3-3-2010

Average Run Length of Two-Span Moving Sum Algorithms

Swarnendu Kar
Syracuse University, swkar@syr.edu

Kishan G. Mehrotra
Syracuse University, mehrotra@syr.edu

Pramod Varshney
Syracuse University, varshney@syr.edu

SYR-EECS-2010-02 Mar. 3, 2010

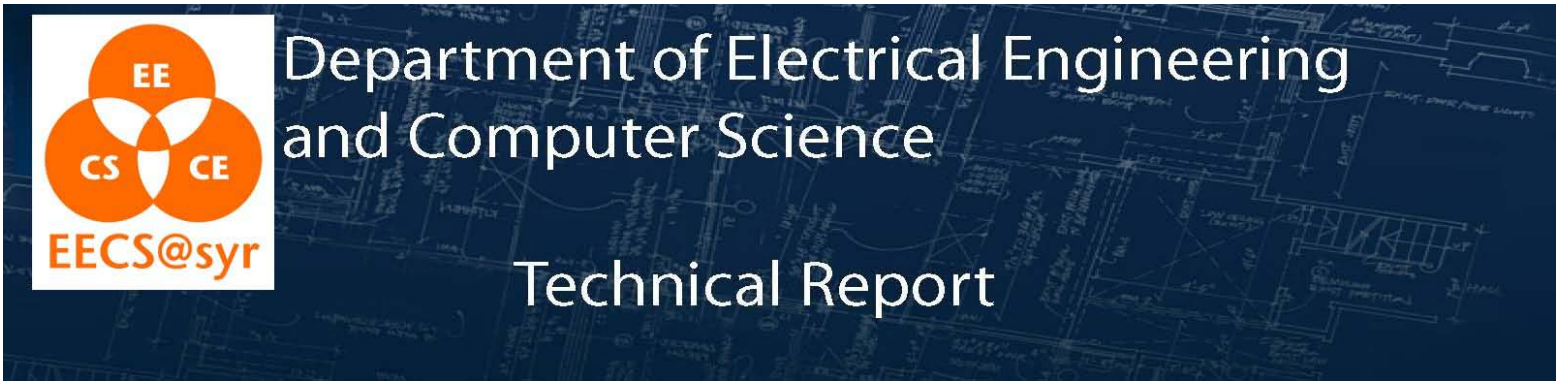
Average Run Length of Two-Span Moving Sum Algorithms

Swarnendu Kar swkar@syr.edu
Kishan G. Mehrotra mehrotra@syr.edu
Pramod Varshney varshney@syr.edu

ABSTRACT: Among the various procedures used to detect potential changes in a stochastic process the moving sum algorithms are very popular due to their intuitive appeal and good statistical performance. One of the important design parameters of a change detection algorithm is the expected interval between false positives, also known
as the average run length (ARL). In this paper, we have derived closed form expressions of ARL for two special cases – namely the two-span moving sum and filtered derivative algorithms. We have assumed that the random variables are uniformly distributed.

KEYWORDS: Average Run Length, Change Detection, Moving Average, Filtered Derivative, Control Charts

Syracuse University - Department of EECS,
4-206 CST, Syracuse, NY 13244
(P) 315.443.2652 (F) 315.443.2583
http://eecs.syr.edu



SYR-EECS-2010-02 Mar. 3, 2010

Average Run Length of Two-Span Moving Sum Algorithms

Swarnendu Kar swkar@syr.edu
Kishan G. Mehrotra mehrotra@syr.edu
Pramod Varshney varshney@syr.edu

ABSTRACT: Among the various procedures used to detect potential changes in a stochastic process the moving sum algorithms are very popular due to their intuitive appeal and good statistical performance. One of the important design parameters of a change detection algorithm is the expected interval between false positives, also known as the average run length (ARL). In this paper, we have derived closed form expressions of ARL for two special cases - namely the two-span moving sum and filtered derivative algorithms. We have assumed that the random variables are uniformly distributed.

KEYWORDS: Average Run Length, Change Detection, Moving Average, Filtered Derivative, Control Charts

Syracuse University - Department of EECS,
4-206 CST, Syracuse, NY 13244
(P) 315.443.2652 (F) 315.443.2583
http://eecs.syr.edu

Average run length of two-span moving sum algorithms

Swarnendu Kar*, Kishan G. Mehrotra, Pramod K. Varshney

Department of Electrical Engineering and Computer Science
Syracuse University, Syracuse, NY, 13244

Abstract

Among the various procedures used to detect potential changes in a stochastic process the moving sum algorithms are very popular due to their intuitive appeal and good statistical performance. One of the important design parameters of a change detection algorithm is the expected interval between false positives, also known as the average run length (ARL). In this paper, we have derived closed form expressions of ARL for two special cases - namely the two-span moving sum and filtered derivative algorithms. We have assumed that the random variables are uniformly distributed.

Key words: Average Run Length, Change Detection, Moving Average, Filtered Derivative, Control Charts

1. Introduction

The problem of detecting a change in the mean value of a process, when both the change point and change magnitude are unknown, is of great importance in various disciplines such as econometrics, engineering and quality control. The optimal scheme, which involves Maximum-Likelihood estimation of both the change point and the change magnitude, is computationally prohibitive. Hence, various simple but suboptimal methods like ordinary moving average (MA) and filtered derivative (FD) are used in practice. For data streams of dynamic nature, MA and FD algorithms are particularly useful in generating synopses that approximate the most recent data to answer queries or discover patterns. For example, MA has been used in the context of speech recognition (Li et al., 2002) and technical analysis of financial data (Murphy, 1999). FD has been used in the context of edge detection (Basseville, 1981).

Change detection schemes are assessed on the basis of the statistical distribution of the run length, i.e., the number of test samples taken before a false positive is detected. For most practical purposes, the distribution function of run length is adequately summarized by its expected (or average) value, also known as the average run length (ARL) (e.g. see Basseville and Nikiforov (1993)). Unfortunately, for most practical schemes closed form expression for ARL is difficult to obtain. For the MA scheme, the bulk of the research so far has been dedicated to either tabulating numerical results through Monte Carlo simulations (SAS/QC® 9.1 User's Guide, 2004) or deriving bounds using multivariate probability distribution functions (MPDFs) (Bohm and Hackl, 1990). Very little work has been done to date regarding the ARL of the FD scheme. If there are closed form expressions to be found, the most logical place to look for them is in the simplest problem. For this, our proxy is the class of algorithms with span $k = 2$. The random variables are assumed to be uniformly distributed. Since the cumulative distribution function of any random variable is uniformly distributed, this assumption may not be too restrictive.

The ARL can be written, as we shall see in Section 2, as the sum of an infinite number of MPDFs of increasingly higher dimensions. But MPDFs, in general, can be computed only numerically and the computational intensity increases with the dimension of the multivariate vector. While addressing the problem of approximating the ARL, it was observed by Robinson and Ho (1978) that the ratio of the successive MPDFs converge as the dimension increases. This fact was used to propose a series based approach of approximating the ARL. For the moving sum

*Corresponding author. Tel: (315) 751-1370 Fax: (315) 443-4745
Email addresses: swkar@syr.edu (Swarnendu Kar), mehrotra@syr.edu (Kishan G. Mehrotra), varshney@syr.edu (Pramod K. Varshney)

algorithm with positive weights and by using only k^{th} order MPDFs, upper and lower bounds were proposed by Lai (1974) and improved by Bohm and Hackl (1990). This was based on the idea that an MPDF of dimension larger than k can be bound (both above and below) by products of lower order MPDFs. Kar et al. (2009) derived the ARL for two-span MA and FD algorithms, where the threshold is the mean of the test statistic. In this paper, we generalize this result further, considering arbitrary thresholds. For uniformly distributed random variables, we derive closed form expressions for the multivariate probabilities and subsequently the ARLs.

2. Main Results

Let $X_1, X_2, \ldots X_m, \ldots$ be a sequence of observations obtained from a discrete time random process. We assume that X_i's are independently distributed random variables with variance σ^2. It is assumed that the mean of X_i's possibly changes from ν to θ at some point, here σ^2, ν and $\theta > \nu$ are unknown parameters and possible time of change is also unknown. For detecting a change in the mean value of X_i we need to formulate an appropriate linear test statistic, say $Y_m = \sum_{i=m-k+1}^{m} c_{m-i} X_i$, where c_i are constants and compare it against some threshold. Roberts (1959) has considered MA where all the weights are equal. A generalization of MA was considered by Bohm and Hackl (1990), where $c_i \geq 0$ for all i's. There are other applications where all the weights need not be positive. For example, in the context of edge detection, Basseville (1981) uses the following test statistic

$$Y_m = \sum_{i=m-k+1}^{m-k/2} X_i - \sum_{i=m-k/2+1}^{m} X_i,$$

where k is assumed to be an even number. This is also known as the filtered derivative (FD) algorithm, since we take the difference of averaged (filtered) blocks of samples.

To test whether the process mean is ν or has shifted to θ, the test statistic Y_m is monitored for successive values of m and compared against an upper threshold, say t. The time elapsed before Y_m exceeds the thresholds for the first time is also known as the run length (RL) or stopping time.

$$RL(t) = \max\{m : Y_m < t\}$$

In this paper, we are interested in the average run length (ARL), namely the expectation of RL, i.e., in

$$L(t) = \mathrm{E}(RL(t)).$$

Let

- $p_n(t)$ denote the probability that RL$= n$, i.e., the test passes $n-1$ consecutive times but fails at the n^{th} instant, i.e.

$$p_n(t) = \mathrm{P}(Y_{k+1}, Y_{k+2}, \ldots, Y_{k+n-1} < t, Y_{k+n} > t).$$

- $q_n(t)$ denote the probability that the test passes n consecutive times, i.e.

$$q_n(t) = \mathrm{P}(Y_{k+1}, Y_{k+2}, \ldots, Y_{k+n} < t).$$

Throughout the paper, we refer to $q_n(t)$ as the n^{th} survival probability. It follows from these definitions that $p_n(t) = q_{n-1}(t) - q_n(t)$. Since the evaluations start at index k, the ARL function can be represented as

$$\begin{aligned} L(t) &= k - 1 + \sum_{n=1}^{\infty} n p_n(t) \\ &= k + \sum_{n=1}^{\infty} q_n(t). \end{aligned} \tag{1}$$

In this paper, we use (1) to derive closed form expressions for two-span moving sum algorithms, i.e., when $k = 2$. In particular, we consider two cases, namely the moving average algorithm, where $c_0 = c_1 = 1$, and the filtered derivative algorithm, where $c_0 = 1, c_1 = -1$. The results are stated in the following section.

2.1. The Survival Probabilities

In this section, we obtain closed form expressions for the survival probabilities.

Theorem 1. *Let $X_1, X_2, \ldots X_n, \ldots$ be independent and uniformly distributed in $[0,1]$ random variables. Then,*

1. *For the two-span filtered derivative algorithm with threshold $t \in [-1, 0)$,*

$$q_n(t) = \begin{cases} \frac{(1+nt)^{n+1}}{(n+1)!} & \text{if } n < |t|^{-1} \\ 0 & \text{if } n \geq |t|^{-1} \end{cases} \tag{2}$$

2. *For the two-span moving average algorithm with threshold $t \in [0, 1]$,*

$$q_n(t) = t^{n+1} \frac{G^{(n+1)}(0)}{(n+1)!} \tag{3}$$

where $G(z) = \sec(z) + \tan(z)$

Proof. **1.** Consider any $t < 0$. From the definition of filtered alternatives for $k = 2$, it follows that

$$q_n(t) = \mathrm{P}(X_i - X_{i+1} < t \text{ for } i = 1, \ldots, n).$$

For $q_n(t)$ to be non-zero, since all the inequalities must be satisfied, therefore, we get $X_{i+1} \geq X_i - t$ for $i = 1, \ldots, n$. In addition, because all variables are uniform between [0, 1] we have $0 \leq X_i \leq 1$ for all i. Consequently, for $q_n(t)$ to be non-zero, the following must be satisfied:

$$1 \geq X_{n+1} > X_n - t > X_{n-1} - 2t > \cdots > X_1 - nt, \text{ and } -nt \leq 1 \tag{4}$$

Hence $q_n(t) = 0$ for $n \geq |t|^{-1}$. To derive the non-zero expression for $q_n(t)$, assume $n < |t|^{-1}$. From (4), we get the following integral representation of $q_n(t)$

$$\begin{aligned} q_n(t) &= \int_0^1 \ldots \int_0^1 \int_0^1 \mathrm{I}(x_1 - x_2 < t, x_2 - x_3 < t, \ldots, x_n - x_{n+1} < t)\, \mathrm{d}x_1\, \mathrm{d}x_2 \ldots\, \mathrm{d}x_{n+1} \\ &= \int_{-nt}^1 \int_{-(n-1)t}^{x_{n+1}+t} \ldots \int_{-2t}^{x_4+t} \int_{-t}^{x_3+t} \int_0^{x_2+t} \mathrm{d}x_1\, \mathrm{d}x_2\, \mathrm{d}x_3 \ldots\, \mathrm{d}x_n\, \mathrm{d}x_{n+1}, \end{aligned} \tag{5}$$

where I(.) denotes the identity function. Let us define $q'_n(y, t)$ by

$$q'_n(y,t) = \int_{-nt}^{y} \int_{-(n-1)t}^{x_{n+1}+t} \ldots \int_{-2t}^{x_4+t} \int_{-t}^{x_3+t} \int_0^{x_2+t} \mathrm{d}x_1\, \mathrm{d}x_2\, \mathrm{d}x_3 \ldots\, \mathrm{d}x_n\, \mathrm{d}x_{n+1}. \tag{6}$$

Clearly, we can easily obtain $q'_n(1, t) = q_n(t)$. To prove (2), it suffices to show that

$$q'_n(y,t) = \frac{(y+nt)^{n+1}}{(n+1)!}, \tag{7}$$

which we shall prove now by mathematical induction.

The base case $q'_1(y, t)$ is readily verified as

$$q'_1(y,t) = \int_{-t}^{y} \int_0^{x_2+t} \mathrm{d}x_1\, \mathrm{d}x_2 = \frac{(y+t)^2}{2}.$$

For $n > 1$, assume that (7) is true for $n - 1$. From (6) and (5), we obtain

$$\begin{aligned} q'_n(y,t) &= \int_{-nt}^{y} q'_{n-1}(x_{n+1} + t, t)\, \mathrm{d}x_{n+1} \\ &= \int_{-nt}^{y} \frac{(x_{n+1}+nt)^n}{n!}\, \mathrm{d}x_{n+1} \\ &= \frac{(y+nt)^{n+1}}{(n+1)!}, \end{aligned}$$

thereby proving the first part of the Lemma.

2. For $t = 1$, the following result was derived in (Kar et al., 2009).

$$q_n(1) = \frac{G^{(n+1)}(0)}{(n+1)!}$$

Hence, to prove (3), it suffices to show that

$$q_n(t) = ct^{n+1} \tag{8}$$

for some constant c that depends only on n and not on t. From the definition of q_n it follows that for the moving average case,

$$\begin{aligned} q_n(t) &= \mathrm{P}(X_1 + X_2 < t, X_2 + X_3 < t, \ldots, X_n + X_{n+1} < t) \\ &= \int_0^t \int_0^{t-x_{n+1}} \cdots \int_0^{t-x_3} \int_0^{t-x_2} \mathrm{d}x_1 \, \mathrm{d}x_2 \ldots \mathrm{d}x_n \, \mathrm{d}x_{n+1}. \end{aligned} \tag{9}$$

Let us define

$$q'_n(y,t) = \int_0^{t-y} \int_0^{t-x_{n+1}} \cdots \int_0^{t-x_3} \int_0^{t-x_2} \mathrm{d}x_1 \, \mathrm{d}x_2 \ldots \mathrm{d}x_n \, \mathrm{d}x_{n+1}. \tag{10}$$

Clearly, we can find $q_n(t)$ by substituting $y = 0$ in $q'_n(y,t)$. To prove (8), it suffices to show that $q'_n(y,t)$ is an $(n+1)^{\text{th}}$ order homogeneous polynomial in y and t, i.e.,

$$q'_n(y,t) = \sum_{j=0}^{n+1} c_j y^j t^{n+1-j}, \tag{11}$$

for some constants $c_1, c_2, \ldots, c_{n+1}$ that depend only on n and not on t. From (11), one can readily obtain $q_n(t)$ as follows

$$q_n(t) = q'_n(0,t) = c_0 t^{n+1},$$

thereby proving (8). We shall prove (11) by mathematical induction. The base case $q'_1(y,t)$ is readily verified as

$$q'_1(y,t) = \int_0^{t-y} \int_0^{t-x_2} \mathrm{d}x_1 \, \mathrm{d}x_2 = \frac{t^2 - y^2}{2},$$

which is a homogeneous polynomial of order 2. For $n > 1$, we assume that (11) is true for $n-1$, i.e., $q'_{n-1}(y,t)$ is an n^{th} degree homogeneous polynomial. From (10) and (11), we obtain

$$\begin{aligned} q'_n(y,t) &= \int_0^{t-y} q'_{n-1}(x_{n+1}, t) \, \mathrm{d}x_{n+1} \\ &= \int_0^{t-y} \left(\sum_{j=0}^{n} c_j x_{n+1}^j t^{n-j} \right) \mathrm{d}x_{n+1} \\ &= \sum_{j=0}^{n} c_j t^{n-j} \left(\int_0^{t-y} x_{n+1}^j \, \mathrm{d}x_{n+1} \right) \\ &= \sum_{j=0}^{n} c_j t^{n-j} \frac{(t-y)^{j+1}}{j+1}, \end{aligned}$$

which is a $(n+1)^{\text{th}}$ order homogeneous polynomial, thereby completing the induction. □

The following lemma relates values of $q_n(t)$ at two different points t and t' in the case of two-span moving sum algorithms. The result is applicable in general, as long as the distribution of X_i is symmetric.

Lemma 1. *Let $X_i, i \in \{1, 2, \ldots\}$ be i.i.d. random variables, the pdf being symmetric w.r.t. mean μ. Consider a two-span moving sum algorithm with weights c_0, c_1 and two thresholds t, t' such that $t + t' = 2\mu(c_0 + c_1)$. Then the survival probabilities at t and t' are related by*

$$q_n(t') = \sum_{m=0}^{n} (-1)^m Q_{n,m}(t), \qquad \text{where}$$

$$Q_{n,m}(t) = \left(\sum_{\substack{l_1, l_2, \cdots, l_m \geq 0 \\ \sum_k k l_k = m \\ l' = n-m+1-\sum l_k}} \binom{n-m+1}{l', l_1, \cdots, l_m} \prod_{k=1}^{m} q_k^{l_k}(t) \right). \tag{12}$$

In particular, when $n = 1, 2$, the above expression simplifies to the well known expressions $q_1(t') = 1 - q_1(t)$ and $q_2(t') = 1 - 2q_1(t) + q_2(t)$.

Proof. Recall that, in the moving average case we calculate $Y_i = c_1 X_i + c_0 X_{i+1}$. By definition

$$q_n(t') = \mathrm{P}\left[\cap_{i=1}^{n} \left(c_1 X_i + c_0 X_{i+1} < t'\right)\right] = \mathrm{P}\left[\cap_{i=1}^{n} \left(Y_i < t'\right)\right]. \tag{13}$$

Let us define the random variable $X_i' = 2\mu - X_i$. We note that X_i and X_i' have identical pdf, since

$$\begin{aligned} \mathrm{P}(X_i' < t) &= \mathrm{P}(2\mu - X_i < t) \\ &= \mathrm{P}(X_i > 2\mu - t) \\ &= \mathrm{P}(X_i < t), \end{aligned}$$

where the last equality is because X_i is symmetric w.r.t. mean μ. We also note that since $X_1, X_2, \ldots, X_n$ are mutually independent, so are $X_1', X_2', \ldots, X_n'$. Substituting X_i' for X_i in (13), we can thus write

$$\begin{aligned} q_n(t') &= \mathrm{P}\left[\cap_{i=1}^{n} \left(c_1 X_i' + c_0 X_{i+1}' < t'\right)\right] \\ &= \mathrm{P}\left[\cap_{i=1}^{n} \left(c_1(2\mu - X_i) + c_0(2\mu - X_{i+1}) < t'\right)\right] \\ &= \mathrm{P}\left[\cap_{i=1}^{n} \left(c_1 X_i + c_0 X_{i+1} > 2\mu(c_0 + c_1) - t'\right)\right] \\ &= \mathrm{P}\left[\cap_{i=1}^{n} \left(Y_i > t\right)\right]. \end{aligned} \tag{14}$$

In order to relate $q_n(t')$ to $q_n(t)$ it remains to apply the inclusion-exclusion principle (e.g. Stanley (1997)) to equation (14). Therefore,

$$\begin{aligned} q_n(t') &= 1 - \mathrm{P}[\cup_{i=1}^{n} (Y_i < t)] \\ &= 1 - \sum_{m=1}^{n} (-1)^{m-1} Q_{n,m}, \end{aligned} \tag{15}$$

where $Q_{n,m}$ denotes the sum of the probabilities associated with selecting m distinct events out of n in the union in equation (15). The selected events are some times contiguous and some times there are not. A collection of j contiguous events will result in a $q_j(t)$, and in $Q_{n,m}(t)$ we may have more than one such situation. In order to simplify $Q_{n,m}(t)$ we proceed as follow.

Define the index set $\mathcal{I} = \{1, 2, \ldots, n\}$. For a subset of indices (sorted for unique representation) $\mathcal{J} \subseteq \mathcal{I}$, define the probability $q_{\mathcal{J}}$ as follows

$$q_{\mathcal{J}}(t) = P\left[\bigcap_{j \in \mathcal{J}} (Y_j < t)\right].$$

This notation is not to be confused with $q_n(\cdot)$ where the subscript n has to be an integer. We shall see shortly that these notations are indeed related. For a contiguous subset $\mathcal{J}$, it readily follows that $q_{\mathcal{J}}(t) = q_{|\mathcal{J}|}(t)$. Any non-contiguous subset $\mathcal{J} \subseteq \mathcal{I}$ can be uniquely decomposed

into sorted, disjoint and contiguous subsets. In general, we can have as many as l_k sets of k contiguous indices, such that $l_k \geq 0$ for all $k = 1, 2, \ldots, m$, $\sum_{k=1}^{m} l_k \times k = m$. The vectors of random variables associated with distinct subsets of noncontiguous indices are independent, leading to the factorization

$$Q_{n,m}(t) = \left(\sum_{\substack{l_1, l_2, \ldots, l_m \geq 0 \\ \sum k l_k = m}} N(l_1, l_2, \ldots, l_m) \prod_{k=1}^{m} q_k^{l_k}(t) \right). \tag{16}$$

To compute $N(l_1, l_2, \ldots, l_m)$, let us consider this combinatorial problem involving *dots* and *bars*. Assume we have $n-m$ dots and L bars of m different colors (l_k each of same color). The bars are to be placed in the space between the dots or at the sides so that no two bars are together. We can chose the L places in $\binom{n-m+1}{L}$ ways. These L places can be filled by $l_1, l_2, \ldots, l_m$ identical bars of m different colors in $\binom{L}{l_1, l_2, \ldots, l_m}$ ways. Hence the number of choices are

$$\binom{n-m+1}{L} \binom{L}{l_1, l_2, \ldots, l_m} = \binom{n-m+1}{l', l_1, \ldots, l_m},$$

where $l' = n - m + 1 - L$. Once a choice is made, we replicate the k^{th} color bar k times, so that the total number of objects (dots and bars combined) are $n - m + \sum_{k=1}^{m} k l_k = n$. As per the order of appearance, we assign an index to each of these objects. Denote the sorted index set of the bars as $\mathcal{J}$. It is easily seen that decomposing $\mathcal{J}$ yields exactly l_k contiguous fragments of size k, for $1 \leq k \leq m$, and also that the two combinatorial problems are equivalent. Hence

$$N(l_1, l_2, \ldots, l_m) = \binom{n-m+1}{l', l_1, \ldots, l_m},$$

which, along with (16), completes the proof of (12). □

2.2. Exact Evaluation of ARL

In this section, we calculate the exact value of ARL for uniformly distributed random variables. Our results apply to the moving average and filtered derivative when the span is 2.

Let $X_i, i \in \{1, 2, \ldots\}$ be i.i.d. random variables, the pdf being symmetric w.r.t. mean μ. Consider a two-span moving sum algorithm with weights c_0, c_1 and two thresholds t, t' such that $t + t' = 2\mu(c_0 + c_1)$ and $L(t) < \infty$. Then, by Lemma 1,

$$\begin{aligned} L(t') &= 2 + \sum_{n=1}^{\infty} q_n(t') \\ &= 2 + \sum_{n=1}^{\infty} \sum_{m=0}^{n} (-1)^m Q_{n,m}(t) \\ &= 1 + \sum_{n=0}^{\infty} \sum_{m=0}^{n} (-1)^m Q_{n,m}(t), \end{aligned} \tag{17}$$

where the last equality follows by defining $Q_{0,0} \triangleq 1$.

Conjecture 1.

$$\sum_{n=0}^{\infty} \sum_{m=0}^{n} (-1)^m Q_{n,m}(t) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} (-1)^m Q_{n+m,m}(t) \tag{18}$$

Discussion. The above assertion follows from the rearrangement described in Figure 1. The series in (17) is only conditionally convergent. Despite that, we conjecture that the rearrangement preserves the limiting value. We shall derive Lemma 2 based on this assumption and verify the results for the uniform and normal random variables using simulation studies.

Next, we define by $\widetilde{q}(t)$, the limit of the alternating series

$$\widetilde{q}(t) = 1 - q_1(t) + q_2(t) - \cdots = 1 + \sum_{i=1}^{\infty} (-1)^i q_i(t). \tag{19}$$

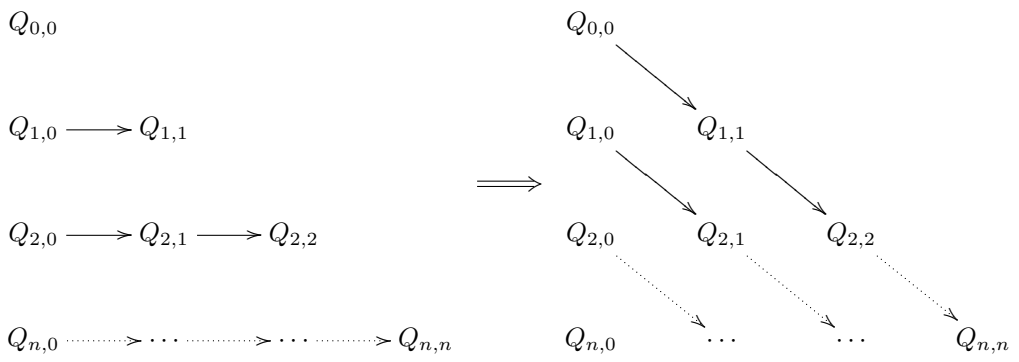


Figure 1: This rearrangement is conjectured to preserve the limiting value

We note that, by assumption, $L(t) = 2 + \sum_{i=1}^{\infty} q_i(t) < \infty$. Hence (19) converges absolutely and $\widetilde{q}(t)$ is well defined. It is easy to see that $0 \leq \widetilde{q}(t) \leq 1$. More generally, we conclude that $(\widetilde{q}(t))^n$ converges absolutely for $1 \leq n < \infty$, and hence any rearrangement of the multinomial expansion of $(\widetilde{q}(t))^n$ converges. In particular, we define

$$\mathcal{L} = \left\{ (l', l_1, l_2, \cdots, l_m) : \sum_{k=1}^{m} k l_k = m, l' = n - \sum l_k \right\},$$

and consider the following arrangement

$$\begin{aligned}
(\widetilde{q}(t))^n &= (1 - q_1(t) + q_2(t) - \cdots)^n \\
&= \sum_{m=0}^{\infty} \left(\sum_{\mathcal{L}} \binom{n}{l', l_1, \cdots, l_m} \prod_{k=1}^{m} \left((-1)^k q_k(t) \right)^{l_k} \right) \\
&= \sum_{m=0}^{\infty} \left(\sum_{\mathcal{L}} \binom{n}{l', l_1, \cdots, l_m} (-1)^{\sum k l_k} \prod_{k=1}^{m} q_k^{l_k}(t) \right) \\
&= \sum_{m=0}^{\infty} (-1)^m \left(\sum_{\mathcal{L}} \binom{n}{l', l_1, \cdots, l_m} \prod_{k=1}^{m} q_k^{l_k}(t) \right) \\
&= \sum_{m=0}^{\infty} (-1)^m Q_{n+m-1,m}, \qquad (20)
\end{aligned}$$

where the last step follows from the definition in (12). Lemma 2 expresses $L(t')$ in terms of $\widetilde{q}(t)$.

Lemma 2. *Let $X_i, i \in \{1, 2, \ldots\}$ be i.i.d. random variables, the pdf being symmetric w.r.t. mean μ. Consider a two-span moving sum algorithm with weights c_0, c_1 and two thresholds t, t' such that $t + t' = 2\mu(c_0 + c_1)$ and $L(t) < \infty$. Then*

$$L(t') = \frac{1}{1 - \widetilde{q}(t)}. \qquad (21)$$

Proof.

$$\begin{aligned}
L(t') &\overset{(a)}{=} 1 + \sum_{n=1}^{\infty} \sum_{m=0}^{\infty} (-1)^m Q_{n+m-1,m}(t) \\
&\overset{(b)}{=} 1 + \sum_{n=1}^{\infty} (\widetilde{q}(t))^n \\
&= \frac{1}{1 - \widetilde{q}(t)},
\end{aligned}$$

where (a) follows from (17) and (18) and (b) follows from (20), thereby completing the derivation of (21). □

The following result is the main contribution of this paper.

Theorem 2. *Let $X_1, X_2, \ldots X_n, \ldots$ be independent random variables, all uniformly distributed in $[0,1]$. Then,*

1. *For the two-span filtered derivative algorithm ($c_0 = 1, c_1 = -1$),*

$$L(t) = \begin{cases} 2 + \sum_{n=1}^{\lfloor |t|^{-1} \rfloor} \frac{(1+nt)^{n+1}}{(n+1)!} & \text{if } -1 < t < 0 \\ \exp(1) & \text{if } t = 0 \\ \left(\sum_{n=1}^{\lfloor t^{-1} \rfloor} (-1)^{n-1} \frac{(1-nt)^{n+1}}{(n+1)!} \right)^{-1} & \text{if } 0 < t < 1 \end{cases} \tag{22}$$

2. *For the two-span moving average algorithm ($c_0 = c_1 = 1$),*

$$L(t) = \begin{cases} \sec(t) + \tan(t) + 1 - t & \text{if } 0 < t \leq 1 \\ (\sec(2-t) - \tan(2-t) + 1 - t)^{-1} & \text{if } 1 < t < 2 \end{cases} \tag{23}$$

Proof. The results follow mostly from the application of Theorem 1 and Lemma 2 to the definition of ARL in (1). Note that the random variable X_i is symmetric w.r.t. mean $\mu = \frac{1}{2}$.

1. For $-1 < t < 0$, the result follows by substituting (2) in (1). The result for $t = 0$, which was obtained by Kar et al. (2009), can also be verified to be the limiting value of $L(t)$ from both directions. For $0 < t < 1$, note that $2\mu(c_0 + c_1) = 2 \cdot \frac{1}{2} \cdot (1 - 1) = 0$ and we define $t' = -t$ so that Lemma 2 can be applied. Also note that $-1 < t' < 0$ so that $q_n(t')$ can be calculated using (2). Hence $q_n(t') = (1 - nt)^{n+1}/(n + 1)!$ and the result follows by interchanging t, t' in (21).

2. For $0 < t < 1$, we proceed by substituting (3) in (1) and complete the derivation as follows

$$\begin{aligned} L(t) &= 2 + \sum_{i=1}^{\infty} t^{i+1} \frac{G^{(i+1)}(0)}{(i+1)!} \\ &= 2 + \sum_{i=2}^{\infty} t^{i} \frac{G^{(i)}(0)}{i!} \\ &= 2 - (G(0) + tG'(0)) + \sum_{i=0}^{\infty} t^{i} \frac{G^{(i)}(0)}{i!} \\ &\overset{(a)}{=} 2 - (G(0) + tG'(0)) + G(t) \\ &\overset{(b)}{=} 1 - t + G(t) \\ &= \sec(t) + \tan(t) + 1 - t, \end{aligned}$$

where (a) follows from the Taylor series expansion of $G(t)$ and (b) is because $G(0) = 1$ and $G'(0) = 1$, both of which can be easily verified from the definition of $G(\cdot)$. For $1 < t < 2$, note that $2\mu(c_0 + c_1) = 2 \cdot \frac{1}{2} \cdot (1 + 1) = 2$ and we define $t' = 2 - t$ so that Lemma 2 can be applied. Also note that $0 < t' < 1$ so that $q_n(t')$ can be calculated using (3). We proceed by interchanging t, t' in (21) and complete the derivation as follows

$$\begin{aligned} (L(t))^{-1} &= \sum_{i=1}^{\infty} (-1)^{i-1} q_i(t') \\ &= \sum_{i=1}^{\infty} (-1)^{i-1} (t')^{i+1} \frac{G^{(i+1)}(0)}{(i+1)!} \\ &= \sum_{i=1}^{\infty} (-t')^{i+1} \frac{G^{(i+1)}(0)}{(i+1)!} \\ &= -(G(0) - t'G'(0)) + \sum_{i=0}^{\infty} (-t')^{i} \frac{G^{(i)}(0)}{i!} \\ &\overset{(a)}{=} -(1 - t') + G(-t') \\ &= \sec(2-t) - \tan(2-t) + 1 - t, \end{aligned}$$

where (a) is because $G(0) = 1$, $G'(0) = 1$ and also because of the Taylor series expansion of $G(-t')$. □

3. Simulation studies

We have performed simulation studies that provide further evidence in support of Conjecture 1. We have run simulations for various thresholds and using uniform $\mathcal{U}[0,1]$ and normal $\mathcal{N}(0,1)$ random variables and displayed the results in Table 1. We have specified the thresholds as $p_1(t)$, i.e. the tail probability of the random variable Y_i. The Monte carlo estimate of ARL, denoted by $\widehat{L(t)}$, was obtained as the mean of 10^7 sample runs. For uniform random variables, we compare $\widehat{L(t)}$ with the closed form expressions derived in Theorem 2. For normal random variables, we compute $q_n(t)$ numerically using the technique in (Genz, 1992) and the software available from author's website. Since we can obtain only finite number of terms this way, we have chosen to compute $q_n(t)$ for $n = 1, 2, \ldots, 6$ only. We use the following formula $L_6 = (q_1 - q_2 + \cdots - q_6)^{-1}$. By Conjecture 1, $L_n(t)$ converges to $L(t)$ for large n. Since we use $n = 6$ only, for the approximation $L_6(t) \approx L(t)$ to be valid, $q_n(t)$ has to converge rapidly. This is the reason we have shown the results only for sufficiently high thresholds (small $p_1(t)$) for normally distributed case.

(a) ARL for uniform distributed random variables

	FD		MA	
$p_1(t)$	$\widehat{L(t)}$	$L(t)$	$\widehat{L(t)}$	$L(t)$
0.99	2.01	2.01	2.01	2.01
0.9	2.10	2.10	2.14	2.14
0.7	2.33	2.33	2.60	2.60
0.5	2.72	2.72	3.41	3.41
0.3	3.67	3.67	5.13	5.12
0.1	10.00	10.00	13.05	13.04
0.01	99.98	100.00	109.53	109.49
0.001	998.78	1000.00	1029.00	1029.87
0.0001	9905.75	10000.00	9988.02	10094.34

(b) ARL for normal distributed random variables

	FD		MA	
$p_1(t)$	$\widehat{L(t)}$	$L_6(t)$	$\widehat{L(t)}$	$L_6(t)$
0.1	10.07	10.07	13.64	13.64
0.01	99.96	100.00	113.98	114.05
0.001	998.70	1000.00	1055.55	1056.67
0.0001	9900.86	10000.00	10130.66	10238.03

Table 1: Comparison of Monte carlo ARL with those predicted by Conjecture 1.

4. Conclusion

In this paper, we have derived closed form expressions of the ARL for two-span moving sum algorithms. Though final expressions for ARL are only derived for uniformly distributed random variables, some intermediate theorems are applicable to a more general class of symmetric distributions. A likely future research direction is to investigate algorithms with higher span sizes ($k = 3, 4 \ldots$). Together, these expressions of ARL are likely to be useful to practitioners designing change detection algorithms for diverse applications.

References

Basseville, M., 1981. Edge detection using sequential methods for change in level-part 2: Sequential detection of change in mean. Acoustics, Speech and Signal Processing, IEEE Transactions on 29 (1), 32–50.

Basseville, M., Nikiforov, I. V., 1993. Detection of Abrupt Changes - Theory and Application. Prentice-Hall, New Jersey.

Bohm, W., Hackl, P., 1990. Improved bounds for the average run length of control charts based on finite weighted sums. The Annals of Statistics 18 (4), 1895–1899.

Genz, A., 1992. Numerical computation of multivariate normal probabilities. J. Comp. Graph Stat. 1, 141–149.

Kar, S., Mehrotra, K. G., Varshney, P. K., 2009. Approximation of average run length of moving sum algorithms using multivariate probabilities. (under review in Statistics and Probability Letters) [arXiv:0908.2954].

Lai, T. L., 1974. Control charts based on weighted sums. The Annals of Statistics 2 (1), 134–147.

Li, Q., Zheng, J., Tsai, A., Zhou, Q., 2002. Robust endpoint detection and energy normalization for real-time speech and speaker recognition. IEEE Trans. Speech Audio Process. 10 (3), 146–157.

Murphy, J. J., 1999. Technical analysis of the financial markets: A comprehensive guide to trading methods and applications. New York Institute of Finance, New York.

Roberts, S. W., 1959. Control charts based on geometric moving averages. Technometrics 1, 239–250.

Robinson, P. B., Ho, T. Y., 1978. Average run lengths of geometric moving average charts by numerical methods. Technometrics 20 (1), 85–93.

Stanley, R. P., 1997. Enumerative combinatorics 1. Cambridge University Press, Cambridge.

SAS/QC® 9.1 User's Guide, 2004. SAS Institute Inc., Cary, NC.