MARKOV CHAINS: A VERY SHORT COURSE

LECTURER: MANJUNATH KRISHNAPUR

1. MOTIVATIONS

▶ Probability is used to model various real life situations, and the model may be fit the real situation to a greater or lesser extent. For example, when we consider the children born in a hospital and whether they are male or female, we may model the outcomes as tosses of a fair coin. Here it is understood that the tosses are independent of each other, i.e., the outcome of one does not influence the outcome of another. For the first time around 1905, Markov introduced a general class of models that could be used to describe a sequence of uncertain outcomes, where each outcome may be influenced by previous ones. This was the genesis of Markov chains, although a few special models studied before Markov could later be seen as special cases of it.

▶ Take the example of a text. On Mathematica, we invoked the text of *Alice in Wonderland*, and simplified it by removing all punctuation and capitalization and extra spaces to get a text of length 49217, made of the 27 characters a–z and space. We take this as the corpus and try to generate a new piece of text of length 80 by using it as follows.

Attempt-1: Count the frequencies of characters in the corpus, and sample independently according to those frequencies (for example, "e" occurs 5000 times, so each character is equal to "e" with probability $5000/49217$. In one experiment, the output was

c fl fgro gkooeyalegyrhptaesd fcroye atbi oetswenhls cda te s ttfrmsko i aalssuem .

Attempt-2: As the text is built up, if the last character so far is "e", we look in the corpus for all the 5000 times "e" occured, and count how many times it was followed by various characters, and sample the next character in proportion to those frequencies. For example, "ea" occurred 274 times, so whenever "e" occurs, the next character will be "a" with probability $274/5000$. In one experiment, the output was

wasasebleyoond thenthevitheale peting aning me anor fin t d keg whi walld the maca.

Attempt-3: Like before, but now we keep track of the last two characters in the new text so far and use that to choose the next one. For example, "th" occurred 1261 times in the corpus, and in 870 of these occurrences, it was followed by "e". Therefore, whenever the last two letters in our text are "th", the next letter will be "e" with probability $870/1261$. In one experiment, the output was

a low tre every ever cauckedly don a ged noice cat the ge fore warged sating anxio .

Attempt-4: Now we keep track of the last 3 or 4 characters to fill in the next. We got the outputs

e mustory made shards aftere of can animall it ve open a mome to know no said to eit *ld front on the cake more plate tried out of it said that thing voice whether whole w*

▶ Clearly, the words start looking more like actual words or other reasonable looking ones, when we keep track of more and more characters. In other words, text is probably better modeled by a sequence with some memory, rather than a sequence of independent characters thrown together. But the size of the memory is not too large, 3 or 4 characters appear to be sufficient.

2. Markov chains: definition

▶ There are two ingredients to defining a Markov chain: (1) A *state space* S, which is just a finite or countably infinite set. (2) A *transition matrix* (also called a *stochastic matrix*) $P = (p_{i,j})_{i,j\in S}$, which is an $S\times S$ (or $n\times n$ if you prefer) matrix whose entries are positive (i.e., $p_{i,j}\geq 0$) and whose row-sums are all equal to 1 (i.e., $\sum_j p_{i,j}=1$ for all i).

▶ A Markov chain with state space S and transition matrix P is a sequence $(X_0, X_1, X_2, \ldots)$ of S-valued random variables (on some probability space) with the property that

$$\mathbf{P}\{X_{t+1}=j \mid X_0=i_0, X_1=i_1,\ldots,X_t=i_t\} = p_{i,j}, \qquad \text{where } i=i_t,$$

for all $t\geq 0$ and all $i_0,\ldots,i_t,j\in S$. In other words, if the "current state" (namely X_t) is equal to i, then irrespective of the past, the chance that $X_{t+1}=j$ is $p_{i,j}$.

3. Examples of Markov chains

▶ **A general 2-state MC**: Let $S=\{1,2\}$ (or any 2-element set) and $P=\begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$ for some $0\leq \alpha,\beta\leq 1$. This will be a useful workhorse for us, to carry out computations explicitly before going to the general case.

▶ **Simple random walk on a graph**: Let $G=(V,E)$ be a finite graph with vertex set V and edge set E. Then SRW on G is the Markov chain with state space V and transition probabilities

$$p_{i,j} = \begin{cases} \frac{1}{\deg(i)} & \text{if } j\sim i, \\ 0 & \text{otherwise.} \end{cases}$$

In words, if the current state is $i\in V$, the next state will be one of the vertices adjacent to i, and it will be picked with uniform probability.

A special case is when G is d-regular. Then $P=\frac{1}{d}A$, where A is the adjacency matrix of G.

▶ **Random walk on a weighted graph**: Let $G=(V,E,w)$, where $w: E\to\mathbb{R}_+$. We say that $w(e)$ is the weight of the edge e. We define

$$p_{i,j} = \frac{w(i,j)}{\sum_{k\sim i} w(i,k)} \text{ if } j\sim i, \quad p_{i,j}=0 \text{ if } j\not\sim i. \tag{1}$$

The corresponding Markov chain is called a random walk on G w.r.t weights/conductances w. If we take $w(e)=1$ for all e, we get back the SRW on G.

▶ **Ehrenfest chain**: The original model comes from statistical physics. Imagine a box containing some gas. Inside the box is a partition with a tiny hole in it. At each instant of time, we imagine that one of the gas molecules (picked at random) is transferred to the other side of the partition.

▶ **Card shuffling mechanisms**: Consider a deck of k cards labelled $1, 2 \ldots, k$. The state space S is the set of all arrangements or permutations of the cards. Any mechanism for shuffling cards leads to a Markov chain on S. Let us consider one simple one, *random transposition*. A transposition is any permutation of the form $\tau_{i,j} = (i,j)$, the permutation that interchanges card i and card j, and leaves everything else untouched. There are $\binom{n}{2}$ such transpositions. Now consider a Markov chain defined by $X_{t+1} = \tau_{I_t,J_t}(X_t)$, whre (I_t, J_t) is picked uniformly at random from the set of pairs $\{(i,j) : i < j\}$ uniformly. In words, we pick two distinct cards and interchange their positions.

▶ **A model for queues** Imagine a queue in a cafeteria. Between time t and time $t+1$, the first person in the queue (if any was there at time t) gets served and leaves the queue. In the same time duration a random number Z_t of people join the queue. A simple model is to assume that $Z_0, Z_1, Z_2, \ldots$ are independent and identically distributed with $\mathbf{P}\{Z_t = m\} = a_m$ for $m \geq 0$. Let X_t denote the number of people in the queue at time t. Then $X = (X_0, X_1, \ldots)$ is a Markov chain with state space $\mathbb{N} = \{0, 1, 2, \ldots\}$ with transitions

$$p_{i,j} = \begin{cases} a_j & \text{if } i = 0, \\ a_{j+1} & \text{if } i \geq 1. \end{cases}$$

▶ **Random walk on directed graphs**: If $G = (V, E, w)$ is a directed graph, then the corresponding random walk is the Markov chain with transitions given by (1), except that now $w(i,j)$ means the weight of the edge $i \to j$. But this is not a special case but all Markov chains!

In fact, given any S and P, form a directed graph with vertex set S, edge set $\{(i,j) : p_{i,j} > 0\}$ and weights $w(i,j) = p_{i,j}$. The random walk on this weighted directed graph is precisely the Markov chain on S with transition matrix P.

As we shall see later, random walks on undirected (weighted or unweighted) graphs are special!

4. Three questions

We raise three questions of interest. What needs to be kept in mind is that the transition matrix tells us how the chain moves in one step. The questions here are not about one step, but over the fullness of time. This is similar in spirit to classical Mechanics where Newton's laws (given as a system of differential equations) tell us how a system of bodies move in an infinitesimal amount of time. The question of how to calculate the positions and velocities after a year or a century is the problem of integrating the differential equation. It can be difficult. But in principle, no more information is needed than the infinitesimal law. Analogously, any question about the behaviour of a Markov chain should be possible to answer in terms of the transition matrix. Again, this is true in principle, but can be difficult in practise.

Q1 (Long-term behaviour of the chain): What is the distribution of X_t for large t? That is, for $i, j \in S$, does $\lim_{t\to\infty} \mathbf{P}\{X_t = j \mid X_0 = i\}$ exist? What is it and how to compute it from the given transition probabilities? Does it depend on the starting point i?

Q2 (Recurrence v/s Transience): Does the Markov chain visit every state in S? If so, does it visit infinitely often? Does the answer depend on the starting point? In symbols, the first question is asking if $\mathbf{P}\{\cup_{t\geq 0}\{X_t = j\}\} = 1$ for all $j \in S$?

If the answer is Yes, we say that the chain is recurrent. Otherwise we say that the chain is transient. How to decide whether a chain is recurrent or transient based on the transition matrix?

Q3 (Harmonic measure): Fix two disjoint subsets A, B of the state space S. For $i \in S$, what is the probability that a Markov chain started at i (i.e., given $X_0 = i$) hits A before B?

This problem is also called the problem of gambler's ruin because of the following special case. Imagine two gamblers with a and b rupees each playing a sequence of gambling games. After each game, the loser pays 1 rupee to the winner. Imagine that both have equal chance of winning any game, and that the games are independent of each other. The profit of the second gambler can be thought of as a SRW on $\mathbb{Z}$ which moves up or down with equal probability (i.e., $p_{i,i+1} = \frac{1}{2} = p_{i,i-1}$). The chain starts at 0, and stops when it hits $+a$ (the first gambler goes bankrupt) or when it hits $-b$ (the second gambler goes bankrupt), whichever happens first. It is of interest to find the chances of the two possibilities. Clearly, this is a harmonic measure problem for the SRW on $\mathbb{Z}$ with $X_0 = 0$ and $A = \{+a\}$ and $B = \{-b\}$ (it makes no difference if we take $A = \{a, a+1, a+2, \ldots\}$ and $B = \{-b, -b-1, -b-2, \ldots\}$).

5. Multi-step transitions and the question of long-term behaviour

► From the definition, we can see that

$$\begin{aligned}
&\mathbf{P}\{X_1 = i_1, X_2 = i_2, \ldots, X_t = i_t \mid X_0 = i_0\} \\
&= \mathbf{P}\{X_1 = i_1 \mid X_0 = i_0\}\mathbf{P}\{X_2 = i_2 \mid X_1 = i_1, X_0 = i_0\} \ldots \mathbf{P}\{X_t = i_t \mid X_{t-1} = i_{t-1}, \ldots, X_0 = i_0\} \\
&= p_{i_0,i_1} p_{i_1,i_2} \ldots p_{i_{t-1},i_t}.
\end{aligned}$$

► What is $\mathbf{P}\{X_t = j \mid X_0 = i\}$? We can sum the above probabilities over all possible trajectories that lead from i to j in exactly t steps. Thus,

$$\mathbf{P}\{X_t = j \mid X_0 = i\} = \sum_{i_1,\ldots,i_{t-1}\in S} p_{i,i_1} p_{i_1,i_2} \ldots p_{i_{t-2},i_{t-1}} p_{i_{t-1},j}.$$

A little thought shows that this is nothing but the (i, j) entry of P^t (the matrix P multiplied with itself t times). Thus[1],

$$\mathbf{P}\{X_t = j \mid X_0 = i\} = P^t_{i,j}.$$

▶ Look back at Q1, the question of long-term behaviour. It was precisely to compute the above probability, as $t \to \infty$. Have we solved the question? We have reduced it to a question about matrices: What are the entries of P^t, for large t?

▶ One can easily generate a few stochastic matrices on a computer and raise them to high powers. Quite often, one sees that all rows become equal! In other words, it seems to be the case that as $P^t_{i,j} \to \pi_j$ as $t \to \infty$, where $\pi_1, \ldots, \pi_n$ are some positive numbers that sum to 1. This is still a numerical observation, to be proved or disproved!

▶ But this is not always true. Here are two cases,

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } Q = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Then $P^t = P$ for all t and $Q^t = Q$ for odd t and $Q^t = P$ for even t. Thus the limit of Q^t does not exist. The limit of P^t does exist, but the rows are not all equal.

▶ Moral: We need to understand the conditions on P under which $P^t_{i,j}$ converges for all i, j and the limit is free of i (depends only on j). We must also ask how these π_j values are to be found from P?

▶ Let us analyse the 2-state chain with $0 < \alpha, \beta < 1$. We claim that with $r = 1 - \alpha - \beta$,

$$P^t = \frac{1}{\alpha + \beta} \begin{bmatrix} \alpha r^t + \beta & \alpha - \alpha r^t \\ \beta - \beta r^t & \alpha + \beta r^t \end{bmatrix} \text{ for } t \in \mathbb{N}.$$

The claim can be easily verified by induction. Note that $|r| < 1$, therefore,

$$P^t \to \frac{1}{\alpha + \beta} \begin{bmatrix} \beta & \alpha \\ \beta & \alpha \end{bmatrix} \text{ as } t \to \infty.$$

In other words, irrespective of whether X_0 is 0 or 1, the probabilities that X_t is equal to 0 or 1 converge to $\frac{\beta}{\alpha+\beta}$ and $\frac{\alpha}{\alpha+\beta}$ respectively.

All of this work as long as $|r| < 1$. Otherwise, $\alpha = \beta = 0$ or $\alpha = \beta = 1$. In either case, P^t alternates between $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ as t varies. Thus, $\lim_{t \to \infty} P^t_{i,j}$ does not exist.

▶ For general P, how do we guess at P^t? Clearly a better way must be found. *Suppose* we can diagonalize P as $P = XDX^{-1}$, where X is an invertible matrix and $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$, where $n = |S|$. Then, $P^t = XD^tX^{-1}$. Here $D^t = \text{diag}(\lambda_1^t, \ldots, \lambda_n^t)$. Since $P\mathbf{1} = \mathbf{1}$, we may take

[1] It would be less ambiguous but more cumbersome to write $(P^t)_{i,j}$. Do not confuse $P^t_{i,j}$ with $(P_{i,j})^t$.

$\lambda_1 = 1$ (and the first column of X to be $\mathbf{1}$). Now additionally *suppose* that $|\lambda_j| < 1$ for $j \geq 2$. Then $D^t \to E_{1,1}$, where $E_{1,1} = e_1 e_1^\dagger$ has $(1,1)$ entry equal to 1 and all other entries are zero. Thus, $P^t \to X\mathrm{diag}(1, 0, \ldots, 0)X^{-1}$. It is easy to see that the (i,j) entry of the limiting matrix is $X_{i,1}(X^{-1})_{1,j}$. But we may take $X_{i,1} = 1$ (as $P\mathbf{1} = \mathbf{1}$), hence if we define $\pi_j = (X^{-1})_{1,j}$, then $P^t_{i,j} \to \pi_j$ for all i, j. Thus the limits exist, and do not depend on the starting state.

▶ Not all matrices are diagonalizable. But every matrix has a Jordan decomposition. To recall this, define the Jordan block

$$J_m(\lambda) = \begin{bmatrix} \lambda & 1 & 0 & \ldots & \ldots & 0 \\ 0 & \lambda & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & 0 & \lambda & 1 \\ 0 & \ldots & \ldots & \ldots & 0 & \lambda \end{bmatrix}_{m \times m}.$$

For $m = 1$, we simply take $J_1(\lambda) = [\lambda]$. Then any matrix is conjugate to a matrix which is block diagonal with each block being a Jordan block. For each eigenvalue λ, there may be one or more Jordan blocks, $J_{m_1}(\lambda), \ldots, J_{m_k}(\lambda)$, where $m_1 + \ldots + m_k$ is the arithmetic multiplicity of λ (and k equals the geometric multiplicity of λ).

▶ Write $P = XDX^{-1}$ where D is block diagonal with each block being a Jordan block. Again, $P\mathbf{1} = \mathbf{1}$, therefore, 1 is an eigenvalue. *Suppose* we assume that 1 is a simple eigenvalue, and that all other eigenvalues satisfy $|\lambda_j| < 1$. Then D is of the form

$$D = \begin{bmatrix} 1 & 0 & 0 & \ldots & \ldots & 0 \\ 0 & J_{m_1}(\lambda_1) & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & J_{m_2}(\lambda_2) & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & 0 & & 01 \\ 0 & \ldots & \ldots & \ldots & 0 & J_{m_k}(\lambda_k) \end{bmatrix}.$$

From this it is clear that $D^t \to E_{1,1}$ and therefore $P^t \to XE_{1,1}X^{-1}$. Thus if we choose X so that $X_{i,1} = 1$ and write $\pi_j = (X^{-1})_{1,j}$, then $P^t_{i,j} \to \pi_j$. As $P^t_{i,j} \geq 0$, we must have $\pi_j \geq 0$. Further $(X^{-1}X)_{1,1} = 1$ imples that $\sum_j \pi_j = 1$. Thus π is a probability distribution on S.

▶ We have almost solved the question of long-term behaviour. If P has 1 as a simple eigenvalue and all other eigenvalues are strictly smaller than 1 in absolute value, then there is a probability vector $\pi = (\pi_j)_{j \in S}$ such that $P^t_{i,j} \to \pi_j$ for all $i, j \in S$. When there are other eigenvalues of absolute value 1 (as in the 2-state example), then $P^t_{i,j}$ may not converge. Still, it would be nice to have a simpler way to determine if 1 is a simple eigenvalue of P (instead of having to compute all the eigenvalues). Such a criterion is given by the *Perron-Frobenius theorem*.

▶ Assume that P is irreducible (i.e., for each $i, j \in S$, there is some t such that $P^t_{i,j} > 0$) and aperiodic (i.e., g.c.d.$\{t : P^t_{i,i} > 0\} = 1$ for some (and then for all) $i \in S$. Then, 1 is a simple eigenvalue of P, and all other eigenvalues have absolute values strictly smaller than 1. As a corollary, there is a probability vector $\pi = (\pi_i)_{i \in S}$ such that $P^t_{i,j} \to \pi_j$ for all $i, j \in S$, as $t \to \infty$.

Proof: The assumption of irreducibility and aperiodicity imply that there is some T such that $P^T_{i,j} > 0$ for all $i, j \in S$ (why?). Applying the Perron-Frobenius theorem to P^T, we see that 1 is a simple eigenvalue and all other eigenvalues have absolute value strictly smaller than 1. But the eigenvalues of P^T are $\lambda_1^T = 1$ and λ_j^T, $j \geq 2$, hence this shows that $|\lambda_j| < 1$ for $j \geq 2$.

6. Stationary distribution

▶ For an irreducible, aperiodic MC, we have see that $P^t_{i,j} \to \pi_j$ for all $i, j \in S$. What are these π_js and how to compute them?

▶ From the proof, we have see that if $P = XDX^{-1}$ is the Jordan decomposition of P and $X_{i,1} = 1$ for all i, then $\pi_j = (X^{-1})_{1,j}$. Thus, π is the first column of $(X^{-1})^\dagger$, which by the equation $P^\dagger = (X^{-1})^\dagger D^\dagger X$ show that π is the eigenvector of $P^\dagger$ for the eigenvalue 1. Further, $(X^{-1}X)_{1,1} = 1$ shows that $\sum_j \pi_j = 1$. Since it arises as the limit of $P^t_{i,j}$, we must have $\pi_j \geq 0$. But this is also part of the Perron-Frobenius theorem (applied to P^T it shows that $\pi_j > 0$ for all j).

▶ Thus π can be found as the top eigenvector of $P^\dagger$ normalized so that $\sum_j \pi_j = 1$. Even if P is not aperiodic (but 1 is a simple eigenvalue of P and hence of $P^\dagger$), we can define π like this. It is called the *stationary distribution* of the Markov chain. Finding it means solving the linear equations $(P^\dagger - I)\pi = 0$, which need not be easy (numerically one can find it on a computer fairly efficiently).

▶ There are two cases when we can find π explicitly.

(1) We say that P is doubly stochastic if its column sums are also 1. Then $\mathbf{P}^\dagger \mathbf{1} = \mathbf{1}$, hence $\pi_j = \frac{1}{n}$ for all $j \in S$.

(2) We say that P is *reversible* w.r.t. the numbers $(\sigma_i)_{i \in S}$ if $\sigma_i > 0$ and $\sigma_i P_{i,j} = \sigma_j P_{j,i}$ for all $i, j \in S$. In such as case $(P^\dagger \sigma)_i = \sum_j P^\dagger_{i,j} \sigma_j = \sum_j P_{j,i} \sigma_i = \sigma_i$. Therefore, σ is an eigenvector of $P^\dagger$ for the eigenvalue 1, and hence the stationary distribution is given by $\pi_i = \sigma_i / \sum_k \sigma_k$.

▶ Consider SRW on a finite graph $G = (V, E)$. Then $P_{i,j} = \frac{1}{\deg(i)}$ if $j \sim i$, and $P_{i,j} = 0$ otherwise. Hence if $\sigma_i = \deg(i)$, then $\sigma_i P_{i,j} = \sigma_j P_{j,i}$ for all $i \sim j$. For $i \not\sim j$, both sides are zero, hence equal. Thus P is reversible and $\pi_i = \frac{\deg(i)}{\sum_{k \in V} \deg(k)}$. The MC spends time at each vertex proportional to its degree.

More generally, for random walk on the weighted graph $G = (V, E, w)$, the chain is reversible with $\sigma(i) = \sum_{j \sim i} w(i, j)$ ("generalized degree"). Hence $\pi_i = \frac{\sigma(i)}{\sum_k \sigma(k)}$.

▶ Consider the Ehrenfest chain with $S = \{0, 1, \ldots, N\}$ and $P_{i,i+1} = \frac{N-i}{N}$ and $P_{i,i-1} = \frac{i}{N}$ and all other $P_{i,j} = 0$. Let $\sigma_i = \binom{N}{i}$. Then

$$\sigma_i P_{i,i+1} = \frac{N!}{i!(N-i)!}\frac{N-i}{N} = \frac{(N-1)!}{i!(N-i-1)!},$$
$$\sigma_{i+1} P_{i+1,i} = \frac{N!}{(i+1)!(N-i-1)!}\frac{i+1}{N} = \frac{(N-1)!}{i!(N-i-1)!}.$$

From this we see that P is reversible w.r.t. σ and hence $\pi_i = \frac{1}{2^N}\binom{N}{i}$ for $0 \le i \le N$.

▶ Consider the Tsetlin library. It is not reversible, because there are $\pi, \sigma \in \mathcal{S}_n$ such that $P_{\pi,\sigma} > 0$ and $P_{\sigma,\pi} = 0$. However, every permutation has n inward edges and n outward edges, from which we see that P is doubly stochastic. Hence π is uniform on $\mathcal{S}_n$.

7. Two remarks on reversible chains

▶ We saw that random walks on wighted (undirected) graphs are reversible. In fact, all reversible Markov chains can be thought of as random walks on weighted graphs.

Suppose S is a finite or countable state space, and P is a transition matrix that is reversible w.r.t. σ, i.e., $\sigma_i > 0$ for all $i \in S$ and $\sigma_i p_{i,j} = \sigma_j p_{j,i}$ for all $i, j \in S$. Then form the graph G with vertex set S, edge set $\{\{i,j\} : p_{i,j} > 0\}$ (the edges are undirected because $\sigma_i p_{i,j} = \sigma_j p_{j,i}$ and $\sigma_i, \sigma_j > 0$ shows that $p_{i,j} > 0$ if and only if $p_{j,i} > 0$). Define the edge-weights, $w(i,j) = \sigma_i p_{i,j}$ (so $w(j,i) = w(i,j)$, the weights don't depend on the direction). Check that the random walk on G is precisely the MC with transition matrix P.

▶ Rate of convergence: Let S be finite with $n = |S|$, and P be irreducible and aperiodic. We know that $P^t_{i,j} \to \pi_j$ for all $i, j \in S$. How fast is the convergence? We restrict ourselves to the special case when $P_{i,j} = P_{j,i}$, i.e., P is reversible w.r.t. $\sigma_i = 1$ (we can extend all the considerations here to general reversible chains, see the end of this section).

▶ Write the spectral decomposition $P = \lambda_1 \mathbf{v}_1\mathbf{v}_1^\dagger + \ldots \lambda_n \mathbf{v}_n\mathbf{v}_n^\dagger$, where $\{\mathbf{v}_i\}$ is an orthonormal basis of $\mathbb{R}^n$ and λ_i are the corresponding eigenvalues (real because $P^\dagger = P$). Take $\lambda_1 = 1$ and $\mathbf{v}_1 = \frac{1}{\sqrt{n}}\mathbf{1}$. Then $-1 < \lambda_2, \ldots, \lambda_n < +1$. Let $\alpha = \max_{j \ge 2} |\lambda_j|$. Then

$$P^t = \frac{1}{n}\mathbf{1}\mathbf{1}^\dagger + \lambda_2^t \mathbf{v}_2\mathbf{v}_2^\dagger + \ldots + \lambda_n^t \mathbf{v}_n\mathbf{v}_n^\dagger$$

is the spectral decomposition of P^t (eigenvectors remain the same, eigenvalues are powers of eigenvalues of P). Thus, for any $i, j \in S$, we have

$$P^t_{i,j} - \frac{1}{n} = \lambda_2 \mathbf{v}_2(i)\mathbf{v}_2(j) + \ldots + \lambda_n^t \mathbf{v}_n(i)\mathbf{v}_n(j)$$

from wheich we can see that

$$\sum_j |P^t_{i,j} - \frac{1}{n}| \le \alpha^t \sum_{k=2}^n \sum_j |\mathbf{v}_k(i)||\mathbf{v}_k(j)| \le C_n \alpha^t.$$

As $|\mathbf{v}_k(i)| \le 1$, we can take $C_n = n^2$ (or do better but we don't care here). This means that the probability vector $(P^t_{i,j})_{j\in S}$ is within $C_n\alpha^t$ distance on the uniform distribution π (in the ℓ^1 distance). In particular, this shows that the distance between them decays exponentially fast. The smaller α is, faster is the decay. If α is close to 1, the decay is slower.

▶ The quantity $1-\alpha$ is called the spectral gap. And the larger it is, faster the convergence of the chain to stationarity. For example, with $P = \frac{1}{d}A$ for the SRW on a d-regular graph, this motivates the question of finding a sequence G_n of d-regular graphs (with d fixed, say $d=3$) so that the spectral gap remains bounded away from zero. Such graphs are called *expander graphs*. Constructing a sequence of expander graphs is not easy!

8. Recurrence and Transience

▶ As always, consider an irreducible chain with state space S and transition P. For a state $j \in S$, define the hitting time random variable $T_j = \min\{t \ge 1 : X_t = j\}$ (infinite if the set is empty). If $\mathbf{P}\{T_j < \infty \bigm| X_0 = i\} < \infty$ for all $i, j \in S$, we say that the MC is *recurrent*. As the chain is irreducible, one can show that this is equivalent to $\mathbf{P}\{T_j < \infty \bigm| X_0 = j\}$ for one single $j \in S$ (we skip the argument here, but it is similar to the one that comes next).

▶ If S is finite and P is irreducible, the chain is recurrent.

Proof: Then there is a $t_0 < \infty$ and $\delta > 0$ such that for every $i, j \in S$, there is some $t \le t_0$ such that $P^t_{i,j} \ge \delta$. Now break the chain into steps of t_0. Fix a target state j. In each block of t_0 steps, there is a probability of at least δ that the chain hits the state j, conditional on all that happened before. Therefore, the probability that $T_j > kt_0$ steps is bounded by $(1-\delta)^k$, which goes to 0 as $k \to \infty$. Therefore $T_j < \infty$ w.p.1.

▶ Thus, for finite state spaces, the recurrence question is not interesting. But one can ask quantitative questions that are more interesting, for example, the distribution or the tails or moments of T_j. A very interesting and important fact is that for a finite state space, irreducible, aperiodic chain,

$$\mathbf{E}[T_j \bigm| X_0 = j] = \frac{1}{\pi_j}.$$

▶ Let us explain one consequence of it. Consider the Ehrenfest chain on $S = \{0, 1, \dots, N\}$. As it is irreducible, the chain visits every state, including 0. In terms of the interpretation in terms of gas molecules, this says that there will come a time when all the air in a room moves to the right half, leaving a vacuum in the left half. But this is not something that anyone has experienced, so it looks like a contradiction. However, the paradox gets resolved when we look at it quantitatively.

Recall that $\pi_j = \frac{1}{2^N}\binom{N}{j}$. Hence $\mathbf{E}[T_0 \mid X_0 = 0] = 2^N$. Contrast with $j = \frac{N}{2}$, for which $\mathbf{E}[T_j \mid X_0 = j] \sim c\sqrt{N}$ (because $\binom{N}{N/2}\frac{1}{2^N} \sim \frac{1}{c\sqrt{N}}$ by Stirlings' approximation (recall that $f(s) \sim g(s)$ means that $f(f)/g(s)$ goes to 1 as $s \to \infty$). Both $\sqrt{N}$ and 2^N are large, but when we plug in $N = 10^{26}$ (realistic for number of molecules of air in a room), the quantities are 10^{13} and $2^{10^{26}}$. If 10^6 moves happen in a second, then the first one is less than half a year, while the second one is larger than the age of the universe! It is not inconsistent with observation.

▶ Now let us come to the problem of deciding whether an irreducible Markov chain on an infinite state space is recurrent or transient. Here is the key theorem.

Theorem: A chain is recurrent or transient according as $\sum_{t=0}^{\infty} P_{i,i}^t$ is divergent or convergent for some $i \in S$ (the convergence/divergence does not depend on i).

▶ Consider random walk on $\mathbb{Z}$ with $p_{i,i+1} = p$ and $p_{i,i-1} = 1-p$ where $0 < p < 1$. Let $i = 0$. Starting from 0, the walk can return to 0 only in an even number of steps. And if $t = 2s$, then $P_{0,0}^t = \binom{2s}{s}(p(1-p))^s$, since there must be s steps to the right and s steps to the left, but they can occur in any order. By Stirlings' approximation $\binom{2s}{s} \sim \frac{2^{2s}}{c\sqrt{2s}}$ as already noted above. Hence $P^{2s}(0,0) \sim (4p(1-p))^s/c\sqrt{s}$ as $s \to \infty$.

If $p = 1/2$, then $P^{2s}(0,0) \sim \frac{1}{c\sqrt{2s}}$ and hence $\sum_t P_{0,0}^t = \infty$. The chain is recurrent.

If $p \neq \frac{1}{2}$, then $4p(1-p) < 1$, hence $P_{0,0}^t$ decays exponentially fast in t. Hence $\sum_t P_{0,0}^t < \infty$. The chain is transient.

▶ Consider SRW on $\mathbb{Z}^2$. Again $P_{0,0}^t$ (here 0 stands for the origin) is zero for odd t. For $t = 2s$, to return to the origin, the walk must make k steps to the right and left each, and $s-k$ steps up and down each (for some $0 \leq k \leq s$). Therefore,

$$
\begin{aligned}
P_{0,}^{2s} &= \sum_{k=0}^{s} \frac{(2s)!}{(k!(s-k)!)^2}\frac{1}{4^{2s}} = \frac{(2s)!}{(s!)^2 4^{2s}}\sum_{k=0}^{s}\binom{s}{k}^2 \\
&= \frac{(2s)!}{(s!)^2 4^{2s}}\binom{2s}{s} = \left(\binom{2s}{s}\frac{1}{2^{2s}}\right)^2 \sim \frac{1}{c^2 s}.
\end{aligned}
$$

Thus, $\sum_t P_{0,0}^t = \infty$, hence the chain is recurrent.

▶ **Exercise**: Show that SRW on the infinite b-regular tree (for $b \geq 3$) is transient.

9. Further reading

▶ We have just touched some of the most basic aspects of Markov chains. Most books on probability, have a chapter on Markov chains, covering all the basics in greater detail than we have done here. Here are some:

(1) W. Feller, *An Introduction to Probability Theory and Its Applications: v. 1*, John Wiley & Sons; 3rd Edition (1968).

(2) Grimmett and Stirzaker, *Probability and random processes*, Oxford university press, 4th ed (2020).

(3) P. Brémaud, *Markov chains*, Springer (2020).

BASED ON LECTURES GIVEN IN *IWM mini-course* (NOV 2024) ORGANIZED BY PROFESSORS ARCHANA MORYE AND ABHAY SOMAN OF UNIVERSITY OF HYDERABAD AND THE WORKSHOP *Graphs, Matrices and Applications* (SEP 2025) ORGANIZED BY PROFESSOR RAJESH KANNAN OF IIT, HYDERABAD.