Markov Processes, Markov Chains and Stationary Distributions

with a worked example of the PageRank algorithm

1 Markov processes

Throughout, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and T is an index set ("time"), typically $T = \mathbb{N}_0 = \{0, 1, 2, \dots\}$ (discrete time) or $T = [0, \infty)$ (continuous time). The process takes values in a measurable space $(\mathcal{S}, \mathcal{E})$, the *state space*.

Definition 1.1 (Stochastic process, natural filtration)**.** A *stochastic process* is a family $X = (X_t)_{t \in T}$ of random variables $X_t : \Omega \to \mathcal{S}$. Its *natural filtration* is $\mathcal{F}_t = \sigma(X_s : s \le t)$, the information generated by the path up to time t.

Definition 1.2 (Markov process)**.** A stochastic process $(X_t)_{t \in T}$ with state space $(\mathcal{S}, \mathcal{E})$ is a *Markov process* if for all $s, t \in T$ with $s \le t$ and every $A \in \mathcal{E}$,

$$\mathbb{P}\big(X_t \in A \,\big|\, \mathcal{F}_s\big) \;=\; \mathbb{P}\big(X_t \in A \,\big|\, X_s\big) \qquad \text{almost surely.} \tag{1}$$

This is the *Markov property*: conditionally on the present state X_s, the future $(X_t)_{t \ge s}$ is independent of the past $(X_u)_{u < s}$.

Intuition. The present state is a *sufficient statistic* for predicting the future. Knowing how the process arrived at its current state adds nothing to the prediction: the process is "memoryless" with respect to its history, though not with respect to its current position.

The general definition allows arbitrary state spaces (e.g. $\mathcal{S} = \mathbb{R}^d$, as for Brownian motion). In the rest of this document we specialise to the case that is most important in applications such as PageRank: *discrete time and discrete state space*.

2 Markov chains

Definition 2.1 (Discrete-time Markov chain)**.** Let $\mathcal{S}$ be a *countable* set (finite or countably infinite), equipped with the power set σ-algebra. A sequence of $\mathcal{S}$-valued random variables $(X_n)_{n \in \mathbb{N}_0}$ is a *(discrete-time) Markov chain* if for every $n \ge 0$ and all states $i_0, \dots, i_{n-1}, i, j \in \mathcal{S}$,

$$\mathbb{P}\big(X_{n+1} = j \,\big|\, X_n = i,\, X_{n-1} = i_{n-1}, \dots, X_0 = i_0\big) \;=\; \mathbb{P}\big(X_{n+1} = j \,\big|\, X_n = i\big), \tag{2}$$

whenever the conditioning event has positive probability.

Because $\mathcal{S}$ is countable, conditioning on single states is meaningful and (2) is equivalent to (1) for $T = \mathbb{N}_0$. One distinguishes two cases.

Definition 2.2 (Finite state space Markov chain)**.** A Markov chain is a *finite state Markov chain* if $|\mathcal{S}| = N < \infty$. Without loss of generality $\mathcal{S} = \{1, 2, \dots, N\}$.

Definition 2.3 (Countably infinite state space Markov chain)**.** A Markov chain has a *countably infinite state space* if $\mathcal{S}$ is in bijection with $\mathbb{N}$, for example $\mathcal{S} = \mathbb{N}_0$ or $\mathcal{S} = \mathbb{Z}$.

Example 2.4 (Finite: a two-state chain)**.** Let $\mathcal{S} = \{0, 1\}$ (e.g. "machine works" / "machine broken"). From 0 the chain jumps to 1 with probability α; from 1 it jumps to 0 with probability β, where $\alpha, \beta \in (0, 1]$. Otherwise it stays put.

Example 2.5 (Countably infinite: random walk on $\mathbb{Z}$)**.** Let $\mathcal{S} = \mathbb{Z}$, $p \in (0, 1)$ and $X_{n+1} = X_n + \xi_{n+1}$ where the ξ_k are i.i.d. with $\mathbb{P}(\xi_k = +1) = p$, $\mathbb{P}(\xi_k = -1) = 1 - p$. Then (X_n) is a Markov chain on the countably infinite space $\mathbb{Z}$.

Remark 2.6. The finite and countably infinite cases share all the algebra below (sums over $\mathcal{S}$ simply become infinite series of non-negative terms). They differ in their *long-run behaviour*: a finite irreducible chain always has a stationary distribution, whereas a countably infinite one may not (the random walk of Example 2.5 has none). Probability mass can "escape to infinity".

3 Transition probabilities and the transition matrix

Definition 3.1 (Transition probabilities)**.** For a Markov chain (X_n) the *one-step transition probabilities at time n* are

$$p_{ij}(n) \;=\; \mathbb{P}\big(X_{n+1} = j \,\big|\, X_n = i\big), \qquad i, j \in \mathcal{S},$$

and the *m-step transition probabilities from time n* are

$$p_{ij}(n, n+m) \;=\; \mathbb{P}\big(X_{n+m} = j \,\big|\, X_n = i\big).$$

Definition 3.2 (Time-homogeneous Markov chain)**.** A Markov chain is *time homogeneous* if its one-step transition probabilities do not depend on n:

$$\mathbb{P}\big(X_{n+1} = j \,\big|\, X_n = i\big) \;=\; p_{ij} \qquad \text{for all } n \geq 0,\ i, j \in \mathcal{S}.$$

Then also $p_{ij}(n, n+m) =: p_{ij}^{(m)}$ depends only on the elapsed time m, not on the starting time n.

Intuition. Time homogeneity means the "rules of the game" do not change over time. The probability of moving from i to j is the same on Monday as on Friday. All chains below are assumed time homogeneous unless stated otherwise.

Definition 3.3 (State transition matrix)**.** For a time-homogeneous chain, the *(state) transition matrix* is

$$P = (p_{ij})_{i,j \in \mathcal{S}}, \qquad p_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i).$$

Row i is the conditional distribution of the next state given the current state i. For finite $\mathcal{S} = \{1, \ldots, N\}$, P is an $N \times N$ matrix. For countably infinite $\mathcal{S}$ it is an infinite matrix indexed by $\mathcal{S} \times \mathcal{S}$.

Definition 3.4 (Stochastic matrix)**.** A finite or infinite matrix $P = (p_{ij})_{i,j \in \mathcal{S}}$ is *(row-)stochastic* if

$$\text{(i)}\ \ p_{ij} \geq 0 \ \ \forall i, j, \qquad \text{(ii)}\ \ \sum_{j \in \mathcal{S}} p_{ij} = 1 \ \ \forall i.$$

Every transition matrix is stochastic. Conversely, every stochastic matrix P together with an initial distribution μ_0 defines a Markov chain via

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) \;=\; \mu_0(i_0)\, p_{i_0 i_1}\, p_{i_1 i_2} \cdots p_{i_{n-1} i_n}.$$

Definition 3.5 (Time-inhomogeneous transition matrices)**.** If the chain is not time homogeneous one obtains a sequence $P(n) = (p_{ij}(n))$ of stochastic matrices, one for every time step. The m-step matrix is then the ordered product $P(n, n+m) = P(n)\, P(n+1) \cdots P(n+m-1)$, which generally depends on n. Time homogeneity is exactly the statement $P(n) \equiv P$, in which case $P(n, n+m) = P^m$.

Example 3.6 (Matrices for the examples)**.** For Example 2.4 and Example 2.5,

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}, \qquad P = \begin{pmatrix} \ddots & \ddots & & & \\ \ddots & 0 & p & & \\ & 1-p & 0 & p & \\ & & 1-p & 0 & \ddots \\ & & & \ddots & \ddots \end{pmatrix},$$

i.e. in the random walk $p_{i,i+1} = p$, $p_{i,i-1} = 1-p$, and $p_{ij} = 0$ otherwise.

Distributions as row vectors. Write the distribution of X_n as a row vector $\mu_n = (\mu_n(i))_{i\in\mathcal{S}}$ with $\mu_n(i) = \mathbb{P}(X_n = i)$. Conditioning on X_n gives

$$\mu_{n+1}(j) = \sum_{i\in\mathcal{S}} \mathbb{P}(X_n = i)\,\mathbb{P}(X_{n+1} = j \mid X_n = i) = \sum_{i\in\mathcal{S}} \mu_n(i)\,p_{ij}, \quad \text{i.e.} \quad \mu_{n+1} = \mu_n P,$$

and by induction $\mu_n = \mu_0 P^n$. (For infinite $\mathcal{S}$, the sums converge because all terms are non-negative and $\sum_i \mu_n(i) = 1$.)

4 The Chapman–Kolmogorov equations

Theorem 4.1 (Chapman–Kolmogorov equations)**.** *Let (X_n) be a Markov chain on a countable state space $\mathcal{S}$. For all times $k \le l \le n$ and all states $i, j \in \mathcal{S}$,*

$$p_{ij}(k,n) \;=\; \sum_{r\in\mathcal{S}} p_{ir}(k,l)\,p_{rj}(l,n). \tag{3}$$

In the time-homogeneous case this reads, for all $m, n \ge 0$,

$$p_{ij}^{(m+n)} \;=\; \sum_{r\in\mathcal{S}} p_{ir}^{(m)}\,p_{rj}^{(n)}, \qquad \textit{equivalently} \qquad P^{(m+n)} = P^{(m)}P^{(n)}, \tag{4}$$

and consequently $P^{(n)} = P^n$: the n-step transition matrix is the n-th matrix power of P.

Proof. Partition the event $\{X_n = j\}$ according to the state at the intermediate time l and use the Markov property:

$$\begin{aligned} p_{ij}(k,n) &= \mathbb{P}(X_n = j \mid X_k = i) = \sum_{r\in\mathcal{S}} \mathbb{P}(X_n = j,\, X_l = r \mid X_k = i) \\ &= \sum_{r\in\mathcal{S}} \mathbb{P}(X_l = r \mid X_k = i)\,\mathbb{P}(X_n = j \mid X_l = r,\, X_k = i) \\ &= \sum_{r\in\mathcal{S}} \mathbb{P}(X_l = r \mid X_k = i)\,\mathbb{P}(X_n = j \mid X_l = r) = \sum_{r\in\mathcal{S}} p_{ir}(k,l)\,p_{rj}(l,n), \end{aligned}$$

where the third equality is the Markov property (the past value X_k is irrelevant once X_l is known). Terms with $\mathbb{P}(X_l = r \mid X_k = i) = 0$ are simply omitted. In the homogeneous case, $p_{ij}(k,n) = p_{ij}^{(n-k)}$, which yields (4). Since $P^{(1)} = P$, induction gives $P^{(n)} = P^n$. □

Intuition. To go from i to j in $m+n$ steps you must be *somewhere*, say at r, after m steps. Because of the Markov property, the first m steps and the last n steps are "glued together" at r independently, so the probabilities multiply, and then one sums over all intermediate states r. Matrix multiplication is exactly this "sum over all intermediate stops".

Remark 4.2 (Continuous time)**.** For a continuous-time Markov process with transition probabilities $p_{ij}(s,t) = \mathbb{P}(X_t = j \mid X_s = i)$ the same argument gives $P(s,u) = P(s,t)P(t,u)$ for $s \le t \le u$. In the homogeneous case $P(t+s) = P(t)P(s)$, so $(P(t))_{t\ge 0}$ is a semigroup; differentiating leads to the Kolmogorov forward and backward equations $P'(t) = P(t)Q$ and $P'(t) = QP(t)$ with generator $Q = P'(0)$.

5 Stationary and equilibrium distributions

We need some structural notions to state existence and uniqueness results.

Definition 5.1 (Irreducibility, period, recurrence)**.** Let P be the transition matrix of a homogeneous chain.

(a) State j is *accessible* from i ($i \to j$) if $p_{ij}^{(n)} > 0$ for some $n \ge 0$. States i, j *communicate* if $i \to j$ and $j \to i$. The chain is *irreducible* if all states communicate.

(b) The *period* of i is $d(i) = \gcd\{n \ge 1 : p_{ii}^{(n)} > 0\}$. State i is *aperiodic* if $d(i) = 1$. In an irreducible chain all states have the same period.

(c) Let $T_i = \inf\{n \ge 1 : X_n = i\}$ be the first return time to i. State i is *recurrent* if $\mathbb{P}_i(T_i < \infty) = 1$ and *transient* otherwise. A recurrent state is *positive recurrent* if $m_i := \mathbb{E}_i[T_i] < \infty$ and *null recurrent* if $m_i = \infty$.

Here $\mathbb{P}_i$ and $\mathbb{E}_i$ refer to the chain started in $X_0 = i$.

Definition 5.2 (Stationary (invariant) distribution)**.** A probability vector $\pi = (\pi_i)_{i\in\mathcal{S}}$, i.e. $\pi_i \ge 0$ and $\sum_{i\in\mathcal{S}} \pi_i = 1$, is a *stationary distribution* of P if

$$\pi P = \pi, \qquad \text{i.e.} \qquad \pi_j = \sum_{i\in\mathcal{S}} \pi_i\, p_{ij} \quad \text{for all } j \in \mathcal{S}. \tag{5}$$

Equivalently, π is a left eigenvector of P with eigenvalue 1 whose entries are non-negative and sum to one.

If $X_0 \sim \pi$ then $\mu_1 = \pi P = \pi$, and inductively $\mu_n = \pi$ for every n. In fact the whole process is then *stationary*: $(X_{n+k})_{n\ge 0}$ has the same law as $(X_n)_{n\ge 0}$ for every k.

Definition 5.3 (Equilibrium (limiting) distribution)**.** A probability vector π is an *equilibrium* or *limiting distribution* of the chain if, for every initial distribution μ_0,

$$\lim_{n\to\infty} \mu_0 P^n = \pi \quad \text{(componentwise)}, \qquad \text{in particular} \qquad \lim_{n\to\infty} p_{ij}^{(n)} = \pi_j \ \ \forall i, j \in \mathcal{S}.$$

For finite $\mathcal{S}$ the last statement says that P^n converges to the matrix $\mathbf{1}\pi$ all of whose rows equal π.

Remark 5.4 (Stationary versus equilibrium)**.** An equilibrium distribution is always stationary: if $\mu_0 P^n \to \pi$ then $\pi = \lim \mu_0 P^{n+1} = (\lim \mu_0 P^n)P = \pi P$ (passing the limit inside is justified for finite $\mathcal{S}$, and in general by dominated convergence). The converse fails:

- In Example 2.4 with $\alpha = \beta = 1$ the chain deterministically alternates $0, 1, 0, 1, \ldots$. Then $\pi = (\frac{1}{2}, \frac{1}{2})$ is stationary, but $p_{00}^{(n)}$ alternates between 1 and 0 and has no limit (period 2).
- A reducible chain may have many stationary distributions (one per closed class and all convex combinations), hence no limit independent of the start.

Many texts use "equilibrium distribution" as a synonym for "stationary distribution". The theorems below say precisely when the two notions coincide.

Theorem 5.5 (Existence and uniqueness)**.** *Let P be irreducible.*

(i) P has a stationary distribution if and only if it is positive recurrent. In that case the stationary distribution is unique and

$$\pi_i \;=\; \frac{1}{\mathbb{E}_i[T_i]} \qquad \text{for all } i \in \mathcal{S}.$$

(ii) If $\mathcal{S}$ is finite, P is automatically positive recurrent, so a unique stationary distribution exists, and all $\pi_i > 0$.

Theorem 5.6 (Convergence to equilibrium)**.** *Let P be irreducible, aperiodic and positive recurrent with stationary distribution π. Then for every initial distribution μ_0,*

$$\lim_{n\to\infty} \mathbb{P}(X_n = j) = \pi_j \quad \text{for all } j, \qquad \text{indeed} \qquad \sum_{j\in\mathcal{S}} \big|(\mu_0 P^n)_j - \pi_j\big| \xrightarrow[n\to\infty]{} 0.$$

So π is the equilibrium distribution. For finite $\mathcal{S}$ the convergence is geometric: there are $C < \infty$, $\rho < 1$ with $\sum_j |(\mu_0 P^n)_j - \pi_j| \le C\rho^n$.

Theorem 5.7 (Ergodic theorem)**.** *Let P be irreducible and positive recurrent with stationary distribution π. For any initial distribution and any $j \in \mathcal{S}$,*

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbf{1}\{X_k = j\} \;\xrightarrow[n\to\infty]{a.s.}\; \pi_j.$$

(Aperiodicity is not *needed here.)*

Remark 5.8 (Why aperiodicity is not needed in Theorem 5.7)**.** Theorems 5.6 and 5.7 answer different questions. Theorem 5.6 concerns the distribution *at a fixed time n*: does $\mathbb{P}(X_n = j)$ converge? Theorem 5.7 concerns the *fraction of time* a single trajectory spends in j: does $\frac{1}{n}\#\{k < n : X_k = j\}$ converge? Periodicity affects *when* the chain can be in a state, not *how often* it is there, and a time average smooths out the timing.

Example. For the deterministic flip chain $P = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ started at 0, the path is $0, 1, 0, 1, \ldots$. Here $\mathbb{P}(X_n = 0)$ alternates $1, 0, 1, 0, \ldots$ and does not converge, but the fraction of time spent in 0 up to time n equals $\frac{1}{2}$ or $\frac{1}{2} + \frac{1}{2n}$, which tends to $\frac{1}{2} = \pi_0$. This is the familiar fact that the sequence $1, 0, 1, 0, \ldots$ diverges while its running averages converge to $\frac{1}{2}$.

Remark 5.9 (Why Theorem 5.6 does need aperiodicity)**.** If an irreducible chain has period $d \ge 2$, its state space splits into d *cyclic classes* $C_0 \to C_1 \to \cdots \to C_{d-1} \to C_0$, visited in a fixed rotation. For $i \in C_0$ and $j \in C_r$ one has $p_{ij}^{(m)} = 0$ unless $m \equiv r \pmod d$, while along the correct phase $p_{ij}^{(nd+r)} \to d\,\pi_j$. The chain never forgets its starting phase, so $\mathbb{P}(X_n = j)$ cannot converge. Averaged over one period, the values $0, \ldots, 0, d\pi_j$ give exactly π_j, which is what the time average in Theorem 5.7 sees.

Taking expectations in Theorem 5.7 (by bounded convergence, since $0 \le V_n/n \le 1$) yields the *Cesàro convergence*

$$\frac{1}{n}\sum_{k=0}^{n-1} p_{ij}^{(k)} \;\xrightarrow[n\to\infty]{}\; \pi_j \qquad \text{for all } i, j \in \mathcal{S},$$

for every irreducible positive recurrent chain, periodic or not. Aperiodicity is precisely the extra condition that upgrades this averaged convergence to convergence of $p_{ij}^{(n)}$ itself.

Example 5.10 (The two-state chain)**.** For Example 2.4, $\pi P = \pi$ reads $\pi_0(1-\alpha) + \pi_1\beta = \pi_0$, i.e. $\alpha\pi_0 = \beta\pi_1$. With $\pi_0 + \pi_1 = 1$:

$$\pi = \Big(\frac{\beta}{\alpha+\beta},\ \frac{\alpha}{\alpha+\beta}\Big).$$

If $\alpha + \beta < 2$ the chain is aperiodic and $P^n \to \begin{pmatrix}\pi\\ \pi\end{pmatrix}$; in fact $P^n = \mathbf{1}\pi + (1-\alpha-\beta)^n\big(I - \mathbf{1}\pi\big)$, so the error decays like $|1-\alpha-\beta|^n$.

6 What the stationary distribution tells us

Intuition: four ways to read π.

1. **Invariance ("statistical equilibrium").** If the state is drawn from π, then after one step (and after any number of steps) it is again distributed as π. Individual trajectories keep moving, but the *distribution* no longer changes. Think of a large population of independent walkers: the fraction of walkers in each state stays constant even though each walker keeps wandering.

2. **Probability flow balance.** Writing $\pi_j = \sum_i \pi_i p_{ij}$ and using $\sum_i p_{ji} = 1$ gives

$$\underbrace{\sum_{i\neq j} \pi_j\, p_{ji}}_{\text{flow out of } j} \;=\; \underbrace{\sum_{i\neq j} \pi_i\, p_{ij}}_{\text{flow into } j}.$$

   In equilibrium, as much probability flows into each state per step as flows out. (The stronger condition $\pi_i p_{ij} = \pi_j p_{ji}$, *detailed balance*, says the flow balances across every single edge. It characterises reversible chains.)

3. **Long-run fraction of time.** By the ergodic theorem, π_j is the long-run proportion of time a single trajectory spends in j. Equivalently $\pi_j = 1/\mathbb{E}_j[T_j]$: a state visited often (short mean return time) gets a large weight. π therefore measures the *long-term importance* or *centrality* of each state under the dynamics.

4. **Forgetting the initial condition.** If the chain is also aperiodic, $\mathbb{P}(X_n = j) \to \pi_j$ whatever X_0 was. After a long time, the chain has "forgotten" where it started. π is the only distribution that can describe the chain in the far future.

Interpretation 3 is the key to PageRank: a web page is important if a random surfer spends a large fraction of time on it.

7 Worked example: the PageRank algorithm

7.1 The random surfer model

Model the web as a directed graph on N pages, with an edge $i \to j$ if page i links to page j. Let $L(i)$ be the number of out-links of i. A *random surfer* moves as follows:

- with probability d (the *damping factor*, classically $d = 0.85$) she clicks a uniformly random out-link of the current page;
- with probability $1 - d$ she gets bored and *teleports* to a page chosen uniformly from all N pages.

Let S be the *link matrix*, $S_{ij} = 1/L(i)$ if $i \to j$ and 0 otherwise. (A page with no out-links, a *dangling node*, gets the row $\frac{1}{N}\mathbf{1}^\top$.) The surfer is a time-homogeneous Markov chain with transition matrix, the *Google matrix*,

$$G \;=\; d\,S \;+\; \frac{1-d}{N}\,\mathbf{1}\mathbf{1}^\top.$$

Since S is stochastic, so is G. Moreover every entry of G is at least $(1-d)/N > 0$, so the chain is *irreducible and aperiodic* on a finite space. By Theorems 5.5 and 5.6, G has a unique stationary distribution π, which is also the equilibrium distribution. The **PageRank** of page j is π_j. Componentwise,

$$\pi_j \;=\; \frac{1-d}{N} \;+\; d\sum_{i:\, i\to j} \frac{\pi_i}{L(i)}. \tag{6}$$

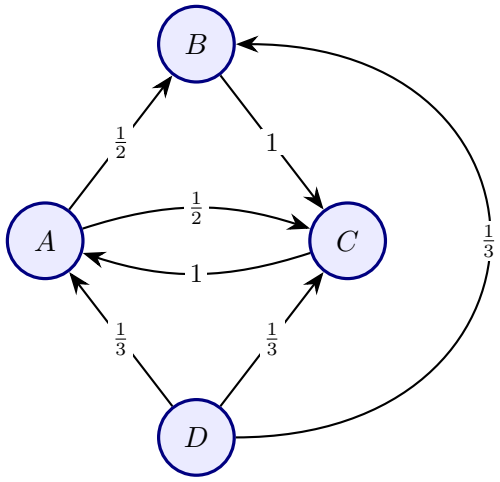
A page is important if important pages link to it, with each page's "vote" split equally among its out-links.

Purpose of the teleportation term. Without it, π might not be unique (the web graph is far from irreducible), might not attract all starting distributions (periodicity), or might put all its mass on "sink" clusters that can be entered but not left. Teleportation makes G strictly positive, which guarantees all good properties at once (Perron–Frobenius).

7.2 A four-page web

Consider pages A, B, C, D with links

$$A \to B, \quad A \to C, \qquad B \to C, \qquad C \to A, \qquad D \to A, \quad D \to B, \quad D \to C.$$



Ordering the states as (A, B, C, D), the link matrix is

$$S = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix}.$$

Step 1: without damping ($d = 1$). First look at the pure link chain S. Page D has no in-links, so once the surfer leaves D she never returns: D is transient, and $\{A, B, C\}$ is a closed class. Solving $\pi S = \pi$:

$$\pi_A = \pi_C + \tfrac{1}{3}\pi_D, \quad \pi_B = \tfrac{1}{2}\pi_A + \tfrac{1}{3}\pi_D, \quad \pi_C = \tfrac{1}{2}\pi_A + \pi_B + \tfrac{1}{3}\pi_D, \quad \pi_D = 0,$$

so $\pi_A = \pi_C$, $\pi_B = \frac{1}{2}\pi_A$, and normalising,

$$\pi^{(d=1)} = \big(\tfrac{2}{5},\ \tfrac{1}{5},\ \tfrac{2}{5},\ 0\big).$$

Page D gets rank 0. Here the result is still unique (there is just one closed class, which is aperiodic since it contains cycles $A \to C \to A$ of length 2 and $A \to B \to C \to A$ of length 3). On the real web, however, there are many closed classes and dangling nodes, and S alone would not give a well-defined ranking.

Step 2: the Google matrix with $d = 0.85$. With $N = 4$, $(1-d)/N = 0.15/4 = \frac{3}{80} = 0.0375$, and

$$G = 0.85\,S + 0.0375\,\mathbf{1}\mathbf{1}^{\top} = \begin{pmatrix} \frac{3}{80} & \frac{37}{80} & \frac{37}{80} & \frac{3}{80} \\ \frac{3}{80} & \frac{3}{80} & \frac{71}{80} & \frac{3}{80} \\ \frac{71}{80} & \frac{3}{80} & \frac{3}{80} & \frac{3}{80} \\ \frac{77}{240} & \frac{77}{240} & \frac{77}{240} & \frac{3}{80} \end{pmatrix} \approx \begin{pmatrix} 0.0375 & 0.4625 & 0.4625 & 0.0375 \\ 0.0375 & 0.0375 & 0.8875 & 0.0375 \\ 0.8875 & 0.0375 & 0.0375 & 0.0375 \\ 0.3208 & 0.3208 & 0.3208 & 0.0375 \end{pmatrix}.$$

Each row sums to 1 and every entry is positive.

Step 3: solving $\pi G = \pi$ exactly. Using (6) with $d = \frac{17}{20}$:

$$\begin{aligned} \pi_A &= \tfrac{3}{80} + \tfrac{17}{20}\big(\pi_C + \tfrac{1}{3}\pi_D\big), \\ \pi_B &= \tfrac{3}{80} + \tfrac{17}{20}\big(\tfrac{1}{2}\pi_A + \tfrac{1}{3}\pi_D\big), \\ \pi_C &= \tfrac{3}{80} + \tfrac{17}{20}\big(\tfrac{1}{2}\pi_A + \pi_B + \tfrac{1}{3}\pi_D\big), \\ \pi_D &= \tfrac{3}{80}. \end{aligned}$$

(Equation (6) already uses $\sum_i \pi_i = 1$.) Nobody links to D, so it only receives teleportation mass: $\pi_D = \frac{3}{80} = 0.0375$. Then $\frac{3}{80} + \frac{17}{20} \cdot \frac{1}{3} \cdot \frac{3}{80} = \frac{77}{1600} =: c$, and the system becomes

$$\pi_A = c + 0.85\,\pi_C, \qquad \pi_B = c + 0.425\,\pi_A, \qquad \pi_C = c + 0.425\,\pi_A + 0.85\,\pi_B.$$

Substituting π_B into π_C: $\pi_C = 1.85\,c + 0.78625\,\pi_A$. Substituting into π_A:

$$\begin{aligned} & \pi_A = c + 0.85\,(1.85\,c + 0.78625\,\pi_A) \\ \Longrightarrow \quad & (1 - 0.6683125)\,\pi_A = 2.5725\,c \\ \Longrightarrow \quad & \pi_A = \frac{2.5725 \cdot 77/1600}{0.3316875} \approx 0.37325. \end{aligned}$$

Back-substitution gives $\pi_B \approx 0.20676$ and $\pi_C \approx 0.38250$. In exact form:

$$\pi = \Big(\tfrac{52822}{141520},\ \tfrac{29260}{141520},\ \tfrac{54131}{141520},\ \tfrac{5307}{141520}\Big) \approx (0.3732,\ 0.2068,\ 0.3825,\ 0.0375)$$

The entries sum to $141520/141520 = 1$, and one checks directly that $\pi G = \pi$.

Step 4: power iteration. For the real web (N in the billions) one cannot solve the linear system directly. Instead one exploits Theorem 5.6: start from any distribution, e.g. uniform $\mu_0 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, and iterate $\mu_{k+1} = \mu_k G$, which only needs sparse matrix–vector products with S since $\mu_k G = d\,\mu_k S + \frac{1-d}{N}\mathbf{1}^{\top}$.

k	$\mu_k(A)$	$\mu_k(B)$	$\mu_k(C)$	$\mu_k(D)$	$\sum$
0	0.2500	0.2500	0.2500	0.2500	1
1	0.3208	0.2146	0.4271	0.0375	1
2	0.4111	0.1845	0.3669	0.0375	1
3	0.3600	0.2229	0.3797	0.0375	1
4	0.3708	0.2011	0.3905	0.0375	1
5	0.3801	0.2057	0.3767	0.0375	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
∞	0.3732	0.2068	0.3825	0.0375	1

The iterates oscillate around and converge to π. The error shrinks at least by a factor $d = 0.85$ per step, because the second-largest eigenvalue of G satisfies $|\lambda_2| \leq d$. This explains the choice of d: smaller d means faster convergence but less weight on the actual link structure.

Step 5: interpretation.

- **Ranking:** C (0.383) > A (0.373) > B (0.207) > D (0.0375).
- C ranks highest: it receives links from *all* other pages, including the full vote of B.
- A is a close second although it has only two in-links (from C and D): it receives the *entire* vote of the top page C, which has only one out-link. Quality of in-links matters, not just quantity.
- B receives only half of A's vote plus a third of D's.
- D has no in-links and receives only the teleportation mass $(1-d)/N$, the minimal possible PageRank.
- In the long-run (ergodic) reading of π: a random surfer following this model spends about 38% of her time on C, 37% on A, 21% on B and under 4% on D. Equivalently, she returns to C on average every $1/0.3825 \approx 2.6$ clicks, but to D only every $1/0.0375 \approx 26.7$ clicks.
- Compared to the undamped solution $(0.4, 0.2, 0.4, 0)$, damping lifts D from 0, pulls the scores slightly towards uniform, and breaks the tie between A and C.

8 Ten influential applications of Markov chains

The concepts above (transition matrices, stationary distributions, convergence to equilibrium, ergodicity) appear across a remarkable range of fields. Below are ten of the most influential applications, with two key references for each.

8.1 Language and Markov's own first application

Markov introduced his chains partly to show that the law of large numbers does not need independence. In 1913 he applied them to the first 20,000 letters of Pushkin's *Eugene Onegin*, modelling the text as a two-state chain on {vowel, consonant} and estimating its transition probabilities [1]. Shannon later used higher-order Markov approximations of English to illustrate information sources. He defined the *entropy rate* of a stationary Markov source as $H = -\sum_i \pi_i \sum_j p_{ij} \log p_{ij}$, where the stationary distribution π weights the uncertainty of each row [2]. This idea underlies data compression and, via n-gram models, early statistical language modelling.

8.2 Markov chain Monte Carlo: the Metropolis–Hastings algorithm

Instead of computing the stationary distribution of a given chain, MCMC turns the problem around: *construct* a chain whose stationary distribution is a target π that is hard to sample from directly. Metropolis and co-authors introduced the idea to simulate a system of particles in statistical mechanics [3]. Hastings generalised it to arbitrary proposal distributions [4]. The acceptance rule enforces detailed balance, so π is stationary, and the ergodic theorem (Theorem 5.7) justifies estimating expectations under π by time averages along the chain. MCMC is one of the most widely used computational methods in science.

8.3 Gibbs sampling and Bayesian computation

The Gibbs sampler updates one coordinate at a time by drawing it from its conditional distribution given all the others. Geman and Geman introduced it for Bayesian image restoration with Markov random field priors [5]. Gelfand and Smith showed how it makes the computation of posterior marginals feasible for general Bayesian models [6]. This sparked the "MCMC revolution" in statistics, making Bayesian inference practical for complex hierarchical models.

8.4 Hidden Markov models and speech recognition

In a hidden Markov model (HMM) the Markov chain (X_n) is not observed directly; one only sees noisy emissions Y_n whose distribution depends on X_n. Baum and co-authors developed the forward–backward and Baum–Welch (expectation–maximisation) algorithms for estimating the transition matrix and emission probabilities [7]. Rabiner's tutorial made HMMs the standard tool in speech recognition for decades [8]. HMMs are also used for part-of-speech tagging, handwriting recognition and gene finding.

8.5 The Viterbi algorithm and error-correcting codes

A convolutional encoder is a finite-state machine, so the transmitted sequence follows a path through a Markov chain observed through a noisy channel. Viterbi introduced a dynamic programming algorithm that finds the most likely state sequence in time linear in the sequence length [9]. Forney showed that it is exactly maximum-likelihood sequence estimation on a trellis [10]. The Viterbi algorithm has decoded signals from deep-space probes and runs in billions of mobile phones, and it is the standard decoder for HMMs.

8.6 Markov decision processes and reinforcement learning

A Markov decision process adds actions and rewards to a Markov chain: the transition matrix depends on the action chosen in the current state. Bellman formulated the model and the dynamic programming (Bellman) equations for optimal policies [11]. This is the mathematical framework of reinforcement learning. A landmark example is the deep Q-network of Mnih et al., which learned to play Atari games at human level directly from pixels [12].

8.7 Diffusion models for generative AI

Diffusion models generate images, audio and video by reversing a Markov chain. A forward chain gradually adds Gaussian noise to data until it reaches the stationary distribution $\mathcal{N}(0, I)$. A neural network then learns the transitions of the reverse chain, which turns noise back into data. Sohl-Dickstein et al. introduced the idea, inspired by non-equilibrium thermodynamics [13]. Ho, Jain and Abbeel's denoising diffusion probabilistic models made it competitive with the best generative models [14], and the approach now underlies most state-of-the-art image generators.

8.8 The Ehrenfest urn model and the arrow of time

Paul and Tatjana Ehrenfest proposed a simple chain to resolve apparent paradoxes in Boltzmann's statistical mechanics [15]. N balls are distributed over two urns, and at each step a uniformly chosen ball switches urns. The state is the number of balls in the first urn, and the stationary distribution is Binomial$(N, \frac{1}{2})$. Kac analysed the model in detail [16]. Since $\pi_i = 1/\mathbb{E}_i[T_i]$, the mean time to return to the state with all balls in one urn is 2^N, astronomically large for physical N. The system *does* return, but practically never, which reconciles recurrence with observed irreversibility. The chain also has period 2, a natural example of the distinction between stationary and limiting distributions.

8.9 Molecular evolution and phylogenetics

DNA substitution is modelled by a continuous-time Markov chain on $\{A, C, G, T\}$ running along each branch of an evolutionary tree. Jukes and Cantor proposed the simplest such model, with equal substitution rates and a uniform stationary distribution [17]. Felsenstein showed how to compute the likelihood of a whole tree efficiently with his pruning algorithm, which made maximum likelihood phylogenetics practical [18]. These models are used to reconstruct the tree of life and to track the evolution and spread of viruses.

8.10 Card shuffling and the cutoff phenomenon

Shuffling a deck of cards is a random walk on the group of 52! permutations, whose stationary distribution is uniform. The question "how many shuffles are enough?" is a question about the speed of convergence in Theorem 5.6. Aldous and Diaconis developed strong stationary times to bound such mixing times and observed the *cutoff phenomenon*: the distance to uniformity stays near 1 and then drops abruptly [19]. Bayer and Diaconis analysed the riffle shuffle exactly and concluded that about seven riffle shuffles suffice for a 52-card deck [20]. Mixing-time theory is now central to the analysis of MCMC algorithms.

References

[1] A. A. Markov. An example of statistical investigation of the text *Eugene Onegin* concerning the connection of samples in chains. *Science in Context*, 19(4):591–600, 2006. English translation of the 1913 original in *Bulletin de l'Académie Impériale des Sciences de St.-Pétersbourg*.

[2] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423 and 27(4):623–656, 1948.

[3] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[4] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[5] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[6] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[7] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[8] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[9] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

[10] G. D. Forney, Jr. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[11] R. Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, 6(5):679–684, 1957.

[12] V. Mnih, K. Kavukcuoglu, D. Silver, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[13] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, PMLR 37:2256–2265, 2015.

[14] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020.

[15] P. Ehrenfest and T. Ehrenfest. Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem. *Physikalische Zeitschrift*, 8:311–314, 1907.

[16] M. Kac. Random walk and the theory of Brownian motion. *The American Mathematical Monthly*, 54(7):369–391, 1947.

[17] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

[18] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[19] D. Aldous and P. Diaconis. Shuffling cards and stopping times. *The American Mathematical Monthly*, 93(5):333–348, 1986.

[20] D. Bayer and P. Diaconis. Trailing the dovetail shuffle to its lair. *The Annals of Applied Probability*, 2(2):294–313, 1992.