This research was supported by the National Institutes of Health, Institute of General Medical Sciences, Grant No. GM-12868-04.

ON THE KIEFER-WOLFOWITZ PROCESS AND SOME OF ITS MODIFICATIONS

Mark Allyn Johnson
University of North Carolina

Institute of Statistics Mimeo Series No. 644

September 1969

ABSTRACT

JOHNSON, MARK ALLYN. On the Kiefer-Wolfowitz Process and Some of Its Modifications. (Under the direction of JAMES E. GRIZZLE).

The Kiefer-Wolfowitz (K-W) process is a stochastic approximation procedure used to estimate the value of a controlled variable that maximizes the expected response (M(x)) of a corresponding dependent random variable.

An extension of a theorem on the asymptotic distribution of a certain class of stochastic approximation procedures, which includes the K-W process, is given. This extension admits an increasing number of observations to be taken on each step, constrains the observed random variable to lie between predetermined bounds (not necessarily finite), and extends the class of admissable norming sequences. Each of these extensions is examined in the context of their usefulness to the application of the K-W process.

Approximations of the small sample mean and variance of the K-W process are given and studied through an examination of various special cases of the response function. Traditionally the estimation of the variance of the process involves the estimation of $M''(\theta)$. It is shown that an efficient estimator of $M''(\theta)$ based on the observations generated by the process provides a consistent estimator of $M''(\theta)$.

A modification of the K-W process is developed for which the dependent random variables influence the direction, but not the magnitude of the correction on each step. In addition to obtaining the

asymptotic distribution of this process, an approximation for the small sample variance is given. An estimator of this variance is given which does not entail the estimation of unknown parameters. The estimator is shown to be consistent.

Finally, examples illustrating the use of these procedures are given.

BIOGRAPHY

The author was born September 30, 1943, in Weld County, Colorado. He was reared on an irrigated farm near Eaton, Colorado. He graduated from Eaton High School in 1956 and went directly to college.

He received the Bachelor of Science degree with a major in zoology at Colorado State University in 1960. On receiving a grant from the National Institute of Health, he studied experimental statistics in the graduate school of the same university for two years. In September, 1966, he entered the Graduate School of North Carolina State University at Raleigh to continue studying experimental statistics. He has accepted a position as a statistical consultant for the Upjohn Company in Kalamazoo, Michigan.

The author married Miss Martha McLendon Willey from Nashville, North Carolina, in 1968. As yet they have no children.

ACKNOWLEDGMENTS

The author wishes to thank his advisor, Dr. James Grizzle, who suggested the topic of this dissertation, for his freely given counsel throughout his doctoral program. He also wishes to thank Professors N. L. Johnson, Pranab K. Sen, Harry Smith, Jr., and H. R. van der Vaart for serving on his advisory committee and offering assistance during this study.

The author wishes to thank Mrs. Gay Goss for typing the manuscript and to thank Mrs. Mary Donelan for assisting in the typing of the rough draft of this manuscript.

Finally, the author wishes to thank his parents, Mr. and Mrs. Walter F. Johnson for their inspiration throughout his academic career, and to thank his dearest wife, Martha, for her patience and her assistance in the proofreading of each manuscript and in the typing of the rough draft.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1. INTRODUCTION AND REVIEW OF LITERATURE

1.1 Introduction

The problem of finding the maximum of a response curve is often encountered in applied statistics. Two methods for locating this maximum have been developed extensively. The method of steepest ascent, which relies on multiple regression principles, is the most widely accepted. The second method is a sequential procedure. Heretofore, this sequential procedure has been developed mainly along asymptotic lines. We shall develop two modifications of this latter method with special emphasis given to the practical application of this method.

With this in mind, let X be a variable that can be controlled by the experimenter, and let Y(X) be an associated random variable such that $E[Y(X)] = M(X)$. In other words, let M(x) be a function of x, and assume Y(x) is an observable random variable with a distribution function $F(\cdot|x)$ such that

$$M(x) = \int_{-\infty}^{\infty} y dF(y|x).$$

It will be assumed throughout this paper that M(x) has a unique maximum at $x = \theta$ and that M(x) is increasing or decreasing as x is less than or greater than θ.

The foundations of the method of estimation, which Kiefer and Wolfowitz [8] adapted for the particular problem at hand, are contained in a paper by Robbins and Monro [9]. The spirit of this estimation procedure is to generate a sequence of estimates $\{X_n\}$ that converge to θ. The procedure is such that the n'th estimate (X_n) equals X_{n-1} plus a correction term that is a function of an observable random variable $(Y(X_{n-1})$ depending on X_i for $i = 1, \ldots n-1$.

These procedures have certain advantages and disadvantages when they are compared to the usual methods of response surface methodology. They are simple to execute and easy to evaluate. They are generally applicable in practice, and they admit a complete mathematical formulation. On the negative side, they have the disadvantage of most sequential procedures in that they require evaluation after each step. Moreover, their variance often depends upon parameters for which a good estimator cannot be calculated from the observations.

The stochastic approximation procedure proposed by Kiefer and Wolfowitz [8] was later called the Kiefer-Wolfowitz process. In this paper it will be referred to as the K-W process. Its definition follows.

<u>Definition</u>: Let $\{a_n\}$ and $\{c_n\}$ be sequences of positive numbers, and suppose Y(x) is a random variable with distribution function $F(y|x)$ where $M(x) = \int_{-\infty}^{\infty} y dF(y|x)$. Let X_1 be arbitrary, and let the random variable sequence $\{X_n\}$ be defined by

$$X_{n+1} = X_n - a_n c_n^{-1} (Y(X_n - c_n) - Y(X_n + c_n)), \tag{1.1}$$

where $Y(X_n - c_n)$ and $Y(X_n + c_n)$ are random variables which, conditional on $Y(X_1 + c_1), \ldots, Y(X_{n-1} + c_{n-1}), X_1, \ldots, X_{n-1}$, and $X_n = x_n$, are distributed independently as $F(y|x_n - c_n)$ and $F(y|x_n + c_n)$.

Before summarizing the literature on this process, some motivation for the assumptions on the sequences $\{a_n\}$ and $\{c_n\}$ will be given. It follows from the definition that

$$E(X_{n+1} - \theta|X_n = x_n) = (x_n - \theta) - a_n c_n^{-1} (M(x_n - c_n) - M(x_n + c_n)) \quad (1.2)$$

If $x_n - c_n \geq \theta$, then $M(x_n - c_n) - M(x_n + c_n)$ is positive, and if $x_n + c_n \leq \theta$, then $M(x_n - c_n) - M(x_n + c_n)$ is negative. Obviously, under suitable regularity assumptions on $M(x)$, there exists a point $\mu(c_n)$ such that $M(x_n - c_n) - M(x_n + c_n)$ is greater than, equal to, or less than zero as x_n is greater than, equal to, or less than $\mu(c_n)$. See Figure 1 for an illustration.

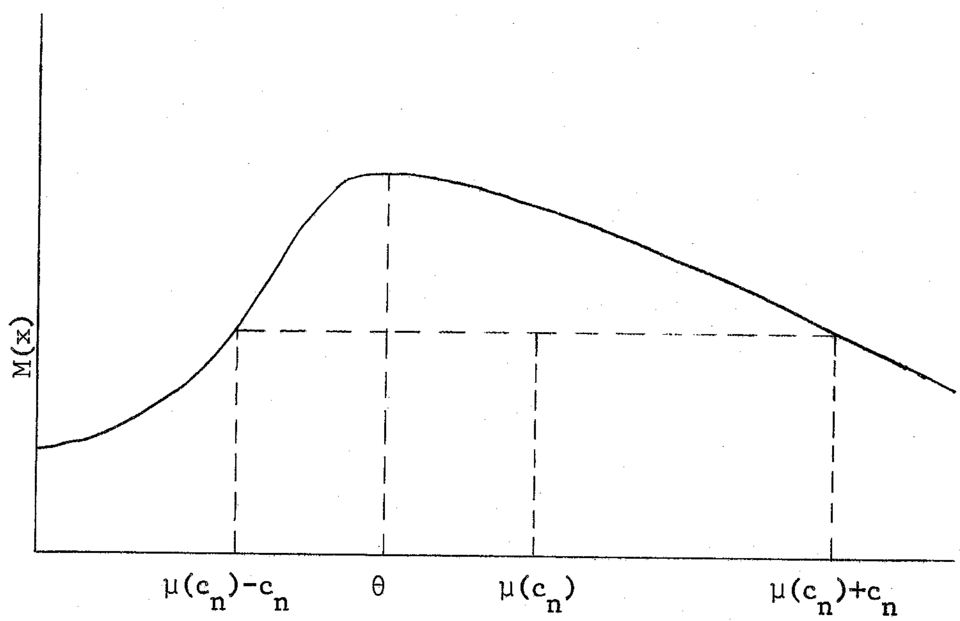

Figure 1. Graphical Definition of $\mu(c_n)$

It is equally clear that if $\mu(c_n)$ exists, then $|\mu(c_n) - \theta| \leq c_n$, and that if $M(x)$ is symmetric about θ, then $\mu(c_n) = \theta$. If $c_n \equiv c$, intuition tells us that X_n should tend to $\mu(c)$ rather than θ. Thus, if $M(x)$ is not required to be symmetric about θ, it is usually assumed that $c_n \to 0$.

Another common assumption is that $\Sigma a_n = \infty$. The necessity of this is seen by assuming that $\Sigma a_n < \infty$, and that for $(x - \theta) > 0$, there exists a $\delta > 0$ such that $|M'(x)| < \delta$. It follows from (1.2) that for $x_1 - \theta$ large enough

$$E(X_2-\theta|X_1=x_1) = (x_1-\theta) - a_1c_1^{-1}(M(x_1-c_1) - M(x_1+c_1))$$
$$>(x_1-\theta) - 2\delta a_1.$$

Moreover,

$$E(X_3-\theta|X_1=x_1) = E(E(X_3-\theta|X_2,X_1)|X_1=x_1)$$
$$> E(x_2-\theta-2a_2\delta|X_1=x_1)$$
$$> (x_1-\theta)-2\delta \sum_{m=1}^{2} a_m.$$

By induction

$$E(X_{n+1}-\theta|X_1=x_1) > (x_1-\theta) - 2\delta \sum_{m=1}^{n} a_m.$$

If $\Sigma a_m < \infty$, it is possible to choose $(x_1-\theta) > 2\delta \sum_1^\infty a_m$ implying an asymptotic bias; thus it is necessary to assume that $\Sigma a_n = \infty$.

The last common assumption is that $\Sigma a_n^2 c_n^{-2} < \infty$. This assumption is somewhat less intuitive. However, it is easy to see why $a_n^2 c_n^{-2}$ must tend to zero. In fact, it follows from (1.1) and the inequalities

$$E(X_{n+1}-\theta)^2 \geq E(X_{n+1}-E(X_{n+1}))^2$$
$$\geq E(X_{n+1}-E(X_{n+1}|X_n))^2$$
$$= E(-a_n c_n^{-1}[(Y(x_n-c_n) - M(x_n-c_n)) - (Y(x_n+c_n) - M(x_n+c_n))])^2.$$

These three assumptions have almost always been made in theorems implying $X_n \to \theta$ in any usual mode of convergence. The exception is that c_n does not have to converge to zero if $M(x)$ is symmetric about θ. The validity of these assumptions will always be assumed in the following review.

1.2 Summary of Literature

In their introduction of K-W process, Kiefer and Wolfowitz [8] proved that $X_n \to \theta$ in probability if $M(x)$ is bounded by two quadratic functions, $E(Y(x))^2 < \infty$, and $\Sigma a_n c_n^{-1} < \infty$. In later papers (J. R. Blum [1], D. L. Burkholder [2], and A. Dvoretzky [4]) the assumption that $\Sigma a_n c_n^{-1} < \infty$ was dropped, and the conclusion was strengthened to that of convergence with probability one. Burkholder's theorem relaxes the assumption of quadratic bounds on $M(x)$ with the assumption which implies for practical purposes that $|M'(x)| > 0$ if $x \neq \theta$. Burkholder also noted that this assumption admits response functions taking values on bounded value sets (e.g. $\exp(-x^2)$).

The first significant results on the asymptotic distribution of X_n were derived by Burkholder [2] who used methods suggested by K. L. Chung [3]. Burkholder established that the moments of $n^{\xi}(X_n-\theta)$ converged to those of a central normal random variable if the sequences $\{a_n\}$ and $\{c_n\}$ are of the form $a_n = 0(n^{-1})$ and $c_n = 0(n^{-\omega})$, and $\xi = 1/2-\omega$. Burkholder also assumed that $\omega > (4+2\eta)^{-1}$, where $|\mu(\varepsilon)-\theta| = 0(\varepsilon^{1+\eta})$. He proved that, if the third derivative of M(x) is continuous, then $\eta \geq 1$. Using a method developed by Hodges and Lehman [7], he was able to drop the assumption of the quadratic bounds, but he had to weaken the conclusion to that of convergence in distribution.

Sachs [10] used a different approach to show that $n^{1/3}(\ln n)^{-1}(X_n-\theta)$ was asymptotically normal under similar restrictions on the sequence $\{a_n\}$ and the assumption of quadratic bounds on M(x). He did this by considering a slightly different class of sequences $\{c_n\}$. He also proved that, in general, $n^{1/3}(X_n-\theta)$ is not asymptotically normal under the assumptions on the sequences $\{a_n\}$ and $\{c_n\}$ discussed in the introduction.

More recently, Fabian [6] proved a very general theorem about which he stated, ". . . [the theorem] implies in a simple way all known results on asymptotic normality in various cases of stochastic approximation." He used his result to obtain the asymptotic normality of his generalization of the K-W process (Fabian [5]) which can be constructed to obtain an asymptotic variance of order $(n^{-(1-\varepsilon)})$ for any $\varepsilon > 0$.

Fabian achieves this smaller variance by taking observations at more than two points on each step. His results have interesting possibilities, especially if a very precise estimate of θ is needed. However, the K-W process seems to have the better small sample properties unless one has substantial a priori knowledge as to the location of θ.

J. H. Venter [12] has proposed that observations be taken at $X_n - c_n$, X_n, and $X_n + c_n$ on each iteration and thereby estimate $M''(\theta)$. He noted that this estimate could be used to minimize the variance of $n^{\xi}(X_n-\theta)$ if a_n is of the form An^{-1}. Again, this is mainly an asymptotic result and will not be considered here.

The only small sample results relevant to the K-W process seem to be those of B. Springer's [11] simulation study on a process proposed by Wilde [13]. This process is a modification of the K-W process in which the sign of $Y(X_n-c_n) - Y(X_n+c_n)$ is used rather than the difference. In this study he used a geometric sequence for $\{a_n\}$ which obviously does not satisfy the usual assumption that Σa_n diverge. Then he unjustifiably criticized the process as a whole for the bias that he observed.

1.3 Summary of Results

The K-W process has been largely of theoretical interest, and the behavior of this process has been studied mainly through its asymptotic properties. With the exception of some remarks by a few writers, little knowledge of use to the consulting statistician can be gleaned from a review of the literature. It is the intent of this paper to fill this gap.

Chapter 2 will be devoted to a generalization of Burkholder's [2] theorem on the asymptotic distribution of a certain class of stochastic approximation procedures. These extensions will admit sequences $\{a_n\}$ of the form $O(n^{-\alpha})$ where $1/2 < \alpha \leq 1$, admit the use of constrained random variables (to be defined later), and allow the experimenter to take an increasing number of observations at each step. The usefulness of these extensions will be discussed in the following chapters as they arise.

In the third chapter, a generalization of the K-W process will be proposed and its asymptotic properties will be obtained from the results of Chapter 2. Approximate bounds will be given for the small sample variance of the process when $M(x)$ is a quadratic function, and this result will be used to obtain approximate bounds in cases of more general functions. The effects of the response function, the value of α, the variance of $Y(x)$, and the starting value (X_1) on the mean square error will be studied in some detail.

In the fourth chapter we shall develop the modification of the K-W process proposed by Wilde [13]. Various asymptotic and small sample properties will be established. In addition, a method of estimating the variance of the estimator will be developed.

In the final chapter two examples will be given to illustrate some of the results in the third and fourth chapters.

CHAPTER 2. STOCHASTIC APPROXIMATION PROCEDURES OF TYPE A_o^T

2.1 Constrained Stochastic Approximation Procedures

In this chapter we shall extend Burkholder's [2] Theorem 3 on the asymptotic normality of a class of stochastic approximation procedures. This extension will improve the practicality and the flexibility of these procedures. The main discussion of the usefulness of this extension will be made in Chapter 2 and in Chapter 3 in which more concrete examples will be examined.

In the development that follows, $F^{[r]}(\cdot)$ will denote the distribution function of the mean of [r] ([r] denotes the greatest integer less than or equal to r) independent random variables with distribution function $F(\cdot)$. We shall denote by Y^T the random variable defined by

$$Y^T = \begin{cases} T, & Y > T \\ Y, & -T \leq Y \leq T \\ -T, & Y < -T \end{cases}$$

where Y is any random variable. The random variable (Y^T) will be called a constrained random variable, and if R(Y) is any function or functional of Y, then $R^T(Y)$ will denote the corresponding function of functional of Y^T. (*E.g.*, if $M(x) = E(Y(x))$, then $M^T(x) = E(Y^T(x))$.) When it is possible to do so without confusion, the random variable sequence defined by

$$Y_n^{T_n} = \begin{cases} T_n, & Y_n > T_n \\ Y_n, & -T_n \leq Y_n \leq T_n \\ -T_n, & Y_n < -T_n \end{cases}$$

will be denoted by Y_n^T. We shall say that a function sequence $\{D_n(x)\}$ is continuously convergent to Λ at $x = \theta$ if for every $\varepsilon > 0$, there exists a $\delta(\varepsilon) > 0$ and an $N(\varepsilon) > 0$ such that if $|x-\theta| < \delta$ and $n > N(\varepsilon)$, then $|D_n(x)-\Lambda| < \varepsilon$. The usual symbols $O(\cdot)$ and $o(\cdot)$ will be used to indicate the order of convergence.

Definition (2.1): Let $\{a'_n\}$ and $\{T_n\}$ be sequences of positive numbers such that if $T_n \neq \infty$ for every n greater than some n_o, then $\inf T_n > 0$. Let $Z_n(x)$ denote a random variable with distribution function $G_n(\cdot|x)$. Let X_1 be fixed, and define the random variable sequence $\{X_n\}$ by

$$X_{n+1} = X_n - a'_n Z_n^T(X), \tag{2.1}$$

where $Z_n^T(x)$ is a random variable with conditional distribution $G_n^T(\cdot|x)$ given $Z_1^T, \ldots, Z_{n-1}^T, X_1, \ldots, X_{n-1}$, and $X_n = x$. The stochastic approximation sequence X_n will be said to be of Type A_o^T.

This definition is, in fact, Burkholder's [2] definition of a Type A_o process. However, the superscript T is added to emphasize the fact that, although the sequence $\{X_n\}$ depends on constrained random variables, it is usually more reasonable to make assumptions on the unconstrained random variables.

In order to get an appreciation for this class of processes and for some functions that will be defined later, we shall derive a few well-known results for the K-W process.

That the K-W process is of Type A_o^T can be seen by letting $a_n' = a_n c_n^{-1}$, $T_n \equiv \infty$, and $Z_n^T(X) = Y(X_n - c_n) - Y(X_n + c_n)$. Let

$R_n(x) = E(Z_n(x))$. Then $R_n(x) = M(x-c_n) - M(x+c_n)$, and, as was noted in the introduction, under suitable regularity assumptions there exists a sequence $\{\mu_n\}$ such that $R_n(\mu_n) = M(\mu_n - c_n) - M(\mu_n + c_n) = 0$, and $(x-\mu_n)\, R_n(x) > 0$ if $x \neq \mu_n$. By considering the Taylor series expansion of $R_n(x)$ about μ_n, we can show that $R_n(x) = -2c_n(x-\mu_n)M''(\mu_n) + o(c_n(\mu_n - \theta))$. Letting $D_n(x) = R_n(x)/c_n(x-\mu_n)$ for $x \neq \mu_n$, $= -2M''(\theta)$ for $x = \mu_n$, we find that the function sequence $\{D_n(x)\}$ is continuously convergent to $-2M''(\theta)$ at $x = \theta$ since $|\mu_n - \theta| \to 0$ as $c_n \to 0$.

2.2 Constrained Random Variables

The following lemmas establish certain relationships between constrained and unconstrained random variables that are used in Theorem (2.1).

Lemma (2.1): Let Y be a random variable with finite variance (σ^2). Then $Var(Y^T) \leq \sigma^2$.

Proof: Without loss of generality assume that $E(Y) = 0$. Define the random variable (Y') by

$$Y' = \begin{cases} T, & Y > T \\ Y, & Y \leq T. \end{cases}$$

Then

$$\mathrm{Var}(Y') = E(Y')^2 - (EY')^2 \leq E(Y')^2 \leq EY^2 = \sigma^2.$$

Therefore

$$\mathrm{Var}(Y^T) \leq \mathrm{Var}(Y') \leq \sigma^2.$$

<u>Lemma</u> (2.2): Let $\{T_n\}$ and $\{r_n\}$ be sequences satisfying Definition (2.1), and let $\{Z_n(x)\}$ denote a sequence of random variables with distribution functions $\{F^{[r_n]}(\cdot|x)\}$ such that $R(x) = \int_{-\infty}^{\infty} z dF(z|x)$. Assume that there exists a finite number (B) and an even integer ($K \geq 2$) such that

$$\sup_x \int_{-\infty}^{\infty} (Z - R(x))^k \, dF(Z|x) \leq B.$$

Then for $0 < C < 1$ there exist positive constants D_1, D_2, and D_3 such that

a. if $|R(x)| > CT_n$, then

$$R^{T_n}(x)\mathrm{sgn}(R(x)) > D_1 T_n - D_2 r_n^{-K/2} \; T_n^{-(K-1)} \quad ,$$

b. and if $|R(x)| \leq CT_n$, then

$$|R(x) - R^{T_n}(x)| \leq D_3 r_n^{-K/2} \; T_n^{-(K-1)} \quad .$$

Proof: Case (a): Without loss of generality, assume $R(x) > 0$. Choose C' such that $0 < C' < C$. We have then

$$E(Z_n^T(x)) = -T_n P(Z_n(x) < -T_n) + \int_{-T_n}^{T_n} z dF^{[r_n]}(z|x)$$

$$+ T_n P(Z_n(x) > T_n)$$

$$\geq -T_n P(Z_n(x) < (C-C')T_n)$$

$$+ (C - C')T_n P(Z_n(x) \leq (C-C')T_n)$$

$$= (C-C')T_n - (1+C-C')T_n P(Z_n(x) \leq (C-C')T_n).$$

But since $R_n(x) > CT_n$ we have

$$P(Z_n(x) < (C-C')T_n) \leq P(|Z_n(x) - R_n(x)| > C'T_n)$$

$$\leq (C'T_n)^{-K} E(Z_n(x) - R(x))^K$$

$$\leq (C'T_n)^{-K} r_n^{-K/2} E(Z(x) - R(x))^K$$

$$\leq B(c'T_n)^{-K} r_n^{-K/2}.$$

We have then

$$E(Z_n^T(x)) \geq (C-C')T_n - (1+C-C')^{-K} BT_n^{-(K-1)} r_n^{-K/2}$$

The desired result is obtained by letting $D_1 = (C-C')$ and $D_2 = (1+C-C')(C')^{-K}$.

Case b: Again we can assume $R(x) > 0$.

We have that

$$|R(x) - R^{T_n}(x)| = |\int_{T_n}^{\infty} (z-T_n)dF^{[r_n]}(z|x) + \int_{-\infty}^{-T_n} (z+T_n)dF^{[r_n]}(z|x)|$$

$$\leq (|R(x)| + T_n)\, P([Z_n(x) < -T_n] \cup [Z_n(x) > T_n])$$

$$+ \int_{T_n}^{\infty} |z-R(x)|\, dF^{[r_n]}(z|x) + \int_{-\infty}^{-T_n} |z-R(x)|\, dF^{[r_n]}(z|x)$$

$$\leq (|R(x)| + T_n)\, P(|Z(x) - R(x)| > T_n - R(x))$$

$$+ |T_n - R(x)|^{-(K-1)} \int_{T_n}^{\infty} |Z-R(x)|^K dF^{[r_n]}(z|x)$$

$$+ |T_n + R(x)|^{-(K-1)} \int_{-\infty}^{-T_n} |Z-R(x)|^K dF^{[r_n]}(z|x)$$

since $|R(x)| \leq CT_n < T_n$. Again using Chebyshev bounds, we see that

$$|R(x) - R^{T_n}| \leq (1+C)T_n(B(1-C)T_n)^{-K} r_n^{-K/2} + |(1-C)T_n|^{-(K-1)} r_n^{-K/2} B.$$

The desired result is obtained by letting $D_3 = 2B(1 + C)(1 - C)^{-K}$.

2.3 Asymptotic Distribution of a Type A_o^T Process

We shall now give an extension of Burkholder's [2] Theorem 3. It is helpful to recall the discussion on the K-W process preceding Lemma (2.1) and to bear in mind that the assumptions are based on the unconstrained random variables even though the theorem concerns a Type A_o^T process.

Before giving Theorem (2.1) we shall give a theorem by Burkholder [2] and a theorem by Fabian [6] which play integral parts in the proof of Theorem (2.1).

Theorem (Burkholder [2]): Suppose $\{X_n^*\}$ is a stochastic approximation process of Type A_o and θ is a real number such that

(1) there is a function Q from the positive real numbers into the integers such that if $\varepsilon > 0$, $|x-\theta| > \varepsilon$, and $n > Q(\varepsilon)$, then $(x-\theta)\, R_n^*(x) > 0$;

(2) $\sup_{n,x}(|R_n^*(x)|/(1+|x|)) < \infty$;

(3) $\sup_{n,x} \mathrm{Var}(Z_n^*(x)) < \infty$;

(4) if $0 < \delta_1, < \delta_2 < \infty$, then $\Sigma a_n^*(\inf_{S_1 \leq |x-\theta| \leq \delta_2} |R_n^*(x)|) = \infty$;

(5) $\Sigma(a_n^*)^2 < \infty$;

(6) if n is a positive integer, then $R_n^*(x)$ and $\mathrm{Var}(Z_n^*(x))$ are Borel measurable.

Then $P(\lim X_n = \theta) = 1$.

It will be useful to note that condition (1) can be replaced by the condition that

(1') there is a function Q from the positive real numbers into the positive integers such that if $\varepsilon > 0$, $|x-\theta| > \varepsilon$, and $n > Q(x)$, then $(x-\theta)R_n^*(x) > \varepsilon_n(x) + \delta_n(x)$ where $\varepsilon_n(x) > 0$, and $\Sigma a_n^* \sup_x |\delta_n(x)| < \infty$.

The following theorem is a one-dimensional version of Fabian's [6] theorem on asymptotic normality. In the theorem let R denote the real numbers. Let P denote the set of all measurable transformations from the measurable space (Ω,S) to R. Let $\chi[A]$ denote the indicator function of the set A, and, unless indicated otherwise, all convergence notions will mean convergence almost everywhere.

Theorem (Fabian): Suppose K is a positive integer, S_n a non-decreasing sequence of σ-fields, $S_n \subset S$; suppose $U_n, V_n, T_n, \Gamma_n, \phi_n \varepsilon P$, Σ, $\Gamma, \phi \varepsilon R$, and $\Gamma > 0$. Suppose Γ_n, ϕ_{n-1}, V_{n-1} are S_n- measurable, C, α, β, εR, and

$$\Gamma_n \to \Gamma,\ \phi_n \to \phi,\ T_n \to T, \text{ or } E|T_n - T| \to 0, \tag{2.2}$$

$$E(V_n|S_n) = 0,\ C > |E(V_n^2|S_n) - \Sigma| \to 0, \tag{2.3}$$

and, with $\sigma^2_{j,r} = E(\chi[V^2 \geq rj\ ]V^2)$, let either

$$\lim_{j\to\infty} \sigma^2_{j,r} = 0 \text{ for every } r > 0, \tag{2.4}$$

or $\alpha = 1$, $\lim_{n\to\infty} n^{-1} \sum_{j=1}^{n} \sigma^2_{j,r} = 0$ for every $r > 0$. Suppose that

$\beta_+ = \beta$ if $\alpha = 1$, $\beta_+ = 0$ if $\alpha \neq 1$,

$$0 < \alpha \leq 1,\ 0 \leq \beta,\ \beta_+ < 2\Gamma, \tag{2.5}$$

and

$$U_{n+1} = (I - n^{-\alpha}\Gamma_n)\ U_n + n^{-(\alpha+\beta)/2}\ \phi_n V_n + n^{-\alpha-\beta/2}\ T_n. \tag{2.6}$$

Then the asymptotic distribution of $n^{\beta/2}\ U_n$ is normal with mean $(\Gamma - \beta_+/2)^{-1}$ T and variance $\Sigma\ \phi^2/(2\Gamma - \beta_+)$.

Theorem (2.1): Let $\{X_n\}$ be a stochastic approximation process of Type A_o^T with $EZ_n(x) = R_n(x)$ and $G_n(\cdot|x) = G^{[n^\rho]}\ (\cdot|x)$. Suppose that $\{\mu_n\}$ is a real number sequence, $\{c_n\}$ is a positive number sequence, θ is a real number, ρ, ω, and ζ are non-negative numbers, and each of n_o, A, C, σ^2, ν, α, and Λ is a positive number such that

i. $\{R_n(x)\}$ is a function sequence such that if $n > n_o$, then $R_n(\mu_n) = 0$, and if $x \neq \mu_n$, then $(x - \mu_n) R_n(x) > 0$;

ii. $\sup_{n,x} n^{\omega}|R_n(x)|/ (1 + |x|) < \infty$;

iii. if $0 < \delta_1 < \delta_2 < \infty$, then $\Sigma a_n'(\inf_{\delta_1 \leq |x-\theta| \leq \delta_2} |R_n(x)|) = \infty$,

and if $T_n \neq \infty$ for all $n > n_o$, then

$$\lim_n (\inf_{\delta_1 \leq |x-\theta| \leq \delta_2} |R_n(x)|) \text{ exists;}$$

iv. $\{D_n(x)\}$ is a function sequence defined by $D_n(x) = R_n(x)/c_n(x-\mu_n)$ if $x \neq \mu_n$, $= \Lambda^*$ if $x = \mu_n$, and is continuously convergent to Λ at $x = \theta$;

v. $\sup \mathrm{Var}(Z(x)) < \infty$, $\mathrm{Var}(Z(x))$ is continuous and converges to σ^2 at $x = \theta$, there exists a finite open interval (I), containing θ such that for $\eta > 0$

$$\lim_{j\to\infty} E\{\chi[(Z(x) - R(x))^2 > \eta j^{\alpha}] (Z(x) - R(x))^2\} \to 0.$$

uniformly for $x\varepsilon I$, and if $T_n \neq \infty$ for every $n > n_o$, then $\sup_x E(|Z(x) - R(x)|^r) < \infty$ for r an even integer ≥ 2;

vi. $G(z|\cdot)$ is Borel measurable for all z;

vii. (a) $1/2 < \alpha \leq 1$, $|\mu_n - \theta| = 0(n^{-\nu})$, $n^{\alpha-\omega} a_n' \to A/C$,

---

*The definition of $D_n(\mu_n)$ is arbitrary except for the restriction that $D_n(\mu_n) \to \Lambda$.

$n^{\alpha}a'_n c_n \to A$, lim inf $n^{-\zeta}T_n > 0$,

$(r\rho/2 + (r-1)\zeta) > \omega + (1 - \alpha)$, and

$\beta/2 = \alpha/2 + \rho/2 - \omega < \nu$;

(b) if $\rho/2 < \omega$, then $\rho > 1 - 2(\alpha - \omega)$;

(c) if $T_n \neq \infty$ for every $n > n_o$, then $\beta/2 < (r\rho/2 + (r-1)\zeta - \omega)$;

(d) and $\beta_+ = 0$ if $\alpha < 1$, $= \beta < 2A\Lambda$ if $\alpha = 1$.

Then $X_n \to \theta$ a.e., and $n^{\beta/2}(X_n - \theta)$ is asymptotically normal $(0, \sigma^2 A^2/C^2(2A\Lambda - \beta_+)$.

Proof: First we shall use Burkholder's theorem to show $X_n \to \theta$ a.e. by considering two cases.

Case I: $\rho/2 < \omega$. Let $a^*_n = a'_n n^{-\rho/2}$ and $Z^*_n(x) = n^{\rho/2} Z^T_n(x)$.

Burkholder has shown that condition (i) implies that, for every $\varepsilon > 0$, there exists a $Q(\varepsilon)$ such that if $|x - \theta| > \varepsilon$, then $(x - \theta)R_n(x) > 0$ for every $n > Q(\varepsilon)$. Thus if $T_n \equiv \infty$,

$$(x - \theta)R^*_n(x) = (x - \theta)n^{\rho/2}R_n(x) > 0.$$

If $T_n \neq \infty$ for every $n > n_o$, then by Lemma (2.2) (with $[r_n] = [n^{\rho}]$) we have that

$$(x - \theta)R^*_n(x) \geq n^{\rho/2} (\min\{(x - \theta)R_n(x),\ D_1 T_n\}$$

$$- 0(n^{-r\rho/2 - (r-1\zeta)} \max\{D_2, D_3\})$$

$$\lim_{n\to\infty}\left(\inf_{\delta_1\le|x-\theta|\le\delta_2}|R_n(x)|\right),$$

by L. This limit exists and is finite by (ii) and (iii). If $L > 0$, Lemma (2.2) and (vii. (a)) give

$$\inf_{\delta_1\le|x-\theta|\le\delta_2}|R_n^T(x)| > D_4 + o(1)$$

for $0 < D_4 < \min\{L, D_1\}$. Since $a_n' = O(n^{-(\alpha-\omega)})$, $\omega \ge 0$, and $\alpha \le 1$, again the series diverges. If $L = 0$, then

$$\inf_{\delta_1\le|x-\theta|\le\delta_2}|R_n(x)| < DT_n$$

for some $0 < D < 1$ and sufficiently large n since $\liminf n^{-\zeta}T_n > 0$. Thus by (b) of Lemma (2.2) and the fact that $(r\rho/2 + (r-1)\zeta) > \omega + (1-\alpha) \ge 0$, the divergence of the series follows from the divergence for the case $T_n \equiv \infty$. Condition (4) therefore holds in all cases.

Condition (5) follows from

$$\Sigma(a_n^*)^2 = \Sigma(a_n' n^{-\rho})^2 = O(\Sigma n^{-2(\alpha-\omega)-\rho}) < \infty$$

by (vii. (b)).

Finally, Burkholder [2] has shown that if $G(z|\cdot)$ is Borel measurable, then $G^{[n^\rho]}(z|\cdot)$ is Borel measurable and if the $\sup_x \mathrm{Var}(Z(x)) < \infty$, then the measurability of $R(x)$ and $Z(x)$ is implied by that of $G(z|\cdot)$. Since the measurability of $G^{[n^\rho]}(z|\cdot)$ clearly implies the same for the distribution function of $n^{\rho/2}Z_n^T(x)$, the measurability of $R_n^*(x)$ and $\mathrm{Var}Z_n^*(x)$ is implied by the fact that

$\sup_{x,n} \text{Var } Z_n^*(x) < \infty$. Thus condition (6) is satisfied.

Therefore, $X_n \to \theta$ a.e.

Case II: $\rho/2 \geq \omega$. Let $a_n^* = a_n'c_n$, and $Z_n^*(x) = c_n^{-1}Z_n^T(x)$.

Conditions (1), (2), (3), (4), and (6) are established as in Case I. Condition (5) follows from

$$\Sigma(a_n^*)^2 = \Sigma(a_n'c_n)^2 = O(\Sigma n^{-2\alpha}) < \infty$$

by (vii. (a)).

Thus, in either case, $X_n \to \theta$ a.e.

We shall now establish the asymptotic normality of $n^{\beta/2}(X_n - \theta)$ using Fabian's theorem.

From the definition of $\{X_n\}$ we have

$$X_{n+1} - \theta = X_n - \theta - a_n'R_n^T(X) - a_n'(Z_n^T(X) - R_n^T(X))$$

$$= X_n - \theta - a_n'R_n(X) - a_n'((R_n^T(X) - R_n(X)) + (Z_n^T(X) - R_n^T(X))).$$

But by conditions (i) and (iv),

$$R_n(x) = c_n(x - \mu_n)\, D_n(x)$$

$$= c_n(\theta - \mu_n)D_n(x) + c_n(X_n - \theta)D_n(x),$$

and so

$$X_{n+1} - \theta = (1 - a_n'c_nD_n(X_n))(X_n - \theta) - a_n'(Z_n^T(X) - R_n^T(X))$$

$$- a_n'((R_n^T(X) - R_n(X)) + c_n(\theta - \mu_n)D_n(X)). \qquad (2.7)$$

Make the correspondence with Fabian's theorem as follows:

$$U_n = X_n - \theta,\ \Gamma_n = n^{\alpha} a_n' c_n D_n(X),$$

$$\phi_n = n^{(\alpha+\beta)/2} [n^{\rho}]^{-\frac{1}{2}} a_n',\ V_n = -[n^{\rho}]^{\frac{1}{2}} (Z_n^T(X) - R_n^T(X)),$$

and

$$T_n = T_n(X) = -n^{\alpha+\beta/2} a_n'((R_n^T(x) - R_n(X)) + c_n D_n(X)\ (\theta - \mu_n)).$$

(Note the possible confusion of $T_n(X)$ with the values at which $Z_n^T(X)$ is constrained.)

Let S_n be the σ-field generated by X_1, Z_1^T, ..., Z_{n-1}^T. Clearly this is the same σ-field generated by Z_1^T, ..., Z_{n-1}^T, X_1, ..., X_n. Moreover, S_n is increasing, and ϕ_{n-1} and V_{n-1} are measurable with respect to S_n. Since $R_n^T(X) = E(Z_n^T(X)|X_n)$, $R_n(X)$ is measurable with respect to S_n implying the same for $D_n(X)$ and therefore Γ_n.

To establish (2.2) we recall that $X_n \overset{a.e.}{\to} \theta$. Since $D_n(X)$ is continuously convergent to Λ at $x = \theta$, it follows that $D_n(X) \to \Lambda$ a.e., and therefore $R_n(X) \overset{a.e.}{\to} 0$. Thus

$$\Gamma_n \overset{a.e.}{\to} \Gamma = A\Lambda$$

by conditions (iv) and (vii. (a)). Again by (vii. (a))

$$\phi_n \overset{a.e.}{\to} n^{(\alpha+\beta)/2-\rho/2-(\alpha-\omega)}\ A/C$$

$$= A/C$$

since $\beta/2 = \alpha/2 + \rho/2 - \omega$. Finally, since $a_n' = 0(n^{-(\alpha-\omega)})$, $T_n(X) \to \theta$

a.e. That

$$n^{\beta/2} (\theta - \mu_n) D_n(X) \to 0 \text{ a.e.}$$

follows from the fact that $|\theta - \mu_n| = 0(n^{-\nu})$ and $\beta/2 < \nu$. Thus $T_n \to 0$ a.e.

We have

$$E(V_n | S_n) = E(E(V_n | X_n) | S_n) = E(0 | S_n) = 0.$$

If $T_n \equiv \infty$, then there exists a $C > 0$ such that (with $\Sigma = \sigma^2$)

$$C > |E(V_n^2 | S_n) - \Sigma| \to 0 \text{ a.e.}$$

by (v). If $T_n \neq \infty$ for every $n > n_o$, we have from (vii. (a)) that $r\rho/2 + (r-1)\zeta > \omega + 1 - \alpha \geq 0$, and thus

$$\text{Var}([n^\rho]\ Z_n^T(X) | S_n) \to \text{Var}\ ([n^\rho]\ Z_n(X)) \text{ a.e.}$$

Hence $|E(V_n^2 | S_n) - \Sigma| \to 0$ a.e. in either case. This establishes (2.3).

To show that $\lim_{j\to\infty} 0_{j,t}^2 = 0$ for every $t > 0$, let I be defined as in (v). Since $E(V_n) = 0$, $\sigma_{j,t}^2 \leq \sup_x E(V_j^2) \leq \sup_x \text{Var } Z(x) < \infty$ by Lemma (2.1) and condition (v). Letting D denote this last upper bound, we obtain

$$\begin{aligned}\sigma_{j,t}^2 &= E\chi\{[V_j^2 \geq tj^\alpha] \cup I\} V_j^2 \\ &\quad + E\chi\{[V_j^2 \geq tj^\alpha] \cup I^c\}\ V_j^2 \\ &\leq E\chi\{[V_j^2 \geq tj^\alpha] \cup I\} V_j^2 + DP(X_j \varepsilon I^c).\end{aligned}$$

The first term goes to zero by (v), and the last term goes to zero since $P(X_n \varepsilon I) \to 1$. Thus condition (2.4) is satisfied with $t = r$.

Finally, by noting that $\Gamma = A\Lambda$ and $\beta < 2A\Lambda$ if $\alpha = 1$, it follows that $n^{\beta/2}(X_n - \theta)$ is asymptotically normal $(0, \sigma^2 A^2/C^2(2A\Lambda - \beta_+))$.

Although Theorem (2.1) does not extend the conclusions of Fabian's theorem to a larger class of stochastic approximation procedures, the assumptions of Theorem (2.1) are more related to the conditions of an experiment in which the use of these procedures may be attractive. Theorem (2.1) also has the added attraction of removing the implied assumption that $X_n \to \theta$ a.e. and including it in the conclusion of the theorem.

Theorem (2.1) extends Burkholder's [2] Theorem 3 in three directions. Firstly, it increases the possible values for α from that of 1 to that of the interval (1/2, 1). Since the constant $a_n (=An^{-\alpha})$ can be thought of as the weight assigned to the n'th observation, this wider range of values for α gives the experimenter more flexibility in weighting the observations. This greater flexibility often enables the experimenter to decrease the mean squared error during the early stages of the search for θ. Secondly, it allows the use of constrained random variables. An example in Chapter 3 points out when such constraining can improve the behavior of the K-W process. And, thirdly, it allows an increasing number of observations to be taken on each step. Again an example in Chapter 3 describes situations in which, by taking more observations on each step, we can obtain the same variance as by taking the same number of observations using more steps. The practical importance of decreasing the number of times the experimenter must obtain observations is quite clear.

Due to the generality of Theorem (2.1), it is difficult to comment on the practical consequences of each assumption with the exception of assumptions (v) and (vi). It is sufficient to guarantee assumption (v) that our observations be bounded. Since this can almost always be assumed in practice, assumption (5) is not practically restrictive for any finite r. Thus if either ζ or ρ is greater than zero, the assumption that

$$r\rho/2 + (r - 1)\,\zeta > \omega + (1 - \alpha)$$

and

$$\beta/2 < (r\rho/2 + (r - 1)\,\zeta - \omega)$$

are almost always satisfied. Assumption (vi) seems never to be questioned in practice.

Theorem (2.1) expresses the order of the variance in terms of the number of steps rather than the number of observations. Let N(n) be the total number of observations necessary to obtain X_n. Then

$$N(n) = \sum_{m=1}^{n-1} m^{\rho} = 0(n^{\rho+1}).$$

If γ is such that $(N(n))^{\gamma} = n^{\beta/2}$, then

$$(\rho + 1)\gamma = \alpha/2 + \rho/2 - \omega,$$

or

$$\gamma = (\alpha/2 + \rho/2 - \omega)/(\rho + 1).$$

We shall discuss the value of γ for the K-W process in the next chapter, and shall only note here that $\omega = 0$ for the Robbins-Monro [9] process giving $\gamma = (\alpha + \rho)/2(\rho + 1)$.

CHAPTER 3. THE CONSTRAINED KIEFER-WOLFOWITZ PROCESS

3.1 Definition of the Constrained Kiefer-Wolfowitz Process

The Kiefer-Wolfowitz process is not generally accepted as a practical procedure for seeking a maximum in experimental work. The basic reasons for this are: its sequential nature, insufficient knowledge of its small sample properties, lack of a good estimate of its variance, and an unfamiliarity of experimental statisticians with its attributes.

We shall attempt to relieve these deficiencies somewhat. Following the definition of the Modified Kiefer-Wolfowitz process, we shall obtain, under certain mild restrictions, its asymptotic distribution as a corollary to Theorem (2.1). Various examples will be given to study the effect of the parameters of the process and to illustrate its small sample behavior. Finally, bounds will be obtained for expressions related to the small sample variance, and a method of estimating this variance will be discussed.

Definition (3.1): Let $\{a_n\}$, $\{c_n\}$, $\{r_n\}$, and $\{T_n\}$ denote sequences of positive numbers such that $[r_n] \geq 1$ and if $T_n \neq \infty$ for all n greater than some n_o, then $\inf T_n > 0$. Let $Y(x)$ denote a random variable with mean $M(x)$ and distribution function $F(\cdot|x)$. Let X_1 be fixed, and let the random variable sequence $\{X_n\}$ be defined by

$$X_{n+1} = X_n - a_n c_n^{-1}(Y(X_n - c_n) - Y(X_n + c_n))^T, \qquad (3.1)$$

where $Y(X_n - c_n)$ and $Y(X_n + c_n)$ are random variables which, conditional on $Y(X_1 - c_1), \ldots, Y(X_{n-1} + c_{n-1}), X_1, \ldots, X_{n-1}, X_n = x_n$, are distributed independently as $F^{[r_n]}(y|x_n - c_n)$ and $F^{[r_n]}(y|x_n + c_n)$.

This will be called the Constrained Kiefer-Wolfowitz process, and it will be referred to as the CKW process.

The CKW process is a generalization of the K-W process. The CKW process bounds the difference $(Y(X_n - c_n) - Y(X_n + c_n))$ by $\pm T_n$, and it allows the experimenter to vary the number of observations on each step. It clearly reduces to the K-W process when $r_n \equiv r \geq 1$ and $T_n \equiv \infty$.

3.2 Asymptotic Properties of the CKW Process

We shall now examine certain asymptotic properties of the CKW process through the following corollary to Theorem (2.1). In reviewing the corollary, it will be helpful to bear in mind that the assumptions are based on the unmodified random variables.

Corollary (3.1): Let $\{X_n\}$ be a CKW process with $r_n = n^{2a\omega}$. Let θ be a real number, let a and b be non-negative numbers, and let each of n_o, A, C, σ^2, α, ω, and Λ be positive numbers such that

a. $M'(x)$ is continuous, $(x - \theta)\, M'(x) > 0$ if $x \neq \theta$, and $\sup_x(|M'(x)|/(1 + |x|)) < \infty$;

b. the third derivative of $M(x)$ exists and is continuous for x in an open interval (I) containing θ, and $M''(\theta) = -\Lambda/2$;

c. Var $Y(x) = \sigma^2(x)$ is bounded and continuous, there exists an open interval (I) containing θ such that for any $t > 0$

$$\lim_{j\to\infty} E\{\chi[(Y(x) - M(x))^2 > tj^{\alpha}] \; (Y - M(x))^2\} \to 0$$

uniformly for $x \in I$, and if $T_n \neq \infty$ for every $n > n_o$, then $\sup_x E(|Y(x) - M(x)|^r) < \infty$ for r an even integer ≥ 2;

d. $F(y|\cdot)$ is Borel measurable for all y;

e. (1) $1/2 < \alpha \leq 1$, $n^{\alpha} a_n \to A$, $n^{\omega} c_n \to C$, $\liminf n^{-b\omega} T_n > 0$,

$(ra + (r - 1)b) > 1 + (1 - \alpha)/\omega$, and

$\beta/2 = \alpha/2 + (a - 1)\omega < 2^{\omega}$;

(2) if $a < 1$, then $a > 1 - (2\alpha - 1)/2\omega$

(3) if $T_n \neq \infty$ for every $n > n_o$, then

$\beta/2 < (ra + (r - 1) b - 1)\omega$;

(4) and $\beta_+ = 0$ if $\alpha < 1$, $= \beta < 2 A\Lambda$ if $\alpha = 1$.

Then $X_n \to \theta$ almost everywhere, and $n^{\beta/2}(X_n - \theta)$ is asymptotically normal $(0, 2 A^2 \sigma^2/C^2 (2 A\Lambda - \beta_+))$.

The basic ideas of Corollary (3.1) have already been noted by Burkholder [2]. However, since Corollary (3.1) is an extension of his work and since it is instructive to do so, we shall give a proof of this corollary.

<u>Proof</u>: Make the correspondence with Theorem (2.1) as follows.

$a'_n = a_n c_n^{-1}$, $Z_n(x) = (Y(X_n - c_n) - Y(X_n + c_n))$, and $R_n(x) = M(x - c_n) - M(x + c_n)$.

Condition (a) implies that M(x) is continuous, and that M(x) is strictly increasing or decreasing as x is less than or greater than θ, respectively. Thus, there exists a sequence $\{\mu_n\}$ such that

$(x - \mu_n)\ R_n(x) > 0$ for $x \neq \mu_n$.

From conditions (a) and (e.(1)) we obtain that the

$$\sup_{n,x} n^{\omega}|R_n(x)|/(1 + |x|) = \sup_{n,x} n^{\omega}(c_n|0(M'x)|/(1 + |x|)) < \infty.$$

This is condition (ii).

To verify condition (iii) we note that by (a) and (e.(1))

$$\Sigma a_n'(\inf_{\delta_1 \leq (x-\theta) \leq \delta_2} |R_n(x)|) = a_n(\inf_{\delta_1 \leq |x-\theta| \leq \delta_2} |0(M'(x))|)$$

$$= 0(\Sigma a_n) = \infty.$$

Moreover, the $\lim_n \inf_{\delta_1 \leq |x-\theta| \leq \delta_2} |R_n(x)| = \lim_n c_n\ 0(1) = 0$ by the above argument and the fact that $c_n \to 0$.

The function sequence $\{D_n(x)\}$ in condition (iv) is defined by $D_n(x) = c_n^{-1}(x - \mu_n)^{-1}\ (M(x - c_n) - M(x + c_n))$ if $x \neq \mu_n$, $= \Lambda$ at $x = \theta$ using condition (b). For $x \in I$ we have, since $R_n(\mu_n) = 0$,

$$M(x - c_n) - M(x + c_n) = (x - \mu_n)(M'(\mu_n - c_n) - M'(\mu_n + c_n))$$

$$+ 0(x_n - \mu_n)^2(M''(\S_1 - c_n) - M''(\S_2 + c_n))$$

$$= -2\ c_n(x - \mu_n)\ M''(\mu_n) + 0(c_n(x - \mu_n))$$

where $|\S_i - x| < |\mu_n - x|$, $i = 1,2$. Since the third derivative exists for $x \in I$ and $|\mu_n - \theta| \leq c_n \to 0$, it follows that for every $\varepsilon > 0$ there exists a $\delta > 0$ and an N such that for $|x - \theta| < \delta$ and $n > N$, $|D_n(x) - \Lambda| < \varepsilon$. Thus the condition (iv) holds.

Condition (v) is a direct consequence of (c) with $\text{Var}\ Z(x) = 2\sigma^2$.

Finally, Burkholder has noted that condition (d) implies (vi). He has also shown that the existence of a continuous third derivative in an open interval containing θ implies that $|\mu_n - \theta| = 0(c_n^2) = 0(n^{-2\omega})$. Letting $\nu = 2\omega$, $\rho = 2a\omega$, and $\zeta = b\omega$, condition (vii) can be verified using simple arithmetic.

Thus the corollary is proven.

Except in a few pathological cases, an experimenter can assume that the derivatives of M(x) exist, that the r'th central moment of Y(x) exists and is continuous, and that $F(y|x)$ is a continuous function of x for every y. Moreover, it is a rare case when he cannot specify a finite interval which should contain θ. By restricting X_n to this interval, all the assumptions of Corollary (3.1) hold for r arbitrarily large.

If the above assumptions are satisfied, condition (c) is equivalent to e'. $1/2 < \alpha \leq 1$, $n^\alpha a_n \to A$, $n^\omega c_n \to C$, $a + b > 0$,

$$\beta/2 = \alpha/2 + (a - 1)\,\omega < 2\omega, \text{ and}$$

$$\beta_+ = 0 \text{ if } \alpha < 1, = \beta < 2\, A\Lambda \text{ if } \alpha = 1.$$

Thus, except for assumption (e'), the assumptions of Corollary (3.1) should not be looked upon as restrictions, but rather as factors influencing the behavior of the process.

Since in experimental work it can usually be assumed that the supremum over x of any central moment of Y(x) is finite (at least over an interval known to contain θ), the assumptions that

$$(ra + (r - 1)b) > 1 + (1 - \alpha)\omega$$

and

$$\alpha/2 + (a - 1)\omega > (ra + (r - 1)b - 1)/\omega$$

are generally not restrictive. Practically speaking then, the constant a must satisfy

$$\alpha/2 + (a - 1)\omega < 2\omega \tag{3.2}$$

and

$$1 - (2\alpha - 1)/2\omega < a,$$

or, equivalently,

$$1 - (2\alpha - 1)/2\omega < a < 3 - \alpha/2\omega. \tag{3.3}$$

As in Chapter 2, let N(n) denote the number of observations required to obtain X_n, and let γ be defined by $(N(n))^{\gamma} = n^{\beta/2}$. By (2.8)

$$\gamma = (\alpha/2 + (a - 1)\omega)/(2a\omega + 1) \tag{3.4}$$

Thus, if $a = 0$, (i.e., the same number of observations are taken on each step), $\gamma = \alpha/2 - \omega$. Since (3.2) implies

$$\omega > \alpha/2(3 - a), \tag{3.5}$$

we obtain, for $a = 0$,

$$\gamma = \beta/2 < \alpha/2 - \alpha/6 = \alpha/3. \tag{3.6}$$

Suppose $a > 0$. We have

$$(2a\omega + 1)^2 \frac{d\gamma}{d\omega} = (2a\omega + 1)(a - 1) - (\alpha/2 + (a - 1)\omega)2a$$

$$= a(1 - \alpha) - 1.$$

It follows that $d\gamma/d\omega$ is greater than, equal to, or less than zero as a is greater than, equal to, or less than $(1 - \alpha)^{-1}$.

Case I: $a < (1 - \alpha)^{-1}$. In this case γ is maximized by minimizing ω. Thus by (3.4) and (3.5)

$$\begin{aligned} \gamma &< (\alpha/2 + \alpha(a - 1)/2(3 - a))/(a\,\alpha/(3 - a) + 1) \\ &= \alpha/(3 - a(1 - \alpha)) \\ &\leq \alpha/3. \end{aligned} \tag{3.7}$$

The equality sign holds if either $a = 0$ or $\alpha = 1$.

Case II: $a \geq (1 - \alpha)^{-1}$. By (3.3) $(1 - \alpha)^{-1} \leq 3 - \alpha/2\omega$. Since $1/2 < \alpha \leq 1$, this equality is not satisfied if either $\omega \leq 1/4$ or $\alpha > 2/3$.

Assuming these conditions are satisfied, substitution into (3.4) gives

$$\gamma = (\alpha/2 + \alpha\omega/(1 - \alpha))/(2\omega/(1 - \alpha) + 1)$$

$$= \alpha/2 < 1/3.$$

If $\alpha > (1 - \alpha)^{-1}$, γ is maximized by minimizing ω. Taking the limit of (3.4) as $\omega \to \infty$ we obtain

$$\gamma = (a - 1)/2a.$$

By (3.3) $a < 3$, and so $\gamma < 1/3$.

Thus in all cases $\gamma < 1/3$. We might add that a and ω can be adjusted so that γ is arbitrarily close to the bound obtained in each case.

We shall see that the bias is minimized by increasing the number of steps and usually not affected by increasing the number of observations taken on each step. The restriction that $\alpha > 1/2$ implies for Case II that $a \geq 2$ and so this case is usually not of practical significance.

It follows from (3.7) that, for Case I, the order of the variance increases as either a decreases or α increases. Moreover, the effect of a change in a becomes less pronounced as α increases, and it vanishes at $\alpha = 1$. Thus, if the number of steps is sufficient to discount a serious bias, the experimenter loses little efficiency in a statistical sense by increasing moderately the number of observations on each step and decreasing the number of steps when α is close to 1. However, by taking observations at fewer levels of the controlled variable (X), the experimenter decreases the time required for the experiment and possibly the cost.

3.3 Effect of Constraining the K-W Process

It is difficult to study the advantages and disadvantages of constraining the random variables via a strict mathematical treatment. Rather, using an intuitive approach, we shall study an example which illustrates when we may improve the K-W process by constraining the random variables.

Let $Z \sim N(\mu,\sigma^2)$. It is clear by Lemma 2.1 and the symmetry of the normal distribution that

1. $\text{sgn}(\mu^T) = \text{sgn}(\mu)$;
2. $\mu^T \simeq \mu$ if $|\mu| << T$; (3.9)

3. $|\mu^T| < |\mu|$ if $|\mu| > T$;

4. Var $Y^T(x) \cong \sigma^2$ if $|\mu| << T$ and $\sigma << T$;

5. and Var $Y^T(x) < \sigma^2$ if $\sigma > T$.

Partition the x-axis into the intervals $I_1 = (-\infty, \theta - k)$, $I_2 = [\theta - k, \theta + k]$, and $I_3 = (\theta + k, \infty)$. Assume that $M'(a) << |M'(b)|$ for $a \in I_1$ and $b \in I_3$. Assume that $T_n \equiv T$, $c_n \equiv C$, and σ^2 are such that $T \leq \sigma$ and $|M(x - C) - M(x + C)| << T$ for $x \in I_1$. (These assumptions on M(x) are often realized in various threshold phenomena.)

Let X_n and X_n^T denote the K-W process and the CKW process respectively, and assume that $X_1 = X_1^T = x$.

To compare the processes we can use the observations in (3.9) with $Z = Y(x - C) - Y(x + C)$. Using (5) and the fact that $T \leq \sigma$, we obtain

$$\mathrm{Var}(Y(x - C) - Y(x + C))^T < \mathrm{Var}(Y(x - C) - Y(X + C)).$$

It follows that, at least for a few steps, the K-W process is more variable about its expected value. In addition to this, as long as both processes remain in the same interval, we can visualize the following cases.

Case I: $x \in I_1$. Since $|M(x - C) - M(x + C)| << T$ for $x \in I_1$, (2) of (3.9) implies the processes approach θ at the same expected rate. The fact that $M'(a) << M'(b)$ for $a \in I_1$ and $b \in I_3$ implies that this expected rate is much less than the corresponding rate for either process approaching θ from I_3. That is, the probability that either process moves from I_1 to I_2 in a few steps is relatively small.

It is possibly a little larger for the K-W process due to its more erratic nature.

Case II: $x \varepsilon I_2$. While in this region, the behavior of both processes is characterized by random movement since $M(x-C) - M(x+C) \simeq 0 << T$. This movement is more erratic for the K-W process if $T < 2\sigma$. Moreover, if by chance either process leaves this region, the movement away from θ should most likely be greater for the K-W process due to its sensitivity to observations from the tails of the distribution. This difference increases with an increase in σ or a decrease in T.

Case III: $x \varepsilon I_3$. It follows from (3) of (3.9) that the expected rate of approach toward θ is faster for the K-W process. This difference increases with a decrease in T. However, the expected rate of approach toward θ is fast for either process, and so the probability that either process moves from I_3 to I_2 in a few steps is relatively large.

Since both processes tend to move out of the region I_3 more quickly than out of region I_1, we can expect both processes to be biased after a few steps given that X_1 is in a neighborhood of θ. The CKW is less biased for two reasons: Firstly, it tends to move from I_3 into I_2 more slowly, and it tends to move from I_1 into I_2 at about the same rate. Secondly, but probably more importantly, on leaving I_2 the K-W process tends to be thrown farther into regions I_1 and I_3. Due to the usually quick recovery from I_3 and the usually slow recovery from I_1, this can greatly increase the difference in the bias of the estimates.

If $X_1 \varepsilon I_1$, neither process seems to be the better, and although the K-W process seems better if $X_1 \varepsilon I_3$, the difference quickly disappears since both processes approach θ at a rapid rate. Thus the CKW process seems to be preferable for this example.

The essential feature of this example is the existence of two regions, both excluding θ, that are differentiated by the magnitude of $M'(x)$. Other examples with this feature can be constructed, but we note only that regression functions of the form e^{-x^2} have this feature. The advantage of the CKW process is that of lessening the probability of a deep penetration into the region for which $|M'(x)|$ is small.

The optimal choice of T is largely determined by σ^2 and the regions I_2 and I_3. Clearly, if $4\sigma < T$, where $\sigma^2 = \text{Var } Y(x)$, constraining is of little consequence in I_2, and if it is of consequence in I_3, it is a negative factor. Thus, we must choose T so as to decrease the variability in I_2 without sacrificing too much of the information that I_3 contains with respect to the direction of θ. As either more observations are taken at each step, or σ^2 is decreased, this becomes more difficult to do.

Before leaving this subject it should be observed that as $a_n \to 0$, the influence of each iteration goes to zero. This implies that both processes become characterized by their expected rate of approach toward θ. In any small neighborhood of θ, toward which both processes tend, this is very similar.

3.4 Expected Small Sample Behavior of the K-W Process

In this section we shall discuss the effects that the parameters α and ω, the starting value (X_1), and the shape of the response curve have on the behavior of the K-W process. Although the CKW process will not be discussed, it will be seen that most of the results in this section apply to it as well.

Let I denote an interval containing θ, and assume that $M'(x) = B$ for $x \in I^c$. Assume that the distribution function of $Y(x) - M(x)$ is independent of x. Let $Z(X_n) = Y(X_n - c_n) - Y(X_n + c_n)$, and Var $Z(X_n) = 2\,\sigma^2$.

We have by (3.1) that

$$X_{n+1} - \theta = X_n - \theta - a_n c_n^{-1}(M(X_n - c_n) - M(X_n + c_n)) - a_n c_n^{-1} Z(X_n) \qquad (3.10)$$

If $X_k >> \theta$, the process will remain in I^c for a few iterations with a high probability. Thus if $n - k$ is not too large, (3.10) becomes

$$X_{n+1} - \theta = (X_n - \theta) - 2\,B\,a_n - a_n c_n^{-1} Z(X_n)$$

By induction

$$X_{n+1} - \theta = (X_k - \theta) - 2\,B \sum_{m=k}^{n} a_m - \sum_{m=k}^{n} a_m c_m^{-1} Z(X_m),$$

where, conditional on $X_k >> \theta$, $Z(X_m)$, $m=k, \ldots, n$, are independent random variables. It follows that

$$E(X_{n+1} - \theta \mid X_k >> \theta) = (X_k - \theta) - 2\,B \sum_{m=k}^{n} a_m, \qquad (3.11)$$

and

$$\mathrm{Var}(X_{n+1} - \theta | X_k >> \theta) = 2\,\sigma^2 \sum_{m=k}^{n} a_m^2 c_m^{-2}. \tag{3.12}$$

Let $a_n = An^{-\alpha}$ and $c_n = Cn^{-\omega}$. It is a well known result that

$$\sum_{m=k}^{n} m^{-\rho} \simeq \int_k^{n+1} x^{-\rho}\, dx \tag{3.13}$$

of $\rho > 0$. Substituting the expressions for a_n and c_n into (3.11) and (3.12) and applying (3.13), we obtain

$$E(X_{n+1} - \theta | X_k >> \theta) \simeq \begin{cases} X_k - \theta - 2BA(1-\alpha)^{-1}((n+1)^{(1-\alpha)} - k^{(1-\alpha)}), \alpha<1 \\ X_k - \theta - 2BA(\ln(n+1) - \ln k),\ \alpha = 1, \end{cases}$$

and

$$\mathrm{Var}(X_{n+1} - \theta | X_k >> \theta) \simeq \sigma^2 A^2 C^{-2} (\alpha-\omega-1/2)^{-1} (K^{1-2(\alpha-\omega)} - (n+1)^{1-2(\alpha-\omega)}).$$

The above expressions indicate that if we are approaching θ along a constant slope, our expected rate of approach is proportional to the slope and increases very rapidly as α decreases. Furthermore, the variance of our rate of approach decreases as either α increases or ω decreases.

Thus if we expect to approach θ in a region in which $M'(x)$ is fairly constant, we should generally decrease both α and ω. It may be wise to let c_n and a_n be constant until the sequence $\{X_n\}$ no longer shows a trend in a particular direction. This lack of a trend indicates we may be in a neighborhood of θ.

In this example we are trying to decrease the bias for a given number of observations. It is clear from (3.11) that the bias is not decreased by increasing the number of observations on each step. Therefore, unless it is practically restrictive, it is best to take a minimum number of observations on each step until we are in a neighborhood of θ.

The previous example reveals the behavior of the K-W process for a few iterations while in a region in which $M'(x)$ is fairly constant. However, the assumptions of this example are certainly violated in a small neighborhood of θ, or in a region in which $M'(x)$ cannot be approximated by a constant.

A more reasonable approximation, at least in a neighborhood of θ, is to assume that $M(x)$ is a quadratic function of x. That is, assume

$$M(x) = -(B/2)(x-\theta)^2 \tag{3.14}$$

for $B > 0$. Let $a_n = An^{-\alpha}$ and $c_n = Cn^{-\omega}$. Let $Z(X_n) = Y(X_n-c_n) - Y(X_n+c_n) - (M(X_n-c_n) - M(X_n+c_n))$, and assume that the distribution function of $Z(X_n|X_n = x)$ is independent of a. Let $\text{Var}\, Z(X_n) = 2\,\sigma^2$.

Using (3.10) we have

$$\begin{aligned} X_{n+1} - \theta &= X_n - \theta - An^{-\alpha}\, c_n(B/2)(-(X_n-c_n-\theta)^2 - (X_n+c_n-\theta)^2) \\ &\quad - AC^{-1}\, n^{-(\alpha-\omega)}\, Z(X_n) \qquad (3.15) \\ &= (1-2ABn^{-\alpha})(X_n-\theta) - AC^{-1}n^{-(\alpha-\omega)}\, Z(X_n). \end{aligned}$$

Since, conditional on $X_n = x$, $Z(X_n)$ is independent of x, we obtain by induction that

$$X_{n+1} - \theta = \prod_{j=k}^{n} (1 - 2ABj^{-\alpha})(X_k - \theta)$$
$$- AC^{-1} \sum_{m=k}^{n} \sum_{j=m+1}^{n} (1 - 2ABj^{-\alpha})n^{-(\alpha-\omega)}Z(X_m),$$

where $\prod_{j=n+1}^{n} (1 - 2ABj^{-\alpha}) = 1$. It follows that

$$E(X_{n+1} - \theta|X_k) = \prod_{j=k}^{n} (1 - 2ABj^{-\alpha})(X_k - \theta),$$

and

$$E((X_{n+1}-\theta)^2|X_k) = D(k,n)(X_k-\theta)^2 + 2A^2C^{-2}\sigma^2 \sum_{m=k}^{n} D(m+1,n)m^{-2(\alpha-\omega)}, \quad (3.16)$$

where $D(m,n) = \prod_{j=m}^{n} (1 - 2ABj^{-\alpha})^2$. The last two expressions of (3.16) are, conditional on X_k, the square of the bias and the variance of X_n, respectively.

It is another well known result that there exists $\varepsilon_k \to 0$ such that for every k, n

$$(1-\varepsilon_k)\exp\{-4AB \sum_{j=k}^{n} j^{-\alpha}\} \leq D(k,n) \leq (1+\varepsilon_k)\exp\{-4AB \sum_{j=k}^{n} j^{-\alpha}\}.$$

Using results similar to (3.13) it has been shown that there exists $\varepsilon_k \to 0$ such that for every k, n

$$(1-\varepsilon_k)\exp\{-4AB(1-\alpha)^{-1}(n^{1-\alpha}-k^{1-\alpha})\} \leq D(k,n)$$
$$\leq (1+\varepsilon_k)\exp\{-4AB(1-\alpha)^{-1}(n^{1-\alpha}-k^{1-\alpha})\}, \ \alpha < 1, \quad (3.17)$$

and

$$(1-\varepsilon_k)k^{4AB}n^{-4AB} \leq D(k,n) \leq (1+\varepsilon_k)k^{4AB}n^{-4AB}, \ \alpha = 1. \qquad (3.18)$$

We can use the results of (3.17), (3.18) and Corollary (3.1) to obtain asymptotic expressions for the bias and variance of X_n in this example. In fact, it follows from (3.16), (3.17), and (3.18) that

$$E^2(X_{n+1}-\theta|X_k) = \begin{cases} ((X_k-\theta)^2 + o(1)\exp\{-4AB(1-\alpha)^{-1}n^{1-\alpha}\}, & \alpha < 1 \\ ((X_k-\theta)^2 + o(1))n^{-4AB}, & \alpha = 1. \end{cases} \qquad (3.19)$$

Although Corollary (3.1) does not guarantee that $\mathrm{Var}(X_n)$ exists, it does imply that $n^{\beta/2}(X_n - \theta)$, with $\beta/2 = (\alpha - 2\omega)/2$, tends to be distributed as a normal random variable with variance equal to

$$2A^2\ \sigma^2/C^2(2A\ \Lambda - \beta_+) = \begin{cases} A\ \sigma^2/C^2\ 2\ B, & \alpha < 1 \\ 2A^2\ \sigma^2/C^2(4AB - (\alpha - 2\omega)), & \alpha = 1 \end{cases} \qquad (3.20)$$

for $A > (\alpha - 2\omega)/4B$ if $\alpha = 1$. Many writers have noted that the variance expression for $\alpha = 1$ is maximized with respect to A if $A = (\alpha - 2\omega)/2B$.

It is reasonably evident from the nature of D(k,m) that the bias of the K-W procedure is diminished by decreasing α. Asymptotically this is clear from (3.19). However, it follows from (3.6) that $\beta/2 = \alpha - 2\omega < \alpha/3$. Thus the asymptotic variance increases as α decreases. Again, until X_n is in a neighborhood of θ, it is probably best to let α and ω be small. Once X_n is in a neighborhood of θ, α should be increased.

It is important to note that in this example the bias is not a function of c_n, and, as is to be expected, the variance decreases as either C increases or ω decreases. It follows that we can set $\omega = 0$. By doing so γ (see (3.4)) becomes $\alpha/2$. Hence with $\alpha = 1$ we obtain an asymptotic variance of $O(n^{-1})$. As was noted in Chapter 1, it is the symmetry of M(x) about θ that allows us to assume $\omega = 0$. It is clear that if M(x) is reasonably symmetric about θ, and if only a rough estimate of θ is needed, then it is best to let $\omega \simeq 0$.

Suppose we can assume that M(x) is symmetric about θ. This implies that we can set $c_n \equiv C$. Asymptotically, the optimal choice of C maximizes the difference $|M(x-C) - M(x+C)|$ for small deviations of x about θ. It is quite clear then how C should be chosen. We only note here that, for a quadratic response surface, C should be made arbitrarily large, and, for surfaces of the form e^{-x^2}, C is determined as in Figure 2.

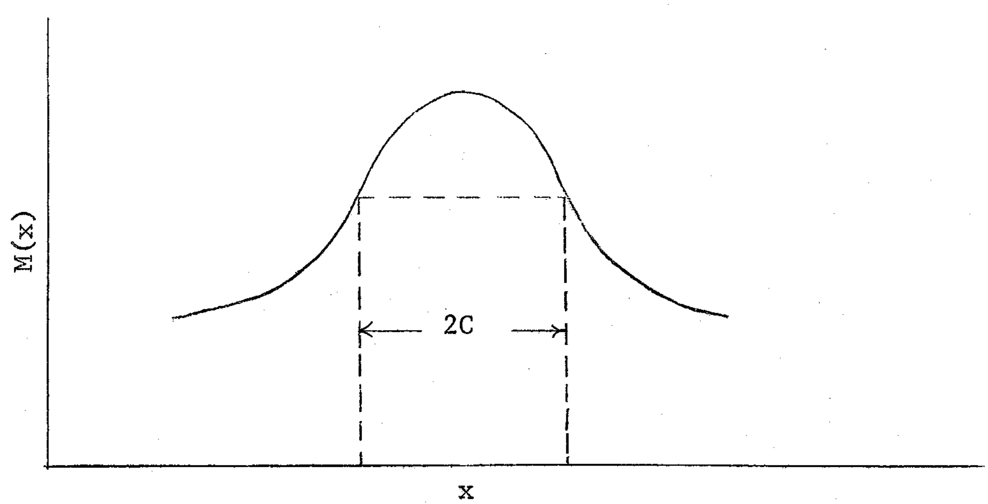

Figure 2. Optimal Choice of $c_n (\equiv C)$

In this example we have assumed that Var $Z(X_n) = 2\sigma^2$. If Var $Z(X_n) = 2[n^{2a\omega}]^{-1}\sigma^2$, (3.10) becomes

$$E(X_{n+1} - \theta | X_k) = D^{\frac{1}{2}}(k,n)(X_k - \theta)$$

and

$$E((X_{n+1} - \theta)^2 = D(k,n)(X_k - \theta)^2$$
$$+ 2AC^{-2}\sigma^2 \sum_{m=k}^{n} D(m+1,n)m^{-2(\alpha+(a-1)\omega)}$$

3.5 Approximations of the Small Sample Variance

The importance of evaluating the last expression in equation (3.21) is quite clear. This expression is, conditional on X_k, the small sample variance of X_n if M(x) is a quadratic function. If each X_m, m=1, ..., n, lies in a neighborhood of θ in which M(x) is reasonably approximated by a quadratic function, it is a close approximation to this variance for more general response functions. At the present, the only simple, but adequate, approximations to this expression rely heavily on asymptotic theory. We now obtain bounds for this expression using an appeal to asymptotic theory that is satisfied for values of k and n usually realized in practice.

Let

$$b_{(m,n)} = \prod_{j=m+1}^{n} (1-Dj^{-\alpha})^2 \, m^{-\phi}, \qquad (3.22)$$

where D, $\phi > 0$. For simplicity, let $b_m = b_{(m,n)}$ keeping in mind the dependence on n.

Using (3.22) we obtain

$$b_m - b_{m-1} = b_m(1-b_m^{-1}\ b_{m-1})$$

$$= b_m(1 - (1-Dm^{-\alpha})^2\ (1-m^{-1})^{-\phi})$$

$$= b_m(2\ Dm^{-\alpha} - \phi m^{-1} + 0(m^{-2\alpha})). \qquad (3.23)$$

Thus, for k sufficiently large, $\{b_m\}$ is an increasing sequence $m \geq k$ if either $\alpha < 1$, or $\alpha = 1$ and $2\ D > \phi$.

Since $m^{-\alpha} - (m-1)^{-\alpha} = 0(n^{-(1+\alpha)})$, the second difference is given by

$$(b_m-b_{m-1}) - (b_{m-1}-b_{m-2}) = (2\ Dm^{-\alpha} - \gamma m^{-1})(b_m-b_{m-1})$$

$$+ 0(m^{-2\alpha})(b_m-b_{m-1}). \qquad (3.24)$$

Thus, for k sufficiently large, $\{b_m-b_{m-1}\}$ is an increasing sequence for $m \geq k$ if either $\alpha < 1$, or $\alpha = 1$ and $2\ D > \gamma$.

The last expression of (3.21) is of the form $\sum_{m=k}^{n} b_m$. It is clear from equations (3.23) and (3.24) that, for sufficiently large k, upper and lower bounds can be obtained for this sum using a geometrical argument. See Figure 3 for a representation of the elements b_m that illustrates this approach. In fact,

$$\sum_{m=k}^{n} b_m \geq \begin{cases} \text{Area } \Delta\ ADE - \text{Area } \Delta\ ABC, & k \geq A \\ \text{Area } \Delta\ ADE, & k < A \end{cases} \qquad (3.25)$$

and

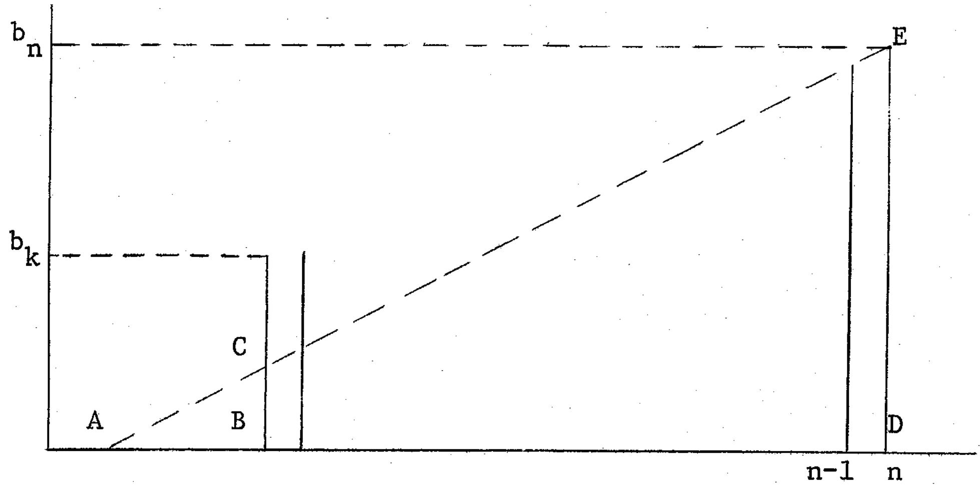

Figure 3. Graph of the Sequence $\{b_n\}$

$$\sum_{m=k}^{n} b_m \leq (n - k)\, b_n .$$

This geometrical approach will be used in only one case in the following lemmas. However, every bound will be obtained by finding sequences of functions $\{f_n(x)\}$ such that either

$$f_n(m) \geq b_m, \quad m = k, \ldots, n$$

or

$$f_n(m) \leq b_m, \quad m = k, \ldots, n.$$

The respective bound will be obtained by integrating $f_n(x)$ over a region such as $(k,n+1)$ or $(k-1,n)$.

Lemma (3.1): Let the sequence $\{b_m\}$ be defined by (3.22) with $\alpha = 1$. Let λ_1, λ_2, ϕ, and D be real numbers such that $\phi < \lambda_1 < 2\,D < \lambda_2$. Then there exists an $N(\lambda_1,\lambda_2)$ such that, for every $k,\ n > N\,(\lambda_1,\lambda_2)$,

$$(\lambda_2 - (\phi - 1))^{-1} n^{-\lambda_2} (n^{\lambda_2-(\phi-1)} - (k-1)^{\lambda_2-(\phi-1)}) \leq \sum_{m=k}^{n} b_m$$

$$\leq (\lambda_1 - (\phi - 1))^{-1} n^{-\lambda_1} ((n+1)^{\lambda_1-(\phi-1)} - k^{\lambda_1-(\phi-1)}).$$

Proof: Consider the function sequence $\{f_n(x,\lambda)\}$ defined by

$$f_n(x,\lambda) = n^{-\lambda} x^{\lambda-\phi}, \tag{3.26}$$

for $x > 0$. Clearly, by (3.22),

$$f_n(n,\lambda) = n^{-\phi} = b_n. \tag{3.27}$$

We have using (3.26) that

$$f_n(m,\lambda) \geq b_m (\leq b_m) \tag{3.28}$$

if and only if

$$n^{-\lambda} \geq m^{\phi-\lambda} b_m \; (\leq m^{\phi-\lambda} b_m). \tag{3.29}$$

Let $d_m(\lambda) = m^{\phi-\lambda} b_m$. We have

$$\begin{aligned} d_{m-1}(\lambda) d_m^{-1}(\lambda) &= ((m-1)/m)^{\phi-\lambda} b_{m-1} b_m^{-1} \\ &= (1-m^{-1})^{\phi-\lambda} (1-Dm^{-1})^2 (1-m^{-1})^{-\phi} \\ &= (1-m^{-1})^{-\lambda} (1-Dm^{-1})^2 \\ &= 1 + (\lambda - 2D) m^{-1} + 0(m^{-2}) \end{aligned}$$

by (3.23). Since $\lambda_1 < 2D < \lambda_2$, there exists an $N = N(\lambda_1,\lambda_2)$ such that the sequences $\{d_m(\lambda_1)\}_{m>N}$ and $\{d_m(\lambda_2)\}_{m>N}$ are increasing and decreasing respectively.

But (3.28), and hence (3.29), holds for $m = n$. Thus, by the monotinicity of the sequences $\{d_m(\lambda_1)\}$ and $\{d_m(\lambda_2)\}$, (3.29), and hence (3.28), holds for all $m > N(\lambda_1,\lambda_2)$.

Since $\phi < \lambda_2 < \lambda_1$, $\{f_n(x,\lambda_1)\}$ and $\{f_n(x,\lambda_2)\}$ are sequences of increasing functions. It follows that, for $k > N(\lambda_1,\lambda_2)$,

$$\int_{k-1}^{n} n^{-\lambda_2} x^{\lambda_2-\phi} dx < \sum_{m=k}^{n} b_m < \int_{k}^{n+1} n^{-\lambda_1} x^{\lambda_1-\phi} dx. \tag{3.30}$$

Lemma (3.1) is obtained by integrating the above expression.

Lemma (3.2): Let the sequence $\{b_m\}$ be defined as in Lemma (3.1). Let λ_1, λ_2, ϕ, and D be such that $0 < \phi - 1 < \lambda_1 < 2D < \lambda_2 < \phi$.

Then there exists an $N(\lambda_1,\lambda_2)$ such that, for every k, $n > N(\lambda_1,\lambda_2)$,

$$(\lambda_2 - (\phi-1))^{-1} n^{-\lambda_2}((n+1)^{\lambda_2-(\phi-1)} - k^{\lambda_2-(\phi-1)}) \leq \sum_{m=k}^{n} b_m$$
$$\leq (\lambda_1 - (\phi-1))^{-1} n^{-\lambda_1} (n^{\lambda_1-(\phi-)} - (k-1)^{\lambda_1-(\phi-1)}).$$

Proof: The proof is a repetition of the proof of Lemma (3.1) except, in this case, $\{f_n(x,\lambda)\}$ is a sequence of decreasing functions. Thus the ranges of integration in (3.30) must be reversed. The restriction that $\lambda_1 > \phi - 1$ is to prevent $f_n(x,\lambda)$ from being of the form x^{-1}.

Lemma (3.3): Let the sequence $\{b_m\}$ be defined by (3.22) with $\alpha < 1$. Let $0 < \lambda < 2D$. Then there exists an $N(\lambda)$ such that

$$b_n^2(b_n-b_{n-1})^{-1} (1 - \exp\{\lambda n^{-\alpha} (k-n)\})/2 \leq \sum_{m=k}^{n} b_m$$
$$\leq \lambda^{-1} n^{-\phi+\alpha}(\exp\{\lambda n^{-\alpha}\} - \exp\{\lambda n^{-\alpha} (k-n)\}).$$

Proof: Consider the function sequence $\{f_n(x,\lambda)\}$ defined by

$$f_n(x,\lambda) = b_n \exp\{\lambda n^{-\alpha} (x-n)\}, \tag{3.31}$$

for $x > 0$. We have

$$f_n(m,\lambda) \geq b_m \tag{3.32}$$

if and only if

$$b_n \exp\{-\lambda n^{1-\alpha}\} \geq b_m \exp\{-\lambda\, mn^{-\alpha}\}. \tag{3.33}$$

Let $d_m = b_m \exp\{-\, mn^{-\alpha}\}$. Then

$$\begin{aligned} d_{m-1}\, d_m^{-1} &= \exp\{\lambda\, n^{-\alpha}\}\, b_{m-1}\, b_m^{-1} \\ &= (1 + \lambda n^{-\alpha} + 0(n^{-2\alpha}))\ (1-2D(m^{-\alpha}) + 0(m^{-1})) \\ &= (1 - (2D- n^{-\alpha}m^{\alpha})m^{-\alpha} + 0(m^{-1})) \end{aligned} \tag{3.34}$$

by (3.23). Since $\lambda < 2D$, there exists an $N(\lambda)$ such that $\{d_m(\lambda)\}_{m>N(\lambda)}$ is an increasing sequence.

But (3.32), and hence (3.33), holds for $m = n$. Thus, by the monotonicity of the sequence $\{d_m(\lambda)\}_{m>N(\lambda)}$, (3.33), and hence (3.32), holds for all $m > N(\lambda)$.

Since $f_n(x,\lambda)$ is an increasing function for all n,

$$\sum_{m=k}^{n} b_m \leq b_n \exp\{-\lambda\, n^{1-\alpha}\} \int_k^{n+1} \exp\{\lambda n^{-\alpha} x\}dx.$$

The right hand inequality of Lemma (3.3) follows by integration and by noting that $b_n = n^{-\phi}$.

To obtain the left hand inequality of Lemma (3.3), we note that the slope of the line AE (see Figure 3) is given by $b_n - b_{n-1}$. Therefore $D - A = b_n(b_n - b_{n-1})^{-1}$ and $(B - A) = (C - B)(b_n - b_{n-1})^{-1}$. But, for $k > N(\lambda)$, (3.31) and (3.32) imply that

$$\begin{aligned} C - B &= b_k \\ &\leq b_n \exp\{\lambda n^{-\alpha}(k-n)\}. \end{aligned}$$

It follows from (3.25) that, for $k > N(\lambda)$,

$$\begin{aligned} \sum_{m=k}^{n} b_m &\geq \text{Area } \Delta \text{ ADE} - \text{Area } \Delta \text{ ABC} \\ &\geq b_n^2 (b_n - b_{n-1})^{-1} (1 - \exp\{\lambda n^{-\alpha}(k-n)\})/2. \end{aligned}$$

Thus Lemma (3.3) is proven.

We now use Lemmas (3.1), (3.2), and (3.3) to calculate asymptotic bounds for the variance.

<u>Theorem</u> (3.2): Let the sequence $\{b_m\}$ be defined by (3.22). Then, for fixed k, if $\alpha < 1$.

$$(4D)^{-1} + o(1) \leq n^{(\phi-\alpha)} \sum_{m=k}^{n} b_m \leq (2D)^{-1} + o(1),$$

and if $\alpha = 1$ and $2D > \phi - 1 > 0$,

$$n^{\phi-\alpha} \sum_{m=k}^{n} b_m = (2D - (\phi-1))^{-1} + o(1).$$

Proof: We see from (3.16) and (3.22) that

$$D(m,n) = m^{\phi} b_{m-1}$$

if $D = 2\,AB$. Using (3.16), (3.19), and the fact that if $\alpha < 1$, then $2D > \phi - 1$, we have $n^{\phi-\alpha} b_k \to 0$, for any fixed k. Thus we need consider only the sum from $m > N$ where N is chosen to satisfy the previous lemmas.

The case in which $\alpha = 1$ and $2D \neq \phi$ follows directly from lemmas (3.1) and (3.2) since λ_1 and λ_2 of these lemmas can be chosen arbitrarily close to 2D. It is true for $\phi = 2D$ since $\sum_{m=k}^{n} b_m$ is a monotonic function of D for every k and n.

Again the fact that λ is arbitrary in Lemma (3.3) implies the right hand inequality for Theorem (3.1) when $\alpha < 1$. The left hand inequality holds from Lemma (3.3) since

$$\begin{aligned} b_n^2(b_n - b_{n-1})^{-1}/2 &= b_n(1 - b_{n-1} b_n^{-1})^{-1}/2 \\ &= n^{-\phi}\,(2Dn^{-\alpha} + 0(n^{-1}))^{-1}/2 \\ &= (4D)^{-1} + o(1) \end{aligned}$$

by (3.22) and (3.23).

We can apply Theorem (3.2) to the example in which M(x) is a quadratic function by letting $D = 2\,AB$ and $\phi = 2(\alpha - \omega)$. Thus, conditional on X_k, the asymptotic variance of $n^{\alpha-2\omega}(X_n - \theta)$ is equal to $p\;2\,A^2\,\sigma^2/C^2(2\,AB - \beta_+)$ where $p = 1$ if $\alpha = 1$, $1/2 \leq p \leq 1$ if $\alpha < 1$. This is in agreement with (3.20). It also points out the important fact that if M(x) is quadratic, the variance of X_n converges to the variance of its distribution given by Corollary (3.1). The same result for a more general class of regression functions has been obtained by many writers using different approaches.

We now consider the extent to which Lemmas (3.1), (3.2), and (3.3) rely on asymptotic theory. We first note that it is not necessary for the sequence $\{d_m\}$ to be monotonic in order that either (3.29) or (3.33) hold.

Consider the upper bounds established in lemmas (3.1) and (3.2). The appeal to asymptotic results lies in the value of N such that, for all m > N,

$$(1-Dm^{-1})^2 > (1-m^{-1})^{\lambda_2}, \tag{3.35}$$

where $\lambda_2 < 2\,D$.

If D = 1, (3.35) holds for N = 1.

Using the Taylor expansion, we find that (3.35) is equivalent to

$$1 - 2Dm^{-1} + D^2m^{-2} < \sum_{j=0}^{\infty} \left(\prod_{i=0}^{j-1} (\lambda_2 - i)\right) \frac{(-m)^{-j}}{j!}. \tag{3.36}$$

By (3.35), the right hand side of (3.36) decreases as λ_2 increases. Thus N is less than or equal to the value of N obtained for $\lambda_2 \simeq 2D$. Assuming $\lambda_2 \simeq 2D$ and considering terms up to orders (m^{-3}), we find N is approximately the solution of

$$D^2x^{-2} = \frac{2D(2D-1)}{2}x^{-2} - \frac{2D(2D-1)(2D-2)}{6}x^{-3}.$$

This solution is given by $x = 2(2D - 1)/3$. Actually, if $2D > 3$, the coefficient of (m^{-4}) is positive implying the value of N is less than $2(2D-1)/3$ up to orders (m^{-4}).

It is clear that the upper bounds in lemmas (3.1) and (3.2) are very practical bounds for moderate D. Moreover, we have already noted that, for $\alpha = 1$, (3.20) is maximized at $A = (1-2\omega)/2B$. The corresponding value of D is $2(1-2\omega)$, giving $N \simeq 2$.

The appeal to asymptotic theory for the upper bound of Lemma (3.3) lies in the minimum value of N such that

$$\exp\{\lambda n^{-\alpha}\}(1-Dm^{-\alpha})^2 \quad (1-m^{-1})^{-\phi} < 1 \tag{3.37}$$

for all $m,n > N$ (see (3.23) and (3.34)).

For the K-W process, $\phi = 2(\alpha + (a-1)\omega)$. Since a is usually less than one, we shall consider the case $\phi < 2$. It follows from (3.23) that if 2D is not significantly less than ϕ, then $(1-Dm^{-\alpha})^2 \ (1-m^{-1})^{-\phi}$ is increasing in m for quite small m. Let N_o be such that this is the case for $m > N_o$. Then, if (3.37) holds for $m = N_o$, it holds for all $m > N$. Let N_1 be the smallest value for which (3.37) holds with $m = n$, and for which $N_1 \geq N_o$. We have, for $m, n \geq N_1$,

$$\exp\{\lambda n^{-\alpha} \ (1-Dm^{-\alpha})^2(1-m^{-1})^{-\phi} < \exp\{\lambda N_1^{-\alpha}\}(1-DN_1^{-\alpha})(1-N_1^{-1})^{-\phi} < 1.$$

Thus, if either $2D \simeq \phi$ or $2D > \phi$, N_1 is a reasonable approximation for the minimum solution of (3.37).

If we let $m = n$ and neglect terms of orders $(n^{-3\alpha})$ in 3.37, we obtain an approximation of N_1 given by the minimum solution of

$1 + (\lambda-2D)x^{-\alpha} + (\lambda^2/2 + D^2 - 2D\lambda)x^{-2\alpha} < 1 - \phi x^{-1} + \phi(\phi-1)x^{-2}$. Since $\phi > 1$ for the K-W process, the coefficient of x^{-2} is positive. We therefore drop this term knowing that the resulting minimum solution is increased.

If we let $\lambda = CD$, for $0 < C < 2$, the above inequality is surely satisfied if x is greater than the minimum solution of

$$((C-2)Dy^{\alpha} + \phi y^{2\alpha-1} + (C^2-4C+2)D^2/2)y^{-2\alpha} < 0,$$

or, equivalently,

$$\phi y^{2\alpha-1} < D(2-C)y^{\alpha} - (C^2-4C+2)D^2/2.$$

For $2 - \sqrt{2} < C < 2$ the quadratic in C is negative. Thus if C is so restrained, our approximation is less than the solution of

$$\phi y^{2\alpha-1} = D(2-C)y^{\alpha},$$

or

$$y = (\phi/(2D-\lambda))^{\frac{1}{1-\alpha}} .$$

3.6 Estimation of $E(X_n - \theta)^2$.

A comparison of (3.11) and (3.12) with (3.16) quickly reveals that (3.16) is generally not the correct expression for $E((X_{n+1} -)^2 | X_k)$. However, it often serves as a reasonable approximation when the variables $\{X_m \pm c_m\}_{m=k_o}^{n}$ lie in a region of θ in which M(x) is reasonably approximated by a quadratic function. We assume in this section that the process has been in such a region for $m = k_o, \ldots, n$.

A natural estimator of $n^{\alpha-2\omega} E(X_n - \theta)^2$ is the right hand expression of (3.16) after the parameters have been replaced by their estimators and the expression has been suitably normalized. Using this estimator, the problem becomes one of choosing a good value for k and estimating the parameters.

Assume, for the moment, that estimates of the parameters are available. The squared bias expression in (3.16) is minimized as either k decreases or X_k approaches θ. This suggests that a good choice for X_k would be close to X_{n+1} (the estimate of θ) and such that $k \simeq k_o$. Using this method, the squared bias expression can be ignored if $n - k_o$ is of any consequence.

A natural estimate of the variance of X_n, conditional on k, would be to substitute the estimate for $-4\,AM''(\theta)$ (= 2D) into the upper bounds given in lemmas (3.1), (3.2), and (3.3). If n-k is moderately large, this estimate (after normalizing by $n^{\alpha-2\omega}$) closely approximates the corresponding estimate of the asymptotic variance given in Corollary (3.1).

The parameters, for which we have assumed estimates exist, are σ^2 and $M''(\theta)$. Carrying the assumption that M(x) is approximately quadratic a little further, we can obtain estimates of these parameters using regression analysis.

It is interesting to consider Fisher's information for $M''(\theta)$ when $Y(x) \sim N(-B(x-\theta)^2, \sigma^2)$. We have

$$-\frac{\partial^2}{\partial B^2} \ln N(\cdot,\cdot) = -\frac{\partial^2}{\partial B^2} (C - (y+B(x-\theta)^2)^2/2\sigma^2)$$

$$= \frac{\partial}{\partial B} ((y + B(x-\theta)^2)(x-\theta)^2/2\,\sigma^2)$$

$$= (x-\theta)^4/2\,\sigma^2.$$

Thus, conditional on X_i, $i = k_o, \ldots, n$, the information is given by (assuming $2[m^{2a\omega}]$ observations are taken on each step)

$$(\sqrt{2}\sigma)^{-2} \sum_{m=k_o}^{n} ((X_m - c_m - \theta)^4 + (X_m + c_m - \theta)^4) m^{2a\omega}$$

$$\geq O(\sum_{m=k_o}^{n} c_m^2\, n^{2a\omega}).$$

If $c_n = n^{-\omega}$ and ω is arbitrarily close to its lower bound $(\alpha/2(3-a))$, the information is bounded below by a quantity of

$$0(n^{1 - \alpha(2-a)/(3-a) - \varepsilon})$$

for arbitrary small ε. This expression increases as either α decreases or as a increases until $a = 2$. Moreover, since $(2-a)/(3-a) < 2/3$ for small a, the usual regression estimators of σ^2 and $M''(\theta)$ will yield consistent estimates if $M(x)$ is quadratic.

CHAPTER 4. THE SIGNED KIEFER-WOLFOWITZ PROCESS

4.1 Definition of the Signed Kiefer-Wolfowitz Process

Chapter 3 is a development of a generalization of the K-W process in which the corrective variable $((Y(X_n-c_n) - Y(X_n+c_n))^{T_n})$ is constrained to lie between $\pm T_n$. This corrective variable can be characterized as influencing both the direction and magnitude of the adjustment at the n'th step. In this chapter we shall develop a modification of the K-W process in which the corrective variable influences only the direction of this adjustment.

Definition (4.1). Let $\{a_n\}$, $\{c_n\}$, and $\{r_n\}$ be sequences of positive numbers where $[r_n] \geq 1$. Let Y(x) denote a random variable with mean M(x) and the distribution function $F(\cdot|x)$. Let

$$\delta(x_n,c_n) = [r_n]^{-1} \sum_{i=1}^{[r_n]} \delta_i(x,c_n),$$

where

$$\delta_i(x,c_n) = \begin{cases} 1 \text{ if } Y_i(x-c_n) > Y_i(x+c_n), \\ 0 \text{ if } Y_i(x-c_n) = Y_i(x+c_n), \\ -1 \text{ if } Y_i(x-c_n) < Y_i(x+c_n), \end{cases}$$

$i = 1, \ldots, [r_n]$. Let X_1 be fixed, and let the random variable sequence $\{X_n\}$ be defined by

$$X_{n+1} = X_n - a_n c_n^{-1} \delta(X_n, c_n), \tag{4.1}$$

where $Y_i(X_n - c_n)$ and $Y_i(X_n + c_n)$, $i = 1, \ldots, [r_n]$, are random variables which, conditional on $(X_1, c_1), \ldots, \delta(X_{n-1}, c_{n-1}), X_1, \ldots, X_{n-1}$, and $X_n = x_n$, are distributed independently as $F(\cdot|x_n - c_n)$ and $F(\cdot|x_n + c_n)$.

This process will be called the Signed Kiefer-Wolfowitz process, and will be referred to as the SKW process. When $[r_n] \equiv r$ (a positive integer) the SKW becomes the process first proposed by Wilde [13]. Wilde did not attempt to develop the process except to note that it seemed less sensitive to the variance of Y(x) than the K-W process. The only other attempt to study this process seems to be Springer's [11] work which was discussed in Chapter 1.

The SKW process is intrinsically different from the K-W process. The K-W process seeks the value (θ) such that $E[Y(\theta)] \geq E[Y(x)]$ for all x. The SKW process seeks the value (θ') such that

$$P[Y(\theta') > Y(x) | Y(\theta') \neq Y(x)] \geq 1/2$$

for all x. Notice that if Y(x) is such that $E[Y(x)] > E[Y(x)]$ for $x \neq x'$ implies that $P[Y(x) > Y(x') | Y(x) \neq Y(x')] > 1/2$, then $\theta = \theta'$. In particular, if $Y(x) \sim N(M(x), \sigma^2(x))$, then $\theta = \theta'$.

However, the behavior of the two processes can be markedly different because of their different emphases on the magnitude of the corrective variable.

When $[r_n] = 1$, the SKW process can be considered as a limiting case of the CKW process. Let $a_n = A^s n^{-\alpha}$ for the SKW process, and let $a_n = An^{-\alpha}$ for the CKW process with $T_n \equiv T$. Then, if $A^s = AR^{-1}$, the corrective term for the CKW process

$$(AT^{-1} n^{-\alpha} c_n^{-1} (Y(X_n - c_n) - Y(X_n + c_n))^T)$$

converges in probability to the correction term of the SKW process $(An^{-\alpha} c_n^{-1} \delta(X_n, c_n))$ for any fixed n as $T \to 0$.

Before deriving any basic results it is helpful to discuss the random variable $(\delta(x, c_n))$. We have

$$E\ \delta(x, c_n) = P[Y(x-c_n) > Y(x+c_n)]$$

$$- P[Y(x-c_n) < Y(x+c_n)]. \qquad (4.2)$$

If Y(x) is a continuous random variable for all x, (4.2) can be expressed as

$$E\ \delta(x, c_n) = \int_{-\infty}^{\infty} \left(\int_{\infty}^{y} dF(z|x+c_n) \right) dF(y|x-c_n)$$

$$- \int_{-\infty}^{\infty} \int_{-\infty}^{y} dF(z|x-c_n) dF(y|x+c_n) \qquad (4.3)$$

We now consider the important special case in which Y(x) is a normal random variable with density function

$$f(y,x) = (2\pi\sigma^2(x))^{-\frac{1}{2}} \exp\{-(1/2)((y - M(x))/\sigma(x))^2\}.$$

Substitution into (4.3) gives

$$E\ \delta(x,\varepsilon) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{y} f(z,x+\varepsilon)dz \right) f(y,x-\varepsilon)dy$$
$$- \int_{-\infty}^{\infty} \left(\int_{-\infty}^{y} f(z,x-\varepsilon)dz \right) f(y,x+\varepsilon)dy. \qquad (4.4)$$

Since we shall be interested in the behavior of $\delta(x, c_n)$ for c_n small and x in a neighborhood of θ, we shall need estimates of

$\frac{\partial}{\partial x}$ E $\delta(x,\varepsilon)$ and $\frac{\partial}{\partial\varepsilon\partial x}$ E $\delta(x,\varepsilon)$.

Assume that

1. $M'(x)$ is continuous and there exists an open interval I containing θ such that for $x \varepsilon I$ $M''(x)$ is continuous,
2. $0 < \sigma^2(x) < \infty$, $\sigma'(x)$ is continuous, and $\sigma''(x)$ is continuous for $x \varepsilon I$.

Then the Taylor expansion of $f(y,x+\varepsilon)$ gives

$$f(y,x+\varepsilon) = f(y,x) + \varepsilon\left(\frac{\partial}{\partial\tau} f(y,x+\ )\right)_{\tau = \xi}$$

for $0 < \xi < \tau$. Since

$$\frac{\partial}{\partial\tau} f(y,\tau) = \frac{(y - M(\tau))^2\sigma'(\tau)}{\sigma^3(\tau)} + \frac{(y - M(\tau))M'(\tau)}{\sigma^2(\tau)} - \frac{\sigma'(\tau)}{\sigma(\tau)} f(y,\tau), \quad (4.6)$$

assumptions (1) and (2) guarantee that the last expression in (4.5) can be dominated by an integrable function for x in a bounded interval. We can, therefore, express (4.4) as

$$E\ \delta(x,\varepsilon) = \int_{-\infty}^{\infty} [\int_{-\infty}^{y} (f(z,x+\varepsilon) - f(z,x-\varepsilon))dz](f(y,x) + \varepsilon g(y,x,\varepsilon)dy \quad (4.7)$$

where $g(y,x,\varepsilon)$ can be dominated by an integrable function.

Since the term in brackets is bounded by $\pm$ 1 and goes to zero for every z as $\varepsilon \to 0$, we have from (4.7) and assumptions (1) and (2), that

$$\lim_{\varepsilon\to 0} \frac{\partial}{\partial\varepsilon} E\delta(x,\varepsilon) = 2 \int_{\infty}^{\infty}\left(\frac{d}{dx} \int_{-\infty}^{y} f(z,x)\ dz\right) f(y,x)dy$$

$$= 2\,\sigma(x) \int_{-\infty}^{\infty} \left(\frac{d}{dx} \left[(z - M(x)/\sigma(x)) \right] \right) f^2(y,x)dy$$

$$= M'(x)/\sqrt{\pi}\,\sigma(x) \tag{4.8}$$

In order to evaluate $\frac{\partial}{\partial\varepsilon}\ \frac{\partial}{\partial}\ E\delta(x,\varepsilon)$ at $x = \theta$ and $\varepsilon = 0$, we note that, for x in a closed subinterval of I, the continuity of $M''(x)$ and $\sigma''(x)$ is sufficient to guarantee the existence and continuity of $\frac{\partial}{\partial\varepsilon}\ \frac{\partial}{\partial}\ E\ \delta(x,\varepsilon)$ and $\frac{\partial}{\partial x}\ \frac{\partial}{\partial\varepsilon}\ E\ \delta(x,\varepsilon)$. Moreover, these second partials are equal, and we have by (4.6), (4.7), and (4.8) that

$$\lim_{c_n \to 0} \frac{\partial}{\partial\varepsilon}\ \frac{\partial}{\partial x}\ E\delta(x,\varepsilon) \Big|_{x=\theta,\ \varepsilon=c_n}$$

$$= - \left(\frac{d}{dx} M'(x)/\sigma(x)\ \sqrt{\pi} \right)_{x=\theta}$$

$$= - M''(\theta)/\sigma(\theta)\sqrt{\pi} \tag{4.9}$$

since $M'(\theta) = 0$.

Thus if $Y(x)$ is a normal random variable, we have

$$E\ \delta(x,c_n) = -c_n\ M'(x)/\sigma(x)\sqrt{\pi} + 0(c_n),$$

and, for $|x - \theta|$ sufficiently small,

$$E\ \delta(x,c_n) = -c_n(x-\theta)M''(\theta)/\sigma(\theta)\sqrt{\pi} + 0(c_n(x-\theta)).$$

If, more generally, $Y(x)$ is a continuous random variable whose density function $f(y(x))$ is continuous in x for all y, then it follows from Definition (4.1) that $E\ \delta_i^2(x,c_n) \equiv 1$ and $E\ \delta_i(x,c_n) \to 0$ as $c_n \to 0$.

Since, conditional on $X_n = x$, $\delta_i(x,c_n)$, $i = 1, \ldots, [r_n]$, are independent random variables, we have

$$\lim_{c_n \to 0} \text{Var } \delta(\theta,c_n) = 1. \quad (4.10)$$

4.2 Asymptotic Theory of the SKW Process

Corollary (4.1): Let $\{X_n\}$ be a SKW process with $r_n = n^{2a\omega}$. Let $\{\mu_n\}$ be a real number sequence, θ be a real number, a and ω be non-negative numbers, and n_o, A^s, C, ν, and α be positive numbers such that

a. if $n > n_o$, then $E\ \delta(\mu_n,c_n) = 0$, and if $x \neq \mu_n$, then $(x-\mu_n)\ E\ \delta(x,c_n) > 0$;

b. for every x, $E\ \delta(x,0) = 0$ and $\frac{\partial}{\partial\varepsilon} E\ \delta(x,\varepsilon)$ is continuous, and if $B = \sup c_n$, then

$$\sup_{|\varepsilon|<B,x} \left|\frac{\partial}{\partial\varepsilon} E\ \delta(x,\varepsilon)\right|/(1+|x|) < \infty$$

c. for every x in an open interval (I) containing θ and for every $|\varepsilon| < B$, $(\delta/\delta x)\ E\ \delta(x,\varepsilon)|_{x=\theta,\ y=0}$ equals zero, $(\delta^2/\delta\varepsilon\delta x)$ $E\ \delta(x,y)$ is continuous, and the sequence $\{(\partial^2/\partial\varepsilon\partial x)\ E\ \delta(x,\varepsilon)|_{x=\theta,\ \varepsilon=c_n}\}$ converges to $\Lambda > 0$;

d. $\text{Var } \delta(x,c_n)$ is continuously convergent to σ^2 at $x = \theta$;

e. $F(y|\cdot)$ is Borel measurable for all y;

f. (1) $1/2 < \alpha \leq 1$, $n^\alpha a_n \to A^s$, $n^\omega c_n \to C$, $|\mu_n - \theta| = 0(n^{-\nu})$, and $\beta/2 = \alpha/2 + (a - 1)\omega < \nu$;

(2) if $a < 1$, then $a > 1 - (2\alpha - 1)/2\omega$;

(3) and $\beta_+ = 0$ if $\alpha < 1$, $= \beta < 2\ A^s\ \Lambda$ if $\alpha = 1$.

Then $X_n \to \theta$ a.e., and $n^{\beta/2}(X_n - \theta) \overset{a}{\sim} N(0, \sigma^2(A^s)^2/C^2(2A^s\Lambda - \beta_+))$.

<u>Proof</u>: Make the correspondence with Theorem (2.1) as follows:

$$a_n' = a_n c_n^{-1},\ Z_n(x) = \delta(x_n, c_n),\ \text{and}\ R_n(x) = E\ \delta(x_n, c_n).$$

To verify that conditions (i) through (vii) of Theorem (2.1) are satisfied, we first note that condition (i) is a direct consequence of condition (a).

It follows from condition (b) that

$$\sup_{n,x} n^{\omega}|E\ \delta(x_n, c_n)|/(1+|x|)$$

$$= \sup_{n,x} n^{\omega} c_n (|\frac{\partial}{\partial\varepsilon} E\ \delta(x_n,\varepsilon)|_{y=\xi_n(x)})/(1+|x|)$$

where $|\xi_n(x)| < c_n$. The last expression is finite by condition (b), and therefore condition (ii) is verified.

To verify condition (iii) we note that condition (b) implies the existence of a function sequence $\{\xi_n(x)\}$ such that

$$|E\ \delta(x_n, c_n)| = c_n|\frac{\partial}{\partial\varepsilon}\ E\ \delta(x_n,\varepsilon)|_{\varepsilon=\xi_n(x)}|$$

where $|\xi_n(x)| < c_n$. By condition (f.(1)), c_n converges to a limit, and hence the function sequence $\{\xi_n(x)\}$ converges to a function, say $\xi(x)$. Moreover, condition (b) implies $\xi(x)$ is continuous, and therefore the function $|(\partial/\partial\varepsilon)\ E\ \delta(x,\varepsilon)|_{\varepsilon=\xi(x)}|$ is continuous and attains its infimum $(k(\delta_1,\delta_2))$ over the set $[\delta_1 \le |x-\theta| \le \delta_2]$.

But $\mu_n \to \theta$ by (f.(1)). Therefore if $0 < \delta_1 \le \delta_2 < \infty$, there exists an N such that for every $n > N$ we have that $[\delta_1 \le |x-\theta| \le \delta_2] \cap [|x-\theta| \le |\mu_n-\theta|] = \phi$. This implies by (a) that $k(\delta_1,\delta_2) > 0$. Thus condition (iii) is verified by noting that the divergence of

$$\Sigma a_n' \inf_{\delta_1 \leq |x-\theta| \leq \delta_2} |R_n(x)|$$

is implied by the divergence of

$$\Sigma a_n' \; c_n \; k(\delta_1,\delta_2) \;\; (= O(\Sigma\, n^{-\alpha})).$$

We have by (a) and (c) that

$$\begin{aligned} E\,\delta(x_n,c_n)/c_n(x-\mu_n) &= c_n^{-1}\left(\frac{\partial}{\partial x}\, E\,\delta(x,c_n)\right)_{x=\eta} \\ &= \left(\frac{\partial^2}{\partial\varepsilon\partial x}\;\; E\,\delta(x,\varepsilon)\right)_{x=\eta,\varepsilon=\xi}, \end{aligned}$$

where $|\eta - \mu_n| < |x - \mu_n|$ and $|\xi| < c_n$. The last expression is continuously convergent to Λ at $x = \theta$ by condition (c). Thus condition (iv) is satisfied.

Condition (v) follows from condition (d) and the fact that $|\delta(x_n,c_n)| \leq 1$.

Since the $P[Y(x-c_n) > Y(x+c_n)]$ can be written as the limit of Lebesgue-Stieltjes sums, which by condition (e) are Borel measurable, it is clear that the distribution function of $\delta(x,c_n)$ is Borel measurable for all (y). Thus condition (vi) is satisfied.

Condition (vii) can be shown to hold by direct substitution for the case $T_n \equiv \infty$ and $\rho = 2a\omega$.

Thus Corollary (4.1) is proven.

The conditions in this corollary are somewhat unintuitive. This arises from the fact that the SKW process explores a response curve

($E\,\delta(x,c_n)$) which, although often closely related, is not the response curve of the observed random variables. However, as in Corollary (3.1), the assumptions of Corollary (4.1) are satisfied in practice almost without exception. This fact will become clearer in the following examples.

Example 1. Suppose $Y(x) \sim N(M(x), \sigma^2(x))$. Let $\{X_n\}$ be a SKW process which satisfies assumptions (1) and (2) preceding (4.5) and assumption (f.) of Corollary (4.1) with $\omega > 0$. We have yet to determine ν.

In addition, assume that

3. $(x-\theta)\,M'(x) < 0$ for $x \neq \theta$, $= 0$ for $x = \theta$,

4. $M'''(x)$ is continuous for $x \in I$;

5. and $\sup_x(|M'(x)| + |\sigma'(x)|)/\sigma(x)(1 + |x|) < \infty$,

We shall show that $n^{\beta/2}(X_n-\theta) \overset{a}{\sim} N(0,(A^s)^2/C^2(2A^s\Lambda - \beta_+))$, where $\Lambda = -M''(\theta)/\sigma(\theta)\sqrt{\pi}$, using Corollary (4.1).

We first note that (3) implies condition (a) since $P[Y(x) > Y(x')] > 1/2$ if and only if $E\,Y(x) > E\,Y(x')$ for independent non-degenerate normal random variables. It then follows by condition (4) and Corollary (3.1) that $\nu = 2\omega$.

Condition (e) is an immediate consequence of the continuity of $M(x)$ and $\sigma(x)$.

We now summarize the relevant results obtained earlier for this special case:

1. $\frac{\partial}{\partial\varepsilon} E\,\sigma(x,\varepsilon)$ is continuous;

2. $\{\text{Var}\,\sigma(x,c_n)\}$ is continuously convergent to 1 at $x = \theta$;

3. for every $x \in I$ and every $|\varepsilon| < B$,

$$\left(\frac{\partial}{\partial x} E\ \delta(x,\varepsilon) \right)_{x=\theta,\ \varepsilon=0} = 0, \text{ the sequence}$$

$$\frac{\partial^2}{\partial\varepsilon\partial x} E\delta(x,\varepsilon)\Big|_{x=\theta,\ \varepsilon=c_n} \text{ converges to}$$

Λ ($= - M''(\theta)/\sigma(\theta)\sqrt{\pi}$), and $\frac{\partial^2}{\partial\varepsilon\partial x} E\ \delta(x,\varepsilon)$ is continuous.

Thus, the assertion of asymptotic normality is established if we can show that the

$$\sup_{|\varepsilon| < B,x} \left|\frac{\partial}{\partial\varepsilon} E\ \delta(x,\varepsilon)\right| / (1 + |x|) < \infty.$$

Taking the partial with respect to ε under the integral signs in (4.3) (The continuity of $M(x)$ and $\sigma(x)$ assures us that this can be done.) we obtain integrals of the form

$$\int_{-\infty}^{\infty} \left(\int_{-\infty}^{y} \frac{\partial}{\partial\varepsilon} f(z,\varepsilon)\ dz \right) f(y,\varepsilon)\ dy$$

and

$$\int_{-\infty}^{\infty} \left(\int_{-\infty}^{y} f(z,\varepsilon)\ dz \right) \frac{\partial}{\partial\varepsilon} f(y,\varepsilon)\ dy.$$

An examination of (4.5) reveals that assumption (5) is sufficient to guarantee that the above supremum is finite.

Thus the desired asymptotic normality of $n^{\beta/2}(X_n - \theta)$ is established.

The variances of the respective normal distributions to which the SKW process and the K-W process converge are respectively

$$(A^s)^2/C^2(2\ A^s|M''(\theta)|/\sigma(\theta)\ \sqrt{\pi} - \beta)$$

and

$$2A^2\sigma^2/C^2(4\ A|M''(\theta)| - \beta)$$

when $\alpha = 1$, and

$$\sqrt{\pi}A^s\sigma(\theta)/2\ C^2|M''(\theta)|$$

and (4.11)

$$A\ \sigma^2(\theta)/\ 2\ C^2|M''(\theta)|$$

when $1/2 < \alpha < 1$. It is evident that the K-W process is asymptotically more sensitive to the variance of the observations.

Before leaving this example we note that if $A^s = A\ \delta(\theta)/\sqrt{\pi}$, the expressions in (4.11) are identical.

Example 2: Suppose $Y(x)$ has an exponential density function given by

$$f(y,x) = \begin{cases} \lambda(x)\ e^{-\lambda(x)y}, & x \geq 0 \\ 0 & , x < 0. \end{cases}$$

Let $\{X_n\}$ be a SKW process satisfying assumption (f) of Corollary (4.1) with $\omega > 0$. Let $M(x) = \lambda^{-1}(x)$ and assume that

1. $M'(x)$ is continuous and $(x-\theta)M'(x) < 0$ by $x \neq \theta$;
2. $\sup_x |M'(x)|/M(x)\ (1 + |x|) < \infty$;
3. and there exists an open interval I containing θ such that $M'''(x)$ is continuous.

Then $n^{\beta/2}\ (X_n-\theta) \overset{a}{\sim} N(0,(A^s)^2/C^2(2\ A^s\Lambda - \beta_+))$ where $\Lambda = -\ M''(\theta)/M(\theta)$.

We have

$$P[Y(x-\varepsilon) < Y(x+\varepsilon)] = 1 - \lambda(x+\varepsilon)/(\lambda(x+\varepsilon) + \lambda(x-\varepsilon))$$

$$= 1 - M(x-\varepsilon)/(M(x+\varepsilon) + M(x-\varepsilon))$$

by the last expression in (4.3).

It follows that

$$E\ \delta(x,\varepsilon) = (M(x-\varepsilon) - M(x+\varepsilon))/(M(x-\varepsilon) + M(x+\varepsilon)). \qquad (4.12)$$

It is clear from (4.12) that the value of ν is less than 2ω as in Corollary (3.1).

With the possible exception of condition (b), the conditions of Corollary (4.1) can be easily verified using assumptions (1) and (2).

We have

$$\sup_{|\varepsilon| < B,x} \left|\frac{\partial}{\partial\varepsilon} E\ \delta(x,\varepsilon)\right| / (1 + |x|)$$

$$= \sup_{|\varepsilon| < B,x} |M'(x)|/M(x)(1 + |x|) < \infty$$

by assumption (2).

It remains to evaluate the Var $\delta(\theta,0)$ and Λ. The $\lim_{c_n \to 0}$ Var $\delta(x,c_n) \to 1$ by (4.9) for all x and, in particular, for $x = \theta$.

By assumption (3) the value of Λ is given by

$$\frac{\partial}{\partial x}\ \frac{\partial}{\partial\varepsilon} E\ \delta(x,\varepsilon)\Big|_{\varepsilon=0\ x=\theta},$$

or

$$\Lambda = - M''(\theta)/M(\theta)$$

$$= - M''(\theta)/\sigma(\theta) \tag{4.13}$$

since $M(\theta) = \lambda^{-1}(\theta) = \sigma(\theta)$.

A comparison of (4.13) with (4.9) suggests that Λ may be generally of the form $- k\, M''(\theta)/\sigma(\theta)$ where k is a constant characterizing a family of distribution functions. The next example gives an important case when this is not true.

Example 3: Suppose $Y(x)$ is a Bernoulli random variable with the $P[Y(x) = 1] = p(x)$. Suppose that assumptions (1) and (3) of Example 2 hold with $M(x)$ replaced by $p(x)$. In addition, assume that

(2) $\sup_x |p'(x)|(1+x) < \infty$.

It is easily seen that these assumptions are sufficient to verify the conditions of both Corollary (4.1) and Corollary (3.1). Moreover, in this case the SKW process and the K-W process are identical. Thus, using the results of Corollary (3.1), we find that

$$n^{\beta/2}(X_n-\theta) \overset{a}{\sim} N(0,(A^s)^2\sigma^2/C^2(2\, A^s\Lambda - \beta_+))$$

where $\Lambda = - 2\, p''(\theta)$, and $\sigma^2 = 2\, p(\theta)\, (1-p(\theta))$.

This example shows that the corrective variable in the SKW process is a transformation of the observed random variable into a discrete, bounded random variable with a regression function that is less than, equal to, and greater than zero as x is less than, equal to, or greater than μ_n. That is, the SKW process is simply a K-W process after this transformation on the observed random variables has been made.

These examples indicate that, at least asymptotically, the SKW process is more dependent than the K-W process on the family of distributions (e.g. normal, exponential) from which the observed random variables come. However, for the special cases of the normal and exponential families, the SKW process is less sensitive to the variance of the parent distribution.

4.3 Sample Properties of the SKW Process

Unlike the K-W process, whose small sample properties are largely determined by $M(x)$ and $\sigma(x)$, the small sample properties of the SKW process often depend on the distribution function in a more intrinsic manner. This makes a general discussion of the small sample properties of the SKW process very difficult.

However, a few general remarks can be made. Firstly, the effect of using the sign of $(Y(X_n-c_n) - Y(X_n+c_n))$ is somewhat similar to that of constraining an observed random variable; both operations lessen the effect of an observation from the tails of the distribution. Secondly, if the SKW process is approaching θ along a constant, relatively small slope, the expected behavior is fairly well approximated by (3.11) and (3.12) in which the constants B and σ^2 are characteristic of the family of observed random variables. And, thirdly, if the SKW process is approaching θ along a steep slope, the approach is very uniform, and the rate is determined by the sequences $\{a_n\}$ and $\{c_n\}$.

It is clear from the nature of the SKW process that is it impossible to consider an example corresponding to $M(x)$ being quadratic for the K-W process. That is, since $\delta(x,c_n)$ is bounded by ± 1, it is impossible for $E\,\delta(x,c_n) = k\,c_n(x-\theta)$ to hold for all x. This implies that the moments of $n^{\beta/2}(X_n-\theta)$ generally do not converge to those of its asymptotic normal distribution unless possibly when X_n is restricted to lie in a bounded interval containing θ. To study the expected rate of approach to θ therefore requires a slight modification of the method used for the K-W process.

Let $\{X_n\}$ be a SKW process. Let I be an open interval of θ such that

$$E\,\delta(x,c_n) = c_n\,T_n(x)(x-\theta) \tag{4.14}$$

where $0 < \underline{T} \leq T_n(x) \leq \bar{T} < \infty$. Let A_k be the event that $X_m \in I$ for $m \geq k$, and let X_n^* denote the random variable X_n conditional on A_k. That is, assume that each member of the random variable sequence $\{X_m\}_{m=k}^{\infty}$ lies in I.

It is not unreasonable to study the random variable sequence $\{X_n^*\}$ for two reasons. Since $X_n \to \theta$ a.e. by Corollary (4.1), $P(A_k) \to 1$ as $k \to \infty$, and since $P(A_k) \to 1$, $n^{\beta/2}(X_n^*-\theta) \to n^{\beta/2}(X_n-\theta)$ in probability and therefore has the same asymptotic distribution.

We have, for $a_n = A^s n^{-\alpha}$ and $c_n = C\,n^{-\omega}$,

$$\begin{aligned} X_{n+1}^* - \theta &= X_n^* - \theta - a_n c_n^{-1}\,\delta(X_n^*,c_n) \\ &= (1 - A^s T_n(X_n^*) n^{-\alpha})(X_n^*-\theta) - A^s C^{-1} n^{-(\alpha-\omega)} Z(X_n^*) \end{aligned} \tag{4.15}$$

where, conditional on X_n^*, $E\,Z(X_n^*) = 0$ and $\text{Var}\,Z(X_n^*) = \text{Var}\,\delta(X_n^*, c_n)$. In much the same manner that we obtained (3.16), we obtain

$$E((X_{n+1}^*-\theta)^2\; X_k) \leq D(k,n)(X_k^*-\theta)^2 + A^2C^{-2}\bar{\sigma}^2 \sum_{m=k}^{n} D(m+1,n)m^{-2(\alpha-\omega)} \qquad (4.16)$$

where $D(k,n) = \prod_{j=m}^{n} (1 - A^s\,\underline{T}\,j^{-\alpha})^2$ and $\bar{\sigma}^2 = \sup_{n,x} \text{Var}\,\delta(x,c_n)$. The reverse inequality holds if $\underline{T}$ and $\bar{\sigma}^2$ were replaced by $\bar{T}$ and $\underline{\sigma}^2$. Quite clearly, since $T_n(X_n^*) \overset{a.e.}{\to} \Lambda$ and $\text{Var}\,\delta(X_n^*, c_n) \overset{a.e.}{\to} \text{Var}\,\delta(\theta,0)$, it could be shown that

$$n^\beta\, E(X_n^*-\theta)^2 \to (A^s)^2C^{-2}\sigma^2(\theta)/(2A^s\,\Lambda - \beta_+) \qquad (4.17)$$

where $\beta = \alpha - 2\omega$.

Since it will be useful in the next section, we shall show that

$$n^{2\beta}\, E(X_n^*-\theta)^4 \to 3\phi^2$$

where

$$\phi = A^2C^{-2}\sigma^2(\theta)/(2\,A\,\Lambda - \beta_+).$$

If we raise (4.15) to the 4'th power and take expectations, we obtain terms of the following forms:

$$O(1)\; E(X_n^*-\theta)^4,\; O(n^{-2(\alpha-\omega)})\; E[X_n^*-\theta)^2\; Z^2(X_n^*)],$$

$$O(n^{-3(\alpha-\omega)})\; E[X_n^*-\theta)\; Z^3(X_n^*)],\ \text{and}\ O(n^{-4(\alpha-\omega)})\; E[Z^4(X_n^*)]$$

since $E[(X_n^*-\theta)^3\; Z_n(X_n^*)|X_n^*] = 0$. Using (4.16) and Hölder's inequality, the last two expressions are of smaller order than the second. Thus we can write

$$E(X_{n+1}^{*}-\theta)^4 \leq D^2(k,n)(X_n^{*}-\theta)^4$$

$$+ 6\ \phi\ (A^s)^2 C^{-2}(\bar{\sigma}^2 + o(1) \sum_{m=k}^{n} D^2(m+1,n) m^{-3\alpha+4\omega} \qquad (4.18)$$

Letting

$$b_m = \prod_{j=m+1}^{n} (1-D\ j^{-\alpha})^{\eta}\ m^{-\phi}, \quad \eta > 0$$

it is seen that Lemmas (3.1), (3.2) and (3.3) as well as Theorem (3.2) hold with 2 D replaced by ηD. Thus the last expression in (4.18) is less than

$$n^{-2(\alpha-2\omega)} 6\ \phi\ (A^s)^2 C^{-2}(\bar{\sigma}^2 + o(1))/(4\ A^s\ \Lambda - 2\ \beta_+)$$

or, equivalently,

$$n^{-2(\alpha-2\omega)}\ 3\ \phi^2\ (1 + o(1)).$$

It should be noted that, when $\alpha = 1$, the restriction that $\alpha - 2\ \omega < 2\ A\ \Lambda$ implies $\phi - 1 = 2(\alpha-2\omega) < 4\ A^s\ \Lambda = \eta\ D$ in the above expressions.

Using (3.17) and (3.18) with $2\ B = A^s\ \Lambda$, we obtain

$$n^{2\beta}\ E(X_n^{*}-\theta)^4 \to 3\ \phi^2, \qquad (4.19)$$

as we might have expected.

4.4 Estimation of $E(X_n - \theta)^2$

The fact that the form of Λ is generally unknown for the SKW process adds a new complexity to the estimation of the variance of X_n. If we can assume a particular distribution for $Y(x)$, we can often evaluate Λ as was illustrated in the preceding examples. An estimate of the variance of X_n could then be calculated using the method developed in Chapter 3. However, the distribution of Y(x) is not always known. We now develop a method of estimating the variance of X_n which does not require this knowledge.

Denote the Var X_n^* by V_n. Let b_m be an increasing subsequence of the integers k, ..., n where b_N=max(b_m). Define $\hat{V}_n$ by

$$\hat{V}_n = N^{-1} \sum_{m=1}^{N} b_m^{\beta} (X_m^* - X_n^*)^2. \tag{4.20}$$

We shall study the properties of $\hat{V}_n$ through a study of the properties of

$$V_n' = N^{-1} \sum_{m=1}^{N} b_m^{\beta} (X_n^* - \theta)^2. \tag{4.21}$$

The asterisk, which indicates that X_m, m=k, ..., n, all lie in a neighborhood of θ in which (4.14) is a reasonable approximation for $E\,\delta(x,c_n)$, will be omitted in the rest of this discussion.

It follows from (4.17) that V_n' is asymptotically unbiased.

To obtain the variance of V_n' we note that

$$\begin{aligned} \mathrm{Var}[V_n'] &= N^{-2} \sum_{m,m'=1}^{N} (b_m\, b_{m'}) \quad \mathrm{Cov}[(X_{b_m} - \theta)^2\ (X_{b_{m'}} - \theta)^2] \\ &= N^{-2} \sum_{m,m'=1}^{N} (b_m\, b_{m'}) \ (E[X_{b_m} - \theta)^2\ (X_{b_{m'}} - \theta)^2] \\ &\qquad - E(X_{b_m} - \theta)^2\, E(X_{b_{m'}} - \theta)^2) \end{aligned}$$

$$\leq 2N^{-2} \sum_{m' \geq m=1}^{N} (b_m b_{m'})^{\beta} \{E[(X_{b_m} - \theta)^2 E^2(X_{b_{m'}} - \theta | X_{b_m})$$

$$+ \mathrm{Var}(X_{b_{m'}} - \theta | X_{b_m})] - E(X_{b_m} - \theta)^2 E(X_{b_{m'}} - \theta)^2\}$$

$$- N^{-2} \sum_{m=1}^{N} b_m^2 E(X_{b_m} - \theta)^4$$

$$\leq 2N^{-2} \sum_{m' \geq m=1}^{N} (b_m b_{m'})^{\beta} D(b_m, b_{m'}) E(X_{b_m} - \theta)^4$$

where $D(b_m, b_{m'}) = \prod_{j=b_m}^{b_{m'}} (1 - A \underline{T} j^{-\alpha})^2$ by (4.16).

Using (4.19) we obtain

$$\mathrm{Var}(V_n') \leq (6 \phi^2 + o(1)) N^{-2} \sum_{m' \geq m}^{N} D(b_m, b_{m'}). \tag{4.22}$$

Consider the case $\alpha = 1$. We have from (3.22) and (3.28) that for every $\lambda < 2 A \underline{T}$ there exists a $N(\lambda)$ such that, for every m, $n > N(\lambda)$,

$$D(b_m, b_{m'}) \leq (b_m / b_{m'})^{\lambda}. \tag{4.23}$$

Let $m' = m + 1$, and suppose that the sequence $\{b_m\}$ is chosen such that (b_m / b_{m+1}) is independent of m. Then, for some constant B,

$$\sum_{i=1}^{m-1} (\ln (b_{i+1}) - \ln b) = (m-1)B$$

or, equivalently,

$$\ln(b_m) - \ln(b_1) = \ln B.$$

Thus b_m is of the form (s^m).

Let b_m be so defined with $s > 1$. Then from (4.22) and (4.23) we obtain

$$\mathrm{Var}(V_n') \leq (6\,\phi^2 + o(1)\; N^{-2} \sum_{m=0}^{N} (N-m)s^{-m\lambda}$$

$$< (6\,\phi^2 + o(1)\; N^{-1}\,(1-s^{-N\lambda})/(1-s^{-\lambda}).$$

Since N is approximately the solution of $s^N = n$, we have

$$\ln(n)\;\mathrm{Var}(V_n') \leq 6\,\phi^2\,\ln(s)/(1-s^{-\underline{T}}) + o(1).$$

Consider the case $\alpha < 1$. We have from (3.22) and (3.32) that for every $\lambda < 2\,A^S\underline{T}$ there exists and $N(\lambda)$ such that for every m, n > $N(\lambda)$

$$D(b_m, b_{m'}) \leq (b_m/b_{m'})^{\phi}\,\exp[\lambda\, b_{m'}^{-\alpha}\,(b_m - b_{m'})] \tag{4.24}$$

where $\phi > 0$.

Let $m' = m + 1$ and suppose the sequence $\{b_m\}$ is chosen such that the term in brackets is asymptotically independent of m. Then, if $b_{m+1} = b_m(1 + o(m^{-1}))$, we have

$$m^{-1}\,b_m^{1-\alpha} \to c.$$

It is easily seen that $b_m = (\tau\,m)^{1-\alpha)^{-1}}$ has the required property.

Let b_m be so defined. Then, for fixed k and $m' = m + k$, the term in brackets becomes $[-\,\lambda\,\tau\,k + o(1)]$. Thus by (4.22) and (4.24) we obtain

$$\mathrm{Var}(V_n') \leq (6\,\phi^2 + o(1))\; N^{-2} \sum_{m=0}^{N} (N-m)(e^{-\lambda\tau})^k$$

$$< (6\,\phi^2 + o(1)\; N^{-1}\,(1-e^{-N\lambda\tau})/(1-e^{-\lambda\tau}).$$

Since N is approximately the solution of

$$(\tau\ N)^{(1-\alpha)^{-1}} = n,$$

or, equivalently,

$$N \simeq \tau^{-1}\ n^{1-\alpha},$$

we have

$$n^{1-\alpha}\ \mathrm{Var}(V_n') \leq 6\ \phi^2\ \tau/(1-e^{-\lambda\tau}) + o(1).$$

Since θ is unknown, it is natural to use $\hat{V}_n$ instead of V_n'.

We have

$$E\ \hat{V}_n = N^{-1} \sum_{m=1}^{N} b_m\ E(X_m - X_n)^2$$

$$= N^{-1} \sum_{m=1}^{N} b_m (E(X_{b_m} - \theta)^2 - 2\ E(X_{b_m} - \theta)(X_n - \theta) + E(X_n - \theta)^2)$$

$$\leq N^{-1} \sum_{m=1}^{N} b_m (E(X_{b_m} - \theta)^2 - 2D^{1/2}(b_m, n) E(X_{b_m} - \theta)^2 + E(X_n - \theta)^2)$$

$$\leq N^{-1} \sum_{m=1}^{N} (\phi + o(1)).$$

Thus $\hat{V}_n$ is asymptotically unbiased. Moreover, for $\alpha < 1$,

$$D^{1/2}(b_m, n) E(b_m - \theta)^2 = 0(b_m^{-\beta} \exp\{-n^{-\alpha}\ (n - b_m)\}$$

$$= o(E(X_n - \theta)^2).$$

Thus for $\alpha < 1$, $\hat{V}_n$ is usually a conservative estimate of V_n.

This method of estimation is obviously very inefficient for α close to one. It does have the advantages of being relatively easy to calculate and of being generally applicable. It should be noted that the order of the variance of $\hat{V}_n$ depends on the number of steps and not on the number of observations. This is undesirable when we wish to economize by taking more observations on each step and taking fewer steps.

CHAPTER 5. TWO EXAMPLES

We now illustrate the application of some of the results of Chapter 3 and of Chapter 4 by means of two hypothetical examples. Both examples entail designing an experiment to emphasize the factors that influence the choice of the sequences $\{a_n\}$, $\{c_n\}$, and $\{T_n\}$.

<u>Example 1</u>: A research worker is studying the relationship between temperature and the rate of reproduction of a certain strain of bacteria. One of his objectives is to estimate the temperature (θ) that maximizes the average rate of reproduction. He emphasizes that he wishes to minimize the $E\,(M(\theta) - M(\theta))^2$ rather than $E(\hat{\theta} - \theta)^2$.

From past experience he is reasonably confident that θ lies between 80 and 110 degrees Fahrenheit and that the average rate of reproduction is reasonably constant in the interval $(\theta-5, \theta+5)$. For temperatures below $\theta - 5$ degrees the average rate of reproduction changes approximately one unit for each 10 degree change in temperature, and for temperatures above $\theta + 5$ degrees the average rate of reproduction changes two to three units for each 10 degree change in temperature. The standard deviation is expected to be approximately three units at temperatures below θ and to moderately increase as temperatures rise in a twenty degree range above θ. He also mentions that for practical reasons he can perform the experiment at only 20 different temperatures.

The above discussion indicates that the relation between the average rate of reproduction and temperature may be represented by Figure 4.

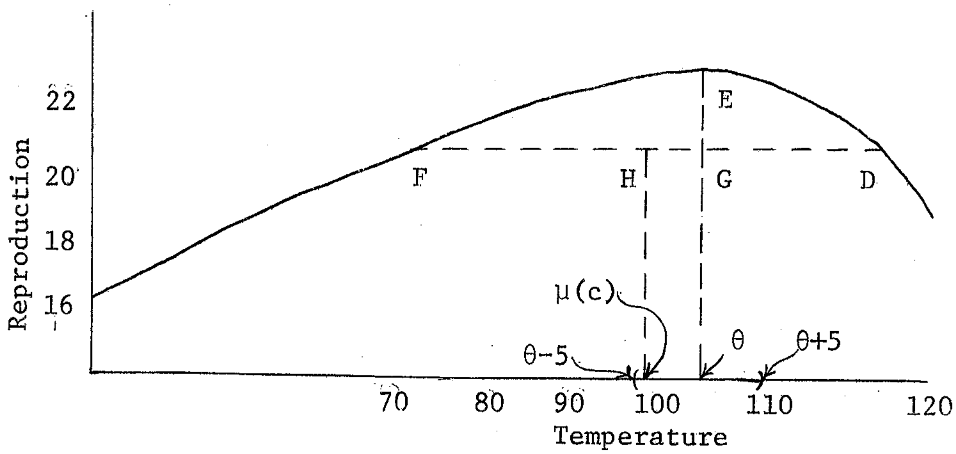

Figure 4. Graph of Reproduction versus Temperature

We shall illustrate the design of this experiment using two CKW processes. In order that the first process (X_n') approaches θ along a moderate slope and that the second process (X_n'') approaches θ along a steep slope, we set $X_1' = 80$ and $X_1'' = 110$. A reason in favor of using two such processes in practice rather than one is that two processes would usually generate observations more uniformly and over a larger range of temperatures. This allows the experimenter to study the whole relationship more reliably.

Since so many of the results in this paper depend on (3.15) being a reasonable approximation of the process, we must define an interval I of θ for which this is true. Then we must design the process such that X_n is in the region at least for the last few steps. Since we are practically restricted to 20 steps, we shall choose the sequence $\{a_n\}$ such that with a high probability X_n has reached this region after 10 steps.

Before doing this we must choose the sequence $\{c_n\}$. This sequence should be chosen with two thoughts in mind. Firstly, c_n should be large as long as $|X_n - \theta| > c_n$; that is, X_n is still searching for a neighborhood of θ. Secondly, when $|X_n - \theta| < c_n$, c_n should be small enough so that the expected correction term always directs the process to a region in which $M(x) \simeq M(\theta)$.

The maximum value of c_n for which the last statement is true can be approximated by considering Figure 4. Let D - F denote the length of the line DF, and let $c = (D - F)/2$. We see from the picture that if $c_n < c$, then $|\mu(c_n) - \theta| < 5$. The experimenter has noted that there is not a biologically significant difference between $M(\mu(c))$ and $M(\theta)$ if $|\mu(c) - \theta| < 5$. Thus we seek the maximum value of c for which $|\mu(c) - \theta| < 5$.

Let ρ be such that $(G - F) = 2\rho c$. Then if s_1 and s_2 represent the slopes of lines FE and ED, we have

$$E - G = 2c\rho s_1 = 2c(1-\rho)s_2,$$

or

$$\rho = s_2/(s_1+s_2).$$

Since

$$5 > |\mu(c) - \theta|$$

$$\simeq 2c(s_2/(s_1+s_2) - 1/2),$$

$$c \simeq 5(s_1+s_2)/(s_2-s_1)$$

for $s_2 > s_1$. Thus, for our example,

$$c \simeq 5(.1 + .2)/(.2 - .1)$$

$$= 15.$$

Consider the choice of c_n for X'_n. Initially c_n should be large enough so that if $x_n + c_n < \theta$, then $X'_{n+1} > X'_n$ with a high probability. Since $P[|Y(x) - M(x)| < \sigma] \simeq .6$, we may choose $c_1(=C)$ as the solution of

$$2\sigma = M(x_1 + C) - M(X_1 - C)$$

$$\simeq .2C,$$

or

$$C \simeq 30.$$

If, in addition, we require that

$$C(10^{-\omega}) = 15,$$

then $\omega \simeq .3$. With these values for C and ω we find that $c_{20} \simeq 11.5$.

Similarly for the choice of c_n for X'_n we obtain

$$2\sigma = M(x_1 + C) - M(x_1 - C)$$

$$\simeq .4C.$$

or

$$C \simeq 15.$$

In this case, c_1 is sufficiently small to discount a serious bias due to $\mu(c_1) \neq \theta$. Thus, we can set $\omega = 0$, or $c_n \equiv 15$.

It should be noted that we could start the process at $n = k > 1$ and let $n = k, \ldots, 20 + k$. By doing so, we can adjust the change in c_n to satisfy $c_1 = D_1$, $c_{10} = D_2$, and $c_{20} = D_3$ where D_1, D_2, and D_3 are chosen constants.

An examination of Figure 4. reveals that if $c_n < 15$ and $|X_n - \theta| < 5$, then (3.15) is a reasonable approximation for X_{n+1}. Equations (3.11) and (3.12) are useful for choosing the sequence $\{a_n\}$ such that X_n reaches the interval I by the n'th step.

In choosing $a_n (=An^{-\alpha})$ for X'_n we note first that this process will be somewhat exploratory due to the small slope. Therefore, we do not want a_n to decrease too rapidly as that would make the process too sensitive to the first few steps. Thus we might set $\alpha = 3/4$. Since $\theta \varepsilon (80, 110)$, $|X_1 - (\theta - 5)| < 25$. Thus a conservative value for A, which reasonably assures that $X_n \varepsilon I$ for some $n \leq 10$, is the solution of

$$2(1.1)A \sum_{m=1}^{10} m^{-.75} \simeq .2(3.8)A$$

$$= 25, \tag{4.25}$$

by (3.11). We therefore set A = 30.

In choosing $a_n (=An^{-\alpha})$ for X''_n, we note first that this process should be in a neighborhood of θ quite early due to the steep slope. Thus, we want the sequence $\{a_n\}$ to optimize the behavior of X''_n under this assumption. Asymptotically this is done by setting $\alpha = 1$ with the restriction that $A > \beta/2\ M''(\theta)|$, where $\beta = 1 - 2\omega = 1$. To obtain a conservative estimate of $|M''(\theta)|$, we note from (3.14) and Figure 4. that

$$-B(\theta-10-\theta) \simeq M'(\theta-10)$$

$$\simeq .1,$$

or

$$B \simeq M''(\theta)$$

$$\simeq .1.$$

Thus $A > 2.5$. Furthermore, by the note following (3.20), asymptotically the optimal choice of A is 5. However, if we require A to be greater than the solution of the equation corresponding to (4.25), we must have

$$A > 25.$$

There are numerous bases on which the sequence $\{T_n\}$ of constraints may be chosen. We illustrate one method for X_n''. In the region of steepest slope the expected correction term is

$$AC^{-1} n^{-1} (M(x-c_n) - M(x+c_n)) \simeq 6AC^{-1} n^{-1}.$$

If we set $T_n = DCA^{-1}n$ then the maximum correction would be D. It is clear that for $D > 6$ this constraint does not seriously affect the approach of X_n' in the region of steep slope. By letting $D < 10$, the possible bias arising from a large movement to the left of θ in the early stages of the experiment is diminished. Obviously, the effect of T_n diminishes very quickly as n gets large.

We can get an approximation for the variance of X_n' and X_n'' using (3.16) and Lemmas (3.1) and (3.3). We have for X_n' that

$$D(m,n) \leq \prod_{j=m+1}^{n} (1 - 2(.1)30j^{-\alpha})^2$$

giving

$$\mathrm{Var}(X_{21} - \theta | X_k) \simeq \frac{2(9)20^{-.75+.6}}{12} (\exp\{6(20)^{-.75}\}$$

$$- \exp\{-(20 - k)6(20)^{-.75}\})$$

by Lemma (3.3). For k = 10, we have

$$\mathrm{Var}(X_{21} - \theta | X_k) \simeq 1.8. \tag{4.26}$$

If $X_{10} - \theta| < 5$, then

$$E^2(X'_{20} - \theta | X_{10}) \leq \prod_{j=11}^{20} (1 - 6j^{-.75})^2\ 25$$

$$\leq (11/20)^{-1.5+.6} \exp\{-12(10)(20)^{-.75}\}25$$

$$\simeq 0.$$

by (4.24) with $b_m = 10$ and $b_{m'} = 20$.

We have for X''_n that

$$D(m,n) \leq \prod_{j=m+1}^{n} (1 - 2(.1)25\ j^{-1})^2.$$

Using Lemma (3.1), we obtain

$$\mathrm{Var}(X''_{21} - \theta | X_{10}) \simeq \frac{(9)(25)^2}{(15)^2(10-1)} (20^{-1} - (.5)^5\ 10^{-1})$$

$$\simeq .14. \tag{4.27}$$

If $|X''_n - \theta| < 5$, then

$$E^2(X''_{20} - \theta | X_{10}) \leq \prod_{j=11}^{20} (1 - 5j^{-1})^2\ 25$$

$$\simeq (11/20)^{10}$$

$$\simeq .03.$$

Before comparing (4.28) and (4.27), we must adjust for the fact that X_n'' approached θ along the steepest slope. The corresponding adjustment in (4.25) changes the restriction on A^s to $A^s > 15$. With this new value for A^s we obtain

$$\mathrm{Var}(X_{21}' - \theta | X_{10}) \simeq .9.$$

We conclude this example by noting that much the same procedure would be followed if the SKW process were considered.

Example 2. Suppose it has been noted that the addition of a moderate amount of a certain medication to the daily diet of some chickens has improved their health, and it is desired to estimate the optimal amount.

In this type of problem, it is often difficult to quantify the observed variable (health). Moreover, the shape of the response function depends on the manner in which we choose to quantify this variable. Thus it seems natural to use a method of estimation that depends on relative magnitudes rather than absolute magnitudes. We therefore illustrate this example using the SKW process.

The corrective variable ($\delta(X_n, c_n)$) is defined to be -1 if the chickens which are given the amount of medication ($X_n - c_n$) are healthier than the chickens which are given the amount of medication ($X_n + c_n$). As we noted in the discussion of Example 3, the random variable $\delta(x, c_n)$ generates a response curve that equals zero at μ_n. We shall suppose that the experimenter is reluctant to assume that

$\mu_n \equiv \theta$. If for no other reason than it is somewhat meaningless to view the expected health as a function of diet, we shall also assume that the experimenter wishes to minimize $E(\hat{\theta} - \theta)^2$.

Since it is not too unreasonable to think of $\mu_n - \theta$ as the bias of X_n, we might let c_n equal some constant (greater than 1) times the bias we are willing to accept. The values C and ω should be chosen such that c_m is large during the period of time we expect the process to be searching for the region of θ, and approximately equal to c_n when in this region.

We have such little knowledge of Λ in this case that is is probably wise to set $\alpha < 1$ and thereby obviate the risk of not satisfying the inequality $2A^s \Lambda > \beta$ for the case $\alpha = 1$. Moreover, due to the unknown nature of Λ, we are restricted to estimating the variance of X_n using the method proposed in the previous section. This latter fact may cause us to diminish α still further and to compensate for the corresponding increase in the variance by decreasing A^s.

Using the principles employed in Example 4, we could evaluate the sequences $\{a_n\}$ and $\{c_n\}$ more specifically.

CHAPTER 6. SUMMARY AND SUGGESTIONS FOR FURTHER RESEARCH

6.1 Summary

A large class of stochastic approximation procedures can be represented by

$$X_{n+1} = X_n - a_n' \, Z(X_n) \qquad (6.1)$$

where, conditioned on X_n, $Z(X_n)$ is an observable random variable independent of X_i and $Z(X_i)$, $i = 1, \ldots, n-1$. In particular the Kiefer-Wolfowitz (K-W) process admits this representation with $a_n' = a_n c_n^{-1}$ and $Z(X_n) = Y(X_n-c_n) - Y(X_n+c_n)$ where, conditional on $X_n = x$, $Y(X_n-c_n)$ and $Y(X_n+c_n)$ are independent random variables with means $M(x-c_n)$ and $M(x+c_n)$ respectively.

An extension of Burkholder's [2] theorem on the asymptotic distribution of the stochastic approximation procedures, which can be represented by (6.1), is given in Chapter 2. This extension allows the number of observations taken on each step to be of $O(n^{-\rho})$, $\rho \geq 0$, allows the random variable $|Z(X_n)|$ to be constrained to lie between bounds of $O(n^{-\zeta})$, $\zeta \geq 0$, and admits a more general sequence of constants a_n'. The practical consequences of this extension for the K-W process are discussed in Chapter 3.

The small sample mean and variance of X_n, conditional on X_k, are studied through a consideration of some specific examples. The case in which $M(x)$ is a quadratic function is considered since $M(x)$

is often closely approximated by a quadratic function in a neighborhood of θ. It is shown by means of some approximations that, after suitable normalization, the variance of the normal random variable to which the K-W process converges in distribution is a conservative estimate of the $\mathrm{Var}(X_n|X_k)$. This asymptotic variance depends on the unknown parameters σ^2 and $M''(\theta)$. It is shown that an efficient estimator of $M''(\theta)$ is also a consistent estimator of $M''(\theta)$ when $Y(x) \sim N(M(x),\sigma^2)$.

A modification of the K-W process first proposed by Wilde [13] is defined in Chapter 4. It satisfies (6.1) with $a'_n = a_n c_n^{-1}$ and with $Z(X_n)$ being the mean of $[r_n]$ independent random variables defined by

$$\delta_1(x,c_n) = \begin{cases} 1 \text{ if } Y_i(X_n-c_n) > Y_i(X_n+c_n), \\ 0 \text{ if } Y_i(X_n-c_n) = Y_i(X_n+c_n), \\ -1 \text{ if } Y_i(X_n-c_n) < Y_i(X_n+c_n), \end{cases}$$

where, conditional on $X_n = x$, $Y_i(X_n-c_n)$ and $Y_i(X_n+c_n)$, $i = 1, \ldots, [r_n]$, are independent random variables with mean $M(x-c_n)$ and $M(x+c_n)$, respectively. This process is termed the Signed Kiefer-Wolfowitz (SKW) process. The asymptotic distribution of the SKW process is obtained under quite general assumptions on $Y(x)$ if the sequences a_n, c_n, and r_n are of the form $n^{-\rho}$, $\rho > 0$. It is seen that the order of the variances of the K-W process and the SKW process are the same, but that the variance of the SKW process shows a greater dependency on the nature of the distribution of $Y(x)$.

The variance of X_n conditional on X_k is again studied by considering specific examples. It is seen that this variance depends on an unknown parameter, the nature of which depends on the form of the distribution of Y(x). The estimator of the variance,

$$\mathrm{Var}(X_n) = N^{-1} \sum_{m=1}^{N} b_m^{\beta} (X_{b_m} - X_n)^2$$

where $\{b_m\}$ is an increasing subsequence of the integers, 1, ..., n, and $n^{\beta}(X_n-\theta)$ is asymptotically normal, is proposed since this estimator does not depend on the unknown parameter. This estimator is shown to be consistent for somewhat restricted class of sequences $\{b_m\}$.

In Chapter 5, two hypothetical examples are given to illustrate the designing of an experiment in which these procedures are used.

6.2 Suggestions for Further Research

Although the results in the paper provide some added insight into the behavior of the K-W and SKW processes and methods of estimating the variance of the estimate of θ are presented, both of these areas need further study. One very promising possibility would be a numerical study comparing these stochastic approximation procedures with the Method of Steepest Ascent.

LIST OF REFERENCES

1. Blum, J. R. 1954. Approximation methods which converge with probability one. Ann. Math. Statist. 25:382-386.

2. Burkholder, D. L. 1956. On a class of stochastic approximation processes. Ann. Math. Statist. 27:1044-1059.

3. Chung, K. L. 1954. On a stochastic approximation method. Ann. Math. Statist. 25:463-483.

4. Dvoretsky, A. 1956. On stochastic approximation. Proc. Third Berkley Symposium. 1:39-56.

5. Fabian, V. 1967. Stochastic approximation with improved asymptotic speed. Ann. Math. Statist. 38:191-200.

6. Fabian, V. 1968. On asymptotic normality in stochastic approximation. Ann. Math. Statist. 39:1327-1332.

7. Hodges, J. L., Jr., and E. L. Lehman. 1956. Two approximations to the Robbins-Monro process. Proc. Third Berkeley Symposium. 1:95-104.

8. Kiefer, J. and J. Wolfowitz. 1952. Stochastic approximation of the maximum of a regression function. Ann. Math. Statist. 23:462-466.

9. Robbins, H. E. and S. Monro. 1951. A stochastic approximation method. Ann. Math. Statist. 22:400-407.

10. Sachs, J. 1958. Asymptotic distribution of stochastic approximation procedures. Ann. Math. Statist. 29:373-405.

11. Springer, B. G. F. 1969. Numerical optimization in the presence of random variability. The single factor case. Biometrika. 56:65-74.

12. Venter, J. H. 1967. An extension of the Robbins-Monro procedure. Ann. Math. Statist. 38:181-190.

13. Wilde, D. J. Optimum seeking methods. Prentice-Hall, Inc. 1964. Englewood Cliffs, N. J.