A STOCHASTIC APPROXIMATION ALGORITHM WITH RANDOM DIFFERENCES*

H. F. Chen

Institute of Systems Science
Beijing 100080, P. R. China
hfchen%iss03@ISS03.iss.ac.cn

T. E. Duncan and **B. Pasik-Duncan**

Department of Mathematics
University of Kansas
Lawrence, KS 66045
duncan@kuhub.cc.ukans.edu
bozenna@kuhub.cc.ukans.edu

Abstract. A simultaneous perturbation algorithm for a Kiefer-Wolfowitz type problem that uses one-sided randomized differences and truncations at randomly varying bounds is given in this paper. At each iteration of the algorithm only two observations are required in contrast to 2ℓ observations, where ℓ is the dimension, in the classical algorithm. The algorithm given here is convergent under only some mild conditions. A rate of convergence and an asymptotic normality of the algorithm are also described. While only an algorithm with one-sided randomized differences is considered here, a corresponding algorithm with two-sided randomized differences has similar results with the same assumptions.

Key words: stochastic approximation, Kiefer-Wolfowitz algorithm, stochastic approximation algorithm with randomized differences, simultaneous perturbation stochastic approximation, constrained simultaneous perturbation stochastic approximation algorithm

1. INTRODUCTION

A Kiefer-Wolfowitz (KW) (1952) algorithm is used to find the extrema of an unknown function $L : \mathbb{R}^\ell \to \mathbb{R}$ which may be observed with some additive noise. If the gradient of L, ∇L, can be observed then the problem can be solved by a Robbins-Monro (RM) algorithm.

Let x_n be the estimate of the unique extremum of L at the nth iteration. One approach to a KW algorithm is to observe L at the following values.

$$x_n^{i+} = [x_n^1, \ldots, x_n^{i-1}, x_n^i + c_n, x_n^{i+1}, \ldots, x_n^\ell]^{\mathrm{T}}$$

$$x_n^{i-} = [x_n^1, \ldots, x_n^{i-1}, x_n^i - c_n, x_n^{i+1}, \ldots, x_n^\ell]^{\mathrm{T}}$$

for $i = 1, 2, \ldots, \ell$ where $c_n \in \mathbb{R} \setminus \{0\}$.

Consider noisy observations of L so that

$$y_{n+1}^{i+} = L(x_n^{i+}) + \xi_{n+1}^{i+}$$

and

$$y_{n+1}^{i-} = L(x_n^{i-}) + \xi_{n+1}^{i-}$$

* Research partially supported by NSF Grant DMS-9305936 and by the National Natural Science Foundation of China.

where ξ_{n+1}^{i+} and ξ_{n+1}^{i-} are observation noises. The ratio

$$\frac{y_{n+1}^{i+} - y_{n+1}^{i-}}{2c_n} \tag{1}$$

can be used as an estimate for the ith component of ∇L. With this approach a KW algorithm requires 2ℓ measurements of L. If ℓ is large, for example, the optimization of weights in a neural network, then this KW algorithm can be rather slow.

To reduce the evaluations for a KW algorithm Spall (1992) replaced the deterministic componentwise differences in (1) by a symmetric random difference. Using the ordinary differential equation (ODE) method (Kushner and Clark, 1978) Spall showed the convergence and the asymptotic normality of the modified KW algorithm though the conditions required are restrictive.

Initially Spall's KW algorithm and the conditions that he uses are described. Let $(\Delta_k^i,\ i = 1,\ldots,\ell,\ k = 1,2,\ldots)$ be a sequence of mutually independent and identically distributed random variables with zero mean. Let Δ_k be given by

$$\Delta_k = [\Delta_k^1,\ldots,\Delta_k^\ell]^{\mathrm{T}}. \tag{2}$$

At each iteration, two measurements are taken:

$$y_{k+1}^+ = L(x_k + c_k\Delta_k) + \xi_{k+1}^+$$
$$y_{k+1}^- = L(x_k - c_k\Delta_k) + \xi_{k+1}^-$$

Then the vector symmetric difference

$$\frac{(y_{k+1}^+ - y_{k+1}^-)g_k}{2c_k} \tag{3}$$

where

$$g_k = \left[\frac{1}{\Delta_k^1},\ldots,\frac{1}{\Delta_k^\ell}\right]^{\mathrm{T}} \tag{4}$$

is used as an estimate for $\nabla L(x_k)$.

The KW algorithm is formed as follows

$$x_{k+1} = x_k + a_k\frac{(y_{k+1}^+ - y_{k+1}^-)g_k}{2c_k}. \tag{5}$$

In this form the algorithm seeks the maximum of L. The minimum of L is found by replacing a_k by $-a_k$.

For the convergence of the algorithm (5) Spall (1992) required the following conditions.

A1. The random variables $(\xi_{k+1}^+ - \xi_{k+1}^-,\ k \in \mathbb{N})$ is a martingale difference sequence (mds) with uniformly bounded second moments.

A2. $\sup_{k\in\mathbb{N}} E(L^2(x_k + c_k\Delta_k)) < \infty$.

A3. The sequence $(x_k,\ k \in \mathbb{N})$ is assumed a priori to be uniformly bounded, that is,

$$\sup_{k\in\mathbb{N}} \|x_k\| < \eta < \infty \quad \text{a.s.}$$

where $\eta \in \mathbb{R}_+$.

A4. The third derivative of L is bounded.

A5. $x^0 \in \mathbb{R}^\ell$ is an asymptotically stable point for the differential equation $\frac{\mathrm{d}x}{\mathrm{d}t} = f(x(t))$ where $f = \nabla L$.

A6. The sequence $(x_k,\ k \in \mathbb{N})$ is infinitely often in a compact set that is contained in the domain of attraction of x^0 given in A5.

A7. The sequences $(a_k,\ k \in \mathbb{N})$ and $(c_k,\ k \in \mathbb{N})$ satisfy $a_k > 0$, $c_k > 0$ for all $k \in \mathbb{N}$, $a_k \to 0$ and $c_k \to 0$ as $k \to \infty$, $\sum_{k=1}^{\infty} a_k = \infty$ and $\sum_{k=1}^{\infty}\left(\frac{a_k}{c_k}\right)^2 < \infty$.

Furthermore, some conditions are imposed on the sequence $(\Delta_k,\ k \in \mathbb{N})$ in (Spall, 1992) but this sequence can be arbitrarily chosen by the user of the algorithm in (Spall, 1992).

In this paper, a one-sided randomized difference is used, that is,

$$\frac{(y_{k+1}^+ - y_{k+1}^0)g_k}{c_k} \tag{6}$$

and

$$g_k = \left[\frac{1}{\Delta_k^1},\ldots,\frac{1}{\Delta_k^\ell}\right]^{\mathrm{T}} \tag{7}$$

is used to estimate $f(x_k) = \nabla L(x_k)$ where

$$y_{k+1}^+ = L(x_k + c_k\Delta_k) + \xi_{k+1}^+ \tag{8}$$
$$y_{k+1}^0 = L(x_k) + \xi_{k+1}^0. \tag{9}$$

By the change in forming the differences and a modification of the algorithm (5) and the use of a direct approach to verify convergence, the conditions A2–A6 are eliminated and A1 is weakened to one that provides not only a sufficient but also a necessary condition for convergence. This result is given in Theorem 1. In Theorem 2 the observation noise is modified to one with only bounded second moments that is independent of $(\Delta_k,\ k \in \mathbb{N})$. A convergence rate and an asymptotic normality of the algorithm are given in Theorems 3 and 4 respectively. The proofs of these results are given in (Chen, *et. al.*, a)

2. THE ALGORITHM AND ITS CONVERGENCE

Initially the algorithm is precisely described. Let $(\Delta_k^i,\ i = 1,\ldots,\ell,\ k \in \mathbb{N})$ be a sequence of independent and identically distributed random variables where $|\Delta_k^i| < a$, $\left|\frac{1}{\Delta_k^i}\right| < b$, $E\left(\frac{1}{\Delta_k^i}\right) = 0$ for all $i \in \{1,\ldots,\ell\}$ and $k \in \mathbb{N}$ and $a, b \in \mathbb{R}_+$. Furthermore let Δ_k be independent of $\mathcal{F}_k^\xi = \sigma(\xi_i^+, \xi_i^-,\ i \in \{0,\ldots,k\})$ for $k \in \mathbb{N}$. Define y_{k+1} and ξ_{k+1} by the following equations

$$y_{k+1} = \frac{(y_{k+1}^+ - y_{k+1}^0)g_k}{c_k}$$
$$\xi_{k+1} = \xi_{k+1}^+ - \xi_{k+1}^0.$$

It follows that

$$y_{k+1} = \frac{(L(x_k + c_k\Delta_k) - L(x_k))g_k}{c_k} + \frac{\xi_{k+1}g_k}{c_k}. \quad (10)$$

Choose $x^* \in \mathbb{R}^\ell$ and fix it. Define the following KW algorithm with randomly varying truncations and randomized differences:

$$x_{k+1} = (x_k + a_k y_{k+1})1_{\{\|x_k + a_k y_{k+1}\| \leqslant M_{\sigma_k}\}} + x^* 1_{\{\|x_k + a_k y_{k+1}\| > M_{\sigma_k}\}} \quad (11)$$

$$\sigma_k = \sum_{i=0}^{k-1} 1_{\{\|x_i + a_i y_{i+1}\| > M_{\sigma_i}\}}, \quad \sigma_0 \equiv 0 \quad (12)$$

where $(M_k,\ k \in \mathbb{N})$ is a sequence of strictly positive, strictly increasing real numbers that diverge to $+\infty$. It is clear that σ_k is the number of truncations that have occurred before time k. Clearly the random vector x_k is measurable with respect to $\mathcal{F}_k := \mathcal{F}_k^\xi \vee \mathcal{F}_{k-1}^\Delta$ where $\mathcal{F}_k^\Delta = \sigma(\Delta_i,\ i \in \{0, \ldots, k\})$. Thus the random vector Δ_k is independent of $\sigma(x_i,\ i \leqslant k)$.

The following conditions are imposed on the algorithm.

H1. The function $\nabla L = f$ is locally Lipschitz continuous. There is a unique maximum of L at x^0 so that $f(x^0) = 0$ and $f(x) \neq 0$ for $x \neq x^0$. There is a $c_0 \in \mathbb{R}_+$ such that $\|x^*\| < c_0$ and $\sup_{\|x\|=c_0} L(x) < L(x^*)$.

H2. The two sequences of strictly positive real numbers $(a_k,\ k \in \mathbb{N})$ and $(c_k,\ k \in \mathbb{N})$ satisfy $c_k \to 0$ as $k \to \infty$, $\sum_{k=1}^{\infty} a_k = \infty$ and there is a $p \in (1, 2]$ such that $\sum_{k=1}^{\infty} a_k^p < \infty$.

Remark. If L is twice continuously differentiable then f is locally Lipschitz continuous. If in H1 x^0 is the unique minimum of L, then in (11,12) a_k should be replaced by $-a_k$.

The following theorem gives necessary and sufficient conditions for the convergence of the algorithm (11).

Theorem 1. *Let* H1 *and* H2 *be satisfied and* $(x_k,\ k \in \mathbb{N})$ *be given by* (11). *The sequence* $(x_k,\ k \in \mathbb{N})$ *satisfies*

$$\lim_{k\to\infty} x_k = x^0 \quad \text{a.s.} \quad (13)$$

if and only if the observation noise ξ_k *in* (10) *can be decomposed into the sum of two parts for each* $j \in \{1, \ldots, \ell\}$ *as*

$$\xi_k = e_k^j + \nu_k^j \quad (14)$$

such that

$$\sum_{k=1}^{\infty} \frac{a_k e_{k+1}^j}{c_k \Delta_k^j} < \infty \quad \text{a.s.} \quad (15)$$

and

$$\lim_{k\to\infty} \frac{\nu_{k+1}^j}{c_k \Delta_k^j} = 0 \quad \text{a.s.} \quad (16)$$

for $j = 1, \ldots, \ell$.

While Theorem 1 gives necessary and sufficient conditions for the convergence of the KW algorithm (11) it may not be apparent if there are useful noise processes that satisfy (14-16). The following theorem gives a large family of noise processes that satisfy (14–16).

Theorem 2. *Let* H1 *and* H2 *be satisfied. If* $\sum_{k=1}^{\infty} \frac{a_k^2}{c_k^2} < \infty$ *and the observation noise* $(\xi_k,\ k \in \mathbb{N})$ *is independent of* $(\Delta_k,\ k \in \mathbb{N})$ *and satisfies one of the following two conditions*

i) $\sup_k |\xi_k| \leqslant \xi$ a.s. *where* ξ *is a random variable;*

ii) $\sup_k E\xi_k^2 < \infty$, *then*

$$\lim_{k\to\infty} x_k = x^0 \quad \text{a.s.} \quad (17)$$

where x_k *is given by* (11).

It is important to note that the random variable ξ_k may have arbitrary dependence on the family of random variables $(\xi_j;\ j \in \mathbb{N},\ j \neq k)$ and may not be zero mean. For example, a sequence of bounded deterministic observation errors satisfies the conditions i) or ii).

Remark. Theorems 1 and 2 can be extended to the case where $f(x) = 0$ for all $x \in J$ and J is not a singleton. In this case H1 is replaced by some conditions on v where $v = -L$ and f is locally Lipschitz continuous [1, 2].

While necessary and sufficient conditions for the convergence of the algorithm (11) are important, it is also important to have some information on the rate of convergence of the algorithm. In the following theorem a rate of convergence of the algorithm (11) is given.

Theorem 3. *Assume the hypotheses of Theorem* 2 *and that*

$$\lim_{n\to\infty} (a_{n+1}^{-1} - a_n^{-1}) = \alpha \geqslant 0 \quad (18)$$

$$c_k = o(a_k^\delta) \quad (19)$$

for some $\delta \in (0, 1)$ *and*

$$f(x) = F(x - x_0) + \delta(x) \quad (20)$$

where $F \in L(\mathbb{R}^\ell, \mathbb{R}^\ell)$, $\delta(x) = o(\|x - x^0\|)$ *and* $F + \alpha\delta I$ *is stable. Then* $(x_n,\ n \in \mathbb{N})$ *given by* (11) *satisfies*

$$\|x_n - x^0\| = o(a_n^\delta) \quad \text{a.s.} \quad (21)$$

for δ *given in* (19).

Remark. If $a_n = \frac{1}{n}$ and $c_n = \frac{1}{n^v}$ for some $v \in \left(0, \frac{1}{2}\right)$ and all $n \in \mathbb{N}$ then the conditions on $(a_n,\ n \in \mathbb{N})$ and $(c_n,\ n \in \mathbb{N})$ in Theorem 3 are satisfied.

The following result is an asymptotic normality property of $(x_n,\ n \in \mathbb{N})$ given by the algorithm (11).

Theorem. 4. *Assume that the conditions of Theorem* 2 *are satisfied and that*

i) $\lim_{n\to\infty}(a_{n+1}^{-1} - a_n^{-1}) = \alpha > 0$ *and* $c_n = a_n^{\gamma}$ *for some* $\gamma \in \left(\frac{1}{4}, \frac{1}{2}\right)$.

ii) $\|f(x) - F(x - x^0)\| \leqslant b\|x - x^0\|^{1+\beta}$ *for some* $\beta > 0$ *and* $b > 0$.

iii) $F + \alpha\mu I$ *is stable for* $\mu = \frac{1}{2} - \gamma$.

iv) $\xi_n = \sum_{i=0}^{r} b_i w_{n-i}$ *where* $w_i = 0$ *for* $i < 0$, $(b_i,\ i \in \mathbb{N})$ *is a sequence of real numbers,* $r \in \mathbb{N}$ *is fixed and* $(w_i,\ \mathcal{F}_i^w,\ i \in \mathbb{N})$ *is a martingale difference sequence that satisfies*

$$E[w_i^2 \mid \mathcal{F}_{i-1}^w] \leqslant \sigma_0$$

for all $i \in \mathbb{N}$ *where* $\sigma_0 \in \mathbb{R}_+$,

$$\lim_{i\to\infty} E[w_i^2 \mid \mathcal{F}_{i-1}^w] = \sigma^2$$

where $\sigma^2 \in \mathbb{R}_+$ *and*

$$\lim_{N\to\infty} \sup_{i\in\mathbb{N}} E[w_i^2 1_{\{|w_i|>N\}}] = 0.$$

Then

$$a_n^{-\mu}(x_n - x^0) \xrightarrow{d} Z$$

where Z *is* $N(0, S)$,

$$S = \sigma^2\sigma_\Delta^2 \left(\sum_{i=0}^{r} b_i\right)^2 \int_0^\infty e^{t(F+\alpha\mu I)} e^{t(F+\alpha\mu I)^T} dt$$

$$\sigma_\Delta^2 = E\left[\frac{1}{(\Delta_i^1)^2}\right]$$

and r *is given in* iv).

A classical Kiefer-Wolfowitz algorithm has been modified in two ways: i) the one-sided randomized differences are used instead of the two-sided deterministic differences, ii) the estimates are truncated at randomly varying bounds. For the convergence analysis, a direct method is used rather than the classical probabilistic method or the ordinary differential equation method. By the algorithm modifications i) and ii) and a different approach to algorithm analysis, the following algorithm improvements have been made: i) some restrictive conditions on the function L or some boundedness assumptions on the estimates have been removed, ii) some restrictive conditions on the noise process have been removed, iii) the number of required observations at each iteration has been reduced. The algorithm has been numerically tested on a number of examples to verify the convergence to the maximum. If the function L has many extrema then the algorithm may become stuck at a local extremum. To obtain the almost sure convergence to the global extrema, some methods that combine search and stochastic approximation are needed, but it seems that there is a lack of a sufficient theory for this approach.

REFERENCES

Chen, H. F., T. E. Duncan and B. Pasik-Duncan, A Kiefer-Wolfowitz algorithm with randomized differences, preprint.

Kiefer, K and J. Wolfowitz (1952). Stochastic approximation of a regression function, *Ann. Math. Stat.*, **23**, 462–466.

Kushner, H. J. and D. A. Clark (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* Springer, New York.

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