Chapter 7

RANDOM WALKS, LARGE DEVIATIONS, AND MARTINGALES

7.1 Introduction

Definition 7.1.1. *Let* $\{X_i;\, i \geq 1\}$ *be a sequence of IID random variables, and let* $S_n = X_1 + X_2 + \cdots + X_n$*. The integer-time stochastic process* $\{S_n;\, n \geq 1\}$ *is called a random walk, or, more precisely, the one-dimensional random walk based on* $\{X_i;\, i \geq 1\}$*.*

For any given n, S_n is simply a sum of IID random variables, but here the behavior of the entire random walk *process*, $\{S_n;\, n \geq 1\}$, is of interest. Thus, for a given real number $\alpha > 0$, we might want to find the probability that the sequence $\{S_n;\, n \geq 1\}$ contains any term for which $S_n \geq \alpha$ (*i.e.*, that a threshold at α is crossed) or to find the distribution of the smallest n for which $S_n \geq \alpha$.

We know that S_n/n essentially tends to $\mathsf{E}[X] = \overline{X}$ as $n \to \infty$. Thus if $\overline{X} < 0$, S_n will tend to drift downward and if $\overline{X} > 0$, S_n will tend to drift upward. This means that the results to be obtained depend critically on whether $\overline{X} < 0$, $\overline{X} > 0$, or $\overline{X} = 0$. Since results for $\overline{X} > 0$ can be easily found from results for $\overline{X} < 0$ by considering $\{-S_n;\, n \geq 1\}$, we usually focus on the case $\overline{X} < 0$.

As one might expect, both the results and the techniques have a very different flavor when $\overline{X} = 0$, since here S_n/n essentially tends to 0 and we will see that the random walk typically wanders around in a rather aimless fashion.[1] With increasing n, σ_{S_n} increases as $\sqrt{n}$ (for X both zero-mean and non-zero-mean), and this is often called diffusion.[2]

[1]When $\overline{X}$ is very close to 0, its behavior for small n resembles that for $\overline{X} = 0$, but for large enough n the drift becomes significant, and this is reflected in the major results.

[2]If we view S_n as our winnings in a zero-mean game, the fact that $S_n/n \to 0$ makes it easy to imagine that a run of bad luck will probably be followed by a run of good luck. However, this is a fallacy here, since the X_n are assumed to be independent. Adjusting one's intuition to understand this at a gut level should be one of the reader's goals in this chapter.

The following three subsections discuss three special cases of random walks. The first two, simple random walks and integer random walks, will be useful throughout as examples, since they can be easily visualized and analyzed. The third special case is that of renewal processes, which we have already studied and which will provide additional insight into the general study of random walks.

After this, Sections 7.2 and 7.3 show how two major application areas, G/G/1 queues and hypothesis testing, can be viewed in terms of random walks. These sections also show why questions related to threshold crossings are so important in random walks.

Section 7.4 then develops the theory of threshold crossings for general random walks and Section 7.5 extends and in many ways simplifies these results through the use of stopping rules and a powerful generalization of Wald's equality known as Wald's identity.

The remainder of the chapter is devoted to a rather general type of stochastic process called martingales. The topic of martingales is both a subject of interest in its own right and also a tool that provides additional insight Rdensage into random walks, laws of large numbers, and other basic topics in probability and stochastic processes.

7.1.1 Simple random walks

Suppose $X_1, X_2, \ldots$ are IID binary random variables, each taking on the value 1 with probability p and -1 with probability $q = 1 - p$. Letting $S_n = X_1 + \cdots + X_n$, the sequence of sums $\{S_n;\, n \geq 1\}$, is called a *simple random walk*. Note that S_n is the difference between positive and negative occurrences in the first n trials, and thus a simple random walk is little more than a notational variation on a Bernoulli process. For the Bernoulli process, X takes on the values 1 and 0, whereas for a simple random walk X takes on the values 1 and -1. For the random walk, if $X_m = 1$ for m out of n trials, then $S_n = 2m - n$, and

$$\Pr\{S_n = 2m - n\} = \frac{n!}{m!(n-m)!}\; p^m (1-p)^{n-m}. \tag{7.1}$$

This distribution allows us to answer questions about S_n for any given n, but it is not very helpful in answering such questions as the following: for any given integer $k > 0$, what is the probability that the sequence $S_1, S_2, \ldots$ ever reaches or exceeds k? This probability can be expressed as[3] $\Pr\{\bigcup_{n=1}^{\infty}\{S_n \geq k\}\}$ and is referred to as the probability that the random walk *crosses a threshold* at k. Exercise 7.1 demonstrates the surprisingly simple result that for a simple random walk with $p \leq 1/2$, this threshold crossing probability is

$$\Pr\left\{\bigcup_{n=1}^{\infty}\{S_n \geq k\}\right\} = \left(\frac{p}{1-p}\right)^k. \tag{7.2}$$

[3]This same probability is often expressed as as $\Pr\{\sup_{n=1} S_n \geq k\}$. For a general random walk, the event $\bigcup_{n\geq 1}\{S_n \geq k\}$ is slightly different from $\sup_{n\geq 1} S_n \geq k$, since $\sup_{n\geq 1} S_n \geq k$ can include sample sequences $s_1, s_2, \ldots$ in which a subsequence of values s_n approach k as a limit but never quite reach k. This is impossible for a simple random walk since all s_k must be integers. It is possible, but can be shown to have probability zero, for general random walks. It is simpler to avoid this unimportant issue by not using the sup notation to refer to threshold crossings.

Sections 7.4 and 7.5 treat this same question for general random walks, but the results are far less simple. They also treat questions such as the overshoot given a threshold crossing, the time at which the threshold is crossed given that it is crossed, and the probability of crossing such a positive threshold before crossing any given negative threshold.

7.1.2 Integer-valued random walks

Suppose next that $X_1, X_2, \dots$ are arbitrary IID integer-valued random variables. We can again ask for the probability that such an integer-valued random walk crosses a threshold at k, *i.e.*, that the event $\bigcup_{n=1}^{\infty}\{S_n \geq k\}$ occurs, but the question is considerably harder than for simple random walks. Since this random walk takes on only integer values, it can be represented as a Markov chain with the set of integers forming the state space. In the Markov chain representation, threshold crossing problems are first passage-time problems. These problems can be attacked by the Markov chain tools we already know, but the special structure of the random walk provides new approaches and simplifications that will be explained in Sections 7.4 and 7.5.

7.1.3 Renewal processes as special cases of random walks

If $X_1, X_2, \dots$ are IID positive random variables, then $\{S_n; n \geq 1\}$ is both a special case of a random walk and also the sequence of arrival epochs of a renewal counting process, $\{N(t); t > 0\}$. In this special case, the sequence $\{S_n; n \geq 1\}$ must eventually cross a threshold at any given positive value α, and the question of whether the threshold is ever crossed becomes uninteresting. However, the trial on which a threshold is crossed and the overshoot when it is crossed are familiar questions from the study of renewal theory. For the renewal counting process, $N(\alpha)$ is the largest n for which $S_n \leq \alpha$ and $N(\alpha) + 1$ is the smallest n for which $S_n > \alpha$, *i.e.*, the smallest n for which the threshold at α is strictly exceeded. Thus the trial at which α is crossed is a central issue in renewal theory. Also the overshoot, which is $S_{N(\alpha)+1} - \alpha$ is familiar as the residual life at α.

Figure 7.1 illustrates the difference between general random walks and positive random walks, *i.e.*, renewal processes. Note that the renewal process in part b) is illustrated with the axes reversed from the conventional renewal process representation. We usually view each renewal epoch as a time (epoch) and view $N(\alpha)$ as the number of renewals up to time α, whereas with random walks, we usually view the number of trials as a discrete-time variable and view the sum of rv's as some kind of amplitude or cost. There is no mathematical difference between these viewpoints, and each viewpoint is often helpful.

7.2 The queueing delay in a G/G/1 queue:

Before analyzing random walks in general, we introduce two important problem areas that are often best viewed in terms of random walks. In this section, the queueing delay in a G/G/1 queue is represented as a threshold crossing problem in a random walk. In the

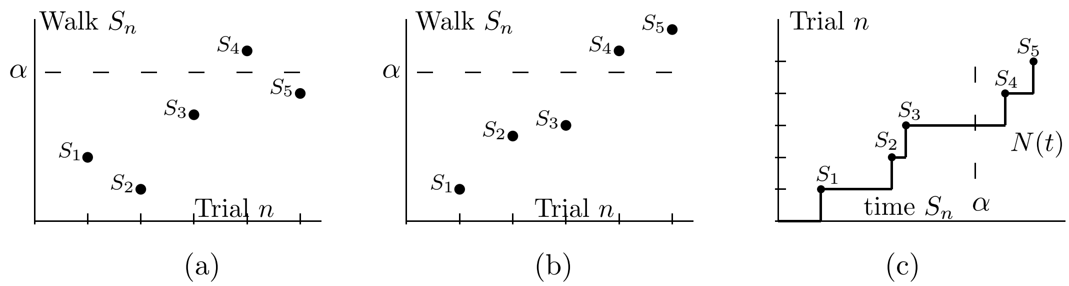


Figure 7.1: The sample function in (a) above illustrates a random walk $S_1, S_2, \dots$, with arbitrary (positive and negative) step sizes $\{X_i;\, i \geq 1\}$. The sample function in (b) illustrates a random walk, $S_1, S_2, \dots$, with only positive step sizes $\{X_i > 0;\, i \geq 1\}$. Thus, $S_1, S_2, \dots$, in (b) are sample renewal points in a renewal process. Note that the axes in (b) are reversed from the usual depiction of a renewal process. The usual depiction, illustrated in (c) for the same sample points, also shows the corresponding counting process. The random walks in parts a) and b) each illustrate a threshold at α, which in each case is crossed on trial 4 with an overshoot $S_4 - \alpha$.

next section, the error probability in a standard type of detection problem is represented as a random walk problem. This detection problem will then be generalized to a sequential detection problem based on threshold crossings in a random walk.

Consider a G/G/1 queue with first-come-first-serve (FCFS) service. We shall associate the probability that a customer must wait more than some given time α in the queue with the probability that a certain random walk crosses a threshold at α. Let $X_1, X_2, \dots$ be the interarrival times of a G/G/1 queueing system; thus these variables are IID with an arbitrary distribution function $\mathsf{F}_X(x) = \Pr\{X_i \leq x\}$. Assume that arrival 0 enters an empty system at time 0, and thus $S_n = X_1 + X_2 + \cdots + X_n$ is the epoch of the n^{th} arrival after time 0. Let $Y_0, Y_1, \dots$, be the service times of the successive customers. These are independent of $\{X_i;\, i \geq 1\}$ and are IID with some given distribution function $\mathsf{F}_Y(y)$. Figure 7.2 shows the arrivals and departures for an illustrative sample path of the process and illustrates the queueing delay for each arrival.

Let W_n be the queueing delay for the nth customer, $n \geq 1$. The system time for customer n is then defined as the queueing delay W_n plus the service time Y_n. As illustrated in Figure 7.2, customer $n \geq 1$ arrives X_n time units after the beginning of customer $n-1$'s system time. If $X_n < W_{n-1} + Y_{n-1}$, *i.e.*, if customer n arrives before the end of customer $n-1$'s system time, then customer n must wait in the queue until n finishes service (in the figure, for example, customer 2 arrives while customer 1 is still in the queue). Thus

$$W_n = W_{n-1} + Y_{n-1} - X_n \qquad \text{if } X_n \leq W_{n-1} + Y_{n-1}. \tag{7.3}$$

On the other hand, if $X_n > W_{n-1} + Y_{n-1}$, then customer $n-1$ (and all earlier customers) have departed when n arrives. Thus n starts service immediately and $W_n = 0$. This is the case for customer 3 in the figure. These two cases can be combined in the single equation

$$W_n = \max[W_{n-1} + Y_{n-1} - X_n,\ 0]; \quad \text{for } n \geq 1; \quad W_0 = 0. \tag{7.4}$$

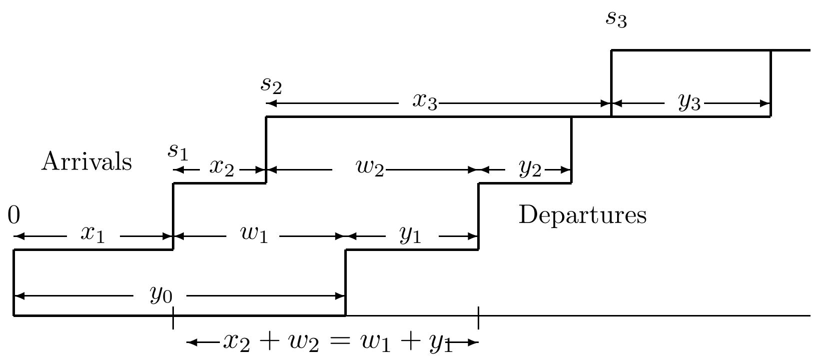


Figure 7.2: Sample path of arrivals and departures from a G/G/1 queue. Customer 0 arrives at time 0 and enters service immediately. Customer 1 arrives at time $s_1 = x_1$. For the case shown above, customer 0 has not yet departed, *i.e.*, $x_1 < y_0$, so customer 1's time in queue is $w_1 = y_0 - x_1$. As illustrated, customer 1's system time (queueing time plus service time) is $w_1 + y_1$.

Customer 2 arrives at $s_2 = x_1 + x_2$. For the case shown above, this is before customer 1 departs at $y_0 + y_1$. Thus, customer 2's wait in queue is $w_2 = y_0 + y_1 - x_1 - x_2$. As illustrated above, x_2+w_2 is also equal to customer 1's system time, so $w_2 = w_1+y_1-x_2$. Customer 3 arrives when the system is empty, so it enters service immediately with no wait in queue, *i.e.*, $w_3 = 0$.

Since Y_{n-1} and X_n are coupled together in this equation for each n, it is convenient to define $U_n = Y_{n-1} - X_n$. Note that $\{U_n;\, n \geq 1\}$ is a sequence of IID random variables. From (7.4), $W_n = \max[W_{n-1} + U_n,\, 0]$, and iterating on this equation,

$$\begin{aligned} W_n &= \max[\max[W_{n-2}+U_{n-1},\, 0]+U_n,\, 0] \\ &= \max[(W_{n-2} + U_{n-1} + U_n),\, U_n,\, 0] \\ &= \max[(W_{n-3}+U_{n-2}+U_{n-1}+U_n),\ (U_{n-1}+U_n),\ U_n,\ 0] \\ &= \cdots \quad \cdots \\ &= \max[(U_1+U_2+\ldots+U_n),\ (U_2+U_3+\ldots+U_n),\ldots,(U_{n-1}+U_n),\ U_n,\ 0]. \end{aligned} \tag{7.5}$$

It is not necessary for the theorem below, but we can understand this maximization better by realizing that if the maximization is achieved at $U_i + U_{i+1} + \cdots + U_n$, then a busy period must start with the arrival of customer $i-1$ and continue at least through the service of customer n. To see this intuitively, note that the analysis above starts with the arrival of customer 0 to an empty system at time 0, but the choice of 0 time and customer number 0 has nothing to do with the analysis, and thus the analysis is valid for any arrival to an empty system. Choosing the largest customer number before n that starts a busy period must then give the correct queueing delay, and thus maximizes (7.5). Exercise 7.2 provides further insight into this maximization.

Define $Z_1^n = U_n$, define $Z_2^n = U_n + U_{n-1}$, and in general, for $i \leq n$, define $Z_i^n = U_n + U_{n-1} + \cdots + U_{n-i+1}$. Thus $Z_n^n = U_n + \cdots + U_1$. With these definitions, (7.5) becomes

$$W_n = \max[0, Z_1^n, Z_2^n, \ldots, Z_n^n]. \tag{7.6}$$

Note that the terms in $\{Z_i^n;\, 1 \leq i \leq n\}$ are the first n terms of a random walk, but it is not the random walk based on $U_1, U_2, \ldots$, but rather the random walk going backward, starting with U_n. Note also that W_{n+1}, for example, is the maximum of a different set of variables, *i.e.*, it is the walk going backward from U_{n+1}. Fortunately, this doesn't matter for the analysis since the ordered variables $(U_n, U_{n-1} \dots, U_1)$ are statistically identical to $(U_1, \dots, U_n)$. The probability that the wait is greater than or equal to a given value α is

$$\Pr\{W_n \geq \alpha\} = \Pr\{\max[0, Z_1^n, Z_2^n, \dots, Z_n^n] \geq \alpha\}. \tag{7.7}$$

This says that, for the n^{th} customer, $\Pr\{W_n \geq \alpha\}$ is equal to the probability that the random walk $\{Z_i^n; 1 \leq i \leq n\}$ crosses a threshold at α by the n^{th} trial. Because of the initialization used in the analysis, we see that W_n is the queueing delay of the n^{th} arrival after the beginning of a busy period (although this n^{th} arrival might belong to a later busy period than that initial busy period).

As noted above, $(U_n, U_{n-1}, \dots, U_1)$ is statistically identical to $(U_1, \dots, U_n)$. This means that $\Pr\{W_n \geq \alpha\}$ is the same as the probability that the first n terms of a random walk based on $\{U_i; i \geq 1\}$ crosses a threshold at α. Since the first $n+1$ terms of this random walk provide one more opportunity to cross α than the first n terms, we see that

$$\cdots \leq \Pr\{W_n \geq \alpha\} \leq \Pr\{W_{n+1} \geq \alpha\} \leq \cdots \leq 1. \tag{7.8}$$

Since this sequence of probabilities is non-decreasing, it must have a limit as $n \to \infty$, and this limit is denoted $\Pr\{W \geq \alpha\}$. Mathematically,[4] this limit is the probability that a random walk based on $\{U_i; i \geq 1\}$ ever crosses a threshold at α. Physically, this limit is the probability that the queueing delay is at least α for any given very large-numbered customer (*i.e.*, for customer n when the influence of a busy period starting n customers earlier has died out). These results are summarized in the following theorem.

Theorem 7.2.1. *Let $\{X_i; i \geq 1\}$ be the IID interarrival intervals of a G/G/1 queue, let $\{Y_i; i \geq 0\}$ be the IID service times, and assume that the system is empty at time 0 when customer 0 arrives. Let W_n be the queueing delay for the n^{th} customer. Let $U_n = Y_{n-1} - X_n$ for $n \geq 1$ and let $Z_i^n = U_n + U_{n-1} + \cdots + U_{n-i+1}$ for $1 \leq i \leq n$. Then for every $\alpha > 0$, and $n \geq 1$, $W_n = \max[0, Z_1^n, Z_2^n, \dots, Z_n^n]$. Also, $\Pr\{W_n \geq \alpha\}$ is the probability that the random walk based on $\{U_i; i \geq 1\}$ crosses a threshold at α on or before the n^{th} trial. Finally, $\Pr\{W \geq \alpha\} = \lim_{n\to\infty} \Pr\{W_n \geq \alpha\}$ is equal to the probability that the random walk based on $\{U_i; i \geq 1\}$ ever crosses a threshold at α.*

Note that the theorem specifies the distribution function of W_n for each n, but says nothing about the joint distribution of successive queueing delays. These are not the same as the distribution of successive terms in a random walk because of the reversal of terms above.

We shall find a relatively simple upper bound and approximation to the probability that a random walk crosses a positive threshold in Section 7.4. From Theorem 7.2.1, this can be applied to the distribution of queueing delay for the G/G/1 queue (and thus also to the M/G/1 and M/M/1 queues).

[4] More precisely, the sequence of queueing delays $W_1, W_2 \dots$, converge in distribution to W, *i.e.*, $\lim_n \mathsf{F}_{W_n}(w) = \mathsf{F}_W(w)$ for each w. We refer to W as the queueing delay in steady state.

7.3 Detection, decisions, and hypothesis testing

Consider a situation in which we make n noisy observations of the outcome of a single discrete random variable H and then guess, on the basis of the observations alone, which sample value of H occurred. In communication technology, this is called a *detection problem*. It models, for example, the situation in which a symbol is transmitted over a communication channel and n noisy observations are received. It similarly models the problem of detecting whether or not a target is present in a radar observation. In control theory, such situations are usually referred to as *decision problems*, whereas in statistics, they are referred to both as *hypothesis testing* and *statistical inference problems*.

The above communication, control, and statistical problems are basically the same, so we discuss all of them using the terminology of hypothesis testing. Thus the sample values of the rv H are called hypotheses. We assume throughout this section that H has only two possible values. Situations where H has more than 2 values can usually be viewed in terms of multiple binary decisions.

It makes no difference in the analysis of binary hypothesis testing what the two hypotheses happen to be called, so we simply denote them as hypothesis 0 and hypothesis 1. Thus H is a binary rv and we abbreviate its PMF as $\mathsf{p}_H(0) = p_0$ and $\mathsf{p}_H(1) = p_1$. Thus p_0 and $p_1 = 1 - p_0$ are the *a priori* probabilities[5] for the two sample values (hypotheses) of the random variable H.

Let $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)$ be the n observations. Suppose, for initial simplicity, that these variables, conditional on $H = 0$ or $H = 1$, have joint conditional densities, $\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0)$ and $\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1)$, that are strictly positive over their common region of definition. The case of primary importance in this section (and the case where random walks are useful) is that in which, conditional on a given sample value of H, the rv's $Y_1, \dots, Y_n$ are IID. Still assuming for simplicity that the observations have densities, say $\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid \ell)$ for $\ell = 0$ or $\ell = 1$, the joint density, conditional on $H = \ell$, is given by

$$\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid \ell) = \prod_{i=1}^{n} \mathsf{f}_{Y|H}(y_i \mid \ell). \tag{7.9}$$

Note that all components of $\boldsymbol{Y}$ are conditioned on the same H, *i.e.*, for a single sample value of H, there are n corresponding sample values of Y.

[5]Statisticians have argued since the earliest days of statistics about the 'validity' of choosing *a priori* probabilities for the hypotheses to be tested. Bayesian statisticians are comfortable with this practice and non-Bayesians are not. Both are comfortable with choosing a probability model for the observations conditional on each hypothesis. We take a Bayesian approach here, partly to take advantage of the power of a complete probability model, and partly because non-Bayesian results, *i.e.*, results that do not depend on the *a priori* probabilies, are often easier to derive and interpret within a collection of probability models using different choices for the *a priori* probabilities. As will be seen, the Bayesian approach also makes it natural to incorporate the results of earlier observations into updated *a priori* probabilities for analyzing later observations. In defense of non-Bayesians, note that the results of statistical tests are often used in areas of significant public policy disagreement, and it is important to give the appearance of a lack of bias. Statistical results can be biased in many more subtle ways than the use of *a priori* probabilities, but the use of an *a priori* distribution is particularly easy to attack.

Assume that it is necessary[6] to select a sample value for H on the basis of the sample value $\boldsymbol{y}$ of the n observations. It is usually possible for the selection (decision) to be incorrect, so we distinguish the sample value h of the actual experimental outcome from the sample value $\hat{h}$ of our selection. That is, the outcome of the experiment specifies h and $\boldsymbol{y}$, but only $\boldsymbol{y}$ is observed. The selection $\hat{h}$, based only on $\boldsymbol{y}$, might be unequal to h. We say that an error has occurred if $\hat{h} \neq h$

We now analyze both how to make decisions and how to evaluate the resulting probability of error. Given a particular sample of n observations $\boldsymbol{y} = y_1, y_2, \dots, y_n$, we can evaluate $\Pr\{H{=}0 \mid \boldsymbol{y}\}$ as

$$\Pr\{H{=}0 \mid \boldsymbol{y}\} = \frac{p_0 f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0)}{p_0 f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0) + p_1 f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1)}. \tag{7.10}$$

We can evaluate $\Pr\{H{=}1 \mid \boldsymbol{y}\}$ in the same way The ratio of these quantities is given by

$$\frac{\Pr\{H{=}0 \mid \boldsymbol{y}\}}{\Pr\{H{=}1 \mid \boldsymbol{y}\}} = \frac{p_0 f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0)}{p_1 f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1)}. \tag{7.11}$$

For a given $\boldsymbol{y}$, if we select $\hat{h} = 0$, then an error is made if $h = 1$. The probability of this event is $\Pr\{H = 1 \mid \boldsymbol{y}\}$. Similarly, if we select $\hat{h} = 1$, the probability of error is $\Pr\{H = 0 \mid \boldsymbol{y}\}$. Thus the probability of error is minimized, for a given $\boldsymbol{y}$, by selecting $\hat{h}{=}0$ if the ratio in (7.11) is greater than 1 and selecting $\hat{h}{=}1$ otherwise. If the ratio is equal to 1, the error probability is the same whether $\hat{h}{=}0$ or $\hat{h}{=}1$ is selected, so we arbitrarily select $\hat{h} = 1$ to make the rule deterministic.

The above rule for choosing $\hat{H}{=}0$ or $\hat{H}{=}1$ is called the *Maximum a posteriori probability* detection rule, abbreviated as the MAP rule. Since it minimizes the error probability for each $\boldsymbol{y}$, it also minimizes the overall error probability as an expectation over $\boldsymbol{Y}$ and H.

The structure of the MAP test becomes clearer if we define the *likelihood ratio* $\Lambda(\boldsymbol{y})$ for the binary decision as

$$\Lambda(\boldsymbol{y}) = \frac{f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0)}{f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1)}.$$

The likelihood ratio is a function only of the observation $\boldsymbol{y}$ and does not depend on the *a priori* probabilities. There is a corresponding rv $\Lambda(\boldsymbol{Y})$ that is a function[7] of $\boldsymbol{Y}$, but we will tend to deal with the sample values. The MAP test can now be compactly stated as

$$\Lambda(\boldsymbol{y}) \begin{cases} > p_1/p_0 \quad ; & \text{select } \hat{h}{=}0 \\ \le p_1/p_0 \quad ; & \text{select } \hat{h}{=}1. \end{cases} \tag{7.12}$$

[6]In a sense, it is in making decisions where the rubber meets the road. For real-world situations, a probabilistic analysis is often connected with studying the problem, but eventually one technology or another must be chosen for a system, one person or another must be hired for a job, one law or another must be passed, etc.

[7]Note that in a non-Bayesian formulation, $\boldsymbol{Y}$ is a different random vector, in a different probability space, for $H = 0$ than for $H = 1$. However, the corresponding conditional probabilities for each exist in each probability space, and thus $\Lambda(\boldsymbol{Y})$ exists (as a different rv) in each space.

For the primary case of interest, where the n observations are IID conditional on H, the likelihood ratio is given by

$$\Lambda(\boldsymbol{y}) = \prod_{i=1}^{n} \frac{f_{Y|H}(y_i \mid 0)}{f_{Y|H}(y_i \mid 1)}.$$

The MAP test then takes on a more attractive form if we take the logarithm of each side in (7.12). The logarithm of $\Lambda(\boldsymbol{y})$ is then a sum of n terms, $\sum_{i=1}^{n} z_i$, where z_i is given by

$$z_i = \ln \frac{f_{Y|H}(y_i \mid 0)}{f_{Y|H}(y_i \mid 1)}.$$

The sample value z_i is called the log-likelihood ratio of y_i for each i, and $\ln \Lambda(\boldsymbol{y})$ is called the log-likelihood ratio of $\boldsymbol{y}$. The test in (7.12) is then expressed as

$$\sum_{i=1}^{n} z_i \begin{cases} > \ln(p_1/p_0) & ; \qquad \text{select } \hat{h}{=}0 \\ \le \ln(p_1/p_0) & ; \qquad \text{select } \hat{h}{=}1. \end{cases} \tag{7.13}$$

As one might imagine, log-likelihood ratios are far more convenient to work with than likelihood ratios when dealing with conditionally IID rv's.

Example 7.3.1. Consider a Poisson process for which the arrival rate λ might be λ_0 or might be λ_1. Assume, for $i \in \{0, 1\}$, that p_i is the *a priori* probability that the rate is λ_i. Suppose we observe the first n arrivals, $Y_1, \dots Y_n$ and make a MAP decision about the arrival rate from the sample values $y_1, \dots, y_n$.

The conditional probability densities for the observations $Y_1, \dots, Y_n$ are given by

$$f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid i) = \prod_{i=1}^{n} \lambda_i e^{-\lambda_i y_i}.$$

The test in (7.13) then becomes

$$n \ln(\lambda_0/\lambda_1) + \sum_{i=1}^{n} (\lambda_1 - \lambda_0) y_i \begin{cases} > \ln(p_1/p_0) & ; \qquad \text{select } \hat{h}{=}0 \\ \le \ln(p_1/p_0) & ; \qquad \text{select } \hat{h}{=}1. \end{cases}$$

Note that the test depends on $\boldsymbol{y}$ only through the time of the nth arrival. This should not be surprising, since we know that, under each hypothesis, the first $n-1$ arrivals are uniformly distributed conditional on the nth arrival time.

The MAP test in (7.12), and the special case in (7.13), are examples of *threshold tests*. That is, in (7.12), a decision is made by calculating the likelihood ratio and comparing it to a threshold $\eta = p_1/p_0$. In (7.13), the log-likelihood ratio $\ln(\Lambda(\boldsymbol{y}))$ is compared with $\ln \eta = \ln(p_1/p_0)$.

There are a number of other formulations of hypothesis testing that also lead to threshold tests, although in these alternative formulations, the threshold $\eta > 0$ need not be equal to

p_1/p_0, and indeed *a priori* probabilities need not even be defined. In particular, then, the threshold test at η is defined by

$$\Lambda(\boldsymbol{y}) \begin{cases} > \eta & ; \quad \text{select } \hat{h}{=}0 \\ \le \eta & ; \quad \text{select } \hat{h}{=}1. \end{cases} \tag{7.14}$$

For example, *maximum likelihood* (ML) detection selects hypothesis $\hat{h} = 0$ if $f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0) > f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1)$, and selects $\hat{h} = 1$ otherwise. Thus the ML test is a threshold test with $\eta = 1$. Note that ML detection is equivalent to MAP detection with equal *a priori* probabilities, but it can be used in many other cases, including those with undefined *a priori* probabilities.

In many detection situations there are unequal costs, say C_0 and C_1, associated with the two kinds of errors. For example one kind of error in a medical prognosis could lead to serious illness and the other to an unneeded operation. A *minimum cost* decision could then minimize the expected cost over the two types of errors. As shown in Exercise 7.5, this is also a threshold test with the threshold $\eta = (C_1p_1)/(C_0p_0)$. This example also illustrates that, although assigning costs to errors provides a rational approach to decision making, there might be no widely agreeable way to assign costs.

Finally, consider the situation in which one kind of error, say $\Pr\{\mathsf{e} \mid H{=}1\}$ is upper bounded by some tolerable limit α and $\Pr\{\mathsf{e} \mid H{=}0\}$ is minimized subject to this constraint. The solution to this is called the *Neyman-Pearson rule*. The Neyman-Pearson rule is of particular interest since it does not require any assumptions about *a priori* probabilities. The next subsection shows that the Neyman-Pearson rule is essentially a threshold test, and explains why one rarely looks at tests other than threshold tests.

7.3.1 The error curve and the Neyman-Pearson rule

Any test, *i.e.*, any deterministic rule for selecting a binary hypothesis from an observation $\boldsymbol{y}$, can be viewed as a function[8] mapping each possible observation $\boldsymbol{y}$ to 0 or 1. If we define A as the set of n-vectors $\boldsymbol{y}$ that are mapped to hypothesis 1 for a given test, then the test can be identified by its corresponding set A.

Given the test A, the error probabilities, given $H = 0$ and $H = 1$ respectively, are given by

$$\Pr\{\boldsymbol{Y} \in A \mid H = 0\}\,; \qquad \Pr\{\boldsymbol{Y} \in A^c \mid H = 1\}\,.$$

Note that these conditional error probabilities depend only on the test A and not on the *a priori* probabilities. We will abbreviate these error probabilities as

$$q_0(A) \;=\; \Pr\{\boldsymbol{Y} \in A \mid H = 0\}\,; \qquad q_1(A) \;=\; \Pr\{\boldsymbol{Y} \in A^c \mid H = 1\}\,.$$

For given *a priori* probabilities, p_0 and p_1, the overall error probability is

$$\Pr\{\mathsf{e}(A)\} \;=\; p_0q_0(A) + p_1q_1(A).$$

[8]By assumption, the decision must be made on the basis of the observation $\boldsymbol{y}$, so a deterministic rule is based solely on $\boldsymbol{y}$, *i.e.*, is a function of $\boldsymbol{y}$. We will see later that randomized rather than deterministic rules are required for some purposes, but such rules are called randomized rules rather than tests.

If A is a threshold test, with threshold η, the set A is given by

$$A = \left\{ \boldsymbol{y} : \frac{\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0)}{\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1)} \leq \eta \right\}.$$

Since threshold tests play a very special role here, we abuse the notation by using η in place of A to refer to a threshold test at η. We can now characterize the relationship between threshold tests and other tests. The following lemma is illustrated in Figure 7.3.

Lemma 7.3.1. *Consider a two dimensional plot in which the error probabilities for each test A are plotted as $(q_1(A), q_0(A))$. Then for any theshold test η, $0 < \eta < \infty$, and any A, the point $(q_1(A), q_0(A))$ is on the closed half plane above and to the right of a straight line of slope $-\eta$ passing through the point $(q_1(\eta), q_0(\eta))$.*

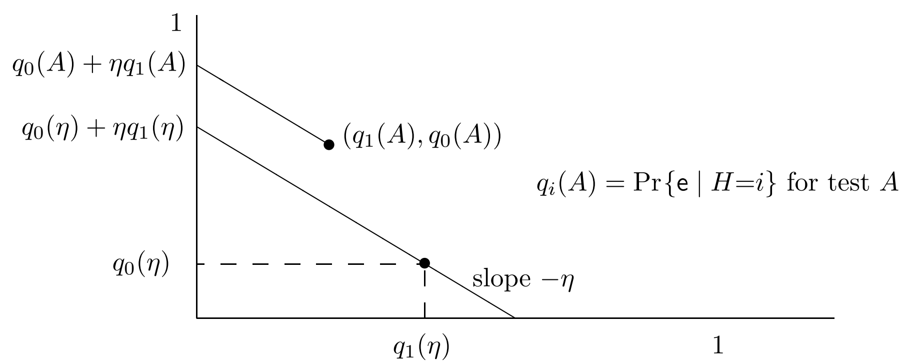


Figure 7.3: Illustration of Lemma 7.3.1

Proof: For any given η, consider the *a priori* probabilities[9] (p_0, p_1) for which $\eta = p_1/p_0$. The overall error probability for test A using these *a priori* probabilities is then

$$\Pr\{\mathsf{e}(A)\} \;=\; p_0 q_0(A) + p_1 q_1(A) \;=\; p_0 \big[q_0(A) + \eta q_1(A)\big].$$

Similarly, the overall error probability for the threshold test η using the same *a priori* probabilities is

$$\Pr\{\mathsf{e}(\eta)\} \;=\; p_0 q_0(\eta) + p_1 q_1(\eta) \;=\; p_0 \left[q_0(\eta) + \eta q_1(\eta)\right].$$

This latter error probability is the MAP error probability for the given p_0, p_1, and is thus the minimum overall error probability (for the given p_0, p_1) over all tests. Thus

$$q_0(\eta) + \eta q_1(\eta) \;\leq\; q_0(A) + \eta q_1(A).$$

As shown in the figure, these are the points at which the lines of slope $-\eta$ from $(q_1(A), q_0(A))$ and $(q_1(\eta), q_0(\eta))$ respectively cross the ordinate axis. Thus all points on the first line

[9]Note that the lemma does not assume any *a priori* probabilities. The MAP test for each *a priori* choice (p_0, p_1) determines $q_0(\eta)$ and $q_1(\eta)$ for $\eta = p_1/p_0$. Its interpretation as a MAP test depends on a model with *a priori* probabilities, but its calculation depends only on p_0, p_1 viewed as parameters.

(including $(q_1(A), q_0(A))$) lie in the closed half plane above and to the right of all points on the second, completing the proof. □

The straight line of slope $-\eta$ through the point $(q_1(\eta), q_0(\eta))$ has the equation $f_\eta(\alpha) = q_0(\eta) + \eta(q_1(\eta) - \alpha)$. Since the lemma is valid for all η, $0 < \eta < \infty$, the point $(q_1(A), q_0(A))$ for an arbitrary test lies above and to the right of the entire family of straight lines that, for each η, pass through $(q_1(\eta), q_0\eta))$ with slope $-\eta$. This family of straight lines has an upper envelope called the error curve, $u(\alpha)$, defined by

$$u(\alpha) = \sup_{0 \leq \eta < \infty} q_0(\eta) + \eta(q_1(\eta) - \alpha). \tag{7.15}$$

The lemma then asserts that for every test A (including threshold tests), we have $u(q_1(A)) \leq q_0(A)$. Also, since every threshold test lies on one of these straight lines, and therefore on or below the curve $u(\alpha)$, we see that the error probabilities $(q_1(\eta), q_0(\eta)$ for each threshold test must lie on the curve $u(\alpha)$. Finally, since each straight line defining $u(\alpha)$ forms a tangent of $u(\alpha)$ and lies on or below $u(\alpha)$, the function $u(\alpha)$ is convex.[10] Figure 7.4 illustrates the error curve.

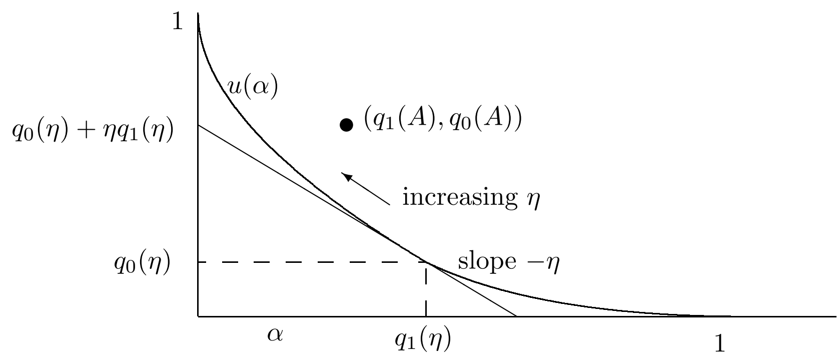


Figure 7.4: Illustration of the error curve $u(\alpha)$ (see (7.15)). Note that $u(\alpha)$ is convex, lies on or above its tangents, and on or below all tests. It can also be seen, either directly from the curve above or from the definition of a threshold test, that $q_1(\eta)$ is non-increasing in η and $q_0(\eta)$ is non-decreasing.

The error curve essentially gives us a tradeoff between the probability of error given $H = 0$ and that given $H = 1$. Threshold tests, since they lie on the error curve, provide optimal points for this tradeoff.

The argument above lacks one important feature. Although all threshold tests lie on the error curve, we have not shown that all points on the error curve correspond to threshold tests. In fact, as the following example shows, it is sometimes necessary to generalize threshold tests by randomization in order to reach all points on the error curve.

Before proceeding to this example, note that Lemma 7.3.1 and the definition of the error curve apply to a broader set of models than discussed so far. First, the lemma still holds if

[10]A convex function of a real variable is sometimes defined as a function with a nonnegative second derivative, but defining it as a function lying on or above all its tangents is more general, allowing for step discontinuities in the first derivative of f. We see the need for this generality shortly.

$\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid \ell)$ is zero over an arbitrary set of $\boldsymbol{y}$ for one or both hypotheses ℓ. The likelihood ratio $\Lambda(\boldsymbol{y})$ is infinite where $\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0) > 0$ and $\mathsf{f}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1) = 0$, but this does not affect the proof of the lemma. Exercise 7.6 helps explain how this situation can affect the error curve.

In addition, it can be seen that the lemma also holds if $\boldsymbol{Y}$ is an n-tuple of discrete rv's. The following example further explains the discrete case and also shows why not all points on the error curve correspond to threshold tests. What we assume throughout is that $\Lambda(\boldsymbol{Y})$ is a rv conditional on $H = 1$ and is a possibly-defective[11] rv conditional on $H = 0$.

Example 7.3.2. About the simplest example of a detection problem is that with a one-dimensional binary observation Y. Assume then that

$$\mathsf{p}_{Y|H}(0 \mid 0) = \mathsf{p}_{Y|H}(1 \mid 1) = \frac{2}{3}; \qquad \mathsf{p}_{Y|H}(0 \mid 1) = \mathsf{p}_{Y|H}(1 \mid 0) = \frac{1}{3}.$$

Thus the observation Y equals H with probability 2/3. The 'natural' decision rule would be to select $\hat{h}$ to agree with the observation y, thus making an error with probability 1/3, both conditional on $H = 1$ and $H = 0$. It is easy to verify from (7.12) (using PMF's in place of densities) that this 'natural' rule is the MAP rule if the *a priori* probabilities are in the range $1/3 \le p_1 < 2/3$.

For $p_1 < 1/3$, the MAP rule can be seen to be $\hat{h} = 0$, no matter what the observation is. Intuitively, hypothesis 1 is too unlikely *a priori* to be overcome by the evidence when $Y = 1$. Similarly, if $p_1 \ge 2/3$, then the MAP rule selects $\hat{h} = 1$, independent of the observation.

The corresponding threshold test (see (7.14)) selects $\hat{h} = y$ for $1/2 \le \eta < 2$. It selects $\hat{h} = 0$ for $\eta < 1/2$ and $\hat{h} = 1$ for $\eta \ge 2$. This means that, although there is a threshold test for each η, there are only 3 resulting error probability pairs, *i.e.*, $(q_1(\eta), q_0(\eta)$ can be (0, 1) or (1/3, 1/3), or (1, 0). The first pair holds for $\eta \ge 2$, the second for $1/2 \le \eta < 2$, and the third for $\eta < 1/2$. This is illustrated in Figure 7.5.

We have just seen that there is a threshold test for each η, $0 < \eta < \infty$, but those threshold tests map to only 3 distinct points $(q_1(\eta), q_0(\eta))$. As can be seen, the error curve joins these 3 points by straight lines.

Let us look more carefully at the tangent of slope -1/2 through the point (1/3, 1/3. This corresponds to the MAP test at $\mathsf{p}_1 = 1/3$. As seen in (7.12), this MAP test selects $\hat{h} = 1$ for $y = 1$ and $\hat{h} = 0$ for $y = 0$. The selection of $\hat{h} = 1$ when $y = 1$ is a don't-care choice in which selecting $\hat{h} = 0$ would yield the same overall error probability, but would change $(q_1(\eta), q_0(\eta)$ from (1/3, 1/3) to (1, 0).

It is not hard to verify (since there are only 4 tests, *i.e.*, deterministic rules, for mapping a binary variable to another binary variable) that no test can achieve the tradeoff between $q_1(A)$ and $q_0(A)$ indicated by the interior points on the straight line between (1/3, 1/3) and (1, 0). However, if we use a randomized rule, mapping 1 to 1 with probability θ and into 0 with probability $1 - \theta$ (along with always mappng 0 to 0), then all points on the straight

[11]If $\mathsf{p}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1) = 0$ and $\mathsf{p}_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0) > 0$ for some $\boldsymbol{y}$, then $\Lambda(\boldsymbol{y})$ is infinite, and thus defective, conditional on $H = 0$. Since the given $\boldsymbol{y}$ has zero probability conditional on $H = 0$, we see that $\Lambda(\boldsymbol{Y})$ is not defective conditional on $H = 1$.

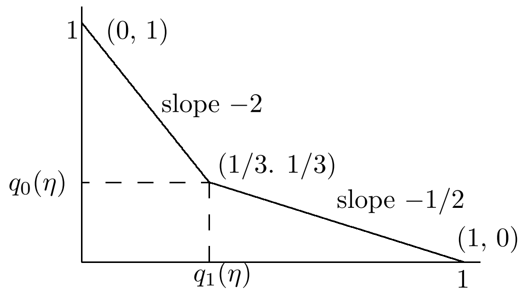


Figure 7.5: Illustration of the error curve $u(\alpha)$ for Example 7.3.2. For all η in the range $1/2 \le \eta < 2$, the threshold test selects $\hat{h} = y$. The corresponding error probabilities are then $q_1(\eta) = q_0(\eta) = 1/3$. For $\eta < 1/2$, the threshold test selects $\hat{h} = 0$ for all y, and for $\eta > 2$, it selects $\hat{h} = 1$ for all y. The error curve (see (7.15)) for points to the right of (1/3, 1/3) is maximized by the straight line of slope -1/2 through (1/3, 1/3). Similarly, the error curve for points to the left of (1/3, 1/3) is maximized by the straight line of slope -2 through (1/3, 1/3). One can visualize the tangent lines as an inverted see-saw, first see-sawing around (0,1), then around (1/3, 1/3), and finally around (1, 0).

line from (1/3, 1/3) to (1, 0) are achieved as θ goes from 0 to 1. In other words, a don't-care choice for MAP becomes an important choice in the tradeoff between $q_1(A)$ and $q_0(A)$.

In the same way, all points on the straight line from (0, 1) to (1/3, 1/3) can be achieved by a randomized rule that maps 0 to 0 with given probability θ (along with always mapping 1 to 1).

In the general case, the error curve can contain straight line segments whenever the distribution function of the likelihood ratio, conditional on $H = 1$, is discontinuous. This is always the case for discrete observations, and, as illustrated in Exercise 7.6, might also occur with continuous observations. To understand this, suppose that $\Pr\{\Lambda(\boldsymbol{Y}) = \eta \mid H{=}1\} = \beta > 0$ for some η, $0 < \eta < \infty$. Then the MAP test at η has a don't-care region of probability β given $H = 1$. This means that if the MAP test is changed to resolve the don't-care case in favor of $H = 0$, then the error probability q_1 is increased by β and the error probability q_0 is decreased by $\eta\beta$. Lemma 7.3.1 is easily seen to be valid whichever way the don't-care cases are resolved, and thus both (q_1, q_0) and $(q_1{+}\beta,\ q_0{-}\eta\beta)$ lie on the error curve. Since all tests lie above and to the right of the straight line of slope $-\eta$ through these points, the error curve has a straight line segment between these points. As mentioned before, any pair of error probabilities on this straight line segment can be realized by using a randomized threshold test at η.

The Neyman-Pearson rule is the rule (randomized where needed) that realizes any desired error probability pair (q_1, q_0) on the error curve, *i.e.*, where $q_0 = u(q_1)$. To be specific, assume the constraint that $q_1 = \alpha$ for any given α, $0 < \alpha \le 1$. Since $\Pr\{\Lambda(\boldsymbol{Y}) > \eta \mid H = 1\}$ is a complementary distribution function, it is non-increasing in η, perhaps with step discontinuities. At $\eta = 0$, it has the value $1 - \Pr\{\Lambda(\boldsymbol{Y}) = 0\}$. As η increases, it decreases to 0, either for some finite η or as $\eta \to \infty$. Thus if $0 < \alpha \le 1 - \Pr\{\Lambda(\boldsymbol{Y}) = 0\}$, we see that α is either equal to $\Pr\{\Lambda(\boldsymbol{y}) > \eta \mid H = 1\}$ for some η or α is on a step discontinuity at some

η. Defining $\eta(\alpha)$ as an η where one of these alternatives occur,[12] there is a solution to the following:

$$\Pr\{\Lambda(\boldsymbol{y})>\eta(\alpha) \mid H{=}1\} \;\le\; \alpha \;\le\; \Pr\{\Lambda(\boldsymbol{y})\ge\eta(\alpha) \mid H{=}1\}\,. \tag{7.16}$$

Note that (7.16) also holds, with $\eta(\alpha) = 0$ if $1 - \Pr\{\Lambda(\boldsymbol{Y}){=}0\} \;<\; \alpha \;\le\; 1$.

The Neyman-Pearson rule, given α, is to use a threshold test at $\eta(\alpha)$ if $\Pr\{\Lambda(\boldsymbol{y}){=}\eta(\alpha) \mid H{=}1\} = 0$. If $\Pr\{\Lambda(\boldsymbol{y}){=}\eta(\alpha) \mid H{=}1\} > 0$, then a randomized test is used at $\eta(\alpha)$. When $\Lambda(\boldsymbol{y}) = \eta(\alpha)$, $\hat{h}$ is chosen to be 1 with probability θ where

$$\theta = \frac{\alpha - \Pr\{\Lambda(\boldsymbol{y}) > \eta(\alpha) \mid H = 1\}}{\Pr\{\Lambda(\boldsymbol{y}) = \eta(\alpha) \mid H = 1\}}. \tag{7.17}$$

This is summarized in the following theorem:

Theorem 7.3.1. *Assume that the likelihood ratio* $\Lambda(\boldsymbol{Y})$ *is a rv under* $H = 1$. *For any* α, $0 < \alpha \le 1$, *the constraint that* $\Pr\{\mathsf{e} \mid H = 1\} \le \alpha$ *implies that* $\Pr\{\mathsf{e} \mid H = 0\} \ge u(\alpha)$ *where* $u(\alpha)$ *is the error curve. Furthermore,* $\Pr\{\mathsf{e} \mid H = 0\} = u(\alpha)$ *if the Neyman-Pearson rule, specified in (7.16) and (7.17), is used.*

Proof: We have shown that the Neyman-Pearson rule has the stipulated error probabilities. An arbitrary decision rule, randomized or not, can be specified by a finite set $A_1, \dots, A_k$ of deterministic decision rules along with a rv V, independent of H and $\boldsymbol{Y}$, with sample values $1, \dots, k$. Letting the PMF of V be $(\theta_1, \dots, \theta_k)$, an arbitrary decision rule, given $\boldsymbol{Y} = \boldsymbol{y}$ and $V = i$ is to use decision rule A_i if $V = i$. The error probabilities for this rule are $\Pr\{\mathsf{e} \mid H{=}1\} = \sum_{i=1}^{k} \theta_i q_1(A_i)$ and $\Pr\{\mathsf{e} \mid H{=}0\} = \sum_{i=1}^{k} \theta_i q_0(A_i)$. It is easy to see that Lemma 7.3.1 applies to these randomized rules in the same way as to deterministic rules. □

All of the decision rules we have discussed are threshold rules, and in all such rules, the first part of the decision rule is to find the likelihood ratio $\Lambda(\boldsymbol{y})$ from the observation $\boldsymbol{y}$. This simplifies the hypothesis testing problem from processing an n-dimensional vector $\boldsymbol{y}$ to processing a single number $\Lambda(\boldsymbol{y})$, and this is typically the major component of making a threshold decision. Any intermediate calculation from $\boldsymbol{y}$ that allows $\Lambda(\boldsymbol{y})$ to be calculated is called a *sufficient statistic*. These usually play an important role in detection, especially for detection in noisy environments.

There are some hypothesis testing problems in which threshold rules are in a sense inappropriate. In particular, the cost of an error under one or the other hypothesis could be highly dependent on the observation $\boldsymbol{y}$. A minimum cost threshold test for each $\boldsymbol{y}$ would still be appropriate, but a threshold test based on the likelihood ratio might be inappropriate since different observations with very different cost structures could have the same likelihood ratio. In other words, in such situations, the likelihood ratio is no longer sufficient to make a minimum cost decision.

[12]In the discrete case, there can be multiple solutions to (7.16) for some values of α, but this does not affect the decision rule. One can choose the smallest value of η to satisfy (7.16) if one wants to eliminate this ambiguity.

So far we have assumed that a decision is made after a predetermined number n of observations. In many situations, observations are costly and introduce an unwanted delay in decisions. One would prefer, after a given number of observations, to make a decision if the resulting probability of error is small enough, and to continue with more observations otherwise. Common sense dictates such a strategy, and the branch of probability theory analyzing such strategies is called *sequential analysis*. In a sense, this is a generalization of Neyman-Pearson tests, which employed a tradeoff between the two types of errors. Here we will have a three-way tradeoff between the two types of errors and the time required to make a decision.

We now return to the hypothesis testing problem of major interest where the observations, conditional on the hypothesis, are IID. In this case the log-likelihood ratio after n observations is the sum $S_n = Z_1 + \cdots Z_n$ of the n IID individual log likelihood ratios. Thus, aside from the question of making a decision after some number of observations, the sequence of log-likelihood ratios $S_1, S_2, \dots$ is a random walk.

Essentially, we will see that an appropriate way to choose how many observations to make, based on the result of the earlier observations, is as follows: The probability of error under either hypothesis is based on $S_n = Z_1 + \cdots + Z_n$. Thus we will see that an appropriate rule is to choose $H{=}0$ if the sample value s_n of S_n is more than some positive threshold α, to choose H_1 if $s_n \leq \beta$ for some negative threshold β, and to continue testing if the sample value has not exceeded either threshold. In other words, the decision time for these sequential decision problems is the time at which the corresponding random walk crosses a threshold. An error is made when the wrong threshold is crossed.

We have now seen that both the queueing delay in a G/G/1 queue and the above sequential hypothesis testing problem can be viewed as threshold crossing problems for random walks. The following section begins a systemic study of such problems.

7.4 Threshold crossing probabilities in random walks

Let $\{X_i;\, i \geq 1\}$ be a sequence of IID random variables (rv's), each with the distribution function $\mathsf{F}_X(x)$, and let $\{S_n;\, n \geq 1\}$ be a random walk with $S_n = X_1 + \cdots + X_n$. Let $g_X(r) = \mathsf{E}\left[e^{rX}\right]$ be the MGF of X and let r_- and r_+ be the upper and lower ends of the interval over which $g(r)$ is finite. We assume throughout that $r_- < 0 < r_+$. Each of r_- and r_+ might be contained in the interval where $g_X(r)$ is finite, and each of r_- and r_+ might be infinite.

The major objective of this section is to develop results about the probability that the sequence $\{S_n;\; n \geq 1\}$ ever crosses a given threshold $\alpha > 0$, *i.e.*, $\Pr\{\bigcup_{n=1}^{\infty}\{S_n \geq \alpha\}\}$. We assume throughout this section that $\overline{X} < 0$ and that X takes on both positive and negative values with positive probability. Under these conditions, we will see that $\Pr\{\bigcup_{n=1}^{\infty}\{S_n \geq \alpha\}\}$ is essentially bounded by an exponentially decreasing function of α. In this section and the next, this exponent is evaluated and shown to be exponentially tight with increasing α. The next section also finds an exponentially tight bound on the probability that a threshold at α is crossed before another threshold at $\beta < 0$ is crossed.

7.4.1 The Chernoff bound

The Chernoff bound was derived and discussed in Section 1.4.3. It was shown in (1.62) that for any $r \in (0, r_+)$ and any $a > \overline{X}$, that

$$\Pr\{S_n \geq na\} \leq \exp\left(n[\gamma_X(r) - ra]\right). \tag{7.18}$$

where $\gamma_X(r) = \ln g_X(r)$ is the semi-invariant MGF of X. The tightest bound of this form is given by

$$\Pr\{S_n \geq na\} \leq \exp[n\mu_X(a)] \qquad \text{where } \mu_X(a) = \inf_{r \in (0,r_+)} \gamma_X(r) - ra.$$

This optimization is illustrated in Figure 7.6 for the case where $\gamma_X(r)$ is finite for all real r. The optimization is simplified by the fact (shown in Exercise 1.24) that $\gamma''(r) > 0$ for $r \in (0, r_+)$. Lemma 1.4.1 (which is almost obvious from the figure) shows that $\mu_X(a) < 0$ for $a > \overline{X}$. This implies that for fixed a, this bound decreases exponentially in n.

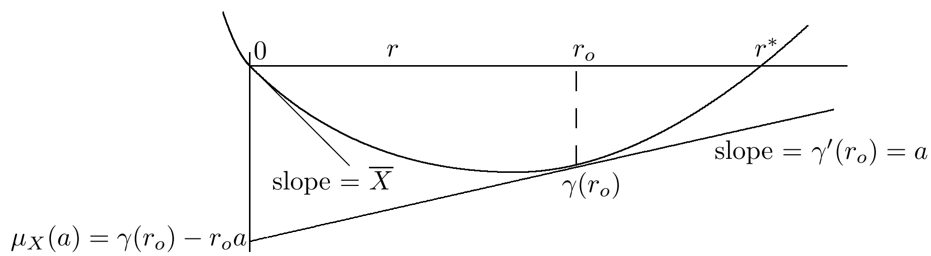


Figure 7.6: Graphical minimization of $\gamma(r) - ar$. For any r, $\gamma(r) - ar$ is the vertical axis intercept of a line of slope a through the point $(r, \gamma(r))$. The minimum occurs when the line of slope a is tangent to the curve, *i.e.*, for the r such that $\gamma'(r) = a$.

Expressing the optimization in the figure algebraically, we try to minimize $\gamma(r) - ar$ by setting the derivative with respect to r equal to 0. Assuming for the moment that this has a solution and the solution is at some value r_o, we have

$$\Pr\{S_n \geq na\} \leq \exp\left\{n\left[\gamma(r_o) - r_o\gamma'(r_o)\right]\right\} \qquad \text{where } \gamma'(r_o) = a. \tag{7.19}$$

This can be expressed somewhat more compactly by substituting $\gamma'(r_o)$ for a in the left side of (7.19). Thus, the Chernoff bound says that for each $r \in (0, r_+)$,

$$\Pr\left\{S_n \geq n\gamma'(r)\right\} \leq \exp\left\{n\left[\gamma(r) - r\gamma'(r)\right]\right\}. \tag{7.20}$$

In principle, we could now bound the probability of threshold crossing, $\Pr\{\bigcup_{n=1}^{\infty}\{S_n \geq \alpha\}\}$, by using the union bound over n and then bounding each term by (7.19). This would be quite tedious and would also give us little insight. Instead, we pause to develop the concept of tilted probabilities. We will use these tilted probabilities in three ways, first to get a better understanding of the Chernoff bound, second to prove that the Chernoff bound is exponentially tight in the limit of large n, and third to prove the Wald identity in the next section. The Wald identity in turn will provide an upper bound on $\Pr\{\bigcup_{n=1}^{\infty}\{S_n \geq \alpha\}\}$ that is essentially exponentially tight in the limit of large α.

7.4.2 Tilted probabilities

As above, let $\{X_n;\, n \geq 1\}$ be a sequence of IID rv's and assume that $g_X(r)$ is finite for $r \in (r_-, r_+)$. Initially assume that X is discrete with the PMF $\mathsf{p}_X(x)$ for each sample value x of X. For any given $r \in (r_-,\, r_+)$, define a new PMF (called a tilted PMF) on X by

$$\mathsf{q}_{X,r}(x) = \mathsf{p}_X(x)\exp[rx - \gamma(r)]. \tag{7.21}$$

Note that $\mathsf{q}_{X,r}(x) \geq 0$ for all sample values x and $\sum_x \mathsf{q}_{X,r}(x) = \sum_x \mathsf{p}_X(x)e^{rx}/\mathsf{E}\left[e^{rx}\right] = 1$, so this is a valid probability assignment.

Imagine a random walk, summing IID rv's X_i, using this new probability assignment rather than the old. We then have the same mapping from sample points of the underlying sample space to sample values of rv's, but we are dealing with a new probability space, *i.e.*, we have changed the probability model, and thus we have changed the probabilities in the random walk. We will further define this new probability measure so that the rv's $X_1, X_2, \dots$ in this new probability space are IID.[13] For every event in the old walk, the same event exists in the new walk, but its probability has changed.

Using (7.21), along with the assumption that $X_1, X_2, \dots$ are independent in the tilted probability assignment, the tilted PMF of an n tuple $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$ is given by

$$\mathsf{q}_{\boldsymbol{X}^n,r}(x_1, \dots, x_n) = \mathsf{p}_{\boldsymbol{X}^n}(x_1, \dots, x_n)\exp(\sum_{i=1}^{n}[rx_i - \gamma(r)]. \tag{7.22}$$

Next we must related the PMF of the sum, $\sum_{i=1}^n X_i$, in the original probability measure to that in the tilted probability measure. From (7.22), note that for every n-tuple $(x_1, \dots, x_n)$ for which $\sum_{i=1}^n x_i = s_n$ (for any given s_n) we have

$$\mathsf{q}_{\boldsymbol{X}^n,r}(x_1, \dots, x_n) = \mathsf{p}_{\boldsymbol{X}^n}(x_1, \dots, x_n)\exp[rs_n - n\gamma(r)].$$

Summing over all $\boldsymbol{x}^n$ for which $\sum_{i=1}^n x_i = s_n$, we then get a remarkable relation between the PMF for S_n in the original and the tilted probability measures:

$$\mathsf{q}_{\boldsymbol{S}_n,r}(s_n) = \mathsf{p}_{S_n}(s_n)\exp[rs_n - n\gamma(r)]. \tag{7.23}$$

This equation is the key to large deviation theory applied to sums of IID rv's . The tilted measure of S_n, for positive r, increases the probability of large values of S_n and decreases that of small values. Since S_n is an IID sum under the tilted measure, however, we can use the laws of large numbers and the CLT around the tilted mean to get good estimates and bounds on the behavior of S_n far from the mean for the original measure.

[13]One might ask whether $X_1, X_2, \dots$ are the same rv's in this new probability measure as in the old. It is usually convenient to view them as being the same, since they correspond to the same mapping from sample points to sample values in both probability measures. However, since the probability space has changed, one can equally well view them as different rv's. It doesn't make any difference which viewpoint is adopted, since all we use the relationship in (7.21), and other similar relationships, to calculate probabilities in the original system.

We now denote the mean of X in the tilted measure as $\mathsf{E}_r[X]$. Using (7.21),

$$\begin{aligned}
\mathsf{E}_r[X] &= \sum_x x\mathsf{q}_{X,r}(x) = \sum_x x\mathsf{p}_X(x)\exp[rx-\gamma(r)] \\
&= \frac{1}{g_X(r)}\sum_x \frac{d}{dr}\mathsf{p}_X(x)\exp[rx] \\
&= \frac{g'_X(r)}{g_X(r)} = \gamma'(r). \qquad (7.24)
\end{aligned}$$

Higher moments of X under the tilted measure can be calculated in the same way, but, more elegantly, the MGF of X under the tilted measure can be seen to be $\mathsf{E}_r[\exp(tX)] = g_X(t+r)/g_X(r)$.

The following theorem uses (7.23) and (7.24) to show that the Chernoff bound is exponentially tight.

Theorem 7.4.1. *Let $\{X_n; n \geq 1\}$ be a discrete IID sequence with a finite MGF for $r \in (r_-, r_+)$ where $r_- < 0 < r_+$. Let $S_n = \sum_{i=1}^n x_i$ for each $n \geq 1$. Then for any $r \in (0, r_+)$, and any $\epsilon > 0$ and $\delta > 0$ there is an m such that for all $n \geq m$,*

$$\Pr\{S_n \geq n(\gamma'(r)-\epsilon)\} \;\geq\; (1-\delta)\exp[n(\gamma(r) - r\gamma'(r) - r\epsilon)]. \qquad (7.25)$$

Proof: The weak law of large numbers, in the form of (1.76), applies to the tilted measure on each S_n. Writing out the limit on n there, we see that for any ϵ, δ, there is an m such that for all $n \geq m$,

$$\begin{aligned}
1-\delta &\leq \sum_{(\gamma'(r)-\epsilon)n \,\leq\, s_n \,\leq\, (\gamma'(r)+\epsilon)n} \mathsf{q}_{S_n,r}(s_n) \\
&= \sum_{(\gamma'(r)-\epsilon)n \,\leq\, s_n \,\leq\, (\gamma'(r)+\epsilon)n} \mathsf{p}_{S_n}(s_n)\exp[rs_n - n\gamma(r)] \qquad (7.26) \\
&\leq \sum_{(\gamma'(r)-\epsilon)n \,\leq\, s_n \,\leq\, (\gamma'(r)+\epsilon)n} \mathsf{p}_{S_n}(s_n)\exp[n(r\gamma'(r) + r\epsilon - \gamma(r))] \qquad (7.27) \\
&\leq \sum_{(\gamma'(r)-\epsilon)n \,\leq\, s_n} \mathsf{p}_{S_n}(s_n)\exp[n(r\gamma'(r) + r\epsilon - \gamma(r))] \qquad (7.28) \\
&= \exp[n(r\gamma'(r) + r\epsilon - \gamma(r))]\,\Pr\{S_n \geq n(\gamma'(r)-\epsilon)\}. \qquad (7.29)
\end{aligned}$$

The equality (7.26) follows from (7.23) and the inequality (7.27) follows because $s_n \leq \gamma'(r) + \epsilon$ in the sum. The next inequality is the result of adding additional positive terms into the sum, and (7.29) simply replaces the sum over a PMF with the probability of the given event. Since (7.29) is equivalent to (7.25), the proof is complete. □

The structure of the above proof can be used in many situations. A tilted probability measure is used to focus on the tail of a distribution, and then some known result about the tilted distribution is used to derive the desired result about the given distribution.

7.4.3 Back to threshold crossings

The Chernoff bound is convenient (and exponentially tight) for understanding $\Pr\{S_n \geq na\}$ as a function of n for fixed a. It does not give us as much direct insight into $\Pr\{S_n \geq \alpha\}$ as a function of n for fixed α, which tells us something about *when* a threshold at α is most likely to be crossed. Additional insight can be achieved by substituing $\alpha/\gamma'(r_0)$ for n in (7.19), getting

$$\Pr\{S_n \geq \alpha\} \leq \exp\left\{\alpha\left[\frac{\gamma(r_o)}{\gamma'(r_o)} - r_o\right]\right\} \qquad \text{where } \gamma'(r_o) = \alpha/n. \tag{7.30}$$

A nice graphic interpretation of this equation is given in Figure 7.7. Note that the exponent in α, namely $\gamma(r_0)/\gamma'(r_0) - r_0$ is the negative of the horizontal axis intercept of the line of slope $\gamma'(r_0) = \alpha/n$ in Figure 7.7.

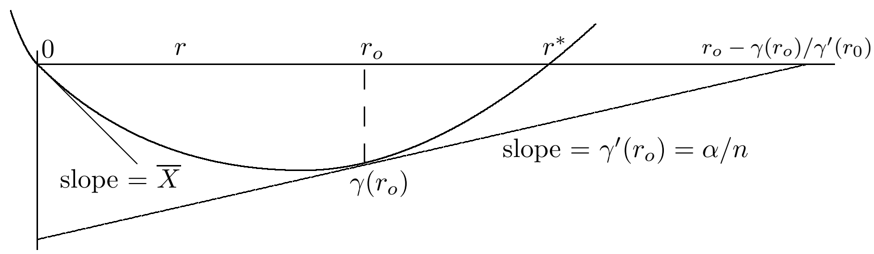


Figure 7.7: The exponent in α for $\Pr\{S_n \geq \alpha\}$, minimized over r. The minimization is the same as that in Figure 7.6, but $\gamma(r_o)/\gamma'(r_o) - r_o$ is the negative of the horizontal axis intercept of the line tangent to $\gamma(r)$ at $r = r_0$.

For fixed α, then, we see that for very large n, the slope α/n is very small and this horizontal intercept is very large. As n is decreased, the slope increases, r_0 increases, and the horizontal intercept decreases. When r_0 increases to the point labeled r^* in the figure, namely the $r > 0$ at which $\gamma(r) = 0$, then the exponent decreases to r^*. When n decreases even further, r_0 becomes larger than r^* and the horizontal intercept starts to increase again.

Since r^* is the minimum horizontal axis intercept of this class of straight lines, we see that the following bound holds for all n,

$$\Pr\{S_n \geq \alpha\} \leq \exp(-r^*\alpha) \qquad \text{for arbitrary } \alpha > 0,\ n \geq 1, \tag{7.31}$$

where r^* is the positive root of $\gamma(r) = 0$.

The graphical argument above assumed that there is some $r^* > 0$ such that $\gamma_X(r^*) = 0$. However, if $r_+ = \infty$, then (since X takes on positive values with positive probability by assumption) $\gamma(r)$ can be seen to approach ∞ as $r \to \infty$. Thus r^* must exist because of the continuity of $\gamma(r)$ in (r_-, r_+). This is summarized in the following lemma.

Lemma 7.4.1. *Let $\{S_n = X_1 + \cdots + X_n;\ n \geq 1\}$ be a random walk where $\{X_i;\ i \geq 1\}$ is an IID sequence where $g_X(r)$ exists for all $r \geq 0$, where $\overline{X} < 0$, and where X takes on both positive and negative values with positive probability. Then for all integer $n > 0$ and* *all* $\alpha > 0$

$$\Pr\{S_n \geq \alpha\} \;\leq\; \exp\{\alpha[\gamma(r_0)n/\alpha - r_0]\} \;\leq\; \exp(-r^*\alpha). \tag{7.32}$$

where $\gamma'(r_o) = \alpha/n$.

Next we briefly consider the situation where $r_+ < \infty$. There are two cases to consider, first where $\gamma(r_+)$ is infinite and second where it is finite. Examples of these cases are given in Exercise 1.22, parts b) and c) respectively. For the case $\gamma(r_+) = \infty$, Exercise 1.23 shows that $\gamma(r)$ goes to ∞ as r increases toward r_+. Thus r^* must exist in this case 1 by continuity.

The case $\gamma(r_+) < \infty$ is best explained by Figure 7.8. As explained there, if $\gamma'(r) = \alpha/n$ for some $r_0 < r_+$, then (7.19) and (7.30) hold as before. Alternatively, if $\alpha/n > \gamma'(r)$ for all $r < r_+$, then the Chernoff bound is optimized at $r = r_+$, and we have

$$\Pr\{S_n \geq \alpha\} \leq \exp\{n[\gamma(r_+) - r_+\alpha/n]\} = \exp\{\alpha[\gamma(r_+)n/\alpha - r_+]\}. \tag{7.33}$$

If we modify the definition of r^* to be the supremum of $r > 0$ such that $\gamma(r) < 0$, then $r^* = r_+$ in this case, and (7.32) is valid in general with the obvious modification of r_o.

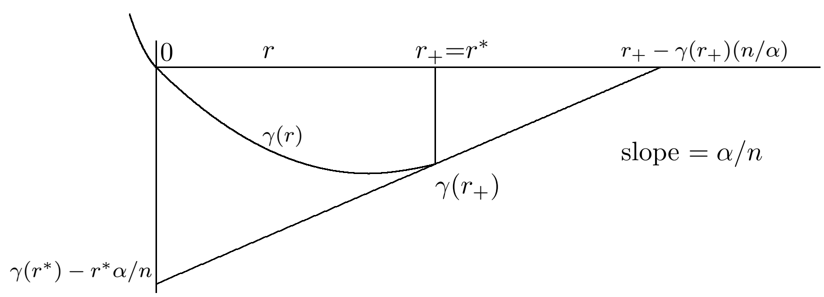


Figure 7.8: Graphical minimization of $\gamma(r) - (\alpha/n)r$ for the case where $\gamma(r^+) < \infty$. As before, for any $r < r_+$, we can find $\gamma(r) - r\alpha/n$ by drawing a line of slope (α/n) from the point $(r, \gamma(r))$ to the vertical axis. If $\gamma'(r) = \alpha/n$ for some $r < r_+$, the minimum occurs at that r. Otherwise, as shown in the figure, it occurs at $r = r_+$.

We could now use this result, plus the union bound, to find an upper bound on the probabiity of threshold crossing, *i.e.*, $\Pr\{\bigcup_{n=1}^{\infty}\{S_n \geq \alpha\}\}$. The coefficients in this are somewhat messy and change according to the special cases discussed above. It is far more instructive and elegant to develop Wald's identity, which shows that $\Pr\{\bigcup_n \{S_n \geq \alpha\}\} \leq \exp[-\alpha r^*]$. This is slightly stronger than the Chernoff bound approach in that the bound on the probability of the union is the same as that on a single term in the Chernoff bound approach. The main value of the Chernoff bound approach, then, is to provide assurance that the bound is exponentially tight.

7.5 Thresholds, stopping rules, and Wald's identity

The following lemma shows that a random walk with both a positive and negative threshold, say $\alpha > 0$ and $\beta < 0$, eventually crosses one of the thresholds. Figure 7.9 illustrates two sample paths and how they cross thresholds. More specifically, the random walk first crosses a threshold at trial n if $\beta < S_i < \alpha$ for $1 \leq i < n$ and either $S_n \geq \alpha$ or $S_n \leq \beta$. For now we make no assumptions about the mean or MGF of S_n.

The trial at which a threshold is first crossed is a possibly-defective rv J. The event $J = n$ (*i.e.*, the event that a threshold is first crossed at trial n), is a function of $S_1, S_2, \dots, S_n$, and thus, in the notation of Section 4.5, J is a possibly-defective stopping trial. The following lemma shows that J is actually a stopping trial, namely that stopping (threshold crossing) eventually happens with probability 1.

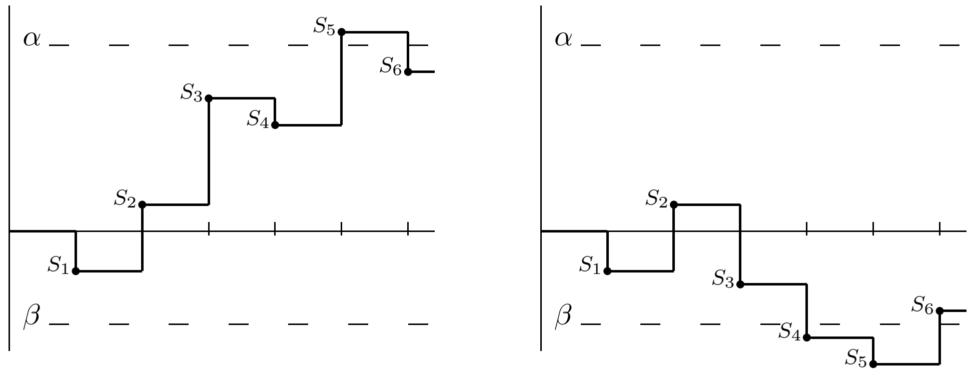


Figure 7.9: Two sample paths of a random walk with two thresholds. In the first, the threshold at α is crossed at $J = 5$. In the second, the threshold at β is crossed at $J = 4$

Lemma 7.5.1. *Let $\{X_i; i \geq 1\}$ be IID rv's, not identically 0. For each $n \geq 1$, let $S_n = X_1 + \cdots + X_n$. Let $\alpha > 0$ and $\beta < 0$ be arbitrary, and let J be the smallest n for which either $S_n \geq \alpha$ or $S_n \leq \beta$. Then J is a random variable (*i.e., *$\lim_{m\to\infty} \Pr\{J \geq m\} = 0$) and has finite moments of all orders.*

Proof: Since X is not identically 0, there is some n for which either $\Pr\{S_n \leq -\alpha + \beta\} > 0$ or for which $\Pr\{S_n \geq \alpha - \beta\} > 0$. For any such n, define ϵ by

$$\epsilon = \max[\Pr\{S_n \leq -\alpha + \beta\}\,,\ \Pr\{S_n \geq \alpha - \beta\}].$$

For any integer $k \geq 1$, given that $J > n(k-1)$, and given any value of $S_{n(k-1)}$ in (β, α), a threshold will be crossed by time nk with probability at least ϵ. Thus,

$$\Pr\{J > nk \mid J > n(k-1)\} \leq 1 - \epsilon,$$

Iterating on k,

$$\Pr\{J > nk\} \leq (1 - \epsilon)^k.$$

This shows that J is finite with probability 1 and that $\Pr\{J \geq j\}$ goes to 0 at least geometrically in j. It follows that the moment generating function $\mathsf{g}_J(r)$ of J is finite in a region around $r = 0$, and that J has moments of all orders. □

7.5.1 Wald's identity for two thresholds

Theorem 7.5.1 (Wald's identity). *Let $\{X_i;\, i \geq 1\}$ be IID and let $\gamma(r) = \ln\{\mathsf{E}\left[e^{rX}\right]\}$ be the semi-invariant MGF of each X_i. Assume $\gamma(r)$ is finite in the open interval (r_-, r_+) with $r_- < 0 < r_+$. For each $n \geq 1$, let $S_n = X_1 + \cdots + X_n$. Let $\alpha > 0$ and $\beta < 0$ be arbitrary real numbers, and let J be the smallest n for which either $S_n \geq \alpha$ or $S_n \leq \beta$. Then for all $r \in (r_-, r_+)$,*

$$\mathsf{E}\left[\exp(rS_J - J\gamma(r))\right] = 1. \tag{7.34}$$

The following proof is a simple application of the tilted probability distributions discussed in the previous section..

Proof: We assume that X_i is discrete for each i with the PMF $\mathsf{p}_X(x)$. For the arbitrary case, the PMF's must be replaced by distribution functions and the sums by Stieltjes integrals, thus complicating the technical details but not introducing any new ideas.

For any given $r \in (r_-, r_+)$, we use the tilted PMF $\mathsf{q}_{X,r}(x)$ given in (7.21) as

$$\mathsf{q}_{X,r}(x) = \mathsf{p}_X(x) \exp[rx - \gamma(r)].$$

Taking the X_i to be independent in the tilted probability measure, the tilted PMF for the n-tuple $\boldsymbol{X}^n = (X_1, X_2, \dots, X_n)$ is given by

$$\mathsf{q}_{\boldsymbol{X}^n,r}(\boldsymbol{x}^n) = \mathsf{p}_{\boldsymbol{X}^n}(\boldsymbol{x}^n) \exp[rs_n - n\gamma(r)] \qquad \text{where } s_n = \sum_{i=1}^{n} x_i.$$

Now let $\mathcal{T}_n$ be the set of n-tuples $X_1, \dots, X_n$ such that $\beta < S_i < \alpha$ for $1 \leq i < n$ and either $S_n \geq \alpha$ or $S_n \leq \beta$. That is, $\mathcal{T}_n$ is the set of $\boldsymbol{x}^n$ for which the stopping trial J has the sample value n. The PMF for the stopping trial J in the tilted measure is then

$$\begin{aligned}
\mathsf{q}_{J,r}(n) &= \sum_{\boldsymbol{x}^n \in \mathcal{T}_n} \mathsf{q}_{\boldsymbol{X}^n,r}(\boldsymbol{x}^n) = \sum_{\boldsymbol{x}^n \in \mathcal{T}_n} \mathsf{p}_{\boldsymbol{X}^n}(\boldsymbol{x}^n) \exp[rs_n - n\gamma(r)] \\
&= \mathsf{E}\left[\exp[rS_n - n\gamma(r) \mid J{=}n\right] \; \Pr\{J = n\}\,.
\end{aligned} \tag{7.35}$$

Lemma 7.5.1 applies to the tilted PMF on this random walk as well as to the original PMF, and thus the sum of $\mathsf{q}_J(n)$ over n is 1. Also, the sum over n of the expression on the right is $\mathsf{E}\left[\exp[rS_J - J\gamma(r)]\right]$, thus completing the proof. □

After understanding the details of this proof, one sees that it is essentially just the statement that J is a non-defective stopping rule in the tilted probability space.

We next give a number of examples of how Wald's identity can be used.

7.5.2 The relationship of Wald's identity to Wald's equality

The trial J at which a threshold is crossed in Wald's identity is a stopping trial in the terminology of Chapter 4. If we take the derivative with respect to r of both sides of (7.34), we get

$$\mathsf{E}\left[[S_J - J\gamma'(r)\right] \; \exp\{rS_J - J\gamma(r)\} = 0.$$

Setting $r = 0$ and recalling that $\gamma(0) = 0$ and $\gamma'(0) = \overline{X}$, this becomes Wald's equality as established in Theorem 4.5.1,

$$\mathsf{E}[S_J] = \mathsf{E}[J]\,\overline{X}. \tag{7.36}$$

Note that this derivation of Wald's equality is restricted[14] to a random walk with two thresholds (and this automatically satisfies the constraint in Wald's equality that $\mathsf{E}[J] < \infty$). The result in Chapter 4 is more general, applying to any stopping trial such that $\mathsf{E}[J] < \infty$.

The second derivative of (7.34) with respect to r is

$$\mathsf{E}\left[\left[(S_J - J\gamma'(r))^2 - J\gamma''(r)\right]\right] \exp\{rS_J - J\gamma(r)\} = 0.$$

At $r = 0$, this is

$$\mathsf{E}\left[S_J^2 - 2JS_J\overline{X} + J^2\overline{X}^2\right] = \sigma_X^2\,\mathsf{E}[J]. \tag{7.37}$$

This equation is often difficult to use because of the cross term between S_J and J, but its main application comes in the case where $\overline{X} = 0$. In this case, Wald's equality provides no information about $\mathsf{E}[J]$, but (7.37) simplifies to

$$\mathsf{E}\left[S_J^2\right] = \sigma_X^2\,\mathsf{E}[J]. \tag{7.38}$$

7.5.3 Zero-mean simple random walks

As an example of (7.38), consider the simple random walk of Section 7.1.1 with $\Pr\{X{=}1\} = \Pr\{X{=}-1\} = 1/2$, and assume that $\alpha > 0$ and $\beta < 0$ are integers. Since S_n takes on only integer values and changes only by ± 1, it takes on the value α or β before exceeding either of these values. Thus $S_J = \alpha$ or $S_J = \beta$. Let q_α denote $\Pr\{S_J = \alpha\}$. The expected value of S_J is then $\alpha q_\alpha + \beta(1 - q_\alpha)$. From Wald's equality, $\mathsf{E}[S_J] = 0$, so

$$q_\alpha = \frac{-\beta}{\alpha - \beta}; \qquad 1 - q_\alpha = \frac{\alpha}{\alpha - \beta}. \tag{7.39}$$

From (7.38),

$$\sigma_X^2\,\mathsf{E}[J] = \mathsf{E}\left[S_J^2\right] = \alpha^2 q_\alpha + \beta^2(1 - q_\alpha). \tag{7.40}$$

Using the value of q_α from (7.39) and recognizing that $\sigma_X^2 = 1$,

$$\mathsf{E}[J] = -\beta\alpha/\sigma_X^2 = -\beta\alpha. \tag{7.41}$$

As a sanity check, note that if α and β are each multiplied by some large constant k, then $\mathsf{E}[J]$ increases by k^2. Since $\sigma_{S_n}^2 = n$, we would expect S_n to fluctuate with increasing n, with typical values growing as $\sqrt{n}$, and thus it is reasonable that the expected time to reach a threshold increases with the product of the distances to the thresholds.

[14] This restriction is quite artificial and made simply to postpone any consideration of generalizations until discussing martingales.

We also notice that if β is decreased toward $-\infty$, while holding α constant, then $q_\alpha \to 1$ and $\mathsf{E}[J] \to \infty$. This helps explain Example 4.5.2 where one plays a coin tossing game, stopping when finally ahead. This shows that if the coin tosser has a finite capital b, *i.e.*, stops either on crossing a positive threshold at 1 or a negative threshold at $-b$, then the coin tosser wins a small amount with high probability and loses a large amount with small probability.

For more general random walks with $\overline{X} = 0$, there is usually an overshoot when the threshold is crossed. If the magnitudes of α and β are large relative to the range of X, however, it is often reasonable to ignore the overshoots, and $-\beta\alpha/\sigma_X^2$ is then often a good approximation to $\mathsf{E}[J]$. If one wants to include the overshoot, then the effect of the overshoot must be taken into account both in (7.39) and (7.40).

7.5.4 Exponential bounds on the probability of threshold crossing

We next apply Wald's identity to complete the analysis of crossing a threshold at $\alpha > 0$ when $\overline{X} < 0$.

Corollary 7.5.1. *Under the conditions of Theorem 7.5.1, assume that* $\overline{X} < 0$ *and that* $r^* > 0$ *exists such that* $\gamma(r^*) = 0$. *Then*

$$\Pr\{S_J \geq \alpha\} \leq \exp(-r^*\alpha). \tag{7.42}$$

Proof: Wald's identity, with $r = r^*$, reduces to $\mathsf{E}[\exp(r^*S_J)] = 1$. We can express this as

$$\Pr\{S_J \geq \alpha\}\,\mathsf{E}[\exp(r^*S_J) \mid S_J \geq \alpha] + \Pr\{S_J \leq \beta\}\,\mathsf{E}[\exp(r^*S_J) \mid S_J \leq \beta] = 1. \tag{7.43}$$

Since the second term on the left is nonnegative,

$$\Pr\{S_J \geq \alpha\}\,\mathsf{E}[\exp(r^*S_J) \mid S_J \geq \alpha] \leq 1. \tag{7.44}$$

Given that $S_J \geq \alpha$, we see that $\exp(r^*S_J) \geq \exp(r^*\alpha)$. Thus

$$\Pr\{S_J \geq \alpha\}\exp(r^*\alpha) \leq 1. \tag{7.45}$$

which is equivalent to (7.42). □

This bound is valid for all $\beta < 0$, and thus is also valid in the limit $\beta \to -\infty$ (see Exercise 7.10 for a more detailed demonstration that (7.42) is valid without a lower threshold). We see from this that the case of a single threshold is really a special case of the two threshold problem, but as seen in the zero-mean simple random walk, having a second threshold is often valuable in further understanding the single threshold case. Equation (7.42) is also valid for the case of Figure 7.8, where $\gamma(r) < 0$ for all $r \in (0, r_+]$ and $r^* = r_+$.

The exponential bound in (7.31) shows that $\Pr\{S_n \geq \alpha\} \leq \exp(-r^*\alpha)$ for each n; (7.42) is stronger than this. It shows that $\Pr\{\bigcup_n\{S_n \geq \alpha\}\} \leq \exp(-r^*\alpha)$. This also holds in the limit $\beta \to -\infty$.

When Corollary 7.5.1 is applied to the G/G/1 queue in Theorem 7.2.1, (7.42) is referred to as the *Kingman Bound*.

Corollary 7.5.2 (Kingman Bound). *Let $\{X_i;\ i \geq 1\}$ and $\{Y_i;\ i \geq 0\}$ be the interarrival intervals and service times of a G/G/1 queue that is empty at time 0 when customer 0 arrives. Let $\{U_i = Y_{i-1} - X_i;\ i \geq 1\}$, and let $\gamma(r) = \ln\{\mathsf{E}\left[e^{Ur}\right]\}$ be the semi-invariant moment generating function of each U_i. Assume that $\gamma(r)$ has a root at $r^* > 0$. Then W_n, the queueing delay of the nth arrival, and W, the steady state queueing delay, satisfy*

$$\Pr\{W_n \geq \alpha\} \leq \Pr\{W \geq \alpha\} \leq \exp(-r^*\alpha) \quad ; \qquad \textit{for all } \alpha > 0. \tag{7.46}$$

In most applications, a positive threshold crossing for a random walk with a negative drift corresponds to some exceptional, and often undesirable, circumstance (for example an error in the hypothesis testing problem or an overflow in the G/G/1 queue). Thus an upper bound such as (7.42) provides an assurance of a certain level of performance and is often more useful than either an approximation or an exact expression that is very difficult to evaluate. Since these bounds are exponentially tight, they also serve as rough approximations.

For a random walk with $\overline{X} > 0$, the exceptional circumstance is $\Pr\{S_J \leq \beta\}$. This can be analyzed by changing the sign of X and β and using the results for a negative expected value. These exponential bounds do not work for $\overline{X} = 0$, and we will not analyze that case here other than the result in (7.38).

Note that the simple bound on the probability of crossing the upper threshold in (7.42) (and thus also the Kingman bound) is an upper bound (rather than an equality) because, first, the effect of the lower threshold was eliminated (see (7.44)), and, second, the overshoot was bounded by 0 (see (7.45). The effect of the second threshold can be taken into account by recognizing that $\Pr\{S_J \leq \beta\} = 1 - \Pr\{S_J \geq \alpha\}$. Then (7.43) can be solved, getting

$$\Pr\{S_J \geq \alpha\} = \frac{1 - \mathsf{E}\left[\exp(r^*S_J) \mid S_J{\leq}\beta\right]}{\mathsf{E}\left[\exp(r^*S_J) \mid S_J{\geq}\alpha\right] - \mathsf{E}\left[\exp(r^*S_J) \mid S_J{\leq}\beta\right]}. \tag{7.47}$$

Solving for the terms on the right side of (7.47) usually requires analyzing the overshoot upon crossing a barrier, and this is often difficult. For the case of the simple random walk, overshoots don't occur, since the random walk changes only in unit steps. Thus, for α and β integers, we have $\mathsf{E}\left[\exp(r^*S_J) \mid S_J{\leq}\beta\right] = \exp(r^*\beta)$ and $\mathsf{E}\left[\exp(r^*S_J) \mid S_J{\geq}\alpha\right] = \exp(r^*\alpha)$. Substituting this in (7.47) yields the exact solution

$$\Pr\{S_J \geq \alpha\} = \frac{\exp(-r^*\alpha)[1 - \exp(r^*\beta)]}{1 - \exp[-r^*(\alpha - \beta)]}. \tag{7.48}$$

Solving the equation $\gamma(r^*) = 0$ for the simple random walk with probabilities p and q yields $r^* = \ln(q/p)$. This is also valid if X takes on the three values -1, 0, and $+1$ with $p = \Pr\{X = 1\}$, $q = \Pr\{X = -1\}$, and $1 - p - q = \Pr\{X = 0\}$. It can be seen that if α and $-\beta$ are large positive integers, then the simple bound of (7.42) is almost exact for this example.

Equation (7.48) is sometimes used as an approximation for (7.47). Unfortunately, for many applications, the overshoots are more significant than the effect of the opposite threshold. Thus (7.48) is only negligibly better than (7.42) as an approximation, and has the further disadvantage of not being a bound.

If $\Pr\{S_J \geq \alpha\}$ must actually be calculated, then the overshoots in (7.47) must be taken into account. See Chapter 12 of [8] for a treatment of overshoots.

7.5.5 Binary hypotheses testing with IID observations

The objective of this subsection and the next is to understand how to make binary decisions on the basis of a variable number of observations, choosing to stop observing additional data when a sufficiently good decision can be made. This initial subsection lays the groundwork for this by analyzing the large deviation aspects of binary detection with a large but fixed number of IID observations.

Consider the binary hypothesis testing problem of Section 7.3 in which H is a binary hypothesis with apriori probabilities $\mathsf{p}_H(0) = p_0$ and $\mathsf{p}_H(1) = p_1$. The observation $Y_1, Y_2, \dots,$ conditional on $H = 0$, is a sequence of IID rv's with the probability density $\mathsf{f}_{Y|H}(y \mid 0)$. Conditional on $H = 1$, the observations are IID with density $\mathsf{f}_{Y|H}(y \mid 1)$. For any given number n of sample observations, $y_1, \dots, y_n$, the likelihood ratio was defined as

$$\Lambda_n(\boldsymbol{y}) = \prod_{i=1}^{n} \frac{\mathsf{f}_{Y|H}(y_n \mid 0)}{\mathsf{f}_{Y|H}(y_n \mid 1)}$$

and the log-likelihood-ratio was defined as

$$s_n = \sum_{i=1}^{n} z_n; \qquad \text{where } z_n = \ln \frac{\mathsf{f}_{Y|H}(y_n \mid 0)}{\mathsf{f}_{Y|H}(y_n \mid 1)}. \tag{7.49}$$

The MAP test gives the maximum aposteriori probability of correct decision based on the n observations $(y_1, \dots, y_n)$. It is defined as the following threshold test:

$$s_n \begin{cases} > \ln(p_1/p_0) \quad ; & \text{select } \hat{h}{=}0 \\ \le \ln(p_1/p_0) \quad ; & \text{select } \hat{h}{=}1. \end{cases}$$

We can use the Chernoff bound to get an exponentially tight bound on the probability of error using the the MAP test and given $H = 1$ (and similarly given $H = 0$). That is, $\Pr\{e \mid H = 1\}$ is simply $\Pr\{S_n \ge \ln(p_1/p_0) \mid H{=}1\}$ where S_n is the rv whose sample value is given by (7.49), *i.e.*, $S_n = \sum_{i=1}^{n} Z_n$ where $Z_n = \ln[f_{Y|H}(Y \mid 0)/f_{Y|H}(Y \mid 1)]$. The semi-invariant MGF $\gamma_1(r)$ of Z given $H = 1$ is

$$\begin{aligned} \gamma_1(r) &= \ln \int_y \mathsf{f}_{Y|H}(y \mid 1) \exp\left\{ r \left[\ln \frac{\mathsf{f}_{Y|H}(y \mid 0)}{\mathsf{f}_{Y|H}(y \mid 1)} \right] \right\} dy \\ &= \ln \int_y [\mathsf{f}_{Y|H}(y \mid 1)]^{1-r} [\mathsf{f}_{Y|H}(y \mid 0)]^r \, dy \end{aligned} \tag{7.50}$$

Surprisingly, we see that $\gamma_1(1) = \ln \int_y \mathsf{f}_{Y|H}(y \mid 0)\, dy = 0$, so that $r^* = 1$ for $\gamma_1(r)$. Since $\gamma_1(r)$ has a positive second derivative, this shows that $\gamma_1'(0) < 0$ and $\gamma_1'(1) > 0$. Figure 7.10 illustrates the optimized exponent in the Chernoff bound,

$$\Pr\{e \mid H{=}1\} \le \exp\left\{ n \left[[\min_r \gamma_1(r) - ra \right] \right\} \qquad \text{where } a = \frac{1}{n} \ln(p_1/p_0)$$

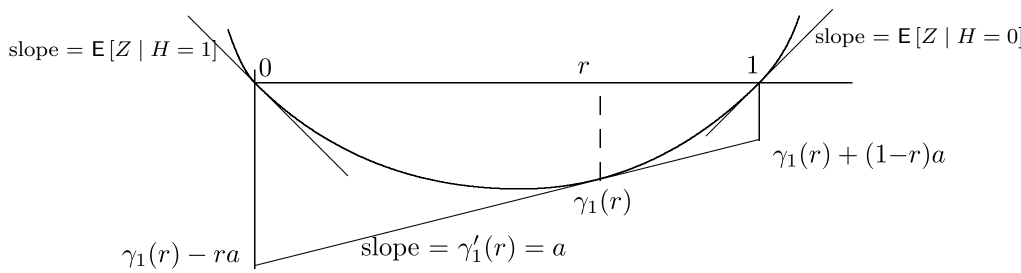


Figure 7.10: Graphical description of the exponents in the optimized Chernoff bounds for $\Pr\{e \mid H{=}1\} \leq \exp\{n[\gamma_1(r) - ra]\}$ and $\Pr\{e \mid H{=}0\} \leq \exp\{n[\gamma_1(r) + (1-r)a]\}$. The slope a is equal to $\frac{1}{n}\ln(p_1/p_0)$. For an arbitrary threshold test with threshold η, the slope a is $\frac{1}{n}\ln\eta$.

We can find the Chernoff bound for $\Pr\{e \mid H{=}0\}$ in the same way. The semi-invariant MGF for Z conditional on $H = 0$ is given by

$$\gamma_0(r) = \ln \int_y [\mathsf{f}_{Y|H}(y \mid 1)]^{-r} [\mathsf{f}_{Y|H}(y \mid 0)]^{1+r}\, dy = \gamma_1(r{-}1) \tag{7.51}$$

We can see from this that the distribution of Z conditional on $H = 0$ is the tilted distribution (at $r = 1$) of Z conditional on $H = 1$. Since $\gamma_0(r) = \gamma_1(r{-}1)$, the optimized Chernoff bound for $\Pr\{e \mid H{=}0\}$ is given by

$$\Pr\{e \mid H{=}0\} \leq \exp\left\{n\left[[\min_r \gamma_1(r) + (1{-}r)a\right]\right\} \qquad \text{where } a = \frac{1}{n}\ln(p_1/p_0)$$

In general, the threshold p_1/p_0 can be replaced by an arbitrary threshold η. As illustrated in Figure 7.10, the slope $\gamma_1'(r) = \frac{1}{n}\ln(\eta)$ at which the optimized bound occurs increases with η, varying from $\gamma'(0) = \mathsf{E}[Z \mid H{=}1]$ at $r = 0$ to $\gamma'(1) = \mathsf{E}[Z \mid H{=}0]$ at $r = 1$. The tradeoff between the two error exponents are seen to vary as the two ends of an inverted see-saw. One could in principle achieve a still larger magnitude of exponent for $\Pr\{()\, e \mid H{=}1)$ by using $r > 1$, but this would be somewhat silly since $\Pr\{e \mid H{=}0\}$ would then be very close to 1 and it would usually make more sense to simply decide on $H = 1$ without looking at the observed data at all.

We can view the tradeoff of exponents above as a large deviation form of the Neyman-Pearson criterion. That is, rather than fixing the allowable value of $\Pr\{e \mid H{=}1\}$ and choosing a test to minimize $\Pr\{e \mid H{=}0\}$, we fix the allowable exponent for $\Pr\{e \mid H{=}1\}$ and then optimize the exponent for $\Pr\{e \mid H{=}1\}$. This is essentially the same as the Neyman-Pearson test, and is a threshold test as before, but it allows us to ignore the exact solution of the Neyman-Pearson problem and focus on the exponentially tight exponents.

7.5.6 Sequential decisions for binary hypotheses

Common sense tells us that it should be valuable to have the ability to make additional observations when the current observations do not lead to a clear choice. With such an ability, we have 3 possible choices at the end of each observation: choose $\hat{H} = 1$, choose $\hat{H} = 0$, or continue with additional observations. It is evident that we need a stopping time J to decide when to stop. We assume for the time being that stopping occurs with probability 1, both under hypothesis $H = 1$ and $H = 0$, and the indicator rv $\mathbb{I}_{J=n}$ for each $n \geq 1$ is a function of $Z_1, \dots, Z_n$.

Given $H = 1$, the sequence of LLR's $\{S_n; n \geq 1\}$ (where $S_n = \sum_{i=1}^n Z_i$) is a random walk. Assume for now that the stopping rule is to stop when the random walk crosses either a threshold[15] at $\alpha > 0$ or a threshold at $\beta < 0$. Given $H = 0$, $S_n = \sum_{i=1}^n Z_n$ is also a random walk (with a different probability measure). The same stopping rule must be used, since the decision to stop at n can be based only on $Z_1, \dots, Z_n$ and not on knowledge of H. We saw that the random walk based on $H = 0$ uses probabilities tilted from those based on $H = 1$, a fact that is interesting but not necessary here.

We assume that when stopping occurs, the decision $\hat{h} = 0$ is made if $S_J \geq \alpha$ and $\hat{h} = 1$ is made if $S_J \leq \beta$. Given that $H = 1$, an error is then made if $S_J \geq \alpha$. Thus, using the probability distribution for $H = 1$, we apply (7.42), along with $r^* = 1$, to get

$$\Pr\{e \mid H{=}1\} \;=\; \Pr\{S_J \geq \alpha \mid H{=}1\} \;\leq\; \exp{-\alpha} \tag{7.52}$$

Similarly, conditional on $H = 0$, an error is made if the threshold at β is crossed before that at α and the error probability is

$$\Pr\{e \mid H{=}0\} \;=\; \Pr\{S_J \leq \beta \mid H{=}0\} \;\leq\; \exp\beta \tag{7.53}$$

The error probabilities can be made as small as desired by increasing the magnitudes of α and β, but the cost of increasing α is to increase the number of observations required when $H = 0$. From Wald's equality,

$$\mathsf{E}[J \mid H{=}0] \;=\; \frac{\mathsf{E}[S_J \mid H{=}0]}{\mathsf{E}[Z \mid H{=}0]} \;\approx\; \frac{\alpha + \mathsf{E}[\text{overshoot}]}{\mathsf{E}[Z \mid H{=}0]}$$

In the approximation, we have ignored the possibility of S_J crossing the threshold at β conditional on $H = 0$ since this is a very small-probability event when α and β have large magnitudes. Thus we see that the expected number of observations (given $H = 0$) is essentially linear in α.

We next ask what has been gained quantitatively by using the sequential decision procedure here. Suppose we use a fixed-length test with $n = \alpha/\mathsf{E}[Z \mid H{=}0]$. Referring to Figure 7.10, we see that if we choose the slope $a = \gamma'(1)$, then the (exponentially tight) Chernoff bound on $\Pr\{e \mid H{=}1\}$ is given by $e^{-\alpha}$, but the exponent on $\Pr\{e \mid H{=}0\}$ is 0. In other words, by using a sequential test as described here, we simultaneously get the error exponent for

[15] It is unfortunate that the word 'threshold' has a universally accepted meaning for random walks (*i.e.*, the meaning we are using here), and the word 'threshold test' has a universally accepted meaning for hypothesis testing. Stopping when S_n crosses either α or β can be viewed as an extension of an hypothesis testing threshold test in the sense that both the stopping trial and the decision is based only on the LLR, and is based on the LLR crossing either a positive or negative threshold, but the results about threshold tests in Section 7.3 are not necessarily valid for this extension.

$H = 1$ that a fixed test would provide if we gave up entirely on an error exponent for $H = 0$, and vice versa.[16]

A final question to be asked is whether any substantial improvement on this sequential decision procedure would result from letting the thresholds at α and β vary with the number of observations. Assuming that we are concerned only with the expected number of observations, the answer is no. We will not carry this argument out here, but it consists of using the Chernoff bound as a function of the number of observations. This shows that there is a typical number of observations at which most errors occur, and changes in the thresholds elsewhere can increase the error probability, but not substantially decrease it.

7.5.7 Joint distribution of crossing time and barrier

Next we look at $\Pr\{J \geq n,\ S_J \geq \alpha\}$, where again we assume that $\overline{X} < 0$ and that $\gamma(r^*) = 0$ for some $r^* > 0$. For any r in the region where $\gamma(r) \leq 0$ (*i.e.*, $0 \leq r \leq r^*$), we have $-J\gamma(r) \geq -n\gamma(r)$ for $J \geq n$. Thus, from the Wald identity, we have

$$\begin{aligned}
1 &\geq \mathsf{E}\left[\exp[rS_J - J\gamma(r)] \mid J \geq n,\ S_J \geq \alpha\right] \Pr\{J \geq n,\ S_J \geq \alpha\} \\
&\geq \exp[r\alpha - n\gamma(r)] \Pr\{J \geq n, S_J \geq \alpha\} \\
\Pr\{J \geq n, S_J \geq \alpha\} &\leq \exp[-r\alpha + n\gamma(r)]\ ; \quad \text{for } \ r \in [0, r^*].
\end{aligned} \tag{7.54}$$

Since this is vaid for all $r \in (0, r^*]$, we can obtain the tightest bound of this form by minimizing the right hand side of (7.54). This is the same minimization (except for the constraint $r \leq r^*$) as in Figures 7.7. Assume that $r^* < r_+$ (*i.e.*, that the exceptional situation[17] of Figure 7.8) does not occur. Define n^* as $\alpha/\gamma'(r^*)$. The result is then

$$\Pr\{J \geq n, S_J \geq \alpha\} \leq \begin{cases} \exp\left[n\gamma(r_o) - r_o\alpha\right] & \text{for } n > n^*,\ \alpha/n = \gamma'(r_o) \\ \\ \exp\left[-r^*\alpha\right] & n \leq n^*. \end{cases} \tag{7.55}$$

The interpretation of (7.55) is that n^* is an estimate of the typical value of J given that the threshold at α is crossed. For n greater than this typical value, (7.55) provides a tighter bound on $\Pr\{J \geq n, S_J \geq \alpha\}$ than the bound on $\Pr\{S_J \geq \alpha\}$ in (7.42), whereas (7.55) provides nothing new for $n \leq n^*$. In Section 7.8, we shall derive the slightly stronger result that $\Pr\left\{\bigcup_{i \geq n}[S_i \geq \alpha]\right\}$ is also upper bounded by the right hand side of (7.55).

[16] In the communication context, decision rules are used to detect sequentially transmitted data. The use of a sequential decision rule usually requires feedback from receiver to transmitter, and also requires a variable rate of transmission. Thus the substantial reductions in error probability are accompanied by substantial system complexity.

[17] The exceptional situation where $\gamma(r)$ is negative for $r \leq r_+$ and discontinuous at $r^* = r_+$, *i.e.*, the situation in Figure 7.8), will be treated in the exercises, and is quite different from the case here. In particular, as can be seen from the figure, the optimized Chernoff bound on $\Pr\{S_n \geq \alpha\}$ is optimized at $n = 1$.

An almost identical upper bound to $\Pr\{J \le n,\ S_J \ge \alpha\}$ can be found (again assuming that $r^* < r_+$) by using the Wald identity for $r > r^*$. Here $\gamma(r) > 0$, so $-J\gamma(r) \ge -n\gamma(r)$ for $J \le n$. The result is

$$\Pr\{J \le n, S_J \ge \alpha\} \le \begin{cases} \exp\left[n\gamma(r_o) - r_o\alpha\right] & \text{for } n < n^*,\ \alpha/n = \gamma'(r_o) \\ \\ \exp\left[-r^*\alpha\right] & n \ge n^*. \end{cases} \tag{7.56}$$

This strengthens the interpretation of n^* as the typical value of J conditional on crossing the threshold at α. That is, (7.56) provides information on the lower tail of the distribution of J (conditional on $S_J \ge \alpha$), whereas (7.55) provides information on the upper tail.

7.6 Martingales

A martingale is defined as an integer-time stochastic process $\{Z_n;\ n \ge 1\}$ with the properties that $\mathsf{E}\left[|Z_n|\right] < \infty$ for all $n \ge 1$ and

$$\mathsf{E}\left[Z_n \mid Z_{n-1}, Z_{n-2}, \dots, Z_1\right] = Z_{n-1}; \qquad \text{for all } n \ge 2. \tag{7.57}$$

The name martingale comes from gambling terminology where martingales refer to gambling strategies in which the amount to be bet is determined by the past history of winning or losing. If one visualizes Z_n as representing the gambler's fortune at the end of the n^{th} play, the definition above means, first, that the game is fair (in the sense that the expected increase in fortune from play $n-1$ to n is zero), and, second, that the expected fortune on the n^{th} play depends on the past only through the fortune on play $n-1$.

The important part of the definition of a martingale, and what distinguishes martingales from other kinds of proceesses, is the form of dependence in (7.57). However, the restriction that $\mathsf{E}\left[|Z_n|\right] < \infty$ is also important, particularly since martingales are so abstract and general that one often loses the insight to understand intuitively when this restriction is important. Students are advised to ignore this restriction when first looking at something that might be a martingale, and to check later after acquiring some understanding.

There are two interpretations of (7.57); the first and most straightforward is to view it as shorthand for $\mathsf{E}\left[Z_n \mid Z_{n-1}{=}z_{n-1},\ Z_{n-2}{=}z_{n-2},\ \dots,\ Z_1{=}z_1\right] = z_{n-1}$ for all possible sample values $z_1, z_2, \dots, z_{n-1}$. The second is that $\mathsf{E}\left[Z_n \mid Z_{n-1}{=}z_{n-1},\ \dots,\ Z_1{=}z_1\right]$ is a function of the sample values $z_1, \dots, z_{n-1}$ and thus $\mathsf{E}\left[Z_n \mid Z_{n-1}, \dots, Z_1\right]$ is a random variable which is a function of the random variables $Z_1, \dots, Z_{n-1}$ (and, for a martingale, a function only of Z_{n-1}). Students are encouraged to take the first viewpoint initially and to write out the expanded type of expression in cases of confusion. The second viewpoint, however, is very powerful, and, with experience, is the more useful viewpoint.

It is important to understand the difference between martingales and Markov chains. For the Markov chain $\{X_n;\ n \ge 1\}$, each rv X_n is conditioned on the past only through X_{n-1}, whereas for the martingale $\{Z_n;\, n \geq 1\}$, it is only the expected value of Z_n that is conditioned on the past only through Z_{n-1}. The rv Z_n itself, conditioned on Z_{n-1}, can also be dependent on all the earlier Z_i's. It is very surprising that so many results can be developed using such a weak form of conditioning.

In what follows, we give a number of important examples of martingales, then develop some results about martingales, and then discuss those results in the context of the examples.

7.6.1 Simple examples of martingales

Example 7.6.1 (Random walks). One example of a martingale is a zero-mean random walk, since if $Z_n = X_1 + X_2 + \cdots + X_n$, where the X_i are IID and zero mean, then

$$\mathsf{E}[Z_n \mid Z_{n-1}, \dots, Z_1] = \mathsf{E}[X_n + Z_{n-1} \mid Z_{n-1}, \dots, Z_1] \tag{7.58}$$

$$= \mathsf{E}[X_n] + Z_{n-1} = Z_{n-1}. \tag{7.59}$$

Extending this example, suppose that $\{X_i;\, i \geq 1\}$ is an arbitrary sequence of IID random variables with mean $\overline{X}$ and let $\widetilde{X}_i = X_i - \overline{X}$. Then $\{S_n;\, n \geq 1\}$ is a random walk with $S_n = X_1 + \cdots + X_n$ and $\{Z_n;\, n \geq 1\}$ is a martingale with $Z_n = \widetilde{X}_1 + \cdots + \widetilde{X}_n$. The random walk and the martingale are simply related by $Z_n = S_n - n\overline{X}$, and thus general results about martingales can easily be applied to arbitrary random walks.

Example 7.6.2 (Sums of dependent zero-mean variables). Let $\{X_i;\, i \geq 1\}$ be a set of dependent random variables satisfying $\mathsf{E}[X_i \mid X_{i-1}, \dots, X_1] = 0$. Then $\{Z_n;\, n \geq 1\}$, where $Z_n = X_1 + \cdots + X_n$, is a zero mean martingale. To see this, note that

$$\begin{aligned} \mathsf{E}[Z_n \mid Z_{n-1}, \dots, Z_1] &= \mathsf{E}[X_n + Z_{n-1} \mid Z_{n-1}, \dots, Z_1] \\ &= \mathsf{E}[X_n \mid X_{n-1}, \dots, X_1] + \mathsf{E}[Z_{n-1} \mid Z_{n-1}, \dots, Z_1] = Z_{n-1}. \end{aligned}$$

This is a more general example than it appears, since given any martingale $\{Z_n;\, n \geq 1\}$, we can define $X_n = Z_n - Z_{n-1}$ for $n \geq 2$ and define $X_1 = Z_1$. Then $\mathsf{E}[X_n \mid X_{n-1}, \dots, X_1] = 0$ for $n \geq 2$. If the martingale is zero mean (*i.e.*, if $\mathsf{E}[Z_1] = 0$), then $\mathsf{E}[X_1] = 0$ also.

Example 7.6.3 (Product-form martingales). Another example is a product of unit mean IID random variables. Thus if $Z_n = X_1 X_2 \dots X_n$, we have

$$\mathsf{E}[Z_n \mid Z_{n-1}, \dots, Z_1] = \mathsf{E}[X_n, Z_{n-1} \mid Z_{n-1}, \dots, Z_1]$$

$$= \mathsf{E}[X_n]\,\mathsf{E}[Z_{n-1} \mid Z_{n-1}, \dots, Z_1] \tag{7.60}$$

$$= \mathsf{E}[X_n]\,\mathsf{E}[Z_{n-1} \mid Z_{n-1}] = Z_{n-1}. \tag{7.61}$$

A particularly simple case of this product example is where $X_n = 2$ with probability 1/2 and $X_n = 0$ with probability 1/2. Then

$$\Pr\{Z_n = 2^n\} = 2^{-n}; \qquad \Pr\{Z_n = 0\} = 1 - 2^{-n}; \qquad \mathsf{E}[Z_n] = 1. \tag{7.62}$$

Thus $\lim_{n\to\infty} Z_n = 0$ with probability 1, but $\mathsf{E}[Z_n] = 1$ for all n and $\lim_{n\to\infty} \mathsf{E}[Z_n] = 1$. This is an important example to keep in mind when trying to understand why proofs about martingales are necessary and non-trivial. This type of phenomenon will be clarified somewhat by Lemma 7.8.1 when we discussed stopped martingales in Section 7.8.

An important example of a product-form martingale is as follows: let $\{X_i;\, i \geq 1\}$ be an IID sequence, and let $\{S_n = X_1 + \cdots + X_n;\, n \geq 1\}$ be a random walk. Assume that the semi-invariant generating function $\gamma(r) = \ln\{\mathsf{E}[\exp(rX)]\}$ exists in some region of r around 0. For each $n \geq 1$, let Z_n be defined as

$$\begin{aligned} Z_n &= \exp\{rS_n - n\gamma(r)\} && (7.63)\\ &= \exp\{rX_n - \gamma(r)\}\exp\{rS_{n-1} - (n-1)\gamma(r)\} \\ &= \exp\{rX_n - \gamma(r)\}Z_{n-1}. && (7.64)\end{aligned}$$

Taking the conditional expectation of this,

$$\begin{aligned} \mathsf{E}[Z_n \mid Z_{n-1}, \dots, Z_1] &= \mathsf{E}[\exp(rX_n - \gamma(r))]\,\mathsf{E}[Z_{n-1} \mid Z_{n-1}, \dots, Z_1] \\ &= Z_{n-1}. && (7.65)\end{aligned}$$

where we have used the fact that $\mathsf{E}[\exp(rX_n)] = \exp(\gamma(r))$. Thus we see that $\{Z_n;\, n \geq 1\}$ is a martingale of the product-form.

7.6.2 Scaled branching processes

A final simple example of a martingale is a "scaled down" version of a branching process $\{X_n;\, n \geq 0\}$. Recall from Section 5.5 that, for each n, X_n is defined as the aggregate number of elements in generation n. Each element i of generation n, $1 \leq i \leq X_n$ has a number $Y_{i,n}$ of offspring which collectively constitute generation $n+1$, *i.e.*, $X_{n+1} = \sum_{i=1}^{X_n} Y_{i,n}$. The rv's $Y_{i,n}$ are IID over both i and n.

Let $\overline{Y} = \mathsf{E}[Y_{i,n}]$ be the mean number of offspring of each element of the population. Then $\mathsf{E}[X_n \mid X_{n-1}] = \overline{Y}X_{n-1}$, which resembles a martingale except for the factor of $\overline{Y}$. We can convert this branching process into a martingale by scaling it, however. That is, define $Z_n = X_n/\overline{Y}^n$. It follows that

$$\mathsf{E}[Z_n \mid Z_{n-1}, \dots, Z_1] = \mathsf{E}\left[\frac{X_n}{\overline{Y}^n} \mid X_{n-1}, \dots, X_1\right] = \frac{\overline{Y}X_{n-1}}{\overline{Y}^n} = Z_{n-1}. \qquad (7.66)$$

Thus $\{Z_n;\, n \geq 1\}$ is a martingale. We will see the surprising result later that this implies that Z_n converges with probability 1 to a limiting rv as $n \to \infty$.

7.6.3 Partial isolation of past and future in martingales

Recall that for a Markov chain, the states at all times greater than a given n are independent of the states at all times less than n, conditional on the state at time n. The following lemma shows that at least a small part of this independence of past and future applies to martingales.

Lemma 7.6.1. *Let $\{Z_n;\, n \geq 1\}$ be a martingale. Then for any $n > i \geq 1$,*

$$\mathsf{E}\left[Z_n \mid Z_i, Z_{i-1}, \dots, Z_1\right] = Z_i. \tag{7.67}$$

Proof: For $n = i+1$, $\mathsf{E}\left[Z_{i+1} \mid Z_i, \dots, Z_1\right] = Z_i$ by the definition of a martingale. Now consider $n = i+2$. Then $\mathsf{E}\left[Z_{i+2} \mid Z_{i+1}, \dots, Z_1\right]$ is a rv equal to Z_{i+1}. We can view $\mathsf{E}\left[Z_{i+2} \mid Z_i, \dots, Z_1\right]$ as the conditional expectation of the rv $\mathsf{E}\left[Z_{i+2} \mid Z_{i+1}, \dots, Z_1\right]$ over Z_{i+1}, conditional on $Z_i, \dots, Z_1$.

$$\begin{aligned} \mathsf{E}\left[Z_{i+2} | Z_i, \dots, Z_1\right] &= \mathsf{E}\left[\mathsf{E}\left[Z_{i+2} \mid Z_{i+1}, Z_i, \dots, Z_1\right] \mid Z_i, \dots, Z_1\right] \\ &= \mathsf{E}\left[Z_{i+1} \mid Z_i, \dots, Z_1\right] = Z_i. \end{aligned} \tag{7.68}$$

For $n = i+3$, (7.68), with i incremented, shows us that the rv $\mathsf{E}\left[Z_{i+3} \mid Z_{i+1}, \dots, Z_1\right] = Z_{i+1}$. Taking the conditional expectation of this rv over Z_{i+1} conditional on $Z_i, \dots, Z_1$, we get

$$\mathsf{E}\left[Z_{i+3} \mid Z_i, \dots, Z_1\right] = Z_i.$$

This argument can be applied successively to any $n > i$. □

This lemma is particularly important for $i = 1$, where it says that $\mathsf{E}\left[Z_n \mid Z_{n-1}, \dots, Z_1\right] = Z_1$. The left side of this is a rv which is a function (in fact the identity function) of Z_1. Thus, by taking the expected value of each side, we see that

$$\mathsf{E}\left[Z_n\right] = \mathsf{E}\left[Z_1\right] \qquad \text{for all } n > 1. \tag{7.69}$$

An important application of this is to the product form martingale in (7.65). This says that

$$\begin{aligned} \mathsf{E}\left[\exp(rS_n - n\gamma(r)\right] &= \mathsf{E}\left[\exp(rX - \gamma(r))\right] \\ &= \mathsf{E}\left[\exp(rX)\right]/g(r) \;=\; 1. \end{aligned} \tag{7.70}$$

We will come back later to relate this to Wald's identity.

7.7 Submartingales and supermartingales

Submartingales and supermartingales are simple generalizations of martingales that provide many useful results for very little additional work. We will subsequently derive the Kolmogorov submartingale inequality, which is a powerful generalization of the Markov inequality. We use this both to give a simple proof of the strong law of large numbers and also to better understand threshold crossing problems for random walks.

Definition 7.7.1. *A submartingale is an integer-time stochastic process $\{Z_n;\, n \geq 1\}$ that satisfies the relations*

$$\mathsf{E}\left[|Z_n|\right] < \infty \quad ; \quad \mathsf{E}\left[Z_n \mid Z_{n-1}, Z_{n-2}, \dots, Z_1\right] \geq Z_{n-1} \quad ; \quad n \geq 1. \tag{7.71}$$

A supermartingale is an integer-time stochastic process $\{Z_n;\, n \geq 1\}$ that satisfies the relations

$$\mathsf{E}\left[|Z_n|\right] < \infty \quad ; \quad \mathsf{E}\left[Z_n \mid Z_{n-1}, Z_{n-2}, \dots, Z_1\right] \leq Z_{n-1} \quad ; \quad n \geq 1. \tag{7.72}$$

In terms of our gambling analogy, a submartingale corresponds to a game that is at least fair, *i.e.*, where the expected fortune of the gambler either increases or remains the same. A *supermartingale* is a process with the opposite type of inequality.[18]

Since a martingale satisfies both (7.71) and (7.72) with equality, a martingale is both a submartingale and a supermartingale. Note that if $\{Z_n;\, n \geq 1\}$ is a submartingale, then $\{-Z_n;\, n \geq 1\}$ is a supermartingale, and conversely. Thus, some of the results to follow are stated only for submartingales, with the understanding that they can be applied to supermartingales by changing signs as above.

Lemma 7.6.1, with the equality replaced by inequality, also applies to submartingales and supermartingales. That is, if $\{Z_n;\, n \geq 1\}$ is a submartingale, then

$$\mathsf{E}\left[Z_n \mid Z_i, Z_{i-1}, \dots, Z_1\right] \geq Z_i \quad ; \quad 1 \leq i < n, \tag{7.73}$$

and if $\{Z_n;\, n \geq 1\}$ is a supermartingale, then

$$\mathsf{E}\left[Z_n \mid Z_i, Z_{i-1}, \dots, Z_1\right] \leq Z_i \quad ; \quad 1 \leq i < n. \tag{7.74}$$

Equations (7.73) and (7.74) are verified in the same way as Lemma 7.6.1 (see Exercise 7.18). Similarly, the appropriate generalization of (7.69) is that if $\{Z_n;\, n \geq 1\}$ is a submartingale, then

$$\mathsf{E}\left[Z_n\right] \geq \mathsf{E}\left[Z_i\right] \quad ; \quad \text{for all } i,\ 1 \leq i < n. \tag{7.75}$$

and if $\{Z_n;\, n \geq 1\}$ is a supermartingale, then

$$\mathsf{E}\left[Z_n\right] \leq \mathsf{E}\left[Z_i\right] \quad ; \quad \text{for all } i,\ 1 \leq i < n. \tag{7.76}$$

A random walk $\{S_n;\, n \geq 1\}$ with $S_n = X_1 + \cdots + X_n$ is a submartingale, martingale, or supermartingale respectively for $\overline{X} \geq 0$, $\overline{X} = 0$, or $\overline{X} \leq 0$. Also, if X has a semi-invariant moment generating function $\gamma(r)$ for some given r, and if Z_n is defined as $Z_n = \exp(rS_n)$, then the process $\{Z_n;\, n \geq 1\}$ is a submartingale, martingale, or supermartingale respectively for $\gamma(r) \geq 0$, $\gamma(r) = 0$, or $\gamma(r) \leq 0$. The next example gives an important way in which martingales and submartingales are related.

Example 7.7.1 (Convex functions of martingales). Figure 7.11 illustrates the graph of a convex function h of a real variable x. A function h of a real variable is defined to be *convex* if, for each point x_1, there is a real number c with the property that $h(x_1) + c(x - x_1) \leq h(x)$ for all x.

Geometrically, this says that every tangent to $h(x)$ lies on or below $h(x)$. If $h(x)$ has a derivative at x_1, then c is the value of that derivative and $h(x_1) + c(x - x_1)$ is the tangent line at x_1. If $h(x)$ has a discontinuous slope at x_1, then there might be many choices for c; for example, $h(x) = |x|$ is convex, and for $x_1 = 0$, one could choose any c in the range -1 to $+1$.

A simple condition that implies convexity is a nonnegative second derivative everywhere. This is not necessary, however, and functions can be convex even when the first derivative

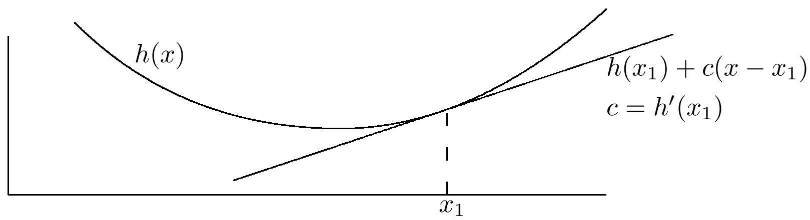


Figure 7.11: Convex functions: For each x_1, there is a value of c such that, for all x, $h(x_1) + c(x - x_1) \leq h(x)$. If h is differentiable at x_1, then c is the derivative of h at x_1.

does not exist everywhere. For example, the function $\gamma(r)$ in Figure 7.8 is convex even though it it is finite at $r = r_+$ and infinite for all $r > r_+$.

Jensen's inequality states that if h is convex and X is a random variable with an expectation, then $h(\mathsf{E}[X]) \leq \mathsf{E}[h(X)]$. To prove this, let $x_1 = \mathsf{E}[X]$ and choose c so that $h(x_1) + c(x - x_1) \leq h(x)$. Using the random variable X in place of x and taking expected values of both sides, we get Jensen's inequality. Note that for any particular event A, this same argument applies to X conditional on A, so that $h(\mathsf{E}[X \mid A]) \leq \mathsf{E}[h(X) \mid A]$. Jensen's inequality is very widely used; it is a minor miracle that we have not required it previously.

Theorem 7.7.1. *Assume that h is a convex function of a real variable, that $\{Z_n;\, n \geq 1\}$ is a martingale or submartingale, and that $\mathsf{E}[|h(Z_n)|] < \infty$ for all n. Then $\{h(Z_n);\, n \geq 1\}$ is a submartingale.*

Proof: For any choice of $z_1, \dots, z_{n-1}$, we can use Jensen's inequality with the conditioning probabilities to get

$$\mathsf{E}[h(Z_n) | Z_{n-1}{=}z_{n-1}, \dots, Z_1{=}z_1] \geq h(\mathsf{E}[Z_n \mid Z_{n-1}{=}z_{n-1}, \dots, Z_1{=}z_1]) = h(z_{n-1}). \quad (7.77)$$

For any choice of numbers $h_1, \dots, h_{n-1}$ in the range of the function h, let $z_1, \dots, z_{n-1}$ be arbitrary numbers satisfying $h(z_1){=}h_1, \dots, h(z_{n-1}){=}h_{n-1}$. For each such choice, (7.77) holds, so that

$$\begin{aligned} \mathsf{E}[h(Z_n) \mid h(Z_{n-1}){=}h_{n-1}, \dots, h(Z_1){=}h_1] &\geq h(\mathsf{E}[Z_n \mid h(Z_{n-1}){=}h_{n-1}, \dots, h(Z_1){=}h_1]) \\ &= h(z_{n-1}) = h_{n-1}. \end{aligned} \quad (7.78)$$

completing the proof. □

Some examples of this result, applied to a martingale $\{Z_n;\, n \geq 1\}$, are as follows:

$$\{|Z_n|;\, n \geq 1\} \text{ is a submartingale} \quad (7.79)$$

$$\{Z_n^2;\, n \geq 1\} \text{ is a submartingale if } \mathsf{E}\left[Z_n^2\right] < \infty \quad (7.80)$$

$$\{\exp(rZ_n);\, n \geq 1\} \text{ is a submartingale for } r \text{ such that } \mathsf{E}[\exp(rZ_n)] < \infty. \quad (7.81)$$

[18]The easiest way to remember the difference between a submartingale and a supermartingale is to remember that it is contrary to what common sense would dictate.

A function of a real variable $h(x)$ is defined to be concave if $-h(x)$ is convex. It then follows from Theorem 7.7.1 that if h is concave and $\{Z_n;\, n \geq 1\}$ is a martingale, then $\{h(Z_n);\, n \geq 1\}$ is a supermartingale (assuming that $\mathsf{E}[|h(Z_n)|] < \infty$). For example, if $\{Z_n;\, n \geq 1\}$ is a positive martingale and $\mathsf{E}[|\ln(Z_n)|] < \infty$, then $\{\ln(Z_n);\, n \geq 1\}$ is a supermartingale.

7.8 Stopped processes and stopping trials

The definition of stopping trials in Section 4.5 applies to arbitrary integer-time processes $\{Z_n;\, n \geq 1\}$ as well as to IID sequences. Recall that J is a stopping trial for a sequence $\{Z_n;\, n \geq 1\}$ of rv's if $\mathbb{I}_{J=n}$ is a function of $Z_1, \dots, Z_n$ and if J is a rv.

If $\mathbb{I}_{J=n}$ is a function of $Z_1, \dots, Z_n$ and J is a defective rv, then J is called a defective stopping trial. For some of the results to follow, it is unimportant whether J is a random variable or a defective random variable (*i.e.*, whether or not the process stops with probability 1). If it is not specified whether J is a random variable or a defective random variable, we refer to the stopping trial as a *possibly-defective stopping trial*; we consider J to take on the value ∞ if the process does not stop.

Definition 7.8.1. *A stopped process $\{Z_n^*;\, n \geq 1\}$ for a possibly-defective stopping trial J relative to a process $\{Z_n;\, n \geq 1\}$ is the process for which $Z_n^* = Z_n$ for $n \leq J$ and $Z_n^* = Z_J$ for $n > J$.*

As an example, suppose Z_n models the fortune of a gambler at the completion of the nth trial of some game, and suppose the gambler then modifies the game by deciding to stop gambling under some given circumstances (*i.e.*, at the stopping trial). Thus, after stopping, the fortune remains constant, so the stopped process models the gambler's fortune in time, including the effect of the stopping trial.

As another example, consider a random walk with a positive and negative threshold, and consider the process to stop after reaching or crossing a threshold. The stopped process then stays at that point beyond the threshold as an artifice to simplify analysis. The use of stopped processes is similar to the artifice that we employed in Section 3.5 for first-passage times in Markov chains; recall that we added an artificial trapping state after the desired passage to simplify analysis.

We next show that the possibly-defective stopped process of a martingale is itself a martingale; the intuitive reason is that, before stopping, the stopped process is the same as the martingale, and, after stopping, $Z_n^* = Z_{n-1}^*$. The following theorem establishes this and the corresponding results for submartingales and supermartingales.

Theorem 7.8.1. *Given a stochastic process $\{Z_n;\, n \geq 1\}$ and a possibly-defective stopping trial J for the process, the stopped process $\{Z_n^*;\, n \geq 1\}$ is a submartingale if $\{Z_n;\, n \geq 1\}$ is a submartingale, is a martingale if $\{Z_n;\, n \geq 1\}$ is a martingale, and is a supermartingale if $\{Z_n;\, n \geq 1\}$ is a supermartingale.*

Proof: First we show that, for all three cases, the stopped process satisfies $\mathsf{E}[|Z_n^*|] < \infty$ for any given $n \geq 1$. Conditional on $J = i$ for some $i < n$, we have $Z_n^* = Z_i$, so

$$\mathsf{E}[|Z_n^*| \mid J = i] = \mathsf{E}[|Z_i| \mid J = i] < \infty \quad \text{for each } i < n \text{ such that } \Pr\{J = i\} > 0.$$

The reason for this is that if $\mathsf{E}[|Z_i| \mid J = i] = \infty$ and $\Pr\{J = i\} > 0$, then $\mathsf{E}[|Z_i|] = \infty$, contrary to the assumption that $\{Z_n;\, n \geq 1\}$ is a martingale, submartingale, or supermartingale. Similarly, for $J \geq n$, we have $Z_n^* = Z_n$ so

$$\mathsf{E}[|Z_n^*| \mid J \geq n] = \mathsf{E}[|Z_n| \mid J \geq n] < \infty \qquad \text{if } \Pr\{J \geq n\} > 0.$$

Averaging,

$$\mathsf{E}[|Z_n^*|] = \sum_{i=1}^{n-1} \mathsf{E}[|Z_n^*| \mid J{=}i] \Pr\{J{=}i\} + \mathsf{E}[|Z_n^*| \mid J \geq n] \Pr\{J \geq n\} < \infty.$$

Next assume that $\{Z_n;\, n \geq 1\}$ is a submartingale. For any given $n > 1$, consider an arbitrary initial sample sequence $(Z_1 = z_1,\, Z_2 = z_2,\, \dots,\, Z_{n-1} = z_{n-1})$. Note that z_1 specifies whether or not $J = 1$. Similarly, (z_1, z_2) specifies whether or not $J = 2$, and so forth up to $(z_1, \dots, z_{n-1})$, which specifies whether or not $J = n-1$. Thus $(z_1, \dots, z_{n-1})$ specifies the sample value of J for $J \leq n-1$ and specifies that $J \geq n$ otherwise.

For $(z_1, \dots, z_{n-1})$ such that $\mathbb{I}_{J \geq n} < n$, we have $z_n^* = z_{n-1}^*$. For all such sample values,

$$\mathsf{E}\left[Z_n^* | Z_{n-1}^*{=}z_{n-1}^*, \dots, Z_1^*{=}z_1^*\right] = z_{n-1}^*. \tag{7.82}$$

For the remaining case, where $(z_1, \dots, z_{n-1})$ is such that $\mathbb{I}_{J \geq n} \geq n$, we have $z_n^* = z_n$. Thus

$$\mathsf{E}\left[Z_n^* | Z_{n-1}^*{=}z_{n-1}^*, \dots, Z_1^*{=}z_1^*\right] \geq z_{n-1}^*. \tag{7.83}$$

The same argument works for martingales and supermartingales by replacing the inequality in (7.83) by equality for the martingale case and the opposite inequality for the supermartingale case. □

Theorem 7.8.2. Given a stochastic process $\{Z_n;\, n \geq 1\}$ and a possibly-defective stopping trial J for the process, the stopped process $\{Z_n^*;\, n \geq 1\}$ satisfies the following conditions for all $n \geq 1$ if $\{Z_n;\, n \geq 1\}$ is a submartingale, martingale, or supermartingale respectively:

$$\mathsf{E}[Z_1] \;\leq\; \mathsf{E}[Z_n^*] \;\leq\; \mathsf{E}[Z_n] \qquad \text{(submartingale)} \tag{7.84}$$

$$\mathsf{E}[Z_1] \;=\; \mathsf{E}[Z_n^*] \;=\; \mathsf{E}[Z_n] \qquad \text{(martingale)} \tag{7.85}$$

$$\mathsf{E}[Z_1] \;\geq\; \mathsf{E}[Z_n^*] \;\geq\; \mathsf{E}[Z_n] \qquad \text{(supermartingale).} \tag{7.86}$$

Proof: Since a process cannot stop before epoch 1, $Z_1 = Z_1^*$ in all cases. First consider the case in which $\{Z_n;\, n \geq 1\}$ is a submartingale. Theorem 7.8.1 shows that $\{Z_n^*;\, n \geq 1\}$ is a submartingale, and from (7.75), $\mathsf{E}[Z_1] \leq \mathsf{E}[Z_n^*]$ for all $n \geq 1$. This establishes the first half of (7.84) and we next prove the second half. First condition on the set of sequences for which the initial segment $(z_1, \dots, z_i)$ specifies that $\mathbb{I}_{J \geq n} = i$ for any given $i < n$. Then $\mathsf{E}[Z_n^*] = z_i$. From (7.73), $\mathsf{E}[Z_n] \geq z_i$, proving (7.84) for this case. For those sequences not having such an initial segment, $Z_n^* = Z_n$, establishing (7.84) in that case. Averaging over these two cases gives (7.84) in general.

Finally, if $\{Z_n;\, n \geq 1\}$ is a supermartingale, then $\{-Z_n;\, n \geq 1\}$ is a submartingale, verifying (7.86). Since a martingale is both a submartingale and supermartingale, (7.85) follows and the proof is complete. □

Consider a (non-defective) stopping trial J for a martingale $\{Z_n;\, n \geq 1\}$. Since the stopped process is also a martingale, we have

$$\mathsf{E}[Z_n^*] = \mathsf{E}[Z_1^*] = \mathsf{E}[Z_1]\,;\; n \geq 1. \tag{7.87}$$

Since $Z_n^* = Z_J$ for all $n \geq J$ and since J is finite with probability 1, we see that $\lim_{n\to\infty} Z_n^* = Z_J$ with probability 1. Unfortunately, in general, $\mathsf{E}[Z_J]$ is unequal to $\lim_{n\to\infty} \mathsf{E}[Z_n^*] = \mathsf{E}[Z_1]$. An example in which this occurs is the binary product martingale in (7.62). Taking the stopping trial J to be the smallest n for which $Z_n = 0$, we have $Z_J = 0$ with probability 1, and thus $\mathsf{E}[Z_J] = 0$. But $Z_n^* = Z_n$ for all n, and $\mathsf{E}[Z_n^*] = 1$ for all n. The problem here is that, given that the process has not stopped by time n, Z_n and Z_n^* each have the value 2^n. Fortunately, in most situations, this type of bizarre behavior does not occur and $\mathsf{E}[Z_J] = \mathsf{E}[Z_1]$. To get a better understanding of when $\mathsf{E}[Z_J] = \mathsf{E}[Z_1]$, note that for any n, we have

$$\begin{aligned} \mathsf{E}[Z_n^*] &= \sum_{i=1}^{n} \mathsf{E}[Z_n^* \mid J = i] \Pr\{J = i\} + \mathsf{E}[Z_n^* \mid J > n] \Pr\{J > n\} && (7.88) \\ &= \sum_{i=1}^{n} \mathsf{E}[Z_J \mid J = i] \Pr\{J = i\} + \mathsf{E}[Z_n \mid J > n] \Pr\{J > n\}. && (7.89) \end{aligned}$$

The left side of this equation is $\mathsf{E}[Z_1]$ for all n. If the final term on the right converges to 0 as $n \to \infty$, then the sum must converge to $\mathsf{E}[Z_1]$. If $\mathsf{E}[|Z_J|] < \infty$, then the sum also converges to $\mathsf{E}[Z_J]$. Without the condition $\mathsf{E}[|Z_J|] < \infty$, the sum might consist of alternating terms which converge, but whose absolute values do not converge, in which case $\mathsf{E}[Z_J]$ does not exist (see Exercise 7.21 for an example). Thus we have established the following lemma.

Lemma 7.8.1. Let J be a stopping trial for a martingale $\{Z_n;\, n \geq 1\}$. Then $\mathsf{E}[Z_J] = \mathsf{E}[Z_1]$ if and only if

$$\lim_{n\to\infty} \mathsf{E}[Z_n \mid J > n] \Pr\{J > n\} = 0 \qquad \text{and} \quad \mathsf{E}[|Z_J|] < \infty. \tag{7.90}$$

Example 7.8.1 (Random walks with thresholds). Recall the generating function product martingale of (7.63) in which $\{Z_n = \exp[rS_n - n\gamma(r)];\, n \geq 1\}$ is a martingale defined in terms of the random walk $\{S_n = X_1 + \cdots + X_n;\, n \geq 1\}$. From (7.85), we have $\mathsf{E}[Z_n] = \mathsf{E}[Z_1]$, and since $\mathsf{E}[Z_1] = \mathsf{E}[\exp\{rX_1 - \gamma(r\}] = 1$, we have $\mathsf{E}[Z_n] = 1$ for all n. Also, for any possibly-defective stopping trial J, we have $\mathsf{E}[Z_n^*] = \mathsf{E}[Z_1] = 1$. If J is a non-defective stopping trial, and if (7.90) holds, then

$$\mathsf{E}[Z_J] = \mathsf{E}[\exp\{rS_J - J\gamma(r)\}] = 1. \tag{7.91}$$

If there are two thresholds, one at $\alpha > 0$, and the other at $\beta < 0$, and the stopping rule is to stop when either threshold is crossed, then (7.91) is just the Wald identity, (7.34).

The nice part about the approach here is that it also applies naturally to other stopping rules. For example, for some given integer n, let J_{n+} be the smallest integer $i \geq n$ for which $S_i \geq \alpha$ or $S_i \leq \beta$. Then, in the limit $\beta \to -\infty$, $\Pr\{S_{J_{n+}} \geq \alpha\} = \Pr\{\cup_{i=n}^{\infty}(S_i \geq \alpha)\}$. Assuming $\overline{X} < 0$, we can find an upper bound to $\Pr\{S_{J_{n+}} \geq \alpha\}$ for any $r > 0$ and $\gamma(r) \leq 0$ (*i.e.*, for $0 < r \leq r^*$) by the following steps

$$\begin{aligned} 1 = \mathsf{E}\left[\exp\{rS_{J_{n+}} - J_{n+}\gamma(r)\}\right] &\geq \Pr\{S_{J_{n+}} \geq \alpha\}\exp[r\alpha - n\gamma(r)] \\ \Pr\{S_{J_{n+}} \geq \alpha\} &\leq \exp[-r\alpha + n\gamma(r)]; \qquad 0 \leq r \leq r^*. \end{aligned} \tag{7.92}$$

This is almost the same result as (7.54), except that it is slightly stronger; (7.54) bounded the probability that the *first* threshold crossing crossed α at some epoch $i \geq n$, whereas this includes the possibility that $S_m \geq \alpha$ and $S_i \geq \alpha$ for some $m < n \leq i$.

7.9 The Kolmogorov inequalities

We now use the previous theorems to establish Kolmogorov's submartingale inequality, which is a major strengthening of the Markov inequality. Just as the Markov inequality in Section 1.7 was used to derive the Chebychev inequality and then the weak law of large numbers, the Kolmogorov submartingale inequality will be used to strengthen the Chebychev inequality, from which the strong law of large numbers will follow.

Theorem 7.9.1 (Kolmogorov's submartingale inequality). *Let $\{Z_n;\, n \geq 1\}$ be a nonnegative submartingale. Then for any positive integer m and any $a > 0$,*

$$\Pr\left\{\max_{1\leq i\leq m} Z_i \geq a\right\} \leq \frac{\mathsf{E}[Z_m]}{a}. \tag{7.93}$$

Proof: Given a nonnegative submartingale $\{Z_n;\, n \geq 1\}$, given $a > 0$, and given a positive integer m, let J be the stopping trial defined as the smallest $n \leq m$ such that $Z_n \geq a$. If $Z_n < a$ for all $n \leq m$, then $J = m$. Thus the process must stop by time m, and $Z_J \geq a$ if and only if $Z_n \geq a$ for some $n \leq m$. Thus

$$\Pr\left\{\max_{1\leq n\leq m} Z_n \geq a\right\} = \Pr\{Z_J \geq a\} \leq \frac{\mathsf{E}[Z_J]}{a}. \tag{7.94}$$

where we have used the Markov inequality. Finally, since the process must be stopped by time m, we have $Z_J = Z_m^*$. From (7.84), $\mathsf{E}[Z_m^*] \leq \mathsf{E}[Z_m]$, so the right hand side of (7.94) is less than or equal to $\mathsf{E}[Z_m]/a$, completing the proof. □

The following simple corollary shows that (7.93) takes a simpler form for nonnegative martingales.

Corollary 7.9.1 (Nonnegative martingale inequality). *Let $\{Z_n;\, n \geq 1\}$ be a nonnegative martingale. Then*

$$\Pr\left\{\sup_{n\geq 1} Z_n \geq a\right\} \leq \frac{\mathsf{E}[Z_1]}{a}; \qquad \textit{for all } a > 0. \tag{7.95}$$

Proof For a martingale, $\mathsf{E}[Z_m] = \mathsf{E}[Z_1]$. Thus, from (7.93), $\Pr\{\max_{1\le i\le m} Z_i \ge a\} \le \frac{\mathsf{E}[Z_1]}{a}$ for all $m > 1$. Passing to the limit $m \to \infty$ essentially yields (7.95). Exercise 7.22 illustrates why the limiting operation here is a little tricky, and shows that it is valid. □

The following corollary bears the same relationship to the submartingale inequality as the Chebychev inequality does to the Markov inequality.

Corollary 7.9.2 (Kolmogorov's martingale inequality). *Let $\{Z_n;\, n \ge 1\}$ be a martingale with $\mathsf{E}\left[Z_n^2\right] < \infty$ for all $n \ge 1$. Then*

$$\Pr\left\{\max_{1\le n\le m} |Z_n| \ge b\right\} \le \frac{\mathsf{E}\left[Z_m^2\right]}{b^2}; \text{ for all integer } m \ge 2, \text{ all } b > 0. \tag{7.96}$$

Proof: Since $\{Z_n;\, n \ge 1\}$ is a martingale and Z_n^2 is a convex function of Z_n, it follows from Theorem 7.7.1 that $\{Z_n^2;\, n \ge 1\}$ is a submartingale. Since Z_n^2 is nonnegative, we can use the Kolmogorov submartingale inequality to see that

$$\Pr\left\{\max_{n\le m} Z_n^2 \ge a\right\} \le \mathsf{E}\left[Z_m^2\right]/a \quad \text{for any } a > 0.$$

Substituting b^2 for a, we get (7.96). □

Corollary 7.9.3 (Kolmogorov's random walk inequality). *Let $\{S_n;\, n \ge 1\}$ be a random walk with $S_n = X_1 + \cdots + X_n$ where $\{X_i;\, i \ge i\}$ is a set of IID random variables with mean $\overline{X}$ and variance σ^2. Then for any positive integer m and any $\epsilon > 0$,*

$$\Pr\left\{\max_{1\le n\le m} |S_n - n\overline{X}| \ge m\epsilon\right\} \le \frac{\sigma^2}{m\epsilon^2}. \tag{7.97}$$

Proof: $\{Z_n = S_n - n\overline{X};\, n \ge 1\}$ is a zero mean random walk, and thus a martingale. Since $\mathsf{E}\left[Z_m^2\right] = m\sigma^2$, (7.97) follows by substituting $m\epsilon$ for b in 7.96). □

Recall that the simplest form of the weak law of large numbers was given in (1.75) as $\Pr\left\{|S_m/m - \overline{X}| \ge \epsilon\right\} \le \sigma^2/(m\epsilon^2)$. This is strengthened in (7.97) to upper bound the probability that any of the first m terms deviate from the mean by more than $m\epsilon$. It is this strengthening that will allow us to prove the strong law of large numbers assuming only a finite variance.

The following corollary yields essentially the same result as (7.56), but is included here as another example of the use of the Kolmogorov submartingale inequality.

Corollary 7.9.4. *Let $\{S_n;\, n \ge 1\}$ be a random walk, $S_n = X_1 + \cdots + X_n$ where each X_i has mean $\overline{X} < 0$ and semi-invariant moment generating function $\gamma(r)$. For any $r > 0$ such that $0 < \gamma(r) < \infty$ (*i.e., *for $r > r^*$), and for any $a > 0$.*

$$\Pr\left\{\max_{1\le i\le n} S_i \ge \alpha\right\} \le \exp\{-r\alpha + n\gamma(r)\}. \tag{7.98}$$

Proof: For $r > r^*$, $\{\exp(rS_n);\, n \geq 1\}$ is a submartingale. Taking a $= \exp(r\alpha)$ in (7.93), we get (7.98). □

The following theorem about supermartingales is, in a sense, the dual of the Kolmogorov submartingale inequality. Note, however, that it applies to the terms $n \geq m$ in the supermartingale rather than $n \leq m$.

Theorem 7.9.2. *Let $\{Z_n;\, n \geq 1\}$ be a nonnegative supermartingale. Then for any positive integer m and any $a > 0$,*

$$\Pr\left\{\bigcup_{i \geq m} \{Z_i \geq a\}\right\} \leq \frac{\mathsf{E}[Z_m]}{a}. \tag{7.99}$$

Proof: For given $m \geq 1$ and $a > 0$, let J be a possibly-defective stopping trial defined as the smallest $i \geq m$ for which $Z_i \geq a$. Let $\{Z_n^*;\, n \geq 1\}$ be the corresponding stopped process, which is also nonnegative and is a supermartingale from Theorem 7.8.1. For any $k > m$, note that $Z_k^* \geq a$ iff $\max_{m \leq i \leq k} Z_i \geq a$. Thus

$$\Pr\left\{\max_{m \leq i \leq k} Z_i \geq a\right\} = \Pr\{Z_k^* \geq a\} \leq \frac{\mathsf{E}[Z_k^*]}{a}.$$

Since $\{Z_n^*;\, n \geq 1\}$ is a supermartingale, (7.76) shows that $\mathsf{E}[Z_k^*] \leq \mathsf{E}[Z_m^*]$. On the other hand, $Z_m^* = Z_m$ since the process can not stop before epoch m. Thus $\Pr\{\max_{m \leq i \leq k} Z_i \geq a\}$ is at most $\mathsf{E}[Z_m]/a$. Since k is arbitrary, we can pass to the limit, getting (7.99) and completing the proof. □

7.9.1 The strong law of large numbers (SLLN)

We now proceed to prove the strong law of large numbers assuming only a second moment. Recall that we proved the SLLN under the assumption of a finite fourth moment in Section 4.2.1. Here we use the Kolmogorov martingale inequality to show that only a second moment is required. The theorem is also true assuming only a first absolute moment, but the truncation argument we used for the weak law in Theorem 1.5.3 does not carry over simply here.

Theorem 7.9.3 (SLLN). *Let $\{X_i;\, i \geq 1\}$ be a sequence of IID random variables with mean $\overline{X}$ and standard deviation $\sigma < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then for any $\epsilon > 0$,*

$$\Pr\left\{\lim_{n\to\infty} \frac{S_n}{n} = \overline{X}\right\} = 1 \qquad \text{and} \tag{7.100}$$

$$\lim_{n\to\infty} \Pr\left\{\bigcup_{m>n} \left|\frac{S_m}{m} - \overline{X}\right| > \epsilon\right\} = 0. \tag{7.101}$$

Proof: From Section 4.2.2, (7.100) and (7.101) are equivalent, so we establish (7.101). As n increases, successive terms are droppped out of the union above, so the probability is non-increasing with n. Thus we can restrict attention to n of the form 2^k for integer k. For any given k, the union above can be separated into blocks as follows:

$$\begin{aligned}
\Pr\left\{\bigcup_{m>2^k}\left\{\left|\frac{S_m}{m}-\overline{X}\right|>\epsilon\right\}\right\} &= \Pr\left\{\bigcup_{j=k}^{\infty}\bigcup_{m=2^j+1}^{2^{j+1}}\left\{\left|\frac{S_m}{m}-\overline{X}\right|>\epsilon\right\}\right\} \\
&\le \sum_{j=k}^{\infty}\Pr\left\{\bigcup_{m=2^j+1}^{2^{j+1}}\left\{\left|\frac{S_m}{m}-\overline{X}\right|>\epsilon\right\}\right\} && (7.102)\\
&= \sum_{j=k}^{\infty}\Pr\left\{\bigcup_{m=2^j+1}^{2^{j+1}}\left\{|S_m-m\overline{X}|>\epsilon m\right\}\right\} \\
&\le \sum_{j=k}^{\infty}\Pr\left\{\bigcup_{m=2^j+1}^{2^{j+1}}\left\{|S_m-m\overline{X}|>\epsilon 2^j\right\}\right\} && (7.103)\\
&= \sum_{j=k}^{\infty}\Pr\left\{\max_{2^j+1\le m\le 2^{j+1}}|S_m-m\overline{X}|>\epsilon 2^j\right\} \\
&\le \sum_{j=k}^{\infty}\Pr\left\{\max_{1\le m\le 2^{j+1}}|S_m-m\overline{X}|>\epsilon 2^j\right\} && (7.104)\\
&\le \sum_{j=k}^{\infty}\frac{2^{j+1}\sigma^2}{\epsilon^2 2^{2j}} = \frac{2^{-k+2}\sigma^2}{\epsilon^2}. && (7.105)
\end{aligned}$$

In (7.102), we used the union bound on the union over j. In (7.103), we used the fact that $m \ge 2^j$ to increase the size of the sets in the union. In (7.104), we upper bounded by adding additional terms into the maximization, and in (7.105) we used the Kolmogorov martingale inequality. The proof is completed by noting that the upper bound in (7.105) goes to 0 with increasing k. □

It should be noted that, although the proof consists of a large number of steps, the steps are all quite small and familiar. Basically the proof is a slick way of upper bounding the probability of the high-order terms in a sequence by using the Kolmogorov martingale inequality, which bounds the low-order terms.

7.9.2 The martingale convergence theorem

Another famous result that follows from the Kolmogorov submartingale inequality is the martingale convergence theorem. This states that if a martingale $\{Z_n;\, n \ge 1\}$ has the property that there is some finite M such that $\mathsf{E}[|Z_n|] \le M$ for all n, then $\lim_{n\to\infty} Z_n$ exists (and is finite) with probability 1. This is a powerful theorem in more advanced work, but it is not quite as useful as it appears, since the restriction $\mathsf{E}[|Z_n|] \le M$ is more than a technical restriction; for example it is not satisfied by a zero-mean random walk. We prove the theorem with the additional restriction that there is some finite M such that $\mathsf{E}\left[Z_n^2\right] \le M$ for all n.

Theorem 7.9.4 (Martingale convergence theorem). *Let $\{Z_n;\, n \geq 1\}$ be a martingale and assume that there is some finite M such that $\mathsf{E}\left[Z_n^2\right] \leq M$ for all n. Then there is a random variable Z such that, for all sample sequences except a set of probability 0, $\lim_{n\to\infty} Z_n = Z$.*

Proof*: From Theorem 7.7.1 and the assumption that $\mathsf{E}\left[Z_n^2\right] \leq M$, $\{Z_n^2;\, n \geq 1\}$ is a submartingale. Thus, from (7.75), $\mathsf{E}\left[Z_n^2\right]$ is nondecreasing in n, and since $\mathsf{E}\left[Z_n^2\right]$ is bounded, $\lim_{n\to\infty} \mathsf{E}\left[Z_n^2\right] = M'$ for some $M' \leq M$. For any integer k, the process $\{Y_n = Z_{k+n} - Z_k;\, n \geq 1\}$ is a zero mean martingale (see Exercise 7.29). Thus from Kolmogorov's martingale inequality,

$$\Pr\left\{\max_{1\leq n\leq m} |Z_{k+n} - Z_k| \geq b\right\} \leq \mathsf{E}\left[(Z_{k+m} - Z_k)^2\right]/b^2. \tag{7.106}$$

Next, observe that $\mathsf{E}\left[Z_{k+m}Z_k \mid Z_k = z_k, Z_{k-1} = z_{k-1}, \dots, Z_1 = z_1\right] = z_k^2$, and therefore, $\mathsf{E}\left[Z_{k+m}Z_k\right] = \mathsf{E}\left[Z_k^2\right]$. Thus $\mathsf{E}\left[(Z_{k+m} - Z_k)^2\right] = \mathsf{E}\left[Z_{k+m}^2\right] - \mathsf{E}\left[Z_k^2\right] \leq M' - \mathsf{E}\left[Z_k^2\right]$. Since this is independent of m, we can pass to the limit, obtaining

$$\Pr\left\{\sup_{n\geq 1} |Z_{k+n} - Z_k| \geq b\right\} \leq \frac{M' - \mathsf{E}\left[Z_k^2\right]}{b^2}. \tag{7.107}$$

Since $\lim_{k\to\infty} \mathsf{E}\left[Z_k^2\right] = M'$, we then have, for all $b > 0$,

$$\lim_{k\to\infty} \Pr\left\{\sup_{n\geq 1} |Z_{k+n} - Z_k| \geq b\right\} = 0. \tag{7.108}$$

This means that with probability 1, a sample sequence of $\{Z_n;\, n \geq 1\}$ is a Cauchy sequence, and thus approaches a limit, concluding the proof. $\square$

This result can be relatively easily interpreted for branching processes. For a branching process $\{X_n;\, n \geq 1\}$ where $\overline{Y}$ is the expected number of offspring of an individual, $\{X_n/\overline{Y}^n;\, n \geq 1\}$ is a martingale that satisfies the above conditions. If $\overline{Y} \leq 1$, the branching process dies out with probability 1, so $X_n/\overline{Y}^n$ approaches 0 with probability 1. For $\overline{Y} > 1$, however, the branching process dies out with some probability less than 1 and approaches ∞ otherwise. Thus, the limiting random variable Z is 0 with the probability that the process ultimately dies out, and is positive otherwise. In the latter case, for large n, the interpretation is that when the population is very large, a law of large numbers effect controls its growth in each successive generation, so that $X_n/\overline{Y}^n$ tends to change in a random way for small n, and then changes increasingly little as n increases.

7.10 Markov modulated random walks

Frequently it is useful to generalize random walks to allow some dependence between the variables being summed. The particular form of dependence here is the same as the Markov reward processes of Section 3.5. The treatment in Section 3.5 discussed only expected rewards, whereas the treatment here focuses on the random variables themselves. Let $\{Y_m;\, m \geq 0\}$ be a sequence of (possibly dependent) rv's, and let

$$\{S_n;\, n \geq 1\} \quad \text{where } S_n = \sum_{m=0}^{n-1} Y_m. \tag{7.109}$$

be the process of successive sums of these random variables. Let $\{X_n;\, n \geq 0\}$ be a Markov chain, and assume that each Y_n can depend on X_n and X_{n+1}. Conditional on X_n and X_{n+1}, however, Y_n is independent of $Y_{n-1}, \dots, Y_1$, and of X_i for all $i \neq n, n{-}1$. Assume that Y_n, conditional on X_n and X_{n+1} has a distribution function $\mathsf{F}_{ij}(y) = \Pr\{Y_n \leq y \mid X_n = i, X_{n+1} = j\}$. Thus each rv Y_n depends only on the associated transition in the Markov chain, and this dependence is the same for all n.

The process $\{S_n;\, n \geq 1\}$ is called a *Markov modulated random walk*. If each Y_m is positive, it is also the sequence of epochs in a semi-Markov process. For each m, Y_m is associated with the transition in the Markov chain from time m to $m+1$, and S_n is the aggregate reward up to but not including time n. Let $\overline{Y}_{ij}$ denote $\mathsf{E}[Y_n \mid X_n = i, X_{n+1} = j]$ and $\overline{Y}_i$ denote $\mathsf{E}[Y_n \mid X_n = i]$. Let $\{P_{ij}\}$ be the set of transition probabilities for the Markov chain, so $\overline{Y}_i = \sum_j P_{ij}\overline{Y}_{ij}$. We may think of the process $\{Y_n;\, n \geq 0\}$ as evolving along with the Markov chain. The distributions of the variables Y_n are associated with the transitions from X_n to X_{n+1}, but the Y_n are otherwise independent random variables.

In order to define a martingale related to the process $\{S_n;\, n \geq 1\}$, we must subtract the mean reward from $\{S_n\}$ and must also compensate for the effect of the state of the Markov chain. The appropriate compensation factor turns out to be the relative-gain vector defined in Section 3.5.

For simplicity, consider only finite-state irreducible Markov chains with M states. Let $\pi = (\pi_1, \dots, \pi_{\mathsf{M}})$ be the steady-state probability vector for the chain, let $\overline{\boldsymbol{Y}} = (\overline{Y}_1, \dots, \overline{Y}_{\mathsf{M}})^{\mathsf{T}}$ be the vector of expected rewards, let $g = \pi\overline{Y}$ be the steady-state gain per unit time, and let $\boldsymbol{w} = (w_1, \dots, w_{\mathsf{M}})^{\mathsf{T}}$ be the relative-gain vector. From Theorem 3.5.1, $\boldsymbol{w}$ is the unique solution to

$$\boldsymbol{w} + g\boldsymbol{e} = \overline{\boldsymbol{Y}} + [P]\boldsymbol{w} \quad ; \quad w_1 = 0. \tag{7.110}$$

We assume a fixed starting state $X_0 = k$. As we now show, the process $Z_n;\, n \geq 1$ given by

$$Z_n = S_n - ng + w_{X_n} - w_k \quad ; \quad n \geq 1 \tag{7.111}$$

is a martingale. First condition on a given state, $X_{n-1} = i$.

$$\mathsf{E}[Z_n \mid Z_{n-1}, Z_{n-2}, \dots, Z_1, X_{n-1} = i]. \tag{7.112}$$

Since $S_n = S_{n-1} + Y_{n-1}$, we can express Z_n as

$$Z_n = Z_{n-1} + Y_{n-1} - g + w_{X_n} - w_{X_{n-1}}. \tag{7.113}$$

Since $\mathsf{E}[Y_{n-1} \mid X_{n-1} = i] = \overline{Y}_i$ and $\mathsf{E}[w_{X_n} \mid X_{n-1} = i] = \sum_j P_{ij} w_j$, we have

$$\mathsf{E}[Z_n \mid Z_{n-1}, Z_{n-2}, \dots, Z_1, X_{n-1} = i] = Z_{n-1} + \overline{Y}_i - g + \sum_j P_{ij} w_j - w_i. \tag{7.114}$$

From (7.110) the final four terms in (7.114) sum to 0, so

$$\mathsf{E}[Z_n \mid Z_{n-1}, \dots, Z_1, X_{n-1} = i] = Z_{n-1}. \tag{7.115}$$

Since this is valid for all choices of X_{n-1}, we have $\mathsf{E}[Z_n \mid Z_{n-1}, \dots, Z_1] = Z_{n-1}$. Since the expected values of all the reward variables $\overline{Y}_i$ exist, we see that $\mathsf{E}[|Y_n|] < \infty$, so that $\mathsf{E}[|Z_n|] < \infty$ also. This verifies that $\{Z_n;\ n \geq 1\}$ is a martingale. It can be verified similarly that $\mathsf{E}[Z_1] = 0$, so $\mathsf{E}[Z_n] = 0$ for all $n \geq 1$.

In showing that $\{Z_n; n \geq 1\}$ is a martingale, we actually showed something a little stronger. That is, (7.115) is conditioned on X_{n-1} as well as $Z_{n-1}, \dots, Z_1$. In the same way, it follows that for all $n > 1$,

$$\mathsf{E}[Z_n \mid Z_{n-1}, X_{n-1}, Z_{n-2}, X_{n-2}, \dots, Z_1, X_1] = Z_{n-1}. \tag{7.116}$$

In terms of the gambling analogy, this says that $\{Z_n; n \geq 1\}$ is fair for each possible past sequence of states. A martingale $\{Z_n; n \geq 1\}$ with this property (*i.e.*, satisfying (7.116)) is said to be a *martingale relative to the joint process* $\{Z_n, X_n; n \geq 1\}$. We will use this martingale later to discuss threshold crossing problems for Markov modulated random walks. We shall see that the added property of being a martingale relative to $\{Z_n, X_n\}$ gives us added flexibility in defining stopping times.

As an added bonus to this example, note that if $\{X_n; n \geq 0\}$ is taken as the embedded chain of a Markov process (or semi-Markov process), and if Y_n is taken as the time interval from transition n to $n+1$, then S_n becomes the epoch of the nth transition in the process.

7.10.1 Generating functions for Markov random walks

Consider the same Markov chain and reward variables as in the previous example, and assume that for each pair of states, i, j, the moment generating function

$$\mathsf{g}_{ij}(r) = \mathsf{E}[\exp(rY_n) \mid X_n = i, X_{n+1} = j]. \tag{7.117}$$

exists over some open interval (r_-, r_+) containing 0. Let $[\Gamma(r)]$ be the matrix with terms $P_{ij}\mathsf{g}_{ij}(r)$. Since $[\Gamma(r)]$ is an irreducible nonnegative matrix, Theorem 3.4.1 shows that $[\Gamma(r)]$ has a largest real eigenvalue, $\rho(r) > 0$, and an associated positive right eigenvector, $\nu(r) = (\nu_1(r), \dots, \nu_{\mathsf{M}}(r))^T$ that is unique within a scale factor. We now show that the process $\{M_n(r); n \geq 1\}$ defined by

$$M_n(r) = \frac{\exp(rS_n)\nu_{X_n}(r)}{\rho(r)^n \nu_k(r)}. \tag{7.118}$$

is a product type Martingale for each $r \in (r_-, r_+)$. Since $S_n = S_{n-1} + Y_{n-1}$, we can express $M_n(r)$ as

$$M_n(r) = M_{n-1}(r)\ \frac{\exp(rY_{n-1})\,\nu_{X_n}(r)}{\rho(r)\nu_{X_{n-1}}(r)}. \tag{7.119}$$

The expected value of the ratio in (7.119), conditional on $X_{n-1} = i$, is

$$E\left[\frac{\exp(rY_{n-1})\nu_{X_n}(r)}{\rho(r)\nu_i(r)} \mid X_{n-1}{=}i\right] = \frac{\sum_j P_{ij}\mathsf{g}_{ij}(r)\nu_j(r)}{\rho(r)\nu_i(r)} = 1. \tag{7.120}$$

where, in the last step, we have used the fact that $\nu(r)$ is an eigenvector of $[\Gamma(r)]$. Thus, $\mathsf{E}[M_n(r) \mid M_{n-1}(r), \dots, M_1(r), X_{n-1} = i] = M_{n-1}(r)$. Since this is true for all choices of i, the condition on $X_{n-1} = i$ can be removed and $\{M_n(r);\, n \geq 1\}$ is a martingale. Also, for $n > 1$,

$$\mathsf{E}[M_n(r) \mid M_{n-1}(r), X_{n-1}, \dots, M_1(r), X_1] = M_{n-1}(r). \tag{7.121}$$

so that $\{M_n(r);\, n \geq 1\}$ is also a martingale relative to the joint process $\{M_n(r), X_n;\, n \geq 1\}$.

It can be verified by the same argument as in (7.120) that $\mathsf{E}[M_1(r)] = 1$. It then follows that $\mathsf{E}[M_n(r)] = 1$ for all $n \geq 1$.

One of the uses of this martingale is to provide exponential upper bounds, similar to (7.18), to the probabilities of threshold crossings for Markov modulated random walks. Define

$$\widetilde{M}_n(r) = \frac{\exp(rS_n)\min_j(\nu_j(r))}{\rho(r)^n \nu_k(r)}. \tag{7.122}$$

Then $\widetilde{M}_n(r) \leq M_n(r)$, so $\mathsf{E}\left[\widetilde{M}_n(r)\right] \leq 1$. For any $\mu > 0$, the Markov inequality can be applied to $\widetilde{M}_n(r)$ to get

$$\Pr\left\{\widetilde{M}_n(r) \geq \mu\right\} \leq \frac{1}{\mu}\mathsf{E}\left[\widetilde{M}_n(r)\right] \leq \frac{1}{\mu}. \tag{7.123}$$

For any given α, and any given r, $0 \leq r < r_+$, we can choose $\mu = \exp(r\alpha)\rho(r)^{-n}\min_j(\nu_j(r))/\nu_k(r)$, and for $r > 0$. Combining (7.122) and (7.123),

$$\Pr\{S_n \geq \alpha\} \leq \rho(r)^n \exp(-r\alpha)\nu_k(r)/\min_j(\nu_j(r)). \tag{7.124}$$

This can be optimized over r to get the tightest bound in the same way as (7.18).

7.10.2 stopping trials for martingales relative to a process

A martingale $\{Z_n;\, n \geq 1\}$ relative to a joint process $\{Z_n, X_n;\, n \geq 1\}$ was defined as a martingale for which (7.116) is satisfied, *i.e.*, $\mathsf{E}[Z_n \mid Z_{n-1}, X_{n-1}, \dots, Z_1, X_1] = Z_{n-1}$. In the same way, we can define a *submartingale or supermartingale* $\{Z_n;\, n \geq 1\}$ *relative to a joint process* $\{Z_n, X_n;\, n \geq 1\}$ as a submartingale or supermartingale satisfying (7.116) with the $=$ sign replaced by $\geq$ or $\leq$ respectively. The purpose of this added complication is to make it easier to define useful stopping rules.

As generalized in Definition 4.5.2, a generalized stopping trial $\mathbb{I}_{J\geq n}$ for a sequence of pairs of rv's $(Z_1, X_1), (Z_2, X_2)\dots$, is a positive integer-valued rv such that, for each $n \geq 1$, $\mathbb{I}_{\{J=n\}}$ is a function of $Z_1, X_1, Z_2, X_2, \dots, Z_n, X_n$.

Theorems 7.8.1, 7.8.2 and Lemma 7.8.1 all carry over to martingales (submartingales or supermartingales) relative to a joint process. These theorems are stated more precisely in Exercises 7.23 to 7.26. To summarize them here, assume that $\{Z_n;\, n \geq 1\}$ is a martingale (submartingale or supermartingale) relative to a joint process $\{Z_n, X_n;\, n \geq 1\}$ and assume that J is a stopping trial for $\{Z_n;\, n \geq 1\}$ relative to $\{Z_n, X_n;\, n \leq 1\}$. Then the stopped process is a martingale (submartingale or supermartingale) respectively, (7.84 —7.86) are satisfied, and, for a martingale, $\mathsf{E}[Z_J] = \mathsf{E}[Z_1]$ is satisfied iff (7.90) is satisfied.

7.10.3 Markov modulated random walks with thresholds

We have now developed two martingales for Markov modulated random walks, both conditioned on a fixed initial state $X_0 = k$. The first, given in (7.111), is $\{Z_n = S_n - ng + w_{X_n} - w_k;\, n \geq 1\}$. Recall that $\mathsf{E}[Z_n] = 0$ for all $n \geq 1$ for this martingale. Given two thresholds, $\alpha > 0$ and $\beta < 0$, define J as the smallest n for which $S_n \geq \alpha$ or $S_n \leq \beta$. The indicator function $\mathbb{I}_{J=n}$ of $\{J = n\}$, is 1 iff $\beta < S_i < \alpha$ for $1 \leq i \leq n-1$ and either $S_n \geq \alpha$ or $S_n \leq \beta$. Since $S_i = Z_i + ig - w_{X_i} + w_k$, S_i is a function of Z_i and X_i, so the stopping trial is a function of both Z_i and X_i for $1 \leq i \leq n$. It follows that J is a stopping trial for $\{Z_n;\, n \geq 1\}$ relative to $\{Z_n, X_n;\, n \geq 1\}$. From Lemma 7.8.1, we can assert that $\mathsf{E}[Z_J] = \mathsf{E}[Z_1] = 0$ if (7.90) is satisfied, *i.e.*, if $\lim_{n\to\infty} \mathsf{E}[Z_n \mid J > n] \Pr\{J > n\} = 0$ is satisfied. Using the same argument as in Lemma 7.5.1, we can see that $\Pr\{J > n\}$ goes to 0 at least geometrically in n. Conditional on $J > n$, $\beta < S_n < \alpha$, so S_n is bounded independent of n. Also w_{X_n} is bounded, since the chain is finite state, and ng is linear in n. Thus $\mathsf{E}[Z_n \mid J > n]$ varies at most linearly with n, so (7.90) is satisfied, and

$$0 = \mathsf{E}[Z_J] = \mathsf{E}[S_J] - \mathsf{E}[J]\, g + \mathsf{E}[w_{X_n}] - w_k. \tag{7.125}$$

Recall that Wald's equality for random walks is $\mathsf{E}[S_J] = \mathsf{E}[J]\, \overline{X}$. For Markov modulated random walks, this is modified, as shown in (7.125), by the relative-gain vector terms.

The same arguments can be applied to the generating function martingale of (7.118). Again, let J be the smallest n for which $S_n \geq \alpha$ or $S_n \leq \beta$. As before, S_i is a function of $M_i(r)$ and X_i, so $\mathbb{I}_n$ is a function of $M_i(r)$ and X_i for $1 \leq i \leq n-1$. It follows that J is a stopping trial for $\{M_n(r);\, n \geq 1\}$ relative to $\{M_n(r), X_n;\, n \geq 1\}$ Next we need the following lemma:

Lemma 7.10.1. *For the martingale $\{M_n(r);\, n \geq 1\}$ relative to $\{M_n(r), X_n;\, n \geq 1\}$ defined in (7.118), where $\{X_n;\, n \geq 0\}$ is a finite-state Markov chain, and for the above stopping trial J,*

$$\lim_{n\to\infty} \mathsf{E}[M_n(r) \mid J > n] \Pr\{J > n\} = 0. \tag{7.126}$$

Proof: From lemma 4, slightly modified for the case here, there is a $\delta > 0$ such that for all states i, j, and all $n > 1$ such that $\Pr\{J = n, X_{n-1} = i, X_n = j\} > 0$,

$$\mathsf{E}[\exp(rS_n \mid J = n, X_{n-1} = i, X_n = j] \geq \delta. \tag{7.127}$$

Since the stopped process, $\{M_n^*(r);\, n \geq 1\}$, is a martingale, we have for each m,

$$1 = \mathsf{E}[M_m^*(r)] \geq \sum_{n=1}^{m} \frac{\mathsf{E}[\exp(rS_n)\nu_{X_n}(r) \mid J = n]}{\rho(r)^n \nu_k(r)}. \tag{7.128}$$

From (7.127), we see that there is some $\delta' > 0$ such that

$$\mathsf{E}\left[\exp(rS_n)\nu_{X_n}(r)\right]/\nu_k(r) \mid J = n] \geq \delta'$$

for all n such that $\Pr\{J = n\} > 0$. Thus (7.128) is bounded by

$$1 \geq \delta' \sum_{n \leq m} \rho(r)^n \Pr\{J = n\}\,.$$

Since this is valid for all m, it follows by the argument in the proof of theorem 7.5.1 that $\lim_{n\to\infty} \rho(r)^n \Pr\{J > n\} = 0$. This, along with (7.127), establishes (7.126), completing the proof. □

From Lemma 7.8.1, we have the desired result:

$$\mathsf{E}\left[M_J(r)\right] = E\left[\frac{\exp(rS_J)\nu_{X_J}(r)}{[\rho(r)]^J \nu_k(r)}\right] = 1; \qquad r_- < r < r_+. \tag{7.129}$$

This is the extension of the Wald identity to Markov modulated random walks, and is used in the same way as the Wald identity. As shown in Exercise 7.28, the derivative of (7.129), evaluated at $r = 0$, is the same as (7.125).

7.11 Summary

Each term in a random walk $\{S_n; n \geq 1\}$ is a sum of IID random variables, and thus the study of random walks is closely related to that of sums of IID variables. The focus in random walks, however, as in most of the processes we have studied, is more in the relationship between the terms (such as which term first crosses a threshold) than in the individual terms. We started by showing that random walks are a generalization of renewal processes, are central to studying the queueing delay for G/G/1 queues, and to sequential analysis for hypothesis testing.

A major focus of the chapter was on estimating the probabilities of very unlikely events, a topic known as large deviation theory. We started by studying the Chernoff bound to $\Pr\{S_n \geq \alpha\}$ for $\alpha > 0$ and $\mathsf{E}[X] < 0$. We then developed the Wald identity, which can be used to find tight upper bounds to the probability that a threshold is ever crossed by a random walk. One of the insights gained here was that if a threshold at α is crossed, it is likely to be crossed at a time close to $n^* = \alpha/\gamma'(r^*)$, where r^* is the positive root of $\gamma(r)$. We also found that r^* plays a fundamental role in the probability of threshold crossings. For questions of typical behavior, the mean and variance of a random variable are the major quantities of interest, but when interested in atypically large deviations, r^* is the major parameter of interest.

We next introduced martingales, submartingales and supermartingales. These are sometimes regarded as somewhat exotic topics in mathematics, but in fact they are very useful in a large variety of relatively simple processes. For example, we showed that all of the random walk issues of earlier sections can be treated as a special case of martingales, and that martingales can be used to model both sums and products of random variables. We also showed how Markov modulated random walks can be treated as martingales.

Stopping trials, as first introduced in chapter 4, were then applied to martingales. We defined a stopped process $\{Z_n^*; n \geq 1\}$ to be the same as the original process $\{Z_n; n \geq 1\}$ up to the stopping point, and then constant thereafter. Theorems 7.8.1 and 7.8.2 showed that the stopped process has the same form (martingale, submartingale, or supermartingale) as the original process, and that the expected values $\mathsf{E}[Z_n^*]$ are between $\mathsf{E}[Z_1]$ and $\mathsf{E}[Z_n]$. We also looked at $\mathsf{E}[Z_J]$ and found that it is equal to $\mathsf{E}[Z_1]$ iff (7.90) is satisfied. The Wald identity can be viewed as $\mathsf{E}[Z_J] = \mathsf{E}[Z_1] = 1$ for the Wald martingale, $Z_n = \exp\{rS_n - n\gamma(r)\}$. We then found a similar identity for Markov modulated random walks. In deriving results for Markov modulated random walks, it was necessary to define martingales relative to other processes in order to find suitable stopping trials, also defined on martingales relative to other processes. This added restriction on martingales is useful in other contexts.

The Kolmogorov inequalities were next developed. They are analogs of the Markov inequality and Chebyshev inequality, except they bound initial segments of submartingales and martingales rather than single rv's. They were used, first, to prove the SLLN with only a second moment and, second,the martingale convergence theorem.

A standard reference on random walks, and particularly on the analysis of overshoots is [Fel66]. Dembo and Zeitouni, [5] develop large deviation theory in a much more general and detailed way than the introduction here. The classic reference on martingales is [6], but [4] and [16] are more accessible.

7.12 Exercises

Exercise 7.1. Consider the simple random walk $\{S_n;\, n \geq 1\}$ of Section 7.1.1 with $S_n = X_1 + \cdots + X_n$ and $\Pr\{X_i = 1\} = p$; $\Pr\{X_i = -1\} = 1 - p$; assume that $p \leq 1/2$.

a) Show that $\Pr\left\{\bigcup_{n\geq 1} \{S_n \geq k\}\right\} = \left[\Pr\left\{\bigcup_{n\geq 1} \{S_n \geq 1\}\right\}\right]^k$ for any positive integer k. Hint: Given that the random walk ever reaches the value 1, consider a new random walk starting at that time and explore the probability that the new walk ever reaches a value 1 greater than its starting point.

b) Find a quadratic equation for $y = \Pr\left\{\bigcup_{n\geq 1} \{S_n \geq 1\}\right\}$. Hint: explore each of the two possibilities immediately after the first trial.

c) For $p < 1/2$, show that the two roots of this quadratic equation are $p/(1-p)$ and 1. Argue that $\Pr\left\{\bigcup_{n\geq 1} \{S_n \geq 1\}\right\}$ cannot be 1 and thus must be $p/(1-p)$.

d) For $p = 1/2$, show that the quadratic equation in part c) has a double root at 1, and thus $\Pr\left\{\bigcup_{n\geq 1} \{S_n \geq 1\}\right\} = 1$. Note: this is the very peculiar case explained in the section on Wald's equality.

e) For $p < 1/2$, show that $p/(1-p) = \exp(-r^*)$ where r^* is the unique positive root of $g(r) = 1$ where $g(r) = \mathsf{E}\left[e^{rX}\right]$.

Exercise 7.2. Consider a G/G/1 queue with IID arrivals $\{X_i;\ i \geq 1\}$, IID FCFS service times $\{Y_i;\ i \geq 0\}$, and an initial arrival to an empty system at time 0. Define $U_i = X_i - Y_{i-1}$ for $i \geq 1$. Consider a sample path where $(u_1, \ldots, u_6) = (1, -2, 2, -1, 3, -2)$.

a) Let $Z_i^6 = U_6 + U_{6-1} + \ldots + U_{6-i+1}$. Find the queueing delay for customer 6 as the maximum of the 'backward' random walk with elements $0, Z_1^6, Z_2^6, \ldots, Z_6^6$; sketch this random walk.

b) Find the queueing delay for customers 1 to 5.

c) Which customers start a busy period (*i.e.*, arrive when the queue and server are both empty)? Verify that if Z_i^6 maximizes the random walk in part a), then a busy period starts with arrival $6 - i$.

d) Now consider a forward random walk $V_n = U_1 + \cdots + U_n$. Sketch this walk for the sample path above and show that the queueing delay for each customer is the difference between two appropriately chosen values of this walk.

Exercise 7.3. A G/G/1 queue has a deterministic service time of 2 and interarrival times that are 3 with probability p and 1 with probability $1 - p$.

a) Find the distribution of W_1, the wait in queue of the first arrival after the beginning of a busy period.

b) Find the distribution of W_∞, the steady-state wait in queue.

c) Repeat parts a) and b) assuming the service times and interarrival times are exponentially distributed with rates μ and λ respectively.

Exercise 7.4. A sales executive hears that one of his sales people is routing half of his incoming sales to a competitor. In particular, arriving sales are known to be Poisson at rate one per hour. According to the report (which we view as hypothesis 1), each second arrival is routed to the competition; thus under hypothesis 1 the interarrival density for successful sales is $f(y|\mathsf{H}_1) = ye^{-y}$; $y \geq 0$. The alternate hypothesis (H_0) is that the rumor is false and the interarrival density for successful sales is $f(y|\mathsf{H}_0) = e^{-y}$; $y \geq 0$. Assume that, *a priori*, the hypotheses are equally likely. The executive, a recent student of stochastic processes, explores various alternatives for choosing between the hypotheses; he can only observe the times of successful sales however.

a) Starting with a successful sale at time 0, let S_i be the arrival time of the i^{th} subsequent successful sale. The executive observes $S_1, S_2, \dots, S_n (n \geq 1)$ and chooses the maximum aposteriori probability hypothesis given this data. Find the joint probability density $f(S_1, S_2, \dots, S_n|\mathsf{H}_1)$ and $f(S_1, \dots, S_n|\mathsf{H}_0)$ and give the decision rule.

b) This is the same as part a) except that the system is in steady state at time 0 (rather than starting with a successful sale). Find the density of S_1 (the time of the first arrival after time 0) conditional on H_0 and on H_1. What is the decision rule now after observing $S_1, \dots, S_n$.

c) This is the same as part b), except rather than observing n successful sales, the successful sales up to some given time t are observed. Find the probability, under each hypothesis, that the first successful sale occurs in $(s_1, s_1 + \Delta]$, the second in $(s_2, s_2 + \Delta], \dots$, and the last in $(s_{N(t)}, s_{N(t)} + \Delta]$ (assume Δ very small). What is the decision rule now?

Exercise 7.5. For the hypothesis testing problem of Section 7.3, assume that there is a cost C_0 of choosing H_1 when H_0 is correct, and a cost C_1 of choosing H_0 when H_1 is correct. Show that a threshold test minimizes the expected cost using the threshold $\eta = (C_1 p_1)/(C_0 p_0)$.

Exercise 7.6. Consider a binary hypothesis testing problem where H is 0 or 1 and a one dimensional observation Y is given by $Y = H + U$ where U is uniformly distributed over [-1, 1] and is independent of H.

a) Find $\mathsf{f}_{Y|H}(y \mid 0)$, $\mathsf{f}_{Y|H}(y \mid 1)$ and the likelihood ratio $\Lambda(y)$.

b) Find the threshold test at η for each η, $0 < \eta < \infty$ and evaluate the conditional error probabilities, $q_0(\eta)$ and $q_1(\eta)$.

c) Find the error curve $u(\alpha)$ and explain carefully how $u(0)$ and $u(1/2)$ are found (hint: $u(0) = 1/2$).

d) Describe a decision rule for which the error probability under each hypothesis is 1/4. You need not use a randomized rule, but you need to handle the don't-care cases under the threshold test carefully.

Exercise 7.7. a) For given α, $0 < \alpha \leq 1$, let η^* achieve the supremum $\sup_{0\leq\eta<\infty} q_1(\eta) + \eta(q_0(\eta) - \alpha)$. Show that $\eta^* \leq 1/\alpha$. Hint: Think in terms of Lemma 7.3.1 applied to a very simple test.

b) Show that the magnitude of the slope of the error curve $u(\alpha)$ at α is at most $1/\alpha$.

Exercise 7.8. Define $\gamma(r)$ as $\ln[\mathsf{g}(r)]$ where $\mathsf{g}(r) = \mathsf{E}[\exp(rX)]$. Assume that X is discrete with possible outcomes $\{a_i;\ i \geq 1\}$, let p_i denote $\Pr\{X = a_i\}$, and assume that $\mathsf{g}(r)$ exists in some open interval (r_-, r_+) containing $r = 0$. For any given r, $r_- < r < r_+$, define a random variable X_r with the same set of possible outcomes $\{a_i;\ i \geq 1\}$ as X, but with a probability mass function $q_i = \Pr\{X_r = a_i\} = p_i \exp[a_i r - \gamma(r)]$. X_r is not a function of X, and is not even to be viewed as in the same probability space as X; it is of interest simply because of the behavior of its defined probability mass function. It is called a tilted random variable relative to X, and this exercise, along with Exercise 7.9 will justify our interest in it.

a) Verify that $\sum_i q_i = 1$.

b) Verify that $\mathsf{E}[X_r] = \sum_i a_i q_i$ is equal to $\gamma'(r)$.

c) Verify that Var $[X_r] = \sum_i a_i^2 q_i - (\mathsf{E}[X_r])^2$ is equal to $\gamma''(r)$.

d) Argue that $\gamma''(r) \geq 0$ for all r such that $\mathsf{g}(r)$ exists, and that $\gamma''(r) > 0$ if $\gamma''(0) > 0$.

e) Give a similar definition of X_r for a random variable X with a density, and modify parts a) to d) accordingly.

Exercise 7.9. Assume that X is discrete, with possible values $\{a_i;\ i \geq 1)$ and probabilities $\Pr\{X = a_i\} = p_i$. Let X_r be the corresponding tilted random variable as defined in Exercise 7.8. Let $S_n = X_1 + \cdots + X_n$ be the sum of n IID rv's with the distribution of X, and let $S_{n,r} = X_{1,r} + \cdots + X_{n,r}$ be the sum of n IID tilted rv's with the distribution of X_r. Assume that $\overline{X} < 0$ and that $r > 0$ is such that $\gamma(r)$ exists.

a) Show that $\Pr\{S_{n,r}{=}s\} = \Pr\{S_n{=}s\} \exp[sr - n\gamma(r)]$. Hint: first show that

$$\Pr\{X_{1,r}{=}v_1, \dots, X_{n,r}{=}v_n\} = \Pr\{X_1{=}v_1, \dots, X_n{=}v_n\} \exp[sr - n\gamma(r)]$$

where $s = v_1 + \cdots + v_n$.

b) Find the mean and variance of $S_{n,r}$ in terms of $\gamma(r)$.

c) Define $a = \gamma'(r)$ and $\sigma_r^2 = \gamma''(r)$. Show that $\Pr\{|S_{n,r} - na| \leq \sqrt{2n}\sigma_r\} > 1/2$. Use this to show that

$$\Pr\Big\{|\ S_n - na\ | \leq \sqrt{2n}\,\sigma_r\ \Big\} > (1/2)\exp[-r(an + \sqrt{2n}\ \sigma_r) + n\gamma(r)].$$

d) Use this to show that for any ϵ and for all sufficiently large n,

$$\Pr\{S_n \geq n(\gamma'(r) - \epsilon)\} > \frac{1}{2}\exp[-rn(\gamma'(r) + \epsilon) + n\gamma(r)].$$

Exercise 7.10. Consider a random walk with thresholds $\alpha > 0$, $\beta < 0$. We wish to find $\Pr\{S_J \geq \alpha\}$ in the absence of a lower threshold. Use the upper bound in (7.42) for the probability that the random walk crosses α before β.

a) Given that the random walk crosses β first, find an upper bound to the probability that α is now crossed before a yet lower threshold at 2β is crossed.

b) Given that 2β is crossed before α, upperbound the probability that α is crossed before a threshold at 3β. Extending this argument to successively lower thresholds, find an upper bound to each successive term, and find an upper bound on the overall probability that α is crossed. By observing that β is arbitrary, show that (7.42) is valid with no lower threshold.

Exercise 7.11. This exercise verifies that Corollary 7.5.1 holds in the situation where $\gamma(r) < 0$ for all $r \in (r_-, r_+)$ and where r^* is taken to be r_+ (see Figure 7.7).

a) Show that for the situation above, $\exp(rS_J) \leq \exp(rS_J - J\gamma(r))$ for all $r \in (0, r^*)$.

b) Show that $\mathsf{E}[\exp(rS_J)] \leq 1$ for all $r \in (0, r^*)$.

c) Show that $\Pr\{S_J \geq \alpha\} \leq \exp(-r\alpha)$ for all $r \in (0, r^*)$. Hint: Follow the steps of the proof of Corollary 7.5.1.

d) Show that $\Pr\{S_J \geq \alpha\} \leq \exp(-r^*\alpha)$.

Exercise 7.12. a) Use Wald's equality to show that if $\overline{X} = 0$, then $\mathsf{E}[S_J] = 0$ where J is the time of threshold crossing with one threshold at $\alpha > 0$ and another at $\beta < 0$.

b) Obtain an expression for $\Pr\{S_J \geq \alpha\}$. Your expression should involve the expected value of S_J conditional on crossing the individual thresholds (you need not try to calculate these expected values).

c) Evaluate your expression for the case of a simple random walk.

d) Evaluate your expression when X has an exponential density, $\mathsf{f}_X(x) = a_1 e^{-\lambda x}$ for $x \geq 0$ and $\mathsf{f}_X(x) = a_2 e^{\mu x}$ for $x < 0$ and where a_1 and a_2 are chosen so that $\overline{X} = 0$.

Exercise 7.13. A random walk $\{S_n; n \geq 1\}$, with $S_n = \sum_{i=1}^{n} X_i$, has the following probability density for X_i

$$\mathsf{f}_X(x) = \begin{cases} \frac{e^{-x}}{e - e^{-1}} \; ; & -1 \leq x \leq 1 \\ \\ = 0 \; ; & \text{elsewhere.} \end{cases}$$

a) Find the values of r for which $\mathsf{g}(r) = \mathsf{E}[\exp(rX)] = 1$.

b) Let P_α be the probability that the random walk ever crosses a threshold at α for some $\alpha > 0$. Find an upper bound to P_α of the form $P_\alpha \leq e^{-\alpha A}$ where A is a constant that does not depend on α; evaluate A.

c) Find a lower bound to P_α of the form $P_\alpha \geq Be^{-\alpha A}$ where A is the same as in part b) and B is a constant that does not depend on α. Hint: keep it simple — you are not expected to find an elaborate bound. Also recall that $\mathsf{E}\left[e^{r^*S_J}\right] = 1$ where J is a stopping trial for the random walk and $\mathsf{g}(r^*) = 1$.

Exercise 7.14. Let $\{X_n;\, n \geq 1\}$ be a sequence of IID integer valued random variables with the probability mass function $P_X(k) = Q_k$. Assume that $Q_k > 0$ for $|k| \leq 10$ and $Q_k = 0$ for $|k| > 10$. Let $\{S_n;\, n \geq 1\}$ be a random walk with $S_n = X_1 + \cdots + X_n$. Let $\alpha > 0$ and $\beta < 0$ be integer valued thresholds, let J be the smallest value of n for which either $S_n \geq \alpha$ or $S_n \leq \beta$. Let $\{S_n^*;\, n \geq 1\}$ be the stopped random walk; *i.e.*, $S_n^* = S_n$ for $n \leq J$ and $S_n^* = S_J$ for $n > J$. Let $\pi_i^* = \Pr\{S_J = i\}$.

a) Consider a Markov chain in which this stopped random walk is run repeatedly until the point of stopping. That is, the Markov chain transition probabilities are given by $P_{ij} = Q_{j-i}$ for $\beta < i < \alpha$ and $P_{i0} = 1$ for $i \leq \beta$ and $i \geq \alpha$. All other transition probabilities are 0 and the set of states is the set of integers $[-9+\beta, 9+\alpha]$. Show that this Markov chain is ergodic.

b) Let $\{\pi_i\}$ be the set of steady-state probabilities for this Markov chain. Find the set of probabilities $\{\pi_i^*\}$ for the stopping states of the stopped random walk in terms of $\{\pi_i\}$.

c) Find $\mathsf{E}[S_J]$ and $\mathsf{E}[J]$ in terms of $\{\pi_i\}$.

Exercise 7.15. a) Conditional on H_0 for the hypothesis testing problem, consider the random variables $Z_i = \ln[f(Y_i|\mathsf{H}_1)/f(Y_i|\mathsf{H}_0)]$. Show that r^*, the positive solution to $\mathsf{g}(r) = 1$, where $g(r) = \mathsf{E}[\exp(rZi)]$, is given by $r^* = 1$.

b) Assuming that Y is a discrete random variable (under each hypothesis), show that the tilted random variable Z_r with $r = 1$ has the PMF $P_Y(y|\mathsf{H}_1)$.

Exercise 7.16. a) Suppose $\{Z_n;\, n \geq 1\}$ is a martingale. Verify (7.69); *i.e.*, $\mathsf{E}[Z_n] = \mathsf{E}[Z_1]$ for $n > 1$.

b) If $\{Z_n;\, n \geq 1\}$ is a submartingale, verify (7.75), and if a supermartingale, verify (7.76).

Exercise 7.17. Suppose $\{Z_n;\, n \geq 1\}$ is a martingale. Show that

$$\mathsf{E}\left[Z_m \mid Z_{n_i}, Z_{n_{i-1}}, \ldots, Z_{n_1}\right] = Z_{n_i} \quad \text{for all } 0 < n_1 < n_2 < \ldots < n_i < m.$$

Exercise 7.18. a) Assume that $\{Z_n;\, n \geq 1\}$ is a submartingale. Show that

$$\mathsf{E}[Z_m \mid Z_n, Z_{n-1}, \ldots, Z_1] \geq Z_n \text{ for all } n < m.$$

b) Show that

$$\mathsf{E}\left[Z_m \mid Z_{n_i}, Z_{n_{i-1}}, \ldots, Z_{n_1}\right] \geq Z_{n_i} \text{ for all } \; 0 < n_1 < n_2 < \ldots < n_i < m.$$

c) Assume now that $\{Z_n;\, n \geq 1\}$ is a supermartingale. Show that parts a) and b) still hold with $\geq$ replaced by $\leq$.

Exercise 7.19. Let $\{Z_n = \exp[rS_n - n\gamma(r)];\ n \geq 1\}$ be the generating function martingale of (7.63) where $S_n = X_1 + \cdots + X_n$ and $X_1, \dots X_n$ are IID with mean $\overline{X} < 0$. Let J be the possibly-defective stopping trial for which the process stops after crossing a threshold at $\alpha > 0$ (there is no negative threshold). Show that $\exp[r^*\alpha]$ is an upper bound to the probability of threshold crossing by considering the stopped process $\{Z_n^*;\ n \geq 1\}$.

The purpose of this exercise is to illustrate that the stopped process can yield useful upper bounds even when the stopping trial is defective.

Exercise 7.20. This problem uses martingales to find the expected number of trials $\mathsf{E}[J]$ before a fixed pattern, $a_1, a_2, \dots, a_k$, of binary digits occurs within a sequence of IID binary random variables $X_1, X_2, \dots$ (see Exercises 4.32 and 3.28 for alternate approaches). A mythical casino and set of gamblers who follow a prescribed strategy will be used to determine $\mathsf{E}[J]$. The casino has a game where, on the ith trial, gamblers bet money on either 1 or 0. After bets are placed, X_i above is used to select the outcome 0 or 1. Let $p(1) = \mathsf{p}_X(1)$ and $p(0) = 1 - p(1) = \mathsf{p}_X(0)$. If an amount s is bet on 1, the casino receives s if $X_i = 0$, and pays out $s/p(1) - s$ (plus returning the bet s) if $X_i = 1$. If s is bet on 0, the casino receives s if $X_i = 1$, and pays out $s/p(0) - s$ (plus the bet s) if $X_i = 0$.

a) Assume an arbitrary pattern of bets by various gamblers on various trials (some gamblers might bet arbitrary amounts on 0 and some on 1 at any given trial). Let Y_i be the net gain of the casino on trial i. Show that $\mathsf{E}[Y_i] = 0$ (*i.e.*, show that the game is fair). Let $Z_n = Y_1 + Y_2 + \cdots + Y_n$ be the aggregate gain of the casino over n trials. Show that for the given pattern of bets, $\{Z_n;\ n \geq 1\}$ is a martingale.

b) In order to determine $\mathsf{E}[J]$ for a given pattern $a_1, a_2, \dots, a_k$, we program our gamblers to bet as follows:

i) Gambler 1 has an initial capital of 1 which is bet on a_1 at trial 1. If he wins, his capital grows to $1/p(a_1)$, which is bet on a_2 at trial 2. If he wins again, he bets his entire capital, $1/[p(a_1)p(a_2)]$, on a_3 at trial 3. He continues, at each trial i, to bet his entire capital on a_i until he loses at some trial (in which case he leaves with no money) or he wins on k successive trials (in which case he leaves with $1/[p(a_1) \dots p(a_k)]$.

ii) Gambler j, $j > 1$, follows exactly the same strategy but starts at trial j. Note that if the pattern $a_1, \dots, a_k$ appears for the first time at $J = n$, then gambler $n - k + 1$ leaves at time n with capital $1/[p(a_1) \dots p(a_k)]$ and gamblers $j < n - k + 1$ all lose their capital.

Suppose the string $(a_1, \dots, a_k)$ is $(0, 1)$. Let for the above gambling strategy. Given that $J = 3$ (*i.e.*, that $X_2 = 0$ and $X_3 = 1$), note that gambler 1 loses his money at either trial 1 or 2, gambler 2 leaves at time 3 with $1/[p(0)p(1)]$ and gambler 3 loses his money at time 3. Show that $Z_J = 3 - 1/[p(0)p(1)]$ given $J = 3$. Find Z_J given $J = n$ for arbitrary $n \geq 2$ (note that the condition $J = n$ uniquely specifies Z_J).

c) Find $\mathsf{E}[Z_J]$ from part a). Use this plus part b) to find $\mathsf{E}[J]$.

d) Repeat parts b) and c) using the string $(a_1, \dots, a_k) = (1, 1)$. Be careful about gambler 3 for $J = 3$. Show that $\mathsf{E}[J] = 1/[p(1)p(1)] + 1/p(1)$

e) Repeat parts b) and c) for $(a_1, \dots, a_k) = (1, 1, 1, 0, 1, 1)$.

Exercise 7.21. **a)** This exercise shows why the condition $\mathsf{E}[|Z_J|] < \infty$ is required in Lemma 7.8.1. Let $Z_1 = -2$ and, for $n \geq 1$, let $Z_{n+1} = Z_n[1 + X_n(3n+1)/(n+1)]$ where $X_1, X_2, \dots$ are IID and take on the values $+1$ and -1 with probability 1/2 each. Show that $\{Z_n;\, n \geq 1\}$ is a martingale.

b) Consider the stopping trial J such that J is the smallest value of $n > 1$ for which Z_n and Z_{n-1} have the same sign. Show that, conditional on $n < J$, $Z_n = (-2)^n/n$ and, conditional on $n = J$, $Z_J = -(-2)^n(n-2)/(n^2 - n)$.

c) Show that $\mathsf{E}[|Z_J|]$ is infinite, so that $\mathsf{E}[Z_J]$ does not exist according to the definition of expectation, and show that $\lim_{n\to\infty} \mathsf{E}[Z_n | J > n] \Pr\{J > n\} = 0$.

Exercise 7.22. This exercise shows why the sup of a martingale can behave markedly differently from the maximum of an arbitrarily large number of the variables. More precisely, it shows that $\Pr\left\{\sup_{n\geq 1} Z_n \geq a\right\}$ can be unequal to $\Pr\left\{\bigcup_{n\geq 1}\{Z_n \geq a\}\right\}$.

a) Consider a martingale where Z_n can take on only the values 2^{-n-1} and $1 - 2^{-n-1}$, each with probability 1/2. Given that Z_n, conditional on Z_{n-1}, is independent of $Z_1, \dots Z_{n-2}$, find $\Pr\{Z_n | Z_{n-1}\}$ for each n so that the martingale condition is satisfied.

b) Show that $\Pr\left\{\sup_{n\geq 1} Z_n \geq 1\right\} = 1/2$ and show that $\Pr\{\bigcup_n \{Z_n \geq 1\}\} = 0$.

c Show that for every $\epsilon > 0$, $\Pr\left\{\sup_{n\geq 1} Z_n \geq a\right\} \leq \frac{\overline{Z}_1}{a-\epsilon}$. Hint: Use the relationship between $\Pr\left\{\sup_{n\geq 1} Z_n \geq a\right\}$ and $\Pr\{\bigcup_n \{Z_n \geq a\}\}$ while getting around the issue in part b).

d Use part c) to establish (7.95).

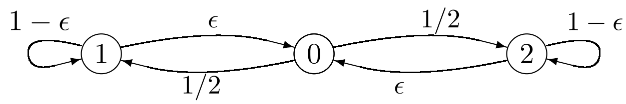
Exercise 7.23. Show that Theorem 7.7.1 is also valid for martingales relative to a joint process. That is, show that if h is a convex function of a real variable and if $\{Z_n;\, n \geq 1\}$ is a martingale relative to a joint process $\{Z_n, X_n;\, n \geq 1\}$, then $\{h(Z_n);\, n \geq 1\}$ is a submartingale relative to $\{h(Z_n), X_n;\, n \geq 1\}$.

Exercise 7.24. Show that if $\{Z_n;\, n \geq 1\}$ is a martingale (submartingale or supermartingale) relative to a joint process $\{Z_n, X_n;\, n \geq 1\}$ and if J is a stopping trial for $\{Z_n;\, n \geq 1\}$ relative to $\{Z_n, X_n;\, n \geq 1\}$, then the stopped process is a martingale (submartingale or supermartingale) respectively relative to the joint process.

Exercise 7.25. Show that if $\{Z_n;\, n \geq 1\}$ is a martingale (submartingale or supermartingale) relative to a joint process $\{Z_n, X_n;\, n \geq 1\}$ and if J is a stopping trial for $\{Z_n;\, n \geq 1\}$ relative to $\{Z_n, X_n;\, n \geq 1\}$, then the stopped process satisfies (7.84), (7.85), or (7.86) respectively.

Exercise 7.26. Show that if $\{Z_n;\, n1\}$ is a martingale relative to a joint process $\{Z_n, X_n;\, n \geq 1\}$ and if J is a stopping trial for $\{Z_n;\, n \geq 1\}$ relative to $\{Z_n, X_n;\, n \geq 1\}$, then $\mathsf{E}[Z_J] = \mathsf{E}[Z_1]$ if and only f (7.90) is satisfied.

Exercise 7.27. Consider the Markov modulated random walk in the figure below. The random variables Y_n in this example take on only a single value for each transition, that value being 1 for all transitions from state 1, 10 for all transitions from state 2, and 0 otherwise. $\epsilon > 0$ is a very small number, say $\epsilon < 10^{-6}$.



a) Show that the steady-state gain per transition is $5.5/(1+\epsilon)$. Show that the relative-gain vector is $\boldsymbol{w} = (0, (\epsilon - 4.5)/[\epsilon(1+\epsilon)], (10\epsilon + 4.5)/[\epsilon(1+\epsilon)])$.

b) Let $S_n = Y_0 + Y_1 + \cdots + Y_{n-1}$ and take the starting state X_0 to be 0. Let J be the smallest value of n for which $S_n \geq 100$. Find $\Pr\{J = 11\}$ and $\Pr\{J = 101\}$. Find an estimate of $\mathsf{E}[J]$ that is exact in the limit $\epsilon \to 0$.

c) Show that $\Pr\{X_J = 1\} = (1 - 45\epsilon + o(\epsilon))/2$ and that $\Pr\{X_J = 2\} = (1 + 45\epsilon + o(\epsilon))/2$. Verify, to first order in ϵ that (7.125) is satisfied.

Exercise 7.28. Show that (7.125) results from taking the derivative of (7.129) and evaluating it at $r = 0$.

Exercise 7.29. Let $\{Z_n;\, n \geq 1\}$ be a martingale, and for some integer m, let $Y_n = Z_{n+m} - Z_m$.

a) Show that $\mathsf{E}[Y_n \mid Z_{n+m-1} = z_{n+m-1}, Z_{n+m-2} = z_{n+m-2}, \dots, Z_m = z_m, \dots, Z_1 = z_1] = z_{n+m-1} - z_m$.

b) Show that $\mathsf{E}[Y_n \mid Y_{n-1} = y_{n-1}, \dots, Y_1 = y_1] = y_{n-1}$

c) Show that $\mathsf{E}[|Y_n|] < \infty$. Note that **b)** and **c)** show that $\{Y_n;\, n \geq 1\}$ is a martingale.

Exercise 7.30. a) Show that Theorem 7.7.1 is valid if $\{Z_n;\, n \geq 1\}$ is a submartingale rather than a martingale. Hint: Simply follow the proof of Theorem 7.7.1 in the text.

b) Show that the Kolmogorov martingale inequality also holds if $\{Z_n;\, n \geq 1\}$ is a submartingale rather than a martingale.

Exercise 7.31 (Continuation of continuous-time branching). This exercse views the continuous-time branching process of Exercise 6.15 as a stopped random walk. Recall that the process was specified there as a Markov process such that for each state j, $j \geq 0$, the transition rate to $j+1$ is $j\lambda$ and to $j-1$ is $j\mu$. There are no other transitions, and in particular, there are no transitions out of state 0, so that the Markov process is reducible. Recall that the embedded Markov chain is the same as the embedded chain of an M/M/1 queue except that there is no transition from state 0 to state 1.

a) To model the possible extinction of the population, convert the embedded Markov chain abve to a stopped random walk, $\{S_n;\, n \geq 0$. The stopped random walk starts at $S_0 =$

0 and stops on reaching a threshold at -1. Before stopping, it moves up by one with probability $\frac{\lambda}{\lambda+\mu}$ and downward by 1 with probability $\frac{\mu}{\lambda+\mu}$ at each step. Give the (very simple) relationship between the state X_n of the Markov chain and the state S_n of the stopped random walk for each $n \geq 0$.

b) Find the probability that the population eventually dies out as a function of λ and μ. Be sure to consider all three cases $\lambda > \mu$, $\lambda < \mu$, and $\lambda = \mu$.