# Markov Chain Monte Carlo

# Why Stochastic Sampling?

- We want to compute the expectation of some function relative to some “difficult” distribution (usually posterior distribution, p(Y|e))
- We could approximate the expectation by sampling from the distribution
- But directly sampling from the distribution is intractable
- We need independent samples to estimate the desired distribution

# Why Stochastic Sampling?

- We can use rejection sampling but it is inefficient
- We can use likelihood weighting but
  - Evidence nodes affect sampling only for their descendants
  - When evidence is mostly at the leaf nodes, we effectively sample from the prior distribution which can be different from posterior distribution
  - Therefore likelihood weighting introduces bias toward the prior in the sample

# Why Stochastic Sampling?

- We will introduce another sampling method, Markov Chain Monte Carlo (MCMC) that uses a Markov chain to generate samples
- Understanding MCMC well depends on a basic understanding of Markov chains
- We will give a brief introduction today to Markov chains before continuing with MCMC on Friday

# Definition

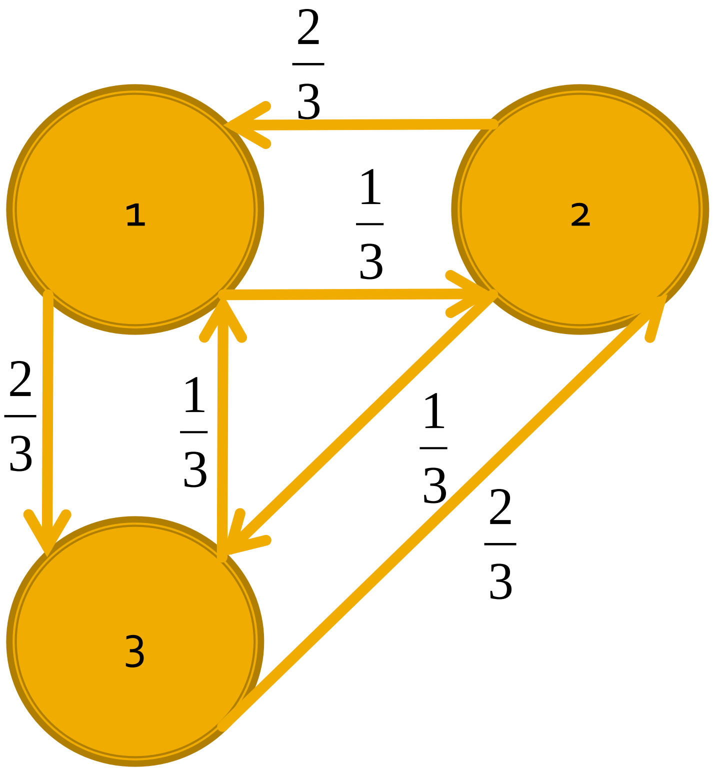
- A discrete time Markov chain is a sequence of random variables whose distributions are related in a particular way (a *stochastic process*)

$$X_0, X_1 \ldots X_n = (X_n)_{n \geq 0}$$

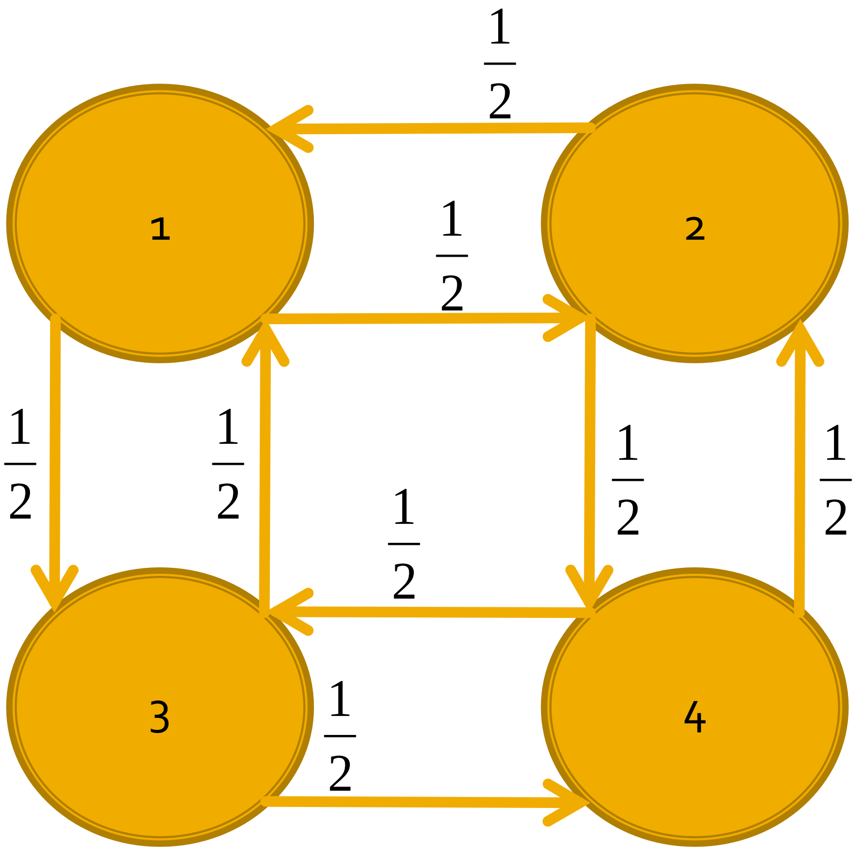
- All of these random variables are drawn from the same set, called the *state space*.

$$X_0, X_1 \ldots X_n \in I$$

# Example

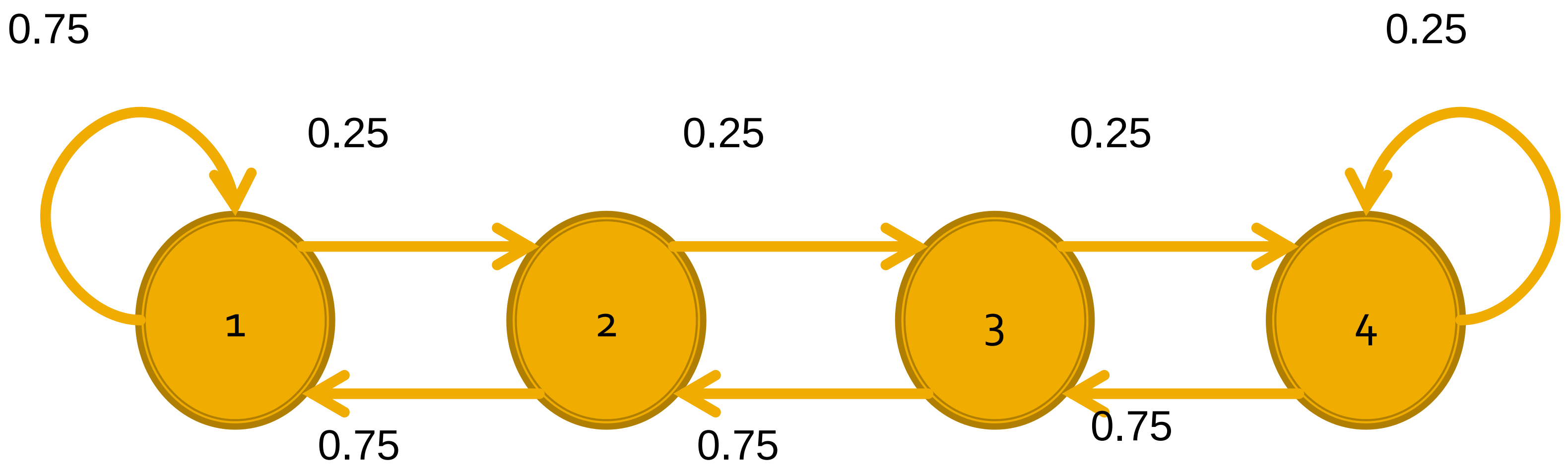


# Example



$\frac{1}{2}$
1
2
$\frac{1}{2}$
$\frac{1}{2}$
$\frac{1}{2}$
$\frac{1}{2}$
$\frac{1}{2}$
$\frac{1}{2}$
3
4
$\frac{1}{2}$

# Example



# Definition

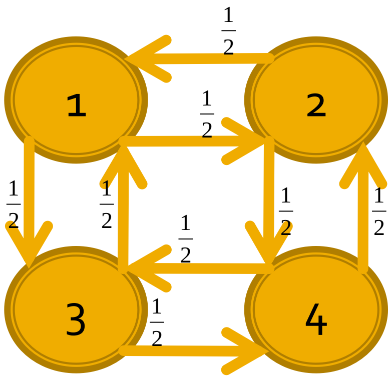
- Each Markov chain is defined by an initial distribution vector and a stochastic transition matrix.

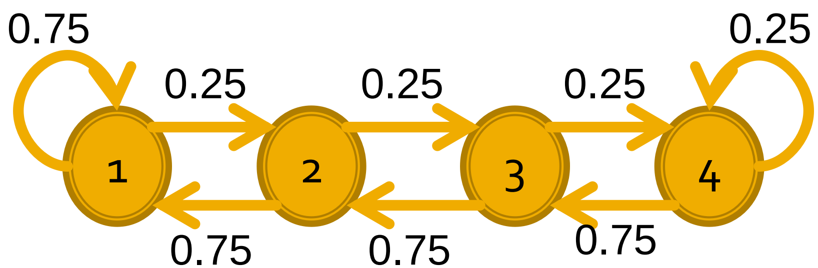
$$(X_n)_{n\geq 0} \sim Markov(\lambda, \mathrm{P})$$

$$\sum_{i\in I} \lambda_i = 1$$

$$\sum_{j\in I} P_{ij} = 1 \text{ for } i \in I$$

# Example




$$\lambda = [0.25 \quad 0.25 \quad 0.25 \quad 0.25]$$

$$P = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0.5 & 0.5 & 0 \end{bmatrix}$$

$$\lambda = [1 \quad 0 \quad 0 \quad 0]$$

$$P = \begin{bmatrix} 0.75 & 0.25 & 0 & 0 \\ 0.75 & 0 & 0.25 & 0 \\ 0 & 0.75 & 0 & 0.25 \\ 0 & 0 & 0.75 & 0.25 \end{bmatrix}$$

# Definition

- A set of random variables is Markov if:

$$P(X_0 = i) = \lambda_i$$

- and

$$P(X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \ldots, X_0 = i_0) = \mathrm{P}_{i_n i_{n+1}}$$

# Markov Property

- Given a Markov process $(X_n)_{n\geq 0} \sim Markov(\lambda, \mathrm{P})$
- Conditioned on current state $X_m = i$

$$(X_{m+n})_{n\geq 0} = (X_m, X_{m+1,} \ldots, X_{m+n}) \sim Markov(\delta_i, \mathrm{P})$$

- We can forget the past. The Markov process has no memory. Only the current state matters for deciding the future.

$$P(X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \ldots, X_0 = i_0) = P(X_{n+1} = i_{n+1} \mid X_n = i_n)$$

# Transition Matrix

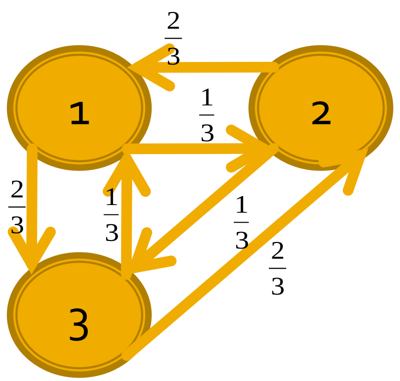
- Multiplying the current distribution by the transition matrix gives the distribution of being in the next state

$$P(X_t) = \vec{\gamma} \Rightarrow P(X_{t+1}) = \vec{\gamma}\mathrm{P}$$

$$P(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \lambda \mathrm{P}_{i_0,i_1} \mathrm{P}_{i_1,i_2} \ldots \mathrm{P}_{i_{n-1},i_n}$$

$$P(X_n) = \lambda \mathrm{P}^n$$

# Example




$$\lambda = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

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

$$\lambda P = \begin{bmatrix} 0 & 1/3 & 2/3 \end{bmatrix}$$

$$\lambda P^2 = \begin{bmatrix} 4/9 & 4/9 & 1/9 \end{bmatrix}$$

$$\lambda P^3 = \begin{bmatrix} 3/9 & 2/9 & 4/9 \end{bmatrix}$$

$$\lambda P^4 = \begin{bmatrix} 8/27 & 11/27 & 8/27 \end{bmatrix}$$

$$\lambda P^5 = \begin{bmatrix} 10/27 & 8/27 & 9/27 \end{bmatrix}$$

$$\lambda P^6 = \begin{bmatrix} 25/81 & 28/81 & 28/81 \end{bmatrix}$$

# Review

- A Markov process is defined by a stochastic transition matrix and an initial distribution.
- In a Markov chain, the future is independent of the past. It only depends on the present.
- Multiplying a distribution on the state space by the transition matrix gives the distribution after one transition.

# Remembering Our Goal

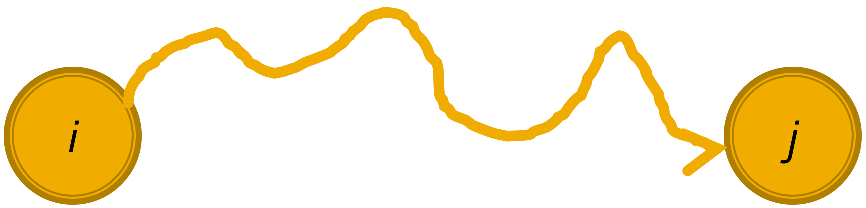
- We are going to generate samples from a Markov process.
- Ideally, the Markov process should behave in "predictable" ways.
- In a sense, we are trying to make a "stable" Markov chain.
- We are going to discuss several properties of a Markov chain that will make it become "stable".

# Irreducibility

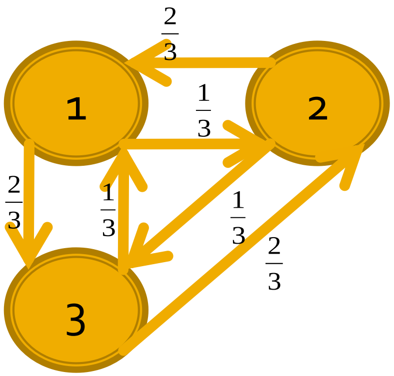
- P is irreducible if

$$\forall i, j \in I; \exists\, m > 0 \;\; s.t \;\; \mathrm{P}_{ij}^{m} > 0$$

$$\mathrm{P}^{m} = \left[\begin{array}{c|cccccc} & 1 & 2 & \dots & j & \dots & N \\ \hline 1 & & & & & & \\ 2 & & & & & & \\ \vdots & & & & & & \\ i & & & & & & \\ \vdots & & & & & & \\ N & & & & & & \end{array}\right]^{m}$$



# Example

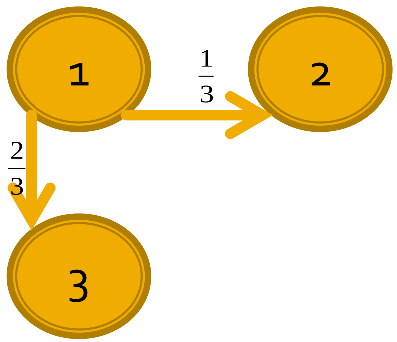



$$P = \begin{bmatrix} 0 & 1/3 & 2/3 \\ 2/3 & 0 & 1/3 \\ 1/3 & 2/3 & 0 \end{bmatrix}$$

$$P^2 = \begin{bmatrix} 4/9 & 4/9 & 1/9 \\ 1/9 & 4/9 & 4/9 \\ 4/9 & 1/9 & 4/9 \end{bmatrix}$$

This process is irreducible

# Example




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

$$P^2 = P^3 = P^n = \begin{bmatrix} 0 & 1/3 & 2/3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

This process is reducible

# Recurrence vs. Transient

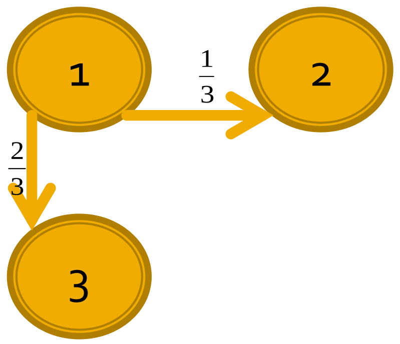
- State *i* is called recurrent if we keep coming back to it.

$$P_i\left(X_n = i \text{ for infinitely many } n\right) = 1$$

- And *i* is called transient if we eventually leave it forever.

$$P_i\left(X_n = i \text{ for infinitely many } n\right) = 0$$

# Example




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

$$P^2 = P^3 = P^n = \begin{bmatrix} 0 & 1/3 & 2/3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

State 1 is transient

# Positive Recurrent

- State *i* is called positive recurrent if the expected return time to the state is finite.

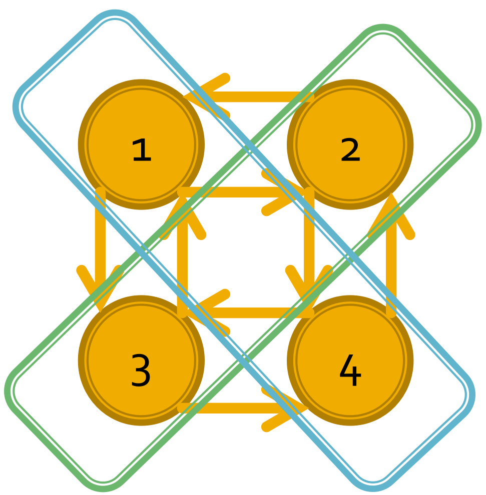
$$E_i(T_i) < \infty$$

- Where $T_i = \inf\{n \geq 0 : X_n = i\}$

- This always true for a recurrent Markov process with finite states, so we won't give examples

# Aperiodicity

- State *i* is called aperiodic if we could return back to the same state with any number of transitions for sufficiently large number of transitions $P_{ii}^{(n)} > 0$




$$P^{2n+1} = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0.5 & 0.5 & 0 \end{bmatrix} \qquad P^{2n} = \begin{bmatrix} 0.5 & 0 & 0 & 0.5 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \end{bmatrix}$$

# Review

- Properties of processes
  - Irreducibility – after waiting a certain time, we have a non-zero probability of getting from any state to any state
- Properties of states
  - Recurrence – we can always return to this state
  - Positive recurrence –we expect to return to this state in a finite amount of time
  - Aperiodicity – we can return to the same state after any number of transitions (after a certain m)

# Invariant (Equilibrium) Distribution

- We can imagine a Markov process that gets "stuck" in a single distribution

$$\pi = \pi \mathrm{P} = \pi \mathrm{P}^n$$

- Not all Markov processes have an invariant distribution
  - States must be positive recurrent
- The invariant distribution is an eigenvector of the transition matrix with eigenvalue one
- The invariant distribution is unique

# Invariant (Equilibrium) Distribution

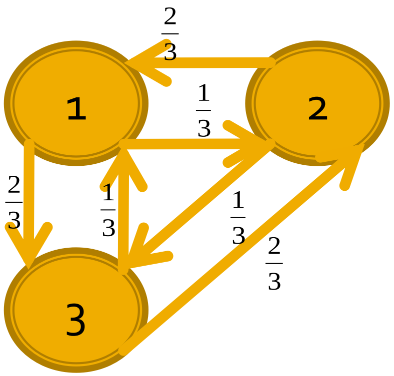
- Since we are sampling from a Markov process, it would be nice if we could design that process to get "stuck" in a distribution we desire
- If the Markov process is in its equilibrium distribution, every sample from it will be an independent sample from the invariant distribution

# Convergence to Equilibrium

- If a Markov process is irreducible and aperiodic, it will always converge to its equilibrium distribution

$$P(X_n = j) \to \pi_j \text{ as } n \to \infty$$

# Example




$$\lambda = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

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

$$\lambda P = \begin{bmatrix} 0 & .3333 & .6667 \end{bmatrix}$$

$$\lambda P^2 = \begin{bmatrix} .4444 & .4444 & .1111 \end{bmatrix}$$

$$\lambda P^3 = \begin{bmatrix} .3333 & .2222 & .4444 \end{bmatrix}$$

$$\lambda P^4 = \begin{bmatrix} .2963 & .4074 & .2963 \end{bmatrix}$$

$$\lambda P^5 = \begin{bmatrix} .3704 & .2963 & .3333 \end{bmatrix}$$

$$\lambda P^6 = \begin{bmatrix} .3086 & .3457 & .3457 \end{bmatrix}$$

$$\lambda P^n = \pi = \begin{bmatrix} .3333 & .3333 & .3333 \end{bmatrix}$$ For large enough n

# Review

- Provided we have an irreducible aperiodic positive recurrent Markov process, it will converge to an equilibrium distribution and stay in that distribution

# Again Remembering Our Goal

- So we can converge to an equilibrium distribution if our Markov process is designed correctly.
- If we could make a process converge to a specific distribution, we could sample from that distribution.
- How can we design a Markov process with a desired equilibrium distribution?

# Time Reversal

- What happens if we run a Markov chain at equilibrium in reverse?

$$(X_n)_{n\geq 0} \sim Markov(\pi, \mathrm{P})$$

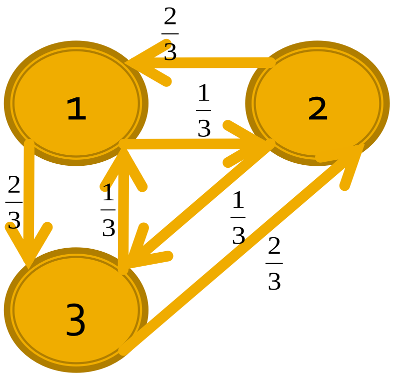
$$Y_n = X_{N-n} \sim Markov(\pi, \hat{\mathrm{P}})$$

$$\pi_j \hat{\mathrm{P}}_{ji} = \pi_i \mathrm{P}_{ij}$$

- Proof: need to show $\pi\hat{\mathrm{P}} = \pi$

$$(\pi\hat{\mathrm{P}})_i = \sum_{j\in I} \pi_j \hat{\mathrm{P}}_{ji} = \sum_{j\in I} \pi_i \mathrm{P}_{ij} = \pi_i \sum_{j\in I} \mathrm{P}_{ij} = \pi_i \times 1 = \pi_i$$

# Example




$$\pi = \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix} \quad \mathrm{P} = \begin{bmatrix} 0 & 1/3 & 2/3 \\ 2/3 & 0 & 1/3 \\ 1/3 & 2/3 & 0 \end{bmatrix}$$

$$\hat{\mathrm{P}} = \begin{bmatrix} 0 & 2/3 & 1/3 \\ 1/3 & 0 & 2/3 \\ 2/3 & 1/3 & 0 \end{bmatrix}$$

$$\pi \mathrm{P}^n = \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix}$$

$$\pi \hat{\mathrm{P}}^n = \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix}$$

# Detailed Balance

- What if we look for a Markov chain such that

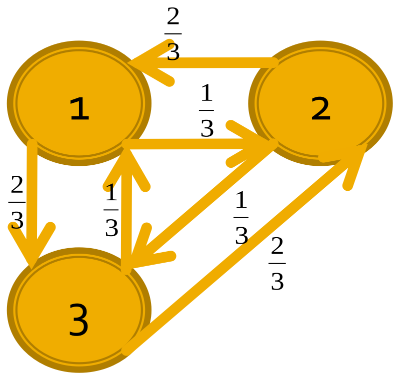
$$\hat{\mathrm{P}} = \mathrm{P}$$

- The Markov chain is the same whether we run it in the forward or reverse direction
- We call such a Markov process *reversible*
- Reversibility implies

$$\pi_j \mathrm{P}_{ji} = \pi_i \mathrm{P}_{ij}$$

- Such a Markov chain is said to be in *detailed balance*

# Example




$$\pi = \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix} \quad \mathrm{P} = \begin{bmatrix} 0 & 1/3 & 2/3 \\ 2/3 & 0 & 1/3 \\ 1/3 & 2/3 & 0 \end{bmatrix}$$

$$\hat{\mathrm{P}} = \begin{bmatrix} 0 & 2/3 & 1/3 \\ 1/3 & 0 & 2/3 \\ 2/3 & 1/3 & 0 \end{bmatrix}$$

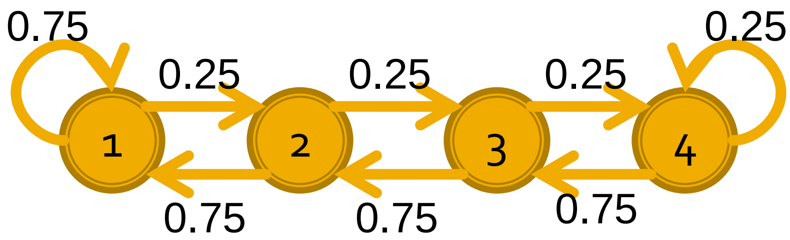
$$\pi \mathrm{P}^n = \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix}$$

$$\pi \hat{\mathrm{P}}^n = \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix}$$

$$\mathrm{P} \neq \hat{\mathrm{P}}$$

This process is *not* reversible (symmetric in time)

# Example



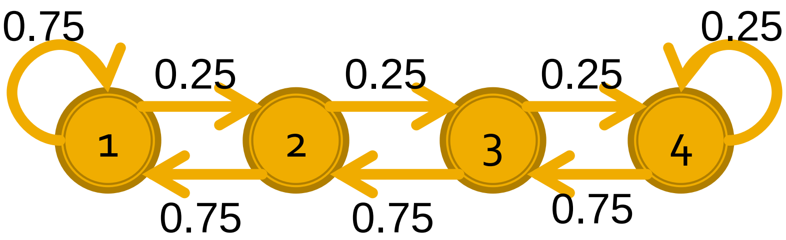

$$\pi_1 P_{12} = \pi_2 P_{21}$$

$$\pi_2 P_{23} = \pi_3 P_{32}$$

$$\pi_3 P_{34} = \pi_4 P_{43}$$

$$\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$$

# Example



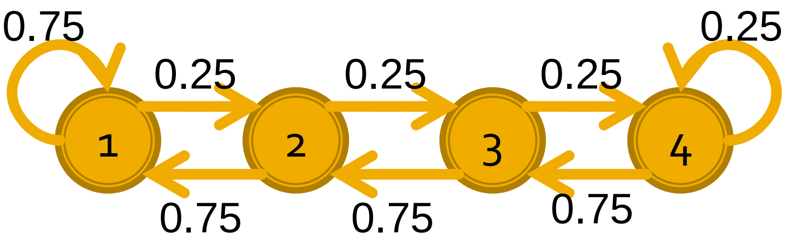

$$0.25\pi_1 = 0.75\pi_2$$

$$0.25\pi_2 = 0.75\pi_3$$

$$0.25\pi_3 = 0.75\pi_4$$

$$\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$$

# Example




$$\pi_1 = 3\pi_2$$
$$\pi_1 = 9\pi_3$$
$$\pi_1 = 27\pi_4$$
$$\pi_1 + \frac{1}{3}\pi_1 + \frac{1}{9}\pi_1 + \frac{1}{27}\pi_1 = 1$$

$$\pi_1 = 27/40$$
$$\pi_2 = 9/40$$
$$\pi_3 = 3/40$$
$$\pi_4 = 1/40$$

# Detailed Balance

- As long as our transition matrix is in detailed balance with our desired distribution, our Markov process will eventually converge to our desired distribution
- Maybe we can sample from such a process, even though *direct sampling from the desired distribution is intractable* (due to the large state space)

$$\pi_j \mathrm{P}_{ji} = \pi_i \mathrm{P}_{ij}$$

# Review

- A Markov process is defined by an initial distribution and a transition matrix

$$Markov(\lambda, \mathrm{P})$$

- The Markov property states that the future depends only on the present

$$P(X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \ldots, X_0 = i_0) = P(X_{n+1} = i_{n+1} \mid X_n = i_n)$$

# Review

- A Markov process that is irreducible, aperiodic, and positively recurrent will converge to its equilibrium distribution

$$\pi = \pi \mathrm{P} = \pi \mathrm{P}^n$$

$$P(X_n = j) \to \pi_j \text{ as } n \to \infty$$

- A reversible Markov chain is in detailed balance

$$\pi_j \mathrm{P}_{ji} = \pi_i \mathrm{P}_{ij}$$

# Next time…

- We will discuss techniques to create Markov processes that are in detailed balance with a specified distribution.
- In this way, we can solve our sampling problem.