Markov Chains to Stationary Distributions

Warm-up, Key Properties, and Why Detailed Balance Matters

Warm-up: the Markov property

Suppose $(Y_1, Y_2, Y_3, \dots)$ is a sequence of random variables taking values in a state space $\mathcal{S}$.

Markov property (principle):

$$\mathbb{P}(Y_{n+1} = j \mid Y_n = i, Y_{n-1}, \dots, Y_1) = \mathbb{P}(Y_{n+1} = j \mid Y_n = i).$$

- **Interpretation:** the *future* depends only on the *immediate present*.
- Everything before time n is "summarised" by the current state Y_n.

States: $\{1, 2, 3\}$. Transition probabilities are:

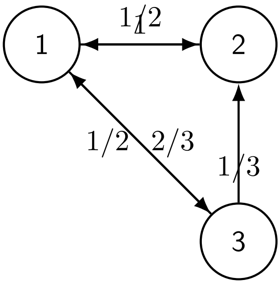
- From state 1: $1 \to 2$ with probability $1/2$, $1 \to 3$ with probability $1/2$, and $1 \to 1$ with probability 0.
- From state 2: $2 \to 1$ with probability 1, and $2 \to 2, 2 \to 3$ with probability 0.
- From state 3: choose probabilities that **sum to 1**. (We fix the earlier mismatch by setting)

$$3 \to 1 \text{ with } 2/3, \qquad 3 \to 2 \text{ with } 1/3, \qquad 3 \to 3 \text{ with } 0.$$

If we use the convention $P_{ij} = \mathbb{P}(Y_{n+1} = j \mid Y_n = i)$, then the transition matrix is

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \\ 2/3 & 1/3 & 0 \end{pmatrix}.$$

Warm-up picture: the directed graph of transitions



Reminder:
each row sums to 1 (outgoing probabilities).

Powers of P: what does P^2 mean?

Key fact (discrete-time Markov chains):

$$(P^n)_{ij} = \mathbb{P}(Y_{t+n} = j \mid Y_t = i).$$

- $P^2 = PP$ gives **two-step** transition probabilities.
- $P^3 = P^2 P$ gives **three-step** transition probabilities.
- More generally, P^n tells you what can happen in n **steps**.

How to think about $(P^2)_{ij}$: sum over all intermediate states k:

$$(P^2)_{ij} = \sum_{k \in \mathcal{S}} P_{ik} P_{kj}.$$

Definition: period and aperiodicity

For a state i, define its **return times**:

$$\{n \geq 1 : (P^n)_{ii} > 0\}.$$

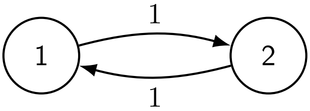
Period of state i:

$$d(i) = \gcd\{n \geq 1 : (P^n)_{ii} > 0\}.$$

- State i is **aperiodic** if $d(i) = 1$.
- A Markov chain is **aperiodic** if *every* state is aperiodic.

Intuition: if $d(i) = 2$, you can only return to i in even numbers of steps $\Rightarrow$ rigid cycling.

Example: a periodic 2-state chain (period 2)



From 1, you *must* go to 2, then *must* return to 1.
So return times to 1 are $\{2, 4, 6, \dots\}$ and $d(1) = 2$ (similarly $d(2) = 2$).

Definition: irreducibility

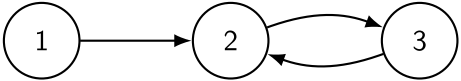
A Markov chain on $\mathcal{S}$ is **irreducible** if for every pair of states (i, j), there exists some $n \in \mathbb{N}$ such that

$$(P^n)_{ij} > 0.$$

- **In words:** it is possible to move from *any* state to *any other* state in a finite number of steps.
- Equivalently: the transition graph is **strongly connected**.

Why we care (sampling viewpoint): if the chain is not irreducible, it may get trapped in one region and *never visit* other valid states.

Example: not irreducible (a state you can never return to)



If you start in $\{2, 3\}$ you can never reach state 1.
So for example $(P^n)_{21} = 0$ for all $n \Rightarrow$ not irreducible.

Back to King Markov: why it is irreducible

Recall the islands are arranged in a ring: $1, 2, \ldots, N$.

Proposal step:

- If on island i, propose to move to $i + 1$ or $i - 1 \pmod N$ by flipping a coin.

Back to King Markov: why it is irreducible

Recall the islands are arranged in a ring: $1, 2, \ldots, N$.

Proposal step:

- If on island i, propose to move to $i + 1$ or $i - 1$ (mod N) by flipping a coin.

Acceptance step:

- Accept with some probability that depends on island sizes (but is *not* identically zero).

Why irreducible?

- You can try to move clockwise many times to reach any target island.
- Each step has **positive probability** of being proposed *and* accepted.
- Therefore for any pair (i, j), there exists some n with $(P^n)_{ij} > 0$.

Back to King Markov: why it is aperiodic

For periodicity, the main danger is a strict "alternating" structure (like bipartite graphs).
King Markov avoids a fixed rhythm because:

Back to King Markov: why it is aperiodic

For periodicity, the main danger is a strict "alternating" structure (like bipartite graphs). King Markov avoids a fixed rhythm because:

- There are multiple possible path lengths to return to the same island.
- For many islands, you can return in (say) 2 steps *and* also in 3 steps with nonzero probability.
- Once you have two return times with $\gcd = 1$, the period becomes 1.

A very common sufficient condition (easy to check):

$$\exists i \text{ such that } p_{ii} > 0 \quad \Longrightarrow \quad \text{aperiodic.}$$

Back to King Markov: why it is aperiodic

For periodicity, the main danger is a strict "alternating" structure (like bipartite graphs).
King Markov avoids a fixed rhythm because:

- There are multiple possible path lengths to return to the same island.
- For many islands, you can return in (say) 2 steps *and* also in 3 steps with nonzero probability.
- Once you have two return times with $\gcd = 1$, the period becomes 1.

A very common sufficient condition (easy to check):

$$\exists i \text{ such that } p_{ii} > 0 \quad \Longrightarrow \quad \text{aperiodic.}$$

(If you sometimes "stay put" with positive probability, you break any rigid cycle.)

A practical trick to guarantee aperiodicity: “lazy” chains

If you’re unsure about periodicity, you can **make the chain lazy**:

$$P_{\text{lazy}} = \frac{1}{2}I + \frac{1}{2}P.$$

A practical trick to guarantee aperiodicity: “lazy” chains

If you’re unsure about periodicity, you can **make the chain lazy**:

$$P_{\text{lazy}} = \tfrac{1}{2}I + \tfrac{1}{2}P.$$

Meaning: at each step

- with probability $\frac{1}{2}$ you **stay where you are**,
- with probability $\frac{1}{2}$ you take the original transition from P.

Then $p_{ii}^{\text{lazy}} \geq \frac{1}{2} > 0$ for every state i, so the chain is **automatically aperiodic**.

Design principle for MCMC

When designing a Markov chain to sample from a posterior distribution:

Target

We want the chain to be **irreducible** and **aperiodic**.

Design principle for MCMC

When designing a Markov chain to sample from a posterior distribution:

Target

We want the chain to be **irreducible** and **aperiodic**.

- **Irreducible** $\Rightarrow$ we can explore the whole support of the posterior.
- **Aperiodic** $\Rightarrow$ we avoid deterministic oscillations and get stable convergence behaviour.

(Next: we connect this to the existence/uniqueness of a stationary distribution and convergence to it.)

Mini-checklist you can apply to any example

Given a proposed MCMC transition rule, ask:

(1) Irreducible?

- Can I reach any state from any start, at least with a tiny (but nonzero) probability?
- Is the state-space graph connected under positive-probability moves?

Mini-checklist you can apply to any example

Given a proposed MCMC transition rule, ask:

(1) Irreducible?

- Can I reach any state from any start, at least with a tiny (but nonzero) probability?
- Is the state-space graph connected under positive-probability moves?

(2) Aperiodic?

- Do I ever have $p_{ii} > 0$ for some i? (Often enough to guarantee aperiodic.)
- Or can I return to a state in two different lengths that are not all multiples of the same integer?

If either fails, the chain can produce misleading samples (even if the code "runs").

What we want for MCMC-style sampling

When designing a Markov chain to explore a **target distribution** (e.g. a posterior):

- **Irreducible:** can reach all relevant parts of the state space.
- **Aperiodic:** avoids deterministic cycling; encourages "random-looking" exploration.

Sampling interpretation:

- If the chain is not irreducible, you may never generate samples from some states with nonzero posterior probability.
- If the chain is periodic, your samples can have strong regularity (poor mixing), which is not what we want.

From an initial distribution to later distributions

Let q be a distribution over states (think of it as a row vector):

$$q = (q_1, \ldots, q_{|\mathcal{S}|}), \qquad q_i = \mathbb{P}(Y_1 = i).$$

Given a transition matrix P with entries $P_{ij} = \mathbb{P}(Y_{n+1} = j \mid Y_n = i)$:

$$\mathbb{P}(Y_2 = j) = \sum_{i \in \mathcal{S}} q_i P_{ij}.$$

Matrix form:

$$q^{(2)} = qP,$$

where $q^{(2)}$ is the distribution of Y_2.

Iterating the chain: $q^{(n)} = qP^{n-1}$

By the Markov property, each step only depends on the current distribution and P:

$$q^{(3)} = q^{(2)}P = (qP)P = qP^2,$$

and more generally

$$q^{(n)} = qP^{n-1}.$$

- P^{n-1} encodes all $(n-1)$-step transition probabilities.
- If P^n "settles down" as n grows, then $q^{(n)}$ can converge to a stable distribution.

This is exactly the mechanism behind Markov chain Monte Carlo: run a chain long enough so that its distribution becomes (approximately) the target.

Proposition: a transition matrix has eigenvalue 1

Proposition 5.1 (informal statement). A transition matrix P always has at least one eigenvalue equal to 1.

Reason (linear algebra intuition):

- For a (row-stochastic) transition matrix, each row sums to 1.
- This implies $P\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is the all-ones column vector.
- So 1 is an eigenvalue (with eigenvector $\mathbf{1}$).

Takeaway: eigenvalue 1 is "built in" to probability transitions.

Definition: stationary distribution

A distribution π over $\mathcal{S}$ (row vector) is called **stationary** for P if

$$\pi P = \pi.$$

- **Meaning:** if $Y_1 \sim \pi$, then $Y_2 \sim \pi$, $Y_3 \sim \pi$, ...
- The distribution is "stable" under the transitions.

Definition 5.5 (as used in the lecture notes): If P has a unique eigenvalue equal to 1 (in the relevant sense), then there is a unique distribution π satisfying $\pi P = \pi$. This π is the **stationary distribution**.

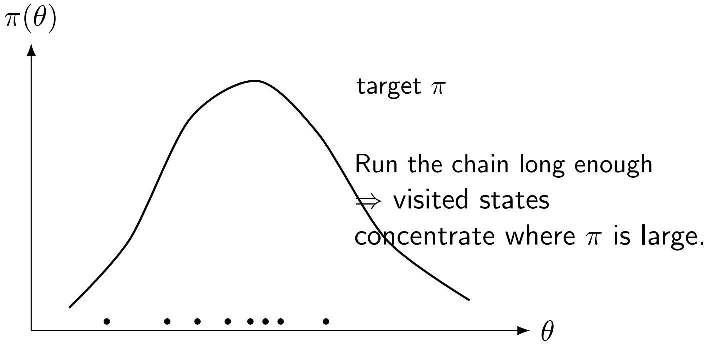
Why stationary distributions matter for posterior sampling

Goal in MCMC:

$$\text{Design } P \text{ so that } \pi \text{ (stationary) } = \text{(posterior distribution we want).}$$

Workflow idea:

- Build a Markov chain $Y_1, Y_2, \ldots$ with transition matrix P.
- Run it for a long time, record visited states.
- The *empirical histogram* of visited states approximates π.


$\pi(\theta)$
target π
Run the chain long enough
$\Rightarrow$ visited states
concentrate where π is large.
θ

Convergence checklist: what we typically need

To argue that a Markov chain converges to a stationary distribution (and is useful for sampling), we want:

1. **Irreducible** (can reach any state from any state),
2. **Aperiodic** (no rigid cycling),
3. **A stationary distribution exists** π with $\pi P = \pi$.

Important nuance:

- Existence of π does not automatically imply convergence from every starting point.
- Irreducibility $+$ aperiodicity are the structural conditions that rule out getting stuck in disconnected parts or cycling forever.

Theorem 5.1: Irreducible $\Rightarrow$ unique stationary distribution (finite case)

Statement. If a *finite* Markov chain is **irreducible**, then it has a **unique** stationary distribution π.

What this really means (intuition).

- **Stationary distribution** π is a probability vector with

$$\pi = \pi P, \qquad \sum_{i \in \mathcal{S}} \pi_i = 1, \ \pi_i \geq 0.$$

- Think of π as a **long-run balance of flow**: for each state, the total probability mass flowing *into* it equals the mass flowing *out*.
- **Irreducible** means the chain is "one connected piece": from any state you can reach any other state (eventually, with positive probability).
- In a finite, connected system, there is only **one** consistent long-run balance π.

Key caveat. Existence/uniqueness of π does *not* yet say the chain *converges* to π from every start.

Definition 5.8: Ergodicity = convergence to π (from any start)

Ergodic (finite) Markov chain. For all $i, j \in \mathcal{S}$,

$$\lim_{n\to\infty} \mathbb{P}(Y_n = i \mid Y_0 = j) = \pi_i.$$

Matrix form (often easiest to remember).

- Let q be any initial distribution (row vector).
- The distribution after n steps is qP^n.
- Ergodicity says:

$$\lim_{n\to\infty} qP^n = \pi \qquad \text{for } \textit{every} \text{ initial } q.$$

Interpretation.

- **Stationarity** ($\pi = \pi P$) means: if you *start* in π, you *stay* in π forever.
- **Ergodicity** means: even if you start elsewhere, you *forget* where you started and **converge** to π.

Why irreducible alone is not enough: a 2-state counterexample

Consider the two-state chain $\mathcal{S} = \{1, 2\}$ with

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Meaning: from 1 you always go to 2, and from 2 you always go to 1.

Step 1: Irreducible? Yes.

- From 1 you can reach 2 in one step.
- From 2 you can reach 1 in one step.
- So every state communicates with every other state.

Step 2: Stationary distribution? Yes, and unique. Try $\pi = (1/2, 1/2)$. Then

$$\pi P = (1/2, 1/2) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = (1/2, 1/2) = \pi.$$

So π is stationary (and by Theorem 5.1, unique).

But it is not ergodic: the chain never settles

Assume we start in state 1, i.e. $Y_0 = 1$.

What happens? The chain alternates deterministically:

$$1 \to 2 \to 1 \to 2 \to \cdots$$

So:

$$\mathbb{P}(Y_{2n} = 1 \mid Y_0 = 1) = 1, \qquad \mathbb{P}(Y_{2n+1} = 1 \mid Y_0 = 1) = 0.$$

No convergence.

- If the chain were ergodic with stationary distribution $\pi = (1/2, 1/2)$, we would need

$$\mathbb{P}(Y_n = 1 \mid Y_0 = 1) \to \pi_1 = 1/2.$$

- But the sequence is $1, 0, 1, 0, \ldots$ so it has **no limit**.

Takeaway: irreducible + stationary distribution does not guarantee convergence.

The missing piece: aperiodicity (why oscillations happen)

Period of a state. For state i,

$$d_i = \gcd\{n \geq 1 : \mathbb{P}(Y_n = i \mid Y_0 = i) > 0\}.$$

- If $d_i = 1$, the state is **aperiodic**.
- If $d_i > 1$, returns to i can only happen in multiples of d_i steps (a built-in "rhythm").

In the 2-state example:

- Starting from 1, you can return to 1 only in $n = 2, 4, 6, \ldots$ steps.
- Hence $d_1 = 2$. Similarly $d_2 = 2$.
- This explains the alternating behaviour: the chain is **stuck in a 2-cycle**.

Moral: aperiodicity is what kills strict cycles and allows true convergence.

Theorem 5.2: Ergodic $\Leftrightarrow$ irreducible + aperiodic (finite case)

Statement. A finite Markov chain is **ergodic** if and only if it is

irreducible and **aperiodic**.

Why these two conditions are exactly right (intuition).

- **Irreducible** ensures the chain can explore the entire state space (no disconnected "islands" that trap you forever).
- **Aperiodic** ensures the exploration does not get locked into a strict cycle.
- Together (in the finite setting), they guarantee:

$$P^n \to \mathbf{1}\pi \quad \text{(rows of } P^n \text{ approach } \pi\text{).}$$

What you should remember:

- Theorem 5.1: irreducible $\Rightarrow$ unique π.
- Theorem 5.2: irreducible + aperiodic $\Rightarrow$ convergence to π from any start.

Why we care: this is the backbone of MCMC

In MCMC, we build a Markov chain whose states are *parameter values* θ, and we want the chain to **spend time according to a target distribution** $\pi(\theta)$ (typically the posterior).

What ergodicity gives us.

- If the chain is ergodic with stationary distribution π, then:

$$\theta^{(n)} \text{ eventually behaves like a draw from } \pi.$$

- "No matter where we start", the chain will **forget the initialisation** and (eventually) sample according to π.

Important practical note.

- Ergodicity is an *asymptotic* guarantee ($n \to \infty$).
- In practice, we also care about **how fast** the chain approaches π (mixing time / burn-in), but that is a separate issue.

Checklist: will my finite Markov chain converge to the desired π?

To have convergence to a stationary distribution from any starting point, check:

1. **Irreducible?**
   - Can I reach any state from any other (eventually, with positive probability)?
2. **Aperiodic?**
   - Is the chain free of strict cycles (period $= 1$)?
   - A common easy sufficient condition: *at least one state has a self-loop* ($p_{ii} > 0$) in an irreducible chain $\Rightarrow$ all states are aperiodic.
3. **Is the target π stationary for P?**
   - Verify $\pi = \pi P$ (or use a sufficient condition like detailed balance later).

If (1)+(2) hold: the chain converges to *its* unique stationary distribution.
If also (3) holds: that stationary distribution is exactly your *desired* π.

Definition: detailed balance equations

Definition 5.6. A Markov chain with transition matrix P satisfies the **detailed balance equations** with respect to distribution π if for all states $i, j \in \mathcal{S}$,

$$\pi_i P_{ij} = \pi_j P_{ji}.$$

Interpretation (flow balance):

- Think of $\pi_i P_{ij}$ as the *long-run flow* from i to j.
- Detailed balance says the flow from i to j equals the flow from j to i.
- This is a form of **reversibility**.

Theorem: detailed balance implies stationarity

Theorem 5.1 (Detailed balance theorem). If P satisfies detailed balance with respect to π, then π is stationary:

$$\pi P = \pi.$$

Proof idea (one line per component): For each j,

$$(\pi P)_j = \sum_{i \in \mathcal{S}} \pi_i P_{ij} = \sum_{i \in \mathcal{S}} \pi_j P_{ji} = \pi_j \sum_{i \in \mathcal{S}} P_{ji} = \pi_j.$$

Why this is helpful in practice:

- Checking eigenvalues of a huge (or continuous) transition operator is hard.
- Designing P to satisfy detailed balance is often *doable*.
- This is the doorway to algorithms like Metropolis–Hastings (next lecture).

Why we care: discrete grids vs continuous posteriors

So far we often picture a finite state space:

$$\pi = (\pi_1, \pi_2, \ldots, \pi_m), \qquad P \in \mathbb{R}^{m \times m}.$$

But most posteriors are **continuous** (infinitely many states).

Practical implication:

- You cannot literally build an "infinite matrix" and compute eigenvalues.
- Detailed balance generalises cleanly to continuous settings (with densities and kernels), and provides a route to proving stationarity without heavy linear algebra.

Today: understand the concept. *Next: build a concrete P that targets a posterior.*

Example intuition: "King Markov" (sampling proportional to island size)

Story model (used as intuition for MCMC design):

- States are islands $1, 2, \ldots, N$.
- Desired long-run visitation: $\mathbb{P}(\text{island } i) \propto i$ (larger islands visited more).
- Move rule: propose a neighbouring island (e.g. $i \to i \pm 1$) using a coin flip, then accept/reject based on a probability that depends on relative sizes.

Why it tends to be good for sampling:

- **Irreducible:** from any island you can eventually reach any other island (nonzero probability path).
- **Aperiodic:** acceptance/rejection introduces randomness that breaks rigid cycles.

This is the exact kind of idea we formalise next lecture.

What you should leave with today

- Markov property: future depends only on present.
- Transition matrix P and its powers: P^n gives n-step probabilities.
- Two structural properties we like for sampling:
  - irreducible (no unreachable regions),
  - aperiodic (no rigid cycles).
- Stationary distribution π: $\pi P = \pi$.
- Detailed balance: $\pi_i P_{ij} = \pi_j P_{ji} \Rightarrow$ stationarity.

Next lecture: we will write down a concrete transition rule P_{ij} that targets a chosen posterior (and see why detailed balance is the key proof tool).