Bayesian Computation and Markov Chains

BEE 6940 Lecture 7

March 06, 2023

Table of Contents

Review of the Bayesian Statistics

Last Class: Bayes' Theorem and Bayesian Statistics

$$\underbrace{p(\theta|y)}_{\text{posterior}} = \frac{\overbrace{p(y|\theta)}^{\text{likelihood}}}{\underbrace{p(y)}_{\text{normalization}}} \overbrace{p(\theta)}^{\text{prior}}$$

Key Takeaways

- No distinction between parameters and unobserved data in Bayesian framework;
- Key emphasis on conditioning on data;
- Likelihood is given by the probability model for the data-generating process;
- Posterior as "compromise" between prior and likelihood;
- Prior can be influential in posterior inferences.

Bayesian Computation

Goals of Bayesian Computation

1. Sampling from the *posterior* distribution

$$p(\theta|\mathbf{y})$$

2. Sampling from the *posterior predictive* distribution

$$p(\tilde{y}|\mathbf{y})$$

by generating data.

Bayesian Computation and Monte Carlo

In other words, Bayesian computation involves Monte Carlo simulation from the posterior (predictive) distribution.

These samples can then be analyzed to identify estimators, credible intervals, etc.

Sampling from the Posterior

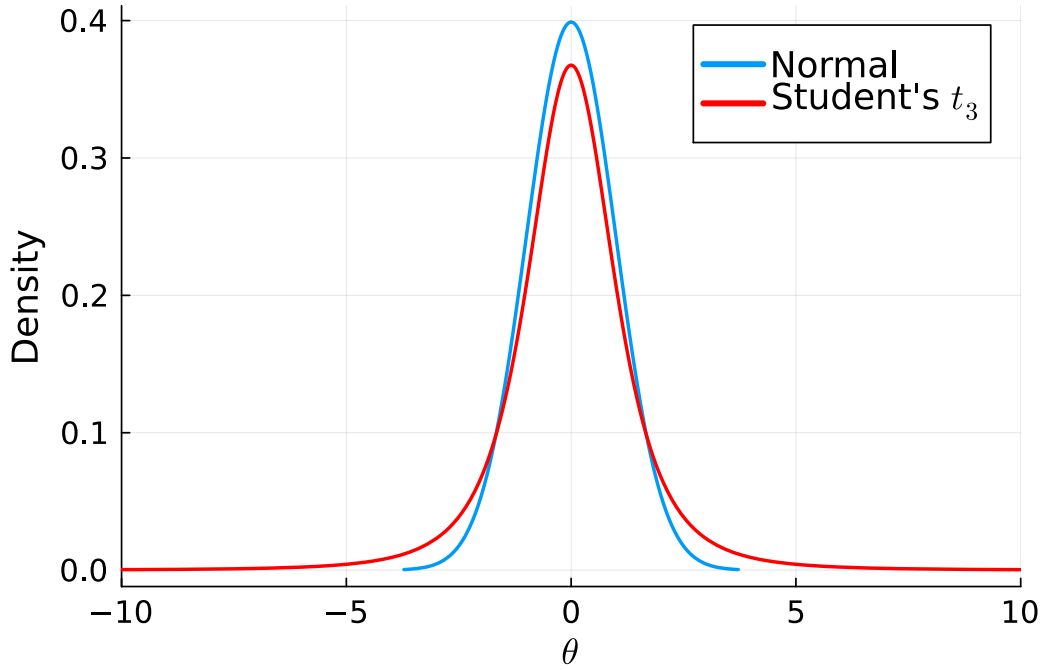
Trivial for extremely simple problems: low-dimensional and with "conjugate" priors (which make the posterior a closed-form distribution).

What to do when problems are more complex and/or we don't want to rely on priors for computational convenience?

A First Algorithm: Rejection Sampling

Idea:

1. Generate proposed samples from another distribution $g(\theta)$ which covers the target $p(\theta|\mathbf{y})$;
2. Accept those proposals based on the ratio of the two distributions.



Rejection Sampling Algorithm

Suppose $p(\theta|\mathbf{y}) \leq Mg(\theta)$ for some $1 < M < \infty$.

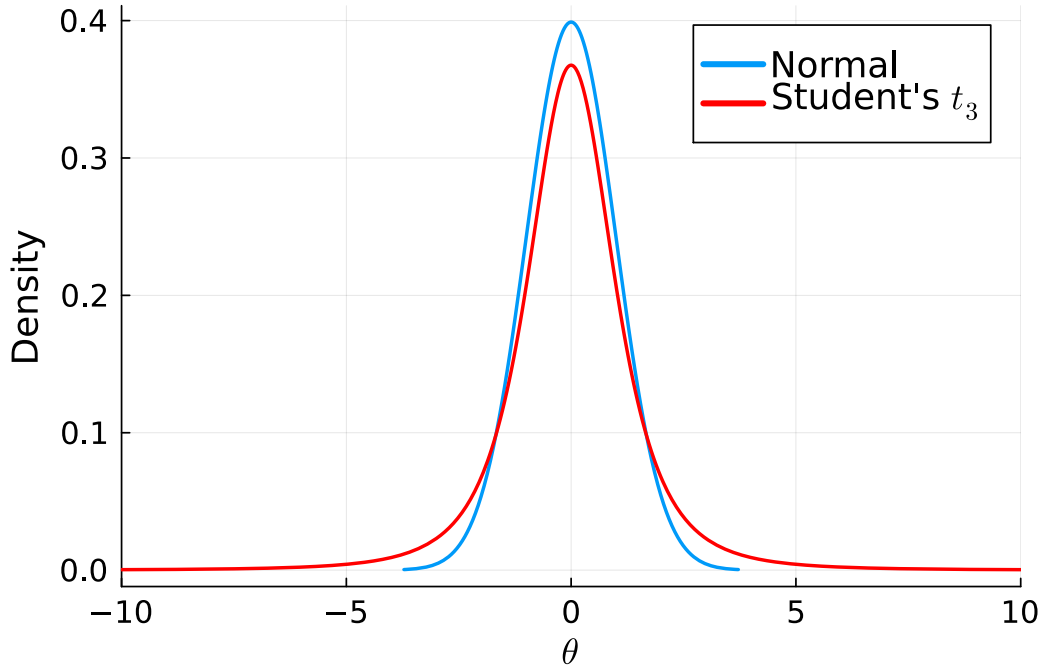
1. Simulate $u \sim \text{Unif}(0, 1)$.
2. Simulate a proposal $\hat{\theta} \sim g(\theta)$.
3. If

$$u < \frac{p(\hat{\theta}|\mathbf{y})}{Mg(\hat{\theta})},$$

accept $\hat{\theta}$. Otherwise reject.

Rejection Sampling Intuition

We want to keep more samples from the areas where $g(\theta) < p(\theta|\mathbf{y})$ and reject where g is heavier (in this case, the tails).



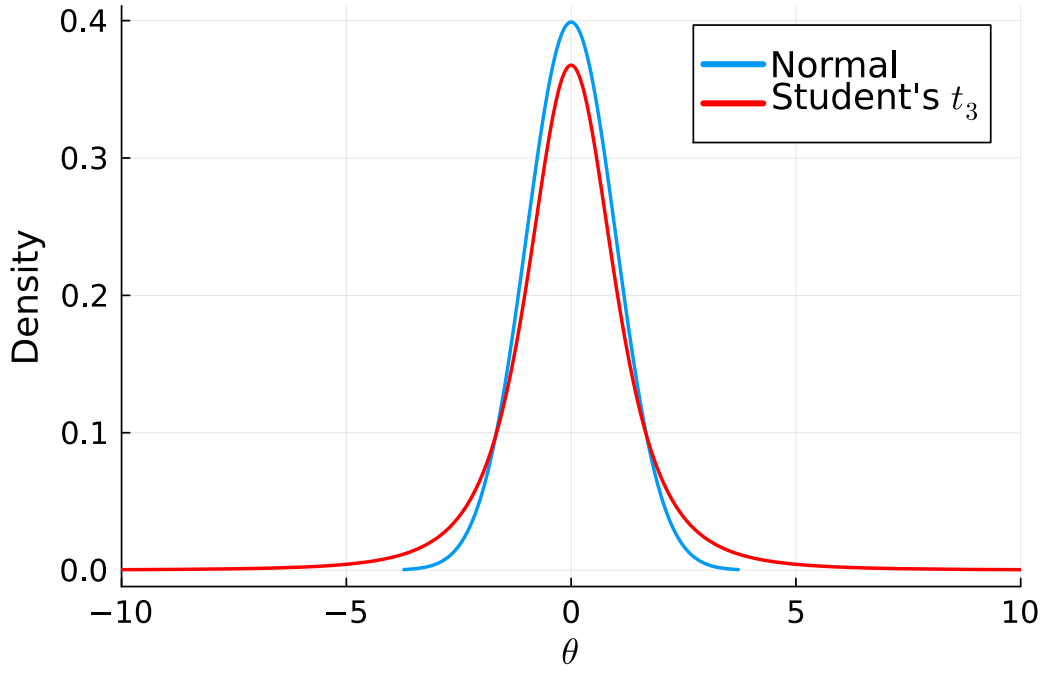
Rejection Sampling Intuition

In the bulk: $Mg(\theta)$ is closer to $p(\theta|\mathbf{y})$, so the acceptance ratio is higher.

In the tails:

$$Mg(\theta) \gg p(\theta|\mathbf{y}),$$

so the acceptance ratio is much lower.



Rejection Sampling Considerations

1. Probability of accepting a sample is $1/M$, so the "tighter" the proposal distribution coverage the more efficient the sampler.
2. Need to be able to compute M.

Finding a good proposal and computing M may not be easy (or possible) for complex posteriors!

How can we do better?

Idea of Better Approach

The fundamental problem with rejection sampling is that we don't know the properties of the posterior. So we don't know that we have the appropriate coverage. But...

What if we could construct an proposal/acceptance/rejection scheme that necessarily converged to the target distribution, even without *a priori* knowledge of its properties?

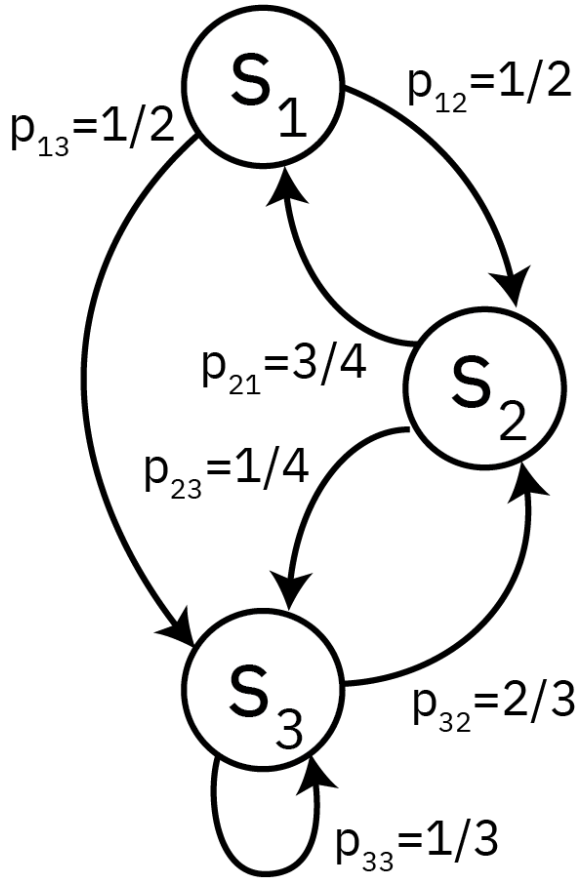
Idea: Develop a stochastic process based on **Markov chains**.

Markov Chains

What is a Markov Chain?

Consider a stochastic process $\{X_t\}_{t \in \mathcal{T}}$, where

- $X_t \in \mathcal{S}$ is the state at time t, and
- $\mathcal{T}$ is a time-index set (can be discrete or continuous)
- $\mathbb{P}(s_i \to s_j) = p_{ij}$.



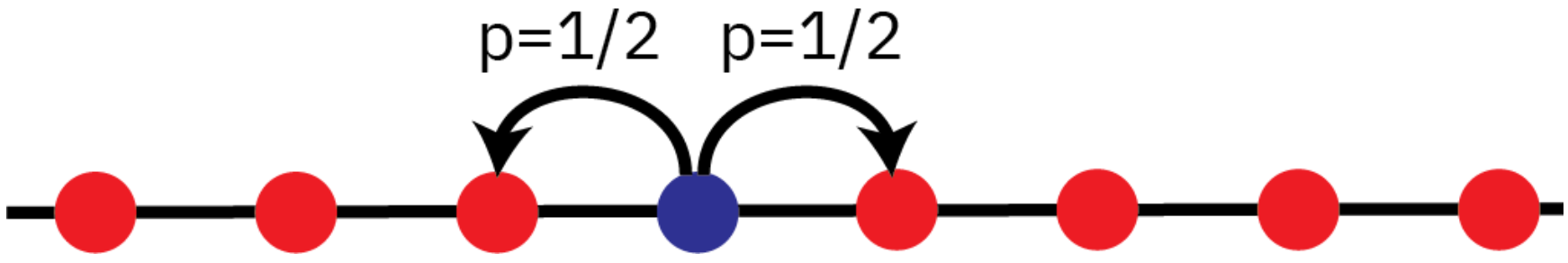
What is a Markov Chain?

This stochastic process is a **Markov chain** if it satisfies the **Markovian (or memoryless) property**:

$$\mathbb{P}(X_{T+1} = s_i | X_1 = x_1, \ldots, X_T = x_T) = \mathbb{P}(X_{T+1} = s_i | X_T = x_T)$$

In other words: the probability of being in any given state x_i at time $T + 1$ only depends on the prior state X_T, not the previous history.

Example: "Drunkard's Walk"



Consider a process where we can "stumble" to the left or right with equal probability.

The *unconditional* probability $\mathbb{P}(X_T = s_i)$ can be modeled by a sum of coin flips from the initial state X_0, but the *conditional* probability $\mathbb{P}(X_T = s_i | X_{T-1} = x_{T-1})$ only depends on the current node, not how we got there.

Weather Example

Let's look at a more interesting example. Suppose the weather can be foggy, sunny, or rainy.

Based on past experience, we know that:

1. There are never two sunny days in a row;
2. Even chance of two foggy or two rainy days in a row;
3. A sunny day occurs 1/4 of the time after a foggy or rainy day.

Aside: Higher-Order Markov Chains

Suppose that today's weather depends on the prior *two* days.

1. Can we write this as a Markov chain?

Aside: Higher-Order Markov Chains

Suppose that today's weather depends on the prior *two* days.

1. Can we write this as a Markov chain?
2. What are the states?

Weather Example: Transition Matrix

We can summarize these probabilities in a **transition matrix** P:

$$P = \begin{array}{c} \begin{matrix} F & S & R \end{matrix} \\ \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix} \end{array} \begin{matrix} F \\ S \\ R \end{matrix}$$

Rows are the current state, columns are the next step, so $\sum_i p_{ij} = 1$.

Weather Example: State Probabilities

Denote by λ^t a probability distribution over the states at time t.

Then $\lambda^t = \lambda^{t-1} P$:

$$\begin{pmatrix} \lambda_F^t & \lambda_S^t & \lambda_R^t \end{pmatrix} = \begin{pmatrix} \lambda_F^{t-1} & \lambda_S^{t-1} & \lambda_R^{t-1} \end{pmatrix} \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}$$

Multi-Transition Probabilities

Notice that

$$\lambda^{t+i} = \lambda^t P^i,$$

so multiple transition probabilities are P-exponentials.

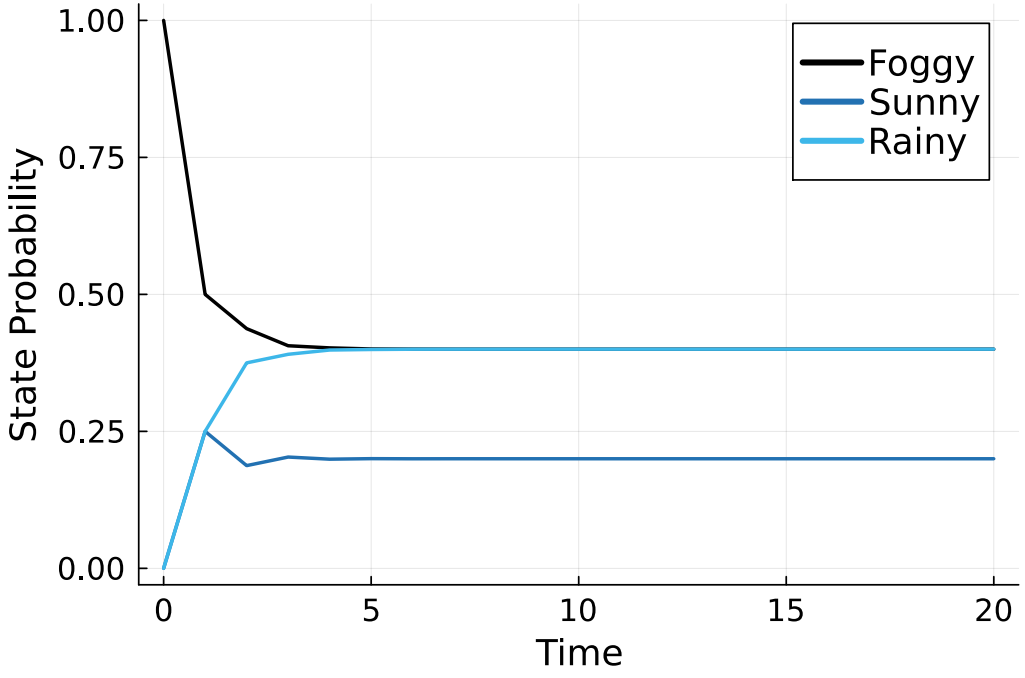
$$P^3 = \begin{array}{c} \begin{matrix} \color{red}{F} & \color{red}{S} & \color{red}{R} \end{matrix} \\ \begin{pmatrix} 26/64 & 13/64 & 25/64 \\ 26/64 & 12/64 & 26/64 \\ 26/64 & 13/64 & 26/64 \end{pmatrix} \end{array} \begin{matrix} \color{red}{F} \\ \color{red}{S} \\ \color{red}{R} \end{matrix}$$

Long-Run Probabilities

What happens if we let the system run for a while starting from an initial sunny day?

Long-Run Probabilities

What happens if we let the system run for a while starting from an initial sunny day?



Notice that the probabilities eventually stabilize.

Stationary Distributions

This stabilization always occurs when the probability distribution is an eigenvector of P with eigenvalue 1:

$$\pi = \pi P.$$

This is called an *invariant* or a *stationary* distribution.

Stationary Distributions

Does every Markov chain have a stationary distribution?

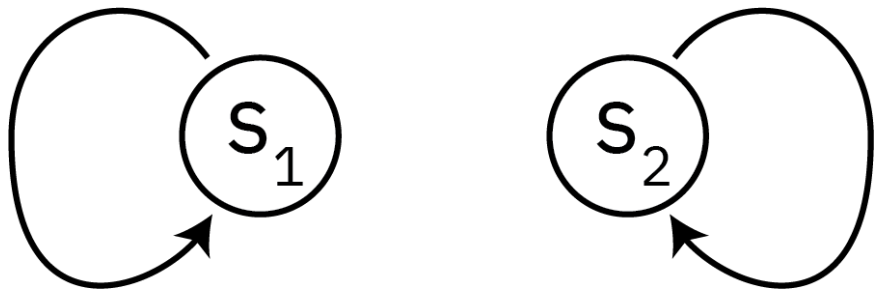
Not necessarily! The key is two properties:

- Irreducible
- Aperiodicity

Irreducibility

A Markov chain is **irreducible** if every state is accessible from every other state, *e.g.* for every pair of states s_i and s_j there is some $k > 0$ such that $P_{ij}^k > 0$.

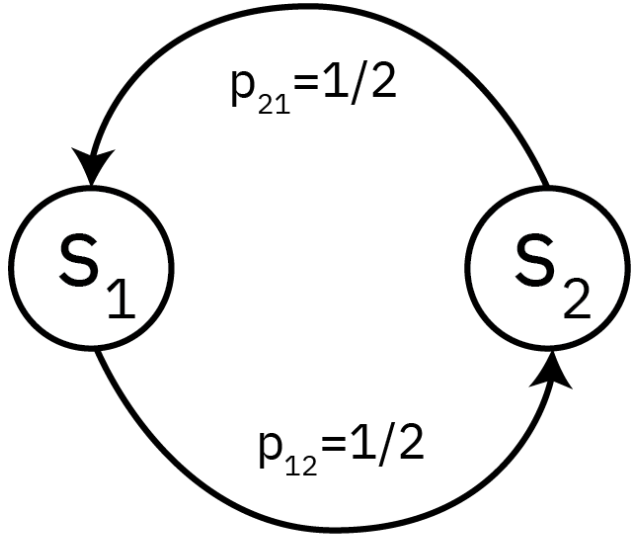
Here is an example of a reducible Markov chain:



Aperiodicity

The period of a state s_i is the greatest common divisor k of all t such that $P_{ii}^t > 0$. In other words, if a state s_i has period k, all returns must occur after time steps which are multiples of k.

A Markov chain is **aperiodic** if all states have period 1.



Ergodic Markov Chains

A Markov chain is **ergodic** if it is aperiodic and irreducible.

Ergodic Markov chains have a *limiting* distribution which is the limit of the time-evolution of the chain dynamics, *e.g.*

$$\pi_j = \lim_{t \to \infty} \mathbb{P}(X_t = s_j).$$

Key: this limit is *independent* of the initial state probability.

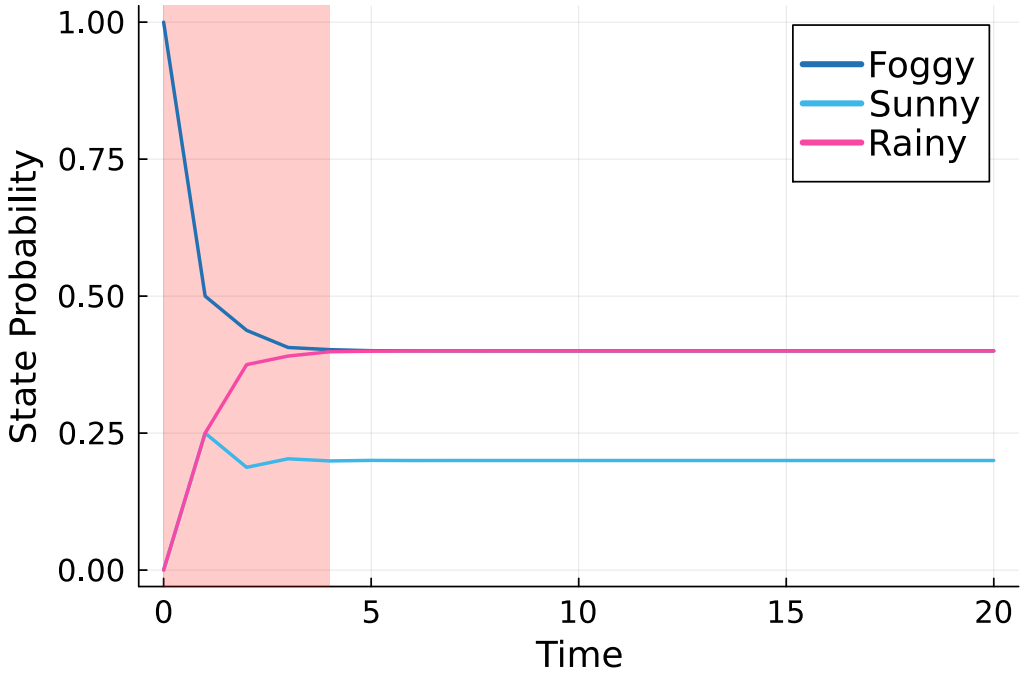
Intuition: Ergodicity means we can exchange thinking about *time-averages* and *ensemble-averages*.

Limiting Distributions Are Stationary

For an ergodic chain, the limiting distribution is the unique stationary distribution (we won't prove uniqueness):

$$\begin{aligned}
\pi_j &= \lim_{t\to\infty} \mathbb{P}(X_t = s_j | X_0 = s_i) \\
&= \lim_{t\to\infty} (P^{t+1})_{ij} = \lim_{t\to\infty} (P^t P)_{ij} \\
&= \lim_{t\to\infty} \sum_d (P^n)_{id} P_{dj} \\
&= \sum_d \pi_d P_{dj}
\end{aligned}$$

Transient Portion of the Chain



The portion of the chain prior to convergence to the stationary distribution is called the **transient** portion.

This will be important next week!

Detailed Balance

The last important concept is **detailed balance**.

Let $\{X_t\}$ be a Markov chain and let π be a probability distribution over the states. Then the chain is in detailed balance with respect to π if

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

Detailed Balance

The last important concept is **detailed balance**.

Let $\{X_t\}$ be a Markov chain and let π be a probability distribution over the states. Then the chain is in detailed balance with respect to π if

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

Detailed balance implies **reversibility**: the chain's dynamics are the same when viewed forwards or backwards in time.

Detailed Balance and Stationary Distributions

Detailed balance is a sufficient but not necessary condition for the existence of a stationary distribution (namely π):

$$\begin{aligned}(\pi P)_i &= \sum_j \pi_j P_{ji} \\ &= \sum_j \pi_i P_{ij} \\ &= \pi_i \sum_j P_{ij} = \pi_i\end{aligned}$$

Intuition About Detailed Balance

What does detailed balance mean? Let's compare with the definition of a stationary distribution.

- The existence of a stationary distribution is a *global* condition: the sum of all probability out of any given node has to equal the total incoming probability.

Intuition About Detailed Balance

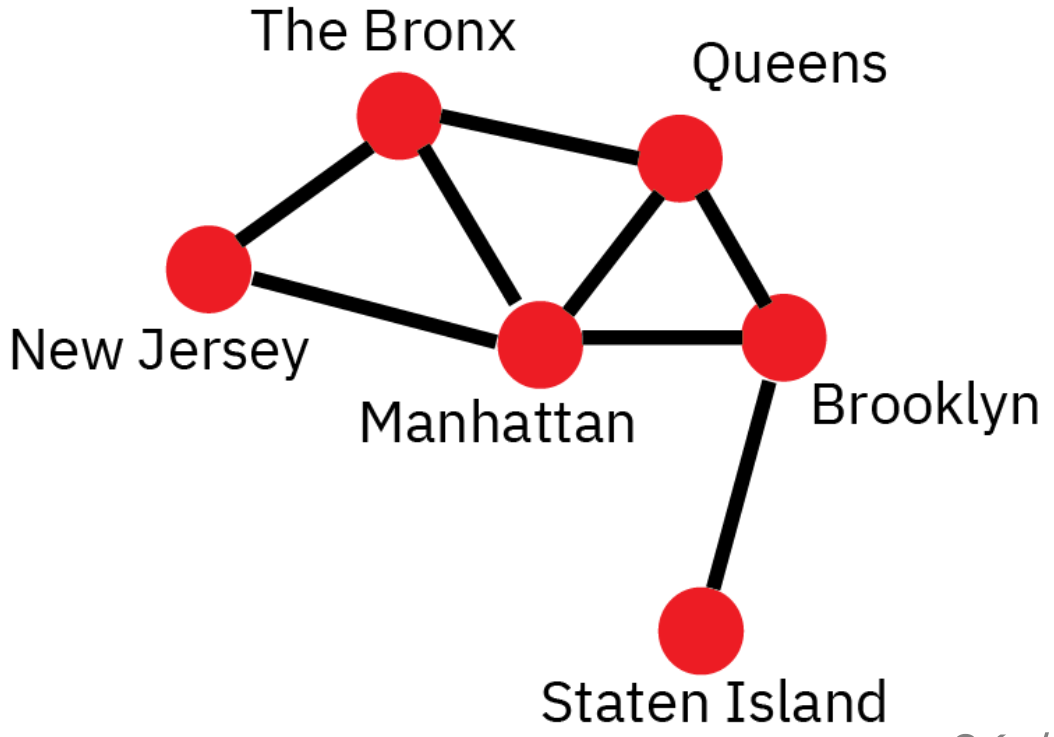
What does detailed balance mean? Let's compare with the definition of a stationary distribution.

- The existence of a stationary distribution is a *global* condition: the sum of all probability out of any given node has to equal the total incoming probability.
- Detailed balance is a stronger *local* condition: not just that the total probability in and out of all nodes, but that the flow of probability must be balanced across every transition.

Detailed Balance Analogy

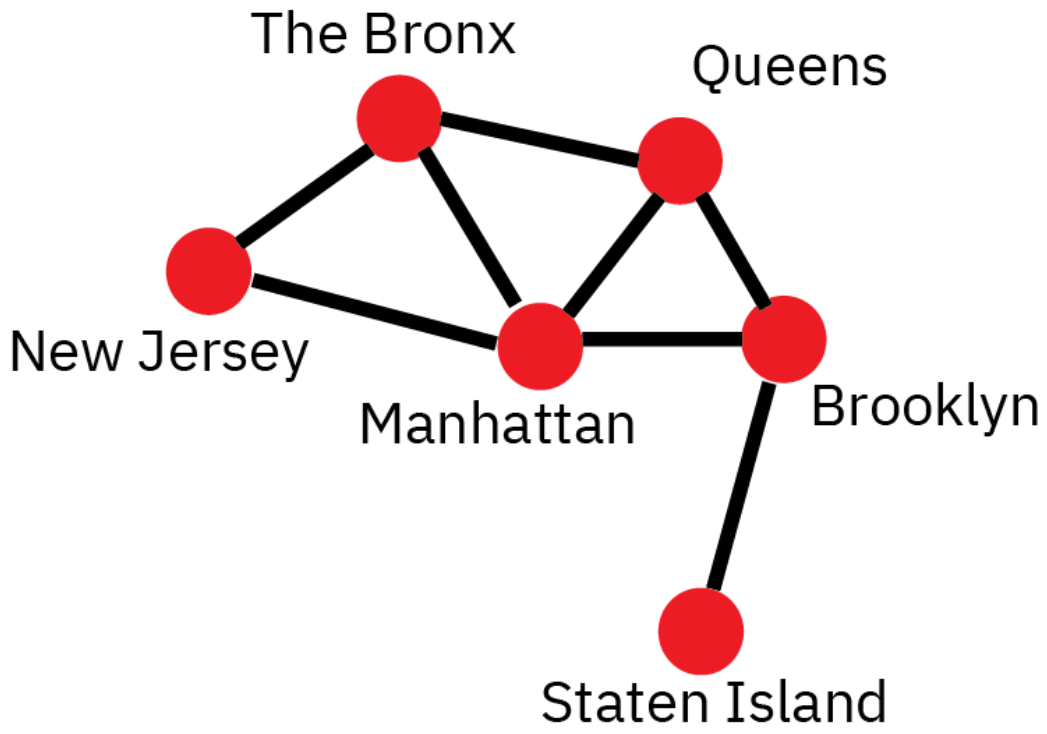
A nice analogy (from Miranda Holmes-Cerfon) is traffic flow.

Consider NYC and its surroundings: each borough/region can be thought of as a node, and population transitions occur across bridges/tunnels.



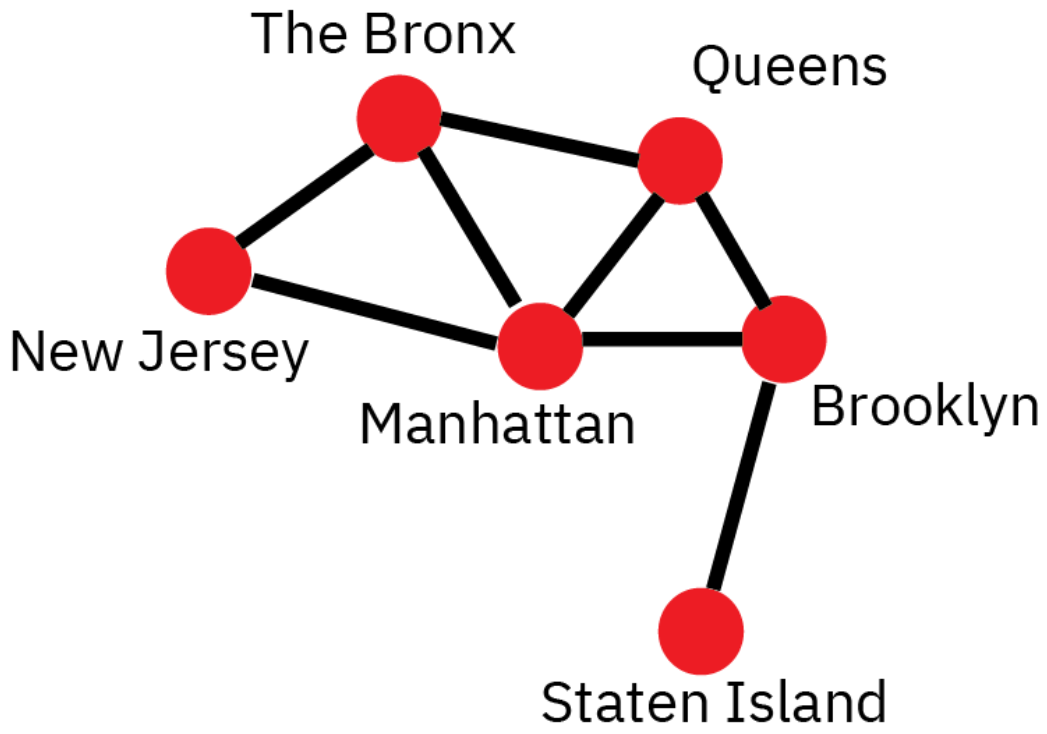
Detailed Balance Analogy

The stationary criterion means that the number of cars per unit time leaving each borough is the same as those entering, regardless of how they move.



Detailed Balance Analogy

Detailed balance means traffic must be balanced across each bridge or tunnel per unit time.



Idea of Sampling Algorithm

The idea of our sampling algorithm (which we will discuss next time) is to construct an ergodic Markov chain from the detailed balance equation for the target distribution.

- Detailed balance implies that the target distribution is the stationary distribution.
- Ergodicity implies that this distribution is unique and can be obtained as the limiting distribution of the chain's dynamics.

Idea of Sampling Algorithm

In other words:

- Generate an appropriate Markov chain,
- Run its dynamics long enough to converge to the stationary distribution,
- Use the resulting ensemble of states as a Monte Carlo sample.

Key Takeaways

Key Takeaways: Markov Chains

- Markov chains are a very useful class of stochastic processes.
- If a chain is ergodic, a stationary distribution exists.
- The stationary distribution is the limit of the time-evolution of the ensemble of states.
- Can split Markov chain dynamics into "transient" and stationary portion.
- Our goal: construct a Markov chain whose stationary distribution is the posterior of our model (this is Markov chain Monte Carlo).
- Today's notation focused on chains on discrete state spaces, but everything maps directly to continuous spaces.

Upcoming Schedule

Upcoming Schedule

Wednesday: Discussion of Oppenheimer et al (2008)

Next Monday: Markov chain Monte Carlo