# Lecture #7: Markov Chain Monte Carlo

## AM 207: Advanced Scientific Computing

### Stochastic Methods for Data Analysis, Inference and Optimization

Fall, 2020

IACS

VE RI TAS

## Outline

1. Gibbs Sampler for a Discrete Distribution
2. Definition and Properties of Markov Chains
3. Markov Chain Monte Carlo

## Motivation

Recall that the Gibbs sampler was a sampling technique that we introduced along with rejection sampling and inverse CDF sampling.

We applied this sampler to sample from the posterior of a semi-conjugate Bayesian model (normal likelihood and Inverse-Gamma prior on the mean parameter). Using rejection sampling on this posterior would have been quite difficult (recall your experiments tuning rejection sampling for the non-conjugate Bayesain model for birth weights in *In-Class Exercise 09.17*).

But unlike in the case of rejection sampling and inverse CDF sampling, we never proved the correctness of this sampler!

# Gibbs Sampler for a Discrete Distribution

# Gibbs Sampler for a Bivariate Discrete Distribution

Suppose we have two independent random variables $X \sim Ber(0.2)$ and $Y \sim Ber(0.6)$. Their joint distribution is a categorical distribution:

$$p(X, Y) = [0.12 \quad 0.48 \quad 0.08 \quad 0.32]$$

over the set of possible outcomes $(X = 1, Y = 1), (X = 0, Y = 1), (X = 1, Y = 0), (X = 1, Y = 1)$.

A Gibbs sampler for $\overset{P(X,Y)}{\cancel{p(X)}}$ will start with a sample $(X = X_n, Y = Y_n)$ and then generate a sample $(X = X_{n+1}, Y = Y_{n+1})$ by

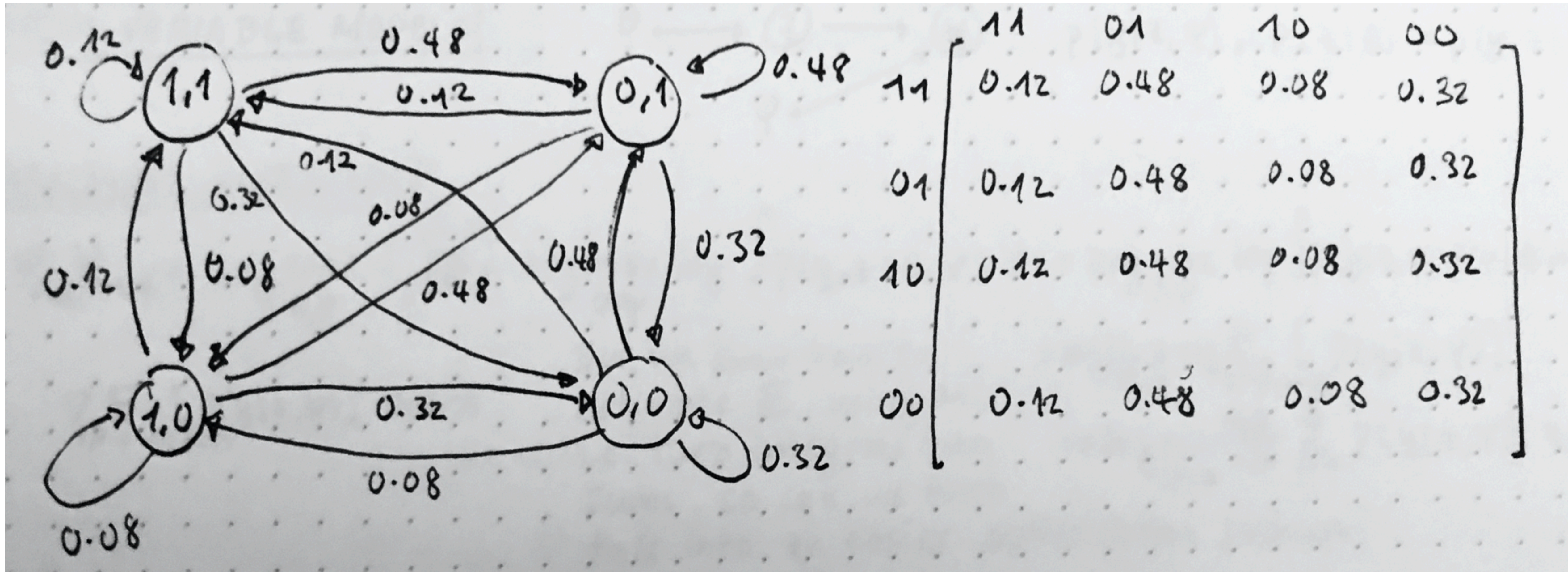
1. sampling $X_{n+1}$ from
   $$p(X|Y = Y_n)$$
2. sampling $Y_{n+1}$ from
   $$p(Y|X = X_{n+1})$$

We want to compute what is the distribution of the generated sample $(X = X_{n+1}, Y = Y_{n+1})$ given $(X = X_n, Y = Y_n)$, i.e. we want $p(X = X_{n+1}, Y = Y_{n+1} | X = X_n, Y = Y_n)$, but there are $4^2$ number of these probabilities! How do we succinctly represent them?

# Gibbs Sampler as Transition Matrix and State Diagram

We can represent the $p(X = X_{n+1}, Y = Y_{n+1} | X = X_n, Y = Y_n)$ as a $4 \times 4$ matrix, $T$, where the $i, j$-th entry is the probability of starting with sample $i$ and generating sample $j$.

Alternatively, we can visualize how the Gibbs sampler moves around in the sample space $(X, Y)$ with a diagram.



# Limiting Distribution

We see that computing the probability of the next sample given the current sample $(X = 1, Y = 1)$

$$p(X_{n+1}, Y_{n+1} | X_n = 1, Y_n = 1)$$

is equivalent to multiplying the vector $[1 \quad 0 \quad 0 \quad 0]$ with the matrix $T$. **Can you see why?**

When we do, we get the distribution $[0.12 \quad 0.48 \quad 0.08 \quad 0.32]$ over the next sample. But this distribution looks just like the joint distribution $p(X, Y)$!

This means that if we start at $(X = 1, Y = 1)$, the next sample the Gibbs sampler returns will be from the joint distribution.

In fact, you can start with any point in the samples space $(X, Y)$ or any distribution over the sample space, the next sample the Gibbs sampler returns will be from the joint distribution. I.e. any vector times $T$ will return $[0.12 \quad 0.48 \quad 0.08 \quad 0.32]$.

This proves the correctness of the Gibbs sampler for $p(X, Y)$!

# Definition and Properties of Markov Chains

# Markov Chains in Discrete and Continuous Spaces

A ***discrete-time stochastic process*** is set of random variables $\{X_0, X_1, \ldots\}$, where each random variable takes value in $\mathcal{S}$. The set $\mathcal{S}$ is called the ***state space*** and can be continuous or finite.

A stochastic process satisfies the ***Markov property*** if $X_n$ depends only on $X_{n-1}$ (i.e. $X_n$ is independent of $X_1, \ldots, X_{n-2}$). A stochastic process that satisfies the Markov property is called a ***Markov chain***.

We will assume that $p(X_n | X_{n-1})$ is the same for all $n$.

**Exercise:** Give an example of a stochastic process that is not a Markov chain. Given an example of a stochastic process that is a Markov chain.

# Transition Matrices and Kernels

The Markov property ensure that we can describe the dynamics of the entire chain by describing how the chain ***transitions*** from state $i$ to state $j$. **Why?**

If the state space is finite, then we can represent the transition from $X_{n-1}$ to $X_n$ as a ***transition matrix*** $T$, where $T_{ij}$ is the probability of the chain transitioning from state $i$ to $j$:

$$T_{ij} = \mathbb{P}[X_n = j | X_{n-1} = i].$$

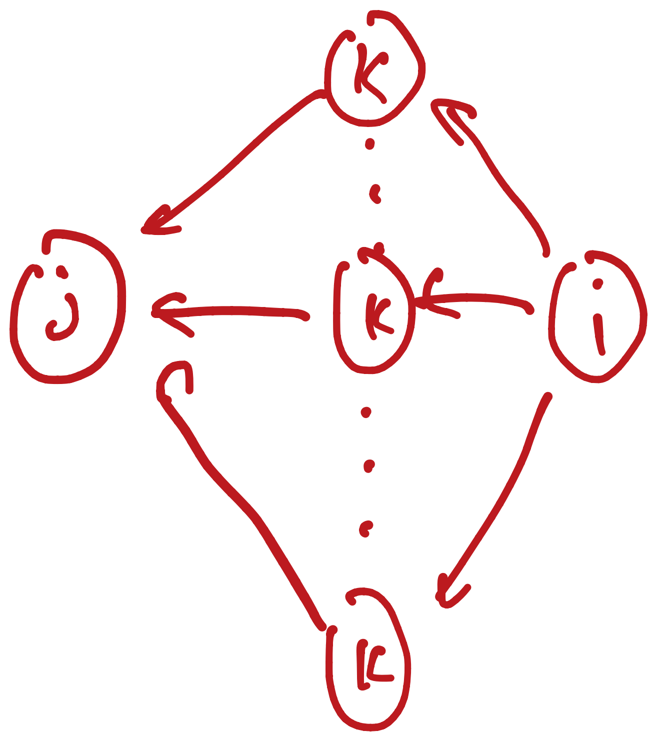
The transition matrix can be represented visually as a ***finite state diagram***.

If the state space is continuous, then we can represent the transition from $X_{n-1}$ to $X_n$ as ***transition kernel pdf***, $T(x, x')$, representing the likelihood of transitioning from state $X_{n-1} = x$ to state $X_n = x'$. The probability of transitioning into a region $A \subset S$ from state $x$ is given by

$$\mathbb{P}[X_n \in A | X_{n-1} = x] = \int_A T(x, y) dy,$$

such that $\int_S T(x, y) dy = 1$.

# Chapman-Kolmogorov Equations: Dynamics as Matrix Multiplication




If the state space is finite, the probability of the $n = 2$ state, given the initial $n = 0$ state. can be computed by the ***Chapman-Kolmogorov equation***:

$$\mathbb{P}[X_2 = j|X_0 = i] = \sum_{k \in \mathcal{S}} \mathbb{P}[X_1 = k|X_0 = i]\mathbb{P}[X_2 = j|X_1 = k] = \sum_{k \in \mathcal{S}} T_{ik}T_{kj}$$

We recognize $\sum_{k \in \mathcal{S}} T_{ik}T_{kj}$ as the $ij$-the entry in the matrix $TT$. Thus, the Chapman-Kolmogorov equation gives us that the matrix $T^{(n)}$ for a $n$-step transition is

$$T^{(n)} = \underbrace{T \dots T}_{n \text{ times}}$$

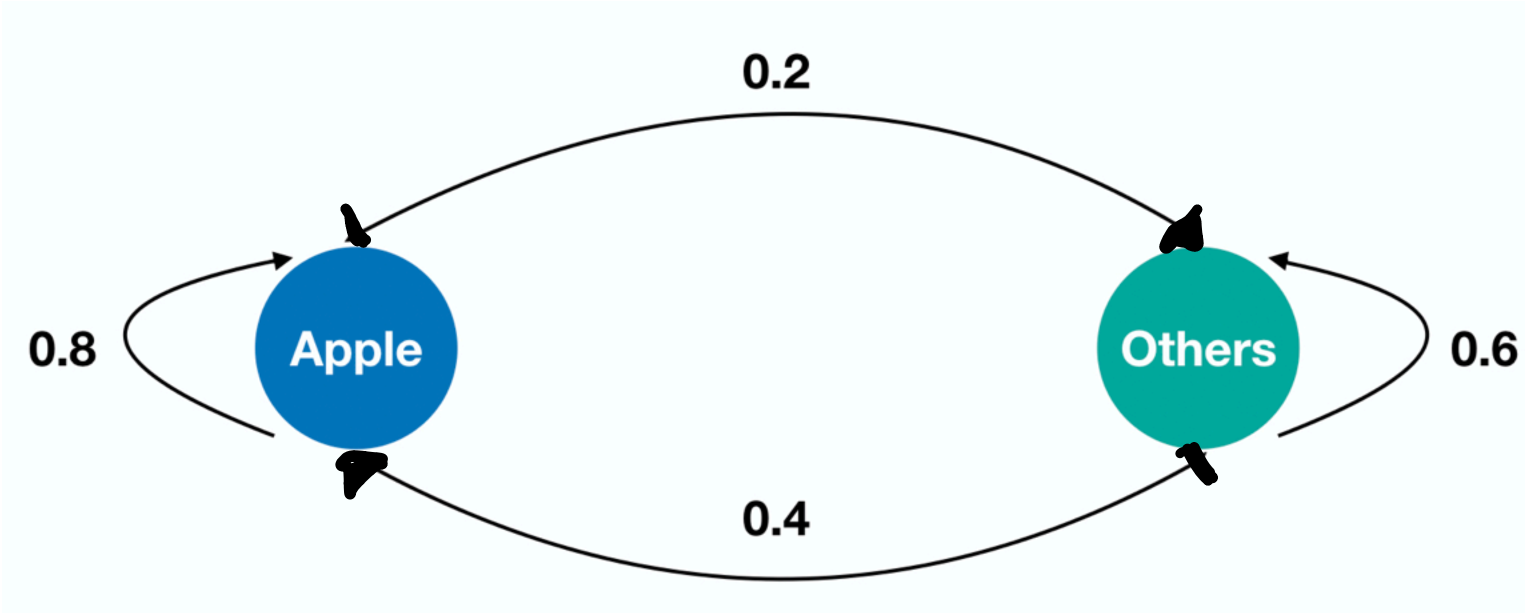
In particular, when we have the initial distribution $\pi^{(0)}$ over states, then the unconditional distribution $\pi^{(1)}$ over the next state is given by:

$$\mathbb{P}[S_1 = i] = \sum_{k \in \mathcal{S}} \mathbb{P}[X_1 = i|X_0 = k]\mathbb{P}[X_0 = k]$$

That is, $\pi^{(1)} = \pi^{(0)}T$.

# Example: Smart Phone Market Model

Consider a simple model of the dynamics of the smart phone market, where we model the customer loyalty as follows:




The transition matrix for the Markov chain is:

$$T = \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix}$$

Say that the market is initially $\pi^{(0)} = [0.7 \;\; 0.3]$, i.e. 70% Apple. What is the market distribution in the long term?

# Example: Smart Phone Market Model

In [127]:

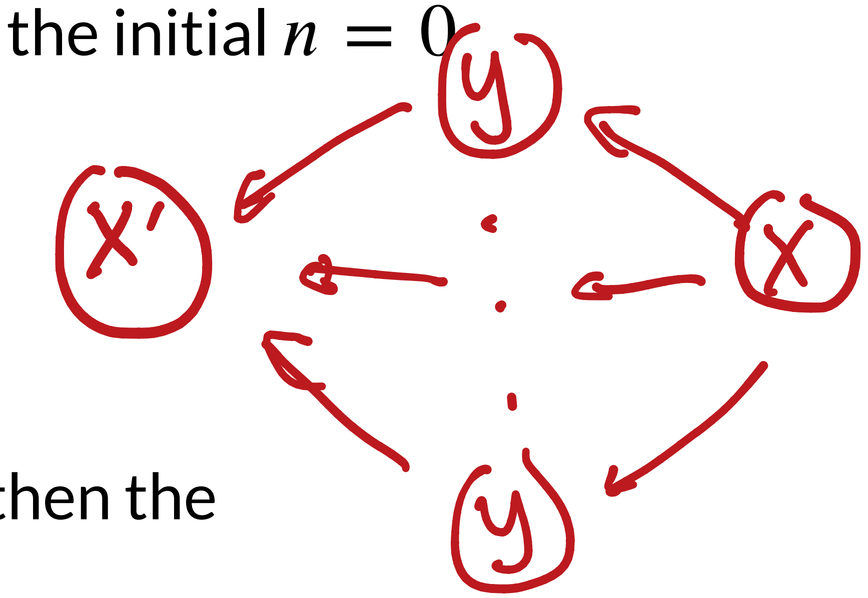
```
#transition matrix
T = np.array([[0.8, 0.2],
              [0.4, 0.6]])
#initial distribution
pi_0 = np.array([0.7, 0.3]).reshape((1, -1))
#time
N = 500

pi_0.dot(np.linalg.matrix_power(T, N))
#try different values of N and different inital distributions!
```

Out[127]:

```
array([[0.66666667, 0.33333333]])
```

## Chapman-Kolmogorov Equations: Continuous State Space

If the state space is continuous, the likelihood of the $n = 2$ state, given the initial $n = 0$ state, can be computed by the ***Chapman-Kolmogorov equation***:



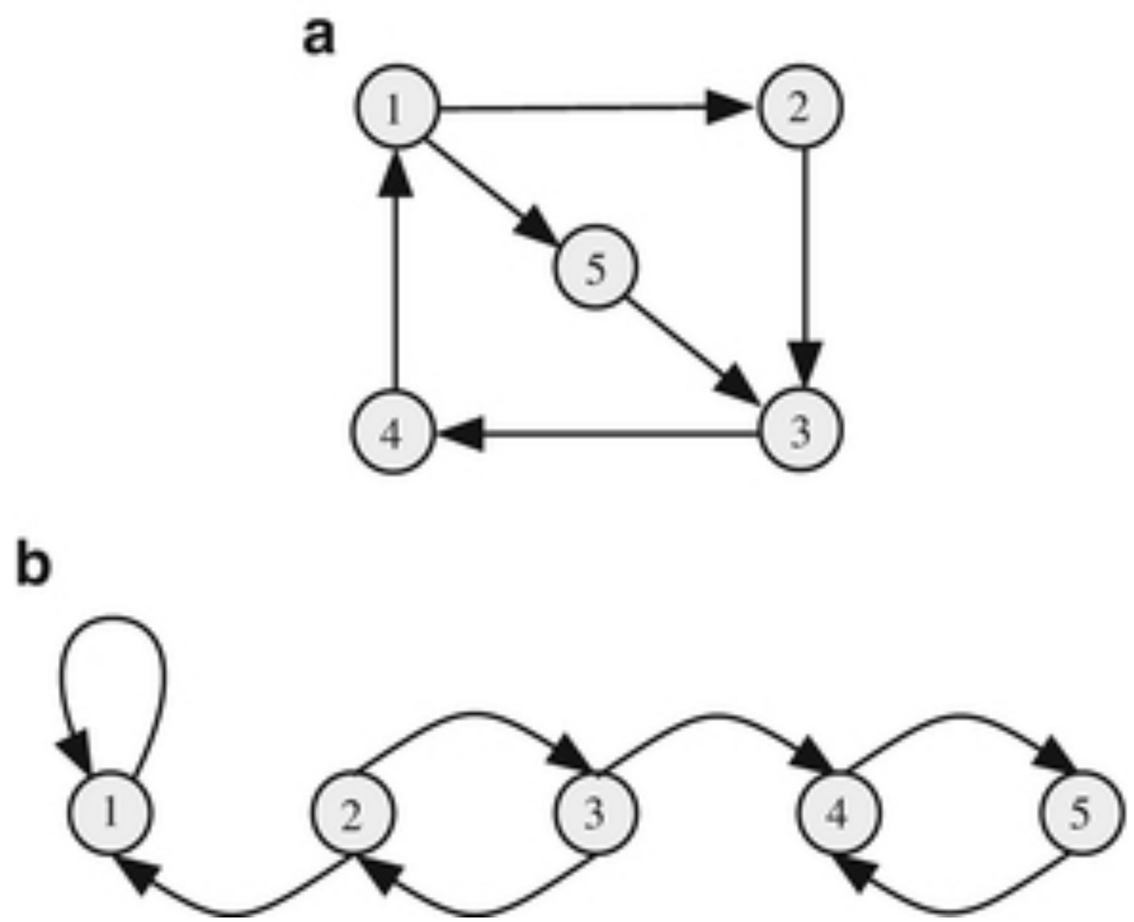

$$T^{(2)}(x, x') = \int_S T(x, y)\, T(y, x') dy.$$

In particular, when we have the initial distribution $\pi^{(0)}(x)$ over states, then the unconditional distribution $\pi^{(1)}(x)$ over the next state is given by:

$$\pi^{(1)}(x) = \int_S T(y, x)\, \pi^{(0)}(y) dy.$$

## Properties of Markov Chains: Irreducibility

A Markov chain is called ***irreducible*** if every state can be reached from every other state in finite time.

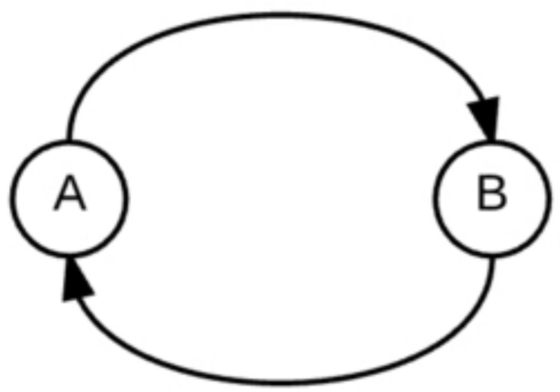


# Properties of Markov Chains: Aperiodicity

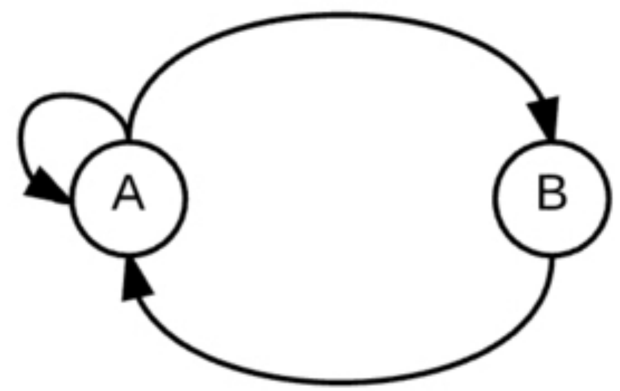
A state $s \in \mathcal{S}$ is has period $t$ if one can only return to $s$ in multiples of $t$ steps.

A Markov chain is called ***aperiodic*** if the period of each state is 1.

Period = 2




Period = 1



## Properties of Markov Chains: Stationary Distributions

A distribution $\pi$ over the finite state space $S$ is a ***stationary distribution*** of the Markov Chain with transition matrix $T$ if

$$\pi = \pi T,$$

i.e. performing the transition matrix doesn't change the distribution.

The equivalent condition for continuous state space $S$ is:

$$\pi(x) = \int_S T(y, x)\, \pi(y) dy.$$

# Properties of Markov Chains: Limiting Distributions

We are often interested in what happens to a distribution after many transitions,

$$\pi^{(n)} = \pi^{(0)} T^{(n)}, \quad \text{or} \quad \pi^{(n)}(x) = \int_S T^{(n)}(y, x)\, \pi^{(0)}(y) dy$$

If $\pi^{(\infty)} = \lim_{n \to \infty} \pi^{(n)}$ exists (with some caveats in the continuous state case), we call it the ***limiting distribution*** of the Markov chain.

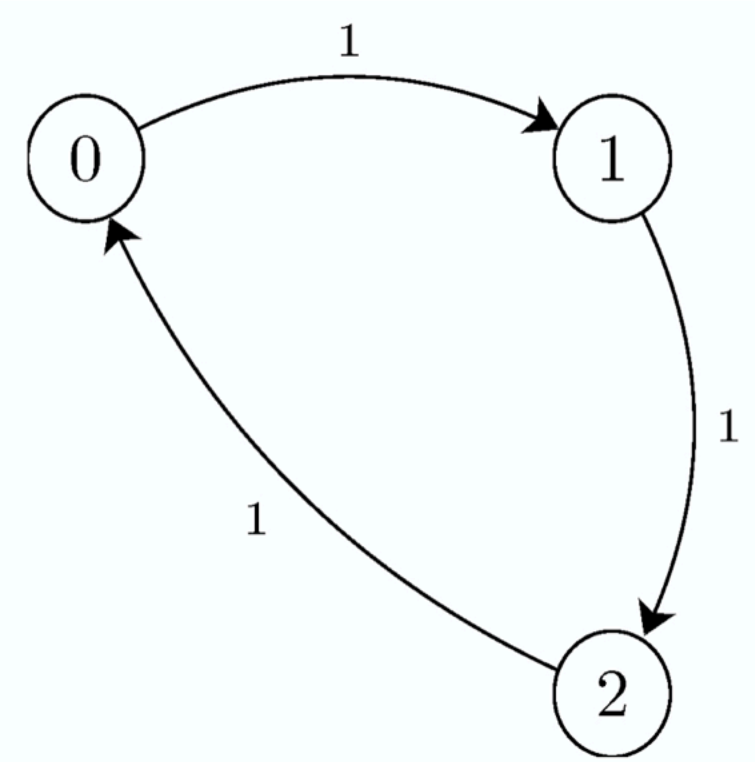
# Fundamental Theorm of Markov Chains

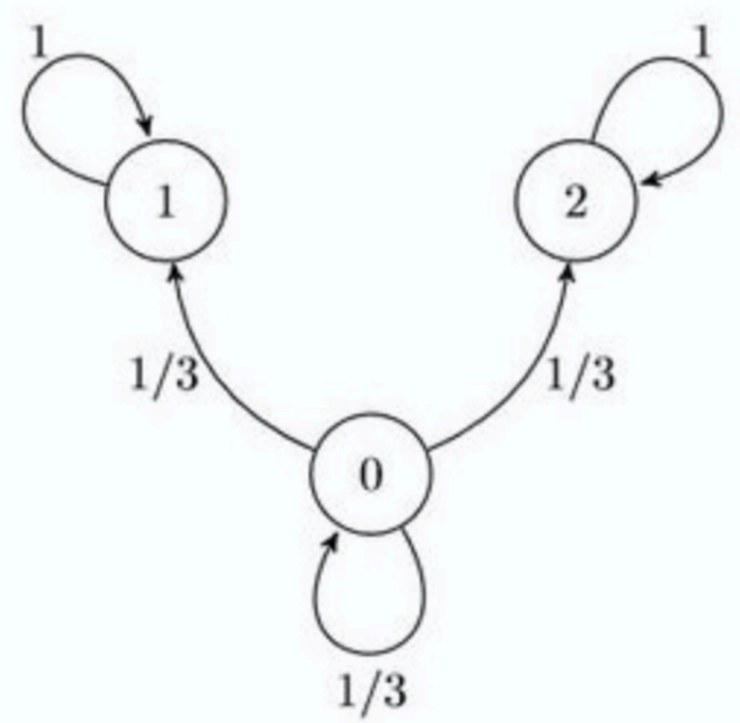
Now we are ready to relate all these properties of Markov chains in a single theorem:

**Fundamental Theorem of Markov Chains:** if a Markov chain is irreducible and aperiodic, then it has a *unique* stationary distribution $\pi$ and $\pi^{\infty} = \lim_{n \to \infty} \pi^{(n)} = \pi$.

In practice, the theorem says you can start with any initial distribution over the state space $\mathcal{S}$, asymptotically, you will always obtain the distribution $\pi$.

While we unfortunately can't prove the theorem, we can indicate why both conditions are necessary.



# Properties of Markov Chains: Reversibility

A Markov chain is called ***reversible*** with respect to a distribution $\pi$ over a finite state space $\mathcal{S}$ if the following holds:

$$\pi T = T \pi^\top$$

The above translates to $\pi_i T_{i,j} = \pi_j T_{j,i}$.

For a continuous state space, the condition is:

$$\pi(x) T(x, y) = T(y, x) \pi(y).$$

The condition for reversibility is often called the ***detailed balance*** condition.

$$\pi = \pi T$$

$$\pi(x) = \int_y T(y,x)\pi(y)dy$$

## Reversibility and Stationary Distributions

Using reversibility, we have another way to characterize a stationary distribution.

**Theorem:** If a Markov chain, with transition matrix or kernel pdf $T$, is reversible with respect to $\pi$. Then $\pi$ is a stationary distribution of the chain.

***Proof:*** We will give the proof for the case of a continuous state space $S$. Supoose that $\pi(x)T(x,y) = T(y,x)\pi(y)$, then

$$\int_S \pi(x)T(x,y)dx = \int_S \pi(y)T(y,x)dx = \pi(y)\int_S T(y,x)dx = \pi(y)\cdot 1 = \pi(x).$$

# Markov Chain Monte Carlo

# Markov Chain Monte Carlo Samplers

Every sampler for a distribution $p(\theta)$ over the domain $\Theta$ defines a stochastic process $\{X_0, X_1, \ldots, \}$, where the state space is $\Theta$.

If the sampler defines a Markov chain whose unique stationary and limiting distribution is $p$, we call it a ***Markov Chain Monte Carlo (MCMC)*** sampler.

That is, for every MCMC sampler, we have that

1. **Stationary:** $pT = p$
2. **Limiting:** $\lim_{n\to\infty} \pi^{(n)} = p$, for any $\pi^{(0)}$

where $T$ is the transition matrix or kernel pdf defined by the sampler.

## What Do We Need to Prove to get $pT = p$ and $\lim_{n\to\infty} \pi^{(n)} = p$?

1. Prove that the sampler is ***irreducible*** and ***aperiodic***. Then, there is a unique stationary distribution $\pi$ such that
$$\pi T = \pi.$$
2. Prove that the sampler is ***reversible*** or ***detailed balanced*** with respect to $p$. Then,
$$\pi = p.$$

# Gibbs as MCMC

We've seen an example where the Gibbs sampler for a discrete target distribution defines a MCMC sampler.

But what about Gibbs samplers for a continuous target distribution $\mathsf{p}$? Certainly, the samples $X_n$ obtained by the sampler defines a Markov Chain: the distribution over the next sample depends only on the current sample.

But, in order to be a MCMC sampler, we need to prove that $\mathsf{p}$ is the stationary and limiting distribution of the sampler?