \mid \mathcal{G}_n) = Hf(\vartheta)(X_n - \vartheta) \qquad \text{a.s.}$$

Define Ω_n as above. Since (13) holds in this case, too, we obtain, similarly as in (14),

$$D_n = Hf(\vartheta)(X_n - \vartheta) + o(\|X_n - \vartheta\|^{p-1}) + \big(\overline{D}_n \mathbb{1}_{\Omega_n} - E\big(\overline{D}_n \mathbb{1}_{\Omega_n} \mid \mathcal{G}_n\big)\big) + \big((D_n - \overline{D}_n)\mathbb{1}_{\Omega_n} - E\big((D_n - \overline{D}_n)\mathbb{1}_{\Omega_n} \mid \mathcal{G}_n\big)\big) + o(c_n^{p-1}).$$

Step 2. Recursion in standard form. Using $U_n := X_n - \vartheta$ and the expansion of D_n of the first step, recursion (1) can be rewritten in the form of (7.2) in [5],

$$U_{n+1} = U_n - a_n Y_n = U_n - a_n D_n + \frac{a_n}{c_n} K(\Delta_n) W_n$$
$$= (I_d - a_n A_n) U_n + a_n n^{\gamma} (V_n + n^{-1/2} T_n),$$

where $\gamma = 1/(2p)$ or $\gamma = 1/(2\lfloor p \rfloor)$ in the case of Theorem 3 or Corollary 5, respectively, and, with notation slightly different from that that appears after Condition $(\widetilde{\mathrm{C}})$,

$$V_n = V_{n,1} + V_{n,2},$$
$$V_{n,1} = c^{-1} K(\Delta_n) W_n,$$
$$V_{n,2} = D_n \mathbb{1}_{\Omega_n} - E\big(D_n \mathbb{1}_{\Omega_n} \mid \mathcal{G}_n\big)$$

and, if $p \in [2, 3)$,

$$A_n = Hf(\vartheta) + o(\|X_n - \vartheta\|^{p-2}), \qquad T_n = o(1) \tag{15}$$

or, if $p \geq 3$,

$$A_n = Hf(\vartheta) + o(c_n^{p-2}) + O(\|X_n - \vartheta\|), \qquad T_n = c^{\lfloor p \rfloor - 1} \widehat{T} + o(1). \tag{16}$$

Step 3. Rates of convergence of (X_n). Since $X_n \to \vartheta$ a.s., we obtain $A_n \to A$ and $T_n \to T$ a.s. This is sufficient for (7.5), (7.8) and (7.9) in [5]. Conditions (7.10) and (7.11) in [5] imposed on $(V_{n,1})$ follow directly from Assumption (D) which is included in $(\widetilde{\mathrm{D}})$. Note that $V_{n,2}$ is $\mathcal{G}_{n+1}$-measurable and $E(V_{n,2} \mid \mathcal{G}_n) = 0$ a.s. Assume $p \geq 3$. Relationships (10), (11) and (12), independence of X_n and Δ_n, and required consistency of (X_n) imply

$$\begin{aligned} &\|E(V_{n,2} \otimes V_{n,2} \mid \mathcal{G}_n)\| \\ &\quad = O(c_n^{-2}) E\big(\|\overline{t}(c_n, \Delta_n)\|^2 \mid \mathcal{G}_n\big) + O(c_n^{-2}) E\big(\|\overline{s}(X_n, c_n, \Delta_n)\|^2 \mathbb{1}_{\Omega_n} \mid \mathcal{G}_n\big) \\ &\qquad + E\big(\|K(\Delta_n)\|^2 \big(o(c_n^{2p-2}) + o(c_n^{2p-4}) \|X_n - \vartheta\|^2 \\ &\qquad\qquad + O(\|X_n - \vartheta\|^4)\big) \mathbb{1}_{\Omega_n} \mid \mathcal{G}_n\big) \\ &\quad = O(c_n^{2\lfloor p \rfloor - 2}) + O(\|X_n - \vartheta\|^2) \mathbb{1}_{\Omega_n} \to 0, \qquad n \to \infty, \text{ a.s.}, \end{aligned} \tag{17}$$

which yields

$$\sup_n E(\|V_{n,2}\|^2 \mid \mathcal{G}_n) < \text{const} < \infty \qquad \text{a.s.} \tag{18}$$

It is easy to see that the same statement is valid in the case $p \in [2, 3)$ too. Now Lemma 7.1(b) in [5] asserts $X_n - \vartheta = O(n^{\gamma}\sqrt{a_n})$ almost in $L^2(P)$. Recall the assumptions on the sequence (a_n) and $p > 2$ as required in Theorem 3. Then the last result together with (15) or (16) implies validity of $A_n - Hf(\vartheta) = o(1/\sqrt{na_n})$ almost in $L^2(P)$. This gives Condition (7.7) in [5].

Step 4. Asymptotic normality of (X_n). Let $B_{n,j}(t) := 1/\sqrt{n}(\sum_{i=1}^{\lfloor nt \rfloor} V_{i,j} + (nt - \lfloor nt \rfloor)V_{\lfloor nt \rfloor + 1, j})$, $t \in [0, 1]$, $j \in \{1, 2\}$. Due to Condition $(\widetilde{\mathrm{D}})$ the process $B_{n,1}$ converges in distribution to a Brownian motion B in $C([0, 1], \mathbb{R}^d)$ with covariance S of $B(1)$.

To prove $B_{n,2} \to_P 0$ in $C([0, 1], \mathbb{R}^d)$, we apply an invariance principle for martingale difference sequences of Berger [1]. Its assumptions can be met by referring to relationships (17), (18) and

$$\underset{r>0}{\forall} \qquad E\big(\|V_{n,2}\|^2 \mathbb{1}_{[\|V_{n,2}\|^2 \geq rn]} \mid \mathcal{G}_n\big) \to 0, \qquad n \to \infty, \text{ a.s.},$$

which is a consequence of (17). Slutsky's theorem yields $B_n := B_{n,1} + B_{n,2} \to_{\mathcal{D}} B$; hence Condition (7.3) in [5] is fulfilled.

Assumption (7.4) in [5] is satisfied if $B_{n,j}(1) = O(1)$ almost in $L^1(P)$. For $j = 1$, this is Condition $(\widetilde{\mathrm{E}})$. For $j = 2$ this follows with (18) from

$$E\,\|B_{n,2}(1)\|^2 = \frac{1}{n}\sum_{i=1}^{n} E\,\|V_{i,2}\|^2 \leq \frac{1}{n}\sum_{i=1}^{n} E\Big(\sup_i E(\|V_{i,2}\|^2 \mid \mathcal{G}_i)\Big) < \infty.$$

Now the assertion of Lemma 7.1(a) in [5] applies. This proves Theorem 3 and Corollary 5. □

PROOF OF THEOREM 2 AND COROLLARY 4. We adopt the first step of the proof of Theorem 3. For the second step, we check the assumptions of Theorem 1 in Walk [20]. Using $U_n := X_n - \vartheta$ and the expansion of D_n of the first step, recursion (1) can be rewritten in the form

$$\begin{aligned} U_{n+1} &= U_n - a_n Y_n = U_n - a_n D_n + \frac{a_n}{c_n} K(\Delta_n) W_n \\ &= \Big(I_d - \frac{1}{n} a A_n\Big) U_n + n^{-(1+\beta)/2} a V_n + n^{-1-\beta/2} a T_n, \end{aligned}$$

where $\beta = 1 - 2\gamma$ and A, A_n, B, B_n, T, T_n and V_n are defined as in the last proof. (Notice that Theorem 1 in [20] treats the case $a = 1$, but the extension to general $a > 0$ is trivial.) Proposition 1 implies $A_n \to A = Hf(\vartheta)$ a.s. Due to the assumption on λ_0, $\operatorname{spec}(aA) > \beta/2$ holds. Distributional convergence of B_n to B is established as above. In this case the covariance of $aB(1)$ is equal to $a^2 S$. Finally, we have $T_n \to T$ a.s. Hence Theorem 1 in [20] can be applied. For $t = 1$ this yields the assertions of Theorem 2 and Corollary 4. □

PROOF OF REMARK 6(C). $(\widehat{\mathrm{D}}) \Rightarrow (\widetilde{\mathrm{D}})$. To show that the process B_n, as defined in $(\widetilde{\mathrm{D}})$, converges in distribution to a Brownian motion B in $C([0,1],\mathbb{R}^d)$ with covariance S of $B(1)$, we apply an invariance principle for martingale difference sequences found in Berger [1].

Since

$$E(V_n \mid \mathcal{F}_n) = K(\Delta_n)\, E(W_n \mid \mathcal{F}_n) = 0 \qquad \text{a.s.}$$

and V_n is $\mathcal{F}_{n+1}$-measurable, (V_n) is a martingale difference sequence with respect to $(\mathcal{F}_{n+1})$. Kolmogorov's strong law of large numbers can be used to show

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} E(V_i \otimes V_i \mid \mathcal{F}_i) &= \frac{1}{n}\sum_{i=1}^{n} K(\Delta_i) \otimes K(\Delta_i)\, E(W_i^2 \mid \mathcal{F}_i) \\ &= \sigma^2 \frac{1}{n}\sum_{i=1}^{n} K(\Delta_i) \otimes K(\Delta_i) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} K(\Delta_i) \otimes K(\Delta_i)\big(E(W_i^2 \mid \mathcal{F}_i) - \sigma^2\big) \\ &\to \sigma^2 E\big(K(\Delta) \otimes K(\Delta)\big) = S, \qquad n \to \infty, \text{ a.s.}\end{aligned}$$

Furthermore, for any $r > 0$, we obtain

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} E\big(\|V_i\|^2 \mathbb{1}_{[\|V_i\|^2 \geq rn]} \mid \mathcal{F}_i\big) \\ &\quad \leq r^{-\varepsilon/2} \frac{1}{n^{1+\varepsilon/2}} \sum_{i=1}^{n} E\big(\|K(\Delta_i) W_i\|^{2+\varepsilon} \mid \mathcal{F}_i\big) \\ &\quad \leq \left(r^{-\varepsilon/2} \frac{1}{n}\sum_{i=1}^{n} \|K(\Delta_i)\|^{2+\varepsilon}\right) \\ &\qquad \times \left(\frac{1}{n^{\varepsilon/2}} \sup_{i \in \mathbb{N}} E(\|W_i\|^{2+\varepsilon} \mid \mathcal{F}_i)\right) \xrightarrow{P} 0, \qquad n \to \infty,\end{aligned}$$

since $\frac{1}{n}\sum_{i=1}^{n} \|K(\Delta_i)\|^{2+\varepsilon}$ converges a.s. and $\sup_{i\in\mathbb{N}} E(\|W_i\|^{2+\varepsilon} \mid \mathcal{F}_i) < \infty$ is bounded a.s. This proves the assumptions of the central limit theorem mentioned above. □

PROOF OF REMARK 6(C). $(\widehat{\mathrm{D}}) \wedge (\widehat{\mathrm{E}}) \Rightarrow (\widetilde{\mathrm{E}})$. Since (V_n) is a martingale difference sequence, we have

$$E\|B_n(1)\|^2 = \frac{1}{n}\sum_{i=1}^{n} E\|V_i\|^2 = \frac{1}{n}\sum_{i=1}^{n} E\big(\|K(\Delta_i)\|^2 W_i^2\big) = O(1). \tag{19}$$

From this we can find sufficient conditions for (19) as those mentioned in $(\widehat{\mathrm{E}})$. □

REFERENCES

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

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

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

[4] CHEN, H. F., DUNCAN, T. E. and PASIK-DUNCAN, B. (1999). A Kiefer–Wolfowitz algorithm with randomized differences. *IEEE Trans. Automat. Control* **44** 442–453.

[5] DIPPON, J. and RENZ, J. (1997). Weighted means in stochastic approximation of minima. *SIAM J. Control Optim.* **35** 1811–1827.

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

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

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

[9] GERENCSÉR, L. (1999). Convergence rate of moments in stochastic approximation with simultaneous perturbation gradient approximation and resetting. *IEEE Trans. Automat. Control* **44** 894–905.

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

[11] KUSHNER, H. J. and CLARK, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer, New York.

[12] KUSHNER, H. J. and YIN, G. G. (1997). *Stochastic Approximation Algorithms and Applications*. Springer, New York.

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

[14] POLYAK, B. T. and TSYBAKOV, A. B. (1990). Optimal orders of accuracy for search algorithms of stochastic optimization. *Problems Inform. Transmission* **26** 126–133.

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

[16] RUPPERT, D. (1991). Stochastic approximation. In *Handbook of Sequential Analysis* (B. K. Ghosh and P. K. Sen, eds.) 503–529. Dekker, New York.

[17] SPALL, J. C. (1988). A stochastic approximation algorithm for large-dimensional systems in the Kiefer–Wolfowitz setting. In *Proc. Conference on Decision and Control* 1544–1548. IEEE, New York.

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

[19] SPALL, J. C. (1997). A one-measurement form of simultaneous perturbation stochastic approximation. *Automatica J. IFAC* **33** 109–112.

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

FACHBEREICH MATHEMATIK
UNIVERSITÄT STUTTGART
70550 STUTTGART
GERMANY
E-MAIL: dippon@mathematik.uni-stuttgart.de