Math-UA.233: Theory of Probability
Lecture 25

Tim Austin

Markov chains (Ross Sec 9.2, roughly). Not for assessment!

Some of the most important and powerful probabilistic models describe *random evolution as time passes*.

Imagine a system that can exist in many possible states, and moves between states as time passes. The movement is not exactly predictable, and we wish to attach probability values to different possible behaviours.

The underlying 'experiment' is 'watching the states that the system passes through for a long interval of time' — or possibly 'forever'.

Some simple examples:

- Each day, we observe whether or not it rained (just two states, very simple model).
- Before each play of a game, we observe a gambler's remaining fortune.
- Each minute, a store observes how many customers are waiting in line at the counter.

To set up a mathematical model, we first choose a **state space**: the set of all possible states that our system may be in.

(WARNING: not the same as the *sample space*, which will consist of whole *sequences* of states, because we're going to observe the system *repeatedly*.)

For simplicity, let's suppose there are finitely many states, and label then $1, \ldots, M$.

Now let $X_0, X_1, X_2, \ldots$ be RVs defined by:

$$X_n = i \quad \text{if the system is in the } i^{\text{th}} \text{ state at time } n.$$

(It's conventional to start the clock at time 0, not 1.)

This sequence of RVs indicates how the state evolves with time.

(Sometimes it's more realistic to let time be a continuous parameter. There's a theory for that too, but it's trickier.)

A sequence such as $X_0, X_1, \ldots$ is called a **stochastic process.**

There are many associated probabilities: for example,

$$P(X_n = i) = \text{probability that the system is in state } i \text{ at time } n,$$

$$P(X_n = i, X_{n+1} = j) = \begin{array}{l}\text{probability that the system is in state } i \\ \text{at time } n\text{, and then moves} \\ \text{to state } j \text{ at the next time,}\end{array}$$

and so on.

There are many different models for such an evolution, some very complicated.

Simplest model: suppose that $X_1, X_2, \ldots$ are *independent*. Easy to compute with, but often not realistic: in this model the past has *no influence of any kind on the future*!

Markov chains are the *next* simplest possibility, but are often much more realistic.

The basic assumption:

> *The past influences the future only through the present.*

Put another way: at each time n, the *present* state may influence what happens in the future, but it doesn't matter *how the system got to the present state*.

Mathematically, this is a statement about conditional probabilities. Precise version:

Definition

The stochastic process $X_0, X_1, \ldots$ is a ***Markov chain*** *if*

$$\underbrace{P(X_{n+1} = j \mid X_n = i_n, \ \ldots, \ X_0 = i_0)}_{\text{conditional probability given entire past}} = \underbrace{P(X_{n+1} = j \mid X_n = i_n)}_{\text{cond. prob. given present}}$$

for all times $n \geqslant 1$ and all possible states $i_0, i_1, \ldots, i_n$ and j.

This equation is often called the **Markov property**.

How do we set up a Markov chain model?

Here are the basic data we need:

1. The **initial distribution**

$$p_{X_0}(i) = P(X_0 = i) \quad \text{for } i = 1, 2, \ldots, M.$$

2. The **transition probabilities**

$$p_{X_{n+1}|X_n}(j|i) = P(X_{n+1} = j \,|\, X_n = i)$$

for each time n and pair of states $i, j = 1, 2, \ldots, M$.

Using these data, we find the probabilities of more complicated events using the *multiplication rule*:

$$P(X_0 = i_0,\ X_1 = i_1,\ \ldots,\ X_n = i_n)$$

$$\begin{aligned} &= P(X_n = i_n \,|\, X_{n-1} = i_{n-1},\ \ldots,\ x_0 = i_0) \\ &\quad \times P(X_{n-1} = i_{n-1} \,|\, X_{n-2} = i_{n-2},\ \ldots,\ x_0 = i_0) \\ &\quad \times \cdots \times P(X_1 = i_1 \,|\, X_0 = i_0) \times P(X_0 = i_0) \end{aligned}$$

$$\begin{aligned} &= \underbrace{P(X_n = i_n \,|\, X_{n-1} = i_{n-1})}_{\text{transition prob.}} \times \underbrace{P(X_{n-1} = i_{n-1} \,|\, X_{n-2} = i_{n-2})}_{\text{transition prob.}} \\ &\quad \times \cdots \times \underbrace{P(X_1 = i_1 \,|\, X_0 = i_0)}_{\text{transition prob.}} \times \underbrace{P(X_0 = i_0)}_{\text{initial dist.}} \end{aligned}$$

Another common assumption: the chain is **time-homogeneous** if the probabilities

$$P(X_{n+1} = j \mid X_n = i)$$

depend on i and j but *not* on n. Then we write $p_{i,j}$ for this number (no need to write n).

This case is simpler, and we focus on it here.

Definition

*The **transition matrix** of our time-homogeneous Markov chain is the $M \times M$ matrix*

$$\mathbf{P} = [p_{i,j}] = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,M} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,M} \\ \vdots & \vdots & & \vdots \\ p_{M,1} & p_{M,2} & \cdots & p_{M,M} \end{bmatrix}.$$

The transition matrix contains all the information needed to describe the probabilities of how the system evolves, if we know its state initially.

Why a *matrix*, not just an *array*? Suggests linear algebra...

BASIC FACTS:

The transition matrix **P** is a matrix of conditional probabilities, and so

$$\boxed{p_{i,j} \geqslant 0 \quad \text{for all states } i, j.}$$

Also, if we fix a state i, then

$$\underbrace{\sum_{j=1}^{M} P(X_1 = j \mid X_0 = i) = P(X_1 = \text{anything} \mid X_0 = i)}_{\text{axiom 3 applied to cond. probs.}} = 1.$$

That is, all row-sums of the transition matrix are equal to 1:

$$\boxed{\sum_{j=1}^{M} p_{i,j} = 1 \quad \text{for all states } i.}$$

Example (Ross E.g. 9.2a)

Each day it either rains ('state 0*') or doesn't ('state* 1*'). If it rains today, then the probability of rain tomorrow is* α*. If it doesn't rain today, then the probability of rain tomorrow is* β*.*

Transition matrix:

$$\begin{bmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{bmatrix}.$$

(Each row can be completed, because row-sums must equal 1*.)*

In this example, time-homogeneity means that α and β don't change with time. That is, this model *ignores climate change*.

Example (The gambler's ruin chain; Ross E.g. 9.2b)

Suppose a gambler places win/lose bets, and wins each of them independently with probability p. He starts with a fortune between 1 and $M - 1$, and quits if it ever sinks to 0 or reaches M. (So there are $M + 1$ states.)

Let X_n be his fortune at time n. Given the event $\{X_n = i\}$ for some $1 \leqslant i \leqslant M - 1$, we know that X_{n+1} equals either $i + 1$ with probability p, or $i - 1$ with probability $1 - p$.

If $X_n = 0$ or $X_n = M$ at some time n, then the process stays in that state forever after.

In this e.g., the outcomes of successive games are independent. But the gambler's successive fortunes are *not*, because X_n is an accumulation over all the games up to time n.

(Gambler, cont.)

Transition matrix:

$$
\begin{bmatrix}
1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
1-p & 0 & p & 0 & \cdots & 0 & 0 & 0 \\
0 & 1-p & 0 & p & \cdots & 0 & 0 & 0 \\
0 & 0 & 1-p & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & 0 & 1-p & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \\
0 & 0 & 0 & 0 & \cdots & 1-p & 0 & p \\
0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1
\end{bmatrix}
$$

In terms of the transition matrix $\mathbf{P}$, we obtain probabilities of longer sequences of states by multiplying:

$$P(X_0 = i_0,\ X_1 = i_1,\ \ldots,\ X_n = i_n) \\ = P(X_0 = i_0) p_{i_0,i_1} p_{i_1,i_2} \cdots p_{i_{n-1},i_n}$$

(some formula as before in new notation).

But what if we only care about the initial state i_0 and the state i_n at some later time, not what happens in between?

For example, consider $n = 2$. By axiom 3 and previous formula:

$$\begin{aligned}
P(X_2 = j \mid X_0 = i) &= \frac{1}{P(X_0 = i)} P(X_2 = j,\ X_0 = i) \\
&= \frac{1}{P(X_0 = i)} \sum_{k=1}^{M} P(X_2 = j,\ X_1 = k,\ X_0 = i) \\
&= \frac{1}{P(X_0 = i)} \sum_{k=1}^{M} P(X_0 = i) p_{i,k} p_{k,j} \\
&= \underbrace{\sum_{k=1}^{M} p_{i,k} p_{k,j}}_{\text{matrix multiplication!}}
\end{aligned}$$

These are the **two-step transition probabilities.** They are given by the entries of the matrix $\mathbf{P} \times \mathbf{P} = \mathbf{P}^2$.

In general, the *n***-step transition probabilities** are the numbers defined by

$$\underbrace{p_{i,j}^{(n)}}_{\text{notation}} = P(X_n = j \mid X_0 = i) \quad \text{for all possible states } i, j.$$

They are given by the matrix product:

$$[p_{i,j}^{(n)}] = \underbrace{\mathbf{P} \times \mathbf{P} \times \cdots \times \mathbf{P}}_{\text{matrix multiplication } n \text{ times}} = \mathbf{P}^n$$

(proof: just generalization of calculation above).

By time-homogeneity, we can *start* this calculation at any time later than 0, and get the same answer: that is,

$$P(X_{m+n} = j \mid X_m = i) = p_{i,j}^{(n)}$$

for any n and m and i and j (so this number doesn't depend on the start time m).

Most general form of this calculation:

Theorem (Chapman-Kolmogorov equations, Ross Prop 9.2.1)

Let $0 < r < n$ *and let* i, j *be any two states. Then the transition probabilities satisfy*

$$\boxed{p_{i,j}^{(n)} = \sum_{k=1}^{M} p_{i,k}^{(r)} p_{k,j}^{(n-r)}.}$$

PROOF: just writing the matrix product in two different ways:

$$\mathbf{P}^n = \mathbf{P}^r \times \mathbf{P}^{n-r}.$$

The n-step transition probabilities $p_{i,j}^{(n)} = P(X_n = j \mid X_0 = i)$ are *conditional* probabilities. But now suppose we just want the distribution (PMF) of X_n by itself. We get this from LOTP:

$$P(X_n = j) = \sum_{i=1}^{M} P(X_0 = i) P(X_n = j \mid X_0 = i).$$

This also depends on the initial distribution $P(X_0 = i)$ — that is, on the PMF of X_0.

If we know the initial distribution, then this formula tells us the distribution of X_n at all subsequent times.

Again:

$$P(X_n = j) = \sum_{i=1}^{M} P(X_0 = i) P(X_n = j \mid X_0 = i).$$

This is also a linear algebra formula. Let's write probability distributions on $\{1, \ldots, M\}$ as row vectors: for example, let

$$\mathbf{t} = (t_1, \ldots, t_M) = (P(X_0 = 1), \ldots, P(X_0 = M))$$

be the initial distribution. Then the above formula gives

$$P(X_n = j) = \sum_{i=1}^{M} t_i p_{i,j}^{(n)} = (\mathbf{t}\mathbf{P}^n)_j.$$

So the distribution at time n is the vector-matrix product $\mathbf{tP}^n$.

Stationary distributions and convergence

Some of the most basic questions about Markov chains concern their *long-run average* behaviour.

To explore this, we must look again at the distributions of the RVs X_n. In general, they are not identically distributed. The distribution $\mathbf{t}\mathbf{P}^n$ changes with time. But for some special choices of $\mathbf{t}$ the distribution *doesn't* change.

Definition

Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ be a probability distribution on $\{1, 2, \ldots, M\}$. It is ***stationary*** *(or an '****equilibrium distribution****' or a '****steady state****') if*

$$\boxed{\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}.}$$

In this case, $\boldsymbol{\pi}$ also satisfies $\boldsymbol{\pi}\mathbf{P}^n = \boldsymbol{\pi}$ for every $n \in \mathbb{N}$, by induction.

Probabilistic meaning: if the initial distribution is $\boldsymbol{\pi}$, then the distribution of every subsequent X_n is also still $\boldsymbol{\pi}$.

This DOES NOT mean that the X_ns are equal! The state is still changing all the time. It's only the *probability values* that are 'stationary'.

Stationary distributions are important because they answer many basic questions about ‘long-run’ behaviour.

The full analysis involves several different cases depending on the matrix $\mathbf{P}$ (specifically, on where there are zeroes in that matrix).

Here we’ll restrict to a setting which avoids those issues.

Definition

*The Markov chain is **ergodic** if there is an $n \geqslant 1$ such that*

$$\boxed{p_{i,j}^{(n)} > 0 \quad \textit{for all states } i, j.}$$

MEANING: By time n, there's some positive chance of ending up anywhere, no matter where you started from.

(For a general Markov chain, one can decompose the state space into 'ergodic pieces' and analyze them separately, but the details are tricky.)

Theorem (Ross Theorem 9.2.1)

If **P** *is the transition matrix for an ergodic Markov chain, then*

1. *it has a unique stationary distribution* π*, and*
2. *we have*

$$p_{i,j}^{(n)} \longrightarrow \pi_j \quad \text{as } n \longrightarrow \infty \text{ for all states } i, j.$$

Importantly, π_j doesn't depend on i. So after waiting a long time, the probability of being in state j is always roughly π_j, no matter which state i we started from.

SLOGAN: An ergodic Markov chain eventually forgets where it started.

From this theorem, one can derive various aspects of the long-run behaviour. For example:

Theorem (Markov-chain LLN; Ross p401)

Let $X_0, X_1, \ldots,$ be an ergodic Markov chain with initial distribution $\mathbf{t}$ *and transition matrix* $\mathbf{P}$*. Let* π *be the stationary distribution. For any state j, let*

$$N(j,n) = \textit{number of visits to state } j \textit{ before time } n.$$

Then for any $\varepsilon > 0$ we have

$$P\left(\left|\frac{N(j,n)}{n} - \pi_j\right| > \varepsilon\right) \longrightarrow 0 \quad \textit{as } n \longrightarrow \infty.$$

INTERPRETATION: The long-run fraction of time spent in state j is (approximately) π_j (with high probability).

Another example:

Theorem (Return-times formula)

Same setting as previous theorem.

The expected time that it takes the chain to return to state j, given that it starts in state j, is $1/\pi_j$. In notation:

$$E[\textit{least } n \geqslant 1 \textit{ such that } X_n = j \mid X_0 = i] = 1/\pi_j.$$

OBSERVE: This generalizes the fact that the mean of the geometric(p) distribution is $1/p$. How?

Example (Rain example again, Ross E.g. 9.2a)

Transition matrix for raining/not raining Markov chain:

$$\mathbf{P} = \begin{bmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{bmatrix}.$$

Let's find the stationary distribution $\boldsymbol{\pi} = (\pi_0, \pi_1)$:

$$\pi_0 = \alpha\pi_0 + \beta\pi_1 \quad \text{and} \quad \pi_0 + \pi_1 = 1$$

$$\implies \quad \pi_0 = \frac{\alpha}{1+\beta-\alpha}, \quad \pi_1 = \frac{1-\alpha}{1+\beta-\alpha}.$$

Interpretation: these fractions are the long-run proportion of the time spent raining/not raining.