# Lecture 13: Markov Chains and Random Walks

October 8, 2025

## 1 Introduction

A **stochastic process** is a time-evolving sequence of random variables: $X_0, X_1, X_2, \ldots, X_t, \ldots$ where $X_t$ is the state of the system at time $t$ and $X_0$ is the initial state (which can itself be a random one). One can view it also as an evolving randomized algorithm. A Markovian process or a Markov chain is a particular type of stochastic process where $X_t$ depends only on $X_{t-1}$. To formalize this, we assume that $X_0, X_1, \ldots$ are random variables over some state space $\Omega$.

**Definition 1.** A stochastic process is **Markovian** if:

$$\Pr[X_t = u_t \mid X_{t-1} = u_{i-1}, X_{t-2} = u_{t-2}, \ldots, X_0 = u_0] = \Pr[X_t = u_t \mid X_{t-1} = u_{t-1}]$$

We will be mostly concerned with finite state Markov chains, although countable state chains are also quite relevant. A finite state Markov chain is easy to visualize as a directed graph $G = (V, E)$ with non-negative edge-weights. Note that we allow self-loops.

- $V$ represents the states. We usually use $n$ for $|V|$ and assume the states are numbered 1 to $n$.
- For state (vertex) $i$, the weighted outgoing edges from $i$ represent the probability with which state $j$ is reached from state $i$ in one step. Thus the weight of the edge $p(i,j) \in [0,1]$. If no edge $(i,j)$ is present, then the probability is implicitly 0.
- Self-loops are allowed since we want to allow the process to remain in the same state.

We associate an $n \times n$ probability transition matrix $P$ with an $n$-state Markov chain:

$$P_{ij} = \Pr[X_t = j \mid X_{t-1} = i]$$

Clearly we need:

$$\sum_{j \in [n]} P_{ij} = 1 \quad \forall i \in [n]$$

and $P_{ij} \geq 0$.

The advantage of the graphical representation is that it allows us to understand the properties of the chain via graph-theoretical aspects.

**Evolution of the Process:** Suppose $\pi(0)$ is an initial probability distribution over the state space $V$. How does the process evolve? This is the central question in Markov chains. We represent probability distributions over the states by $n$-dimensional *row* vectors. We let $\pi_i$ denote the probability of being in state $i$. Suppose we start with $\pi(0)$ as the starting distribution (can be deterministic in that we have $\pi(0)_i = 1$ for some fixed vertex $i$). Then it is not difficult to see that after one step the distribution is:

$$\pi(1) = \pi(0)P$$

since $\Pr[X_1 = j] = \sum_i \Pr[X_0 = i]P_{ij}$ due to the Markovian property. After $t$ steps, we see by induction that:

$$\pi(t) = \pi(0)P^t$$

We want to understand properties of the chain as it evolves and in particular the long-term behaviour of the chain. A central result in the theory is that the process converges to a stable stationary distribution under reasonable conditions. These conditions are natural and relatively easy to understand from a graph-theoretic viewpoint.

**Transient states and irreducability:** Suppose $G$ is not strongly connected. Then it is not hard to see that the process will get stuck in a sink component of the underlying meta-graph (i.e., the strongly connected component graph). There could be multiple such sink components and the process will reach one of them and will not leave. Thus any state that is not in one of those sink components is going to be *transient.* Thus, for understanding long-term behaviour it suffices to focus on strongly connected chains (sometimes we simply say connected). This motivates the following:

**Definition 2.** A Markov chain is **irreducible** if the underlying graph is strongly connected.

If a chain is irreducible then any state $i$ can reach any other state $j$ with some probability $\epsilon > 0$ since there is an $i$ to $j$ path with non-zero probabilities on each of the edges. We can then prove the following.

**Lemma 1.** *Let $h_{ij}$ be the expected time to hit state $j$ for the first time starting in state $i$; if $j = i$ we think of the first time when $i$ is revisited. Then $h_{ij} < 0$.*

**Periodicity:** A second issue that comes up is periodicity or oscillatory behaviour.

**Definition 3** (Period)**.** For a finite state Markov chain defined by a graph $G = (V, E)$ and a state $i \in V$, period$(i)$ is the largest non-negative integer $d$ such that $d$ divides the length of any closed walk containing $i$:

$$\text{period}(i) = \gcd\{|W| : W \text{ is a closed walk containing } i\}$$

**Lemma 2.** *Suppose $G$ is strongly connected. Then period$(i)$ = period$(j)$ for all $i, j \in V$.*

*Proof.* Exercise. □

The preceding lemma implies that there is a period $d$ for a strongly connected graph.

**Lemma 3.** *Suppose period is $d \geq 1$. Then $V$ can be partitioned into $V_0, V_1, \ldots, V_{d-1}$ such that $(u, v) \in E$ implies $u \in V_i$ and $v \in V_{(i+1) \bmod d}$.*

**Definition 4.** An irreducible finite state Markov chain is **aperiodic** if period $d = 1$.

**Definition 5.** An irreducible and aperiodic Markov chain is called **ergodic**.

**Lazy Random Walk** Suppose we have a Markov chain which is irreducible but periodic. We can make it aperiodic by adding a self-loop to each $i$ and making $i$ stay in state $i$ with probability $p > 0$ (say $\frac{1}{2}$) and take the original transition with probability $(1 - p)$. In other words, we are changing the transition matrix from $P$ to $(1 - p)P + pI$ where $I$ is the identity matrix. The new walk is called a **lazy version** of the original walk and retains the essential properties.

An important and fundamental notion is the following.

**Definition 6.** A distribution $\pi$ is a **stationary distribution** of a Markov chain with transition matrix $P$ if $\pi P = \pi$.

A stationary distribution is a stable distribution.

# 2 Fundamental Theorem of Markov Chains

**Theorem 1.** *Suppose $P$ corresponds to a finite state irreducible Markov chain. Then there exists a unique stationary distribution $\pi$ for it.*

1. *For any $i$, $\pi_i = \frac{1}{h_{ii}}$ where $h_{ii}$ is the expected time for the chain to revisit $i$ if started in $i$.*

2. *Let $N_i(t)$ be the number of times the chain visits $i$ in $t$ steps. Then $\lim_{t\to\infty} \frac{N_i(t)}{t} = \pi_i$.*

*Moreover if $P$ is also aperiodic, and hence the chain is ergodic, for any starting distribution $\pi(0)$, $\lim_{t\to\infty} \pi(0)P^t = \pi$.*

## 2.1 Proof via Perron-Frobenius Theorem

**Connection to eigenvalues:** $\pi P = \pi$ implies $\pi$ is a left eigenvector of $P$ with eigenvalue 1. Thus, to prove existence of $\pi$ we can try to do it via linear algebra. We are typically used to right eigenvectors: If $A$ is an $n \times n$ matrix, $Ax = \lambda x$ has a non-zero solution $x$ iff $\lambda$ is an eigenvalue and $x$ is a corresponding eigenvector. $\det(A - \lambda I)$ is the characteristic polynomial; its roots are eigenvalues. In general, eigenvalues need not be real. There are two well-known situations when $A$ has real eigenvalues.

- If $A$ is symmetric, then all eigenvalues are real.
- If $A$ is a symmetric positive semi-definite matrix then all eigenvalues are $\geq 0$.

But $P$ is not symmetric in the general case. However $P$ is non-negative. Note that $P\mathbf{1} = \mathbf{1}$ since $P$ is a stochastic matrix and hence $P$ has an eigenvalue of 1 and a right eigenvector which is the all ones vector but we are looking for a left eigen vector. Note that for a matrix $A$ the eigenvalues of $A$ and its transpose $A^T$ are the same (due to the characteristic polynomial characterization) and the right eigenvectors of $A$ are the left eigenvectors of $A^T$.

**Theorem 2** (Perron)**.** *Let $A$ be a non-negative matrix (i.e., $A_{ij} \geq 0$ for all $i, j$). Then:*

1. *$A$ has a real positive eigenvalue $\lambda_0 > 0$ and corresponding positive eigenvector $v > 0$ such that any other eigenvalue $\lambda$ (can be complex) satisfies $|\lambda| < \lambda_0$; hence $\lambda_0$ is the unique largest eigenvalue.*

2. *$v$ is the unique non-negative vector (up to scaling) such that $Av = \lambda_0 v$.*

3. *Every other eigenvector has at least one non-positive coordinate.*

Perron's theorem requires $A > 0$ (strictly positive) while $P$ for a Markov chain satisfies $P \geq 0$. Although one can derive properties for non-negative matrices via Perron's theorem from limits of positive matrices, there are subtelties for general non-negative matrices. Frobenius generalized Perron's theorem to a class of matrices including ones relevant to us.

**Definition 7.** $A$ is an $n \times n$ matrix with $A \geq 0$ (i.e., $A_{ij} \geq 0$ for all $i, j$). We say $A$ is **irreducible** if the corresponding weighted directed graph is strongly connected.

**Theorem 3** (Perron-Frobenius)**.** *Let $A \geq 0$ and irreducible. Then $A$ has a positive eigenvalue $\lambda_0 > 0$ and all other eigenvalues $\lambda$ satisfy $|\lambda| \leq \lambda_0$. There is a positive eigenvector $v > 0$ such that $Av = \lambda_0 v$, and the following hold for $\lambda_0$ and $v$:*

1. *$v$ is the unique eigenvector associated with $\lambda_0$. That is, if $Ax = \lambda_0 x$, then $x = \alpha v$ for some scalar $\alpha \geq 0$.*

2. *$v$ is the only eigenvector with strictly positive coordinates.*

**Corollary 1.** *The largest real eigenvalue of an irreducible matrix $A \geq 0$ has a positive left eigenvector $\pi$. $\pi$ is unique up to scaling and is the only non-zero vector that satisfies $\pi A = \lambda_0 \pi$.*

*Proof.* Consider $A^T$. $A^T \geq 0$ and irreducible. $A^T$ has same eigenvalues as $A$. $\pi$ is the right eigenvector of $A^T$. □

Now we can apply the preceding to Markov chains. Consider stochastic matrix $P$ from an irreducible Markov chain: We saw that $\lambda = 1$ is a eigenvalue since $P\mathbf{1} = \mathbf{1}$. Since $P$ is stochastic we can see that $Ax \leq x$ and hence 1 is the largest positive real eigenvalue. Thus the left eigenvector $\pi > 0$ corresponding to the eigenvalue 1 is unique by the preceding theorem/corollary. If $\pi > 0$ we can noramalize it to be a probability distribution. Uniqueness of $\pi$ follows from the fact that the eigenvector for $\lambda = 1$ is unique up to scaling.

**Periodic and aperiodic chains:** We established existence of a stationary distribution via irreducability. When we have aperiodicity we can prove a stronger property. For this we note that if $P$ is aperiodic and connected then for some sufficiently large $t$, $P^t > 0$. This is because if the gcd of the walk lengths in $G$ is 1 then there is some integer $K$ such that for all $t \geq K$, there is a walk of length $t$ from $i$ to $j$ for any $i, j$. This implies that $P^t > 0$. Such matrices are called primitive. When $P^t$ is strictly positive we can use Perron's theorem which is stronger and guarantees that all eigen values are strictly smaller than $\lambda_0$ (they could be complex). One can then use this gap to show that $\pi(0)P^t$ converges to $\pi$ as $t \to \infty$.

Perron's theorem is classical and there are many sources. See Kents notes for one. It is not that long.

## 2.2 A second proof

This is from the book by Blum, Hopcroft, Kannan.

Let $\pi_t$ be the distribution after $t$ steps starting with $\pi(0)$. We have $\pi(t) = \pi(0)P^t$

Define $a(t) = \frac{1}{t+1}(\pi(0) + \pi(1) + \cdots + \pi(t-1))$ be the time-averaged distribution. Note that this is also a probability distribution.

**Theorem 4.** *Let $G$ be an irreducible Markov chain. There is a unique probability distribution $\pi$ such that $\pi P = \pi$. Moreover, for any starting distribution, $\lim_{t \to \infty} a(t)$ exists and equals $\pi$.*

**Lemma 4.** *Let $P$ be the transition matrix of a connected Markov chain. The matrix $A = [P - I \mid \mathbf{1}]$ obtained by augmenting $P - I$ by an all-ones column has rank $n$.*

*Proof.* Note that $A$ is a $n \times (n+1)$ matrix. Suppose $\mathrm{rank}(A) < n$. Then let $S$ be the null space of $A$, i.e., $S = \{y \in R^{n+1} : Ay = 0\}$. Then $\dim(S) \geq 2$. Each row of $P$ sums to 1, so each row of $P - I$ sums to 0. Thus $(\mathbf{1}, 0) \in S$. If $\dim(S) \geq 2$, there exists a vector $(x, \alpha) \in S$ orthogonal to $(\mathbf{1}, 0)$.

Orthogonality implies $\sum_i x_i = 0$. Since $A \cdot (x, \alpha) = 0$, we have $(P - I)x = \alpha\mathbf{1}$. This means $x_i = \sum_j P_{ij}x_j + \alpha$ for each $i \in [n]$.

Let $x_k$ have the max value among $x_1, x_2, \ldots, x_n$. Some $\ell \neq k$ has $x_\ell < x_k$ since $\sum_i x_i = 0$ and $x \neq 0$. By connectedness of $G$, there exists some $\ell$ such that $(k, \ell)$ is an edge and $x_\ell < x_k$. But we have $x_k = \sum_j P_{kj}x_j + \alpha$ which implies that $\alpha > 0$. Similarly, by considering the min value among $x_1, x_2, \ldots, x_n$ we can derive that $\alpha < 0$. This is a contradiction. □

Consider $a(t)P - a(t)$.

$$\begin{aligned} a(t)P - a(t) &= \frac{1}{t}(\pi(0)P + \pi(1)P + \ldots + \pi(t-1)P) - a(t) \\ &= \frac{1}{t}(\pi(1) + \pi(2) + \ldots + \pi(t)) - \frac{1}{t}(\pi(0) + \ldots + \pi(t-1)) \\ &= \frac{1}{t}(\pi(t) - \pi(0)) \end{aligned}$$

Define $b(t) = a(t)(P - I) = \frac{1}{t}(\pi(t) - \pi(0))$. Then $\|b(t)\|_\infty \leq \frac{2}{t} \to 0$ as $t \to \infty$.

By the preceding lemma, $A = [P - I \mid \mathbf{1}]$ has rank $n$. Since $a(t)(P - I) = b(t)$, we have $a(t)[P - I \mid \mathbf{1}] = [b(t) \mid 1]$. Consider the $n \times n$ sub-matrix $B$ of $A$ obtained by ignoring the first column of $A$; $B$ has rank $n$. Let $c(t)$ be obtained from $b(t)$ by removing the first entry. Then $a(t)B = [c(t, 1]$. Since $B$ is invertible, we have $a(t) = [c(t), 1]B^{-1}$. Since $b(t) \to 0$, we have $c(t) \to 0$ and hence $a(t) \to [\mathbf{0}, 1]B^-1$.

Thus, $\lim_{t\to\infty} a(t) = \pi$ where $\pi = [\mathbf{0}, 1]B^-1$. $\pi$ is unique because we showed that starting with any distribution $\pi_0$, $a(t)$ converges to $\pi$. If there existed another stationary distribution $\pi'$, then starting at $\pi'$ we would have $a(t) = \pi'$ for all $t$.

A useful lemma is the following.

**Lemma 5.** *Suppose $P$ corresponds to an irreducible chain. If we have a distribution $\pi$ such that $\pi_i P_{ij} = \pi_j P_{ji}$ for all $i, j$, then $\pi P = \pi$.*

*Proof.* Exercise. □

## 2.3 Another Proof

This is from the Levin-Peres book.

For $i \in V$, define:

$$\tau_i = \min\{t \geq 0 : X_t = i\}$$

as the first hitting time for $i$, and

$$\tau_i^+ = \min\{t \geq 1 : X_t = i\}$$

which is the first hitting time not counting the initial state.

When $X_0 = i$, we call $\tau_i^+$ the first return time. The expected value $h_{ij} = E_i[\tau_j^+]$ is the expected time to reach $j$ starting at $X_0 = i$. In the rest of this section we will be interested in various quantities conditioned on $X_0 = i$. We use $E_i$ and $\Pr_i$ to denote these quantities.

**Lemma 6.** *For any states $i, j$ of an irreducible chain, $h_{ij} = E_i[\tau_j^+] < \infty$.*

*Proof.* Since the chain is strongly connected and $P_{ij} > 0$ for $(i,j) \in E$, there exists $\epsilon > 0$ and an integer $r$ such that for any two states $u, v$, $\Pr_u[X_\ell = v] \geq \epsilon$. This is because we can take a path of length at most $n$ from $u$ to $v$ and multiply the probabilities along that path to see that it is some non-zero value.

Thus, for any value of $X_t$, the probability of hitting state $j$ between $t$ and $t + r$ is at least $\epsilon$. Hence, for $k \geq 0$, we have:

$$\Pr_i[\tau_j > kr] \leq (1-\epsilon)\Pr_i[\tau_j > (k-1)r]$$

Therefore:

$$\Pr_i[\tau_j > kr] \leq (1-\epsilon)^k$$

Now, if $Z$ is a non-negative random variable:

$$E[Z] = \sum_{t=0}^{\infty} \Pr[Z > t]$$

We have $\Pr[\tau_j > t]$ is a decreasing function of $t$. Hence:

$$\begin{aligned} E_i[\tau_j] &= \sum_{t=0}^{\infty} \Pr_i[\tau_j > t] \\ &\leq \sum_{k=0}^{\infty} r \Pr_i[\tau_j > kr] \\ &\leq r \sum_{k=0}^{\infty} (1-\epsilon)^k < \infty \end{aligned}$$

□

**Existence of a stationary distribution:** This proof has the nice feature that we can construct the stationary distribution somewhat explicitly, which also implies that $\pi_i = \frac{1}{h_{ii}}$ for each $i$ which is intuitive.

Let $k$ be an *arbitrary state* of the irreducible chain. For any $i \in V$, define:

$$\pi'_i = E_k[\text{number of visits to } i \text{ before returning to } k] = \sum_{t=0}^{\infty} \Pr_k[X_t = i, \tau_j^+ > t]$$

Note that $\pi'_k = 1$ by the above definition.

**Lemma 7.** *Let $\pi'$ be defined as above. Then $\pi'$ satisfies $\pi' P = \pi'$ and $\pi'/h_{kk}$ is a stationary distribution.*

In particular it shows that $\pi'_k/h_{kk} = 1/h_{kk}$. Note that the lemma applies for every $k$. It does not directly prove that we get the same stationary distribution if we use different states $k$ but if you knew that the stationary distribution is unique then you would be able to conclude that $\pi_i = 1/h_{ii}$ for all $i$.

Now we prove the lemma. For any $i$ we have $\pi'_i \leq h_{kk} < \infty$ (which we have established previously). We will check that $\pi'$ is stationary. Fix state $j$. From definition of $\pi'_i$,

We have

$$\sum_i \pi'_i P(i,j) = \sum_i \sum_{t=0}^{\infty} \Pr_k[X_t = i, \tau_k^+ > t] P(i,j)$$

Since the event $\{\tau_k^+ \geq t+1\} = \{\tau_k^+ > t\}$ is determined by $X_0, \ldots, X_t$,

$$\Pr_k[X_t = i, X_{t+1} = j, \tau_k^+ \geq t+1] = \Pr_k[X_t = i, \tau_k^+ \geq t+1] P(i,j)$$

Reversing the order of summation in the first equality and using the preceding identity we get

$$\sum_i \pi'_i P(i,j) = \sum_{t=0}^{\infty} \Pr_k[X_{t+1} = j, \tau_k^+ \geq t+1] = \sum_{t=1}^{\infty} \Pr_k[X_t = j, \tau_k^+ \geq t]$$

The expression $\sum_{t=1}^{\infty} \Pr_k[X_t = j, \tau_k^+ \geq t]$ is very similar to the definition of $\pi'_j$ and our goal is to show that it is indeed the same which would verify the stationarity of $\pi'$.

$$\begin{aligned}
\sum_{t=1}^{\infty} \Pr_k[X_t = j, \tau_k^+ \geq t] &= \pi'_j - \Pr_k[X_0 = j, \tau_k^+ \geq 0] + \sum_{t=1}^{\infty} \Pr_k[X_t = j, \tau_k^+ = t] \\
&= \pi'_j - \Pr_k[X_0 = j] + \Pr_k[X_{\tau_k^+} = j] \\
&= \pi'_j
\end{aligned}$$

We want to justify the last inequality by considering two cases.

- $j = k$. Since $X_0 = k$ and $X_{\tau_k^+} = k$, the terms $\Pr_k[X_0 = j]$ and $\Pr_k[X_{\tau_k^+} = j]$ are 1 and cancel out.

- $j \neq k$. Both terms are 0.

Finally, to get a probability measure we normalize by $\sum_i \pi'_i = E_k[\tau_k^+] = h_{kk}$. Thus $\pi = \pi'/h_{kk}$ is a stationary distribution.

## 3 Application to PageRank

The early approach of Google to rank web pages was based on using the link information that was in the pages. This was done to avoid the deficiencies of previous approaches that were based on manually classifying (Yahoo and AltaVista and others) and deficiencies of keyword search due to spam and other reasons.

The web graph is a directed graph where each web page corresponds to a node in a graph (this is a somewhat crude approximation) and a link in a page to another page creates a natural arc. Links encode information that gives information on how important pages are. The goal is to create a ranking of webpages globally; use ranking + query words later for personalized ranking.

How should one rank webpages in terms of importance? A simple approach is assign a score based on how many other links point to a webpage. $\text{score}(u) = \sum_{v:(v,u)\in E} 1$ This is easy to spam. A better approach is to score based on the importance of incoming pages:

$$\text{score}(u) = \sum_{v:(v,u)\in E} \frac{\text{score}(v)}{\text{out-deg}(v)}$$

This is a recursive definition. Main question: Does score($v$) exist? Can normalize scores (since scaling does not violate the equation). Assume $\sum_{u \in V} x_u = 1$, so $x$ is a probability distribution. This looks like the stationary distribution of a random walk on the web graph! But the web graph may not be ergodic.

The Brin-Page trick is to create an ergodic chain by considering a new graph $H$ which is a convex combination of the webgraph and a complete bipartite graph. Thus, if $P$ is the matrix corresponding to $G$ and $Q$ is the matrix corresponding to the complete directed graph on $V$ we let

$$P' = (1 - \epsilon)P + \epsilon Q$$

This corresponding to a random walk where with probability $\epsilon$, jump to a random webpage; with probability $1 - \epsilon$, follow a random outgoing link of the web graph. $P'$ corresponds to an ergodic chain, so there exists a unique stationary distribution $\pi$ for any fixed $\epsilon > 0$. This $\pi$ is the ranking.

**Computing PageRank:** How to compute $\pi$? Use the power method: $\pi(0)P^t \rightarrow \pi$ for any $\pi(0)$. Start with $\pi(0) = \frac{1}{n}\mathbf{1}$. Graph is sparse (average out-degree is 8-10), hence computation is not onerous. Mostly matrix-vector multiplication and numerical linear algebra and the process converges after a few iterations to a reasonable vector - note that the goal is not to actually compute a stationary distribution but only to find a ranking.