REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
	Technical Report	-

4. TITLE AND SUBTITLE
Convergence Rates of Finite Difference Stochastic ApproximationAlgorithms

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W911NF-04-D-0003

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Liyi Dai

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
North Carolina State University
2701 Sullivan Drive
Admin Services III; Box 7514
Raleigh, NC 27695 -7514

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
62483-CS-SR.17

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Recently there has been renewed interests in derivative free approaches to stochastic optimization. In this paper, we examine the rates of convergence for the Kiefer-Wolfowitz algorithm and the mirror descent algorithm, under various updating schemes using finite dfferences as gradient approximations. It is shown that the convergence of these algorithms can be accelerated by controlling the implementation of the finite differences. Particularly, it is shown that the rate can be increased to n^{-2/5} in general and to n^{-1/2} in Monte Carlo optimization for a broad class of problems, in the iteration number n.

15. SUBJECT TERMS
stochastic approximation, Kiefer-Wolfowitz algorithm, mirror descent algorithm,♀finite-difference approximation, Monte Carlo methods

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Liyi Dai
UU	UU	UU	UU		19b. TELEPHONE NUMBER 919-549-4350

Report Title

Convergence Rates of Finite Difference Stochastic ApproximationAlgorithms

ABSTRACT

Recently there has been renewed interests in derivative free approaches to stochastic optimization. In this paper, we examine the rates of convergence for the Kiefer-Wolfowitz algorithm and the mirror descent algorithm, under various updating schemes using finite dfferences as gradient approximations. It is shown that the convergence of these algorithms can be accelerated by controlling the implementation of the finite differences. Particularly, it is shown that the rate can be increased to n^{-2/5} in general and to n^{-1/2} in Monte Carlo optimization for a broad class of problems, in the iteration number n.

Convergence Rates of Finite Difference Stochastic Approximation Algorithms [1]

Liyi Dai
Army Research Office
Research Triangle Park, NC 27703
liyi.dai.civ@mail.mil

Abstract

Recently there has been renewed interests in derivative free approaches to stochastic optimization. In this paper, we examine the rates of convergence for the Kiefer-Wolfowitz algorithm and the mirror descent algorithm, under various updating schemes using finite differences as gradient approximations. It is shown that the convergence of these algorithms can be accelerated by controlling the implementation of the finite differences. Particularly, it is shown that the rate can be increased to $n^{-2/5}$ in general and to $n^{-1/2}$ in Monte Carlo optimization for a broad class of problems, in the iteration number n.

Keywords. stochastic approximation, Kiefer-Wolfowitz algorithm, mirror descent algorithm, finite-difference approximation, Monte Carlo methods

[1] *This work is supported in part by the U.S. Army Research Office under agreement W911NF-04-D-0003.*

1. Introduction. Let R denote the set of real numbers. Consider a real-valued function $J(\theta)$ of the form $J(\theta) = E_X[L(X(\theta))]$ where θ is a parameter, or a vector of parameters, $L(X)$ is a real-valued function, and $X(\theta)$ is a random variable that depends on θ. For simplicity, throughout this paper we assume that θ is a scalar and $\theta \in \Theta \subset R$, and $X(\theta)$ is of the form $X(\theta) = X(\theta, \xi)$, where ξ is a random variable independent of θ. In such a formulation, $X(\theta)$ is parameterized on an underlying probability space that is independent of θ. For any two random variables η and ξ, there exists a Borel function ϕ such that $\eta = \phi(\xi)$ [Shiryayev (1984), p.172]. Such a representation for $X(\theta, \xi)$ is always possible. Therefore, $J(\theta)$ can be written as $J(\theta) = E_\xi[L(X(\theta, \xi))]$. We are particularly interested in finding an optimal parameter $\theta^* \in \Theta$ to optimize, say minimize, $J(\theta)$. This is a challenging problem since the analytical form of $J(\theta)$ is usually unavailable for most problems of interest. What is obtainable are the sample measurements of the random value of $L(X(\theta, \xi))$. We have to use the information on $L(X(\theta, \xi))$ to find θ^*. Such stochastic optimization problems can be found in many applications. The main approach to finding the optimal solution is to successively approximate θ^* via algorithms of *stochastic approximation*. This is a classical and standard approach that has been adopted in practice for decades. The *Robbins-Monro* (RM) algorithm, the *Kiefer-Wolfowitz* (KW) algorithm, and the relatively recent mirror descent (MD) algorithm are the most popular algorithms of this class.

The RM algorithm, introduced by Robbins and Monro (1951), finds θ^* in the following way. Let θ_0 be selected and $\{a_n\}$ a sequence of positive numbers. For each integer $n \geq 0$, let

$$\theta_{n+1} = \theta_n - a_n g_n \tag{1}$$

where g_n is an unbiased estimate of the derivative $J'(\theta)$ of $J(\theta)$ with respect to θ. Assume that $J'(\theta)$ exists on Θ and the variance of g_n is uniformly bounded for all n. Assume $(\theta - \theta^*)J'(\theta) > 0$ for all $\theta \neq \theta^*$,

$$\sum_n a_n = \infty, \quad \sum_n a_n^2 < \infty,$$

and that several other technical conditions are satisfied. Then $\{\theta_n\}$ converges to θ^* with probability one. The convergence rate (in terms of root mean square error) is $n^{-1/2}$. Note that this is the best possible rate of convergence for algorithms of the form (1) for stochastic optimization [see, e.g., Fabian (1971)].

The KW algorithm, introduced by Kiefer and Wolfowitz (1952), is a modification of the RM algorithm by approximating the gradient using a finite difference and finds θ^* recursively

by

$$\theta_{n+1} = \theta_n - a_n h_n, \tag{2}$$

where

$$h_n = \frac{L(X(\theta_n + \delta_n, \xi_{1,n})) - L(X(\theta_n - \delta_n, \xi_{2,n}))}{2\delta_n}, \tag{3}$$

$\{\delta_n\}$ is a sequence of positive numbers, $L(X(\theta_n + \delta_n, \xi_{1,n}))$ and $L(X(\theta_n - \delta_n, \xi_{2,n}))$ are two measurements of $L(X(\theta, \xi))$ at $\theta_n + \delta_n$ and $\theta_n - \delta_n$, and $\xi_{1,n}, \xi_{2,n}$ are corresponding samples of ξ. Kiefer-Wolfowitz (1952) proved that if $J(\theta)$ is decreasing for $\theta < \theta^*$ and increasing for $\theta > \theta^*$, and if

$$\delta_n \to 0, \ \sum_n a_n = \infty, \ \sum_n a_n \delta_n < \infty, \ \sum_n a_n^2/\delta_n^2 < \infty,$$

the sequence $\{\theta_n\}$ converges to θ^* with probability one under some additional minor conditions. If all entries in $\{\xi_{i,n}\}$ are mutually independent, the best possible convergence rate for the KW algorithm (2) is $n^{-1/3}$ which is achieved by choosing $a_n = an^{-1}, \delta_n = dn^{-1/6}$ with $a, d > 0$ constants [e.g. Burkholder (1956); Fabian (1971); Sacks (1958)]. The rate $n^{-1/3}$ is regarded not satisfactory compared to the best possible rate $n^{-1/2}$ for the RM algorithm.

The MD algorithm, introduced by Nemirovski and Yudin (1983), improves the robustness of gradient based optimization algorithms. At iteration $n \geq 0$, θ_{n+1} is updated via solving

$$\theta_{n+1} = \text{argmin}_{\theta \in \Theta} \left\{ < h_n, \theta > + \frac{1}{a_n} D_\psi(\theta, \theta_n) \right\}, \tag{4}$$

where h_n is an estimate of the derivative $J'(\theta)$, $D : \Theta \times \Theta \to R^+$ is a Bregman distance defined as

$$D(\theta, \tau) := \psi(\theta) - \psi(\tau) - < \psi'(\tau), \theta - \tau > \geq \kappa ||\theta - \tau||^2, \tag{5}$$

where $\psi(.)$ is a distance generating function and $\kappa > 0$ is a constant. In (5), $||.||$ is a general norm on R^m (and on R in this paper). It has been established by Nemirovski et al. (2009) and Duchi et al. (2012,2013) that if $J(\theta)$ is convex, Lipschitz continuous and

$$a_n \to 0, \ \sum_n a_n = \infty,$$

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n \theta_i \text{ or } \hat{\theta}_n = \sum_{i=1}^n \nu_i \theta_i, \ \ \nu_i = \frac{a_i}{\sum_{j=1}^i a_j}, i = 1, 2, ...$$

then $J(\hat{\theta}_n)$ converges to the minimum of $J(\theta)$ and the rate of convergence is $n^{-1/2}$ under mild technical conditions that will be specified in Section 3.

The convergence of these algorithm is fairly understood [Burkholder (1956); Fabian (1971); Kushner and Clark (1978); Chung (1954); Dupač (1957); Dvoretzky (1956); Sacks

(1958); Wasan (1969), Nemirovski et al. (2009), Duchi et al. (2012, 2013)]. The conditions for the convergence of these algorithms can be made substantially weaker than those we have previously mentioned [e.g. Kushner and Clark (1978); Wasan (1969)]. The convergence rate for the RM algorithm is much faster than that for the KW algorithm. This is not surprising if we note that the KW algorithm uses the finite difference h_n as an approximation to the derivative $J'(\theta)$, while the RM algorithm uses an unbiased estimate of $J'(\theta)$. Therefore, the faster rate is achieved at the cost of obtaining an unbiased estimate of the derivative $J'(\theta)$ that is often challenging in practice. On the other hand, although its convergence is slower, the KW algorithm requires no detailed information on the function $J(\theta)$. It is simple to use and applicable to a wide range of problems. Kesten (1958) suggests that the stepsize a_n be chosen according to the fluctuation in the signs of g_n and h_n. A few of other techniques for the acceleration of stochastic approximation algorithms can be found in Wasan (1986). None of these accelerating techniques can improve the rate of convergence of the algorithms under study.

In this paper, we are interested in the acceleration of the KW algorithm and the MD algorithm through controlling the estimation of the derivative using finite differences. Furthermore, we consider the employment of the scheme of common random numbers (CRN) for improving the convergence of the algorithms — that is, the random factors $\xi_{1,n}$ and $\xi_{2,n}$ are chosen in such a manner that $\xi_{1,n} = \xi_{2,n} = \xi_n$. Implementation of CRN in Monte Carlo optimization is rather straightforward. The term "Monte Carlo optimization" is used here to refer to the procedure of finding the optimal solutions through computer simulation where the random factors, represented through a sequence of psuedo-random numbers, can be controlled [see Bratley et al. (1983)]. Computer simulation is often necessary when the form of $L(X(\theta, \xi))$ is too complicated. This is the case when $L(X(\theta, \xi))$ represents a performance measure of a stochastic system such as queueing systems, manufacturing systems, transportation systems, and communications networks [see, e.g. Ho and Cao (1991); Bratley et al. (1983); Law and Kelton (1982)]. In Section 6, we will give an example where the scheme of common random numbers is feasible.

We use the term CRN in a much narrow sense. The term CRN has more general and sometimes ill-posed meaning than we intend in this paper [see Glasserman and Yao (1992)]. In this paper, CRN simply refers to that simulation experiments be performed with the same stream of random numbers. As far as the KW algorithm (2) or the MD algorithm (4) is concerned, CRN requires that estimates of $J(\theta + \delta)$ and $J(\theta - \delta)$ be obtained from

simulation experiments using the same stream of random numbers $\{\xi_n\}$. Let $F(\theta, x)$ denote the distribution function of $X(\theta, \xi)$. Then any experiments with h_n constructed in the following form conform the CRN requirement:

$$\frac{L(Y_1(\theta_n, \delta_n, \xi_n)) - L(Y_2(\theta_n, \delta_n, \xi_n))}{2\delta_n} \tag{6}$$

where the marginal distributions of $Y_1(\theta_n, \delta_n, \xi_n)$ and $Y_2(\theta_n, \delta_n, \xi_n)$ are $F(\theta_n + \delta_n, x)$ and $F(\theta_n - \delta_n, x)$, respectively. Note that the joint distribution of $Y_1(\theta_n, \delta_n, \xi_n)$ and $Y_2(\theta_n, \delta_n, \xi_n)$ is left open, which may be used to improve the estimation variance. For a distribution function $F(\theta, x)$, its inverse function is defined as $F^{-1}(\theta, x) \stackrel{\text{def}}{=} \inf\{u \mid F(\theta, u) > x\}$. Cambanis and Simons (1976) and Whitt (1976) proved that the variance of (6) is minimized when $Y_1(\theta_n, \delta_n, \xi_n) = F^{-1}(\theta_n + \delta_n, \xi_n)$ and $Y_2(\theta_n, \delta_n, \xi_n) = F^{-1}(\theta_n - \delta_n, \xi_n)$. In this paper we assume that the form of $L(X(\theta, \xi))$ is given and fixed. The term CRN merely refers to the special choice of $\xi_{1,n} = \xi_{2,n}$. We will show that the use of CRN can significantly increase the rate of convergence for the KW algorithm (2) or the MD algorithm (4) from $n^{-1/3}$ to at least $n^{-2/5}$. For a large class of functions, the rate can be increased to $n^{-1/2}$, the best possible rate for stochastic approximation algorithms. CRN increases the rate of convergence by reducing the variance of h_n. Let $Var[X]$ denote the mathematical variance of a random variable X. Assume that $Var[L(X(\theta, \xi))]$ is continuous in θ, is bounded from below by a positive constant and from above by a constant. Then if $\xi_{1,n}$ and $\xi_{2,n}$ are independent, the variance of h_n is

$$(Var[L(X(\theta_n + \delta_n, \xi_{1,n}))] + Var[L(X(\theta_n - \delta_n, \xi_{2,n}))])/(2\delta_n)^2 = O(1/\delta_n^2)$$

which grows quadratically as δ_n goes to zero. We say a variable $f(s) = O(s)$ if $|f(s)/s| \leq C, C > 0$ is a constant independent of s ($f(s) = o(s)$ if $\lim |f(s)/s| = 0$ when s goes to zero or infinity depending on the context). It is such a large variance of h_n that slows down the convergence rate since, when δ_n is suitably chosen, the rate would be $n^{-1/2}$ if the variance of h_n is bounded. As we will show later, the convergence rate for (2) depends on how fast the variance of h_n goes to infinity. The slower the variance goes to infinity, the faster the convergence rate for (2) is. CRN has been observed effective for variance reduction for decades. It is perhaps the most popular method for variance reduction [Bratley et al. (1983); Conway (1963); Fishman (1974); Hammersley and Handscomb (1964); Heikes et al. (1976); Kleijnen (1974); Law and Kelton (1982)].

The rest of the paper is arranged as follows: In Section 2 we examine the rates of convergence for the KW algorithm under a very general setting that covers many interesting

situations. the analysis is extended to the MD algorithms in Section 4. In Section 4 we show that the use of CRN can reduce the variance of h_n by orders of magnitude, which in turn accelerates the convergence of the KW algorithm. In Section 5, we examine the rate of convergence for the MD algorithm under CRN. In Section 6, we extend the results to multivariates. A practical example is given to illustrate the feasibility of applying CRN in practice. Finally, a summary is provided in Section 7.

2. Rates of convergence for the KW algorithm. In this section, we examine the rates of convergence for the KW algorithm (2) under general assumptions on h_n. We do not assume that h_n is of the form (3). We will see later that such a treatment covers several important cases.

Assume that $\delta_n > 0$ goes to zero as $n \to \infty$ and, for $n \geq n_0 > 1$, h_n satisfies the following assumptions:

$$E[h_n|\theta_n] = J'(\theta_n) + \Delta_n, \quad |\Delta_n| \leq b\delta_n^\beta, \tag{7}$$

and

$$Var[h_n|\theta_n] \leq c\delta_n^\gamma, \tag{8}$$

where b, c, β are real nonnegative numbers, $\gamma \in R$. The form of (7) assures that h_n is an asymptotically unbiased estimate of $J'(\theta)$ when $\beta > 0$. When $\gamma > 0$, the variance of the estimate goes to zero as $n \to \infty$. This is generally impossible in practice. When $\gamma = 0$ such as in the RM algorithm, the variance is bounded. In the case that $\gamma < 0$, e.g. $\gamma = -2$ if h_n is defined by (3) and if $\xi_{1,n}$ and $\xi_{2,n}$ are independent, the variance of h_n goes to infinity. Next we examine the convergence and the rate of convergence for the KW algorithm (2). The commonly used criterion for measuring the convergence of a stochastic sequence $\{\theta_n\}$ is the *root mean square error* (RMSE) defined as

$$RMSE_{\theta_n} = (E[(\theta_n - \theta^*)^2])^{1/2}.$$

If $RMSE_{\theta_n} = O(n^{-s}), s > 0$, we say that $\{\theta_n\}$ converges at the rate of n^{-s} or the convergence rate for $\{\theta_n\}$ is n^{-s}.

We need the next lemma that was due to Chung (1954) and was formulated in the present form by Fabian (1971).

LEMMA 1. *Let s, t, B, A_n, b_n be real numbers, $0 < s \leq 1$, $t \geq 0$, $B > 0$. Define $b_+ = 0$ if $s < 1$ and $b_+ = t$ if $s = 1$ and assume that $c = \lim_{n\to\infty} A_n - b_+$ exists and is finite. If for*

$n \geq n_0$,

$$b_{n+1} \leq b_n(1 - \frac{A_n}{n^s}) + \frac{B}{n^{s+t}}$$

and if $c > 0$, *then*

$$\lim_{n\to\infty} \sup n^t b_n \leq B/c.$$

The statement remains valid if all the inequalities are reversed and lim sup *is replaced by* lim inf.

The following Theorems 1 and 2 give the convergence rate for the KW algorithm (2) with h_n satisfying (7)-(8):

THEOREM 1. *Assume that* $\{\theta_n\}$ *is determined by (2) and*

(A1). $a_n = an^{-\alpha}, \delta_n = dn^{-\eta}$, $0 < \alpha \leq 1$, $\eta > 0$, $a, d > 0$;

(A2). $J(\theta)$ *is increasing for* $\theta < \theta^*$ *and decreasing for* $\theta > \theta^*$, *and there exist two constants* K_1, K_2, $0 < K_1 \leq K_2 < \infty$, *such that for all* $\theta \in \Theta$,

$$K_1|\theta - \theta^*| \leq |J'(\theta)| \leq K_2|\theta - \theta^*|;$$

(A3). conditioned on θ_n, h_n *at the* n*th iteration is independent of those at the other iterations.*

Then, if $\sigma = (1/2)\min\{\alpha + \gamma\eta, 2\beta\eta\}$ *and* $0 < \sigma < aK_1$, *we have*

$$\lim_{n\to\infty} \sup n^{2\sigma} E[(\theta_n - \theta^*)^2] \leq C \tag{9}$$

where $C > 0$ *is a constant. The convergence rate for* $RMSE_{\theta_n}$ *is at least* $n^{-\sigma}$.

PROOF. Without loss of generality, we assume that $\theta^* = 0$. Then

$$\begin{aligned} E[\theta_{n+1}^2] &= E[\theta_n^2] - 2a_n E[\theta_n h_n] + a_n^2 E[h_n^2] \\ &= E[\theta_n^2] - 2a_n E[\theta_n(J'(\theta_n) + \Delta_n)] + a_n^2((E[h_n])^2 + Var[h_n]). \end{aligned}$$

According to (7)-(8), we have

$$E[\theta_{n+1}^2] \leq E[\theta_n^2] - 2a_n E[\theta_n J'(\theta_n)] + 2ba_n\delta_n^\beta E[|\theta_n|] + 2a_n^2(E[J'(\theta_n)]^2 + b^2\delta_n^{2\beta}) + ca_n^2\delta_n^\gamma. \tag{10}$$

By Assumption (A2), $\theta_n J'(\theta_n) \geq 0$ and

$$\theta_n J'(\theta_n) \geq K_1\theta_n^2, \;\; (J'(\theta_n))^2 \leq K_2^2\theta_n^2. \tag{11}$$

Furthermore, for any $\epsilon_n > 0$, $|\theta_n| \leq \epsilon_n + \theta_n^2/\epsilon_n$ and consequently

$$E[|\theta_n|] \leq \epsilon_n + \frac{1}{\epsilon_n} E[\theta_n^2].$$

By setting $0 < \epsilon < 1$ and

$$\epsilon_n = \frac{2b\delta_n^\beta}{K_1\epsilon},$$

we have

$$E[|\theta_n|] \leq \frac{2b\delta_n^\beta}{K_1\epsilon} + \frac{K_1\epsilon}{2b\delta_n^\beta} E[\theta_n^2]. \tag{12}$$

Substituting (11) and (12) into (10), we obtain

$$E[\theta_{n+1}^2] \leq E[\theta_n^2][1 - (2-\epsilon)K_1 a_n + 2K_2^2 a_n^2] + 2b^2 a_n^2 \delta_n^{2\beta} + c a_n^2 \delta_n^\gamma + \frac{4b^2}{K_1\epsilon} a_n \delta_n^{2\beta}. \tag{13}$$

According to (13), also noting Assumption (A1), we can choose an $n_1 \geq n_0 > 1$ such that for all $n \geq n_1$

$$E[\theta_{n+1}^2] \leq E[\theta_n^2](1 - \frac{A_n}{n^\alpha}) + \frac{B}{n^{\alpha+2\sigma}}$$

where

$$A_n = (2-\epsilon)aK_1 - \frac{2K_2^2 a^2}{n^\alpha}, \quad B = 2a^2 b^2 d^{2\beta} + c a^2 d^\gamma + \frac{4ab^2 d^{2\beta}}{K_1\epsilon}.$$

If $aK_1 > \sigma$, we can always choose $\epsilon > 0$ so small that $(2-\epsilon)aK_1 > 2\sigma$. Applying Lemma 1, we obtain (9) with $C = B/((2-\epsilon)aK_1)$ if $\alpha < 1$ and $C = B/((2-\epsilon)aK_1 - 2\sigma)$ if $\alpha = 1$. ∎

It follows directly from Theorem 1 that $\{\theta_n\}$ converges to θ^* as long as $\sigma > 0$, or equivalently, as long as $\alpha + \gamma\eta > 0$. When $\alpha + \gamma\eta \leq 0$ which is possible only when $\gamma < 0$, the variance of h_n grows to infinity at the rate of n^t with $t = -\gamma\eta \geq \alpha$. It is obvious from (2) that $\{\theta_n\}$ does not converge. Another extreme case is that $\gamma > 0$. In this case, σ can be made arbitrarily large by choosing appropriate η. The convergence rate for $\{\theta_n\}$ can be made arbitrarily large if η can take any value. In fact, by setting $\eta \to \infty$ in (13) (or equivalently, $\delta_n \to 0$) and $a_n = a$ such that $0 < q = 1 - (2-\epsilon)K_2 a + 2K_2^2 a^2 < 1$ for sufficiently large n, we have

$$E[\theta_{n+1}^2] \leq qE[\theta_n^2].$$

The convergence rate for the sequence $\{\theta_n\}$ is that of a geometric progression. Unfortunately, this is a very special case. One should not expect $\gamma > 0$ in practice. Both of the situations $\gamma > 0$ and $\alpha + \gamma\eta \leq 0$ are too special to deserve further study. The most interesting case is when γ satisfies $-\alpha/\eta < \gamma \leq 0$.

Theorem 1 shows that, when h_n satisfies (7)-(8), $\{\theta_n\}$ converges with probability one to the optimal parameter θ^* at a rate of at least $n^{-\sigma}$. We can further prove that $\{\theta_n\}$ converges exactly at this rate as interpreted in the following Theorem 2.

THEOREM 2. *Assume that Assumptions (A1)-(A3) are satisfied.*

1. *If* $\alpha+\gamma\eta < 2\beta\eta$, $aK_2 > \sigma$, $E[h_n|\theta_n] = J'(\theta_n)+\Delta_n$, $|\Delta_n| \le b\delta_n^\beta$, *and* $Var[h_n|\theta_n] \ge c\delta_n^\gamma$, *we have*

$$\lim_{n\to\infty} \inf n^{2\sigma} E[(\theta_n - \theta^*)^2] \ge C_1$$

where $C_1 > 0$ *is a constant.*

2. *If* $\alpha+\gamma\eta \ge 2\beta\eta$, $E[h_n|\theta_n] = J'(\theta_n) + b\delta_n^\beta(1+\varepsilon_n)$, *and* $J'(\theta_n) = (\theta_n-\theta^*)(K_3+\tau_n)$, $\varepsilon_n = o(1)$ *and* $\tau_n = o(1)$ *uniformly as* $n\to\infty$, $K_3 = J''(\theta^*) > 0$, $\sigma < aK_3$, *then*

$$\lim_{n\to\infty} \sup n^{\sigma} E[\theta_n - \theta^*] \le -C_2$$

where $C_2 > 0$ *is a constant.*

PROOF. Let consider the first statement. For simplicity and without loss of generality, we assume $\theta^* = 0$. Parallel to the derivation of (10) we have

$$E[\theta_{n+1}^2] = E[\theta_n^2] - 2a_n E[\theta_n(J'(\theta_n)+\Delta_n)] + a_n^2\{(E[h_n])^2 + Var[h_n]\}$$

which implies

$$E[\theta_{n+1}^2] \ge E[\theta_n^2] - 2a_n E[\theta_n J'(\theta_n)] - 2ba_n\delta_n^\beta E[|\theta_n|] + ca_n^2\delta_n^\gamma.$$

Assumption (A2) implies that $0 \le \theta_n J'(\theta_n) \le K_2\theta_n^2$ which, together with (12) where K_1 is replaced by K_2, shows that

$$E[\theta_{n+1}^2] \ge E[\theta_n^2](1-(2-\epsilon)K_2a_n) + ca_n^2\delta_n^\gamma - \frac{4b^2}{K_2\epsilon}a_n\delta_n^{2\beta}. \tag{14}$$

If $\alpha+\gamma\eta < 2\beta\eta$, there exists an $n_0 > 1$ such that when $n \ge n_0$

$$ca_n^2\delta_n^\gamma - \frac{4b^2}{K_2\epsilon}a_n\delta_n^{2\beta} \ge \frac{1}{2}ca_n^2\delta_n^\gamma.$$

Therefore, we know from (14) that when $n \ge n_0$

$$E[\theta_{n+1}^2] \ge E[\theta_n^2](1-\frac{A_n}{n^\alpha}) + \frac{ca^2d^\gamma}{2n^{\alpha+2\sigma}}.$$

Since $aK_2 > \sigma$, we can always choose $\epsilon > 0$ so small that $(2-\epsilon)aK_2 > 2\sigma$. The first statement of the theorem follows from applying Lemma 1 with $C_1 = ca^2d^\gamma/(2(2-\epsilon)aK_2)$ if $\alpha < 1$ and $C_1 = ca^2d^\gamma/(2(2-\epsilon)aK_2 - 4\sigma)$ if $\alpha = 1$.

If $\alpha + \gamma\eta \geq 2\beta\eta$, we know that $\sigma = \beta\eta > 0$ and

$$\begin{aligned} E[\theta_{n+1}] &= E[\theta_n] - a_n E[J'(\theta_n)] - ba_n\delta_n^\beta(1+\varepsilon_n) \\ &= E[\theta_n](1-K_3a_n) + a_nE[\tau_n\theta_n] - ba_n\delta_n^\beta(1+E[\varepsilon_n]) \end{aligned} \tag{15}$$

Define $z_n = n^\sigma E[\theta_n]$. Then (15) shows that

$$\begin{aligned} z_{n+1} &= n^\sigma(1+\frac{1}{n})^{n^\sigma}[E[\theta_n](1-K_3a_n) + a_nE[\tau_n\theta_n] - ba_n\delta_n^\beta(1+\varepsilon_n)] \\ &= z_n[1+\frac{\sigma}{n} - K_3a_n - \frac{\sigma}{n}K_3a_n + O(\frac{1}{n^2})] + (1+\frac{1}{n})^\sigma n^\sigma(a_nE[\tau_n\theta_n] - ba_n\delta_n^\beta(1+E[\varepsilon_n])). \end{aligned}$$

Denote

$$A_n = 1 + \frac{\sigma}{n} - K_3a_n - \frac{\sigma}{n}K_3a_n + O(\frac{1}{n^2}),$$

$$B_n = (1+\frac{1}{n})^\sigma n^\sigma(ba_n\delta_n^\beta(1+E[\varepsilon_n]) - a_nE[\tau_n\theta_n]).$$

Then

$$z_{n+1} = A_nz_n - B_n. \tag{16}$$

Note that $a_n = an^{-\alpha}, 0 < \alpha \leq 1, \delta_n = dn^{-\eta}, aK_3 > \sigma$. We may choose $\tilde{A}_1, \tilde{A}_2 > 0, n_1 > 1$ such that, for all $n \geq n_1$,

$$0 \leq 1 - \frac{\tilde{A}_1}{n^\alpha} \leq A_n = 1 + \frac{\sigma}{n} - \frac{aK_3}{n^\alpha} - \frac{a\sigma K_3}{n^{1+\alpha}} + O(\frac{1}{n^2}) \leq 1 - \frac{\tilde{A}_2}{n^\alpha}. \tag{17}$$

Since Assumptions (A1)-(A3) in Theorem 1 are satisfied, $\lim_{n\to\infty}\sup n^{2\sigma}E[\theta_n^2] \leq C$ which implies that

$$\lim_{n\to\infty}\sup n^\sigma|E[\theta_n]| \leq \lim_{n\to\infty}\sup(n^{2\sigma}E[\theta_n^2])^{1/2} \leq \sqrt{C}.$$

According to the assumptions that $\epsilon_n = o(1)$, $\tau_n = o(1)$ uniformly as $n \to \infty$, and $\delta_n^\beta = d^\beta n^{-\sigma}$. There exists an $n_2 > 1$ such that, when $n \geq n_2$,

$$\begin{aligned} B_n &= (1+\frac{1}{n})^\sigma n^\sigma(ba_n\delta_n^\beta(1+E[\varepsilon_n]) - a_nE[\tau_n\theta_n]) \\ &\geq (1+\frac{1}{n})^\sigma n^\sigma(ba_n\delta_n^\beta(1+E[\varepsilon_n]) - a_nE[|\tau_n|]E[|\theta_n|]) \\ &\geq (1+\frac{1}{n})^\sigma n^\sigma\frac{1}{2}ba_n\delta_n^\beta \geq \frac{1}{2}n^\sigma ba_n\delta_n^\beta. \end{aligned} \tag{18}$$

Let $n_0 = \max\{n_1, n_2\}$. Then, from (16) we know that for all $n \geq n_0$

$$z_n = z_{n_0} \prod_{i=n_0}^{n} A_i - \sum_{i=n_0}^{n-1} B_i \prod_{j=i+1}^{n} A_j - B_n. \tag{19}$$

Since $0 < \alpha \leq 1$, (17) shows that

$$0 \leq \lim_{n\to\infty} \prod_{i=n_0}^{n} A_i \leq \lim_{n\to\infty} \prod_{i=n_0}^{n} (1 - \frac{\tilde{A}_2}{i^\alpha}) = 0. \tag{20}$$

Furthermore, $\lim_{n\to\infty} B_n = 0$, and (17) and (18) imply that

$$\sum_{i=n_0}^{n-1} B_i \prod_{j=i+1}^{n} A_j \geq \sum_{i=n_0}^{n-1} \frac{abd^\beta}{2i^\alpha} \prod_{j=i+1}^{n} A_j \geq \frac{abd^\beta}{2n^\alpha} \sum_{i=n_0}^{n-1} A_n^{n-i} \geq \frac{abd^\beta}{2n^\alpha} \sum_{i=n_0}^{n-1} (1 - \frac{\tilde{A}_1}{n^\alpha})^{n-i}. \tag{21}$$

On the other hand,

$$\lim_{n\to\infty} \frac{1}{n^\alpha} \sum_{i=n_0}^{n-1} (1 - \frac{\tilde{A}_1}{n^\alpha})^{n-i} = \begin{cases} 1, & \text{if } 0 < \alpha < 1 \\ 1 - e^{-\tilde{A}_1}, & \text{if } \alpha = 1. \end{cases}$$

Substituting the preceding inequality, (20) and (21) into (19), we see that

$$\lim_{n\to\infty} \sup z_n \leq -C_2$$

with $C_2 = (1/2)abd^\beta(1 - e^{-\tilde{A}_1}) > 0$. This is exactly what we want to prove. ∎

Theorems 1 and 2 show that the convergence rate for $\{\theta_n\}$ is generally $n^{-\sigma}$. If we are free to choose the positive numbers α, η, it follows directly from Theorem 1 that

COROLLARY 1. *Assume that h_n satisfies (7)-(8) and $\gamma \leq 0$. Under Assumptions (A2)-(A3) in Theorem 1, the best possible convergence rate for the KW algorithm (2) is $n^{-\beta/(2\beta-\gamma)}$ which is achieved by setting $\alpha = 1$, $\eta = 1/(2\beta - \gamma)$, and by choosing appropriate $a, d > 0$.*

For the KW algorithm (2) with h_n defined by (3), assume that $J(\theta)$ is continuously differentiable of order up to three and the third order derivative $J'''(\theta)$ is uniformly bounded on Θ, we have

$$E[h_n|\theta_n] = \frac{J(\theta_n + \delta_n) - J(\theta_n - \delta_n)}{2\delta_n} = J'(\theta_n) + \frac{1}{6}J'''(\tilde{\theta}_n)\delta_n^2 = J'(\theta_n) + O(\delta_n^2)$$

where $\tilde{\theta}_n \in [\theta_n - \delta_n, \theta_n + \delta_n]$. In this case, $\beta = 2$. If the assumptions (8) and (A1)-(A3) are satisfied and if the positive number a is chosen sufficiently large, we know from Theorems 1 and 2 that

$$\sigma = \frac{1}{2} \min\{\alpha + \gamma\eta, 4\eta\}. \tag{22}$$

If we use the one-sided finite-difference approximation in (2):

$$h_n = \frac{L(X(\theta_n + \delta_n, \xi_{1,n})) - L(X(\theta_n, \xi_{2,n}))}{\delta_{2,n}} \tag{23}$$

and if $J(\theta)$ is twice continuously differentiable and the second order derivative $J''(\theta)$ is bounded on Θ, then for any θ_n, δ_n there exists a $\hat{\theta}_n \in [\theta_n, \theta_n + \delta_n]$ such that

$$E[h_n|\theta] = \frac{J(\theta_n + \delta_n) - J(\theta_n)}{\delta_n} = J'(\theta_n) + \frac{1}{2}J''(\hat{\theta}_n)\delta_n = J'(\theta_n) + O(\delta_n).$$

Therefore, $\beta = 1$. Under the same conditions as those in the previous case we know that

$$\sigma = \frac{1}{2}\min\{\alpha + \gamma\eta, 2\eta\}. \tag{24}$$

It is clear from (22) and (24) that, under the same condition for the variance $Var[h_n|\theta_n]$, the convergence rate of the KW algorithm is faster when symmetric differences are used than that when one-sided differences are used. Corollary 1 shows that the best possible convergence rate depends on two factors—how fast the bias decreases to zero and how slow the variance increases to infinity. Using symmetric finite difference (3) instead of the one-sided finite difference (23) can reduce the bias of h_n. To summarize, we have the following conclusion which will be used later.

COROLLARY 2. *Suppose that (A1)-(A3) are satisfied. If*

(A4). $J(\theta)$ is continuously differentiable of order up to three and the third order derivative $J'''(\theta)$ is bounded on Θ,

then the best possible convergence rate for the KW algorithm (2) with h_n defined in (3) is $n^{-2/(4-\gamma)}$. If

(A5). $J(\theta)$ is twice continuously differentiable and the second order derivative $J''(\theta)$ is bounded on Θ,

then the best possible convergence rate for the KW algorithm (2) with h_n defined in (23) is $n^{-1/(2-\gamma)}$.

3. Rates of convergence for the MD algorithm. The rate of convergence of the MD algorithms was established by Nemivoski et al. (2009) when the h_n in (4) is an unbiased estimate of the derivative, and by Duchi et al. (2012, 2013) when the h_n is approximated by

the one-sided finite difference (23). In this section, we examine the rate of convergence of the MD algorithm for general h_n. Again, we only assume that h_n satisfies (7)-(8). For notational consistence, the norm $||.||$ in (4) is taken as the l_2 norm. Its dual norm $||x||_* := \sup_{||y||\leq 1} y^T x$ is also the l_2 norm. Define

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\theta_i.$$

We next examine the convergence of $J(\hat{\theta}_n)$.

THEOREM 3. *Assume that $\{\theta_n\}$ is determined by (4), and*

(B1). $\psi(\theta)$ is strongly convex, Θ is compact and convex, and there exists $r > 0$ such that $D(\theta^*, \theta) \leq (1/2)r^2, r > 0$ for all $\theta \in \Theta$;

(B2). $L(X)$ is closed convex, and there exist two constants $K_1, K_2, 0 < K_1 \leq K_2 < \infty$, such that for all $\theta \in \Theta$,

$$K_1|\theta - \theta^*| \leq |J'(\theta)| \leq K_2|\theta - \theta^*|;$$

(B3). conditioned on θ_n, h_n at the nth iteration is independent of those at the other iterations.

Then

$$E[J(\hat{\theta}_n) - J(\theta^*)] \leq \frac{C_1}{na_n} + \frac{C_2}{n}\sum_{i=1}^{n} a_i\delta_i^{\gamma} + \frac{C_3}{n}\sum_{i=1}^{n} a_i + \frac{C_4}{n}\sum_{i=1}^{n} a_i\delta_i^{2\beta} + \frac{C_5}{n}\sum_{i=1}^{n}\delta_i^{\beta}, \tag{25}$$

where

$$C_1 = \frac{r^2}{2},\ \ C_2 = \frac{c}{2\kappa},\ \ C_3 = \frac{K_2^2 r^2}{2\kappa^2},\ \ C_4 = \frac{b^2}{\kappa},\ \ C_5 = \frac{br}{\sqrt{2\kappa}}.$$

PROOF. Under Assumptions (B1)-(B3), we know from Duchi et al. (2013), eqn. (13), that

$$J(\hat{\theta}_n) - J(\theta^*) \leq \frac{r^2}{2na_n} + \frac{1}{2n\kappa}\sum_{i=1}^{n} a_i h_i^2 - \frac{1}{n}\sum_{i=1}^{n}\Delta_i(\theta_i - \theta^*). \tag{26}$$

Therefore,

$$E[J(\hat{\theta}_n) - J(\theta^*)] \leq \frac{r^2}{2na_n} + \frac{1}{2n\kappa}\sum_{i=1}^{n} a_i E[h_i^2] + \frac{1}{n}\sum_{i=1}^{n} E[|\Delta_i(\theta_i - \theta^*)|]. \tag{27}$$

The assumptions (7)-(8) give

$$E[h_i^2] = Var[h_i] + (E[J'(\theta_i) + \Delta_i])^2 \leq c\delta_i^{\gamma} + 2(E[J'(\theta_i)])^2 + 2b^2\delta_i^{2\beta} \tag{28}$$

On the other hand, according to Assumptions (B1)-(B2),

$$(E[J'(\theta_i)])^2 \leq (K_2 E[|\theta_i - \theta^*|])^2 \leq \frac{K_2^2}{2\kappa} r^2 \tag{29}$$

and

$$E[|\Delta_i||\theta_i - \theta^*|] \leq \frac{b\delta_i^\beta r}{\sqrt{2\kappa}}. \tag{30}$$

By combining (28)-(30) with (27), we obtain

$$E[J(\hat{\theta}_n) - J(\theta^*)] \leq \frac{r^2}{2na_n} + \frac{c}{2n\kappa}\sum_{i=1}^n a_i\delta_i^\gamma + \frac{K_2^2 r^2}{2n\kappa^2}\sum_{i=1}^n a_i + \frac{b^2}{n\kappa}\sum_{i=1}^n a_i\delta_i^{2\beta} + \frac{br}{n\sqrt{2\kappa}}\sum_{i=1}^n \delta_i^\beta,$$

which is exactly (25). ∎

A special case of interest is $\gamma = 0$, which corresponds to bounded variance of derivative estimation. The following Corollary 3 provides a bound on the convergence of the MD algorithm for this case with properly chosen $\{a_n\}$, $\{\delta_n\}$.

COROLLARY 3. *Suppose that (A4), (B1)-(B3) are satisfied. Let* $\gamma = 0$, $a_n = an^{-1/2}$, $\delta_n = dn^{-1}$, $a > 0, d > 0$.
(a) For h_n *defined in (23),*

$$E[J(\hat{\theta}_n) - J(\theta^*)] \leq (C_1 + 2C_2 + 2C_3)\frac{\max(a, a^{-1})}{\sqrt{n}} + (2.5C_4 d^2)\frac{a}{n} + (C_5 d)\frac{1+\log n}{n}. \tag{31}$$

(b) For h_n *defined in (3),*

$$E[J(\hat{\theta}_n) - J(\theta^*)] \leq (C_1 + 2C_2 + 2C_3)\frac{\max(a, a^{-1})}{\sqrt{n}} + (9C_4 d^4/7)\frac{a}{n} + (2C_5 d^2)\frac{1}{n}. \tag{32}$$

PROOF. For $\gamma = 0$, $a_n = an^{-1/2}$, $\delta_n = dn^{-1}$, combining the first three terms in (25) gives the first term in (31). For h_n defined by the one-sided finite difference (23), under Assumption (A4), we have $\beta = 1$. Consequently, the fourth term in (25) is

$$\frac{C_4}{n}\sum_{i=1}^n a_i\delta_i^2 = \frac{C_4 a d^2}{n}\sum_{i=1}^n i^{-2.5} \leq \frac{C_4 a d^2 2.5}{n}.$$

The last term is

$$\frac{C_5}{n}\sum_{i=1}^n \delta_i = \frac{C_5 d}{n}\sum_{i=1}^n i^{-1} \leq \frac{C_5 d(1+\log n)}{n}.$$

Combing the previous two inequalities with (25) gives (31).

For h_n defined by the symmetric finite difference (3), under Assumption (A4), we have $\beta = 2$. Consequently, the fourth term in (25) is

$$\frac{C_4}{n}\sum_{i=1}^{n} a_i\delta_i^4 = \frac{C_4 a d^4}{n}\sum_{i=1}^{n} i^{-4.5} \leq \frac{C_4 a d^4 (9/7)}{n}.$$

The last term is

$$\frac{C_5}{n}\sum_{i=1}^{n} \delta_i^2 = \frac{C_5 d^2}{n}\sum_{i=1}^{n} i^{-1} \leq \frac{2C_5 d^2}{n}.$$

Combing the previous two inequalities with (25) gives (32). ∎

Duchi et al. (2013) investigated the convergence of the MD algorithm using the one-sided finite difference (23) as an approximation to the derivative. The bound (31) is technically the same as that in Duchi et al. (2013). When the symmetric finite difference (3) is used, the $\log n$ factor disappears in the last term of (32), which indicates that the symmetric finite-difference approximation (3) leads to a tighter bound under similar assumptions, which is due to that the symmetric finite difference (3) typically provides more accurate estimate of the mean of the derivative than the one-sided ones do. Note that Duchi et al. (2012, 2013) implicitly assumes that CRN is used in calculating the finite difference (3) or (23) that will be covered in Sections 4 and 5.

It's worth of noting that the rate of convergence for the MD algorithm, as given by (25), is $n^{-1/2}$ which is not affected by the choice of finite-difference approximation, either symmetric or one-sided, to the derivative. This is by design since the MD algorithm was originally proposed for improving the robustness in the choice of stepsizes at the cost of slower convergence.

When a finite difference is used to approximate the derivative, it is desirable to have $\delta_n \to 0$ as $n \to \infty$ to ensure asymptotically unbiased estimate of the derivative. In this case, it is possible (and likely in practice!) that the variance of the estimate goes to infinity. This is a special case of (25) with $\gamma < 0$. Therefore, Theorem 3 allows flexibility to cover general cases.

A special situation is when $L(X(\theta_n + \delta_n, \xi_{1,n}))$ and $L(X(\theta_n - \delta_n, \xi_{2,n}))$ or $L(X(\theta_n, \xi_{2,n}))$ in (3) or (23) are sampled independently. In this case, $\gamma = -2$. Assume further that $\{a_n\}$ and $\{\delta_n\}$ are specified as in Assumption (A1). Then the right hand side of (25) becomes

$$H(n) := \frac{C_1}{na_n} + \frac{C_2}{n}\sum_{i=1}^{n} a_i\delta_i^{\gamma} + \frac{C_3}{n}\sum_{i=1}^{n} a_i + \frac{C_4}{n}\sum_{i=1}^{n} a_i\delta_i^{2\beta} + \frac{C_5}{n}\sum_{i=1}^{n} \delta_i^{\beta}$$

$$= \frac{C_1}{an^{1-\alpha}} + \frac{C_2}{n}\sum_{i=1}^{n} a\delta^{-2} i^{-\alpha+2\eta} + \frac{C_3}{n}\sum_{i=1}^{n} a i^{-\alpha} + \frac{C_4}{n}\sum_{i=1}^{n} a\delta^{2\beta} i^{-\alpha-\beta\eta} + \frac{C_5}{n}\sum_{i=1}^{n} \delta i^{-\beta\eta}$$

$$= O(n^{-1+\alpha}) + O(n^{-\alpha+2\eta}) + O(n^{-\alpha}) + O(n^{-\alpha-\beta\eta}) + O(n^{-\beta\eta})$$

$$= O(n^{-\sigma}),$$

where

$$\sigma = \min\{1-\alpha, \alpha-2\eta, \alpha, \alpha+2\beta\eta, \beta\eta\} = \min\{1-\alpha, \alpha+2\eta, \beta\eta\}.$$

For the one-sided finite difference (23), $\beta = 1$. Then

$$\sigma = \min\{1-\alpha, \alpha+2\eta, \eta\} \leq 1/4.$$

For the symmetric finite difference (3), $\beta = 2$. Then

$$\sigma = \min\{1-\alpha, \alpha+2\eta, 2\eta\} \leq 1/3.$$

The previous discussion can be summarized in the following Corollary 4.

COROLLARY 4. *Assume that Assumptions (A1), (A4), (B1)-(B3) are satisfied, and that $L(X(\theta_n+\delta_n, \xi_{1,n}))$ and $L(X(\theta_n-\delta_n, \xi_{2,n}))$ in (3) (or $L(X(\theta_n, \xi_{2,n}))$ in (23)) are independent. Then*

(i) the best possible rate of convergence for the upper bound $H(n)$ is $n^{-1/4}$ when the one-sided finite difference (23) is used,

(ii) the best possible rate of convergence for the upper bound $H(n)$ is $n^{-1/3}$ when the symmetric finite difference (3) is used.

Note that the rates of convergence are only upper bounds of $E[J(\hat{\theta}_n)]$. Such rates of convergence are consistent with those for $\{\theta_n\}$.

4. The KW algorithm with CRN. In this section, we will show how CRN can accelerate the convergence of the KW algorithm. For clarity and without getting trapped into unnecessary tediousness of details, we focus our attention on the case in which $\xi \in R$ is a real one-dimensional random variable. In Monte Carlo optimization, ξ is usually a psuedo-random number generated by a computer. For most applications, such a pseudo-random number is sufficiently good to be regarded as a random number uniformly distributed on $[0, 1)$. In Section 6, we extend the results to general situations.

To avoid repetition, we only consider the h_n defined as in (3) with $\xi_{1,n} = \xi_{2,n} = \xi_n$. The analysis is applicable to the one-sided finite-difference approximation (23) without any difficult. For a given θ_n, h_n is a finite-difference approximation to the derivative $J'(\theta)$ at $\theta = \theta_n$. For simplicity, we omit the subscript n. Then

$$h = \frac{L(X(\theta + \delta, \xi)) - L(X(\theta - \delta, \xi))}{2\delta}. \tag{33}$$

The mean of h is

$$E[h] = \frac{J(\theta + \delta) - J(\theta - \delta)}{2\delta}$$

which is the same as that of (3) without the use of CRN. However, the variance of (33), as we will show, is generally smaller than that of (3) without the use of CRN when $\delta > 0$ is sufficiently small. We will also show that the reduction in the variance of h may have a significant impact on the convergence rate for the KW algorithm for Monte Carlo optimization. Toward that end, we need to specify the generation of the random variable $X(\theta, \xi)$ with a given distribution $F(\theta, x)$. Next we examine the variance of (33) for several popular random number generation methods. Note that

$$Var[h] = \frac{1}{(2\delta)^2}\{E[(L(X(\theta + \delta, \xi)) - L(X(\theta - \delta, \xi)))^2] + (J(\theta + \delta) - J(\theta - \delta))^2\}.$$

If $J(\theta)$ is continuously differentiable on Θ with bounded derivatives, then

$$Var[h] = \frac{1}{(2\delta)^2}E[(L(X(\theta + \delta, \xi)) - L(X(\theta - \delta, \xi)))^2] + O(1). \tag{34}$$

4.1. Inversion method. Inversion is one of the most popular methods for random variable generation. Let $F(\theta, x)$ be the distribution function of $X(\theta, \xi)$. The inversion method generates the random variable $X(\theta, \xi)$ in the following way:

1. Generate a random number ξ uniformly distributed on $[0, 1)$.

2. Set $X(\theta, \xi) = F^{-1}(\theta, \xi)$.

Then it is straightforward to verify that $X(\theta, \xi)$ has the desired distribution. Note that the mapping $F(\theta, x) : R \to R$ is not one to one in general. To ensure its existence for general distribution functions, the inverse function is defined as

$$F^{-1}(\theta, \xi) = \min\{x \mid F(\theta, x) > \xi,\ x \in R\}$$

which is different from the usual definition [see Krantz (1991)]. It coincides with the usual definition if $F(\theta, x)$ is continuous and strictly increasing. Such a definition of the inverse function covers both continuous and discrete random variables. For example, consider a discrete random variable $X(\theta, \xi) = x_i$ with probability $p_i(\theta)$. Define $\rho_0(\theta) = 0, \rho_i(\theta) = \sum_{j=1}^{i} p_j(\theta)$ for $i \geq 1$. Let ξ be uniformly distributed on $[0, 1)$. The inversion method gives $F^{-1}(\theta, \xi) = x_i$ if $\xi \in [\rho_{i-1}(\theta), \rho_i(\theta))$. Then direct verification shows that $X(\theta, \xi)$ obeys the desired distribution. This is a discrete version of the inversion method.

In order to proceed with our discussion, let us first examine the properties of distribution functions. A distribution $F(\theta, x)$ is a nondecreasing and right-continuous function of x. $F(\theta, x)$ has at most countably many points of discontinuity on R and all of the discontinuities are of the first kind — that is, for any $x \in R$, $F(\theta, x^-) = \lim_{y \uparrow x} F(\theta, y)$ and $F(\theta, x^+) = \lim_{y \downarrow x} F(\theta, y)$ exist and are finite [e.g. Krantz (1991),149-150]. Therefore, we can divide R into $\bigcup_i B_i(\theta) = R$, where $B_i(\theta) = [b_i(\theta), b_{i+1}(\theta))$, such that, for each i, $F(\theta, x)$ is continuous on $B_i(\theta)$, but jumps at $b_i(\theta)$. Assume that, for each i, $F(\theta, x)$ is piecewise differentiable on $B_i(\theta)$. Then $F'_x(\theta, x) > 0$ whenever it exists. We further divide the interval $B_i(\theta)$ into subintervals according to whether the derivative of $F(\theta, x)$ with respect to x is zero or not. For simplicity, we assume that $B_i(\theta) = B_i^0(\theta) \bigcup B_i^+(\theta)$ such that $F'_x(\theta, x) = 0$ on $B_i^0(\theta) = [b_i(\theta), c_i(\theta)]$ and $F(\theta, x) = F_i(\theta, x)$ is continuously differentiable with strictly positive derivatives on $B_i^+(\theta) = (c_i(\theta), b_{i+1}(\theta))$. It is possible that $b_i(\theta) = c_i(\theta)$. On $B_i^0(\theta)$, the derivatives $F'_x(\theta, x)$ should be understood as the right and the left derivatives at $b_i(\theta), c_i(\theta)$, respectively. It is possible that $F(\theta, x)$ is not differentiable at $c_i(\theta)$. The inverse $F_i^{-1}(\theta, \xi)$ is defined in the usual sense. It is continuous, strictly increasing, and differentiable on $(F(\theta, c_i(\theta)), F(\theta, b_{i+1}^-(\theta)))$.

Under the preceding decomposition, $F(\theta, x)$ is discontinuous at $b_i(\theta)$, is a constant on $B_i^0(\theta)$, and is strictly increasing and differentiable on $B_i^+(\theta)$.

The following Lemma 2 follows directly from the definition of the inverse function and the decomposition of $F(\theta, x)$.

LEMMA 2. *Let $X(\theta, \xi)$ be defined by the inverse function $X(\theta, \xi) = F^{-1}(\theta, \xi)$. Let $\Xi_i(\theta) = [F(\theta, b_i^-(\theta)), F(\theta, b_{i+1}^-(\theta)))$. Then $\Xi_i(\theta) \subset [0, 1)$ and for any $\xi \in \Xi_i(\theta)$*

$$X(\theta, \xi) = \begin{cases} b_i(\theta), & \text{if } \xi \in [F(\theta, b_i^-(\theta)), F(\theta, c_i(\theta))), \\ c_i(\theta), & \text{if } \xi = F(\theta, c_i(\theta)), \\ F_i^{-1}(\theta, \xi), & \text{if } \xi \in (F(\theta, c_i(\theta)), F(\theta, b_{i+1}^-(\theta))). \end{cases} \tag{35}$$

We need the following result.

LEMMA 3. *Assume that*

(C1). $L(X)$ and $L'_X(X)$ are bounded, $J(\theta)$ is continuously differentiable on Θ;

(C2). for each i, $F_i(\theta, x)$ is continuously differentiable on $B_i^+(\theta)$ with strictly positive derivatives with respect to x, and

$$\sum_i E[(\max_\theta (F'_{i\theta}(\theta, x))^2 / F'_{ix}(\theta, x)) I_{B_i^+(\theta)}] < \infty;$$

(C3). $b_i(\theta)$ is continuously differentiable in θ, and $\sum_i \max_\theta (b'_i(\theta))^2 < \infty$;

(C4). for each i, the functions $F(\theta, c_i(\theta))$ and $F(\theta, b_i^-(\theta))$ are continuously differentiable in θ, and $\sum_i \max_\theta |F'(\theta, c_i(\theta))| < \infty$, $\sum_i \max_\theta |F'(\theta, b_i^-(\theta))| < \infty$,

Define $M_1(\theta) = 2\sum_i (L(c_i(\theta)) - L(b_i(\theta)))^2 |F'(\theta, c_i(\theta))|$. Then $M_1(\theta) \geq 0$ is bounded for all θ. If $M_1(\theta) > 0$, we have

$$E[(L(X(\theta+\delta, \xi)) - L(X(\theta-\delta, \xi)))^2] = M_1(\theta)\delta + o(\delta) \tag{36}$$

as $\delta > 0$ goes to zero.

PROOF. We calculate

$$\begin{aligned} &\lim_{\delta \to 0} \frac{1}{\delta} E[(L(X(\theta+\delta, \xi)) - L(X(\theta-\delta, \xi)))^2] \\ &\quad = \lim_{\delta \to 0} \frac{1}{\delta} \sum_i E[(L(X(\theta+\delta, \xi)) - L(X(\theta-\delta, \xi)))^2 I_{\Xi_i(\theta-\delta)}]. \end{aligned}$$

Let

$$R_i(\theta, \delta) = \frac{1}{\delta} E[(L(X(\theta+\delta, \xi)) - L(X(\theta-\delta, \xi)))^2 I_{\Xi_i(\theta-\delta)}].$$

Then

$$\lim_{\delta \to 0} \frac{1}{\delta} E[(L(X(\theta+\delta, \xi)) - L(X(\theta-\delta, \xi)))^2] = \lim_{\delta \to 0} \sum_i R_i(\theta, \delta). \tag{37}$$

Next, we prove that the limit and the summation commute. Define $D_{i,1} = \Xi_i(\theta-\delta) \bigcap [0, F(\theta+\delta, b_i^-(\theta+\delta)))$, $D_{i,2} = \Xi_i(\theta-\delta) \bigcap \Xi_i(\theta+\delta)$, and $D_{i,3} = \Xi_i(\theta-\delta) \bigcap [F(\theta+\delta, b_{i+1}^-(\theta+\delta)), 1)$. It is possible for each of $D_{i,j}, j = 1, 2, 3$, to be empty. Then $\Xi_i(\theta-\delta) = \bigcup D_{i,j}$ and

$$R_i(\theta, \delta) = \sum_{j=1}^3 R_{i,j}, \quad R_{i,j} = \frac{1}{\delta} E[(L(X(\theta+\delta, \xi)) - L(X(\theta-\delta, \xi)))^2 I_{D_{i,j}}]. \tag{38}$$

By Assumption (C1), there exist $N_1, N_2 > 0$ such that $|L(X)| \leq N_1$, $|L'_X(X)| \leq N_2$. Therefore,

$$\begin{aligned} R_{i,1} &\leq (2N_1)^2 |F(\theta+\delta, b_i^-(\theta+\delta)) - F(\theta-\delta, b_i^-(\theta-\delta))|/\delta \\ &\leq 2(2N_1)^2 \max_\theta |F'(\theta, b_i^-(\theta))|, \end{aligned}$$

$$\begin{aligned} R_{i,3} &\leq (2N_1)^2 |F(\theta+\delta, b_{i+1}^-(\theta+\delta)) - F(\theta-\delta, b_{i+1}^-(\theta-\delta))|/\delta \\ &\leq 2(2N_1)^2 \max_\theta |F'(\theta, b_{i+1}^-(\theta))|, \end{aligned}$$

Without loss of generality, assume that $F(\theta+\delta, b_i^-(\theta+\delta))) \geq F(\theta-\delta, b_i^-(\theta-\delta)))$ and $F(\theta+\delta, b_{i+1}^-(\theta+\delta))) \leq F(\theta-\delta, b_{i+1}^-(\theta-\delta)))$. If $F(\theta+\delta, c_i(\theta+\delta)) > F(\theta-\delta, c_i(\theta-\delta))$,

$$\begin{aligned} R_{i,2} &= \frac{1}{\delta} \int_{F(\theta+\delta, b_i^-(\theta+\delta))}^{F(\theta-\delta, c_i(\theta-\delta))} (L(X(\theta+\delta,\xi)) - L(X(\theta-\delta,\xi)))^2 d\xi \\ &\quad + \frac{1}{\delta} \int_{F(\theta-\delta, c_i(\theta-\delta))}^{F(\theta+\delta, c_i(\theta+\delta))} (L(X(\theta+\delta,\xi)) - L(X(\theta-\delta,\xi)))^2 d\xi \\ &\quad + \frac{1}{\delta} \int_{F(\theta+\delta, c_i(\theta+\delta))}^{F(\theta+\delta, b_{i+1}^-(\theta+\delta))} (L(X(\theta+\delta,\xi)) - L(X(\theta-\delta,\xi)))^2 d\xi \\ &= \frac{1}{\delta} \int_{F(\theta+\delta, b_i^-(\theta+\delta)))}^{F(\theta-\delta, c_i(\theta-\delta))} (L(b_i(\theta+\delta)) - L(b_i(\theta-\delta)))^2 d\xi \\ &\quad + \frac{1}{\delta} \int_{F(\theta-\delta, c_i(\theta-\delta))}^{F(\theta+\delta, c_i(\theta+\delta))} (L(b_i(\theta+\delta)) - L(F_i^{-1}(\theta-\delta,\xi)))^2 d\xi \\ &\quad + \frac{1}{\delta} \int_{F(\theta+\delta, c_i(\theta+\delta))}^{F(\theta+\delta, b_{i+1}^-(\theta+\delta))} (L(F_i^{-1}(\theta+\delta,\xi)) - L(F_i^{-1}(\theta-\delta,\xi)))^2 d\xi \end{aligned} \tag{39}$$

The first two terms of (39) are bounded respectively by

$$4N_2^2 \max_\theta (b_i'(\theta))^2 \delta \quad \text{and} \quad 2(2N_1)^2 \max_\theta |F'(\theta, c_i(\theta))|.$$

The third term of (39) can be rewritten as

$$\begin{aligned} &\frac{1}{\delta} \int_{F(\theta+\delta, c_i(\theta+\delta))}^{F(\theta, c_i(\theta))} (L(F_i^{-1}(\theta+\delta,\xi)) - L(F_i^{-1}(\theta-\delta,\xi)))^2 d\xi \\ &+ \frac{1}{\delta} \int_{F(\theta, b_{i+1}^-(\theta))}^{F(\theta+\delta, b_{i+1}^-(\theta+\delta))} (L(F_i^{-1}(\theta+\delta,\xi)) - L(F_i^{-1}(\theta-\delta,\xi)))^2 d\xi \\ &+ \frac{1}{\delta} \int_{F(\theta, c_i(\theta))}^{F(\theta, b_{i+1}^-(\theta))} (L(F_i^{-1}(\theta+\delta,\xi)) - L(F_i^{-1}(\theta-\delta,\xi)))^2 d\xi \\ &\leq 2(2N_1)^2 \max_\theta |F'(\theta, c_i(\theta))| + 2(2N_1)^2 \max_\theta |F'(\theta, b_{i+1}^-(\theta))| \end{aligned}$$

$$+4N_2^2 E[(\max_\theta (F'_{i\theta}(\theta,x))^2/F'_{ix}(\theta,x))I_{B_i^+(\theta)}]\delta.$$

Therefore, $R_{i,2}$ is bounded by

$$4N_2^2 \max_\theta (b_i'(\theta))^2\delta + 4(2N_1)^2 \max_\theta |F'(\theta, c_i(\theta))|$$

$$+2(2N_1)^2 \max_\theta |F'(\theta, b_{i+1}^-(\theta))| + 4N_2^2 E[(\max_\theta (F'_{i\theta}(\theta,x))^2/F'_{ix}(\theta,x))I_{B_i^+(\theta)}]\delta$$

Similarly, we can prove that if $F(\theta+\delta, c_i(\theta+\delta)) \le F(\theta-\delta, c_i(\theta-\delta))$,

$$\begin{aligned} R_{i,2} &= \frac{1}{\delta}\int_{F(\theta+\delta,b_i^-(\theta+\delta)))}^{F(\theta+\delta,c_i(\theta+\delta)} (L(b_i(\theta+\delta)) - L(b_i(\theta-\delta)))^2 d\xi \\ &\quad + \frac{1}{\delta}\int_{F(\theta+\delta,c_i(\theta+\delta))}^{F(\theta-\delta,c_i(\theta-\delta))} (L(F_i^{-1}(\theta+\delta,\xi)) - L(b_i(\theta-\delta)))^2 d\xi \\ &\quad + \frac{1}{\delta}\int_{F(\theta-\delta,c_i(\theta-\delta))}^{F(\theta+\delta,b_{i+1}^-(\theta+\delta)))} (L(F_i^{-1}(\theta+\delta,\xi)) - L(F_i^{-1}(\theta-\delta,\xi)))^2 d\xi \\ &\le 4N_2^2 \max_\theta (b_i'(\theta))^2\delta + 4(2N_1)^2 \max_\theta |F'(\theta, c_i(\theta))| \\ &\quad + 2(2N_1)^2 \max_\theta |F'(\theta, b_{i+1}^-(\theta))| \\ &\quad + 4N_2^2 E[(\max_\theta (F'_{i\theta}(\theta,x))^2/F'_{ix}(\theta,x))I_{B_i^+(\theta)}]\delta. \end{aligned} \tag{40}$$

Substituting the upper bounds for $R_{i,j}, j = 1,2,3$, into (37), also noting the assumptions (C2)-(C4), we see that $R_i(\theta,\delta)$ is uniformly bounded with respect to δ. Therefore, $\sum_i R_i(\theta,\delta)$ converges uniformly in $(0,\delta_0)$ for any $\delta_0 > 0$. By the Weierstrass M-test [Krantz (1991),211], we know that the limit and the summation commute. From (37),

$$\lim_{\delta\to 0}\frac{1}{\delta}E[(L(X(\theta+\delta,\xi)) - L(X(\theta-\delta,\xi)))^2] = \lim_{\delta\to 0}\sum_i R_i(\theta,\delta) = \sum_i \lim_{\delta\to 0} R_i(\theta,\delta). \tag{41}$$

We next calculate $\lim_{\delta\to 0} R_i(\theta,\delta)$. For each i, there exists a $\delta_i > 0$ such that for any $\delta \le \delta_i$

$$\begin{aligned} &|F(\theta+\delta, b_j^-(\theta+\delta))) - F(\theta-\delta, b_j^-(\theta-\delta)))| \\ &\quad \le \frac{1}{4}\min_{j=i-1,i,i+1,i+2}\{F(\theta-\delta, b_{j+1}^-(\theta-\delta)) - F(\theta-\delta, b_j^-(\theta-\delta))\}. \end{aligned}$$

Note that $D_{i,1} = \Xi_i(\theta-\delta)\bigcap \Xi_{i-1}(\theta+\delta)$ and $D_{i,3} = \Xi_i(\theta-\delta)\bigcap \Xi_{i+1}(\theta+\delta)$ when $\delta \le \delta_i$. Therefore, by taking into account that each of $D_{i,1}$ and $D_{i,3}$ may be empty, we have

$$\begin{aligned} R_{i,1} &\le \int_{F(\theta-\delta,b_i^-(\theta-\delta))}^{F(\theta+\delta,b_i^-(\theta+\delta))} (L(b_i(\theta+\delta)) - L(F_{i-1}^{-1}(\theta-\delta,\xi)))^2 d\xi \\ &\le \max_\theta |F'(\theta, b_i^-(\theta))| (L(b_i(\theta+\delta)) - L(F_{i-1}^{-1}(\theta-\delta,\tilde{\xi})))^2 \\ &= o(1) \end{aligned}$$

where $\tilde{\xi} \in [F(\theta-\delta, b_i^-(\theta-\delta)), F(\theta+\delta, b_i^-(\theta+\delta)))$. Similarly, $R_{i,3} = o(1)$. Hence, $\lim_{\delta\to 0} D_{i,j} = 0$ for $j = 1, 3$. Also, the analysis of (39) and (40) shows that

$$\lim_{\delta\to 0} D_{i,2} = 2(L(c_i(\theta)) - L(b_i(\theta)))^2 |F'(\theta, c_i(\theta))| \tag{42}$$

Substituting (42) into (41) we get

$$\lim_{\delta\to 0} \frac{1}{\delta} E[(L(X(\theta+\delta,\xi)) - L(X(\theta-\delta,\xi)))^2] = M_1(\theta)$$

which is exactly what we want to prove. ∎

The proof of Lemma 3 shows that Assumptions (C2) and (C3) guarantee that the inverse function $F_i^{-1}(\theta, \xi)$ is sufficiently smooth. Assumption (C4) ensures the existence of $M_1(\theta)$. Assumptions (C2)-(C4) are mild. Assumption (C1) guarantees the smoothness of the function $L(X)$. The boundedness of $L(X)$ and $L'_X(X)$ can be removed if there are only a finite number of sets of $B_i(\theta)$. The finiteness of $B_i(\theta)$ can also relax the assumptions (C2)-(C4).

The case of $M_1(\theta) = 0$ can only occur when either $b_i(\theta) = c_i(\theta)$ or $F'(\theta, c_i(\theta)) = 0$. The situation of $b_i(\theta) = c_i(\theta)$ (assuming that $L(X)$ is not a constant) happens when $F(\theta, x)$ is strictly increasing. A repetition of the proof of Lemma 3 yields that

COROLLARY 5. *If Assumption (C1) is satisfied and*

$$E[(F'_\theta(\theta, x))^2 / F'_x(\theta, x)] < \infty, \tag{43}$$

then $E[(L(X(\theta+\delta,\xi)) - L(X(\theta-\delta,\xi)))^2] = O(\delta^2)$.

Corollary 5 recovers a result obtained by Glasserman and Yao (1992) under the assumption of Lipschitz continuity of $L(F^{-1}(\theta, \xi))$. When $F'(\theta, c_i(\theta)) = 0$ for all i, using the same arguments as that of Corollary 5 we can establish that

COROLLARY 6. *In addition to Assumptions (C1)-(C4), assume that* $F(\theta, c_i(\theta))$ *is continuously twice differentiable for all* i *with*

$$0 < \sum_i (L(c_i(\theta)) - L(b_i(\theta)))^2 |F''(\theta, c_i(\theta))| < \infty. \tag{44}$$

Then, $E[(L(X(\theta+\delta,\xi)) - L(X(\theta-\delta,\xi)))^2] = O(\delta^2)$.

The following Theorem 4 is the main conclusion of this subsection.

THEOREM 4. *Assume that Assumptions (A1)-(A4) and (C1)-(C4) are satisfied. If* $M_1(\theta) > 0$ *for all* θ*, then the best convergence rate for the KW algorithm (2) with* h_n *defined by (33) is* $n^{-2/5}$*. This rate is attained by choosing* $a_n = an^{-1}$*,* $a > 2/(5K_1)$*, and* $\delta_n = n^{-1/5}$*.*

PROOF. Under Assumption of (C1)-(C4) and $M_1(\theta) > 0$, we know from Lemma 3 that $E[(L(X(\theta+\delta,\xi)) - L(X(\theta-\delta,\xi)))^2] = M_1(\theta)\delta + o(\delta)$. According to (34), the variance of h_n is of order $Var[h_n|\theta_n] = M_1(\theta_n)/\delta_n + o(1/\delta_n)$. Lemma 3 shows that $M_1(\theta)$ is bounded. Therefore, $\gamma = -1$ in (8). Since (A1)-(A4) are satisfied, Corollary 2 shows that the best convergence rate is $n^{-2/(4-\gamma)} = n^{-2/5}$. ∎

The following Theorem 5 summerizes the rate of convergence of the KW algorithm (2) when (3) is replaced with one-sided finite difference approximation with CRN. The proofs are omitted since they are very similar to that of Theorem 4.

THEOREM 5. *(I) Under the same conditions as those of Theorem 4 but the estimate* h *is replaced by the following one-sided finite difference with the use of CRN*

$$h = \frac{L(X(\theta+\delta,\xi)) - L(X(\theta,\xi))}{\delta}, \tag{45}$$

the best convergence rate is $n^{-1/3}$ *which is achieved by setting* $a_n = an^{-1}, a > 1/3K_1$*, and* $\delta_n = dn^{-1/3}$*.*

(II) Assume all the assumptions of Theorem 4 except that $M_1(\theta) = 0$*. Then Corollaries 5 and 6 show that* $E[(L(X(\theta+\delta,\xi)) - L(X(\theta-\delta,\xi)))^2] = O(\delta^2)$ *if either of (43) or (44) holds. Hence,* $Var[h_n|\theta_n] = O(1)$ *for* h_n *defined by either (33) or (45). The best convergence rate for the KW algorithm (2) is* $n^{-1/2}$*. This rate can be attained by setting* $a_n = an^{-1}, a > 1/2K_1$*, and* $\delta_n = dn^{-\eta}, \eta \geq 1/2$*.*

We would like to emphasize that the assumptions in Corollaries 5 and 6 are satisfied for a broad class of stochastic optimization problems [see Glasserman and Yao (1992) for a discussion]. Theorems 4 and 5 state that, when the inversion method is used in the generation of random variables and when h is defined by (33), the convergence rate for the KW algorithm with CRN is $n^{-2/5}$ in general and is $n^{-1/2}$ for a large class of problems that satisfy the assumptions in Corollaries 5 and 6. The improvement is signficant since the best possible rate for the same KW algorithm without CRN is $n^{-1/3}$.

4.2. Rejection method. Let $f(\theta, x)$ be the density function of $X(\theta,\xi)$. Assume that, for all $\theta \in \Theta$, $f(\theta, x)$ is zero outside a finite interval $[a, b]$ and is bounded by $0 \leq f(\theta, x) \leq c$,

$c > 0$ is a constant. The rejection method generates $X(\theta, \xi)$ according to the following three steps:

1. Generate ξ_1 uniformly distributed on $[a, b]$.
2. Generate ξ_2 uniformly distributed on $[0, c]$.
3. If $\xi_2 \leq f(\theta, \xi_1)$, then set $X(\theta, \xi) = \xi_1$; otherwise go to 1.

The rejection method uses at least two random numbers ξ_1 and ξ_2 to generate $X(\theta, \xi)$. The total number of random numbers ξ_1, ξ_2 required before outputing $X(\theta, \xi)$ is a random value. The rejection method does not accurately meet the CRN requirements since it is impossible to define $X(\theta+\delta, \xi)$ and $X(\theta-\delta, \xi)$ using a fixed set of uniform random numbers [Bratley et al. (1983); Franta (1975)]. Therefore, we modify the definition of CRN in the sense defined by the following procedure for the generation of a paired random variables:

Generation of $X(\theta + \delta, \xi)$ *and* $X(\theta - \delta, \xi)$:

1. Generate ξ_1 uniformly distributed on $[a, b]$.
2. Generate ξ_2 uniformly distributed on $[0, c]$.
3. If $\xi_2 \leq f(\theta - \delta, \xi_1)$ and $\xi_2 \leq f(\theta + \delta, \xi_1)$, then set $X(\theta - \delta, \xi) = X(\theta + \delta, \xi) = \xi_1$.
4. If $\xi_2 \leq f(\theta - \delta, \xi_1)$ and $\xi_2 > f(\theta + \delta, \xi_1)$, then set $X(\theta - \delta, \xi) = \xi_1$ and generate a $X(\theta + \delta, \xi) = \xi_3$ by the rejection method.
5. If $\xi_2 > f(\theta - \delta, \xi_1)$ and $\xi_2 \leq f(\theta + \delta, \xi_1)$, then set $X(\theta + \delta, \xi) = \xi_1$ and generate a $X(\theta - \delta, \xi) = \xi_4$ by the rejection method.
6. If $\xi_2 > f(\theta - \delta, \xi_1)$ and $\xi_2 > f(\theta + \delta, \xi_1)$, go to 1.

This is essentially a coupling procedure [see Devroye (1990) for a discussion on coupling]. Such a modification is necessary to mimic the scheme of CRN using the rejection method. We will soon see that even such a loosely defined scheme can accelerate the convergence of the KW algorithm. Let $X(\theta - \delta, \xi), X(\theta + \delta, \xi)$ be generated by the preceding procedure. It is obvious that $E[h]$ for h in (33) remains the same as that in the inversion method.

THEOREM 6. *Suppose that* $f(\theta, x)$ *is zero outside* $[a, b]$, $0 \leq f(\theta, x) \leq c$ *for all* $x \in [a, b]$, *and* $X(\theta - \delta, \xi), X(\theta + \delta, \xi)$ *are generated by the previously described procedure. Assume that*

(H1). $Var[L(X(\theta,\xi))]$ *is continuous in* $\theta \in \Theta$*;*

(H2). $f(\theta,x)$ *is differentiable in* θ *for each* $x \in [a,b]$*,* $f(\theta,x)$ *satisfies the Lipschitz condition with respect to* θ*, i.e., there is a* $K(x)$ *such that* $|f(\theta+\delta,x)-f(\theta,x)| \le K(x)\delta$*, and that* $\int_a^b K(x)dx < \infty$*.*

Define

$$M_2(\theta) = \frac{Var[L(X(\theta,\xi))]}{2c(b-a)} \int_a^b |f'_\theta(\theta,x)|dx.$$

Then $0 \le M_2(\theta) < \infty$*. If* $M_2(\theta) > 0$ *for all* θ*,* $Var[L(X(\theta,\xi))]$ *is bounded,* h *is defined by (33), and the assumptions (A1)-(A4) are satisfied, then the convergence rate for the KW algorithm with CRN is* $n^{-2/5}$*.*

Proof. We see from the procedure of generating $X(\theta-\delta,\xi)$ and $X(\theta+\delta,\xi)$ that, conditioned on either $\xi_2 \le f(\theta-\delta,\xi_1)$ or $\xi_2 \le f(\theta+\delta,\xi_1)$, $X(\theta+\delta,\xi) = X(\theta-\delta,\xi) = \xi_1$ when $\xi_2 \le f(\theta-\delta,\xi_1)$ and $\xi_2 \le f(\theta+\delta,\xi_1)$; otherwise $X(\theta+\delta,\xi) = \xi_3$ and $X(\theta-\delta,\xi) = \xi_4$. Note that ξ_3 and ξ_4 are independent. Therefore,

$$Var[h] = \frac{1}{4\delta^2}(Var[L(\xi_3)] + Var[L(\xi_4)])\frac{1}{c}E[|f(\theta+\delta,\xi_1) - f(\theta-\delta,\xi_1)|] \tag{46}$$

Under Assumption (H1), $Var[L(\xi_3)]+Var[L(\xi_4)] = 2Var[L(X(\theta,\xi))]+o(1)$. By Assumption (H2),

$$\frac{1}{\delta}E[|f(\theta+\delta,\xi_1) - f(\theta-\delta,\xi_1)|] \le \frac{2}{b-a}\int_a^b K(x)dx$$

and $K(x)$ is integrable on $[a,b]$. According to the Weierstrass M-test, (46) implies that

$$\begin{aligned} Var[h] &= \frac{1}{2\delta^2}Var[L(X(\theta,\xi))]\frac{1}{c}E[|f(\theta+\delta,\xi_1) - f(\theta-\delta,\xi_1)|] + o(\frac{1}{\delta^2}) \\ &= \frac{1}{2\delta}Var[L(X(\theta,\xi))]\frac{1}{c(b-a)}\int_a^b |f'_\theta(\theta,x)|dx + o(\frac{1}{\delta}) \\ &= \frac{M_2(\theta)}{\delta} + o(\frac{1}{\delta}). \end{aligned}$$

Thus, we know from Corollary 2 where $\gamma = -1$ that the conclusion follows. ∎

For simplicity, we only consider the simplest form of the rejection method and the case in which $f(\theta,x)$ is continuous. An analysis similar to the one used in the proof of Theorem 6 shows that $Var[h] = O(1/\delta)$ remains valid in the following three situations: (i) The estimate h is replaced by the one-sided finite difference (45); (ii) The density function $f(\theta,x)$ is piecewise differentiable; (iii) The rejection method is replaced by the following *generalized rejection method*. Assume that there exist a positive constant A and a density function $g(x)$ such that $f(\theta,x) \le Ag(x)$ for all θ and for all $x \in [a,b]$. Then

1. generate ξ_1 with the density function $g(x)$;

2. generate ξ_2 uniformly distributed on $[0, Ag(\xi_1)]$;

3. if $\xi_2 \leq f(\theta, \xi_1)$, then set $X(\theta, \xi) = \xi_1$; otherwise go to 1.

It is easy to verify that $X(\theta, \xi)$ has the desired distribution. The density function $g(x)$ should be chosen such that it is easier to generate a random variable with $g(x)$ than those with $f(\theta, x)$.

Generally speaking, the convergence rates for the KW algorithm are the same when either the inversion method or the rejection method is used in the generation of the random variable $X(\theta, \xi)$. However, the rate corresponding to the use of the rejection method is universally true for any function: It can be seen from its definition that $M_2(\theta)$ is always positive except when $Var[L(X)] = 0$ or when $f(\theta, x)$ is independent of θ. Both cases are of little practical relevance. Furthermore, assume that the assumptions in Theorem 6 are satisfied and, in addition, $f(\theta, x)$ is strictly positive on (a, b). Then the best possible convergence rate for the KW algorithm is $n^{-2/5}$ when the rejection method is used in generating $X(\theta, \xi)$. On the other hand, the distribution function $F(\theta, x) = \int_a^x f(\theta, u)du$ is continuously differentiable and strictly increasing on $[a, b]$. Theorem 5 shows that the convergence rate for the KW algorithm is $n^{-1/2}$ if the inversion method is used in generating $X(\theta, \xi)$. Therefore, as far as the convergence of the KW algorithm is concerned, the inversion method leads to faster convergence than the rejection method. This conclusion is in favor of the argument that the inversion method is superior to the rejection method [c.f. Bratley et al. (1983), 141].

4.3. Composition method. Assume that the distribution function $F(\theta, x)$ of $X(\theta, \xi)$ is of the form

$$F(\theta, x) = \sum_{i=1}^{m} p_i(\theta) F_i(\theta, x)$$

where $p_i(\theta) > 0, m \leq \infty, \sum_i p_i(\theta) = 1$, and for each i, $F_i(\theta, x)$ is a distribution function. The composition method generates the random variable $X(\theta, \xi)$ in the following way:

1. Generate a random variable Y with distribution $Prob\{Y = i\} = p_i(\theta)$.

2. If $Y = i$, generate $X(\theta, \xi)$ according to distribution $F_i(\theta, x)$.

In the composition method, there is no specification on the method for the generation of random variables at each step. Any method such as inversion and rejection can be used.

As an example, we consider the case in which random variables are generated using the inversion method which is superior to the rejection method, as we have argued in the previous subsection. Define $\rho_0(\theta) = 0, \rho_i(\theta) = \sum_{j=1}^{i} p_j(\theta)$ for $i \geq 1$. The following procedure is the actual composition method we are considering.

1. Generate a random number ξ_1 uniformly distributed on $[0, 1)$.

2. If $\xi_1 \in [\rho_{i-1}(\theta), \rho_i(\theta))$, then generate a random number ξ_2 uniform on $[0, 1)$ and set $X(\theta, \xi) = F_i^{-1}(\theta, \xi_2)$.

In this algorithm, we need two uniform random numbers in the generation of $X(\theta, \xi)$. Actually we can do with only one random number by setting $\xi_2 = (\xi_1 - \rho_{i-1}(\theta))/p_i(\theta)$. Direct verification shows that, conditional on $\xi_1 \in [\rho_{i-1}(\theta), \rho_i(\theta))$, ξ_2 is uniform on $[0, 1)$. In the composition method, we regard that $X(\theta - \delta, \xi)$ and $X(\theta + \delta, \xi)$ conform the CRN requirement if they are generated by the preceding procedure using the same $\xi = (\xi_1, \xi_2)$. We can prove that it can accelerate the convergence of the KW algorithm.

For simplicity, we assume that, for each i, $F_i(\theta, x) = F_i(x)$ is independent of θ, and the number of distribution component is finite, i.e., $m < \infty$. The case in which $F(\theta, x)$ is of general form can be treated parallel to the proof of Theorem 4. Our aim here is to find special features of the decomposition method rather than to develop the complete theory which is not difficult to derive. We first consider the situation where ξ_1 and ξ_2 are independent. We then examine the case where $\xi_2 = (\xi_1 - \rho_{i-1}(\theta))/p_i(\theta)$.

THEOREM 7. *Suppose that ξ_1 and ξ_2 are independent , $p_i(\theta)$ is differentiable in θ, and*

$$M_3(\theta) = \sum_{i=1}^{m} E[(L(F_{i+1}^{-1}(\xi)) - L(F_i^{-1}(\xi)))^2]|\rho_i'(\theta)|$$

exits and is finite. If $M_3(\theta) > 0$ is bounded from above for all θ, h is defined by (33), and the assumptions (A1)-(A4) are satisfied, then the convergence rate for the KW algorithm is $n^{-2/5}$.

PROOF. The proof is a simplified version of that of Lemma 3 since the assumptions here are stronger. According to the generation scheme of $X(\theta + \delta, \xi)$ and $X(\theta - \delta, \xi)$, we have

$$E[(L(X(\theta + \delta, \xi)) - L(X(\theta - \delta, \xi)))^2] \tag{47}$$

$$= \sum_{i=1}^{m} \int_{\rho_{i-1}(\theta-\delta)}^{\rho_i(\theta-\delta)} E_{\xi_2}[(L(X(\theta + \delta, \xi)) - L(X(\theta - \delta, \xi)))^2] d\xi_1$$

Note that m is finite and, for each i, $\rho_i(\theta)$ is continuous. There exists a δ_0 such that when $\delta \le \delta_0$

$$|\rho_i(\theta+\delta) - \rho_i(\theta-\delta)| \le \frac{1}{4}\min_j\{\rho_{j+1}(\theta-\delta) - \rho_j(\theta-\delta)\}, \quad \text{for all } i. \tag{48}$$

Let us first consider the case in which $\rho_{i-1}(\theta+\delta) > \rho_{i-1}(\theta-\delta)$ and $\rho_i(\theta+\delta) \le \rho_i(\theta-\delta)$. When $\delta \le \delta_0$, (48) ensures that (47) can be rewritten as

$$\begin{aligned}
&\sum_{i=1}^{m}\int_{\rho_{i-1}(\theta-\delta)}^{\rho_{i-1}(\theta+\delta)} E_{\xi_2}[(L(X(\theta+\delta,\xi)) - L(X(\theta-\delta,\xi)))^2]d\xi_1 \\
&+\sum_{i=1}^{m}\int_{\rho_{i-1}(\theta+\delta)}^{\rho_i(\theta+\delta)} E_{\xi_2}[(L(X(\theta+\delta,\xi)) - L(X(\theta-\delta,\xi)))^2]d\xi_1 \\
&+\sum_{i=1}^{m}\int_{\rho_i(\theta+\delta)}^{\rho_i(\theta-\delta)} E_{\xi_2}[(L(X(\theta+\delta,\xi)) - L(X(\theta-\delta,\xi)))^2]d\xi_1 \\
&=\sum_{i=1}^{m}\int_{\rho_{i-1}(\theta-\delta)}^{\rho_{i-1}(\theta+\delta)} E_{\xi_2}[(L(F_{i-1}^{-1}(\xi_2)) - L(F_i^{-1}(\xi_2)))^2]d\xi_1 \\
&\quad+\sum_{i=1}^{m}\int_{\rho_i(\theta+\delta)}^{\rho_i(\theta-\delta)} E_{\xi_2}[(L(F_{i+1}^{-1}(\xi_2)) - L(F_i^{-1}(\xi_2)))^2]d\xi_1 \\
&\quad=\sum_{i=1}^{m} E[(L(F_{i-1}^{-1}(\xi) - L(F_i^{-1}(\xi)))^2]\rho'_{i-1}(\theta)2\delta \\
&\quad+\sum_{i=1}^{m} E[(L(F_{i+1}^{-1}(\xi)) - L(F_i^{-1}(\xi)))^2]\rho'_i(\theta)2\delta + o(\delta)
\end{aligned}$$

By considering every case of $\rho_{i-1}(\theta+\delta) \le \rho_{i-1}(\theta-\delta)$ and $\rho_i(\theta+\delta) \le \rho_i(\theta-\delta)$, $\rho_{i-1}(\theta+\delta) > \rho_{i-1}(\theta-\delta)$ and $\rho_i(\theta+\delta) > \rho_i(\theta-\delta)$, and $\rho_{i-1}(\theta+\delta) \le \rho_{i-1}(\theta-\delta)$ and $\rho_i(\theta+\delta) > \rho_i(\theta-\delta)$, we obtain that

$$E[(L(X(\theta+\delta,\xi)) - L(X(\theta-\delta,\xi)))^2] = M_3(\theta)\delta + o(\delta).$$

It follows from (34) that $Var[h] = (1/4)M_3(\theta)/\delta + o(1/\delta)$. Applying Corollary 2, it is easy to see that the convergence rate for the KW algorithm is $n^{-2/5}$. We have thus completed the proof. ∎

Similarly, we can prove that the rate $n^{-2/5}$ remains valid for the case in which the random variable $X(\theta,\xi) = F_i^{-1}(\theta,\xi_2)$ is generated by setting $\xi_2 = (\xi_1 - \rho_{i-1}(\theta))/p_i(\theta)$.

THEOREM 8. *Suppose that $p_i(\theta) > 0$ is differentiable and the following function exists and is finite for all θ:*

$$M_4(\theta) = \sum_{i=1}^{m}[L(F_{i+1}^{-1}(1^-)) - L(F_i^{-1}(0^+))]^2|\rho'_i(\theta)|,$$

where $F_{i+1}^{-1}(1^-) = \lim_{\xi\uparrow 1} F_{i+1}^{-1}(\xi)$ *and* $F_i^{-1}(0^+) = \lim_{\xi\downarrow 0} F_i^{-1}(\xi)$. *If* $M_4(\theta) > 0$ *is bounded for all* θ *and (A1)-(A4) are satisfied, then the convergence rate for the KW algorithm is* $n^{-2/5}$.

The previous Theorems 7 and 8 show that the convergence rate for the KW algorithm is $n^{-2/5}$ when the composition method is used. This rate does not depend on how many random numbers are used in the generation of random variables. We would emphasize that it is unlikely for each of $M_3(\theta)$, $M_4(\theta)$ to be zero in practice.

5. The MD algorithm with CRN. In this section, we examine the rates of convergence for the MD algorithm under CRN. As shown in the previous section, the use of CRN largely affects $E[h_n^2]$ and thus the reduction of the variance $Var[h_n]$. The analysis in the previous Section 4 provides direct information on $E[h_n^2]$. Therefore, in this section we directly work $E[h_n^2]$ without going through $Var[h_n]$. We may represent $E[h_n^2]$ in the following form.

$$E[h_n^2] \le \tilde{c}\delta_n^\gamma. \tag{49}$$

Recall that $\gamma = -2$ for independent samplings of $X(\theta, \xi)$ without CRN. With CRN, $\gamma = -1$ if $M_1(\theta) > 0$ and $\gamma = 0$ if $M_1(\theta) = 0$. By following the same arguments as in the proof of Theorem 3 and applying (49) directly for $E[h_i^2]$ in (27), we obtain the following theorem.

THEOREM 9. *Assume (B1)-(B3) and (49). Then we have*

$$E[J(\hat{\theta}_n) - J(\theta^*)] \le \frac{C_1}{na_n} + \frac{\tilde{C}_2}{n}\sum_{i=1}^n a_i\delta_i^\gamma + \frac{C_5}{n}\sum_{i=1}^n \delta_i^\beta, \tag{50}$$

where $\tilde{C}_2 = \tilde{c}/(2\kappa)$, C_1 *and* C_5 *are specified in Theorem 3.*

The following Corollary 7 summarizes the rates of convergence for the MD algorithm with using CRN in calculating the finite difference (33) and (45).

COROLLARY 7. *Assume (B1)-(B3) and (49). Denote*

$$\tilde{H}_n = \frac{C_1}{na_n} + \frac{\tilde{C}_2}{n}\sum_{i=1}^n a_i\delta_i^\gamma + \frac{C_5}{n}\sum_{i=1}^n \delta_i^\beta.$$

Then

(i) if $\gamma = -1$, *the best possible rate of convergence for the upper bound* $H(n)$ *is* $n^{-1/3}$ *when the one-sided finite difference (45) is used,*

(ii) if $\gamma = -1$, *the best possible rate of convergence for the upper bound* $H(n)$ *is* $n^{-2/5}$ *when the symmetric finite difference (33) is used.*

(iii) if $\gamma = 0$, the best possible rate of convergence for the upper bound $H(n)$ is $n^{-1/2}$ when either the one-sided finite difference (45) or the symmetric finite difference (33) is used,

6. Generalization and applications. In Sections 4-5, all the results are obtained for one dimensional random variables only. In this section, we extend the results to a case of multivariates, which is not difficult but very tedious. Assume that $J(\theta) = E_\xi[L(X(\theta,\xi))]$, where the multidimensional random variable $X(\theta,\xi) = [X_1(\theta,\xi), X_2(\theta,\xi), ..., X_m(\theta,\xi)]^T \in R^m$. For each i, $X_i(\theta,\xi) = X_i(\theta,\xi_i) \in R$, ξ_i is uniform on $[0,1)$. We only consider the case in which each $X_i(\theta,\xi_i)$ is generated from ξ_i using the inversion method. To avoid repetition, we list the result without proof which is very similar to that of Theorem 5.

Assume that $J(\theta) \in R$ and $\theta \in \Theta$. For each i, $1 \le i \le m < \infty$, let $F_i(\theta,x)$ be the distribution function of $X_i(\theta,\xi_i)$ with the decomposition that

$$\frac{dF_i(\theta,x)}{dx} = \begin{cases} 0, & \text{if } x \in B^0_{i,j}(\theta) = [b_{i,j}(\theta), c_{i,j}(\theta)] \\ f_{i,j}(\theta,x), & \text{if } x \in B^+_{i,j}(\theta) = (c_{i,j}(\theta), b_{i,j+1}(\theta)), \end{cases}$$

where $\bigcup_j \{B^0_{i,j}(\theta) \bigcup B^+_{i,j}(\theta)\} = R$ for all i, $f_{i,j}(\theta,x) > 0$ for any $x \in B^+_{i,j}(\theta)$. It is possible that $F_i(\theta,x)$ is discontinuous at $b_{i,j}(\theta)$.

THEOREM 10. *Assume Assumptions (A1)-(A4) and, in addition,*

(C1)'. $L(X)$ is continuously differentiable in X, $L(X)$ and $L'_{X_i}(X)$ are bounded for all i;

(C2)'. for each i,

$$\sum_j E[(\max_\theta \left(\frac{\partial F_{i,j}(\theta,x)}{\partial \theta}\right)^2 / \frac{\partial F_{i,j}(\theta,x)}{\partial x}) I_{B^+_{i,j}(\theta)}] < \infty;$$

(C3)'. for all i,j, $b_{i,j}(\theta)$ is continuously differentiable in θ, and $\sum_j \max_\theta (b'_{i,j}(\theta))^2 < \infty$;

(C4)'. for all i,j, the functions $F_i(\theta, c_{i,j}(\theta))$ and $F_i(\theta, b^-_{i,j}(\theta))$ are continuously differentiable in θ, and $\sum_j \max_\theta |F'_i(\theta, c_{i,j}(\theta))| < \infty$, $\sum_j \max_\theta |F'_i(\theta, b^-_{i,j}(\theta))| < \infty$,

Define $\tilde{M}_1(\theta) = \sum_{i,j}(L(c_{i,j}(\theta)) - L(b_{i,j}(\theta)))^2 |F'_i(\theta, c_{i,j}(\theta))|$. Then $\tilde{M}_1(\theta) \ge 0$ is bounded. If $\tilde{M}_1(\theta) > 0$ for all θ, the best possible convergence rate for the KW algorithm (2) with h_n defined by (24) is $n^{-2/5}$. This rate is attained by choosing $a_n = an^{-1}$, $a > 2/(5K_1)$, and $\delta_n = n^{-1/5}$.

Similar results can be obtained if other methods are used in the generation of random variables or if $L(X)$ is a piecewise continuous function of X. The analysis can be applied to

general problems such as Monte Carlo optimization of queueing systems and other general systems. Although such a generalization is not trivial, the basic idea is the same except that the analysis becomes tedious and lengthy. Next we illustrate an application of Theorem 10 to the optimization of queueing systems [see, e.g. Kleinrock (1976)].

EXAMPLE 1. GI/G/1 QUEUE WITH SINGLE CLASS OF CUSTOMERS. In a GI/G/1 queue, there is one server (such as a teller in a bank) and one queue. Upon its arrival, a customer enters the server for service if the server is free, otherwise it joins the queue and waits for its turn. The service discipline is first-come-first-serve. The server cannot be free if there is at least one customer waiting in the queue. Assume that the distribution of interarrival times is $G_a(t)$ and the distribution of service times is $G_s(\theta,t) = p(\theta)G_s^1(t) + (1-p(\theta))G_s^2(t)$. For simplicity, we assume that $G_a(t)$, $G_s^1(t)$, and $G_s^2(t)$ are independent of θ and $\int t^2 dG_s^j(t) < +\infty, j = 1,2$, $p(\theta)$ is continuously differentiable in θ, $G_a(t), G_s^j(t), j = 1,2$, are strictly increasing and continuously differentiable in t. In queueing theory, the system time of a customer is defined as the time period from its arrival till departure. Let $L(X(\theta,\xi))$ be the average system time of the first N customers

$$L(X(\theta,\xi)) = \frac{1}{N}\sum_{i=1}^{N} T_i(\theta,\xi),$$

where $T_i(\theta,\xi)$ is the system time of the ith customer. Then $J(\theta) = E[L(X(\theta,\xi))]$ is the mean system time of the first N customers. We want to find the optimal parameter θ^* to minimize $J(\theta)$. It is known that the analytical form of $J(\theta)$ is not available for general $G_a(t)$, $G_s^1(t)$, and $G_s^2(t)$ [e.g. Kleinrock (1976)]. So we find θ^* via the KW algorithm. Assume that the queue is initially empty. According to Lindley's equation [e.g. Kleinrock (1976)]:

$$T_i(\theta,\xi) = \max\{T_{i-1}(\theta,\xi) - A_i, 0\} + S_i, \;\; T_0(\theta,\xi) = 0, \tag{51}$$

where A_i is the interarrival time between the $(i-1)$th and the ith customer, S_i is the service time of the ith customer. The distributions of A_i and S_i are respectively $G_a(t)$ and $G_s(\theta,t)$. We consider two scenarios.

Case 1. We find θ^* through computer simulation. We write a program to simulate the GI/G/1 queue. At the nth iteration, we perform two experiments with the same ξ_n to obtain a h_n that is defined by (33). Consider that the inversion method is used in the generation of random variables $A_i = G_a^{-1}(u_i), S_i = G_s^{-1}(\theta, v_i), i = 1,2,...,N$. Define the random factor as $\xi = [u_1, u_2, ..., u_N, v_1, v_2, ..., v_N]^T$, $A(\xi) = [A_1, A_2, ..., A_N]$, $S(\theta,\xi) =$

$[S_1, S_2, ..., S_N]$, and $X(\theta, \xi) = [A(\xi), S(\theta, \xi)]^T$. Since the function $\max\{x, 0\}$ is continuous in x, $L(X)$ is continuous in X. According to Theorem 10, we know that $\tilde{M}_1(\theta) = 0$ since both $G_a(t)$ and $G_s(\theta, t)$ are strictly increasing and continuously differentiable in t. Note that $L(X)$ is not differentiable in X. However, $L(X)$ is left and right differentiable with bounded one-sided derivatives. A simple modification of the proof of Corollary 3 shows that $Var[h_n] = O(1)$. Therefore, the convergence rate for the KW algorithm is $n^{-1/2}$. If the composition method is used in the generation of $S(\theta, \xi)$ according to the distribution $G_s(\theta, t)$, then from Theorems 7 and 8 (which is applicale to the case of multivariates) we know that the rate of convergence is $n^{-2/5}$.

Case 2. Assume that this is a real system and we want to perform on-line parameter adjustment. Let visualize n as the nth day of service. Suppose that the server serves more than N customers each day. At the nth day, the server serves customers with parameter value θ_n and simultaneously collects information of $X(\theta_n, \xi_n)$ which simply is a record of interarrival times $\{A_i^n\}$ and service times $\{S_i^n\}$. At the end of the nth day, the server calculates

$$v_i = G_s(\theta_n, S_i^n), i = 1, 2, 3, ..., N.$$

It is easy to verify that each v_i is uniform on $[0, 1)$. Then the server defines ξ_n from the preceding $v_i, i = 1, 2, ..., N$, takes a $\delta_n > 0$, and

$$S(\theta_n + \delta_n, \xi_n) = [S_1^{n,1}, S_2^{n,1}, ..., S_N^{n,1}], S_i^{n,1} = G_s^{-1}(\theta_n + \delta_n, G_s(\theta_n, S_i^n)), \quad i = 1, 2, ..., N;$$

$$S(\theta_n - \delta_n, \xi_n) = [S_1^{n,2}, S_2^{n,2}, ..., S_N^{n,2}], S_i^{n,2} = G_s^{-1}(\theta_n - \delta_n, G_s(\theta_n, S_i^n)), \quad i = 1, 2, ..., N.$$

If $G_s(\theta, t) = 1 - e^{-t/\theta}$ is exponential, then $S_i^{n,1} = (\theta_n + \delta_n)S_i^n/\theta_n$, $S_i^{n,2} = (\theta_n - \delta_n)S_i^n/\theta_n$. With the values of $A(\xi_n), S(\theta_n + \delta_n, \xi_n), S(\theta_n + \delta_n, \xi_n)$, from (40) and the form of $L(X(\theta, \xi))$, the server computes $L(X(\theta_n + \delta_n, \xi_n))$ and $L(X(\theta_n - \delta_n, \xi_n))$, which determines a h_n. With this h_n, the server updates the parameter θ_{n+1} according to the KW algorithm (2) for the next $(n+1)$th day. In such a way, we have formulated an on-line optimization problem that mimics the Monte Carlo optimization. Its convergence can be analyzed similarly to that of Case 1. Our purpose here is simply to point out that the results of this paper are not restricted to Monte Carlo optimization.

7. Summary. So far, we have examined several variations of the KW algorithm and the MD algorithm under the symmetric finite difference, the one-sided finite difference, and

the use of CRN when different methods are used in the generation of random variables. The results of this paper, together with previous results on the KW algorithm without the use of CRN [c.f. Fabian (1971); Kushner and Clark (1978)], provide a complete view toward the rates of convergence for the KW algorithm. For the ease of comparison, we summarize all the results in the following table.

Table I. Rates of convergence for the KW/MD algorithm

	with CRN $h(33)$	with CRN $h(45)$	without CRN $h(3)$	without CRN $h(23)$
inversion: $M_1(\theta) \neq 0$	$n^{-2/5}$	$n^{-1/3}$	$n^{-1/3}$	$n^{-1/4}$
inversion: $M_1(\theta) = 0$	$n^{-1/2}$	$n^{-1/2}$	$n^{-1/3}$	$n^{-1/4}$
rejection: general	$n^{-2/5}$	$n^{-1/3}$	$n^{-1/3}$	$n^{-1/4}$
composition: general	$n^{-2/5}$	$n^{-1/3}$	$n^{-1/3}$	$n^{-1/4}$

In Table I, $h(3)$, $h(23)$, $h(33)$, and $h(45)$ refer to the finite-difference approximation h_n defined by (3), (23), (33), and (45), respectively. The phrase "without CRN" refers to using independent samples in calculating the finite difference $\{h_n\}$, which excludes sampling schemes that may lead to correlations between the samples. In other words, "without CRN" simply means that $\xi_{1,n}$ and $\xi_{2,n}$ are independent in (3) and (23). When the inversion method is used in the generation of random variables and when $M_1(\theta) = 0$, we assume that Corollaries 3 and 4 are applicable. Results pertaining to the KW algorithm without the use of CRN can be found in, for example, Fabian (1971), and Kushner and Clark (1978).

Generally speaking, the use of CRN is always helpful in accelerating the convergence of the KW algorithm or the MD algorithm. In some cases, such as when $M_1(\theta) = 0$ in Theorem 5, CRN helps a lot. In some of other cases, CRN may help less much. When the inversion method is used and when $M_1(\theta) = 0$, the convergence rate can reach the best possible rate for the two types of stochastic approximation algorithms. The remark at the end of Subsection 3.2 shows that, as far as the convergence rate of the KW algorithm is concerned, the inversion method is superior to the rejection method. Note that inversion can also be used to generate random variables with distributions of the form $\sum_i p_i(\theta)F_i(\theta, x)$. A comparison

of Theorem 4 and Theorems 7 and 8 shows that inversion is also superior to composition. When the distribution function $F(\theta, x)$ of $X(\theta, \xi)$ is strictly increasing and continuous, a close examination of the inversion, rejection, and composition methods shows that $X(\theta, \xi)$ is continuous in θ if it is generated from inversion. However, $X(\theta, \xi)$ is discontinuous in θ if it is generated from either rejection or composition. It is such a distinction of continuity that determines the rates of the convergence for the KW algorithm.

REFERENCES

1. P. Bratley, B. Fox, and L. Schrage, *A Guide to Simulation*, Springer-Verlag, New York, 1983

2. D.L. Burkholder, *On a class of stochastic approximation processes,* Annals of Mathematical Statistics, 27 (1956), pp. 1044-1059.

3. S. Cambanis, G. Simons, and W. Stout, *Inequalities for $Ek(X, Y)$ when marginals are fixed,* Z. Whar. Geb. 36 (1976), pp. 285-294.

4. K.L. Chung, *On a stochastic approximation method,* Annals of Mathematical Statistics, 25 (1954), pp. 463-483.

5. R. W. Conway, *Some tactical problems in digital simulation,* Management Science, 10 (1963), pp. 47-61.

6. L. Devroye, *Coupled samples in simulation,* Operations Research, 38 (1990), pp. 115-126.

7. V. Dupač, *On the Kiefer-Wolfowitz approximation method,* Casopis Pest. Math, 82 (1957), pp. 47-75.

8. J.C. Duchi, A. Agarwal, M. Johansson, and M.I. Jordan, *Ergodic Mirror Descent,* SIAM Journal on Optimization, 22 (2012), pp. 1549-1578.

9. J. Duchi, M.I. Jordan, M. Wainwright, and A. Wibisono, *Finite sample convergence rates of zero-order stochastic optimization methods,* In *Advances in Neural Information Processing Systems (NIPS)*, P. Bartlett, F. Pereira, L. Bottou and C. Burges (Eds.), 2013.

10. A. Dvoretzky, *On stochastic approximation,* Proc. Third Berkeley Symp. Math. Statis. Prob., 1 (1956), pp. 39-56.

11. V. Fabian, *Stochastic approximation,* In *Optimizing Methods in Statistics*, J.S. Rustagi (ed.), Academic Press, New York, 1971.

12. G.S. Fishman, *Correlated simulation experiments,* Simulation, 23 (1974), pp. 177-180.

13. W.R. Franta, *The Process View of Simulation.* North Holland, New York, 1975.

14. P. Glasserman and D. Yao, *Some guidelines and guarantees for common random numbers,* Management Sciences, 38(1992), pp. 884-908.

15. J.M. Hammersley and D.C. Handscomb, *Monte Carlo Methods*, Methuen, London, 1964.

16. R.G. Heikes, D.C. Montogomery, and R.L. Rardin, *Using common random numbers in simulation experiments,* Simulation, 27 (1976), pp. 81-85.

17. Y.C. Ho and X.R. Cao, *Perturbation Analysis of Discrete Event Dynamic Systems,* Kluwer Academic Publishers, Boston, 1991.

18. H. Kesten, *Accelerated stochastic approximation,* Annals of Mathematical Statistics, 29 (1958), pp. 41-59.

19. J. Kiefer and J. Wolfowitz. *Stochastic estimation of the maximum of a regression function,* Annals of Mathematical Statistics, 23 (1952), pp. 462-466.

20. J.P.C. Kleijnen, *Statistical Techniques in Simulation,* Marcel Dekker, New York, 1974.

21. L. Kleinrock, *Queueing Systems*, Vol.I, Wiley, New York, 1976.

22. S.G. Krantz, *Real Analysis and Foundations*, CRC Press, Bocan Raton, FL, 1991.

23. H.J. Kushner and D.S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag. New York, 1978.

24. A.M. Law and W.D. Kelton.*Simulation Modeling and Analysis*, McGraw-Hill, New York, 1982.

25. A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming,* SIAM Journal on Optimization, 19 (2009), pp. 1574-1609.

26. A. Nemirovski, and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley, New York, 1983.

27. H. Robbins, and S. Monro, *A stochastic approximation method,* Annals of Mathematical Statistics, 22 (1951), pp. 400-407.

28. J. Sacks, *Asymptotic distribution of stochastic approximation procedures,* Annals of Mathematical Statistics, 29 (1958), pp. 373-405.

29. A.N. Shiryayev, *Probability*, Springer-Verlag, New York, 1984.

30. M.T. Wasan, *Stochastic Approximation,* Cambridge University Press, Cambridge, England, 1969.

31. W. Whitt, *Bivariate distributions with given marginals,* Ann. Math. Stat., 4 (1976), pp. 1280-1289.