Unit 1: Introduction to Martingale Theory

Instructor: Quan Zhou

1.1 Martingales and stopping times

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space. Let $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 = \{0, 1, 2, \dots\}$.

Definition 1.1. $(\mathcal{F}_n)_{n \in \mathbb{N}_0}$ is called a filtration, if it is an non-decreasing sequence of σ-algebras; that is, each $\mathcal{F}_n$ is a σ-algebra on Ω and $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}$. Define $\mathcal{F}_\infty = \sigma\left(\bigcup_n \mathcal{F}_n\right)$. We say $(\Omega, \mathcal{F}, (\mathcal{F}_n), \mathsf{P})$ is a filtered probability space.

Definition 1.2. A sequence of random variables $(X_n)_{n \geq 0}$ is said to be adapted to $(\mathcal{F}_n)_{n \geq 0}$ if $X_n \in \mathcal{F}_n$ for each n.

Definition 1.3. A sequence of random variables $(X_n)_{n \geq 0}$ adapted to $(\mathcal{F}_n)_{n \geq 0}$ is said to be a martingale w.r.t $(\mathcal{F}_n)_{n \geq 0}$, if for each $n \in \mathbb{N}_0$, we have (i) $\mathsf{E}|X_n| < \infty$, and (ii) $\mathsf{E}[X_{n+1} \,|\, \mathcal{F}_n] = X_n$, a.s.

Definition 1.4. Let $T \colon \Omega \to \mathbb{N}_0 \cup \{\infty\}$ be measurable. We say T is a stopping time w.r.t. $(\mathcal{F}_n)_{n \geq 0}$ if $\{T \leq n\} \in \mathcal{F}_n$ for each $n \in \mathbb{N}_0$.

1.2 Examples in probability theory

Example 1.1. Let's flip a fair coin. What is the expected number of flips needed to get the sequence HHHH (i.e., four consecutive heads)?

A martingale solution. Imagine that we are betting on heads/tails in a casino. For each dollar I bet, I get an additional dollar if I am correct, and I lose that dollar if I am wrong. Since the coin is assumed fair, this game is fair, which intuitively means that I should not be able to make or lose money in expectation. Here is my strategy: at each flip, I bet all the money I have from previous bets plus one additional dollar on heads. Now let T denote the first time that the sequence HHHH first happens. My expected profit at time T should be zero. Hence,

$$\mathsf{E}(T) = 16 + 8 + 4 + 2 = 30.$$

In the above calculation, $\mathsf{E}(T)$ is the expected total number of dollars that come out of my pocket. It should be equal to the money that I have after the

T-th flip, which is equal to 30. This solution is extremely simple, compared to a more standard approach based on Markov chains.

Is this calculation really rigorous? The filtered probability space for this question can be constructed as follows: Ω is the collection of all possible outcomes of infinitely many flips, $\mathcal{F}$ is the power set on Ω, P is the probability measure under which all flips are fair and independent, and $\mathcal{F}_n$ is the σ-algebra generated by the first n flips. Let Y_n denote my net profit after n flips. It is not difficult to prove that Y_n is a martingale, and thus $\mathsf{E}(Y_n) = 0$ for each n. However, in the above calculation, we actually assumed that $\mathsf{E}(Y_T) = 0$, where T is a stopping time (T is random, not deterministic). This step was not justified, and for now, it is unclear why it is true. If you think replacing n with a stopping time T is always harmless, see the next simple counterexample.

Example 1.2. Let $Z_1, Z_2, \dots$ be i.i.d. random variables such that $\mathsf{P}(Z_n = 1) = \mathsf{P}(Z_n = -1) = 1/2$. Define $\mathcal{F}_0 = \{\Omega, \emptyset\}$, $\mathcal{F}_n = \sigma(Z_1, \dots, Z_n)$ for each $n \geq 1$, and $X_n = Z_1 + \cdots + Z_n$. Clearly, X_n is a martingale w.r.t. $(\mathcal{F}_n)$ and $\mathsf{E}(X_n) = 0$ for each n. Now define $T = \min\{t \colon X_t = 1\}$. It can be shown that $\mathsf{P}(T < \infty) = 1$, and thus $\mathsf{E}(X_T) = 1 \neq 0$.

Intuitive explanation. In a gambling context, we can think of Z_n as my net profit from the n-th bet. Each individual bet is fair. And now if I choose to stop betting at time T, my expected profit is 1. This is of course too good to be true. Indeed, we can show that $\mathsf{E}(T) = \infty$, which suggests that implementing this strategy in practice may be problematic. This and the previous example motivate us to study a key result in martingale theory, optional sampling theorem. It tells us that, for a martingale (X_n) and stopping time T, when we have $\mathsf{E}(X_T) = \mathsf{E}(X_0)$.

Example 1.3. Let $\{Z_{n,i} \colon n \in \mathbb{N}_0, \, i \in \mathbb{N}\}$ be a collection of i.i.d. random variables taking values in $\mathbb{N}_0$. Let $X_0 = 1$, and

$$X_n = \sum_{i=1}^{X_{n-1}} Z_{n-1,i}, \quad \forall \, n \in \mathbb{N}.$$

We say $(X_n)_{n \in \mathbb{N}_0}$ is a Galton-Watson branching process. It can be interpretd as the evolution of a population, which at time 0 only has $X_0 = 1$ individual. Each individual only lives for one unit of time, and the i-th individual at time n has $Z_{n,i}$ offspring. So X_n is the number of individuals at time n. To avoid uninteresting cases, we assume $\mu = \mathsf{E}(Z_{0,1}) \in (0, \infty)$.

Constructing a martingale. Define $\mathcal{F}_n = \sigma(\{Z_{k,i} \colon k < n,\, i \in \mathbb{N}\})$; you can check that (X_n) is adapted to $(\mathcal{F}_n)$. Define $W_n = X_n/\mu^n$, and it can be quickly verified that W_n is a martingale such that $\mathsf{E}(W_n) = 1$ for each n.

Question. Does $\lim_{n\to\infty} W_n$ exist a.s.? This probably is not very clear at first glance. We will see that a fundamental martingale convergence result immediately shows that $W_\infty = \lim_{n\to} W_n$ exists a.s. Can we further say anything about $\mathsf{E}(W_\infty)$?

Example 1.4. This problem is known as the Mabinogion sheep. Consider a magical flock of sheep; some are black, and the others are white. At each time $n \in \mathbb{N}$, a sheep is drawn randomly (with equal probability) from the whole flock and bleats. If the bleating sheep is white, one black sheep becomes white instantly; if the bleating sheep is black, a white sheep becomes black. Of course when all sheep are black or all are white, this magical process stops. Now suppose that we are allowed to do the following: at each time $n \in \mathbb{N}_0$, we can remove any number of white sheep from the flock. How to maximize the expected final number of black sheep?

Solution. Let w be the number of white sheep and b be the number of black ones. Here is the optimal policy: if $b > w$ or $b = 0$, we do nothing; if $w \geq b$, reduce w to $b - 1$. The optimality of this policy can be proved by using martingale theory (the calculation is somewhat complicated). Actually, this type of problems is called stochastic control, for which martingale theory is a fundamental tool. A more famous example in stochastic control is call/put option pricing in mathematical finance, which you can easily find online.

1.3 Examples in mathematical statistics

Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathsf{E}(X_1) = 0$ and $\mathrm{Var}(X_1) = 1$. Define $S_n = X_1 + \cdots + X_n$. We know that

$$\begin{aligned} \text{SLLN}: &\quad S_n/n \overset{\text{a.s.}}{\to} 0, \\ \text{LIL}: &\quad \limsup_{n\to\infty} S_n/\sqrt{2n\log(\log n)} = 1, \text{a.s.} \\ \text{CLT}: &\quad S_n/\sqrt{n} \overset{\text{D}}{\to} N(0,1). \end{aligned}$$

'SLLN' denotes the strong law of large numbers, 'LIL' denotes the law of the iterated logarithm, and 'CLT' denotes the central limit theorem. In mathematical statistics, we often want to establish the above three types of results for some estimator, and very often the independence assumption becomes too restrictive. In the seminal book [3], the authors extended these results to the case where S_n is a zero-mean martingale, and they argued that the limit theory for martingales essentially covers the limit theory for

(1) processes with independent increments,

(2) Markov processes,

(3) stationary processes,

(4) processes with "asymptotically independent" increments.

Thus, we can probably claim that martingale theory provides the principal tool for building the statistical theory concerning dependent data.

Example 1.5. Let $X_1, X_2, \dots$ be i.i.d. random elements. Let H be a bivariate function such that (i) $H(x, y) = H(y, x)$, (ii) $\mathsf{E}[H(X_1, X_2)] = 0$, and (iii) $\mathsf{E}[H(X_1, X_2) \mid X_1] = 0$ a.s. Then we say

$$U_n = \sum_{1 \le i < j \le n} H(X_i, X_j)$$

is a centered, degenerate U-statistic. To construct a martingale, observe that

$$\begin{aligned} U_n = \; & H(X_1, X_2) + \\ & H(X_1, X_3) + H(X_2, X_3) + \\ & H(X_1, X_4) + H(X_2, X_4) + H(X_3, X_4) + \\ & \quad \vdots \\ & H(X_1, X_n) + H(X_2, X_n) + \cdots + H(X_{n-1}, X_n). \end{aligned}$$

So $U_n = \sum_{j=2}^{n} Y_j$, where

$$Y_j = \sum_{i=1}^{j-1} H(X_i, X_j).$$

Letting $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$, we have $\mathsf{E}(Y_n \mid \mathcal{F}_{n-1}) = 0$. Hence, $(U_n)_{n \ge 2}$ is a martingale w.r.t. $(\mathcal{F}_n)$. Martingale CLT then can be used to obtain CLT for U-statistics. See [2] for a classical application to kernel density estimation.

Example 1.6. Let $X_1, X_2, \dots$ denote our (possibly dependent) observations. The underlying probability measure, P_θ, depends on an unknown parameter θ. Let $L_n(\theta)$ denote the likelihood function with first n observations, and let

$$s_n(\theta) = \frac{\mathrm{d} \log L_n(\theta)}{\mathrm{d}\theta}.$$

As usual, let $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$. Set $L_0 = 1$ and $s_0 = 0$. Under some regularity conditions, we can show that (s_n) is a martingale w.r.t. $(\mathcal{F}_n)$ under the measure P_θ. Define the "conditional" Fisher information by

$$I_n(\theta) = \sum_{i=1}^{n} \mathsf{E}_\theta \left[(s_i(\theta) - s_{i-1}(\theta))^2 \,|\, \mathcal{F}_{i-1} \right].$$

Note that when observations are independent, $I_n(\theta)$ is the standard Fisher information. Martingale theory can be used to show that $s_n(\theta)/\sqrt{I_n(\theta)} \overset{\mathrm{D}}{\to} N(0,1)$ under certain regularity conditions, generalizing the score test for i.i.d. observations, and to study the asymptotic efficiency of maximum likelihood estimators. See [3].

1.4 An example in machine learning

Example 1.7. Consider a function $M(\theta)$ such that $M(\theta) = 0$ has a unique solution at $\theta = \theta^*$. Suppose that we cannot observe $M(\theta)$ but we have access to an unbiased estimator $Y(\theta)$ for each θ. The famous Robbins-Monro algorithm begins with an initial guess θ_0 and then update it iteratively by

$$\theta_{n+1} = \theta_n - a_n Y_n(\theta_n),$$

where (a_n) is a decreasing sequence of positive constants, and $Y_n(\theta_n)$ is the unbiased estimator for $M(\theta_n)$ generated in the n-th iteration independently of the previous iterations. Under some conditions, we have $\theta_n \overset{\text{a.s.}}{\to} \theta$. Now imagine that we have a large number of independent observations whose distributions depend on an uknown parameter θ, and $M(\theta)$ is the log-likelihood function. Then, we can let $Y(\theta)$ be the log-likelihood of θ evaluated by only using a random subsample of the entire data set, which can be quickly shown to be unbiased. In this context, the Robbins-Monro algorithm is known as the stochastic gradient descent (SGD), which is widely used in machine learning to find the maximum likelihood estimator.

Martingale proof. The martingale theory can be used to study the convergence of θ_n. The key idea is to rewrite the Robbins-Monro update by

$$\theta_{n+1} = \theta_n - a_n M(\theta_n) - a_n Z_{n+1},$$

where $Z_{n+1} = Y_n(\theta_n) - M(\theta_n)$. Note that $\mathsf{E}(Z_{n+1} \,|\, \theta_1, \dots, \theta_n) = 0$; i.e., (Z_n) is a martingale difference sequence.

References

[1] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[2] Peter Hall. Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of multivariate analysis*, 14(1):1–16, 1984.

[3] Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

[4] Achim Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

[5] Sheldon M Ross. *Stochastic processes*. John Wiley & Sons, 1995.

[6] David Williams. *Probability with martingales*. Cambridge university press, 1991.