Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds.

AN EMPIRICAL SENSITIVITY ANALYSIS OF THE KIEFER-WOLFOWITZ ALGORITHM AND ITS VARIANTS

Marie Chau
Huashuai Qu

Department of Mathematics
University of Maryland, College Park
College Park, MD, 20742 USA

Michael C. Fu
Ilya O. Ryzhov

Robert H. Smith School of Business
University of Maryland, College Park
College Park, MD, 20742 USA

ABSTRACT

We investigate the mean-squared error (MSE) performance of the Kiefer-Wolfowitz (KW) stochastic approximation (SA) algorithm and two of its variants, namely the scaled-and-shifted KW (SSKW) in Broadie, Cicek, and Zeevi (2011) and Kesten's rule. We conduct a sensitivity analysis of KW with various tuning sequences and initial start values and implement the algorithms for two contrasting functions. From our numerical experiments, SSKW is less sensitive to initial start values under a set of pre-specified parameters, but KW and Kesten's rule outperform SSKW if they begin with well-tuned parameter values. We also investigate the tightness of an MSE bound for quadratic functions, a relevant issue for determining how long to run an SA algorithm. Our numerical experiments indicate the MSE bound for quadratic functions for the KW algorithm is sensitive to the noise level.

1 INTRODUCTION

Consider the stochastic optimization problem

$$\max_{x\in\Theta} f(x) = E[\tilde{f}(x)], \tag{1}$$

where $\tilde{f}(x)$ is a noisy observation of $f(x)$, and the objective is to find x^* maximizing f. Various iterative methods such as Robbins-Monro (RM), Kiefer-Wolfowitz (KW), and simultaneous perturbation stochastic approximation (SPSA) have been used to estimate x^*; see Robbins and Monro (1951), Kiefer and Wolfowitz (1952) and Spall (1992) for details. Each of these SA algorithms follows the underlying recursion

$$X_{n+1} = \Pi_\Theta\left(X_n + a_n\hat{\nabla}f(X_n)\right), \tag{2}$$

when finding the zero of $\nabla f(x)$ in (1), where Π_Θ is a projection of X_{n+1} back into the feasible region Θ if $X_{n+1}\notin\Theta$, a_n is the step size or gain size, and $\hat{\nabla}f(X_n)$ is an estimate of $\nabla f(X_n)$. The projection operator Π_Θ is particularly important in the constrained optimization setting. Initially, the asymptotic theory underlying SA considered functions that satisfy specific global conditions; however, later research has shown it is only necessary for the requirements to hold on a compact set Θ that contains the optimum. Since the optimum is unknown, the compact set must be large enough to increase the likelihood that $x^*\in\Theta$; however, this may increase the potential of an algorithm to perform poorly (Andradóttir 1995). The gain sequence $\{a_n\}$ could be deterministic (as with the popular rule $a_n = \theta_a/n$ where $\theta_a\in\mathbb{R}$) or adaptive. Adaptive rules adjust the step size based on the ongoing performance of the algorithm; one well-known example is the rule by Kesten (1958), which decreases the step size only when there is a directional change in the iterates. See George and Powell (2006) for an extensive review of both deterministic and stochastic step sizes. The choice of

$\{a_n\}$ has a significant impact on the performance of the algorithm, and this impact is quite difficult to characterize theoretically.

Asymptotic convergence properties of the KW algorithm and its variations have been a major research focus in SA. Dupac (1957), Derman (1956), Fabian (1967), Tsybakov and Polyak (1990), and Polyak and Juditsky (1992) have all proven convergence in MSE for various assumptions and modifications of the KW algorithm. However, in practice, where the run-time is finite, a good finite-time bound for the MSE is useful. Broadie, Cicek, and Zeevi (2011) derived such a bound using techniques similar to those in Dupac (1957). The MSE bound depends on certain problem-dependent constants, which are typically difficult to calculate in practice. However, in the special case where f is quadratic, the bound can be computed in closed form, allowing us to observe its tightness.

We focus on the one-dimensional KW algorithm, which generates $\hat{\nabla} f(X_n)$ in (2) using finite differences. Although theoretical convergence can be guaranteed by satisfying certain requirements, practical performance depends on the choice of tuning sequences. In addition to selecting a gain sequence $\{a_n\}$ in (2), the KW algorithm requires an additional task of choosing a finite difference step-size sequence $\{c_n\}$ for the gradient. The finite-time performance of KW depends on both sequences $\{a_n\}$ and $\{c_n\}$. Because of the sensitivity of the KW algorithm to the tuning sequences, it is essential to choose an appropriate pair. In practice, KW could have the following shortcomings: long oscillatory period if the gain sequence $\{a_n\}$ is "too large," degraded convergence rate if $\{a_n\}$ is "too small," and poor gradient estimates if the gradient estimation step-size sequence $\{c_n\}$ is "too small."

In this paper, we conduct an empirical investigation of the sensitivity of KW and two of its adaptive variants, namely Kesten's rule and the scaled and shifted KW (or SSKW) algorithm of Broadie et al. (2011). Our goal is to identify problem characteristics that exert a strong impact on algorithm performance, even in the presence of theoretical guarantees. For example, in the numerical results reported in Broadie et al. (2011), SSKW outperforms the KW algorithm in terms of both MSE and oscillatory behavior in finite time; however, this result is obtained using what seem to be nearly worst-case parameter setting for KW. We replicate these results, but we also find that the performance of KW can be significantly improved over a fairly wide choice of parameter settings. Although the worst-case performance of SSKW is much better than that of KW, it is also the case that KW provides the best performance in a significant proportion of problem instances. In addition, we find that Kesten's rule performs similar to KW, and sometimes better, when both algorithms begin with the same initial start value. We also investigate the finite-time MSE bound in Broadie et al. (2011) and characterize instances where this bound is tight. These results underscore the well-known difficulty of tuning, even for adaptive versions of KW. We hope that the results of this analysis will provide guidance to practitioners on the challenges involved in implementing KW or similar algorithms.

2 THE KW ALGORITHM AND ITS VARIANTS

In this paper, we focus on the truncated KW algorithm and its variants to find x^* using symmetric differences to estimate the gradient. Both algorithms use the underlying recursion

$$X_{n+1} = \Pi_\Theta \left(X_n + a_n \left(\frac{\tilde{f}(X_n + c_n) - \tilde{f}(X_n - c_n)}{c_n} \right) \right),$$

where X_1 is an arbitrary starting point, $\tilde{f}(x) \sim H(\cdot|x)$ and the tuning sequences $\{a_n\}$ and $\{c_n\}$ satisfy a set of assumptions. The accuracy and precision of the algorithms are highly reliant on the choice of the tuning sequences in finite-time. For the KW algorithm, the $\{a_n\}$ and $\{c_n\}$ are predetermined as opposed to the SSKW, where the sequences are dynamically adjusted throughout the algorithm.

In Section 2.1, we discuss the KW convergence result from Kiefer and Wolfowitz (1952). We introduce the finite-time MSE bound of KW derived in Broadie, Cicek, and Zeevi (2011) for quadratic functions in Section 2.2. We describe two adaptive algorithms, Kesten's rule and SSKW, in Sections 2.3 and 2.4, respectively.

2.1 Kiefer-Wolfowitz Algorithm

Kiefer and Wolfowitz (1952) proved that the KW algorithm converges asymptotically to the true solution. Let $\{a_n\}$ and $\{c_n\}$ be positive tuning sequences satisfying the conditions

$$c_n \to 0,\ \sum a_n = \infty,\ \sum a_n c_n < \infty,\ \sum a_n^2 c_n^{-2} < \infty.$$

Also suppose that the function $f(x)$ is strictly increasing for $x < x^*$, strictly decreasing for $x > x^*$, $\int_{-\infty}^{\infty}(y - f(x))^2 dH(y \mid x) < \infty$ and satisfies the following regularity conditions:

1) There exist positive constants β and B such that

$$|x' - x^*| + |x'' - x^*| < \beta \Longrightarrow |f(x') - f(x'')| < B|x' - x''|.$$

2) There exist positive ρ and R such that

$$|x' - x''| < \rho \Longrightarrow |f(x') - f(x'')| < R.$$

3) For every $\delta > 0$ there exists a positive $\pi(\delta)$ such that

$$|x - x^*| > \delta \Longrightarrow \inf_{\frac{\delta}{2} > \varepsilon > 0} \frac{|f(x+\varepsilon) - f(x-\varepsilon)|}{\varepsilon} > \pi(\delta).$$

Then X_n converges to x^* a.s.

The regularity conditions require $f(x)$ to be locally Lipschitz in a neighborhood of x^*, preventing $f(x)$ from changing drastically in the feasible region and from being very flat outside a neighborhood of x^* so the iterates approach the optimum. Although the KW algorithm converges asymptotically, its finite-time performance is dependent on the choice of tuning sequences, $\{a_n\}$ and $\{c_n\}$. If the current X_n is in a relatively flat region of the function and the a_n is small, then the convergence will be slow. On the other hand, the X_n is located in a very steep region of the function and $\{a_n\}$ is large, then the iterates will experience a long oscillation period. If $\{c_n\}$ is too small, the gradient estimates using finite differences could be extremely noisy.

2.2 Finite-time MSE Bound

Broadie, Cicek, and Zeevi (2011) derived a finite-time bound for the MSE of the KW algorithm by applying a similar technique as in Dupac (1957) used to prove convergence in MSE. We briefly summarize the bound as follows. First, we make the following assumptions on the function $f(x)$:

F1) There exist positive constants K_0, K_1, and C_0 such that for every $c \in [0, C_0]$,

$$-K_1(x - x^*)^2 \le \frac{f(x+c) - f(x-c)}{c}(x - x^*) \le -K_0(x - x^*)^2.$$

F2) $f'(x)(x - x^*) < 0$ for all $x \in R \backslash \{x^*\}$.

We also assume that the tuning sequences satisfy:

S1) $a_n / c_n^2 \le (a_{n+1}/c_{n+1}^2)(1 + A a_{n+1})$ for all $n \ge 1$,
S2) $a_n \to 0$ as $n \to \infty$,

with $0 < A < 2K_0$. Then,

$$E(X_{n+1} - x^*)^2 \le C a_n / c_n^2 \text{ for all } n \ge 1, \tag{3}$$

where C is a constant explicitly defined as

$$C = \max\left\{\frac{\sigma^2}{\xi}, \max_{1\le n\le n_0}\left\{\frac{c_n^2}{a_n}B_{n+1}\right\}\right\},$$

and

$$\begin{aligned}
D_n &= K_1^2 A a_n^2 + (K_1^2 - 2AK_0)a_n - 2K_0 - A,\\
n_0 &= \begin{cases} 1 & \text{if } D_n < 0 \text{ for all } n \ge 1\\ \sup\left\{n \ge 1 : (K_1^2 - 2AK_0)a_n + K_1^2 A a_n^2 \ge 2K_0 - A\right\} + 1 & \text{otherwise}\end{cases},\\
\xi &= -\sup\{A - 2K_0 + (K_1^2 - 2AK_0)a_n + K_1^2 A a_n^2 : n \ge n_0\},\\
B_n &= X_1^2 \prod_{i=1}^{n} p_i + \sum_{i=2}^{n-1} q_i \prod_{j=i+1}^{n} p_j + q_n,\\
p_i &= 1 - 2a_i K_0 + K_1^2 a_i^2, \text{ for } i = 1,2,\ldots,n,\\
q_i &= \frac{a_i^2}{c_i^2}\sigma^2, \text{ for } i = 1,2,\ldots,n,\\
\sigma^2 &= \sup_{x\in\Theta} \mathrm{Var}[\tilde{f}(X_n + c_n) - \tilde{f}(X_n - c_n)|X_n = x].
\end{aligned}$$

Equation (3) does not guarantee convergence in MSE, but rather establishes a finite bound for each iteration. The bound is thus more useful when it is tight. In Section 3, we investigate the tightness of (4) by comparing the bound with the exact MSE of simple quadratic functions of the form $f(x) = \alpha x^2$ where $\alpha < 0$ and the optimal $x^* = 0$. The exact MSE can be computed as follows:

$$E(X_{n+1} - x^*)^2 = X_1^2 \Pi_{i=1}^{n}(1 + 2\alpha a_i)^2 + \frac{\sigma^2}{2}\sum_{k=1}^{n-1}\frac{a_k^2}{c_k^2}\Pi_{j=k+1}^{n}(1 + 2\alpha a_j)^2 + \frac{a_n^2\sigma^2}{2c_n^2}. \tag{4}$$

2.3 Kesten's Rule

Kesten (1958) proposes a stochastic step-size, which decreases when there is a directional change in the iterates, i.e. $(X_{n+1} - X_n)(X_n - X_{n-1}) < 0$. The idea behind this adaptive step-size is that, if the iterates continue in the same direction, there is reason to believe they are approaching the optimum and should not decrease the momentum. In our numerical experiments, we consider $a_n = \theta_a/n$, so the gain size a_n changes to a_{n+1} only if there is a change in direction.

2.4 Scaled and Shifted Kiefer-Wolfowitz Algorithm

SSKW attempts to prevent slow convergence in finite-time by using adaptive tuning sequences $\{a_n\}$ and $\{c_n\}$ in the algorithm. In general, the SSKW has two phases: scaling and shifting.

Scaling Phase

- Step 0. Specify the following parameters:
  - h_0 = number of forced boundary hits
  - γ = scale up factor for $\{c_n\}$
  - k_a = maximum number of shifts of $\{a_n\}$
  - v_a = initial upper bound of shift
  - ϕ_a = maximum scale up factor for $\{a_n\}$
  - k_c = maximum number of scale ups for $\{c_n\}$

- g_0 = maximum number of gradient estimates in scaling phase
- m_{max} = maximum number of adaptive iterations

Choose $X_1 \in [l+c_1, u-c_1]$. Initialize $sh = 0, sc = 0$. Let $n = 1$, $m = 1$ and $g = 1$.

- Step 1. For $m \leq h_0$ and $g \leq g_0$, generate an estimate $\hat{\nabla} f(X_n)$ using symmetric differences and compute X_{n+1} using the recursion in (2). If $X_{n+1} \in (l+c_n, X_n)$, go to Step 2. If $X_{n+1} \in (X_n, u-c_n,)$, go to Step 3. If $X_{n+1} \notin (l+c_n, u-c_n)$, go to Step 4.
- Step 2. Scale $\{a_n\}$ up by $\alpha = \min(\phi_a, (u-c_{n+1}-X_n)/(X_{n+1}-X_n))$ and use $\{\alpha a_n\}$ for the remaining iterations. Set $X_{n+1} = l+c_n$. Let $n = n+1$, $m = m+1$, $g = g+1$ and go to Step 1.
- Step 3. Scale $\{a_n\}$ up by $\alpha = \min(\phi_a, (l+c_{n+1}-X_n)/(X_{n+1}-X_n))$ and use $\{\alpha a_n\}$ for the remaining iterations. Set $X_{n+1} = u-c_n$. Let $n = n+1$, $m = m+1$, $g = g+1$ and go to Step 1.
- Step 4. Scale $\{c_n\}$ up by γ and use γc_n for the remaining iterations. Set $X_{n+1} = \min\{u-c_{n+1}, \max\{X_{n+1}, l+c_n\}\}$. Let $n = n+1$, $g = g+1$ and go to Step 1.

Shifting Phase

- Step 1. For $n \leq m_{max}$, generate an estimate $\hat{\nabla} f(X_n)$ using symmetric differences and compute X_{n+1} using (2). If $X_{n+1} > u-c_{n+1}$ and $X_n = l+c_n$ or if $X_{n+1} < l+c_{n+1}$ and $X_n = u-c_n$, go to Step 2. If $X_{n+1} > X_n = u-c_n$ or $X_{n+1} < X_n = l+c_n$, go to Step 3.
- Step 2. If $sh \leq k_a$, find the smallest integer β' such that $X_{n+1} \in (l+c_n, u-c_n)$ with $a_{n+\beta'}$. Set $\beta = \min(v_a, \beta')$ and shift $\{a_n\}$ to $\{a_{n+\beta}\}$. If $\beta = v_a$, set $v_a = 2v_a$. Let $sh = sh+1$. Go to Step 4.
- Step 3. If $(sc \leq k_c)$, scale $\{c_n\}$ up by γ and use $\{\gamma c_n\}$ for the remaining iterations. Let $sc = sc+1$.
- Step 4. Set $X_{n+1} = \min\{u-c_{n+1}, \max\{X_{n+1}, l+c_n\}\}$ and let $n = n+1$ and go to Step 1.

The scaling and shifting phases adjust the tuning sequences in hopes of improving the finite-time performance. In the scaling phase, the $\{a_n\}$ is scaled up by α, i.e. $\{a_n\}$ to $\{\alpha a_n\}$, so the iterates can move from one boundary to the other. In addition, $\{c_n\}$ is increased by scaling up by γ, i.e. $\{c_n\}$ to $\{\gamma c_n\}$, to minimize the noise of the gradient estimate if the iterates fall outside the truncation interval due to an incorrect gradient direction. In the shifting phase, the sequence $\{a_n\}$ is decreased by shifting or "skipping" a finite number (β) of terms, i.e. $\{a_n\}$ to $\{a_{n+\beta}\}$, when the iterates fall outside of the feasible region when the sign of the gradient is correct. In addition, c_n is scaled up by γ if the previous iterate is at the boundary and the update falls outside the feasible region but in the wrong direction. These adjustments do not affect the asymptotic convergence, since the scaling phase only scales the sequences by a constant and the shifting phase only scales up the $\{c_n\}$ finitely many times and skips a finite number of terms in $\{a_n\}$.

3 NUMERICAL EXPERIMENTS

We conduct two sets of numerical experiments. The first is to investigate the tightness of the finite-time MSE bound derived in Broadie, Cicek, and Zeevi (2011), and the second is to compare the MSE performance between KW and two of its variants described in Section 2.2, Kesten's rule and SSKW. All experiments were implemented with $a_n = \theta_a/n,\ c_n = \theta_c/n^s$ where $s \in \{1/4, 1/2\},\ \theta_a > 0,\ \theta_c > 0$, 10,000 iterations, and 1,000 sample paths.

3.1 Tightness of the Finite-time MSE Bound for Quadratics

We generated the MSE bound in (3) and the exact MSE in (4) for quadratic functions with various noise levels and initial starting values for three different cases: 1) $f(x) = -.001x^2,\ c_n = 1/n^{1/2},\ f(x) = -.15x^2,\ c_n = 1/n^{1/4}$ and $f(x) = -.15x^2,\ c_n = 1/n^{1/2}$. The MSE bound is a function of constants that are not unique, satisfying S1, S3, A1, and A2. We picked the largest K_0 and smallest K_1 satisfying A1 and A slightly less than $2K_0$. Table 3.1 lists the constants used in our calculations for the MSE bound in (3), and the exact MSE and MSE bound are listed in Table 3.1. The exact MSE (4) is a sum of three components. The first term on the right hand side (RHS) of (4) is independent of σ and is dominated by the initial starting value,

X_1. The second and third terms in (4) are dominated by σ. When $\sigma \in \{0.001, 0.01, 0.1, 1.0\}$, both terms are small ($< 1$), but when $\sigma = 10$, the RHS is dominated by the second term. Therefore, the exact MSE increases with X_1 and σ. Using the parameters in Table 3.1, the constant C in the MSE bound can be expressed as

$$C = \max\left\{\frac{\sigma^2}{\xi}, \frac{c_1^2}{a_1}\left(X_1(1-2a_1K_0+K_1^2a_1^2)(1-2a_2K_0+K_1^2a_2^2)+\frac{a_2^2}{c_2^2}\sigma^2\right)\right\}. \quad (5)$$

The first term in (5) dominates when σ is large since $\xi = 0.001, 0.1$ for $f(x) = -.001x^2, -.15x^2$, respectively. Therefore, the MSE bound and difference between the exact MSE and MSE bound increases significantly when σ increases from 1.0 to 10.0. Otherwise, the MSE bound is equal to the second term, which increases with X_1 and σ. Table 3.1 contains the exact MSE and MSE bound for three different parameter settings and for $\sigma \in \{0.001, 0.01, 0.1, 1.0, 10.0\}$. The first column in Table 3.1 lists/presents results for $f(x) = -.001x^2,\ c_n = 1/n^{1/2}$. In the presence of more noise, i.e. $\sigma = 10.0$, the MSE bound is 99,999.20, which is the first term in (5) for each initial starting value. The difference between this bound and the exact MSE is significant with a difference greater than 97,500. For $\sigma = 1.0$, the MSE bound only takes the second term in (5), when the starting position is farther from the optimum, i.e. $X_1 = -40$ and is tight. However, when the initial starting value is closer to the optimum, i.e. $X_1 = 0, -5, -10, -20$, the MSE bound is equal to 999.99, which is the first term in (5), and thus the MSE bound is significantly greater than the exact MSE. The MSE bound is very tight for rest of the cases with the exception of when $\sigma = 0.1$ and $X_1 = 0$.

		$f(x) = -.001x^2$ $c_n = 1/n^{1/2}$		$f(x) = -.15x^2$, $c_n = 1/n^{1/4}$		$f(x) = -.15x^2$ $c_n = 1/n^{1/2}$	
σ	X_1	Exact	Bound	Exact	Bound	Exact	Bound
0.001	0	0.00	0.00	0.00	0.00	0.00	0.00
	-5	24.04	24.85	0.06	8.85	0.06	0.09
	-10	96.16	99.40	0.24	35.40	0.24	0.35
	-20	384.64	397.61	0.95	141.61	0.95	1.42
	-40	1538.56	1590.42	3.78	566.44	3.78	5.66
0.01	0	0.00	0.10	0.00	0.00	0.00	0.00
	-5	24.04	24.85	0.06	8.85	0.06	0.09
	-10	96.16	99.40	0.24	35.40	0.24	0.35
	-20	384.64	397.61	0.95	141.61	0.95	1.42
	-40	1538.56	1590.42	3.78	566.44	3.78	5.66
0.1	0	0.05	10.0	0.01	0.10	0.00	0.00
	-5	24.09	24.86	0.07	8.86	0.06	0.09
	-10	96.21	99.41	0.24	35.41	0.24	0.35
	-20	384.69	397.61	0.95	141.62	0.95	1.42
	-40	1538.61	1590.43	3.79	566.45	3.78	5.66
1.0	0	4.8	999.99	0.83	10.00	0.03	0.10
	-5	28.84	999.99	0.89	10.00	0.09	0.10
	-10	100.96	999.99	1.07	35.90	0.27	0.36
	-20	389.44	999.99	1.78	142.11	0.98	1.42
	-40	1543.36	1590.92	4.61	566.94	3.81	5.67
10.0	0	480.08	99999.20	83.23	999.79	3.20	9.99
	-5	504.12	99999.20	83.29	999.79	3.26	9.99
	-10	576.24	99999.20	83.47	999.79	3.43	9.99
	-20	864.72	99999.20	84.18	999.79	4.14	9.99
	-40	2018.64	99999.20	87.01	999.79	6.98	9.99

Table 1: Finite-time MSE bound and exact MSE for KW with $n = 10,000,\ a_n = 1/n$.

$f(x)$	a_n	c_n	K_0	K_1	A	n_0	ξ
$-.001x^2$	$1/n$	$1/n^{1/2}$	0.002	0002	0.003	1	0.001
$-.15x^2$	$1/n$	$1/n^{1/2}$	0.3	0.3	0.5	1	0.1
$-.15x^2$	$1/n$	$1/n^{1/4}$	0.3	0.3	0.5	1	0.1

Table 2: Finite-time MSE Bound Parameters for KW

For the second column with $f(x) = -.15x^2$, $c_n = 1/n^{1/2}$, the MSE bound is significantly greater than the exact MSE across the board. The third column reports results for $f(x) = -.15x^2, c_n = 1/n^{1/2}$ the MSE bound is tight for all cases with the exception of the case with $\sigma = 10.0$. It would seem that the bound is a useful guideline for problems with low variance, but becomes less tight as the noise level increases.

3.2 Numerical Experiment: Comparison of KW and its Variants

Not surprisingly, the performance of SSKW relative to KW heavily depends on the chosen parameters such as truncated interval length, initial starting value, and tuning sequences. Our analysis replicates the results of Broadie, Cicek, and Zeevi (2011), where SSKW performs significantly better than KW in terms of MSE and oscillatory period, but we find that the chosen parameters for this experiment are among the worst possible parameters for KW as illustrated in Figure 1 with KW and SSKW under $\theta_a = \theta_c = 1$. By choosing a different initial starting position, the performance of KW can be significantly improved, as demonstrated in Table 3 for two functions $f(x) = -0.001x^2$ and $f(x) = 100e^{-.006x^2}$. To offer a contrast with the quadratic function, the second function considered is very steep and has flat tails.

Broadie, Cicek, and Zeevi (2011) compared the SSKW performance with that of KW whose MSE is highly reliant on the tuning sequences and initial start value.

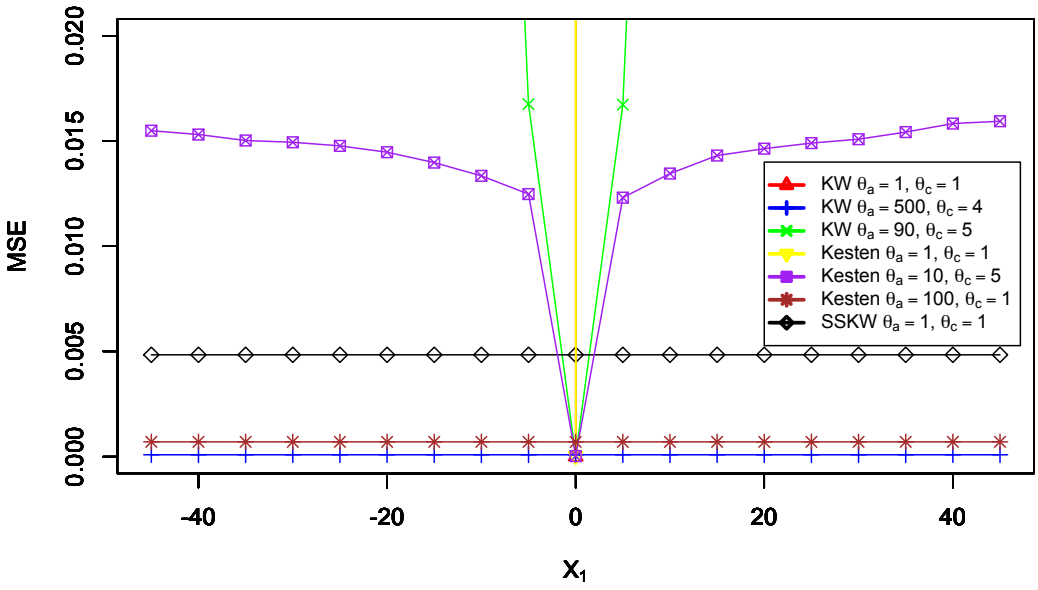

Figure 1: MSE of the 10,000th iterate of KW and Kesten for three parameter settings and SSKW for $f(x) = -.001x^2$, $\sigma = 0.001$, $a_n = \theta_a/n$, $c_n = \theta_c/n^{1/4}$.

The MSE performance results for $f(x) = -.001x^2$ using KW in Broadie, Cicek, and Zeevi (2011) were poor because the initial position was chosen to be far from the optimum and the gain size a_n was too small to make any noticeable progress towards it after 10,000 iterations, so the iterates hover around the initial position. In our numerical experiments, we also consider $a_n = \frac{\theta_a}{n}$ and $c_n = \frac{\theta_c}{n^{1/4}}$ for $\theta_a, \theta_c > 0$. If $\theta_a = \theta_c = 1$ as in Broadie, Cicek, and Zeevi (2011), but the initial start value is 0.01 instead of 30, then the MSE from KW is significantly lower compared to SSKW. The first column in Table 3 compares the MSE all three algorithms with $X_1 = 0.01$, and clearly, KW outperforms SSKW in almost all cases. Of course, a practitioner would have no way of knowing whether or not the starting iterate was close to the true optimum, so these results do not indicate that KW will always perform well. They do indicate, however, that KW exhibits substantial variation in performance.

We also conduct a sensitivity analysis for $f(x) = -.001x^2$ with various starting positions X_1 and multiplicative constants, θ_a, and θ_c and implement SSKW and KW using Kesten's rule. For the sensitivity analysis, we considered a wide selection of parameters: 19 initial starting values uniformly spaced within the truncated interval $X_1 \in \{-50+5k \mid k = 1,2,...19\}$, 45 different θ_a values parametrized by $\theta_a \in \{10^s k \mid k = 1,2,\ldots,9, s = 0,1,\ldots,4\}$ and 10 different θ_c values parametrized by $\theta_c \in \{10^s k \mid k = 1,2,...,5, s = 0,1\}$. In total, there are 8550 possible combinations of parameters.

The results show that KW and Kesten's rule are sensitive to the parameter choice, but near-optimal performance can be obtained with tuning. Figure 1 plots the MSE of KW for $f(x) = -.001x^2, \sigma = 0.001$ against the initial starting values X_1 for different sets of parameter choices. These cases serve as a good representation of the majority of the MSE behaviors among the entire set of results. The case with $\theta_a = \theta_c = 1$ is among the worst for KW and Kesten's rule. The MSE is represented by a nearly vertical line for both algorithms. For this parameter setting, SSKW beats KW and Kesten's rule significantly for all initial values with the exception of $X_1 = 0$. For the case where $\theta_a = 90, \theta_c = 5$, KW outperforms SSKW in a neighborhood around the optimum. However, there are cases such as $\theta_a = 500,\ \theta_c = 4$ for KW and $\theta_a = 100,\ \theta_c = 1$ for Kesten's rule that outperform SSKW for all initial start values. Of the 8550 combinations varying all parameters and 450 combinations with $X_1 = 30$, KW performs better than SSKW in 4275 and 215 cases, respectively, suggesting that KW requires some tuning to perform well, but that there is a fairly wide range of tunable parameters that yield good performance. If KW performs better than KW, the difference is not as pronounced as when SSKW outperforms KW, but careful tuning can partially mitigate KW's sensitivity to parameters such as the initial iterate.

		$f(x) = -0.001x^2$ [-50, 50] $X_1 = .01$			$f(x) = 100e^{-0.006x^2}$ [-50, 50] $X_1 = 30$		
σ	Algorithm	100	1000	10000	100	1000	10000
0.001	SSKW	$5.10x10^{-2}$	$1.70x10^{-2}$	$5.00x10^{-3}$	$5.07x10^{-2}$	$1.68x10^{-2}$	$4.84x10^{-3}$
	KW	10^{-4}	10^{-4}	10^{-4}	763.8	653.3	431.4
	Kesten	$1.12x10^{-4}$	$1.08x10^{-4}$	$1.04x10^{-4}$	10^{-7}	$3x10^{-8}$	10^{-8}
0.01	SSKW	$5.10x10^{-2}$	$1.70x10^{-2}$	$5.00x10^{-3}$	5.07	1.68	$4.90x10^{-1}$
	KW	10^{-4}	10^{-4}	10^{-4}	763.8	653.3	431.2
	Kesten	$2.10x10^{-3}$	$2.11x10^{-3}$	$2.05x10^{-3}$	$9.54x10^{-6}$	$2.76x10^{-6}$	$8.41x10^{-7}$
0.1	SSKW	$5.10x10^{-2}$	$1.70x10^{-2}$	$5.00x10^{-3}$	165.8	57.4	16.0
	KW	10^{-4}	10^{-4}	10^{-4}	763.4	651.4	418.2
	Kesten	$2.01x10^{-1}$	$2.03x10^{-1}$	$1.97x10^{-1}$	$5.65x10^{-2}$	$2.76x10^{-4}$	$8.41x10^{-5}$
1.0	SSKW	$5.10x10^{-2}$	$1.70x10^{-2}$	$5.00x10^{-3}$	187.2	57.8	18.7
	KW	10^{-4}	10^{-4}	10^{-4}	722.5	562.5	415.7
	Kesten	20.1	20.3	19.7	456.9	315.1	239.7

Table 3: MSE of the 100th, 1,000th, and 10,000th iteration for KW and its variates with $a_n = 1/n$, $c_n = 1/n^{1/4}$.

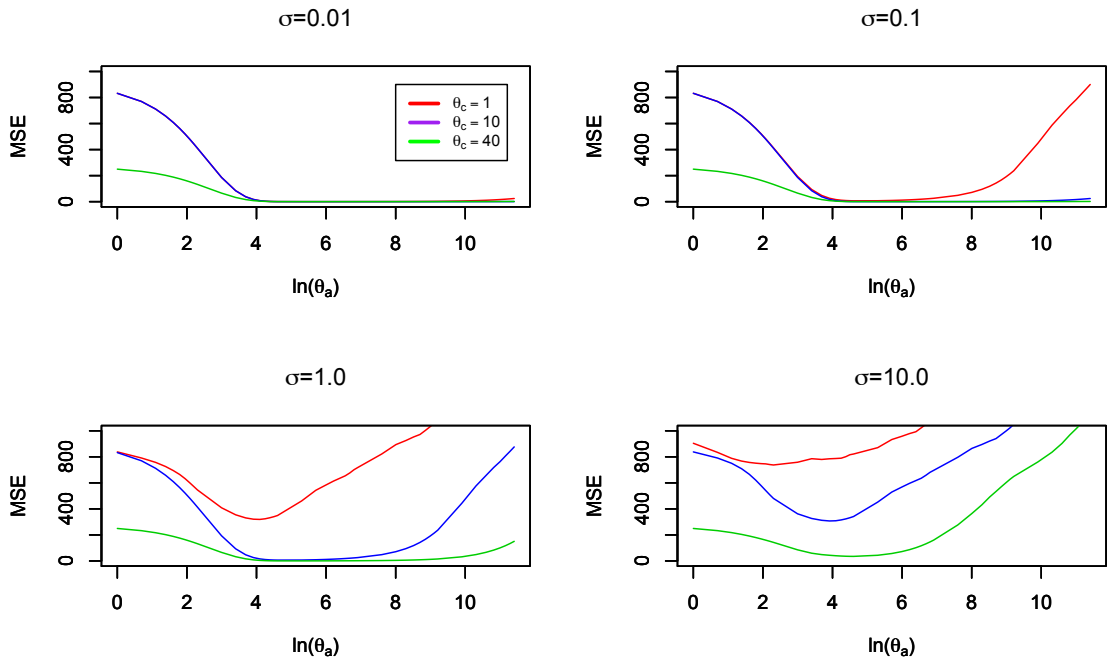

Figure 2: Sensitivity of KW to θ_a for $f(x) = -.001x^2$ $a_n = \theta_a/n$, $c_n = \theta_c/n^{1/4}$, $n = 10{,}000$.

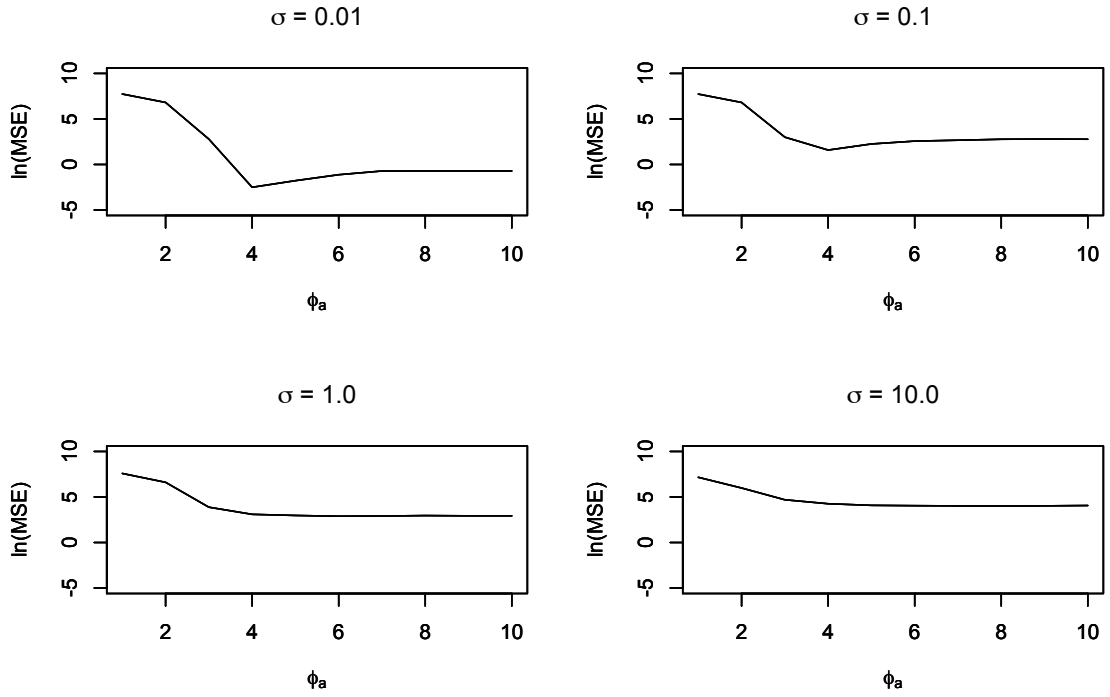

Figure 3: Sensitivity of SSKW for $f(x) = -.001x^2$ as a function of ϕ_a, $n = 10{,}000$.

Figure 2 plots the MSE $f(x) = -.001x^2, a_n = \theta_a/n, c_n = \theta_c/n^{1/4}$of the 10,000th iterate as a function of $\log\theta_a$ given $\theta_c = 1, 10, 40$. The case where $\sigma = 0.001$ is omitted, because the results are similar to those for $\sigma = 0.01$. For $\log\theta_a < 4$, the MSE decreases for each given value of θ_c. However, for $\log\theta_a \geq 4$, the MSE behaves differently for all noise levels. But, the overall behavior as a function of θ_c is similar across noise levels. The MSE decreases for all θ_a as θ_c increases, so in the case where $\theta_c = 40$, there is a wide range of θ_a values where the MSE of KW is lower than that of SSKW. But the MSE of KW could also be extremely high if the tuning sequences are not chosen well. Moreover, we investigate the sensitivity of SSKW to ϕ_a, which is the upper bound of the scale up factor for $\{a_n\}$ as depicted in Figure 3. The MSE decreases until ϕ_a, the maximum scale up factor for $\{a_n\}$, is equal to 4 and increases for $\sigma = 0.01, 0.1$ while it levels off for $\sigma \in \{1.0, 10.0\}$ thereafter. It seems that for lower noise levels,i.e. $\sigma \in \{0.01, 0.1\}$, $\phi_a = 4$ is a better choice, while $\phi_a = 10$ leads to a lower MSE for $\sigma \in \{1.0, 10.0\}$.

In addition, we implement KW and its variants using the same parameters $(a_n = 1/n, c_n = 1/n^{1/4}, X_1 = 30)$ as in Broadie, Cicek, and Zeevi (2011) on $f(x) = 100e^{-.006x^2}$ to test the algorithms under the same setting for a different function. Figure 4 plots the MSE of the 10,000th iterate as a function of the initial start value. KW and Kesten's rule outperform SSKW within certain intervals around the optimum for $\sigma \in \{0.001, 0.01, 0.1, 1.0\}$ and Kesten's better performance intervals overlap the intervals of KW. However, the KW using the deterministic step-size $1/n$ performs better than using Kesten's step-size where the intervals overlap, which can be seen in Figure 4. Unfortunately, outside of those intervals, both algorithms have a tendency to perform poorly. However, for the other four noise levels, the intervals where KW and Kesten's rule outperform SSKW are larger. However, there is a tradeoff, since if by chance the initial start value is closer to the boundary, the difference in performance can be drastic.

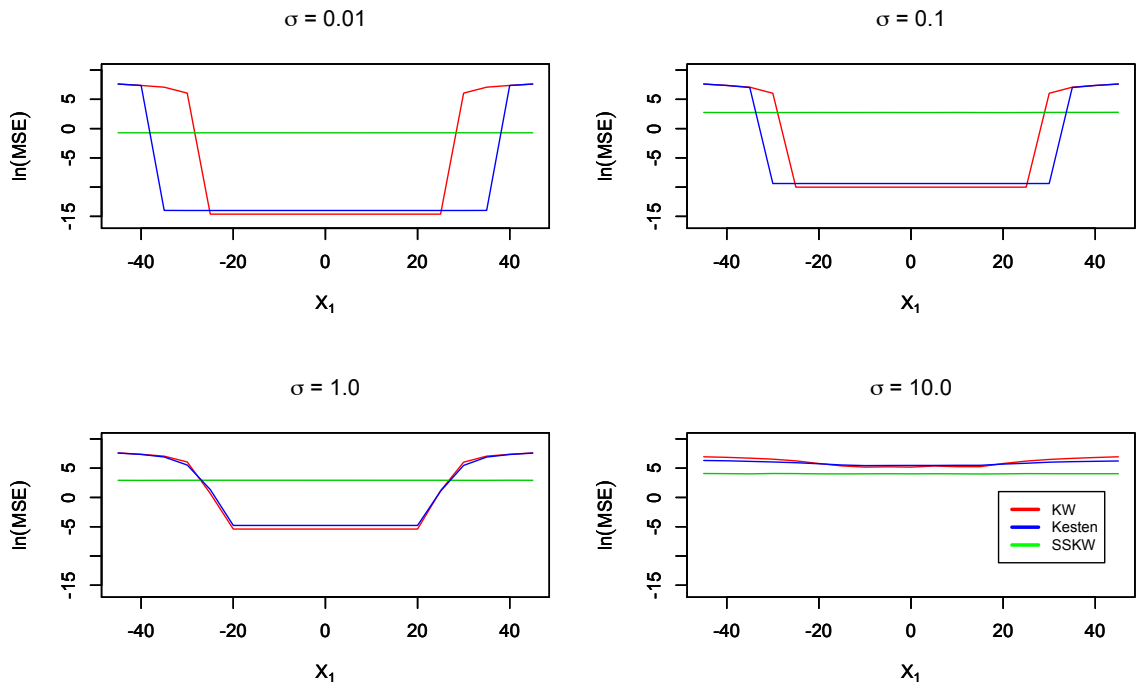

Figure 4: MSE Comparison of KW, Kesten, and SSKW for $f(x) = 100e^{-.006x^2}$, $a_n = 1/n$, $c_n = 1/n^{1/4}$, $n = 10,000$.

4 CONCLUSION

Our objective was to further investigate the MSE performance of KW and its variants and to test the quality of a finite-time MSE bound for the KW algorithm. From our numerical experiments, SSKW is insensitive to the initial start value; however, finite-time performance could be further improved by implementing KW or Kesten with well-tuned parameters, which allows both algorithms to be less sensitive to the initial start value. An advantage of KW or Kesten's rule is the ability to fine-tune only two parameters to achieve a lower MSE compared to a total of eight parameters for SSKW; however, the tradeoff is the potential to perform extremely poorly if the wrong parameters are chosen. SSKW is more conservative; its MSE is higher than that of KW or Kesten under good parameter choices but not nearly as high as the MSE of KW or Kesten when they perform poorly. The performance of the algorithms is heavily dependent on the chosen parameters as well as the geometry of the functions. Furthermore, our numerical results regarding the finite-time MSE bound indicate the bound is tight for functions with less noise. In practice, the success or failure of a KW-like algorithm hinge on problem-dependent factors such as the geometry of the function and the level of simulation noise. These issues prove difficult to overcome even for adaptive stepsizes. We hope that this study will shed light on some of these practical factors and help guide practitioners in their choice of SA algorithm.

REFERENCES

Andradóttir, S. 1995. "A Stochastic Approximation Algorithm with Varying Bounds". *Operations Research* 43 (6): 1037–1048.

Broadie, M., D. Cicek, and A. Zeevi. 2011. "General Bounds and Finite-Time Improvement for the Kiefer-Wolfowitz Stochastic Approximation Algorithm". *Operations Research* 59 (5): 1211–1224.

Derman, C. 1956. "An Application of Chung's Lemma to the Kiefer-Wolfowitz Stochastic Approximation Procedure". *The Annals of Mathematical Statistics* 27 (2): 532–536.

Dupac, V. 1957. "On the Kiefer-Wolfowitz approximation method". *Časopis pro péstovánť Matematiky* 082 (1): 47–75.

Fabian, V. 1967. "Stochastic Approximation of Minima with Improved Asymptotic Speed". *The Annals of Mathematical Statistics* 38 (1): 191–200.

George, A. P., and W. B. Powell. 2006, October. "Adaptive stepsizes for recursive estimation with applications in approximate dynamic programming". *Machine Learning* 65 (1): 167–198.

Kesten, H. 1958. "Accelerated Stochastic Approximation". *The Annals of Mathematical Statistics* 29 (1): 41–59.

Kiefer, K., and J. Wolfowitz. 1952. "Stochastic estimation of the maximum of a regression function". *The Annals of Mathematical Statististics* 23 (3): 462–466.

Polyak, B., and A. Juditsky. 1992. "Acceleration of Stochastic Approximation by Averaging". *SIAM Journal on Control and Optimization* 30 (4): 838–855.

Robbins, H., and S. Monro. 1951. "A Stochastic Approximation Method". *The Annals of Mathematical Statistics* 22:400–407.

Spall, J. 1992. "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation". *IEEE Transactions on Automatic Control* 37 (3): 332–341.

Tsybakov, A., and B. Polyak. 1990. "Optimal order of accuracy of search algorithms in stochastic optimization". *Problemy Peredachi Informatsii* 26 (2): 45–63.

AUTHOR BIOGRAPHIES

MARIE CHAU is a Ph.D. candidate in Applied Mathematics, Statistics, and Scientific Computation at the University of Maryland. She has a M.S. in applied math and degrees in math, finance and economics from UMCP. Her research interests lie in stochastic optimization and financial engineering. Her email address

is mchau@math.umd.edu.

HUASHUAI QU is a Ph.D. candidate in Applied Mathematics, Statistics, and Scientific Computation at the University of Maryland, where he received a Gold Medal in Teaching Award from the Mathematics Department in 2010. His research interests lie in the broad areas of optimal learning and simulation optimization. He received the 2012 INFORMS Computing Society Student Paper Award and the Best Theoretical Paper Award at WSC2012. His email address is huashuai@math.umd.edu.

MICHAEL C. FU is Ralph J. Tyser Professor of Management Science in the Robert H. Smith School of Business, with a joint appointment in the Institute for Systems Research and affiliate faculty appointment in the Department of Electrical and Computer Engineering, all at the University of Maryland. His research interests include simulation optimization and applied probability, with applications in supply chain management and financial engineering. He has a Ph.D. in applied math from Harvard and degrees in math and EECS from MIT. He served as WSC2011 Program Chair, NSF Operations Research Program Director, *Management Science* Stochastic Models and Simulation Department Editor, and *Operations Research* Simulation Area Editor. He is a Fellow of **INFORMS** and **IEEE**. His email address is mfu@umd.edu.

ILYA O. RYZHOV is an Assistant Professor in the Robert H. Smith School of Business at the University of Maryland. He received a Ph.D. in Operations Research and Financial Engineering from Princeton University. His research deals with optimal learning and the broader area of stochastic optimization, with applications in disaster relief, energy, and revenue management. He was a recipient of WSC's Best Theoretical Paper Award in 2012. His work has appeared in *Operations Research*, and he is a co-author of the book *Optimal Learning*, published in 2012 by John Wiley & Sons. His email address is iryzhov@rhsmith.umd.edu.