Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds

Odalric-Ambrym Maillard

Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds

Odalric-Ambrym Maillard ODALRIC.MAILLARD@INRIA.FR
Inria Lille – Nord Europe
SequeL team
40 Avenue Halley, Villeneuve d'Ascq

Editors: Aurélien Garivier and Satyen Kale

Abstract

We consider change-point detection in a fully sequential setup, when observations are received one by one and one must raise an alarm as early as possible after any change. We assume that both the change points and the distributions before and after the change are unknown. We consider the class of piecewise-constant mean processes with sub-Gaussian noise, and we target a detection strategy that is uniformly good on this class (this constrains the false alarm rate and detection delay). We introduce a novel tuning of the GLR test that takes here a simple form involving scan statistics, based on a novel sharp concentration inequality using an extension of the Laplace method for scan-statistics that holds doubly-uniformly in time. This also considerably simplifies the implementation of the test and analysis. We provide (perhaps surprisingly) the first fully non-asymptotic analysis of the detection delay of this test that matches the known existing asymptotic orders, with fully explicit numerical constants. Then, we extend this analysis to allow some changes that are not-detectable by any uniformly-good strategy (the number of observations before and after the change are too small for it to be detected by any such algorithm), and provide the first robust, finite-time analysis of the detection delay.

1. Change-point setups

Detecting a change of measure in a sequence of observations $Y_1, \dots Y_n$ is a classical problem that received a lot of attention from various areas of mathematical statistics, information theory and computer science over the past century. We refer to Basseville et al. (1993); Brodsky and Darkhovsky (1993); Jie and Gupta (2000); Tartakovsky (1991); Csörgö and Horváth (1997); Wu (2007) for classical textbooks on the topic, while the more recent references Blazek et al. (2001); Moustakides (2004); Bouzebda (2016); Vaswani (2005); Downey (2008); Khaleghi and Ryabko (2014); Frick et al. (2014); Garreau and Arlot (2016); Celisse et al. (2017) give a flavor of the increasing diversity of works in the field. One should however be careful as the same terminology encompasses different setups. In the batch setup, one has access to all the data points ahead of time and decision is output after seeing the full batch n of data. In this case, one cares about estimating the change point locations as precisely as possible using all points. In stark contrast, in a sequential setup data points are received one by one and one must raise an alarm as early as possible after any change τ, only based on past data. Hence, we cannot wait to see one or several other change points to estimate changes a posteriori (from this standpoint, all batch strategies suffer a huge detection delay) but must detect the latest change-point as fast as possible (hence detection and estimation are two different problems). A second crucial point is the available information that considerably modifies the achievable performance and proposed strategies: an important body of work has focused on the sce-

nario when the distributions before and after a change are perfectly known and only the time of the change is unknown (leading to CUSUM-like algorithms, see Page (1954); Shiryaev (1963); Roberts (1966); Lorden (1971); Shiryaev (1978); Pollak (1985)), while more recent works address the less informative setup of parametric distributions when parameters may be unknown before and after the change (Siegmund and Venkatraman (1995); Pollak and Siegmund (1991); Mei (2006, 2008); Lai and Xing (2010)). In all these setups, it is crucial to relate the magnitude of changes (e.g. difference of means in a change of mean setup) to the number of available observations between changes. We recall below the most emblematic change-point detection strategies.

CUSUM One of the most famous change-point detection algorithm is the CUSUM strategy from Page (1954) that is based on likelihood ratio thresholding: Assuming that $Y_1, \dots, Y_\tau$ is i.i.d. from the distribution p_0 and $Y_{\tau+1}, \dots, Y_n$ is i.i.d. from the distribution p_1, where both p_0 and p_1 are perfectly known and $\tau \in \mathbb{N}$ is the unknown change point, the original CUSUM change-point detection procedure takes a positive constant $c \in \mathbb{R}^+$ as input parameter and builds the following quantity:

$$\text{(CUSUM)} \quad \tau(c; p_0, p_1) = \min\Big\{t \in [1,n] : \max_{s\in[0,t)} L_{s:t} \geqslant c\Big\} \quad \text{where } L_{s:t} = \sum_{t'=s+1}^{t} \log \frac{p_1(Y_{t'})}{p_0(Y_{t'})}\,. \quad (1)$$

This quantity is a stopping time and enjoys nice theoretical properties: Let $\mathbb{E}_\tau$ and $\mathbb{P}_\tau$ denote the expectation and probability with respect to the process that changes from p_0 to p_1 at change-point $\tau + 1$. CUSUM minimizes the worst-case delay $\max_\tau \mathbb{E}_\tau(\widehat{\tau} - \tau | \widehat{\tau} \geqslant \tau)$ amongst all algorithms outputting $\widehat{\tau}$ for which $\mathbb{E}_0(\widehat{\tau}) = \mathbb{E}_0(\tau(c; p_0, p_1))$, see e.g. Blazek et al. (2001). On the other hand, this procedure is restricted to the case when p_0 and p_1 are known. The same criticism applies to the Shiryaev-Pollak stopping time $\min\{t \in [1,n] : \log \sum_{s=0}^{t-1} \exp(L_{s:t}) \geqslant c\}$.

GLR When p_0, p_1 are unknown, it is natural to replace the log-likelihood ratios with a generalized likelihood ratio (GLR). While initially introduced for the case when p_0 is known and p_1 is not, Lai and Xing (2010) extends the GLR to the case when both distributions are unknown, assuming they come from the same canonical exponential family. Namely, for the density model $p_\theta(y) = \exp(\theta^\top y - \psi(\theta))$ with log-partition function ψ defining the exponential family $\mathcal{E} = \{p_\theta : \psi(\theta) < \infty\}$ it writes

$$\text{(GLR)} \quad \tau_n(c; \mathcal{E}) \;=\; \min\Big\{t \in [1,n] : \max_{s\in[0,t)} G^{\mathcal{E}}_{1:s:t} \geqslant c\Big\} \quad (2)$$

$$\begin{aligned} \text{where} \quad G^{\mathcal{E}}_{t_0:s:t} &= \sup_{\theta_1,\theta_2} \sum_{t'=t_0}^{s} \log p_{\theta_1}(Y_{t'}) + \sum_{t'=s+1}^{t} \log p_{\theta_2}(Y_{t'}) - \sup_{\theta} \sum_{t'=t_0}^{t} \log p_\theta(Y_{t'}) \\ &= (s - t_0 + 1)\psi_\star(\mu_{t_0:s}) + (t-s)\psi_\star(\mu_{s+1:t}) - (t - t_0 + 1)\psi_\star(\mu_{t_0:t})\,, \end{aligned}$$

in which we introduced the empirical means and Fenchel-Legendre dual notations

$$\mu_{t':t} = \frac{1}{t - t' + 1} \sum_{s=t'}^{t} Y_s, \quad \psi_\star(\mu) = \sup_\theta \{\theta^\top \mu - \psi(\theta)\}\,.$$

Example 1 *For the family $\mathcal{N}_1 = \{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$ of standard univariate Gaussian distributions, the GLR statistics simplifies to*

$$G^{\mathcal{N}_1}_{t_0:s:t} = (s - t_0 + 1)(t - s)(\mu_{t_0,s} - \mu_{s+1,t})^2/(t - t_0 + 1)\,,$$

thus leading to the stopping time

$$(\mathcal{N}_1\text{-GLR}) \quad \tau_n(c; \mathcal{N}_1) = \min\Big\{t \in [1,n] : \max_{s\in[t_0,t)} \frac{(s-t_0+1)(t-s)}{t - t_0 + 1}(\mu_{t_0,s} - \mu_{s+1,t})^2 \geqslant c\Big\}\,. \quad (3)$$

Sequential setup The previous formulation is in the batch setup, however both CUSUM and GLR (and their variants) can be phrased in the sequential setup as well (see Downey (2008)). In this case, at time t, upon having observed $Y_{t_0+1}, \dots, Y_t$, an alert is raised according to the boolean test

$$\mathrm{CUSUM}(t_0, t) = \mathbb{I}\{\max_{s\in[t_0,t)} L_{s:t} \geqslant c\}, \quad \text{or to} \quad \mathrm{GLR}^{\mathcal{E}}(t_0, t) = \mathbb{I}\{\max_{s\in[t_0,t)} G^{\mathcal{E}}_{t_0:s:t} \geqslant c\}$$

where c may now depend on t. Here, we note that the observations $Y_{t'}$ for $t' \in [t+1, n]$ are not available at time t. Further, n is not assumed to be known (or is considered infinite). Hence, we want a strategy that is anytime in the sense it does not depend on the total number of the observations.

Delay and false alarms We measure the quality of a detection algorithm using the two following notions: First the *probability of false alarm*, that is of detecting a change at some time t while there is no change: For GLR this quantity is $\mathbb{P}(\exists t \in \mathbb{N} : \max_{s\in[t_0,t)} G^{\mathcal{E}}_{t_0:s:t} \geqslant c)$. Second the *detection delay*, that is the difference between the first time step when an algorithm detects a change and $\tau+1$. For GLR, this is the random variable $\tau_t(c, \mathcal{E}) - \tau - 1$ for $t > \tau$, that can be studied in expectation or high probability. A natural question is then how to choose the threshold c.

The classical literature only provides an asymptotic control of the expected delay (e.g. expressed for the limiting case when the probability of false alarm tends to 0). We show we can be more precise, by requesting non-asymptotic results that hold for each t, each δ and each τ. Further, unlike the classical literature that studies the expected detection delay, in this document, we seek a high probability control. We request sequential change-point detection procedures that are uniformly-good in the following sense:

Definition 1 (Uniformly-good detection strategies) *A sequential change-point detection strategy $\mathcal{A}$ is called δ-uniformly-good on a class of processes $\mathcal{D}$ if*

$$\textit{(False alarm)} \quad \forall \nu \in \mathcal{D},\ \textit{change-point } \tau, \quad \mathbb{P}_\nu\big(\exists t \in [t_0, \tau], \mathcal{A}(t_0, t) = 1\big) \leqslant \delta.$$

It is (Δ, δ)-uniformly-optimal if for a change of magnitude Δ, its detection delay $d_\nu(t_0, \tau + 1, \Delta, \delta)$ at probability level $1-\delta$ is minimal amongst the uniformly-good strategies. It is uniformly-optimal if (Δ, δ)-uniformly-optimal for all Δ, δ.

Remark 2 (Undetectable changes) *When the delay $d^\star_\nu(t_0, \tau + 1, \Delta, \delta)$ of a uniformly-optimal detection strategy is positive, this means any uniformly-good strategy will not detect (on an event of probability higher than $1-\delta$) a change that is happening at $\tau + 1$, with magnitude less that Δ: We naturally call such changes "undetectable".*

2. Outline and contributions

In this paper, we consider a sequential change-point detection problem and analyze the GLR strategy for the class sub-σ of distributions with σ-sub-Gaussian observation noise, that is we make the following mild assumption on the sequence $(Y_t)_t$ of real-valued observations

Assumption 1 (Sub-Gaussian observation noise) *The sequence $(Y_t)_t$ has σ-sub-Gaussian noise, meaning that*

$$\forall t, \forall \lambda \in \mathbb{R}, \quad \log \mathbb{E}[\lambda(Y_t - \mathbb{E}[Y_t])] \leqslant \frac{\lambda^2\sigma^2}{2}.$$

Bounded observations in $[0, 1]$ corresponds to $\sigma = 1/2$. For clarity, we further restrict to the case of change in the mean only and assume piecewise i.i.d. data (all our results extend trivially to piecewise predictable processes). In this case the magnitude Δ of a change is simply the absolute difference of means before and after a change.

In section 3, we first provide a refined concentration inequality on the scan-statistics of the GLR test (Theorem 4, Theorem 5) that improves on more classical bounds derived from applications of the Bonferroni inequality (aka union bound). This result is used to show that choosing the threshold $c = (1 + \frac{1}{t-t_0+1}) 2 \ln \left[\frac{2(t-t_0)\sqrt{t-t_0+2}}{\delta} \right]$ for a given confidence level $\delta \in (0,1)$ ensures GLR$^{\text{sub-}\sigma}$ is uniformly-good, with a fully explicit detection delay in finite time, that asymptotically matches the known lower-bounds on the detection delay.

In section 5, we extend these results to multiple change points and relax the classical assumption that all changes are abrupt enough to be detectable: We allow undetectable changes and analyze in this context the robustness of the detection guarantees when input observations are perturbed by undetectable change, see Theorem 11, and 12. Up to our knowledge, this is the first time such a robust, non-asymptotic analysis is obtained.

3. Concentration of scan-statistics and false alarm probability of GLR

In order to define the threshold c so as to control the false alarm probability of the GLR test in (3), we now study the concentration properties of the scan statistics $\mu_{t_0,s} - \mu_{s+1,t}$ uniformly over s and t.

Uniform confidence bounds Let $\mu_t = \frac{1}{t} \sum_{t'=1}^{t} Y_{t'}$ denote the empirical mean with t observations. We first recall the following uniform confidence bound that is obtained by an application of the Laplace method (method of mixtures for sub-Gaussian variables) in the i.i.d sub-Gaussian case.

Lemma 3 (Time-uniform concentration bound) *Under Assumption 1, the following holds*

$$\textit{(Laplace method)} \quad \mathbb{P}\left(\exists t \in \mathbb{N},\ \mu_t - \mathbb{E}[\mu_t] \geqslant \sigma \sqrt{\frac{2(1+\frac{1}{t}) \ln\left(\sqrt{t+1}/\delta\right)}{t}} \right) \leqslant \delta\,, \tag{4}$$

which we reprove in Appendix A for completeness. Note that this holds simultaneously over all t. We highlight the fact that this time-uniform concentration inequality is much sharper than what could be obtained for instance from a simple union bound (or even a peeling argument); we discuss this fact in more detail in Appendix B.

3.1. Doubly time uniform concentration of scan-statistics

In order to handle changes of the mean, it is natural to study the concentration of $\mu_{1:s} - \mu_{s+1:t}$. A simple way to achieve time-uniform confidence bounds for such quantities is to make use of uniform concentration inequalities for $\mu_{1:s}$ and $\mu_{s+1:t}$ separately, and combine them with a simple union bound. This leads to the bound $b_{t_0}^{\text{disjoint}}(s,t,\delta)$ given in the following Theorem 5. A better approach is to handle the concentration of the terms in $z_{1:s:t} := \mu_{1:s} - \mu_{s+1:t}$ jointly, since it is a sum of t independent random variables, s of which are σ/s-sub-Gaussian, and the others are $\sigma/(t-s)$-sub-Gaussian. This however requires to extend the Laplace method, which we do now:

Theorem 4 (Time-uniform joint concentration) *Under Assumption 1, for each $s \in \mathbb{N}$, $\delta \in [0,1]$,*

$$\textit{(Extended Laplace)} \quad \mathbb{P}\left(\exists t > s,\, z_{1:s:t} - \mathbb{E}[z_{1:s:t}] \geqslant \sigma \sqrt{\left(\frac{1}{s} + \frac{1}{t-s}\right)\left(1+\frac{1}{t}\right) 2 \ln\left(\sqrt{t+1}/\delta\right)} \right) \leqslant \delta\,. \tag{5}$$

We prove this result in Appendix A. This is non-trivial as the proof builds a quantity that is *not a super-martingale*, contrary to the proof of the Laplace method. Note this result is uniform in t, for a sum of t independent but not i.i.d. variables. We obtain the bound $b_{t_0}^{\text{disjoint}}(s,t,\delta)$ in the following Theorem 5 upon using some additional union bound over s:

Theorem 5 (Doubly-time-uniform concentration) *Let $Y_1, \dots Y_t$ be a sequence of t independent real-valued random variables satisfying Assumption 1. Let $\mu_{t_1+1:t_2} = \frac{1}{t_2-t_1} \sum_{s=t_1+1}^{t_2} Y_s$ be the empirical mean estimate on the time interval $[t_1+1, t_2]$. Then, for each $t_0 \in \mathbb{N}_\star$, for all $\delta \in (0,1)$,*

$$\mathbb{P}\Big(\exists t\in\mathbb{N}_\star,\, s\in[t_0:t),\, \big|\mu_{t_0:s}-\mu_{s+1:t}-\mathbb{E}[\mu_{t_0:s}-\mu_{s+1:t}]\big|\geqslant b_{t_0}(s,t,\delta)\Big)\leqslant\delta \quad \textit{where } b_{t_0} \textit{ is either}$$

$$b^{\textit{disjoint}}_{t_0}(s,t,\delta) = \sqrt{2}\sigma\left[\sqrt{\frac{1+\frac{1}{s-t_0+1}}{s-t_0+1}\ln\left[\frac{4\sqrt{s-t_0+2}}{\delta}\right]}+\sqrt{\frac{1+\frac{1}{t-s}}{t-s}\ln\left[\frac{4(t-t_0)\sqrt{t-s+1}}{\delta}\right]}\right] \textit{ or}$$

$$b^{\textit{joint}}_{t_0}(s,t,\delta) = \sigma\sqrt{\Big(\frac{1}{s-t_0+1}+\frac{1}{t-s}\Big)\Big(1+\frac{1}{t-t_0+1}\Big)2\ln\left[\frac{2(t-t_0)\sqrt{t-t_0+2}}{\delta}\right]}\,.$$

In the sequel, we choose $b_{t_0}(s,t,\delta)=b^{\text{joint}}_{t_0}(s,t,\delta)$ as it is generally tighter than $b^{\text{disjoint}}_{t_0}(s,t,\delta)$.

4. Non-asymptotic detection delay of non-parametric GLR

We now make use of the confidence bounds in order to tune the GLR change-point detection procedure in the sub-Gaussian setting, that we define now. The next result (main result of this section) bounds its detection delay.

$$\text{GLR}^{\text{sub-}\sigma}(t_0,t)=\mathbb{I}\{\exists s\in[t_0:t):\left|\mu_{t_0:s}-\mu_{s+1:t}\right|\geqslant b_{t_0}(s,t,\delta)\}$$

Theorem 6 (Detection delay) *Let $Y_{t_0},\dots Y_\tau$ be a sequence of τ i.i.d. real-valued random variables with mean μ_1 and $Y_{\tau+1},\dots Y_t$ be a sequence of $t-\tau$ i.i.d. real-valued random variables with mean μ_2, both satisfying Assumption 1. Let $\Delta=|\mu_2-\mu_1|$, $\delta\in(0,1)$. Then, the procedure GLR$^{\text{sub-}\sigma}$ started at time t_0 and using $b^{\textit{joint}}_{t_0}(s,t,\delta)$ for each time t satisfies:*
(i) With probability higher than $1-\delta$, no false alarm occurs on the time interval $[t_0,\tau]$.
(ii) For all $\varepsilon\in[0,1]$, the change point occurring at $\tau+1$ is detected at time $t=\tau+1+d_\varepsilon(t_0,\tau+1,\Delta)$ (hence with delay not exceeding $d(t_0,\tau+1,\Delta)$) where

$$\textit{(Delay)}\quad d_\varepsilon(t_0,\tau+1,\Delta)=\min\left\{d'\in\mathbb{N}:\, d'>\frac{2(1+\varepsilon)^2\sigma^2(1+\frac{1}{\tau-t_0+1})\ln\left[\frac{2x_{d'}}{\delta}\right]}{\left(\Delta^2-\frac{2(1+\varepsilon)^2\sigma^2}{\tau-t_0+1}\ln\left[\frac{2x_{d'}}{\delta}\right]\right)_+}-1\right\},$$

where we introduced the short-hand notation $x_d=(d+\tau-t_0+1)\sqrt{d+\tau-t_0+3}$, and $(x)_+=\max\{x,0\}$, with probability higher than $1-\delta_t(\varepsilon)$ where

$$\delta_t(\varepsilon)=2(t-t_0)\left(\frac{\delta}{2(t-t_0)\sqrt{t-t_0+2}}\right)^{\varepsilon^2(1+\frac{1}{t-t_0+1})} \textit{ and } \delta_t(1)=\delta\,.$$

For illustration, Figure 1 depicts the delay function $d_1(t_0,\tau+1,\Delta)$ for $\sigma=0.5$ and $\delta=0.05$.
(iii) if $\tau=\tau_c$ is undetectable (see Remark 2) in the sense that no algorithm can detect the change before time τ_{c+1} using only data from time $[\tau_{c-1}+1,\tau_{c+1}]$, where $t_0-1=\tau_{c-1}<\tau_c<\tau_{c+1}$, then the gap must be of magnitude $\Delta\leqslant\overline{\Delta}(\tau_{c-1}+1,\tau_c+1,\tau_{c+1})$ where

$$\textit{(Gap)}\quad \overline{\Delta}(t_0,\tau+1,t)=\sigma\sqrt{\frac{(t-t_0+2)}{(t-\tau)(\tau-t_0+1)}8\ln\left[\frac{2(t-t_0)\sqrt{t-t_0+2}}{\delta}\right]}\,.$$

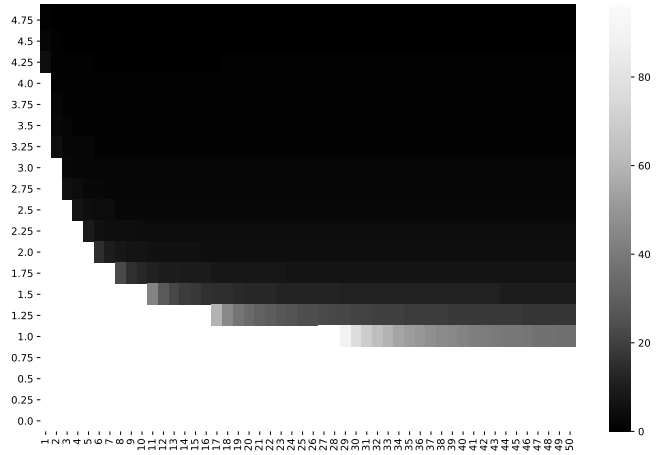

Figure 1: Detection delay as a function of the change time τ (x axis) and gap Δ (y axis).

Discussion about asymptotic optimality We compare this result with the existing lower-bound on the detection delay, see e.g. Theorem 3.1 in Lai and Xing (2010), that are stated in the asymptotic regime when $\frac{\tau-t_0+1}{\log(1/\delta)} \to \infty$ but $\log(\tau - t_0 + 1) = o(\log(1/\delta))$ when $\delta \to 0$ (note that this requires $\tau - t_0 + 1 \to \infty$). Such a regime may look a bit odd at first sight, since in our sequential setup both δ and the change points are fixed. It can be considered as the regime of the "easy case", when we have a lot of observations before a change occurs. For that reason it enables to disregard the effect of undetected change points and perhaps gives a simplified (but less informative) result. In this easy regime, then

$$\begin{aligned} d_\varepsilon(t_0, \tau+1, \Delta) &= \frac{(1+\varepsilon)^2 2\sigma^2 \ln\left[\frac{2x_d}{\delta}\right]}{\Delta^2} + \frac{4(1+\varepsilon)^4\sigma^4}{\tau-t_0+1} \ln^2\left[\frac{2x_d}{\delta}\right] + o(1) \\ &= \frac{(1+\varepsilon)^2 2\sigma^2 \ln\left[\frac{2x_d}{\delta}\right]}{\Delta^2} + o(\ln(1/\delta))\,. \end{aligned}$$

Note that the term $\Delta^2/2\sigma^2$ coincides with the Kullback-Leibler divergence of Gaussian distributions; while for bounded distributions in $[0,1]$, it becomes $2\Delta^2$, which also relates to the Kullback-Leibler divergence by Pinsker inequality. Further in this situation, the condition $\log(\tau - t_0 + 1) = o(\log(1/\delta))$ ensures that for all ε, $\delta_t(\varepsilon) \to 0$ as $\delta \to 0$, hence that detection occurs, and that

$$d_\varepsilon(t_0, \tau+1, \Delta) = \left(\frac{(1+\varepsilon)^2}{2\Delta^2} + o(1)\right) \log(1/\delta) \text{ as } \delta \to 0.$$

This shows that the detection delay of GLR$^{\text{sub-}\sigma}$ is also asymptotically optimal in the sense of Lai and Xing (2010) (up to Pinsker inequality tightness; it is exactly order optimal for Gaussian distributions). Now, Theorem 6 is more precise as it provides a fully explicit and non-asymptotic bound valid for each δ, Δ, τ. In particular, it enables to understand what happens not only in the (easy) asymptotic regime but also beyond that, in all and difficult situations. Up to our knowledge, it is surprisingly the first time such precise bounds on the detection delay are obtained.

Remark 7 (Scaling) *It is intuitive that the detection delay may not be bounded for change points of too small magnitude (difficult case). More precisely, taking $t_0 = 1$ in Theorem 6 shows that if the number of observations τ before the change point and the magnitude of the change point Δ are such that $\Delta < \sigma\sqrt{\frac{8}{\tau}\ln\left(2\tau\sqrt{\tau+2}/\delta\right)} = \tilde{O}(\frac{\sigma}{\sqrt{\tau}})$, then no change is detected (the detection delay is infinite). Now for larger Δ, Theorem 6 shows that the delay of the detection scales essentially as $O(\frac{\sigma^2}{\Delta^2}\ln(\tau + \frac{\sigma}{\Delta}))$.*

Remark 8 (Other work) *Theorem 6 is also coherent with the analysis from (Garreau and Arlot, 2016, Theorem 3.1) albeit in the fairly different batch setting with kernels. Now, Theorem 6 improves the constants and log scalings thanks to the Laplace method.*

5. Robustness to undetectable changes

While Theorem 6 applies to the ideal situation when at most one single change point occurs between time t_0 and time t, we now address the case when multiple (possibly undetectable) changes occur. Note that undetectability may alter the theoretical guarantees of a detection strategy obtained in the pure situation with a single change; hence this extension is non trivial.

To this end, we first need a few more definitions related to the notion of undetectable changes. Indeed detectability of a change point $(\tau+1, \Delta)$ not only depends on the change point but also on the number of available observations from the previous change point location and to the next one, hence is linked to the change-point sequence $(\tau_c)_c$.

Definition 9 (Detectable change points) *A change point τ_c is undetectable w.r.t the sequence $(\tau_c)_c$ if no uniformly-good algorithm can detect the change (w.p. $1-\delta$) before time τ_{c+1} using only data from time $[\tau_{c-1}+1, \tau_{c+1}]$.*

A change point τ_c is (ε, d)-detectable w.r.t the sequence $(\tau_c)_c$ for the delay detection function d and fraction $\varepsilon \in [0,1]$ if

$$d(\tau_{c-1}+1+\varepsilon\ell_{c-1}, \tau_c+1, \Delta_c) < \ell_c \quad \text{where } \ell_{c-1} = \tau_c - \tau_{c-1}\,.$$

That is, using only a fraction $1-\varepsilon$ of the ℓ_{c-1}-many observations available between the previous change point and τ_c+1, the change can be detected at a time not exceeding the next change point.

Remark 10 (Interpretation) *The term ε is a slack variable. For $\varepsilon = 0$, we recover that when starting at the previous change point, the current change point must be detected before the next one. The smaller ε, the smaller the starting time $\tau_{c-1}+1+\varepsilon\ell_{c-1}$, thus the larger the available number of observations before the change, and the smaller the detection delay.*

Instead of restricting to the ideal but rather unrealistic situation when all change points are detectable (which is the scope of the great majority of works on this topic), in this paper, we now investigate the case when some change points are not. These undetectable change points obviously create some perturbations and without further assumption it may be hard to make use of the procedure GLR$^{\text{sub-}\sigma}$ in this context. We now provide two approaches in order to handle the case when undetectable change points occur. The first approach is to modify the procedure so as to improve its robustness to undetectable changes, which however comes at a higher computational price. The second approach is to keep same the procedure but impose further restrictions (see Assumption 2) on the change points that are not detectable.

The basic idea behind the robust GLR change point detection procedure detailed below is that if a change τ_c is $(0,d)$-detectable then there exists a previous time $r = \tau_{c-1}+1$ from which, using observations between r on, one can detect this change before the next change point, no matter whether the previous change point was undetectable or not. This suggests the following modified procedure:

$$\text{rGLR}^{\text{sub-}\sigma}(t_0,t) = \mathbb{I}\{s\in[t_0:t) : \exists r\in[t_0-1,s)\ \left|\mu_{r+1:s}-\mu_{s+1:t}\right| \geqslant b_{t_0}(r,s,t,\delta)\} \quad \text{where}$$

$$b_{t_0}(r,s,t,\delta) = \sigma\sqrt{\left(\frac{1}{s-r}+\frac{1}{t-s}\right)\left(1+\frac{1}{t-r}\right)2\ln\left[\frac{(t-t_0)(t-t_0+1)\sqrt{t-t_0+2}}{\delta}\right]}.$$

The main result that justifies this procedure is the following uniform concentration inequality

$$\mathbb{P}\left(\exists t\in\mathbb{N}_\star, s\in[t_0:t), r\in[t_0-1:s), \left|\mu_{r+1:s}-\mu_{s+1:t}-\mathbb{E}[\mu_{r+1:s}-\mu_{s+1:t}]\right| \geqslant b_{t_0}(r,s,t,\delta)\right) \leqslant \delta\,. \quad (6)$$

Theorem 11 (Robust Detection delay) *Consider the procedure rGLR$^{\text{sub-}\sigma}$ started at time t_0, run for each subsequent time t. Then the following holds under Assumption 1:*
(i) When no change point occurs on $[t_0,\tau]$, then on an event of probability higher than $1-\delta$, no false alarm is raise for all $t\leqslant\tau$.
(ii) When a $(0,d)$-detectable change point occurs at τ_c+1 with magnitude Δ_c, no matter whether there were undetectable change points before τ_c, it is detected with probability higher than $1-\delta$, with a delay not exceeding $d_1(\max\{t_0,\tau_{c-1}+1\},\tau_c+1,\Delta_c)$.

The procedure rGLR$^{\text{sub-}\sigma}$ thus improves on GLR$^{\text{sub-}\sigma}$ in terms of detection delay in the case when there are undetectable change points. It is however more costly than GLR$^{\text{sub-}\sigma}$, since it checks whether a change point occurred using all starting points r and not only point t_0. One can reduce this overhead a little bit: Indeed the means $(\mu_{s:s'})_{t_0\leqslant s,s'\leqslant t}$ can be stored rather than recomputed for each pair. Then, updating this matrix requires computing $t-t_0+1$ new pairs at time t, where each can be computed at cost $O(1)$ from the means computed at time $t-1$. Thus computing all the means at time t only requires a $O(t)$ update computations. Now, one still needs to perform all the $O((t-t_0)^2)$ tests. The required memory is $O((t-t_0)^2)$ until the first change point is detected, and it can be freed after the second change is detected. We deduce that the computational time and the memory usage scale at time t as $O\big((\tau_{c_{i+1}}-\tau_{c_{i-1}})^2\big)$ where c_i is the i^{th} $(0,d)$-detectable change point and i is such that $\tau_{c_i}<t\leqslant\tau_{c_{i+1}}$. This may or not be affordable in practice depending on the application, but when it is the case, we suggest to use this procedure.

When this procedure is considered too costly, one may stick to the procedure GLR$^{\text{sub-}\sigma}$ (that only requires $O(t-t_0)$ update computations). It is however less robust to a small undetectable change point. Indeed, an undetectable change point causes a perturbation in the apparent mean of the process, which may then cause an additional detection delay. More precisely, the last part of Theorem 6 gives an upper bound on the maximal gap of an *undetectable* change point τ_c: It is immediate to see from this expression that this gap is maximal for extreme values of τ_c close to either τ_{c-1} or τ_{c+1} and minimal for τ_c close to the mid-point of this interval. For a large interval, the maximal and minimal value may thus be very different. This happens for instance when $\tau_{c+1}-\tau_c=1$ and $\tau_{c+1}-\tau_{c-1}=100$. In order to avoid such imbalanced situations, we introduce the following assumption; it applies exclusively to the time occurrence of *undetectable* change points (this does not restrict time of detectable change points):

Assumption 2 (Centered undetectable changes) *There exists some $\eta\in[0,1/2)$ such that for all* undetectable *change point τ_c, τ_c is approximately centered in $[\tau_{c-1}+1,\tau_{c+1}]$ in the sense that*

$$\frac{\min\{\tau_{c+1}-\tau_c,\tau_c-\tau_{c-1}\}}{\tau_{c+1}-\tau_{c-1}}>\eta\,.$$

We further assume that η is away from 0.

The performance of the procedure in such a context is summarized below:

Theorem 12 (Detection delay with perturbation) *Let τ_c be the first (ε,d)-detectable change point after t_0. Consider the procedure GLR$^{\text{sub-}\sigma}$ started at time t_0, run for each subsequent time and that first detects a change point at time t. Then the following holds under Assumptions 1 and 2:*

(iv) If $t > \tau_c$ (no detection occurred before τ_c+1) and previous changes are undetectable, then with probability higher than $1-\delta$, the change is detected with a delay not exceeding

$$d_1(\max\{t_0, \tau_{c-1}+1\}, \tau_c+1, \Delta_c - \Gamma_\eta(t_0,\tau_c)), \quad \textit{where}$$

$$\Gamma_\eta(t_0,\tau_c) \;=\; \sigma\sqrt{\frac{(1-\eta)(\tau_c-t_0+2)}{\eta(\tau_c-t_0+1)^2}\,8\ln\left[\frac{2(\tau_c-t_0)\sqrt{\tau_c-t_0+2}}{\delta}\right]}\,.$$

Remark 13 *Theorem 11 does not require Assumption 2, unlike Theorem 12.*

The last two results show how change points of small-magnitude may cause delay in the detection of a detectable change point. More precisely, both lemmas focus on the delay in the case when the first detection occurs after the first detectable change point. In order to have a complete picture, one should also handle the possible nasty situation when a detection occurs before the first detectable change point. This situation is ruled out under the following two simple assumptions:

Assumption 3 (Separation) *All change points are either (ε, d)-detectable, or undetectable.*

Assumption 4 (Isolation) *For all c, if τ_c is undetectable, then τ_{c+1} and τ_{c-1} are (ε, d)-detectable.*

Assumption 3 is a rather classical *separation* condition that enables to obtain non-trivial and not too technical detection guarantees. Assumption 4 ensures that we cannot have a series of successive undetectable change points. Indeed it may be the case that a sequence of successive undetectable changes causes enough cumulative perturbation to trigger a detection event. In this situation, the detection may occur late after the last undetectable change point, say τ_c and soon before the detectable change point τ_{c+1}. This is a nasty situation since when this happens, any algorithm that resets the detection procedure immediately after a change point detection event may perform badly. Indeed, the fresh new instance will only observe a small fraction of the available observations before τ_{c+1}, leading to a possibly large detection delay. Assumption 4 thus restricts the situation to the case of sequences of arbitrarily many successive detectable change points, but isolated undetectable change points. Although still a bit restrictive, this encompasses many situations and is considerably weaker than the more classical scenario that assumes all changes are detectable. Armed with these assumptions, we now complete our understanding of the detection procedures:

Theorem 14 (Robust false alert) *Let either GLR$^{\text{sub-}\sigma}$ or its robust version be started at time t_0. Under Assumptions 1, 3 and 4, on an event of probability higher than $1-\delta$, no (false) detection occurs before the first (ε, d)-detectable change point time following time t_0.*

6. Discussion and extensions

In this section, we now illustrate the performance of the main detection procedure. We note that, while it is desirable to compare to other state-of-the-art methods, it unfortunately turns out to be challenging for the reason that many strategies are heavily designed for the batch setup plus assume prior knowledge of quantities such as the minimum mean gap Δ_{min} or minimum time $\Delta\tau_{\min}$ between two changes; This is for instance the case of the strategy from Frick et al. (2014), that explicitly uses the total number of observations in all confidence bounds, and resorts to Monte-Carlo estimates and optimization steps involving Δ_{min} and $\Delta\tau_{\min}$ in order to tune their threshold parameter; these many tricks hinder the reproducibility of their results and adaptation to the sequential setup. Now it can be checked that their procedure is based only on bounds that hold for each time horizon n, and not uniformly over all time steps; they also provide much weaker theoretical results (asymptotic optimality of the *rate* only).

The closest competitor to our setup and our proposed strategy is the one considered in Lai and Xing (2010), that is designed for the sequential setup and enjoys asymptotic uniformly-good strategies; however the criterion they consider is the expected delay, while we consider a high-probability control of the delay. More importantly, they make use of a peeling strategy instead of a Laplace method for their analysis: This made them design a modified GLR strategy using peeling that requires additional parameters whose tuning is only based on asymptotic requirements; further the threshold value is tuned in practice using Monte-Carlo estimates plus heuristics due to some complications that they discuss; this hinders reproducibility. In stark contrast, thanks to our use of the Laplace method instead of of a Peeling technique, we can keep the GLR strategy untouched and greatly simplify both the analysis and tuning; actually our threshold is fully explicit.

A baseline comparison Since existing analyses resort to a simple union bound, or at best a peeling argument, and since the main contribution of this work is to make use of the refined Laplace method for the tuning of parameter c, we provide below an illustrative experiment, showing a sequential change-point detection task where we compare $\mathrm{GLR}^{\text{sub-}\sigma}$ with the simple GLR test $\mathrm{glr}^{\text{sub-}\sigma}$ obtained by standard union bounds:

$$\mathrm{glr}^{\text{sub-}\sigma}(t_0,t) = \mathbb{I}\Big\{\exists s\in[t_0,t) : |\mu_{t_0:s}-\mu_{s+1,t}| \geqslant \sqrt{2}\sigma\Big[\sqrt{\frac{1}{s-t_0+1}\ln\Big[\frac{4(s-t_0+1)(s-t_0+2)}{\delta}\Big]}+\sqrt{\frac{1}{t-s}\ln\Big[\frac{4(t-t_0)(t-s+1)(t-s)}{\delta}\Big]}\Big]\Big\}.$$

We performed 200 experiments of $n = 300$ $\sigma = 0.1$-Gaussian observations each with pieces of randomly generated length $\in [10, 30]$ and changes of magnitude randomly chosen in $[0, 1]$, with confidence level $\delta = 0.01$. Figure 2 plots a histogram of the difference of detection delays between the two procedures, aggregated over all changes in each sequence of observations then on all sequences (this results in about 2.10^3 to 6.10^3 many delay-differences). Note that the histogram does not distinguish between changes of large or small magnitude (Intuitively, for changes of large magnitude, both methods should have same detection delay, while for changes of small magnitude, they should differ more significantly). However, it clearly highlights the benefit of the refined procedure that reduces the detection delays and detects more changes (see the red bar).

In order to complement this illustration, we provide in Figure 3 scatter plots of the detection delays when running the two strategies on the same sequences (we used 20 instead of 200 experiments for readability purpose). Each blue circle corresponds to a detection event for a change point c: the x coordinate is the number of observations the algorithms received since its last detection event, the y coordinate is the mean gap of the change point, the size of the point is the delay. Further, we added at same locations red/green circles whose size is the difference between the detection delay of $\mathrm{glr}^{\text{sub-}\sigma}$ and $\mathrm{GLR}^{\text{sub-}\sigma}$ strategy. A red color indicates that $\mathrm{GLR}^{\text{sub-}\sigma}$ improves on $\mathrm{glr}^{\text{sub-}\sigma}$, and a green color the other case. The scatter plots clearly highlight the massive benefit of the novel strategy over the one built from simple union bounds in the difficult regime (bottom, left). It also shows that in the easy cases (top, right), which was the only regime covered by the existing theory, the two strategies detect changes equally well.

Beyond piecewise-iid models? Theorem 6 illustrates the properties of the GLR test in the i.i.d. setting with change of means. Note that, as is usual in such sequential setups, it is immediate to extend these results from i.i.d. to predictive sequence (see also Lai (1998)).

More interestingly, it is natural to ask whether a similar result can be derived for other type of changes, such as a change of variance, say in a family of centered Gaussians $\{\mathcal{N}(0,\theta) : \theta \in \mathbb{R}\}$,

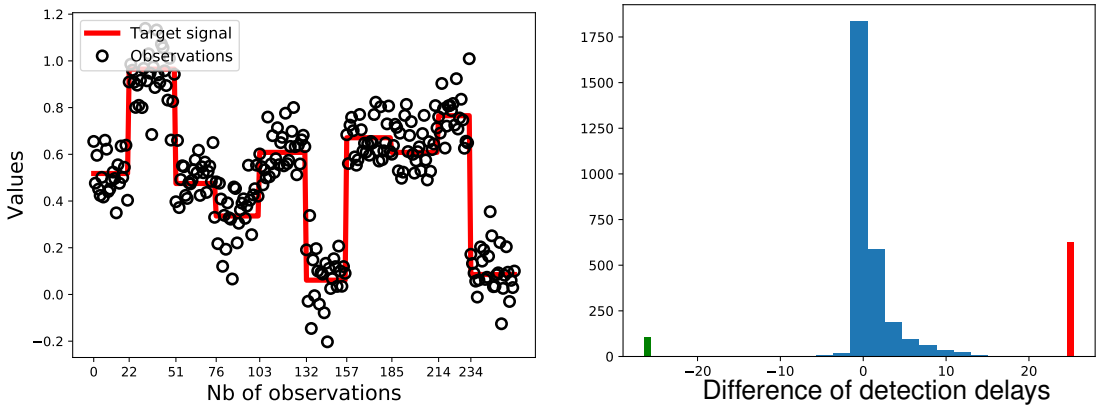

Figure 2: Left: example of input signal, Right: Histogram of the difference between the detection delay of $\text{glr}^{\text{sub-}\sigma}$ and $\text{GLR}^{\text{sub-}\sigma}$. The red bar counts change points missed by $\text{glr}^{\text{sub-}\sigma}$ but detected by $\text{GLR}^{\text{sub-}\sigma}$; the green bar counts the reverse case.

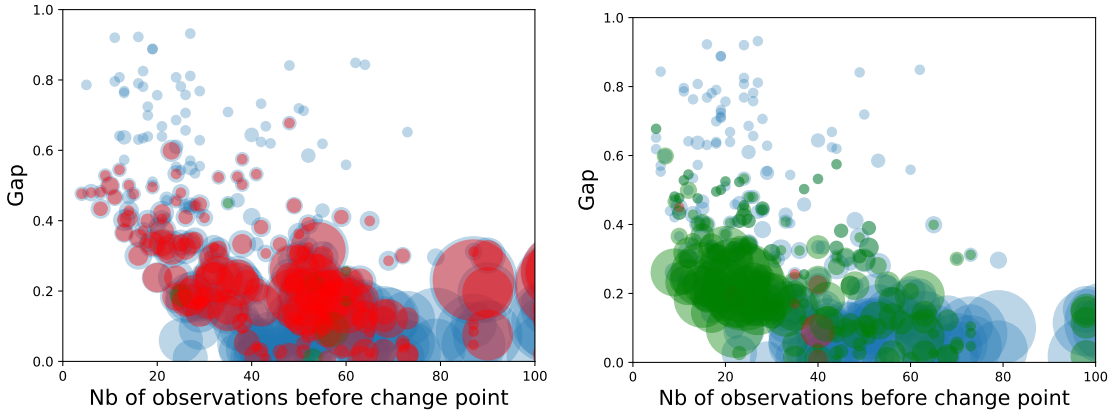

Figure 3: Scatter plots showing the detection delays of $\text{glr}^{\text{sub-}\sigma}$ (left) versus $\text{GLR}^{\text{sub-}\sigma}$ (right). Red/Green circles: difference of detection delays (red is positive, green is negative).

or to higher order parameters: The Laplace method of mixture gives exact bounds in the (sub)-Gaussian case; getting similar results for arbitrary families is an open question. For arbitrary parametric distribution of dimension 1, one may still use the (less tight) peeling method; an extension beyond dimension 1 is more challenging and related to long-lasting open questions in boundary crossing probabilities (see Maillard (2018)).

On the other hand, the Laplace method can be used to produce self-normalized inequalities parametric linear regression setup (with vector-valued martingales) (see Abbasi-Yadkori et al. (2011) or Peña et al. (2008)). Hence one may want to extend the above strategy to such setups. We leave the derivation of (doubly-time uniform) joint concentration inequalities in this context as an open question.

Localization When considering a sequence of multiple change points, coupling detection techniques with estimation (localization) can be beneficial. Indeed a detection procedure only raises an alarm if a change is detected, and is not intended to be informative about the time when the change occurred. On the other hand, a localization procedure builds, after a change point is detected by the detection procedure, an interval, say $[\tau^-, \tau^+]$ that contains the change point τ with high probability. Since localization errors can be of lower magnitude than detection delay, this makes the use of localization interesting in sequential decision making. We also leave this interesting related problem for future work.

Conclusion

In this paper, we have revisited the GLR change-point detection method for the class of mean piecewise process with sub-Gaussian observation noise. We provided two refinements. The first one is to make use of the Laplace method of mixture to derive novel concentration inequalities for the scan-statistics of the GLR test (Theorem 5). This enables to obtain results that hold, with high probability, doubly-uniformly on all time steps, thanks to a non-trivial extension of the Laplace method (Theorem 4) that is of independent interest. More importantly, this enables to obtain sharp, explicit and non-asymptotic bound on the detection delay, under a high probability control of the false alarm probability, hence giving insights much beyond the asymptotic case (that corresponds to easy detection regime). This technique also leads to an explicit tuning, that ensures the strategy asymptotically matches the lower achievable bounds for this problem, while enjoying finite-time detection guarantees. It also enables to avoid resorting to the more tedious and less optimal peeling technique previously used in other methods, hence simplifying both the analysis and threshold tuning.

Finally, in stark contrast with the existing analyses, we study the practically-relevant situation when undetectable changes may happen, and study the robustness of the GLR procedure to such undetectable changes in Theorem 11 and Theorem 12. This seems to be novel. We also believe such results to be especially useful in applications.

Acknowledgments

This work has been supported by CPER Nord-Pas de Calais/FEDER DATA Advanced data science and technologies 2015-2020, the French Ministry of Higher Education and Research, Inria Lille – Nord Europe, CRIStAL, and the French Agence Nationale de la Recherche (ANR), under grant ANR-16-CE40-0002 (project BADASS).

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Michèle Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs, 1993.

Rudolf B Blazek, Hongjoong Kim, Boris Rozovskii, and Alexander Tartakovsky. A novel approach to detection of denial-of-service attacks via adaptive sequential and batch-sequential change-point detection methods. In *Proceedings of IEEE systems, man and cybernetics information assurance workshop*, pages 220–226. Citeseer, 2001.

S Bouzebda. Some applications of the strong approximation of the integrated empirical copula processes. *Mathematical Methods of Statistics*, 25(4):281–303, 2016.

E Brodsky and Boris S Darkhovsky. *Nonparametric Methods in Change-Point Problems*, volume 243. Springer, 1993.

Alain Celisse, Guillemette Marot, Morgane Pierre-Jean, and Guillem Rigaill. New efficient algorithms for multiple change-point detection with kernels. *arXiv preprint arXiv:1710.04556*, 2017.

Miklós Csörgö and Lajos Horváth. *Limit theorems in change-point analysis*, volume 18. John Wiley & Sons Inc, 1997.

Allen B Downey. A novel changepoint detection algorithm. *arXiv preprint arXiv:0812.1237*, 2008.

Klaus Frick, Axel Munk, and Hannes Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580, 2014.

Aurélien Garivier. Informational confidence bounds for self-normalized averages and applications. In *Information Theory Workshop (ITW), 2013 IEEE*, pages 1–5. IEEE, 2013.

Damien Garreau and Sylvain Arlot. Consistent change-point detection with kernels. *arXiv preprint arXiv:1612.04740*, 2016.

Chen Jie and AK Gupta. Parametric statistical change point analysis. *Birkh User*, 2000.

Azadeh Khaleghi and Daniil Ryabko. Asymptotically consistent estimation of the number of change points in highly dependent time series. In *International Conference on Machine Learning*, pages 539–547, 2014.

Tze Leung Lai. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory*, 44(7):2917–2929, 1998.

Tze Leung Lai and Haipeng Xing. Sequential change-point detection when the pre-and post-change parameters are unknown. *Sequential analysis*, 29(2):162–175, 2010.

Gary Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, pages 1897–1908, 1971.

O-A Maillard. Boundary crossing probabilities for general exponential families. *Mathematical Methods of Statistics*, 27(1):1–31, 2018.

Odalric-Ambrym Maillard. Basic concentration properties of real-valued distributions. 2017.

Yajun Mei. Suboptimal properties of page's cusum and shiryayev-roberts procedures in change-point problems with dependent observations. *Statistica Sinica*, pages 883–897, 2006.

Yajun Mei. Is average run length to false alarm always an informative criterion? *Sequential Analysis*, 27(4):354–376, 2008.

George V Moustakides. Optimality of the cusum procedure in continuous time. *Annals of Statistics*, pages 302–315, 2004.

Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.

Moshe Pollak. Optimal detection of a change in distribution. *The Annals of Statistics*, pages 206–227, 1985.

Moshe Pollak and David Siegmund. Sequential detection of a change in a normal mean when the initial value is unknown. *The Annals of Statistics*, pages 394–416, 1991.

SW Roberts. A comparison of some control chart procedures. *Technometrics*, 8(3):411–430, 1966.

Albert N Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8(1):22–46, 1963.

AN Shiryaev. Statistical analysis of sequential processes. optimal stopping rules, 1978.

David Siegmund and ES Venkatraman. Using the generalized likelihood ratio statistic for sequential detection of a change-point. *The Annals of Statistics*, pages 255–271, 1995.

AG Tartakovsky. Sequential methods in the theory of information systems, 1991.

Namrata Vaswani. The modified cusum algorithm for slow and drastic change detection in general hmms with unknown change parameters. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 4, pages iv–701. IEEE, 2005.

Yanhong Wu. *Inference for change point and post change means after a CUSUM test*, volume 180. Springer, Lecture Notes in Statistics, 2007.

Appendix A. Uniform confidence bounds from the Laplace method

Lemma 15 (Uniform confidence intervals) *Let $Y_1, \dots Y_t$ be a sequence of t i.i.d. real-valued random variables with mean μ, such that $Y_t - \mu$ is σ-sub-Gaussian. Let $\mu_t = \frac{1}{t}\sum_{s=1}^{t} Y_s$ be the empirical mean estimate. Then, for all $\delta \in (0,1)$, it holds*

$$\mathbb{P}\left(\exists t \in \mathbb{N}, \quad \mu_t - \mu \geqslant \sigma\sqrt{\left(1+\frac{1}{t}\right)\frac{2\ln\left(\sqrt{t+1}/\delta\right)}{t}}\right) \leqslant \delta$$
$$\mathbb{P}\left(\exists t \in \mathbb{N}, \quad \mu - \mu_t \geqslant \sigma\sqrt{\left(1+\frac{1}{t}\right)\frac{2\ln\left(\sqrt{t+1}/\delta\right)}{t}}\right) \leqslant \delta\,.$$

(The "Laplace" method refers to using the Laplace method of integration for optimization)

Proof of Lemma 15, equation (4):

We introduce for a fixed $\delta \in [0,1]$ the random variable

$$\tau = \min\left\{t \in \mathbb{N} : \mu_t - \mu \geqslant \sigma\sqrt{\left(1+\frac{1}{t}\right)\frac{2\ln\left(\sqrt{1+t}/\delta\right)}{t}}\right\}.$$

This quantity is a random stopping time for the filtration $\mathcal{F} = (\mathcal{F}_t)_t$, where $\mathcal{F}_t = \sigma(Y_1, \dots, Y_t)$, since $\{\tau \leqslant m\}$ is $\mathcal{F}_m$-measurable for all m. We want to show that $\mathbb{P}(\tau < \infty) \leqslant \delta$. To this end, for any λ, and t, we introduce the following quantity

$$M_t^\lambda = \exp\left(\sum_{s=1}^{t}(\lambda(Y_s - \mu) - \frac{\lambda^2\sigma^2}{2})\right).$$

By assumption, the centered random variables are σ-sub-Gaussian and it is immediate to show that $\{M_t^\lambda\}_{t\in\mathbb{N}}$ is a non-negative super-martingale that satisfies $\ln \mathbb{E}[M_t^\lambda] \leqslant 0$ for all t. It then follows that $M_\infty^\lambda = \lim_{t\to\infty} M_t^\lambda$ is almost surely well-defined and so, M_τ^λ as well. Further, let us introduce the stopped version $Q_t^\lambda = M^\lambda_{\min\{\tau,t\}}$. An application of Fatou's lemma shows that $\mathbb{E}[M_\tau^\lambda] = \mathbb{E}[\liminf_{t\to\infty} Q_t^\lambda] \leqslant \liminf_{t\to\infty} \mathbb{E}[Q_t^\lambda] \leqslant 1$. Thus, $\mathbb{E}[M_\tau^\lambda] \leqslant 1$.

The next step is to introduce the auxiliary variable $\Lambda = \mathcal{N}(0, \sigma^{-2})$, independent of all other variables, and study the quantity $M_t = \mathbb{E}[M_t^\Lambda | \mathcal{F}_\infty]$. Note that the standard deviation of Λ is σ^{-1} due to the fact we consider σ-sub-Gaussian random variables. We immediately get $\mathbb{E}[M_\tau] = \mathbb{E}[\mathbb{E}[M_\tau^\Lambda|\Lambda]] \leqslant 1$. For convenience, let $S_t = t(\mu_t - \mu)$. By construction of M_t, we have

$$\begin{aligned}
M_t &= \frac{1}{\sqrt{2\pi\sigma^{-2}}}\int_{\mathbb{R}} \exp\left(\lambda S_t - \frac{\lambda^2\sigma^2 t}{2} - \frac{\lambda^2\sigma^2}{2}\right)d\lambda \\
&= \frac{1}{\sqrt{2\pi\sigma^{-2}}}\int_{\mathbb{R}} \exp\left(-\left[\lambda\sigma\sqrt{\frac{t+1}{2}} - \frac{S_t}{\sigma\sqrt{2(t+1)}}\right]^2 + \frac{S_t^2}{2\sigma^2(t+1)}\right)d\lambda \\
&= \exp\left(\frac{S_t^2}{2\sigma^2(t+1)}\right)\frac{1}{\sqrt{2\pi\sigma^{-2}}}\int_{\mathbb{R}} \exp\left(-\lambda^2\sigma^2\frac{t+1}{2}\right)d\lambda \\
&= \exp\left(\frac{S_t^2}{2\sigma^2(t+1)}\right)\frac{\sqrt{2\pi\sigma^{-2}/(t+1)}}{\sqrt{2\pi\sigma^{-2}}}\,.
\end{aligned}$$

Thus, we deduce that

$$S_t = \sigma\sqrt{2(t+1)\ln\left(\sqrt{t+1}M_t\right)}\,.$$

We conclude by applying a simple Markov inequality:

$$\mathbb{P}\left(\tau(\mu_\tau - \mu) \geqslant \sigma\sqrt{2(\tau+1)\ln\left(\sqrt{\tau+1}/\delta\right)}\right) = \mathbb{P}(M_\tau \geqslant 1/\delta) \leqslant \mathbb{E}[M_\tau]\delta\,.$$

□

We now consider an extension of this powerful result.

Proof of Theorem 4 (equation (5)):

To this end, we first note that $(t-t')(\mu_{1:t'} - \mu_{t'+1:t} - \mathbb{E}[\mu_{1:t'} - \mu_{t'+1:t}]) = \sum_{s=1}^{t} Z_s$ is the sum of t independent random variables, t' of which are $\sigma(t-t')/t'$-sub-Gaussian, and $t-t'$ are σ-sub-Gaussian. We denote Z_s the sth term of the sum and introduce $\sigma_s \overset{def}{=} \sigma(t-t')/t'$ if $s \leqslant t'$, and $\sigma_s = \sigma$ for $s > t'$.

We then form, for each λ, the following quantity

$$\begin{aligned}
M_t^\lambda &= \exp\left(\sum_{s=1}^{t}\lambda Z_s - \frac{\lambda^2\sigma_s^2}{2}\right)\\
&= \exp\left((\sum_{s=1}^{t}\lambda Z_s) - \sum_{s=1}^{t'}\frac{\lambda^2\sigma^2}{2t'^2}(t-t')^2 - \sum_{s=t'+1}^{t}\frac{\lambda^2\sigma^2}{2}\right)\\
&= \exp\left((\sum_{s=1}^{t}\lambda Z_s) - \frac{\lambda^2\sigma^2}{2t'}(t-t')^2 - \frac{\lambda^2\sigma^2}{2}(t-t')\right)\\
&= \exp\left((\sum_{s=1}^{t}\lambda Z_s) - \frac{\lambda^2\sigma^2(t/t'-1)t}{2}\right).
\end{aligned}$$

where we used in the last line that $\frac{(t-t')^2}{t'} + (t-t') = (t-t')(t/t') = (t/t'-1)t$ Note that M_t^λ is *not* a super-martingale due to the fact that σ_s depends on t for $s < t'$. However, it satisfies for each λ and each $t > t'$ $\ln\mathbb{E}[M_t^\lambda] \leqslant 0$ thanks to the sub-Gaussian assumption. More importantly, it can be shown that

$$\mathbb{E}[M_{t+1}^\lambda|\mathcal{F}_t] \leqslant \underbrace{\exp\left[\lambda(\mu_{1:t'} - \mathbb{E}[\mu_{1:t'}]) - \frac{\lambda^2\sigma^2}{2t'}\right]\exp\left(-\lambda^2\sigma^2\frac{(t-t')}{t'}\right)}_{\alpha_{t',t}} M_t^\lambda\,.$$

This positive factor $\alpha_{t',t}$ is $\mathcal{F}_{t'}$ measurable and satisfies $\mathbb{P}\big(\alpha_{t',t} \geqslant 1\big) \leqslant \exp\left(-\lambda^2\sigma^2\frac{(t-t')}{t'}\right)$. Hence $\mathbb{E}[M_{t+1}^\lambda|\mathcal{F}_t] \leqslant M_t^\lambda$ holds on an $\mathcal{F}_{t'}$-measurable event whose probability tends to 1 exponentially fast as $t \to \infty$. This ensures that, although M_t^λ is not a super-martingale, it behaves

asymptotically as such and that $M_\infty^\lambda = \lim_{t\to\infty} M_t^\lambda$ is still almost surely well-defined (following the same steps as for Doob's upcrossing lemma, and using that $\sup_n \sum_{s=t'}^n \mathbb{E}[(\alpha_{t',s}-1)_+]$ is finite). This ensures that M_τ^λ is also well defined, whether or nor τ is finite.

We now apply similar steps as for the proof of Theorem 5, but for a slightly different quantity. Namely, we introduce the auxiliary variable $\Lambda_t = \mathcal{N}(0, v_t^{-2})$, where $v_t = \sigma\sqrt{t/t'-1}$ independent of all other variables, and study the quantity $M_t = \mathbb{E}[M_t^{\Lambda_t}|\mathcal{F}_\infty]$. We note that $\mathbb{E}[M_t] \leqslant 1$ for all t, since for each t, $\mathbb{E}[M_t^\lambda] \leqslant 1$ holds for all $\lambda \in \mathbb{R}$. Further, let us introduce the stopped version $Q_t = M_{\min\{\tau,t\}}$. An application of Fatou's lemma shows that $\mathbb{E}[M_\tau] = \mathbb{E}[\liminf_{t\to\infty} Q_t] \leqslant \liminf_{t\to\infty} \mathbb{E}[Q_t] \leqslant 1$. Thus, $\mathbb{E}[M_\tau] \leqslant 1$.

For convenience, let $S_t = (t-t')(\mu_{1:t'} - \mu_{t'+1:t} - \mathbb{E}[\mu_{1:t'} - \mu_{t'+1:t}])$. By construction of M_t, we have

$$\begin{aligned}
M_t &= \frac{1}{\sqrt{2\pi v_t^{-2}}}\int_{\mathbb{R}} \exp\left(\lambda S_t - \frac{\lambda^2\sigma^2(t/t'-1)t}{2} - \frac{\lambda^2 v_t^2}{2}\right)d\lambda \\
&= \frac{1}{\sqrt{2\pi v_t^{-2}}}\int_{\mathbb{R}} \exp\Big(- \Big[\lambda\sigma\sqrt{\frac{(t/t'-1)(t+1)}{2}} - \frac{S_t}{\sigma\sqrt{2(t/t'-1)(t+1)}}\Big]^2 \\
&\quad + \frac{S_t^2}{2\sigma^2((t/t'-1)(t+1))}\Big)d\lambda \\
&= \exp\left(\frac{S_t^2}{2\sigma^2((t/t'-1)(t+1))}\right)\frac{1}{\sqrt{2\pi v_t^{-2}}}\int_{\mathbb{R}} \exp\Big(- \lambda^2\sigma^2\frac{(t/t'-1)(t+1)}{2}\Big)d\lambda \\
&= \exp\left(\frac{S_t^2}{2\sigma^2((t/t'-1)(t+1))}\right)\frac{\sqrt{2\pi v_t^{-2}/(t+1)}}{\sqrt{2\pi v_t^{-2}}}\,.
\end{aligned}$$

Thus, we deduce that

$$S_t = \sigma\sqrt{2((t/t'-1)(t+1))\ln\left(\sqrt{t+1}M_t\right)}\,.$$

applying a simple Markov inequality, and reorganize the terms yields

$$\mathbb{P}\left(\exists t > t', \mu_{1:t'} - \mu_{t'+1:t} - \mathbb{E}[\mu_{1:t'} - \mu_{t'+1:t}]) \geqslant \sigma\sqrt{2\frac{(t/t'-1)(t+1)}{(t-t')^2}\ln\left(\sqrt{t+1}/\delta\right)}\right) \leqslant \delta$$

We conclude by applying a similar argument to control the reverse inequality, and by remarking that

$$\frac{(t/t'-1)(t+1)}{(t-t')^2} = \frac{t+1}{t'(t-t')} = (\frac{1}{t'} + \frac{1}{t-t'})(1+\frac{1}{t})\,.$$

□

Proof of Theorem 5:

We focus only on the joint case (the proof of the disjoint case uses similar steps). Using insights from the proof of Theorem 4, we denote $S_{s:t} = (t-s)(\mu_{1:s} - \mu_{s+1:t} - \mathbb{E}[\mu_{1:s} - \mu_{s+1:t}])$. Let us introduce the following quantity

$$\begin{aligned}\tau &= \min\left\{t \in \mathbb{N} : \exists s < t : S_{s:t} \geqslant \sigma\sqrt{2((t/s-1)(t+1))\ln(t\sqrt{t+1}/\delta)}\right\} \\ &= \min\left\{t \in \mathbb{N} : \exists s < t : M_{s:t} \geqslant t/\delta\right\},\end{aligned}$$

where the second line holds since $S_{s:t} = \sigma\sqrt{2((t/s-1)(t+1))\ln(\sqrt{t+1}M_{s:t})}$. τ is random stopping time. Let us remark that since $\mathbb{E}[M_{s:t}] \leqslant 1$ holds for all s, t, then for all t, we also have $\mathbb{E}\big[\max_{s<t} \frac{M_{s:t}}{t}\big] \leqslant \mathbb{E}\big[\sum_{s=0}^{t-1} \frac{M_{s:t}}{t}\big] \leqslant 1$. Similarly to $M_{s:t}$, $\lim_{t\to\infty} \sum_{s=0}^{t-1} \frac{M_{s:t}}{t}$ is almost surely well-defined, since the quantity behaves asymptotically like a super-martingale. Hence $\sum_{s=0}^{\tau-1} \frac{M_{s:\tau}}{\tau}$ and $\max_{s<\tau} \frac{M_{s:\tau}}{\tau}$ are also well-defined, whether or not τ is finite. Further, it can be shown that $\mathbb{E}\left[\max_{s<\tau} \frac{M_{s:\tau}}{\tau}\right] \leqslant 1$. Hence, we deduce that

$$\begin{aligned}\mathbb{P}\left(\forall t, \exists s < t : S_{s:t} \geqslant \sigma\sqrt{2((t/s-1)(t+1))\ln(t\sqrt{t+1}/\delta)}\right) &= \mathbb{P}(\exists s < \tau : M_{s:\tau} \geqslant \tau/\delta) \\ &\leqslant \mathbb{P}\Big(\frac{\max_{s<\tau} M_{s:\tau}}{\tau} \geqslant 1/\delta\Big) \\ &\leqslant \mathbb{E}\left[\frac{\max_{s<\tau} M_{s:\tau}}{\tau}\right]\delta \leqslant \delta\,.\end{aligned}$$

□

Appendix B. Other time-uniform concentration bounds

It is worth discussing how the Laplace bound of Lemma 15 relates to other alternatives. We show below the Laplace method leads to confidence bounds up to twice smaller than more common alternative bounds, such as

$$\text{(Union bound)} \qquad \mathbb{P}\left(\exists t \in \mathbb{N}, \quad \mu_t - \mathbb{E}[\mu_t] \geqslant \sigma\sqrt{\frac{2\ln\big(t(t+1)/\delta\big)}{t}}\right) \leqslant \delta\,,$$

$$\text{(Peeling method)} \qquad \mathbb{P}\left(\exists t \in \mathbb{N}, \quad \mu_t - \mathbb{E}[\mu_t] \geqslant \sigma\sqrt{\frac{2(1+\eta)}{t}\ln\left(\frac{\ln(t)\ln(t(1+\eta))}{\delta\ln^2(1+\eta)}\right)}\right) \leqslant \delta\,,$$

where $\eta \leqslant 1$ is any fixed constant not depending on t. Indeed the bound obtain by a union bound is very crude, and even the *a priori appealing* $\ln\ln(t)$ scaling of the bound obtained by the peeling method is however not better than the one derived by the Laplace method, unless for huge times t ($t \geqslant 10^6$, for $\delta = 0.05$ and any η, see also Figure B). This should not be surprising, since neither methods make use of the fact that the variables are sub-Gaussian, contrary to the Laplace method.

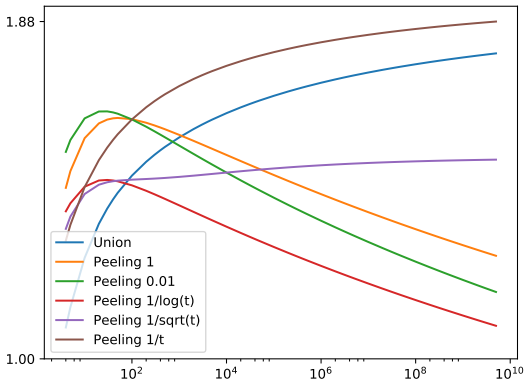

Figure 4: Ratio of different time-uniform concentration bounds over that of the Laplace method, as a function of t, for a confidence level $\delta = 0.01$ and various choice of $\eta = \eta(t)$. This indicates that all other bounds are larger by a multiplicative factor up to 1.88 here, and none is smaller until at least time $t = 10^{10}$. Notice the logarithmic scale.

Remark 16 (Comments regarding the alternative results) *Let us provide a few hints about the last two inequalities for the interested reader. The one called (union bound) is trivially obtained by a union bound argument, using that* $\sum_t \frac{1}{t(t+1)} = 1$. *The second one uses a peeling technique. Results for peeling have been obtained traditionally for the control of partial sums* $S_n = \sum_{i=1}^n Y_i \xi_i$, *where* Y_i *is a real-valued random variable and* $\xi = (\xi_i)_i$ *is a predictable sequence of* $\{0,1\}$*-valued random variables. Introducing the random count* $N_n = \sum_{i=1}^n \xi_i$, *and localizing it in slices of exponentially increasing size* $\alpha^k < N_n < \alpha^{k+1}$, *for some* $\alpha > 0$, *the typical scaling obtained is of the form*

$$\mathbb{P}\left(\frac{1}{N_n}\sum_{i=1}^n (Y_i - \mu_i)\xi_i \geqslant \sigma\sqrt{\frac{2\alpha}{N_n}\ln\left(\left\lceil\frac{\ln(n)}{\ln(\alpha)}\right\rceil\frac{1}{\delta}\right)}\right) \leqslant \delta\,.$$

See Garivier (2013) for similar results and a few variants. For large (huge) n, *we reach the (asymptotic)* $\ln\ln(n)$ *regime of the law of iterated logarithm, and thus these bounds are un-improvable in this sense. However, although making appear a* $\ln\ln(n)$ *dependency instead of* $\ln(\sqrt{n})$, *these bounds are actually less tight than using Laplace method, unless* n *and* N_n *are huge. This is due to the non-adaptive parameter* α. *Now, these bounds use the fact that* N_n *is a random stopping time for the filtration induced by the* Y_i, *and, crucially, that* N_n *is bounded by* n. *To obtain the (Peeling method) inequality, one should study instead* $\sum_{i=1}^{\tau} Y_i$ *for a random stopping time* τ *that is unfortunately not necessarily bounded a priori. The peeling method can be amended by summing now over an infinite number of slices instead of* $\left\lceil\frac{\ln(n)}{\ln(\alpha)}\right\rceil$ *and avoiding to use an a priori bound* n. *Revisiting the proof, one can easily derive the desired result, by replacing* n *with* τ *which comes at the price of slightly increasing the* $\ln$ *dependency to* $\ln^2$. *See more details in Maillard (2017).*

Appendix C. Detection

C.1. Detection without perturbation

Proof of Theorem 6:

i) False detection By definition of the detection procedure, a detection occurs a time t if $\exists s \in [t_0 : t)$ such that

$$|\mu_{t_0:s} - \mu_{s+1:t}| \quad > \quad b_{t_0}(s,t,\delta)\,.$$

In the first case (i), since there is no change point before τ, then for all $s, t \leqslant \tau$, $\mathbb{E}[\mu_{t_0:s} - \mu_{s+1:t}] = 0$. Now, we observe that, thanks to the uniform concentration inequality, it holds

$$\mathbb{P}\Big(\exists t \in \mathbb{N}_\star,\, s \in [t_0:t),\, \Big|\mu_{t_0:s} - \mu_{s+1:t} - \mathbb{E}[\mu_{t_0:s} - \mu_{s+1:t}]\Big| \geqslant b_{t_0}(s,t,\delta)\Big) \leqslant \delta\,.$$

We deduce that on an event of probability higher than $1-\delta$, no detection occurs for any $t \leqslant \tau$.
ii) Detection delay We now turn to the second case (ii). In the sequel, we consider that $t_0 = 1$. On the same event, it holds for all $t > \tau$ and $s < t$,

$$\begin{aligned}
\mu_{t_0:s} - \mu_{s+1:t} &\geqslant \mathbb{E}[\mu_{t_0:s} - \mu_{s+1:t}] - b_{t_0}(s,t,\delta) \\
\mu_{s+1:t} - \mu_{t_0:s} &\geqslant \mathbb{E}[\mu_{s+1:t} - \mu_{t_0:s}] - b_{t_0}(s,t,\delta)\,,
\end{aligned}$$

which implies that $|\widehat{\Delta}_{s,t}| - \geqslant |\Delta_{s,t}| - b_{t_0}(s,t,\delta)$, where

$$\stackrel{def}{=} \widehat{\Delta}_{s,t} = \mu_{t_0:s} - \mu_{s+1:t}, \qquad \Delta_{s,t} = \mathbb{E}[\mu_{t_0:s} - \mu_{s+1:t}]|\,.$$

At this point, note that we have the relations

$$\begin{aligned}
\forall t' > \tau, \quad \mu_{t_0:t'} &= \frac{\tau - t_0 + 1}{t' - t_0 + 1}\mu_{t_0:\tau} + \frac{t' - \tau}{t' - t_0 + 1}\mu_{\tau+1:t'} \\
\forall t' < \tau < t, \quad \mu_{t'+1:t} &= \frac{\tau - t'}{t - t'}\mu_{t'+1:\tau} + \frac{t - \tau}{t - t'}\mu_{\tau+1:t}\,.
\end{aligned}$$

Thus, taking the expectation on each side, this means that

$$\mathbb{E}[\mu_{t_0:t'} - \mu_{t'+1:t}] = \begin{cases} \frac{t-\tau}{t-t'}(\mu_1 - \mu_2) & \text{if } t' \leqslant \tau \leqslant t \\ \frac{\tau - t_0 + 1}{t' - t_0 + 1}(\mu_1 - \mu_2) & \text{if } \tau \leqslant t' \leqslant t\,. \end{cases}$$

Hence, we get the following expression

$$|\Delta_{s,t}| = \Big(\frac{t-\tau}{t-s}\mathbb{I}\{s \leqslant \tau\} + \frac{\tau - t_0 + 1}{s - t_0 + 1}\mathbb{I}\{s > \tau\}\Big)\Delta\,.$$

Now, we note that for each δ', then

$$\mathbb{P}\Big(\exists s \in [t_0, t) : |\widehat{\Delta}_{s,t} - \Delta_{s,t}| \geqslant \tilde{b}_{t_0}(s,t,\delta')\Big) \leqslant \delta'$$

$$\text{where} \qquad \tilde{b}_{t_0}(s,t,\delta') = \sigma\sqrt{\Big(\frac{1}{s - t_0 + 1} + \frac{1}{t - s}\Big) 2\ln\left[\frac{2(t-t_0)}{\delta}\right]}\,.$$

Hence, using this expression we deduce that with probability higher than $1 - \delta_t$, a detection occurs at most at time $t > \tau$ if for some $s < t$, $\Delta_{s,t} > b_{t_0}(s,t,\delta) + \tilde{b}_{t_0}(s,t,\delta_t)$, where

$$b_{t_0}(s,t,\delta) = \sigma\sqrt{\Big(\frac{1}{s - t_0 + 1} + \frac{1}{t - s}\Big)\Big(1 + \frac{1}{t - t_0 + 1}\Big) 2\ln\left[\frac{2(t-t_0)\sqrt{t - t_0 + 2}}{\delta}\right]}\,.$$

It is convenient to choose δ_t such that $\tilde{b}_{t_0}(s,t,\delta_t) = \varepsilon b_{t_0}(s,t,\delta)$ for some $\varepsilon > 0$. This corresponds to choosing

$$\delta_t(\varepsilon) = 2(t-t_0)\left(\frac{\delta}{2(t-t_0)\sqrt{t-t_0+2}}\right)^{\varepsilon^2(1+\frac{1}{t-t_0+1})}, \text{ and } \delta_t(1) = \delta\,.$$

For $s > \tau$, this corresponds to the condition

$$\begin{aligned}(\tau-t_0+1)\Delta \;\; &> \;\; (1+\varepsilon)\sigma\sqrt{2\min_{s\in[\tau+1:t-1]}\Big(s-t_0+1+\frac{(s-t_0+1)^2}{t-s}\Big)}\times \\ &\qquad \sqrt{\Big(1+\frac{1}{t-t_0+1}\Big)\ln\left[\frac{2(t-t_0)\sqrt{t-t_0+2}}{\delta}\right]}\end{aligned}$$

which can be simplified into

$$\frac{(\tau-t_0+1)^2\Delta^2}{(\tau-t_0+2)2(1+\varepsilon)^2\sigma^2} > \Big(1+\frac{\tau-t_0+2}{t-\tau-1}\Big)\Big(1+\frac{1}{t-t_0+1}\Big)\ln\left[\frac{2(t-t_0)\sqrt{t-t_0+2}}{\delta}\right]$$

Introducing the delay $d = t-(\tau+1)$, it comes

$$\frac{(\tau-t_0+1)^2\Delta^2}{(\tau-t_0+2)2(1+\varepsilon)^2\sigma^2} > \Big(\frac{d+\tau-t_0+3}{d}\Big)\ln\left[\frac{2(d+\tau-t_0+1)\sqrt{d+\tau-t_0+3}}{\delta}\right]$$

Thus, if we introduce now the minimal integer $d \in \mathbb{N}$ (if any) that satisfies

$$d > \frac{2(1+\varepsilon)^2\sigma^2(1+\frac{2}{\tau-t_0+1})\ln\left[\frac{2(d+\tau-t_0+1)\sqrt{d+\tau-t_0+3}}{\delta}\right]}{\frac{(\tau-t_0+1)\Delta^2}{(\tau-t_0+2)} - \frac{2(1+\varepsilon)^2\sigma^2}{\tau-t_0+1}\ln\left[\frac{2(d+\tau-t_0+1)\sqrt{d+\tau-t_0+3}}{\delta}\right]}\,, \tag{7}$$

then a detection occurs at time $t = \tau+1+d$ with probability higher than $1-\delta_{d+\tau+1}(\varepsilon)$. We now detail the case of $s \leqslant \tau$ that corresponds to the condition

$$(t-\tau)^2\Delta^2 > 2(1+\varepsilon)^2\sigma^2 \min_{s\in[t_0:\tau]}(t-s)\frac{t-t_0+2}{s-t_0+1}\ln\left[\frac{2(t-t_0)\sqrt{t-t_0+2}}{\delta}\right].$$

which shows a detection occurs for the minimal t (if any) that satisfies

$$\frac{(t-\tau)\Delta^2}{2(1+\varepsilon)^2\sigma^2} > \frac{t-t_0+2}{\tau-t_0+1}\ln\left[\frac{2(t-t_0)\sqrt{t-t_0+2}}{\delta}\right].$$

Thus, the detection delay must satisfy in this case

$$\frac{(\tau-t_0+1)\Delta^2}{2(1+\varepsilon)^2\sigma^2} > \Big(1+\frac{\tau-t_0+2}{d+1}\Big)\ln\left[\frac{2(d+\tau-t_0+1)\sqrt{d+\tau-t_0+3}}{\delta}\right],$$

That is

$$d > \frac{2(1+\varepsilon)^2\sigma^2(1+\frac{1}{\tau-t_0+1})\ln\left[\frac{2(d+\tau-t_0+1)\sqrt{d+\tau-t_0+3}}{\delta}\right]}{\Delta^2 - \frac{2(1+\varepsilon)^2\sigma^2}{\tau-t_0+1}\ln\left[\frac{2(d+\tau-t_0+1)\sqrt{d+\tau-t_0+3}}{\delta}\right]} - 1\,. \tag{8}$$

Combining inequalities (7) and (8), the detection delay $d = t - (\tau + 1)$, is not larger than

$$\min\Big\{d' \in \mathbb{N}\colon d' \text{ satisfies (7) or (8)}\Big\} \leqslant \min\Big\{d' \in \mathbb{N}\colon d' > \frac{2(1+\varepsilon)^2\sigma^2(1+\frac{1}{\tau-t_0+1})\ln\left[\frac{2x_{d'}}{\delta}\right]}{\Delta^2 - \frac{2(1+\varepsilon)^2\sigma^2}{\tau-t_0+1}\ln\left[\frac{2x_{d'}}{\delta}\right]} - 1\Big\}.$$

where $x_d = (d+\tau-t_0+1)\sqrt{d+\tau-t_0+3}$, and detected with probability $1 - \delta_t(\varepsilon)$.

iii) Maximal no-detection gap It remains to handle the maximal not-detectable gap. Proceeding with similar steps, we have obtained that if a change occurs at $\tau + 1$, and $t > \tau$ is such that

$$\begin{aligned} \text{either} \quad \frac{\Delta^2}{2(1+\varepsilon)^2\sigma^2} &> \frac{(\tau-t_0+2)}{(\tau-t_0+1)^2}\frac{t-t_0+2}{t-\tau-1}\ln\left[\frac{2(t-t_0)\sqrt{t-t_0+2}}{\delta}\right] \\ \text{or} \quad \frac{\Delta^2}{2(1+\varepsilon)^2\sigma^2} &> \frac{t-t_0+2}{(t-\tau)(\tau-t_0+1)}\ln\left[\frac{2(t-t_0)\sqrt{t-t_0+2}}{\delta}\right]. \end{aligned}$$

then a detection occurs with probability higher than $1 - \delta_t(\varepsilon)$. Note also that if we replace $\delta_t(\varepsilon)$ with δ, the same bounds holds with $\varepsilon = 1$. Hence, looking at the minimum of the left-hand side quantities, we deduce that if a change occurring at $\tau + 1$ is not detectable at level δ, where $\tau = \tau_c$, then the change must be of magnitude $\Delta \leqslant \min\{\overline{\Delta}(\tau_{c-1}+1, \tau_c+1, t), t \in [\tau_c+1, \tau_{c+1}]\}$, where we introduced the quantity

$$\overline{\Delta}(t_0, \tau+1, t) = \sigma\sqrt{\frac{(t-t_0+2)}{(t-\tau)(\tau-t_0+1)}8\ln\left[\frac{2(t-t_0)\sqrt{t-t_0+2}}{\delta}\right]}.$$

□

C.2. Detection with perturbation due to undetectable change points

Proof of inequality (6):

In order to prove inequality (6), we use a union bound over all $s \in [t_0, t)$ and $r \in [t_0 - 1, s)$ of the bound given by (5). Indeed, for each $s \in [t_0 : t), r \in [t_0 - 1 : s)$, we obtain by (5) that

$$\mathbb{P}\Bigg(\exists t \in \mathbb{N}_\star, \left|\mu_{r+1:s} - \mu_{s+1:t} - \mathbb{E}[\mu_{r+1:s} - \mu_{s+1:t}]\right| \geqslant \sigma\sqrt{\Big(\frac{1}{s-r} + \frac{1}{t-s}\Big)\Big(1+\frac{1}{t-r}\Big)2\ln\left[\frac{2\sqrt{t-r+1}}{\delta}\right]}\Bigg) \leqslant \delta\,.$$

By a union bound argument, it thus remains to control the following sum to conclude:

$$\sum_{s=t_0}^{t-1}\sum_{r=t_0-1}^{s-1}\sqrt{t-r+1} \leqslant \sum_{s=t_0}^{t-1}(s-t_0+1)\sqrt{t-t_0+2} = \sum_{s=1}^{t-t_0} s\sqrt{t-t_0+2}\,.$$

□

Proof of Theorem 6:

First of all, (i) is a direct consequence of the time uniform concentration inequality. In order to prove (ii), let us first consider the case when $t_0 \leqslant \tau_{c-1}+1$. Since the algorithm checks for a detection event for each starting time r, then a test is performed in particular for $r = \tau_{c-1} \geqslant t_0 - 1$. Since only observations from time $\tau_{c-1}+1$ are considered by such a test, the detection delay is thus at most $d(\tau_{c-1}+1, \tau_c+1, \Delta_c)$ in this case. Now when $\tau_c + 1 \geqslant t_0 > \tau_{c-1}+1$, then observations from previous pieces do not perturb the detection started at time t_0, and thus the delay is at most $d(t_0, \tau_c+1, \Delta_c)$. □

Proof of Theorem 12:

We restart from point iii) in Theorem 6, that shows no change is detected before time t (with high probability), if the change at time $\tau+1$ is of magnitude $\Delta \leqslant \overline{\Delta}(t_0, \tau+1, t)$. Let μ_1 be the mean before the change point $\tau+1$, μ_2 the mean on $[\tau+1, \tau_c]$, and μ_3 the mean after τ_c. If there were no change point at $\tau+1$, the mean of all observations on $[t_0, \tau_c]$ would simply be μ_2. Due to the undetectable change point $\tau+1$, it is instead $\mu_2 + \frac{\tau-t_0+1}{\tau_c-t_0+1}(\mu_1-\mu_2)$. We thus observe a perturbation of the mean μ_2. The maximal perturbation on the means that does not trigger a detection event can then be computed. Since τ is undetectable by assumption, $\Delta < \min_{t\in[\tau+1,\tau_c]} \overline{\Delta}(t_0, \tau+1, t) = \overline{\Delta}(t_0, \tau+1, \tau_c)$ and the maximal perturbation of the mean μ_2 that does not trigger a detection event up to time τ_c is thus

$$\max_{\tau} \frac{\tau-t_0+1}{\tau_c-t_0+1}\overline{\Delta}(t_0, \tau+1, \tau_c) = \max_{\tau\in[t_0,\tau_c)} \sigma\sqrt{\frac{(\tau-t_0+1)(\tau_c-t_0+2)}{(\tau_c-\tau)(\tau_c-t_0+1)^2} 8\ln\left[\frac{2(\tau_c-t_0)\sqrt{\tau_c-t_0+2}}{\delta}\right]}.$$

Now, under Assumption 2 and Assumption 4, we know that since τ is undetectable, it must be that $\tau_c - \tau \geqslant \eta(\tau_c - t_0 + 1)$. We can thus restrict the maximum in the above equality to such values of τ. We deduce that

$$\max_{\tau} \frac{\tau-t_0+1}{\tau_c-t_0+1}\overline{\Delta}(t_0, \tau+1, \tau_c) \leqslant \Gamma_\eta(t_0, \tau_c) \text{ where}$$

$$\Gamma_\eta(t_0, \tau_c) \quad = \quad \sigma\sqrt{\frac{(1-\eta)(\tau_c-t_0+2)}{\eta(\tau_c-t_0+1)^2} 8\ln\left[\frac{2(\tau_c-t_0)\sqrt{\tau_c-t_0+2}}{\delta}\right]}.$$

The maximal perturbation $\Gamma(t_0, \tau_c)$ of μ_2 is thus of order $\tilde{O}(\frac{\sigma}{\sqrt{\eta(\tau_c-t_0+1)}})$, and the detection delay for τ_c of the procedure started at time t_0 is at most $d(\tau+1, \tau_c+1, \Delta_c - \Gamma_\eta(t_0, \tau_c))$. □