Change-point Problems
IMS Lecture Notes – Monograph Series (Volume 23, 1994)

DESIGN OF CONTROL CHARTS FOR DETECTING THE CHANGE POINT

BY YANHONG WU
University of Toronto

The discrete time change-point detecting problem is considered. The main purpose is to review some accurate approximations for the operating characteristics (ARL_0 and ARL_1) for three well-known detecting procedures: CUSUM, EWMA, and Shiryayev-Roberts procedures, based on the boundary correction technique. These approximations are shown to be very accurate compared with simulation and numerical values. The results can be used for the design of these control charts.

1. Introduction. Suppose $X_1, \ldots, X_{\theta-1}, X_\theta, \ldots, X_n, \ldots$ are a sequence of independent random variables, where $X_1, \ldots, X_{\theta-1}$ are iid with the density function $f_0(x)$, $X_\theta, \ldots, X_n, \ldots$ are iid with the density function $f_1(x)$, and θ is the change-point and assumed to be unknown. The purpose is to find a detecting procedure in order to raise an alarm as soon as possible after the change occurs. The change-point problem has many applications in a variety of areas such as the surveillance of a system, monitoring the quality of production processes, and alarming for a flood etc..

For the convenience of discussion, we shall use the terminology from quality control. And for simplicity, we consider the normal case with f_0 following N(0,1), and f_1 $N(\mu, 1)$ with $\mu > 0$ unknown. Denote $E_\theta[.]$ as the expectation when the change is at θ. In particular, E_∞ and E_1 denote the probability and expectation calculated when the change-point is at infinity and at the beginning, respectively.

For a stopping rule τ as the alarming time associated with a detecting procedure, two mostly used operating characteristics are the average in-control run length(ARL_0) and the average out-of-control run length(ARL_1), defined by

$$ARL_0 = E_\infty[\tau], ARL_1(\mu) = E_1[\tau].$$

AMS 1991 Subject Classification: Primary 62N10.

Key words and phrases: Average run length, boundary correction, CUSUM, EWMA, Shiryayev-Roberts procedure.

ARL_0 and ARL_1 are used to be evaluated by simulation as drawn by nomograms. Recent approaches are mostly through numerical methods such as the Markov chain method when the detecting process is Markovian, see Brook and Evans (1971) for the CUSUM procedure and Lucas and Saccucci (1990) for the EWMA procedure. The goal of this paper is to provide another simple method based on the boundary correction technique as discussed in Siegmund (1985) for the CUSUM procedure. The basic idea is to correct the control limit by adding the average overshoot at the alarming time. This method has several advantages. First, it has a clear relationship with the result in the continuous time case which usually has a simple close form. Second, it gives very simple formula which is extremely useful for the design purpose. Third, it has quite satisfactory accuracy for practical use. Three detecting procedures, i.e. CUSUM, EWMA and Shiryayev-Roberts procedures will be discussed. In Section 2, we first give the formulas for the approximations for ARL's when the control limit is large. The emphasis is for ARL_0 since it is crucial for the design. Comparisons with numerical values are made to show their accuracies. The main contribution of this paper is to give a simple design method for the EWMA procedure. In Section 3, we give some general guidelines on the practical use of these three procedures. Section 4 gives the technical details for the proofs of the related results in two subsections. First, we check the accuracy of the approximations for the ARL's in the CUSUM procedure. Then, we give the boundary correction results for the EWMA procedure.

2. Approximation for ARL's. In this section, we give the approximations of ARL's for the three procedures and checking their accuracies by comparing with the numerical values. We begin with the CUSUM procedure.

2.1. CUSUM Procedure. When μ is unknown, we usually select a reference value δ to form a simple procedure. The CUSUM process for detecting the shift δ is defined by

$$Y_n = \max\left(0, Y_{n-1} + X_n - \frac{\delta}{2}\right) \tag{1}$$

where Y_0 is usually taken as 0. An alarm is made at

$$\tau = \min\{n \geq 1 : Y_n \geq d\}.$$

Siegmund (1985) gives the corrected diffusion approximations of ARL_i's for the CUSUM procedure under the exponential family. In the normal case,

$$ARL_0 \approx \frac{e^{\delta(d+2\rho)} - 1 - \delta(d+2\rho)}{\delta^2/2}, \tag{2}$$

as $\delta \to 0$, $d \to \infty$, and $\delta d \to$ *constant*, where $\rho = 0.583$ for relatively small δ.

Theorem 10.16 of Siegmund (1985) shows that the error of approximation (2) is on the order of $o(1/\delta)$. A careful reevaluation shows that the approximation is accurate in the order of $o(1)$ as $\delta \to 0$, see Section 4.1. The same conclusion is true for ARL_1 where the approximation is given by

$$ARL_1(\mu) \approx \frac{e^{-2(\mu-\delta/2)(d+2\rho)} - 1 + 2(\mu - \delta/2)(d + 2\rho)}{2(\mu - \delta/2)^2}, \tag{3}$$

for $\mu > \delta/2$. A more general discussion for the second order approximation for the ARL's of the CUSUM procedure can be seen in Pollak and Siegmund (1986).

One may note that if we ignore the correction factor 2ρ, (2) and (3) are the results for the continuous time CUSUM procedure (Taylor (1975)).

Van Dobben de Bruyn (1968) uses several numerical methods to evaluate ARL's for the CUSUM procedure. Table 1 gives the comparison of the approximations (2) and (3) with the numerical values for $\delta \leq 1.2$ and $d = 2, 2.5, 3, 4, 5, 6$ respectively. We find that the relative error is within 2%. For large δ, higher order expansions for the mean overshoot will be necessary for a more accurate approximation, see Section 4.1 for the first order expansion.

2.2. Shiryayev-Roberts Procedure. The discrete time Shiryayev-Roberts procedure was first given by Roberts (1966). The procedure was formally discussed by Pollak and Siegmund (1985) in the continuous time case and shown to have similar behaviors as the CUSUM procedure. The Shiryayev-Roberts process is defined as

$$R_n = (1 + R_{n-1})e^{\delta x_n - \delta^2/2}, \tag{4}$$

with $R_0 = 0$ for the chosen reference value δ. An alarm is made at

$$\tau = \inf\{n > 0 : R_n > T\}.$$

The approximation for ARL_0 is given by Pollak (1987) for the exponentially family, from which, a simple approximation is given by

$$ARL_0 \approx Te^{\rho\delta}, \tag{5}$$

where ρ is given as before.

The accuracy of approximation (5) is even better than that for the CUSUM procedure. Table 2 gives the comparison with the simulated results for $T = 100, 300, 500$ respectively. The simulation is replicated 10,000 times. From the table, we see that for $\delta \leq 2.0$, the approximation is very satisfactory.

For $ARL_1(\mu)$, the approximation appears slightly complicated which is given by

$$ARL_1(\mu) = \frac{1}{\delta(\mu - \frac{\delta}{2})}(\ln T + E\ln\left(1 + \sum_{n=1}^{\infty} e^{-\delta S_n - n\delta(\mu-\delta/2)}\right) + o(1)),$$

see Pollak (1987).

2.3. EWMA Procedure. It is not difficult to derive the CUSUM and Shiryayev-Roberts procedures based on the likelihood principle and Bayesian approach, respectively. A more intuitive idea is to smooth the previous data in order to reduce the effect of noises in the sampling procedure. Two common methods are the moving average and the exponential smoothing (Roberts (1966)). The EWMA procedure based on the exponential smoothing has several quite interesting features. First, it gives the estimation of current mean, and thus can be used as a detecting process. Second, it is the optimal predicted value for the IMA(0,1,1) process which has been used for modeling the quality characteristic process for gradually increasing variation (Box and Jenkins (1963)). Third, it is Markovian and thus one can evaluated its operating characteristics quite easily (Lucas and Saccucci (1990)).

Define the exponential smoothing of $X_1, \ldots, X_n$ by

$$Z_n = (1 - \beta)Z_{n-1} + \beta X_n,$$

with $Z_0 = 0$ and $0 < \beta \leq 1$. The limiting variance of Z_n can be easily found as

$$\lim \mathrm{Var}_\infty(Z_n) = \beta/(2 - \beta) = \sigma_\infty^2.$$

To detect a positive shift of mean, the usual EWMA procedure is defined to make an alarm at

$$\tau = \inf\{n \geq 1 : Z_n \geq B = b\sigma_\infty\}.$$

From the definition of Z_n, we see that after the change, the mean of Z_n exponentially increases to the true mean δ. The design of EWMA procedure is rather complicated. In order to be able to detect the shift efficiently, $\sigma_\infty b$ should be taken less that δ, and as small as possible in order to have small average delay time. However, b and β have to be chosen to satisfy the condition for ARL_0. Thus, there is an optimal design problem of choosing β and b which minimizes the average delay time for given ARL_0.

In the continuous time case, Srivastava and Wu (1993) have considered this optimal design problem in terms of the stationary average delay time (Shiryayev (1963)). The main result shows that as $ARL_0 = T \to \infty$, in the

first order, the optimal smoothing parameter β^* and control limit b* satisfies the following two equations:

$$\beta = 2c^*\delta^2/b^{*2}, \tag{6}$$

$$\frac{1}{\beta*}\int_0^{b*}[\phi(x)]^{-1}\Phi(x)dx = T = ARL_0, \tag{7}$$

where $c^* = 0.5117$. An approximation for β^* can be obtained as

$$\beta^* \approx 0.5117\delta^2/\ln[0.4083\delta^2 T(2\ln(0.4083\delta^2 T))^{1/2}] \tag{8}$$

where obviously, the right hand side is required to be positive. Furthermore, it follows from (6) that as $T \to \infty$,

$$B = b\sigma_\infty \approx b(\beta/2)^{1/2} \approx 0.715\delta.$$

This implies that the optimal control limit for Z_n is approximately set at 0.715δ.

Under this optimal design, for $\mu > 0.715\delta$, we have

$$ARL_1(\mu) = \frac{1}{\delta^2}\left[\frac{-\ln(1-0.715\delta/\mu)}{2\times 0.715^2}b*^2 - \frac{\delta^2}{4\mu^2}\left(\frac{1}{(1-0.715\delta/\mu)^2}-1\right)+o\left(\frac{1}{b*^2}\right)\right],$$

see Section 4.2 for the exact results in the discrete time case. One can see that comparing to the CUSUM and Shiryayev-Roberts procedures, the approximation for the EWMA procedure is less accurate as the error is on the order of $1/\ln T$ rather than $\ln T/T$ for the other two. Also when $\mu = \delta$, it is not difficult to check that the EWMA procedure is not as efficient as the other two procedures as the same $ARL_0 \to \infty$ (Srivastava and Wu (1993)).

The approximation (7) is too crude to be acceptable in the discrete time case. In Section 4.2, a more accurate approximation is obtained by adding the mean overshoot $ER = E(Z_\tau - b\sigma_\infty)$ to $b\sigma_\infty$, which is roughly estimated as

$$ER \approx \beta^*\rho, \tag{9}$$

where $\rho \approx 0.583$ as before. By this correction, ARL_0 is about

$$ARL_0 \approx e^{0.834\delta}T,$$

in the first order, where T is given by (7), see Section 4.2.

In order to detect a shift value δ with $ARL_0 = T$, the design for an EWMA procedure can be done in the following way. First, select a β based on

(8) with T replaced by $Te^{-0.834\delta}$. Then, let $B = 0.715\delta - 0.583\beta$. Although this does not give the ideal optimal design, it is quite good enough for practical use.

An interesting method based on the Edgeworth expansion of the crossing probability is given by Robins and Ho (1978) by calculating the first four moments recursively. Table 3 gives some comparisons between the corrected diffusion approximation, the numerical values given in Robins and Ho (1978) , and the lower bound given by (16) of Section 4.2 which is the approximation by ignoring the overshoot. We see that the corrected boundary approximation gives much improved values than the lower bound. Ironically, the corrected diffusion approximation achieves its best accuracies around the region of $\{0.1 \leq \beta \leq 0.25\}$, which, according to Montgomery (1991), is the most desirable region for the value of β in practice.

Table 1: Comparison of ARL_0 and ARL_1 for CUSUM

d	δ	ARL_0		ARL_1	
		Num	Approx	Num	Approx
2.0	0.0	10.0	10.2	10.0	10.2
	0.4	15.9	16.02	6.86	6.85
	0.8	28.0	28.30	5.06	5.04
	1.2	54	55.37	3.96	3.92
2.5	0.0	13.4	13.44	13.4	13.44
	0.4	23.3	23.34	8.73	8.71
	0.8	46.1	46.40	6.24	6.21
	1.2	104	105.54	4.79	4.74
3.0	0.0	17.3	17.36	17.3	17.36
	0.4	32.8	32.83	10.7	10.69
	0.8	73.6	74.01	7.44	7.40
	1.2	195	197.63	5.62	5.56
4.0	0.0	26.6	26.69	26.6	26.69
	0.4	60.3	60.37	14.9	14.91
	0.8	178	178.81	9.88	9.84
	1.2	660	673.81	7.28	7.22
5.0	0.0	38.1	38.02	38.1	38.02
	0.4	104	103.92	19.4	19.39
	0.8	414	415.11	12.4	12.31
6.0	0.0	51.6	51.35	51.6	51.35
	0.4	171	171.34	24.0	24.04
	0.8	940	944.06	14.9	14.80

*Numerical values are taken from Van Dobben de Bruyn (1968)

Table 2: Comparison of ARL_0 for the Shiryayev-Roberts Procedure

δ	$T = 100$		$T = 300$		$T = 500$	
	Approx	Sim	Approx	Sim	Approx	Sim
0.1	106.00	106.58 (0.49)	318.01	316.31 (1.99)	530.02	532.48 (3.74)
0.2	112.37	113.43 (0.77)	337.10	333.94 (2.74)	561.84	562.86 (4.87)
0.5	133.84	136.12 (1.23)	401.53	400.63 (3.83)	669.84	682.72 (6.46)
1.0	179.14	181.18 (1.75)	537.42	532.72 (5.12)	895.70	905.27 (8.97)
1.5	239.77	238.14 (2.39)	719.30	724.18 (7.16)	1198.84	1194.40 (12.0)
2.0	320.91	314.08 (3.17)	962.74	950.90 (9.46)	1604.57	1559.69 (15.62)

Table 3: Comparison of Approximations of ARL_0 for EWMA procedure

b/β		0.05	0.10	0.25
2.0	LB	203.31	98.98	36.25
	Num.	244.99	149.15	80.51
	Approx	268.77	142.85	74.28
2.25	LB	319.86	155.72	57.03
	Num.	391.62	244.62	137.32
	Approx.	432.19	230.10	125.86
2.50	LB	526.66	256.40	93.90
	Num.	660.84	424.14	248.42
	Approx	730.20	450.25	226.21
2.75	LB	915.99	445.94	163.32
	Num.	1183.63	778.81	477.26
	Approx	1307.16	833.96	433.75
3.0	LB	1694.79	825.09	302.18
	Num.	2242.04	1504.96	974.68
	Approx	2493.02	1651.16	890.44
3.5	LB	7103.38	3458.18	1266.52
	Num.	9136.09	6582.11	4936.30
	Approx	11090.78	6017.31	4627.41
4.0	LB	39349.64	19156.82	7015.98
	Num.	45821.86	37821.50	33197.59
	Approx	64773.33	34875.90	31729.11

*LB: Lower bound; Num: Robins and Ho (1978); Approx: corrected boundary approx.

3. General Discussion. (1) In the above section, we emphasized the approximations for ARL_0 since it is critical for the design of these control charts. The traditional approximation by ignoring the overshoot significantly underestimates the true value as we can see from Tables 1–3. The approximation of ARL_1 is also important if we want to compare these three procedures and to consider the economic design. The comparisons among the three procedures have been done by many authors by a variety of methods. Roberts (1966) compared several charts by simulation. The currently most used method in quality control literature is the Markov chain method, see Lucas and Saccucci (1990) for the comparison between the CUSUM procedure and the EWMA procedure. This method is obviously very space-consuming. The theoretical comparisons have been done by Pollak and Siegmund (1985) and Srivastava and Wu (1993) under the continuous time model. A more recent study by Pollak and Siegmund (1991) also considered the case when the initial level is unknown.

(2) Only comparing ARL_1 may be misleading as it considers only the case when the change occurs at the beginning. A typical example is the FIR (fast initial response) technique, and also see Srivastava and Wu (1993) for an example in the one-sided EWMA procedure. Thus, more reasonable measures for the average delay time should be chosen. Three interesting ones are the conditional stationary average delay time, the unconditional stationary average delay time and the maximum conditional delay time (Pollak and Siegmund (1985), Shiryayev (1963), and Lorden (1971)). The CUSUM procedure is optimal in the worst case and the Shiryayev-Roberts procedure is optimal in the stationary case. The asymptotic behaviors of the conditional and the unconditional stationary delay time are almost same as ARL_0 is large. The comparisons among these three procedures can be seen in the literature mentioned above.

(4) In this note, we only considered the one-sided shift case. The two-sided shift case can be similarly discussed (Siegmund (1985), Pollak and Siegmund (1991), Lucas and Saccucci (1990)). A theoretical treatment for the two-sided as well as the multivariate EWMA procedures will be discussed in another communication.

4. Technical Results. In this section, we give some technical details for the results given in Section 2. The readers are assumed to be familiar with Wald's likelihood identity and Stone's strong renewal theorem. We refer to Siegmund (1985) for a more detailed discussion. We have two objectives. One is to show that the approximations given in (2) and (3) are accurate in the second order of δ under appropriate conditions. The other is to give a heuristic argument for the approximation of ARL_0 for the EWMA procedures by using

the boundary correction technique.

4.1. Error Checking for the Approximations (2) and (3). We only give the details for ARL_0. For ARL_1 as well as the stationary average delay time, Pollak and Siegmund (1986) have given a more general discussion in the exponential family.

Define

$$N = \inf\{n > 0 : S_n < 0 \text{ or } > d\},$$

and

$$\tau_d = \inf\{n > 0 : S_n > d\}, \text{ and } \tau_{-(+)} = \inf\{n > 0 : S_n < (>)0\}.$$

Then

$$ARL_0 = \frac{E_0 N}{P_0(S_N > d)} = \frac{E_0 S_N}{-\delta/2 P_0(S_N > d)}.$$

The key to guarantee the accuracy of the following approximations is the strong renewal theorem which states that as $d \to \infty$, uniformly for $\delta > 0$ and $y \geq 0$,

$$|P_1(S_{\tau_d} - d > y) - P_1(R > y)| = o(e^{-rd}),$$

for a positive constant r, where $P_1(R \in dy) = P_1(S_{\tau_+} > y)/E_1 S_{\tau_+} dy$, see Siegmund (1979) and some refined result by Lotov (1991).

LEMMA 1.

$$P_0(S_N > d) = \frac{(E_1 e^{-\delta R})^{-1} P_1(\tau_- = \infty)}{e^{\delta d}(E_1 e^{-\delta R})^{-2} - 1}(1 + o(e^{-rd})),$$

$$P_1(S_N > d) = P_1(\tau_- = \infty)/(1 - e^{-\delta d}(E_1 e^{-\delta R})^2)(1 + o(e^{-rd})).$$

The proof can be obtained similar to Lemma 4.

LEMMA 2.

$$P_1(\tau_- = \infty)/E_0(S_{\tau_-}) = -\delta E_1 e^{-\delta R}.$$

LEMMA 3.

$$E_0 S_N = E_0(S_{\tau_-}) + E_1 R P_0(S_N > d)(1 + o(e^{-rd})) + E_0(S_N; S_N > d).$$

PROOF. Note that

$$E_0 S_N = E_0(S_N; S_N < 0) + E_0(S_N; S_N > d).$$

On the other hand,

$$\begin{aligned} E_0(S_{\tau_-}) &= E_0(S_{\tau_-}; S_N < 0) + E_0(S_{\tau_-}; S_N > d) \\ &= E_0(S_N; S_N < 0) + E_0[E_0[S_{\tau_-}|S_N]; S_N > d] \\ &= E_0(S_N; S_N < 0) + E_1 R P_0(S_N > d)(1 + o(e^{-rd})). \end{aligned}$$

Based on Lemmas 1–3, we get

$$\begin{aligned} ARL_0 &= \frac{1}{-\delta/2}\Big[\frac{(e^{\delta d}(E_1(e^{-\delta R}))^{-2} - 1)ES_{\tau_-}}{(E_1 e^{-\delta R})^{-1} P_1(\tau_- = \infty)} \\ &\qquad + E_1 R + E_0(S_N|S_N > d)\Big](1 + o(e^{-rd})) \\ &= \frac{1}{\delta/2}\Big(\frac{1}{\delta}(e^{\delta d}(E_1 e^{-\delta R})^{-2} - 1) \\ &\qquad - (E_1 R + E_0(S_N|S_N > d))\Big)(1 + o(e^{-rd})\Big). \end{aligned}$$

The only remaining thing is to evaluate $E_0(S_N|S_N > d)$. This is given by

LEMMA 4.

$$E_0(S_N|S_N > d) = d + \frac{E_1 R e^{-\delta R}}{E_1 e^{-\delta R}}(1 + o(e^{-rd})).$$

PROOF. By noting that

$$E_0(S_{\tau_d} - d; \tau_d < \infty) = E_0(S_N - d; S_N > d) + E_0(S_{\tau_d} - d; \tau_d < \infty, S_N < 0),$$

we have

$$\begin{aligned} E_0(S_N - d; S_N > d) &= E_1((S_{\tau_d} - d)e^{-\delta S_{\tau_d}}) \\ &\quad - E_0[E_1(S_{\tau_d} - d)e^{-\delta S_{\tau_d}}|S_N]; S_N < 0] \\ &= (e^{-\delta d} E_1 R e^{-\delta R} - e^{-\delta d} E_0[E_1 R e^{-\delta R} e^{\delta S_N}; S_N < 0])(1 + o(e^{-rd})) \\ &= e^{-\delta d} E_1 R e^{-\delta R} P_1(S_N > d)(1 + o(e^{-rd})) \\ &= \frac{E_1 R e^{-\delta R}}{E_1 e^{-\delta R}} P_0(S_N > d)(1 + o(e^{-rd})), \end{aligned}$$

where Lemma 1 is used in the last step. ∎

Finally, we have

THEOREM 1. As $d \to \infty$,

$$ARL_0 = \frac{e^{\delta d}(E_1 e^{-\delta R})^{-2} - 1 - \delta(d + E_1 R + (E_1 e^{-\delta R})^{-1} E_1 R e^{-\delta R})}{\delta^2/2} + o(e^{-rd}).$$

By using Hölder inequality and the fact that $(E_1 e^{-\delta R})^{-1} E_1 R e^{-\delta R} \geq E_1 R$, we get

$$ARL_0 \leq \frac{e^{\delta(d+2E_1R)} - 1 - \delta(d + 2E_1R)}{\delta^2/2} + o(e^{-rd}) \tag{10}$$

As δ is small, we may replace $E_1 R$ by ρ when the drift is taken as zero, which gives us (2). Thus, (2) slightly overestimates the true ARL_0, which is also confirmed from Table 1. A similar argument shows that (3) slightly underestimates the true ARL_1. In the following, we shall show that (2) is actually accurate in the order of o(1) as $\delta \to 0$ and $d \to \infty$ such that δd remains bounded. By taking Taylor series expansion, it follows that

$$E_1 e^{-\delta R} = 1 - \delta E_1 R + \frac{\delta^2}{2} E_1 R^2 + o(\delta^2). \tag{11}$$

A result from Problem 10.2 of Siegmund (1985) gives that

$$E_1 R = \rho + \frac{\delta}{2}(\rho_2 - \rho^2) + o(\delta), \tag{12}$$

where $\rho_2 = E_1 R^2$. Substituting (12) into (11) and simplifying it, we get

$$E_1 e^{-\delta R} = e^{-\delta\rho}(1 + o(\delta^2)). \tag{13}$$

On the other hand,

$$\begin{aligned}(E_1 e^{-\delta R})^{-1} E_1 R e^{-\delta R} &= E_1 R - \delta(E_1 R^2 - (E_1 R)^2) + o(\delta) \\ &= \rho - \frac{\delta}{2}(\rho_2 - \rho^2) + o(\delta).\end{aligned}$$

Combining with (12), we get

$$E_1 R + (E_1 e^{-\delta R})^{-1} E_1 R e^{-\delta R} = 2\rho + o(\delta). \tag{14}$$

Substituting (13) and (14) into (10), we see that the approximation (2) is actually accurate up to order o(1), which slightly improves Theorem 10.16 of Siegmund (1985) and also confirms accuracy of the approximation.

The above argument can be easily adapted to the exponential family case. In the nonsymmetric case, the second order approximations involve the third moment of R, see Pollak and Siegmund (1986) for some specific results. When δ is too large, even this second order approximation may not be satisfactory. In this case, we should use the result of Theorem 1 and take more terms in the expansion for $E_1 R$ in δ.

4.2. Boundary Correction for EWMA Procedure. In this subsection, we discuss in detail the approximations for the EWMA procedure. The key is to form a martingale for the EWMA process Z_n which only involves Z_n and the

time n. We write $B = b\sigma_\infty$ as the control limit for the EWMA process. The following lemma is the key for our discussion.

Denote $\psi(u) = \sum_1^\infty c((1-\beta)^{n-1}\beta u)$ as the cumulant generating function for Z_∞, where c(u) is the cumulant generating function for X_1.

From Novikov (1990), it is known that

LEMMA 5.

$$Y_n = \int_0^\infty \frac{1}{u}(e^{uZ_n} - 1)e^{-\psi(u)}du + n\ln(1-\beta)$$

is a martingale.

In the normal case,

$$\psi(u) = \frac{u^2}{2}\frac{\beta}{2-\beta} + u\mu,$$

when the mean is μ after the change. By changing the order of integrals, we have

$$ARL_0 = \frac{1}{-\ln(1-\beta)}E\int_0^{Z_\tau/(\frac{\beta}{2-\beta})^{1/2}} [\phi(x)]^{-1}\Phi(x)dx, \tag{15}$$

and

$$ARL_1(\mu) = \frac{1}{-\ln(1-\beta)}E\int_0^{Z_\tau/(\frac{\beta}{2-\beta})^{1/2}} \cdot \left[\phi\left(x - \mu\left(\frac{2-\beta}{\beta}\right)^{1/2}\right)\right]^{-1}\Phi\left(x - \mu\left(\frac{2-\beta}{\beta}\right)^{1/2}\right)dx.$$

If we ignore the overshoot, an obvious lower bound is given by

$$ARL_0 \geq \frac{1}{-\ln(1-\beta)}\int_0^b [\phi(x)]^{-1}\Phi(x)dx, \tag{16}$$

which is similar to (7) in the continuous time case by simply changing $-\ln(1-\beta)$ to β.

As we showed in Table 3, this lower bound is too crude for practical use. We thus consider the effect of the overshoot. Similar to the random walk case, we define the following ladder variables based on the EWMA process Z_n:

$$\tau^{(1)} = \inf\{n > 0 : Z_n > 0\},$$
$$\tau^{(2)} = \inf\{n > 0 : Z_{n+\tau^{(1)}} > Z_{\tau^{(1)}}\},$$

generally,

$$\tau^{(k)} = \inf\{n > 0 : Z_{n+\tau^{(1)}+\cdots+\tau^{(k-1)}} > Z_{\tau^{(1)}+\cdots+\tau^{(k-1)}}\}$$

Thus, $\tau = \tau^{(1)} + \cdots + \tau^{(N)}$ with

$$N = \inf\{k > 0 : Z_{\tau^{(1)}+\cdots+\tau^{(k)}} > B\}.$$

By writing

$$Z_n^{(k)} = Z_{n+\tau^{(1)}+\cdots+\tau^{(k-1)}} - (1-\beta)^n Z_{\tau^{(1)}+\cdots+\tau^{(k-1)}},$$

we see that $Z_n^{(k)}$ and Z_n are distributionally equivalent. If we approximate $Z_{\tau^{(1)}+\cdots+\tau^{(N-1)}}$ by B, then the mean overshoot can be approximated by

$$E[Z_\tau - B] \approx E[Z_\nu - (1-(1-\beta)^\nu)B],$$

where

$$\nu = \min\{n > 0 : Z_n > (1-(1-\beta)^n)B\}.$$

Thus, the calculation of the mean overshoot is transformed into another boundary crossing problem with a curved boundary.

In the following, we look at the mean overshoot behavior as $\beta \to 0$ as required under the optimal design. As $\beta \to 0$, Z_n/β behaves like a random walk with drift 0, and on the other hand,

$$(1-(1-\beta)^n)/\beta \to n.$$

Locally speaking, the ladder variables for the EWMA process can be approximated as a sequence of boundary crossing times for a random walk with randomly increasing drift parameters. Thus, a better approximation can be obtained as

$$ER \approx \beta\rho(-B)$$

as $\beta \to 0$, where $\rho(-B) = ES_{\tau_+}^2/2ES_{\tau_+}$ is evaluated with mean -B. As $\delta \to 0$, $\rho(-B) \to \rho = 0.583$. The numerical comparison with the numerical values given in Table 3 shows that this approximation is generally good. For example, for $\beta = 0.10$ and $b = 2.0$, the corrected diffusion approximation gives 149 which is very close to the numerical value 143 given by Robins and Ho (1978).

To see how much effect of this overshoot on ARL_0, we consider the first order approximation as $b \to \infty$. Under the optimal design, recall that $b(\beta/(2-\beta))^{1/2} \to 0.715\delta$. Based on this, the following rough approximation can be

obtained.

$$\begin{aligned} ARL_0 &= \frac{1}{-\ln(1-\beta)} E \int_0^{b+R(\beta(2-\beta))^{1/2}+o(\beta^{1/2})} [\phi(x)]^{-1}\Phi(x)dx \\ &= 1 - \ln(1-\beta) E \int_0^{b+\frac{1.43\delta R}{b}+o(\frac{1}{b})} [\phi(x)]^{-1}\Phi(x)dx \\ &= \frac{1}{-\ln(1-\beta)} E e^{1.43\delta R} \int_0^b [\phi(x)]^{-1}\Phi(x)dx \\ &\approx e^{0.834\delta} T, \end{aligned}$$

as δ is small. Therefore, in the first order, the correction factor for the EWMA procedure lies between the correction factors of CUSUM and Shiryayev-Roberts procedures.

Similar to the other two procedures, this correction may become less satisfactory when $B = b\sigma_\infty$ becomes larger. A more accurate approximation can be obtained by taking the second order expansion for $\rho(\mu)$ in μ. Partial results have been given by Chang (1989). For example,

$$E_\mu S_{\tau_+} = \frac{1}{\sqrt{2}}\left(1 + \mu\rho + \frac{\mu^2}{2}\rho^2 + o(\mu^2)\right).$$

Finally, we point out that the above argument is only heuristic, a formal treatment can be done similar to the method used by Pollak (1987) by considering the behaviour of Z_n in the region $(B(1-\epsilon), B)$ with a properly chosen small ϵ. More specifically, let

$$\tau_{B-\epsilon} = \inf\{n > 0; Z_n > B(1-\epsilon)\}.$$

Consider the process $\{Z_{\tau_{B(1-\epsilon)}+n}\}$ for $n \geq 1$. Then it will either cross the boundary B soon, or "not at all" in a near future. If it does, then locally, it behaves like a random walk; if it doesn't, wait until next time it crosses $B - \epsilon$ again. Details will be presented somewhere else.

A different approach may be to use the method of Siegmund (1985, Chap. 4) based on the time-scale transformation.

Acknowledgement: I am grateful to the referee for many helpful suggestions. This research is partially supported by a Postdoctoral Fellowship from NSERC.

REFERENCES

Box, G. E. P and Jenkins, G. M. (1963). Further contributions to adaptive quality control: simultaneous estimation of dynamics: Non-zero costs. *Bull. Internat. Statist. Inst.* **34**, 943–974.

Brook, D. and Evans, D. A. (1971). An approach to the probability distribution of CUSUM lengths. *Biometrika*, **59** 539–549.

Chang, T. (1989). *Random Walks, Moderate Deviations, and the CUSUM Procedure*, Ph.D. dissertation, Stanford University.

Van Dobben de Bruyn, C. S. (1968). *Cumulative Sum Tests.* Griffin, London.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *Ann. Math. Statist.* **42**, 1897–1908.

Lotov, V. I. (1991). On random walks in a strip. *Theory Prob. Appl.* **36**, 165–170.

Lucas, J. M. and Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics*, **32**, 1–29.

Montgomery, D. C. (1991). *Introduction to Statistical Quality Control.* 2nd ed., John Wiley and Sons, New York.

Novikov, A. (1990). On the first passage time of an autoregressive process over a level and an application to a "disorder" problem. *Theory Prob. Appl.* **35**, 269–279.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, **41**, 100-114.

Pollak, M. (1987) Average run lengths of an optimal method of detecting a change in distribution. *Ann. Statist.*, **15**, 749–779.

Pollak, M. and Siegmund, D. (1985). A diffusion process and its applications to detecting a change in the drift of Brownian motion. *Biometrika*, **72**, 267–80.

Pollak, M. and Siegmund, D. (1986). Approximations to the ARL of CUSUM tests. Technical Report, Department of Statistics, Stanford University.

Pollak, M. and Siegmund, D. (1991). Sequential detection of a change in a normal mean when the initial value is unknown. *Ann. Statist.*, **19**, 394–416.

Roberts, S. W. (1959). Control chart tests based on geometric moving average. *Technometrics*, **1**, 239–250.

Roberts, S. W. (1966). A comparison of some control chart procedures. *Technometrics*, **8**, 411–430.

ROBINS, P. B. and Ho, T. Y. (1978). Average run length of geometric moving average charts by numerical methods. *Technometrics*, **10**, 85–93.

SHIRYAYEV, A. N. (1963). On optimum methods in quickest detection problems. *Theory Prob. Appl.* **13**, 22–46.

SIEGMUND, D. (1979). Corrected diffusion approximations in certain random walk problems. *Adv. Appl. Prob.* **11**, 701–719.

SIEGMUND, D. (1985). *Sequential Analysis: Tests and Confidence Intervals.* Springer, Berlin.

SRIVASTAVA, M. S. and YANHONG WU (1993). Comparison of EWMA, CUSUM and Shiryayev-Roberts procedures for detecting a shift in the mean. (To appear in *Ann. Statist.*)

TAYLOR, H. M. (1975). A stopped Brownian motion formula. *Ann. Prob.*, **3**, 234–246.

WU, YANHONG (1991). *Some Contributions to On-line Quality Control.* Ph.D. Thesis, University of Toronto.

DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
TORONTO, ONTARIO, CANADA M5S 1A1