2

Change Detection Algorithms

In this chapter, we describe the simplest change detection algorithms. We consider a sequence of *independent* random variables $(y_k)_k$ with a probability density $p_\theta(y)$ depending upon only one *scalar* parameter. Before the *unknown change time* t_0, the parameter θ is equal to θ_0, and after the change it is equal to $\theta_1 \neq \theta_0$. The problems are then to detect and estimate this change in the parameter.

The main **goal** of this chapter is to introduce the reader to the design of *on-line* change detection algorithms, basically assuming that the parameter θ_0 before change is *known*. We start from elementary algorithms originally derived using an intuitive point of view, and continue with conceptually more involved but practically not more complex algorithms. In some cases, we give several possible derivations of the same algorithm. But the key point is that we introduce these algorithms within a general statistical framework, based upon likelihood techniques, which will be used throughout the book. Our conviction is that the early introduction of such a general approach in a simple case will help the reader to draw up a unified mental picture of change detection algorithms in more complex cases. In the present chapter, using this general approach and for this simplest case, we describe several on-line algorithms of increasing complexity. We also discuss the *off-line* point of view more briefly. The main example, which is carried through this chapter, is concerned with the detection of a change in the mean of an independent Gaussian sequence.

The **tools** for reaching this goal are as follows. First, our description of all the algorithms of this chapter is based on a concept that is very important in mathematical statistics, namely the logarithm of the likelihood ratio, defined by

$$s(y) = \ln \frac{p_{\theta_1}(y)}{p_{\theta_0}(y)} \tag{2.0.1}$$

and referred to as the log-likelihood ratio. The key statistical property of this ratio is as follows : Let $\mathbf{E}_{\theta_0}$ and $\mathbf{E}_{\theta_1}$ denote the expectations of the random variables under the two distributions p_{θ_0} and p_{θ_1}, respectively. Then,

$$\mathbf{E}_{\theta_0}(s) < 0 \ \text{ and } \ \mathbf{E}_{\theta_1}(s) > 0 \tag{2.0.2}$$

In other words, *a change in the parameter θ is reflected as a change in the sign of the mean value of the log-likelihood ratio*. This property can be viewed as a kind of detectability of the change. Because the Kullback information $\mathbf{K}$ is defined by $\mathbf{K}(\theta_1, \theta_0) = \mathbf{E}_{\theta_1}(s)$, we also have that the difference between the two mean values is

$$\mathbf{E}_{\theta_1}(s) - \mathbf{E}_{\theta_0}(s) = \mathbf{K}(\theta_1, \theta_0) + \mathbf{K}(\theta_0, \theta_1) > 0 \tag{2.0.3}$$

From this, we deduce that the detectability of a change can also be defined with the aid of the Kullback information between the two models before and after change. These concepts are used throughout the book.

Second, even for this simple case, it is of interest to classify all possible practical problem statements with respect to two different issues :

- The first possible classification is with respect to assumptions about the unknown change time t_0. In some applications, it is useful to consider t_0 as a nonrandom unknown value, or a random unknown value with unknown distribution. In other words, we deal with a nonparametric approach as far as this change time t_0 is concerned. This assumption is useful because very often in practice, either it is very difficult to have *a priori* information about the distribution of the change times, or this distribution is nonstationary. This point of view is taken in sections 2.1, 2.2, and 2.4 for on-line algorithms and in section 2.6 for off-line algorithms. In some applications, it is possible to use *a priori* information about the distribution of the change time, taking a Bayesian point of view. Such *a priori* information can be available from life-time estimations made in reliability investigations. This point of view is used in section 2.3.
- The second possible classification of algorithms is with respect to the available information about the value θ_1 of the parameter after change, as we discussed in section 1.4. We first consider that this value is known : This is the case of sections 2.1, 2.2, and 2.3. The case of unknown value for θ_1 is investigated in section 2.4 for on-line algorithms and in section 2.6 for off-line algorithms.

Before proceeding, let us add one comment concerning the performances of these algorithms and the detectability of a given change. The criteria for the performance evaluation of these algorithms were introduced in section 1.4 from an intuitive point of view. The performances of the *on-line* algorithms presented in the present chapter are investigated in detail in chapter 5 with the aid of the formal definition of these criteria, given in section 4.4. These performance evaluations can be computationally complex, even in the present simple case. For this reason, it is also of interest to consider a kind of weak performance index, the positivity of which simply states the detectability of a change (with no more indication on the properties of the detection). The Kullback information is a good candidate for such a weak index, both because of the above-mentioned inequalities and because, as shown in chapter 4, it is an adequate index of separability between two probability measures. This mutual information is zero only when the parameters are equal, and can be shown to be an increasing function of the Euclidean distance between the parameters θ_0 and θ_1 when this distance is small. This detectability definition is investigated in detail in more complex cases in chapters 7, 8, and 9.

2.1 Elementary Algorithms

In this section, we describe several simple and well-known algorithms. Most of the algorithms presented here work on samples of data with *fixed* size; only one uses a growing memory. In the next section, we deal basically with a random-size sliding window algorithm. In quality control, these elementary algorithms are usually called *Shewhart control charts* and finite or infinite *moving average control charts*. We also introduce another elementary algorithm, called a *filtered derivative* algorithm, which is often used in image edge detection. We place these algorithms in our general likelihood framework, and consider the case in which the only unknown value is the change time t_0. Recall that all the key mathematical concepts are described in chapters 3 and 4.

2.1.1 Limit Checking Detectors and Shewhart Control Charts

Let us first introduce the initial idea used in quality control under the name of continuous inspection. Samples with fixed size N are taken, and at the end of each sample a decision rule is computed to test between

the two following hypotheses about the parameter θ :

$$\begin{aligned} \mathbf{H}_0 &: \theta = \theta_0 \\ \mathbf{H}_1 &: \theta = \theta_1 \end{aligned} \tag{2.1.1}$$

As long as the decision is taken in favour of $\mathbf{H}_0$, the sampling and test continue. Sampling is stopped after the first sample of observations for which the decision is taken in favor of $\mathbf{H}_1$.

We introduce the following notation, which is used throughout this and the subsequent chapters. Let

$$\begin{aligned} S_j^k &= \sum_{i=j}^{k} s_i \\ s_i &= \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \end{aligned} \tag{2.1.2}$$

be the log-likelihood ratio for the observations from y_j to y_k. We refer to s_i as the *sufficient statistic* for reasons that are explained in section 4.1.

The following statement is a direct consequence of the Neyman-Pearson lemma, which we recall in chapter 4. For a fixed sample size N, the optimal decision rule d is given by

$$d = \begin{cases} 0 & \text{if} \quad S_1^N < h; \quad \mathbf{H}_0 \text{ is chosen} \\ 1 & \text{if} \quad S_1^N \geq h; \quad \mathbf{H}_1 \text{ is chosen} \end{cases} \tag{2.1.3}$$

where h is a conveniently chosen threshold. The sum S_1^N is said to be the *decision function.* The decision is taken with the aid of what is called a stopping rule, which in this case is defined by

$$t_a = N \cdot \min\{K : d_K = 1\} \tag{2.1.4}$$

where d_K is the decision rule for the sample number K (of size N) and t_a is the *alarm time.* In other words, the observation is stopped after the first sample of size N for which the decision is in favor of $\mathbf{H}_1$.

Example 2.1.1 (Change in mean). *Let us now consider the particular case where the distribution is Gaussian with mean value μ and constant variance σ^2. In this case, the changing parameter θ is μ. The probability density is*

$$p_\theta(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \tag{2.1.5}$$

and the sufficient statistic s_i is

$$s_i = \frac{\mu_1 - \mu_0}{\sigma^2} \left(y_i - \frac{\mu_0 + \mu_1}{2} \right) \tag{2.1.6}$$

which we shall write as

$$\begin{aligned} s_i &= \frac{b}{\sigma} \left(y_i - \frac{\mu_0 + \mu_1}{2} \right) \\ &= \frac{b}{\sigma} \left(y_i - \mu_0 - \frac{\nu}{2} \right) \end{aligned} \tag{2.1.7}$$

where

$$\nu = \mu_1 - \mu_0 \tag{2.1.8}$$

is the change magnitude *and*

$$b = \frac{\mu_1 - \mu_0}{\sigma} \tag{2.1.9}$$

is the signal-to-noise ratio. *Therefore, the decision function (2.1.2) is*

$$S_1^N = \frac{b}{\sigma} \sum_{i=1}^{N} \left(y_i - \mu_0 - \frac{\nu}{2} \right) \tag{2.1.10}$$

The stopping rule for the change detection algorithm is as in (2.1.4), with the decision rule defined by

$$d = \begin{cases} 0 & \text{if} \quad S_1^N(K) < h \\ 1 & \text{if} \quad S_1^N(K) \geq h \end{cases} \tag{2.1.11}$$

where

$$S_1^N(K) = S_{N(K-1)+1}^{NK} \tag{2.1.12}$$

with S_i^j defined in (2.1.2). This change detection algorithm is one of the oldest and most well-known algorithms for continuous inspection, and is called Shewhart control chart *[Shewhart, 1931]. For this control chart, when $\mu_1 > \mu_0$, the alarm is set the first time at which*

$$\bar{y}(K) \geq \mu_0 + \kappa \frac{\sigma}{\sqrt{N}} \tag{2.1.13}$$

where

$$\bar{y}(K) = \frac{1}{N} \sum_{i=N(K-1)+1}^{NK} y_i \tag{2.1.14}$$

Note that the threshold is related to the standard deviation of the left side of this inequality. This stopping rule is standard in quality control, where the name for the right side of this inequality is the upper control limit. *The tuning parameters of this Shewhart control chart are κ and N. The behavior of this chart, when applied to the signal of figure 1.1, is depicted in figure 2.1.*

It is often more useful to detect deviations from μ_0 in both directions, namely increases and decreases. In this case, assume that the mean value after the change is either $\mu_1^+ = \mu_0 + \nu$ or $\mu_1^- = \mu_0 - \nu$. Then the alarm is set the first time at which

$$|\bar{y}(K) - \mu_0| \geq \kappa \frac{\sigma}{\sqrt{N}} \tag{2.1.15}$$

where $\mu_0 - \kappa \frac{\sigma}{\sqrt{N}}$ is the lower control limit. *This is depicted in the figure 2.2. The tuning parameters of this algorithm are κ and N again. The optimal tuning of these parameters can be obtained with the aid of an a priori information concerning the change magnitude ν.*

Let us add one comment about a slightly different use of control charts [S.Roberts, 1966]. To prevent false alarms and to obtain more reliable detection results, the intuitive idea consists of deciding a change when a preassigned number of crossings in (2.1.15) occur among several successive data samples of size N. This idea is known as a *run test* in quality control, and sometimes as a *counter* in the engineering literature. Various types of run tests have been used to supplement Shewhart control charts, as explained in [S.Roberts, 1966]. A similar idea is also used for another change detection algorithm in subsection 2.1.4.

2.1.2 Geometric Moving Average Control Charts

Two key ideas underlie the geometric moving average (GMA) algorithm. The first idea is related to the above-mentioned behavior of the log-likelihood ratio (2.0.1). The second deals with the widespread intuitive idea of exponential weighting of observations. As usual in nonstationary situations, because of the unknown

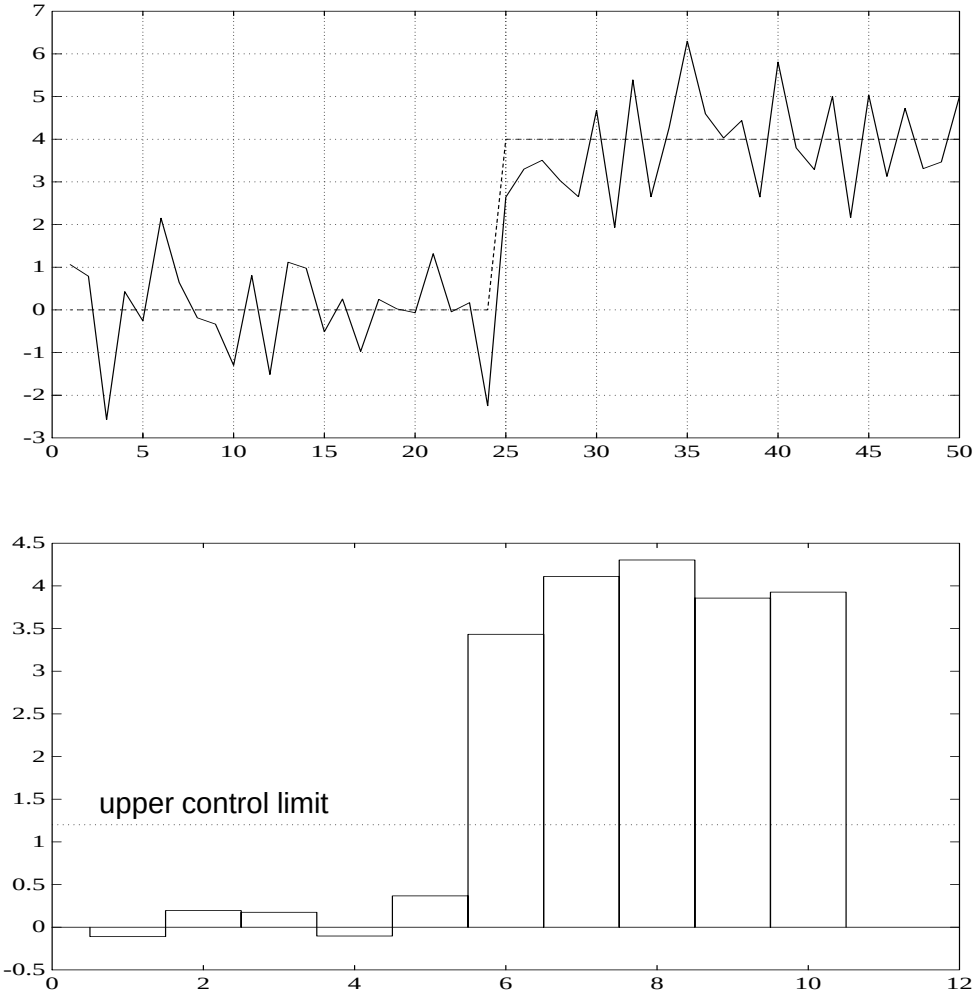

Figure 2.1 A Shewhart control chart corresponding to a change in the mean of a Gaussian sequence with constant variance.

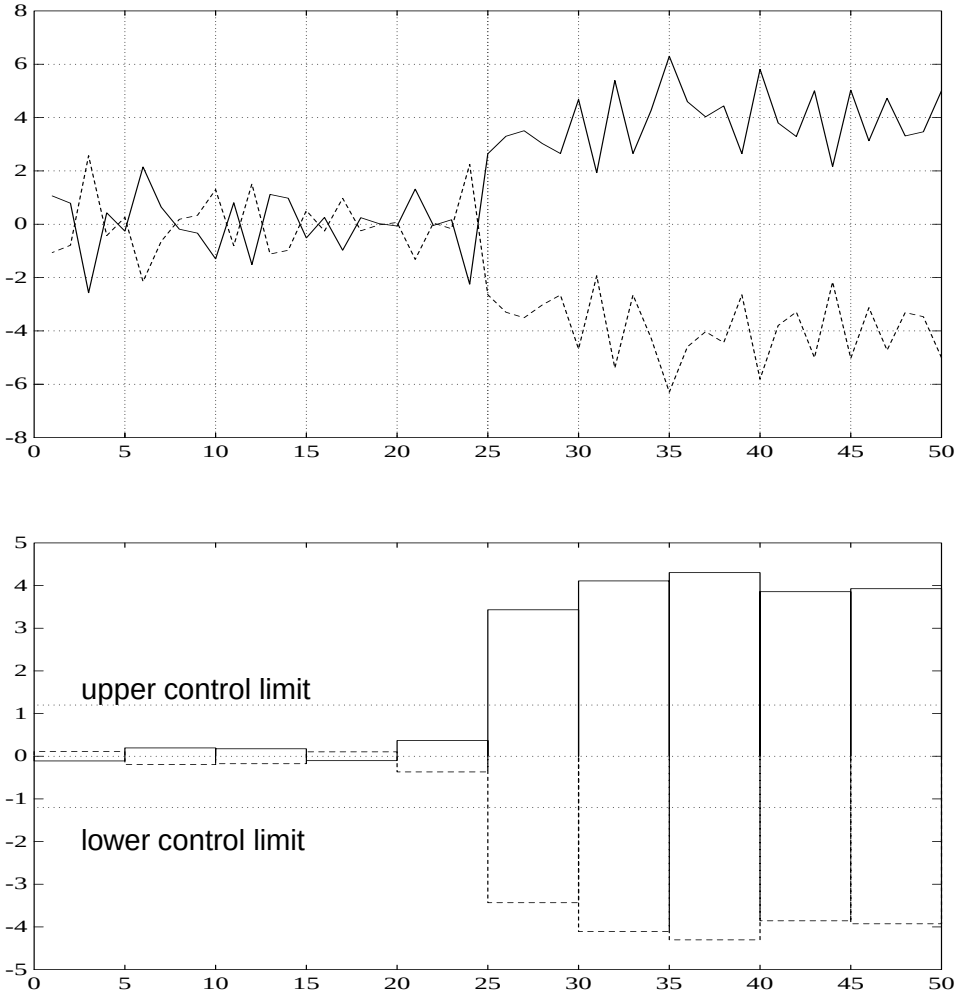

Figure 2.2 A two-sided Shewhart control chart.

change time t_0, it is of interest to use higher weights on recent observations and lower weights on past ones. Therefore, the following decision function is relevant [S.Roberts, 1959, Hines, 1976a, Hines, 1976b] :

$$\begin{aligned} g_k &= \sum_{i=0}^{\infty} \gamma_i \ln \frac{p_{\theta_1}(y_{k-i})}{p_{\theta_0}(y_{k-i})} \\ &= \sum_{i=0}^{\infty} \gamma_i s_{k-i} \end{aligned} \tag{2.1.16}$$

where the weights γ_i are exponential, namely

$$\gamma_i = \alpha(1-\alpha)^i, \, 0 < \alpha \leq 1 \tag{2.1.17}$$

The coefficient α acts as a forgetting factor. This decision function can be rewritten in a recursive manner as

$$g_k = (1-\alpha)\, g_{k-1} + \alpha\, s_k, \;\; \text{with:} \;\; g_0 = 0 \tag{2.1.18}$$

The alarm time is defined by the following stopping rule :

$$t_a = \min\{k : g_k \geq h\} \tag{2.1.19}$$

where h is a conveniently chosen threshold.

Example 2.1.2 (Change in mean - contd.). *In the case of a change in the mean of an independent Gaussian sequence, s_k is given by (2.1.6), and the GMA decision function is*

$$\tilde{g}_k = (1-\alpha)\, \tilde{g}_{k-1} + \alpha\, (y_k - \mu_0), \;\; \textit{with:} \;\; \tilde{g}_0 = 0 \tag{2.1.20}$$

where $\tilde{g}$ and g are related through

$$\tilde{g}_k = \frac{\sigma^2}{\mu_1 - \mu_0}\, g_k - \frac{\mu_1 - \mu_0}{2} \tag{2.1.21}$$

The behavior of this decision function, when applied to the signal of figure 1.1, is depicted in figure 2.3. In the corresponding two-sided situation, the stopping rule is

$$t_a = \min\{k : |\tilde{g}_k| \geq h\} \tag{2.1.22}$$

Example 2.1.3 (Change in variance). *In the case of a change in the variance σ^2, which is relevant in quality control, as explained in example 1.2.1, we have*

$$s_k = \ln \frac{\sigma_0}{\sigma_1} + \left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2} \right) \frac{(y_k - \mu)^2}{2} \tag{2.1.23}$$

Therefore, the relevant decision function can be written as

$$\tilde{g}_k = \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2}\, g_k - \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2} \ln \frac{\sigma_0}{\sigma_1} \tag{2.1.24}$$

where g_k is defined in (2.1.18). In a recursive form, this becomes

$$\tilde{g}_k = (1-\alpha)\, \tilde{g}_{k-1} + \alpha\, (y_k - \mu)^2, \;\; \textit{with:} \;\; \tilde{g}_0 = 0 \tag{2.1.25}$$

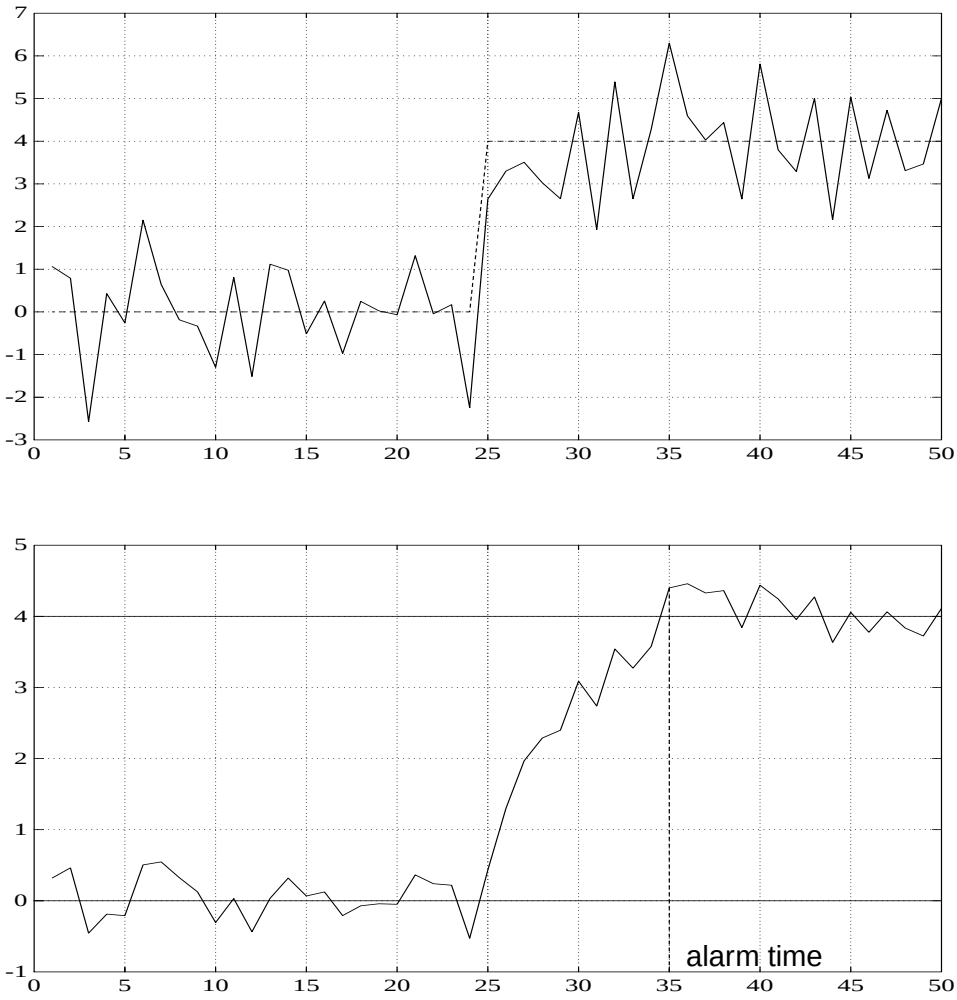

Figure 2.3 A geometric moving average algorithm corresponding to a change in the mean of a Gaussian sequence with constant variance.

2.1.3 Finite Moving Average Control Charts

A similar idea to the previous control charts consists in replacing the exponential forgetting operation by a finite memory one, and thus in using a finite set of weights, which are no longer assumed to form a geometric sequence. For defining this new detector, which is called finite moving average (FMA) algorithm, let us follow the derivation of the geometric moving average control charts. First, consider the following variant of the causal filtering (2.1.16) used in these charts :

$$g_k = \sum_{i=0}^{N-1} \gamma_i \ln \frac{p_{\theta_1}(y_{k-i})}{p_{\theta_0}(y_{k-i})} \tag{2.1.26}$$

where the weights γ_i are any weights for causal filters. The stopping rule is as in the previous control chart :

$$t_a = \min\{k : g_k \geq h\} \tag{2.1.27}$$

Example 2.1.4 (Change in mean - contd.). *In the case of an* increase *in the mean, this stopping rule can be computed as follows. Using (2.1.6), the decision function g_k in (2.1.26) can be expressed as*

$$g_k = \sum_{i=0}^{N-1} \gamma_i (y_{k-i} - \mu_0) \tag{2.1.28}$$

In the two-sided case, g_k is the same, and the stopping rule is

$$t_a = \min\{k : |g_k| \geq h\} \tag{2.1.29}$$

2.1.4 Filtered Derivative Algorithms

In the case of a change in the mean of a Gaussian sequence, the filtered derivative algorithms are based on the following very intuitive idea. Ideally, that is, in a no noise situation, a change in the mean level of a sequence of observations is locally characterized by a great absolute value of the (discrete) derivative of the sample observations. Because the derivative operator acts in a very poor manner as soon as noise is present in observations, a more realistic detector should use a filtering operation before derivation. This explains the title of this subsection. The typical behavior of this algorithm is depicted in figure 2.4 for the ideal and realistic situations. Now, because of the smoothing operation on the jump, several alarms are to occur in the neighborhood of t_0. An elementary way to increase the robustness of this detector is to count the number of threshold crossings during a fixed time interval before deciding the change actually occurred.

Let us now put this intuition-based detector into our more formal framework for change detection algorithms. We use again the derivation of the finite moving average control charts :

$$g_k = \sum_{i=0}^{N-1} \gamma_i \ln \frac{p_{\theta_1}(y_{k-i})}{p_{\theta_0}(y_{k-i})} \tag{2.1.30}$$

where the weights γ_i are again any weights for causal filters, and we consider the discrete derivative of g_k :

$$\nabla g_k = g_k - g_{k-1} \tag{2.1.31}$$

and the following stopping rule :

$$t_a = \min\{k : \sum_{i=0}^{N-1} \mathbf{1}_{\{\nabla g_{k-i} \geq h\}} \geq \eta\} \tag{2.1.32}$$

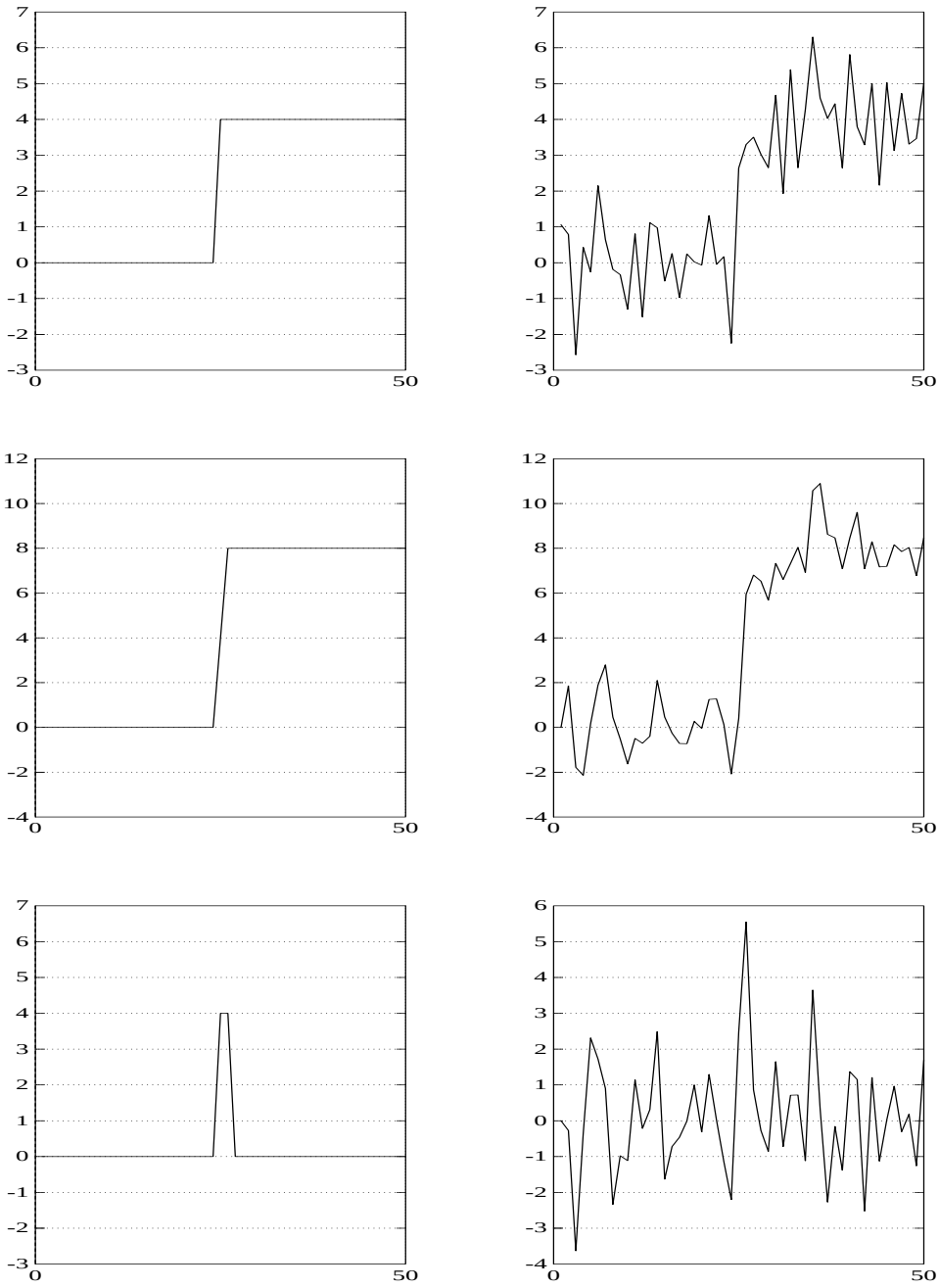

Figure 2.4 Ideal (left) and realistic (right) behaviors of a filtered derivative algorithm corresponding to a change in the mean of a Gaussian sequence with constant variance : signal (first row), filtered signal (second row), and filtered and derivate signal (third row).

where $\mathbf{1}_{\{x\}}$ is the indicator of event $\{x\}$. In this formula, h is the threshold for the derivative, and η is a threshold for the number of crossings of h. This threshold η is used for decreasing the number of alarms in the neighborhood of the change due to the smoothing operation. It turns out that, in practice, $\eta = 2$ is often a convenient value for achieving this goal.

Example 2.1.5 (Change in mean - contd.). *In the case of an* increase *in the mean, the decision function g_k corresponding to (2.1.30) can again be taken as*

$$g_k = \sum_{i=0}^{N-1} \gamma_i (y_{k-i} - \mu_0) \tag{2.1.33}$$

The stopping rule is as in (2.1.32). In the two-sided case of jump in mean in an unknown direction, the stopping rule is

$$t_a = \min\{k : \sum_{i=0}^{N-1} \mathbf{1}_{\{|\nabla g_{k-i}| \geq h\}} \geq \eta\} \tag{2.1.34}$$

Two elementary choices of smoothing filters in (2.1.30) are as follows :

- *An integrating filter with N constant unit weights γ_i, which results in*

$$\nabla g_k = y_k - y_{k-N}$$

- *A triangular filter with impulse response of triangular form, namely $\gamma_{p+i} = \gamma_{p-i} = i$ for $0 \leq i \leq p$, where $N - 1 = 2p$, which results in*

$$\nabla g_k = \sum_{i=0}^{p-1} y_{k-i} - \sum_{i=p}^{2p-1} y_{k-i}$$

In other words, the corresponding stopping rules are based upon the difference between either sample values or local averages of sample values.

2.2 CUSUM Algorithm

We now introduce the cumulative sum (CUSUM) algorithm, which was first proposed in [Page, 1954a]. We describe four different derivations. The first is more intuition-based, and uses ideas connected to a simple integration of signals with *adaptive threshold*. The second derivation is based on a more formal on-line statistical approach, similar to the approach used before for introducing control charts, and based upon a *repeated use of the sequential probability ratio test*. The third derivation comes from the use of the off-line point of view for a *multiple hypotheses testing* approach. This derivation is useful for the introduction of the geometrical interpretation of the CUSUM algorithm with the aid of a V-mask. The fourth derivation is based upon the concept of open-ended tests.

2.2.1 Intuitive Derivation

As we mentioned in the previous section, the typical behavior of the log-likelihood ratio S_k shows a negative drift before change, and a positive drift after change, as depicted in figure 2.5, again for the signal of figure 1.1. Therefore, the relevant information, as far as the change is concerned, lies in the difference

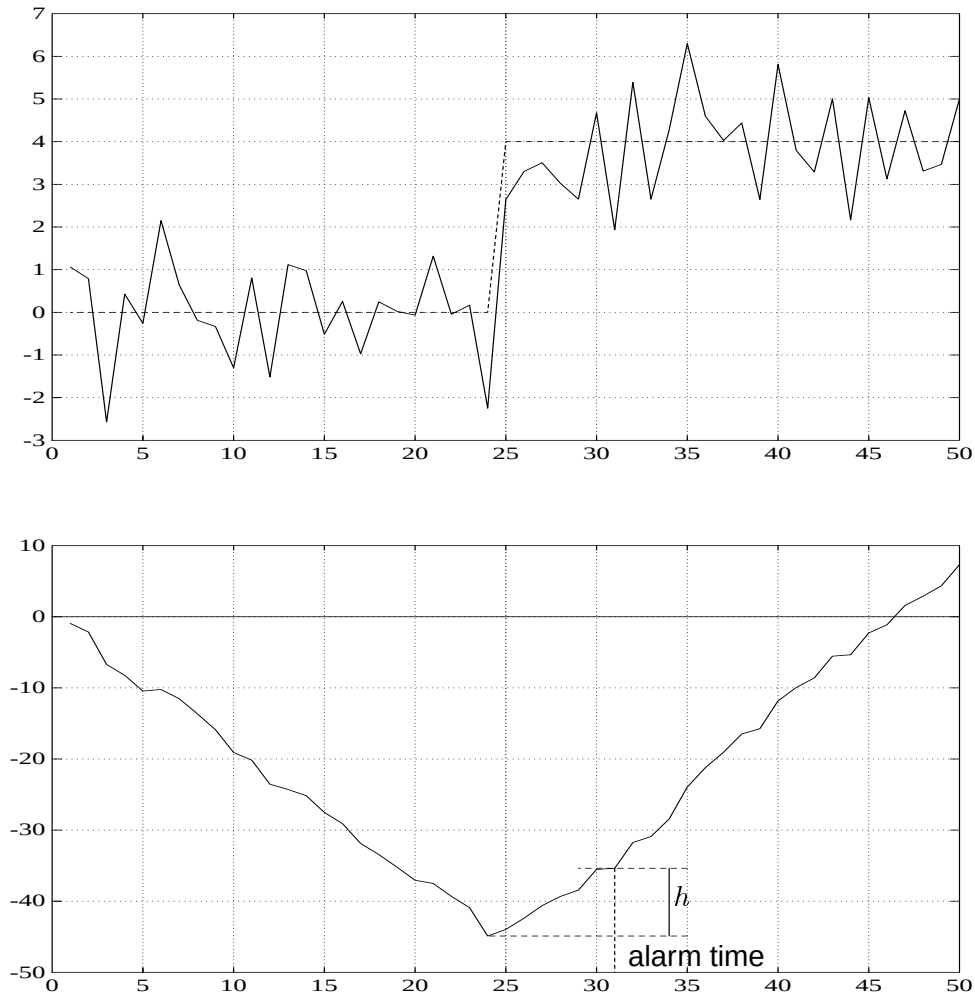

Figure 2.5 Typical behavior of the log-likelihood ratio S_k corresponding to a change in the mean of a Gaussian sequence with constant variance : negative drift before and positive drift after the change.

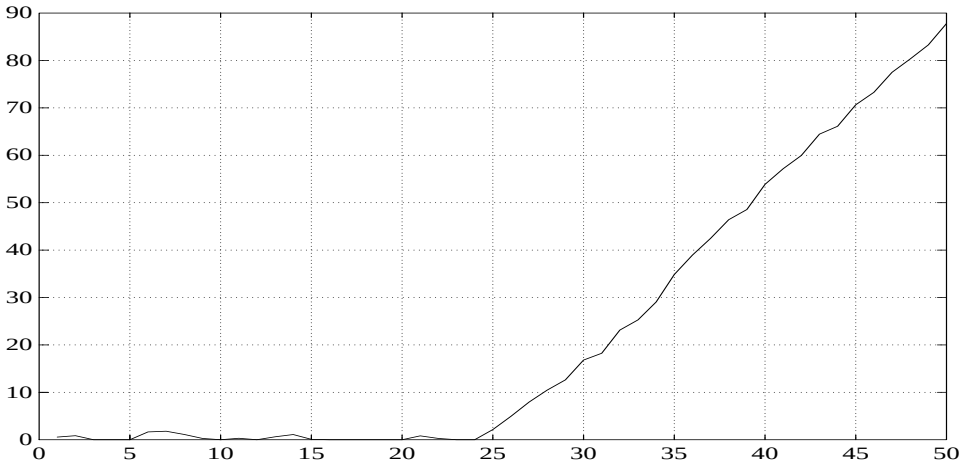

Figure 2.6 Typical behavior of the CUSUM decision function g_k.

between the value of the log-likelihood ratio and its current minimum value; and the corresponding decision rule is then, at each time instant, to compare this difference to a threshold as follows :

$$g_k = S_k - m_k \geq h \tag{2.2.1}$$

where

$$\begin{aligned} S_k &= \sum_{i=1}^{k} s_i \\ s_i &= \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \\ m_k &= \min_{1 \leq j \leq k} S_j \end{aligned} \tag{2.2.2}$$

The typical behavior of g_k is depicted in figure 2.6. The stopping time is

$$t_a = \min\{k : g_k \geq h\} \tag{2.2.3}$$

which can be obviously rewritten as

$$t_a = \min\{k : S_k \geq m_k + h\} \tag{2.2.4}$$

Now it becomes clear that this detection rule is nothing but a comparison between the cumulative sum S_k and an *adaptive threshold* $m_k + h$. Because of m_k, this threshold not only is modified on-line, but also keeps *complete* memory of the entire information contained in the past observations. Moreover, it is obvious from (2.1.6) that, in the case of change in the mean of a Gaussian sequence, S_k is a standard *integration* of the observations.

2.2.2 CUSUM Algorithm as a Repeated Sequential Probability Ratio Test

Page suggested the use of repeated testing of the two simple hypotheses :

$$\mathbf{H}_0 \quad : \quad \theta = \theta_0 \tag{2.2.5}$$

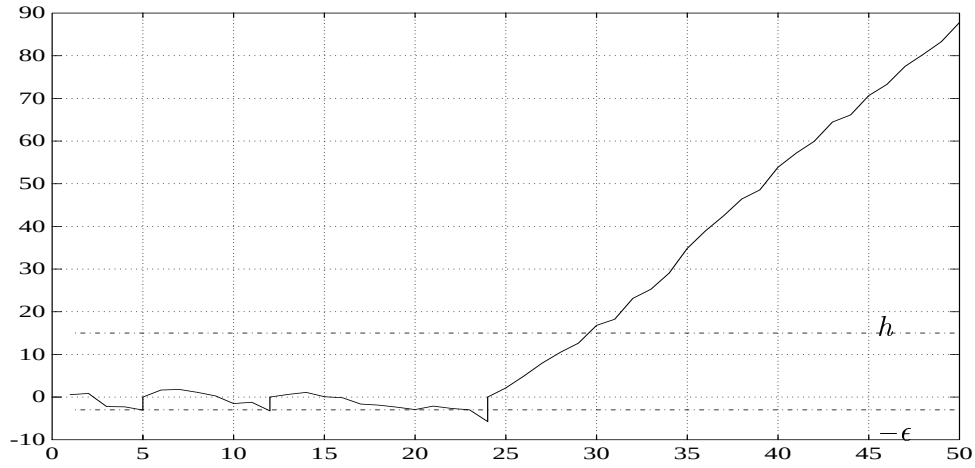

Figure 2.7 Repeated use of SPRT. $T_i = 5, 12, 24,$ and 30 are the stopping times in each successive cycle, and $d_i = 0, 0, 0,$ and 1 are the corresponding decision rules.

$$\mathbf{H}_1 \quad : \quad \theta = \theta_1$$

with the aid of the *sequential probability ratio test (SPRT)*. Let us first define a single use of the SPRT algorithm. The SPRT is defined with the aid of the pair (d, T) where d is the decision rule and T is a stopping time, exactly as the Neyman-Pearson rule is defined with the aid of the decision rule d. The stopping time T is the time at which the final decision is taken and thus at which observation is stopped. The definition of the SPRT is thus

$$d = \begin{cases} 0 & \text{if} \quad S_1^T \leq -\epsilon \\ 1 & \text{if} \quad S_1^T \geq h \end{cases} \tag{2.2.6}$$

where T is the exit time :

$$T = T_{-\epsilon,h} = \min\{k : (S_1^k \geq h) \cup (S_1^k \leq -\epsilon)\} \tag{2.2.7}$$

where $\epsilon \geq 0$ and $h > 0$ are conveniently chosen thresholds. Now, as in section 2.1, we use repeated SPRT until the decision $d = 1$ is taken. The typical behavior of this repeated use of the SPRT is depicted in figure 2.7, where $T_i = 5, 12, 24,$ and 30 are the stopping times in each successive cycle, and $d_i = 0, 0, 0,$ and 1 are the corresponding decision rules. The key idea of Page was to *restart the SPRT algorithm as long as the previously taken decision is* $d = 0$. The first time at which $d = 1$, we stop observation and do not restart a new cycle of the SPRT. This time is then the *alarm time* at which the change is detected.

Using an intuitive motivation, Page suggested that the optimal value of the lower threshold ϵ should be zero. This statement was formally proven later [Shiryaev, 1961, Lorden, 1971, Moustakides, 1986, Ritov, 1990] and is discussed in section 5.2. Starting from the repeated SPRT with this value of lower threshold, the resulting decision rule can be rewritten in a recursive manner as

$$g_k = \begin{cases} g_{k-1} + \ln \frac{p_{\theta_1}(y_k)}{p_{\theta_0}(y_k)} & \text{if} \quad g_{k-1} + \ln \frac{p_{\theta_1}(y_k)}{p_{\theta_0}(y_k)} > 0 \\ 0 & \text{if} \quad g_{k-1} + \ln \frac{p_{\theta_1}(y_k)}{p_{\theta_0}(y_k)} \leq 0 \end{cases} \tag{2.2.8}$$

where $g_0 = 0$. Remembering the definition of s_k in (2.1.2), this can be compacted into

$$g_k = (g_{k-1} + s_k)^+ \tag{2.2.9}$$

where $(x)^+ = \sup(0, x)$. Finally, the stopping rule and alarm time are defined by

$$t_a = \min\{k : g_k \geq h\} \tag{2.2.10}$$

where g_k is given in (2.2.9). The typical behavior of this decision function is depicted in figure 2.6. It is easy to prove that this form of decision rule is equivalent to the other form that we presented in (2.2.4). On the other hand, it can also be written as

$$g_k = \left(S_{k-N_k+1}^k\right)^+ \tag{2.2.11}$$

where

$$N_k = N_{k-1} \cdot \mathbf{1}_{\{g_{k-1}>0\}} + 1 \tag{2.2.12}$$

$\mathbf{1}_{\{x\}}$ is the indicator of event x, and t_a is defined in (2.2.10). In this formula, N_k is the number of observations after re-start of the SPRT. The formula (2.2.11) can be interpreted as an integration of the observations over a *sliding window with random size*. This size is chosen according to the behavior of the entire past observations.

2.2.3 Off-line Statistical Derivation

As we discussed in chapter 1, when taking an off-line point of view, it is convenient to introduce the following hypotheses about the observations $y_1, ..., y_k$:

$$\begin{array}{rrlll} & \mathbf{H}_0: & \theta = \theta_0 & \text{for} & 1 \leq i \leq k \\ \text{for } 1 \leq j \leq k, & \mathbf{H}_j: & \theta = \theta_0 & \text{for} & 1 \leq i \leq j-1 \\ & & \theta = \theta_1 & \text{for} & j \leq i \leq k \end{array} \tag{2.2.13}$$

The likelihood ratio between the hypotheses $\mathbf{H}_0$ and $\mathbf{H}_j$ is

$$\Lambda_1^k(j) = \frac{\prod_{i=1}^{j-1} p_{\theta_0}(y_i) \cdot \prod_{i=j}^{k} p_{\theta_1}(y_i)}{\prod_{i=1}^{k} p_{\theta_0}(y_i)} \tag{2.2.14}$$

(where $\prod_{i=1}^{0} = 1$). Thus, the log-likelihood ratio is

$$S_j^k = \sum_{i=j}^{k} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \tag{2.2.15}$$

When the change time j is unknown, the standard statistical approach consists of estimating it by using the maximum likelihood principle, which leads to the following decision function :

$$g_k = \max_{1 \leq j \leq k} S_j^k \tag{2.2.16}$$

This decision function is the same as those obtained in formulas (2.2.4) and (2.2.9). It can also be written as

$$t_a = \min\{k : \max_{1 \leq j \leq k} S_j^k \geq h\} \tag{2.2.17}$$

Up to now, we have discussed only the *detection* issue in change detection problems. Let us now consider the *estimation of the change time* t_0. It follows from equation (2.2.16) that the maximum likelihood estimate of t_0 *after detection* is equal to the time j at which the maximum in (2.2.16) is reached. This estimate can be computed using the following formula :

$$\hat{t}_0 = t_a - N_{t_a} + 1 \tag{2.2.18}$$

We discuss this formula in section 2.6.

Example 2.2.1 (Change in mean - contd.). *We now continue the discussion about the simple example of a change in the mean value μ of an independent Gaussian random sequence, with known variance σ^2. We first consider the one-sided case of an increase in the mean, namely $\mu_1 > \mu_0$. In this case, (2.1.6) holds, and the decision function g_k introduced in (2.2.1), (2.2.9), and (2.2.16) becomes in the first formulation,*

$$\begin{aligned} g_k &= S_1^k - \min_{1\leq j\leq k} S_1^j \\ S_1^j &= \frac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^{j}\left(y_i - \frac{\mu_1+\mu_0}{2}\right) \end{aligned} \tag{2.2.19}$$

and in the second formulation,

$$g_k = \left[g_{k-1} + \frac{\mu_1-\mu_0}{\sigma^2}\left(y_k - \frac{\mu_1+\mu_0}{2}\right)\right]^+ \tag{2.2.20}$$

and finally

$$g_k = \max_{1\leq j\leq k} S_j^k \tag{2.2.21}$$

in the third formulation. It is obvious from the formula for S_1^j that the observations are first processed through an ordinary integration; and then, as stated before, an adaptive threshold is used.

2.2.4 Parallel Open-ended Tests

Now let us emphasize the connection between formulas (2.2.15)-(2.2.17) and an idea due to [Lorden, 1971] which turns out to be very useful for the design and the analysis of change detection algorithms. The CUSUM stopping time t_a can be interpreted using a set of *parallel* so-called open-ended SPRT, which are activated at each possible change time $j = 1, \ldots, k$, and with upper threshold h and lower threshold $-\epsilon = -\infty$. Each of these SPRT stops at time k if, for some $j \leq k$, the observations $y_j, \ldots, y_k$ are significant for accepting the hypothesis about change. Let us formalize this in the following way. Let T_j be the stopping time for the open-ended SPRT activated at time j :

$$T_j = \min\{k \geq j \ : S_j^k \geq h\} \tag{2.2.22}$$

where we use the convention that $T_j = \infty$ when this minimum is never reached. Lorden defined the following *extended stopping time* as the minimum of the T_j :

$$T^* = \min_{j=1,2,\ldots}\{T_j\} \tag{2.2.23}$$

The comparison between (2.2.17) and (2.2.22)-(2.2.23) shows that $t_a = T^*$. We continue this discussion when describing the geometrical interpretation after.

2.2.5 Two-sided CUSUM Algorithm

Let us now investigate further the situation discussed in section 2.1 where the mean value after change is either $\mu_1^+ = \mu_0 + \nu$ or $\mu_1^- = \mu_0 - \nu$, with ν known. In this case, it is relevant [Page, 1954a] to use two CUSUM algorithms together; the first for detecting an increase in the mean, and the second for detecting a decrease in the mean. The resulting alarm time is

$$\begin{aligned} t_a &= \min\{k : (g_k^+ \geq \bar{h}) \cup (g_k^- \geq \bar{h})\} \\ g_k^+ &= \left(g_{k-1}^+ + y_k - \mu_0 - \frac{\nu}{2}\right)^+ \\ g_k^- &= \left(g_{k-1}^- - y_k + \mu_0 - \frac{\nu}{2}\right)^+ \end{aligned} \tag{2.2.24}$$

In these formulas, we canceled the multiplicative term $\frac{\mu_1-\mu_0}{\sigma^2}$, which can be incorporated in the threshold $\bar{h}$ in an obvious manner. Formula (2.2.24) corresponds to the well-known *cumulative sum control chart* widely used in continuous inspection for quality control.

Let us add some comments about ν. When introducing this chapter, we discussed the availability of information about θ_1, or, equivalently from an on-line point of view, about the change magnitude ν. In most practical cases, little is known about this parameter. However, three possible *a priori* choices can be made for using the CUSUM algorithm in this case. The first consists of choosing ν as a minimum possible magnitude of jump. In the second, we choose *a priori* the most likely magnitude of jump. The third choice for ν is a kind of worst-case value from the point of view of the cost of a nondetected change. In these three cases, the resulting change detection algorithm is optimal for only *one* possible jump magnitude equal to ν. Notice that an *a posteriori* choice of the most likely magnitude leads to the GLR algorithm, which is introduced in subsection 2.4.3, and leads to the almost optimal algorithm in such a case.

From the point of view of minimum magnitude of change, the limit case is $\nu = 0$. In other words, this situation occurs when all possible jumps are to be detected, whatever their magnitude. It is useful to note [Nadler and Robbins, 1971] that, for this situation, the double CUSUM algorithm presented before in formula (2.2.24) is equivalent to

$$t_a = \min\{k : R_k \geq \bar{h}\} \tag{2.2.25}$$

where

$$R_k = \max_{j \leq k} \sum_{i=1}^{j} (y_i - \mu_0) - \min_{j \leq k} \sum_{i=1}^{j} (y_i - \mu_0) \tag{2.2.26}$$

2.2.6 Geometrical Interpretation in the Gaussian Case

If we rewrite the decision function (2.2.21), we obtain

$$g_k = \max_{1 \leq j \leq k} \sum_{i=j}^{k} \left(y_i - \mu_0 - \frac{\nu}{2} \right) \tag{2.2.27}$$

In the corresponding decision rule, the alarm is set the first time k at which there exists a time instant j_0 such that

$$\sum_{i=j_0}^{k} \left(y_i - \mu_0 - \frac{\nu}{2} \right) \geq \bar{h} \tag{2.2.28}$$

At each time k, this can be seen as a SPRT with reverse time and only one (upper) threshold $\bar{h}$ [Lorden, 1971, Page, 1954a]. For this purpose, look at figure 2.8 upside down. This can be geometrically interpreted, as depicted in figure 2.9. In this figure the cumulative sum

$$\tilde{S}_1^k = \frac{1}{\sigma} \sum_{i=1}^{k} (y_i - \mu_0) \tag{2.2.29}$$

is plotted in the case $\mu_0 = 0$. Because this cumulative sum does not contain the term $-\frac{\nu}{2}$, the corresponding threshold is no longer a constant value, but a straight line with slope $\omega \tan(\alpha)$, where ω is the horizontal distance between successive points in terms of a unit distance on the vertical scale, and α is the angle between this line and the horizontal one. It is obvious that

$$\tan(\alpha) = \frac{\nu}{2\omega} \tag{2.2.30}$$

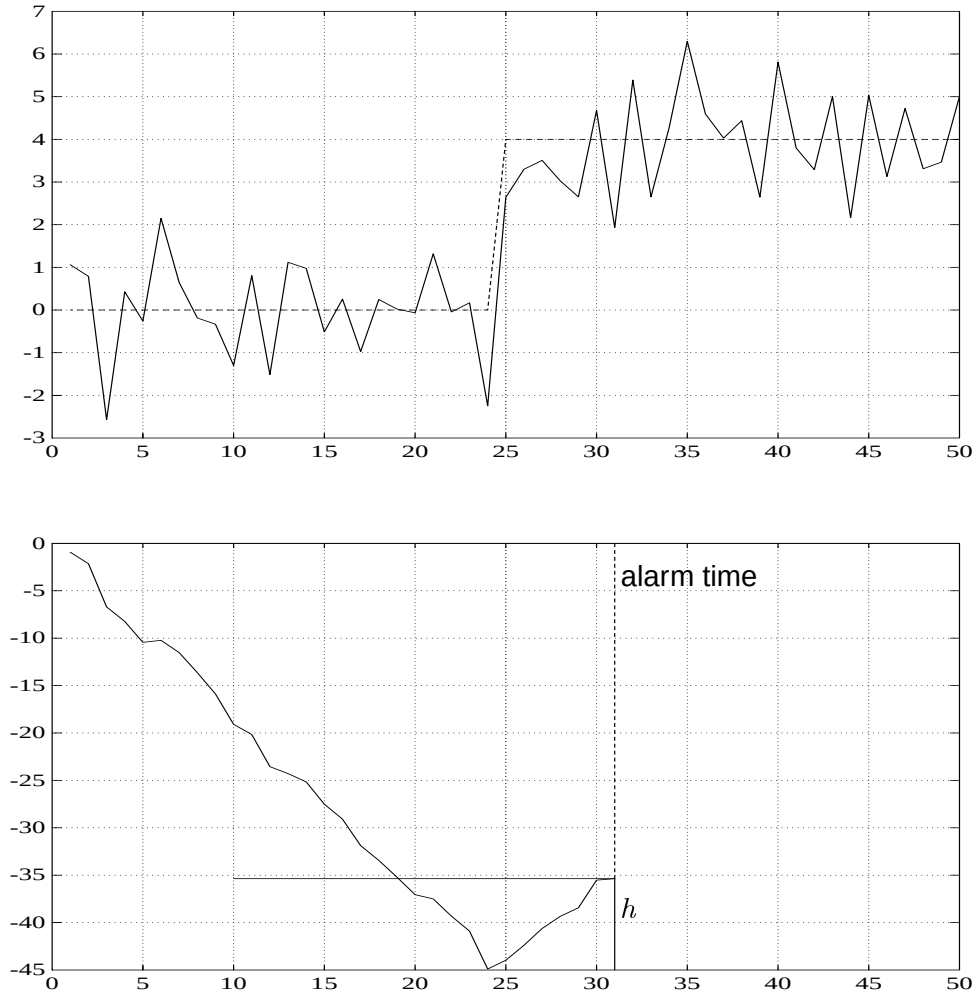

Figure 2.8 Behavior of S_j^k as a SPRT with reverse time (look upside down).

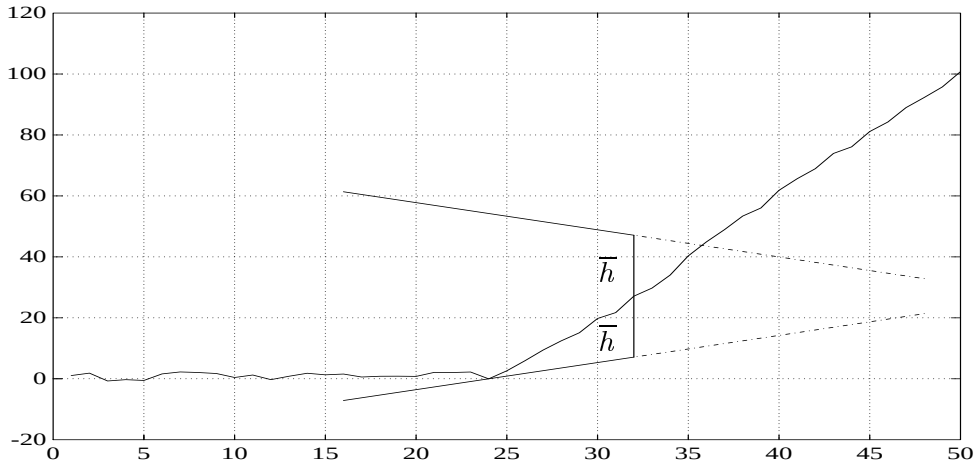

Figure 2.9 The cumulative sum $\tilde{S}_1^k$ intersected by a V-mask, in the case $\mu_0 = 0, \sigma = 1$.

This defines half a V-mask, as depicted in figure 2.9. Let $d = \bar{h}/\tan(\alpha)$ be the distance between the current sample point y_k and the vertex of the V-mask plotted forward. Then equation (2.2.28) can be rewritten in terms of these parameters :

$$\sum_{i=j_0}^{k} [y_i - \mu_0 - \omega \tan(\alpha)] \geq d \ \tan(\alpha) \tag{2.2.31}$$

Notice that, because of (2.2.30), the size of the angle α of the V-mask decreases with the magnitude ν of the jump. This concludes the geometrical interpretation for one-sided CUSUM algorithms. The geometrical interpretation of two-sided CUSUM algorithms is obtained with the aid of a symmetry of the previous picture with respect to the horizontal line, which gives rise to the so-called V-mask. The decision rule is then simply to stop when the boundaries of this mask cover any point already plotted.

The geometrical interpretation of the CUSUM algorithm when viewed as a set of open-ended SPRT is based on figure 2.10, again for the signal of figure 1.1. In this figure are depicted the cumulative sum $\tilde{S}_1^k$, several upper thresholds for the open-ended SPRT, and a standard V-mask. Note that the center of local coordinates for the SPRT beginning at time k is placed at $(k-1, y_{k-1})$. It is obvious that the slope of the upper thresholds of the parallel one-sided SPRT is the same as the slope $\omega \tan(\alpha)$ of the V-mask. This figure shows that the stopping time t_a in (2.2.17) or T^* in (2.2.23) is attained when the decision function of the one-sided SPRT reaches the upper threshold or when the cumulative sum in reverse time reaches the V-mask.

2.3 Bayes-type Algorithms

In this section, we continue to investigate the problem of detecting a change in the scalar parameter of an independent random sequence. As stated in the introduction, we discuss the Bayesian approach in which *a priori* information about the distribution of the change time is available. We assume that this information is in the form of an *a priori* probability distribution for the change time t_0. This approach was first investigated in [Girshick and Rubin, 1952] for continuous inspection of a technological process with known transition probabilities between the two (normal and abnormal) functioning modes. The theoretical derivation of opti-

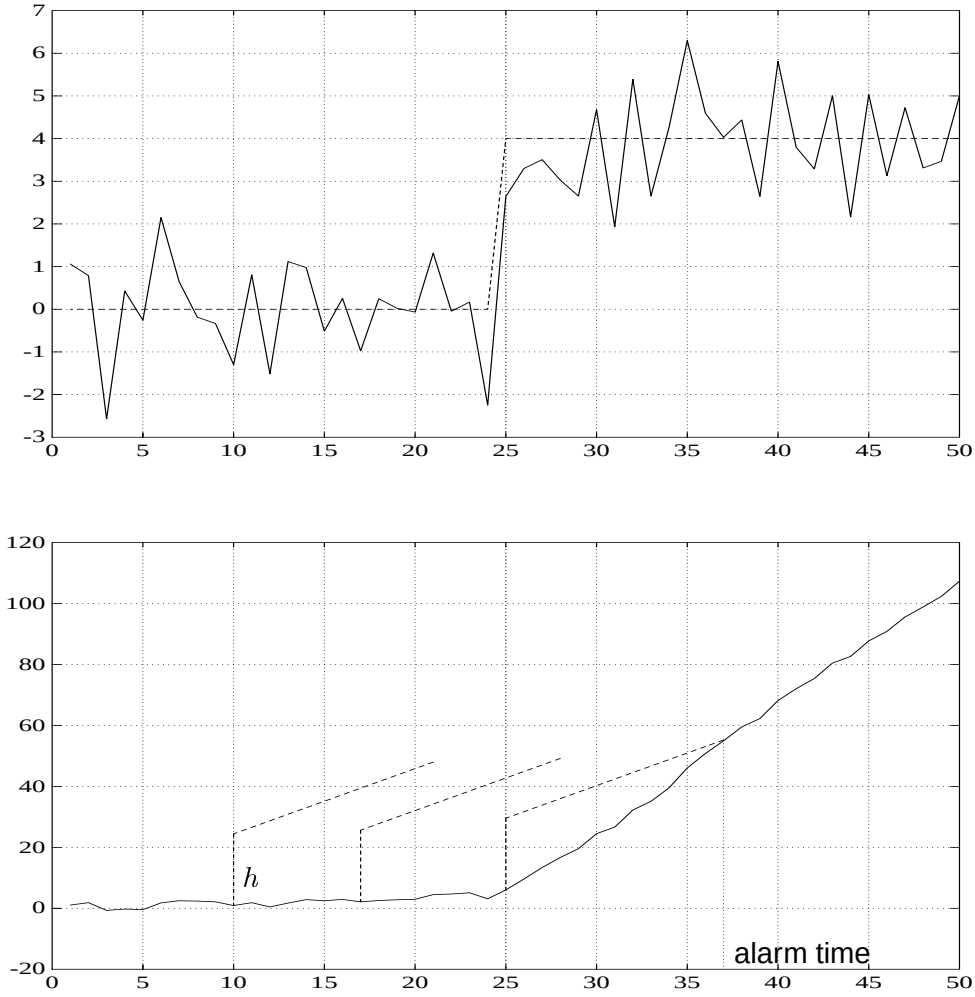

Figure 2.10 The CUSUM algorithm as a set of open-ended SPRT.

mal Bayesian algorithms for change detection was obtained in [Shiryaev, 1961]. This pioneering work was the starting point and theoretical background of a great number of other papers about Bayes-type algorithms.

The main (classical Bayesian) idea consists of deciding that a change has occurred when the *a posteriori* probability of a change exceeds a conveniently chosen threshold. We assume here that the *a priori* distribution of the change time t_0 is geometric :

$$\mathbf{P}(t_0 = k) = \varrho\ (1-\varrho)^{k-1}\ \text{, for } k > 0$$

We assume that the change from θ_0 to θ_1 in the probability density $p_\theta(y_k)$ of our independent sequence can be modeled by a Markov chain with two states, 0 and 1. The transition matrix of this Markov chain is

$$P = \begin{pmatrix} p(0|0) & p(0|1) \\ p(1|0) & p(1|1) \end{pmatrix} = \begin{pmatrix} 1-\varrho & 0 \\ \varrho & 1 \end{pmatrix} \tag{2.3.1}$$

where $p(i|j)$ is the probability of a transition from state j to state i. The probability of the initial state is given by $p(0) = 1 - \pi$ and $p(1) = \pi$. Note that the expectation of the change time is $\mathbf{E}(t_0|t_0 > 0) = \frac{1}{\varrho}$.

Let π_k be the *a posteriori probability* of state 1 of this Markov chain. It results from Bayes' rule that

$$\pi_k = \frac{\pi_{k-1}\ p_{\theta_1}(y_k) + (1-\pi_{k-1})\ \varrho\ p_{\theta_1}(y_k)}{\pi_{k-1}\ p_{\theta_1}(y_k) + (1-\pi_{k-1})\ \varrho\ p_{\theta_1}(y_k) + (1-\pi_{k-1})(1-\varrho)\ p_{\theta_0}(y_k)} \tag{2.3.2}$$

For simplicity, we will deal with a monotonic function of π_k instead of π_k alone, because it will be more convenient for recursive computations. This function is

$$\varpi_k = \frac{\pi_k}{1-\pi_k} \tag{2.3.3}$$

The recursive formula for ϖ_k is

$$\varpi_k = \frac{1}{1-\varrho}(\varpi_{k-1} + \varrho)\ \frac{p_{\theta_1}(y_k)}{p_{\theta_0}(y_k)} \tag{2.3.4}$$

To deal with the log-likelihood ratio as in the previous sections, we rewrite this formula as follows :

$$g_k = \ln(\varrho + e^{g_{k-1}}) - \ln(1-\varrho) + \ln\frac{p_{\theta_1}(y_k)}{p_{\theta_0}(y_k)} \tag{2.3.5}$$

where

$$g_k = \ln \varpi_k \tag{2.3.6}$$

The last term is the log-likelihood ratio, which basically contains the updating information available at time k. Because g_k is an increasing function of π_k, the Bayesian stopping rule becomes :

$$t_a = \min\{k\ : g_k \geq h\} \tag{2.3.7}$$

exactly as in the previous sections (remember (2.2.10)).

Example 2.3.1 (Change in mean - contd.). *Let us return to our basic example. We assume here that the mean values μ_0, μ_1, and the constant variance σ^2 are known. In this case, the log-likelihood ratio is given in (2.1.6), and consequently the decision function g_k is*

$$g_k = \ln(\varrho + e^{g_{k-1}}) - \ln(1-\varrho) + \frac{\mu_1 - \mu_0}{\sigma^2}\left(y_k - \frac{\mu_0 + \mu_1}{2}\right) \tag{2.3.8}$$

The behavior of this decision function is depicted in figure 2.11, again for the signal of figure 1.1. In this figure, the influence of the choice of the parameter ϱ of the geometric distribution is emphasized. The solid line corresponds to the ideal case where we know the true value 0.05 of this parameter. The two other lines correspond to cases where the tuning value of ϱ is different from this true value.

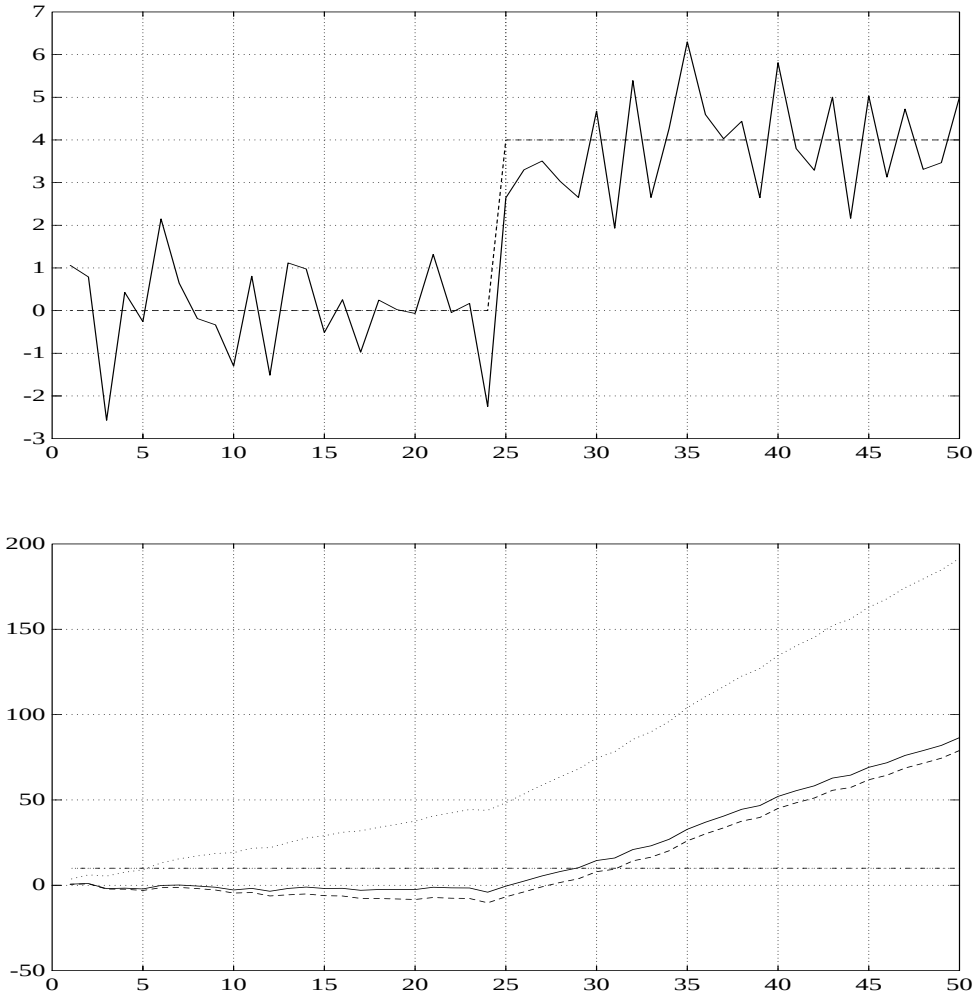

Figure 2.11 Typical behavior of a Bayesian decision function : ϱ chosen to be the true value $\varrho = 0.05$ (solid line); noncorrect but acceptable choice of $\varrho = 0.001$ (dashed line); nonacceptable choice of $\varrho = 0.9$ (dotted line).

Notice that, in some sense, the Bayesian decision rule is not of the same type as the other ones before, because it assumes the availability of the parameter ϱ of the geometric *a priori* distribution of the change time t_0, and of the initial probability π which is implicit in g_0. For this reason, the practical implementation of this decision rule is not so simple and requires a preliminary investigation of this question of *a priori* information. The effect of the choice of the parameter ϱ on the behavior of g_k is depicted in figure 2.11.

2.4 Unknown Parameter After Change

We now discuss the case where the parameter θ_1 after change is unknown. Without loss of generality in our on-line framework, the parameter θ_0 before change is assumed to be known.

2.4.1 Introduction

It follows from the previous discussion that a sequential change detection algorithm can be interpreted as a set of "parallel" open-ended tests. We begin the present discussion with these tests.

As explained in [Wald, 1947], two possible solutions exist in the present case. The first one consists of weighting the likelihood ratio with respect to all possible values of the parameter θ_1, using a weighting function $dF(\theta_1)$, where $F(\theta_1)$ may be interpreted as the cumulative distribution function of a probability measure. In the second solution, the unknown parameter θ_1 is replaced by its maximum likelihood estimate, which results in the generalized likelihood ratio (GLR) algorithm. In other words, for known θ_1, change detection algorithms are based on the likelihood ratio :

$$\Lambda_n = \frac{p_{\theta_1}(y_1, \ldots, y_n)}{p_{\theta_0}(y_1, \ldots, y_n)} \tag{2.4.1}$$

and for unknown θ_1 we must replace Λ_n by other statistic. More precisely, the first solution is based upon the weighted likelihood ratio :

$$\tilde{\Lambda}_n = \int_{-\infty}^{\infty} \frac{p_{\theta_1}(y_1, \ldots, y_n)}{p_{\theta_0}(y_1, \ldots, y_n)}\, dF(\theta_1) \tag{2.4.2}$$

and the second one uses the GLR :

$$\hat{\Lambda}_n = \frac{\sup_{\theta_1} p_{\theta_1}(y_1, \ldots, y_n)}{p_{\theta_0}(y_1, \ldots, y_n)} \tag{2.4.3}$$

We investigate these two solutions in subsections 2.4.2 and 2.4.3, respectively.

2.4.2 Weighted CUSUM Algorithm

Let us now explain in detail the algorithm resulting from the idea of weighting the unknown parameter.

2.4.2.1 Derivation of the Algorithm

We follow Lorden's idea introduced before, which explains the CUSUM algorithm as an extended stopping time associated with a family of open-ended SPRT. The weighted-CUSUM algorithm was derived for change detection in [Pollak and Siegmund, 1975], and is a direct extension of the CUSUM stopping time. It is defined as follows. Let

$$\tilde{\Lambda}_j^k = \int_{-\infty}^{\infty} \frac{p_{\theta_1}(y_j, \ldots, y_k)}{p_{\theta_0}(y_j, \ldots, y_k)} dF(\theta_1) \tag{2.4.4}$$

be the weighted likelihood ratio for the observations from time j up to time k. Then the stopping time is

$$t_a = \min\{k : \max_{1 \le j \le k} \ln \tilde{\Lambda}_j^k \ge h\} \tag{2.4.5}$$

Typical choices of the weighting function $F(\theta)$ are the following. The most simple choices involve using the uniform distribution over a specified interval that contains all possible values of the parameter θ_1, or Dirac masses on some specified values. Another useful choice is the Gaussian distribution. Note that this type of algorithm *cannot* be written in a recursive manner as the simple CUSUM algorithm (2.2.9) that we describe in section 2.2.

Example 2.4.1 (χ^2-CUSUM algorithm). *Let us now discuss the problem of detecting a change in the mean of a Gaussian sequence with known variance σ^2, in the special case where the distribution $F(\theta) = F(\mu)$ is concentrated on two points, $\mu_0 - \nu$ and $\mu_0 + \nu$. In this case, the weighted likelihood ratio is easily shown to be*

$$\tilde{\Lambda}_j^k = \int_{-\infty}^{\infty} \exp\left[b\tilde{S}_j^k - \frac{b^2}{2}(k-j+1)\right] dF(\nu) \tag{2.4.6}$$

where

$$b = \frac{\nu}{\sigma} \tag{2.4.7}$$

is the signal-to-noise ratio, and

$$\tilde{S}_j^k = \frac{1}{\sigma}\sum_{i=j}^{k}(y_i - \mu_0) \tag{2.4.8}$$

This reduces to

$$\begin{aligned} \tilde{\Lambda}_j^k &= \cosh\left(b\tilde{S}_j^k\right) e^{-\frac{b^2}{2}(k-j+1)} \\ &= \cosh\left[b(k-j+1)\chi_j^k\right] e^{-\frac{b^2}{2}(k-j+1)} \end{aligned} \tag{2.4.9}$$

where

$$\chi_j^k = \frac{1}{k-j+1}|\tilde{S}_j^k| \tag{2.4.10}$$

Note that $\tilde{\Lambda}_j^k$ in (2.4.9) is the likelihood ratio for testing the noncentrality parameter of a χ^2 distribution with one degree of freedom, between the values 0 and $(k-j+1)\ b^2$. This fact explains the name of the χ^2-CUSUM algorithm.

The stopping time is thus

$$t_a = \min\{k : g_k \ge h\} \tag{2.4.11}$$

where

$$g_k = \max_{1 \le j \le k}\left[\ln\cosh\left(b\tilde{S}_j^k\right) - \frac{b^2}{2}(k-j+1)\right] \tag{2.4.12}$$

As we said before, this algorithm cannot be written in a recursive manner because it is derived from Lorden's open-ended test. However, using Page's and Shiryaev's interpretation of the CUSUM algorithm as a repeated SPRT with lower threshold equal to 0 and upper threshold equal to h as discussed in subsection 2.2.2, it is possible to design a slightly modified decision rule which is written in a recursive manner.

This results in

$$g_k = (\check{S}^k_{k-N_k+1})^+ \tag{2.4.13}$$

$$\check{S}^k_{k-N_k+1} = -\frac{1}{2}N_k\ b^2 + \ln\cosh(b\tilde{S}^k_{k-N_k+1}) \tag{2.4.14}$$

$$\bar{S}_k = \tilde{S}^k_{k-N_k+1} \tag{2.4.15}$$

$$\bar{S}_k = \bar{S}_{k-1}\mathbf{1}_{\{g_{k-1}>0\}} + \frac{y_k - \mu_0}{\sigma} \tag{2.4.16}$$

where $N_k = N_{k-1}\mathbf{1}_{\{g_{k-1}>0\}} + 1$.

This CUSUM algorithm can be used in the same situations as the two-sided CUSUM algorithm. The multidimensional parameter counterpart of this algorithm is investigated in section 7.2, case 3.

2.4.2.2 Geometrical Interpretation in the Gaussian Case

We continue to investigate the detection of a change in the mean of a Gaussian sequence, and give now the geometrical interpretation of the weighted CUSUM (2.4.4) and χ^2-CUSUM (2.4.9) algorithms in this case. We discuss first a one-sided weighted CUSUM algorithm, and then a two-sided one. We finish with the geometrical interpretation of the χ^2-CUSUM algorithm.

Let us assume that the probability measure $F(\mu)$ is confined to the interval $[\mu_0, \infty)$. The weighted CUSUM algorithm is based upon the stopping time :

$$t_a = \min\{k : g_k = \max_{1\le j\le k} \ln\tilde{\Lambda}^k_j \ge h\} \tag{2.4.17}$$

where the weighted likelihood ratio is

$$\tilde{\Lambda}^k_j = \int_0^\infty \exp\left[\frac{\nu}{\sigma}\tilde{S}^k_j - \frac{\nu^2}{2\sigma^2}(k-j+1)\right] dF(\nu) \tag{2.4.18}$$

Let us define the following function :

$$f(x,l) = \ln\int_0^\infty \exp\left(\frac{\nu}{\sigma}x - \frac{\nu^2}{2\sigma^2}l\right) dF(\nu) \tag{2.4.19}$$

Because F defines a probability measure on $(\mathbf{R}, \mathcal{R})$, the function $f(x,l)$ is an increasing function of x. It is obvious that the decision rule involves stopping the first time k at which the cumulative sum $\tilde{S}^k_j$ reaches the curve line threshold $\tilde{c}_{k-j+1}$, where $\tilde{c}_l$ is the unique positive solution of the equation $f(x,l) = h$ [Robbins, 1970]. This threshold $\tilde{c}_l$ is the half lower part of the curve in figure 2.12 and is called a U-mask. The geometrical interpretation is now the same as for the CUSUM algorithm.

If we now assume that F is a symmetric distribution over $(-\infty, \infty)$, then

$$f(x,l) \ge h \text{ if and only if } |x| \ge \tilde{c}_l \tag{2.4.20}$$

Therefore, the geometrical interpretation of the two-sided weighted CUSUM algorithm is obtained from the one-sided one, with the aid of a symmetry with respect to the horizontal line drawn at the last observation point, as depicted in the figure 2.12, and as for the ordinary CUSUM algorithm before.

Finally, let us assume that F is concentrated on two points, which corresponds to the χ^2-CUSUM algorithm. In this case, the function f can be written as

$$f(x,l) = \ln\cosh(bx) - \frac{b^2}{2}l \tag{2.4.21}$$

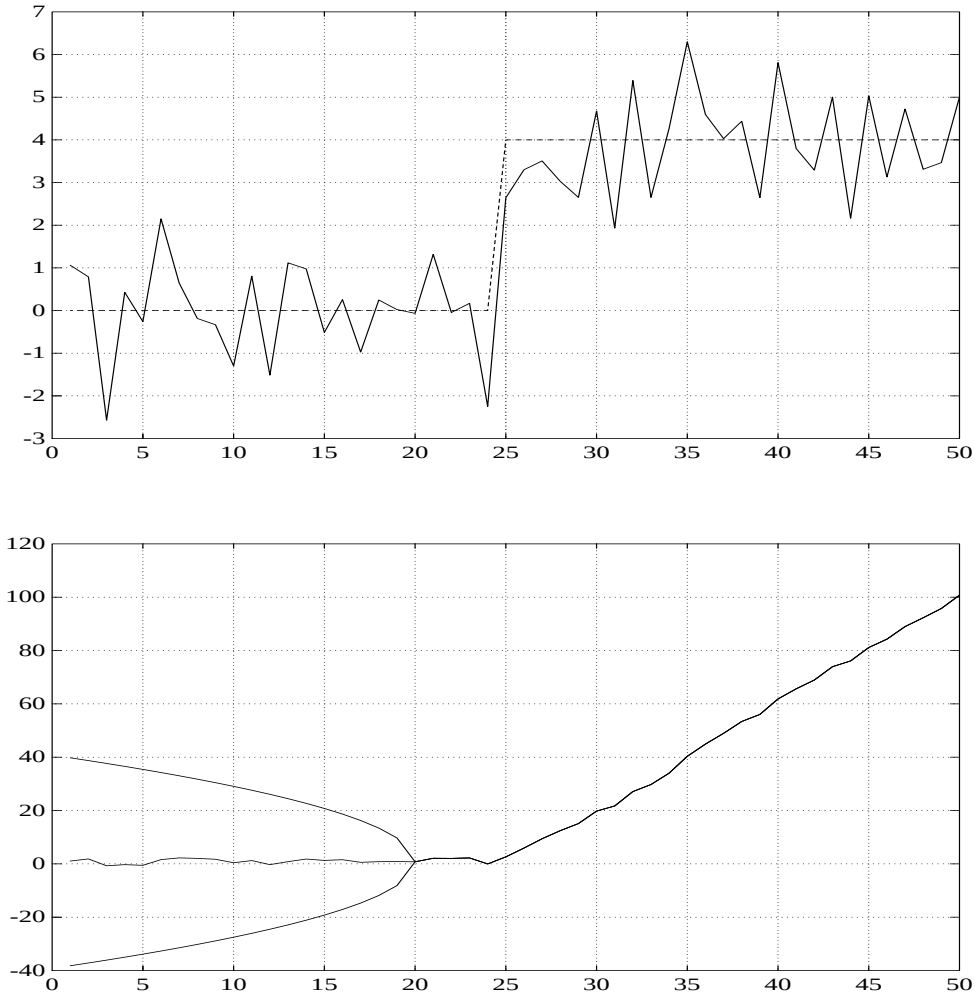

Figure 2.12 U-mask for the weighted CUSUM algorithm.

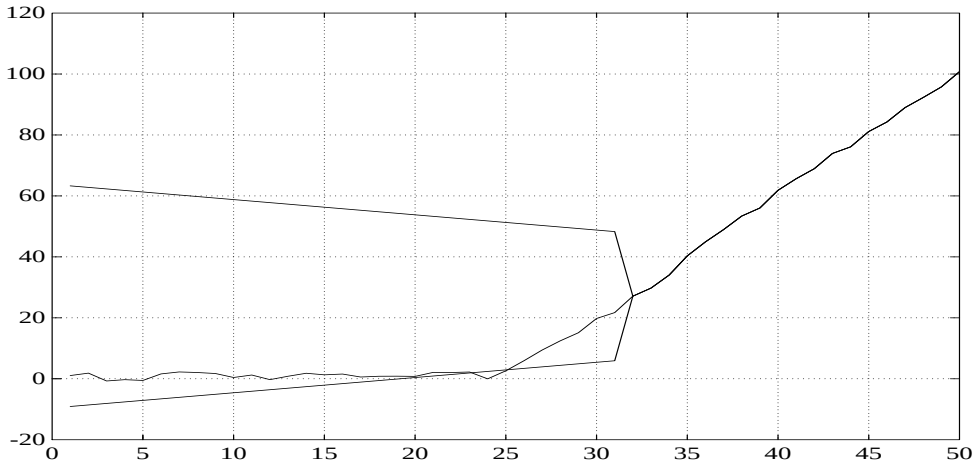

Figure 2.13 Mask for the χ^2-CUSUM algorithm.

and we wish to find $\tilde{c}_l$ such that

$$f(\tilde{c}_l, l) = h \tag{2.4.22}$$

For $v \geq 0$, the equation $\ln \cosh |u| = v$ has a unique positive solution, which is given by

$$|u| = \ln(e^v + \sqrt{e^{2v} - 1}) = v + \ln(1 + \sqrt{1 - e^{-2v}}) \tag{2.4.23}$$

From this solution the boundary $\tilde{c}_l$ is

$$|\tilde{c}_l| = \frac{1}{b}\left(h + \ln\left\{1 + \sqrt{1 - \exp\left[-2\left(h + \frac{b^2 l}{2}\right)\right]}\right\}\right) + \frac{b}{2}\, l \tag{2.4.24}$$

When l goes to infinity, the two asymptotes of this boundary have the equation

$$c_l = \pm\left(\frac{h + \ln 2}{b} + \frac{b}{2} l\right) \tag{2.4.25}$$

This fact is depicted in figure 2.13. From these formulas the difference between the boundary and its asymptotes decreases very quickly when h increases for all l. In other words,

$$\tilde{c}_l - c_l = O(e^{-2h}) \tag{2.4.26}$$

when h goes to infinity. Therefore, the stopping boundary for the χ^2-CUSUM algorithm is made nearly of straight lines, and thus is very close to the stopping boundary of the two-sided CUSUM algorithm. We continue this discussion in section 11.1.

Example 2.4.2 (Change in mean - contd.). *Let us again discuss the problem of detecting a change in the mean of a Gaussian sequence with* unit *variance, in another special case where the distribution $F(\theta) = F(\mu)$ is Gaussian with mean μ_0 and known variance σ^2. In this case, the weighted likelihood ratio can be written as*

$$\tilde{\Lambda}_j^k = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[\nu \tilde{S}_j^k - \frac{\nu^2}{2}(k - j + 1)\right] \exp\left[-\frac{\nu^2}{2\sigma^2}\right] d\nu \tag{2.4.27}$$

or

$$\ln \tilde{\Lambda}_j^k = \frac{\sigma^2}{2[\sigma^2(k-j+1)+1]} \left(\tilde{S}_j^k\right)^2 - \frac{1}{2}\ln[\sigma^2(k-j+1)+1] \tag{2.4.28}$$

where $\tilde{S}_j^k$ is defined in (2.4.8). The function $f(x,l)$ can be written as

$$f(x,l) = \frac{\sigma^2}{2(\sigma^2 l+1)}\, x^2 - \frac{1}{2}\ln(\sigma^2 l+1) \tag{2.4.29}$$

and satisfies (2.4.20). The equation $f(|x|,l) = h$ has a unique positive solution from which we deduce that the boundary $\tilde{c}_l$ is

$$\tilde{c}_l = \pm\sqrt{2(l+\sigma^{-2})\left[h + \frac{1}{2}\ln(\sigma^2 l+1)\right]} \tag{2.4.30}$$

2.4.3 GLR Algorithm

We continue to discuss the case where the parameter θ_1 after change is unknown. The parameter θ_0 before change is again assumed to be known. The derivation of the GLR algorithm proceeds in the same way as the third derivation of the CUSUM algorithm. Actually we follow [Lorden, 1971], except that we use the widely accepted term "generalized likelihood ratio" (GLR) instead of "maximum likelihood."

2.4.3.1 Derivation of the Algorithm

We now describe Wald's second solution for the case of unknown parameter after change. Let us start from the generalized likelihood ratio given in equation (2.4.3). As before, the log-likelihood ratio for the observations from time j up to time k is

$$S_j^k(\theta_1) = \sum_{i=j}^{k} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \tag{2.4.31}$$

In the present case, θ_1 is unknown; therefore, this ratio is a function of two unknown independent parameters : the change time and the value of the parameter after change. The standard statistical approach is to use the maximum likelihood estimates of these two parameters, and thus the *double* maximization :

$$g_k = \max_{1\le j\le k} \ln \hat{\Lambda}_j^k = \max_{1\le j\le k} \sup_{\theta_1} S_j^k(\theta_1) \tag{2.4.32}$$

The precise statement of the conditions on the probability densities p_{θ_i} under which this double maximization can be performed is found in [Lorden, 1971]. Actually, the densities should belong to the so-called Koopman-Darmois family of probability densities :

$$p_\theta(y) = e^{\theta T(y) - d(\theta)} h(y) \tag{2.4.33}$$

where d is strictly concave upward and infinitely differentiable over an interval of the real line. This family is discussed in detail in chapter 4. The corresponding stopping rule is the same as in (2.2.10). As we said before, this algorithm cannot be written in a recursive manner.

Now let us discuss further the issue of level of available *a priori* information about the parameter after change. In many applications, it is possible to know a minimum magnitude ν_m of the changes of interest

in the parameter θ. In this case, the second maximization in the GLR algorithm can be achieved using this minimum magnitude of change as follows :

$$g_k = \max_{1\leq j\leq k} \sup_{\theta_1:|\theta_1-\theta_0|\geq\nu_m>0} S_j^k(\theta_1) \tag{2.4.34}$$

If information about a maximum possible magnitude of change is also available, the decision function is modified accordingly in an obvious manner.

Let us now discuss the *estimation issue*. In the present case, two unknown values have to be estimated after a change has been detected : the change time t_0 and the magnitude of the jump $(\theta_1 - \theta_0)$. As far as t_0 is concerned, the estimation is the same as before in the third derivation of the CUSUM algorithm, namely the maximum likelihood estimate which is given by (2.2.18). The conditional maximum likelihood estimates of the change magnitude and time are given by

$$(\tilde{j}, \tilde{\theta}_1) = \arg \max_{1\leq j\leq t_a} \sup_{\theta_1:|\theta_1-\theta_0|\geq\nu_m>0} \sum_{i=j}^{t_a} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \tag{2.4.35}$$

and $\hat{t}_0 = \tilde{j}$.

Example 2.4.3 (Change in mean - contd.). *Let us return to the example of change in the mean of an independent Gaussian sequence. In this case, the mean μ_0 before change is known, and the mean μ_1 after change is unknown. The constant variance σ^2 is also known. The corresponding cumulative sum can be rewritten as*

$$S_j^k = \frac{\mu_1 - \mu_0}{\sigma^2} \sum_{i=j}^{k} \left(y_i - \frac{\mu_1 + \mu_0}{2} \right) \tag{2.4.36}$$

Let us introduce $\nu = \mu_1 - \mu_0$. Then equation (2.4.34) can be rewritten as

$$g_k = \max_{1\leq j\leq k} \sup_{\nu:|\nu|\geq\nu_m>0} \sum_{i=j}^{k} \left[\frac{\nu(y_i - \mu_0)}{\sigma^2} - \frac{\nu^2}{2\sigma^2} \right] \tag{2.4.37}$$

In the present independent Gaussian case, the constrained maximization over ν is explicit :

$$g_k = \max_{1\leq j\leq k} \sum_{i=j}^{k} \left[\frac{\hat{\nu}_j(y_i - \mu_0)}{\sigma^2} - \frac{\hat{\nu}_j^2}{2\sigma^2} \right] \tag{2.4.38}$$

where the absolute value of the constrained *change magnitude estimate is*

$$|\hat{\nu}_j| = \left(\frac{1}{k-j+1} \sum_{i=j}^{k} |y_i - \mu_0| - \nu_m \right)^+ + \nu_m \tag{2.4.39}$$

and its sign is the same as the sign of the mean value $\frac{1}{k-j+1}\sum_{i=j}^{k}(y_i - \mu_0)$ of the last centered observations or "innovations." Note that the second term $\frac{\nu^2}{2\sigma^2}$ on the right side of (2.4.37) is nothing but the Kullback information between the two laws before and after the change.

Note also that, when $\nu_m = 0$, the decision function is

$$g_k = \frac{1}{2\sigma^2} \max_{1\leq j\leq k} \frac{1}{k-j+1} \left[\sum_{i=j}^{k} (y_i - \mu_0) \right]^2 \tag{2.4.40}$$

The above property of explicit maximization over the unknown parameter θ_1 after change can be exploited in more complex situations, as explained in section 7.2.4. Furthermore, (2.4.38) can be viewed as a correlation *between the innovation $(y_i - \mu_0)$ and the "signature" of the change $\hat{\nu}_k$. This correlation property, which is typical for matched-filtering operations, is recovered in (7.2.118) for the more general situation of additive changes in state-space models.*

Finally, let us comment further on the asymptotic equivalence, in the Gaussian case again, between the three algorithms, which we describe for the case of unknown parameter after change. As we explain in the previous subsection, the χ^2-CUSUM algorithm is asymptotically equivalent to the two-sided CUSUM algorithm when the threshold goes to infinity. But it should be clear that the two-sided CUSUM algorithm is nothing but the GLR algorithm corresponding to the degenerate situation where $\mu_1 = \mu_0 \pm \nu$.

2.4.3.2 Geometrical Interpretation in the Gaussian Case

We describe the geometrical interpretation of the GLR algorithm in the same way we described the CUSUM algorithm, namely starting from the reverse time interpretation of the decision function. We begin with a one-sided GLR algorithm, and we use a symmetry with respect to the horizontal line for the two-sided case as before. From the decision function (2.4.32), it follows that the stopping rule can be rewritten in reverse time as follows. There exists a time instant l such that the following inequality holds :

$$\sup_{\nu:\nu\geq\nu_m>0} \quad \sum_{i=1}^{l} \left[\nu(y_i - \mu_0) - \frac{\nu^2}{2}\right] \geq h\sigma^2 \tag{2.4.41}$$

This can be rewritten as

$$\tilde{S}_1^l = \frac{1}{\sigma}\sum_{i=1}^{l}(y_i - \mu_0) \geq \inf_{\nu:\nu\geq\nu_m>0} \quad \left(\frac{h\sigma}{\nu} + \frac{\nu}{2\sigma}l\right) \tag{2.4.42}$$

Let us now introduce the lower boundary $\hat{c}_l$ for the cumulative sum $\tilde{S}_1^l$:

$$\hat{c}_l = \inf_{\nu:\nu\geq\nu_m>0} \quad \left(\frac{h\sigma}{\nu} + \frac{\nu}{2\sigma}l\right) \tag{2.4.43}$$

and discuss this minimization. We distinguish two situations for the parameter ν: $\nu = \nu_m$ and $\nu > \nu_m$. For the situation $\nu = \nu_m$, and from the discussion in section 2.2 about the geometrical interpretation of the stopping rule in terms of the V-mask, we find that, for large l, the boundary in (2.4.43) is the straight line with minimal angle with respect to the horizontal line, as depicted in figure 2.14. For $\nu > \nu_m$, the boundary is a curve, as we explain now. Let us consider again the reverse time SPRT with one threshold h. Because of the Wald's identity (which we explain in detail in chapter 4), for a SPRT with threshold h, the average number of samples until the threshold is reached is asymptotically

$$\mathbf{E}(l) \approx \frac{h}{\mathbf{K}(\nu)} \tag{2.4.44}$$

where $\mathbf{K}$ is the Kullback information. In the Gaussian case, it is well known that $\mathbf{K}(\nu) = \frac{\nu^2}{2\sigma^2}$. It follows that, for $l \geq \frac{h}{\mathbf{K}(\nu_m)}$, the minimum in equation (2.4.43) is then reached for $\nu = \nu_m$. On the other hand, for small values of l, the minimum in equation (2.4.43) is then reached for ν such that $l\ \mathbf{K}(\nu) = h$. Inserting this value in equation (2.4.43), we obtain

$$\hat{c}_l = \sqrt{2hl} \tag{2.4.45}$$

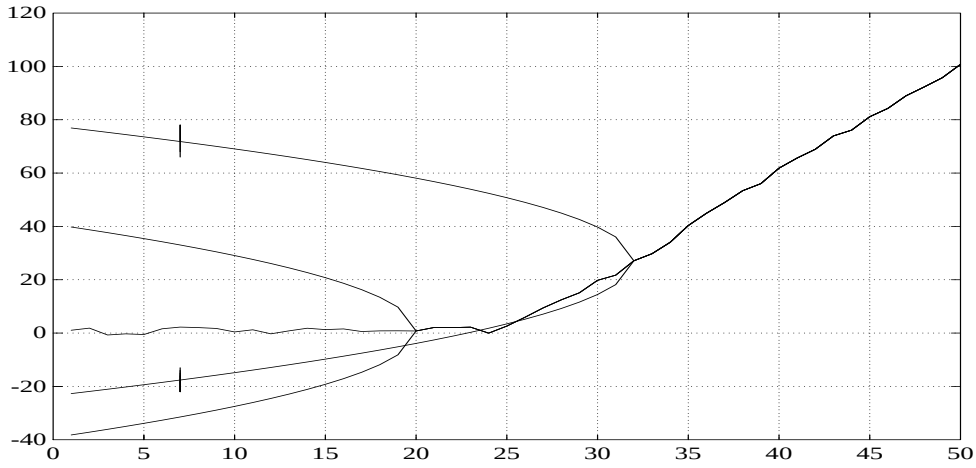

Figure 2.14 U-mask for the GLR algorithm : boundary with equation (2.4.46).

which is the equation of a parabola, leading to the so-called U-mask depicted in figure 2.14. This parabola is inscribed in the V-mask discussed before, because the points of tangency between the straight line and the parabola have the abscissa $l = \frac{2h\sigma^2}{\nu_m^2}$ as depicted by vertical segments in this figure. In summary, the equation of the boundary is

$$\hat{c}_l = \begin{cases} \sqrt{2hl} & \text{if } l \leq \frac{2h\sigma^2}{\nu_m^2} \\ \frac{h\sigma}{\nu_m} + \frac{\nu_m l}{2\sigma} & \text{otherwise} \end{cases} \tag{2.4.46}$$

The explanation for the upper boundary is the same.

As we explained before, the GLR algorithm is computationally complex. Approximations of this algorithm, with lower computational cost, are thus of interest. In [Lorden and Eisenberger, 1973], a possible approximation of the GLR algorithm dealing with the joint use of two CUSUM algorithms is proposed. These two algorithms are designed to detect changes with large and small magnitudes, respectively. The geometrical interpretation of this approximation is that a U-mask can be approximated by the intersection of two V-masks, as depicted in figure 2.15. This point is further discussed in chapter 11.

2.5 Change Detection and Tracking

In this section, we do not introduce any other derivations of change detection algorithms. Instead we explain an example of the use of one of the previously described algorithms in the framework of adaptive identification, for improving the tracking capability of adaptive identification algorithms.

Let us consider the simple example of a piecewise constant sequence perturbed by a white Gaussian noise ε. In other words, we consider the multiple change times counterpart of the above widely discussed example, modeled as

$$y_k = \varepsilon_k + \mu(k) \tag{2.5.1}$$

where $\mu(k)$ is an unknown piecewise constant function of time, as depicted in figure 2.16. The standard recursive estimation of the mean value can be written as

$$\bar{y}_k = \frac{k-1}{k}\, \bar{y}_{k-1} + \frac{1}{k}\, y_k \tag{2.5.2}$$

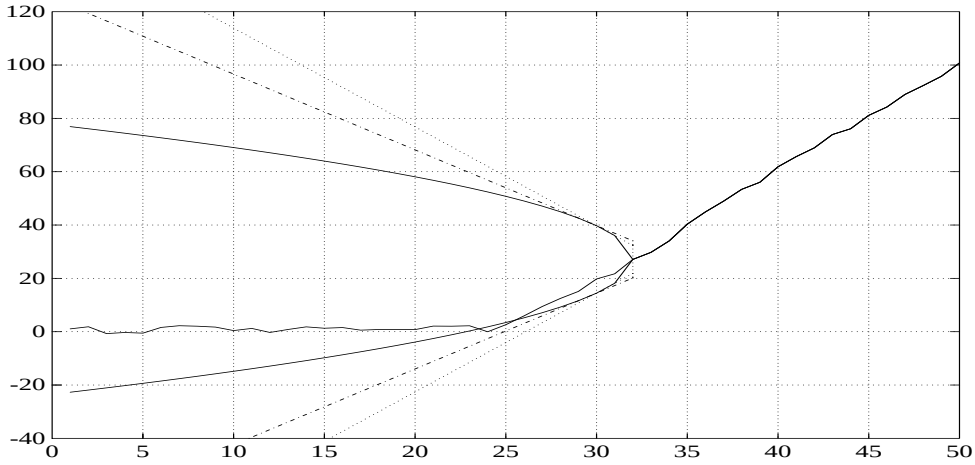

Figure 2.15 Two V-masks (dotted lines) approximating one U-mask (solid curve) : how a GLR algorithm can be approximated by two CUSUM algorithms for detecting changes with small and large magnitudes, respectively.

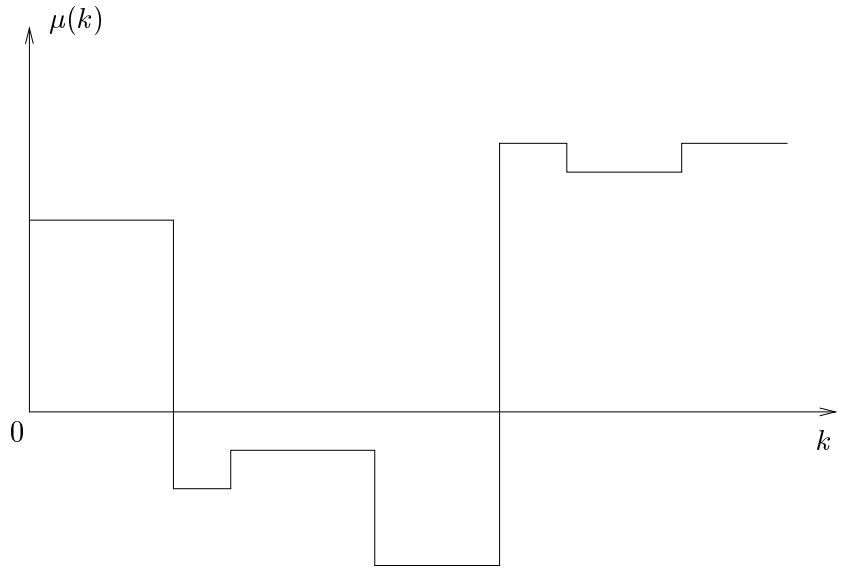

Figure 2.16 Piecewise constant signal.

This estimation is known to be efficient provided that the underlying unknown mean value is constant. Our suggestion is to use change detection algorithms for checking this assumption. We assume that the time duration between successive jumps is bounded from below. This assumption is necessary for the initial estimation of the mean to be used in the subsequent detection of change. The joint use of the estimation and detection algorithms results in cycles made of the following steps :

1. Initial estimation of the mean, during a fixed size time interval during which the detection algorithm is switched off; let $\bar{y}_N$ be this estimated mean value.
2. Carrying on the estimation and activation of the change detection algorithm using $\mu_0 = \bar{y}_k$ for $k \geq N$.
3. Updating the initial estimation after a change has been detected. This updating can take place either at the alarm time if no other information is provided by the change detection algorithm, or at the estimated change time $\hat{t}_0$ if this information is available. Similarly, the updating can include the possible estimate $\hat{\nu}$ of the magnitude of the jump. If both values $\hat{t}_0$ and $\hat{\nu}$ are available, returning to step 1 after a change has been detected is not necessary; the cycle restarts from step 2.

The two main types of relevant change detection algorithms to be used in such a cycle are the CUSUM and GLR algorithms introduced before. The main reason is that these are the only algorithms that can provide us with an estimate of the change time t_0 in addition to an alarm time t_a.

Let us add some comments about the tuning of change detection algorithms in such a framework. Minimum values ν_m of jump magnitudes (for the CUSUM and GLR algorithms) and thresholds are required. Minimum values of jumps must be close to the precision of the estimation algorithm, for example, of the order of magnitude of the corresponding standard deviation of the estimate. On the other hand, the threshold has to be chosen in such a way that the mean time between false alarms should not be too much less than the mean time between successive jumps in the piecewise function.

2.6 Off-line Change Detection

In this section, we introduce two new tasks, which were mentioned in subsection 1.1.2 :

1. *Off-line hypotheses testing* between the hypotheses "without change" and "with change."
2. *Off-line estimation of the unknown change time*.

The main difference between this section and the previous ones is that now the complete sample of observations is available before beginning the investigation for a change.

This task was first investigated in [Page, 1957], using basically the same type of ideas that he used for the CUSUM algorithm, which are described in subsection 2.2.3. The problem of off-line estimation of the change time was investigated in [Hinkley, 1970, Hinkley, 1971], including precision issues and the distribution of the estimation error.

2.6.1 Off-line Hypotheses Testing

Let $(y_k)_{1 \leq k \leq N}$ be a sequence of independent random observations with density $p_\theta(y)$. Two situations are possible. Either all the observations in this sample have the same density, characterized by $\tilde{\theta}_0$, or there exists an *unknown change time* $1 < t_0 \leq N$ such that, before t_0, the parameter θ is equal to θ_0, and after the change it is equal to $\theta_1 \neq \theta_0$. Let us first assume that $\tilde{\theta}_0$, θ_0, and θ_1 are known. As discussed in subsection 2.2.3, it is convenient to introduce the following hypotheses about this sequence of observations :

$$\begin{array}{rrll} & \mathbf{H}_0 : & \theta = \tilde{\theta}_0 & \text{for} \quad 1 \leq k \leq N \\ \text{for } 1 \leq j \leq N, & \mathbf{H}_j : & \theta = \theta_0 & \text{for} \quad 1 \leq k \leq j-1 \\ & & \theta = \theta_1 & \text{for} \quad j \leq k \leq N \end{array} \tag{2.6.1}$$

The problem is to test between the hypothesis $\mathbf{H}_0$ and the composite hypothesis :

$$\mathcal{H}_1 = \cup_{j \geq 1} \mathbf{H}_j \tag{2.6.2}$$

Note that the estimation of the change time is *not* included in this problem statement, and that the unknown change time may be interpreted here as a *nuisance* parameter. The estimation of the change time is discussed in the next subsection.

The likelihood ratio corresponding to the hypotheses $\mathbf{H}_0$ and $\mathbf{H}_j$ is

$$\Lambda_1^N(j) = \frac{\prod_{i=1}^{j-1} p_{\theta_0}(y_i) \cdot \prod_{i=j}^{N} p_{\theta_1}(y_i)}{\prod_{i=1}^{N} p_{\tilde{\theta}_0}(y_i)} \tag{2.6.3}$$

(where $\prod_{i=1}^{0} = 1$). The standard statistical approach in this situation consists of replacing the unknown parameter t_0 by its *maximum likelihood estimate* (M.L.E.). Therefore, we consider the following statistic :

$$\Lambda_N = \max_{1 \leq j \leq N} \Lambda_1^N(j) \tag{2.6.4}$$

and the decision rule d such that $d = 0$ (1), according to which hypothesis $\mathbf{H}_0$ $(\mathcal{H}_1)$ is chosen, is given by

$$d = \begin{cases} 0 & \text{if} \quad \ln \Lambda_N < h \\ 1 & \text{if} \quad \ln \Lambda_N \geq h \end{cases} \tag{2.6.5}$$

When the parameters $\tilde{\theta}_0$, θ_0 and θ_1 are unknown, they are also replaced by their M.L.E. This results in the following decision function :

$$\tilde{\Lambda}_N = \max_{1 \leq j \leq N} \sup_{\tilde{\theta}_0} \sup_{\theta_0} \sup_{\theta_1} \Lambda_1^N(j, \tilde{\theta}_0, \theta_0, \theta_1) \tag{2.6.6}$$

2.6.2 Off-line Estimation of the Change Time

We consider the same hypotheses as in the previous subsection. We assume the existence of a change point (typically this assumption is the result of the previous hypotheses testing) and the problem is now to estimate the change time. In the present case, all the parameters θ_0, θ_1, and t_0 are assumed to be unknown. Therefore, the corresponding M.L.E. algorithm is

$$(\hat{t}_0, \hat{\theta}_0, \hat{\theta}_1) = \arg \max_{1 \leq k \leq N} \sup_{\theta_0} \sup_{\theta_1} \ln \left[\prod_{i=1}^{k-1} p_{\theta_0}(y_i) \prod_{i=k}^{N} p_{\theta_1}(y_i) \right] \tag{2.6.7}$$

which can be condensed into

$$\hat{t}_0 = \arg \max_{1 \leq k \leq N} \ln \left[\prod_{i=1}^{k-1} p_{\hat{\theta}_0}(y_i) \prod_{i=k}^{N} p_{\hat{\theta}_1}(y_i) \right] \tag{2.6.8}$$

where $\hat{\theta}_0$ is the M.L.E. estimate of θ_0 based on the observations $y_1, ..., y_{k-1}$, and $\hat{\theta}_1$ is the M.L.E. estimate of θ_1 based upon the observations $y_k, ..., y_N$. When θ_0 and θ_1 are assumed to be known, this can be simplified to

$$\hat{t}_0 = \arg \max_{1 \leq k \leq N} \ln \left[\prod_{i=1}^{k-1} p_{\theta_0}(y_i) \prod_{i=k}^{N} p_{\theta_1}(y_i) \right] \tag{2.6.9}$$

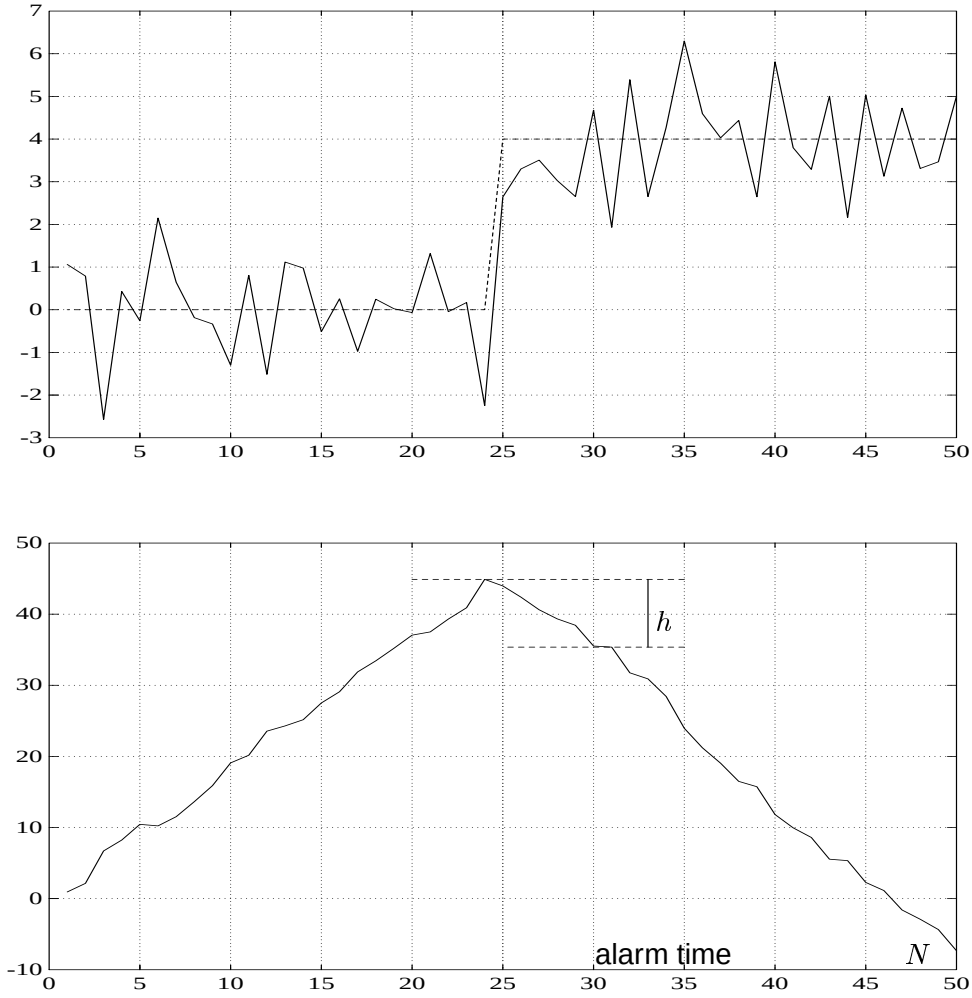

Figure 2.17 Estimation of the change time. The MLE of the change time is the abscissa of the maximum value of the cumulative sum S_k^N.

and rewritten as

$$\hat{t}_0 = \arg \max_{1 \leq k \leq N} \left[\ln \frac{\prod_{i=k}^{N} p_{\theta_1}(y_i)}{\prod_{i=k}^{N} p_{\theta_0}(y_i)} + \ln \prod_{i=1}^{N} p_{\theta_0}(y_i) \right] \tag{2.6.10}$$

The second term on the right of this equation is constant for a given sample. Therefore, the estimate of the change time is

$$\hat{t}_0 = \arg \max_{1 \leq k \leq N} \sum_{i=k}^{N} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \tag{2.6.11}$$

The geometrical interpretation of this estimation method is depicted in figure 2.17, in which we plot the cumulative sum :

$$S_k^N = \sum_{i=k}^{N} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \tag{2.6.12}$$

The figure shows that the M.L.E. of t_0 is the abscissa of the maximum value of this sum. Let us add some further comments about the relationship between this algorithm and the CUSUM algorithm described in subsection 2.2.3. Formula (2.6.11) can be rewritten as

$$\hat{t}_0 = \arg \min_{1 \leq k \leq N} \quad \sum_{i=1}^{k-1} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \tag{2.6.13}$$

which has the following geometrical interpretation. Let us return once more to figure 2.5. From the previous formula, it is obvious that the estimate $\hat{t}_0$ is one plus the abscissa of the minimum value of the cumulative sum plotted in this figure. On the other hand, the on-line CUSUM algorithm can be geometrically interpreted with the aid of figure 2.17 in the following manner. The alarm of this on-line algorithm is set when the deviation of the cumulative sum S_k^N with respect to its current maximum value is greater than the threshold h. If you look at figure 2.17 *both upside down and from the back*, you see that you exactly recover the picture of figure 2.5. From this explanation, it is obvious that estimate (2.6.13) can be rewritten as in (2.2.18).

Example 2.6.1 (Change in mean - contd.). *We continue the investigation of the Gaussian independent case, and we assume that the variance σ^2 is known, but that the two mean values μ_0 before and μ_1 after the change are unknown. In this case, the M.L.E. formula (2.6.8) can be written as*

$$\hat{t}_0 = \arg \max_{1 \leq k \leq N} \left\{ - \left[\sum_{i=1}^{k-1} (y_i - \hat{\mu}_0)^2 + \sum_{i=k}^{N} (y_i - \hat{\mu}_1)^2 \right] \right\} \tag{2.6.14}$$

where we canceled the terms that do not modify the argument of the maximization. By replacing the estimates by their values, which are the relevant empirical means of the observations,

$$\hat{\mu}_0 = \frac{1}{k-1} \sum_{i=1}^{k-1} y_i \tag{2.6.15}$$

and

$$\hat{\mu}_1 = \frac{1}{N-k+1} \sum_{i=k}^{N} y_i \tag{2.6.16}$$

we obtain, after straightforward manipulations,

$$\hat{t}_0 = \arg \max_{1 \leq k \leq N} \left[-(k-1)(N-k+1)(\hat{\mu}_0 - \hat{\mu}_1)^2 \right] \tag{2.6.17}$$

The geometrical interpretation is the same as before in figure 2.17.

Let us give a further interpretation of (2.6.14) in terms of least-squares estimation. This equation can be rewritten as

$$\hat{t}_0 = \arg \min_{1 \leq k \leq N} \inf_{\mu_0, \mu_1} \left[\sum_{i=1}^{k-1} (y_i - \mu_0)^2 + \sum_{i=k}^{N} (y_i - \mu_1)^2 \right] \tag{2.6.18}$$

In other words, we use a least-squares estimation algorithm for the following piecewise regression problem :

$$y_k = \mu(k) + \varepsilon_k \tag{2.6.19}$$

where $\mathcal{L}(\varepsilon_k) = \mathcal{N}(0, \sigma^2)$ and

$$\mu(k) = \begin{cases} \mu_0 & \text{if} \quad k < t_0 \\ \mu_1 & \text{if} \quad k \geq t_0 \end{cases} \tag{2.6.20}$$

as depicted in figure 2.18. This problem is the simplest case of the more complex problem of choice of segments for piecewise approximation, which is also called two-phase regression. More details can be found in [Quandt, 1958, Quandt, 1960, Hinkley, 1969, Hinkley, 1971, Seber, 1977].

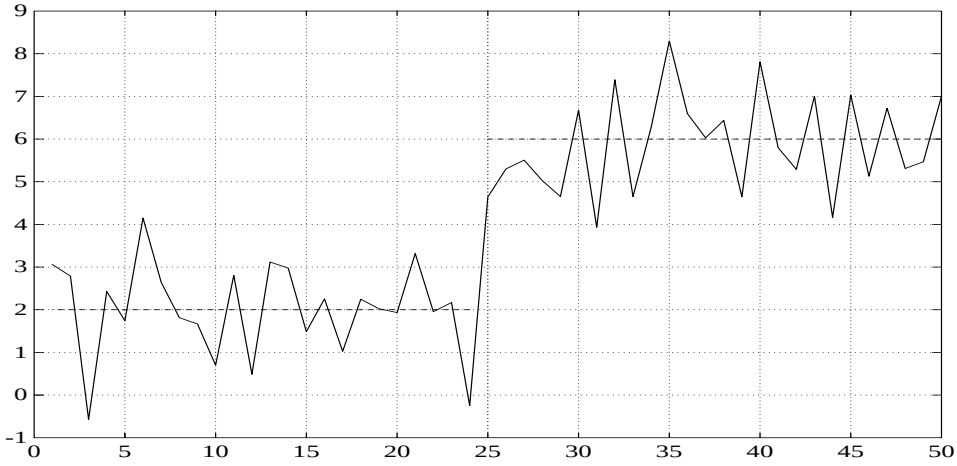

Figure 2.18 Least-squares regression : piecewise constant mean (dotted line), and corresponding Gaussian signal (solid line).

2.7 Notes and References

Section 2.1

All these algorithms were introduced for solving problems in quality control [Duncan, 1986], which is the origin of the word "chart", as used in this context. The first proposed algorithm was Shewhart's control chart [Shewhart, 1931], which was investigated further in [Page, 1954c]. The geometric moving average algorithm was introduced in [S.Roberts, 1959] as a more efficient alternative to Shewhart's chart in many cases. Another alternative, finite moving average chart, was introduced in [Page, 1954a, Laï, 1974]. A close although essentially different algorithm, the filtered derivative algorithm, was introduced in [Basseville *et al.*, 1981]; this algorithm is similar to the gradient techniques used for edge detection in image processing [L.Roberts, 1965].

Section 2.2

The CUSUM algorithm was introduced in [Page, 1954a]. The literature concerning this algorithm is quite extensive [Phillips, 1969, Woodward and Goldsmith, 1964, Van Dobben De Bruyn, 1968, Hinkley, 1969, Hinkley, 1970, Hinkley, 1971]. One reason for this situation is the optimal property of this algorithm, which was proved in [Lorden, 1971]. This algorithm is also often referred to as Shiryaev's SPRT [Shiryaev, 1961].

Section 2.3

Bayesian techniques for change detection were introduced in [Girshick and Rubin, 1952], further developed and investigated in [Shiryaev, 1961, Shiryaev, 1963, Shiryaev, 1965, S.Roberts, 1966], and more recently in [Shiryaev, 1978, Pollak, 1985, Pollak, 1987]. They were initially the result of the first attempt to solve change detection problems in quality control with the aid of a formal mathematical problem statement. The optimal properties of these algorithms were obtained *before* the proof of optimality of CUSUM techniques, and with the aid of slightly different criteria.

Section 2.4

In the case of an unknown parameter after change, the GLR algorithm was derived in [Lorden, 1971] as a generalization of the CUSUM algorithm for this situation. The interest in this algorithm is justified by its "uniformly optimal properties" [Lorden, 1971, Lorden, 1973]. This algorithm is less efficient than the CUSUM algorithm because it does not require the precise knowledge of the parameter after change. Furthermore, the possibility of adapting it to more complex situations makes this algorithm quite attractive. Another less sensitive algorithm is the weighted CUSUM algorithm introduced in [Pollak and Siegmund, 1975]. The χ^2-CUSUM algorithm was introduced in [Nikiforov, 1980, Nikiforov, 1986].

Section 2.5

To our knowledge, the idea of using a change detection algorithm to improve the performance of an adaptive identification algorithm was introduced in [Willsky and Jones, 1976], which is an extension of the work in [MacAulay and Denlinger, 1973]. For earlier investigations concerning the joint use of detection and identification, the reader is referred to [Lainiotis, 1971]. In the present framework of a change in a scalar parameter, the CUSUM algorithm was used in [Perriot-Mathonna, 1984, Favier and Smolders, 1984, Bivaikov, 1991]. Similar attempts, although not based on the same detection algorithms, can be found in [Hägglund, 1983, Chen and Norton, 1987, Mariton *et al.*, 1988].

Section 2.6

The off-line hypotheses testing problem was first addressed in [Page, 1957]. Other investigations can be found in [Deshayes and Picard, 1986, Siegmund, 1985b]. The off-line estimation of a change time was originally obtained in [Page, 1957]. The literature on this issue is extensive [Hinkley, 1969, Hinkley, 1970, Hinkley, 1971, Kligiene and Telksnys, 1983, Picard, 1985, Deshayes and Picard, 1986].

2.8 Summary

Main notation :

$$\begin{aligned} s_i &= \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \\ S_j^k &= \sum_{i=j}^{k} s_i; \qquad S_k = S_1^k \\ t_a &= \min\{k : g_k \geq h\} \end{aligned}$$

For the basic example of a change in the mean μ of a Gaussian distribution with constant variance σ^2, we also use the notation :

$$\begin{aligned} b &= \frac{\mu_1 - \mu_0}{\sigma} \\ s_i &= \frac{b}{\sigma}\left(y_i - \frac{\mu_0 + \mu_1}{2}\right) \\ \tilde{S}_j^k &= \frac{1}{\sigma}\sum_{i=j}^{k}(y_i - \mu_0) \end{aligned}$$

Elementary Algorithms

Shewhart control chart

$$g_{KN} = S_1^N(K) = S_{N(K-1)+1}^{NK}$$

where K is the sample number. The tuning parameters are the size N of the sample of observations tested and the threshold h.

GMA algorithm

$$g_k = (1-\alpha)g_{k-1} + \alpha s_k, \text{ with: } g_0 = 0$$

The tuning parameters are the weight $0 < \alpha \leq 1$ and the threshold h.

FMA algorithm

$$g_k = \sum_{i=0}^{N} \gamma_i \ln \frac{p_{\theta_1}(y_{k-i})}{p_{\theta_0}(y_{k-i})}$$

The tuning parameters are the size N of the sliding window, the weights γ_i, which are any weights for causal filters, and the threshold h.

Filtered derivative algorithm

$$\begin{aligned} \nabla g_k &= g_k - g_{k-1} \\ t_a &= \min\{k : \sum_{i=0}^{N} \mathbf{1}_{\{\nabla g_{k-i} \geq h\}} \geq \eta\} \end{aligned}$$

The tuning parameters are again the size N of the sliding window, the weights γ_i, which are any weights for causal filters, the threshold h, and the counter of alarms η. For the basic example, two useful choices are

$$\begin{aligned} \nabla g_k &= y_k - y_{k-N} \\ \nabla g_k &= \sum_{i=0}^{N-1} y_{k-i} - \sum_{i=N}^{2N-1} y_{k-i} \end{aligned}$$

CUSUM Algorithm

Intuitive derivation of the CUSUM algorithm

$$\begin{aligned} g_k &= S_k - m_k \\ m_k &= \min_{1 \leq j \leq k} S_j \end{aligned}$$

The stopping rule can thus be rewritten as

$$t_a = \min\{k : S_k \geq m_k + h\}$$

or equivalently as an integrator compared to an adaptive threshold.

CUSUM as a repeated SPRT The CUSUM algorithm can be recursively written as

$$g_k = (g_{k-1} + s_k)^+$$

or equivalently as

$$\begin{aligned} g_k &= \left(S^k_{k-N_k+1}\right)^+ \\ N_k &= N_{k-1} \cdot \mathbf{1}_{\{g_{k-1}>0\}} + 1 \end{aligned}$$

The CUSUM algorithm can thus be seen as a random size sliding window algorithm.

Off-line derivation

$$g_k = \max_{1\le j\le k} S^k_j$$

The estimate of the change time is

$$\hat{t}_0 = t_a - N_{t_a} + 1$$

Two-sided CUSUM algorithm For the basic example,

$$\begin{aligned} t_a &= \min\{k : (g^+_k \ge \bar{h}) \cup (g^-_k \ge \bar{h})\} \\ g^+_k &= \left(g^+_{k-1} + y_k - \mu_0 - \frac{\nu}{2}\right)^+ \\ g^-_k &= \left(g^-_{k-1} - y_k + \mu_0 - \frac{\nu}{2}\right)^+ \end{aligned}$$

Bayes-type Algorithms

$$g_k = \ln(\varrho + e^{g_{k-1}}) - \ln(1-\varrho) + \ln\frac{p_{\theta_1}(y_k)}{p_{\theta_0}(y_k)}$$

The tuning parameters of this Bayes-type algorithm are the *a priori* probability ϱ of a change, the initial probability π implicit in g_0, and the threshold h.

Unknown Parameter After Change

χ^2-CUSUM algorithm For the basic example,

$$g_k = \max_{1\le j\le k}\left[\ln\cosh\left(b\tilde{S}^k_j\right) - \frac{b^2}{2}(k-j+1)\right]$$

GLR algorithm

$$g_k = \max_{1\le j\le k} \sup_{\theta_1} S^k_j(\theta_1)$$

For the basic example, the second maximization is explicit :

$$\begin{aligned} g_k &= \max_{1\le j\le k} \sum_{i=j}^{k}\left[\frac{\hat{\nu}_j(y_i-\mu_0)}{\sigma^2} - \frac{\hat{\nu}_j^2}{2\sigma^2}\right] \\ \hat{\nu}_j &= \frac{1}{k-j+1}\sum_{i=j}^{k}(y_i-\mu_0) \end{aligned}$$

Off-line Change Detection

Off-line hypotheses testing

$$\Lambda_N = \max_{1 \le j \le N} \Lambda_1^N(j)$$

$$\tilde{\Lambda}_N = \max_{1 \le j \le N} \sup_{\tilde{\theta}_0} \sup_{\theta_0} \sup_{\theta_1} \Lambda_1^N(j, \tilde{\theta}_0, \theta_0, \theta_1)$$

Off-line estimation

$$\hat{t}_0 = \arg \max_{1 \le k \le N} \sum_{i=k}^{N} \ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)}$$