Markov Chains & MCMC: From Definitions to Metropolis–Hastings

Warm-up, detailed balance, MH algorithm, and a worked example

Warm-up: match the key definitions

The warm-up was just to match the definitions about Markov properties (or Markov chains).

- **Markov property:** the probability of being at any state in the future (or the immediate future) only depends on the present, and nothing else that happened before.

Warm-up: match the key definitions

The warm-up was just to match the definitions about Markov properties (or Markov chains).

- **Markov property:** the probability of being at any state in the future (or the immediate future) only depends on the present, and nothing else that happened before.
- **Irreducibility:** we can go from state i to state j — there is a positive probability of going from i to j in n steps (for some n).
    - So it's not like we ever get stuck.
    - We can reach any state from any other state (eventually, with positive probability).

Warm-up: match the key definitions

The warm-up was just to match the definitions about Markov properties (or Markov chains).

- **Markov property:** the probability of being at any state in the future (or the immediate future) only depends on the present, and nothing else that happened before.
- **Irreducibility:** we can go from state i to state j — there is a positive probability of going from i to j in n steps (for some n).
  - So it's not like we ever get stuck.
  - We can reach any state from any other state (eventually, with positive probability).
- **Stationary distribution:** $\pi P = \pi$.
  - We have a distribution π and a transition matrix P.
  - If we apply the transition matrix, we get back the same distribution again.

Warm-up: match the key definitions

The warm-up was just to match the definitions about Markov properties (or Markov chains).

- **Markov property:** the probability of being at any state in the future (or the immediate future) only depends on the present, and nothing else that happened before.
- **Irreducibility:** we can go from state i to state j — there is a positive probability of going from i to j in n steps (for some n).
  - So it's not like we ever get stuck.
  - We can reach any state from any other state (eventually, with positive probability).
- **Stationary distribution:** $\pi P = \pi$.
  - We have a distribution π and a transition matrix P.
  - If we apply the transition matrix, we get back the same distribution again.
- **Aperiodic:** no state is periodic.
  - There is no fixed rhythm like "visit every three steps" or "every ten steps".

Where this is going: sampling from posteriors

Where we were on Monday (King Markov and the islands) and yesterday (definitions about Markov chains) is working towards a way of sampling from posterior distributions.

Suppose we have a posterior distribution for θ and it's not very nice.

Goal: set up a Markov chain so that, no matter where we start, the states we visit eventually behave like samples from the posterior.

Where this is going: sampling from posteriors

Where we were on Monday (King Markov and the islands) and yesterday (definitions about Markov chains) is working towards a way of sampling from posterior distributions.

Suppose we have a posterior distribution for θ and it's not very nice.

Goal: set up a Markov chain so that, no matter where we start, the states we visit eventually behave like samples from the posterior.

- Start somewhere, then move: here, here, here, here, here again...
- We "walk around" the state space.
- We visit states proportional to the posterior:
  - more likely to visit where the posterior is large,
  - less likely to visit where the posterior is small.
- If we collect all visited states, we can build a histogram that represents the posterior.

Reminder: detailed balance

We also had detailed balance at the end of yesterday.

Detailed balance:

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

- This is a reversibility condition.
- Intuition: the long-run "flow" from i to j equals the flow from j to i.

Reminder: detailed balance

We also had detailed balance at the end of yesterday.

Detailed balance:

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

- This is a reversibility condition.
- Intuition: the long-run "flow" from i to j equals the flow from j to i.

We even had the detailed balance theorem yesterday (Theorem 5.1):

If P satisfies detailed balance w.r.t. π, then $\pi P = \pi$.

That is: detailed balance $\Rightarrow$ π is stationary.

Theorem 5.1: detailed balance $\Rightarrow$ stationarity (proof)

We didn't do the proof yesterday. The proof isn't too bad.

Let's look at the j-th component of πP:

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

Theorem 5.1: detailed balance $\Rightarrow$ stationarity (proof)

We didn't do the proof yesterday. The proof isn't too bad.

Let's look at the j-th component of πP:

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

By detailed balance, $\pi_i P_{ij} = \pi_j P_{ji}$, so:

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

Theorem 5.1: detailed balance $\Rightarrow$ stationarity (proof)

We didn't do the proof yesterday. The proof isn't too bad.

Let's look at the j-th component of πP:

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

By detailed balance, $\pi_i P_{ij} = \pi_j P_{ji}$, so:

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

Now π_j does not depend on i, so it comes out:

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

Theorem 5.1: detailed balance $\Rightarrow$ stationarity (proof)

We didn't do the proof yesterday. The proof isn't too bad.

Let's look at the j-th component of πP:

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

By detailed balance, $\pi_i P_{ij} = \pi_j P_{ji}$, so:

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

Now π_j does not depend on i, so it comes out:

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

But probabilities in a row sum to 1:

$$\sum_{i \in \mathcal{S}} P_{ji} = 1.$$

So:

$$(\pi P)_j = \pi_j \qquad \text{for all } j,$$

Putting the pieces together (what we need for MCMC)

Let's put the pieces together from yesterday:

- aperiodic Markov chains,
- irreducible Markov chains,
- and now detailed balance.

Putting the pieces together (what we need for MCMC)

Let's put the pieces together from yesterday:

- aperiodic Markov chains,
- irreducible Markov chains,
- and now detailed balance.

So we now have **three statements** that we want for our Markov chain to:

- travel around the posterior state space,
- and visit states proportional to the posterior density.

Putting the pieces together (what we need for MCMC)

Let's put the pieces together from yesterday:

- aperiodic Markov chains,
- irreducible Markov chains,
- and now detailed balance.

So we now have **three statements** that we want for our Markov chain to:

- travel around the posterior state space,
- and visit states proportional to the posterior density.

Takeaway: if a Markov chain is

aperiodic + **irreducible** + **satisfies detailed balance**,

then we are OK to use it to sample from our posterior distribution.

Next question: how do we do that in practice?

From theory to practice: what are the P_{ij}'s?

We now want to talk about these P_{ij}'s: what actually are they?

The method we will use is **Section 5.2: Metropolis–Hastings**.

Metropolis–Hastings tells you how to set up the transition probabilities so that:

- we have detailed balance,
- the chain is irreducible,
- and the chain is aperiodic.

From theory to practice: what are the P_{ij}'s?

We now want to talk about these P_{ij}'s: what actually are they?

The method we will use is **Section 5.2: Metropolis–Hastings**.

Metropolis–Hastings tells you how to set up the transition probabilities so that:

- we have detailed balance,
- the chain is irreducible,
- and the chain is aperiodic.

If you think back to the story: King Markov was already doing the MH pattern:

- propose a move (coin flip to choose a neighbour),
- accept/reject (his weird seashells-and-stones rule),
- repeat forever.

Metropolis–Hastings algorithm (high level)

Goal: sample from a posterior distribution $\pi(\theta \mid y)$ using Metropolis–Hastings (MH).

Algorithm (one iteration):

1. Set an initial value $\theta^{(0)}$ (start somewhere).
2. For $i = 1, 2, \ldots$:
   1. Propose θ' from a proposal distribution $q(\cdot \mid \theta^{(i-1)})$.
   2. Accept $\theta^{(i)} = \theta'$ with probability

$$p_{\text{acc}} = \min\left(1,\ \frac{\pi(\theta' \mid y)\, q(\theta^{(i-1)} \mid \theta')}{\pi(\theta^{(i-1)} \mid y)\, q(\theta' \mid \theta^{(i-1)})}\right).$$

   Otherwise reject and set $\theta^{(i)} = \theta^{(i-1)}$.

Metropolis–Hastings algorithm (high level)

Goal: sample from a posterior distribution $\pi(\theta \mid y)$ using Metropolis–Hastings (MH).

Algorithm (one iteration):

1. Set an initial value $\theta^{(0)}$ (start somewhere).
2. For $i = 1, 2, \ldots$:
    1. Propose θ' from a proposal distribution $q(\cdot \mid \theta^{(i-1)})$.
    2. Accept $\theta^{(i)} = \theta'$ with probability

    $$p_{\text{acc}} = \min\left(1, \frac{\pi(\theta' \mid y)\, q(\theta^{(i-1)} \mid \theta')}{\pi(\theta^{(i-1)} \mid y)\, q(\theta' \mid \theta^{(i-1)})}\right).$$

    Otherwise reject and set $\theta^{(i)} = \theta^{(i-1)}$.

This becomes "MH" very quickly because it's a lot to write.

What the acceptance probability is doing (intuition)

Each iteration:

- we propose a new value from q (like proposal distributions in rejection sampling),
- then we accept or reject.

What the acceptance probability is doing (intuition)

Each iteration:

- we propose a new value from q (like proposal distributions in rejection sampling),
- then we accept or reject.

The acceptance probability compares:

- how plausible the proposed value is under the posterior,
- and how “fair” the proposal mechanism is forward vs backward.

What the acceptance probability is doing (intuition)

Each iteration:

- we propose a new value from q (like proposal distributions in rejection sampling),
- then we accept or reject.

The acceptance probability compares:

- how plausible the proposed value is under the posterior,
- and how "fair" the proposal mechanism is forward vs backward.

Sometimes the ratio is bigger than 1, so we cap it at 1:

$$p_{\text{acc}} = \min(1, \text{ratio}).$$

Connection to King Markov:

- proposal $=$ coin flip to pick a neighbouring island,
- accept/reject $=$ his rule that makes him spend more time on bigger islands,
- repeat $=$ long-run visitation matches the target.

A key Bayesian trick: the normalising constant cancels

Inside p_{acc} we have a ratio:

$$\frac{\pi(\theta' \mid y)}{\pi(\theta \mid y)}.$$

A key Bayesian trick: the normalising constant cancels

Inside p_{acc} we have a ratio:

$$\frac{\pi(\theta' \mid y)}{\pi(\theta \mid y)}.$$

But Bayesian posteriors are usually only known up to proportionality:

$$\pi(\theta \mid y) \propto p(y \mid \theta)\, p(\theta).$$

A key Bayesian trick: the normalising constant cancels

Inside p_{acc} we have a ratio:

$$\frac{\pi(\theta' \mid y)}{\pi(\theta \mid y)}.$$

But Bayesian posteriors are usually only known up to proportionality:

$$\pi(\theta \mid y) \propto p(y \mid \theta)\, p(\theta).$$

So:

$$\frac{\pi(\theta' \mid y)}{\pi(\theta \mid y)} = \frac{p(y \mid \theta')\, p(\theta')/p(y)}{p(y \mid \theta)\, p(\theta)/p(y)} = \frac{p(y \mid \theta')\, p(\theta')}{p(y \mid \theta)\, p(\theta)}.$$

Takeaway: we never need to evaluate $p(y)$ (the nasty normalising constant). That's exactly why MH is useful when the posterior is hard to normalise.

Example 5.3: radioactive decay data (setup)

We have a counter that monitors the time until an atom decays. We collect data $X_1, \ldots, X_n$.

Model:

$$X_i \mid \lambda \sim \text{Exponential}(\lambda), \qquad i = 1, \ldots, n.$$

Example 5.3: radioactive decay data (setup)

We have a counter that monitors the time until an atom decays. We collect data $X_1, \ldots, X_n$.

Model:

$$X_i \mid \lambda \sim \text{Exponential}(\lambda), \qquad i = 1, \ldots, n.$$

The time until an atom decays is very short (less than one second): these are highly radioactive items.

Prior: we assume $\lambda \in (0, 1)$ and use a Beta prior:

$$\lambda \sim \text{Beta}(\alpha, \beta).$$

Example 5.3: radioactive decay data (setup)

We have a counter that monitors the time until an atom decays. We collect data $X_1, \ldots, X_n$.

Model:

$$X_i \mid \lambda \sim \text{Exponential}(\lambda), \qquad i = 1, \ldots, n.$$

The time until an atom decays is very short (less than one second): these are highly radioactive items.

Prior: we assume $\lambda \in (0, 1)$ and use a Beta prior:

$$\lambda \sim \text{Beta}(\alpha, \beta).$$

First step: derive the posterior (up to proportionality). I'll give you a few minutes to write down the likelihood, the prior, and the posterior.

Example 5.3: likelihood, prior, posterior (worked)

Likelihood:

$$p(x \mid \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i \right).$$

Example 5.3: likelihood, prior, posterior (worked)

Likelihood:

$$p(x \mid \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right).$$

Prior (up to proportionality):

$$p(\lambda) \propto \lambda^{\alpha-1}(1-\lambda)^{\beta-1}, \qquad 0 < \lambda < 1.$$

Example 5.3: likelihood, prior, posterior (worked)

Likelihood:

$$p(x \mid \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right).$$

Prior (up to proportionality):

$$p(\lambda) \propto \lambda^{\alpha-1}(1-\lambda)^{\beta-1}, \qquad 0 < \lambda < 1.$$

Posterior (up to proportionality):

$$\pi(\lambda \mid x) \propto \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) \lambda^{\alpha-1}(1-\lambda)^{\beta-1}.$$

Example 5.3: likelihood, prior, posterior (worked)

Likelihood:

$$p(x \mid \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right).$$

Prior (up to proportionality):

$$p(\lambda) \propto \lambda^{\alpha-1}(1-\lambda)^{\beta-1}, \qquad 0 < \lambda < 1.$$

Posterior (up to proportionality):

$$\pi(\lambda \mid x) \propto \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) \lambda^{\alpha-1}(1-\lambda)^{\beta-1}.$$

Tidying:

$$\pi(\lambda \mid x) \propto \lambda^{n+\alpha-1}(1-\lambda)^{\beta-1} \exp\left(-\lambda \sum_{i=1}^{n} x_i\right).$$

This is not a "nice" standard distribution with a closed form sampler, so it's an ideal place to use MH.

Example 5.3: MH with a random-walk proposal

Step 1: initialise. Choose $\lambda^{(0)} \in (0,1)$. For concreteness, start in the middle:

$$\lambda^{(0)} = 1/2.$$

Example 5.3: MH with a random-walk proposal

Step 1: initialise. Choose $\lambda^{(0)} \in (0,1)$. For concreteness, start in the middle:

$$\lambda^{(0)} = 1/2.$$

Step 2: propose. A common choice is a random walk:

$$\lambda' = \lambda^{(i-1)} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,\sigma^2).$$

Example 5.3: MH with a random-walk proposal

Step 1: initialise. Choose $\lambda^{(0)} \in (0,1)$. For concreteness, start in the middle:

$$\lambda^{(0)} = 1/2.$$

Step 2: propose. A common choice is a random walk:

$$\lambda' = \lambda^{(i-1)} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Because the prior has support only on $(0,1)$, if $\lambda' \notin (0,1)$ we can reject immediately.

Example 5.3: MH with a random-walk proposal

Step 1: initialise. Choose $\lambda^{(0)} \in (0,1)$. For concreteness, start in the middle:

$$\lambda^{(0)} = 1/2.$$

Step 2: propose. A common choice is a random walk:

$$\lambda' = \lambda^{(i-1)} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Because the prior has support only on $(0,1)$, if $\lambda' \notin (0,1)$ we can reject immediately.

Step 3: accept/reject.

$$p_{\text{acc}} = \min\left(1,\ \frac{\pi(\lambda' \mid x)}{\pi(\lambda^{(i-1)} \mid x)} \cdot \frac{q(\lambda^{(i-1)} \mid \lambda')}{q(\lambda' \mid \lambda^{(i-1)})}\right).$$

Example 5.3: simplify the proposal ratio (Gaussian RW)

For the Gaussian random-walk proposal,

$$q(\lambda' \mid \lambda) = \varphi(\lambda';\, \lambda, \sigma^2), \qquad q(\lambda \mid \lambda') = \varphi(\lambda;\, \lambda', \sigma^2).$$

Example 5.3: simplify the proposal ratio (Gaussian RW)

For the Gaussian random-walk proposal,

$$q(\lambda' \mid \lambda) = \varphi(\lambda';\, \lambda, \sigma^2), \qquad q(\lambda \mid \lambda') = \varphi(\lambda;\, \lambda', \sigma^2).$$

But

$$(\lambda - \lambda')^2 = (\lambda' - \lambda)^2,$$

so the Gaussian density is symmetric:

$$q(\lambda \mid \lambda') = q(\lambda' \mid \lambda).$$

Example 5.3: simplify the proposal ratio (Gaussian RW)

For the Gaussian random-walk proposal,

$$q(\lambda' \mid \lambda) = \varphi(\lambda';\, \lambda, \sigma^2), \qquad q(\lambda \mid \lambda') = \varphi(\lambda;\, \lambda', \sigma^2).$$

But

$$(\lambda - \lambda')^2 = (\lambda' - \lambda)^2,$$

so the Gaussian density is symmetric:

$$q(\lambda \mid \lambda') = q(\lambda' \mid \lambda).$$

Therefore the proposal ratio cancels:

$$\frac{q(\lambda \mid \lambda')}{q(\lambda' \mid \lambda)} = 1.$$

Example 5.3: simplify the proposal ratio (Gaussian RW)

For the Gaussian random-walk proposal,

$$q(\lambda' \mid \lambda) = \varphi(\lambda';\, \lambda, \sigma^2), \qquad q(\lambda \mid \lambda') = \varphi(\lambda;\, \lambda', \sigma^2).$$

But

$$(\lambda - \lambda')^2 = (\lambda' - \lambda)^2,$$

so the Gaussian density is symmetric:

$$q(\lambda \mid \lambda') = q(\lambda' \mid \lambda).$$

Therefore the proposal ratio cancels:

$$\frac{q(\lambda \mid \lambda')}{q(\lambda' \mid \lambda)} = 1.$$

So the acceptance probability becomes:

$$p_{\mathsf{acc}} = \min\left(1,\ \frac{\pi(\lambda' \mid x)}{\pi(\lambda \mid x)}\right),$$

(where λ means the current state).

Example 5.3: posterior ratio (explicit form)

Using

$$\pi(\lambda \mid x) \propto \lambda^{n+\alpha-1}(1-\lambda)^{\beta-1} \exp\left(-\lambda \sum_{i=1}^{n} x_i\right),$$

the posterior ratio is:

$$\frac{\pi(\lambda' \mid x)}{\pi(\lambda \mid x)} = \left(\frac{\lambda'}{\lambda}\right)^{n+\alpha-1} \left(\frac{1-\lambda'}{1-\lambda}\right)^{\beta-1} \exp\left(-(\lambda'-\lambda) \sum_{i=1}^{n} x_i\right).$$

Example 5.3: posterior ratio (explicit form)

Using

$$\pi(\lambda \mid x) \propto \lambda^{n+\alpha-1}(1-\lambda)^{\beta-1} \exp\left(-\lambda \sum_{i=1}^{n} x_i\right),$$

the posterior ratio is:

$$\frac{\pi(\lambda' \mid x)}{\pi(\lambda \mid x)} = \left(\frac{\lambda'}{\lambda}\right)^{n+\alpha-1} \left(\frac{1-\lambda'}{1-\lambda}\right)^{\beta-1} \exp\left(-(\lambda'-\lambda) \sum_{i=1}^{n} x_i\right).$$

So:

$$p_{\text{acc}} = \min\left(1,\ \left(\frac{\lambda'}{\lambda}\right)^{n+\alpha-1} \left(\frac{1-\lambda'}{1-\lambda}\right)^{\beta-1} \exp\left(-(\lambda'-\lambda) \sum_{i=1}^{n} x_i\right)\right).$$

Accept λ' with this probability; otherwise keep the current λ. Repeat for many iterations to get a histogram of visited λ values.

How this looks in R (implementation pattern)

Key practical tip: compute acceptance on the **log scale** to avoid numerical underflow.

```
N_iter <- 10000
lambda_store <- numeric(N_iter)
lambda <- 0.5 # initial value
n <- 20 # number of atoms
sum_x <- 67.6 # sum of observed decay times
alpha <- 1; beta <- 1 # prior parameters
sigma2 <- 0.1 # RW variance
for (i in 1:N_iter) {
  lambda_prop <- rnorm(1, mean = lambda, sd = sqrt(sigma2))

  if (lambda_prop > 0 && lambda_prop < 1) {

    # log acceptance ratio (posterior ratio; proposal cancels for symmetric RW)
    log_acc <- (n + alpha - 1) * log(lambda_prop / lambda) +
               (beta - 1) * log((1 - lambda_prop) / (1 - lambda)) -
               (lambda_prop - lambda) * sum_x

    if (log(runif(1)) < min(0, log_acc)) {
      lambda <- lambda_prop
    }
```

Diagnostics and summaries (what you compute afterwards)

Once you have `lambda_store` (all visited states):

- **Trace plot:** the value of the chain at each iteration.
  - In the King Markov video: if it looks like a "hairy fat caterpillar", you've got a happy Markov chain.
  - Or: if it looks like white noise, it's doing the right kind of exploration.

Diagnostics and summaries (what you compute afterwards)

Once you have `lambda_store` (all visited states):

- **Trace plot:** the value of the chain at each iteration.
  - In the King Markov video: if it looks like a "hairy fat caterpillar", you've got a happy Markov chain.
  - Or: if it looks like white noise, it's doing the right kind of exploration.
- **Posterior histogram/density:** the empirical distribution of visited λ values.

Diagnostics and summaries (what you compute afterwards)

Once you have `lambda_store` (all visited states):

- **Trace plot:** the value of the chain at each iteration.
  - In the King Markov video: if it looks like a "hairy fat caterpillar", you've got a happy Markov chain.
  - Or: if it looks like white noise, it's doing the right kind of exploration.
- **Posterior histogram/density:** the empirical distribution of visited λ values.
- **Posterior mean:** $\hat{\lambda}_{\text{mean}} = \texttt{mean(lambda\_store)}$ (example: 0.311).

Diagnostics and summaries (what you compute afterwards)

Once you have `lambda_store` (all visited states):

- **Trace plot:** the value of the chain at each iteration.
  - In the King Markov video: if it looks like a "hairy fat caterpillar", you've got a happy Markov chain.
  - Or: if it looks like white noise, it's doing the right kind of exploration.
- **Posterior histogram/density:** the empirical distribution of visited λ values.
- **Posterior mean:** $\hat{\lambda}_{\text{mean}} = \texttt{mean(lambda\_store)}$ (example: 0.311).
- **Credible interval:** e.g. 95% via quantiles:

```
quantile(lambda_store, c(0.025, 0.975))
```

  giving (example) $[0.193,\ 0.458]$.

You can compute mean, median, mode, variance, one-sided intervals — whatever you want from the posterior samples.

A Bayesian aside: why do this at all?

Just to stress:

- Bayesian inference is often **more computationally difficult**.
  - We run Metropolis–Hastings,
  - and we need Markov chain theory.

A Bayesian aside: why do this at all?

Just to stress:

- Bayesian inference is often **more computationally difficult**.
  - We run Metropolis–Hastings,
  - and we need Markov chain theory.
- In a frequentist maximum-likelihood approach, you might:
  - take logs,
  - differentiate,
  - set equal to 0,
  - and you're done.

A Bayesian aside: why do this at all?

Just to stress:

- Bayesian inference is often **more computationally difficult**.
  - We run Metropolis–Hastings,
  - and we need Markov chain theory.
- In a frequentist maximum-likelihood approach, you might:
  - take logs,
  - differentiate,
  - set equal to 0,
  - and you're done.
- But then you mostly get a point estimate (the MLE).

A Bayesian aside: why do this at all?

Just to stress:

- Bayesian inference is often **more computationally difficult**.
  - We run Metropolis–Hastings,
  - and we need Markov chain theory.
- In a frequentist maximum-likelihood approach, you might:
  - take logs,
  - differentiate,
  - set equal to 0,
  - and you're done.
- But then you mostly get a point estimate (the MLE).
- With Bayesian inference you get the whole posterior distribution:
  - uncertainty quantification,
  - credible intervals,
  - whatever functional of the posterior you care about.

It's harder work — but you get much more information.

Back to theory: why does Metropolis–Hastings work?

Some of you may be thinking: how do we know Metropolis–Hastings gives us the stationary distribution we want?

Fortunately, we spent the last week building theory:

- when a Markov chain converges to a stationary distribution,
- and how to prove stationarity.

Back to theory: why does Metropolis–Hastings work?

Some of you may be thinking: how do we know Metropolis–Hastings gives us the stationary distribution we want?

Fortunately, we spent the last week building theory:

- when a Markov chain converges to a stationary distribution,
- and how to prove stationarity.

Now we apply that theory to MH.

Proposition 5.2: the Markov chain generated by Metropolis–Hastings satisfies detailed balance with respect to the posterior distribution.

So the stationary distribution is exactly the posterior we care about.

Proposition 5.2: set up the detailed balance statement

Let the current state be θ, and the proposed state be θ'.

Detailed balance (what we want to show) is:

$$\pi(\theta \mid y)\, P(\theta \to \theta') = \pi(\theta' \mid y)\, P(\theta' \to \theta).$$

In MH, the move probability factors into:

$$P(\theta \to \theta') = q(\theta' \mid \theta)\, a(\theta, \theta'),$$

where q is the proposal density and a is the acceptance probability.

Proposition 5.2: set up the detailed balance statement

Let the current state be θ, and the proposed state be θ'.

Detailed balance (what we want to show) is:

$$\pi(\theta \mid y)\, P(\theta \to \theta') = \pi(\theta' \mid y)\, P(\theta' \to \theta).$$

In MH, the move probability factors into:

$$P(\theta \to \theta') = q(\theta' \mid \theta)\, a(\theta, \theta'),$$

where q is the proposal density and a is the acceptance probability.

So we want to show:

$$\pi(\theta \mid y)\, q(\theta' \mid \theta)\, a(\theta, \theta') = \pi(\theta' \mid y)\, q(\theta \mid \theta')\, a(\theta', \theta).$$

Proposition 5.2: plug in the MH acceptance rule

MH acceptance probability:

$$a(\theta,\theta') = \min\left(1,\ \frac{\pi(\theta' \mid y)\, q(\theta \mid \theta')}{\pi(\theta \mid y)\, q(\theta' \mid \theta)}\right).$$

Proposition 5.2: plug in the MH acceptance rule

MH acceptance probability:

$$a(\theta,\theta') = \min\left(1,\ \frac{\pi(\theta' \mid y)\, q(\theta \mid \theta')}{\pi(\theta \mid y)\, q(\theta' \mid \theta)}\right).$$

Multiply by $\pi(\theta \mid y)\, q(\theta' \mid \theta)$:

$$\pi(\theta \mid y)\, q(\theta' \mid \theta)\, a(\theta,\theta') = \min\Big(\pi(\theta \mid y)\, q(\theta' \mid \theta),\ \pi(\theta' \mid y)\, q(\theta \mid \theta')\Big).$$

Proposition 5.2: plug in the MH acceptance rule

MH acceptance probability:

$$a(\theta, \theta') = \min\left(1,\ \frac{\pi(\theta' \mid y)\, q(\theta \mid \theta')}{\pi(\theta \mid y)\, q(\theta' \mid \theta)}\right).$$

Multiply by $\pi(\theta \mid y)\, q(\theta' \mid \theta)$:

$$\pi(\theta \mid y)\, q(\theta' \mid \theta)\, a(\theta, \theta') = \min\Big(\pi(\theta \mid y)\, q(\theta' \mid \theta),\ \pi(\theta' \mid y)\, q(\theta \mid \theta')\Big).$$

Now do the same on the other side:

$$\pi(\theta' \mid y)\, q(\theta \mid \theta')\, a(\theta', \theta) = \min\Big(\pi(\theta' \mid y)\, q(\theta \mid \theta'),\ \pi(\theta \mid y)\, q(\theta' \mid \theta)\Big).$$

Proposition 5.2: plug in the MH acceptance rule

MH acceptance probability:

$$a(\theta, \theta') = \min\left(1,\ \frac{\pi(\theta' \mid y)\, q(\theta \mid \theta')}{\pi(\theta \mid y)\, q(\theta' \mid \theta)}\right).$$

Multiply by $\pi(\theta \mid y)\, q(\theta' \mid \theta)$:

$$\pi(\theta \mid y)\, q(\theta' \mid \theta)\, a(\theta, \theta') = \min\Big(\pi(\theta \mid y)\, q(\theta' \mid \theta),\ \pi(\theta' \mid y)\, q(\theta \mid \theta')\Big).$$

Now do the same on the other side:

$$\pi(\theta' \mid y)\, q(\theta \mid \theta')\, a(\theta', \theta) = \min\Big(\pi(\theta' \mid y)\, q(\theta \mid \theta'),\ \pi(\theta \mid y)\, q(\theta' \mid \theta)\Big).$$

These are the same expression (just written in the opposite order), hence:

$$\pi(\theta \mid y)\, q(\theta' \mid \theta)\, a(\theta, \theta') = \pi(\theta' \mid y)\, q(\theta \mid \theta')\, a(\theta', \theta).$$

So MH satisfies detailed balance w.r.t. the posterior.

What this proof buys us (and what's next)

From Proposition 5.2:

- The MH chain satisfies detailed balance w.r.t. the posterior.
- Therefore the posterior is stationary for the MH chain.
- With the other conditions (irreducibility + aperiodicity), we are guaranteed to converge to the posterior distribution.

What this proof buys us (and what's next)

From Proposition 5.2:

- The MH chain satisfies detailed balance w.r.t. the posterior.
- Therefore the posterior is stationary for the MH chain.
- With the other conditions (irreducibility + aperiodicity), we are guaranteed to converge to the posterior distribution.

So everything is OK: we can use Metropolis–Hastings to sample from posterior distributions.