Derivation of results regarding long-time properties and stationary distribution

Thursday, September 25, 2008
11:59 AM

For detailed derivations of the results with complete proofs, see:

- Resnick Secs. 2.12-2.15: derivations are stochastic/probabilistically based like what we will sample in class. This derivation carries over to the case where the Markov chain has countably infinite states (which we'll do in a few weeks)
- Karlin & Taylor Secs. 3.1 & 3.2: Matrix-based approach (old-fashioned)

In the lecture, I will provide an overview of the techniques and carry out one detailed derivation.

To get intuition for why the various theorems about long-time properties of Markov chains hold, it's helpful to first think about matrix properties.

Consider an eigenvalue/eigenvector decomposition of the probability transition matrix P:

$$P = \sum_{j=1}^{M} \lambda^{(j)} \vec{u}^{(j)} \otimes \vec{v}^{(j)} \qquad \left(= \sum_{j=1}^{M} \lambda^{(j)} |\vec{u}^{(j)}\rangle\langle\vec{v}^{(j)}| \right)$$

where $P\vec{u}^{(j)} = \lambda^{(j)}\vec{u}^{(j)}$ (right eigenvector)

$\vec{v}^{(j)}P = \lambda^{(j)}\vec{v}^{(j)}$ (left eigenvector)

$P^{\dagger}\vec{v}^{(j)}$ ''

$$\vec{u}^{(i)}\cdot\vec{v}^{(j)} = \delta_{ij} = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{if } i\neq j \end{cases}$$

provided that the matrix is diagonalizable. Note that we need to use a biorthogonal (biorthonormal) representation in terms of left and right eigenvectors because the matrix P is not necessarily symmetric. One place to see this development is in Appendix B of Karlin and Taylor.

$$\vec{u}\otimes\vec{v} = \begin{pmatrix} u_1v_1 & u_1v_2 & \cdots & u_1v_M \\ \vdots & & & \\ u_Mv_1 & \cdots & & u_Mv_M \end{pmatrix}$$

$$P = S^{-1}\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_M \end{pmatrix} S$$

transformation matrix; rows = left eigenvectors

No one actually guarantees that the probability transition matrix is diagonalizable (it's not symmetric!) but if it's not, similar ideas apply but you have to work with Jordan forms.

Let's consider now what it would mean to have a unique stationary distribution. Looking back the definition, it means that there is only one (left) eigenvector with eigenvalue 1 that has nonnegative entries (up to rescaling). It's easy to see from the fact that

$$P \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} P_{11} + P_{12} + \cdots + P_{1M} \\ \vdots \\ P_{M1} + P_{M2} + \cdots + P_{MM} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

This shows that the probability transition matrix always has a right eigenvector with eigenvalue 1. This guarantees the existence of at least one left eigenvector with eigenvalue 1. But how do you know that one of these left eigenvectors has all nonnegative entries and that it's unique?

This is given by the Perron-Frobenius theorem for matrices with nonnegative entries. (Karlin and Taylor Appendix 2):

- If for some $m>0$, P^m has all positive entries, then the eigenvalue of P with the largest modulus is simple (only one eigenvector corresponds to it), is real and positive, and the corresponding eigenvector can be expressed with all nonnegative entries, and all other eigenvalues have strictly smaller modulus.

Applying this to our present case, irreducibility and aperiodicity of Markov chain can be shown to imply that for some $m>0$, P^m has all positive entries.

Irreducible and aperiodic imply that I have a simple eigenvalue

$$\lambda_1 = 1 \quad \text{and} \quad |\lambda_j| < 1 \quad \text{for } j = 2, \ldots, M.$$

$$\phi_i^{(j)} = P(X_j = i)$$

$$\vec{\phi}^{(j)} = \vec{\phi}^{(0)} P^j$$

$$\lim_{j\to\infty} \vec{\phi}^{(j)} = \lim_{j\to\infty} \vec{\phi}^{(0)} \cdot P^j$$

$$= \lim_{j\to\infty} \vec{\phi}^{(0)} \left(\sum_{i=1}^{M} \left(\lambda^{(i)}\right)^j \vec{u}_i \otimes \vec{v}_i \right)$$

$$= \vec{\phi}^{(0)} \cdot \underbrace{\lambda^{(1)}}_{=1} \underbrace{\vec{u}_1}_{=\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}} \otimes \underbrace{\vec{v}_1}_{=\vec{\pi}}$$

$$= \vec{\pi}$$

That explains why the stationary distribution is the limit distribution.

Uniqueness of the stationary distribution follows from the fact that 1 is a simple eigenvalue.

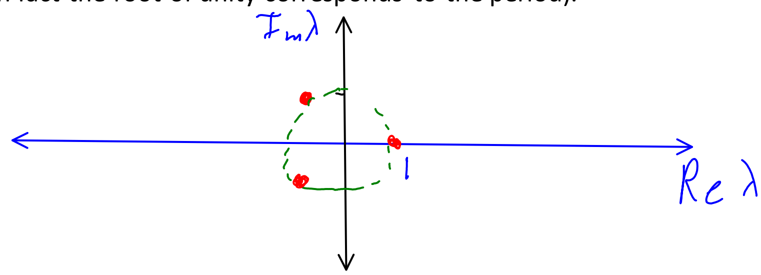
What about if we just have irreducibility but not aperiodicity? Then another variation of the Perron-Frobenius tells you that the matrix has all of the properties described above except you do not know that the other eigenvalues have strictly smaller modulus than the eigenvalue with the largest modulus.

Irreducibility but maybe periodicity:

$\lambda^{(1)} = 1$ (simple) (true but not immediately clear how to prove this with Perron-Frobenius!)

$$|\lambda^{(j)}| \le 1 \quad \text{for } j = 2, \ldots, M$$

What happens for periodic, irreducible Markov chains that do not have a limit distribution is that they have eigenvalues which are roots of unity (in fact the root of unity corresponds to the period).



The part of the matrix that corresponds to the eigenvalues with unit modulus (on the unit circle in the complex plane) do not decay, but oscillate. So this spoils convergence to the limit distribution but does not spoil the law of large numbers because the oscillations average out when you take a time average.

Full details: Karlin and Taylor Appendix B

Intuitive description: Lawler Ch. 1

These arguments involving Perron-Frobenius theorem don't carry over to the case of Markov chains with infinite states.

Let's look at the long-time properties concerning stationary distribution from a more probabilistic perspective (as in Resnick), which will give us also the connection between stationary distribution and the following formulas:

$$\pi_j = \frac{1}{\mathbb{E}[\tau_j(1) \mid X_0 = j]} \leftarrow \text{average first return time to state } j$$

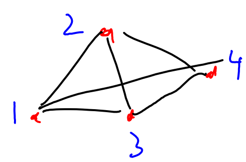
$$\text{where } \tau_j(1) = \min\{n > 0 : X_n = j\}$$

$$\frac{\pi_k}{\pi_j} = \mathbb{E}\left[T_{j,k} \mid X_0 = j\right] \qquad I = \begin{cases} 1 & \text{if true} \\ 0 & \text{if false} \end{cases}$$

$$\text{where } T_{j,k} = \sum_{n=0}^{\tau_j(1)-1} I\{X_n = k\}$$

$$= \#\text{ visits to state } k \text{ before visiting state } j$$

Intuition behind these results:



2 2 3 2 1 2 2 3 4 1 2 2 3 1 2 2

$$\pi_1 = 1/5$$
$$\pi_2 = \frac{1}{2}$$
$$\pi_3 = \frac{1}{5}$$
$$\pi_4 = \frac{1}{10}$$

The proof is mostly important for the techniques it uses, rather than the overall structure of the argument. (Resnick)

Strategy: Hypothesize that the stationary distribution has the following form:

$$\vec{\tilde{\pi}} \text{ where } \tilde{\pi}_k = \frac{\mathbb{E}[T_{j,k} \mid X_0 = j]}{\mathbb{E}[\tau_j(1) \mid X_0 = j]}$$

Here j is any fixed reference state in the Markov chain. Idea is that by looking at the epochs between successive visits to reference state j, the right hand side is intuitively the fraction of those epochs in which state k is visited. (Resnick Sec. 2.5 calls this sort of trick of chopping up the trajectory of a Markov chain into useful blocks of sequences of epochs (here into blocks of epochs between successive visits to a reference state j) the dissection principle).

Let's check that this hypothesis actually does satisfy the equations for the stationary distribution, and then appeal to some argument for uniqueness which then would yield the claimed results. The relationships between the stationary distribution and the trajectory properties described above follow directly from the above hypothesis, using for example that $T_{j,j} = 1$.

So we need to check the following:

a) $\tilde{\pi}_k \geq 0$ ✓

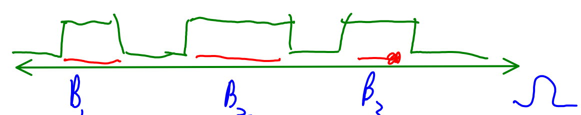
b) $\sum_{S} \tilde{\pi}_k = 1$ ✓

c) $\tilde{\vec{\pi}} \cdot P = \vec{\pi}$

a): obvious other than having to check that neither the numerator or denominator is infinite. This can be verified through a technical argument by using irreducibility (see Lawler and his exercise 1.7) but this point becomes much more subtle and important when we consider Markov chains with infinite states, where we need to check more than irreducibility.

b)
$$\sum_{k \in S} T_{j,k} = \sum_{k \in S} \left(\sum_{n=0}^{\tau_j(1)-1} I\{X_n = k\} \right)$$
$$= \sum_{n=0}^{\tau_j(1)-1} \sum_{k \in S} I\{X_n = k\}$$
$$= \sum_{n=0}^{\tau_j(1)-1} I\left(\bigsqcup_{k \in S} \{X_n = k\} \right)$$

$$\left(I\left\{ \bigsqcup_{k} B_k \right\} = \sum_{k} I\{B_k\} \right)$$



$$= \sum_{n=0}^{\tau_j(1)-1} I(X_n \in S) = \sum_{n=0}^{\tau_j(1)-1} 1 = \tau_j(1)$$

$$\sum_{k \in S} \tilde{\pi}_k = \sum_{k \in S} \frac{\mathbb{E}[T_{j,k} \mid X_0 = j]}{\mathbb{E}[\tau_j(1) \mid X_0 = j]}$$

$$= \mathbb{E}\left[\sum T_{j,k} \mid X_0 = j \right]$$

$$= \frac{\mathbb{E}\left[\sum_{k\in S} T_{j,k} \,\middle|\, X_0 = j\right]}{\mathbb{E}\left[\tau_j(1) \,\middle|\, X_0 = j\right]}$$

$$= \frac{\mathbb{E}\left[\tau_j(1) \,\middle|\, X_0 = j\right]}{\mathbb{E}\left[\tau_j(1) \,\middle|\, X_0 = j\right]} = 1. \checkmark$$

To prove c), enough to show:

$$\vec{r} \cdot P = \vec{r} \quad \text{where } r_k = \mathbb{E}\left[T_{j,k} \,\middle|\, X_0 = j\right]$$

because $\vec{\pi} = c\,\vec{r}$

$$c = \frac{1}{\mathbb{E}\left[\tau_j(1) \,\middle|\, X_0 = j\right]}$$

(normalization constant)

($\vec{r}$ is an invariant distribution: $r_k \geq 0$, $\vec{r} \cdot P = \vec{r}$)

To do this we first will reprocess the expression for υ so that it will be easier to analyze.

$$r_k = \mathbb{E}\left[\sum_{n=0}^{\tau_j(1)-1} I\{X_n = k\} \,\middle|\, X_0 = j\right]$$

convert random sum into a deterministic summation

$$= \mathbb{E}\left[\sum_{n=0}^{\infty} I\{X_n = k, \tau_j(1) > n\} \,\middle|\, X_0 = j\right]$$

makes terms w/ $n \geq \tau_j(1)$ be 0.

Randomness has been moved from the sum and now resides entirely in the indicator function (rather than being in two places).

$$= \sum_{n=0}^{\infty} \mathbb{E}\left[I\{X_n = k, \tau_j(1) > n\} \,\middle|\, X_0 = j\right]$$

Commuting expectation w/ deterministic sum

$$\mathbb{E}\, I(B) = P(B)$$

$$\mathbb{E}\, I(B) = 1 \cdot P(B) + 0 \cdot P(\Omega \setminus B) = P(B)$$

$$r_k = \sum_{n=0}^{\infty} P\left(X_n = k,\ T_j(1) > n \mid X_0 = j \right)$$