Probability I
Proof of Ergodic Theorem

Daniel Yu

October 31, 2024

Contents

1 Proof of Ergodic Theorem

Note. The ergodic theorem for finite state markov chain states the following. Let P be a the transition matrix. Then $\lim_{n\to\infty}$ exists has constant columns and each row is made up of the unique stationary vectors of P. Furthermore, for any intial probability distribution v, $\lim_{n\to\infty} P^n v = \pi$ the unique stationary distribution

Proof. $\forall i, j \in \Omega$, if x_n is irreducible, aperiodic, markov chain,

$$\lim_{n\to\infty} P[X_n = j \mid X_0 = i] = \pi_j.$$

where π is the stationary distribution vector. If $x_0 \sim \pi$ i.e. $P[x_0 = j] = \pi_j$. What is the distribution of the following?

$$\begin{aligned} P[X_1 = k] &= \sum_{j\in\Omega} P[X_1 = k \mid X_0 = j] P[X_0 = j] \\ &= \sum_{j\in\Omega} P_{j,k}\pi \\ &= (\pi P)_k \cdot \pi_k. \end{aligned}$$

If the intial distribution is π then the distribution at all times is also π. Let's consider if $v \neq \pi$.

Let (X_n, Y_n) be two pariticles evolving with the same markov chain dynamics (i.e. same transition matrix) where $X_0 = i$ and $Y_0 \sim \pi$. So X_n evolves in a complicated way but $Y_n \sim \pi$. At any time n, every state i chooses another state k with probability $P_{i,k}$, every particle that was at state i at times n moves to state k at time $n+1$ since the transition probability is the same for X_n and Y_n. This is known as a **coalescing walk**, when the two particles meet, they now have the same distribution at any time $n+1$ after.

Let $T = \min\{n : X_n = Y_n\}$ at any time after $T, X_n = Y_n$ since they follow the same set of instructions.Let's compute,

$$P[X_n = j \mid X_0 = i, T \le n] = P[Y_n = j \mid X_0 = i, T \le n] = \pi_j.$$

By law of total probability:

$$\begin{aligned} P[X_n = j \mid X_0 = i] &= P[X_n = j \mid n < T, x_0 = i] \cdot P[n < T \mid X_0 = i] + \\ &\quad P[X_n = j \mid n \ge T, X_0 = i] \cdot (1 - P[n < T \mid X_0 = i]) \\ &= P[X_n = j \mid n < T, X_0 = i] P[n < T \mid X_0 = i] + \pi_j (1 - P[n < T \mid X_0 = i]) \\ &= \pi_j + P[n < T \mid X_0 = i](P[X_n = j \mid n < T, X_0 = i] - \pi_j) \\ P[X_n = j \mid X_0 = i] - \pi_j &= P[n < T \mid X_0 = i](P[X_n = j \mid n < T, X_0 = i] - \pi_j) \\ \|P[X_n = j \mid X_0 = i] - \pi_j\| &= P[n < T \mid X_0 = i] \|(P[X_n = j \mid n < T, X_0 = i] - \pi_j)\| \\ &\le P[n < T \mid X_0 = i] \end{aligned}$$

.

We want to show that $P[n < T \mid X_0 = i] \to 0$ as $n \to \infty$ since this implies that $P[X_n = j \mid X_0 = i] = \pi_j$ and the random variable X_n approaches the stationary distribution despite the intial distribution.

Consider the $\mid \Omega \mid \times \mid \Omega \mid$ joint state space (X_n, Y_n) that describes the markov chain dynamics of X_n, Y_n where

$$(i, j) = (X_n, Y_n).$$

Since P is irreducible and aperiodic, then we can reach any pair (i, j) from any other state (k, l). This means that $P[$ everything absorbed eventually$] = 1$. Equivalently, $P[T < \infty] = 1$ and

$$\lim_{n\to\infty} P[n < T] = P[T = \infty] = 0.$$

and we conclude that $\|P[X_n = j \mid X_0 = i\| \to 0$ as $n \to \infty$. In fact, $\exists p > 0$ that $P[n < T] = P[n < T] \le (1-p)^n$. $\square$

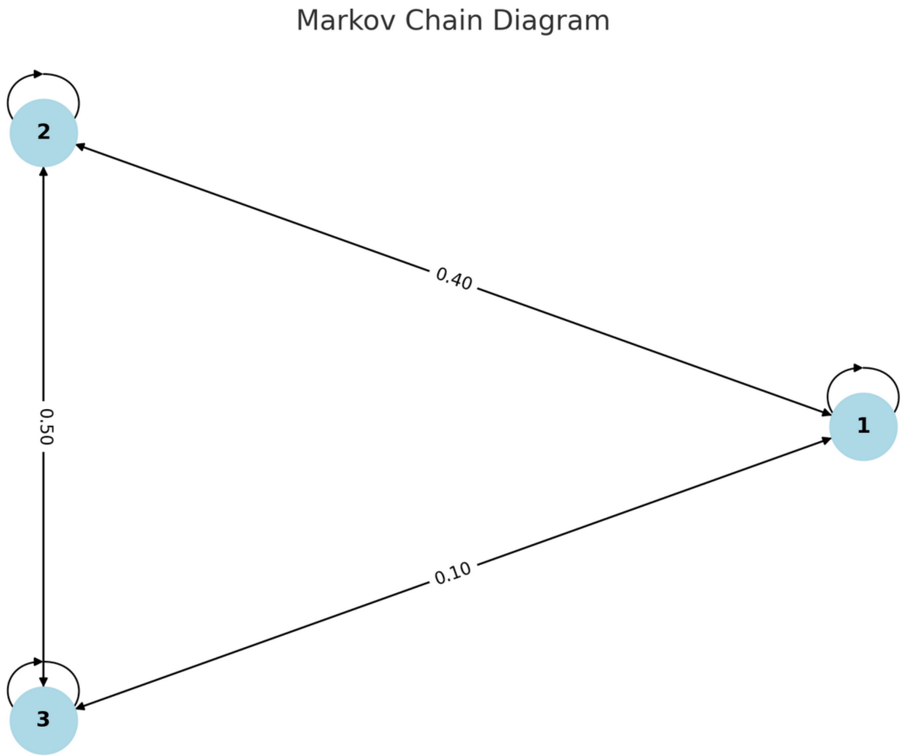
2 Mixing Times

Note. How can we compare distributions? We know that eventually the distirbution will converge to π but what is the distribution at $n = 100$, what about $n = 1000$? It will be almost π but not quite.

Definition 1. Let Ω be finite, and u, v be two rwo vectors. WE can denote the **variation distance** as

$$d_{tv}(u,v) = \frac{1}{2}\sum_{i\in\Omega} \|u_i - v_i\|.$$

Consider the following example:



The transition matrix P is

$$P = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.4 & 0.2 \\ 0.1 & 0.5 & 0.4 \end{bmatrix}, \quad \pi^{(0)} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Lets consider some intial vector $v = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ We can compute the variation distance after two steps (how far our distribution v is from the stationary distribution) as follows.

1. Solve for $\pi P = \pi$, we get

$$\pi = \begin{bmatrix} 0.267 \\ 0.4 \\ 0.333 \end{bmatrix}$$

2. Compute P^2 and $vP^2 = v^2$.

$$P^2 = P \times P = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.4 & 0.2 \\ 0.1 & 0.5 & 0.4 \end{bmatrix} \times \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.4 & 0.2 \\ 0.1 & 0.5 & 0.4 \end{bmatrix} = \begin{bmatrix} 0.39 & 0.37 & 0.24 \\ 0.38 & 0.38 & 0.24 \\ 0.29 & 0.43 & 0.28 \end{bmatrix}$$

, so we get $v^2 = vP^2 = \begin{bmatrix} 0.39 & 0.37 & 0.24 \end{bmatrix}$

3. We get that

$$d_{tv}(v^2,\pi) = \frac{1}{2}\sum_{i=1}^{3}\left|v_i^{(2)} - \pi_i\right| = \frac{1}{2}\left(|0.39 - 0.267| + |0.37 - 0.4| + |0.24 - 0.333|\right) = 0.123$$

Thus, the total variation distance after 2 steps is approximately 0.123.

Definition 2. Define $DTM(n) =$ distance to mixing time at n $= \sup_i d_{tv}(v_n^i, \pi)$. If v_n is the distribution of v_n with any intial conditions, $d_{tv}(v_n, \pi) \leq DTM(n)$.

We define the mixing time with threshold ε as

$$\min\{n \mid DTM(n) \leq \varepsilon\}.$$

i.e. the minimum number of steps to get ε close to the stationary distribution π. The markov chain after this timestep is consider **mixed**

Let's continue our previous example. What is the mixing time with $\varepsilon = .1$?

$$\begin{aligned}
\text{Step 0:}\quad & v^{(0)} = \begin{bmatrix}1 & 0 & 0\end{bmatrix} \\
& d_{\mathrm{TV}}(v^{(0)},\pi) = \frac{1}{2}\sum_{i=1}^{3}\left|\pi_i^{(0)} - \pi_i\right| = 0.733 \\
\\
\text{Step 1:}\quad & v^{(1)} = v^{(0)}P = \begin{bmatrix}0.5 & 0.3 & 0.2\end{bmatrix} \\
& d_{\mathrm{TV}}(v^{(1)},\pi) = 0.233 \\
\\
\text{Step 2:}\quad & v^{(2)} = v^{(1)}P = \begin{bmatrix}0.39 & 0.37 & 0.24\end{bmatrix} \\
& d_{\mathrm{TV}}(v^{(2)},\pi) = 0.123 \\
\\
\text{Step 3:}\quad & v^{(3)} = v^{(2)}P = \begin{bmatrix}0.35 & 0.39 & 0.26\end{bmatrix} \\
& d_{\mathrm{TV}}(v^{(3)},\pi) = 0.1 \\
\\
\text{Step 4:}\quad & v^{(4)} = \pi^{(3)}P = \begin{bmatrix}0.33 & 0.4 & 0.27\end{bmatrix} \\
& d_{\mathrm{TV}}(v^{(4)},\pi) = 0.0953
\end{aligned}$$

Thus, the mixing time is 4.

Theorem 1. Perron-Frobenius Theorem
If P is the transition matrix of an irreducible markov chain then $\lambda = 1$ is always an eigenvalue with row-eigenvalue π of strictly positive entries. All other eigenvalues $\{\lambda_i\}_{i=2}^n$ have $\mid \lambda \mid < 1$. If α is a row eigenvector of P with eigenvalue $\neq 1$, then $\sum_{i=1}^n \alpha_i = 0$.

Note. Eigenvalues can be complex.

Corollary 1. Any vector P with $p_i \geq 0$ and $\sum_{i=1}^n P_i = 1$ can be written as $P = \pi + \sum_{i=1}^n c_i\vec{\alpha_i}$ hwere $c_i \in \mathbb{R}$ and α_i are row vectors of P with eigenvalue $\neq 1$.

3 Card Shuffling

Consider the n card deck is shuffled by moving the top card to a uniformly chosen random position. $\Omega =$ arrangement of n cards. Clearly, $\mid \Omega \mid = n!$. Every row of $P_{|\Omega|\times|\Omega|}$ would have n positive entries, with probability to be $\frac{1}{n}$ and all entries 0. It will be a very sparse matrix.

It turns out that this matrix is doubly stochastic! The stationary distribution is just the uniform distribution. By the ergodic theorem, this deck is well-shuffled, because $n \to \infty,^n = \pi$ where π is the uniform distribution.