Chapter 14
Discrete Time Martingales and Concentration Inequalities

For an independent sequence of random variables $X_1, X_2, \ldots$, the conditional expectation of the present term of the sequence given the past terms is the same as its unconditional expectation. Martingales let the conditional expectation depend on the past terms, but in a special way. Thus, similar to Markov chains, martingales act as natural models for incorporating dependence into a sequence of observed data. But the value of the theory of martingales is much more than simply its modeling value. Martingales arise, as natural byproducts of the mathematical analysis in an amazing variety of problems in probability and statistics. Therefore, results from martingale theory can be immediately applied to all these situations in order to make deep and useful conclusions about numerous problems in probability and statistics. A particular modern set of applications of martingale methods is in the area of concentration inequalities, which place explicit bounds on probabilities of large deviations of functions of a set of variables from their mean values. This chapter gives a glimpse into some important concentration inequalities, and explains how martingale theory enters there. Martingales form a nearly indispensable tool for probabilists and statisticians alike.

Martingales were introduced into the probability literature by Paul Lévy, who was interested in finding situations beyond the iid case where the strong law of large numbers holds. But its principal theoretical studies were done by Joseph Doob. Two extremely lucid expositions on martingales are Doob (1971) and Heyde (1972). Some other excellent references for this chapter are Karlin and Taylor (1975), Chung (1974), Hall and Heyde (1980), Williams (1991), Karatzas and Shreve (1991), Fristedt and Gray (1997), and Chow and Teicher (2003). Other references are provided in the sections.

14.1 Illustrative Examples and Applications in Statistics

We start with a simple example, which nevertheless captures the spirit of the idea of a martingale sequence of random variables.

A. DasGupta, *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics*, Springer Texts in Statistics, DOI 10.1007/978-1-4419-9634-3_14,

Example 14.1 (Gambler's Fortune). Consider a gambler repeatedly playing a fair game in a casino. Thus, a fair coin is tossed. If heads show, the player wins \$1; if it is tails, the house wins \$1. He plays repeatedly. Let $X_1, X_2, \ldots$ be the players's sequence of wins. Thus, the X_i are iid with the common distribution $P(X_i = \pm 1) = \frac{1}{2}$. The player's fortune after n plays is $S_n = S_0 + \sum_{i=1}^{n} X_i, n \geq 1$. If we take the player's initial fortune S_0 to be just zero, then $S_n = \sum_{i=1}^{n} X_i$. Suppose now the player has finished playing for n times, and he is looking ahead to what his fortunes will be after he plays the next time. In other words, he wants to find $E(S_{n+1} \,|S_1, \ldots, S_n)$. But,

$$\begin{aligned} & E(S_{n+1} \,|S_1, \ldots, S_n) \\ &= E(S_n + X_{n+1} \,|S_1, \ldots, S_n) = S_n + E(X_{n+1} \,|S_1, \ldots, S_n) \\ &= S_n + E(X_{n+1}) = S_n + 0 = S_n. \end{aligned}$$

In the above, $E(X_{n+1} \,|S_1, \ldots, S_n)$ equals the unconditional expectation of X_{n+1} because X_{n+1} is independent of $(X_1, X_2, \ldots, X_n)$, and hence, independent of $(S_1, \ldots, S_n)$.

Notice that the sequence of fortunes $S_1, S_2, \ldots$ is not an independent sequence. There is information in the past sequence of fortunes for predicting the current fortune. But the players's forecast for what his fortune will be after the next round of play is simply what his fortunes are right now, no more and no less. This is basically what the martingale property means, and is the reason for equating martingales with *fair games*.

Here is the definition. Rigorous treatment of martingales requires use of measure theory. For the most part, our treatment avoids measure-theory terminology.

Definition 14.1. Let $X_n, n \geq 1$ be a sequence of random variables defined on a common sample space Ω such that $E(|X_n|) < \infty$ for all $n \geq 1$. The sequence $\{X_n\}$ is called a *martingale adapted to itself* if for each $n \geq 1$, $E(X_{n+1} \,|X_1, X_2, \ldots, X_n) = X_n$ with probability one.

The sequence $\{X_n\}$ is called a *supermartingale* if for each $n \geq 1$, $E(X_{n+1} \,|X_1, X_2, \ldots, X_n) \leq X_n$ with probability one. The sequence $\{X_n\}$ is called a *submartingale* if for each $n \geq 1$, $E(X_{n+1} \,|X_1, X_2, \ldots, X_n) \geq X_n$ with probability one.

Remark. We generally do not mention the *adapted to itself* qualification when that is indeed the case. It is sometimes useful to talk about the martingale property with respect to a different sequence of random variables. This concept is defined below and Example 14.8 is an example of such a martingale sequence.

Note that X_n is a submartingale if and only if $-X_n$ is a supermartingale, and that it is a martingale if and only if it is both a supermartingale and a submartingale. Also notice that for a martingale sequence $X_n, E(X_{n+m}) = E(X_n)$ for all $n, m \geq 1$; in other words, $E(X_n) = E(X_1)$ for all n.

Definition 14.2. Let $X_n, n \geq 1$ and $Y_n, n \geq 1$ be sequences of random variables defined on a common sample space Ω such that $E(|X_n|) < \infty$ for all $n \geq 1$.

The sequence $\{X_n\}$ is called a *martingale adapted to the sequence* $\{Y_n\}$ if for each $n \geq 1$, X_n is a function of $Y_1, \ldots, Y_n$, and $E(X_{n+1}|Y_1, Y_2, \ldots, Y_n) = X_n$ with probability one.

Some elementary examples are given first.

Example 14.2 (Partial Sums). Let $Z_1, Z_2, \ldots$ be independent zero mean random variables, and let S_n denote the partial sum $\sum_{i=1}^{n} Z_i$. Then, clearly, $E(S_{n+1}|S_1, \ldots, S_n) = S_n + E(Z_{n+1}|S_1, \ldots, S_n) = S_n + E(Z_{n+1}) = S_n$, and so $\{S_n\}$ forms a martingale. More generally, if the common mean of the Z_i is some number μ, then $S_n - n\mu$ is a martingale.

Example 14.3 (Sums of Squares). Let $Z_1, Z_2, \ldots$ be iid $N(0, 1)$ random variables, and let $X_n = (Z_1 + \cdots + Z_n)^2 - n = S_n^2 - n$, where $S_n = Z_1 + \cdots + Z_n$. Then,

$$\begin{aligned}
&E(X_{n+1}|X_1, X_2, \ldots, X_n) \\
&= E[(Z_1 + \cdots + Z_n)^2 + 2Z_{n+1}(Z_1 + \cdots + Z_n) \\
&\quad + Z_{n+1}^2|X_1, X_2, \ldots, X_n] - (n+1) \\
&= X_n + n + 2(Z_1 + \cdots + Z_n)E(Z_{n+1}|X_1, X_2, \ldots, X_n) \\
&\quad + E(Z_{n+1}^2|X_1, X_2, \ldots, X_n) - (n+1) \\
&= X_n + n + 0 + 1 - (n+1) = X_n,
\end{aligned}$$

and so $\{X_n\}$ forms a martingale sequence.

Actually, we did not use the normality of the Z_i at all, and the martingale property holds without the normality assumption. That is, if $Z_1, Z_2, \ldots$ are iid with mean zero and variance σ^2, then $S_n^2 - n\sigma^2$ is a martingale.

Example 14.4. Suppose $X_1, X_2, \ldots$ are iid $N(0, 1)$ variables and $S_n = \sum_{i=1}^{n} X_i$. Because $S_n \sim N(0, n)$, its mgf $E(e^{tS_n}) = e^{nt^2/2}$. Now let $Z_n = e^{tS_n - nt^2/2}$, where t is a fixed real number. Then, $E(Z_{n+1}|Z_1, \ldots, Z_n) = e^{-(n+1)t^2/2} E(e^{tS_n} e^{tX_{n+1}}|S_n) = e^{-(n+1)t^2/2} e^{tS_n} e^{t^2/2} = Z_n$. Therefore, for any real t, the sequence $e^{tS_n - nt^2/2}$ forms a martingale.

Once again, a generalization beyond the normal case is possible; see the chapter exercises for a general result.

Example 14.5 (Matching Problem). Consider the matching problem. For example, suppose N people, each wearing a hat, have gathered in a party and at the end of the party, the N hats are returned to them at random. Those that get their own hats back then leave the room. The remaining hats are distributed among the remaining guests at random, and so on. The process continues until all the hats have been given away. Let X_n denote the number of guests still present after the nth round of this hat returning process.

At each round, we expect one person to get his own hat back and leave the room. In other words, $E(X_n - X_{n+1}) = 1 \quad \forall n$. In fact, with a little calculation, we even have

$$\begin{aligned} E(X_{n+1}|X_1, \ldots, X_n) &= E(X_{n+1} - X_n + X_n|X_1, \ldots, X_n) \\ &= E(X_{n+1} - X_n|X_1, \ldots, X_n) + X_n = -1 + X_n. \end{aligned}$$

This immediately implies that $E(X_{n+1}+n+1\,|X_1, \ldots, X_n) = -1+(n+1)+X_n = X_n + n$. Hence the sequence $\{X_n + n\}$ is a martingale.

Example 14.6 (Pólya's Urn). The Pólya urn scheme is defined as follows. Initially, an urn contains a white and b black balls, a total of $a + b$ balls. One ball is drawn at random from among all the balls in the urn. It, together with c more balls of its color is returned to the urn, so that after the first draw, the urn has $a + b + c$ balls. This process is repeated.

Suppose X_i is the indicator of the event A_i that a white ball is drawn at the ith trial, and for given $n \geq 1, S_n = X_1 + \cdots + X_n$, which is the total number of times that a white ball has been drawn in the first n trials. For the sake of notational simplicity, we take $c = 1$. Then, the proportion of white balls in the urn just after the nth trial has been completed is $R_n = \frac{a+S_n}{a+b+n}$.

Elementary arguments show that

$$P(X_{n+1} = 1\,|X_1 = x_1, \ldots, X_n = x_n) = \frac{a + x_1 + \cdots + x_n}{a + b + n}.$$

Thus,

$$\begin{aligned} E(S_{n+1}\,|S_1, \ldots, S_n) &= E(S_{n+1}\,|S_n) = S_n + \frac{a + S_n}{a + b + n} \\ \Rightarrow E(R_{n+1}\,|R_1, \ldots, R_n) &= \frac{a}{a + b + n + 1} + \frac{1}{a + b + n + 1} \\ &\quad [(a + b + n)R_n - a + R_n] = R_n. \end{aligned}$$

We therefore have the interesting result that in the Pólya urn scheme, the sequence of proportions of white balls in the urn forms a martingale.

Example 14.7 (The Wright–Fisher Markov Chain). Consider the stationary Markov chain $\{X_n\}$ on the state space $\{0, 1, 2, \ldots, N\}$ with the one-step transition probabilities

$$p_{ij} = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}.$$

This is the Wright–Fisher chain in population genetics (see Chapter 10). We show that X_n is a martingale adapted to itself. Indeed, by direct calculation,

$$E(X_{n+1}\,|X_1,\ldots,X_n) = E(X_{n+1}\,|X_n)$$
$$= \sum_{j=0}^{N} j\binom{N}{j}\left(\frac{X_n}{N}\right)^j\left(1-\frac{X_n}{N}\right)^{N-j} = N\frac{X_n}{N} = X_n.$$

Example 14.8 (Likelihood Ratios). Suppose $X_1, X_2, \ldots, X_n$ are iid with a common density function f, which is one of f_0, and f_1, two different density functions. The statistician is supposed to choose from the two densities f_0, f_1, the one that is truly generating the observed data $x_1, x_2, \ldots, x_n$. One therefore has the *null hypothesis* H_0 that $f = f_0$, and the *alternate hypothesis* that $f = f_1$. The statistician's decision is commonly based on the *likelihood ratio*

$$\Lambda_n = \prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)}.$$

If Λ_n is large for the observed data, then one concludes that the data values come from a high-density region of f_1 and a low-density region of f_0, and therefore concludes that the true f generating the observed data is f_1.

Suppose now the null hypothesis is actually true; that is, truly, $X_1, X_2, \ldots$ are iid with the common density f_0. Now,

$$E_{f_0}[\Lambda_{n+1}\,|\Lambda_1,\ldots,\Lambda_n] = E_{f_0}\left[\frac{f_1(X_{n+1})}{f_0(X_{n+1})}\Lambda_n\,|\Lambda_1,\ldots,\Lambda_n\right]$$
$$= \Lambda_n E_{f_0}\left[\frac{f_1(X_{n+1})}{f_0(X_{n+1})}\,|\Lambda_1,\ldots,\Lambda_n\right]$$
$$= \Lambda_n E_{f_0}\left[\frac{f_1(X_{n+1})}{f_0(X_{n+1})}\right]$$

(because the sequence $X_1, X_2, \ldots$ are independent)

$$= \Lambda_n \int_{\mathcal{R}} \frac{f_1(x)}{f_0(x)} f_0(x)dx = \Lambda_n \int_{\mathcal{R}} f_1(x)dx$$
$$= \Lambda_n \times 1 = \Lambda_n.$$

Therefore, the sequence of likelihood ratios forms a martingale under the null hypothesis (i.e., if the true f is f_0).

Example 14.9 (Bayes Estimates). Suppose random variables $Y, X_1, X_2, \ldots$ are defined on a common sample space Ω. For given $n \geq 1$, $(X_1, X_2, \ldots, X_n)$ has the joint conditional distribution $P_{\theta,n}$ given that $Y = \theta$. From a statistical point of view, Y is supposed to stand for an unknown parameter, which is formally treated

as a random variable, and $X^{(n)} = (X_1, X_2, \ldots, X_n)$ for some specific n, namely the actual sample size, is the data that the statistician has available to estimate the unknown parameter. The *Bayes estimate* of the unknown paramter is the posterior mean $E(Y\,|X^{(n)})$ (see Chapter 3).

Denote for each $n \geq 1$, $E(Y\,|X^{(n)}) = Z_n$. We show that Z_n forms a martingale sequence with respect to the sequence $X^{(n)}$; that is, $E(Z_{n+1}\,|X^{(n)}) = Z_n$. However, this follows on simply observing that by the iterated expectation formula,

$$Z_n = E\left(Y\,|X^{(n)}\right) = E_{X_{n+1}\,|X^{(n)}}\left[E\left(Y|X^{(n)}, X_{n+1}\right)\right] = E\left(Z_{n+1}\,|X^{(n)}\right).$$

Example 14.10 (Square of a Martingale). Suppose X_n, defined on some sample space Ω is a positive submartingale sequence. For simplicity, let us consider the case when it is adapted to itself. Thus, for any $n \geq 1$, $E(X_{n+1}\,|X_1, \ldots, X_n) \geq X_n$ (with probability one). Therefore, for any $n \geq 1$,

$$\begin{aligned} E(X_{n+1}^2\,|X_1, \ldots, X_n) &\geq [E(X_{n+1}\,|X_1, \ldots, X_n)]^2 \\ &\geq X_n^2. \end{aligned}$$

Therefore, if we let $Z_n = X_n^2$, then Z_n is a submartingale sequence.

If we inspect this example carefully, then we realize that we have only used a very special case of Jensen's inequality to establish the needed submartingale property for the Z_n sequence. Furthermore, if the original $\{X_n\}$ sequence is a martingale, rather than a submartingale, then the positivity restriction on the X_n is no longer necessary. Thus, by simply following the steps of this example, we in fact have the following simple but widely useful general result.

Theorem 14.1 (Convex Function Theorem). *Let $X_n, n \geq 1$ be defined on a common sample space Ω. Let f be a convex function on $\mathcal{R}$, and let $Z_n = f(X_n)$.*

(a) Suppose $\{X_n\}$ is a martingale adapted to some sequence *$\{Y_n\}$. Then $\{Z_n\}$ is a submartingale adapted to $\{Y_n\}$.*
(b) Suppose $\{X_n\}$ is a submartingale adapted to some sequence $\{Y_n\}$. Assume that f is in addition nondecreasing. Then $\{Z_n\}$ is a submartingale adapted to $\{Y_n\}$.

14.2 Stopping Times and Optional Stopping

The *optional stopping theorem* is one of the most useful results in martingale theory. It can be explained in gambling terms. Consider a gambler playing a fair game repeatedly, so that her sequence of fortunes forms a martingale. One might think that by gaining experience as the game proceeds, and by quitting at a cleverly chosen opportune time based on the gambler's experience, a fair game could be turned into a favorable game. The optional stopping theorem says that this is in fact not possible, if the gambler does not have unlimited time on her hands and the house

has limits on how much she can put up on the table. Mathematical formulation of the optional stopping theorem requires use of *stopping times*, which were introduced in Chapter 11 in the context of random walks. We redefine stopping times and give additional examples below before introducing optional stopping.

14.2.1 Stopping Times

Definition 14.3. Let $X_1, X_2, \ldots$ be a sequence of random variables, all defined on a common sample space Ω. Let τ be a nonnegative integer-valued random variable, also defined on Ω. We call τ a *stopping time adapted to the sequence* $\{X_n\}$ if $P(\tau < \infty) = 1$, and if for each $n \geq 1$, $I_{\{\tau \leq n\}}$ is a function of only $X_1, X_2, \ldots, X_n$.

In other words, τ is a stopping time adapted to $\{X_n\}$ if for any $n \geq 1$, whether or not $\tau \leq n$ can be determined by only knowing $X_1, X_2, \ldots, X_n$, and provided that τ cannot be infinite with a positive probability.

We have seen some examples of stopping times in Chapter 11. We start with a few more illustrative examples.

Example 14.11 (Sequential Tests in Statistics). Suppose to start with we have an infinite sequence of random variables $X_1, X_2, \ldots$ on a common sample space Ω, and let S_n denote the nth partial sum, $S_n = \sum_{i=1}^{n} X_i, n \geq 1$. The X_n need not be independent. Fix numbers $-\infty < l < u < \infty$. Then τ defined as

$$\tau = \inf\{n : S_n < l \text{ or } S_n > u\},$$

and $\tau = \infty$ if $l \leq S_n \leq u \; \forall n \geq 1$, is a stopping time adapted to the sequence $\{S_n\}$.

A particular case of this arises in sequential testing of hypotheses in statistics. Suppose an original sequence $Z_1, Z_2, \ldots$ is iid from some density f, which equals either f_0 or f_1. Then, as we have seen above, the *likelihood ratio* is

$$\Lambda_n = \frac{\prod_{i=1}^{n} f_1(Z_i)}{\prod_{i=1}^{n} f_0(Z_i)}.$$

The *Wald sequential probability ratio test* (SPRT) continues sampling as long as Λ_n remains between two specified numbers a and $b, a < b$, and stops and decides in favor of f_1 or f_0 the first time $\Lambda_n > b$ or $< a$. If we denote $l = \log a, u = \log b$, then Wald's test waits till the first time $\log \Lambda_n = \sum_{i=1}^{n} \log \frac{f_1(Z_i)}{f_0(Z_i)} = \sum_{i=1}^{n} X_i$ (say) goes above u or below l, and thus the *sampling number* of Wald's SPRT is a stopping time.

Example 14.12 (Combining Stopping Times). This example shows a few ways that we can make new stopping times out of given ones. Suppose τ is a stopping time (adapted to some sequence $\{X_n\}$) and n is a prespecified positive integer. Then $\tau_n = \min(\tau, n)$ is a stopping time (adapted to the same sequence). This is because

$$\{\tau_n \leq k\} = \{\tau \leq k\} \cup \{n \leq k\},$$

and therefore, τ being a stopping time adapted to $\{X_n\}$, for any given k, deciding whether $\tau_n \leq k$ requires the knowledge of only $X_1, \ldots, X_k$.

Suppose τ_1, τ_2 are both stopping times, adapted to some sequence $\{X_n\}$. Then $\tau_1 + \tau_2$ is also a stopping time adapted to the same sequence. To prove this, note that

$$\{\tau_1 + \tau_2 \leq k\} = \cup_{i=0}^{k} \cup_{j=0}^{i} \{\tau_1 = j, \tau_2 = i - j\} = \cup_{i=0}^{k} \cup_{j=0}^{i} A_{ij}.$$

and whether any A_{ij} occurs depends only on $X_1, \ldots, X_k$.

For the sake of reference, we collect a set of such facts about stopping times in the next result. They are all easy to prove.

Theorem 14.2. *(a) Let τ be a stopping time adapted to some sequence $\{X_n\}$. Then, for any given $n \geq 1$, $\min(\tau, n)$ is also a stopping time adapted to $\{X_n\}$.*

(b) Let τ_1, τ_2 be stopping times adapted to $\{X_n\}$. Then each of $\tau_1 + \tau_2, \min(\tau_1, \tau_2)$, $\max(\tau_1, \tau_2)$ is a stopping time adapted to $\{X_n\}$.

(c) Let $\{\tau_k, k \geq 1\}$ be a countable family of stopping times, each adapted to $\{X_n\}$. Let

$$\underline{\tau} = \inf_k \tau_k; \overline{\tau} = \sup_k \tau_k; \tau = \lim_{k \to \infty} \tau_k,$$

where $\underline{\tau}, \overline{\tau}$, and τ are defined pointwise, and it is assumed that the limit τ exists almost surely. Then each of $\underline{\tau}, \overline{\tau}$ and τ is a stopping time adapted to $\{X_n\}$.

14.2.2 Optional Stopping

The most significant derivative of introducing the concept of stopping times is the *optional stopping theorem*. At the expense of using some potentially hard to verify conditions, stronger versions of our statement of the optional stopping theorem can be stated. We choose to opt for simplicity of the statement over greater generality, and refer to more general versions (which are useful!). The main message of the optional stopping theorem is that a gambler cannot convert a fair game into a favorable one by using clever quitting strategies.

Theorem 14.3 (Optional Stopping Theorem). *Let $\{X_n, n \geq 0\}$ be a submartingale adapted to some sequence $\{Y_n\}$, and τ a stopping time adapted to the same sequence. For $n \geq 0$, let $\tau_n = \min(\tau, n)$. Then $\{X_{\tau_n}\}$ is also a submartingale adapted to $\{Y_n\}$, and for each $n \geq 0$,*

$$E(X_0) \leq E(X_{\tau_n}) \leq E(X_n).$$

In particular, if

$$\{X_n\} \text{ is a martingale, } E(|X_\tau|) < \infty, \text{ and } \lim_{n \to \infty} E(X_{\tau_n}) = E(X_\tau),$$

then

$$E(X_\tau) = E(X_0).$$

Remark. It is of course unsatisfactory to simply demand that $E(|X_\tau|) < \infty$ and $\lim_{n\to\infty} E(X_{\tau_n}) = E(X_\tau)$. What we need are simple sufficient conditions that a user can verify relatively easily. This is addressed following the proof of the above theorem.

Proof of Theorem. For simplicity, we give the proof only for the case when $\{X_n\}$ is adapted to itself. The main step involved is to notice the identity

$$W_n = X_{\tau_n} = \sum_{i=0}^{n-1} X_i I_{\{\tau=i\}} + X_n I_{\{\tau\geq n\}}, \quad (*)$$

for all $n \geq 0$. It follows from this identity and the submartingale property of the $\{X_n\}$ sequence that

$$\begin{aligned}
&E(W_{n+1}\,|X_0,\ldots,X_n)\\
&= \sum_{i=0}^{n} E(X_i I_{\{\tau=i\}}\,|X_0,\ldots,X_n) + E(X_{n+1} I_{\{\tau>n\}}\,|X_0,\ldots,X_n)\\
&= \sum_{i=0}^{n} X_i I_{\{\tau=i\}} + I_{\{\tau>n\}} E(X_{n+1}\,|X_0,\ldots,X_n)\\
&\geq \sum_{i=0}^{n} X_i I_{\{\tau=i\}} + X_n I_{\{\tau>n\}} = X_{\tau_n} = W_n.
\end{aligned}$$

Thus, as claimed, $W_n = \{X_{\tau_n}\}$ is a submartinagle adapted to the original $\{X_n\}$ sequence. It follows that

$$E(X_{\tau_n}) = E(W_n) \geq E(W_0) = E(X_0).$$

To complete the proof of the theorem, we need the reverse inequality $E(W_n) \leq E(X_n)$. This too follows from the same identity $(*)$ given at the beginning of the proof of this theorem, and on using the additional inequality

$$\begin{aligned}
&E(X_n I_{\{\tau=i\}}\,|X_0,\ldots,X_i)\\
&= I_{\{\tau=i\}} E(X_n\,|X_0,\ldots,X_i) \geq I_{\{\tau=i\}} X_i,
\end{aligned}$$

because $\{X_n\}$ is a submartingale. If this bound on $X_i I_{\{\tau=i\}}$ is plugged into our basic identity $(*)$ above, the reverse inequality follows.

The remaining claim, when $\{X_n\}$ is in fact a martingale, follows immediately from the two inequalities $E(X_0) \leq E(W_n) \leq E(X_n)$. □

14.2.3 Sufficient Conditions for Optional Stopping Theorem

Easy examples show that the assertion $E(X_\tau) = E(X_0)$ for a martingale sequence $\{X_n\}$ cannot hold without some control on the stopping time τ. We first provide a simple example where the assertion of the optional stopping theorem fails. In looking for such counterexamples, it is useful to construct the stopping time in a way that when we stop, the value of the stopped martingale is a constant; that is, X_τ is a constant.

Example 14.13 (An Example Where the Optional Stopping Theorem Fails). Consider again the gambling example, or what really is the simple symmetric random walk, with X_i iid having the common distribution $P(X_i = \pm 1) = \frac{1}{2}$, and $S_n = \sum_{i=1}^n X_i, n \geq 1$. We define $S_0 = 0$. We know S_n to be a martingale. Consider now the stopping time

$$\tau = \inf\{n > 0 : S_n = 1\}.$$

We know from Chapter 11 that the one-dimensional simple symmetric random walk is recurrent; thus, $P(\tau < \infty) = 1$. Note that $S_\tau = 1$, and so, $E(S_\tau) = 1$. However, $E(S_0) = E(S_n) = 0$. So, the assertion of the optional stopping theorem does not hold.

What is going on in this example is that we do not have enough control on the stopping time τ. Although the random walk visits all its states (infinitely often) with probability one, the recurrence times are infinite on the average. Thus, τ can be uncontrollably large. Indeed, the assumption

$$\lim_{n\to\infty} E(S_{\min(\tau,n)}) = E(S_\tau)(= 1)$$

does not hold. Roughly speaking, $P(\tau > n)$ goes to zero at the rate $\frac{1}{\sqrt{n}}$ and if the random walk still has not reached positive territory by time n, then it has traveled to some distance roughly of the order of $-\sqrt{n}$. These two now exactly balance out, so that $E(S_{\min(\tau,n)})I_{\{\tau>n\}}$ does not go to zero. This causes the assumption $\lim_{n\to\infty} E(S_{\min(\tau,n)}) = E(S_\tau) = 1$ to fail.

Thus, our search for sufficient conditions in the optional stopping theorem should be directed at finding nice enough conditions that ensure that the stopping time τ cannot get too large with a high probability. The next two theorems provide such a set of aesthetically attractive sufficient conditions. It is not hard to prove these two theorems. We refer the reader to Fristedt and Gray (1997, Chapter 24) for proofs of these two theorems.

Theorem 14.4. *Suppose $\{X_n, n \geq 0\}$ is a martingale, adapted to itself, and τ a stopping time adapted to the same sequence. Suppose any one of the following conditions holds.*

(a) For some $n < \infty$, $P(\tau \leq n) = 1$.
(b) For some nonnegative random variable Z with $E(Z) < \infty$, the martingale sequence $\{X_n\}$ satisfies $|X_n| \leq Z$ for all $n \geq 0$.

(c) For some positive and finite c, $|X_{n+1} - X_n| \le c$ *for all* $n \ge 0$, $E(|X_0|) < \infty$, *and* $E(\tau) < \infty$.
(d) For some finite constant c, $E(X_n^2) \le c$ *for all* $n \ge 0$.

Then $E(X_\tau) = E(X_0)$.

Remark. It is important to keep in mind that none of these four conditions is necessary for the equality $E(X_\tau) = E(X_0)$ to hold. We recall from our discussion of *uniform integrability* in Chapter 7 that conditions (b) and (d) in Theorem 14.4 each imply that the sequence $\{X_n\}$ is uniformly integrable. In fact, it may be shown that under the *weaker condition* that our martingale sequence $\{X_n\}$ is uniformly integrable, the equality $E(X_\tau) = E(X_0)$ holds. The important role played by uniform integrability in martingale theory reappears when we discuss convergence of martingales.

An important case where the equality holds with essentially the minimum requirements is the special case of a random walk. We precisely describe this immediately below. The point is that the four sufficient conditions are all-purpose conditions. But if the martingale has a special structure, then the conditions can sometimes be weakened. Here is such a result for a special martingale, namely the random walk.

Theorem 14.5. *Let* $Z_1, Z_2, \ldots$ *be an iid sequence such that* $E(|Z_1|) < \infty$. *Let* $S_n = \sum_{i=1}^n Z_i, n \ge 1$. *Let* τ *be any stopping time adapted to* $\{S_n\}$ *such that* $E(\tau) < \infty$. *Consider the martingale sequence* $X_n = S_n - n\mu, n \ge 1$, *where* $\mu = E(Z_1)$. *Then the equality* $E(X_\tau) = E(X_1) = 0$ *holds.*

Remark. The special structure of the random walk martingale allows us to conclude the assertion of the optional stopping theorem, without requiring the bounded increments condition $|X_{n+1} - X_n| \le c$, which was included in the all-purpose sufficient condition in Theorem 14.4.

Example 14.14 (Weighted Rademacher Series). Let $X_1, X_2, \ldots$ be a sequence of iid *Rademacher variables* with common distribution $P(X_i = \pm 1) = \frac{1}{2}$. For $n \ge 1$, let $S_n = \sum_{i=1}^n \frac{X_i}{i^\alpha}$, where $\alpha > \frac{1}{2}$. Because X_i are independent and $E(\frac{X_i}{i^\alpha}) = 0$ for all i, S_n forms a martingale sequence (see Example 14.2). On the other hand,

$$E(S_n^2) = \text{Var}(S_n) = \sum_{i=1}^n \frac{\text{Var}(X_i)}{i^{2\alpha}}$$
$$= \sum_{i=1}^n \frac{1}{i^{2\alpha}} \le \sum_{i=1}^\infty \frac{1}{i^{2\alpha}} = \zeta(2\alpha) < \infty,$$

where $\zeta(z)$ is the Riemann zeta function $\zeta(z) = \sum_{n=1}^\infty \frac{1}{n^z}, z > 1$. Therefore, if $\alpha > \frac{1}{2}$, $E(S_n^2) \le c = \zeta(2\alpha)$ for all n, and hence by our theorem above, $E(S_\tau) = 0$ holds for any stopping time τ adapted to $\{S_n\}$.

Example 14.15 (The Simple Random Walk). Consider the one-dimensional random walk with iid steps X_i, having the common distribution $P(X_i = 1) = p, P(X_i = -1) = q, 0 < p < 1, p + q = 1$. Then, $E(X_i) = p - q = \mu$ (say), and $S_n - n\mu$, where $S_n = \sum_{i=1}^{n} X_i$, is a martingale. We also have, for any n,

$$|S_{n+1} - (n+1)\mu - (S_n - n\mu)| = |X_{n+1} - \mu| \le 2.$$

Furthermore, $E(|S_1 - \mu|)$ is clearly finite. Therefore, for any stopping time τ with a finite expectation, by using our theorem above, the equality $E(S_\tau - \mu\tau) = 0$, or equivalently, $E(S_\tau) = \mu E(\tau)$ holds. Recall from Chapter 11 that this is a special case of *Wald's identity*. Wald's identity is revisited in the next section.

14.2.4 Applications of Optional Stopping

We provide a few applications of the optional stopping theorem. The optional stopping theorem also has important applications to martingale inequalities, which is our topic in the next section.

Perhaps the two best general applications of the optional stopping theorem are two identities, known as *Wald identities*. Of these, the first Wald identity is already known to us; see Chapter 11. We connect that identity to martingale theory and present a second identity, which was not presented in Chapter 11.

Theorem 14.6 (Wald's First and Second Identity). *Let $X_1, X_2, \ldots$ be a sequence of iid random variables, defined on a common sample space Ω. Let $S_n = \sum_{i=1}^{n} X_i, n \ge 1$. Let τ be a stopping time adapted to the sequence $\{S_n\}$ and suppose that $E(\tau) < \infty$.*

(a) Suppose $E(|X_1|) < \infty$ and $E(X_1) = \mu$ (which need not be zero). Then $E(S_\tau) = \mu(E\tau)$.
(b) Suppose $E(X_1) = 0, E(X_1^2) = \sigma^2 < \infty$. Then $Var(S_\tau) = \sigma^2(E\tau)$.

Proof. Both parts of this theorem follow from Theorem 14.5. For part (a), apply Theorem 14.5 to the martingale sequence $S_n - n\mu$ to conclude that $E(S_\tau - \tau\mu) = 0 \Rightarrow E(S_\tau) = \mu(E\tau)$. For part (b), because $\mu = E(X_1)$ has now been assumed to be zero, by applying part (a) of this theorem,

$$\text{Var}(S_\tau) = E(S_\tau - E[S_\tau])^2 = E(S_\tau - 0)^2 = E(S_\tau^2).$$

Next note that because the X_i are independent,

$$\begin{aligned} \text{Var}(S_{n+1} \,|S_1, \ldots, S_n) &= \text{Var}(X_{n+1}) = \sigma^2 \\ \Rightarrow E(S_{n+1}^2 \,|S_1, \ldots, S_n) &= S_n^2 + \sigma^2 \\ \Rightarrow E(S_{n+1}^2 - (n+1)\sigma^2 \,|S_1, \ldots, S_n) &= S_n^2 - n\sigma^2; \end{aligned}$$

that is, $S_n^2 - n\sigma^2$ is a martingale sequence adapted to the S_n sequence. From here, it follows that $E(S_\tau^2 - \tau\sigma^2) = E(S_1^2 - \sigma^2) = 0$, which means

$$\text{Var}(S_\tau) = E(S_\tau^2) = \sigma^2(E\tau),$$

which is what part (b) says. □

Example 14.16 (Expected Hitting Times for a Random Walk). The Wald identity may be used to evaluate the expected hitting time of a given level by a random walk. Specifically, let S_n be the one-dimensional simple symmetric random walk with the iid steps having the common distribution $P(X_i = \pm 1) = \frac{1}{2}$, Let x be any given positive integer and consider the first passage time

$$\tau_x = \inf\{n > 0 : S_n = x\}.$$

We know from general random walk theory (Chapter 11) that $P(\tau_x < \infty) = 1$. Also, obviously $E(|X_1|) = 1 < \infty$, and $\mu = E(X_1) = 0$. Therefore, if $E(\tau_x)$ is finite, Wald's identity $E(S_{\tau_x}) = \mu E(\tau_x)$ will hold. However, $S_{\tau_x} = x$ with probability one, and hence, $E(S_{\tau_x}) = x$. It follows that the equality $x = 0 \times E(\tau_x)$ cannot hold for any finite value of $E(\tau_x)$. In other words, for any positive x, the expected hitting time of x must be infinite for the simple symmetric random walk. The same argument also works for negative x.

Example 14.17 (Gambler's Ruin). Now let us revisit the so-called *gambler's ruin problem*, wherein the gambler quits when he either goes broke, or attains a prespecified amount of fortune (see Chapter 10). In other words, the gambler waits for the random walk S_n to hit one of two integers $0, b, b > 0$. Suppose $a < b$ is the amount of money with which the gambler walked in, so that the gambler's sequence of fortunes is the random walk $S_n = \sum_{i=1}^n X_i + S_0$, where $S_0 = a$, and the steps are still iid with $P(X_i = \pm 1) = \frac{1}{2}$. Formally, let

$$\tau = \tau_{\{a,b\}} = \inf\{n > 0 : S_n \in \{0, b\}\}.$$

By applying the optional stopping theorem,

$$E(S_\tau) = 0 \times P(S_\tau = 0) + b[1 - P(S_\tau = 0)] = E(S_0) = a;$$

note that we have implicitly assumed the validity of the optional stopping theorem in the last step (which is true in this example; why?). Rearranging terms, we deduce that $P(S_\tau = 0) = \frac{b-a}{b}$, or equivalently, $P(S_\tau = b) = \frac{a}{b}$.

Example 14.18 (Generalized Wald Identity). The two identities of Wald given above assume only the existence of the first and the second moment of X_i, respectively. If we make the stronger assumption that the X_i have a finite mgf, then a more embracing martingale identity can be proved, from which the two Wald identities given above fall out as special cases. This generalized Wald identity is presented in this example.

The basic idea is the same as before, which is to think of a suitable martingale, and apply the optional stopping theorem to it. Suppose then that $X_1, X_2, \ldots$ is an iid sequence, with the mgf $\psi(t) = E(e^{tX_i})$, which we assume to be finite in some nonempty interval containing zero. The martingale that works for our purpose in this example is

$$Z_n = [\psi(t)]^{-n} e^{tS_n}, \quad n \geq 0,$$

where, as usual, $S_n = \sum_{i=1}^{n} X_i$, and we take $S_0 = 0$. The number t is fixed, and is often cleverly chosen in specific applications.

The special normal case of this martingale was seen in Example 14.4. Exactly the same proof works in order to show that Z_n as defined above is a martingale in general, not just the normal case. Formally, therefore, whenever we have a stopping time τ such that the optional stopping theorem is valid for this martingale sequence Z_n, we have the identity

$$E(Z_\tau) = E[(\psi(t))^{-\tau} e^{tS_\tau}] = E(Z_0) = 1.$$

Once we have this general identity, we can manipulate it for special stopping times τ to make useful conclusions in specific applications.

Example 14.19 (Error Probabilities of Wald's SPRT). As a specific application of historical importance in statistics, consider again the example of Wald's SPRT (Example 14.11). The setup is that we are acquiring iid observations $Z_1, Z_2, \ldots$ from a parametric family of densities $f(x|\theta)$, and we have to decide between the two hypotheses $H_0 : \theta = \theta_0$ (the null hypothesis), and $H_1 : \theta = \theta_1$ (the alternative hypothesis). As was explained in Example 14.11, we continue sampling as long as $l < S_n < u$ for some $l, u, l < u$, and stop and decide in favor of H_1 or H_0 when for the first time $S_n \geq u$ or $S_n \leq l$; here, S_n is the log likelihood ratio

$$\begin{aligned} S_n &= \log \Lambda_n = \log \frac{\prod_{i=1}^{n} f(Z_i\,|\theta_1)}{\prod_{i=1}^{n} f(Z_i\,|\theta_0)} \\ &= \sum_{i=1}^{n} \log \frac{f(Z_i\,|\theta_1)}{f(Z_i\,|\theta_0)} = \sum_{i=1}^{n} X_i \text{ say.} \end{aligned}$$

Therefore, in this particular case, the relevant stopping time is

$$\tau = \inf\{n > 0 : S_n \notin (l, u)\}.$$

The *type I error probability* of our test is the probability that the test would reject H_0 if H_0 happened to be true. Denoting the type I error probability as α, we have $\alpha = P_{\theta=\theta_0}(S_\tau \geq u)$. We use Wald's generalized identity to approximate α. Exact calculation of α is practically impossible except in stray cases.

To proceed with this approximation, suppose there is a number $t \neq 0$ such that $E_{\theta=\theta_0}(e^{tX_i}) = 1$. In our notation for the generalized Wald identity, this makes $\psi(t) = 1$ for this judiciously chosen t. If we now make the assumption (of some faith) that when S_n leaves the interval (l, u), it does not overshoot the limits l, u by too much, we should have

$$S_\tau \approx uI_{\{S_\tau \geq u\}} + lI_{\{S_\tau \leq l\}}.$$

Therefore, by applying Wald's generalized identity,

$$\begin{aligned} 1 = E_{\theta=\theta_0}(e^{tS_\tau}) &\approx e^{tu}\alpha + e^{tl}(1-\alpha) \\ \Rightarrow \alpha &\approx \frac{1-e^{tl}}{e^{tu}-e^{tl}}. \end{aligned}$$

This is the classic *Wald approximation to the type I error probability of the SPRT (sequential probability ratio test)*. A similar approximation exists for the type II error probability of the SPRT, which is the probability that the test will accept H_0 if H_0 happens to be false.

14.3 Martingale and Concentration Inequalities

The optional stopping theorem is also the main tool in proving a collection of important inequalities involving martingales. To provide a little context for such inequalities, consider the special martingale of a random walk, namely $S_n = \sum_{i=1}^n X_i, n \geq 1$, where we assume the X_i to be iid mean zero random variables with a finite variance σ^2. If we take any fixed n, and any fixed $\lambda > 0$, then simply by Chebyshev's inequality, $P(|S_n| \geq \lambda) \leq \frac{E(S_n^2)}{\lambda^2}$. Kolmogorov's inequality (see Chapter 8) makes the stronger assertion $P(\max_{1\leq k\leq n} |S_k| \geq \lambda) \leq \frac{E(S_n^2)}{\lambda^2}$. A fundamental inequality in martingale theory says that such an inequality holds for more general martingales, and not just the special martingale of a random walk.

14.3.1 Maximal Inequality

Theorem 14.7 (Martingale Maximal Inequality).

(a) Let $\{X_n, n \geq 0\}$ be a nonnegative submartingale adapted to some sequence $\{Y_n\}$, and λ any fixed positive number. Then, for any $n \geq 0$,

$$P\left(\max_{0\leq k\leq n} X_k \geq \lambda\right) \leq \frac{E(X_n)}{\lambda}.$$

(b) Let $\{X_n, n \geq 0\}$ be a martingale adapted to some sequence $\{Y_n\}$, and λ any fixed positive number. Suppose $p \geq 1$ is such that $E(|X_k|^p) < \infty$ for any $k \geq 0$. Then, for any $n \geq 0$,

$$P\left(\max_{0\leq k\leq n} |X_k| \geq \lambda\right) \leq \frac{E\left(|X_n|^p I_{\{\max_{0\leq k\leq n}|X_k|\geq\lambda\}}\right)}{\lambda^p} \leq \frac{E(|X_n|^p)}{\lambda^p}.$$

Proof. Note that the final inequality in part (b) follows from part (a) by use of Theorem 14.1 because $f(z) = |z|^p$ is a nonnegative convex function, and therefore if $\{X_n\}$ is a martingale adapted to some sequence $\{Y_n\}$, then for $p \geq 1$, $\{|X_n|^p\}$ is a nonnegative submartingale (adapted to that same sequence $\{Y_n\}$). The first inequality in part (b) is proved by partitioning the event $\{\max_{0\leq k\leq n} |X_k| \geq \lambda\}$ into disjoint events of the form $\{|X_0| < \lambda, \ldots |X_i| < \lambda, |X_{i+1}| \geq \lambda\}$, and then using simple bounds on each of these partitioning sets. This is left as an exercise.

For proving part (a) of this theorem, define the stopping time

$$\tau = \inf\{k \geq 0 : X_k > \lambda\},$$

and $\tau_n = \min(\tau, n)$.

Then, by the optional stopping theorem,

$$\begin{aligned} E(X_n) &\geq E(X_{\tau_n}) = E\left(X_{\tau_n} I_{\{\max_{0\leq k\leq n} X_k\geq\lambda\}}\right) + E(X_{\tau_n} I_{\{\max_{0\leq k\leq n} X_k<\lambda\}}) \\ &\geq E\left(X_{\tau_n} I_{\{\max_{0\leq k\leq n} X_k\geq\lambda\}}\right) \end{aligned}$$

(since the $\{X_n\}$ sequence has been assumed to be nonnegative)

$$\geq \lambda E\left[I_{\{\max_{0\leq k\leq n} X_k\geq\lambda\}}\right] = \lambda P\left(\max_{0\leq k\leq n} X_k \geq \lambda\right),$$

which is what part (a) of this theorem says. □

Part (a) of the theorem above assumes the submartingale $\{X_n\}$ to be nonnegative. This assumption is in fact not needed. In addition, the inequality itself can be somewhat strengthened. The following improved version of the maximal inequality can be proved by minor modifications of the argument given above; we record the stronger version, which is important for applications.

Theorem 14.8 (A Better Maximal Inequality). *Let $\{X_n, n \geq 0\}$ be a submartingale adapted to some sequence $\{Y_n\}$, and λ any fixed positive number. Then, for any $n \geq 0$,*

$$P\left(\max_{0\leq k\leq n} X_k \geq \lambda\right) \leq \frac{E(X_n^+)}{\lambda} \leq \frac{E(|X_n|)}{\lambda},$$

where for any real number x, $x^+ = \max(x, 0) \leq |x|$.

Example 14.20 (Sharper Bounds Near Zero). The bounds in Theorem 14.7 and Theorem 14.8 are not useful unless λ is large, because the upper bounds blow up as $\lambda \to 0$. However, if we work a little harder, then useful bounds can be derived at least in some cases even when λ is near zero. This example illustrates such a calculation.

Let $\{X_n\}$ be a zero mean martingale, and suppose $\sigma_k^2 = \text{Var}(X_k) < \infty$ for all k. For $n \geq 0$, denote $M_n = \max_{0 \leq k \leq n} X_k$. Fix a constant $c > 0$; the constant c is chosen later suitably. By Theorem 14.1, $\{(X_k + c)^2\}$ is a submartingale, and therefore, by Theorem 14.8,

$$P(M_n \geq \lambda) = P(M_n + c \geq \lambda + c) = P\left(\max_{0 \leq k \leq n} (X_k + c) \geq \lambda + c\right)$$
$$\leq \frac{E(X_n + c)^2}{(\lambda + c)^2} = \frac{c^2 + \sigma_n^2}{c^2 + 2c\lambda + \lambda^2}.$$

Therefore,

$$P(M_n \geq \lambda) \leq \inf_{c > 0} \frac{c^2 + \sigma_n^2}{c^2 + 2c\lambda + \lambda^2}.$$

The function $\frac{c^2 + \sigma_n^2}{c^2 + 2c\lambda + \lambda^2}$ is uniquely minimized at the root of the derivative equation

$$\frac{c}{c^2 + \sigma_n^2} - \frac{c + \lambda}{c^2 + 2c\lambda + \lambda^2} = 0$$
$$\Leftrightarrow c^2\lambda + c(\lambda^2 - \sigma_n^2) - \lambda\sigma_n^2 = 0 \Leftrightarrow c = \frac{\sigma_n^2}{\lambda}.$$

Plugging this value of c, we get

$$P(M_n \geq \lambda) \leq \inf_{c > 0} \frac{c^2 + \sigma_n^2}{c^2 + 2c\lambda + \lambda^2}$$
$$= \frac{\sigma_n^2}{\lambda^2 + \sigma_n^2},$$

for any $\lambda > 0$. Clearly, this bound is strictly smaller than one for any $\lambda > 0$.

Example 14.21 (Bounds on the Moments of the Maximum). Here is a clever application of Theorem 14.7 to bounding the moments of $M_n = \max_{0 \leq k \leq n} |X_k|$ in terms of the same moment of $|X_n|$ for a martingale sequence $\{X_n\}$. The example is a very nice illustration of the art of putting simple things together to get a pretty end result.

Suppose that $\{X_n, n \geq 0\}$ is a martingale sequence, and $p > 1$ is such that $E(|X_k|^p) < \infty$ for every k. The proof of the result in this example makes use of Holder's inequality $E(|XY|) \leq (E|X|^\alpha)^{1/\alpha}(E|Y|^\beta)^{1/\beta}$, where $\alpha, \beta > 1$, and $\beta = \frac{\alpha}{\alpha - 1}$ (see Chapter 1).

Proceeding,

$$E(M_n^p) = \int_0^\infty p\lambda^{p-1} P(M_n > \lambda) d\lambda$$
$$\leq \int_0^\infty p\lambda^{p-1} \frac{E\left(|X_n| I_{\{M_n \geq \lambda\}}\right)}{\lambda} d\lambda$$

(by using part (b) of Theorem 14.7)

$$= \int_0^\infty p\lambda^{p-2} E\left(|X_n| I_{\{M_n \geq \lambda\}}\right) d\lambda = E\left[p|X_n| \left(\int_0^{M_n} \lambda^{p-2} d\lambda\right)\right]$$

(by Fubini's theorem)

$$= E\left[p|X_n| \frac{M_n^{p-1}}{p-1}\right] = \frac{p}{p-1} E\left(|X_n| M_n^{p-1}\right)$$
$$\leq \frac{p}{p-1} [E|X_n|^p]^{1/p} [E(M_n^p)]^{(p-1)/p}$$

(by using Holder's inequality with $\alpha = p, \beta = \frac{p}{p-1}$).

Transferring $[E(M_n^p)]^{(p-1)/p}$ to the left side,

$$[E(M_n^p)]^{1/p} \leq \frac{p}{p-1} [E|X_n|^p]^{1/p}.$$

In particular, for a square integrable martingale, by using $p = 2$ in the inequality we just derived,

$$[E(M_n^2)]^{1/2} \leq 2[E(X_n^2)]^{1/2} \Rightarrow E(M_n^2) \leq 4E(X_n^2),$$

a very pretty and useful inequality.

14.3.2 ∗ Inequalities of Burkholder, Davis, and Gundy

The previous two examples indicated applications of various versions of the maximal inequality to obtaining bounds on the moments of the maximum $M_n = \max_{0 \leq k \leq n} |X_k|$ for a martingale sequence $\{X_n\}$. The maximal inequality tells us how to obtain bounds on the moments from bounds on the tail probability. In particular, if the martingale is square integrable, that is, if $E(X_k^2) < \infty$ for any k, then the maximal inequality leads to a bound on the second moment of M_n in terms of the second moment of the last term, namely $E(X_n^2)$.

There is a useful connection between $E(X_n^2)$ and $E(D_n^2)$ for a general square integrable martingale $\{X_n\}$, where $D_n^2 = \sum_{i=1}^n (X_i - X_{i-1})^2$. The connection, which we prove below, is the neat identity $E(X_n^2) - E(X_0^2) = E(D_n^2)$, so that if $X_0 = 0$, then $E(X_n^2)$ and $E(D_n^2)$ are equal. Therefore, we can think of the maximal inequality and the implied moment bounds in terms of $E(D_n^2)$, because $E(D_n^2)$ and $E(X_n^2)$ are, after all, equal. It was shown in Burkholder (1973), Davis (1970), and Burkholder, Davis, and Gundy (1972) that one can bound expectations of far more general functions of M_n in terms of expectations of the same functions of D_n; in particular, one can bound the pth moment of M_n from both directions by multiples of the pth moment of D_n for general $p \geq 1$. In some sense, the moments of M_n and the moments of D_n grow in the same order; if one can control the increments of the martingale sequence, then one can control the maximum. Three such important bounds are presented in this section for reference and completeness. But first, we demonstrate the promised connection between $E(X_n^2)$ and $E(D_n^2)$, an interesting result in its own right.

Proposition. *Suppose $\{X_n, n \geq 0\}$ is a martingale. Let $V_i = X_i - X_{i-1}, i \geq 1$, and $D_n^2 = \sum_{i=1}^n V_i^2$. Suppose $E(X_k^2) < \infty$ for each $k \geq 0$. Then, for any $n \geq 1$,*

$$E(D_n^2) = E(X_n^2) - E(X_0^2).$$

Proof.

$$\begin{aligned}
E(D_n^2) &= \sum_{i=1}^n E[(X_i - X_{i-1})^2] = \sum_{i=1}^n E[X_i(X_i - X_{i-1}) - X_{i-1}(X_i - X_{i-1})] \\
&= \sum_{i=1}^n E(E[X_i(X_i - X_{i-1}) \,|X_0, \ldots, X_{i-1}]) \\
&\quad - \sum_{i=1}^n E(E[X_{i-1}(X_i - X_{i-1}) \,|X_0, \ldots, X_{i-1}]) \\
&= \sum_{i=1}^n \{E(E[X_i^2 \,|X_0, \ldots, X_{i-1}]) - E(X_{i-1}E[X_i \,|X_0, \ldots, X_{i-1}])\} \\
&\quad - \sum_{i=1}^n E(X_{i-1}E[X_i \,|X_0, \ldots, X_{i-1}] - X_{i-1}^2) \\
&= \sum_{i=1}^n \{E(X_i^2) - E(X_{i-1}^2)\} - \sum_{i=1}^n E(X_{i-1}^2 - X_{i-1}^2) \\
&= E(X_n^2) - E(X_0^2).
\end{aligned}$$

□

Remark. In view of this result, we can restate part (b) of Theorem 14.7 for the case $p = 2$ in the following manner.

Theorem 14.9. *Let $\{X_n, n \geq 0\}$ be a martingale such that $X_0 = 0$ and $E(X_k^2) < \infty$ for all $k \geq 1$. Let λ be any fixed positive number, and for any $n \geq 1$, $M_n = \max_{0 \leq k \leq n} |X_k|$. Then,*

$$P(M_n \geq \lambda) \leq \frac{E(D_n^2)}{\lambda^2}.$$

The inequalities of Burkholder, Davis, and Gundy show how to establish bounds on moments of M_n in terms of the same moments of D_n. To describe some of these bounds, we first need a little notation.

Given a real-valued random variable X, and a positive number p, the *L_p norm of X* is defined as $||X||_p = [E(|X|^p)]^{\frac{1}{p}}$, assuming that $E(|X|^p) < \infty$. Obviously, if X is already a nonnegative random variable, then $||X||_p = [E(X^p)]^{\frac{1}{p}}$, Here are two specific bounds on the L_p norms of M_n in terms of the L_p norms of D_n. Of these, the case $p > 1$ was considered in works of Donald Burkholder (e.g., Burkholder (1973)); the case $p = 1$ needed a separate treatment, and was dealt with in Davis (1970).

Theorem 14.10. *(a) Suppose $\{X_n, n \geq 0\}$ is a martingale, with $X_0 = 0$. Suppose for some given $p > 1$, $||X_k||_p < \infty$ for all $k \geq 1$. Then, for any $n \geq 1$,*

$$\frac{p-1}{18p^{3/2}}||D_n||_p \leq ||M_n||_p \leq \frac{18p^{3/2}}{(p-1)^{3/2}}||D_n||_p.$$

(b) There exist universal positive constants c_1, c_2 such that

$$c_1||D_n||_1 \leq ||M_n||_1 \leq c_2||D_n||_1.$$

Moreover, the constant c_2 may be taken to be $\sqrt{3}$.

For $p \geq 1$, the functions $x \to |x|^p$ are convex. It was shown in Burkholder, Davis, and Gundy (1972) that bounds of the same nature as in the theorem above hold for general convex functions. The exact result says the following.

Theorem 14.11. *Suppose $\{X_n, n \geq 0\}$ is a martingale with $X_0 = 0$ and $\phi : \mathcal{R} \to \mathcal{R}$ a convex function. Then there exist universal positive constants $c_\phi, C_\phi, c_\phi \leq C_\phi$, depending only on the function ϕ, such that for any $n \geq 1$,*

$$c_\phi E(\phi(D_n)) \leq E(\phi(M_n)) \leq C_\phi E(\phi(D_n)).$$

Remark. Note that apart from the explicit constants, both parts of Theorem 14.10 follow as special cases of this theorem. To our knowledge, no explicit choices of c_ϕ, C_ϕ are known.

14.3.3 Inequalities of Hoeffding and Azuma

The classical inequality of Hoeffding (Hoeffding (1963); see Chapter 8) gives bounds on the probability of a large deviation of a partial sum of bounded iid random variables from its mean value. The message of that inequality is that if the iid summands can be controlled, then the deviations of the sum from its mean can be controlled. Inequalities on probabilities of the form $P(|f(X_1, X_2, \ldots, X_n) - E(f(X_1, X_2, \ldots, X_n))| > t)$ are called *concentration inequalities*. An equally classic concentration inequality of K. Azuma (Azuma (1967)) shows that a Hoeffding type inequality holds for a martingale sequence, provided that the increments $X_k - X_{k-1}$ vary in bounded intervals. The analogy between the iid case and the martingale case is then clear. In the iid case, we can control $S_n = \sum_{i=1}^n X_i$ if we can control the summands X_i; in the martingale case, we can control $X_n - X_0 = \sum_{i=1}^n (X_i - X_{i-1})$ if we can control the summands $X_i - X_{i-1}$. Here is Azuma's inequality in its classic form; a more general form is given afterwards.

Theorem 14.12 (Azuma's Inequality). *Suppose $\{X_n, n \geq 0\}$ is a martingale such that $V_i = |X_i - X_{i-1}| \leq c_i$, where c_i are positive constants. Then, for any positive number t and any $n \geq 1$,*

(a) $P(X_n - X_0 \geq t) \leq e^{-\frac{t^2}{2\sum_{i=1}^n c_i^2}}$.

(b) $P(X_n - X_0 \leq -t) \leq e^{-\frac{t^2}{2\sum_{i=1}^n c_i^2}}$.

(c) $P(|X_n - X_0| \geq t) \leq 2e^{-\frac{t^2}{2\sum_{i=1}^n c_i^2}}$.

The proof of part (b) is exactly the same as that of part (a), and part (c) is an immediate consequence of parts (a) and (b). So only part (a) requires a proof. For this, we need a classic convexity lemma, originally used in Hoeffding (1963), and then a generalized version of it. Here is the first lemma.

(Hoeffding's Lemma). *Let X be a zero mean random variable such that $P(a \leq X \leq b) = 1$, where a, b are finite constants. Then, for any $s > 0$,*

$$E(e^{sX}) \leq e^{\frac{s^2(b-a)^2}{8}}.$$

Remark. It is important to note that the bound in this lemma depends only on $b - a$ and the mean zero assumption, but not on the individual values of a, b.

Proof of Hoeffding's Lemma. The proof uses convexity of the function $x \to e^{sx}$, and a calculus inequality on the function $\phi(u) = -pu + \log(1 - p + pe^u), u \geq 0$, where p is a fixed number in $(0, 1)$.

First, by the convexity of $x \to e^{sx}$, for $a \le x \le b$,

$$e^{sx} \le \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa}.$$

Taking an expectation,

$$E(e^{sX}) \le pe^{sb} + (1-p)e^{sa}. \ (*)$$

where $p = \frac{-a}{b-a}$; note that p belongs to $[0, 1]$. It now remains to show that $pe^{sb} + (1-p)e^{sa} \le e^{\frac{s^2(b-a)^2}{8}}$. Towards this, write

$$\begin{aligned} pe^{sb} + (1-p)e^{sa} &= e^{sa}\left[1 - p + pe^{s(b-a)}\right] = e^{-sp(b-a)}\left[1 - p + pe^{s(b-a)}\right] \\ &= e^{-sp(b-a)+\log(1-p+pe^{s(b-a)})} = e^{-pu+\log(1-p+pe^{u})}, \end{aligned}$$

writing u for $s(b-a)$.

A relatively simple calculus argument shows that the function $\phi(u) = -pu + \log(1-p+pe^u)$ is bounded above by $\frac{u^2}{8}$ for all $u > 0$. Plugging this bound in $(*)$ results in the bound in the lemma.

(Generalized Hoeffding Lemma). *Let V, Z be two random variables such that*

$$E(V|Z) = 0, \text{ and } P(f(Z) \le V \le f(Z) + c) = 1$$

for some function $f(Z)$ of Z and some positive constant c. Then, for any $s > 0$,

$$E(e^{sV}|Z) \le e^{\frac{s^2c^2}{8}}.$$

The generalized Hoeffding lemma has the same proof as Hoeffding's lemma itself. Refer to the remark that we made just before the proof of Hoeffding's lemma. It is the generalized Hoeffding lemma that gives us Azuma's inequality.

Proof of Azuma's Inequality. Still using the notation $V_i = X_i - X_{i-1}$, then, with $s > 0$,

$$\begin{aligned} P(X_n - X_0 \ge t) &= P\left(e^{s(X_n - X_0)} \ge e^{st}\right) \le e^{-st} E\left(e^{s(X_n - X_0)}\right) \\ &= e^{-st} E\left(e^{s\sum_{i=1}^{n} V_i}\right) = e^{-st} E\left(e^{s\sum_{i=1}^{n-1} V_i + sV_n}\right) \\ &= e^{-st} E\left(e^{s\sum_{i=1}^{n-1} V_i} E\left[e^{sV_n}|X_0, \ldots, X_{n-1}\right]\right) \\ &\le e^{-st} E\left(e^{s\sum_{i=1}^{n-1} V_i} e^{\frac{s^2(2c_n)^2}{8}}\right) \end{aligned}$$

(because $E(V_n \,|X_0, \ldots, X_{n-1}) = 0$ by the martingale property of $\{X_n\}$, and then by applying the generalized Hoeffding lemma)

$$= e^{-st} e^{\frac{s^2 c_n^2}{2}} E\left(e^{s\sum_{i=1}^{n-1} V_i}\right) \leq e^{-st} e^{\frac{s^2 \sum_{i=1}^{n} c_i^2}{2}},$$

by repeating the same argument.

This latest inequality is true for any $s > 0$. Therefore, by minimizing the bound over $s > 0$,

$$P(X_n - X_0 \geq t) \leq \inf_{s>0} e^{-st} e^{\frac{s^2 \sum_{i=1}^{n} c_i^2}{2}} = e^{-\frac{t^2}{2\sum_{i=1}^{n} c_i^2}},$$

where the infimum over s is easily established by a simple calculus argument. This proves Azuma's inequality. □

14.3.4 ∗ Inequalities of McDiarmid and Devroye

McDiarmid (1989) and Devroye (1991) use novel martingale techniques to derive concentration inequalities and variance bounds for potentially complicated functions of independent random variables. The only requirement is that the function should not change by arbitrarily large amounts if all but one of the coordinates remain fixed. The first result below says that functions of certain types are concentrated near their mean value with a high probability.

Theorem 14.13. *Suppose $X_1, \ldots, X_n$ are independent random variables, and $f(x_1, \ldots, x_n)$ is a function such that for each $i, 1 \leq i \leq n$, there exist finite constant $c_i = c_{i,n}$ such that*

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i$$

for all $x_1, \ldots, x_i, x_i', \ldots, x_n$. Let t be any positive number. Then,

(a) $P(f - E(f) \geq t) \leq e^{-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}}$.

(b) $P(f - E(f) \leq -t) \leq e^{-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}}$.

(c) $P(|f - E(f)| \geq t) \leq 2e^{-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}}$.

Proof. Once again, only part (a) is proved, because (b) is proved exactly analogously, and (c) follows by adding the inequalities in (a) and (b). For notational convenience, we take $E(f)$ to be zero; this allows us to write f in place of $f - E(f)$ below.

The trick is to decompose f as $f = \sum_{k=1}^{n} V_k$, where $\{V_k\}$ is a martingale difference sequence such that it can be bounded in both directions, $Z_k \leq V_k \leq W_k$, in a manner so that $W_k - Z_k \leq c_k, k = 1, 2, \ldots, n$. Then, Azuma's inequality applies and the inequality of this theorem falls out. Construct the random variables V_k, Z_k, W_k as follows.

Define

$$\begin{aligned}
\eta(x_1, \ldots, x_k) &= E[f(X_1, \ldots, X_n) \,|X_1 = x_1, \ldots, X_k = x_k];\\
V_k &= \eta(X_1, \ldots, X_k) - \eta(X_1, \ldots, X_{k-1}) \text{ for } k \geq 2, \text{ and } V_1 = \eta(X_1);\\
Z_k &= \inf_{x_k} \eta(X_1, \ldots, X_{k-1}, x_k) - \eta(X_1, \ldots, X_{k-1}) \text{ for } k \geq 2,\\
&\quad \text{and } Z_1 = \inf_{x_1} \eta(x_1);\\
W_k &= \sup_{x_k} \eta(X_1, \ldots, X_{k-1}, x_k) - \eta(X_1, \ldots, X_{k-1}) \text{ for } k \geq 2,\\
&\quad \text{and } W_1 = \sup_{x_1} \eta(x_1).
\end{aligned}$$

Now observe the following facts.

(a) By construction, $Z_k \leq V_k \leq W_k$ for each k.
(b) By hypothesis, $W_k - Z_k \leq c_k$ for each k.
(c) $f(X_1, \ldots, X_n) = \sum_{k=1}^{n} V_k$.
(d) $\{V_k\}_1^n$ forms a martingale difference sequence.

Therefore, we can once again apply the generalized Hoeffding lemma and simply repeat the proof of Azuma's inequality to obtain the inequality in part (a) of this theorem. □

An interesting feature of McDiarmid's inequality is that martingale methods were used to derive a probability inequality involving independent random variables. It turns out that martingale methods may also be used to derive variance bounds for functions of independent random variables. The following variance bound is taken from Devroye (1991).

Theorem 14.14. *Suppose $X_1, \ldots, X_n$ are independent random variables and $f(x_1, \ldots, x_n)$ is a function that satisfies the conditions of Theorem 14.13. Then,*

$$Var(f(X_1, \ldots, X_n)) \leq \frac{\sum_{i=1}^{n} c_i^2}{4}.$$

Proof. We use the same notation as in the proof of Theorem 14.13. The proof consists of showing $\text{Var}(f) = E(\sum_{i=1}^{n} V_i^2)$ and $E(V_i^2) \leq (c_i^2/4)$.

To prove the first fact, we use the martingale decomposition as in Theorem 14.13 to get

$$\begin{aligned}
\mathrm{Var}(f) &= \mathrm{Var}\left(\sum_{i=1}^{n} V_i\right) = E\left[\left(\sum_{i=1}^{n} V_i\right)^2\right] \\
&= \sum_{i=1}^{n} E[V_i^2] + 2\sum\sum_{i<j} E[V_i V_j] \\
&= \sum_{i=1}^{n} E[V_i^2] + 2\sum\sum_{i<j} E(V_i E[V_j \,|X_1, \ldots, X_{j-1}]) \\
&= \sum_{i=1}^{n} E[V_i^2] + 2\sum\sum_{i<j} E(V_i \times 0) = \sum_{i=1}^{n} E[V_i^2].
\end{aligned}$$

To prove the second fact, we use an extremal property of two-point distributions, namely that the two-point distribution placing probability $\frac{1}{2}$ at each of a, b maximizes the variance among all distributions supported on $[a, b]$, and that this two-point distribution has variance $\frac{(b-a)^2}{4}$. From the proof of Theorem 14.13, $Z_i \le V_i \le W_i \le Z_i + c_i$. Therefore, the conditional variance of V_i given $X_1, \ldots, X_{i-1}$ is at most $\frac{c_i^2}{4}$, and the conditional mean is zero. Putting these two facts together, we get our desired bound $E(V_i^2) \le \frac{c_i^2}{4}$, which gives the variance bound stated in this theorem. □

The two theorems in this section give useful probability and variance bounds in many complicated problems in which direct evaluation would be essentially impossible.

Example 14.22 (The Kolmogorov–Smirnov Statistic). Suppose $X_1, X_2, \ldots, X_n$ are iid observations from some continuous CDF $F(x)$ on the real line. It is sometimes of interest in statistics to test the hypothesis that $F = F_0$, some specific CDF on the real line. By the Glivenko–Cantelli theorem (see Chapter 7), the empirical CDF F_n converges uniformly to the true CDF with probability one. So a measure of discrepancy of the observed data from the postulated CDF F_0 is $\Delta_n = \sup_x |F_n(x) - F_0(x)|$. The *Kolmogorov–Smirnov statistic* is $D_n = \sqrt{n}\Delta_n$. Exact calculations with D_n are very cumbersome, because of the complicated nature of its distribution for given n. The purpose of this example is to use the inequalities of McDiarmid and Devroye to get useful bounds on its tail probabilities and the variance.

The function f to which we would apply the inequalities of McDiarmid and Devroye is $f(X_1, \ldots, X_n) = \sup_x |F_n(x) - F_0(x)|$. We need to show that if just one data value changes, then the function f cannot change by too large an amount. Indeed, consider two datasets, $\{X_1, \ldots, X_i, \ldots, X_n\}$ and $\{X_1, \ldots, X_i', \ldots, X_n\}$, where in the second set the X_i' value is different from X_i. Let the corresponding

empirical CDFs be F_n, F_n'. Fix an x. The number of observations $\le x$ in the two datasets can differ by at most one, and therefore $|F_n(x) - F_n'(x)| \le \frac{1}{n}$. This holds for any x. By the triangular inequality,

$$|\sup_x |F_n(x) - F_0(x)| - \sup_x |F_n'(x) - F_0(x)|| \le \sup_x |F_n(x) - F_n'(x)| \le 1/n.$$

Thus, we may use $c_i = c_{i,n} = \frac{1}{n}$ in the inequalities of McDiarmid and Devroye. First, by simply plugging $c_i = \frac{1}{n}$ in Theorem 14.13, we get

$$\begin{aligned} &P(|\Delta_n - E(\Delta_n)| \ge t) \le 2e^{-2nt^2} \\ \Rightarrow\ &P(|D_n - E(D_n)| \ge t) \le 2e^{-2t^2}. \end{aligned}$$

This concentration inequality holds for every fixed n and $t > 0$, and we do not need to deal with the exact distribution of D_n to arrive at this inequality.

Again plugging $c_i = \frac{1}{n}$ in Theorem 14.14, we get

$$\text{Var}(\Delta_n) \le \frac{1}{4n} \Rightarrow \text{Var}(D_n) \le \frac{1}{4},$$

for all $n \ge 1$. Once again, this is an attractive variance bound that is valid for every n, and we do not need to work with the exact distribution of D_n to arrive at this bound.

14.3.5 The Upcrossing Inequality

A final key inequality in martingale theory that we present is *Doob's upcrossing inequality*. The inequality is independently useful for studying fluctuations in the trajectory of a martingale (submartingale) sequence. It is also the result we need in the next section for establishing the fundamental convergence theorem for martingales (submartingales).

Given the discrete time process $\{X_n, n \ge 0\}$, fix an integer $N > 0$, and two numbers $a, b, a < b$. We now track the time instants at which this process crosses b from below, or a from above. Formally, let $T_0 = \inf\{k \ge 0 : X_k \le a\}$. If $X_0 > a$, then this is the first *downcrossing* of a. If $X_0 \le a$, then $T_0 = 0$. Now we wait for the first *upcrossing* of b after the time T_0. Formally, $T_1 = \inf\{k > T_0 : X_k \ge b\}$. We continue tracking the down and the upcrossings of the two levels a, b in this fashion. Here then is the formal definition for the entire sequence of stopping times T_n:

$$\begin{aligned} T_0 &= \inf\{k \ge 0 : X_k \le a\}; \\ T_{2n+1} &= \inf\{k > T_{2n} : X_k \ge b\}, \quad n \ge 0; \\ T_{2n+2} &= \inf\{k > T_{2n+1} : X_k \le a\}, \quad n \ge 0. \end{aligned}$$

The times $T_1, T_3, \ldots$ are then the instants of upcrossing, and the times $T_0, T_2, \ldots$ are the instants of downcrossing. The upcrossing inequality places a bound on the expected value of $U_{a,b,N}$, the number of upcrossings up to the time N. Note that this is simply the number of odd labels $2n+1$ for which $T_{2n+1} \leq N$.

Theorem 14.15. *Let $\{X_n, n \geq 0\}$ be a submartingale. Then for any $a, b, N (a < b)$,*

$$E[U_{a,b,N}] \leq \frac{E[(X_N - a)^+] - E[(X_0 - a)^+]}{b-a} \leq \frac{E(|X_N|) + |a|}{b-a}.$$

Proof. The second inequality follows from the first inequality by the pointwise inequality $(x-a)^+ \leq x^+ + |a| \leq |x| + |a|$, and so, we prove only the first inequality.

First make the following reduction. Define a new nonnegative submartingale as $Y_n = (X_n - a)^+, n \geq 0$. This shifting by a is going to result in a useful reduction. There is a functional identity between the upcrossing variable that we are interested in, namely $U_{a,b,N}$ and the number of upcrossings $V_{0,b-a,N}$ of this new process $\{Y_n\}_0^N$ of the two new levels 0 and $b-a$. Indeed, $U_{a,b,N} = V_{0,b-a,N}$. So we need to show that $E[V_{0,b-a,N}] \leq \frac{E(Y_N - Y_0)}{b-a}$.

The key to proving this inequality is to write a clever decomposition

$$Y_N - Y_0 = \sum_{i=0}^{N} (Y_{\tau_i} - Y_{\tau_{i-1}}),$$

such that three things happen:

(i) The τ_i are increasing stopping times, so that the submartingale property is inherited by the Y_{τ_i} sequence.

(ii) The sum over the odd labels in this decomposition satisfy the pointwise inequality

$$\sum_{i:0 \leq i \leq N, i \text{ odd}} (Y_{\tau_i} - Y_{\tau_{i-1}}) \geq (b-a) V_{0,b-a,N}.$$

(iii) The sum over the even labels satisfy the inequality

$$E\left[\sum_{i:0 \leq i \leq N, i \text{ even}} (Y_{\tau_i} - Y_{\tau_{i-1}})\right] \geq 0.$$

If we put (ii) and (iii) together, we immediately get

$$E(Y_N - Y_0) \geq (b-a) E[V_{0,b-a,N}],$$

which is the needed result.

What are these stopping times τ_i, and why are (ii) and (iii) true? The stopping times $\tau_0 \leq \tau_1 \leq \ldots$ are defined in the following way. Analogous to the downcrossing and upcrossing times $T_0, T_1, \ldots$ of (a, b) for the original $\{X_n\}$ process, let

$T_0', T_1', \ldots$ be the downcrossing and upcrossing times of $(0, b-a)$ for the new $\{Y_n\}$ process. Now define $\tau_i = \min(T_i', N)$. The τ_i are increasing, that is, $\tau_0 \le \tau_1 \le \ldots$, because the T_i' are. Note that these τ_i are stopping times adapted to $\{Y_n\}$.

Now look at the sum over the odd labels, namely $(Y_{\tau_1} - Y_{\tau_0}) + (Y_{\tau_3} - Y_{\tau_2}) + \cdots$. Break this sum further into two subsets of labels, $i \le V = V_{0,b-a,N}$, and $i > V$. For each label i in the first subset, $(Y_{\tau_{2i+1}} - Y_{\tau_{2i}}) \ge b - a$, because $Y_{\tau_{2i+1}} \ge b$ and $Y_{\tau_{2i}} \le a$. Adding over these labels, of which there are V many, we get the sum to be $\ge (b-a)V$. The labels in the other subset can be seen to give a sum ≥ 0 (just think of what V means, and a little thinking shows that the rest of the labels produce a sum ≥ 0). So, now adding over the two subsets of labels, we get our claimed inequality in (ii) above.

The claim in (iii) is automatic by the optional stopping theorem, because for each individual i, we will have $E(Y_{\tau_{i-1}}) \le E(Y_{\tau_i})$ (actually, this is a slightly stronger demand than what the optional stopping theorem says; but it is true).

As was explained above, this completes the argument for the upcrossing inequality. □

14.4 Convergence of Martingales

14.4.1 The Basic Convergence Theorem

Paul Lévy initiated his study of martingales in his search for laws of large numbers beyond the case of means in the iid case. It turns out that martingales often converge to a limiting random variable, and even convergence of the means or higher moments can be arranged, provided that our martingale sequence is not allowed to fluctuate or grow out of control. To see why some such conditions would be needed, consider the case of the simple symmetric random walk $S_n = \sum_{i=1}^n X_i$, where the X_i are iid taking the values ± 1 with probability $\frac{1}{2}$ each. We know that the simple symmetric random walk is recurrent, and therefore it comes back infinitely often to every integer value x with probability one. So S_n, although a martingale, does not converge to some S_∞. The expected value of $|S_n|$ in the simple symmetric random walk case is of the order of $c\sqrt{n}$ for some constant c, and $c\sqrt{n}$ diverges as $n \to \infty$. A famous result in martingale theory says that if we can keep $E(|X_n|)$ in control (i.e., bounded away from ∞), then a martingale sequence $\{X_n\}$ will in fact converge to some suitable X_∞. Furthermore, some such condition is also essentially necessary for the martingale to converge. We start with an example.

Example 14.23 (Convergence of the Likelihood Ratio). Consider again the likelihood ratio $\Lambda_n = \prod_{i=1}^n \frac{f_1(X_i)}{f_0(X_i)}$, where f_0, f_1 are two densities and the sequence $X_1, X_2, \ldots$ is iid from the density f_0. We have seen that Λ_n is a martingale (see Example 14.8).

The likelihood ratio Λ_n gives a measure of the support in the first n data values for the density f_1. We know f_0 to be the true density from which the data values

are coming, therefore we would like the support for f_1 to diminish as more data are accumulated. Mathematically, we would like Λ_n to converge to zero as $n \to \infty$. We recognize that this is therefore a question about convergence of a martingale sequence, because Λ_n, after all, is a martingale if the true density is f_0.

Does Λ_n indeed converge (almost surely) to zero? Indeed, it does, and we can verify it directly, without using any martingale convergence theorems that we have not yet encountered. Here is why we can verify the convergence directly.

Assume that f_0, f_1 are strictly positive for the same set of x values; that is, $\{x : f_1(x) > 0\} = \{x : f_0(x) > 0\}$. Since $u \to \log u$ is a strictly concave function on $(0, \infty)$, by Jensen's inequality,

$$m = E_{f_0}\left[\log \frac{f_1(X)}{f_0(X)}\right] < \log\left(E_{f_0}\left[\frac{f_1(X)}{f_0(X)}\right]\right) = \log 1 = 0.$$

Because $Z_i = \log \frac{f_1(X_i)}{f_0(X_i)}$ are iid with mean m, by the usual SLLN for iid random variables,

$$\frac{1}{n}\log \Lambda_n = \frac{1}{n}\sum_{i=1}^{n} Z_i \overset{\text{a.s.}}{\to} m < 0 \Rightarrow \log \Lambda_n \overset{\text{a.s.}}{\to} -\infty$$
$$\Rightarrow \Lambda_n \overset{\text{a.s.}}{\to} 0.$$

So, in this example, the martingale Λ_n does converge with probability one to a limiting random variable Λ_∞, and Λ_∞ happens to be a constant random variable, equal to zero. We remark that the martingale Λ_n satisfies $E(|\Lambda_n|) = E(\Lambda_n) = 1$ and so, a fortiori, $\sup_n E(|\Lambda_n|) < \infty$. This has something to do with the fact that Λ_n converges in this example, although the random walk, also a martingale, failed to converge. This is borne out by the next theorem, a famous result in martingale theory. The proof of this next theorem requires the use of two basic facts in measure theory, which we state below.

Theorem 14.16 (Fatou's Lemma). *Let $X_n, n \geq 1$ and X be random variables defined on a common sample space Ω. Suppose each X_n is nonnegative with probability one, and suppose $X_n \overset{\text{a.s.}}{\to} X$. Then,* $\liminf_n E(X_n) \geq E(X)$.

Theorem 14.17 (Monotone Convergence Theorem). *Let $X_n, n \geq 1$ and X be random variables defined on a common sample space Ω. Suppose each X_n is nonnegative with probability one, that $X_1 \leq X_2 \leq X_3 \leq \ldots$, and $X_n \overset{\text{a.s.}}{\to} X$. Then $E(X_n) \uparrow E(X)$.*

Theorem 14.18 (Submartingale Convergence Theorem). *(a) Let $\{X_n\}$ be a submartingale such that $\sup_n E(X_n^+) = c < \infty$. Then there exists a random variable $X = X_\infty$, almost surely finite, such that $X_n \overset{\text{a.s.}}{\to} X$.*

(b) Let $\{X_n\}$ be a nonnegative supermartingale, or a nonpositive submartingale. Then there exists a random variable $X = X_\infty$, almost surely finite, such that $X_n \overset{\text{a.s.}}{\to} X$.

Proof. The proof uses the upcrossing inequality, the monotone convergence theorem, and Fatou's lemma. The key idea is first to show that under the hypothesis of the theorem, the process $\{X_n\}$ cannot fluctuate indefinitely between two given numbers $a, b, a < b$. Then a standard analytical technique of approximation by rationals, and use of the monotone convergence theorem and Fatou's lemma produces the submartingale convergence theorem. Here are the steps of the proof. Define

$$\begin{aligned} U_{a,b,N} &= \text{Number of upcrossings of (a,b) by } X_0, X_1, \ldots, X_N; \\ U_{a,b} &= \text{Number of upcrossings of (a,b) by } X_0, X_1, \ldots; \\ \Theta_{a,b} &= \{\omega \in \Omega : \liminf_n X_n \le a < b \le \limsup_n X_n\}; \\ \Theta &= \{\omega \in \Omega : \liminf_n X_n < \limsup_n X_n\}. \end{aligned}$$

First, by the monotone convergence theorem, $E[U_{a,b,N}] \to E[U_{a,b}]$ as $N \to \infty$, because $U_{a,b,N}$ converges monotonically to $U_{a,b}$ as $N \to \infty$. Therefore, by the upcrossing inequality,

$$\begin{aligned} E[U_{a,b,N}] \le \frac{E(|X_N|) + |a|}{b-a} &\Rightarrow E[U_{a,b}] = \lim_N E[U_{a,b,N}] \\ &\le \frac{\limsup_N E(|X_N|) + |a|}{b-a} < \infty. \end{aligned}$$

This means that $U_{a,b}$ must be finite with probability one (i.e., it cannot equal ∞ with a positive probability).

Next, note that $\Theta \subseteq \cup_{\{a<b,\, a,\, b \text{ rational}\}} \Theta_{a,b}$, and because we now have that $P(\Theta_{a,b}) = 0$ for any specific pair a, b, $P(\cup_{\{a<b,\, a,\, b \text{ rational}\}} \Theta_{a,b})$ must also be zero. This then implies that $P(\Theta) = 0$, which establishes the existence of an almost sure limit for the sequence X_n.

However, a subtle point still remains. The limit, X, could be ∞ or $-\infty$ with a positive probability. We use Fatou's lemma to rule out that possibility. Indeed, by Fatou's lemma,

$$E(|X|) \le \liminf_n E(|X_n|) \le \sup_n E(|X_n|) < \infty,$$

and so X must be finite with probability one. This finishes the proof of part (a) of the submartingale convergence theorem.

Part (b) is an easy consequence of part (a). For example, if $\{X_n\}$ is a nonpositive submartingale, then

$$\sup_n E(|X_n|) = \sup_n E(-X_n) = -\inf_n E(X_n) = -E(X_1) < \infty,$$

and so convergence of X_n to an almost surely finite X follows from part (a). □

14.4.2 Convergence in L_1 and L_2

The basic convergence theorem that we just proved says that an L_1 bounded submartingale converges to some random variable X. It is a bit disappointing that the apparently strong hypothesis that the submartingale is L_1 bounded is not strong enough to ensure convergence of the expectations: $E(X_n)$ need not converge to $E(X)$ in spite of the L_1 bounded assumption. A slightly stronger control on the growth of the submartingale sequence is needed to ensure convergence of expectations, in addition to the convergence of the submartingale itself. For example, $\sup_n E(|X_n|^p) < \infty$ for some $p > 1$ will suffice. A condition of this sort immediately reminds us of uniform integrability. Indeed, if $\sup_n E(|X_n|^p) < \infty$ for some $p > 1$, then $\{X_n\}$ will be uniformly integrable. It turns out that uniform integrability will be enough to assure us of convergence of the expectations in the basic convergence theorem, and it is almost the minimum that we can get away with. Statisticians are often interested in convergence of variances also. That is a stronger demand, and requires a stronger hypothesis. The next theorem records the conclusions on these issues. For reasons of space, this next theorem is not proved. One can see a proof in Fristedt and Gray (1997, p. 480).

Theorem 14.19. *Let $\{X_n, n \geq 0\}$ be a submartingale.*

(a) Suppose $\{X_n\}$ is uniformly integrable. Then there exists an X such that $X_n \overset{\text{a.s.}}{\to} X$, and $E(|X_n - X|) \to 0$ as $n \to \infty$.

(b) Conversely, suppose there exists an X such that $E(|X_n - X|) \to 0$ as $n \to \infty$. Then $\{X_n\}$ must be uniformly integrable, and moreover, X_n necessarily converges almost surely to this X.

(c) If $\{X_n\}$ is a martingale, and is L_2 bounded (i.e., $\sup_n E(X_n^2) < \infty$), then there exists an X such that $X_n \overset{\text{a.s.}}{\to} X$, and $E(|X_n - X|^2) \to 0$ as $n \to \infty$.

Example 14.24 (Pólya's Urn). We previously saw that the proportion of white balls in Pólya's urn, namely $R_n = \frac{a+S_n}{a+b+n}$ forms a martingale (see Example 14.6). This is an example in which the various convergences that we may want come easily. Because R_n is obviously a uniformly bounded sequence, by the theorem stated above, R_n converges almost surely and in L_2 (and therefore, in L_1) to a limiting random variable R, taking values in $[0, 1]$.

Neither the basic (sub)martingale convergence theorem nor the theorem in this section helps us in any way to identify the ditribution of R. In fact, in this case, R has a nondegenerate distribution, which is a Beta distribution with parameters a and b. As a consequence of this, $E(R_n) \to \frac{a}{a+b}$ and $\text{Var}(R_n) \to \frac{ab}{(a+b)^2(a+b+1)}$ as $n \to \infty$. A proof that R has a Beta distribution with parameters a, b is available in DasGupta (2010).

Example 14.25 (Bayes Estimates). We saw in Example 14.9 that the sequence of Bayes estimates (namely, the mean of the posterior distribution of the parameter)

is a martingale adapted to the sequence of data values $\{X_n\}$. Continuing with the same notation as in Example 14.9, $Z_n = E(Y \mid X^{(n)})$ is our martingale sequence. Assume that the prior distribution for the parameter has a finite variance; that is, $E(Y^2) < \infty$. Then, by using Jensen's inequality for conditional expectations,

$$E(Z_n^2) = E[(E(Y \mid X^{(n)}))^2] \leq E[E(Y^2 \mid X^{(n)})] = E(Y^2).$$

Hence, by the theorem above in this section, the sequence of Bayes estimates Z_n converges to some Z almost surely, and moreover the mean and the variance of Z_n converge to the mean and the variance of Z.

A natural followup question is what exactly is this limiting random variable Z. We can only give partial answers in general. For example, for each n, $E(Z \mid X^{(n)}) = Z_n$ with probability one. It is tempting to conclude from here that Z is the same as Y with probability one. This will be the case if knowledge of the entire infinite data sequence $X_1, X_2, \ldots$ pins down Y completely, that is, if it is the case that someone who knows the infinite data sequence also knows Y with probability one.

14.5 ∗ Reverse Martingales and Proof of SLLN

Partial sums of iid random variables are of basic interest in many problems in probability, such as the study of random walks, and as we know, the sequence of centered partial sums forms a martingale. On the other hand, the sequence of sample means is of fundamental interest in statistics; but the sequence of means does not form a martingale. Interestingly, if we measure time *backwards*, then the sequence of means does form a martingale, and then the rich martingale theory once again comes into play. This motivates the concept of a *reverse martingale*.

Definition 14.4. A sequence of random variables $\{X_n, n \geq 0\}$ defined on a common sample space Ω is called a *reverse submartingale* adapted to the sequence $\{Y_n, n \geq 0\}$, defined on the same sample space Ω, if $E(|X_n|) < \infty$ for all n and $E(X_n \mid Y_{n+1}, Y_{n+2}, \ldots) \geq X_{n+1}$ for each $n \geq 0$. The sequence $\{X_n\}$ is called a *reverse supermartingale* if $E(X_n \mid Y_{n+1}, Y_{n+2}, \ldots) \leq X_{n+1}$ for each n.

The sequence $\{X_n\}$ is called a *reverse martingale* if it is both a reverse submartingale and a reverse supermartingale with respect to the same sequence $\{Y_n\}$, that is, if $E(X_n \mid Y_{n+1}, Y_{n+2}, \ldots) = X_{n+1}$ for each n.

Example 14.26 (Sample Means). Let $X_1, X_2, \ldots$ be an infinite *exchangeable* sequence of random variables: for any $n \geq 2$ and any permutation π_n of $(1, 2, \ldots, n)$, $(X_1, X_2, \ldots, X_n)$ and $(X_{\pi_n(1)}, X_{\pi_n(2)}, \ldots, X_{\pi_n(n)})$ have the same joint distribution. For $n \geq 1$, let $\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} = \frac{S_n}{n}$ be the sequence of sample means.

Then, by the exchanageability property of the $\{X_n\}$ sequence, for any given n, and any $k, 1 \le k \le n$,

$$\overline{X}_n = E(\overline{X}_n \,|S_n, S_{n+1}, \ldots) = \frac{1}{n}\sum_{i=1}^{n} E(X_i \,|S_n, S_{n+1}, \ldots)$$
$$= \frac{1}{n} n E(X_k \,|S_n, S_{n+1}, \ldots) = E(X_k \,|S_n, S_{n+1}, \ldots).$$

Consequently,

$$E(\overline{X}_{n-1} \,|S_n, S_{n+1}, \ldots) = \frac{1}{n-1}\sum_{k=1}^{n-1} E(X_k \,|S_n, S_{n+1}, \ldots)$$
$$= \frac{1}{n-1}(n-1)\overline{X}_n = \overline{X}_n,$$

which shows that the sequence of sample means is a reverse martingale (adapted to the sequence of partial sums).

There is a useful convex function theorem for reverse martingales as well, which is straightforward to prove.

Theorem 14.20 (Second Convex Function Theorem). *Let $\{X_n\}$ be a sequence of random variables defined on some sample space Ω, and f a convex function. Let $Z_n = f(X_n)$.*

(a) If $\{X_n\}$ is a reverse martingale, then $\{Z_n\}$ is a reverse submartingale.
(b) If $\{X_n\}$ is a reverse submartingale, and f is also nondecreasing, then $\{Z_n\}$ is a reverse submartingale.
(c) If $\{X_{n,m}\}, m = 1, 2, \ldots$ is a countable family of reverse submartingales, defined on the same space Ω and all adapted to the same sequence, then $\{\sup_m X_{n,m}\}$ is also a reverse submartingale, adapted to the same sequence.

Example 14.27 (A Paradoxical Statistical Consequence). Suppose Y is some real-valued random variable with mean η, and that we do not know the true value of η. Thus, we would like to estimate η. But, suppose that we cannot take any observations on the variable Y (for whatever reason). We can, however, take observations on a completely unrelated random variable X, where $E(|X|) < \infty$. Suppose we do take n iid observations on X. Call them $X_1, X_2, \ldots, X_n$ and let $\overline{X}_n$ be their mean. Then, by part (a) of the second convex function theorem, $|\overline{X}_n - \eta|$ forms a reverse submartingale, and hence $E(|\overline{X}_n - \eta|)$ is monotone nonincreasing in n. In other words, $E(|\overline{X}_{n+1} - \eta|) \le E(|\overline{X}_n - \eta|)$ for all n, and so taking more observations on the useless variable X is going to be beneficial for estimating the mean of Y, a comical conclusion.

Note that there is really nothing special about using the absolute difference $|\overline{X}_n - \eta|$ as the criterion for the accuracy of estimation of η. The standard terminology in statistics for the criterion to be used is a *loss function*, and loss functions $L(\overline{X}_n, \eta)$ with a convexity property with respect to $\overline{X}_n$ for any fixed η will result in the same paradoxical conclusion. One needs to make sure that $E[L(\overline{X}_n, \eta)]$ is finite.

Reverse martingales possess a universal special property that is convenient in applications. The property is that a reverse martingale always converges almost surely to some finite random variable. The convergence property also holds for reverse submartingales, but the limiting random variable may equal $+\infty$ or $-\infty$ with a positive probability. An important and interesting consequence of this universal convergence property is a proof of the SLLN in the iid case by using martingale techniques. This is shown seen below as an example. The convergence property of reverse martingales is stated below.

Theorem 14.21 (Reverse Martingale Convergence Theorem). *(a) Let $\{X_n\}$ be a reverse martingale adapted to some sequence. Then it is necessarily uniformly integrable, and there exists a random variable X, almost surely finite, such that $X_n \stackrel{\text{a.s.}}{\to} X$, and $E|X_n - X|) \to 0$ as $n \to \infty$.*

(b) Let $\{X_n\}$ be a reverse submartingale adapted to some sequence. Then there exists a random variable X taking values in $[-\infty, \infty]$ such that $X_n \stackrel{\text{a.s.}}{\to} X$.

See Fristedt and Gray (1997, pp. 483–484) for a proof using uniform integrability techniques. Here is an important application of this theorem.

Example 14.28 (Proof of Kolmogorov's SLLN). Let $X_1, X_2, \ldots$ be iid random variables, with $E(|X_1|) < \infty$, and let $E(X_1) = \mu$. The goal of this example is to show that the sequence of sample means, $\overline{X}_n$, converges almost surely to μ.

We use the reverse martingale convergence theorem to obtain a proof. Because we have already shown that $\{\overline{X}_n\}$ forms a reverse martingale sequence, by the reverse martingale convergence theorem we are assured of a finite random variable Y such that $\overline{X}_n$ converges almost surely to Y, and we are also assured that $E(Y) = \mu$. The only task that remains is to show that Y equals μ with probability one.

This is achieved by establishing that $P(Y \geq y) = [P(Y \geq y)]^2$ for all real y (i.e., $P(Y \geq y)$ is 0 or 1 for any y), which would force Y to be degenerate and therefore degenerate at μ. To prove that $P(Y \geq y) = [P(Y \geq y)]^2$ for all real y, define the double sequence

$$Y_{m,n} = \frac{X_{m+1} + X_{m+2} + \cdots + X_{m+n}}{n},$$

$m, n \geq 1$. Note that $\overline{X}_k$ and $Y_{m,n}$ are independent for any $m, k \leq m$, and any n, and that, furthermore, for any fixed m, $Y_{m,n}$ converges almost surely to Y (the same Y as above) as $n \to \infty$. These two facts together imply

$$\begin{aligned} P\left(Y \geq y, \max_{1 \leq k \leq m} \overline{X}_k \geq y\right) &= P(Y \geq y) P\left(\max_{1 \leq k \leq m} \overline{X}_k \geq y\right) \\ &\Rightarrow P(Y \geq y) = P(Y \geq y) P(Y \geq y) = [P(Y \geq y)]^2, \end{aligned}$$

which is what we needed to complete the proof.

14.6 Martingale Central Limit Theorem

For an iid mean zero sequence of random variables $Z_1, Z_2, \ldots$ with variance one, the central limit theorem says that for large n, $\frac{Z_1+\cdots+Z_n}{\sqrt{n}}$ is approximately standard normal. Suppose now that we consider a mean zero martingale (adapted to some sequence $\{Y_n\}$) $\{X_n, n \geq 0\}$ with $X_0 = 0$ and write $Z_i = X_i - X_{i-1}, i \geq 1$. Then, obviously we can write

$$X_n = X_n - X_0 = \sum_{i=1}^{n} (X_i - X_{i-1}) = \sum_{i=1}^{n} Z_i.$$

The summands Z_i are certainly no longer independent; however, they are uncorrelated (see the chapter exercises). The martingale central limit theorem says that under certain conditions on the growth of the conditional variances $\text{Var}(Z_n \,|Y_0, \ldots, Y_{n-1})$, $\frac{X_n}{\sqrt{n}}$ will still be approximately normally distributed for large n.

The area of martingale central limit theorems is a bit confusing due to an overwhelming variety of central limit theorems, each known as a martingale central limit theorem. In particular, the normalization of X_n can be deterministic or random. Also, there can be a double array of martingales and central limit theorems for them, analogous to Lyapounov's central limit theorem for the independent case. The best source and exposition of martingale central limit theorems is the classic book by Hall and Heyde (1980). We present two specific martingale central limit theorems in this section.

First, we need some notation. Let $\{X_n, n \geq 0\}$ be a zero mean martingale adapted to some sequence $\{Y_n\}$, with $X_0 = 0$. Let

$$\begin{aligned} Z_i &= X_i - X_{i-1}, i \geq 1; \\ \sigma_j^2 &= \text{Var}(Z_j \,|Y_0, \ldots, Y_{j-1}) = E(Z_j^2 \,|Y_0, \ldots, Y_{j-1}); \\ V_n^2 &= \sum_{j=1}^{n} \sigma_j^2; \\ s_n^2 &= E(V_n^2) = E(X_n^2) = \text{Var}(X_n); \end{aligned}$$

(see Section 14.3.2 for the fact that $E(V_n^2)$ and $E(X_n^2)$ are equal if $X_0 = 0$).

The desired result is that $\frac{X_n}{s_n}$ converges in distribution to $N(0, 1)$. The question is under what conditions can one prove such an asymptotic normality result. The conditions that we use are very similar to the corresponding Lindeberg–Lévy conditions in the independent case. Here are the two conditions we assume.

(A) Concentration Condition

$$\frac{V_n^2}{s_n^2} = \frac{V_n^2}{E(V_n^2)} \xrightarrow{P} 1.$$

(B) Martingale Lindeberg Condition

$$\text{For any } \epsilon > 0, \quad \frac{\sum_{j=1}^{n} E(Z_j^2 I_{\{|Z_j| \geq \epsilon s_n\}})}{s_n^2} \xrightarrow{P} 0.$$

Under condition (A), the Lindeberg condition (B) is nearly equivalent to the uniform asymptotic negligibility condition that $\frac{\max_{1 \leq j \leq n} \sigma_j^2}{s_n^2} \xrightarrow{P} 0$. We commonly see such uniform asymptotic negligibility conditions in the independent case central limit theorems. See Hall and Heyde (1980) and Brown (1971) for much additional discussion on the exact role of the Lindeberg condition in martingale central limit theorems. Here is our basic martingale CLT.

Theorem 14.22 (Basic Martingale CLT). *Suppose conditions (A) and (B) hold. Then* $\frac{X_n}{s_n} \overset{\mathcal{L}}{\Rightarrow} Z$, *where* $Z \sim N(0,1)$.

The proof of the Lindeberg–Lévy theorem for the independent case has to be suitably adapted to the martingale structure in order to prove this theorem. The two references above can be consulted for a proof. The Lindeberg condition can be difficult to verify. The following simpler version of martingale central limit theorems suffices for many applications. For this, we need the additional notation

$$\tau_t = \inf \left\{ n > 0 : \sum_{j=1}^{n} \sigma_j^2 \geq t \right\}.$$

Here is our simpler version of the martingale CLT.

Theorem 14.23. *Assume that*

$$|Z_i| \leq K < \infty \quad \textit{for all } i \textit{ and some } K;$$

$$\sum_{j=1}^{\infty} \sigma_j^2 = \infty \textit{ almost surely};$$

$$\frac{t}{\tau_t} \xrightarrow{\text{a.s.}} \sigma^2 \quad \textit{for some finite and positive constant } \sigma^2.$$

Then $\frac{X_n}{\sqrt{n}} \overset{\mathcal{L}}{\Rightarrow} W$, *where* $W \sim N(0, \sigma^2)$.

Exercises

Exercise 14.1. Suppose $\{X_n, n \geq 1\}$ is a martingale adapted to some sequence $\{Y_n\}$. Show that $E(X_{n+m} | Y_1, \ldots, Y_n) = X_n$ for all $m, n \geq 1$.

Exercise 14.2. Suppose $\{X_n, n \geq 1\}$ is a martingale adapted to some sequence $\{Y_n\}$. Fix $m \geq 1$ and define $Z_n = X_n - X_m, n \geq m+1$. Is it true that $\{Z_n\}$ is also a martingale?

Exercise 14.3 (Product Martingale). Let $X_1, X_2, \ldots$ be iid nonnegative random variables with a finite positive mean μ. Identify a sequence of constants c_n such that $Z_n = c_n(\prod_{i=1}^{n} X_i), n \geq 1$ forms a martingale.

Exercise 14.4. Let $\{U_n\}, \{V_n\}$ be martingales, adapted to the same sequence $\{Y_n\}$. Identify, with proof, which of the following are also submartingales, and for those that are not necessarily submartingales, give a counterexample.

(a) $|U_n - V_n|$.
(b) $U_n^2 + V_n^2$.
(c) $U_n - V_n$.
(d) $\min(U_n, V_n)$.

Exercise 14.5 (Bayes Problem). Suppose given p, $X_1, X_2, \ldots$ are iid Bernoulli variables with a parameter p, and the marginal distribution of p is $U[0,1]$. Let $S_n = X_1 + \cdots + X_n, n \geq 1$, and $Z_n = \frac{S_n+1}{n+2}$. Show that $\{Z_n\}$ is a martingale with respect to the sequence $\{X_n\}$.

Exercise 14.6 (Bayes Problem). Suppose given λ, $X_1, X_2, \ldots$ are iid Poisson variables with some mean λ, and the marginal density of λ is $\frac{\beta^\alpha e^{-\beta\lambda}\lambda^{\alpha-1}}{\Gamma(\alpha)}$, where $\alpha, \beta > 0$ are constants. Let $S_n = X_1 + \cdots + X_n, n \geq 1$, and $Z_n = \frac{S_n+\alpha}{n+\beta}$. Show that $\{Z_n\}$ is a martingale with respect to the sequence $\{X_n\}$.

Exercise 14.7 (Bayes Problem). Suppose given μ, $X_1, X_2, \ldots$ are iid $N(\mu, 1)$ variables, and that the marginal distribution of μ is standard normal. Let $S_n = X_1 + \cdots + X_n, n \geq 1$, and $Z_n = \frac{S_n}{n+1}$. Show that $\{Z_n\}$ is a martingale with respect to the sequence $\{X_n\}$.

Exercise 14.8. Suppose $\{X_n\}$ is known to be a submartingale with respect to some sequence $\{Y_n\}$. Show that $\{X_n\}$ is also a martingale if and only if $E(X_n) = E(X_m)$ for all m, n.

Exercise 14.9. Let $X_1, X_2, \ldots$ be a sequence of iid random variables such that $E(|X_1|) < \infty$. For $n \geq 1$, let $X_{n:n} = \max(X_1, \ldots, X_n)$. Show that $\{X_{n:n}\}$ is a submartingale adapted to itself.

Exercise 14.10 (Random Walk). Consider a simple asymmetric random walk with iid steps distributed as $P(X_i = 1) = p, P(X_i = -1) = 1 - p, p < \frac{1}{2}$. Let $S_n = X_1 + \cdots + X_n, n \geq 1$, Show that

(a) $V_n = (\frac{1-p}{p})^{S_n}$ is a martingale.
(b) Show that with probability one, $\sup_n S_n < \infty$.

Exercise 14.11 (Branching Process). Let $\{Z_{ij}\}$ be a double array of iid random variables with mean μ and variance $\sigma^2 < \infty$. Let $X_0 = 1$ and $X_{n+1} = \sum_{j=1}^{X_n} Z_{nj}$. Show that

(a) $W_n = \frac{X_n}{\mu^n}$ is a martingale.
(b) $sup_n E(W_n) < \infty$.
(c) Is $\{W_n\}$ uniformly integrable? Prove or disprove it.

Remark. The process W_n is commonly called a branching process and is important in population studies.

Exercise 14.12 (A Time Series Model). Let $Z_0, Z_1, \ldots$ be iid standard normal variables. Let $X_0 = Z_0$, and for $n \geq 1, X_n = X_{n-1} + Z_n h_n(X_0, \ldots, X_{n-1})$, where for each $n, h_n(x_0, \ldots, x_{n-1})$ is an absolutely bounded function.

Show that $\{X_n\}$ is a martingale adapted to some sequence $\{Y_n\}$, and explicitly identify such a sequence $\{Y_n\}$.

Exercise 14.13 (Another Time Series Model). Let $Z_0, Z_1, \ldots$ be a sequence of random variables such that $E(Z_{n+1} \,|Z_0, \ldots, Z_n) = cZ_n + (1-c)Z_{n-1}, n \geq 1$, where $0 < c < 1$. Let $X_0 = Z_0, X_n = \alpha Z_n + Z_{n-1}, n \geq 1$. Show that α may be chosen to make $\{X_n, n \geq 0\}$ a martingale with respect to $\{Z_n\}$.

Exercise 14.14 (Conditional Centering of a General Sequence). Let $Z_0, Z_1, \ldots$ be a general sequence of random variables, not necessarily independent, such that $E(|Z_k|) < \infty$ for all k. Let $V_n = \sum_{i=1}^{n} [Z_i - E(Z_i \,|Z_0, \ldots, Z_{i-1})], n \geq 1$. Show that $\{V_n\}$ is a martingale with respect to the sequence $\{Z_n\}$.

Exercise 14.15 (The Cross-Product Martingale). Let $X_1, X_2, \ldots$ be independent random variables, with $E(|X_i|) < \infty$ and $E(X_i) = 0$ for all i. For a fixed $k \geq 1$, let $V_{k,n} = \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} X_{i_1} \ldots X_{i_k}, n \geq k$. Show that $\{V_{k,n}\}$ is a martingale with respect to $\{X_n\}$.

Exercise 14.16 (The Wright–Fisher Markov Chain). Consider the Wright-Fisher Markov chain of Example 14.7. Let

$$V_n = \frac{X_n(N - X_n)}{(1 - \frac{1}{N})^n}, \quad n \geq 0.$$

Show that $\{V_n\}_0^N$ is a martingale.

Exercise 14.17 (An Example of Samuel Karlin). Let f be a continuous function defined on $[0, 1]$ and $U \sim U[0, 1]$. Let $X_n = \frac{\lfloor 2^n U \rfloor}{2^n}$, and $V_n = \frac{f(X_n + 2^{-n}) - f(X_n)}{2^{-n}}$. Show that $\{V_n\}$ is a martingale with respect to the sequence $\{X_n\}$.

Exercise 14.18. Let $X_1, X_2, \ldots$ be iid symmetric random variables with mean zero, and let $S_n = \sum_{i=1}^{n} X_i, n \geq 1$, and $S_0 = 0$. Let $\psi(t)$ be the characteristic function of X_1, and $V_n = [\psi(t)]^{-n} e^{itS_n}, n \geq 0$. Show that the real part as well as the imaginary part of $\{V_n\}$ is a martingale.

Exercise 14.19 (Stopping Times). Consider the simple symmetric random walk S_n with $S_0 = 0$. Identify, with proof, which of the following are stopping times, and which among them have a finite expectation.

(a) $\inf\{n > 0 : |S_n| > 5\}$.
(b) $\inf\{n \geq 0 : S_n < S_{n+1}\}$.
(c) $\inf\{n > 0 : |S_n| = 1\}$.
(d) $\inf\{n > 0 : |S_n| > 1\}$.

Exercise 14.20. Let τ be a nonnegative integer-valued random variable, and $\{X_n, n \geq 0\}$ a sequence of random variables, all defined on a common sample space Ω. Prove or disprove that τ is a stopping time adapted to $\{X_n\}$ if and only if for every $n \geq 0$, $I_{\{\tau=n\}}$ is a function of only $X_0, \ldots, X_n$.

Exercise 14.21. Suppose τ_1, τ_2 are both stopping times with respect to some sequence $\{X_n\}$. Is $|\tau_1 - \tau_2|$ necessarily a stopping time with respect to $\{X_n\}$?

Exercise 14.22 (Condition for Optional Stopping Theorem). Suppose $\{X_n, n \geq 0\}$ is a martingale, and τ a stopping time, both adapted to a common sequence $\{Y_n\}$. Show that the equality $E(X_\tau) = E(X_0)$ holds if $E(|X_\tau|) < \infty$, and $E(X_{\min(\tau,n)} I_{\{\tau>n\}}) \to 0$ as $n \to \infty$.

Exercise 14.23 (The Random Walk). Consider the asymmetric random walk $S_n = \sum_{i=1}^n X_i$, where $P(X_i = 1) = p, P(X_i = -1) = q = 1 - p, p > \frac{1}{2}$, and $S_0 = 0$. Let x be a fixed positive integer, and $\tau = \inf\{n > 0 : S_n = x\}$. Show that for $0 < s < 1, E(s^\tau) = (\frac{1-\sqrt{1-4pqs^2}}{2qs})^x$.

Exercise 14.24 (The Random Walk; continued). For the stopping time τ of the previous exercise, show that

$$E(\tau) = \frac{x}{p-q} \quad \text{and} \quad \text{Var}(\tau) = \frac{x[1-(p-q)^2]}{(p-q)^3}.$$

Exercise 14.25 (Gambler's Ruin). Consider the general random walk $S_n = \sum_{i=1}^n X_i$, where $P(X_i = 1) = p \neq \frac{1}{2}, P(X_i = -1) = q = 1 - p$, and $S_0 = 0$. Let a, b be fixed positive integers, and $\tau = \inf\{n > 0 : S_n = b \text{ or } S_n = -a\}$. Show that

$$E(\tau) = \frac{b}{p-q} - \frac{a+b}{p-q} \frac{[1-(\frac{p}{q})^b]}{[1-(\frac{p}{q})^{a+b}]},$$

and that by an application of L'Hospital's rule, this gives the correct formula for $E(\tau)$ even when $p = \frac{1}{2}$.

Exercise 14.26 (Martingales for Patterns). Consider the following martingale approach to a geometric distribution problem. Let $X_1, X_2, \ldots$ be iid Bernoulli variables, with $P(X_i = 1) = p, P(X_i = 0) = q = 1-p$. Let $\tau = \min\{k \geq 1 : X_k = 0\}$, and $\tau_n = \min(\tau, n), n \geq 1$.

Define $V_n = \frac{1}{q}\sum_{i=1}^{n} I_{\{X_i=0\}}, n \geq 1$.

(a) Show that $\{V_n - n\}$ is a martingale with respect to the sequence $\{X_n\}$.
(b) Show that $E(V_{\tau_n}) = E(\tau_n)$ for all n.
(c) Hence, show that $E(\tau) = E(V_\tau) = \frac{1}{q}$.

Exercise 14.27 (Martingales for Patterns). Let $X_1, X_2, \ldots$ be iid Bernoulli variables, with $P(X_i = 1) = p$, $P(X_i = 0) = q = 1 - p$. Let τ be the first k such that X_{k-2}, X_{k-1}, X_k are each equal to one (e.g., the number of tosses of a coin necessary to first obtain three consecutive heads), and $\tau_n = \min(\tau, n), n \geq 3$.

Define

$$V_n = \frac{1}{p^3}\sum_{i=1}^{n-2} I_{\{X_i=X_{i+1}=X_{i+2}=1\}} + \frac{1}{p^2} I_{\{X_n=X_{n-1}=1\}} + \frac{1}{p} I_{\{X_n=1\}}, \quad n \geq 3.$$

(a) Show that $\{V_n - n\}$ is a martingale with respect to the sequence $\{X_n\}$.
(b) Show that $E(V_{\tau_n}) = E(\tau_n)$ for all n.
(c) Hence, show that

$$E(\tau) = E(V_\tau) = \frac{1}{p} + \frac{1}{p^2} + \frac{1}{p^3}.$$

(d) Generalize to the case of the expected waiting time for obtaining r consecutive 1s.

Exercise 14.28. Let $\{X_n, n \geq 0\}$ be a martingale.

(a) Show that $\lim_{n\to\infty} E(|X_n|)$ exists.
(b) Show that for any stopping time τ, $E(|X_\tau|) \leq \lim_{n\to\infty} E(|X_n|)$.
(c) Show that if $\sup_n E(|X_n|) < \infty$, then $E(|X_\tau|) < \infty$ for any stopping time τ.

Exercise 14.29 (Inequality for Stopped Martingales). Let $\{X_n, n \geq 0\}$ be a martingale, and τ a stopping time adapted to $\{X_n\}$. Show that $E(|X_\tau|) \leq 2\sup_n E(X_n^+) - E(X_1) \leq 3\sup_n E(|X_n|)$.

Exercise 14.30. Let $X_1, X_2, \ldots$ be iid random variables such that $E(|X_1|) < \infty$. Consider the random walk $S_n = \sum_{i=1}^{n} X_i, n \geq 1$ and $S_0 = 0$. Let τ be a stopping time adapted to $\{S_n\}$. Show that if $E(|S_\tau|) = \infty$, then $E(\tau)$ must also be infinite.

Exercise 14.31. Let $\{X_n, n \geq 0\}$ be a martingale, with $X_0 = 0$. Let $V_i = X_i - X_{i-1}, i \geq 1$. Show that for any $i \neq j$, V_i and V_j are uncorrelated.

Exercise 14.32. Let $\{X_n, n \geq 1\}$ be some sequence of random variables. Suppose $S_n = \sum_{i=1}^{n} X_i, n \geq 1$, and that $\{S_n, n \geq 1\}$ forms a martingale. Show that for any $i \neq j$, $E(X_i X_j) = 0$.

Exercise 14.33. Let $\{X_n, n \geq 0\}$ and $\{Y_n, n \geq 0\}$ both be square integrable martingales, adapted to some common sequence. Let $X_0 = Y_0 = 0$. Show that $E(X_n Y_n) = \sum_{i=1}^{n} E[(X_i - X_{i-1})(Y_i - Y_{i-1})]$ for any $n \geq 1$.

Exercise 14.34. Give an example of a submartingale $\{X_n\}$ and a convex function f such that $\{f(X_n)\}$ is not a submartingale.

Remark. Such a function f cannot be increasing.

Exercise 14.35 (Characterization of Uniformly Integrable Martingales). Let $\{X_n\}$ be uniformly integrable and a martingale with respect to some sequence $\{Y_n\}$. Show that there exists a random variable Z such that $E(|Z|) < \infty$ and such that for each n, $E(Z\,|Y_1, \ldots, Y_n) = X_n$ with probability one.

Exercise 14.36 (L_p-Convergence of a Martingale). Let $\{X_n, n \geq 0\}$ be a martingale, or a nonnegative submartingale. Suppose for some $p > 1$, $\sup_n E(|X_n|^p) < \infty$. Show that there exists a random variable X, almost surely finite, such that $E(|X_n - X|^p) \to 0$ and $X_n \overset{\text{a.s.}}{\to} X$ as $n \to \infty$.

Exercise 14.37. Let $\{X_n\}$ be a nonnegative martingale. Suppose $E(X_n) \to 0$ as $n \to \infty$. Show that $X_n \overset{\text{a.s.}}{\to} 0$.

Exercise 14.38. Let $X_1, X_2, \ldots$ be iid normal variables with mean zero and variance σ^2. Show that $\sum_{n=1}^{\infty} \frac{\sin(n\pi x)}{n} X_n$ converges with probability one for any given real number x.

Exercise 14.39 (Generalization of Maximal Inequality). Let $\{X_n, n \geq 0\}$ be a nonnegative submartingale, and $\{b_n, n \geq 0\}$ a nonnegative nonincreasing sequence of constants such that $b_n \to 0$ as $n \to \infty$, and $\sum_{n=0}^{\infty}[b_n - b_{n+1}]E(X_n)$ converges.

(a) Show that for any $x > 0$,

$$P(\sup_{n\geq 0} b_n X_n \geq x) \leq \frac{1}{x} \sum_{n=0}^{\infty} [b_n - b_{n+1}]E(X_n).$$

(b) Derive the Kolmogorov maximal inequality for nonnegative submartingales as a corollary to part (a).

Exercise 14.40 (Decomposition of an L_1-Bounded Martingale). Let $\{X_n\}$ be an L_1-bounded martingale adapted to some sequence $\{Y_n\}$, that is, $\sup_n E(|X_n|) < \infty$.

(a) Define $Z_{m,n} = E[|X_{m+1}|\,|Y_1, \ldots, Y_n]$. Show that $Z_{m,n}$ is nondecreasing in m.
(b) Show that for fixed n, $Z_{m,n}$ converges almost surely.
(c) Let $U_n = \lim_m Z_{m,n}$. Show that $\{U_n\}$ is an L_1-bounded martingale.
(d) Show that X_n admits the decomposition $X_n = U_n - V_n$, where both U_n, V_n are nonnegative L_1-bounded martingales.

References

Azuma, K. (1967). Weighted sums of certain dependent random variables, *Tohoku Math. J.*, 19, 357–367.

Brown, B.M. (1971). Martingale central limit theorems, *Ann. Math. Statist.*, 42, 59–66.

Burkholder, D.L. (1973). Distribution function inequalities for martingales, *Ann. Prob.*, 1, 19–42.

Burkholder, D.L., Davis, B., and Gundy, R. F. (1972). Integral inequalities for convex functions of operators on martingales, *Proc. Sixth Berkeley Symp. Math. Statist. Prob*., Vol. II, 223–240, University of California Press, Berkeley.

Chow, Y.S. and Teicher, H. (2003). *Probability Theory: Independence, Interchangeability, Martingales*, Springer, New York.

Chung, K.L. (1974). *A Course in Probability*, Academic Press, New York.

DasGupta, A. (2010). *Fundamentals of Probability: A First Course*, Springer, New York.

Davis, B. (1970). On the integrability of the martingale square function, *Israel J. Math*., 8, 187–190.

Devroye, L. (1991). *Exponential Inequalities in Nonparametric Estimation, Nonparametric Functional Estimation and Related Topics*, 31–44, Kluwer Acad. Publ., Dordrecht.

Doob, J.L. (1971). What is a martingale?, *Amer. Math. Monthly*, 78, 451–463.

Fristedt, B. and Gray, L. (1997). *A Modern Approach to Probability Theory*, Birkhäuser, Boston.

Hall, P. and Heyde, C. (1980). *Martingale Limit Theory and Its Applications*, Academic Press, New York.

Heyde, C. (1972). Martingales: A case for a place in a statistician's repertoire, *Austr. J. Statist*., 14, 1–9.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables, *J. Amer. Statisto Assoc*., 58, 13–30.

Karatzas, I. and Shreve, S. (1991). *Brownian Motion and Stochastic Calculus*, Springer, New York.

Karlin, S. and Taylor, H.M. (1975). *A First Course in Stochastic Processes*, Academic Press, New York.

McDiarmid, C. (1989). *On the Method of Bounded Differences, Surveys in Combinatorics*, London Math. Soc. Lecture Notes, 141, 148–188, Cambridge University Press, Cambridge, UK.

Williams, D. (1991). *Probability with Martingales*, Cambridge University Press, Cambridge, UK.

http://www.springer.com/978-1-4419-9633-6

Probability for Statistics and Machine Learning
Fundamentals and Advanced Topics
DasGupta, A.
2011, XX, 784 p., Hardcover
ISBN: 978-1-4419-9633-6