CHAPTER 3

SEQUENTIAL ANALYSIS

The main reason for including this chapter in the thesis is to provide the theoretical background knowledge required to implement CUSUM as a sequential hypertemporal remote sensing change detection algorithm. It is recommended to first study Section A.1 if the reader is unfamiliar with stochastic calculus. Stochastic calculus provides the mathematical framework needed to understand and study sequential analysis.

In this chapter, different statistical techniques are investigated either to classify observations or to detect changes in the underlying distribution of observation. All the techniques investigated have no pre-determined sample size and are thus purely *sequential* or *on-line*. The study of statistical sequential classification and change detection techniques is known collectively as *sequential analysis*. Good literature reviews can be found in [24, 25] on the subject of sequential analysis. The main advantage of a sequential approach is that on average sequential approaches require fewer observations than a fixed-sample-size approach while maintaining the same probability of error. The reason for this is that sequential algorithms terminate uniquely for each observable sequence. In an ambiguous case the algorithm will take longer to terminate than in an unambiguous case [23]. The chapter starts with Neyman and Pearson's 1933 seminal result [44], which provides an optimal fixed-sample-size classification strategy. Neyman and Pearson's result inspired Wald [45, 46] to develop a sequential solution to the classification problem during the 1940s. Optimality was subsequently proven by Wald and Wolfowitz in 1948 [47]. The sequential classification problem, also known as *sequential detection*, is discussed in two frameworks, namely in Wald's framework (frequentest) (Section 3.3) [46] and in a Bayesian framework (Section 3.4) [48]. From sequential detection the chapter progresses to a group of change detection algorithms grouped under the collective name of *quickest detection*. Quickest detection techniques are statistical techniques capable of detecting a change as quickly as possible

after it occurs (using different measures for the delay). Statistical change detection has its roots in the seminal papers of Shewhart [49] and Page [6]. The quickest detection techniques discussed in this chapter are divided into Bayesian (see Section 3.5) and non-Bayesian (see Section 3.6) approaches. The problem of quickest detection was first cast into a Bayesian framework in 1952 [50], and was subsequently solved in 1963 by Shiryaev [51]. The most famous non-Bayesian change detection algorithm is arguably the CUSUM stopping time, first developed by Page (see Section 3.6.1) [6]. It has been shown that the CUSUM stopping time is asymptotically optimal [52] (when employing the *worst case expected delay* as a performance measure). The asymptotically optimal result was later extended by showing that the CUSUM stopping time is in fact exactly optimal [53, 54]. An alternative to the CUSUM stopping rule was proposed in 1966 and is known as the *Shiryaev-Roberts* stopping time (see Section 3.6.2) [51, 55]. An extension to the Shiryaev-Roberts stopping time known as the *Shiryaev-Roberts-Pollak* stopping time was developed in 1985 by Pollak [56]. The Shiryaev-Roberts-Pollak method is a third-order asymptotically optimal sequential procedure when employing *Pollak's performance measure* (which is less restrictive than the worst case expected delay) [56]. More recently the Shiryaev-Roberts-Pollak stopping time was extended to the *deterministic Shiryaev-Roberts* stopping time [57, 58], which can uniformly outperform both Shiryaev-Roberts and Shiryaev-Roberts-Pollak for appropriately chosen starting conditions [57,58]. For a good theoretical introduction to sequential analysis the reader is referred to [48], while the reader is referred to [59] for an overview that focuses more strongly on implementation specifics.

3.1 NEYMAN-PEARSON

The following section closely follows the notation of [60, 61]. Let $\mathbf{z}^n = \{z_k\}_{\{k=1,2,\cdots,n\}}$ be an independent and identically distributed (i.i.d.) sequence of real observation of size n following one of two hypotheses:

$$\mathscr{H}_0 : z_k \sim Q_0,\ k = 1,2,\cdots,n$$

versus

$$\mathscr{H}_1 : z_k \sim Q_1,\ k = 1,2,\cdots,n;$$

where Q_0 and Q_1 are two probability distributions with associated densities q_0 and q_1, respectively. The *problem* is to determine which hypothesis is true by only looking at the observations. Furthermore, let $q_0(\mathbf{z}^n)$ and $q_1(\mathbf{z}^n)$ denote n-dimensional density functions. Let T be a function of the observations, known as the *test statistic* and let R be the image of T. The image R can be divided into

a *critical region* R_0 and a *region of acceptance* R_1, such that $R_0 \cup R_1 = R$. If $T(\mathbf{z}^n)$ fall into R_0, $\mathscr{H}_0$ is rejected. Constructing a hypothesis test thus requires selecting a test statistic and a critical region. The probability α of rejecting $\mathscr{H}_0$ when it is in fact true is known as the *level of significance* or the *size of the test* and is equal to

$$\alpha = \int_{\{\mathbf{z}^n : T(\mathbf{z}^n) \in R_0\}} q_0(\mathbf{z}^n) d\mathbf{z}^n.$$

The probability α is also known as the probability of a false alarm P_{FA} or the type I error. The *power of the test* $1-\beta$ is defined as the probability of accepting $\mathscr{H}_1$ if it is in fact true, also known as the probability of detection P_{D}, and is equal to

$$1-\beta = \int_{\{\mathbf{z}^n : T(\mathbf{z}^n) \in R_0\}} q_1(\mathbf{z}^n) d\mathbf{z}^n.$$

In other words, β is the probability of accepting $\mathscr{H}_0$ when it is in fact false (type II error).

Neyman and Pearson [44, 61] derived the following theorem, which states that the likelihood ratio Λ is the best possible choice of T. The likelihood ratio maximises the P_{D} given a specific false alarm rate P_{FA}.

Theorem 1 (Neyman-Pearson) *To maximise the P_D for a given P_{FA} decide $\mathscr{H}_1$ if*

$$\Lambda(\mathbf{z}^n) = \frac{q_1(\mathbf{z}^n)}{q_0(\mathbf{z}^n)} > \gamma,$$

where the threshold γ is found from

$$P_{FA} = \int_{\{\mathbf{z}^n : \Lambda(\mathbf{z}^n) > \gamma\}} q_0(\mathbf{z}^n) d\mathbf{z}^n.$$

3.2 KULLBACK-LEIBLER DIVERGENCE

Theorem 1 stipulates that Λ is the optimal test statistic and some of the unique properties of Λ that make it a useful tool when building a classifier or a change detector should therefore be highlighted. Instead of calculating the likelihood ratio

$$\Lambda_k = \prod_{i=1}^{k} \frac{q_1(z_i)}{q_0(z_i)}, \tag{3.1}$$

the log-likelihood ratio

$$S_k = \sum_{i=1}^{k} s_i, \tag{3.2}$$

where

$$s_i = \ln \frac{q_1(z_i)}{q_0(z_i)}, \tag{3.3}$$

could be calculated. The sum S_k is derived from $\ln\Lambda_k$.

In probability theory and information theory, the Kullback-Leibler divergence is a non-symmetric measure of the difference between two probability distributions Q_0 and Q_1 and is defined as

$$\begin{aligned} D_{\mathrm{KL}}(Q_1\|Q_0) &= \int_{-\infty}^{\infty} q_1(z_1) ln \frac{q_1(z_1)}{q_0(z_1)} dz_1 \\ &= \mathbb{E}_1\left[ln\frac{q_1(z_1)}{q_0(z_1)}\right] \\ &= \mathbb{E}_1[s_1]. \end{aligned}$$

The reason why z_1 can be used is that the sequence $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ is a sample path of a discrete, stationary stochastic process. In other words, the density of the first observation of multiple sample paths equals the density from which the sequence $\mathbf{z}$ is drawn. Kullback-Leibler divergence is always positive, implying that S_k will have positive drift under $\mathscr{H}_1$, because $\mathbb{E}_1[s_1] > 0$. Under $\mathscr{H}_0$, S_k will have negative drift, since $\mathbb{E}_0[s_1] = -\int_{-\infty}^{\infty} q_0(z) ln\frac{q_0(z)}{q_1(z)} dz = -D_{\mathrm{KL}}(Q_0\|Q_1) < 0$. It should be perfectly clear that $\ln\Lambda$ is a good statistic to use when building a classifier, since under $\mathscr{H}_1$, S_k experiences positive drift, while under $\mathscr{H}_0$, S_k experiences negative drift [59].

3.3 HYPOTHESIS TESTING: WALD'S FORMULATION

The following section closely follows the notation of [59]. The problem with Neyman-Pearson is that the sample size has to be chosen before the threshold can be computed and therefore the algorithm is non-sequential. In contrast with Neyman-Pearson, Wald's formulation, the Sequential Probability Ratio Test (SPRT), is a sequential approach and is the Uniformly Most Efficient (UME) test among all sequential tests. In general the art of classifying as quickly (no predetermined sample size) and accurately as possible is known as *sequential detection*.

The problem introduced in Section 3.1 is now restated without limiting the sample size. Consider the sequence $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ of i.i.d. real observation following one of two hypotheses:

$$\mathscr{H}_0 : z_k \sim Q_0,\ k = 1,2,\cdots$$

versus

$$\mathscr{H}_1 : z_k \sim Q_1,\ k = 1,2,\cdots;$$

where Q_0 and Q_1 are two probability distributions with associated densities q_0 and q_1, respectively.

Furthermore, let the sequence $\mathbf{z}$ be adapted to the filtration $\mathscr{F}_k = \sigma(\{z_k\}_{\{k=1,2,\cdots\}})$. The *problem* is to determine which hypothesis is true by only looking at the observations.

The above problem can be solved by using a *sequential statistical test*. A sequential statistical test for testing between hypotheses $\mathscr{H}_0$ and $\mathscr{H}_1$ is defined by a *sequential decision rule*, which is the pair (δ, T), where T is a *stopping time* and δ a *decision function*. In the case of the SPRT the stopping time is equal to

$$T = T^{\text{SPRT}}_{\{A,B\}} = \inf\{k \geq 0 | \Lambda_k \notin (A,B)\}, \tag{3.4}$$

where Λ_k was defined in Equation 3.1, and the decision function (after stopping) is

$$\delta_T = \begin{cases} 0 & \text{when } \Lambda_T \leq A \\ 1 & \text{when } \Lambda_T \geq B. \end{cases}$$

In other words, Wald's test keeps on sampling until the likelihood ratio crosses the *exit thresholds* A or B, at which time a decision is made. If Λ_T is less or equal to A, hypothesis $\mathscr{H}_0$ is accepted, if Λ_T is greater or equal to B, hypothesis $\mathscr{H}_1$ is accepted. The type I error α is equal to the probability $P_0(\delta_T = 1)$, where the subscript refers to the fact that $\mathscr{H}_0$ is assumed to be true. The probability of a type II error β is equal to the probability $P_1(\delta_T = 0)$.

The log-likelihood ratio S_k (Equation 3.2) can also be used instead of Λ_k to derive the sequential decision rule, (δ, T). In the log-likelihood domain the SPRT stopping time is equal to

$$T = T^{\text{SPRT}}_{\{-a,h\}} = \inf\{k \geq 0 | S_k \notin (-a,h)\}, \tag{3.5}$$

where $\ln A = -a$ and $\ln B = h$. The decision rule now becomes

$$\delta_T = \begin{cases} 0 & \text{when } S_T \leq -a \\ 1 & \text{when } S_T \geq h. \end{cases}$$

Overshoot is an important concept that is often used to analyse the performance of the SPRT algorithm and should therefore be defined formally. Let

$$\mathscr{O}(T, S_T, -a, h, \delta_T) = \begin{cases} |S_T + a| & \text{when } \delta_T = 0 \\ |S_T - h| & \text{when } \delta_T = 1. \end{cases} \tag{3.6}$$

When inspecting Equation 3.6 it should be clear to the reader that $\mathscr{O}(T, S_T, -a, h, \delta_T)$ is a random variable. If $\mathscr{O}(T, S_T, -a, h, \delta_T) \equiv 0$ then there is no overshoot (S_T always equals one of the boundaries $\{a, h\}$), however when $\mathscr{O}(T, S_T, -a, h, \delta_T) \not\equiv 0$ then overshoot does occur.

As mentioned before, Wald's SPRT is the UME test among all sequential tests. This fact is formally stated by the Wald-Wolfowitz theorem, given below without proof [47,48].

Theorem 2 (Wald-Wolfowitz) *Suppose* (T,δ) *is the Sequential Probability Ratio Test, SPRT*(A,B) *with* $0 < A \leq 1 \leq B < \infty$, *and let* (T',δ') *denote any other sequential decision rule with* $\max\{\mathbb{E}_0[T'],\mathbb{E}_1[T']\} < \infty$, *and satisfying*

$$\alpha' = P_0(\delta'_{T'} = 1) \leq P_0(\delta_T = 1) = \alpha \text{ and } \beta' = P_1(\delta'_{T'} = 0) \leq P_1(\delta_T = 0) = \beta,$$

with

$$P_0(\delta_T = 1) + P_1(\delta_T = 0) < 1.$$

Then

$$\mathbb{E}_0[T'] \geq \mathbb{E}_0[T] \text{ and } \mathbb{E}_1[T'] \geq \mathbb{E}_1[T].$$

At this point the natural question arises, "*How can the thresholds A and B be selected to achieve a certain probability of error?*" It turns out that it is quite complex to find the exact thresholds A and B, but quite simple to find approximations of A and B that typically work well in practice. These practical estimates are known as *Wald's approximations of A and B*. When $\Lambda_k \geq B$, sampling stops and hypothesis $\mathscr{H}_1$ is chosen. Clearly the decision rule leads to [23]

$$\prod_{i=1}^{k} q_1(z_k) \geq B \cdot \prod_{i=1}^{k} q_0(z_k) \implies P_1(\delta = 1) \geq B \cdot P_0(\delta = 1), \tag{3.7}$$

which can be interpreted as the probability of observing z_k under $\mathscr{H}_1$ is at least B times bigger than under $\mathscr{H}_0$. Furthermore, since $\mathscr{H}_1$ was chosen, the type I error is equal to $P_0(\delta = 1)$. Recognising the type II error β to be equal to $P_1(\delta = 0)$, an upper limit for B can be derived by using Equation 3.7 and is equal to

$$B \leq \frac{1-\beta}{\alpha}. \tag{3.8}$$

Similarly a lower limit for A can be derived and is equal to

$$A \geq \frac{\beta}{1-\alpha}. \tag{3.9}$$

Wald's approximations for A and B are derived by replacing the inequalities with equalities in Equation 3.9 and Equation 3.8 and are thus equal to

$$\tilde{A} = \frac{\beta}{1-\alpha} \text{ and } \tilde{B} = \frac{1-\beta}{\alpha}.$$

Wald's approximations for the log-likelihood domain can be found similarly and are equal to

$$-\tilde{a} = \ln\frac{\beta}{1-\alpha} \text{ and } \tilde{h} = \ln\frac{1-\beta}{\alpha}.$$

Alternatively *Wald's approximate error probabilities* can be calculated with the correct exit boundaries A and B, resulting in

$$\tilde{\alpha} = \frac{1-A}{B-A} \text{ and } \tilde{\beta} = A\frac{B-1}{B-A}. \tag{3.10}$$

3.3.1 The OC and ASN functions of the SPRT

The problem stated at the beginning of Section 3.3 is restated with additional information to help in defining the Operating Characteristic (OC) and the Average Sample Number (ASN) functions properly. Consider the sequence $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ of i.i.d. real observations (adapted to the filtration $\mathscr{F}_k$), which obeys one of the two hypotheses,

$$\mathscr{H}_0 : \theta = \theta_0$$

versus

$$\mathscr{H}_1 : \theta = \theta_1;$$

best, where $\theta \in \Theta$ is a unique property of the random variable generating the sequence $\mathbf{z}$, for example θ could represent the mean of a Gaussian random variable and $\Theta \subset \mathbb{R}$ is the possible range of θ. The problem could also be stated for a parameter list $\boldsymbol{\theta}$, but for simplicity it is not done here. If θ is equal to θ_0, the random variable generating the sequence $\mathbf{z}$ will obey density q_0, similarly when θ equals θ_1 the sequence $\mathbf{z}$ will be distributed according to density q_1. Assume now that the sequence $\mathbf{s}$ is derived from $\mathbf{z}$ by using Equation 3.3 and is also i.i.d.. Under $\mathscr{H}_0$ the random variable generating $\mathbf{s}$ will have density f_0 and under $\mathscr{H}_1$ the random variable will have density f_1, which is shorthand for f_{θ_0} and f_{θ_1}, respectively. In general, when the unique property of $\mathbf{z}$ is equal to θ, the sequence $\mathbf{s}$ will be distributed according to density f_θ. In effect the parameter θ determines the density of the random variable generating $\mathbf{s}$.

The probability $\mathscr{Q}(\theta)$ of accepting hypothesis $\mathscr{H}_0$, treated as a function of $\theta \in \Theta$, when the exit thresholds $-a$ and h are fixed, is called the *operating characteristic* function. In other words the type I error α is equal to $1-\mathscr{Q}(\theta_0)$ and the type II error β is equal to $\mathscr{Q}(\theta_1)$ [59].

The *average sample number* $\mathbb{E}_\theta[T]$ is the mean number of sample points required to make a decision when performing a hypothesis test, with fixed exit thresholds $-a$ and h, as a function of $\theta \in \Theta$. In

the cases where θ equals θ_0 or θ_1 the shorthand notations $\mathbb{E}_0[T]$ and $\mathbb{E}_1[T]$ are used. If the sequence $\mathbf{z}$ follows $\mathscr{H}_0$, the expected number of samples required to make a decision is equal to $\mathbb{E}_0[T]$, while $\mathbb{E}_1[T]$ is defined similarly [59].

3.3.2 Wald's approximations

The OC function $\mathscr{Q}(\theta)$ can be approximated by $\tilde{\mathscr{Q}}(\theta)$ with equation

$$\tilde{\mathscr{Q}}(\theta) = \begin{cases} \frac{e^{-\omega_0(\theta)h}-1}{e^{-\omega_0(\theta)h}-e^{\omega_0(\theta)a}} & \text{when } \mathbb{E}_\theta[s_1] \neq 0 \\ \frac{h}{h+a} & \text{when } \mathbb{E}_\theta[s_1] = 0, \end{cases} \tag{3.11}$$

where $\omega_0(\theta)$ is the unique non-zero real number which satisfies

$$\begin{aligned} \mathbb{E}_\theta[e^{-\omega_0(\theta)s_1}] &= \int_{-\infty}^{\infty} e^{-\omega_0(\theta)s_1} f_\theta(s_1) ds_1 \\ &= 1, \end{aligned} \tag{3.12}$$

if a non-zero solution exists, otherwise $\omega_0(\theta) = 0$, and $\mathbb{E}_\theta[s_1]$ is defined as

$$\mathbb{E}_\theta[s_1] = \int_{-\infty}^{\infty} s_1 f_\theta(s_1) ds_1. \tag{3.13}$$

Wald's approximation of $\mathscr{Q}(\theta)$ is derived from the following well-known *identity of Wald* (see Theorem 6) [48]

$$\mathbb{E}_\theta\left[e^{-\omega S_T}\left(\mathbb{E}_\theta\left[e^{-\omega s_1}\right]\right)^{-T}\right] = 1, \tag{3.14}$$

where T is the SPRT stopping time defined in Equation 3.5. Wald's identity is valid for all $\omega \in \{\omega | \mathbb{E}_\theta[e^{-\omega s_1}] < \infty\}$. Equation 3.14 can be transformed into

$$\mathbb{E}_\theta\left[e^{-\omega S_T - T\ln\mathbb{E}_\theta[e^{-\omega s_1}]}\right] = 1, \tag{3.15}$$

trivially. If ω is substituted with $\omega_0(\theta)$, Equation 3.15 reduces to

$$\mathbb{E}_\theta\left[e^{-\omega_0(\theta)S_T}\right] = 1. \tag{3.16}$$

If the excess over the boundaries $-a$ and h is ignored, S_T is approximately equal to either $-a$ or h, and Equation 3.16 becomes

$$e^{-\omega_0(\theta)h}[1 - P_\theta(S_T \leq -a)] + e^{\omega_0(\theta)a} P_\theta(S_T \leq -a) \approx 1, \tag{3.17}$$

where $P_\theta(S_T \leq -a)$ is equal to $\mathcal{Q}(\theta)$. By making $\mathcal{Q}(\theta)$ the subject of Equation 3.17, Equation 3.11 is obtained.

The ASN function $\mathbb{E}_\theta(T)$ can be approximated via $\tilde{\mathbb{E}}_\theta(T)$ with equation

$$\tilde{\mathbb{E}}_\theta[T] = \begin{cases} \frac{-a\hat{\mathcal{Q}}(\theta)+h(1-\hat{\mathcal{Q}}(\theta))}{\mathbb{E}_\theta[s_1]} & \text{when } \mathbb{E}_\theta[s_1] \neq 0 \\ \frac{a^2\hat{\mathcal{Q}}(\theta)+h^2(1-\hat{\mathcal{Q}}(\theta))}{\mathbb{E}_\theta[s_1^2]} & \text{when } \mathbb{E}_\theta[s_1] = 0. \end{cases} \tag{3.18}$$

The approximation of the ASN function was also derived through another *identity of Wald*, namely

$$\mathbb{E}_\theta[T] = \begin{cases} \frac{\mathbb{E}_\theta[S_T]}{\mathbb{E}_\theta[s_1]} & \text{if } \mathbb{E}[s_1] \neq 0 \\ \frac{\mathbb{E}_\theta[S_T^2]}{\mathbb{E}_\theta[s_1^2]} & \text{if } \mathbb{E}[s_1] = 0, \end{cases} \tag{3.19}$$

where $\mathbb{E}_\theta[S_T]$ equals

$$-a\tilde{\mathcal{Q}}(\theta) + h[1-\tilde{\mathcal{Q}}(\theta)], \tag{3.20}$$

and $\mathbb{E}_\theta[S_T^2]$ equals

$$a^2\tilde{\mathcal{Q}}(\theta) + h^2[1-\tilde{\mathcal{Q}}(\theta)], \tag{3.21}$$

if the excess over the boundaries is ignored. The stopping time T used in Wald's identity is again the SPRT stopping time in Equation 3.5. The result of substituting Equation 3.20 and Equation 3.21 into Equation 3.19 is Equation 3.18.

Wald's approximations can be restated in the likelihood domain in which case $\tilde{\mathcal{Q}}(\theta)$ and $\tilde{\mathbb{E}}_\theta[T]$ can be expressed as

$$\tilde{\mathcal{Q}}(\theta) = \begin{cases} \frac{B^{-\omega_0(\theta)}-1}{B^{-\omega_0(\theta)}-A^{-\omega_0(\theta)}} & \text{when } \mathbb{E}_\theta(s_1) \neq 0 \\ \frac{\ln B}{\ln(BA^{-1})} & \text{when } \mathbb{E}_\theta(s_1) = 0, \end{cases}$$

and

$$\tilde{\mathbb{E}}_\theta[T] = \begin{cases} \frac{\ln A\hat{\mathcal{Q}}(\theta)+\ln B(1-\hat{\mathcal{Q}}(\theta))}{\mathbb{E}_\theta[s_1]} & \text{when } \mathbb{E}_\theta[s_1] \neq 0 \\ \frac{(\ln A)^2\hat{\mathcal{Q}}(\theta)+(\ln B)^2(1-\hat{\mathcal{Q}}(\theta))}{\mathbb{E}_\theta[s_1^2]} & \text{when } \mathbb{E}_\theta[s_1] = 0. \end{cases}$$

In the special case when $\theta = \theta_0$ or $\theta = \theta_1$ then $\tilde{\mathcal{Q}}(\theta_0)$ and $\tilde{\mathcal{Q}}(\theta_1)$ reduces to

$$\tilde{\mathcal{Q}}(\theta_0) = \frac{B-1}{B-A} \text{ and } \tilde{\mathcal{Q}}(\theta_1) = A\frac{B-1}{B-A},$$

respectively, which is nothing more than the approximations already stated in Equation 3.10. The function $\tilde{\mathbb{E}}_\theta(T)$ also simplifies in the two special cases $\theta = \theta_0$ and $\theta = \theta_1$ to

$$\tilde{\mathbb{E}}_0[T] = (\mathbb{E}_0[s_1])^{-1}\left[\tilde{\alpha}\ln\left(\frac{1-\tilde{\beta}}{\tilde{\alpha}}\right) + (1-\tilde{\alpha})\ln\left(\frac{\tilde{\beta}}{1-\tilde{\alpha}}\right)\right], \tag{3.22}$$

$$\tilde{\mathbb{E}}_1[T] = (\mathbb{E}_1[s_1])^{-1}\left[(1-\tilde{\beta})\ln\left(\frac{1-\tilde{\beta}}{\tilde{\alpha}}\right) + \tilde{\beta}\ln\left(\frac{\tilde{\beta}}{1-\tilde{\alpha}}\right)\right], \tag{3.23}$$

respectively, where $\tilde{\alpha}$ and $\tilde{\beta}$ are Wald's probability of error approximations [59].

3.3.3 Exact computation

For a given θ, let $P_\theta(-a|y) = P_\theta(y)$ be the probability that S_k (Equation 3.2), starting from y, reaches the lower bound $-a$, and let $\mathbb{E}_\theta[T|y] = N_\theta(y)$ be the expected number of sample points required by the SPRT algorithm to terminate when S_k starts at y, i.e. $S_k = \sum_{i=1}^{k} s_i + y$ [59]. It should now be clear that $\mathscr{Q}(\theta)$ is equal to $P_\theta(0)$ and $\mathbb{E}_\theta[T]$ is equal to $N_\theta(0)$. It is widely known that $P_\theta(y)$ and $N_\theta(y)$ respectively satisfy the following two Fredholm integral equations of the second kind [62],

$$P_\theta(y) = \int_{-\infty}^{-a-y} f_\theta(s_1)ds_1 + \int_{-a}^{h} P_\theta(s_1)f_\theta(s_1-y)ds_1,\ -a \le y \le h, \tag{3.24}$$

$$N_\theta(y) = 1 + \int_{-a}^{h} N_\theta(s_1)f_\theta(s_1-y)ds_1,\ -a \le y \le h, \tag{3.25}$$

which can be solved through a system of linear equations that approximate Equation 3.24 and Equation 3.25 [63]. The derivation of $P_\theta(y)$ and $N_\theta(y)$ is based upon the theory of a random walk with absorbing and reflecting boundaries (barriers) [59]. See Section 3.3.5.2 for further details. Another exact approach is the Markov method of Brook [64].

3.3.4 Simulation

The easiest way to compute $\mathscr{Q}(\theta)$ and $\mathbb{E}_\theta[T]$ is through simulation. The pseudo-code for computing the OC and ASN functions is given in Listing 3.1 (listed at the end of the chapter). The functions obtained via simulation become more accurate as N becomes larger.

3.3.5 Example: Gaussian random variable

Consider the sequence $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ of i.i.d. real observations (adapted to the filtration $\mathscr{F}_k$) generated by a Gaussian random variable with density $\mathscr{N}(\theta,1)$. The problem is to choose one of two

fixed hypotheses so that the data sequence fits that hypothesis best. The following two hypotheses are under consideration:

$$\mathscr{H}_0 : \theta = \theta_0 = 0$$

versus

$$\mathscr{H}_1 : \theta = \theta_1 = 1.$$

For this example the increment sequence $\mathbf{s}$ becomes $\{s_k\}_{\{k=1,2,\cdots\}} = \{z_k - \frac{1}{2}\}_{\{k=1,2,\cdots\}}$, since

$$\begin{aligned} s_k &= \ln \frac{q_1(z_k)}{q_0(z_k)} \\ &= \ln \frac{e^{-(z_k-1)^2}}{e^{-z_k^2}} \\ &= z_k - \frac{1}{2}, \end{aligned}$$

and is thus also an independent Gaussian sequence with density $f_\theta(s_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(s_1-(\theta-\frac{1}{2}))^2}{2}}$. Example realisations of z_k and S_k when $\theta = \theta_0$ or $\theta = \theta_1$ are given in Figure 3.1, while the probability density functions of the random variable generating z_k and s_k under the same assumption of θ are given in Figure 3.2. Note that the example sequences are classified correctly for the exit thresholds equal to $h = 3$ and $-a = -3$ and that the stopping times are equal to $T = 15$ and $T = 11$ in Figure 3.1c and Figure 3.1d, respectively.

The OC and ASN functions offer insight into the problem given above and will be calculated by using Wald's approximation as well as the exact computational approach presented in Section 3.3.3.

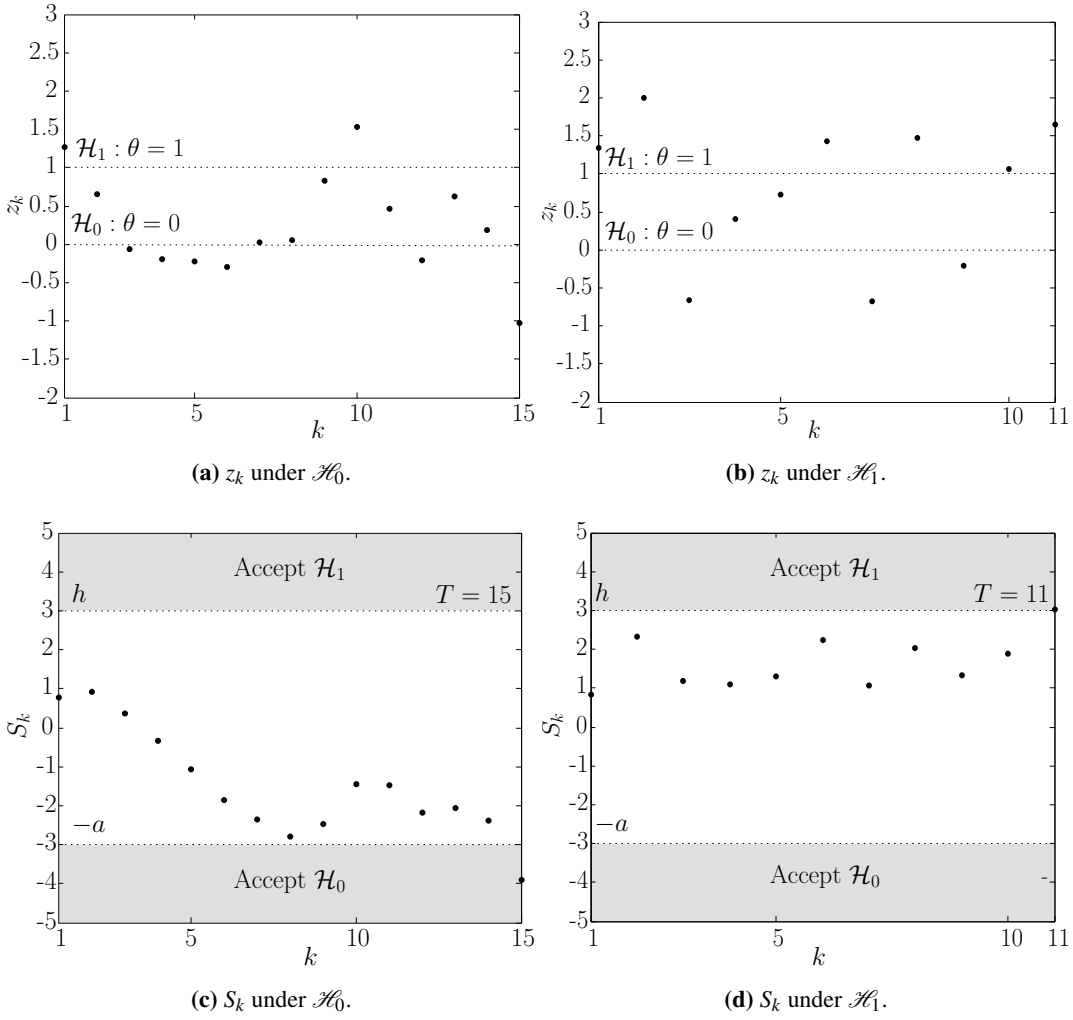
3.3.5.1 Wald's approximation

The first step in calculating the OC and ASN functions is to determine $\omega_0(\theta)$ by using Equation 3.12 and for this example is equal to

$$\begin{aligned} \mathbb{E}[e^{-\omega_0(\theta)s_1}] &= \int_{-\infty}^{\infty} e^{-\omega_0(\theta)s_1} \frac{1}{\sqrt{2\pi}} e^{-\frac{(s_1-(\theta-\frac{1}{2}))^2}{2}} ds_1 \\ &= 1 \\ e^{-\omega_0(\theta)(\theta-\frac{1}{2})+\frac{1}{2}\omega_0^2(\theta)} &= e^0 \\ \omega_0(\theta) &= 2\theta - 1. \end{aligned} \tag{3.26}$$

The next step is to calculate $\mathbb{E}_\theta[s_1]$ by using Equation 3.13, in order to attain

$$\mathbb{E}_\theta[s_1] = \theta - \frac{1}{2}.$$



(a) z_k under $\mathscr{H}_0$.

(b) z_k under $\mathscr{H}_1$.

(c) S_k under $\mathscr{H}_0$.

(d) S_k under $\mathscr{H}_1$.

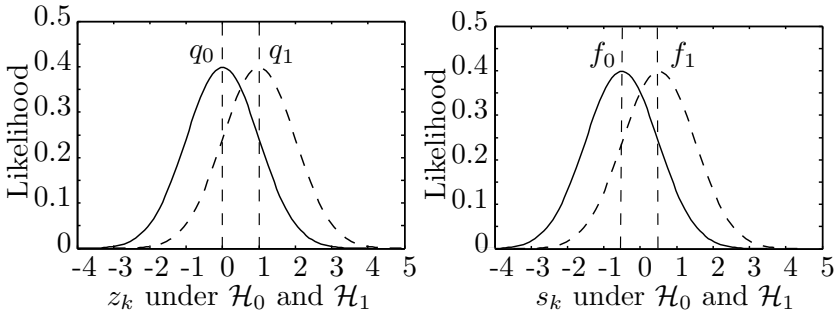
Figure 3.1: Example sequences of z_k and S_k for the unit variance Gaussian example having a mean of $\theta_0 = 0$ and $\theta_1 = 1$ under $\mathscr{H}_0$ and $\mathscr{H}_1$, respectively. The exit thresholds are equal to 3 and -3.

The quantity $\mathbb{E}_\theta[s_1^2]$ is also required and is equal to 1 since $s_1^2 \sim \chi_1^2$. The approximate OC function is determined by substituting Equation 3.26 into Equation 3.11 to obtain

$$\tilde{\mathscr{Q}}(\theta) = \begin{cases} \frac{e^{-(2\theta-1)h}-1}{e^{-(2\theta-1)h}-e^{(2\theta-1)a}} & \text{when } \theta \neq \frac{1}{2} \\ \frac{h}{h+a} & \text{when } \theta = \frac{1}{2}. \end{cases} \tag{3.27}$$

Wald's approximated OC function for the Gaussian example is presented in Figure 3.3.

The approximate ASN function is calculated by substituting $\mathbb{E}_\theta[s_1]$, $\mathbb{E}_\theta[s_1^2]$ and Equation 3.27 into



(a) Probability density function of z_k. **(b)** Probability density function of s_k.

Figure 3.2: Probability density functions of the random variable generating z_k and s_k under $\mathscr{H}_0$ and $\mathscr{H}_1$.

Equation 3.18, which gives

$$\tilde{\mathbb{E}}_\theta[T] = \begin{cases} \frac{1}{\theta-\frac{1}{2}}\left[\frac{1-e^{a(2\theta-1)}}{e^{-h(2\theta-1)}-e^{a(2\theta-1)}}h - \frac{e^{-h(2\theta-1)}-1}{e^{-h(2\theta-1)}-e^{a(2\theta-1)}}a\right] & \text{when } \theta \neq \frac{1}{2} \\ ah & \text{when } \theta = \frac{1}{2}. \end{cases}$$

Wald's approximated ASN function for the Gaussian example is presented in Figure 3.4.

3.3.5.2 Exact computation

By substituting $f_\theta(s_1)$ into Equation 3.24 and Equation 3.25 and applying the method of Gaussian quadrature (Section A.2) [63, 65], Equation 3.24 and Equation 3.25 can be reduced to

$$\tilde{P}(y) = \Phi\left(-a-y-\left(\theta-\frac{1}{2}\right)\right) + \sum_{k=1}^{m} A_k \cdot (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}(y_k-y-(\theta-\frac{1}{2}))^2} \cdot \tilde{P}(y_k), \tag{3.28}$$

$$\tilde{N}(y) = 1 + \sum_{k=1}^{m} A_k \cdot (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}(y_k-y-(\theta-\frac{1}{2}))^2} \cdot \tilde{N}(y_k), \tag{3.29}$$

where $\Phi(y) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{y} e^{-\frac{t^2}{2}} dt$, and, A_k and y_k are, respectively, the weights and roots of the Gaussian quadrature for the interval $[-a, h]$. The θ subscript is dropped to avoid clutter. Equation 3.28 can be replaced by the following system of linear equations

$$A \cdot \tilde{P} = \tilde{B},$$

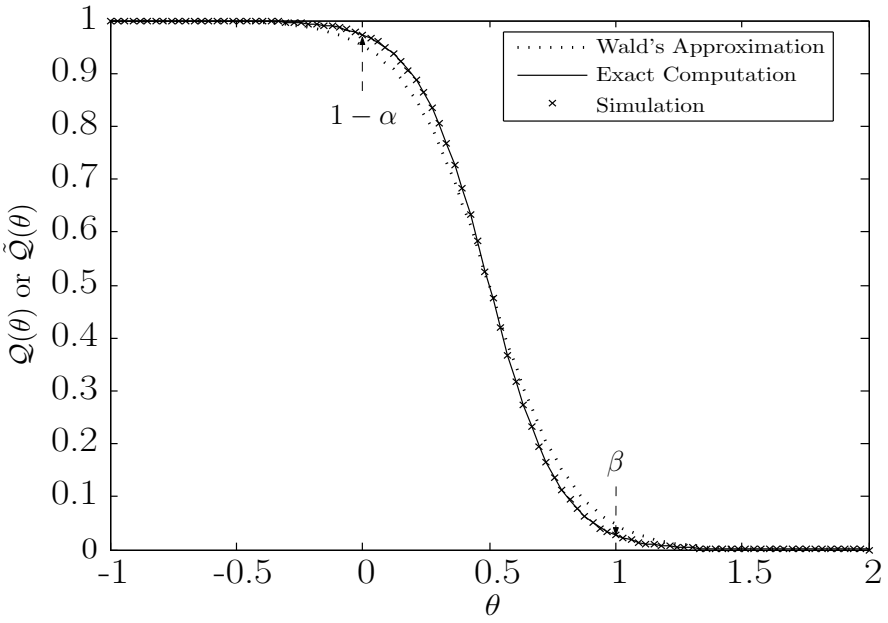


Figure 3.3: The exact OC function and Wald's approximated OC function for the unit variance Gaussian example with mean $\theta = 0$ and $\theta = 1$ under $\mathscr{H}_0$ and $\mathscr{H}_1$ respectively. The exit thresholds are equal to 3 and -3.

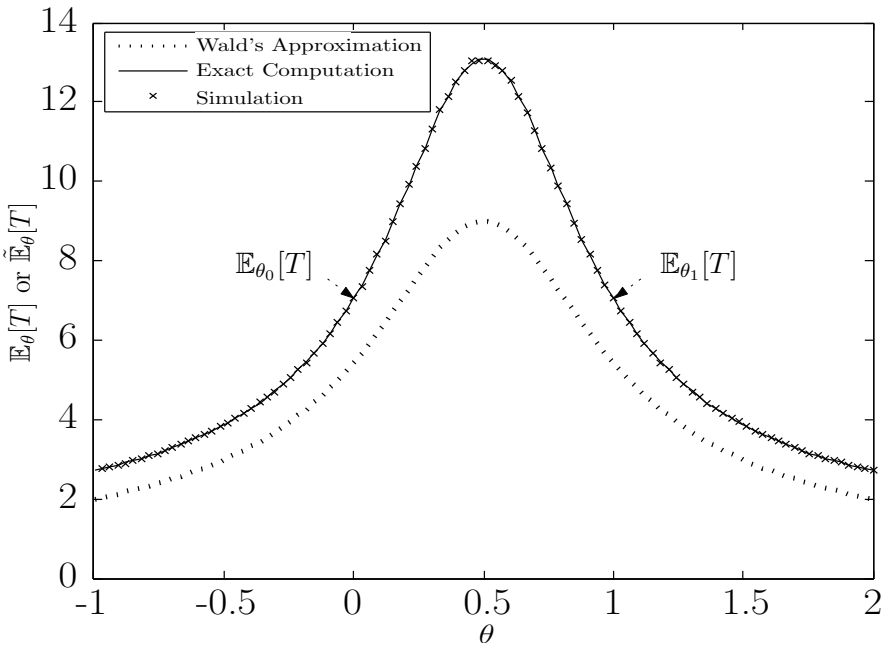


Figure 3.4: The exact ASN function and Wald's approximated ASN function for the unit variance Gaussian example with mean $\theta = 0$ and $\theta = 1$ under $\mathscr{H}_0$ and $\mathscr{H}_1$ respectively. The exit thresholds are equal to 3 and -3.

where the matrix $A(m \times m)$ and column vectors $\tilde{P}(m \times 1)$ and $\tilde{B}(m \times 1)$ are defined by

$$\begin{aligned} A &= (a_{ij}),\ i,j = 1,\cdots,m; \\ \tilde{P}^T &= [\tilde{P}(y_1),\cdots,\tilde{P}(y_m)], \\ \tilde{B}^T &= \left[\Phi\left(-a-y_1-\left(\theta-\frac{1}{2}\right)\right),\cdots,\Phi\left(-a-y_m-\left(\theta-\frac{1}{2}\right)\right)\right], \end{aligned}$$

with

$$\begin{aligned} a_{ij} &= -A_j\psi(y_j,y_i) \text{ for } i \neq j, \\ a_{ii} &= 1 - A_i\psi(y_j,y_i), \\ \psi(y_j,y_i) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_j-y_i-(\theta-\frac{1}{2}))^2}{2}}. \end{aligned}$$

The column vector $\tilde{P}$ can now be solved easily with $\tilde{P} = A^{-1} \cdot \tilde{B}$. The OC function $\mathscr{Q}(\theta)$ is obtained by substituting the column vector $\tilde{P}$ into Equation 3.28 and setting y to naught. Similarly, $\tilde{N}$ can be ascertained by replacing Equation 3.29 with the linear system

$$A \cdot \tilde{N} = I,$$

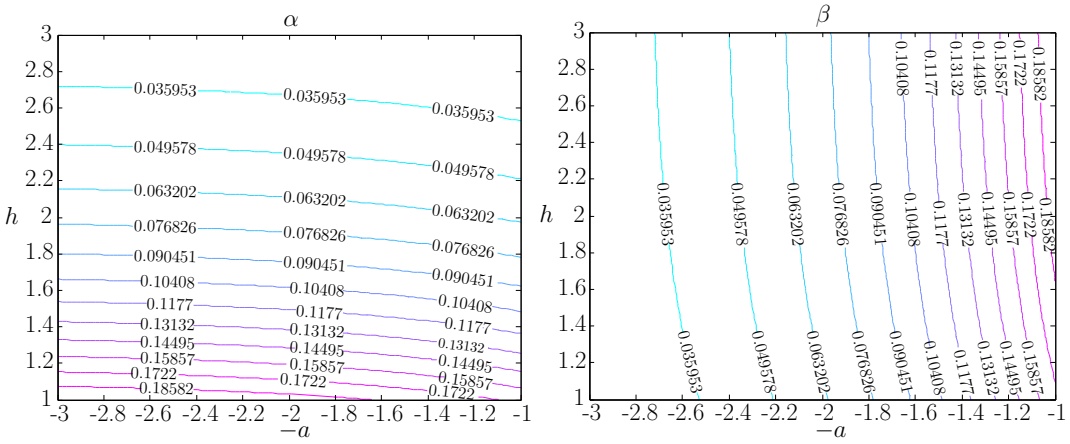
where A is as before, I is an $m \times 1$ unit vector and $\tilde{N}$ is the column vector $\tilde{N}^T = [\tilde{N}(y_1),\cdots,\tilde{N}(y_m)]$. As before, the column vector $\tilde{N}$ can be solved with $\tilde{N} = A^{-1} \cdot I$. The ASN function $\mathbb{E}_\theta[T]$ is obtained by substituting the column vector $\tilde{N}$ into Equation 3.29 and setting y to naught. The exact OC and ASN functions for the Gaussian example are presented in Figure 3.3 and Figure 3.4. Note that the curve obtained through simulation fits precisely on the exact theoretical curve.

The exact OC and ASN functions can be used to do a sweep of the exit boundaries. The type I and type II error, as well as the ASN of the SPRT algorithm, are presented in Figure 3.5 for the unit variance Gaussian example with exit boundaries in the range of $[1,3]$.

3.3.6 Example: Bernoulli random variable

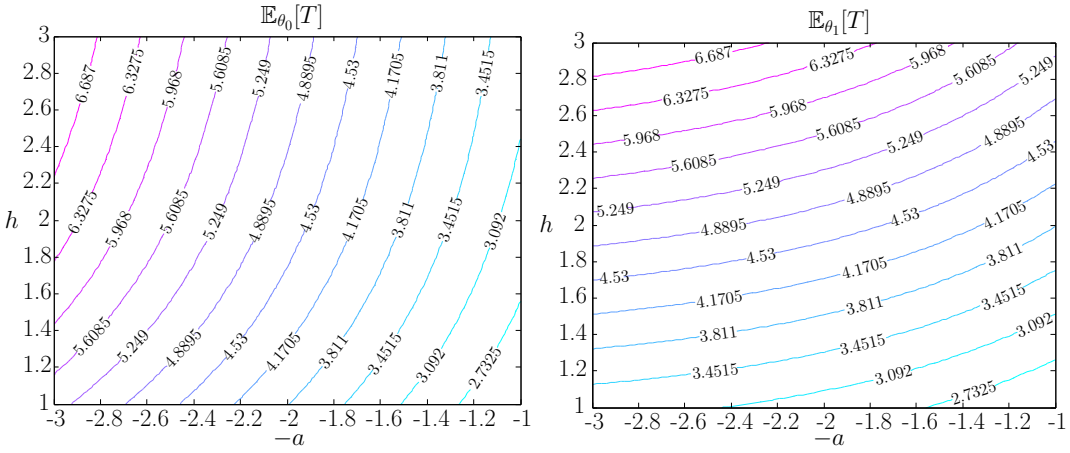
Consider the sequence $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ of i.i.d. real observations (adapted to the filtration $\mathscr{F}_k$) generated by a Bernoulli random variable with probability mass function

$$q_p(z_1) = \begin{cases} p & \text{if } z_1 = 1 \\ 1-p & \text{if } z_1 = 0. \end{cases}$$



(a) Type I error.

(b) Type II error.



(c) $\mathbb{E}_{\theta_0}[T]$.

(d) $\mathbb{E}_{\theta_1}[T]$.

Figure 3.5: The general performance of the unit variance Gaussian example with mean $\theta = 0$ and $\theta = 1$ under $\mathscr{H}_0$ and $\mathscr{H}_1$, respectively, and exit boundaries in the range of $[1,3]$.

The problem is to choose one of two fixed hypotheses so that the data sequence fits that hypothesis best. The following two hypotheses are under consideration:

$$\mathscr{H}_0 : p = p_0 = y = 0.4$$

versus

$$\mathscr{H}_1 : p = p_1 = 1 - y = 0.6.$$

For this example the increment sequence **s** becomes

$$s_k = \begin{cases} \ln \frac{1-y}{y} & \text{if } z_k = 1 \\ \ln \frac{y}{1-y} & \text{if } z_k = 0, \end{cases}$$

with probability mass function equal to

$$f_p(s_1) = \begin{cases} p & \text{if } s_1 = \ln\frac{1-y}{y} \\ 1-p & \text{if } s_1 = \ln\frac{y}{1-y}. \end{cases}$$

and is thus also an i.i.d. Bernoulli sequence. To make the meaning of the OC and ASN functions clearer, they will be calculated in the following sections by using Wald's approximation, which for the above problem also produces the exact solution. The series S_k can only increase or decrease by $\ln\frac{1-y}{y}$ or $-\ln\frac{1-y}{y}$ and as such if the exit thresholds are chosen as integer multiples of $\ln\frac{1-y}{y}$ there will be no overshoot. When there is no overshoot, Wald's approximations are exact as they are derived by ignoring the overshoot.

3.3.6.1 Wald's approximation

As stated before, the first step in calculating the OC and ASN functions is to determine $\omega_0(p)$ by using Equation 3.12. By applying Equation 3.12 the following is attained:

$$\begin{aligned} \mathbb{E}[e^{-\omega_0(p)s_1}] &= e^{-\omega_0(p)\cdot\ln\frac{1-y}{y}}\cdot p + e^{-\omega_0(p)\cdot\ln\frac{y}{1-y}}\cdot(1-p) \\ &= e^{-\mathscr{Y}}\cdot p + e^{\mathscr{Y}}\cdot(1-p) \\ &= 1. \end{aligned} \tag{3.30}$$

Equation 3.30 can be solved by using a simple substitution, namely $e^{\mathscr{Y}} = \mathscr{X}$, as is done below:

$$\begin{aligned} e^{2\mathscr{Y}}\cdot(1-p) - e^{\mathscr{Y}} + p &= 0 \\ \mathscr{X}^2\cdot(1-p) - \mathscr{X} + p &= 0. \end{aligned}$$

The usable value of $\mathscr{X}(p)$ is equal to

$$\mathscr{X}(p) = \begin{cases} \frac{1+\sqrt{1-4(1-p)\cdot p}}{2(1-p)} & \text{if } 0 < p \leq 0.5 \\ \frac{1-\sqrt{1-4(1-p)\cdot p}}{2(1-p)} & \text{if } 0.5 < p < 1 \end{cases} \tag{3.31}$$

and is a direct result of the quadratic formula. From Equation 3.31, $\omega_0(p)$ is obtained trivially, as $e^{\omega_0(p)\cdot\ln\frac{1-y}{y}}$ is equal to $\mathscr{X}(p)$ so that

$$\omega_0(p) = \frac{\ln\mathscr{X}(p)}{\ln\frac{1-y}{y}} \text{ if } 0 < p < 1$$

$$\omega_0(p) \approx 2.47\ln\mathscr{X}(p). \tag{3.32}$$

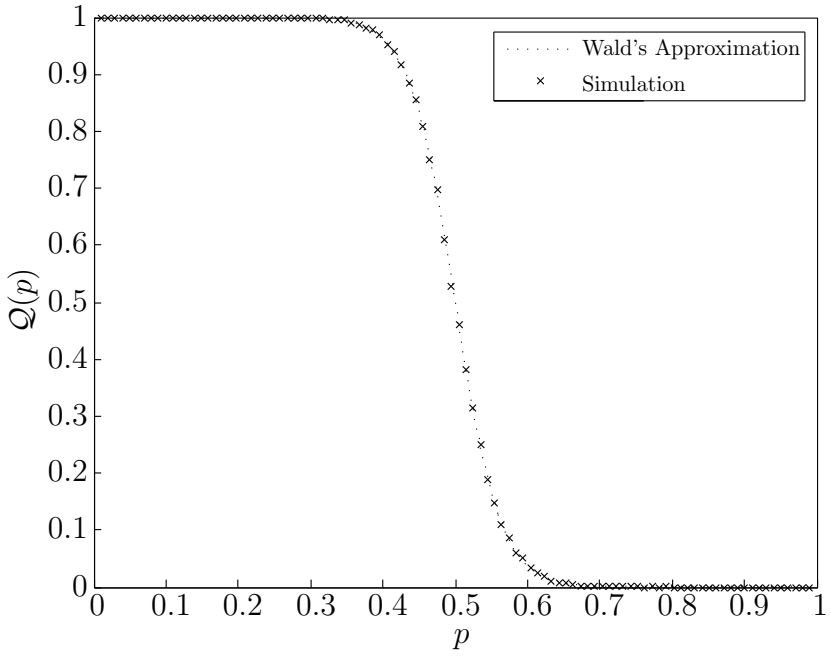


Figure 3.6: The exact OC function derived using Wald's approximation for the Bernoulli example with $p_0 = 0.4$ and $p_1 = 0.6$ under $\mathscr{H}_0$ and $\mathscr{H}_1$ respectively. The exit thresholds are equal to $8\ln\frac{0.6}{0.4}$ and $-8\ln\frac{0.6}{0.4}$.

Note that when $p \to 0$, $\omega_0(p) \to -\infty$ and when $p \to 1$, $\omega_0(p) \to \infty$.

The OC function is calculated by substituting Equation 3.32 into Equation 3.11 to obtain

$$\tilde{\mathscr{Q}}(p) = \begin{cases} \frac{e^{-2.47\ln\mathscr{X}h}-1}{e^{-2.47\ln\mathscr{X}h}-e^{2.47\ln\mathscr{X}a}} & \text{when } p \neq \frac{1}{2} \\ \frac{h}{h+a} & \text{when } p = \frac{1}{2}, \end{cases} \tag{3.33}$$

and is presented in Figure 3.6. Note that the p of $\mathscr{X}(p)$ is implied.

The values $\mathbb{E}_p[s_1]$ and $\mathbb{E}_p[s_1^2]$ need to be calculated before computing the ASN function. For this example, $\mathbb{E}_p[s_1] = (2p+1)\ln\frac{1-y}{y} \approx 0.41(2p+1)$ and $\mathbb{E}_p[s_1^2] = (\ln\frac{1-y}{y})^2 \approx 0.16$. The ASN function is calculated by substituting $\mathbb{E}_p[s_1]$, $\mathbb{E}_p[s_1^2]$ and Equation 3.33 into Equation 3.18 which gives

$$\tilde{\mathbb{E}}_p[T] = \begin{cases} \frac{1}{0.41(2p+1)}\left[\frac{1-e^{2.47\ln\mathscr{X}a}}{e^{-2.47\ln\mathscr{X}h}-e^{2.47\ln\mathscr{X}a}}h - \frac{e^{-2.47\ln\mathscr{X}h}-1}{e^{-2.47\ln\mathscr{X}h}-e^{2.47\ln\mathscr{X}a}}a\right] & \text{when } p \neq \frac{1}{2} \\ \frac{ah}{0.16} & \text{when } p = \frac{1}{2}. \end{cases}$$

The exact ASN function for the Bernoulli example can be found in Figure 3.7.

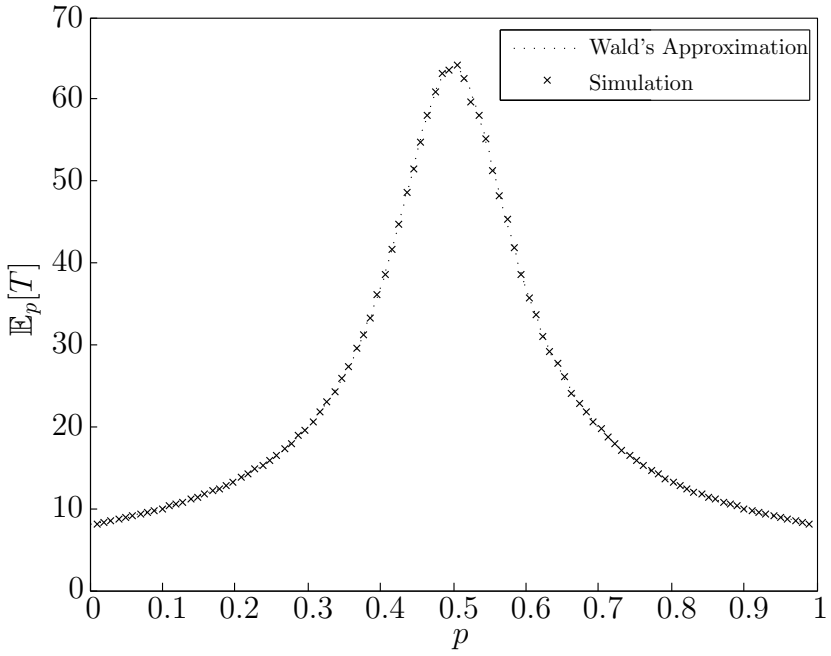


Figure 3.7: The exact ASN function function derived using Wald's approximation for the Bernoulli example with $p_0 = 0.4$ and $p_1 = 0.6$ under $\mathscr{H}_0$ and $\mathscr{H}_1$ respectively. The exit thresholds are equal to $8\ln\frac{0.6}{0.4}$ and $-8\ln\frac{0.6}{0.4}$.

3.4 HYPOTHESIS TESTING: BAYESIAN FORMULATION

The following section closely follows the notation of [48]. Once again, consider the sequence $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ of i.i.d. real observations (adapted to the filtration $\mathscr{F}_k$) following one of two hypotheses:

$$\mathscr{H}_0 : z_k \sim Q_0,\ k = 1,2,\cdots$$

versus

$$\mathscr{H}_1 : z_k \sim Q_1,\ k = 1,2,\cdots;$$

where Q_0 and Q_1 are two probability distributions with associated densities q_0 and q_1, respectively. Further assume that hypothesis $\mathscr{H}_1$ occurs with prior probability π and hypothesis $\mathscr{H}_0$ occurs with prior probability $1-\pi$. Instead of asking what the exit boundaries should be to obtain a *certain probability of error*, as is done in the case of Wald, the problem could be restated in terms of different *costs*, that should be minimised concurrently. Naturally three different costs are important, namely the cost incurred by observing an observation $c \geq 0$, the cost of making a type I error $c_0 > 0$ and the cost of making a type II error $c_1 > 0$. The power of this approach is flexibility, as the objective is not

to have the lowest probability of error but rather to minimise a cost function, which takes into account c, c_0 and c_1. This alternative problem formulation is known as the *Bayesian sequential detection problem*.

As already stated, any sequential test consists of a sequential decision rule (T, δ), where T is a stopping time and δ is a decision function that can be evaluated after each observation. From the sequential decision rule it follows that the *average cost of error* can be expressed as

$$\begin{aligned} c_e(T,\delta) &= (1-\pi)c_0 P_0(\delta_T = 1) + \pi c_1 P_1(\delta_T = 0) \\ &= (1-\pi)c_0\alpha + \pi c_1\beta. \end{aligned} \tag{3.34}$$

Complementary to the average cost of error is the *cost of sampling*, which can be expressed as

$$c\mathbb{E}_\pi[T] = c\cdot[(1-\pi)\mathbb{E}_0[T] + \pi\mathbb{E}_1[T]], \tag{3.35}$$

where $\mathbb{E}_\pi[\cdot]$ denotes expectation under the probability measure $P_\pi = (1-\pi)P_0 + \pi P_1$. The average cost of error reduces to the *average probability of error* P_e when $c_0 = c_1 = 1$. When $c = 1$ the average cost of sampling reduces to the Average Run Length (ARL). Normally the ARL is associated with either hypothesis $\mathscr{H}_0$ or $\mathscr{H}_1$ [6], but here it refers to the general expected run length of the experiment and could therefore be seen as a misuse of terminology. To avoid ambiguity the term ASN (which is closely related to the ARL) is only used when working in Wald's framework.

The *total cost* incurred by (or *Bayes risk* of) any sequential decision rule is thus equal to the sum of the average cost of error and the cost of sampling and is expressed mathematically as

$$c_e(T,\delta) + c\mathbb{E}_\pi[T].$$

Naturally, the best sequential decision rule would be the rule that minimises the total cost, which can be stated as

$$g(\pi) = \inf_{T\in\mathscr{T},\delta\in\mathscr{D}} \left[c_e(T,\delta) + c\mathbb{E}_\pi[T]\right], \tag{3.36}$$

where $g(\pi)$ is known as the *minimal expected cost function*, and $\mathscr{T}$ and $\mathscr{D}$ are the set of all valid stopping times and decision rules, respectively. Through simple mathematical manipulation Equation 3.36 can be reformulated as

$$g(\pi) = \inf_{T\in\mathscr{T}} \mathbb{E}_\pi\left[\min\{c_1\pi_T^\pi, c_0(1-\pi_T^\pi)\} + cT\right], \tag{3.37}$$

where π_k^π is the posterior probability that $\mathscr{H}_1$ is true, given all the information up to observation k, and is expressed as

$$\begin{aligned}\pi_k^\pi &= \frac{\pi\prod_{i=1}^k q_1(z_i)}{\pi\prod_{i=1}^k q_1(z_i)+(1-\pi)\prod_{i=1}^k q_0(z_i)}\\ &= \frac{\pi_{k-1}^\pi q_1(z_k)}{\pi_{k-1}^\pi q_1(z_k)+(1-\pi_{k-1}^\pi)q_0(z_k)},\end{aligned}$$

with $\pi_0^\pi = \pi$. The optimal sequential decision rule satisfying Equation 3.36 or Equation 3.37 is given by the following theorem [23, 48]:

Theorem 3 (Optimal i.i.d. sequential decision rule) *Consider the optimisation problem of Equation 3.36 or Equation 3.37. The optimal solution is given by the sequential decision rule (T,δ) with*

$$T = \inf\{k \geq 0 | \pi_k^\pi \notin (\pi_L, \pi_U)\} \tag{3.38}$$

and

$$\delta_k = \begin{cases} 0 \text{ if } \pi_k^\pi \leq c_0/(c_0+c_1) \\ 1 \text{ if } \pi_k^\pi > c_0/(c_0+c_1), \end{cases}$$

where the exit thresholds π_L and π_U are given by

$$\pi_L = \sup\{0 \leq \pi \leq 1 | g(\pi) = c_1\pi\} \tag{3.39}$$

and

$$\pi_U = \inf\{0 \leq \pi \leq 1 | g(\pi) = c_0(1-\pi)\} \tag{3.40}$$

respectively. That is, the optimal sequential decision rule continues sampling until $\pi_k^\pi \notin (\pi_L, \pi_U)$, at which time it chooses hypothesis $\mathscr{H}_1$ if $\pi_k^\pi \geq \pi_U$ and $\mathscr{H}_0$ otherwise.

The minimal cost function $g(\pi) = \inf_{T\in\mathscr{T}} \mathbb{E}_\pi\{h(\pi_T^\pi) + cT\}$, where $h(\pi) = \min\{c_1\pi, c_0(1-\pi)\}$, can be calculated easily, since $g(\pi)$ is the monotone point-wise limit from above of the sequence of functions

$$g_k(\pi) = \min\{h(\pi), \mathscr{R}g_{k-1}(\pi) + c\},\ k = 1, 2, \cdots \tag{3.41}$$

with $g_0(\pi) = h(\pi)$, and where the operator $\mathscr{R}$ is defined by

$$\begin{aligned}\mathscr{R}r(\pi) &= \mathbb{E}_\pi[r(\pi_1^\pi)]\\ &= \int_{-\infty}^{\infty} r\left(\frac{\pi q_1(z_1)}{\pi q_1(z_1)+(1-\pi)q_0(z_1)}\right)\cdot\left[\pi q_1(z_1)+(1-\pi)q_0(z_1)\right]dz_1,\end{aligned}$$

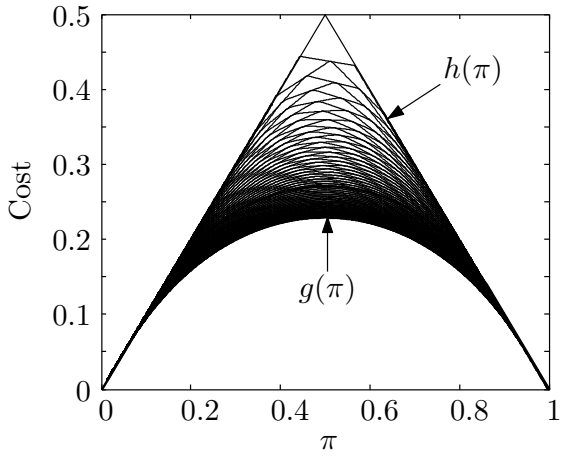


Figure 3.8: The limiting procedure used to calculate $g(\pi)$, in the case where the observations generated by $\mathscr{H}_0$ and $\mathscr{H}_1$ are different Bernoulli random variables.

such that

$$\mathscr{R}g_{k-1}(\pi) = \mathbb{E}_\pi[g_{k-1}(\pi_1^\pi)].$$

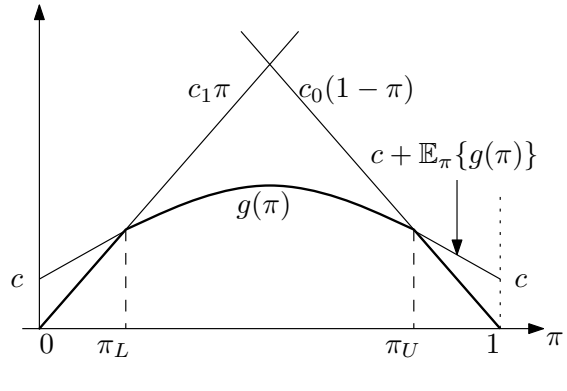
After $g(\pi)$ has been computed, the exit boundaries π_L and π_U are respectively calculated with Equation 3.39 and Equation 3.40. The limiting procedure used to calculate $g(\pi)$ is illustrated in Figure 3.8 [23, 66].

3.4.1 On the structure of the minimal cost function

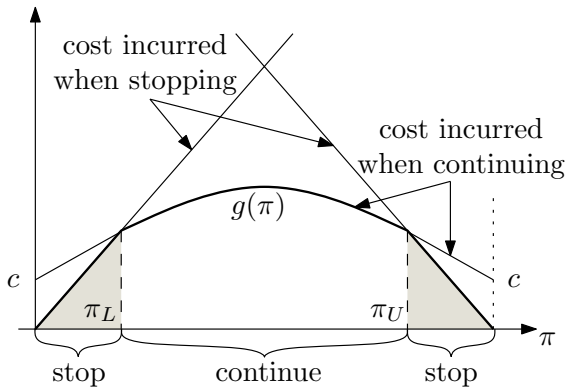
As shown in [48], the minimal cost function $g(\pi)$ is concave, and is bounded by $0 \leq g(\pi) \leq h(\pi)$, where $h(\pi) = \min\{c_1\pi, c_0(1-\pi)\}$, as mentioned in Section 3.4. Furthermore, $g(0) = g(1) = 0$. Interestingly enough, the prior probability that $\mathscr{H}_1$ is true (i.e., π) is not used to determine the minimal cost function. That is, the same $g(\pi)$ is used for any $\pi \in [0,1]$.

A classic minimal cost function is shown in Figure 3.9a, which is symmetric about the line $\pi = 1/2$, since $c_0 = c_1$.

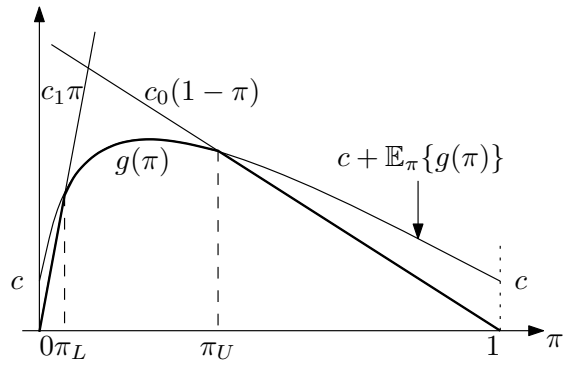
Figure 3.9b indicates that $g(\pi)$ can be divided into a continue sampling region and a stop sampling region. More specifically, $g(\pi)$ represents the minimum between the cost incurred when continuing to sample (corresponding to $c + \mathbb{E}_\pi\{g(\pi)\}$ in Figure 3.9a) and the cost incurred when terminating the experiment. The same applies to Figure 3.9c and Figure 3.9d, where the only difference is that the costs of errors (c_0 and c_1) are no longer equal.



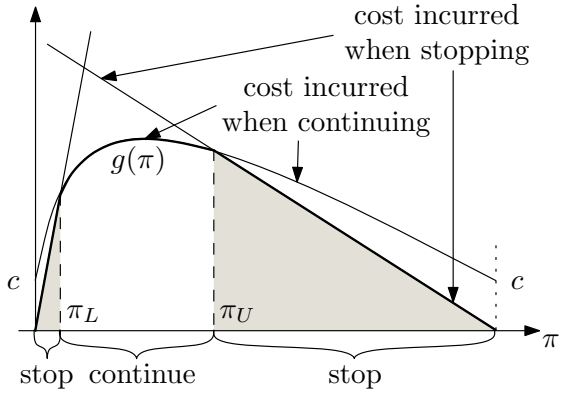
(a) Structure of the cost function, $g(\pi)$, $c_0 = c_1$.



(b) Decision regions of the cost function, $c_0 = c_1$.



(c) Structure of the cost function, $g(\pi)$, $c_0 \neq c_1$.



(d) Decision regions of the cost function, $c_0 \neq c_1$.

Figure 3.9: Typical structure and behaviour of the minimal cost function, $g(\pi)$ (from [23]).

3.4.2 Bayesian versus Wald's formulation

The boundaries π_{L} and π_U can be converted to Wald's exit boundaries with

$$A = \frac{1-\pi}{\pi}\frac{\pi_{\mathrm{L}}}{1-\pi_{\mathrm{L}}} \Longleftrightarrow \pi_{\mathrm{L}} = \frac{\pi A}{1-\pi(1-A)}, \tag{3.42}$$

and

$$B = \frac{1-\pi}{\pi}\frac{\pi_U}{1-\pi_U} \Longleftrightarrow \pi_U = \frac{\pi B}{1-\pi(1-B)}, \tag{3.43}$$

implying that Wald's SPRT stopping time (Equation 3.4) is nothing more than the Bayesian optimal stopping time (Equation 3.38) [23]. The relationship that exists between the Bayesian formulation and Wald's approach makes it possible to express the approximate type I and type II errors as a function of π_L and π_U. The approximate type I error is equal to

$$\tilde{\alpha} = \frac{1-A}{B-A} = \frac{\pi_L - \pi}{\pi - 1}\cdot\frac{\pi_{\mathrm{U-1}}}{\pi_L - \pi_U}, \tag{3.44}$$

while the approximate type II error becomes

$$\tilde{\beta} = A\frac{B-1}{B-A} = \frac{\pi_L}{\pi}\cdot\frac{\pi-\pi_U}{\pi_L-\pi_U}. \tag{3.45}$$

3.4.3 Example

Let $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ be an i.i.d. sequence of real observation (adapted to the filtration $\mathscr{F}_k$) following one of two equiprobable ($\pi = 0.5$) hypotheses:

$$\mathscr{H}_0 : z_k \sim Q_0,\ k = 1,2,\cdots,n$$

versus

$$\mathscr{H}_1 : z_k \sim Q_1,\ k = 1,2,\cdots,n$$

where Q_0 and Q_1 are two probability distributions with associated probability mass functions q_0 and q_1, respectively. The probability mass functions q_0 and q_1 are equal to

$$q_0(z_1) = \begin{cases} 0.4 \text{ if } z_1 = 1 \\ 0.6 \text{ if } z_1 = 0, \end{cases}$$

and

$$q_1(z_1) = \begin{cases} 0.6 \text{ if } z_1 = 1 \\ 0.4 \text{ if } z_1 = 0. \end{cases}$$

For this example $c_0 = 1$ and $c_1 = 1$ and $c \in [0, 0.05]$. When working in the Bayesian framework, the instinctive question arises, what should the values of c_0, c_1 and c be to obtain a certain type I and type II error? The solution to this problem turns out to be quite difficult, as there is no direct link between the costs and the probability of error. Without this link the choice of c_0, c_1 and c is quite arbitrary and of no real practical value. To find this link for the current problem, c will be traversed to determine the effect of c on α and β, while keeping c_1 and c_2 constant. See [23] for a greater variety of examples with different initial conditions. In particular, [23] investigates the case when the hypotheses are not equiprobable, as well as the case when $c_0 \neq c_1$. The focus here is however to provide an extensive example that would enable the reader to link the costs to the probability of error, for an arbitrary choice of π, c_0, c_1 and c. The exit boundaries π_U and π_L are displayed in Figure 3.10 as a function of c for the above-mentioned example. The probability of error P_e (Equation 3.34) and the ARL (Equation 3.35) is displayed in Figure 3.11. The step-like nature of the P_e and the ARL is due to the fact that for the example the exit boundaries can only be discrete functions (limited

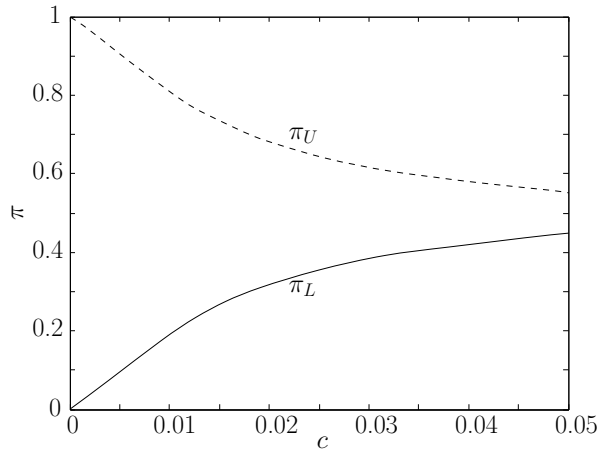
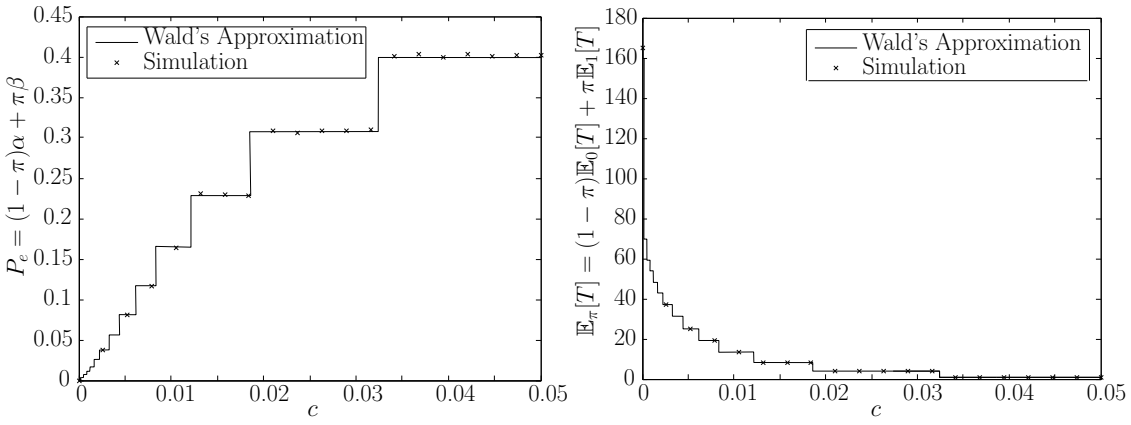


Figure 3.10: The value for π_L and π_U as a function of c with $c_0 = c_1 = 1$ for the Bernoulli example.

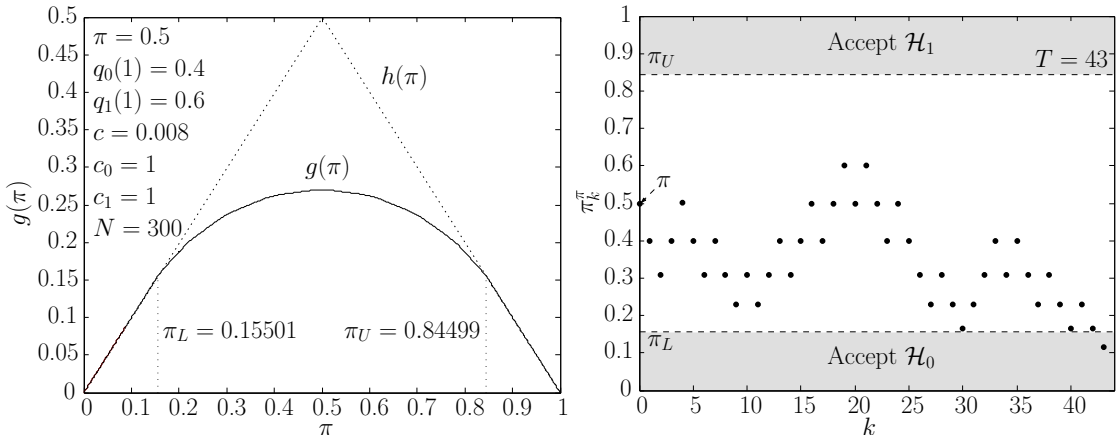


(a) Average probability of error, P_e.

(b) ARL, $\mathbb{E}_\pi[T]$.

Figure 3.11: The P_e and $\mathbb{E}_\pi[T]$ as a function of c for the Bernoulli example.

number of values), which implies that there is no overshoot, causing Wald's approximation to be exact (see Section 3.3.6 for more details). The fact that there is no overshoot should actually be taken into account when computing π_L and π_U, but is not done here for the sake of simplicity and to be compatible with [23]. The value of c is now restricted to 0.008 in order to show the reader how to obtain the curves in Figure 3.10 and Figure 3.11 (where c was traversed). When the value of c is fixed, the values for π_L and π_U are calculated by first determining $g(\pi)$ with Equation 3.41 by letting $k \to \infty$ (in practice 300 iterations were used) and then applying Equation 3.39 and Equation 3.40. The calculated function $g(\pi)$ and thresholds $\pi_L = 0.15501$ and $\pi_U = 0.84499$ for $c_0 = 1, c_1 = 1$ and $c = 0.008$ can be found in Figure 3.12a. The values of π_L and π_U can be converted to A and B with Equation 3.42 and Equation 3.43, which gives 0.1834 and 5.4512, respectively. The exit boundaries A and B need to be converted to $\bar{A}$ and $\bar{B}$, as there is no overshoot for this problem. The first step in



(a) The minimal cost function, $g(\pi)$.

(b) The sequence $\boldsymbol{\pi}^{\pi}$.

Figure 3.12: The minimal cost function and an example posterior sequence for the Bernoulli example.

calculating $\bar{A}$ and $\bar{B}$ is to calculate the constant integer k via

$$k = \left\lceil \frac{\ln B}{\ln \frac{q_1(1)}{q_0(1)}} \right\rceil .$$

Only B is used, as using A in a similar fashion would produce the same integer k. Now

$$\begin{aligned} \bar{A} &= e^{-k*\ln \frac{q_1(1)}{q_0(1)}}, \\ &= \left(\frac{q_1(1)}{q_0(1)}\right)^{-k}, \end{aligned}$$

and

$$\begin{aligned} \bar{B} &= e^{k*\ln \frac{q_1(1)}{q_0(1)}}, \\ &= \left(\frac{q_1(1)}{q_0(1)}\right)^{k}. \end{aligned}$$

Substituting $\bar{A}$ and $\bar{B}$ into Equation 3.44 and Equation 3.45 yields the correct value for $\alpha = 0.1164$ and $\beta = 0.1164$ as there is no overshoot. The average probability of error can now finally be calculated with $P_e = (1-\pi)\alpha + (\pi)\beta$, which is equal to 0.1164. Next compute the averages $\mathbb{E}_0[T]$ and $\mathbb{E}_1[T]$ with Equation 3.22 and Equation 3.23, where $\tilde{\alpha} = \alpha$ and $\tilde{\beta} = \alpha$ is already known and

$$\mathbb{E}_0[s_1] = q_0(0) \times \ln\left(\frac{q_1(0)}{q_0(0)}\right) + q_0(1) \times \ln\left(\frac{q_1(1)}{q_0(1)}\right),$$

$$\mathbb{E}_1[s_1] = q_1(0) \times \ln\left(\frac{q_1(0)}{q_0(0)}\right) + q_1(1) \times \ln\left(\frac{q_1(1)}{q_0(1)}\right).$$

With $\mathbb{E}_0[T]$ and $\mathbb{E}_1[T]$ the ARL is easily calculated and is equal to 14.4290. An example sequence $\boldsymbol{\pi}^{\pi}$ for the current problem under $\mathscr{H}_0$ is displayed in Figure 3.12b.

3.5 BAYESIAN QUICKEST DETECTION

The following section closely follows the notation of [48]. In Section 3.3 and Section 3.4 the focus was on sequential detection or rather classification of an observed sequence with no fixed sample size. A more general problem will be studied in the remainder of the chapter. In the more general case the observed sequence is allowed to switch from one hypothesis to another, and the aim is to detect this change as quickly as possible, while simultaneously minimising the probability of a false alarm. This section is called *Bayesian quickest detection*, since the distribution of the change point is known beforehand. In Section 3.6 the case where the change point distribution is unknown is investigated. The Bayesian quickest detection problem is also known as *Shiryaev's disruption* problem, since Shiryaev solved it [67].

Shiryaev's disruption problem is now introduced formally. Consider the sequence $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ of i.i.d. real observations with a random change point τ. Further assume that conditioned on τ, $\mathbf{z}$ is an independent sequence where $\mathbf{z}^{-\tau} = \{z_1, z_2, \cdots, z_{\tau-1}\}$, is i.i.d. with marginal distribution Q_0, and $\mathbf{z}^{+\tau} = \{z_\tau, z_{\tau+1}, \cdots\}$ is also i.i.d. with marginal distribution Q_1. The associated densities of Q_0 and Q_1 are q_0 and q_1, respectively. A probability distribution P_π is considered that describes both the (prior) distribution of τ and the distribution of $\mathbf{z}$ induced by this prior and above conditional behaviour. Moreover, the observations $\{z_k\}_{\{k=1,2,\cdots\}}$ generate the filtration $\mathscr{F}_k$, with

$$\mathscr{F}_k = \sigma(\{z_k\}_{\{k=1,2,\cdots\}}, \{\tau = 0\}),\ k = 1,2,\cdots$$

and $\mathscr{F}_0$ contains not only Ω (the sample space) but also the set $\{\tau = 0\}$. The case where τ is geometrically distributed will be considered, and consequently,

$$P_\pi\{\tau = k\} = \begin{cases} \pi & \text{if } k = 0 \\ (1-\pi)(1-\rho)^{k-1}\rho & \text{if } k = 1,2,\cdots. \end{cases}$$

Let $T \in \mathscr{T}$ be a stopping time and let $\mathscr{T}$ be the set consisting of all valid stopping times, then T is actually the time at which the alarm is sounded to signal that a change in distribution has occurred. The optimal choice of T is the T that minimises jointly the *probability of a false alarm*

$$P_\pi\{T < \tau\} \tag{3.46}$$

and the *expected delay*

$$\mathbb{E}_\pi[(T-\tau)^+] = \mathbb{E}_\pi[\max\{T-\tau, 0\}], \tag{3.47}$$

where $\mathbb{E}_\pi$ denotes expectation under the probability measure P_π.

A convenient way of implementing a joint minimisation between Equation 3.46 and Equation 3.47 is to seek $T \in \mathscr{T}$ to solve the optimisation problem

$$g(\pi) = \inf_{T \in \mathscr{T}} \left[P_\pi\{T < \tau\} + c \cdot \mathbb{E}_\pi[(T-\tau)^+] \right], \tag{3.48}$$

where $c > 0$ is a constant controlling the relative importance of the two performance indices and $g(\pi)$ is known as the *minimal expected cost*, or simply the *minimal cost function*. Through simple mathematical manipulation Equation 3.48 can be reformulated as

$$g(\pi) = \mathbb{E}_\pi \left[1 - \pi_T^\pi + c \cdot \sum_{k=0}^{T-1} \pi_k^\pi \right], \tag{3.49}$$

where π_k^π is the posterior probability that a change did occur before or at k given all the observations up to k and is expressed as

$$\pi_k^\pi = \frac{[\pi_{k-1}^\pi + (1-\pi_{k-1}^\pi)\rho] q_1(z_k)}{[\pi_{k-1}^\pi + (1-\pi_{k-1}^\pi)\rho] q_1(z_k) + [(1-\pi_{k-1}^\pi)(1-\rho)] q_0(z_k)}, \tag{3.50}$$

with $\pi_0^\pi = \pi$.

The optimal stopping time satisfying Equation 3.48 or Equation 3.49 is given by the following theorem [48]:

Theorem 4 (Bayes optimal stopping time) *Consider the optimisation problem of Equation 3.48 or Equation 3.49. The optimal solution is given by*

$$T = \inf\{k \geq 0 | \pi_k^\pi \geq \pi^*\}$$

where the exit boundary π^* is given by

$$\pi^* = \inf\{0 \leq \pi \leq 1 | g(\pi) = 1 - \pi\}. \tag{3.51}$$

That is continue sampling until $\pi_k^\pi \geq \pi^*$, at which time a change is declared.

The minimal cost function $g(\pi) = \inf_{T \in \mathscr{T}} \mathbb{E}_\pi\{h(\pi_T^\pi) + c \cdot \sum_{k=0}^{T-1} \pi_k^\pi\}$, where $h(\pi) = 1 - \pi$, can be calculated easily, since $g(\pi)$ is the monotone point-wise limit from above of the sequence of functions

$$g_k(\pi) = \min\{h(\pi), \mathscr{R} g_{k-1}(\pi) + c\pi\}, \; k = 1, 2, \cdots$$

with $g_0(\pi) = h(\pi)$, and where the operator $\mathscr{R}$ is defined by

$$\begin{aligned}\mathscr{R}r(\pi) &= \mathbb{E}_\pi[r(\pi_1^\pi)] \\ &= \int_{-\infty}^{\infty} r(\pi_1^\pi) \cdot \big[[\pi + (1-\pi)\rho]q_1(z_1) + (1-\pi)(1-\rho)q_0(z_1)\big]dz_1,\end{aligned}$$

with

$$\pi_1^\pi = \frac{[\pi + (1-\pi)\rho]q_1(z_1)}{[\pi + (1-\pi)\rho]q_1(z_k) + [(1-\pi)(1-\rho)]q_0(z_1)},$$

such that

$$\mathscr{R}g_{k-1}(\pi) = \mathbb{E}_\pi[g_{k-1}(\pi_1^\pi)].$$

Once $g(\pi)$ is known, the exit boundary π^* is calculated with Equation 3.51 [66].

3.6 NON-BAYESIAN QUICKEST DETECTION

This section closely follows the notation from [48, 59]. In this section the quickest detection algorithms have no prior change point distribution. Two measures of *detection delay* are investigated, namely *Lorden's performance measure* [52] and *Pollak's performance measure* [56].

3.6.1 Lorden's performance measure

Consider a measurable space $(\Omega, \mathscr{F})$, consisting of a sample space Ω and a σ-field $\mathscr{F}$ of events [48]. Further consider a family $\{P_\tau | \tau \in [1, 2, \cdots, \infty]\}$ of probability measures on $(\Omega, \mathscr{F})$ and a random sequence $\mathbf{z} = \{z_k; k = 1, 2, \cdots \infty\}$, such that, under P_τ, $\mathbf{z}^{-\tau} = \{z_1, z_2, \cdots, z_{\tau-1}\}$ are independent and identically distributed (i.i.d) with a fixed marginal distribution Q_0 and $\mathbf{z}^{+\tau} = \{z_\tau, z_{\tau+1}, \cdots, \infty\}$ are i.i.d with marginal distribution Q_1 and are independent of $\mathbf{z}^{-\tau}$. The probability densities associated with Q_0 and Q_1 are q_0 and q_1 respectively. A procedure is desired that can detect a change in the underlying distribution of $\mathbf{z}$ (when $\mathbf{z}$ is sampled from Q_1 instead of Q_0), if it occurs (i.e. if $\tau < \infty$), as quickly as possible after it occurs. As a set of detection strategies, it is natural to consider the set $\mathscr{T}$ of all (extended) stopping times with respect to the filtration $\{\mathscr{F}_k\}$ where $\mathscr{F}_k$ denotes the smallest σ-field with respect to which $z_0, z_1, \cdots, z_k$ are measurable. Thus, when the stopping time T takes on the value k, the interpretation is that T has detected the existence of a change point τ at or prior to time k. It is of interest to penalise *expected delay via its worst case* value (also known as *Lorden's performance measure*)

$$d_l(T) = \sup_{\tau \geq 1} \operatorname{ess\,sup} \mathbb{E}_\tau\{(T - \tau + 1)^+ | \mathscr{F}_{\tau-1}\}, \tag{3.52}$$

where $\mathbb{E}_\tau\{\cdot\}$ denotes expectation under the distribution P_τ and $(T-\tau+1)^+ = \max\{T-\tau+1, 0\}$. Note that ess sup $\mathbb{E}_\tau\{(T-\tau+1)^+|\mathscr{F}_{\tau-1}\}$ is the worst case average delay under P_τ, where the worst case is taken over all realization of $\mathbf{z}^{-\tau}$. In other words, it is the same as measuring the average detection delay when the first sample already belongs to the changed distribution. The desire to make $d_l(T)$ small must be balanced with a constraint on the false alarm rate. The fact that false alarms will occur is accepted, however the rate at which they occur is fixed. The false alarm rate is quantified by the *mean time between false alarms*

$$f(T) = \mathbb{E}_\infty\{T\}. \tag{3.53}$$

A useful design criterion is then given by

$$\inf_{T\in\mathscr{T}} d_l(T) \text{ subject to } f(T) \geq \lambda, \tag{3.54}$$

where λ is a positive, finite constant. A stopping time is desired that minimises the worst case expected delay within a lower-bound constraint on the mean time between false alarms. A possible stopping time that meets the requirements of Equation 3.54 is Page's CUSUM stopping time [6]. In particular, for $h \geq 0$ the CUSUM stopping time is defined as

$$T_h^{\text{CUSUM}} = \inf\{k \geq 0 | g_k \geq h\},$$

where

$$g_k = \begin{cases} (g_{k-1} + s_k)^+ & \text{if } k > 0 \\ y \in \mathbb{R}^+ & \text{if } k = 0, \end{cases} \tag{3.55}$$

and

$$s_k = \ln \frac{q_1(z_k)}{q_0(z_k)}. \tag{3.56}$$

Under normal CUSUM operating conditions y is set to 0. As it turns out T_h^{CUSUM} is the optimal choice, as indicated by the theorem below [48, 53, 54]:

Theorem 5 (Optimality of CUSUM) *Choose $h \geq 0$. Then, the stopping time T_h^{CUSUM} solves Equation 3.54 with $\lambda = f(T_h^{CUSUM})$. That is,*

$$f(T) \geq f(T_h^{CUSUM}) \implies d_l(T) \geq d_l(T_h^{CUSUM}).$$

3.6.1.1 The ARL function of CUSUM

As explained in Section 3.3.1, the parameter θ determines the distribution of $\mathbf{z}$ and when $\theta = \theta_0$ density q_0 is obeyed (no change) and when $\theta = \theta_1$ density q_1 is obeyed (change occurred). The

average run length function $\mathscr{L}(\theta)$ is the expected number of samples required for an algorithm (for example CUSUM) to terminate as a function of θ when the exit threshold(s) (for example h) is/are fixed. It turns out that in the case of CUSUM, when $\theta = \theta_0$ then $\mathscr{L}(\theta_0) = f(T_h^{\text{CUSUM}})$ and when $\theta = \theta_1$ then $\mathscr{L}(\theta_1) = d_l(T_h^{\text{CUSUM}})$. The function $\mathscr{L}(\theta)$ can be calculated with [59]

$$\mathscr{L}(\theta) = \frac{N_\theta(0)}{1 - P_\theta(0)}, \tag{3.57}$$

where $N_\theta(0) = \mathbb{E}_\theta[T|0]$ and $P_\theta(0) = P_\theta(-a|0)$ were defined in Section 3.3.3. Equation 3.57 is only valid when $-a = 0$. The exact value of $\mathscr{L}(\theta)$ can thus be calculated by solving Equation 3.24 and Equation 3.25 with $-a = 0$. *Wald's approximation* of $\tilde{\mathscr{L}}(\theta)$ can be derived by evaluating the following limit

$$\tilde{\mathscr{L}}(\theta) = \lim_{a \to 0} \frac{\tilde{E}_\theta[T|0]}{1 - \tilde{P}_\theta(-a|0)}, \tag{3.58}$$

where $\tilde{P}_\theta$ is Wald's approximated OC function and $\tilde{E}_\theta$ is Wald's approximated ASN function with SPRT exit boundaries $-a$ and h. After evaluating the limit, Equation 3.58 becomes [59]

$$\tilde{\mathscr{L}}(\theta) = \begin{cases} \frac{1}{\mathbb{E}_\theta[s_k]}\left(h + \frac{e^{-\omega_0(\theta)h}}{\omega_0(\theta)} - \frac{1}{\omega_0(\theta)}\right) & \text{if } \mathbb{E}_\theta[s_k] \neq 0 \\ \frac{h^2}{\mathbb{E}_\theta[s_k^2]} & \text{if } \mathbb{E}_\theta[s_k] = 0. \end{cases} \tag{3.59}$$

However *Siegmund's approximation* is much better than Wald's approximation, as Siegmund incorporates an approximation of the overshoot. Siegmund's approximation is equal to [59]

$$\hat{\mathscr{L}}(\theta) = \begin{cases} \frac{1}{\mathbb{E}_\theta[s_k]}\left(h + \delta^+ + \delta^- + \frac{e^{-\omega_0(\theta)(h+\delta^+ + \delta^-)}}{\omega_0(\theta)} - \frac{1}{\omega_0(\theta)}\right) & \text{if } \mathbb{E}_\theta[s_k] \neq 0 \\ \frac{(h+\delta^+ + \delta^-)^2}{\mathbb{E}_\theta[s_k^2]} & \text{if } \mathbb{E}_\theta[s_k] = 0, \end{cases} \tag{3.60}$$

where

$$\delta^+ \approx \mathbb{E}_\theta[S_T - h | S_T - h \geq 0],$$

$$\delta^- \approx \mathbb{E}_\theta[S_T | S_T \leq 0].$$

3.6.1.2 Example: Gaussian random variable

Suppose there is an observed sequence $\mathbf{z}$, such that z_k is drawn from density $q_0 \sim \mathcal{N}(0,1)$ before change point τ. From time point τ, z_k is drawn from density $q_1 \sim \mathcal{N}(1,1)$. Assuming $\mathbf{z}$, it follows that $\mathbf{s}$ is also i.i.d and is characterised by density $f_0 \sim \mathcal{N}\left(-\frac{1}{2}, 1\right)$ before the change and $f_1 \sim \mathcal{N}\left(\frac{1}{2}, 1\right)$ after the change occurred. In general $\mathbf{s}$ is characterised by $f_\theta \sim \mathcal{N}(\theta - \frac{1}{2}, 1)$ (see Section 3.3.5). The CUSUM sequence $\mathbf{g}$ is derived from $\mathbf{s}$. As soon as $\mathbf{g}$ crosses h a change can be declared. An example

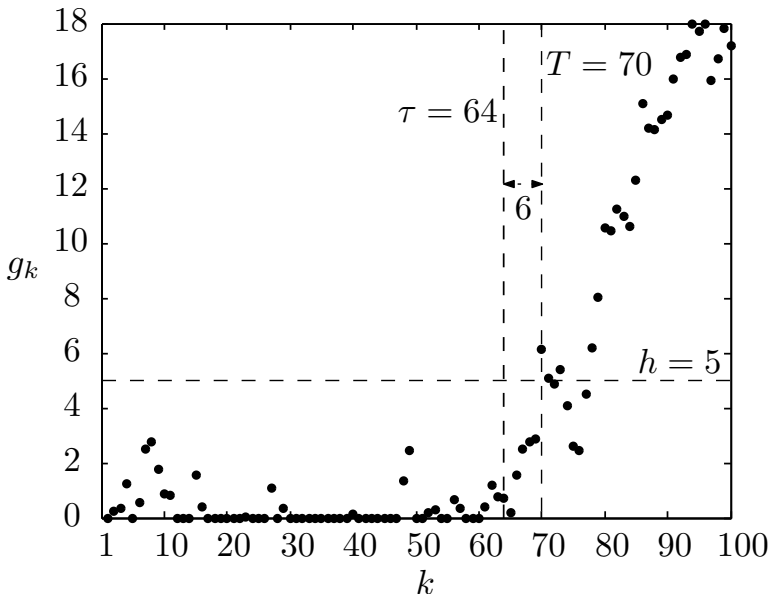


Figure 3.13: An example CUSUM sequence $\mathbf{g}$ with $q_0 \sim \mathcal{N}(0,1)$, $q_1 \sim \mathcal{N}(1,1)$, $h = 5$ and change point $\tau = 64$.

of $\mathbf{g}$ can be found in Figure 3.13. The measures defined in Equation 3.52 and Equation 3.53 can be calculated for every h by respectively setting θ equal to either 0 or 1 in Equation 3.57. The exact values of $\mathscr{L}(0)$ and $\mathscr{L}(1)$ are calculated by using the same approach as discussed in Section 3.3.5.2. The exact values of Equation 3.52 and Equation 3.53 are displayed in Figure 3.14 for $h \in [1,5]$.

Furthermore, $\mathscr{L}(\theta)$ can be calculated by traversing θ in Equation 3.57 and fixing h. By substituting $\mathbb{E}_\theta[s_1], \mathbb{E}_\theta[s_1^2]$ and $\omega_0(\theta)$ (calculated in Section 3.3.5.1) into Equation 3.59 Wald's approximation of $\mathscr{L}(\theta)$ is obtained, which is equal to

$$\tilde{\mathscr{L}}(\theta) = \begin{cases} \frac{e^{-2(\theta-\frac{1}{2})h} - 1 + 2(\theta-\frac{1}{2})h}{2(\theta-\frac{1}{2})^2} & \text{if } \theta \neq \frac{1}{2} \\ h^2 & \text{if } \theta = \frac{1}{2}. \end{cases}$$

Siegmund's approximation is obtained by substituting $\mathbb{E}_\theta[s_1], \mathbb{E}_\theta[s_1^2]$ and $\omega_0(\theta)$ into Equation 3.60, which results in

$$\hat{\mathscr{L}}(\theta) = \begin{cases} \frac{e^{-2[(\theta-\frac{1}{2})h + 1.166(\theta-\frac{1}{2})]} - 1 + 2[(\theta-\frac{1}{2})h + 1.166(\theta-\frac{1}{2})]}{2(\theta-\frac{1}{2})^2} & \text{if } \theta \neq \frac{1}{2} \\ (h+1.166)^2 & \text{if } \theta = \frac{1}{2}. \end{cases} \tag{3.61}$$

Equation 3.61 could be calculated since in the Gaussian case $\delta^+ + \delta^- = 2\varsigma$, where [59]

$$\varsigma = -\pi^{-1} \int_0^\infty x^{-2} \ln\left[\frac{2}{x^2}(1 - e^{-\frac{1}{2}x^2})\right] dx \approx 0.583. \tag{3.62}$$

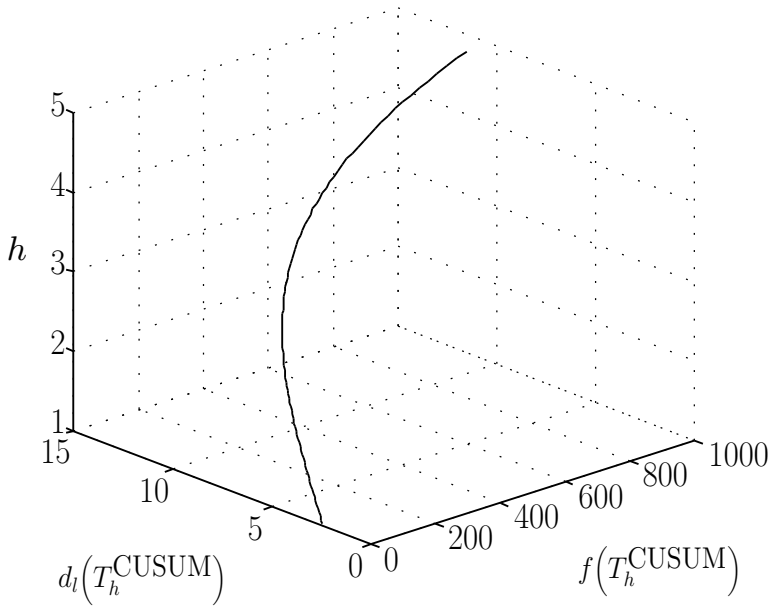


Figure 3.14: Exact values for Equation 3.52 and Equation 3.53 with $q_0 \sim \mathcal{N}(0,1)$, $q_1 \sim \mathcal{N}(1,1)$ and $h \in [1,5]$.

The different ARL functions for $h = 5$ can be found in Figure 3.15.

3.6.2 Pollak's performance measure

The delay measure $d_l(T)$ introduced in Section 3.6.1 can also be replaced by *Pollak's performance measure*, which is equal to [56]

$$d_p(T) = \sup_{1 \leq \tau < \infty} \mathbb{E}_\tau[T - \tau | T \geq \tau], \tag{3.63}$$

which transforms the optimisation problem in Equation 3.54 into

$$\inf_{T \in \mathscr{T}} d_p(T) \text{ subject to } f(T) \geq \lambda. \tag{3.64}$$

A few stopping times have been proposed to solve Equation 3.54. The first stopping time of interest is the *Shiryaev-Roberts* stopping time, which is defined as [51, 55]

$$T_v^{\text{SR}} = \inf\{k \geq 0 | R_k \geq v\},$$

where

$$R_k = \begin{cases} (1 + R_{k-1}) \cdot s_k & \text{if } k > 0 \\ 0 & \text{if } k = 0. \end{cases}$$

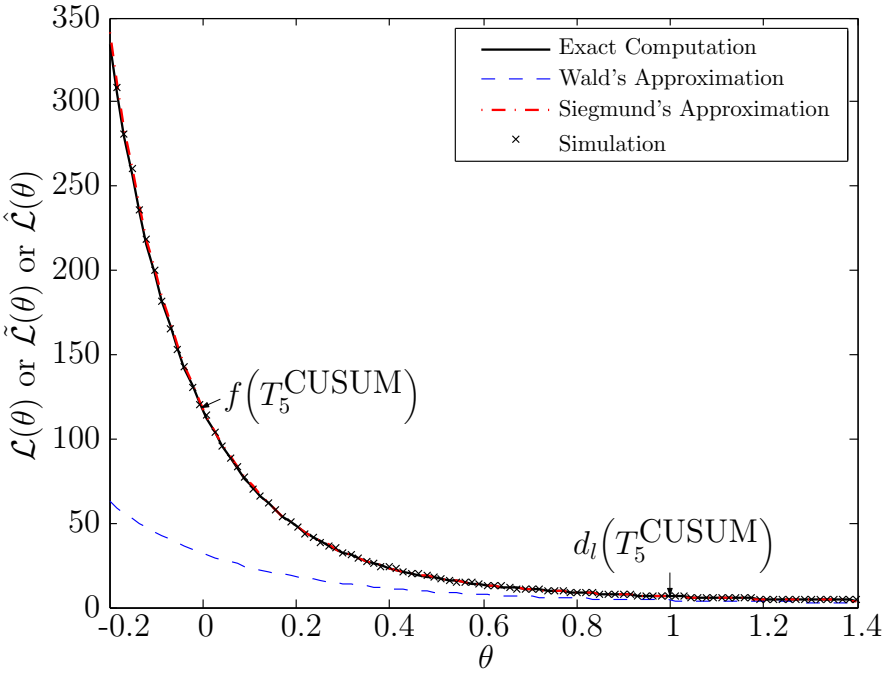


Figure 3.15: The different ARL functions with $q_0 \sim \mathcal{N}(0,1)$, $q_1 \sim \mathcal{N}(1,1)$, $h = 5$ and $\theta = [-0.2, 1.4]$.

The value of R_k can also be calculated non-iteratively via

$$R_k = \sum_{i=1}^{k} \prod_{j=i}^{k} s_j.$$

The *Shiryaev-Roberts-Pollak* stopping time is closely related to Equation 3.65. The only difference is that R_0 is not initialised with 0 but is assumed to be random with distribution equal to the quasi-stationary distribution of R_k. Another stopping time is the *deterministic Shiryaev-Roberts* stopping time. The deterministic Shiryaev-Roberts method once again considers the SR statistic R_0 to be deterministic, but not necessarily equal to zero [58]. It was shown that the Shiryaev-Roberts stopping time and the Shiryaev-Roberts-Pollak stopping time are suboptimal solutions to Equation 3.64 [57]. The Shiryaev-Roberts-Pollak stopping time is however an asymptotically optimal ($\lambda \to \infty$) solution of Equation 3.64 in an $\mathcal{O}(1)$ sense [56].

3.7 CONCLUSION

Many different algorithms were investigated in this chapter, of which only CUSUM (Section 3.6.1) and a Bayesian sequential detection (Section 3.4) variation (called *time-varying maximum likelihood classification* [23]) will be used in the remaining chapters. It is important to realize that even though

only a few of the algorithms are used directly in the remaining chapters, most of the theory in this chapter is important and is provided to derive (or understand) the algorithms that are used in the later chapters. The Neyman-Pearson (Section 3.1) result is critical to include in this chapter as it is the fundamental building block on which sequential analysis rests. The SPRT (Wald's formulation– Section 3.3) must be included as it provides the mathematical background needed to understand CUSUM, which is merely a repeated SPRT [59]. The SPRT algorithm also helps shed light on the Bayesian sequential detection problem. Section 3.5 (Bayesian quickest detection) and Section 3.6.2 (the Shiryaev-Roberts stopping time and its variants) are the only two sections that can be seen as non-critical and are included for the sake of completeness. The Bayesian quickest detection algorithm was not implemented on the datasets in Section 2.8, as the change point of the datasets was not geometrically distributed. As mentioned in Section 6.3 the Shiryaev-Roberts stopping time could still turn out to be useful (in the remote sensing field).

Listing 3.1: The pseudo-code for determining the OC and ASN functions via simulation.

```
N = 100000; %amount of sequences to generate for each theta
%parameter determining the density of the observable sequence
set theta equal to an experimental range;
S = 0; %the sum of the log-likelihood ratios
accept_H_0 = 0; %the amount of times H_0 was accepted
teller = 0; %samples required before a decision is made
%density_H_0(z) and ..._H_1(z) are the density functions
%of H_0 and H_1
delay = zeros(1,N); %vector of delays for each theta
Q = zeros(1,length(theta)); %the OC function
E_T = zeros(1,length(theta)); %the ASN function
fix h; fix a; %upper and lower boundaries of SPRT
for k = 1:length(theta) %iterate through theta
    accept_H_0 = 0; delay = zeros(1,N);
    for n = 1:N %perform N experiments
        exit = false; teller = 0; S = 0;
        while !exit %continue until exit boundaries are crossed
            teller = teller + 1;
            draw a z from density with parameter theta(k);
            s = log(density_H_1(z)/density_H_0(z)); S = S + s;
            if S >= h %crossed upper boundary
                delay(n) = teller; exit = true;
            end%if
            if S <= -a %crossed lower boundary
                accept_H_0 = accept_H_0 + 1;
                delay(n) = teller; exit = true;
            end%if
        end%while
    end%for
    Q(k) = accept_H_0/N; E_T(k) = mean(delay);
end%for
```