Caltech Department of Mathematics

Ma 3/103 KC Border
Introduction to Probability and Statistics Winter 2020

Lecture 26: Introduction to Markov Chains

This is a very brief introduction to Markov chains. To do justice to the topic takes a full quarter or more. But it is a useful and important subject, so I feel that you should be with familiar the main ideas, even if I will not prove very many of the results.

This material is not covered in the textbooks. These notes are still in development. Most of the material here is covered in Chapter 1 of Norris [16].

26.1 ⋆ Stochastic Processes

A **stochastic process** is an indexed family

$$\{X_t : t \in T\}$$

of random variables (or random vectors) on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ that take values in a set S. The set S is called the **state space** of the process. Note that for stochastic processes, I denote the probability measure by $\mathbb{P}$ instead of just plain P. This is traditional, in part because we are soon going to use P to denote a transition matrix.

The set T is called the **index set** of the stochastic process and is interpreted as **time**, so T is a typically an infinite subset of the real line. Thus the X_t's represent a procession of random variables, hence the term process.[1] The index set T might be the natural numbers or integers, a **discrete-time process**; or an interval of the real line, a **continuous-time process**. Feller [10, Chapter 3] uses the term **epoch** to refer to point in T.

Each random variable X_t, $t \in T$ is a function on the probability space Ω. The value $X_t(\omega)$ depends on both ω and t. Thus another way to view a stochastic process is as a **random function** of T: each $\omega \in \Omega$ defines a function $t \mapsto X_t(\omega)$ from T into S. In fact, it is not uncommon to write $X(t)$ instead of X_t, but I personally found this usage confusing.

The Poisson arrival process is a continuous-time process that counts arrival, resulting in discrete jumps in the state space $S = \{0, 1, 2, \dots\}$, at exponentially distributed intervals.

Other important examples of stochastic processes are the **random walk** and its continuous-time counterpart, **Brownian motion**.

26.2 ⋆ Markov chains and transition matrices

A Markov chain is a discrete-time process $X_0, X_1, \dots$ with index set $T = \{0, 1, 2, \dots\}$, with a countable state space S. For many purposes we simply label the states $1, 2, 3, \dots$, and the value of X_t is interpreted as the label of the state. On the other hand, sometimes the state may represent a count, a coordinate, or an amount of money, and thus have intrinsic numerical interest.

A Markov chain works like this: The **initial state** X_0 is chosen at random according the initial distribution λ on S, which assign probability λ_i to state i. That is,

$$\mathbb{P}(X_0 = i) = \lambda_i.$$

[1] There are other indexed families of random variables, where the index is not interpreted as time. For instance, a **random field**, has an index set T interpreted as a region in $\boldsymbol{R}^m$. Such models, such as the *Ising model* are used to model magnetism in metals.

Now for each state $i \in S$, there is another distribution on S, denoted p_i, which assigns probability p_{ij} to state $j \in S$. Consequently, for each $i \in S$,

$$\sum_{j \in S} p_{ij} = 1.$$

We can think of the square array $P = \big[p_{ij}\big]$, where $i, j \in S$, as a matrix, called the **transition matrix**. Note that we allow matrices to have countably many rows and columns. A square matrix P is called a **stochastic matrix** if each $p_{ij} \geqslant 0$, and for each row i,

$$\sum_{j \in S} p_{ij} = 1.$$

We think of a stochastic matrix indexed by S as a mapping from S into the set of probability measures on S: Each state i gets mapped to the probability measure p_i on S that assigns probability p_{ij} to state j.

Now think of the chain as a dance on the state space S. At each step t the dancer finds theirself in some state i, and makes an independent draw of the state at step $t+1$ according to the probability measure p_i. This process repeats itself endlessly.

Aside: Some authors, e.g., Samuel Karlin [13, p. 27] or Joseph Doob [9, p. 170] do not require a transition probabilities to be time-invariant, but then almost immediately restrict attention to the time-invariant case. Other authors, e.g., Kemeny and Snell [14, Definition 2.1.3, p. 25], J. R. Norris [16, p. 2], or Ching and Ng [6, p. 2] make time-invariance part of the definition of a Markov chain. It seems less verbose to add time-invariance to the definition.

Ching and Ng also transpose the transition probabilities, that is, our p_{ij} is their p_{ji}.

What makes a Markov chain a tractable model is what the transition probabilities do not depend upon. The probability distribution of the state at $t+1$ depends only on the state at time t. In this sense a Markov chain has a **one-period memory**.

26.3 ⋆ Examples of Markov chains

Here are some examples of Markov chains:

• A **random walk** is a Markov chain. Let $X_1, \ldots, X_t, \ldots$ be a sequence of independent random variables where,

$$X_t = \begin{cases} 1 & \text{with probability } p \\ -1 & \text{with probability } 1-p \end{cases}$$

For each t, define the running sums

$$S_t = X_1 + \cdots + X_t.$$

We define $S_0 = 0$. The sequence $\{S_t\}$ is the **random walk**. The **simple random walk** is the case where $p = 1/2$. It is discussed in greater detail in Supplement 6.

• The game of Twister can be viewed as a Markov chain, where the sate consists of a specification of whose hands and feet are where.

• 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 an example where the numerical magnitude of the state 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 [2] and its references. Among other things they argue that it takes at least 7 riffle shuffles to get an acceptable degree of randomness.

- 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 a random number of periods. 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 [7, 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

$$P(X \mid Y, Z) = p(X \mid Y).$$

- A Markov chain can be rather degenerate. For example, if $X_t = t$ with probability one, then $X_0, X_1, \dots,$ is a Markov chain.

- Markov chains can exhibit a more complicated dependence on history at the expense of a larger state space. For example, consider the Fibonacci sequence as a degenerate Markov chain with state space $S = \mathbb{N} = 0, 1, \dots$. The chain $F_0, F_1, \dots$ is

$$F_0 = 0,\ F_1 = 1, \text{ and for } t > 1,\ F_t = F_{t-1} + F_{t-2}.$$

In this description F_{t+1} depends on more than F_t—it also depend on F_{t-1}. But if we enlarge the state space to $S = \mathbb{N} \times \mathbb{N}$, then we describe the Fibonacci sequence as

$$X_0 = (F_0, 1) = (0, 1),\ X_1 = (F_1, F_0) = (1, 0),\ X_2 = (F_2, F_1) = (1, 1),$$

$$X_3 = (F_3, F_2) = (2, 1),\ \dots, X_t = (F_t, F_{t-1}) = \big((X_{t-1})_1 + (X_{t-1})_2, (X_{t-1})_1\big),\ \dots$$

That is, the first component of the state at time t is sum of the components at time $t-1$, while the second component of the state at time t is just the first component at time $t-1$. The first component at time t is the t^{th} Fibonacci number.

In fact, if the stochastic process has the feature that the distribution of X_t depends only on the k previous states, by enlarging the state space to S^k, we can represent the stochastic process as a Markov process. The cost is that the state space now has dimension k. This can lead to the **curse of dimensionality**.

We could even go so far as to let the state space be $\bigcup_{n=1}^{\infty} S^n$, to make any discrete-time chain a Markov chain, but that defeats the point. A Markov process is supposed to have a simple structure.

26.4⋆ The distribution of a Markov Chain

The above description of a Markov chain makes no reference to the probability space on which the variables are defined. Perhaps the simplest choice for the underlying probability space is as the set of possible realizations:

$$\Omega = S^{\infty} = S \times S \times \cdots$$

Now define the random variables as follows:

$$\text{For } \omega = (i_0, i_1, i_2, \dots),\ \text{define } X_t(\omega) = i_t.$$

We want the set of events to include every event of the form $(X_0 = i_0, X_1 = i_1, \dots, X_t = i_t)$, so we take $\mathcal{E}$ to be the smallest σ-algebra that includes all of these events.

Now we have to define a probability for each of these events. We start with an **initial distribution** λ of the initial state X_0. That is,

$$\lambda_i = \mathbb{P}(X_0 = i).$$

Now we take a transition matrix P and define the resulting measure $\mathbb{P}$ on Ω as follows.

26.4.1 Definition *Let λ be a distribution on S, where λ_i denotes the probability of state i, and let P be a stochastic matrix indexed by S. The **Markov$(\boldsymbol{\lambda}, \boldsymbol{P})$** chain is defined by the probability measure $\mathbb{P}$ on $\Omega = S^\infty$ defined by*

$$\mathbb{P}(X_0 = i,\, X_1 = i_1,\, \ldots, X_t = i_t) = \lambda_{i_0} \cdot p_{i_0 i_1} \cdot p_{i_1 i_2} \cdots p_{i_{t-1} i_t}. \tag{1}$$

*The distribution λ is the **initial distribution** of the state of the chain.*

The question is, does this completely determine the probability $\mathbb{P}$? The answer is yes. The proof is beyond the scope of this course and relies on the Kolmogorov Extension Theorem, but you may read my on-line notes or consult [1, § 15.6, pp. 519–523].

A consequence of this definition of a Markov chain is that the transition probabilities p_{ij} have the interpretation as conditional probabilities:

$$\mathbb{P}\left(X_{t+1} = j \;\middle|\; X_t = i\right) = p_{ij}.$$

This requires some elementary, but tedious computations, which I shall omit.

Another consequence of the definition is that a Markov chain has the **Markov property**, namely, for every finite sequence of epochs

$$t_1 < t_2 < \cdots < t_n < t_{n+1},$$

we have

$$\mathbb{P}\left(X_{t_{n+1}} = i_{n+1} \;\middle|\; X_{t_n} = i_n,\ X_{t_{n-1}} = i_{n-1},\ \ldots,\ X_{t_1} = i_1\right) = \mathbb{P}\left(X_{t_{n+1}} = x_{n+1} \;\middle|\; X_{t_n} = x_n\right).$$

That is, the future depends on the past only through the present. This too requires some elementary, but tedious computations, which I shall also omit.

26.4.2 Definition *The degenerate probability measure on S that assigns probability one to state i, called the **point-mass** at i, is denoted δ_i.*

The next result is straightforward from the definitions, and may be found in Norris [16, Theorem1.1.2, pp. 3–4].

26.4.3 Theorem (Restarting a Markov chain) *Let $X_0, X_1, \ldots$ be a Markov chain with transition matrix P. Conditional on $X_t = i$, the continuation defined by*

$$\tilde{X}_s = X_{t+s}$$

is the Markov(δ_i, P) chain, and is independent of $X_1, \ldots, X_m$.

Add proof

26.4.4 Definition *The notation $\mathbb{P}_i$ is used to refer to probabilities in a chain conditional on starting in state i,*

$$\mathbb{P}_i(E) = \mathbb{P}(E \mid X_0 = i),$$

or equivalently, the distribution of the Markov(δ_i, P) chain. It depends only on the transition matrix P.

26.4.5 Proposition (Markov chains have independent increments) *For a Markov Chain with a state space $S \subset \boldsymbol{R}$, for every $0 < t_1 < t_2 < \cdots < t_n$ the random variables*

$$X_{t_1} - X_0,\ X_{t_2} - X_{t_1},\ \ldots,\ X_{t_n} - X_{t_{n-1}}$$

are stochastically independent.

26.5 ⋆ Two-step and n-step transition probabilities

The transition matrix tells everything about the evolution of the 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

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

By definition this is just

$$\mathbb{P}\left(X_{t+2} = j \mid X_t = i\right) = \frac{\mathbb{P}\left(X_{t+2} = j \ \&\ X_t = i\right)}{\mathbb{P}\left(X_t = i\right)}. \tag{2}$$

The intermediate state X_{t+1} must take on one of the values $k \in S$. So the event

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

is the disjoint union

$$\bigcup_{k \in S} (X_t = i \ \&\ X_{t+1} = k \ \&\ X_{t+2} = j).$$

Thus we may write

$$\mathbb{P}\left(X_{t+2} = j \mid X_t = i\right) = \frac{\sum_{k \in S} \mathbb{P}\left(X_t = i \ \&\ X_{t+1} = k \ \&\ X_{t+2} = j\right)}{\mathbb{P}\left(X_t = i\right)}. \tag{2′}$$

By the multiplication rule (Section 4.8), for each k,

$$\begin{aligned}\mathbb{P}\left(X_t = i \ \&\ X_{t+1} = k \ \&\ X_{t+2} = j\right)& \\ = \mathbb{P}\left(X_t = i\right) \mathbb{P}\left(X_{t+1} = k \mid X_t = i\right) \mathbb{P}\left(X_{t+2} = j \mid X_{t+1} = k \ \&\ X_t = i\right).&\end{aligned} \tag{3}$$

By the Markov property

$$\mathbb{P}\left(X_{t+2} = j \mid X_{t+1} = k \ \&\ X_t = i\right) = \mathbb{P}\left(X_{t+2} = j \mid X_{t+1} = k\right). \tag{4}$$

Combining (2′), (3), and (4) gives

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

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.

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

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.

26.6 ⋆ Markov chains and graphs

26.6.1 Definition *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 into **communication classes**.*

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 $p_{ij} > 0$. An arrow from node i to node j is used to indicate that the transition from i to j can occur (with nonzero probability) in one step. A loop at a node i indicates that the transition from i back 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.

Make sure to define this SRW.

For instance, the transition matrix P of Example 26.7.1 below corresponds to the graph in Figure 26.1. Figure 26.2 depicts the graph of the simple random walk.

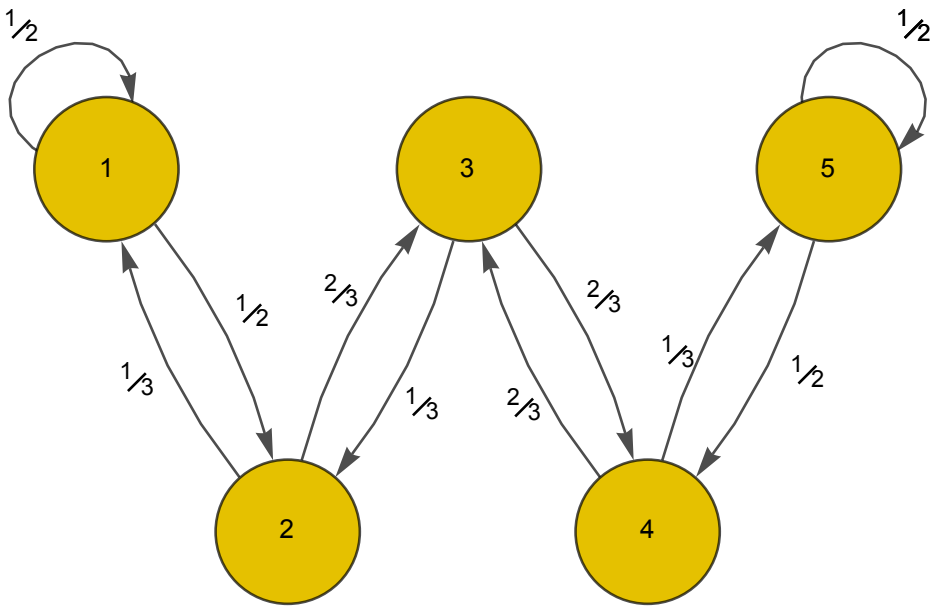

Figure 26.1. The graph of the 5-state Markov chain in Example 26.7.1.

26.6.2 Definition *A communication class C is **closed** if whenever $i \in C$ and j is reachable from i, then $j \in C$. A state i is **absorbing** if $p_{ii} = 1$, that is, if $\{i\}$ is a closed class.*

For instance if you are gambling in a casino that does not extend credit, then a wealth level of 0 is an absorbing state.

26.6.1 ⋆ Mathematica and Markov transition graphs

Since version 11, (and perhaps earlier versions, I don't know) Mathematica will draw the graph of a finite-state Markov chain for you. The first thing you have to do is create a

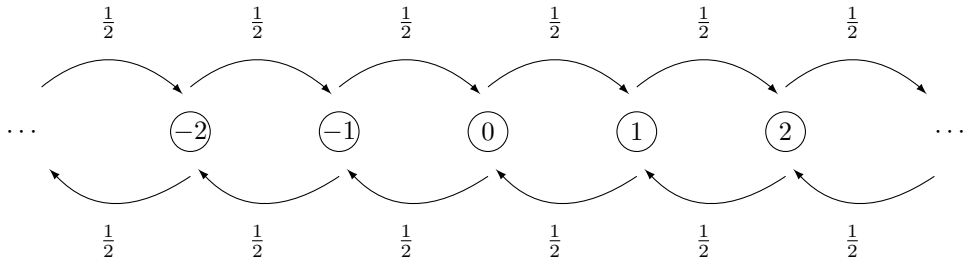

Figure 26.2. The graph of the simple random walk.

`DiscreteMarkovProcess` object. To do this, you need to define a transition matrix `p` and an initial state `s0`, and then

```
mp = DiscreteMarkovProcess[s0, p];
```

creates an object `mp`, and you can use the `Graph` function to create a graph like this:

```
mg = Graph[mp]
```

Then `mg` is a MATHEMATICA `Graphics` object which displays as the transition graph. Each communications class of vertices is given a different color. The graph can be `Exported`ed to a graphics file.

But the plain vanilla graph is probably not what you wanted, since it does not label the edges with their probabilities. I found some code on stackexchange that will do that:

```
mg = Graph[mp,
EdgeLabels -> {DirectedEdge[i_, j_] :>
   MarkovProcessProperties[mp, "TransitionMatrix"][[i, j]]}
]
```

I personally find the vertex labels to need adjustment. You will probably want to fool around with different `GraphLayout` specifications to rearrange the graph to your liking. The graph in Figure 26.1 was produced with

```
Graph[mp,
  EdgeLabels -> {DirectedEdge[i_, j_] :> p[[i, j]]},
  EdgeStyle -> Directive[Black],
  VertexSize -> 0.4,
  VertexCoordinates -> Table[{i, 2 Mod[i, 2]}, {i, 5}],
  BaseStyle -> {FractionBoxOptions -> {Beveled -> True}}
  ];
```

and then the edge labels were manually adjusted.

The `MarkovProcessProperties[mp]` command produces a display of many of the chain's properties.

26.7 ⋆ Irreducible Markov chains

When every state communicates with every other state, the chain is called **irreducible**.[2]

[2] This is the definition of irreducibility in Karlin [13, p. 42]. MATHEMATICA's documentation uses the term to refer to a Markov chain with a single recurrent class (see Section 26.13 ⋆).

26.7.1 Example Here is an example of an irreducible 5-state transition matrix. Its graph is given in Figure 26.1.

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

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-run 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. Every stochastic matrix has an eigenvalue equal to 1 (corresponding to the vector $[1, \dots, 1]$ and all the characteristic roots of a matrix are $\leqslant 1$ in absolute value. 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 [8].

26.8 ⋆ Invariant distributions

Suppose I have a Markov chain and choose the initial state (X_0) according to some probability measure λ on S. Then

$$\lambda P = \left[\textstyle\sum_{i \in S} \lambda_i p_{ij}\right]$$

gives the probability distribution of states at time $t = 1$,

$$\mathbb{P}(X_1 = j) = \sum_{k \in S} \mathbb{P}\left(X_1 = j \;\middle|\; X_0 = i\right) \mathbb{P}(X_0 = i) = \sum_{i \in S} \lambda_i p_{ij} = (\lambda P)_j.$$

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

A probability distribution π on the state space is an **invariant** or **stationary** or **equilibrium** distribution if

$$\pi P = \pi.$$

That is, the unconditional distribution of states at time 1 is the same as the initial distribution π. This also says that π is a left eigenvector of P corresponding to the eigenvalue 1.

26.8.1 Proposition *Every m-state Markov chain has an invariant distribution.*

Proof: The complete proof is beyond the scope of this course, but here's my favorite proof, taken from Debreu and Herstein [8]. 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 nonempty 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 $\bar{x}P = \bar{x}$. ∎

The gap in the above argument is the Brouwer theorem. For a simple proof of the Brouwer Theorem, I'm partial to Border [3], but Franklin [11] and Milnor [15] provide alternate proofs that you may prefer.) There is also a proof based on linear algebra, specifically the Perron–Frobenius Theorem, see, e.g., Debreu and Herstein [8] or Wielandt [18].

Is the invariant distribution unique? Not necessarily.

26.8.2 Example (Non-uniqueness) For the two-state transition matrix

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

every distribution is invariant. The graph of this chain is given in Figure 26.3.

Figure 26.3. A reducible Markov chain.

An interesting aspect of this example is that each state is absorbing. More generally, if a chain has two closed classes, it will not have a unique invariant distribution, since you can find an invariant for each class viewed as a chain in its own right, and then randomize between these. □

N.B. Proposition 26.8.1 used the fact that the are only finitely many states. For infinite state Markov chains, there may be no invariant distribution.

26.8.3 Example (Infinite state chain with no invariant distribution) For instance, if the set of states is $\mathbb{N} = \{0, 1, 2, 3, \dots\}$, the transition probabilities

$$p_{ij} = \begin{cases} 1 & \text{if } j = i + 1 \\ 0 & \text{otherwise} \end{cases}$$

do not admit an invariant distribution—for each n, after $n + 1$ steps the probability of being in state n is zero. The conditions for the existence of an invariant distribution in the general (infinite state space) case are beyond the scope of this course. But if you want a good starting point, try the book by Caltech alumnus Leo Breiman [5]. □

26.8.1 Google's PageRank™ ranking

Ching and Ng [6, pp. 47–59] report that Google's PageRank™ is based on stationary distributions of a transition matrix. The method is named after Larry Page, one of Google's co-founders and was first described in [17]. The state space is the set of web pages. Assume each page has hyperlink to itself. If page i has hyperlinks to n_i pages (including itself), assume the probability of transitioning from page i to a hyperlinked page is $1/n_i$. If this transition matrix is irreducible, it has a stationary distribution π. The rank of page i is π_i, the stationary probability of page i. (My colleague Omer Tamuz tells me this idea had been used before as a measure of centrality in graph theory. Google's contribution was to actually compute it for the web.) Ching and Ng have lots of details on computational techniques, as well as what to do if the transition matrix is not irreducible.

26.9 ⋆ Invariant distributions as limits

Consider the transition matrix

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

Its graph is shown in Figure 26.4. It has the unique invariant distribution $\pi = [1/3, 1/3, 1/3]$.

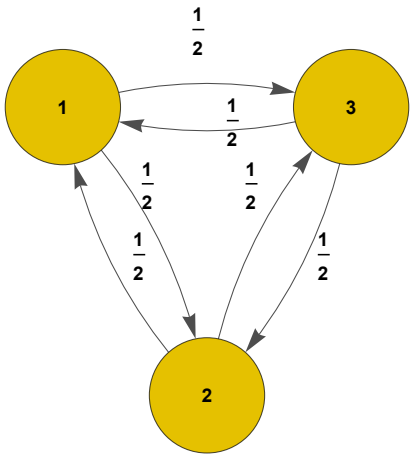

Figure 26.4.

Let δ_1 be the distribution that gives state 1 for sure, $\delta_1 = [1, 0, 0]$. Now consider the sequence $\delta_1 P, \delta_1 P^2, \delta_1 P^3, \ldots$. Some of the terms are reproduced here:

$$\delta_1 P = \begin{bmatrix} 0 \\ 0.5 \\ 0.5 \end{bmatrix}, \quad \delta_1 P^2 = \begin{bmatrix} 0.5 \\ 0.25 \\ 0.25 \end{bmatrix}, \quad \delta_1 P^3 = \begin{bmatrix} 0.25 \\ 0.375 \\ 0.375 \end{bmatrix}, \quad \ldots, \quad \delta_1 P^{20} = \begin{bmatrix} 0.33333 \\ 0.33333 \\ 0.33333 \end{bmatrix}$$

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

26.9.1 Example (P^n does not converge) The transition matrix

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

has period 2. This is easily seen by inspecting its transition graph, see Figure 26.5. This matrix

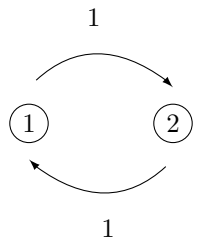

Figure 26.5. A Markov chain of period 2.

has the unique invariant distribution $[1/2, 1/2]$, but P^n does not converge:

$$P^2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

so

$$P^k = \begin{cases} P & \text{for } k \text{ odd,} \\ I & \text{for } k \text{ even.} \end{cases}$$

So setting $\delta_1 = [1, 0]$ gives

$$\delta_1 P^k = \begin{cases} \delta_2 & \text{for } k \text{ odd,} \\ \delta_1 & \text{for } k \text{ even.} \end{cases}$$

These sequences oscillate rather than converge. The problem is that this Markov chain is not **aperiodic**. □

26.9.2 Definition *In a Markov chain the* ***period*** *of state i is the greatest common divisor of $\{n : p_{ii}^{(n)} > 0\}$. If for every n we have $p_{ii}^{(n)} = 0$, we say i has period zero.*

A Markov chain is ***aperiodic*** *if every state has period one.*

In Example 26.9.1, for each state $i = 1, 2$,

$$p_{ii}^{(n)} = \begin{cases} 1 & n \text{ even} \\ 0 & n \text{ odd} \end{cases}$$

so every state has period 2.

You may find the following theorem in Breiman [5, Theorem 6.20, p. 172] or Norris [16, Theorem 1.8.3, p. 41–42].

26.9.3 Theorem (Convergence to the invariant distribution) *For a Markov chain with transition matrix P, if the chain is irreducible and aperiodic, then the invariant distribution π is unique, and for any initial distribution λ, the sequence λP^n converges to π.*

In particular, for any states i and j

$$p_{ij}^{(n)} \to \pi_j \text{ as } n \to \infty.$$

At times it seems that the ratio of definitions to theorems regarding Markov chains is unusually high. Some accessible resources are Breiman [5, Chapter 6], Karlin [13], or Norris [16].

26.10 ⋆ "Infinitely often"

Let $E_1, E_2, \ldots$ be an infinite sequence of events in Ω. Which points belong to infinitely many of the E_n's?

For ω to belong to infinitely many of the E_n's, for each n there must be some $m \geqslant n$ for which $\omega \in E_m$. (Otherwise ω belongs to at most n of the sets.) That is,

$$\omega \in \bigcup_{m=n}^{\infty} E_m.$$

But this must be true for each n, so we must have

$$\omega \in \bigcap_{n=1}^{\infty} \Big(\bigcup_{m=n}^{\infty} E_m \Big).$$

This set is the event where elements in the sequence E_n occur "infinitely often," abbreviated

$$(E_n \text{ i.o.}) = \bigcap_{n=1}^{\infty} \Big(\bigcup_{m=n}^{\infty} E_m \Big).$$

This set is also known as $\limsup_n E_n$

The complementary event, that only finitely many of the events occur is by DeMorgan's laws

$$(E_n \text{ finitely often}) = \bigcup_{n=1}^{\infty} (\bigcap_{m=n}^{\infty} E_m^{\mathrm{c}}).$$

This set is also called the $\liminf_n E_n$.

26.11 ⋆ The Borel–Cantelli Lemma

The Borel–Cantelli Lemma is a useful result concerning the general problem of events occurring infinitely often. This seems like a good place to mention it, but it should probably come earlier. In fact, another concept that should have been introduced earlier is that of mutual independence. A family of events is **mutually independent** if for every finite subcollection $E_1, \ldots, E_n$, we have $P(\cap_{i=1}^n E_i) = P(E_1) \cdots P(E_n)$. This is equivalent to their indicator functions being independent.

26.11.1 Borel–Cantelli Lemma *Let $E_1, E_2, \ldots$ be an infinite sequence of events in the probability space $(\Omega, \mathcal{E}, P)$.*

1. *If $\sum_{n=1}^{\infty} P(E_n) < \infty$, then $P(E_n \text{ i.o.}) = 0$.*
2. *If the events are mutually independent, and if $\sum_{n=1}^{\infty} P(E_n) = \infty$, then $P(E_n \text{ i.o.}) = 1$.*

Proof: (Cf. Jacod and Protter [12, pp. 71–72], or Breiman [4, Lemma 3.14, pp. 41–42].)

(1.) Assume $\sum_{m=1}^{\infty} P(E_m) < \infty$. Then

$$\lim_{n \to \infty} \sum_{m=n}^{\infty} P(E_m) = 0.$$

Now

$$\begin{aligned} P(E_n \text{ i.o.}) &= P\left(\bigcap_{n=1}^{\infty}\Big(\bigcup_{m=n}^{\infty} E_m\Big)\right) \\ &= \lim_{n\to\infty} P\Big(\bigcup_{m=n}^{\infty} E_m\Big) && \text{Proposition 2.4.4} \\ &\leqslant \lim_{n\to\infty} \sum_{m=n}^{\infty} P(E_m) && \text{Boole's Inequality 2.2.2} \\ &= 0. \end{aligned}$$

(2.) Assume the events are mutually independent, and that $\sum_{n=1}^{\infty} P(E_n) = \infty$. Now

$$\begin{aligned} P(E_n \text{ i.o.}) &= P\left(\bigcap_{n=1}^{\infty}\Big(\bigcup_{m=n}^{\infty} E_m\Big)\right) \\ &= \lim_{n\to\infty} P\Big(\bigcup_{m=n}^{\infty} E_m\Big) && \text{Proposition 2.4.4} \\ &= \lim_{n\to\infty} 1 - P\Big(\bigcup_{m=n}^{\infty} E_m\Big)^{\mathrm{c}} && P(E^{\mathrm{c}}) = 1 - P(E) \\ &= 1 - \lim_{n\to\infty} P\Big(\bigcap_{m=n}^{\infty} E_m^{\mathrm{c}}\Big) && \text{de Morgan's Laws.} \end{aligned}$$

So to prove $P(E_n \text{ i.o.})$, it suffices to show that for any n,

$$P\Big(\bigcap_{m=n}^{\infty} E_m^{\mathrm{c}}\Big) = 0. \tag{5}$$

Since the events E_n are mutually independent, so are the events E_n^{c} (Lemma 2.6.3), so

$$P\Big(\bigcap_{m=n}^{\infty} E_m^{\mathrm{c}}\Big) = \lim_{k\to\infty} P\Big(\bigcap_{m=n}^{k} E_m^{\mathrm{c}}\Big) = \lim_{k\to\infty} \prod_{m=n}^{k} P(E_m^{\mathrm{c}}) = \lim_{k\to\infty} \prod_{m=n}^{k} \big(1 - P(E_m)\big).$$

Taking logarithms implies

$$\lim_{k\to\infty} \ln P\Big(\bigcap_{m=n}^{k} E_m^{\mathrm{c}}\Big) = \lim_{k\to\infty} \sum_{m=n}^{k} \ln P(E_m^{\mathrm{c}}) = \lim_{k\to\infty} \sum_{m=n}^{k} \ln\big(1 - P(E_m)\big).$$

We now use the fact that the logarithm is a concave function so by the Subgradient Inequality (see Section 6.9⋆) the first-order Taylor series overestimates the logarithm. That is,

$$\ln(1-x) \leqslant \ln(1) + \ln'(1)(-x) = -x.$$

Thus

$$\sum_{m=n}^{k} \ln\big(1 - P(E_m)\big) \leqslant -\sum_{m=n}^{k} P(E_m).$$

Letting $k \to \infty$ we have

$$\ln P\Big(\bigcap_{m=n}^{\infty} E_m^{\mathrm{c}}\Big) \leqslant -\sum_{m=n}^{\infty} P(E_m) = -\infty,$$

which proves (5). ∎

26.11.2 Remark We can prove Part (1) of the Borel–Cantelli Lemma by the "method of indicators." Let

$$X(\omega) = \#\{i : \omega \in E_i\} = \sum_{i=1}^{\infty} \mathbf{1}_{E_i}.$$

Proposition 8.2.1 says that for each n,

$$\boldsymbol{E} \sum_{i=1}^{n} \mathbf{1}_{E_i} = \sum_{i=1}^{n} P(E_i),$$

so by the Monotone Convergence Theorem 5.13.2,

$$\boldsymbol{E}\, X = \boldsymbol{E} \sum_{i=1}^{\infty} \mathbf{1}_{E_i} = \sum_{i=1}^{\infty} P(E_i).$$

So if $\boldsymbol{E}\, X < \infty$, it must be that $P(X = \infty) = 0$, but the event $(X = \infty) = (E_n \text{ i.o.})$.

26.12 ⋆ Tail events and Zero-One laws

Let $X_1, \dots, X_n, \dots$ is a sequence of random variables. An event E is a **tail event** if it belongs to the σ-algebra generated by the tail sequence $X_n, X_{n+1}, \dots$ for every n. Letting $\sigma(X_n, X_{n+1}, \dots)$ denote the σ-algebra generated by the tail sequence, a tail event is an element of the **tail $\boldsymbol{\sigma}$-algebra**

$$\bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \dots).$$

26.12.1 Kolmogorov Zero-One Law *If $X_1, \dots, X_n, \dots$ is a sequence of mutually independent random variables, and E it a tail event, then either $P(E) = 0$ or $P(E) = 1$.*

For a proof see, e.g., Jacod and Protter [12, Theorem 10.6, p. 72], or Breiman [4, Theorem 3.12, p. 40].

26.13 ⋆ Transience and recurrence

26.13.1 Definition *A state i in a Markov chain is **recurrent** if starting in state i, the chain returns to state i infinitely often with probability one. That is,*

$$\mathbb{P}_i\,(X_t = i \text{ i.o.}) = 1.$$

*A state i in a Markov chain is **transient** if*

$$\mathbb{P}_i\,(X_t = i \text{ i.o.}) = 0.$$

A transient state is one that eventually never reoccurs. Note that recurrence depends only on the transition matrix. It turns out that every state is either recurrent or transient. To frame the proper theorem I need one more definition.

26.13.2 Definition *For a state i in a Markov chain, the **first passage time** to i, denoted T_i is defined to be*

$$T_i(\omega) = \inf\{t : X_t(\omega) = i\},$$

where $\inf \varnothing = \infty$.

26.13.3 Theorem (Recurrence and transience are class properties) *If i is recurrent and i communicates with j, then j is recurrent. If i is transient and i communicates with j, then j is transient.*

We now have the following theorem, which may be found, e.g, in Norris [16, Theorem 1.5.2, p. 26] or Karlin [13, Theorem 5.1, p. 48].

26.13.4 Theorem *For a Markov chain, either*

1. $\mathbb{P}_i(T_i < \infty) = 1$, *in which case i is recurrent and* $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$, *or else*
2. $\mathbb{P}_i(T_i < \infty) < 1$, *in which case i is transient and* $\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty$.

Consequently, every state is either recurrent or transient, and

$$\mathbb{P}_i(T_i < \infty) = 1 \iff \sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty.$$

Bibliography

[1] C. D. Aliprantis and K. C. Border. 2006. *Infinite dimensional analysis: A hitchhiker's guide*, 3d. ed. Berlin: Springer–Verlag.

[2] 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

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

[4] L. Breiman. 1968. *Probability*. Reading, Massachusetts: Addison Wesley.

[5] ——— . 1986. *Probability and stochastic processes: With a view toward applications*, 2d. ed. Palo Alto, California: Scientific Press.

[6] W.-K. Ching and M. K. Ng. 2006. *Markov chains: Models, algorithms and applications*. International Series in Operations Research and Management Science. New York: Springer.

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

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

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

[10] W. Feller. 1968. *An introduction to probability theory and its applications*, 3d. ed., volume 1. New York: Wiley.

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

[12] J. Jacod and P. Protter. 2004. *Probability essentials*, 2d. ed. Berlin and Heidelberg: Springer.

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

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

[15] 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.

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

[17] L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University.

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