Caltech Department of Mathematics

Ma 3/103 KC Border
Introduction to Probability and Statistics Winter 2017

Lecture 15: Markov Chains and Martingales

This material is not covered in the textbooks. These notes are still in development.

15.1 ⋆ Markov chains

Most authors, e.g., Samuel Karlin [9, p. 27] or Joseph Doob [6, p. 170], define a **Markov chain** to be a discrete time stochastic process where the random variables are discrete, and where the following **Markov property** holds:

> For every $t_1 < t_2 < \cdots < t_n < t_{n+1}$, the conditional distribution of $X_{t_{n+1}}$ given $X_{t_n}, \ldots, X_{t_1}$ is the same as that given X_{t_n}. That is,
>
> $$P\left(X_{t_{n+1}} = x_{n+1} \,\middle|\, X_{t_n} = x_n, X_{t_{n-1}} = x_{n-1}, \ldots, X_{t_1} = x_1\right) = P\left(X_{t_{n+1}} = x_{n+1} \,\middle|\, X_{t_n} = x_n\right).$$
>
> That is, the future depends on the past only through the present, not the entire history.

The value of a random variable is usually referred to as the **state** of the chain. There are many examples where the numerical magnitude of X_t is not really of interest, it is just an ID number for the state.

> For many purposes we simply number the states $1, 2, 3, \ldots$, and the value of X_t is interpreted as the number of the state.

Here are some examples of Markov chains:

- A deck of n cards is one of $n!$ states, each state being an order of the deck. Shuffling is a random experiment that changes the state. Assign each order an ID number, and let X_0 denote the original state, and X_t denote the state after t shuffles. Clearly this is a Markov chain, where the numerical magnitude of X_t is not of interest.

If you are interested in the details of card shuffling, I highly recommend the paper by Dave Bayer and Persi Diaconis [1] and its references. Among other things they argue that it takes at least 7 riffle shuffles to get an acceptable degree of randomness.

- A random walk (on a lattice) is a Markov chain.
- Let X_t denote the fortune (wealth) of a gambler after t \$1 bets. If the bets are independent, then this is a Markov chain.
- The **branching process**: Suppose an organism lives one period and produces a random number X progeny during that period, each of whom then reproduces the next period, etc. The population X_n after n generations is a Markov chain.
- Queueing: Customers arrive for service each period according to a probability distribution, and are served in order, which takes one period. The state of the system is the length of the queue, which is a Markov chain.
- In information theory, see, e.g., Thomas Cover and Joy Thomas [4, p. 34], the term Markov chain can refer to a sequence of just three random variables, X, Y, Z if the joint probability can be written as

$$p(x \mid y, z) = p(x \mid y).$$

15.2 ⋆ Markov Chains and Transition Matrices

A Markov chain is **time-invariant**, or **stationary**, if the distribution of $X_{t+s} \,\big|\, X_t$ does not depend on t. (Some authors, e.g., Kemeny and Snell [10, Definition 2.1.3, p. 25] or Norris [13, p. 2], make time-invariance part of the definition of a Markov chain. Others, such as Karlin [9, pp. 19–20, 27], do not.) When $X_t = i$ and $X_{t+1} = j$, we say the that the chain makes a **transition** from state i to state j at time $t+1$, and we often use the notation

$$i \to j$$

to indicate a transition event. This may be a little confusing, since $i \to j$ does not indicate which t we are talking about. But for a time-invariant Markov chain, t doesn't matter in that the probability of this event does not depend on t. In this case we can define the **transition probability**

$$p(i,j) = P\left(X_{t+1} = j \,\big|\, X_t = i\right),$$

which is independent of t. We may also write p_{ij} or $p(i \to j)$ for $p(i,j)$.

For a time-invariant m-state Markov chain the $m \times m$ matrix P of transition probabilities

$$P = \big[p(i,j)\big] = \big[p_{ij}\big]$$

is called the **transition matrix** for the chain. (It's even possible to consider infinite matrices, but we won't do that here.)

For each row i of the transition matrix P, the row sum $\sum_{j=1}^m p_{ij}$ must be equal to one. A nonnegative square matrix with this property is called a **stochastic matrix**, and there is a vast literature on such matrices.

15.2.1 Two-step transition probabilities

The transition matrix tells everything about the evolution of the m-state time-invariant Markov chain from its initial state X_0. If p_{ij} is the probability of transitioning from state i to state j in one step, what is the probability of transitioning from i to j in exactly two steps? That is, what is

$$p_{ij}^{(2)} = P\left(X_{t+2} = j \,\big|\, X_t = i\right)?$$

By definition this is just

$$P\left(X_{t+2} = j \,\big|\, X_t = i\right) = \frac{P\left(X_{t+2} = j \;\&\; X_t = i\right)}{P\left(X_t = i\right)}. \tag{1}$$

The intermediate state X_{t+1} must take on one of the values $k = 1, \dots, m$. So the event

$$\left(X_{t+2} = j \;\&\; X_t = i\right)$$

is the disjoint union

$$\bigcup_{k=1}^m \left(X_t = i \;\&\; X_{t+1} = k \;\&\; X_{t+2} = j\right).$$

Thus we may write

$$P\left(X_{t+2} = j \,\big|\, X_t = i\right) = \frac{\sum_{k=1}^m P\left(X_t = i \;\&\; X_{t+1} = k \;\&\; X_{t+2} = j\right)}{P\left(X_t = i\right)}. \tag{1′}$$

By the multiplication rule (Section 4.6), for each k,

$$\begin{aligned} &P\left(X_t = i \;\&\; X_{t+1} = k \;\&\; X_{t+2} = j\right) \\ &\qquad = P\left(X_t = i\right) P\left(X_{t+1} = k \,\big|\, X_t = i\right) P\left(X_{t+2} = j \,\big|\, X_{t+1} = k \;\&\; X_t = i\right). \end{aligned} \tag{2}$$

By the Markov property

$$P\left(X_{t+2} = j \,\middle|\, X_{t+1} = k \;\&\; X_t = i\right) = P\left(X_{t+2} = j \,\middle|\, X_{t+1} = k\right). \tag{3}$$

Combining (1′), (2), and (3) gives

$$p_{ij}^{(2)} = \sum_{k=1}^{m} p_{ik} p_{kj},$$

but this just

$$\text{the } i,j \text{ entry of the matrix } P^2.$$

Similarly, the probability $p_{ij}^{(n)}$ of transitioning from i to j in n steps is the i,j entry of the matrix P^n. That is, calculating the distribution of future states is just an exercise in matrix multiplication.

$$P\left(X_{t+n} = j \,\middle|\, X_t = i\right) \text{ is the } (i,j) \text{ entry of the matrix } P^n.$$

This provides a powerful tool for studying the behavior of a Markov chain.

I recommend **ACM/EE 116. Introduction to Stochastic Processes and Modeling** if you want to learn more about this, and **CS/EE 147. Network Performance Analysis** for applications.

15.3⋆ Markov chains and graphs

From now on we will consider only time-invariant Markov chains.

The nature of reachability can be visualized by considering the set states to be a **directed graph** where the set of **nodes** or **vertexes** is the set of states, and there is a **directed edge** from i to j if $P(i \to j) = p_{ij} > 0$. An arrow from node i to node j is used to indicate that the transition $i \to j$ can occur (with nonzero probability) in one step. A loop at a node i indicates that the transition $i \to i$ (remaining in state i) has nonzero probability. The edges of the graph are labeled with the probability of the transition. The state j is reachable from i if there is a **path** in the graph from i to j.

For instance, the transition matrix P of Example 15.4.1 corresponds to the graph in Figure 15.1.

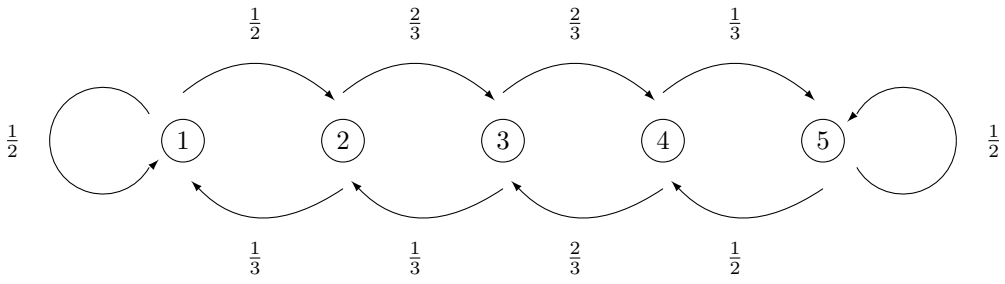

Figure 15.1. The graph of a 5-state Markov chain.

15.4⋆ Irreducible Markov chains

We say that state j is **reachable** from i if $p_{ij}^{(n)} > 0$ for some n. If states i and j are mutually reachable, then we say they **communicate**, denoted $i \leftrightarrow j$. The relation $\leftrightarrow$ is an equivalence relation and partitions the states. When every state communicates with every other state, the chain is called **irreducible** or **indecomposable**.

15.4.1 Example Here is an example of an irreducible 5-state transition matrix. Its graph is given in Figure 15.1.

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & 0 & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

And here are a few successive powers (n-step transitions)

$$P^2 = \begin{bmatrix} \frac{5}{12} & \frac{1}{4} & \frac{1}{3} & 0 & 0 \\ \frac{1}{6} & \frac{7}{18} & 0 & \frac{4}{9} & 0 \\ \frac{1}{9} & 0 & \frac{2}{3} & 0 & \frac{2}{9} \\ 0 & \frac{2}{9} & 0 & \frac{11}{18} & \frac{1}{6} \\ 0 & 0 & \frac{1}{3} & \frac{1}{4} & \frac{5}{12} \end{bmatrix} \qquad P^3 = \begin{bmatrix} \frac{7}{24} & \frac{23}{72} & \frac{1}{6} & \frac{2}{9} & 0 \\ \frac{23}{108} & \frac{1}{12} & \frac{5}{9} & 0 & \frac{4}{27} \\ \frac{1}{18} & \frac{5}{18} & 0 & \frac{5}{9} & \frac{1}{9} \\ \frac{2}{27} & 0 & \frac{5}{9} & \frac{1}{12} & \frac{31}{108} \\ 0 & \frac{1}{9} & \frac{1}{6} & \frac{31}{72} & \frac{7}{24} \end{bmatrix}$$

$$P^{100} = \begin{bmatrix} 0.0952381 & 0.142857 & 0.285715 & 0.285714 & 0.190476 \\ 0.0952380 & 0.142858 & 0.285713 & 0.285715 & 0.190476 \\ 0.0952382 & 0.142857 & 0.285715 & 0.285713 & 0.190476 \\ 0.0952380 & 0.142858 & 0.285713 & 0.285715 & 0.190476 \\ 0.0952381 & 0.142857 & 0.285715 & 0.285714 & 0.190476 \end{bmatrix}$$

$$P^{200} = \begin{bmatrix} 0.0952381 & 0.142857 & 0.285714 & 0.285714 & 0.190476 \\ 0.0952381 & 0.142857 & 0.285714 & 0.285714 & 0.190476 \\ 0.0952381 & 0.142857 & 0.285714 & 0.285714 & 0.190476 \\ 0.0952381 & 0.142857 & 0.285714 & 0.285714 & 0.190476 \\ 0.0952381 & 0.142857 & 0.285714 & 0.285714 & 0.190476 \end{bmatrix}$$

This last matrix has the following interesting property: for any i, i', j, we have

$$p_{ij}^{(200)} \approx p_{i'j}^{(200)}.$$

In other words, the initial state has no effect on the long-rum distribution of states. □

In the example above, it looks as though the powers of P are converging to a limiting matrix. Indeed they are. In fact, you can express each $p^{(n)}(i,j)$ as a linear combination of n^{th} powers of the characteristic roots of P. All the characteristic roots of a stochastic matrix are $\leqslant 1$ in absolute value and every stochastic matrix has an eigenvalue equal to 1 (corresponding to the vector$[1, \ldots, 1]$. If the eigenspace of the eigenvalue 1 has dimension 1, then $P^{(n)}$ necessarily has a limit. For details, see, e.g., Debreu and Herstein [5].

15.5 ⋆ Invariant distributions

Suppose I choose the initial state (X_0) according to some probability vector $x = [x_1, \dots, x_m]$ on states. (That is, $x \in \boldsymbol{R}^{\mathrm{m}}$ and each $x_i \geqslant 0$, and $\sum_{i=1}^m x_i = 1$. We shall assume that x is given as a row vector, or $1 \times m$ matrix.) Then xP gives the probability distribution on states at time $t = 1$,

$$P(X_1 = j) = \sum_{k=1}^m P\left(X_1 = j \,\middle|\, X_0 = k\right) P(X_0 = k) = \sum_{k=1}^m p_{kj} x_k = (xP)_j.$$

Likewise xP^2 is the distribution of states at time $t = 2$, etc.

A probability distribution x on states is **invariant** if

$$xP = x.$$

That is, x is an eigenvector of P corresponding to the eigenvalue 1.

Does every m-state Markov chain have an invariant distribution? The answer is Yes, but the proof is beyond the scope of this course.[1] The main theorem in this regard is known as the Perron–Frobenius Theorem, see, e.g., Debreu and Herstein [5] or Wielandt [15].

Is the invariant distribution unique? Not necessarily. For the two-state transition matrix

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

every distribution is invariant.

N.B. The above result used the fact that the are only finitely many states. For infinite state Markov chains, there may be no invariant distribution. For instance, if the set of states is $\mathbb{N} = \{1, 2, 3, \dots\}$, the transition probabilities

$$p(i, j) = \begin{cases} 1 & \text{if } j = i + 1 \\ 0 & \text{otherwise} \end{cases}$$

do not admit an invariant distribution—for each n, after n steps the probability of being in state n is zero. The conditions for the existence of an invariant distribution in the general case are beyond the scope of this course. But if you want a good starting point, try the book by Leo Breiman [3].

15.6 ⋆ Invariant distributions as limits

Consider the transition matrix

$$P = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}.$$

It has the invariant distribution $x = [1/3, 1/3, 1/3]$. Let y be the distribution that gives state 1 for sure, $y = [1, 0, 0]$. Now consider the sequence $yP, yP^2, yP^3, \dots$. Some of the terms are reproduced here:

$$yP = \begin{bmatrix} 0 \\ 0.5 \\ 0.5 \end{bmatrix}, \quad yP^2 = \begin{bmatrix} 0.5 \\ 0.25 \\ 0.25 \end{bmatrix}, \quad yP^3 = \begin{bmatrix} 0.25 \\ 0.375 \\ 0.375 \end{bmatrix}, \quad \dots, \quad yP^{20} = \begin{bmatrix} 0.33333 \\ 0.33333 \\ 0.33333 \end{bmatrix}$$

[1] Here's my favorite proof, taken from Debreu and Herstein [5]. Let Δ denote the set of probability vectors in $\boldsymbol{R}^{\mathrm{m}}$. Note that it is a closed, bounded, and convex set. If x is a probability vector, then so is xP. Thus the mapping $x \mapsto xP$ maps Δ into itself. It is also a continuous function. The Brouwer Fixed Point Theorem says that whenever a continuous function maps a closed bounded convex subset of $\boldsymbol{R}^{\mathrm{m}}$ into itself, it must have a fixed point. That is, there is an $\bar{x}$ satisfying $\bar{x}P = \bar{x}$. (For a simple proof of the Brouwer Theorem, I'm partial to Border [2], but Franklin [7] and Milnor [11] provide alternatives proofs that you may prefer.)

This sequence is indeed converging to the invariant distribution. But this does not happen for every transition matrix. For instance,

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

has the unique invariant distribution $[1/2, 1/2]$, but setting $y = [1, 0]$ gives $yP^k = [0, 1]$ for k odd, and $yP^k = [1, 0]$ for k even. This sequence oscillates rather than converges. The problem is that this Markov chain is **periodic**.

The **period** of a state i is the greatest common divisor of $\{n : p_{ii}^{(n)} > 0$. The state is **aperiodic** if this gcd is 1. This is equivalent to saying that for all n large enough, $p_{ii}^{(n)} > 0$. If the chain is irreducible and has an a periodic state, then every state is aperiodic [13, Lemma 1.8.2, p. 41]. If every state is aperiodic, we say that the chain is aperiodic.

You may find the following theorem in Breiman [3, Theorem 6.20, p. 172] or Norris [13, Theorem 1.8.3, pp. 41–42].

15.6.1 Theorem *For a finite-state Markov chain with transition matrix P, if the chain is irreducible and aperiodic, then the invariant distribution x is unique, and for any initial distribution y, the sequence yP^n converges to x.*

At times it seems that the ratio of definitions to theorems regarding Markov chains is unusually high. Some accessible resources are Breiman [3, Chapter 6], Karlin [9], or Norris [13].

15.7 ⋆ Martingales

15.7.1 Definition *A **martingale** is a stochastic process $\{X_t : t \in T\}$ such that*

$$\boldsymbol{E}\,|X_t| < \infty \quad \forall t \in T,$$

and for every $t_1 < t_2 < \cdots < t_n < t_{n+1}$ we have

$$\boldsymbol{E}\left(X_{t_{n+1}} \,\middle|\, X_{t_n}, \ldots, X_{t_1}\right) = X_{t_n}.$$

That is, the expectation *in the future conditioned on any past values is simply the current* value*.*

Aside: The definition I gave is not as general as the standard definition. In the standard definition, there is also a collection $\{\mathcal{E}_t : t \in T\}$ of σ-algebras of events that such that if $s < t$, then $\mathcal{E}_s \subset \mathcal{E}_t$, that is, every event in $\mathcal{E}_s$ is also an event in $\mathcal{E}_t$. Such a collection is frequently called a **filtration**.[2] The random variables X_t are required to be **adapted** to the filtration, meaning that each event $(X_t \in [a, b])$ belongs to $\mathcal{E}_t$. Further, conditional expectations are defined with respect to the σ-algebra, a topic we shall not get into.[3] My definition restricts $\mathcal{E}_t$ to be $\sigma(\{X_s : s \leqslant t\})$.

For example, if Y_i, $i = 0, 1, 2, \ldots$ are independent mean-zero random variables, then

$$X_t = \sum_{n=0}^{t} Y_n$$

[2] The term filtration comes from a movement by 20th century French mathematicians to name mathematical objects with ordinary French words. They used the term *filtre* (meaning filter, but think of a funnel-like coffee filter) to indicate a family indexed by a particular kind of partial order.

[3] If X is a random variable and $\mathcal{E}'$ is a σ-subalgebra of $\mathcal{E}$, then $\boldsymbol{E}(X \,|\, \mathcal{E}')$ is a random variable adapted to $\mathcal{E}'$ such that for every event E in $\mathcal{E}'$, $\boldsymbol{E}\left[\boldsymbol{E}(X \,|\, \mathcal{E}')\mathbf{1}_E\right] = \boldsymbol{E}[X\mathbf{1}_E]$.

defines a martingale. Thus the Random Walk is a martingale, and so is the wealth of a gambler during a sequence of fair bets.

Another example of a martingale pops up in learning models. Let Z and Y_i, $i = 0, 1, 2, \dots$ be random variables with finite means (not necessarily independent or mean-zero). Then

$$X_n = \boldsymbol{E}(Z \,\big|\, Y_n, \dots, Y_0).$$

defines a martingale.

A **submartingale** (or **semi-martingale**) is a stochastic process $\{X_t : t \in T\}$ such that

$$\boldsymbol{E}\,|X_t| < \infty \quad \forall t \in T,$$

and for every $t_1 < t_2 < \cdots < t_n < t_{n+1}$ we have

$$\boldsymbol{E}(X_{t_{n+1}} \,\big|\, X_{t_n}, \dots, X_{t_1}) \geqslant X_{t_n}.$$

That is, the expected value in the future is greater than the current value. A **supermartingale** reverses the inequality. (You might think that since the study of these processes resulted from studying the wealth of a gambler, that the definitions ought to be reversed, but the early probabilists worked for the casino.)

The Markov property says that the entire distribution of X_{t+s} depends on the past only through X_t.

In a martingale, only the expectation of X_{t+s} depends on the past only through X_t, but in a very special way.

15.8 ⋆ Martingale Convergence Theorem

One of the most important result on martingales is this.

15.8.1 Martingale Convergence Theorem (Cf. Doob [6, Theorem 4.1, p. 319].) *Let $\{X_n : n = 0, 1, 2, \dots, \}$ be a martingale. If $\lim_{n\to\infty} \boldsymbol{E}\,|X_n| = M < \infty$, then there is a random variable X_∞ with $\boldsymbol{E}\,|X_\infty| \leqslant M$ such that*

$$X_n \xrightarrow[n\to\infty]{a.s.} X_\infty.$$

Moreover, if $X_n \geqslant 0$ for all n, or if $X_n \leqslant 0$ for all n, then $M < \infty$ is satisfied, so the conclusion above follows.

Finally, if for some $q > 1$, we have $\lim_{n\to\infty} \boldsymbol{E}\,|X_n|^q < \infty$, then we may append ∞ to T, so that $\{X_n : n = 1, 2, \dots, \infty\}$ is a martingale, $\boldsymbol{E}\,|X_\infty|^q < \infty$, and $X_n \xrightarrow[n\to\infty]{q} X_\infty$.

15.8.1 On the terminology

If you look up the word martingale in a dictionary, you will find that it may come from the Portuguese *martengau*, meaning in inhabitant of Martigues (in Provence), or perhaps it comes from French via Spanish from the Arabic *al mirta'ah*, meaning rein or check.

The first definition should refer to some kind of harness. Later definitions refer to a betting system. For example, my office dictionary [12] gives definition 1 as "A strap fastened to a horse's girth, passing between his forelegs, and fastened to the bit, or now more commonly ending in two rings, through which the reins pass. It is intended to hold down the head of the horse, and prevent him from rearing." It gives definition 3 as "Any system of betting which, in a series of bets, determines the amount to be wagered after each win or loss. The term is usually applied to

dividing a specified amount desired to be won at one session into smaller unequal parts, arranged in a vertical column. By adding together the top and bottom figures after a loss, canceling them after a win, when all are crossed off, the desired amount is gained."

J. M. Hammersley [8], in a paper on a multidimensional generalization of martingales, offers this explanation of why the equestrian terminology may have been used to describe a stochastic process:

> The idea behind the terminology is the following. In gaming, a martingale is a fair gambling system, and this is probably the immediate source of the stochastic sense of "martingale.'' But in turn, the gaming term seems to have its origin in the equestrian sense of the word "martingale.'' In that sense, a martingale is a strap that prevents a horse from throwing up his head. If the horse is proceeding in the positive sense of the parameter i, and his mouth and breast are at heights y_i and z_i, respectively, above the ground at time i, then he will be moving in a steady horizontal fashion when his breast is now at the same height as his mouth was at the previous moment; thus, $y_i = z_i = y_{i-1}$ in conformity with (2.1). Since the strap checks upward but not downward movements of the head, it comes closer to what a mathematician would call a submartingale. If there are constraints from several different directions, as in (2.4), we may imagine them caused by several different straps, or by a harness.
>
> The history of martingales goes back a long way, and there are elaborate reliefs in the British Museum depicting martingales in the reigns of Tiglath-Pileser III (745-727 B.C.), Sennacherib (705-681 B.C.), and Assurbanipal (668-626 B.C.). Anderson [1] writes of the Assyrians: "They manage their horses with bit and bridle, and later reliefs show a remarkable anticipation of the modern martingale (not used as far as I know by any other ancient people). The reins are attached to a large tassel hanging below the horse's neck, which continues to provide a certain check on the horse's mouth. The rider is thus enabled to use both hands for his weapons, and can shoot the bow at full gallop." Müseler [8] and Hitchcock [4] give information about the various types of modern martingale (the standing, running, and Irish martingales), and the latter author has a colorful passage in which he says: "The standing martingale, which is used as a check to prevent the horse from throwing up his head and hitting the rider in the face, or carrying it too high, is a good remedy for stargazing, or for horses which have ewe-necks. ... This type of martingale is used universally on the polo ground."

[Note: The reference numbers in the above passage refer to Hammersley's bibliography, not this one.]

15.9 ⋆ Bayesian updated beliefs as a martingale

Urn 1 has 2 Black balls and 1 White ball. Urn 2 has 1 Black ball and 4 White balls. I believe that Urn 1 has been chosen with probability p_0 where $0 < p_0 < 1$.

What do I expect now (before any ball has been chosen) my belief p_1 to be after one sample is drawn from the Urn? There are two possible outcomes of the draw: B and W, and

$$P(B) = P(B \mid 1)P(1) + P(B \mid 2)P(2) = \frac{2}{3}p_0 + \frac{1}{5}(1 - p_0) = \frac{1}{5} + \frac{7}{15}p_0$$

and

$$P(W) = P(W \mid 1)P(1) + P(W \mid 2)P(2) = \frac{1}{3}p_0 + \frac{4}{5}(1 - p_0) = \frac{4}{5} - \frac{7}{15}p_0.$$

If B is the outcome, my posterior belief, applying Bayes' Law, is

$$P(1 \mid B) = P(B \mid 1)\frac{P(1)}{P(B)} = \frac{2}{3}\frac{p_0}{\frac{1}{5} + \frac{7}{15}p_0} = \frac{10p_0}{3 + 7p_0}$$

and if W is the outcome, my posterior belief, applying Bayes' Law, is

$$P(1 \mid W) = P(W \mid 1)\frac{P(1)}{P(W)} = \frac{1}{3}\frac{p_0}{\frac{4}{5} - \frac{7}{15}p_0} = \frac{5p_0}{12 - 7p_0}.$$

So

$$\begin{aligned}\boldsymbol{E}\left(p_1 \,\middle|\, p_0\right) &= P(1 \,|\, B)P(B) + P(1 \,|\, W)P(W)\\ &= \frac{10p_0}{3+7p_0}\left(\frac{1}{5}+\frac{7}{15}p_0\right) + \frac{5p_0}{12-7p_0}\left(\frac{4}{5}-\frac{7}{15}p_0\right)\\ &= \frac{2}{3}p_0 + \frac{1}{3}p_0 = p_0.\end{aligned}$$

The same argument applies into the future. That is, the probability p_t, the probability I attach to urn 1 after t independent samples (with replacement) is a martingale: $\boldsymbol{E}\left(p_{t+s} \,\middle|\, p_t\right) = p_t$.

Since probabilities are bounded, the Martingale Convergence Theorem implies that my beliefs will converge with probability one as $t \to \infty$.

15.10 ⋆ Stopping times

Given a discrete-time stochastic process $X_1, \ldots, X_n \ldots$, a **stopping time** is an integer-valued random variable N such that

$$P\left(N < \infty\right) = 1,$$

and

$$\text{the event } \ (N = k) \ \text{ belongs to } \sigma(X_1, X_2, \ldots, X_k),$$

This mean that the indicator function $\mathbf{1}_{(N=k)}$ can be written as some function h of $X_1, \ldots, X_n$. In other words, you can't "peek ahead" to decide whether to stop.

15.11 ⋆ Stopped martingales

If $Z_1, Z_2, \ldots$ is a martingale and N is a stopping time for this martingale, then the **stopped martingale** is

$$\bar{Z}_n = Z_{\min(N,n)}.$$

15.11.1 Theorem *The stopped martingale is a martingale.*

The proof is taken from Sheldon Ross and Erol Peköz [14, Lemma 3.13, p. 88].

Proof: Start with the identity

$$\bar{Z}_n = \bar{Z}_{n-1} + \mathbf{1}_{N \geqslant n}(Z_n - Z_{n-1}).$$

(Verify this.)

Since conditional expectation is a positive linear operator,

$$\begin{aligned}\boldsymbol{E}\left(\bar{Z}_n \,\middle|\, X_1, \ldots, X_{n-1}\right) &= \boldsymbol{E}(\bar{Z}_{n-1} \,|\, X_1, \ldots, X_{n-1}) + \boldsymbol{E}(\mathbf{1}_{N \geqslant n}(Z_n - Z_{n-1} \,|\, X_1, \ldots, X_{n-1})\\ &= \bar{Z}_{n-1} + \mathbf{1}_{N \geqslant n}\, \boldsymbol{E}((Z_n - Z_{n-1} \,|\, X_1, \ldots, X_{n-1})\\ &= \bar{Z}_{n-1}.\end{aligned}$$

∎

A proof of the following theorem may be found in Ross and Peköz [14, Theorem 3.14, pp. 88–89].

15.11.2 Theorem (Martingale Stopping Theorem)

$$\boldsymbol{E}\, Z_N = \boldsymbol{E}\, Z_1$$

if any one of the following sufficient conditions hold:

1. $\bar{Z}_n$ *are uniformly bounded.*

2. N *is bounded.*

3. $\boldsymbol{E}\, N < \infty$ *and there is some* $M < \infty$ *such that for all* n,
$$\boldsymbol{E}\left(Z_{n+1} - Z_n \,\middle|\, Z_n\right) < M.$$

Some of the consequences of this theorem are:

- There are no gambling "systems."
- No parental stopping rule can account for Sen's missing women.

Bibliography

[1] D. Bayer and P. Diaconis. 1992. Trailing the dovetail shuffle to its lair. *Annals of Applied Probability* 2(2):294–313. http://www.jstor.org/stable/2959752

[2] K. C. Border. 1985. *Fixed point theorems with applications to economics and game theory*. New York: Cambridge University Press.

[3] L. Breiman. 1986. *Probability and stochastic processes: With a view toward applications*, 2d. ed. Palo Alto, California: Scientific Press.

[4] T. M. Cover and J. A. Thomas. 2006. *Elements of information theory*, 2d. ed. Hoboken, New Jersey: Wiley–Interscience.

[5] G. Debreu and I. N. Herstein. 1953. Nonnegative square matrices. *Econometrica* 21(4):597–607. http://www.jstor.org/stable/1907925

[6] J. L. Doob. 1953. *Stochastic processes*. New York: Wiley.

[7] J. Franklin. 1980. *Methods of mathematical economics*. Undergraduate Texts in Mathematics. New York: Springer–Verlag.

[8] J. M. Hammersley. 1967. Harnesses. In L. M. Le Cam and J. Neyman, eds., *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 3: Physical Sciences*, pages 89–117. Berkeley and Los Angeles: University of California Press. http://projecteuclid.org/euclid.bsmsp/1200513623

[9] S. Karlin. 1969. *A first course in stochastic processes*. New York & London: Academic Press.

[10] J. G. Kemeny and J. L. Snell. 1960. *Finite Markov chains*. The University Series in Undergraduate Mathmatics. Princeton, New Jersey: D. Van Nostrand.

[11] J. W. Milnor. 1969. *Topology from the differentiable viewpoint*, corrected second printing. ed. Charlottesville: University Press of Virginia. Based on Notes by David W. Weaver.

[12] W. A. Neilson, T. A. Knott, and P. W. Carhart, eds. 1944. *Webster's new international dictionary of the English language*, second unabridged ed. Springfield, Massachusetts: G. & C. Merriam Company.

[13] J. R. Norris. 1998. *Markov chains*. Number 2 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

[14] S. M. Ross and E. A. Peköz. 2007. *A second course in probability*. Boston: Probability-Bookstore.com.

[15] H. Wielandt. 1950. Unzerlegbare, nicht negative Matrizen. *Mathematische Zeitschrift* 52:642–648. DOI: 10.1007/BF02230720