1 Introduction

The problem of detecting abrupt changes in the statistical behavior of an observed signal or time series is a classical one, whose provenance dates at least to work in the 1930s on the problem of monitoring the quality of manufacturing processes [224]. In more recent years, this problem has attracted attention in a wide variety of fields, including climate modeling [15], econometrics [4,5,7,8,34,51], environment and public health [110,115,170,201], finance [7,30,193], image analysis [17,214], medical diagnosis [59,84,85,171,229], navigation [148,159], network security [53,55,125,164,208,213], neuroscience [60,66,217,232], other security applications such as fraud detection and counter-terrorism [86,87,125,202], remote sensing (seismic, sonar, radar, biomedical) [104,143,172], video editing [126,134], and even the analysis of historical texts [50,95,182]. This list, although long, is hardly exhaustive, and other applications can be found, for example, in [6,8,18,19,45,52,87,114,128,131,149,161,162,163,165,230]. These cited references only touch the surface of a very diverse and vibrant field, in which this general problem is known variously as *statistical change detection*, *change-point detection*, or *disorder detection*.

Many of these applications, such as those in image analysis, econometrics, or the analysis of historical texts, involve primarily off-line analyses to detect a change in statistical behavior during a pre-specified frame of time or space. In such problems, it is of interest to estimate the occurrence time of a change, and to identify appropriate statistical models before and after the change. However, it is not usually an objective of these applications to perform these functions in real time.

On the other hand, there are many applications of change detection in which it is of interest to perform on-line (i.e. real-time) detection of such changes in a way that minimizes the delay between the time a change occurs and the time it is detected. This latter type of problem is know as the *quickest detection problem*, and this problem arises in many of the above-noted applications. For example, in seismology, quickest detection can be used to detect the onset of seismic events that may presage earthquakes. It is important that such events be detected as quickly as possible so that emergency action can be taken. Similar issues arise in the monitoring of cardiac patients, the monitoring of the radio spectrum for opportunistic wireless transmission, the analysis of financial indicators to detect fundamental shifts in sector performance or foreign exchange trends, the monitoring of computer networks for faults or security breaches, etc. Again, the list of such problems is quite long and diverse.

This book describes a theoretical basis for the design, analysis and understanding of quickest detection algorithms. There are six chapters (plus bibliography) beyond the current one. These are described briefly as follows.

(2) Probabilistic framework
This chapter provides an overview of the elements of probability theory needed to place the quickest detection problem in a mathematical setting. The topics reviewed include probability spaces, random variables, expectations, Radon–Nikodym derivatives, conditional expectations and independence, properties of random sequences, martingales, stopping times, Brownian motion, and Poisson processes. It is assumed that the reader has prior exposure to most of these ideas, and this chapter is intended primarily as a review and as a mechanism for establishing notation and vocabulary.

(3) Markov optimal stopping theory
This chapter develops the concepts and tools of Markov optimal stopping theory that are necessary to derive and understand optimal procedures for quickest detection. These concepts include the general characterization of optimal stopping procedures in terms of the Snell envelope, and the explicit techniques (e.g. dynamic programming) for computing solutions to Markov optimal stopping problems. Three cases are treated: finite-horizon discrete time, infinite-horizon discrete time, and infinite-horizon continuous time. The emphasis here is on the first two of these cases, with the third case being treated only briefly. Several examples are used to illustrate this theory, including the classical selection problem, option trading, etc.

(4) Sequential detection
This chapter formulates and solves the classical sequential detection problem as an optimal stopping problem. This problem deals with the optimization of decision rules for deciding between two possible statistical models for an infinite, statistically homogeneous sequence of random observations. The optimization is carried out by penalizing, in various ways, the probabilities of error and the average amount of time required to reach a decision. By optimizing separately over the error probabilities with the decision time fixed, this problem becomes an optimal stopping problem that can be treated using the methods of the preceding chapter. As this problem is treated in many sources, the primary motivation for including it here is that it serves as a prototype for developing the tools needed in the related problem of quickest detection.

With this in mind, both Bayesian and non-Bayesian, as well as discrete-time and continuous-time formulations of this problem are treated as well as models that combine both a discrete and a continuous nature. In the course of this treatment, a set of analytical techniques is developed that will be useful in the solution and performance analysis of problems of quickest detection to be treated in subsequent chapters. Specific topics included are Bayesian optimization, the Wald–Wolfowitz theorem, the fundamental identity of sequential analysis, Wald's approximations, and diffusion approximations.

A basic conclusion of this chapter is the general optimality of the sequential probability ratio test (SPRT), which, in its various forms, is a central algorithm for the problem of sequential detection.

(5) Bayesian quickest detection

This chapter treats the 'disorder' problem, first posed by Kolmogorov and Shiryaev, in which the distribution of an observed random sequence changes abruptly at an unknown time (the change point). This change point is assumed to have a known geometric prior distribution, and hence this problem provides a Bayesian framework for quickest detection. The choice of a geometric prior is mathematically convenient, but it also provides a reasonable model for a number of practical applications.

The objective of a detection procedure in this situation is to react as quickly as possible to the change in distribution, within a constraint on the probability of reacting before the change occurs. Thus, the design of such procedures involves the satisfaction of optimization criteria comprised of two performance indices: the mean delay until detection, and the probability of false alarm (i.e. premature detection). As with the classical sequential detection problem, this problem can also be formulated as an optimal stopping problem after a suitable transformation.

We also describe various other formulations of this problem starting with its continuous-time analog of detecting a change in the drift of a Brownian motion, where the prior for the change point is assumed to be exponential. We subsequently consider a Poisson model which combines both continuous-time and discrete-time features in its nature and treatment. We further include a different treatment of this problem that focuses on devising a stopping rule that, with high probability, is as close as possible to the change point. This problem is also formulated as an optimal stopping problem and the tools developed in Chapter 3 are used for its treatment.

After a discussion of several alternative optimization criteria comprised of trade-offs similar to the ones mentioned above, we finally conclude this chapter with a game theoretic approach to the problem of Bayesian quickest detection, in which the change point is viewed as having been selected by an opponent ("nature") playing a competitive game with the designer of the detection procedure.

A central theme of this chapter is the general Bayesian optimality of procedures that announce the presence of a change point at the first upcrossing of a threshold by the posterior probability of a change, given the past and present observations. An exception to this general optimality arises in the game theoretic formulation, for which the optimal solution is the so-called cumulative sum (CUSUM) procedure, also known as Page's test, which plays a central role in non-Bayesian formulations of the quickest detection problem. This latter formalism thus provides a bridge between Bayesian and non-Bayesian problems, and a segue to the next chapter of this book.

(6) Non-Bayesian quickest detection

This chapter treats a non-Bayesian formulation of the quickest detection problem, first proposed by Lorden, in which no prior knowledge of the change point is known. For many applications, this formulation is more useful than the Shiryaev

formulation, since the assumption of a prior distribution for the change point is sometimes unrealistic. Without a prior, however, the performance indices used in the Shiryaev formulation of this problem – namely, mean detection delay and false-alarm probability – are not meaningful since there is an infinite set of possible distributions for the observations, one for each possible value of the change point.

Lorden's formulation deals with this difficulty by replacing the mean detection delay with a worst-case conditional delay, where the conditioning is with respect to the change point, and the worst case is taken over all possible values of the change point and all realizations of the measurements leading up to the change point. False alarms are controlled by placing a lower bound on the allowable mean time between false alarms. Here, considering first the discrete-time case, we present a solution to the problem of minimizing delay within this constraint, again by appealing to a related optimal stopping problem. The above-noted CUSUM algorithm is the optimal solution here, and it is in fact the central (although not the only) algorithm arising in non-Bayesian quickest detection problems.

Results from renewal theory are also used here to relate the performance of optimal detection procedures for this problem to that of the classical sequential detection procedures described in Chapter 4. Through this connection, a number of approximations and bounds for the relevant performance indices are developed.

The non-Bayesian quickest detection problem is also treated under the assumption of several continuous-time models for the observations. The problem is seen once again as an optimal stopping problem, but now we introduce a different approach, based on establishing global lower bounds for the performance of all relevant stopping times, in solving it. In the case of a specific continuous-time model we also discuss the practically important problem of an unknown change after the change point.

We finally give several asymptotic results that are useful in the analysis of the CUSUM algorithm, and in its generalization. We also apply this asymptotic analysis to treat the problem of two-sided alternatives (i.e. changes in the mean of unknown sign) in the context of a specific continuous-time observation model.

(7) Additional topics

This final chapter considers the problem of sequential and quickest detection in several settings that arise from practical considerations not treated in the previous chapters. These include decentralized, robust, and adaptive methods for quickest detection. Decentralized problems arise, for example, in applications involving sensor networks or distributed databases. Robust and adaptive methods are generically of interest when there is uncertainty in the statistical models used to describe observations. Other generalizations and alternative formulations of the quickest-detection problem are also described, notably in connection with problems in which the observations do not form an independent sequence. Such problems arise in applications involving the analysis of time series, for example. All the results of this chapter are cast in a discrete-time framework.

The basic idea in the treatment of the decentralized detection problems is to again formulate them as optimal stopping problems and use the results of Chapter 3 to

solve them. Adopting a Bayesian model for the change point, the similarity of the problems of decentralized sequential and quickest detection to the problems treated in Chapters 4 and 5, in both formulation and solution, is easily seen. However, an additional feature here that does not arise in the earlier formulations is the need of optimizing local decisions in addition to global ones.

As noted above, the problems of robust and adaptive quickest detection treat situations of modeling uncertainty. The two approaches are quite different, as robust procedures seek to provide guaranteed performance in the face of small, but potentially damaging, non-parametric uncertainties in statistical models, whereas adaptive procedures are based on the on-line estimation of parametrized models. In the former case, we describe and solve a minimax formulation of quickest detection, whereas the latter problem is solved using combined detection–estimation procedures. In both case, the approach is essentially non-Bayesian.

Finally, we present the problem of quickest detection in the case of more general dependence models than the independent-sampling models used in earlier chapters, again using a non-Bayesian formulation. After first giving a precise generalization of the optimality of the CUSUM to a class of dependent observation processes, we then turn to a more general approach to change detection in time series models based on a general asymptotic local formulation of change detection, which makes heavy use of diffusion approximations to develop asymptotically optimal procedures.

2 Probabilistic framework

2.1 Introduction

Probability theory provides a useful mathematical setting for problems of optimal stopping and statistical change detection. This chapter provides a brief overview of the concepts from probability that will be used in the sequel. This overview is organized into five sections, a review of basic probability (Section 2.2), a collection of results about martingales and stopping times (Section 2.3), some introductory material on Brownian motion and Poisson processes (Section 2.4), an overview of semimartingales (Section 2.5), and the definition and properties of the stochastic integral (Section 2.6). Those who are already familiar with the basic definitions of probability, expected value, random variables, and stochastic convergence, may want to skip Section 2.2.

2.2 Basic setting

In this section, we define some essential notions from probability theory that will be useful in the sequel. Most of this material can be found in many basic books, including [37,46], or in the first chapter of [225].

2.2.1 Probability spaces

The basic notion in a probabilistic model is that of a *random experiment*, in which outcomes are produced according to some chance mechanism. From a mathematical point of view, this notion is contained in an abstraction – a *probability space*, which is a triple $(\Omega, \mathcal{F}, P)$ consisting of the following elements:

- a *sample space* Ω of elemental outcomes of the random experiment;
- an *event class* $\mathcal{F}$, which is a nonempty collection of subsets of Ω to which we wish to assign probabilities; and
- a *probability measure* (or *probability distribution*) P, which is a real-valued set function that assigns probabilities to the events in $\mathcal{F}$.

In order to be able to manipulate probabilities, we do not allow the event class to be arbitrary, but rather we assume that it is a σ*-field* (or σ*-algebra*); that is, we assume that $\mathcal{F}$ is closed under complementation and under countable unions. The usual algebra

of set operations then assures that $\mathcal{F}$ is closed under arbitrary countable sequences of the operations: union, intersection, and complementation. Such a class necessarily contains the sample space Ω and the null set $\emptyset$. The elements of $\mathcal{F}$ are called *events*. A pair $(\Omega, \mathcal{F})$ consisting of a sample space and event class is called a *measurable space* or a *pre-probability space*.

The probability measure P is constrained to have the following properties, which axiomatize the intuitive notion of what probability means:

- $P(\Omega) = 1$;
- $P(F) \geq 0, \ \forall F \in \mathcal{F}$; and
- $$P\left(\bigcup_{n=1}^{\infty} F_n\right) = \sum_{n=1}^{\infty} P(F_n),$$

  for all sequences $\{F_k; k = 1, 2, \ldots\}$ of elements of $\mathcal{F}$ satisfying $F_m \cap F_n = \emptyset, \ \forall m \neq n$.

That is, P is constrained to be non-negative, normalized, and countably additive.

2.2.2 Random variables

The probability space is a useful abstraction that allows us to think of a chance mechanism underlying more concrete, observable phenomena that we wish to model as being random. Such concrete things can be modeled as being *random variables*.

Mathematically, a random variable is defined to be a measurable mapping from the sample space Ω (endowed with the event class $\mathcal{F}$) to the real line $\mathbb{R}$ (endowed with the usual Borel σ-field $\mathcal{B}$).[1] That is, $X : \Omega \to \mathbb{R}$ is a random variable if

$$X^{-1}(B) \in \mathcal{F}, \ \forall B \in \mathcal{B}, \tag{2.1}$$

where $X^{-1}(B)$ denotes the pre-image under X of $B : \{\omega \in \Omega | X(\omega) \in B\}$.

The measurability of X assures that probabilities can be assigned to all Borel subsets of $\mathbb{R}$ via the obvious assignment:

$$P_X(B) = P\left(X^{-1}(B)\right), \ \forall B \in \mathcal{B}. \tag{2.2}$$

In this way, X induces a probability measure P_X on $(\mathbb{R}, \mathcal{B})$ so that $(\mathbb{R}, \mathcal{B}, P_X)$ is also a probability space. Once P_X is known, the structure of the underlying probability space $(\Omega, \mathcal{F}, P)$ is irrelevant in describing the probabilistic behavior of the random variable X.

[1] Recall that the Borel σ-field in $\mathbb{R}$ is the smallest σ-field that contains all intervals. This is a natural event class to consider on the real line, since the intervals, their complements, unions, and intersections, are the sets of most interest in the context of describing the behavior of observed real phenomena.

The information contained in the probability measure P_X is more succinctly described in terms of the *cumulative probability distribution function* (cdf) of X, defined as

$$F_X(x) = P(X \leq x) = P_X((-\infty, x]), \ x \in \mathbb{R}. \tag{2.3}$$

Either of the two functions F_X or P_X determines the other.

The family of all cdf's is the set of all non-decreasing, right-continuous functions with left limit zero and right limit one. That is, all cdf's satisfy these properties; and, given any function with these properties, one can construct a random variable having that function as its cdf. Random variables are classified according to the nature of their cdf's. Two distinct types of interest are: *continuous* random variables, whose cdf's are absolutely continuous functions; and *discrete* random variables, whose cdf's are piecewise constant.

It is sometimes of interest to generalize the notion of a random variable slightly to permit the values $\pm\infty$ in the range of the random variable. To preserve measurability, the sets of outcomes in Ω for which the variable takes on the values $+\infty$ and $-\infty$ must be in $\mathcal{F}$. This generalization of a random variable is known as an *extended* random variable.

2.2.3 Expectation

The cdf of a random variable completely describes its probabilistic behavior. A coarser description of this behavior can be given in terms of the *expected value* of the random variable.

A *simple* random variable is one taking on only finitely many values. The expected value of a simple random variable X taking on the values $x_1, x_2, \ldots, x_n$, is defined as

$$E\{X\} = \sum_{k=1}^{n} x_k P(F_k), \tag{2.4}$$

where $F_k = \{X = x_k\}$. The expected value of a general non-negative random variable X is defined as the (possibly infinite) value

$$E\{X\} = \sup_{\{\text{simple } Y | P(Y \leq X) = 1\}} E\{Y\}. \tag{2.5}$$

The expected value of an arbitrary random variable is defined if at least one of the non-negative random variables, $X^+ = \max\{0, X\}$ and $X^- = (-X)^+$, has a finite expectation, in which case

$$E\{X\} = E\{X^+\} - E\{X^-\}. \tag{2.6}$$

Otherwise $E\{X\}$ is undefined. The interpretation of $E\{X\}$ is as the average value of X, where the average is taken over all values in the range of X weighted by the probabilities with which these values occur.

When $E\{X\}$ exists, we write it as the integral

$$\int_\Omega X(\omega)P(\mathrm{d}\omega) = \int_\Omega X\mathrm{d}P. \tag{2.7}$$

If $E\{|X|\} < \infty$, we say that X is *integrable*.

The simplest possible non-trivial random variable is the indicator function of an event, say F, which is defined as

$$1_F(\omega) = \begin{cases} 1 & \omega \in F \\ 0 & \omega \notin F \end{cases}. \tag{2.8}$$

Since we have $E\{1_F\} = P(F)$, it follows that knowledge of expectations of all random variables is equivalent to knowledge of the probability distribution P. For an event F and a random variable X whose expectation exists, we write

$$E\{X1_F\} = \int_F X(\omega)P(\mathrm{d}\omega) = \int_F X\mathrm{d}P. \tag{2.9}$$

The integral (2.7) is a Lebesgue–Stieltjes integral, and it equals the Lebesgue–Stieltjes integral

$$\int_{\mathbf{R}} x\,P_X(\mathrm{d}x), \tag{2.10}$$

which in turn equals the Riemann–Stieltjes integral

$$\int_{-\infty}^{\infty} x\mathrm{d}F_X(x), \tag{2.11}$$

whenever this integral converges. For a continuous random variable we thus have

$$E\{X\} = \int_{-\infty}^{\infty} xf_X(x)\mathrm{d}x, \tag{2.12}$$

where $f_X(x) = \mathrm{d}F_X(x)/\mathrm{d}x$ is the *probability density function* (pdf) of X. Similarly, for a discrete random variable X taking the values $x_1, x_2, \ldots$, we have

$$E\{X\} = \sum_{k=1}^{\infty} x_k p_X(x_k), \tag{2.13}$$

where $p_X(x) = P(X = x)$ is the *probability mass function* (pmf) of X.

If X is a random variable and g is a measurable function from $(\mathbb{R}, \mathcal{B})$ to $(\mathbb{R}, \mathcal{B})$, then the composite function $Y = g(X)$ is also a random variable, and its expectation (assuming it exists) is given by

$$E\{Y\} = \int_\Omega g(X(\omega))P(\mathrm{d}\omega) = \int_{\mathbf{R}} g(x)P_X(\mathrm{d}x). \tag{2.14}$$

(The right-hand integral also must equal $\int_{\mathbf{R}} yP_Y(\mathrm{d}y)$, of course.) The following quantities are of interest.

- The *moments* of a random variable:

$$E\{X^n\},\ n = 1, 2, \ldots \tag{2.15}$$

The first moment (which is the expected value) is called the *mean* of X.

- The *central moments* of a random variable:

$$E\left\{(X - E\{X\})^n\right\},\ n = 1, 2, \ldots \tag{2.16}$$

The second central moment is the *variance* of X.

- The function $M_X(t) = E\{e^{tX}\}$ for t complex, which is known as the *moment generating function* if $t \in \mathbb{R}$, and the *characteristic function* if $t = iu$ with $i = \sqrt{-1}$ and $u \in \mathbb{R}$. The characteristic function is sometimes written as $\phi_X(u) = M_X(iu)$. Note that P_X and ϕ_X form a unique pair.

A useful result involving expectations of functions of random variables is *Jensen's inequality:*

$$E\{g(X)\} \geq g\left(E\{X\}\right), \tag{2.17}$$

which holds for convex functions g such that the left-hand side exists. If g is strictly convex, then the inequality in Jensen's inequality is strict unless X is almost surely constant.

2.2.4 Radon–Nikodym derivatives

Suppose P and Q are two probability measures on a measurable space $(\Omega, \mathcal{F})$. Then, we have the following theorems.

- *Lebesgue decomposition theorem.* There exists a random variable f (unique up to sets of P-probability zero), and an event H satisfying $P(H) = 0$, such that

$$Q(F) = \int_F f\,\mathrm{d}P + Q(H \cap F),\ \forall\ F \in \mathcal{F}. \tag{2.18}$$

We say that Q is *absolutely continuous with respect to* P (or that P *dominates* Q) if $P(F) = 0$ implies $Q(F) = 0$. We write $Q \ll P$. If $Q \ll P$ and $P \ll Q$, we say that P and Q are *equivalent* and we write $P \equiv Q$.
A trivial corollary to the Lebesgue decomposition theorem is the following.

- *Radon–Nikodym theorem.* Suppose $Q \ll P$. Then there exists a random variable f such that

$$Q(F) = \int_F f\,\mathrm{d}P,\ \forall\ F \in \mathcal{F}. \tag{2.19}$$

The function f appearing in (2.19) is called the *Radon–Nikodym derivative of* Q with respect to P, and we write

$$f(\omega) = \frac{\mathrm{d}Q}{\mathrm{d}P}(\omega). \tag{2.20}$$