Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 53 (5) (2024), 1484 – 1496
DOI : 10.15672/hujms.1428934

Research Article

On the ARL of CUSUM in multinomial models

Shangchen Yao, Mohammad Kazim Khan*

Department of Mathematical Sciences, Kent State University, Kent, Ohio 44242, USA

Abstract

There is no known closed form expression for the average sample number, also known as average run length, of a multivariate CUSUM procedure $N = \min\{M_1, M_2, \cdots, M_m\}$ for $m \geq 3$, where M_i are univariate CUSUM procedures. The problem is generally considered to be hopelessly complicated for any model. In this paper, for the multinomial model we show, however, that there is a rather simple closed form expression for the average run length of N with an elementary proof. A bit surprisingly, we further show that the average run length of N is related to the average run lengths of M_i the same way as the capacitance of a series network of capacitors is related to the capacitances of its own components.

Mathematics Subject Classification (2020). 60G40, 62L15

Keywords. Average run length, multivariate CUSUM, Markov chain, sequential change point detection

1. Introduction

The cumulative sum (CUSUM) procedure of Page [28] is a univariate continually process monitoring algorithm when the information flow is observable sequentially. It is related to Wald's [40] sequential probability ratio test (SPRT), however, the main goal of the CUSUM procedure is to detect deviations in the process. The SPRT is, on the other hand, designed to test two competing classical statistical hypotheses in a sequential environment. There are several approaches in analyzing the univariate CUSUM procedures [28], such as solving the related integral equations, [15,16], or using the Wiener process approximations, [2, 13, 33] or martingale approach, [17–19]. See also [6, 9, 36, 37, 45] for a more detailed description and further references.

The univariate CUSUM procedure of Page [28] is known to have optimal properties. Lorden [21] used a minimax approach to show an asymptotic optimality of CUSUM. Moustakides [24] showed its optimality for the *iid* information flow and [25] extended the results for some dependent information flow. Ritov [34], and later Beibel [3] provided further results regarding the optimality of the CUSUM procedure in a Bayesian framework, see also [8, 24]. Moustakides [26] provided the corresponding optimality results of the CUSUM procedure in continuous time models. See [14, 30, 43] for further theory and diverse applications.

*Corresponding Author.

Email addresses: shangchenyao@hotmail.com (S. Yao), mkkhan@kent.edu (M. K. Khan)

Received: 31.01.2024; Accepted: 12.09.2024

The univariate CUSUM procedure, N_h, may be described very briefly as follows when $S_n = X_1+X_2+\cdots+X_n$ are the partial sums of an independent and identically distributed sequence of random variables,

$$N_h \ := \ \inf\{n \geq 1 : D_n \geq h\}, \qquad D_n := S_n - \min_{0\leq k\leq n} S_k, \quad S_0 = 0.$$

This stopping rule has the following link with the boundary crossing problem

$$N_h \ = \ \tau_{0,h} + N^* \, I(S_{\tau_{0,h}} \leq 0),$$

where $\tau_{0,h}$ is the boundary crossing stopping rule $\tau_{a,h} = \inf\{n \geq 1 : S_n \notin [a, h)\}$, with $a = 0$, and N^* is another identically distributed CUSUM stopping rule as N_h which is independent of $S_{\tau_{0,h}}$ given $\tau_{0,h}$.

The CUSUM procedure has found a large collection of applications besides its traditional usage in process monitoring and quality control. The monitoring process of the CUSUM stopping rule, D_n, is related to the ladder index concept of queuing theory, [31], as well as some other related fields such as insurance risk, dams, and data communication. The classical trading the line strategy of finance used for fast financial trading platforms can be analyzed by using the boundary crossing stopping rule τ, and its above link with the CUSUM procedure [1]. The Media Access Control (MAC) layer of communication systems contains a back-off protocol, should two clients approach the server exactly at the same time and cause a collision. There is a potential for client misbehavior. Cardenas et. al. [5] showed that the geometric model, if followed by the misbehaving client, leads to the most difficult detection case. Such protocols can be analyzed with the help of the CUSUM procedure as well, as we will describe in the examples at the end of the paper. For some potential applications from a Bayesian perspective see for instance [23].

The CUSUM procedure has been extended to multivariate settings as well [7, 11, 20, 29, 32, 35, 42]. One of the ways is by the use of the log-likelihood ratios, as is the case for setting up the SPRT [11]. This variety of the CUSUM procedure therefore falls into the category of primarily a parametric monitoring procedure, dependent on the underlying model assumptions that the user makes. However, by reducing the underlying model assumptions it can be made comparable to non-parametric setups. One of the main weaknesses of this form of constructing the stopping procedure is that it does not directly by itself identify the likely source of the deviations when this variety of CUSUM gets triggered. An alternative version of the multivariate CUSUM procedure uses a separate univariate CUSUM procedure and gets triggered when any one of the component univariate CUSUM procedures gets triggered [20, 42]. This version allows one to identify the source of the triggering stream. This procedure can also be used both in parametric as well as non-parametric frame works. The non-parametric version of this CUSUM procedure is basically the multinomial CUSUM, which is the subject matter of this paper.

Consider a sequence of independent observations of a m-dimensional multinomial experiment $\mathbf{Y}_n = (Y_{1n}, Y_{2n}, \cdots, Y_{mn})$, $n = 1, 2, \cdots$, with a common distribution, where $Y_{jn} \in \{0, 1\}$ and $\sum_{j=1}^{m} Y_{jn} \in \{0, 1\}$. The probabilities of success of face i being p_i and $p_1+p_2+\cdots+p_m < 1$. The tracking process of face j being $W_{jn} = \max(0, W_{j,n-1}+2Y_{jn}-1)$, $j = 1, 2, \cdots, m$. The starting value W_{j0}, a nonnegative integer value, of the tracking process is sometimes called the amount of head start [22]. The multinomial CUSUM stopping rule with respective head start values $\mathbf{i} := (i_1, i_2, \cdots, i_m)$ and triggering boundaries $\mathbf{h} = (h_1, h_2, \cdots, h_m)$ is defined as

$$N_{\mathbf{h}}^{\mathbf{i}} \ = \ \min\{M_{h_1,1}^{i_1}, M_{h_2,2}^{i_2}, \cdots, M_{h_m,1}^{i_m}\},$$

where $M_{h_j,j}^{i_j} = \inf\{n \geq 1 : W_{jn} \geq h_j\}$ and $W_{j0} = i_j$. When all the triggering boundaries are the same, we will denote the above stopping rule by using non-bold sub and super scripts, N_h^i. To date the average run length (ARL) of this basic CUSUM procedure is not known. The main focus of the paper, Theorem 2.1 below, is to show that this problem can have a closed form solution when all the trigger constants are the same. We further illustrate that our approach can lead to closed form results when the trigger constants differ, albeit with more complicated closed form expressions.

The next section presents the main result and its proof. Section three provides some examples and discussion. The paper concludes with a short summary in the last section.

2. The main result

We will use the Markov chain approach (see [4,41,44]) to obtain closed form expressions of the ARL for the multinomial CUSUM procedure.

Theorem 2.1. *For the m-dimensional CUSUM stopping rule under an m-dimensional multinomial model with corresponding probability vector $\mathbf{p} = (p_1, p_2, \cdots, p_m)$ and $p_1 + p_2 + \cdots + p_m \leq 1$, consider the CUSUM stopping rule $N_h^{\mathbf{i}}$, with the common triggering threshold h and started from the initial state $\mathbf{i} = (i_1, i_2, \cdots, i_m)$, where $0 \leq i_j < h$, and $j = 1, 2, \cdots, m$, and $i_1 + i_2 + \cdots + i_m < h$. Then the average run length is*

$$\begin{aligned} \mathbb{E}_{\mathbf{p}}\left(N_h^{\mathbf{i}}\right) &= \frac{\prod_{j=1}^m A_h(p_j) - \sum_{k=1}^m p_k^{h-i_k} A_{i_k}(p_k) \prod_{j\neq k}^m A_h(p_j)}{\sum_{k=1}^m p_k^h \prod_{j\neq k}^m A_h(p_j)}, \quad \textit{where} \\ A_h(p) &:= (1-p)A_{h-1}(p) + hp^{h-1} = \cdots \\ &= \frac{(1-p)^{h+1} - p^{h+1} - (h+1)(1-2p)p^h}{(1-2p)^2}, \qquad h \geq 1, \end{aligned}$$

and we take $A_0(p) := 0$. In particular, when $\mathbf{i} = \mathbf{0}$, and $p = p_1 = p_2 = \cdots = p_m$, we have the following link with the well known ARL, cf. [17, 27], of the one-dimensional (Bernoulli) CUSUM procedure M_h^0,

$$\mathbb{E}_p\left(N_h^{\mathbf{0}}\right) = \frac{1}{m}\mathbb{E}_p\left(M_h^0\right) = \frac{(1-p)^{h+1} - p^{h+1} - (h+1)(1-2p)p^h}{mp^h(1-2p)^2}. \tag{2.1}$$

For a common head start, $\mathbf{i} = (i, i, \cdots, i)$,

$$\mathbb{E}_p\left(N_h^{\mathbf{i}}\right) = \frac{1}{m}\left\{\mathbb{E}\left(M_h^0(p)\right) - m\mathbb{E}\left(M_i^0(p)\right)\right\}, \qquad p \leq 1/m.$$

Proof: Consider the Markov chain $\{\mathbf{W}_n, n \geq 0\}$ starting from the initial state $(i_1, i_2, \cdots, i_m)$. Any one of the states of the form $(i_1, i_2, \cdots, i_m)$ for which $0 \leq i_j \leq h$, such that $i_1 + i_2 + \cdots + i_m = h$ will trigger the multivariate CUSUM stopping rule, and we consider this set of states as the stopping state. To find the expected first passage times, $\mathbb{E}_{\mathbf{p}}\left(N_h^{\mathbf{i}}\right)$, for all non-stopping states, $(i_1, i_2, \cdots, i_m)$, we arrange these states in some order. Let $\mathbf{R}$ represent the transition probability matrix from any non-stopping state to another non-stopping state, and let $\mathbf{V}$ represent the corresponding vector of expected first passage times starting form the non-stopping states, also arranged in the same order. We need only verify that $\mathbf{V} = \mathbf{1} + \mathbf{VR}$. This system of linear equations can also be obtained by using the Chapman-Kolmogorov equations. It turns out to be easier to just directly verify our proposed solution. For Theorem (2.1) the structure of these equations can be divided into essentially two types. Those equations in which the starting state is "near" to the exiting state and the rest. The exiting states are of the type $(0, 0, \cdots, 0, h, 0, \cdots, 0)$, where h can be in any one of the m coordinates. The m states "near" to the exiting states are of the type $(0, 0, \cdots, 0, h-1, 0, \cdots, 0)$. Note that no state, which has more than one strictly positive entry, can be the one from which the Markov chain can exit in one step. This is due to the fact that all our starting states are assumed to have coordinates less than h and when one coordinate value goes up by 1 all the rest of the entries reduce by one without going below zero.

Since the vector $\mathbf{p} = (p_1, p_2, \cdots, p_m)$ will remain fixed, $\mathbb{E}$ will stand for $\mathbb{E}_{\mathbf{p}}$ from now on. The key idea of the proof can be explained for the simple case when the initial state is the origin. We need to verify that

$$\mathbb{E}\left(N_h^{0,0,\cdots,0}\right) = 1 + \left(1 - \sum_{k=1}^m p_k\right)\mathbb{E}\left(N_h^{0,0,\cdots,0}\right) + \sum_{k=1}^m p_k \mathbb{E}\left(N_h^{0,\cdots,0,1,0,\cdots,0}\right),$$

where, in the last term, the 1 appears in the k-th coordinate. Here if $h = 1$, the last term will become zero. Using the postulated result, the right hand side becomes

$$\begin{aligned} 1 + \left(1 - \sum_{k=1}^m p_k\right) & \frac{\prod_{j=1}^m A_h(p_j)}{\sum_{k=1}^m p_k^h \prod_{j\neq k}^m A_h(p_j)} + \sum_{k=1}^m p_k \left\{\frac{\prod_{j=1}^m A_h(p_j) - p_k^{h-1} A_1(p_k) \prod_{j\neq k}^m A_h(p_j)}{\sum_{k=1}^m p_k^h \prod_{j\neq k}^m A_h(p_j)}\right\} \\ &= 1 + \mathbb{E}\left(N_h^{0,0,\cdots,0}\right) - \sum_{k=1}^m \left\{\frac{p_k^h \prod_{j\neq k}^m A_h(p_j)}{\sum_{k=1}^m p_k^h \prod_{j\neq k}^m A_h(p_j)}\right\}, \end{aligned}$$

which is the postulated form of the left hand side. For other initial states the verification depends on the location of the initial state, and whether $h \leq m$ or $h > m$. We explain the idea when $h > m$ in the following first two cases. In the third case h may be less than m. First note that the ARL expression may be expressed as

$$\begin{aligned} \mathbb{E}\left(N_h^{i_1,\cdots,i_m}\right) &= \mathbb{E}(N_h) - \sum_{k=1}^m \frac{p_k^h A_{i_k}(p_k)}{p_k^{i_k} A_h(p_k)} \left(\frac{\prod_{j=1}^m A_h(p_j)}{\sum_{k=1}^m p_k^h \prod_{j\neq k}^m A_h(p_j)}\right) \\ &= \mathbb{E}(N_h) - \mathbb{E}(N_h) \sum_{k=1}^m \frac{p_k^h A_{i_k}(p_k)}{p_k^{i_k} A_h(p_k)}, \end{aligned}$$

where N_h represents $N_h^{0,0,\cdots,0}$. Its verification has three varieties. Case A. The initial state is a non-boundary state, i.e., all $0 < i_k < h-1$ for all k. For the first case, the Markovian property gives that

$$\begin{aligned} \mathbb{E}(N_h^{i_1,\cdots,i_m}) &= 1 + \left(1 - \sum_{k=1}^m p_k\right)\mathbb{E}(N_h^{i_1-1,\cdots,i_m-1}) \\ &\quad + \sum_{\ell=1}^m p_\ell \mathbb{E}\left(N_h^{i_1-1,\cdots,i_{\ell-1}-1,i_\ell+1,i_{\ell+1}-1,\cdots,i_m-1}\right). \end{aligned}$$

The last term of the right hand side becomes:

$$\begin{aligned} \mathbb{E}(N_h)\left(\sum_{\ell=1}^m p_\ell\right) &- \mathbb{E}(N_h)\sum_{\ell=1}^m p_\ell \left(\sum_{k=1}^m \frac{p_k^h A_{i_k-1}(p_k)}{p_k^{i_k-1} A_h(p_k)}\right. \\ &\quad \left. - \frac{p_\ell^h A_{i_\ell-1}(p_\ell)}{p_\ell^{i_\ell-1} A_h(p_\ell)} + \frac{p_\ell^h A_{i_\ell+1}(p_\ell)}{p_\ell^{i_\ell+1} A_h(p_\ell)}\right) \\ = \mathbb{E}(N_h)\left(\sum_{\ell=1}^m p_\ell\right) &- \mathbb{E}(N_h)\left(\sum_{\ell=1}^m p_\ell\right)\sum_{k=1}^m \frac{p_k^h A_{i_k-1}(p_k)}{p_k^{i_k-1} A_h(p_k)} \\ &+ \mathbb{E}(N_h)\sum_{\ell=1}^m \left(\frac{p_\ell^{h+2} A_{i_\ell-1}(p_\ell)}{p_\ell^{i_\ell} A_h(p_\ell)} - \frac{p_\ell^h A_{i_\ell+1}(p_\ell)}{p_\ell^{i_\ell} A_h(p_\ell)}\right). \end{aligned}$$

Also we have

$$\mathbb{E}\left(N_h^{i_1-1,\cdots,i_m-1}\right) = \mathbb{E}(N_h) - \mathbb{E}(N_h)\sum_{k=1}^m \frac{p_k^h A_{i_k-1}(p_k)}{p_k^{i_k-1} A_h(p_k)}.$$

Using these expressions, along with the fact that

$$p_k A_{i_k-1}(p_k) - A_{i_k}(p_k) - p_k^2 A_{i_k-1} + A_{i_k+1}(p_k) = p_k^{i_k},$$

and some simplification gives the postulated expression for $\mathbb{E}\left(N_h^{i_1,\cdots,i_m}\right)$.

Case B. The initial state is a boundary state, "away" from the stopping state, i.e., for some k, $i_k = 0$ and no $i_j = h-1$. The Markovian property gives that

$$\begin{aligned}\mathbb{E}(N_h^{i_1,\cdots,i_m}) &= 1+\left(1-\sum_{k=1}^m p_k\right)\mathbb{E}(N_h^{[i_1-1]^+,\cdots,[i_m-1]^+}) \\ &\quad +\sum_{\ell=1}^m p_\ell \mathbb{E}\left(N_h^{[i_1-1]^+,\cdots,[i_{\ell-1}-1]^+,i_\ell+1,[i_{\ell+1}-1]^+,\cdots,[i_m-1]^+}\right),\end{aligned}$$

where $[x]^+$ stands for x if $x>0$ and zero otherwise. The last term of the right hand side becomes:

$$\begin{aligned}&\mathbb{E}(N_h)\left(\sum_{\ell=1}^m p_\ell\right)-\mathbb{E}(N_h)\left(\sum_{\ell=1}^m p_\ell\right)\sum_{k=1}^m \frac{p_k^h A_{[i_k-1]^+}(p_k)}{p_k^{[i_k-1]^+}A_h(p_k)} \\ &+\mathbb{E}(N_h)\sum_{\ell=1}^m\left(\frac{p_\ell^{h+1}A_{[i_\ell-1]^+}(p_\ell)}{p_\ell^{[i_\ell-1]^+}A_h(p_\ell)}-\frac{p_\ell^h A_{i_\ell+1}(p_\ell)}{p_\ell^{i_\ell}A_h(p_\ell)}\right).\end{aligned}$$

The other term is

$$\mathbb{E}\left(N_h^{[i_1-1]^+,\cdots,[i_m-1]^+}\right) = \mathbb{E}(N_h)-\mathbb{E}(N_h)\sum_{k=1}^m\frac{p_k^h A_{[i_k-1]^+}(p_k)}{p_k^{[i_k-1]^+}A_h(p_k)}.$$

Plugging these into the last Markovian equation the verification holds provided the following equation holds.

$$1=\mathbb{E}(N_h)\sum_{k=1}^m\frac{p_k^h}{A_h(p_k)}\left\{\frac{A_{[i_k-1]^+}(p_k)}{p_k^{[i_k-1]^+}}-\frac{A_{i_k}(p_k)}{p_k^{i_k}}-\frac{p_k A_{[i_k-1]^+}}{p_k^{[i_k-1]^+}}+\frac{A_{i_k+1}(p_k)}{p_k^{i_k}}\right\}.$$

When an $i_k=0$, the expression in the curly braces becomes

$$\frac{A_{[i_k-1]^+}(p_k)}{p_k^{[i_k-1]^+}}-\frac{A_{i_k}(p_k)}{p_k^{i_k}}-\frac{p_k A_{[i_k-1]^+}}{p_k^{[i_k-1]^+}}+\frac{A_{i_k+1}(p_k)}{p_k^{i_k}} = 1.$$

Also, when $i_k>0$, the expression inside the curly braces becomes

$$\begin{aligned}&\frac{1}{p_k^{i_k}}\left\{p_k A_{i_k-1}(p_k)-A_{i_k}(p_k)-p_k^2 A_{i_k-1}+A_{i_k+1}(p_k)\right\} \\ &= \frac{1}{p_k^{i_k}}\left\{p_k(1-p_k)A_{i_k-1}(p_k)-A_{i_k}(p_k)+A_{i_k+1}(p_k)\right\} \\ &= \frac{1}{p_k^{i_k}}\left\{-A_{i_k+1}(p_k)+(i_k+1)p_k^{i_k}-i_k p_k^{i_k}+A_{i_k+1}(p_k)\right\} \\ &= 1,\end{aligned}$$

and the verification holds.

Case C. The initial state is a boundary state, "next to" the stopping state, i.e., $(0,0,\cdots,h-1,0,\cdots,0))$, where $h>1$. Without loss of generality consider $(h-1,0,\cdots,0))$, and the Markovian equation becomes

$$\begin{aligned}\mathbb{E}(N_h^{h-1,0,\cdots,0}) &= 1+\left(1-\sum_{k=1}^m p_k\right)\mathbb{E}(N_h^{h-2,0,\cdots,0}) \\ &\quad +\sum_{\ell=2}^m p_\ell\mathbb{E}\left(N_h^{h-2,0,\cdots,0,1,0,\cdots,0}\right),\end{aligned}$$

where in the last expression the exponent 1 is in the ℓ-th coordinate. The verification proceeds as in the previous two cases, except here we need to use the recursive property of $A_h(p)$, giving

$$p_1 - \frac{p_1 A_{h-1}(p_1)}{A_h(p_1)} \quad = \quad -\frac{p_1^2 A_{h-2}(p_1)}{A_h(p_1)} + \frac{p_1^3 A_{h-2}(p_1)}{A_h(p_1)} + \frac{p_1^h}{A_h(p_1)}.$$

This completes the proof. □

3. Discussion & examples

In this section we present several examples dealing with some applications. In particular, an electrical engineering analogy of the multinomial CUSUM with a series network of capacitors in a direct current, some extensions of our main theorem, a numerical example, and a comparison of the multinomial CUSUM with the likelihood ratio multivariate CUSUM procedure.

Example 3.1. Theorem (2.1) shows a somewhat surprising similarity with the total capacitance property of capacitors. If there are 2 (or more) capacitors connected in a series network with individual capacitances $\mathbb{E}_{p_1}(M_h^0)$ and $\mathbb{E}_{p_2}(M_h^0)$, then the total capacitance of the circuit is $\mathbb{E}_{p_1,p_2}(N_{h,h}^{0,0})$. This indicates that the multinomial CUSUM procedure can be made more sensitive so that it triggers quickly while the component CUSUM procedures are set to trigger less frequently to avoid false alarms. The triggering aspect may be thought of as an dielectric breakdown when the electric field exceeds the rated maximum. Just as capacitors need charging, as the CUSUM is started it goes into its charging mode. We may now reinterpret the idea of Lucas and Crosier [22] regarding giving a head start to CUSUM. In terms of the capacitors analogy it is essentially as if starting with a charged capacitor. Since the charge, $Q = CV$, where C is the capacitance and V is the voltage, for any fixed capacitance C, the charge is proportional to the voltage applied. In terms of the CUSUM, the trigger constant of a CUSUM procedure controls the "capacitance" aspect, the larger the constant the longer it will take to "charge the capacitor". The head start concept is essentially to start the procedure with an appropriately "charged" setting of the univariate CUSUM. The analog of a triggered CUSUM suggests a self destruction event of the capacitor as too high a voltage is applied to the capacitor. These analogies from electrical engineering of the multinomial CUSUM procedure may be helpful while tuning the multinomial CUSUM.

When the triggering constants start to differ with wider gaps the expressions for the ARL quickly become more complicated. The following results illustrate these aspects for the two dimensional (trinomial) models, starting with the adjacent triggering constants case, for which the analogous conclusions of Theorem (2.1) remain valid. We omit their completely analogous proofs.

Proposition 3.2. *For the* 2*-dimensional multinomial (trinomial) model and the* 2*-dimensional CUSUM stopping rule* $N_{h,h-1}^{i,j}(p_1, p_2)$*, which started from the state* (i, j)*, where* $0 \le i < h$ *and* $0 \le j < h-1$ *with* $i + j < h$*, we have*

$$\begin{aligned}
\mathbb{E}_{p_1,p_2}\left(N_{h,h-1}^{i,j}\right) &= \mathbb{E}_{p_1,p_2}\left(N_{h,h-1}^{0,0}\right)\left\{1 - \frac{p_1^h A_i(p_1)}{p_1^i A_h(p_1)} - \frac{p_2^{h-1} A_j(p_2)}{p_2^j A_{h-1}(p_2)}\right\}, \\
&= \mathbb{E}_{p_1,p_2}\left(N_{h,h-1}^{0,0}\right)\left\{1 - \frac{\mathbb{E}_{p_1}(M_i))}{\mathbb{E}_{p_1}(M_h)} - \frac{\mathbb{E}_{p_2}(M_j)}{\mathbb{E}_{p_2}(M_{h-1})}\right\}. \\
\mathbb{E}_{p_1,p_2}\left(N_{h,h-1}^{0,0}\right) &= \frac{A_h(p_1)\, A_{h-1}(p_2)}{p_1^h A_{h-1}(p_2) + p_2^{h-1} A_h(p_1)}, \qquad h \ge 2.
\end{aligned}$$

Again if there are 2 (or more) capacitors connected in a series network of direct current, with individual capacitances $\mathbb{E}_{p_1}(M_h^0)$ and $\mathbb{E}_{p_2}(M_{h-1}^0)$, then the total capacitance of the circuit is $\mathbb{E}_{p_1,p_2}(N_{h,h-1}^{0,0})$. The complexity of the expressions of ARL increases as the gap size between the triggering constants increases.

Proposition 3.3. *For the two dimensional CUSUM stopping rule in the trinomial model* $N_{h,h-2}^{i,j}(p_1,p_2)$*, where* $0 \le i < h,\ 0 \le j < h-2$ *with* $i+j < h$*, we have*

$$\mathbb{E}_{p_1,p_2}(N_{h,h-2}^{0,\,0}) = \frac{A_{h-2}(p_2)(a_{h-2}A_h(p_1) - (-1)^{\lfloor \frac{h-1}{2} \rfloor}(p_1p_2)^{h-2})}{a_{h-2}A_{h-2}(p_2)p_1^h + p_2^{h-2}(a_{h-2}A_h(p_1) - (-1)^{\lfloor \frac{h-1}{2} \rfloor}(p_1p_2)^{h-2})}$$

$$= \left(\frac{a_{h-2}p_1^h}{a_{h-2}A_h(p_1) - (-1)^{\lfloor \frac{h-1}{2} \rfloor}(p_1p_2)^{h-2}} + \frac{p_2^{h-2}}{A_{h-2}(p_2)} \right)^{-1}.$$

For the CUSUM with head start, if $i+j < h-1$*, the ARL,* $\mathbb{E}_{p_1,p_2}(N_{h,\,h-2}^{i,j})$*, is given by*

$$\mathbb{E}_{p_1,p_2}(N_{h,h-2}^{0,0}) \left(1 - \frac{a_{h-2}p_1^{h-i}A_i(p_1)}{a_{h-2}A_h(p_1) - (-1)^{\lfloor \frac{h-1}{2} \rfloor}(p_1p_2)^{h-2}} - \frac{p_2^{h-2-j}A_j(p_2)}{A_{h-2}(p_2)} \right).$$

For the case $i+j = h-1$*, the ARL,* $\mathbb{E}_{p_1,p_2}(N_{h,h-2}^{i,j})$*, is given by*

$$\mathbb{E}_{p_1,p_2}(N_{h,h-2}^{0,0}) \left(1 - \frac{a_{h-2}p_1^{h-i}A_i(p_1)}{a_{h-2}A_h(p_1) - (-1)^{\lfloor \frac{h-1}{2} \rfloor}(p_1p_2)^{h-2}} - \frac{p_2^{h-2-j}A_j(p_2)}{A_{h-2}(p_2)} \right.$$

$$\left. -(-1)^{\lfloor \frac{h-1}{2} \rfloor + \lfloor \frac{j-1}{2} \rfloor} \cdot \frac{a_j p_1^{h-1} p_2^{h-2-j}}{a_{h-2}A_h(p_1) - (-1)^{\lfloor \frac{h-1}{2} \rfloor}(p_1p_2)^{h-2}} \right)$$

where $a_0 = 0,\ a_1 = 1,\ a_n = p_1p_2a_{n-2} + (-1)^n a_{n-1},\ n \ge 3$*. Or in other words,*

$$\begin{aligned} a_{2k} &= \{(p_1p_2 - \frac{1}{2} + \frac{r}{2})^k - (p_1p_2 - \frac{1}{2} - \frac{r}{2})^k\}/r \\ a_{2k+1} &= \{(p_1p_2 - \frac{1}{2} + \frac{r}{2})^k(r-1) + (p_1p_2 - \frac{1}{2} - \frac{r}{2})^k(r+1)\}/(2r), \end{aligned}$$

and $r = (1 - 4p_1p_2)^{1/2}$*.*

The expression for the ARL when the trigger constants differ by three, analogous results can be derived, however, the complexity of the results increases. We omit the details. Also, using a result of Khan [20], we get

$$\frac{N_{\mathbf{h}}}{h} \overset{a.s.}{\to} \min\left\{ \frac{\beta_1}{\mathbb{E}(X_1)}, \cdots, \frac{\beta_m}{\mathbb{E}(X_m)} \right\}, \qquad \text{as } h \to \infty,$$

and also $\mathbb{E}(\frac{N_{\mathbf{h}}}{h})$ has the same limit, where $\beta_i = \lim_h \frac{h_i}{h}$, $i = 1, 2, \cdots, m$. The asymptotic approximations are of limited value since the trigger constants are usually rather small in practice. However, the larger the dimension, m, of the process being monitored, the larger the values of h_i can be deployed while keeping the chances of a false alarm in a manageable range.

Example 3.4. As a potential application of the multinomial CUSUM procedure, consider detection of a misbehaving client in the 802.11 computer network communication protocol, see van Holt and Huang [12]. Upon collision of two clients trying to access the tower, the two clients must back off and wait a uniformly distributed random amount of time chosen from $[0, w]$. In 802.11 protocol, the starting width is $w = 31$. Since this random number generation mechanism is in the possession of the client, there is a potential for misbehavior (quicker re-attempt to access the tower) and a resulting unfairly higher utilization of the resources. Monitoring the client only by the available information regarding its sequence of choices, gives rise to detecting any change in an underlying symmetric (fair) multinomial model of dimension w. The multinomial CUSUM may be used for "all sided" monitoring or "some special sides" monitoring. When a smaller subset, consisting of $m \leq w$ number of choices need to be monitored for their occurrence above and beyond their fair share, Theorem (2.1) can be used directly where the m+1-th category represents all the remaining categories. Their respective probabilities being $p_1, p_2, \cdots, p_m, p_{m+1}$ with $p_1 + p_2 + \cdots + p_{m+1} = 1$. For the "all sided" alternative, we may use the result of Theorem (2.1) after taking the limit as $p_{m+1} \to 0$. In the following we present some numeric performance statistics while taking small values of m for illustration purposes.

Consider the problem of sequentially testing if an $m+1$ sided die is fair. Now one may define

$$X_{k\ell} \ := \ 2Y_{k\ell} - 1, \qquad \ell = 1, 2, \cdots, m+1, \quad k = 1, 2, \cdots,$$

where $\mathbf{Y_k} = (Y_{k1}, \cdots, Y_{k,m+1})$ is a multinomial random vector representing the outcome of a role of the die, $Y_{k1} + \cdots + Y_{k,m+1} = 1$. Let $S_{n\ell} = S_{0\ell} + \sum_{k=1}^{n} X_{k\ell}$, $\ell = 1, 2, \cdots, m+1$ be the partial sum sequence where $S_{0\ell}$ is used to give a head start. To detect an "upward shift" for a specified subset of faces, say faces $1, 2, \cdots, m$, we may use the resulting multinomial CUSUM stopping rule $N_{\mathbf{h}}^{\mathbf{i}}$ with head start values $\mathbf{i} = (i_1, i_2, \cdots, i_m)$ and $\mathbf{p} = (p_1, \cdots, p_m)$. As an example consider the performance of a five sided die, leading to a multinomial CUSUM for detecting a departure from a uniform model. Our aim is to see which types of departures from the uniform model are more difficult to detect for $m = 4$ for which the five faces/categories are ordered. The null hypothesis is uniform, i.e., $\theta_{i0} = \frac{1}{5}$, $i = 1, 2, 3, 4, 5$. We will take

$$\max_{1 \leq i \leq m+1} |p_{i1} - p_{i0}| \ = \ \Delta,$$

to remain the same for the various alternatives. Table 1 provides the various alternatives that we will compare with.

Table 1. Various alternative hypotheses.

	Various Alternatives				
Name	Δ_1	Δ_2	Δ_3	Δ_4	Δ_5
L-Shape	Δ	-0.25Δ	-0.25Δ	-0.25Δ	-0.25Δ
Tent	-0.5Δ	0	Δ	0	-0.5Δ
Slope	$-\Delta$	-0.5Δ	0	0.5Δ	Δ
Dome	$-\Delta$	0.666Δ	0.668Δ	0.666Δ	$-\Delta$
Ramp	$-\Delta$	$-\Delta$	0	Δ	Δ

We use $h = h_1 = \cdots = h_5$ for various values of h. When h is chosen so that $\mathbb{E}_{H_0}(N_h)$ is sufficiently large so our ability of rejecting the null hypothesis wrongly (type I error) is made virtually impossible, we may compare the table entries $\mathbb{E}_{\Delta}(N_h)$ indicating which types of alternatives take longer time to detect by the multinomial CUSUM. Table 2 gives these expected values under the various alternatives.

Table 2. Average run lengths.

h	Fair	L-Shape	Tent	Slope	Dome	Ramp
3	27	23	23	21	19	18
4	112	78	75	65	59	52
5	453	228	218	184	171	134
6	1818	609	584	489	479	331
7	7279	1527	1479	1252	1328	795

It is clear that from these alternatives the L-Shape is the most difficult one for the multinomial CUSUM to detect.

Example 3.5. While monitoring the inner diameters of ball bearings, the recorded information is whether the diameter is within specification or larger or smaller than the specified range. In the latter case of out of specification range the component can be reworked to bring it back into compliance while in the former case the component is lost and has to be discarded. The i-th observable random vector (Y_{i1}, Y_{i2}) has a trinomial distribution indicating which type of out of specification may have occurred. One may monitor the process with a standard multivariate CUSUM which uses a transformation of the type,

$$X_k = \psi(Y_{k1}, Y_{k2}) := b_1 Y_{k1} + b_2 Y_{k2} - c, \qquad k = 1, 2, \cdots,$$

where b_1, b_2, c are chosen in some optimal way. For instance, when the null hypothesis is $p_1 \leq 0.05$ and $p_2 \leq 0.05$ and the alternative hypothesis is $p_1 = \pi_1 > 0.05$ or $p_2 = \pi_2 > 0.05$ for specified values π_1, π_2 we may use the likelihood ratio method to set the constants b_1, b_2, c. For the case when $\pi_1 = \pi_2 > 0.05$ the likelihood ratio method to determine ψ makes $b_1 = b_2$, and therefore, without loss of generality, can be taken to be any fixed positive value, which we will take to be 2. For our example we will take $c = 1$ since in this case an exact expression for the ARL, $\mathbb{E}(M_h^0)$, is available and is given by (2.1). Here $p = p_1 + p_2$, is the probability of observing X_k equal to 1, where p_1 is the probability of diameter being above the specification and p_2 being the probability of the diameter being below the specification.

The construction of a likelihood based univariate CUSUM depends on the alternative hypothesis. Therefore its optimality properties are guaranteed only in the framework of correctly identifying both the null and the alternative hypotheses. When this is not valid, by converting a multivariate version model into a univariate model, the transformation ψ may lose some information. Alternatively, we may use the results of the main theorem to run a two dimensional CUSUM procedure, $N_h^{i,j}$, by using

$$X_{k\ell} := 2Y_{k\ell} - 1, \qquad \ell = 1, 2, \quad k = 1, 2, \cdots,$$

Now i, j are the head start values of $\mathbf{W}_0$. Theorem (2.1) allows us to compare the performances of the two procedures, $\mathbb{E}_{p_1+p_2}(M_h)$ with $\mathbb{E}_{p_1,p_2}(N_h)$, when $p_1 \geq 0.05$ or $p_2 \geq 0.05$ after the values of h are chosen for the two procedures to give a large roughly equal ARL value under the null hypothesis. As the top left plot of Figure 1 shows, the two procedures give about the same ARL for $p_1 = p_2 = 0.05$ when we take $h = 9$ for M and $h = 7$ for N. The remaining three plots indicate that on average M_9 detects faster than N_7, when p_1, p_2, lie in a neighborhood of 0.05 with both p_1 and p_2 get larger than 0.05, as should be the case. Since the likelihood ratio principle will make $b_1 = b_2$ and hence the resulting procedure M_9 will be sensitive to the combined effects of the two types of shifts. N_7 performs better when one of the p_i, $i = 1, 2$ gets larger than 0.05. Besides the N_7 CUSUM gives one more benefit. It can identify the source of shift from its one dimensional components that triggered the stopping rule. The likelihood ratio CUSUM, however, needs further analysis.

As another application of this model, in randomized clinical trials context it is of interest to detect changes in the response variable that may take place during the trials. For example, De Leval et al. [10] describe the problem of detecting the success or failure of surgeries when the response variable has three levels: death, near miss or success. The 'near miss' situation indicates that certain serious complications occurred that had to be tackled for the recovery of the patient. Similar problems arise not only in various contexts in medical profession, [38, 39], but also in quality control and other disciplines. For instance, in quality control while monitoring the inner diameters of ball bearings, the recorded information is whether the diameter is within specification or larger or smaller than the specified range. In the latter case of out of specification range the component can be reworked to bring it back into compliance while in the former case the component

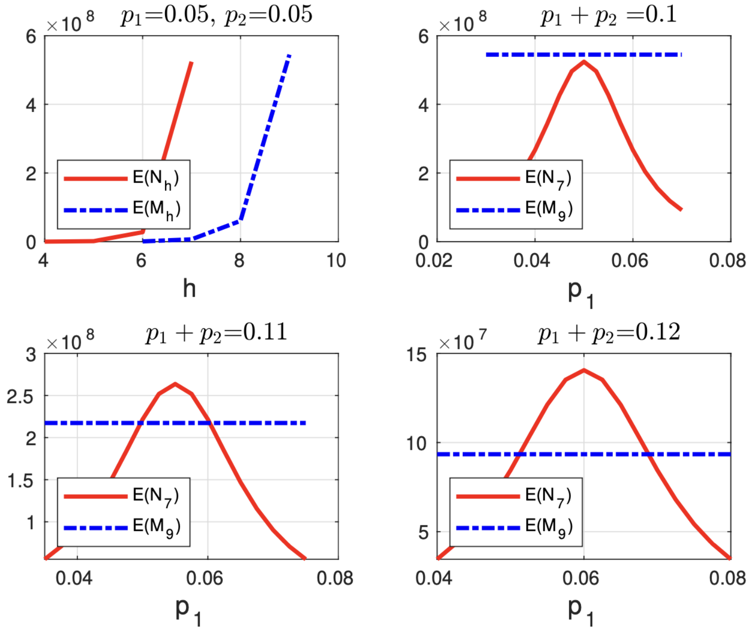

Figure 1. $E(M_9)$ versus $E(N_7)$ for Various Alternatives

is lost and has to be discarded. This sequential monitoring problem is essentially the same as in the above setup of De Leval et al. [10].

4. Summary & open problems

There are several varieties of multivariate CUSUM procedures and, unfortunately, not all are known to be optimal. The likelihood based version of the CUSUM procedure essentially turns the problem into a univariate version and hence the above cited optimality results remain valid. Even when the CUSUM (or another) procedure is optimal, and when it is applied to a wrong model it loses its optimality. Hence, they are sensitive to the parametric model assumption and it is paramount that the assumed model be accurate, which in real life is unlikely to be so. The likelihood ratio based, as well as some other varieties of multivariate CUSUM procedures, are parametric procedures. The multinomial CUSUM procedure, considered in this paper, is a non-parametric procedure. Also it may be used when the information flow is qualitative.

For the multinomial CUSUM procedure there is no known closed form expression for the average run length. The problem is generally assumed to be hopelessly intractable [20]. By providing a closed form expression for the ARL, the paper shows that this commonly held belief may have exceptions after all. Moreover, the paper provides a somewhat simple method of derivation based on a multivariate version of the well-known Markov chain approch, due to Brook and Evans [4].

The results show that the success of the approach used in this paper is dependent upon how close the various trigger constants are to each other. We show that when the trigger constants get farther apart the derivation becomes more complicated. We were unable to find an algorithm that can describe the ARL as the trigger constants start to differ by large amounts. A description of this dependence remains as an open problem.

The univariate CUSUM procedure is known to have rather large variance [1]. Another open problem is the derivation of the variance of the multinomial CUSUM and to find its relationship with its component univaraite CUSUMs that it is based on. Also, it would be interesting to know how the variance of the multinomial CUSUM depends on the dimension of the multinomial process. These are some of the areas where a thorough and large scale simulation study could shed some light on. Of course the continuous time analogs of results proved in this paper is another open question.

Acknowledgements

The authors would like to thank the referees and the editor for their constructive suggestions.

Author contributions. All the co-authors have contributed equally in all aspects of the preparation of this submission.

Conflict of interest statement. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding. This research was not funded by any source.

Data availability. No data was used for the research described in the article.

References

[1] V. Abramov, M.K. Khan and R.A. Khan, *A probablistic analysis of trading the line strategy*, Quant. Finance. **8**, 499-512, 2008.

[2] M. Bagshaw and R.A. Johnson, *The effect of serial correlation on the performance of CUSUM tests II*, Technometrics. **17**, 73-80, 1975.

[3] M. Beibel, *A note on Ritov's Bayes approach to the minimax property of the CUSUM procedure*, Ann. Statist. **24** (4), 1804-1812, 1996.

[4] D. Brook and D.A. Evans, *An approach to the probability distribution of CUSUM run length*, Biometrika. **59**, 539-549, 1972.

[5] A. Cardenas, S. Radosavac and S. Baras, *Performance comparison of detection schemes for MAC layer misbehavior*, Infocom, 2007, 26th IEEE International Conference on Computer Communications, IEEE, Anchorage, AK, 1496-1504, 2007.

[6] Y.S. Chow, H. Robbins and D. Siegmund, *The Theory of Optimal Stopping*, Houghton Miffin Co., 1971.

[7] R.B. Crosier, *Multivariate generalization of cumulative sum quality-control schemes*, Technometrics. **30**, 291-303, 1988.

[8] M. DeGroot, *Optimal Statistical Decisions*, McGraw-Hill Book Co., 1970.

[9] D.S. van Dobben de Bruyn, *Cumulative Sum Tests*, Griffin Publishers, London, UK, 1968.

[10] De Leval, M.R., Francois, K., Bull, C., Brawn, W.B., Spiegelhalter, D., *Analysis of a cluster of surgical failures: Application to a series of neonatal arterial switch operations*, J. Thorac. Cardiovasc. Surg. **107**, 914-924, 1994.

[11] J.D. Healy, *A note on multivariate CUSUM procedures*, Technometrics. **29** (4), 409-412, 1987.

[12] A. van Holt and C.Y. Huang, *802.11 Wireless Networks: Security and Analysis*, Springer, London UK, 2010.

[13] R.A. Johnson and M. Bagshaw, *The effect of serial correlation on the performance of CUSUM tests*, Technometrics. **16**, 103-112, 1974.

[14] E. Kaufmann and W.M. Koolen, *Mixture martingales revisited with applications to sequential tests and confidence intervals*, J. Mach. Learn. Res. **22** (246), 1-44, 2021.

[15] K. Kemp, *Formulae for calculating the operating characteristics and the Average Sample Number of some sequential tests*, J. R. Stat. Soc., B: Stat. Methodol. **20**, 379-386, 1958.

[16] K. Kemp *The average run length of the cumulative sum chart when a V-mask is used*, J. R. Stat. Soc., B: Stat. Methodol. **23**, 149-153, 1961.

[17] D.P. Kennedy, *Some martingales related to cumulative sum tests and single-server queues*, Stoch. Process. Their Appl. **4**, 261-269, 1976.

[18] R.A. Khan, *A note on Page's two-sided cumulative sum procedure*, Biometrika. **68**, 717-719, 1981.

[19] R.A. Khan, *On cumulative sum procedures and the SPRT with applications*, J. R. Stat. Soc., B: Stat. Methodol. **46**, 79-85, 1984.

[20] R.A. Khan, *Detecting changes in probabilities of a multi-component process*, Seq. Anal. **14**, 375-388, 1995.

[21] G. Lorden, *Procedures for reacting to a change in distribution*, Ann. Math. Statist. **42**, 1897-1908, 1971.

[22] J.M. Lucas and R.B. Crosier, *Fast initial response for CUSUM quality control schemes: give your CUSUM a head start*, Technometrics. **24**, 199-205, 1982.

[23] A.F. Martinez and R.H. Mena, *On a nonparametric change point detection model in Markovian regimes*, Bayesian Anal. **9**, 823-858, 2014.

[24] G.V. Moustakides, *Optimal stopping times for detecting changes in distributions*, Ann. Stat. **14**, 1379-1387, 1986.

[25] G.V. Moustakides, *Quickest detection of abrupt changes for a class of random processes*, IEEE Trans. Inform. Theory. **44**, 1965-1968, 1998.

[26] G.V. Moustakides, *Optimality of the CUSUM procedure in continuous time*, Ann. Stat. **32**, 302-315, 2004.

[27] A.G Munford, *A control chart based on cumulative scores*, Appl. Stat. **29**, 252-258, 1980.

[28] E.S. Page, *Continuous inspection schemes*, Biometrika. **41**, 100-115, 1954.

[29] J.J. Pignatiello and G.C. Runger, *Comparisons of multivariate CUSUM charts*, J. Qual. Technol. **22**, 173-186, 1990.

[30] H.V. Poor and O. Hadjiliadis, *Quickest Detection*, Cambridge University Press, 2009.

[31] N.U. Prabhu, *Stochastic Storage Systems: queues, insurance risk, dams, and data communication*, 2nd ed. Springer, New York, 2012.

[32] P. Qiu and D. Hawkins, *A rank based multivariate CUSUM procedure*, Technometrics. **43**, 120-132, 2001.

[33] M.R. Reynolds, *Approximations to the average run length in cumulative sum control charts*, Technometrics. **17**, 65-71, 1975.

[34] Y. Ritov, *Decision theoretic optimality of the CUSUM procedure*, Ann. Statist. **18** (3), 1464-1469, 1990.

[35] G.C. Runger and M. Testik, *Multivariate extensions to cumulative sum control charts*, Qual. Reliab. Engng. Int. **20**, 587-606, 2004.

[36] A.N. Shiryaev, *Optimal Stopping Rules*, Springer, Berlin, 2007.

[37] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals.* Springer-Verlag, New York, 1985.

[38] S.H. Steiner, R.J. Cook and V.T. Farewell, *Monitoring paired binary surgical outcomes using cumulative sum charts*, Statist. Med. **18**, 69-86, 1999.

[39] S.H. Steiner, P.L. Geyer and G.O. Wesolowsky, *Grouped data-sequential probability ratio tests and cumulative sum control charts*, Technometrics. **38**, 230-237, 1996.

[40] A. Wald, *Sequential Analysis*, John Wiley, New York, 1947.

[41] W.H. Woodall, *On the Markov chain approach to the two-sided CUSUM procedure*, Technometrics. **26**, 41-46, 1984.

[42] W.H. Woodall and M.M. Ncube, *Multivariate CUSUM quality-control procedures*, Technometrics. **27** (3), 285-292, 1985.

[43] L. Xie, S. Zou, Y. Xie and V.V. Veeravalli, *Sequential (quickest) change detection: Classical results and new directions*, IEEE J. Selected Areas in Information Theory. **2** 2, 494-514, 2021.

[44] S. Zacks, *The probability distribution and the expected value of a stopping variable associated with one-sided CUSUM procedures for non-negative integer valued random variables*, Commun. Stat. - Theory Methods **A10**, 2245-2258, 1981.

[45] S. Zacks, *Detection and change-point problems*, In: B. K. Ghosh, and P. K. Sen, (Eds), Handbook of Sequential Analysis. Marcel Dekker, New York, 531-562, 1991.