

Markov chains: calculations over finite and indefinite time horizons

Monday, September 15, 2008
11:46 AM

Homework 1 is due Monday, September 22 at 12 PM.

Interested in possibly converting lecture notes into LaTeX for financial compensation (approx. 10 hours/week)? Either during semester or over winter break. Please let me know by Wednesday, September 17 at 5 PM.

Proof of Chapman-Kolmogorov equation (from last time)

$$P(X_m = j | X_n = i) = ?$$

$$\{X_m = j\} = \bigsqcup_{j' \in S} \{X_m = j, X_{n'} = j'\}$$

disjoint union

Useful principle: The rules for probability theory apply equally well to conditional probabilities, provided the conditions are carried along in every probability in the equation.

$$P\left(\bigsqcup_{j' \in J} A_{j'} \mid C\right) = \sum_{j' \in J} P(A_{j'} \mid C)$$

The mathematical justification is that if P is a probability measure, then so is $P(\cdot \mid C)$

provided that $P(C) > 0$ so that the conditional probability is well-defined.

Applying this to our desired conditional probability:

$$\begin{aligned} P(X_m = j \mid X_n = i) &= \sum_{j' \in S} P(X_m = j, X_{n'} = j' \mid X_n = i) \\ &\text{with } n < n' < m. \end{aligned}$$

What we've done here is a common trick in probability theory and stochastic processes -- the computation of a certain probability is expressed in terms of probabilities of events with more detail.

Using the principle of carrying over probability relationships to conditional probabilities, we have:

$$P(A \text{ and } B | C) = P(A | B \text{ and } C) P(B | C)$$

because

$$P(A \text{ and } B) = P(A | B) P(B)$$

Apply this to the summands:

$$P(\overset{A}{\underbrace{X_m = j}} | \overset{B}{\underbrace{X_{n'} = j'}} | \overset{C}{\underbrace{X_n = i}})$$

$$= P(X_m = j | X_{n'} = j', X_n = i) P(X_{n'} = j' | X_n = i)$$

This is another common trick in the analysis of probability and stochastic processes, which is to break up a calculation involving the intersection of two or more events using conditional probabilities -- this can be used to convert the calculation into computing the probability of one event at a time, given that other events happen. Moving events behind the conditioning bar makes them "given" information and often helps with the calculation.

$$P(X_m = j | X_{n'} = j', X_n = i)$$

$$= P(X_m = j | X_{n'} = j')$$

Because of Markov property and $n < n' < m$.

Putting this all together:

$$P(X_m = j | X_n = i) = \sum_{j' \in S} P(X_m = j | X_{n'} = j') P(X_{n'} = j' | X_n = i)$$

which is the Chapman-Kolmogorov equation we wanted to show.

This illustrates a common combination of techniques used in deriving properties of stochastic processes:

- a. Introduce additional useful information (value of the state at an intermediate time n')
- b. Use the definition of conditional probability to simplify the

- resulting probabilities into products of conditional probabilities with only one simple event appearing to the left of conditioning bar
- c. Using Markov property to simplify the portions of the conditional probability behind the conditioning bar.

We showed at the end of last lecture how to calculate arbitrary conditional probabilities:

$$\# \quad P(X_m = j \mid X_n = i) = (P^{m-n})_{ij} \\ \text{for } m-n \geq 0.$$

To calculate arbitrary probabilities for Markov chains over finite time horizons, also need to be able to calculate quantities like:

$$P(X_{n_0} = j_0)$$

$$\begin{aligned} P(X_{n_0} = j_0) &= P\left(\bigsqcup_{i \in S} \{X_{n_0} = j_0, X_0 = i\}\right) \\ &\quad \downarrow \text{(Union of mutually exclusive)} \\ &= \sum_{i \in S} P(X_{n_0} = j_0, X_0 = i) \\ &\quad \downarrow \text{(defn of cond prob)} \\ &= \sum_{i \in S} P(X_{n_0} = j_0 \mid X_0 = i) P(X_0 = i) \\ &= \sum_{i \in S} (P^{n_0})_{ij_0} \phi_i \\ &\quad \uparrow \text{initial prob dist} \\ &\quad P(X_0 = i) = \phi_i \end{aligned}$$

$$\Delta \quad P(X_{n_0} = j_0) = (\vec{\phi} \cdot P^{n_0})_{j_0}$$

Notice that the derivation of this equation again used a similar strategy as in the derivation of Chapman-Kolmogorov (w/o invoking Markov property).

Based on our comments last time



the ability to compute conditional probabilities of the form



and singleepoch probability distributions of the form



the ability to compute conditional probabilities of the form $\#$

and single-epoch probability distributions of the form \triangle

we can compute any probability about what the Markov chain is doing over a fixed finite time horizon.

Now, how do we compute properties of the Markov chain that may take place over an indefinite time horizon? Need to develop techniques for this.

Stationary and limit distributions

How should we choose the probability distribution for the initial state?

$$\theta_j = P(X_0 = j)$$

Sometimes it's prescribed, especially if the system is able to be initialized manually.

So if for example, we know that $X_0 = 7$

then $\vec{\theta} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$

But what if one doesn't really initialize the system but rather simply starts observing it at time 0, after it's been running for a while.

Many times it is natural to initialize the system by its **stationary distribution** which is defined to be a solution $\vec{\pi}$ of the following equations:

$$\vec{\pi} \cdot P = \vec{\pi}$$

$$\sum_{j \in S} \pi_j = 1$$

$$\pi_j \geq 0 \text{ for } j \in S$$

The reason for this definition is that if we initialize the system so that

$$P(X_0 = j) = \pi_j$$

$$\begin{aligned} \text{then } P(X_n = j) &= (\vec{\pi} \cdot P^n)_j \\ &= (\vec{\pi} \cdot P \cdot P^{n-1})_j = (\vec{\pi} \cdot P^{n-1})_j \\ &\dots = \pi_j \end{aligned}$$

If the Markov chain is in a stationary distribution at a given epoch, then it remains in that same stationary distribution for all future epochs.

In some way it is a natural probability distribution for the states of the Markov chain (provided the Markov chain is time-homogenous).

How do we know a stationary distribution exists and whether it's unique...we'll answer those questions after some preparation to define the conditions under which this is guaranteed.

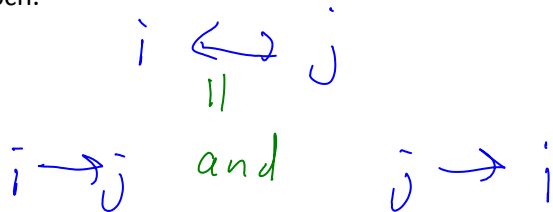
Communication classes of a Markov chain

Two states $i, j \in S$ are said to **communicate** in a Markov chain with probability transition matrix P , provided that there exist some positive integers

$$m, n \in \mathbb{N} = \mathbb{Z}_+ = \{1, 2, \dots\}$$

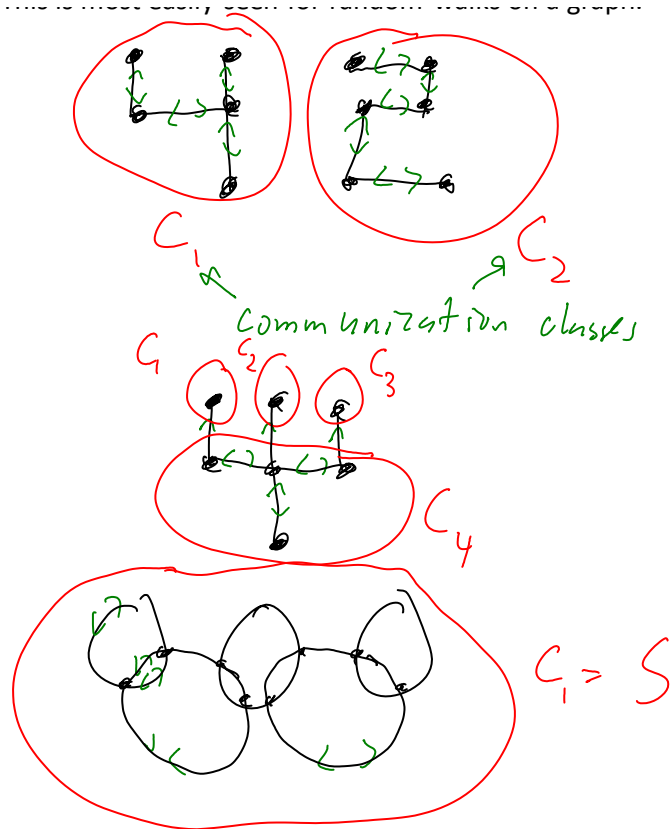
$$\text{Such that } (P^n)_{ij} > 0 \text{ and } (P^m)_{ji} > 0.$$

This is just the mathematical encoding of the intuitive concept that there is some possibility (with nonzero probability) of starting in state i and visiting state j at a later epoch, and of starting in state j and visiting i at a later epoch.



It turns out that communication so defined is an equivalence relationship, which implies that the state space S can be decomposed uniquely into **communication classes**, such that within each class all states communicate with each other, but there is no (two-way) communication between classes.

This is most easily seen for random walks on a graph.



If the entire state space forms a single communication class (all states communicate with each other), then the Markov chain is said to be **irreducible**. Otherwise, the Markov chain is said to be **reducible**.

Proposition: Any finite-state, discrete-time, time-homogenous **irreducible** Markov chain has a **unique stationary distribution** π . Moreover, the values of the stationary distribution have the following interpretation:

$$\pi_j = \frac{1}{\mathbb{E}[\tau_j(1) | X_0 = j]}$$

Where

$$\tau_j(1) = \min \{n : n > 0 \text{ s.t. } X_n = j\}$$

is the **first passage time** to state j . (the first epoch in the future at which state j is visited by the Markov chain.)

Notation for **expectation** of a discrete random variable Y :

$$E Y = \langle Y \rangle \equiv \sum_{j \in S} j P(X=j)$$

Conditional expectation:

$$E [Y | A] = \langle Y | A \rangle \equiv \sum_{j \in S} j P(X=j | A)$$

Note that in fact $E [T_j(1) | X_0=j]$ provided that $P(A) > 0$.

can be interpreted as the average **first return time** to state j , meaning the average time between successive visits to state j (not just from the initial epoch). This statement uses time-homogeneity and strong Markov property which we'll discuss later.

Key result about long-time properties of Markov chains requires the introduction of one more property of Markov chains.

The **period** $d(i)$ of a state $i \in S$ in a Markov chain is defined to be:

$$d(i) = \gcd \{ n \geq 0 : (P^n)_{ii} > 0 \}$$

(the greatest common denominator of all epochs over which state i can return to itself).

If $d(i)=1$, the state is said to be **aperiodic**.

One can show that all states in a given communication class have the same period, so that periodicity is a class property.

Theorem: If $\{X_n\}_{n=0}^{\infty}$

is an **aperiodic, irreducible** finite-state, discrete-time, time-homogenous Markov chain with probability transition matrix P , then the Markov chain has a unique stationary distribution π

And: $\lim_{n \rightarrow \infty} P(X_n = j) = \pi_j$

That is, the stationary distribution also serves as a **limit distribution** which attracts the Markov chain from arbitrary initial distribution.

And the **Law of Large Numbers for Markov chains** holds:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{m=1}^n f(X_m) = \sum_{j \in S} f(j) \pi_j$$

time average *ensemble average*
↓ ↓

with probability 1, for deterministic functions f on the state space.

Long time averages of $f(X_n)$ are equivalent to averages against the stationary distribution. This is also known as an **ergodic property**.