Introduction to Stochastic Processes

Dr. Naomi Feldheim, `naomi.feldheim@biu.ac.il`

Summary by Ari Feiglin

Contents

1 Introduction **1**

2 Markov Chains **1**

2.1 Stationary Distributions and the Convergence of Markov Chains . . . 8

Existence and Uniqueness Theorem . . . 9

1 Introduction

This course will focus on tools which can be used to study random processes. A random process is a sequence of random variables which represent measurements of the process. Examples of random processes are random walks (these are commonly described as the path a drunk man would take while trying to get home), card shuffles (which can be viewed as choosing a card and placing it randomly in the deck), and branching (for example the population of bunnies in a specific area: the random variable being the number of bunnies in each generation).

2 Markov Chains

2.0.1 Definition

A **discrete-time Markov process** is a sequence of random variables $\{X_n\}_{n\geq 0}$. This sequence is called a **Markov chain** on a set of states S if:

(1) For every n, $X_n \in S$ almost surely (meaning $\mathbb{P}(X_n \in S) = 1$),

(2) For every $n \geq 0$ and for every $s_0, \dots, s_{n+1} \in S$,

$$\mathbb{P}(X_{n+1} = s_{n+1} \mid X_0 = s_0, \dots, X_n = s_n) = \mathbb{P}(X_{n+1} = s_{n+1} \mid X_n = s_n)$$

ie. the probability of the next measurement being some arbitrary value is dependent only on the previous measurement. This is only necessary if $\mathbb{P}(X_0 = s_0, \dots, X_n = s_n) > 0$.

In this course S will always be countable. We can also write the second condition using distributive equivalence:

$$X_{n+1}|X_0, \dots, X_n \overset{d}{=} X_{n+1}|X_n$$

Notice how the Markov property can be strengthened in various ways, for example if $n > m$ then

$$\begin{aligned}
&\mathbb{P}(X_n = s_n \mid X_{n-1} = s_{n-1}, \dots, X_m = s_m) \\
&\qquad = \sum_{s_m,\dots,s_0} \mathbb{P}(X_n = s_n \mid X_{n-1} = s_{n-1}, \dots, X_0 = s_0) \cdot \mathbb{P}(X_{m-1} = s_{m-1}, \dots, X_0 = s_0 \mid X_{n-1} = s_{n-1}, \dots, X_m = s_m) \\
&\qquad = \mathbb{P}(X_n = s_n \mid X_{n-1} = s_{n-1}) \cdot \sum \mathbb{P}(X_{m-1} = s_{m-1}, \dots, X_0 = s_0 \mid X_{n-1} = s_{n-1}, \dots, X_m = s_m) \\
&\qquad = \mathbb{P}(X_n = s_n \mid X_{n-1} = s_{n-1})
\end{aligned}$$

This can be viewed as the base case for

$$\mathbb{P}(X_{n+k} = s_{n+k} \mid X_n = s_n, \dots, X_m = s_m) = \mathbb{P}(X_{n+k} = s_{n+k} \mid X_n = s_n, \dots, X_{m'} = s_{m'})$$

where $m' < m$. This is since for $k = 1$, both of these are equal to $\mathbb{P}(X_{n+1} = s_{n+1} \mid X_n = s_n)$. The induction step follows by

$$\begin{aligned}
&\mathbb{P}(X_{n+k+1} = s_{n+k+1} \mid X_n = s_n, \dots, X_m = s_m) \\
&\qquad = \sum_{s_{n+1}} \mathbb{P}(X_{n+k+1} = s_{n+k+1} \mid X_{n+1} = s_{n+1}, \dots, X_m = s_m) \cdot \mathbb{P}(X_{n+1} = s_{n+1} \mid X_n = s_n, \dots, X_m = s_m) \\
&\qquad = \sum_{s_{n+1}} \mathbb{P}(X_{n+k+1} = s_{n+k+1} \mid X_{n+1} = s_{n+1}, \dots, X_{m'} = s_{m'}) \cdot \mathbb{P}(X_{n+1} = s_{n+1} \mid X_n = s_n, \dots, X_{m'} = s_{m'}) \\
&\qquad = \mathbb{P}(X_{n+k+1} = s_{n+k+1} \mid X_n = s_n, \dots, X_{m'} = s_{m'})
\end{aligned}$$

By taking $m' = 0$ and $m = n$ we get $\mathbb{P}(X_{n+k} = s_{n+k} \mid X_n = s_n) = \mathbb{P}(X_{n+k} = s_{n+k} \mid X_n = s_n, \dots, X_0 = s_0)$, or in other words for all $m < n$,

$$\mathbb{P}(X_n = s_n \mid X_m = s_m, \dots, X_0 = s_0) = \mathbb{P}(X_n = s_n \mid X_m = s_m)$$

This can be even further strengthened: let $\varnothing \neq B \subseteq \{0, \dots, n-1\}$ and $m = \max B$ then

$$\mathbb{P}(X_n = s_n \mid \forall i \in B \colon X_i = s_i) = \mathbb{P}(X_n = s_n \mid X_m = s_m)$$

To prove this let $C = \{0, \dots, m\} \setminus B$ then

$$\begin{aligned}
\mathbb{P}(X_n = s_n \mid \forall i \in B \colon X_i = s_i) &= \sum_{(s_i)_{i\in C} \in S^C} \mathbb{P}(X_n = s_n \mid X_m = s_m, \dots, X_0 = s_0) \cdot \mathbb{P}(\forall i \in C \colon X_i = s_i \mid \forall i \in B \colon X_i = s_i) \\
&= \mathbb{P}(X_n = s_n \mid X_m = s_m) \cdot \sum \mathbb{P}(\forall i \in C \colon X_i = s_i \mid \forall i \in B \colon X_i = s_i) \\
&= \mathbb{P}(X_n = s_n \mid X_m = s_m)
\end{aligned}$$

A consequence of this is that if $\{X_n\}_{n\geq 0}$ is a Markov chain and $\{a_n\}_{n\geq 0}$ is strictly monotonic then $Y_n = X_{a_n}$ is also a Markov chain. After all if we let $B = \{a_{n-1}, \ldots, a_0\}$ then $\max B = a_{n-1}$ and so

$$\begin{aligned}\mathbb{P}(Y_n = s_{a_n} \mid Y_{n-1} = s_{a_{n-1}}, \ldots, Y_0 = s_{a_0}) = \mathbb{P}(X_{a_n} = s_{a_n} \mid \forall i \in B \colon X_i = s_i) &= \mathbb{P}(X_{a_n} = s_{a_n} \mid X_{a_{n-1}} = s_{a_{n-1}}) \\ &= \mathbb{P}(Y_n = s_{a_n} \mid Y_{n-1} = s_{a_{n-1}})\end{aligned}$$

as required.

2.0.2 Definition

For a Markov chain $\{X_n\}_{n\geq 0}$ on a finite set of states S, we define the **adjacency matrix** at the nth measurement by

$$P_{ij}^{(n)} = \mathbb{P}(X_n = j \mid X_{n-1} = i)$$

for $i, j \in S$. This is also sometimes written as $P_n(i \to j)$ (the probability measuring i on the $n-1$th measurement gives j on the next). If $P^{(n)}$ is the same for all n, then we say that the chain is **homogeneus in time**, and we generally write P in place of $P^{(n)}$.

For example, suppose a frog is hopping between N leaves. The frog can hopping from every leaf to every other leaf, and it always chooses a leaf in an independent and uniform manner. This defines a Markov chain where the states are the leaves, and X_n is the leaf the frog is on after n hops. This Markov chain is even homogeneus since the frog makes its choices in a manner which does not take the current number of hops into account. The adjacency matrix is defined by

$$P_{ij} = \begin{cases} \frac{1}{N-1} & i \neq j \\ 0 & i = j \end{cases}$$

This is the simple random process on the complete graph of N vertices, K_N.

Suppose $N = 4$, and supppose that at the beginning the frog is on either the first or second leaf with equal probability. What is the probability that after one hop the frog is on the fourth leaf? The following notation will be used: $X \sim (a_0, \ldots, a_n)$ will be used to mean $\mathbb{P}(X = s_i) = a_i$, where s_i is some understood ordering of the set of states S. Then

$$\mathbb{P}\left(X_1 = j \,\middle|\, X_0 \sim \left(\frac{1}{2}, \frac{1}{2}, 0, 0\right)\right) = \mathbb{P}(X_1 = j \mid X_0 = 1) \cdot \frac{1}{2} + \mathbb{P}(X_1 = j \mid X_0 = 2) \cdot \frac{1}{2}$$

as the rest of the terms are zero. For $j = 4$ we get that this is equal to $\frac{1}{3}$. Notice that we can generalize this and get

$$\mathbb{P}(X_{n+1} = j \mid X_n \sim \vec{v}) = \sum_{i\in S} \mathbb{P}(X_{n+1} = j \mid X_n = i) \cdot \mathbb{P}(X_n = i) = \sum_{i\in S} P_{ij}^{(n+1)} \vec{v}_i = (\vec{v} \cdot P^{(n+1)})_j$$

So we have proven the following:

2.0.3 Proposition

If $X_n \sim \vec{v}$ then $X_{n+1}|X_n \sim \vec{v} \cdot P^{(n+1)}$, and so $X_n|X_0 \sim \vec{v} \cdot P^{(n)} \cdots P^{(1)}$. In particular if the Markov chain is homogeneus, $X_n|X_0 \sim \vec{v} \cdot P^n$.

This simplifies dealing with Markov chains, especially homogeneus ones.

2.0.4 Example

Suppose $\{Y_n\}_{n=1}^{\infty}$ is a sequence of random variables which have the distribution $Y_n \sim \mathrm{Ber}\left(\frac{1}{n}\right)$ (recall that $X \sim \mathrm{Ber}(p)$ means that X is 1 with probability p and zero otherwise). And we define $X_n = \chi\{(\exists m \leq n)\, Y_m = 1\}$, the indicator of the set of all values such that there is an index before n where $Y_m = 1$ (χ_S is the *indicator function* of the set S, defined by $\chi_S(x) = 1$ for $x \in S$ and zero otherwise). We will prove X_n is a Markov chain. Notice that

$$X_n = \chi\{(\exists m \leq n)\, Y_m = 1\} = \chi\{(\exists m \leq n-1)\, Y_m = 1\} \vee \chi\{Y_n = 1\} = X_{n-1} \vee \chi\{Y_n = 1\}$$

$\vee$ is bitwise or, or equivalently the maximum. And therefore we get that $X_n = \bigvee_{i=1}^{n} \chi\{Y_i = 1\}$. This means that if $X_{n-1} = 1$ then $X_n = 1$, and if $X_{n-1} = 0$ then $X_n = 1$ if and only if $Y_n = 1$. And so X_n's value depends only on X_{n-1}'s and not any previous X_i. So $\{X_n\}_{n=1}^{\infty}$ is indeed a Markov chain.

Notice that

$$\mathbb{P}(X_n = 0 \mid X_{n-1} = 0) = \mathbb{P}(Y_n = 0) = \frac{n-1}{n}, \quad \mathbb{P}(X_n = 1 \mid X_{n-1} = 0) = \mathbb{P}(Y_n = 1) = \frac{1}{n},$$

$$\mathbb{P}(X_n = 0 \mid X_{n-1} = 1) = 0, \quad \mathbb{P}(X_n = 1 \mid X_{n-1} = 1) = 1$$

And so we get that

$$P^{(n)} = \begin{pmatrix} \frac{n-1}{n} & \frac{1}{n} \\ 0 & 1 \end{pmatrix}$$

2.0.5 Definition

A real $n \times n$ matrix P such that $P_{ij} \geq 0$ for every i, j, and for every row i we have $\sum_{j=1}^{n} P_{ij} = 1$ then P is called an **stochastic matrix**.

Notice that we can draw a diagram for every stochastic matrix and it will be the transition matrix of a Markov chain. Meaning every stochastic matrix is the transition matrix of some Markov chain, and every transition matrix is stochastic. Notice that the second condition for a matrix to be stochastic can be written as $P\mathbf{1} = \mathbf{1}$ where $\mathbf{1} = (1, \dots, 1)^\top$.

2.0.6 Definition

Let $\{X_n\}_{n\geq 0}$ be a Markov chain over a state space S, and let $A \subseteq S$. Then we define the **hitting time** to A to be the random variable

$$T_A = \min\{t \geq 1 \mid X_t \in A\}$$

Note that if X_t is never in A then T_A can be ∞, and so T_A is a function from the probability space to the extended reals: $\Omega \longrightarrow \mathbb{R} \cup \{\infty\}$. This means that $T_A^{-1}\{\infty\}$ must also be measurable (an event).

In the case that A is a singleton $A = \{a\}$ then we write T_a in place of T_A. Notice that T_A measures starting from $t = 1$, while it is possible that the initial condition is in A, ie. $X_0 \in A$. So in the case that $X_0 \in A$, T_A measures the *return time* to A, in particular if $X_0 \sim \delta_a$ where $\delta_a = (0, \dots, 1, \dots, 0)$ (1 is at the index corresponding to the state a). We also use the following notation

$$\mathbb{P}_V(E) = \mathbb{P}(E \mid X_0 \sim V), \qquad \mathbb{P}_{\delta_a}(E) = \mathbb{P}_a(E) = \mathbb{P}(E \mid X_0 = a)$$

If P is the transition matrix of a homogeneus Markov chain, then $P^n(a \to b)$ means $P^n_{ba} = \mathbb{P}(X_n = b \mid X_0 = a)$.

2.0.7 Lemma

If $\{X_n\}$ is a homogeneus Markov chain, then

$$P^n(a \to b) = \sum_{m=1}^{n} \mathbb{P}_a(T_b = m) P^{n-m}(b \to b)$$

$$\begin{aligned} P^n(a \to b) = \mathbb{P}_a(X_n = b) = \mathbb{P}\left(\dot{\bigcup_{m=1}^{n}} \{T_b = m\}, X_n = b \,\middle|\, X_0 = b \right) &= \sum_{m=1} \mathbb{P}(T_b = m, X_n = b \mid X_0 = b) \\ &= \sum_{m=1}^{n} \mathbb{P}(X_n = b \mid T_b = m, X_0 = a) \cdot \mathbb{P}(T_b = m \mid X_0 = a) \end{aligned}$$

Now, $\mathbb{P}(X_n = b \mid T_b = m, X_0 = a) = \mathbb{P}(X_n = b \mid X_m = b, X_{m-1} \neq b, \dots, X_1 \neq b, X_0 = a) = \mathbb{P}(X_n = b \mid X_m = b)$ by the Markov property. Since $\{X_n\}$ is homogeneus this is just equal to $P^{n-m}(b \to b)$. Thus this formula is equal to

$$\sum_{m=1}^{b} \mathbb{P}(X_n = b \mid X_m = b) \cdot \mathbb{P}_a(T_b = m) = \sum_{m=1}^{b} P^{n-m}(b \to b) \cdot \mathbb{P}_a(T_b = m) \qquad \blacksquare$$

Let us introduce some more notation:

$$f_{a \to b} = \mathbb{P}(T_b < \infty \mid X_0 = a), \qquad f_{a \to a} = f_a = \mathbb{P}(T_a < \infty \mid X_0 = a)$$

thus $f_{a \to b}$ is the probability that if we start at a, we eventually reach b.

2.0.8 Lemma

$f_{a\to c} \geq f_{a\to b} \cdot f_{b\to c}$

Notice that $\{T_c < \infty\} = \{(\exists t > 0)X_t = c\} \supseteq \dot\bigcup_{k>0}\{T_b = k, (\exists t > k)X_t = c\}$. Thus we get

$$\begin{aligned}
f_{a\to c} = \mathbb{P}(T_c < \infty \mid X_0 = a) &\geq \sum_{k=1}^{\infty} \mathbb{P}(T_b = k, (\exists t > k)X_t = c \mid X_0 = a) \\
&= \sum_{k=1}^{\infty} \mathbb{P}(T_b = k \mid X_0 = a) \cdot \mathbb{P}((\exists t > k)X_t = c \mid T_b = k, X_0 = a) \\
&= \sum_{k=1}^{\infty} \mathbb{P}(T_b = k \mid X_0 = a) \cdot \mathbb{P}((\exists t > k)X_t = c \mid X_k = b, X_{k-1} \neq b, \dots, X_1 \neq b, X_0 = a) \\
\text{(Markov property)} \quad &= \sum_{k=1}^{\infty} \mathbb{P}(T_b = k \mid X_0 = a) \cdot \mathbb{P}((\exists t > k)X_t = c \mid X_k = b) \\
\text{(homogeneity)} \quad &= \sum_{k=1}^{\infty} \mathbb{P}(T_b = k \mid X_0 = a) \cdot \mathbb{P}((\exists t > 0)X_t = c \mid X_0 = b) \\
&= \sum_{k=1}^{\infty} \mathbb{P}(T_b = k \mid X_0 = a) \cdot f_{b\to c} = f_{a\to b} \cdot f_{b\to c} \qquad \blacksquare
\end{aligned}$$

In particular this means

$$f_a \geq f_{a\to b} \cdot f_{b\to a}$$

For every $a \in S$ we define the random variable $N(a) = \sum_{n=1}^{\infty} \chi\{X_n = a\}$, which is the number of times the state a is visited from time 1 and onward. When $X_0 \sim V$ we write $N_V(a)$. Notice then that $f_{a\to b} = \mathbb{P}(N(b) \geq 1 \mid X_0 = a)$ and so $f_a = \mathbb{P}(N(a) \geq 1 \mid X_0 = a)$.

2.0.9 Proposition

$\mathbb{P}(N(a) \geq k \mid X_0 = a) = f_a^k$

We prove this by induction, for $k = 1$ this is simply what we just said. Now

$$\begin{aligned}
\mathbb{P}(N(a) \geq k+1 \mid X_0 = a) &= \sum_{m=1}^{\infty} \mathbb{P}(T_a = m, |\{j > m \mid X_j = a\}| \geq k \mid X_0 = a) \\
\text{(Markov property)} \quad &= \sum_{m=1}^{\infty} \mathbb{P}(T_a = m \mid X_0 = a) \cdot \mathbb{P}(|\{j > m \mid X_j = a\}| \geq k \mid X_m = a) \\
\text{(homogeneity)} \quad &= \sum_{m=1}^{\infty} \mathbb{P}(T_a = m \mid X_0 = a) \cdot \mathbb{P}_a(N(a) \geq k) \\
\text{(induction)} \quad &= f_a^k \sum_{m=1}^{\infty} \mathbb{P}(T_a = m \mid X_0 = a) = f_a^{k+1} \qquad \blacksquare
\end{aligned}$$

Notice then that

$$\mathbb{P}(N(a) = k \mid X_0 = a) = \mathbb{P}_a(N(a) \geq k) - \mathbb{P}_a(N(a) \geq k+1) = f_a^k - f_a^{k+1} = f_a^k(1 - f_a)$$

Thus $N_a(a) \sim \text{Geo}(1 - f_a) - 1$ (the $+1$ is since $X \sim \text{Geo}(p)$ means $\mathbb{P}(X = k) = p(1-p)^{k-1}$). Thus

$$\mathbb{E}[N_a(a)] = \frac{1}{1 - f_a} - 1 = \frac{f_a}{1 - f_a}$$

2.0.10 Definition

A state $b \in S$ is **recurrent** if $f_b = 1$, equivalently if $\mathbb{P}_b(T_b < \infty)$ (the probability of returning to b is 1). A non-recurrent state is called **transient**. b is **absorbing** if $P(b \to b) = 1$.

Notice that if b is recurrent then if $f_b = 1$, $N_b(b) \sim \text{Geo}(0) - 1$, meaning $\mathbb{P}_b(N(b) = \infty) = 1$. And if b is transient then $N_b(b)$ is a finite geometric variable and so $\mathbb{P}_b(N(b) < \infty) = 1$. And so

$$\begin{aligned} b \text{ is recurrent} &\iff \mathbb{P}(N(b) = \infty \mid X_0 = b) = 1, \\ b \text{ is transient} &\iff \mathbb{P}(N(b) < \infty \mid X_0 = b) = 1 \iff \mathbb{P}(N(b) < \infty \mid X_0 \sim v) = 1 \end{aligned}$$

2.0.11 Definition

Let $a, b \in S$ be states. Then b is **reachable** from a if $f_{a \to b} \neq 0$ or $a = b$, this is denoted $a \to b$. a and b are **connected** if both $a \to b$ and $b \to a$, this is denoted $a \leftrightarrow b$.

This means that $a \to b$ if and only if there exists some $n \geq 0$ such that $P^n(a \to b) > 0$. Furthermore, connectivity is an equivalence relation: it is obviously reflexive and symmetric and if $a \to b$ and $b \to c$, since $f_{a \to c} \geq f_{a \to b} \cdot f_{b \to c} > 0$, we get that reachability and therefore connectivity is transitive. Thus S can be partitioned into *connectivity classes*.

2.0.12 Lemma

If $a \to b$ and $a \neq b$ then $\mathbb{P}(T_b < T_a \mid X_0 = a) > 0$.

Since $a \to b$, there exists a sequence of states $a = s_0, \dots, s_m = b$ such that $P_{s_i s_{i+1}} > 0$ for all i. We can assume that for every $i > 0$, $a \neq s_i$. So we have a sequence whose probability is positive and where the hitting time of b is before that of a, so the probability that $T_b < T_a$ must be positive. ■

2.0.13 Definition

$A \subseteq S$ is **closed** if for every $a \in A$ and every $b \notin A$, b is not reachable from a. A is also called **irreducible** if it is closed and connected.

2.0.14 Theorem

If a is recurrent and $a \to b$, then also $b \to a$ and b is recurrent.

We know

$$f_{a \to b} = \mathbb{P}_a(T_a > T_b) + \mathbb{P}_a(T_a < T_b) \cdot \mathbb{P}(T_b < \infty \mid T_a < T_b)$$

by the above lemma $p = \mathbb{P}_a(T_b < T_a) > 0$ and so by homogeneity

$$= p + (1 - p) \cdot \mathbb{P}(T_b < \infty \mid X_0 = a) = p + (1 - p) f_{a \to b}$$

Thus we get that $p \cdot f_{a \to b} = p$ and since $p \neq 0$, $f_{a \to b} = 1$. Now

$$f_{a \to b}(1 - f_{b \to a}) = \mathbb{P}(X_n \text{ hits } b \text{ and never returns to } a \mid X_0 = a) \leq \mathbb{P}_a(N(a) < \infty) = 0$$

Thus $f_{b \to a} = 1$. Now $f_b \geq f_{b \to a} \cdot f_{a \to b} = 1$ so b is also recurrent. ■

So if $a \leftrightarrow b$, then a is recurrent if and only if b is. If b is reachable from a but a is not reachable from b, then a is transient. And if a is recurrent and $a \to b$ then $\mathbb{P}_b(N(a) = \infty) = 1$.

2.0.15 Theorem

A finite closed set of states $A \subseteq S$ contains a recurrent state.

Suppose A has only transient states. This means that $\mathbb{P}_v(N(a) < \infty) = 1$ for every $a \in A$, and so we get that $\mathbb{P}_v((\forall a \in A) N(a) < \infty) = 1$ (as the intersection of a countable number of events with probability one). And this means $\mathbb{P}_v\left(\sum_{a \in A} N(a) < \infty\right) = 1$ since A is finite. But since A is closed, we can never leave A and so if v's support is in A then $\sum_{a \in A} N_v(a) = \infty$. ■

In particular, since S is closed, if S is finite it contains a recurrent state.

2.0.16 Theorem

If S is a finite state space, then it can be uniquely partitioned into

$$S = T \uplus C_1 \uplus \cdots \uplus C_k$$

where T is the set of all transient states, and C_i are all disjoint irreducible (closed and connected) sets.

So T is the set of all transient states, and for every recurrent state $a \in S \setminus T$ let $C_a = \{b \mid a \to b\}$. By a previous theorem, for every $b \in C_a$, $b \to a$ so and if $b \to b'$ then $a \to b'$ meaning $b' \in C_a$, so C_a is closed. And if $b, b' \in C_a$ then $a \to b$ and $a \to b' \implies b' \to a$ and therefore $b' \to b$, so C_a is connected and therefore irreducible. By taking representatives of each C_a, let $C_i = C_{a_i}$, we get the partition.

This partition is unique: since if $C_1 \uplus \cdots \uplus C_k = C'_1 \uplus \cdots \uplus C'_m$ let $a \in C_1$ then $a \in C'_i$ for some i, without loss of generality assume $a \in C'_1$. Then for every $b \in C_1$, since C_1 is connected $a \to b$ and so $b \in C'_1$ since C'_1 is closed, thus $C_1 = C'_1$. Continuing inductively we get $k = m$ and $C_i = C'_i$ as required. ■

2.0.17 Example

Suppose Elise is in a room 0, and can either stay in the room with probability $1 - p_1 - p_2$, go to room 1 with probability p_1 or go to room 2 with probability p_2. If she goes to a new room, she stays there forever. Knowing that ends up in room 2, what is the expected amount of time she spends waiting in room 0?

So we want to find the expected value of $N_0(0)$ knowing that $T_2 < \infty$. So we will compute

$$\mathbb{P}(N_0(0) = k \mid T_2 < \infty) = \frac{\mathbb{P}(N_0(0) = k, T_2 < \infty)}{\mathbb{P}(T_2 < \infty)} = \frac{\mathbb{P}(X_1 = \cdots = X_k = 0, X_{k+1} = 2)}{\mathbb{P}(T_2 < \infty)}$$

Now, utilizing conditional probability and the Markov property (this is all done under the assumption $X_0 = 0$),

$$\mathbb{P}(X_1 = \cdots = X_k = 0, X_{k+1} = 2) = \mathbb{P}(X_{k+1} = 2 \mid X_k = 0) \cdot \mathbb{P}(X_k = 0 \mid X_{k-1} = 0) \cdots \mathbb{P}(X_1 = 0) = p_2 \cdot (1 - p_1 - p_2)^k$$

And $\mathbb{P}(T_2 < \infty) = \frac{p_2}{p_1 + p_2}$ since to get to room 2 we must visit room 0 an arbitrary number of times, and then go to room 2, so

$$\mathbb{P}(T_2 < \infty) = \sum_{n=0}^{\infty} p_2 \cdot (1 - p_1 - p_2)^n = \frac{p_2}{p_1 + p_2}$$

Thus

$$\mathbb{P}(N_0(0) = k \mid T_2 < \infty) = (p_1 + p_2) \cdot (1 - p_1 - p_2)^k$$

Which means that

$$(N_0(0) \,|\, T_2 < \infty) \sim \mathrm{Geo}(p_1 + p_2) - 1 \implies \mathbb{E}[N_0(0) \mid T_2 < \infty] = \frac{1 - p_1 - p_2}{p_1 + p_2}$$

Notice two things: firstly, by symmetry this means that $(N_0(0) \,|\, T_1 < \infty) \sim \mathrm{Geo}(p_1 + p_2) - 1$ which is the same distribution. And secondly, this is the same distribution as $N_0(0)$, so the expected time Elise waits at room 0 does not change if we know which room she ends up in.

2.0.18 Definition

Let $a \in S \setminus T$ be a recurrent state, then we define its **period** to be

$$d(a) = \gcd\{n \geq 1 \mid P^n(a \to a) > 0\}$$

An irreducible Markov chain is called **periodic** if every state is recurrent and has the same period greater than 1, which is the **period** of the Markov chain.

Notice that if $P(a \to a) > 0$ then $d(a) = 1$, and so a periodic chain can be made aperiodic by adding a self-edge whose probability is nonzero.

2.0.19 Proposition

If P is the transition matrix of some periodic chain with a period of d, then P^d is reducible.

2.0.20 Proposition

If the Markov chain is irreducible then every state has the same period.

Let P be the transition matrix of the chain. Since $x \leftrightarrow y$, there exist natural r, ℓ such that $P^r(x,y), P^\ell(y,x) > 0$. So let $m = r + \ell$ and so

$$P^m(x,x) \geq P^r(x,y) \cdot P^\ell(y,x) > 0, \qquad P^m(y,y) \geq P^\ell(y,x) \cdot P^r(x,y) > 0$$

So let $\tau(a) = \{n \geq 1 \mid P^n(a,a) > 0\}$, and by above we have shown that $m \in \tau(x) \cap \tau(y)$. Now for every $n \in \tau(x)$ we have that $P^{\ell+n+r}(y,y) \geq P^\ell(y,x)P^n(x,x)P^r(x,y) > 0$ and so $n + m \in \tau(y)$. Thus $m + \tau(x) \subseteq \tau(y)$. By definition we have $d(y) = \gcd(\tau(y))$ and since $m \in \tau(y)$ we have $d(y)|m$ and since $m + \tau(x) \subseteq \tau(y)$ we must have that $d(y)|\tau(x)$. Thus $d(y)|d(x)$, and since x, y are arbitrary we get $d(x)|d(y)$ and so $d(x) = d(y)$ as required. ■

This means that every irreducible Markov chain has a period, and if the period is > 1, it is periodic. So in order for an irreducible Markov chain to be periodic, it is sufficient for there to exist a state a with $d(a) > 1$.

A common Markov chain is a random walk on $\mathbb{Z}$, where

$$P(i, i+1) = p, \quad P(i, i-1) = 1 - p, \quad P(i,j) = 0 \text{ for } j \notin \{i \pm 1\}$$

Another way of representing X_n is by $X_n = \sum_{k=1}^n B_k$ where $B_k = 1$ with probability p and $B_k = -1$ with probability $1 - p$. $\{B_k\}$ is independent. If $p = \frac{1}{2}$, the walk is called *fair*.

2.0.21 Theorem

If $p \neq \frac{1}{2}$, every state in $\mathbb{Z}$ is transient.

Since all the states are connected, it is sufficient to show that 0 is transient. So we set $X_0 = 0$ and notice that $\frac{B_k+1}{2} \sim \text{Ber}(p)$ and thus $\frac{X_n+n}{2} \sim \text{Bin}(n,p)$ thus

$$\mathbb{P}(X_{2n} = 0) = \mathbb{P}\left(\frac{X_{2n} + 2n}{2} = n\right) = \binom{2n}{n} p^n (1-p)^n$$

and $\mathbb{P}(X_{2n+1} = 0) = \mathbb{P}\left(\frac{X_{2n+1}+2n+1}{2} = n + \frac{1}{2}\right) = 0$ since binomial distributions take on only integer values. By Stirling's approximation: $k! \in \Theta(k^{k+1/2}e^{-k})$, we get that there exists some $c > 0$ such that

$$\mathbb{P}(X_{2n} = 0) = \frac{(2n)!}{n!n!} p^n (1-p)^n \leq cp^n(1-p)^n \frac{(2n)^{2n+1/2}e^{-2n}}{n^{2n+1}e^{-2n}} = cp^n(1-p)^n \frac{2^{2n+1/2}}{\sqrt{n}} = c' \frac{\big(4p(1-p)\big)^n}{\sqrt{n}}$$

This can be bound by a q^n where $q \in [0,1)$, since $4p(1-p) < 1$ for $p \neq \frac{1}{2}$. Thus we get that $\sum_{k=1}^\infty \mathbb{P}(X_k = 0 \mid X_0 = 0)$ and so by Borel-Cantelli we then get that $\mathbb{P}(X_k = 0 \text{ i.o.} \mid X_0 = 0) = 0$, meaning that the probability $X_k = 0$ an infinite number of times is zero. Thus $\mathbb{P}(N(0) = \infty \mid X_0 = 0) = 0$, and so this means 0 is transient as required. ■

If we have a Markov chain, and $A \subseteq S$, we can ask questions about hitting times in A by removing all the states in A and adding a new state $\hat{A}$. This can only be done if for every $a, a' \in A$ and $b \notin A$, $P(a \to b) = P(a' \to b)$, and we define that the probability $P(\hat{A} \to b) = P(a \to b)$. And $P(b \to \hat{A}) = \sum_{a \in A} P(b \to a)$. In particular this can be done if A is closed.

2.0.22 Example

Suppose we have the following Markov chain:

$$P = \begin{pmatrix} 1 & 0 & 0 \\ p & q & r \\ 0 & 0 & 1 \end{pmatrix}$$

where $p, q, r \geq 0$ and $p + q + r = 1$. If we know that $X_0 = 2$, what is the probability that the chain will be absorbed into 1 or 3?

Let us define

$$\ell_j = \mathbb{P}(T_1 < \infty \mid X_0 = j)$$

since 1 and 3 are absorbing states, $\ell_1 = 1$ and $\ell_3 = 0$. Now, we want to compute ℓ_2:

$$\ell_2 = \mathbb{P}(T_1 < \infty \mid X_0 = 2) = \sum_{j=1}^{3} \mathbb{P}(T_1 < \infty \mid X_1 = j, X_0 = 2) \cdot \mathbb{P}(X_1 = j \mid X_0 = 2)$$
$$= \sum_{j=1}^{3} \mathbb{P}(T_1 < \infty \mid X_1 = j) \cdot P_{2j}$$

where the last step is due to homogeneity. This is equal to $\sum_{j=1}^{3} \ell_j P_{2j} = \ell_1 p + \ell_2 q + \ell_3 r = p + \ell_2 r$. Thus we get that $\ell_2 = p + \ell_2 r$ and so $\ell_2 = \frac{p}{1-r}$. Thus the probability that starting from $X_0 = 2$ we are absorbed into 1 (meaning $T_1 = \infty$) is $1 - \ell_2 = \frac{q}{1-r}$. Since 2 is transient, we are either absorbed into 1 or 3, so the probability of being absorbed into 3 is $\frac{p}{1-r}$.

Let us now ask what the expected time until being absorbed is. By the law of total expectation: Now, $\mathbb{E}\left[T_{\{1,3\}} \mid X_1 = 2\right] = \mathbb{E}\left[T_{\{1,3\}} \mid X_0 = 2\right] + 1$ since it takes one more step, and so

$$= \left(1 + \mathbb{E}\left[T_{\{1,3\}} \mid X_0 = 2\right]\right) \cdot \mathbb{P}_2(X_1 = 2) + \mathbb{P}_2(X_1 = 1) + \mathbb{P}_2(X_1 = 3)$$

So let $x = \mathbb{E}\left[T_{\{1,3\}} \mid X_0 = 2\right]$, we get

$$x = (1 + x)r + p + q = (1 + x)r + (1 - r) \implies x = \frac{1}{1 - r}$$

2.0.23 Example

Suppose we have the following Markov chain:

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ a_1 & a_2 & 0 & a_4 & 0 & 0 \\ 0 & 0 & b_3 & 0 & b_5 & b_6 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

What is the probability of being absorbed into one of the absorbing states $(1, 2, 5, 6)$ if it starts on one of the non-absorbing states $(3, 4)$?

Let us define $\ell_{m,k} = \mathbb{P}_m(T_k < \infty)$. Now, let us notice that

$$\ell_{m,k} = \mathbb{P}(T_k < \infty \mid X_0 = m) = \sum_{j=1}^{6} \mathbb{P}(T_k < \infty \mid X_1 = j) \cdot \mathbb{P}(X_1 = j \mid X_0 = m) = \sum_{j=1}^{6} P_{mj} \ell_{jk}$$

So if we define $L_{ij} = \ell_{ij}$ then we get that $L = PL$ and we can solve for L.

What is the expected time until being absorbed? We can consolidate $A = \{1, 2, 5, 6\}$ to a state we will call 1, then the new transition matrix is

$$P' = \begin{pmatrix} 1 & 0 & 0 \\ a_1 + a_2 & 0 & a_4 \\ b_5 + b_6 & b_3 & 0 \end{pmatrix}$$

Now let us define $r_j = \mathbb{E}[T_1 \mid X_0 = j]$, then we get

$$r_j = \sum_{i=1}^{3} \mathbb{E}[T_1 \mid X_1 = i] P_{ji} = P_{j1} + \sum_{i=2}^{3} (r_i + 1) P_{ji} = P_{j1} + P_{j2} + P_{j3} + r_2 P_{j2} + r_3 P_{j3}$$

Which is a linear system of equations which can be solved.

2.1 Stationary Distributions and the Convergence of Markov Chains

2.1.1 Definition

Suppose $|S| = N$, then a **stationary distribution** of P is a row vector π which represents a distribution (meaning $\pi_i \geq 0$ and $\sum_{i=1}^N \pi_i = 1$) such that $\pi = \pi P$.

A stationary distribution is an eigenvector (or the transpose of one) of $P^\top$ whose eigenvalue is 1. If π is a stationary distribution, then $\pi P = \pi \implies \pi P^n = \pi$ for every $n \geq 0$. This means that if $X_0 \sim \pi$ then $X_n \sim \pi$ for every n (since $\mathbb{P}(X_n = k \mid X_0 \sim \pi) = (\pi P^n)_k = \pi_k$).

For example if $G = (V, E)$ is an undirected graph where $|V| = N$ and the transitions from each state are all uniform (meaning $\mathbb{P}(X_n = v \mid X_{n-1} = u) = \frac{1}{\deg(u)}$ if $v \leftrightarrow u$), then let

$$\tilde{\pi} = (\deg(v_1), \dots, \deg(v_N))$$

Then (using the notation $\delta\varphi$ which is 1 if φ is true and 0 otherwise) we have that $P_{xy} = \frac{1}{\deg(x)}\delta(x \leftrightarrow y)$, so

$$(\tilde{\pi}P)_y = \sum_{x\in V} \tilde{\pi}_x P_{xy} = \sum_{x\in V} \deg(x)\frac{1}{\deg(x)}\delta(x \leftrightarrow y) = \sum_{x\in V}\delta(x \leftrightarrow y) = \deg(y) = \tilde{\pi}_y$$

So $\tilde{\pi}$ is a non-negative row vector, but it must be normalized to become a distribution, so we define

$$\pi_v = \frac{\deg(v)}{\sum_{u\in V}\deg(u)} = \frac{\deg(v)}{2|E|}$$

If the degree of each vertex is constant, suppose $\deg(v) = d$ for all $v \in V$, then $\pi_v = \frac{d}{dN} = \frac{1}{N}$ so π is a uniform distribution.

2.1.2 Theorem (Existence and Uniqueness Theorem)

Let P be the transition matrix of irreducible finite-state Markov chain, then there exists a unique stationary distribution π for P.

We know that $P\mathbf{1} = \mathbf{1}$ and so 1 is an eigenvalue for P, and since P and $P^\top$ are similar, they share eigenvalues. Thus $P^\top$ has an eigenvalue of 1 and therefore must have a stationary distribution. To show that this eigenvector is unique, we will show that the column eigenspace of P has a dimension of one, and since the eigenspaces of a matrix and its transpose are equal (think Jordan normal forms), this is sufficient. So we will show that if $h \in \mathbb{R}^N$ is an eigenvector of P with an eigenvalue of 1, it is of the form $h = (c, \dots, c)^\top$. Because S is finite, there exists a state $a \in S$ such that $h_a = M$ is maximal. Now suppose there exists a $z \in S$ such that $h_z < M$ and $P_{az} > 0$ then

$$h_a = (Ph)_a = \sum_{y\in S} P_{ay}h_y = P_{az}h_z + \sum_{y\neq z} P_{ay}h_y < M\left(\sum_{y\in S} P_{ay}\right) = M = h_a$$

since $P_{az} > 0$ and $h_z < M$, and this is a contradiction. So for every state where $P_{az} > 0$, $h_z = M$. If we continue this proof (since $P^n h = h$), we get that if $a \to z$ then $h_z = M$. Since the Markov chain is irreducible, it is closed and therefore $h_z = M$ for every $z \in S$. ■

Notice that the proof of existence here assumes nothing about S other than it being finite. But in the case that the chain is irreducible, we can also provide a constructive proof of the existence of a stationary distribution. But first, a lemma:

2.1.3 Lemma

For every two states $x, y \in S$ in a finite irreducible state space $\mathbb{E}_x[T_y] < \infty$.

Since S is irreducible and finite, there exists an $\varepsilon > 0$ and a $r \in \mathbb{N}$ such that for every $a, b \in S$, there exists a $j \leq r$ such that $P^j(a, b) > \varepsilon$. This is since S is connected and so between every two states there exists a path of length $\leq r$ (taking the maximum length of all paths, or just N) and so $P^j(a, b) > 0$. Take ε to be less than the minimum of all such $P^j(a, b)$, which we can do since S is finite.

Thus

$$\mathbb{P}((\exists m \in [0, \dots, r])X_m = b \mid X_n = a) > \varepsilon$$

Now we know that $T_b > kr$ if and only if $X_0, \ldots, X_r \neq b$ and then we don't hit b for another $(k-1)r$ rounds, meaning $T_b > (k-1)r$. By homogeneity this means

$$\begin{aligned}\mathbb{P}(T_b > kr \mid X_0 = a) \leq \max_{a'} \mathbb{P}(T_b > (k-1)r \mid X_0 = a')\, \mathbb{P}((\forall m \in [0,r]) X_m \neq b \mid X_0 = a) \\ \leq \max_{a'} \mathbb{P}(T_b > (k-1)r \mid X_0 = a') \cdot (1-\varepsilon)\end{aligned}$$

and so by induction, this is $\leq (1-\varepsilon)^k$. Thus

$$\mathbb{E}[T_b \mid X_0 = a] = \sum_{n=0}^{\infty} \mathbb{P}(T_b > n \mid X_0 = a) \leq r \sum_{k=0}^{\infty} \mathbb{P}(T_b > kr \mid X_0 = a) \leq r \sum_{k=0}^{\infty} (1-\varepsilon)^k < \infty$$

The first inequality is due to the series being decreasing, and so we can take a summand and copy it r times, then take the rth next. ■

Now we can construct a stationary distribution. Let us define

$$\tilde{\pi}_y = \mathbb{E}_{z_0}\left[\begin{array}{c}\text{the number of times } y \text{ is visited,} \\ \text{including at time 0,} \\ \text{before returning to } z_0\end{array}\right] = \sum_{n=0}^{\infty} \mathbb{P}(X_n = y, T_{z_0} > n \mid X_0 = z_0)$$

The last equality is since this probability is equal to the number of visits being $\geq n$. This is well-defined as

$$\tilde{\pi}_y \leq \sum_{n=0}^{\infty} \mathbb{P}(T_{z_0} > n \mid X_0 = z_0) = \mathbb{E}_{z_0}[T_{z_0}]$$

and this is finite by the above lemma, so $\tilde{\pi}_y < \infty$. Now we will compute $(\tilde{\pi}P)_y$:

$$\begin{aligned}(\tilde{\pi}P)_y &= \sum_{x \in S} \tilde{\pi}_x P_{xy} \\ &= \sum_{x \in S} \sum_{n=0}^{\infty} \mathbb{P}_{z_0}(X_n = x, T_{z_0} > n) P_{xy} \\ &= \sum_{n=0}^{\infty} \sum_{x \in S} \mathbb{P}_{z_0}(X_n = x, T_{z_0} \geq n+1)\, \mathbb{P}(X_{n+1} = y \mid X_n = x) \\ &= \sum_{n=0}^{\infty} \sum_{x \in S} \mathbb{P}_{z_0}(X_{n+1} = y, X_n = x, T_{z_0} \geq n+1) \\ &= \sum_{n=0}^{\infty} \mathbb{P}_{z_0}(X_{n+1} = y, T_{z_0} \geq n+1) \\ &= \sum_{k=1}^{\infty} \mathbb{P}_{z_0}(X_k = y, T_{z_0} \geq k) \\ &= \sum_{k=0}^{\infty} \mathbb{P}_{z_0}(X_k = y, T_{z_0} \geq k) + \sum_{k=0}^{\infty} \mathbb{P}_{z_0}(X_k = y, T_{z_0} = k) - \mathbb{P}_{z_0}(X_0 = y, T_{z_0} = 0) \\ &= \tilde{\pi}_y + \sum_{k=0}^{\infty} \mathbb{P}_{z_0}(X_k = y, T_{z_0} = k) - \delta(y = z_0)\end{aligned}$$

Notice that $X_k = y, T_{z_0} = k$ if and only if $T_{z_0} = k$ and $y = z_0$, and so the sum is equal to $\delta(y = z_0)$. So we get that $\tilde{\pi}P = \tilde{\pi}$ as required. So we just need to normalize it by

$$\sum_{x \in S} \tilde{\pi}_S = \mathbb{E}_{z_0}[T_{z_0}]$$

And thus the stationary distribution is

$$\pi_x = \frac{\tilde{\pi}_x}{\mathbb{E}_{z_0}[T_{z_0}]}$$

■

2.1.4 Corollary

If P is irreducible then $\pi_a = \frac{1}{\mathbb{E}_a[T_a]}$.

Since π is unique we can choose any z_0 and get the same result. So we can choose $z_0 = a$ and so

$$\pi_a = \frac{\mathbb{E}\left[\begin{matrix}\text{The number of times we visit } a\\ \text{before returning to } a\\ \text{including } t=0\end{matrix}\;\middle|\; X_0 = a\right]}{\mathbb{E}_a[T_a]}$$

The numerator here is obviously 1, and so $\pi_a = \frac{1}{\mathbb{E}_a[T_a]}$. ■

For example, we showed that for a connected graph where the degree of each vertex is d (a connected d-regular graph), $\pi_v = \frac{1}{N}$ where $N = |V|$. Thus since P is irreducible, we get that

$$\frac{1}{N} = \pi_v = \frac{1}{\mathbb{E}_a[T_a]} \implies \mathbb{E}_a[T_a] = N$$

This is independent of the structure of the graph. But importantly, T_a is dependent on the structure of the graph!

As another example, if P is symmetric then $\mathbf{1}^\top P = (P\mathbf{1})^\top = \mathbf{1}^\top$ and so $\frac{1}{N}\mathbf{1}$ is a stationary distribution of P. And thus $\mathbb{E}_a[T_a] = N$ where $N = |S|$.

2.1.5 Theorem

If $a \in S$ is a transient state and S is finite, then for every stationary distribution π, $\pi_a = 0$.

There are two cases we will consider: that a is connected to only transient states, and that there exists a recurrent state b such that $a \to b$. In the second case we have that $b \not\to a$ since a is transient and b is recurrent. Let $a_0 = a \to a_1 \to \cdots \to a_n \to b$ be the path from a to b, and we can assume that all a_i are transient (as otherwise we could set $b = a_i$ for the minimum i where a_i is recurrent). Let C be the connected component of b and π be a stationary distribution on all of S. Then

$$\sum_{z\in C}\pi_z = \sum_{z\in C}(\pi P)_z = \sum_{z\in C}\left(\sum_{y\in C}\pi_y P(y,z) + \sum_{y\notin C}\pi_y P(y,z)\right) = \sum_{y\in C}\pi_y\sum_{z\in C}P(y,z) + \sum_{z\in C}\sum_{y\notin C}\pi_y P(y,z)$$

Since C is closed and $y \in C$, we have that $\sum_{z\in C} P(y,z) = 1$ and thus we get that the left sum is $\sum_{y\in C}\pi_y$, and since the entire expression is equal to $\sum_{z\in C}\pi_z$, we must have that the right sum is zero. So for every $z \in C$ and $y \notin C$, $\pi_y P(y,z) = 0$.

This must be true in particular for $y = a_n$ and $z = b$, and since $P(a_n, b) > 0$ this means $\pi_{a_n} = 0$. And we claim inductively that $\pi_{a_k} = 0$, since

$$\pi_{a_k} = \sum_{y\in S}\pi_y P(y, a_k)$$

and so if $\pi_{a_k} = 0$ then $\pi_y P(y, a_k) = 0$ for all $y \in S$. Since $P(a_{k-1}, a_k) > 0$ this means $\pi_{a_{k-1}} = 0$. And so in particular we have that $\pi_a = \pi_{a_0} = 0$ as required. ■

Now suppose S is a finite state space, then it can be uniquely partitioned into

$$S = T \uplus C_1 \uplus \cdots \uplus C_n$$

where T is the set of all transient states, and C_i are irreducible components. We showed that for every stationary distribution π, for every $a \in T$ we have $\pi_a = 0$. And we also showed that for every $1 \le i \le n$ there exists a unique stationary distribution π_i whose support is C_i (meaning for every $a \notin C_i$, $\pi_i(a) = 0$). Thus a general stationary distribution is a normalized vector (meaning the sum of its coefficients is one) in $\operatorname{span}\{\pi_1, \ldots, \pi_n\}$. This is since the transition matrix P can be viewed as a block matrix over the partition of S.

For the next lemma, let us state a combinatorical fact: if $A \subseteq \mathbb{N}$ is closed under addition and has a greatest common divisor of 1, then $\mathbb{N} \setminus A$ is finite. This is trivial if $1 \in A$.

2.1.6 Lemma

Suppose P is the transition matrix of an irreducible, aperiodic, finite-state, homogeneus Markov chain. Then there exists an $r_0 > 0$ such that for all $r \geq r_0$ and $a, b \in S$, $P^r(a,b) > 0$.

Let us define as before $\tau(a) = \{n \geq 1 \mid P^n(a,a) > 0\}$. Since P is aperiodic, $d(a) = \gcd \tau(a) = 1$, and $\tau(a)$ is closed under addition since $P^{n+m}(a,a) \geq P^n(a,a)P^m(a,a)$. This means that $\mathbb{N} \setminus \tau(a)$ is finite. This means that $\bigcup_{a\in S}(\mathbb{N} \setminus \tau(a)) = \mathbb{N} \setminus \bigcap_{a\in S}\tau(a)$ is finite as well as the finite union of finite sets. Let t_0 be an upper bound for $\mathbb{N} \setminus \bigcap_{a\in S}\tau(a)$, so for every $t \geq t_0$ we have that $t \in \bigcap_{a\in S}\tau(a)$ meaning $P^t(a,a) > 0$ for all $a \in S$.

Since P is irreducible, for every $a, b \in S$ there exists an $n = n(a,b)$ such that $P^n(a,b) > 0$. Now n is bound by $|S|$ and therefore we can define $n_0 = \max_{a,b\in S} n(a,b)$ and so for every $r \geq t_0 + n_0$ we have that $r - n_0 \geq t_0$ and so $P^{r-n_0}(a,a) > 0$. Thus

$$P^r(a,b) \geq P^{r-n_0}(a,a)P^{n_0}(a,b) > 0$$

so $r_0 = t_0 + n_0$ satisfies the condition. ■

2.1.7 Lemma

Again suppose P is irreducible and aperiodic, and let π be its unique stationary distribution. Then there exists an $0 < \alpha < 1$ and a constant $c > 0$ such that for every $k \in \mathbb{N}$ and every distribution vector v,

$$\left\|vP^k - \pi\right\|_1 \leq c\alpha^k$$

where $\|\cdot\|_1$ is the 1-norm on $\mathbb{R}^n$: $\|u\|_1 = \sum_{k=1}^n |u_i|$.

By the previous lemma, there exists an $r > 0$ such that $P^r > 0$ (meaning every coefficient of P^r is positive). Since P is finite, there exists a $0 < \delta < 1$ such that for every $a, b \in S$: $P^r(a,b) \geq \delta\pi_b$. Let Π be the matrix whose rows are all π. Then let us define the matrix Q by

$$P^r = \delta\Pi + (1-\delta)Q$$

and since $P^r \geq \delta\Pi$ (pointwise), we have $Q \geq 0$ (pointwise). Now notice that Π is stochastic since $(\Pi\mathbf{1})_i = \pi\mathbf{1} = 1$, and so Q is also stochastic:

$$\mathbf{1} = P^r\mathbf{1} = \delta\mathbf{1} + (1-\delta)Q\mathbf{1} \implies (1-\delta)\mathbf{1} = (1-\delta)Q\mathbf{1}$$

and since $\delta < 1$, $1 - \delta \neq 0$. Let us define $\theta := 1 - \delta$ and we will prove by induction that for all $k \geq 1$,

$$P^{rk} = (1-\theta^k)\Pi + \theta^k Q^k$$

for $k = 1$ this is trivial. For the induction step,

$$P^{r(k+1)} = P^{rk}P^r = \left((1-\theta^k)\Pi + \theta^k Q^k\right)P^r = (1-\theta^k)\Pi P^r + \theta^k Q^k P^r$$

Since $\Pi P = \Pi$, we have that $\Pi P^r = \Pi$ and so this is equal to

$$= (1-\theta^k)\Pi + \theta^k\left((1-\theta)Q^k\Pi + \theta Q^{k+1}\right) = (1-\theta^k)\Pi + \theta^k(1-\theta)Q^k\Pi + \theta^{k+1}Q^{k+1}$$

Now since Q^k is stochastic and Π's columns are constant, $Q^k\Pi = \Pi$. And so this is equal to

$$= (1-\theta^k+\theta^k-\theta^{k+1})\Pi + \theta^{k+1}Q^{k+1} = (1-\theta^{k+1})\Pi + \theta^{k+1}Q^{k+1}$$

as required.

And so now we have for all $j \geq 0$, $P^{rk+j} = (1-\theta^k)\Pi + \theta^k Q^k P^j$ and so

$$P^{rk+j} - \Pi = \theta^k\left(Q^k P^j - \Pi\right)$$

Since Q^kP^j and Π are all stochastic matrices and thus their coefficients all are bound by 1, the coefficients of $Q^kP^j - \Pi$ all have an absolute value bound by 1 as well. Now since $(v\Pi)_i$ is equal to v times the ith column of Π which is $\pi_i\mathbf{1}$, we have $(v\Pi)_i = \pi_i v\mathbf{1} = \pi_i$ (since v is a distribution, $v\mathbf{1} = 1$). And so $\pi = v\Pi$, so

$$\left\|vP^{rk+j} - \pi\right\|_1 = \left\|vP^{rk+j} - v\Pi\right\|_1 = \left\|v(P^{rk+j} - \Pi)\right\|_1 = \theta^k\left\|v(Q^kP^j - \Pi)\right\|_1$$

since $Q^kP^j - \Pi$'s coefficients are all bound by 1, the norm is bound by a constant (which is the norm of v times the matrix of all ones, since v is positive). So we have that $\left\|vP^{rk+j} - \pi\right\|_1 \leq c\theta^k$ and finding the appropriate values, we can bound this by some $c'\alpha^{rk+j}$. ■

2.1.8 Theorem

Let P be the transition matrix of an irreducible aperiodic Markov chain, and let π be its unique stationary distribution. Then for every initial distribution v, $vP^n \xrightarrow{n\to\infty} \pi$ pointwise (meaning $(vP^n)_i \xrightarrow{n\to\infty} \pi_i$). Since $(vP^n)_i = \mathbb{P}_v(X_n = i)$, equivalently $\mathbb{P}_v(X_n = i) \xrightarrow{n\to\infty} \pi_i$ or $X_n \xrightarrow{d} \pi$.

So we must simply show that $|(vP^n)_i - \pi_i| \xrightarrow{n\to\infty} 0$. This is an immediate consequence of the previous lemma, which gave us that $\|vP^n - \pi\|_1 \leq c\alpha^n$ and so in particular $\|vP^n - \pi\|_1 \xrightarrow{n\to\infty} 0$. Since $\|vP^n - \pi\|_1 = \sum_{i=1}^N |(vP^n)_i - \pi_i|$, certainly $|(vP^n)_i - \pi_i| \xrightarrow{n\to\infty} 0$, as required. (In general convergence in the p-norms of $\mathbb{R}^N$ is equivalent to pointwise convergence.) ■

2.1.9 Corollary

If P is a stochastic matrix then all of its eigenvalues are bound by 1 (in absolute value).

Let γ be an eigenvalue of P, then there exists a vector v such that $Pv = \lambda v$. Let j be the state in S such that $|v_j| = \max_{i\in S}|v_i|$ and so

$$|\lambda||v_j| = |(Pv)_j| = \left|\sum_{i\in S} P_{ji}v_i\right| \leq \sum_{i\in S} P_{ji}|v_i| \leq |v_j|\sum_{i\in S} P_{ji} = |v_j|$$

Thus $|\lambda| \leq 1$. ■

2.1.10 Definition

Let μ and ν be two be two probability measures over the same σ-algebra $\mathcal{F}$, then we define their **total variation** to be

$$\|\mu - \nu\|_{\text{TV}} = \sup_{A\in\mathcal{F}}|\mu(A) - \nu(A)|$$

This is also denoted $d_{\text{TV}}(\mu, \nu)$, and this is in fact a metric over the space of probability measures on $\mathcal{F}$.

2.1.11 Definition

Let P be the transition matrix of an irreducible Markov chain whose stationary distribution is π, the we define

$$d(k) = \max_{j\in S} d_{\text{TV}}(e_jP^k, \pi)$$

Since e_jP^k and π are both distributions, they can be viewed as probability measures, and so we can discuss their total variation. e_jP^k is the distribution of X_k if $X_0 = j$, and so $d(k)$ gives us the maximum total variation of the distribution of X_k and π over all possible initial states. Let us also define the **mixing time** to be

$$t_{\text{mix}}(\varepsilon) = \min\{k \mid d(k) < \varepsilon\}$$

$t_{\text{mix}}(\varepsilon)$ gives us the minimum k where the total variation of the distribution X_k and π is less than ε, independent of the initial state. Though generally if we talk about the "mixing time" of a Markov chain, we set $\varepsilon = \frac{1}{4}$. And finally we also define

$$\bar{d}(k) = \max_{i,j\in S} d_{\text{TV}}(e_iP^k, e_jP^k)$$

By the triangle inequality, $\bar{d}(k) \leq 2d(k)$. And in fact $d(k) \leq \bar{d}(k)$ so

$$d(k) \leq \max_{i,j\in S}\left\|e_iP^k - e_jP^k\right\|_{\text{TV}}$$

2.1.12 Definition

A **coupling** of two probability measures μ and ν over the same σ-algebra $\mathcal{F}$ is a pair of random variables (X, Y) such that $X \sim \mu$ and $Y \sim \nu$. Formally, a coupling is a new probability space and random variables whose codomain is $\mathcal{F}$ such that for every $A \in \mathcal{F}$, $\mathbb{P}(X \in A) = \mu(A)$ and $\mathbb{P}(Y \in A) = \nu(A)$.

2.1.13 Proposition

If μ and ν are probability measures over the same σ-algebra, then

$$\|\mu - \nu\|_{\text{TV}} \leq \inf\{\mathbb{P}(X \neq Y) \mid (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}$$

Let (X, Y) be a coupling and $A \in \mathcal{F}$ then

$$\mu(A) - \nu(A) = \mathbb{P}(X \in A) - \mathbb{P}(Y \in A) \leq \mathbb{P}(X \in A, Y \notin A) \leq \mathbb{P}(X \neq Y)$$

and taking the infimum over all couplings (X, Y) preserves this inequality. ■

In fact, there is actually an equality here but the other direction is harder to prove.

2.1.14 Theorem

Suppose $\{X_n\}$ and $\{Y_n\}$ are two Markov chains with the same transition matrix P. Further suppose that if $X_s = Y_s$ then $X_t = Y_t$ for all $t \geq s$, then

$$\left\|e_x P^t - e_y P^t\right\|_{\text{TV}} \leq \mathbb{P}(X_t \neq Y_t \mid X_0 = x, Y_0 = y)$$

This is as $e_x P^t$ and $e_y P^t$ are the distributions of X_t and Y_t under the assumption that $X_0 = x$ and $Y_0 = y$. And (X_t, Y_t) is certainly a coupling of these distributions in $\mathbb{P}(\cdot \mid X_0 = x, Y_0 = y)$. ■

This means that if $\{X_n\}$ is a Markov chain, and $\{Y_n\}$ is some other Markov chain with the same transition matrix then $d(k)$ (for either $\{X_n\}$ or $\{Y_n\}$) can be bound by:

$$d(k) \leq \max_{i,j \in S} \mathbb{P}(X_k \neq Y_k \mid X_0 = i, Y_0 = j)$$

2.1.15 Example

What is the mixing time of the random walk on the circle C_N (this is the graph of N nodes, $\{v_1, \ldots, v_N\}$ with the edges $\{v_i, v_i\}$ and $\{v_i, v_{i+1}\}$)? Let us define two Markov chains X_n and Y_n where at every step we choose a random chain with equal probability and that will be the chain which will make the next step. As soon as the two chains intercept, they step together. Let T be the time that the two chains intercept, then by above and Markov's inequality

$$d(t) \leq \max_{x,y} \mathbb{P}_{x,y}(T > t) \leq \max_{x,y} \frac{\mathbb{E}_{x,y}[T]}{t}$$

The