Data-efficient Quickest Change Detection with Unknown Post-change Distribution

Taposh Banerjee and Venugopal V. Veeravalli

Dept. of Electrical & Computer Eng.
University of Illinois, Urbana-Champaign
Urbana, USA
Email: banerje5,vvv@illinois.edu

Abstract—The problem of quickest change detection is studied, where there is an additional constraint on the cost of observations used before the change point and where the post-change distribution is unknown. An algorithm is proposed for the case where there are finite number of possibilities for the unknown post-change distribution. It is shown that if the post-change family of distributions satisfies some additional conditions, then the proposed algorithm is asymptotically optimal uniformly for all possible post-change distributions.

I. INTRODUCTION

The problem of detecting an abrupt change in the statistical properties of a measurement process is encountered in many engineering applications. Applications include detection of the appearance of a sudden fault/stress in a system being monitored, e.g., bridges, historical monuments, power grids, bird/animal habitats, etc. Often in these applications the decision making has to be done in real time, by taking measurements sequentially. In statistics this detection problem is formulated within the framework of quickest change detection (QCD) [1], [2].

In the QCD problem, the objective is to detect an abrupt change in the distribution of a sequence of random variables. The random variables follow a particular distribution in the beginning, and after an unknown point of time, follow another distribution. This problem is well studied in the literature [1], [3], [4]. The objective is to find a stopping time for the random variables so as to minimize a suitable metric on the average detection delay subject to a constraint on a suitable metric on the false alarm rate. When the pre- and post-change distributions are known, the optimal stopping rule is a single threshold test, where a sequence of statistics is computed using the likelihood ratio of the observations, and a change is declared the first time the sequence of statistics crosses a threshold. The threshold is chosen to meet the constraint on the false alarm rate.

In practice, often the post-change distribution is not known or known only to belong to a parametric family of distributions. Moreover, the change occurs rarely and it is of interest to constrain the number of observations used before the change point.

The problem of detecting a change when the post-change distribution is unknown has been well studied in the literature. One approach has been to modify the classical tests by computing the statistic at each time based on either the maximum over the parameter space of the likelihood ratios (GLRT), or based on a weighted average of the likelihood ratios (mixture based). Other approaches include robust formulations of the QCD problem or detection based on adaptive updates of the unknown parameter. See [1], [3] and [4] for a review.

In [5] and [6] we studied the classical QCD problems with an additional constraint on the cost of observations used before the change point. We called these formulations data-efficient quickest change detection (DE-QCD). We showed that two-threshold generalizations of the classical single-threshold QCD tests are asymptotically optimal for the proposed formulations.

In this paper we combine the ideas from [6] and from the QCD literature on unknown post-change distribution to study DE-QCD when the post-change distribution is unknown, but belongs to a finite set of distributions. We propose a modified version of the problem formulation in [6], and propose an algorithm called the MDE-CuSum algorithm. We show that when the post-change family of distributions satisfies some additional conditions, then the MDE-CuSum algorithm is asymptotically optimal for the proposed formulation, uniformly over all possible post-change distributions. The assumption that the post-change distribution belongs to a finite set of distributions is satisfied in many practical applications. For example, it is satisfied in the problem of detecting a power line outage in a power grid [2], or in a multi-channel scenario where the observations are vector valued and a change affects the distribution of only a subset of the components (each component for example may correspond to the output of a distinct sensor on a sensor board) [7], [8]. Also, see [9] for a possible scenario.

II. PROBLEM FORMULATION

A sequence of random variables \( \{X_n\} \) is being observed. Initially, the random variables are i.i.d. with p.d.f. \( f_0 \). At some unknown point of time—denoted by \( \gamma \) and called the change point in the following—the density of the random variables changes to \( f_\theta \). We assume that

\[
\theta \in \Theta = \{\theta_1, \theta_2, \cdots, \theta_M\}.
\]

Thus, we assume that the unknown post-change distribution belongs to a finite set of distributions. We denote by \( P_\gamma^0 \)
the underlying probability measure which governs such a sequence. We use \( \mathbb{E}_\gamma \) to denote the expectation with respect to this probability measure. We use \( \mathbb{P}_\infty (\mathbb{E}_\infty) \) to denote the probability measure (expectation) when change never occurs (i.e., the random variable \( X_n \) has p.d.f. \( f_0, \forall n \)). We wish to detect this change in distribution as quickly as possible subject to a constraint on the false alarm rate. In many applications the change occurs rarely, corresponding to a large \( \gamma \). As a result, we also wish to control the number of observations used for decision making before \( \gamma \). We are interested in the following stochastic optimization problem:

\[ \text{Problem 1:} \]

\[
\begin{align*}
\min_{\Psi} & \quad \text{CADD}\theta(\Psi) \\
\text{subj. to} & \quad \text{FAR}(\Psi) \leq \alpha, \\
\text{and} & \quad \text{PDC}(\Psi) \leq \beta,
\end{align*}
\]

where, \( 0 \leq \alpha, \beta \leq 1 \) are the given constraints.

Note that since \( S_n = 1 \) when \( X_n \) is used for decision making, the PDC metric is the average fraction of time observations are taken before \( \gamma \). We note that \( \text{PDC}(\Psi) \leq 1 \). As a result if \( \beta = 1 \), the above problem reduces to the classical minimax formulation of Pollak [10]. In the above problem we have implicitly restricted the search over those policies for which the CADD is well defined.

We are also interested in the problem where the CADD in Problem 1 is replaced by the following worst case average detection delay (WADD) metric of Lorden [11],

\[
\text{WADD}\theta(\Psi) := \sup_{\gamma \geq 1} \text{ess sup}_{\tau, \phi} \mathbb{E}_\tau(\theta^\tau | X_1, \cdots, X_{\gamma-1})
\]

where, \( x^+ := \max\{0, x\} \):

\[ \text{Problem 2:} \]

\[
\begin{align*}
\min_{\Psi} & \quad \text{WADD}\theta(\Psi) \\
\text{subj. to} & \quad \text{FAR}(\Psi) \leq \alpha, \\
\text{and} & \quad \text{PDC}(\Psi) \leq \beta,
\end{align*}
\]

where, \( 0 \leq \alpha, \beta \leq 1 \) are the given constraints. It can be shown that for any \( \Psi \) [1]

\[
\text{CADD}\theta(\Psi) \leq \text{WADD}\theta(\Psi).
\]

Our objective is to find an algorithm that is a solution to the above problems uniformly for each \( \theta \in \Theta \). However, it is not obvious if such a solution exists, even with \( \beta = 1 \). As a result we seek a solution that is asymptotically optimal, for a given \( \beta \), for each \( \theta \), as \( \alpha \to 0 \).

In the rest of the paper we use \( D(f_0 || f_0) \) and \( D(f_0 || f_0) \) to denote

\[
D(f_0 || f_0) := \mathbb{E}_1^\theta \left[ \log \frac{f_0(X_1)}{f_0(X_1)} \right],
\]

\[
D(f_0 || f_0) := -\mathbb{E}_\infty \left[ \log \frac{f_0(X_1)}{f_0(X_1)} \right].
\]

We assume throughout that both \( D(f_0 || f_0) \) and \( D(f_0 || f_0) \) are finite and positive. We also use \( f_0 \succeq g_0 \) as \( \alpha \to 0 \) to denote \( \lim_{\alpha \to 0} f_0/g_0 \leq 1 \). The notations \( f_0 \asymp g_0 \) and \( f_0 \asymp g_0 \) are similarly defined with \( \prec \) replaced with \( \succ \) and vice versa. We also use the following notation: \( x^+ := \max\{-h, x\} \).

III. BACKGROUND FROM QCD AND DE-QCD

In this section we review the results from [9] and [6] that are relevant to this paper. Let

\[
\Delta_\alpha := \{ \Psi : \text{FAR}(\Psi) \leq \alpha \}.
\]

When the post-change density is \( f_0 \), a universal lower bound on the CADD over the class \( \Delta_\alpha \) is given by [12]

\[
\inf_{\Psi \in \Delta_\alpha} \text{CADD}\theta(\Psi) \geq \frac{\log \alpha}{D(f_0 || f_0)} \text{ as } \alpha \to 0.
\]

By (5), this is a lower bound to the WADD as well.
If $\beta = 1$ and the post-change density is not known, or if the value of $\theta$ is not known, then the stopping rule that achieves the lower bound (6) is studied in [9]. The fundamental component of that stopping rule is the Cumulative Sum (CuSum) algorithm defined as follows [13]

\[
C_0(\theta) = 0,
\]

\[
C_n(\theta) = \left( C_{n-1}(\theta) + \log \frac{f_0(X_n)}{f_0(X_n)} \right) \quad \text{for } n \geq 1,
\]

\[
\tau_c(\theta) = \inf \{ n \geq 1 : C_n(\theta) \geq A \}.
\]

The algorithm that achieves the lower bound in (6) uniformly over $\theta$ is defined as follows: compute $C_n(\theta)$ for each $\theta \in \Theta = \{\theta_1, \ldots, \theta_M\}$ and stop at

\[
\tau_w = \inf \{ n \geq 1 : \max_{\theta \in \Theta} C_n(\theta) \geq A \}.
\]

We refer to this algorithm by MCuSum. The MCuSum algorithm can also be written as:

\[
\tau_w = \min_{\theta \in \Theta} \tau_c(\theta).
\]

It is a well known that when the post-change parameter is known, the CuSum algorithm is asymptotically optimal for both Problem 1 and Problem 2, with $\beta = 1$, as $\alpha \to 0$ [11].

Thus, to detect a change when the post-change parameter is unknown, $M$ CuSum algorithms are executed in parallel, one for each post-change parameter. A change is declared the first time a change is detected in any one of the CuSum algorithms. We note that the PDC of the MCuSum algorithm is equal to 1. The asymptotic optimality of the MCuSum algorithm is proved in [9]. Specifically, setting $A = \log M/\alpha$ ensures that \FAR{} and \CADD{} uniformly over $\theta$ are as follows: \FAR{} for each $\theta \in \Theta = \{\theta_1, \ldots, \theta_M\}$ and stop at

\[
\tau_w = \inf \{ n \geq 1 : \max_{\theta \in \Theta} C_n(\theta) \geq A \}.
\]

We define the ladder variable [14]

\[
\tau_w(\theta) = \inf \{ n \geq 1 : \sum_{k=1}^{n} \log f_0(X_k) < 0 \}.
\]

Then note that $W_{\tau_w}(\theta)$ is the ladder height. In the following theorem, we suppress the dependence on $\theta$ because it is known.

Theorem 3.1: When the post-change density $f_0$ is fixed and known, and $\mu > 0, \theta < \infty$, and $A = |\log \alpha|$, we have

\[
\text{FAR}(\tau_w) \leq \text{FAR}(\tau_c) \leq \alpha,
\]

\[
\text{PDC}(\tau_w) = E_{\infty}[\tau_w] + E_{\infty}[|W_{\tau_w}^h|/\mu].
\]

\[
\text{WADD}(\tau_w) \leq \text{WADD}(\tau_c) \leq \frac{|\log \alpha|}{\mu + D(f_0 || f_0)}.
\]

If $h = \infty$, then

\[
\text{PDC}(\tau_0) \leq \frac{\mu}{\mu + D(f_0 || f_0)}.
\]

Proof: The proof for the FAR and WADD analysis are identical to that provided in [6]. For the PDC the argument is similar to that provided in [6], and is based on the renewal reward theorem and the Wald’s lemma [14].

We note that the expression for the PDC is not a function of the threshold $A$. Also, for any given $h > 0$, the smaller the value of the parameter $\mu$, the smaller the PDC.

With $A = |\log \alpha|$ and $\mu$ and $h$ set to achieve the PDC constraint of $\beta$ (independent of the choice of $A$), the WADD of the MCuSum algorithm achieves the lower bound (6). Hence, we have from (5) that the algorithm is asymptotically optimal for both Problem 1 and Problem 2, for the given $\beta$, as $\alpha \to 0$. Thus, the pre-change observation control can be executed, i.e., any arbitrary but fixed fraction of samples can be dropped before change, without any loss in the asymptotic performance. See [6] for further remarks and numerical results.
IV. ALGORITHM FOR DE-QCD WHEN THE POST-CHANGE PARAMETER $\theta$ IS UNKNOWN

In this section we propose an algorithm for DE-QCD when the post-change distribution is not known, i.e., when $\beta < 1$ and $\theta$ is unknown. Specifically, we propose a modification of the MCuSum algorithm (8) to allow for observation control. We call this new algorithm the MDECuSum algorithm.

Without loss of generality we assume that:

Assumption 1:

$$D(f_{\theta_k} \mid f_0) \leq D(f_{\theta_k} \mid f_0) \forall k.$$  

We note that for $k \geq 1$,

$$E_{\theta_1}^{\theta_k} \left[ \log \frac{f_{\theta_k}(X_1)}{f_0(X_1)} \right] = D(f_{\theta_k} \mid f_0) - D(f_{\theta_k} \mid f_{\theta_1}).$$

We further make an important assumption.

Assumption 2: For every $k \geq 1$,

$$D(f_{\theta_k} \mid f_0) - D(f_{\theta_k} \mid f_{\theta_1}) > 0.$$  

Both the assumptions are satisfied for example when $f_{\theta_1}$ is the least favourable distribution in the family $\{f_{\theta} \}$. Specifically, if the law of $\log f_{\theta_1}(X_1)$ under $\{f_{\theta} \}$ is stochastically bounded by its law under $f_{\theta_1}$ (see Definition 1 in [15]), i.e.,

$$\mathbb{P}_{\theta_1} \left( \log \frac{f_{\theta_1}(X_1)}{f_0(X_1)} > x \right) \geq \mathbb{P}_{\theta_1} \left( \log \frac{f_{\theta_1}(X_1)}{f_0(X_1)} > x \right), \forall k, \forall x.$$  

The latter condition is satisfied for example when $f_{\theta_1} = N(0, 1)$ and $f_{\theta_0} = N(\theta_k, 1)$, with $0 < \theta_1 < \theta_2 < \cdots < \theta_M$.

We now propose the MDECuSum algorithm. Recall that in the MCuSum algorithm, $M$ CuSum algorithms are executed in parallel, one for each post-change parameter (see (9)). A change is declared the first time a change is detected in any of the $M$ CuSum algorithms. In the MDECuSum algorithm also, $M$ algorithms are executed in parallel, one for each post-change parameter, with the difference that the CuSum algorithm corresponding to the parameter $\theta = \theta_1$ is replaced by the DE-CuSum algorithm. Essentially, the post-change density closest to the pre-change density $f_0$ (in the KL divergence sense) is used for observation control, which by Assumption 1 is $f_{\theta_1}$.

Mathematically, the MDECuSum algorithm is described as follows. Fix $\mu > 0$ and $h \geq 0$,

$$V_0(\theta) = 0 \quad \forall \theta \in \Theta,$$

$$V_n(\theta_1) = \left( V_{n-1}(\theta_1) + \log \frac{f_{\theta_1}(X_n)}{f_0(X_n)} \right)^{\frac{1}{\mu}} \quad \text{if } V_{n-1}(\theta_1) \geq 0,$$

$$V_n(\theta_1) = \min \{ 0, V_{n-1}(\theta_1) + \mu \} \quad \text{if } V_{n-1}(\theta_1) < 0,$$

$$V_n(\theta_k) = \left( V_{n-1}(\theta_k) + \log \frac{f_{\theta_k}(X_n)}{f_{\theta_1}(X_n)} \right)^{\frac{1}{\mu}} \quad \text{for } k \geq 2, \text{ if } V_{n-1}(\theta_1) \geq 0,$$

$$V_n(\theta_k) = V_{n-1}(\theta_k) \quad \text{for } k \geq 2, \text{ if } V_{n-1}(\theta_1) < 0,$$

$$\tau_{\text{M}} = \inf \{ n \geq 1 : \max_{\theta \in \Theta} V_n(\theta) \geq A \}. \quad (15)$$

The evolution of the MDECuSum algorithm can be described as follows. In this algorithm $M$ sequences of statistics $\{V_n(\theta_k)\}_{k=1}^{M}$ are computed in parallel. While the entire vector of statistics is used to detect a change, only the statistic $V_n(\theta_1)$ is used for observation control. Specifically, the statistic $V_n(\theta_1)$ is updated using the DECuSum algorithm (12). For each $k \geq 2$, the statistics $V_n(\theta_k)$ is updated using the CuSum algorithm (7) with the difference that when $V_n(\theta_1) < 0$, the statistic is not updated and set to its value on the previous time instant. This is because when $V_n(\theta_1) < 0$, it is incremented by $\mu$ at each time instant, and observations are skipped till this statistic reaches 0 from below. In the absence of any new observation, the CuSum statistics $\{V_n(\theta_k)\}_{k=2}^{M}$ cannot be updated. In this algorithm, by design, while $V_n(\theta_1) < 0$, the CuSum statistics $\{V_n(\theta_k)\}_{k=2}^{M}$ are set to their value in the last time instant.

The Assumption 2 is critical to the working of this algorithm. By this assumption the mean of the log likelihood ratio between $f_{\theta_1}$ and $f_0$ is positive for every possible post-change distribution. This ensures that after the change occurs, and after a finite number of samples (irrespective of the threshold $A$), the DE-CuSum statistic $V_n(\theta_1)$ always remains positive and no more observations are skipped. This allows the statistic $V_n(\theta_1)$ corresponding to the actual post-change parameter $\theta$ to grow with the right 'slope'. If the Assumption 2 is violated, and the post-change parameter is $\theta_k \neq \theta_1$, then the statistic $V_n(\theta_k)$ will be below zero for a longer duration of time, and this time grows to infinity as the threshold $A \rightarrow \infty$. Thus, essentially, the growth of the CuSum statistic corresponding to $\theta_k$ will be intercepted by multiple sojourns of the statistic $V_n(\theta_1)$ below zero. As a result, the change will still be detected, but with a delay larger than the lower bound (6).

V. OPTIMALITY OF THE MDECuSUM ALGORITHM

The evolution of the MDECuSum algorithm is statistically identical to that of the MCuSum algorithm, except of the possible sojourns of the statistic $V_n(\theta_1)$ below 0. This fact will now be used to prove the asymptotic optimality of the MDECuSum algorithm. Let

$$\nu(\theta) = \inf \{ n \geq 1 : V_n(\theta) \geq A \},$$

the first time the statistic $V_n(\theta)$ crosses the threshold $A$. Note that $\nu(\theta_1)$ is statistical identical to the DECuSum algorithm $\tau_{\text{M}}(\theta_1)$.

Theorem 5.1: For any fixed $\mu > 0$ and $h < \infty$ and with $A = \log M/\alpha$ we have

$$\text{FAR}(\tau_{\text{M}}) \leq \text{FAR}(\tau_{\text{M}}) \leq \alpha,$$

$$\text{PDC}(\tau_{\text{M}}) = \text{PDC}(\nu(\theta_1)),$$

$$\text{WADD}(\tau_{\text{M}}) \leq \text{WADD}(\tau_{\text{M}}) \approx \frac{|\log \alpha|}{D(f_0 || f_0)} \quad (16)$$

as $\alpha \rightarrow 0$, for each $\theta \in \Theta$.

Proof: The FAR result follows because of (10) and the fact that the CuSum statistics $\{V_n(\theta_k)\}_{k=2}^{M}$ are set to their past values when the DECuSum statistic $\{V_n(\theta_1)\}$ is below 0. As a result, the mean time to false alarm for the MDECuSum algorithm is on an average equal to the corresponding time for
the MCuSum algorithm plus the average time for which the statistic \( \{V_n(\theta_1)\} \) is below 0.

The PDC result holds straightforwardly because the observation control is governed by the statistic \( \{V_n(\theta_1)\} \).

The evolution of the MDECUsum algorithm is statistically identical to that of the MCuSum algorithm, except for the possible sojourns of the statistic \( \{V_n(\theta_1)\} \) below 0. By Assumption 2, the mean of this sojourn time is finite after \( \gamma \). If \( X_1, \ldots, X_{\gamma-1} \) is such that \( V_{\gamma-1}(\theta_1) \geq 0 \), then the mean sojourn time below 0 of the statistic \( \{V_n(\theta_1)\} \) for \( n \geq \gamma \), is bounded from above by \[ \mathbb{E}^\theta_\gamma((\tau_{\mathrm{MC}} - \gamma)^+) | X_1, \ldots, X_{\gamma-1} \leq \frac{[h/\mu]}{\mathbb{P}^\theta_\gamma((\tau_{\mathrm{MC}} - \gamma) = \infty)} + \mathbb{E}^\theta_\gamma(\tau_{\mathrm{MC}} - 1), \]

where, we have also used (11). Since, this upper bound is not a function of \( \gamma \) and \( X_1, \ldots, X_{\gamma-1} \), the result follows from (10), (11), and the definition of WADD (4).

Thus, from (5), setting the threshold \( \Lambda \) as in the theorem above, and choosing the parameters \( \mu \) and \( h \) to set PDC \( \leq \beta \) ensures that the MDECUsum algorithm is asymptotically optimal for both Problem 1 and Problem 2, for the given \( \beta \), as \( \alpha \to 0 \).

VI. NUMERICAL RESULTS

In Fig. 1 we plot the CADD–FAR trade-off curves obtained using simulations for the MDECUsum algorithm (15), the MCuSum algorithm (8), and the fractional sampling scheme. In the latter the MCuSum algorithm is used and observations are skipped randomly, independent of the observation process. The simulation set used is: \( M = 4 \), \( f_0 = N(0, 1) \), \( f_\theta_1 = N(0.4, 1) \), \( f_\theta_2 = N(0.6, 1) \), \( f_\theta_3 = N(0.8, 1) \), \( f_\theta_4 = N(1, 1) \), \( \mu = 0.08 \) and \( h = \infty \). The post-change parameter is \( \theta = \theta_2 = 0.6 \), and the value of \( \mu \) is chosen using (14) and (16) to achieve a PDC of 0.5 (skipping/saving 50% of the samples). To achieve a PDC of 0.5 through the fractional sampling scheme, every alternate sample is skipped in the MCuSum algorithm. In the figure we see that skipping samples randomly results in a two-fold increase in delay as compared to that of the MCuSum algorithm. However, if we use the MDECUsum algorithm and use the state of the system to skip observations, then there is a small and constant penalty on the delay, as compared to the performance of the MCuSum algorithm. Thus, the MDECUsum algorithm provides a significant gain in performance as compared to the fractional sampling scheme.

VII. CONCLUSIONS AND FUTURE WORK

We have proposed an algorithm for DE-QCD when the post-change distribution is not known. We have shown that if the post-change family of distributions is finite, and has a member that is least favourable, then the MDECUsum algorithm proposed in this paper is asymptotically uniformly optimal, for each post-change parameter. An interesting problem for future work is to extend the results in this paper to the case when the parameter space is infinite.

VIII. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (NSF) under grant DMS 12-22498 and by the Defense Threat Reduction Agency (DTRA) under subcontract 147755 at the University of Illinois, Urbana-Champaign from prime award HDTRA1-10-1-0086. We thank Prof. Alexander Tartakovsky for useful discussions related to this paper.

REFERENCES