The Annals of Statistics
1993, Vol. 21, No. 2, 645–670

COMPARISON OF EWMA, CUSUM AND SHIRYAYEV–ROBERTS PROCEDURES FOR DETECTING A SHIFT IN THE MEAN[1]

By M. S. Srivastava and Yanhong Wu

University of Toronto

Pollak and Siegmund compared the Shiryayev–Roberts procedure with the CUSUM procedure for detecting a change in the drift of a Brownian motion based on the conditional average delay time. In this paper, the exponentially weighted moving average (EWMA) procedure proposed by Roberts is compared with the Shiryayev–Roberts and CUSUM procedures. The comparison is based on the stationary average delay time as advocated by Shiryayev. The optimal design for the EWMA procedure and its asymptotic properties are studied when the average in-control run length is large. The results show that the EWMA procedure is less efficient than the other two procedures.

1. Introduction. An important application of statistics lies in the area of quality control in which we are interested in detecting a shift in the mean of a production process as soon as it occurs. Shewhart's (1931) $\bar{x}$ charts, with various modifications, have been very popular in the past. This procedure, however, has been found to be somewhat inefficient in detecting small shifts. To overcome this shortcoming, several procedures have been developed over the past few decades. Among them, the CUSUM procedure [Page (1954)] seems to be the most popular one. Its properties have been thoroughly studied in the literature [see, for example, Van Dobben de Bruyn (1968)]. Another procedure, called the Shiryayev–Roberts procedure [Shiryayev (1963) and Roberts (1966)], was studied and compared with the CUSUM procedure by Pollak and Siegmund (1985). This comparison is based on the conditional average delay time in detecting the change point, given that no false alarm was made. The results showed that the Shiryayev–Roberts procedure is as powerful as the CUSUM procedure. In a more recent paper, Pollak and Siegmund (1991) further considered the case in which the initial level is unknown.

In this paper, we study the EWMA (exponentially weighted moving average) procedure proposed by Roberts (1959). The EWMA procedure has recently received considerable attention in the literature. A numerical comparison with the CUSUM procedure by Lucas and Saccucci (1990) showed that the EWMA procedure is quite competitive in most practical situations. However, its theoretical properties have not been studied as thoroughly as for the CUSUM and Shiryayev–Roberts procedures.

Received December 1990; revised August 1992.

[1]This research was supported by Natural Sciences & Engineering Research Council of Canada.

AMS 1991 *subject classifications.* Primary 62L10; secondary 62N10.

Key words and phrases. Average run length, Brownian motion, Ornstein–Uhlenbeck process, renewal process, stationary average delay time.

The purpose of this paper is to study the properties of the EWMA procedure under the continuous time model and to compare it with the CUSUM and Shiryayev–Roberts procedures. Our comparison, however, is different from the one made by Pollak and Siegmund (1985), since it will be based on the stationary average delay time (SADT), which was advocated by Shiryayev (1963). Thus, it will be assumed here that the change only occurs after many false alarms. This seems appropriate in quality control when the cost of false alarm is relatively less important than the delay-time in detection. In Section 2, we first introduce several standard notations and some results from diffusion theory used in our study. We shall see that under the continuous time model, all three detecting processes can be formulated as certain diffusion processes with a changed drift parameter. In Section 3, we study the properties of the EWMA procedure. The optimal design which minimizes the SADT for fixed average in-control run length (ARL_0) is considered. Approximate formulae for the optimal weight factor and the corresponding control limit are given. The corresponding results for the CUSUM and Shiryayev–Roberts procedures are given in Section 4. The comparison of the EWMA procedure with the CUSUM and Shiryayev–Roberts procedures is carried out in Section 5. The results show that the EWMA procedure is less efficient than the other two procedures when $\mathrm{ARL}_0 \to \infty$. An interesting result, however, is that the EWMA procedure is less sensitive to the reference value when the shift amount is unknown. Further discussions are given in Section 6 where we shall briefly show that the comparison based on the conditional average delay time (CADT) as done in Pollak and Siegmund (1985) can also be carried out asymptotically.

2. Definitions and some preliminary results.

We consider the following change-point problem in a Brownian motion. Let B_t be the standard Brownian motion with drift 0 and diffusion 1. We shall assume that $B_0 = 0$. Suppose the observation process is given by

$$dW_t = \delta I_{[t \geq \theta]}\, dt + dB_t,$$

where θ is the change point and δ is the amount of shift, both are assumed to be unknown. Let $H_t = \sigma(W_s,\ 0 \leq s \leq t)$ be the history of the observation process up to time t, and τ a stopping time adapted to $\{H_t\}$. At τ, an alarm will be made; if it is a false alarm, a new procedure starts again. This procedure continues until the detection of the change point.

Let N denote the number of false alarms before θ and $\{\tau_i\}$ for $i = 1, 2, \dots, N+1$, be the consecutive alarm intervals until the detection of the change point. Thus,

$$\tau_1 + \cdots + \tau_N < \theta \leq \tau_1 + \cdots + \tau_N + \tau_{N+1}.$$

The average delay time for $\theta = t$ is thus

$$\mathrm{ADT}(t) = E_t[\tau_1 + \cdots + \tau_{N+1} - t],$$

where $E_t[\cdot]$ denotes the expectation when the place of shift is at a fixed time t. When $t = 0$, it becomes the out-of-control average run length

$$\mathrm{ARL}_1 = \mathrm{ADT}(0) = E_0\tau.$$

In this paper, we are mainly interested in the situation in which there are many false alarms before the change point, although in Section 6, we also briefly consider the case in which there is no false alarm. There are two main reasons for this consideration. First, the change-point rarely occurs. Second, the cost of false alarm is relatively small compared to the loss due to delay in detection. Thus, we shall consider the stationary average delay time

$$\mathrm{SADT} = \lim_{t\to\infty} \mathrm{ADT}(t),$$

as the main measure for evaluating the performance of a detecting procedure.

By using renewal theory, we know that

$$\lim_{t\to\infty} P_t(t - (\tau_1 + \cdots + \tau_N) \le u) = \int_0^u P_\infty(\tau > x)\, dx / E_\infty\tau,$$

where $P_\infty(\cdot)$ and $E_\infty(\cdot)$ denote the probability and expectation when there is no change. In particular,

$$\mathrm{ARL}_0 = E_\infty\tau,$$

usually called the average in-control run length. Thus,

$$\begin{aligned}
\mathrm{SADT} &= \lim_{t\to\infty} \int_0^t E_t[\tau_1 + \cdots + \tau_{N+1} - t | t - (\tau_1 + \cdots + \tau_N) = u] \\
&\quad \times dP_t(t - (\tau_1 + \cdots + \tau_N) \le u) \\
&= \lim_{t\to\infty} \int_0^t E_u(\tau_1 - u | \tau_1 > u)\, dP_t(t - (\tau_1 + \cdots + \tau_N) \le u) \\
&= \int_0^\infty \mathrm{CADT}(u) \frac{P_\infty(\tau > u)}{E_\infty(\tau)}\, du,
\end{aligned}$$

where

$$\mathrm{CADT}(u) = E_u[\tau - u | \tau > u],$$

and is called the conditional average delay time. This is another measure used to assess the performance of a detecting procedure, and has been used for comparison in Pollak and Siegmund (1985) when $u \to \infty$ and $\mathrm{ARL}_0 \to \infty$. For simplicity, we shall use CADT to denote CADT(∞). However, it is usually difficult to find CADT(t) and the distribution of τ under P_∞.

To overcome this difficulty, we shall show that all three detecting processes are time homogeneous diffusion processes. Thus, we can use the results of diffusion theory as given in Karlin and Taylor (1981) to evaluate SADT and so on. We shall now define these three procedures for detecting a shift in a Brownian motion.

To define the EWMA procedure under the continuous time model, we first consider the procedure under the discrete time model. Suppose $\{z_k\}$, $k = 1, 2, \dots$, is the observation process which is a normal sequence with unit variance and shifted mean. Then the detecting process is defined by $y_k = \tilde{y}_k / (\mathrm{var}_\infty(\tilde{y}_\infty))^{1/2}$, where

$$\tilde{y}_k = (1 - \beta)\tilde{y}_{k-1} + \beta z_k, \quad \text{for } k \geq 1, \tilde{y}_0 = 0, 0 \leq \beta \leq 1.$$

The process is stopped and checked at the smallest value of n for which y_n exceeds a given value, called the control limit.

The continuous time process corresponding to $\tilde{y}_k$ can be written as

$$d\tilde{Y}_t = -\beta \tilde{Y}_t \, dt + \beta \, dW_t, \quad \text{for } \tilde{Y}_t = 0,$$

which is an Ornstein–Uhlenbeck process when the process is in control. The integral solution for $\tilde{Y}_t$ can be written as $\tilde{Y}_t = \beta \int_0^t e^{-\beta(t-s)} \, dW_s$. The variance of $\tilde{Y}_t$ and its limit when there is no change can easily be obtained as [see Karlin and Taylor (1981), pages 171]

$$\mathrm{Var}_\infty\big(\tilde{Y}_t\big) = \frac{\beta}{2}(1 - e^{-2\beta t}) \quad \text{and} \quad \lim_{t \to \infty} \mathrm{Var}_\infty\big(\tilde{Y}_t\big) = \frac{\beta}{2}.$$

As in the discrete time case, we define the EWMA process $Y_t = \tilde{Y}_t / (\mathrm{Var}_\infty(\tilde{Y}_\infty))^{1/2}$ by normalizing $\tilde{Y}_t$, that is,

$$dY_t = -\beta Y_t \, dt + \sqrt{2\beta} \, dW_t \quad \text{for } Y_0 = 0. \tag{1}$$

Thus, the stopping time is given by

$$\tau = \inf\{t > 0 \colon Y_t > b\},$$

where b is the control limit determined from a specified ARL_0.

The Shiryayev–Roberts procedure is obtained as a limit of a sequence of Bayes procedures for detecting a shift in the mean of a Brownian motion. The detecting process is given by

$$R_t = \int_0^t e^{\delta(W_t - W_s) - (\delta^2/2)(t-s)} \, ds \quad \text{with } R_0 = 0,$$

where the integrand is nothing but the likelihood ratio of the observation up to time t $\{W_s, 0 < s \leq t\}$ when the change is at s with respect to no change. The stopping rule is given by

$$\tau = \inf\{t > 0 \colon R_t \geq T\},$$

for a given control limit T such that ARL_0 is a given specified number. Using Itô's formula, R_t can be written as the following differential form:

$$dR_t = dt + \delta R_t \, dW_t \quad \text{for } R_0 = 0. \tag{2}$$

The CUSUM process is defined as

$$X_t = \max_{0 \le s \le t} \left(W_t - W_s - \frac{\delta}{2}(t - s) \right),$$

which can be obtained from the maximum likelihood ratio process. The stopping rule is given by

$$\tau = \inf\{t > 0 \colon X_t > d\},$$

for a specified control limit d such that ARL_0 is a specified number.

To write the CUSUM process in a stochastic differential form, we use the property for a reflected Brownian motion [see Karlin and Taylor (1981), page 385] which is given in the following lemma.

LEMMA 1. *X_t has the same probability structure as $|\tilde{X}_t|$ where $\tilde{X}_t$ is a diffusion process with the differential form*

$$d\tilde{X}_t = \frac{\delta}{2} \operatorname{sgn}(t - \theta) \operatorname{sgn}(\tilde{X}_t) \, dt + dB_t. \tag{3}$$

Thus, the stopping rule in this case becomes

$$\tau = \inf\{t > 0 \colon |\tilde{X}_t| > d\}.$$

From the above definitions, we see that all three detecting processes are diffusion processes. However, the CUSUM and Shiryayev–Roberts processes are not linear. A more important observation is that the change occurs only in the drift parameters of the three processes while the diffusion parameters remain unchanged. Thus, we consider the following general change-point problem in a diffusion process ξ_t,

$$d\xi_t = \left(\mu_0(\xi_t) I_{[\theta < t]} + \mu_1(\xi_t) I_{[\theta \ge t]} \right) dt + \sigma(\xi_t) \, dB_t,$$

with the drift parameter changed from $\mu_0(\cdot)$ to $\mu_1(\cdot)$ after the change point.

Let

$$\tau = \inf\{t > 0 \colon \xi_t \le a \text{ or } \xi_t \ge b\} \quad \text{for } a \le \xi_0 = x \le b,$$

be the first time ξ_t exits the interval (a, b). Assume that after each stopping, the process is forced back to the initial state x if it is a false alarm. We shall use the Green function for the random stopped process $\{\xi_t;\ 0 < t < \tau\}$.

Let

$$s_i(x) = \exp\left(\int^x -\frac{2\mu_i(u)}{\sigma^2(u)}\,du\right) \quad \text{and} \quad S_i(x) = \int^x s_i(u)\,du, \qquad i = 0, 1.$$

Then the Green function is defined as

$$G_i(x,z) = 2\frac{(S_i(b) - S_i(\max(x,z))(S_i(\min(x,z) - S_i(a))}{(S_i(b) - S_i(a))\sigma^2(z)s_i(z)} \tag{4}$$

for $i = 0, 1$.

Then $G_i(x, z)\,dz$ measures the total expected time spent in the infinitesimal interval $[z, z + dz)$ in a cycle, $i = 0, 1$.

Note that

$$\mathrm{ARL}_i(x) = \int_a^b G_i(x,z)\,dz \quad \text{for } i = 0, 1 \tag{5}$$

are the average run lengths with initial state x when the process is in control and out of control, respectively. When $x = 0$, we shall write $\mathrm{ARL}_i(0) = \mathrm{ARL}_i$, for $i = 0, 1$.

Under our consideration, the instantaneous return process which we shall call the controlled process will be at the stationary state when the change occurs. Using renewal theory, this stationary density can be written as

$$\alpha_0(y|x) = \frac{G_0(x,y)}{\mathrm{ARL}_0(x)} = \frac{G_0(x,y)}{\int_a^b G_0(x,z)\,dz}, \tag{6}$$

which is just the proportion of time spent by ξ_t in state y in a cycle; see Karlin and Taylor [(1981), page 260]. We shall write $\alpha_0(y)$ for $\alpha_0(y|0)$. By using the strong Markov property of a diffusion process, we get

$$\mathrm{SADT}(x) = \int_a^b \alpha_0(y|x)\,\mathrm{ARL}_1(y)\,dy. \tag{7}$$

When $x = 0$, we shall write SADT for SADT(0).

Using the above result, we can find the required characteristics for all three procedures. In this paper, we shall compare the SADT's of the three procedures for fixed $\mathrm{ARL}_0 = T$. In other words, we shall choose the control limits for the three procedures in such a way that they all have the same ARL_0. Our emphasis is, however, on the EWMA procedure.

3. EWMA procedure. In this section, we first obtain the operating characteristics of the EWMA procedure, such as ARL_0, ARL_1 and SADT. We then consider the optimal design for the EWMA procedure by searching for the optimal value of β which minimizes the SADT for fixed $\mathrm{ARL}_0 = T$ when the true shift value δ is assumed to be known. Finally, we study the behavior of

SADT when the true shift value is unknown. The control limit for the EWMA procedure is b, and the stopping time is

$$\tau = \inf\{t > 0: Y_t > b\}.$$

3.1. *Average run lengths and* SADT. To apply the results from diffusion theory given in Section 2, we let $a \to -\infty$. When the process is in control, Y_t has the diffusion parameters

$$\mu(x) = -\beta x \quad \text{and} \quad \sigma(x) = (2\beta)^{1/2}.$$

From (4), the corresponding Green function with $Y_0 = x$ is thus

$$G_0(x, z) = \frac{1}{\beta} e^{-z/2} \int_{\max(x,z)}^{b} e^{u^2/2}\, du.$$

Thus, from (5),

$$\mathrm{ARL}_0(x) = \int_{-\infty}^{b} G_0(x, z)\, dz = \frac{1}{\beta} \int_{x}^{b} [\phi(z)]^{-1} \Phi(z)\, dz,$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal pdf and cdf, respectively. The stationary density for the controlled process can be written from (6) as

$$\alpha_0(y|x) = \frac{\phi(y) \int_{\max(y,x)}^{b} [\phi(u)]^{-1}\, du}{\int_{x}^{b} [\phi(z)]^{-1} \Phi(z)\, dz} \quad \text{for } y \le b, \tag{8}$$

which is functionally free of β.

Now by taking the stationary state y as the new initial state with shifted drift, the process after the change can be written as

$$dY_t = \left(-\beta Y_t + (2\beta)^{1/2}\delta\right) dt + (2\beta)^{1/2}\, dB_t \quad \text{for } Y_0 = y,$$

with

$$\mu(x) = -\beta x + (2\beta)^{1/2}\delta, \qquad \sigma(x) = (2\beta)^{1/2}.$$

Thus, the corresponding Green function can be written from (4) as

$$G_1(y, z) = \frac{1}{\beta} \phi\left(z - \left(\frac{2}{\beta}\right)^{1/2} \delta\right) \int_{\max(y,z)}^{b'} \left[\phi\left(u - \left(\frac{2}{\beta}\right)^{1/2} \delta\right)\right]^{-1} du.$$

From (5), we get

$$\mathrm{ARL}_1(y) = \frac{1}{\beta} \int_{y}^{b} \left[\phi\left(u - \left(\frac{2}{\beta}\right)^{1/2} \delta\right)\right]^{-1} \Phi\left(u - \left(\frac{2}{\beta}\right)^{1/2} \delta\right) du.$$

Finally, the SADT can be obtained from (7) as

$$\begin{aligned}\mathrm{SADT}(x) &= \int_{-\infty}^{b} \mathrm{ARL}_1(y)\alpha_0(y|x)\,dy \\ &= \frac{1}{\beta\int_x^b [\phi(u)]^{-1}\Phi(u)\,du}\int_{-\infty}^{b}\phi(y)\int_{\max(y,x)}^{b}[\phi(u)]^{-1}\,du \\ &\qquad \times\left(\int_{y-\delta(2/\beta)^{1/2}}^{b-\delta(2/\beta)^{1/2}}[\phi(x)]^{-1}\Phi(x)\,dx\right)dy.\end{aligned}$$

Letting $x = 0$, we obtain the following results.

THEOREM 1. *For the EWMA procedure, the* ARL_0, ARL_1 *and* SADT *are, respectively, given by*

$$\mathrm{ARL}_0 = \frac{1}{\beta}\int_0^b [\phi(z)]^{-1}\Phi(z)\,dz, \tag{9}$$

$$\mathrm{ARL}_1 = \frac{1}{\beta}\int_0^b\left[\phi\left(u-\left(\frac{2}{\beta}\right)^{1/2}\delta\right)\right]^{-1}\Phi\left(u-\left(\frac{2}{\beta}\right)^{1/2}\delta\right)du. \tag{10}$$

$$\begin{aligned}\mathrm{SADT} &= \frac{1}{\beta\int_0^b [\phi(u)]^{-1}\Phi(u)\,du}\int_0^b[\phi(u)]^{-1} \\ &\qquad\times\int_{-\infty}^{u}\phi(y)\int_{y-\delta\sqrt{2/\beta}}^{b-\delta\sqrt{2/\beta}}[\phi(x)]^{-1}\Phi(x)\,dx\,dy\,du.\end{aligned} \tag{11}$$

3.2. *Asymptotic properties for optimal design.* For any detection rule, it would be desirable to have the average delay time in detecting the change as small as possible, and among all the procedures with the same ARL_0, the one having the smallest average delay time should be preferred. Traditionally, the comparison between two detecting procedures is usually based on ARL_1 [see Roberts (1966)], and a procedure with smaller ARL_1 is preferred. Thus, many efforts have been made to reduce ARL_1 such as the fast initial response (FIR) technique [see Lucas and Crosier (1982)]. However, in the following theorem, we shall show that for the EWMA procedure, one can always find a sequence of $\beta \to 0$ for which $\mathrm{ARL}_1 \to 0$. Thus, measures, other than ARL_1, such as SADT and CADT, should play more important role in selecting a procedure.

THEOREM 2. *For any fixed* $\mathrm{ARL}_0 = T$, *a sequence of* $\beta \to 0$, *can be chosen such that* $\mathrm{ARL}_1 \to 0$.

PROOF. Let

$$\beta_0 = \frac{\beta}{\delta^2} \quad \text{and} \quad T_0 = T\delta^2.$$

Since δ is assumed known and fixed, we may use β_0 and T_0 instead of β and T whenever convenient. Thus for fixed T, we should choose β_0 to minimize

$$\mathrm{ARL}_1 = \frac{1}{\beta}\int_0^b \left[\phi\left(x - \sqrt{2/\beta_0}\right)\right]^{-1}\Phi\left(x - \sqrt{2/\beta_0}\right)dx,$$

subject to the condition that

$$\int_0^b [\phi(x)]^{-1}\Phi(x)\,dx = T_0\beta_0 = T\beta.$$

Now we let $\beta_0 \to 0$. By expanding the integrand around zero, we find that for fixed $T_0 < \infty$,

$$b = \left(\frac{2}{\pi}\right)^{1/2} T_0\beta_0 + o(\beta_0).$$

In the next step as well as in the proof of following theorems in this section, we shall repeatedly use the following well-known approximation for the tail probability of the standard normal distribution

$$(12) \quad 1 - \Phi(x) = \frac{1}{x}\phi(x)\left(1 - \frac{1}{x^2} + \frac{3}{x^4} - \frac{15}{x^6} + O\left(\frac{1}{x^8}\right)\right) \quad \text{as } x \to \infty,$$

see Feller [(1957), page 179]. Thus, as $\beta_0 \to 0$,

$$\begin{aligned}
\mathrm{ARL}_1 &= \frac{1}{\beta}\int_{-\sqrt{2/\beta_0}}^{b-\sqrt{2/\beta_0}} [\phi(x)]^{-1}\Phi(x)\,dx \\
&= \frac{1}{\beta}\int_{-\sqrt{2/\beta_0}}^{b-\sqrt{2/\beta_0}} \left(-\frac{1}{x} + \frac{1}{x^3} + O\left(\frac{1}{x^5}\right)\right)dx \\
&= \frac{1}{\beta}\log\left(\frac{(2/\beta_0)^{1/2}}{(2/\beta_0)^{1/2} - b}\right) + O(\beta_0^{3/2}) \\
&= \frac{1}{\beta}b\left(\frac{\beta_0}{2}\right)^{1/2} + O(\beta_0^{3/2}) = \left(\frac{\beta_0}{\pi}\right)^{1/2} T + O(\beta_0^{3/2}).
\end{aligned}$$

Thus, as $\beta \to 0$, $\mathrm{ARL}_1 \to 0$, which completes the proof. □

In the following, we obtain the optimum value of β which minimizes the SADT for the EWMA procedure when the shift value δ is known. Our discussion will be mainly focused on the most interesting case in which $\mathrm{ARL}_0 \to \infty$. In the following theorem, we state the main result which gives the asymptotic form for the optimal parameter β^* as well as the corresponding SADT. The proof of the result will be completed after two lemmas.

THEOREM 3. *As* $T \to \infty$, *the optimal* β^* *that minimizes the* SADT *of the* EWMA *procedure is approximately*

$$\beta^* = \frac{\delta^2 c^*}{\log(\delta^2 T)}(1 + o(1))$$

and the corresponding minimum SADT *is approximately*

$$\mathrm{SADT}(\beta^*) = \frac{k^*}{\delta^2}\log(\delta^2 T)(1 + o(1)),$$

where $c^* \approx 0.5117$ *and* $k^* \approx 2.4554$.

Before we prove the theorem, we first give two lemmas which establish certain properties of the optimum β^* for large T.

LEMMA 2. *As* $T \to \infty$, *the optimum* $\beta^* \to 0$.

PROOF. We shall use the negative method to prove the result. From Theorem 1, our objective is to minimize SADT given in (11) under the constraint

$$\int_0^b [\phi(x)]^{-1}\Phi(x)\,dx = T_0\beta_0. \tag{13}$$

We now show that $\beta^* \to 0$, as $T \to \infty$. In fact, if β^* does not go to zero, then there must exist a subsequence of β, such that $\beta \to c > 0$ as $T \to \infty$, where c may be equal to infinity. For notational convenience, we still denote this subsequence as well as its corresponding control limit as β and b, respectively. As $T \to \infty$, from (13), $b \to \infty$. Since the left-hand side in (13) is dominated by $\sqrt{2\pi}\, b e^{b^2/2}$, we have, as $T \to \infty$,

$$b \approx \sqrt{2\log(\beta_0 T_0)}\,.$$

Hence from (11), we have as $T \to \infty$,

$$\begin{aligned}\mathrm{SADT} &= O\left(\frac{1}{\beta}\int_0^{b-\sqrt{2/\beta_0}}[\phi(x)]^{-1}\Phi(x)\,dx\right)\\ &= O\left(\frac{1}{\beta}\int_0^{b-\sqrt{2/c}}[\phi(x)]^{-1}\Phi(x)\,dx\right)\\ &\geq O(\sqrt{T}),\end{aligned}$$

which is obviously not the optimal choice. Thus, $\beta_0\,\sigma\;\;0$. □

LEMMA 3. *As $T \to \infty$, $\beta^* T \to \infty$.*

PROOF. We have already shown that as $T_0 \to \infty$, the optimum $\beta_0^* \to 0$. We again prove the result by contradiction. Suppose $\beta_0 T_0$ does not go to infinity. Then, there must exist a subsequence of β such that $\beta T \le c < \infty$, for some finite number c, as $T \to \infty$. Consequently, there exists a convergent sub-subsequence in the above subsequence. To avoid notational confusion, we shall still denote this sequence as β and its corresponding control limit as b. Thus, we may assume that $\beta_0 T_0 \to c < \infty$. We first show that c cannot be equal to 0. We know from (13) that if $\beta_0 T_0 \to 0$,

$$b = \left(\frac{2}{\pi}\right)^{1/2} \beta_0 T_0 + o(1),$$

and hence $b \to 0$ as well. Thus, since $\beta_0 \to 0$ and $b \to 0$, for $y \le b$, we have from (12)

$$\begin{aligned}
\int_{y-\sqrt{2/\beta_0}}^{b-\sqrt{2/\beta_0}} [\phi(x)]^{-1}\Phi(x)\,dx &= \int_{y-\sqrt{2/\beta_0}}^{b-\sqrt{2/\beta_0}} \left(-\frac{1}{x} + O\left(\frac{1}{x^3}\right)\right) dx \\
&= -\int_{y-\sqrt{2/\beta_0}}^{b-\sqrt{2/\beta_0}} \frac{dx}{x} + O(\beta_0) \\
&= \log\left(\frac{\sqrt{2/\beta_0} - y}{\sqrt{2/\beta_0} - b}\right) + O(\beta_0) \\
&= \sqrt{\beta_0/2}\,(b-y) + O(\beta_0).
\end{aligned} \tag{14}$$

Hence,

$$\begin{aligned}
\text{SADT} &= \frac{\int_0^b [\phi(x)]^{-1} \int_{-\infty}^u \phi(y)\sqrt{\beta_0/2}\,(b-y)\,dy\,du}{\beta \int_0^b [\phi(x)]^{-1}\Phi(x)\,dx} + O(1) \\
&= \frac{1}{\delta^2\sqrt{\beta_0}} \frac{\int_0^b [\phi(x)]^{-1} \int_{-\infty}^u (-y)\phi(y)\,dy\,du}{\int_0^b [\phi(x)]^{-1}\Phi(x)\,dx} + O(1) \\
&= \frac{4\sqrt{2\pi}}{\delta^2\sqrt{\beta_0}} + O(1).
\end{aligned}$$

That means,

$$\frac{\text{SADT}}{\sqrt{T_0}} = \frac{4\sqrt{2\pi}}{(\beta_0 T_0)^{1/2}} + O\left(\frac{1}{\sqrt{T}}\right) \to \infty,$$

which is not the optimal choice. Thus $c > 0$. Next, suppose $\beta_0 T_0 \to c > 0$.

Then b converges to a constant. Thus, by using (14), we have

$$\text{SADT} = \frac{\int_0^b [\phi(x)]^{-1} \int_{-\infty}^u \phi(y)\sqrt{\beta_0/2}\,(b-y)\,dy\,du}{\beta \int_0^b [\phi(x)]^{-1}\Phi(x)\,dx} + O(1)$$

$$= O\left(\frac{1}{\delta^2\sqrt{\beta_0}}\right) = O\left(\frac{\sqrt{T_0}}{\delta^2}\right).$$

This is not a good choice either. Thus, we have shown that $\beta_0^* T_0 \to \infty$. □

The above two lemmas show that the optimal β^* goes to zero in an order slower than $1/T$. The main problem is to find the rate of convergence which is given in Theorem 3.

PROOF OF THEOREM 3. Let us recall that our objective is to minimize (11) under the constraint (13). By carefully checking the expression given in (11) and following the steps of Lemmas 2 and 3, we see that to minimize SADT, b^* and β_0^* must satisfy

$$b - \sqrt{2/\beta_0} \to -\infty,$$

that is,

$$\frac{b^2\beta_0 - 2}{\beta_0} \to -\infty.$$

From Lemma 3, we know that $\beta_0 T_0 \to \infty$ and hence from (13)

$$b \approx (2\log(\beta_0 T_0))^{1/2}.$$

Thus, β_0^* and b^* should satisfy

$$\beta_0 b^2/2 \le 1,$$

when T is sufficiently large. Thus, there must exist a subsequence of β such that $\beta_0 b^2/2 \to c$ with $0 \le c \le 1$. It is easy to see that if we can show that c is unique, then the theorem will be proved. As was the case with the proof of Lemma 3, we still denote this subsequence by β. We first show that $c \ne 0$. In fact, if $c = 0$, then by using (12), we have,

$$\int_{y-\sqrt{2/\beta_0}}^{b-\sqrt{2/\beta_0}} [\phi(x)]^{-1}\Phi(x)\,dx = \int_{y-\sqrt{2/\beta_0}}^{b-\sqrt{2/\beta_0}} \left(-\frac{1}{x} + O\left(\frac{1}{x^3}\right)\right)dx$$

$$= \log\left(\frac{1-\sqrt{\beta_0/2}\,y}{1-\sqrt{\beta_0/2}\,b}\right) + o(\beta_0)$$

$$= \left(\frac{\beta_0 b^2}{2}\right)^{1/2}\left(1 - \frac{y}{b}\right) + o(\beta_0).$$

Thus,

$$\mathrm{SADT} = \frac{(\beta_0 b^2/2)^{1/2} \int_0^b [\phi(u)]^{-1} \int_{-\infty}^u \phi(y)(1 - y/b)\, dy\, du}{\delta^2 \beta_0 \int_0^b [\phi(x)]^{-1} \Phi(x)\, dx} + o(1)$$

$$= \frac{b^2}{(\beta_0 b^2/2)^{1/2}} + o(1).$$

But $b^2 \approx 2 \log T_0$, and thus

$$\frac{\mathrm{SADT}}{(\log T_0)} = \frac{1}{(\beta_0 b^2/2)^{1/2}} + o(1) \to \infty.$$

That means, $c = 0$ will not give the optimal β_0^*. Similarly, we can show that $c \neq 1$. Therefore we must have $0 < c < 1$.

By taking $\beta_0 = 2c/b^2$, we have by using (12) again

$$\int_{y-\sqrt{2/\beta_0}}^{b-\sqrt{2/\beta_0}} [\phi(x)]^{-1} \Phi(x)\, dx = \int_{y-\sqrt{2/\beta_0}}^{b-\sqrt{2/\beta_0}} \left(-\frac{1}{x} + O\left(\frac{1}{x^3}\right)\right) dx$$

$$= \log\left(\frac{1 - \sqrt{\beta_0/2}\, y}{1 - \sqrt{\beta_0/2}\, b} \right) + O\left(\frac{1}{b^2}\right)$$

$$= \log\left(\frac{1 - \sqrt{c}\, y/b}{1 - \sqrt{c}} \right) + O\left(\frac{1}{b^2}\right).$$

Thus,

$$\mathrm{SADT} = \frac{\int_0^b [\phi(u)]^{-1} \int_{-\infty}^u \phi(y) \log\left(\frac{1 - \sqrt{c}\, y/b}{1 - \sqrt{c}} \right) dy\, du}{\delta^2 \beta_0 \int_0^b [\phi(x)]^{-1} \Phi(x)\, dx} + O(1)$$

$$= \frac{b^2}{2\delta^2} \frac{-\log(1 - \sqrt{c})}{c} + O(1).$$

Therefore, to the first order of b^2, we should choose c to minimize $-\log(1 - \sqrt{c})/c$. The minimum value is 2.4554 and the c that minimizes it is obtained by numerical search as $c^* = 0.5117$. The result is obtained by noting that

$$b^* \approx (2 \log T_0)^{1/2}. \qquad \square$$

For moderate values of T, the approximation given in Theorem 3 is often too crude to apply directly as it is only up to the first order. In the following corollary, we give some higher order approximations for calculating SADT and the optimal value of β^*, which will be used in our comparison of the EWMA procedure with the CUSUM and Shiryayev–Roberts procedures.

COROLLARY 1. *For the* EWMA *procedure, as* $T \to \infty$,

$$\beta^* = \frac{2c^*\delta^2}{b^{*2}},$$

$$\text{SADT} = \frac{1}{\delta^2}\left[1.2277b^{*2} - 3.085 + \frac{22.92}{b^{*2}} - \frac{615}{b^{*4}} + O\left(\frac{1}{b^{*6}}\right)\right],$$

with

$$b^{*2} = 2\log\left(\left(\frac{2}{\pi}\right)^{1/2} c^* T_0\right) - \log\left[2\log\left(\left(\frac{2}{\pi}\right)^{1/2} c^* T_0\right)\right] + o(1).$$

PROOF. In the final step of the proof for Theorem 3, if we take higher order expansion for $\Phi(x)$ as $x \to -\infty$ as given in (12), we get for $\beta_0 = 2c/b^2$,

$$\int_{y-\sqrt{2/\beta_0}}^{b-\sqrt{2/\beta_0}} [\phi(x)]^{-1}\Phi(x)\,dx$$

$$= \int_{y-b/\sqrt{c}}^{b(1-1/\sqrt{c})} \left(-\frac{1}{x} + \frac{1}{x^3} - \frac{3}{x^5} + \frac{15}{x^7} + O\left(\frac{1}{x^9}\right)\right) dx$$

$$= -\log\frac{b(1-1/\sqrt{c})}{y-b/\sqrt{c}} - \frac{1}{2}\left(\frac{1}{b^2(1-1/\sqrt{c})^2} - \frac{1}{(y-b/\sqrt{c})^2}\right)$$

$$+ \frac{3}{4}\left(\frac{1}{(b(1-1/\sqrt{c}))^4} - \frac{1}{(y-b/\sqrt{c})^4}\right)$$

$$- \frac{15}{6}\left(\frac{1}{(b(1-\sqrt{c}))^6} - \frac{1}{(y-b/\sqrt{c})^6}\right) + O\left(\frac{1}{b^8}\right)$$

$$= -\log\frac{1-\sqrt{c}}{1-(y/b)\sqrt{c}} - \frac{c}{2b^2}\left(\frac{1}{(1-\sqrt{c})^2} - \frac{1}{(1-(y/b)\sqrt{c})^2}\right)$$

$$+ \frac{3c^2}{4b^4}\left(\frac{1}{(1-\sqrt{c})^4} - \frac{1}{(1-(y/b)\sqrt{c})^4}\right)$$

$$- \frac{15c^3}{6b^6}\left(\frac{1}{(1-\sqrt{c})^6} - \frac{1}{(1-(y/b)\sqrt{c})^6}\right) + O\left(\frac{1}{b^8}\right).$$

Thus,

$$\text{SADT} \approx \frac{b^2}{2c\delta^2}\left[-\log(1-\sqrt{c}) + \frac{\int_0^b [\phi(u)]^{-1}\int_{-\infty}^u \phi(y)\log\left(1-\sqrt{c}\,\frac{y}{b}\right)dy\,du}{\int_0^b [\phi(u)]^{-1}\Phi(u)\,du}\right.$$

$$- \frac{c}{2b^2}\frac{1}{(1-\sqrt{c})^2}$$

$$+\frac{c}{2b^2}\frac{\int_0^b[\phi(u)]^{-1}\int_{-\infty}^u\phi(y)\frac{1}{\left(1-(y/b)\sqrt{c}\right)^2}\,dy\,du}{\int_0^b[\phi(u)]^{-1}\Phi(u)\,du}$$

$$+\frac{3c^2}{4b^4\left(1-\sqrt{c}\right)^4}$$

$$-\frac{3c^2}{4b^4}\frac{\int_0^b[\phi(u)]^{-1}\int_{-\infty}^u\phi(y)\frac{1}{\left(1-(y/b)\sqrt{c}\right)^4}\,dy\,du}{\int_0^b[\phi(u)]^{-1}\Phi(u)\,du}$$

$$\left.-\frac{15c^3}{6b^6\left(1-\sqrt{c}\right)^6}+\frac{5c^3}{6b^6}\right]+O\left(\frac{1}{b^6}\right).$$

Now we take Taylor series expansions for the functions $\log(1-x)$, $1/(1-x)^2$ and $1/(1-x)^4$ around $x=0$, and substitute them in the above equation. Because of the symmetric property of $\phi(x)$, the terms with odd powers of y will go to zero in the exponential order of b. Thus, only the terms with even powers of y remain. By using the l'Hôpital rule, we obtain

$$\begin{aligned}\text{SADT} &= \frac{b^2}{2c\delta^2}\left\{-\log\left(1-\sqrt{c}\right)-\left[\frac{c}{2b^2}+\frac{3c^2}{4b^4}+\frac{30c^3}{6b^6}\right]-\frac{c}{2b^2\left(1-\sqrt{c}\right)^2}\right.\\ &\quad+\frac{c}{2b^2}\left[1+\frac{3c}{b^2}+\frac{15c^2}{b^4}\right]+\frac{3c^2}{4b^4\left(1-\sqrt{c}\right)^4}\\ &\quad\left.-\frac{3c^2}{4b^4}\left(1+\frac{6c}{b^2}\right)-\frac{15c^3}{6b^6\left(1-\sqrt{c}\right)^6}+\frac{15c^3}{6b^6}\right\}\\ &\quad+O\left(\frac{1}{b^6}\right)\\ &=\frac{b^2}{2c\delta^2}\left[-\log\left(1-\sqrt{c}\right)-\frac{c}{2b^2\left(1-\sqrt{c}\right)^2}+\frac{3c^2}{4b^4\left(1-\sqrt{c}\right)^4}\right.\\ &\quad\left.+\frac{c^3}{b^6}\left(\frac{1}{2}-\frac{5}{2\left(1-\sqrt{c}\right)^6}\right)\right]+O\left(\frac{1}{b^6}\right).\end{aligned}$$

By substituting the asymptotic optimal value $c^* = 0.5117$ into the above form, we get the approximate formula for SADT as

$$\mathrm{SADT} \approx \frac{1}{\delta^2}\left[1.2277b^{*2} - 3.085 + \frac{29.221}{b^{*2}} - \frac{615.05}{b^{*4}} + o\left(\frac{1}{b^4}\right)\right].$$

The approximation for b^* is obtained by expanding $\int_0^b[\phi(x)]^{-1}\Phi(x)\,dx$ in the form $(1/b)[\phi(b)]^{-1}[1 + \cdots]$ as $b \to \infty$ and then solving

$$\int_0^{b^*}[\phi(x)]^{-1}\Phi(x)\,dx = \frac{2c^*T_0}{b^{*2}}. \qquad \square$$

We carried out a numerical check and found the approximation for β^* quite satisfactory. However, it can be seen that the approximation for SADT is not good for small T because of the large coefficients and the lower order of convergence. However, we can obtain simple upper and lower bounds for SADT by taking the first term and the first two terms, respectively. The next theorem gives the results for small T.

THEOREM 4. *For the* EWMA *procedure, as* $T \to 0$, *the optimal value* β^* *approximately satisfies*

$$\beta^* \approx \frac{a^*}{T}$$

and the corresponding minimum SADT *is approximately*

$$\mathrm{SADT} \approx A^*T(1 + o(1)),$$

where

$$a^* = \int_0^{b^*}[\phi(x)]^{-1}\Phi(x)\,dx$$

and b^* *is the minimum point for*

$$A^* = \min_b \frac{\int_0^b[\phi(u)]^{-1}\int_{-\infty}^u\phi(y)\int_y^b[\phi(x)]^{-1}\Phi(x)\,dx\,dy\,du}{\left(\int_0^b[\phi(x)]^{-1}\Phi(x)\,dx\right)^2}.$$

PROOF. The proof is similar to Theorem 3. We first note that $\beta_0^* \to \infty$. Otherwise, as in the proof of Theorem 3, we may assume that $\beta \to c < \infty$ for notational convenience. If $\beta_0 \to 0$, then it is easy to see that $b \to 0$. Hence, as in the proof of Lemma 2, we get

$$\mathrm{SADT} \approx \frac{(\beta_0/\pi)^{1/2}}{\delta^2\beta_0} \approx \frac{1}{\delta^2(\pi\beta_0)^{1/2}} \to \infty.$$

Similarly, if $\beta_0 \to c$, a finite constant, then we can easily show that the SADT also goes to a constant, which is not the optimal choice.

Next, we show that $T_0\beta_0^* \to a > 0$. As in the argument above, we show that $T_0\beta_0$ does not go to either zero or infinity. We only give the details in the first case. If $T_0\beta_0 \to 0$, then since $\beta_0 \to \infty$,

$$b \approx T_0\beta_0 \to 0$$

and thus

$$\mathrm{SADT} \approx \frac{1}{\delta^2\beta_0} = \frac{T_0}{\delta^2\beta_0 T_0} = \frac{1}{\delta^2\sqrt{\pi\beta_0}}.$$

Hence,

$$\frac{\mathrm{SADT}}{T_0} \approx \frac{1}{\delta^2\beta_0 T_0} \to \infty,$$

which is not the optimal choice either.

Now by taking $T_0\beta_0 \to a > 0$, that is,

$$\beta_0 = \frac{1}{T_0}\int_0^b [\phi(x)]^{-1}\Phi(x)\,dx,$$

with fixed b. Then, we have,

$$\mathrm{SADT} \approx \frac{T_0}{\delta^2}\frac{\int_0^b [\phi(u)]^{-1}\int_{-\infty}^u \phi(y)\int_y^b [\phi(x)]^{-1}\Phi(x)\,dx\,dy\,du}{\left(\int_0^b [\phi(x)]^{-1}\Phi(x)\,dx\right)^2}.$$

Thus, if we choose b such that it minimizes the ratio of the two integrations, we obtain the optimal value of β and the minimum value for SADT. □

3.3. *The case of δ unknown.* The above discussion and results are based on the assumption that δ is known. Now we consider the case in which δ is unknown. In other words, we study in this subsection the effect of the true unknown μ on the wrong choice of δ, usually called the reference value. Suppose that the value β^* is the optimum choice corresponding to the reference value δ when the true shift is in fact μ. We shall write ARL_μ and SADT_μ to denote the average run length and SADT when the true shift value is μ. Using the method of Section 2, we can obtain:

THEOREM 5. *For the* EWMA *procedure, as* $T \to \infty$,

$$\mathrm{SADT}_\mu = \frac{1}{\delta^2}\left[-\frac{\log\left(1-(\delta/\mu)\sqrt{c^*}\right)}{2c^*}b^{*2} - \frac{\delta^2}{4\mu^2\left(1-(\delta/\mu)\sqrt{c^*}\right)^2} + O\left(\frac{1}{b^{*2}}\right)\right] \quad \text{for } \mu^2 > c^*\delta^2$$

$$= \frac{(2\pi)^{1/2}b^*}{2c^*\delta^2\left(1-(\mu/\delta)/\sqrt{c^*}\right)}\exp\left[\frac{b^{*2}\left(1-(\mu/\delta)/\sqrt{c^*}\right)^2}{2}\right](1+o(1)) \quad \text{for } \mu^2 < c^*\delta^2,$$

where c^* and b^{*2} are as given in Corollary 1.

PROOF. The proof is similar to that of Corollary 1. We first consider the case in which $\mu^2 > \delta^2 c^*$. In this case, $b^* - \mu\sqrt{2/\beta_0^*} \to -\infty$. Thus, for fixed y,

$$\int_{y-(\mu/\delta)\sqrt{2/\beta_0^*}}^{b^*-(\mu/\delta)\sqrt{2/\beta_0^*}} [\phi(x)]^{-1}\Phi(x)\,dx$$

$$= \int_{y-(\mu/\delta)\sqrt{2/\beta_0^*}}^{b^*-(\mu/\delta)\sqrt{2/\beta_0^*}} \left[-\frac{1}{x} + \frac{1}{x^3} + O\left(\frac{1}{x^5}\right)\right] dx$$

$$= \log\left(\frac{1-(\delta/\mu)\sqrt{c^*}\,y/b^*}{1-(\delta/\mu)\sqrt{c^*}}\right)$$

$$-\frac{1}{2}\left(\frac{1}{b^{*2}\left(1-(\delta/\mu)\sqrt{c^*}\right)^2} - \frac{1}{\left(1-(\delta/\mu)(y/b^*)\sqrt{c^*}\right)^2}\right) + O\left(\frac{1}{b^{*4}}\right).$$

Hence,

$$\mathrm{SADT} = \frac{b^{*2}}{2c^*\delta^2}\left[-\log\left(1-\frac{\delta}{\mu}\sqrt{c^*}\right) - \frac{c^*\delta^2}{2b^{*2}\mu^2\left(1-(\delta/\mu)\sqrt{c^*}\right)^2} + O\left(\frac{1}{b^{*4}}\right)\right].$$

In the case of $\mu^2 < \delta^2 c^*$, we note that as $T \to \infty$, $b^* - \mu\sqrt{2/\beta_0^*} \to \infty$. Thus, for fixed y,

$$\int_{y-(\mu/\delta)\sqrt{2/\beta_0^*}}^{b^*-(\mu/\delta)\sqrt{2/\beta_0^*}} [\phi(x)]^{-1}\Phi(x)\,dx$$

$$= \frac{(2\pi)^{1/2}}{b^*\left(1-(\mu/\delta)/\sqrt{c^*}\right)} \exp\left(\frac{\left(b^*\left(1-(\mu/\delta)/\sqrt{c^*}\right)\right)^2}{2}\right)(1+o(1)).$$

The results are obtained by substituting the above approximation into (2). □

REMARK 1. From Theorem 4, we know that as $T \to 0$, the asymptotic result for SADT has the same form up to the first order no matter what the true shift value is.

4. CUSUM and Shiryayev–Roberts procedures. In this section we briefly give the results on CUSUM and Shiryayev–Roberts procedures. Several results are known and some are derived. We first consider the case in which δ is known.

4.1. *δ known case.* In this subsection, we briefly study the other two procedures for the case of known δ. Suppose the control limits for the CUSUM and Shiryayev–Roberts procedures are d and T, respectively. The exact formulae for ARL_1 and SADT can be obtained similarly as for the EWMA procedure. For the CUSUM procedure, we need to use Lemma 1 with $a = -d$, $b = d$. For the Shiryayev–Roberts procedure, we let $a = 0$ and $b = T$.

THEOREM 6. *For the* CUSUM *procedure,*

$$\mathrm{ARL}_0 = T = \frac{e^{\delta d} - 1 - \delta d}{\delta^2/2}, \qquad \mathrm{ARL}_1 = \frac{e^{-\delta d} - 1 + \delta d}{\delta^2/2},$$

$$\mathrm{SADT} = \frac{6 - \delta^2 d^2 - (2\delta d + 3)e^{-\delta d} + (2\delta d - 3)e^{\delta d}}{\delta^2(e^{\delta d} - (1 + \delta d))}.$$

As $T \to 0$,

$$\mathrm{SADT} = \tfrac{5}{6}T(1 + O(\delta^2 T)),$$

$$\mathrm{ARL}_1 = T\left(1 - \tfrac{2}{3}\sqrt{\delta^2 T} + O(\delta^2 T)\right),$$

and as $T \to \infty$,

$$\mathrm{SADT} = \frac{2}{\delta^2}\left(\log\frac{\delta^2 T}{2} - 1.5 + O\left(\frac{2}{\delta^2 T}\log\frac{\delta^2 T}{2}\right)\right),$$

$$\mathrm{ARL}_1 = \frac{2}{\delta^2}\left(\log\frac{\delta^2 T}{2} - 1 + O\left(\frac{2}{\delta^2 T}\log\frac{\delta^2 T}{2}\right)\right).$$

ARL_0 and ARL_1 for the CUSUM procedure are well known [see Taylor (1975)]. SADT for the CUSUM procedure is new. Similarly, we can obtain the corresponding results for the Shiryayev–Roberts procedure, see Shiryayev (1963) or Pollak and Siegmund (1985).

THEOREM 7. *For the Shiryayev–Roberts procedure,*

$$\mathrm{ARL}_0 = T, \qquad \mathrm{ARL}_1 = \frac{2}{\delta^2}e^{2/(\delta^2 T)}\int_{2/\delta^2 T}^{\infty}\frac{e^{-x}}{x}\,dx,$$

$$\mathrm{SADT} = \frac{2}{\delta^2}\left(e^{2/(\delta^2 T)}\int_{2/\delta^2 T}^{\infty}\frac{e^{-z}}{z}\,dz - 1 + \frac{2}{\delta^2 T}\int_0^{\infty}e^{-2z/(\delta^2 T)}\frac{\log(1+z)}{z}\,dz\right).$$

As $T \to 0$,

$$\mathrm{SADT} = \frac{T}{2}(1 + O(\delta^2 T)),$$

$$\mathrm{ARL}_1 = T\left(1 - \frac{\delta^2 T}{2} + o(\delta^2 T)\right);$$

and as $T \to \infty$,

$$\mathrm{SADT} = \frac{2}{\delta^2}\left(\log\frac{\delta^2 T}{2} - 1.57772 + O\left(\frac{2}{\delta^2 T}\log\frac{\delta^2 T}{2}\right)\right),$$

$$\mathrm{ARL}_1 = \frac{2}{\delta^2}\left(\log\frac{\delta^2 T}{2} - 0.57772 + O\left(\frac{2}{\delta^2 T}\log\frac{\delta^2 T}{2}\right)\right).$$

4.2. δ *unknown case.* In this subsection, we discuss some asymptotic properties of the CUSUM and Shiryayev–Roberts procedures. There is no basic difficulty to obtain the exact formulae for ARL_μ and SADT_μ by using the method of Section 2. However, for the purpose of comparison, we only give the second order approximation.

We consider the general change point problem mentioned in Section 2, for a diffusion process ξ_t. If ξ_t has a proper stationary density function $\pi(z)$ when there is no change point, that is,

$$\pi(z) = \lim_{t\to\infty}\frac{dP}{dz}[\xi_t \le z],$$

then we have the following lemma.

LEMMA 4. *Suppose* ξ_t *is a diffusion process with an interval range* $[L, U]$. *Then if its stationary density function* $\pi(z)$ *exits and is proper,*

$$\pi(z) = \lim_{a\to L}\lim_{b\to U}\alpha_0(z|x),$$

where a and b are the exiting boundaries.

The proof is beyond the scope of this paper but can be proved directly by using the Green function or on the lines of Pollak and Siegmund (1985).

From this lemma, we know that as $\mathrm{ARL}_0 \to \infty$, the corresponding control limit for each procedure will go to ∞. Thus, from Lemma 1 of Pollak and Siegmund (1985), we know that SADT and CADT(∞) have the same first order approximation as $T \to \infty$. Therefore, the asymptotic results for the CUSUM and Shiryayev–Roberts procedures will be the same in the first order. These results are given in Theorems 1 and 2 of Pollak and Siegmund (1985). For later comparison, we list the corresponding results for SADT in the following theorem.

THEOREM 8. (i) *For the* CUSUM *procedure, as* $T \to \infty$,

$$\mathrm{SADT}_\mu = \frac{1}{\delta(\mu - \delta/2)}\left[\delta d - 1 - \frac{\delta^2}{2\mu(2\mu - \delta)} + \left(\frac{\delta}{(2\mu - \delta)}\right)e^{\delta d(1 - 2\mu/\delta)} + o(1)\right].$$

(ii) *For the Shiryayev–Roberts procedure, as* $T \to \infty$,

$$\mathrm{SADT}_\mu = \frac{1}{\delta(\mu - \delta/2)}\left[\log\frac{T_0}{2} - 0.5772 - \frac{\delta}{(2(\mu - \delta/2))} + o(1)\right] \quad \text{for } \mu > \frac{\delta}{2}$$

$$= \frac{1}{\delta(\delta/2 - \mu)}\left[\left(\frac{T_0}{2}\right)^{1-2\mu/\delta}\Gamma\left(1 - \frac{2\mu}{\delta}\right) - \log\frac{T_0}{2} + 0.5772 - \left(1 - \frac{2\mu}{\delta}\right) + o(1)\right] \quad \text{for } \mu < \frac{\delta}{2}.$$

REMARK 2. As in the case of EWMA procedure, it can be shown that as $T \to 0$, the SADT for the two procedures will have the same first order expansions no matter what the true value of the shift is.

5. Comparisons. In this section, we compare the three procedures. As showed in Theorem 2, we shall not compare the ARL_1's since it is not a comparable measure for the average delay time in our discussion. We shall therefore concentrate on the comparison of SADT's. The comparison for the case in which δ is known will be considered separately from the case of δ unknown in the following two sections.

5.1. *δ known.* First we give some analytical comparisons. It is known that both the CUSUM and Shiryayev–Roberts procedures have some optimal properties in certain sense. The CUSUM procedure, as shown by Lorden (1971) and further by Moustakides (1986), minimizes the essential supremum of $\mathrm{CADT}(t)$ in t. This minimum value coincidently turns out to be ARL_1 for both the CUSUM and Shiryayev–Roberts procedures. Therefore, the CUSUM procedure has smaller ARL_1 than the Shiryayev–Roberts procedure. However, for the EWMA procedure this coincidence does not happen. This is partially a reason why ARL_1 is not comparable in general.

On the other hand, the Shiryayev–Roberts procedure minimizes the SADT. Pollak and Siegmund (1985) made another comparison with the CUSUM procedure based on approximated values of ARL_1 and CADT. Their results show that there is very little difference between the two procedures. The same conclusion can be obtained from a SADT comparison based on the results of Theorem 8 as $T \to \infty$. However, we should note that for small T, the Shiryayev–Roberts procedure becomes much better than the CUSUM procedure.

We have shown in Theorem 3 that as $T \to \infty$, the EWMA procedure is not as efficient as the other two procedures in the first order of $\log T$. For the EWMA procedure, the coefficient is 2.4554, while it is 2 for the other two procedures. However, if we take into account the remaining term, that is, up to the

TABLE 1
Comparison of SADT's *with* δ *known*

	EWMA				CUSUM		Sh–Rob
T	δ	β^*	b^*	SADT	d	SADT	SADT
100	0.5	0.08	1.82	12.46	5.58	12.71	12.15
	1.0	0.19	2.32	5.26	4.01	5.32	5.16
	1.5	0.34	2.61	2.99	3.18	2.99	2.92
	2.0	0.52	2.80	1.96	2.67	1.95	1.91
500	0.5	0.04	2.40	22.70	8.43	22.77	22.17
	1.0	0.12	2.87	8.47	5.55	8.21	8.05
	1.5	0.24	3.12	4.45	4.23	4.34	4.27
	2.0	0.38	3.28	2.79	3.46	2.72	2.68

constant term, we can see that for the EWMA procedure, the remaining term is much smaller than those of the other two procedures. This can be verified by using the results of Corollary 1. This means that for moderate values of ARL_0, the difference between the EWMA procedure with the other two procedures will be reduced. However, we should note that for the EWMA procedure, the remainder term in the approximation is of the order of $1/\log T_0$, while for the other two procedures, the approximations are accurate to the order of $\log T_0/T_0$. Thus, the approximation for the EWMA procedure is less accurate than for the other two procedures. Also, for small T, the EWMA procedure becomes less efficient than the other two procedures although this case is not of practical significance.

To show these results, we give some numerical comparisons based on the approximations as well as on the exact evaluations of SADT's for the three procedures. Table 1 gives the values of SADT with $T = 100$ and 500 for the three procedures. For the EWMA procedure, we use the approximate formula given in Corollary 1 for the optimum β. However, the SADT is calculated from the exact formula given in Theorem 3. For the other two procedures, we directly evaluate SADT by using the exact formula.

From the table, we observe that in the case of δ known, the EWMA procedure appears to be quite competitive especially for moderate ARL_0's, while the other two procedures are almost indistinguishable.

5.2. *δ unknown case.* In the following, we briefly consider the unknown δ case since we rarely know the exact shift value. One important reason for this comparison is to investigate the sensitivity of the three procedures to the choice of the reference value. The comparison for the three procedures in this case is much more complicated. For simplicity, in Table 2, we only give the values of SADT when the reference values are taken as $\delta = 0.5$ and 1.0 with $T = 100$ and 500. For the EWMA procedure, we choose β^* and b^* as the approximated optimal values given in Table 1. The SADT is calculated by using the approximation for $\mu > \delta$ and exactly for $\mu \leq \delta$. It might be checked

TABLE 2
Comparison of SADT *with* δ *unknown for* $T = 500$

		$T = 100$			$T = 500$		
δ	μ	EWMA	CUSUM	Sh–Rob	EWMA	CUSUM	Sh–Rob
0.5	0.5	12.46	12.71	12.15	22.70	22.77	22.17
	1.0	4.52	4.55	4.31	9.34	8.35	8.59
	1.5	3.33	2.81	2.80	5.94	5.09	5.37
	2.0	2.46	2.03	2.06	4.34	3.66	3.90
1.0	0.5	13.28	16.08	15.47	24.01	30.80	29.17
	1.0	5.26	5.32	5.16	8.47	8.21	8.01
	1.5	3.00	2.84	2.84	4.81	4.36	4.44
	2.0	2.18	1.95	2.00	3.41	2.97	3.07

numerically that for $\mu > \delta$, the approximations are very accurate by investigating the constant term in the approximate formula for SADT. The same argument also applies to the other two procedures by using Theorem 8. However, for $\mu \le \delta$, the approximation is too crude to be satisfactory especially for the EWMA procedure.

First, we note from Theorem 8 that the difference between the CUSUM and the Shiryayev–Roberts procedures is only in the constant term. Roughly speaking, for

$$0.5772 + \frac{\delta}{2(\mu - \delta/2)} \ge 1 + \frac{\delta^2}{2\mu(2\mu - \delta)},$$

that is, $\mu \ge 1.18\delta$, the CUSUM procedure is approximately better than the Shiryayev–Roberts procedure, as noted by Pollak and Siegmund (1985). In the opposite case, the Shiryayev–Roberts procedure is slightly better. This can also be seen from the table. A partial conclusion from this observation is that if we want to balance the two procedures, we should choose smaller reference values for the CUSUM procedure and larger reference values for the Shiryayev–Roberts procedure if μ is totally unknown.

Second, it is easy to check that for $\mu > \delta\sqrt{c^*}$, the SADT for the EWMA procedure is always larger than the ones for the other two procedures in the first order since

$$\frac{1}{\delta(\mu - \delta/2)} \le \frac{-\log\left(1 - (\delta/\mu)\sqrt{c^*}\right)}{\delta^2 c^*}.$$

However, as we can see from the table, the difference is not significant. On the other hand, from Corollary 1 and Theorem 8, we see that for $\mu < \delta\sqrt{c^*}$, SADT $= O(T^{(1-(\mu/\delta)/\sqrt{c^*})^2})$ for the EWMA procedure, while it is $O(T^{1-2\mu/\delta})$ for the other two procedures. Thus, roughly speaking, for $\mu < 0.4085\delta$, the EWMA procedure is better than the other two procedures. The results in Table 2 confirm this point. For example, for $\delta = 1.0$ and $\mu = 0.5$, the EWMA procedure has smaller SADT's than the other two procedures. This suggests that if we try to balance the EWMA with the other two procedures, we should choose relatively larger reference values for the EWMA procedure if the true shift value is totally unknown. For example, for $T = 500$, if the reference value δ is selected as 0.5 for the CUSUM procedure, we may choose $\delta = 1$ for the EWMA procedure. Under this match, the disadvantages for the EWMA procedure can be reduced if we compare the SADT's for all values of μ.

6. Further discussions. In this paper, we have studied the EWMA procedure in the continuous time case. It is shown that when $\mathrm{ARL}_0 \to \infty$, the EWMA procedure is less efficient than the CUSUM and Shiryayev–Roberts procedures, while the latter two procedures are almost indistinguishable. In the following, we discuss several points briefly.

1. Pollak and Siegmund (1985) compared the CUSUM procedure with the Shiryayev–Roberts procedure under a different criterion, that is, the conditional average delay time (CADT), as defined in the introduction. A similar comparison can be carried out for the EWMA procedure with the other two procedures. Consider the general change point problem as discussed in the introduction and write

$$\beta_0(y)\,dy = \lim_{t\to\infty} P_0(\xi_t \in dy \mid \tau > t) \quad \text{for } a \le y \le b,$$

as the quasistationary distribution for ξ_t under the stopping rule τ. Then similar to the derivation of formula (7) for SADT, we can show that

$$\mathrm{CADT} = \int_a^b \beta_0(y)\,\mathrm{ARL}_1(y)\,dy.$$

However, exact evaluation of $\beta_0(y)$ for a diffusion process ξ_t is almost impossible and so is the CADT. Thus, an accurate approximation is usually necessary for a comparison. In our case, all three detecting processes have stationary distributions. Pollak and Siegmund (1986) have shown that a result similar to our Lemma 4 also holds for the quasistationary distribution, that is,

$$\pi(y) = \lim_{a\to L}\lim_{b\to U} \beta_0(y).$$

Obviously, this implies that CADT and SADT have the same first order approximations. A further investigation by using the technique in Pollak and Siegmund (1985) can show that the CADT and SADT have actually the same expansions up to the second order, see Theorem 8 for the CUSUM and Shiryayev–Roberts procedures. The detailed argument will appear elsewhere.
2. As the referee has pointed out, as $\mathrm{ARL}_0 \to \infty$, the optimal $\beta^* \to 0$ in the order of $1/\log \mathrm{ARL}_0$ and hence the EWMA process Y_t goes to stationary state rather slowly. However, for this process $\mathrm{Var}_\infty(Y_t) = 1 - e^{-2\beta t}$ which goes to 1 in the exponential order of t. Thus, the process Y_t will become stationary in an average time of order $\log \mathrm{ARL}_0$, which is very short compared to ARL_0. On the other hand, the other two detecting processes

enter the stationary state in a constant average time as $\mathrm{ARL}_0 \to \infty$. Therefore, compared to the other two procedures, the speed to enter the stationary state is relatively slow for the EWMA procedure.

3. From a practical point of view, an adjustment of the results in the continuous time case is needed in order to match it with the discrete time model. Some techniques for this kind of correction have been developed by Siegmund (1985) and have been used for the CUSUM and Shiryayev–Roberts procedures [see Pollak (1987) and Wu (1991)]. However, for the EWMA procedure, it seems that a different technique is needed since the process cannot be approximated by a random walk. Many efforts have been made to obtain closer lower or upper bound for ARL_0 and ARL_1. Some crude approximation for ARL_0 for the EWMA procedure can be found in Lai (1974). Obviously, for the same control limits, ARL_0 obtained in the continuous time case is a lower bound for the discrete time model, and the same is true for the other two procedures. On the other hand, as $\mathrm{ARL}_0 \to \infty$, we know from Theorem 3 that the optimal weight factor β^* goes to zero in the order of $1/\log \mathrm{ARL}_0$. Thus, a second order adjustment in the discrete time case seems unnecessary. For example, if we compare the optimal values of β^* in Table 1 with the optimal values obtained in Lucas and Saccucci (1990) by numerical search in the discrete time case, we find that they are almost the same.

Acknowledgments. We would like to thank the referee for the suggestion to include the optimal search of β in the EWMA procedure and several helpful comments. We are also grateful to Professor D. Siegmund for some valuable discussions on the EWMA procedure.

REFERENCES

FELLER, W. (1957). *An Introduction to Probability Theory and Its Applications*. Wiley, New York.

KARLIN, S. and TAYLOR, H. M. (1981). *A Second Course In Stochastic Processes*. Academic, New York.

LAI, T. L. (1974). Control charts based on weighted sums. *Ann. Statist.* **2** 134–147.

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

LUCAS, J. M. and CROSIER, P. R. (1982). Fast initial response (FIR) for the cumulative sum quality control schemes. *Technometrics* **24** 199–205.

LUCAS, J. M. and SACCUCCI, M. S. (1990). Exponentially weighted moving average control charts schemes: properties and enhancements. *Technometrics* **32** 1–30.

MOUSTAKIDES, G. V. (1986). Optimal stopping times for detecting changes in distributions. *Ann. Statist.* **14** 1379–1387.

NADLER, J. and ROBBINS, N. B. (1971). Some characteristics of Page's two-sided procedure for detecting a change in a location parameter. *Ann. Math. Statist.* **42** 538–551.

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** 267–280.

POLLAK, M. and SIEGMUND, D. (1986). Convergence of quasi-stationary to stationary distribution for stochastic Markov processes. *J. Appl. Probab.* **23** 215–220.

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.

SHEWHART, W. (1931). *Economic Control of Quality of Manufactured Product.* Van Nostrand, Princeton.

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

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

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

VAN DOBBEN DE BRUYN, C. S. (1968). *Cumulative Sum Tests*. Griffin, London.

WU, Y. (1991). Some contributions to on-line quality control. Ph.D. dissertation, Univ. Toronto.

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