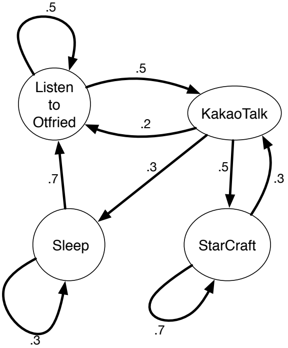
Introduction to MCMC

Eric Vigoda
Georgia Tech
KAIST (Spring '15)

Guest lecture for
KAIST CS 500
Graduate Algorithms
Friday, March 6, 2015

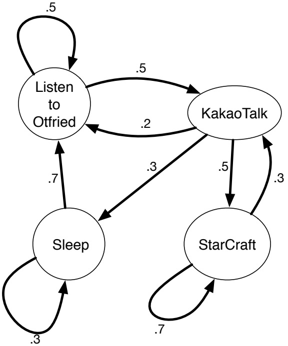
What is a Markov chain?

Example: Life in CS 500, discrete time $t = 0, 1, 2, \ldots$:



What is a Markov chain?

Example: Life in CS 500, discrete time $t = 0, 1, 2, \ldots$:



Each vertex is a state of the Markov chain.

Directed graph, possibly with self-loops.

Edge weights represent probability of a transition, so:
non-negative and sum of weights of outgoing edges $= 1$.

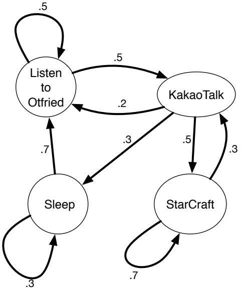
Transition matrix

In general: N states $\Omega = \{1, 2, \ldots, N\}$.

$N \times N$ transition matrix P where:

$\mathrm{P}(i,j) =$ weight of edge $i \to j = \mathbf{Pr}$ (going from i to j)

For earlier example:



$$\mathrm{P} = \begin{bmatrix} .5 & .5 & 0 & 0 \\ .2 & 0 & .5 & .3 \\ 0 & .3 & .7 & 0 \\ .7 & 0 & 0 & .3 \end{bmatrix}$$

P is a stochastic matrix = rows sum to 1.

One-step transitions

Time: $t = 0, 1, 2, \ldots$.
Let X_t denote the state at time t.
X_t is a random variable.

One-step transitions

Time: $t = 0, 1, 2, \ldots$.
Let X_t denote the state at time t.
X_t is a random variable.
For states k and j, $\mathbf{Pr}\left(X_1 = j \mid X_0 = k\right) = \mathrm{P}(k, j)$.

One-step transitions

Time: $t = 0, 1, 2, \ldots$.
Let X_t denote the state at time t.
X_t is a random variable.
For states k and j, $\mathbf{Pr}\left(X_1 = j \mid X_0 = k\right) = \mathrm{P}(k, j)$.
In general, for $t \geq 1$, given:

in state k_0 at time 0, in k_1 at time 1, ..., in k_{t-1} at time $t-1$, what's the probability of being in state j at time t?

One-step transitions

Time: $t = 0, 1, 2, \ldots$.
Let X_t denote the state at time t.
X_t is a random variable.
For states k and j, $\mathbf{Pr}\left(X_1 = j \mid X_0 = k\right) = \mathrm{P}(k, j)$.
In general, for $t \geq 1$, given:

in state k_0 at time 0, in k_1 at time 1, ..., in k_{t-1} at time $t - 1$,
what's the probability of being in state j at time t?

$$\begin{aligned}
\mathbf{Pr}\left(X_t = j \mid X_0 = k_0, X_1 = k_1, \ldots, X_{t-1} = k_{t-1}\right) \\
&= \mathbf{Pr}\left(X_t = j \mid X_{t-1} = k_{t-1}\right) \\
&= \mathrm{P}(k_{t-1}, j).
\end{aligned}$$

One-step transitions

Time: $t = 0, 1, 2, \ldots$.
Let X_t denote the state at time t.
X_t is a random variable.
For states k and j, $\mathbf{Pr}\left(X_1 = j \mid X_0 = k\right) = \mathrm{P}(k, j)$.
In general, for $t \geq 1$, given:

in state k_0 at time 0, in k_1 at time 1, ..., in k_{t-1} at time $t-1$, what's the probability of being in state j at time t?

$$\begin{aligned}\mathbf{Pr}\left(X_t = j \mid X_0 = k_0, X_1 = k_1, \ldots, X_{t-1} = k_{t-1}\right) \\ &= \mathbf{Pr}\left(X_t = j \mid X_{t-1} = k_{t-1}\right) \\ &= \mathrm{P}(k_{t-1}, j).\end{aligned}$$

Process is **memoryless** –

only current state matters, previous states do not matter.

Known as **Markov property**, hence the term Markov chain.

2-step transitions

What's probability *Listen* at time 2 given *Kakao* at time 0?
Try all possibilities for state at time 1.

2-step transitions

What's probability *Listen* at time 2 given *Kakao* at time 0?
Try all possibilities for state at time 1.

$$\begin{aligned}
&\mathbf{Pr}\left(X_2 = \mathit{Listen} \mid X_0 = \mathit{Kakao}\right) \\
&= \quad \mathbf{Pr}\left(X_2 = \mathit{Listen} \mid X_1 = \mathit{Listen}\right) \times \mathbf{Pr}\left(X_1 = \mathit{Listen} \mid X_0 = \mathit{Kakao}\right) \\
&\quad + \mathbf{Pr}\left(X_2 = \mathit{Listen} \mid X_1 = \mathit{Kakao}\right) \times \mathbf{Pr}\left(X_1 = \mathit{Kakao} \mid X_0 = \mathit{Kakao}\right) \\
&\quad + \mathbf{Pr}\left(X_2 = \mathit{Listen} \mid X_1 = \mathit{StarCraft}\right) \times \mathbf{Pr}\left(X_1 = \mathit{StarCraft} \mid X_0 = \mathit{Kakao}\right) \\
&\quad + \mathbf{Pr}\left(X_2 = \mathit{Listen} \mid X_1 = \mathit{Sleep}\right) \times \mathbf{Pr}\left(X_1 = \mathit{Sleep} \mid X_0 = \mathit{Kakao}\right)
\end{aligned}$$

2-step transitions

What's probability *Listen* at time 2 given *Kakao* at time 0?
Try all possibilities for state at time 1.

$$
\begin{aligned}
&\mathbf{Pr}\left(X_2 = \mathit{Listen} \mid X_0 = \mathit{Kakao}\right) \\
= \quad & \mathbf{Pr}\left(X_2 = \mathit{Listen} \mid X_1 = \mathit{Listen}\right) \times \mathbf{Pr}\left(X_1 = \mathit{Listen} \mid X_0 = \mathit{Kakao}\right) \\
& + \mathbf{Pr}\left(X_2 = \mathit{Listen} \mid X_1 = \mathit{Kakao}\right) \times \mathbf{Pr}\left(X_1 = \mathit{Kakao} \mid X_0 = \mathit{Kakao}\right) \\
& + \mathbf{Pr}\left(X_2 = \mathit{Listen} \mid X_1 = \mathit{StarCraft}\right) \times \mathbf{Pr}\left(X_1 = \mathit{StarCraft} \mid X_0 = \mathit{Kakao}\right) \\
& + \mathbf{Pr}\left(X_2 = \mathit{Listen} \mid X_1 = \mathit{Sleep}\right) \times \mathbf{Pr}\left(X_1 = \mathit{Sleep} \mid X_0 = \mathit{Kakao}\right) \\
= \quad & (.5)(.2) + 0 + 0 + (.7)(.3) = .31
\end{aligned}
$$

$$
\mathrm{P} = \begin{bmatrix} .5 & .5 & 0 & 0 \\ .2 & 0 & .5 & .3 \\ 0 & .3 & .7 & 0 \\ .7 & 0 & 0 & .3 \end{bmatrix}
\qquad
\mathrm{P}^2 = \begin{bmatrix} .35 & .25 & .25 & .15 \\ .31 & .25 & .35 & .09 \\ .06 & .21 & .64 & .09 \\ .56 & .35 & 0 & .09 \end{bmatrix}
$$

States: 1=*Listen*, 2=*Kakao*, 3=*StarCraft*, 4=*Sleep*.

k-step transitions

2-step transition probabilities: use P^2.
In general, for states i and j:

$$
\begin{aligned}
&\mathbf{Pr}\left(X_{t+2} = j \mid X_t = i\right) \\
&= \sum_{k=1}^{N} \mathbf{Pr}\left(X_{t+2} = j \mid X_{t+1} = k\right) \times \mathbf{Pr}\left(X_{t+1} = k \mid X_t = i\right) \\
&= \sum_k \mathrm{P}(k,j)\mathrm{P}(i,k) = \sum_k \mathrm{P}(i,k)\mathrm{P}(k,j) = \mathrm{P}^2(i,j)
\end{aligned}
$$

k-step transitions

2-step transition probabilities: use P^2.
In general, for states i and j:

$$\begin{aligned}
&\mathbf{Pr}\left(X_{t+2} = j \mid X_t = i\right) \\
&= \sum_{k=1}^{N} \mathbf{Pr}\left(X_{t+2} = j \mid X_{t+1} = k\right) \times \mathbf{Pr}\left(X_{t+1} = k \mid X_t = i\right) \\
&= \sum_{k} \mathrm{P}(k,j)\mathrm{P}(i,k) = \sum_{k} \mathrm{P}(i,k)\mathrm{P}(k,j) = \mathrm{P}^2(i,j)
\end{aligned}$$

ℓ-step transition probabilities: use P^ℓ.
For states i and j and integer $\ell \geq 1$,

$$\mathbf{Pr}\left(X_{t+\ell} = j \mid X_t = i\right) = \mathrm{P}^\ell(i,j),$$

Random Initial State

Suppose the state at time 0 is not fixed
but is chosen from a probability distribution μ_0.

Notation: $X_0 \sim \mu_0$.

What is the distribution for X_1?

Random Initial State

Suppose the state at time 0 is not fixed
but is chosen from a probability distribution μ_0.

Notation: $X_0 \sim \mu_0$.

What is the distribution for X_1?
For state j,

$$\begin{aligned}
\mathbf{Pr}\left(X_1 = j\right) &= \sum_{i=1}^{N} \mathbf{Pr}\left(X_0 = i\right) \times \mathbf{Pr}\left(X_1 = j \mid X_0 = i\right) \\
&= \sum_{i} \mu_0(i)\mathrm{P}(i,j) = (\mu_0 \mathrm{P})(j)
\end{aligned}$$

So $X_1 \sim \mu_1$ where $\mu_1 = \mu_0 \mathrm{P}$.

Random Initial State

Suppose the state at time 0 is not fixed
but is chosen from a probability distribution μ_0.

Notation: $X_0 \sim \mu_0$.

What is the distribution for X_1?
For state j,

$$\begin{aligned}\mathbf{Pr}\left(X_1 = j\right) &= \sum_{i=1}^{N} \mathbf{Pr}\left(X_0 = i\right) \times \mathbf{Pr}\left(X_1 = j \mid X_0 = i\right) \\ &= \sum_{i} \mu_0(i)\mathrm{P}(i,j) = (\mu_0 \mathrm{P})(j)\end{aligned}$$

So $X_1 \sim \mu_1$ where $\mu_1 = \mu_0 \mathrm{P}$.

And $X_t \sim \mu_t$ where $\mu_t = \mu_0 \mathrm{P}^t$.

Back to CS 500 example: big t?

Let's look again at our CS 500 example:

$$P = \begin{bmatrix} .5 & .5 & 0 & 0 \\ .2 & 0 & .5 & .3 \\ 0 & .3 & .7 & 0 \\ .7 & 0 & 0 & .3 \end{bmatrix}$$

Back to CS 500 example: big t?

Let's look again at our CS 500 example:

$$P = \begin{bmatrix} .5 & .5 & 0 & 0 \\ .2 & 0 & .5 & .3 \\ 0 & .3 & .7 & 0 \\ .7 & 0 & 0 & .3 \end{bmatrix} \qquad P^2 = \begin{bmatrix} .35 & .25 & .25 & .15 \\ .31 & .25 & .35 & .09 \\ .06 & .21 & .64 & .09 \\ .56 & .35 & 0 & .09 \end{bmatrix}$$

Back to CS 500 example: big t?

Let's look again at our CS 500 example:

$$P = \begin{bmatrix} .5 & .5 & 0 & 0 \\ .2 & 0 & .5 & .3 \\ 0 & .3 & .7 & 0 \\ .7 & 0 & 0 & .3 \end{bmatrix} \qquad P^2 = \begin{bmatrix} .35 & .25 & .25 & .15 \\ .31 & .25 & .35 & .09 \\ .06 & .21 & .64 & .09 \\ .56 & .35 & 0 & .09 \end{bmatrix}$$

$$P^{10} = \begin{bmatrix} .247770 & .244781 & .402267 & .105181 \\ .245167 & .244349 & .405688 & .104796 \\ .239532 & .243413 & .413093 & .103963 \\ .251635 & .245423 & .397189 & .105754 \end{bmatrix}$$

Back to CS 500 example: big t?

Let's look again at our CS 500 example:

$$P = \begin{bmatrix} .5 & .5 & 0 & 0 \\ .2 & 0 & .5 & .3 \\ 0 & .3 & .7 & 0 \\ .7 & 0 & 0 & .3 \end{bmatrix} \qquad P^2 = \begin{bmatrix} .35 & .25 & .25 & .15 \\ .31 & .25 & .35 & .09 \\ .06 & .21 & .64 & .09 \\ .56 & .35 & 0 & .09 \end{bmatrix}$$

$$P^{10} = \begin{bmatrix} .247770 & .244781 & .402267 & .105181 \\ .245167 & .244349 & .405688 & .104796 \\ .239532 & .243413 & .413093 & .103963 \\ .251635 & .245423 & .397189 & .105754 \end{bmatrix}$$

$$P^{20} = \begin{bmatrix} .244190 & .244187 & .406971 & .104652 \\ .244187 & .244186 & .406975 & .104651 \\ .244181 & .244185 & .406984 & .104650 \\ .244195 & .244188 & .406966 & .104652 \end{bmatrix}$$

Back to CS 500 example: big t?

Let's look again at our CS 500 example:

$$P = \begin{bmatrix} .5 & .5 & 0 & 0 \\ .2 & 0 & .5 & .3 \\ 0 & .3 & .7 & 0 \\ .7 & 0 & 0 & .3 \end{bmatrix} \qquad P^2 = \begin{bmatrix} .35 & .25 & .25 & .15 \\ .31 & .25 & .35 & .09 \\ .06 & .21 & .64 & .09 \\ .56 & .35 & 0 & .09 \end{bmatrix}$$

$$P^{10} = \begin{bmatrix} .247770 & .244781 & .402267 & .105181 \\ .245167 & .244349 & .405688 & .104796 \\ .239532 & .243413 & .413093 & .103963 \\ .251635 & .245423 & .397189 & .105754 \end{bmatrix}$$

$$P^{20} = \begin{bmatrix} .244190 & .244187 & .406971 & .104652 \\ .244187 & .244186 & .406975 & .104651 \\ .244181 & .244185 & .406984 & .104650 \\ .244195 & .244188 & .406966 & .104652 \end{bmatrix}$$

Columns are converging to

$$\pi = [\ .244186, \ .244186, \ .406977, \ .104651].$$

Limiting Distribution

For big t,

$$P^t \approx \begin{bmatrix} .244186 & .244186 & .406977 & .104651 \\ .244186 & .244186 & .406977 & .104651 \\ .244186 & .244186 & .406977 & .104651 \\ .244186 & .244186 & .406977 & .104651 \end{bmatrix}$$

Limiting Distribution

For big t,

$$\mathrm{P}^t \approx \begin{bmatrix} .244186 & .244186 & .406977 & .104651 \\ .244186 & .244186 & .406977 & .104651 \\ .244186 & .244186 & .406977 & .104651 \\ .244186 & .244186 & .406977 & .104651 \end{bmatrix}$$

Regardless of where it starts X_0, for big t:

$$\begin{aligned} \mathbf{Pr}\left(X_t = 1\right) &= .244186 \\ \mathbf{Pr}\left(X_t = 2\right) &= .244186 \\ \mathbf{Pr}\left(X_t = 3\right) &= .406977 \\ \mathbf{Pr}\left(X_t = 4\right) &= .104651 \end{aligned}$$

Limiting Distribution

For big t,

$$P^t \approx \begin{bmatrix} .244186 & .244186 & .406977 & .104651 \\ .244186 & .244186 & .406977 & .104651 \\ .244186 & .244186 & .406977 & .104651 \\ .244186 & .244186 & .406977 & .104651 \end{bmatrix}$$

Regardless of where it starts X_0, for big t:

$$\begin{aligned} \mathbf{Pr}(X_t = 1) &= .244186 \\ \mathbf{Pr}(X_t = 2) &= .244186 \\ \mathbf{Pr}(X_t = 3) &= .406977 \\ \mathbf{Pr}(X_t = 4) &= .104651 \end{aligned}$$

Let $\pi = [\ .244186,\ \ .244186,\ \ .406977,\ \ .104651]$.

In other words, for big t, $X_t \sim \pi$.

π is called a *stationary distribution*.

Let $\pi = [\ .244186,\ \ .244186,\ \ .406977,\ \ .104651]$.

π is called a *stationary distribution*.

Let $\pi = [\ .244186,\ \ .244186,\ \ .406977,\ \ .104651]$.
π is called a *stationary distribution*.

Once we reach π we stay in π: if $X_t \sim \pi$ then $X_{t+1} \sim \pi$,
in other words, $\pi\mathrm{P} = \pi$.

Let $\pi = [\ .244186,\ \ .244186,\ \ .406977,\ \ .104651]$.
π is called a *stationary distribution*.

Once we reach π we stay in π: if $X_t \sim \pi$ then $X_{t+1} \sim \pi$,
in other words, $\pi \mathrm{P} = \pi$.

Any distribution π where $\pi \mathrm{P} = \pi$ is called a stationary distribution of the Markov chain.

Stationary Distributions

Key questions:

- When is there a stationary distribution?
- If there is at least one, is it unique or more than one?
- Assuming there's a unique stationary distribution:
  - Do we always reach it?
  - What is it?
  - *Mixing time* = Time to reach unique stationary distribution

Algorithmic Goal:

- If we have a distribution π that we want to sample from, can we design a Markov chain that has:
  - Unique stationary distribution π,
  - From every X_0 we always reach π,
  - Fast mixing time.

Irreducibility

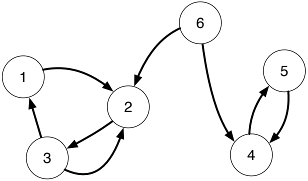
Want a unique stationary distribution π and that
get to it from every starting state X_0.

Irreducibility

Want a unique stationary distribution π and that

get to it from every starting state X_0.

But if multiple strongly connected components (SCCs) then can't go from one to the other:



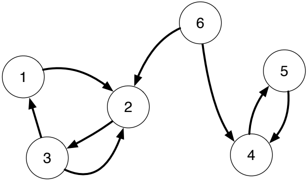
Starting at 1 gets to different distribution than starting at 5.

Irreducibility

Want a unique stationary distribution π and that

get to it from every starting state X_0.

But if multiple strongly connected components (SCCs) then can't go from one to the other:



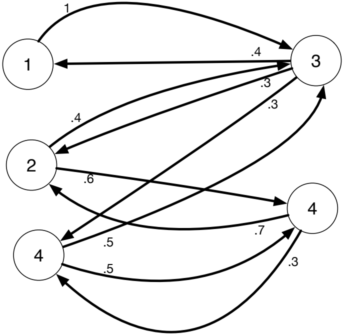
Starting at 1 gets to different distribution than starting at 5.

State i communicates with state j if starting at i can reach j:

there exists t, $\mathrm{P}^t(i,j) > 0$.

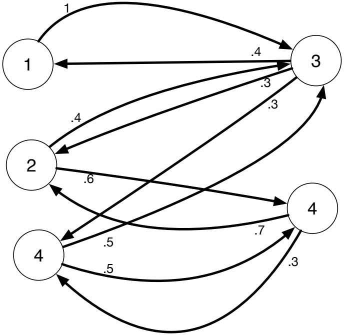
Markov chain is irreducible if all pairs of states communicate..

Example of bipartite Markov chain:



Starting at 1 gets to different distribution than starting at 3.

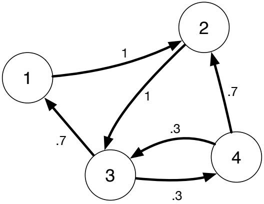
Example of bipartite Markov chain:



Starting at 1 gets to different distribution than starting at 3.

Need that no periodicity.

Aperiodic

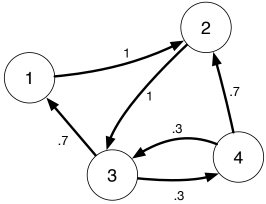


Return times for state i are times $R_i = \{t : \mathrm{P}^t(i,i) > 0\}$.

Above example: $R_1 = \{3, 5, 6, 8, 9, \dots\}$.

Let $r = \gcd(R_i)$ be the period for state i.

Aperiodic



Return times for state i are times $R_i = \{t : \mathrm{P}^t(i,i) > 0\}$.

Above example: $R_1 = \{3, 5, 6, 8, 9, \dots\}$.

Let $r = \gcd(R_i)$ be the period for state i.

If P is irreducible then all states have the same period.
If $r = 2$ then the Markov chain is bipartite.
A Markov chain is aperiodic if $r = 1$.

Ergodic: Unique Stationary Distribution

Ergodic = Irreducible and aperiodic.

Ergodic: Unique Stationary Distribution

Ergodic = Irreducible and aperiodic.

Fundamental Theorem for Markov Chains:
Ergodic Markov chain has a unique stationary distribution π.
And for all initial $X_0 \sim \mu_0$ then:

$$\lim_{t \to \infty} \mu_t = \pi.$$

In other words, for big enough t, all rows of P^t are π.

Ergodic: Unique Stationary Distribution

Ergodic = Irreducible and aperiodic.

Fundamental Theorem for Markov Chains:
Ergodic Markov chain has a unique stationary distribution π.
And for all initial $X_0 \sim \mu_0$ then:

$$\lim_{t \to \infty} \mu_t = \pi.$$

In other words, for big enough t, all rows of P^t are π.

How big does t need to be?

What is π?

Proof idea: Ergodic MC has Unique Stationary Distribution

What is a π?

Proof idea: Ergodic MC has Unique Stationary Distribution

What is a π?

Fix a state i and set $X_0 = i$.
Let T be the first time we visit state i again.
T is a random variable.
For every state j,
let $\rho(j) =$ expected number of visits to j up to time T.
(Note, $\rho(i) = 1$.)

Let $\pi(j) = \rho(j)/Z$ where $Z = \sum_k \rho(k)$.
Can verify that this π is a stationary distribution.

Proof idea: Ergodic MC has Unique Stationary Distribution

What is a π?

Fix a state i and set $X_0 = i$.
Let T be the first time we visit state i again.
T is a random variable.
For every state j,
let $\rho(j) =$ expected number of visits to j up to time T.
(Note, $\rho(i) = 1$.)

Let $\pi(j) = \rho(j)/Z$ where $Z = \sum_k \rho(k)$.
Can verify that this π is a stationary distribution.

Why is it unique and we always reach it?

Proof idea: Ergodic MC has Unique Stationary Distribution

What is a π?

Fix a state i and set $X_0 = i$.
Let T be the first time we visit state i again.
T is a random variable.
For every state j,
let $\rho(j) =$ expected number of visits to j up to time T.
(Note, $\rho(i) = 1$.)

Let $\pi(j) = \rho(j)/Z$ where $Z = \sum_k \rho(k)$.
Can verify that this π is a stationary distribution.

Why is it unique and we always reach it?
Make 2 chains (X_t) and (Y_t):
X_0 is arbitrary, and
Y_0 is chosen from π so that $Y_t \sim \pi$ for all t.
Using irreducibility, can "couple" the transitions of these chains:
for big t we have $X_t = Y_t$ and thus $X_t \sim \pi$.

Determining π: Symmetric Markov Chain

Symmetric if for all pairs i, j: $\mathrm{P}(i,j) = \mathrm{P}(j,i)$.
Then π is uniformly distributed over all of the states $\{1, \ldots, N\}$:

$$\pi(j) = \frac{1}{N} \text{ for all states } j.$$

Determining π: Symmetric Markov Chain

Symmetric if for all pairs i, j: $\mathrm{P}(i,j) = \mathrm{P}(j,i)$.
Then π is uniformly distributed over all of the states $\{1, \ldots, N\}$:

$$\pi(j) = \frac{1}{N} \text{ for all states } j.$$

Proof: We'll verify that $\pi\mathrm{P} = \pi$ for this π.
Need to check that for all states j: $(\pi\mathrm{P})(j) = \pi(j)$.

Determining π: Symmetric Markov Chain

Symmetric if for all pairs i, j: $P(i, j) = P(j, i)$.
Then π is uniformly distributed over all of the states $\{1, \ldots, N\}$:

$$\pi(j) = \frac{1}{N} \text{ for all states } j.$$

Proof: We'll verify that $\pi P = \pi$ for this π.
Need to check that for all states j: $(\pi P)(j) = \pi(j)$.

$$(\pi P)(j) = \sum_{i=1}^{N} \pi(i) P(i, j)$$

Determining π: Symmetric Markov Chain

Symmetric if for all pairs i, j: $\mathrm{P}(i,j) = \mathrm{P}(j,i)$.
Then π is uniformly distributed over all of the states $\{1, \ldots, N\}$:

$$\pi(j) = \frac{1}{N} \text{ for all states } j.$$

Proof: We'll verify that $\pi\mathrm{P} = \pi$ for this π.
Need to check that for all states j: $(\pi\mathrm{P})(j) = \pi(j)$.

$$\begin{aligned}(\pi\mathrm{P})(j) &= \sum_{i=1}^{N} \pi(i)\mathrm{P}(i,j) \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathrm{P}(i,j)\end{aligned}$$

Determining π: Symmetric Markov Chain

Symmetric if for all pairs i, j: $\mathrm{P}(i,j) = \mathrm{P}(j,i)$.
Then π is uniformly distributed over all of the states $\{1, \ldots, N\}$:

$$\pi(j) = \frac{1}{N} \text{ for all states } j.$$

Proof: We'll verify that $\pi\mathrm{P} = \pi$ for this π.
Need to check that for all states j: $(\pi\mathrm{P})(j) = \pi(j)$.

$$\begin{aligned}
(\pi\mathrm{P})(j) &= \sum_{i=1}^{N} \pi(i)\mathrm{P}(i,j) \\
&= \frac{1}{N} \sum_{i=1}^{N} \mathrm{P}(i,j) \\
&= \frac{1}{N} \sum_{i=1}^{N} \mathrm{P}(j,i) \text{ since P is symmetric}
\end{aligned}$$

Determining π: Symmetric Markov Chain

Symmetric if for all pairs i, j: $P(i,j) = P(j,i)$.
Then π is uniformly distributed over all of the states $\{1, \ldots, N\}$:

$$\pi(j) = \frac{1}{N} \text{ for all states } j.$$

Proof: We'll verify that $\pi P = \pi$ for this π.
Need to check that for all states j: $(\pi P)(j) = \pi(j)$.

$$\begin{aligned}
(\pi P)(j) &= \sum_{i=1}^{N} \pi(i) P(i,j) \\
&= \frac{1}{N} \sum_{i=1}^{N} P(i,j) \\
&= \frac{1}{N} \sum_{i=1}^{N} P(j,i) \quad \text{since P is symmetric} \\
&= \frac{1}{N} \quad \text{since rows of P always sum to 1}
\end{aligned}$$

Determining π: Symmetric Markov Chain

Symmetric if for all pairs i, j: $P(i,j) = P(j,i)$.
Then π is uniformly distributed over all of the states $\{1, \ldots, N\}$:

$$\pi(j) = \frac{1}{N} \text{ for all states } j.$$

Proof: We'll verify that $\pi P = \pi$ for this π.
Need to check that for all states j: $(\pi P)(j) = \pi(j)$.

$$\begin{aligned}
(\pi P)(j) &= \sum_{i=1}^{N} \pi(i) P(i,j) \\
&= \frac{1}{N} \sum_{i=1}^{N} P(i,j) \\
&= \frac{1}{N} \sum_{i=1}^{N} P(j,i) \quad \text{since P is symmetric} \\
&= \frac{1}{N} \quad \text{since rows of P always sum to 1} \\
&= \pi(j) \qquad \square
\end{aligned}$$

Determining π: Reversible Markov Chain

Reversible with respect to π if for all pairs i, j:

$$\pi(i)\mathrm{P}(i,j) = \pi(j)\mathrm{P}(j,i).$$

If can find such a π then it is the stationary distribution.

Determining π: Reversible Markov Chain

Reversible with respect to π if for all pairs i, j:

$$\pi(i)\mathrm{P}(i,j) = \pi(j)\mathrm{P}(j,i).$$

If can find such a π then it is the stationary distribution.

Proof: Similar to the symmetric case.
Need to check that for all states j: $(\pi \mathrm{P})(j) = \pi(j)$.

Determining π: Reversible Markov Chain

Reversible with respect to π if for all pairs i, j:

$$\pi(i)\mathrm{P}(i,j) = \pi(j)\mathrm{P}(j,i).$$

If can find such a π then it is the stationary distribution.

Proof: Similar to the symmetric case.
Need to check that for all states j: $(\pi\mathrm{P})(j) = \pi(j)$.

$$(\pi\mathrm{P})(j) = \sum_{i=1}^{N} \pi(i)\mathrm{P}(i,j)$$

Determining π: Reversible Markov Chain

Reversible with respect to π if for all pairs i, j:

$$\pi(i)\mathrm{P}(i,j) = \pi(j)\mathrm{P}(j,i).$$

If can find such a π then it is the stationary distribution.

Proof: Similar to the symmetric case.
Need to check that for all states j: $(\pi\mathrm{P})(j) = \pi(j)$.

$$\begin{aligned}(\pi\mathrm{P})(j) &= \sum_{i=1}^{N} \pi(i)\mathrm{P}(i,j) \\ &= \sum_{i=1}^{N} \pi(j)\mathrm{P}(j,i) \quad \text{since P is reversible}\end{aligned}$$

Determining π: Reversible Markov Chain

Reversible with respect to π if for all pairs i, j:

$$\pi(i)\mathrm{P}(i,j) = \pi(j)\mathrm{P}(j,i).$$

If can find such a π then it is the stationary distribution.

Proof: Similar to the symmetric case.
Need to check that for all states j: $(\pi\mathrm{P})(j) = \pi(j)$.

$$\begin{aligned}
(\pi\mathrm{P})(j) &= \sum_{i=1}^{N} \pi(i)\mathrm{P}(i,j) \\
&= \sum_{i=1}^{N} \pi(j)\mathrm{P}(j,i) \quad \text{since P is reversible} \\
&= \pi(j)\sum_{i=1}^{N} \mathrm{P}(j,i)
\end{aligned}$$

Determining π: Reversible Markov Chain

Reversible with respect to π if for all pairs i, j:

$$\pi(i)\mathrm{P}(i,j) = \pi(j)\mathrm{P}(j,i).$$

If can find such a π then it is the stationary distribution.

Proof: Similar to the symmetric case.
Need to check that for all states j: $(\pi\mathrm{P})(j) = \pi(j)$.

$$\begin{aligned}(\pi\mathrm{P})(j) &= \sum_{i=1}^{N}\pi(i)\mathrm{P}(i,j) \\ &= \sum_{i=1}^{N}\pi(j)\mathrm{P}(j,i) \quad \text{since P is reversible} \\ &= \pi(j)\sum_{i=1}^{N}\mathrm{P}(j,i) \\ &= \pi(j)\end{aligned}$$

□

Some Examples

Random walk on a d-regular, connected undirected graph G:
- What is π?

Some Examples

Random walk on a d-regular, connected undirected graph G:

- What is π?
  - Symmetric: for edge (i,j), $\mathrm{P}(i,j) = \mathrm{P}(j,i) = 1/d$.
  - So π is uniform: $\pi(i) = 1/n$.

Some Examples

Random walk on a d-regular, connected undirected graph G:

- What is π?
  - Symmetric: for edge (i,j), $\mathrm{P}(i,j) = \mathrm{P}(j,i) = 1/d$.
  - So π is uniform: $\pi(i) = 1/n$.

Random walk on a general connected undirected graph G:

- What is π?

Some Examples

Random walk on a d-regular, connected undirected graph G:

- What is π?
  - Symmetric: for edge (i,j), $\mathrm{P}(i,j) = \mathrm{P}(j,i) = 1/d$.
  - So π is uniform: $\pi(i) = 1/n$.

Random walk on a general connected undirected graph G:

- What is π?
  - Consider $\pi(i) = d(i)/Z$ where
    - $d(i) =$ degree of vertex i and
    - $Z = \sum_{j \in V} d(j)$. (Note, $Z = 2m = 2|E|$.)
  - Check it's reversible: $\pi(i)\mathrm{P}(i,j) = \frac{d(i)}{Z}\frac{1}{d(i)} = \frac{1}{Z} = \pi(j)\mathrm{P}(j,i)$.

Some Examples

Random walk on a d-regular, connected undirected graph G:

What is π?

Symmetric: for edge (i,j), $\mathrm{P}(i,j) = \mathrm{P}(j,i) = 1/d$.

So π is uniform: $\pi(i) = 1/n$.

Random walk on a general connected undirected graph G:

What is π?

Consider $\pi(i) = d(i)/Z$ where

$d(i) =$ degree of vertex i and

$Z = \sum_{j \in V} d(j)$. (Note, $Z = 2m = 2|E|$.)

Check it's reversible: $\pi(i)\mathrm{P}(i,j) = \frac{d(i)}{Z}\frac{1}{d(i)} = \frac{1}{Z} = \pi(j)\mathrm{P}(j,i)$.

What if G is a directed graph?

Random walk on a d-regular, connected undirected graph G:

What is π?

Symmetric: for edge (i,j), $P(i,j) = P(j,i) = 1/d$.

So π is uniform: $\pi(i) = 1/n$.

Random walk on a general connected undirected graph G:

What is π?

Consider $\pi(i) = d(i)/Z$ where

$d(i) =$ degree of vertex i and

$Z = \sum_{j \in V} d(j)$. (Note, $Z = 2m = 2|E|$.)

Check it's reversible: $\pi(i)P(i,j) = \frac{d(i)}{Z}\frac{1}{d(i)} = \frac{1}{Z} = \pi(j)P(j,i)$.

What if G is a directed graph?

Then it may not be reversible, and if it's not reversible:

then usually we can't figure out the stationary distribution since typically N is HUGE.

PageRank

PageRank is an algorithm devised by Brin and Page 1998:
determine the “importance” of webpages.

PageRank is an algorithm devised by Brin and Page 1998:
determine the “importance” of webpages.

Webgraph:
V = webpages
E = directed edges for hyperlinks

Let $\pi(x) =$ “rank” of page x.
We are trying to define $\pi(x)$ in a sensible way.

PageRank is an algorithm devised by Brin and Page 1998:
determine the “importance” of webpages.

Webgraph:
V = webpages
E = directed edges for hyperlinks

Notation:
For page $x \in V$, let:

$$\begin{aligned} \text{Out}(x) &= \{y : x \to y \in E\} = \text{outgoing edges from } x \\ \text{In}(x) &= \{w : w \to x \in E\} = \text{incoming edges to } x \end{aligned}$$

Let $\pi(x) =$ “rank” of page x.
We are trying to define $\pi(x)$ in a sensible way.

First Ranking Idea

First idea for ranking pages: like academic papers

use citation counts

Here, citation = link to a page.

So set $\pi(x) = |\mathrm{In}(x)| =$ number of links to x.

Refining the Ranking Idea

What if:

Georgia Tech's webpage has 500 links, one is to Eric's page.
KAIST's webpage has only 5 links, one is to Otfried's page.

Which link is more valuable?

Refining the Ranking Idea

What if:

Georgia Tech's webpage has 500 links, one is to Eric's page.
KAIST's webpage has only 5 links, one is to Otfried's page.

Which link is more valuable?

Academic papers: If a paper cites 50 other papers, then each reference gets $1/50$ of a citation.

Refining the Ranking Idea

What if:

Georgia Tech's webpage has 500 links, one is to Eric's page.
KAIST's webpage has only 5 links, one is to Otfried's page.

Which link is more valuable?

Academic papers: If a paper cites 50 other papers, then each reference gets $1/50$ of a citation.

Webpages: If a page y has $|\mathrm{Out}(y)|$ outgoing links, then:
each linked page gets $1/|\mathrm{Out}(y)|$.

New solution:

$$\pi(x) = \sum_{y \in \mathrm{In}(x)} \frac{1}{|\mathrm{Out}(y)|}.$$

Further Refining the Ranking Idea

Previous:

$$\pi(x) = \sum_{y \in \mathrm{In}(x)} \frac{1}{|\mathrm{Out}(y)|}.$$

But if *CNN* has a link to a page that's more important than if *KAIST CS* has a link to it.

Further Refining the Ranking Idea

Previous:

$$\pi(x) = \sum_{y \in \mathrm{In}(x)} \frac{1}{|\mathrm{Out}(y)|}.$$

But if *CNN* has a link to a page that's more important than if *KAIST CS* has a link to it.

Solution: define $\pi(x)$ recursively.
Page y has importance $\pi(y)$.
A link from y gets $\pi(y)/|\mathrm{Out}(y)|$ of a citation.

$$\pi(x) = \sum_{y \in \mathrm{In}(x)} \frac{\pi(y)}{|\mathrm{Out}(y)|}.$$

Random Walk

Importance of page x:

$$\pi(x) = \sum_{y \in \mathrm{In}(x)} \frac{\pi(y)}{|\mathrm{Out}(y)|}.$$

Random Walk

Importance of page x:

$$\pi(x) = \sum_{y \in \mathrm{In}(x)} \frac{\pi(y)}{|\mathrm{Out}(y)|}.$$

Recursive definition of π, how do we find it?

Importance of page x:

$$\pi(x) = \sum_{y \in \mathrm{In}(x)} \frac{\pi(y)}{|\mathrm{Out}(y)|}.$$

Recursive definition of π, how do we find it?

Look at the random walk on the webgraph $G = (V, E)$.
From a page $y \in V$, choose a random link and follow it.
This is a Markov chain.
For $y \to x \in E$ then:

$$\mathrm{P}(y, x) = \frac{1}{|\mathrm{Out}(y)|}$$

What is the stationary distribution of this Markov chain?

Random Walk

Random walk on the webgraph $G = (V, E)$.
For $y \to x \in E$ then:

$$P(y, x) = \frac{1}{|\text{Out}(y)|}$$

What is the stationary distribution of this Markov chain?

Random walk on the webgraph $G = (V, E)$.
For $y \to x \in E$ then:

$$\mathrm{P}(y, x) = \frac{1}{|\mathrm{Out}(y)|}$$

What is the stationary distribution of this Markov chain?
Need to find π where $\pi = \pi\mathrm{P}$.
Thus,

$$\pi(x) = \sum_{y \in V} \pi(y)\mathrm{P}(y, x) = \sum_{y \in \mathrm{In}(x)} \frac{\pi(y)}{|\mathrm{Out}(y)|}.$$

This is identical to the definition of the importance vector π.

Summary: the stationary distribution of the random walk on the webgraph gives the importance $\pi(x)$ of a page x.

Random Walk on the Webgraph

Random walk on the webgraph $G = (V, E)$.

Is π the only stationary distribution?
In other words, is the Markov chain ergodic?

Random Walk on the Webgraph

Random walk on the webgraph $G = (V, E)$.

Is π the only stationary distribution?
In other words, is the Markov chain ergodic?

Need that G is strongly connected – it probably is not.
And some pages have no outgoing links...

then hit the "random" button!

Random Walk on the Webgraph

Random walk on the webgraph $G = (V, E)$.

Is π the only stationary distribution?
In other words, is the Markov chain ergodic?

Need that G is strongly connected – it probably is not.
And some pages have no outgoing links...
then hit the "random" button!

Solution to make it ergodic:
Introduce "damping factor" α where $0 < \alpha \leq 1$.
(in practice apparently use $\alpha \approx .85$)

From page y,
with prob. α follow a random outgoing link from page y.
with prob. $1 - \alpha$ go to a completely random page
(uniformly chosen from all pages V).

Random Surfer

Let $N = |V|$ denote number of webpages.
Transition matrix of new Random Surfer chain:

$$\mathrm{P}(y,x) = \begin{cases} \frac{1-\alpha}{N} & \text{if } y \to x \notin E \\ \frac{1-\alpha}{N} + \frac{\alpha}{|\mathrm{Out}(y)|} & \text{if } y \to x \in E \end{cases}$$

This new Random Surfer Markov chain is ergodic.
Thus, unique stationary distribution is the desired π.

Random Surfer

Let $N = |V|$ denote number of webpages.
Transition matrix of new Random Surfer chain:

$$\mathrm{P}(y,x) = \begin{cases} \frac{1-\alpha}{N} & \text{if } y \to x \notin E \\ \frac{1-\alpha}{N} + \frac{\alpha}{|\mathrm{Out}(y)|} & \text{if } y \to x \in E \end{cases}$$

This new Random Surfer Markov chain is ergodic.
Thus, unique stationary distribution is the desired π.

How to find π?

Take last week's π, and compute $\pi \mathrm{P}^t$ for big t.

What's a big enough t?

Mixing Time

How fast does an ergodic MC reach its unique stationary π?

Mixing Time

How fast does an ergodic MC reach its unique stationary π?

Need to measure distance from π, use total variation distance.
For distributions μ and ν on set Ω:

$$d_{TV}(\mu, \nu) = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|.$$

How fast does an ergodic MC reach its unique stationary π?

Need to measure distance from π, use total variation distance.

For distributions μ and ν on set Ω:

$$d_{TV}(\mu, \nu) = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|.$$

Example: $\Omega = \{1, 2, 3, 4\}$.
μ is uniform: $\mu(1) = \mu(2) = \mu(3) = \mu(4) = .25$.
And ν has: $\nu(1) = .5, \nu(2) = .1, \nu(3) = .15, \nu(4) = .25$.

$$d_{TV}(\mu, \nu) = \frac{1}{2}(.25 + .15 + .1 + 0) = .25$$

Mixing Time

Consider ergodic MC with states Ω, transition matrix P, and unique stationary distribution π.
For state $x \in \Omega$, time to mix from x:

$$T(x) = \min\{t : \mathrm{d_{TV}}(\mathrm{P}^t(x,\cdot),\pi) \leq 1/4.$$

Mixing Time

Consider ergodic MC with states Ω, transition matrix P, and unique stationary distribution π.
For state $x \in \Omega$, time to mix from x:

$$T(x) = \min\{t : \mathrm{d_{TV}}(\mathrm{P}^t(x, \cdot), \pi) \leq 1/4.$$

Then, mixing time $\mathrm{T_{mix}} = \max_x T(x)$.
Summarizing in words:

mixing time is time to get within distance $\leq 1/4$ of π from the worst initial state X_0.

Mixing Time

Consider ergodic MC with states Ω, transition matrix P, and unique stationary distribution π.
For state $x \in \Omega$, time to mix from x:

$$T(x) = \min\{t : \mathrm{d}_{\mathrm{TV}}(\mathrm{P}^t(x, \cdot), \pi) \leq 1/4.$$

Then, mixing time $\mathrm{T}_{\mathrm{mix}} = \max_x T(x)$.
Summarizing in words:

mixing time is time to get within distance $\leq 1/4$ of π from the worst initial state X_0.

Choice of constant $1/4$ is somewhat arbitrary.
Can get within distance $\leq \epsilon$ in time $O(\mathrm{T}_{\mathrm{mix}} \log(1/\epsilon))$.

Mixing Time of Random Surfer

Coupling proof:
Consider 2 copies of the Random Surfer chain (X_t) and (Y_t).

Choose Y_0 from π. Thus, $Y_t \sim \pi$ for all t.
And X_0 is arbitrary.

Mixing Time of Random Surfer

Coupling proof:
Consider 2 copies of the Random Surfer chain (X_t) and (Y_t).

Choose Y_0 from π. Thus, $Y_t \sim \pi$ for all t.
And X_0 is arbitrary.

If $X_{t-1} = Y_{t-1}$ then they choose the same transition at time t.
If $X_{t-1} \neq Y_{t-1}$ then with prob. $1 - \alpha$ choose the same random page z for both chains.

Therefore,

$$\mathbf{Pr}\left(X_t \neq Y_t\right) \leq \alpha^t.$$

Coupling proof:
Consider 2 copies of the Random Surfer chain (X_t) and (Y_t).

Choose Y_0 from π. Thus, $Y_t \sim \pi$ for all t.
And X_0 is arbitrary.

If $X_{t-1} = Y_{t-1}$ then they choose the same transition at time t.
If $X_{t-1} \neq Y_{t-1}$ then with prob. $1 - \alpha$ choose the same random page z for both chains.

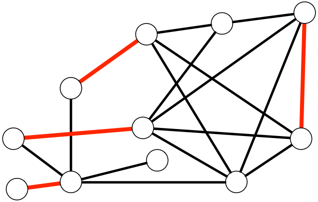
Therefore,

$$\mathbf{Pr}\left(X_t \neq Y_t\right) \leq \alpha^t.$$

Setting: $t \geq -2/\log(\alpha)$ we have $\mathbf{Pr}\left(X_t \neq Y_t\right) \leq 1/4$.
Therefore, mixing time:

$$\mathrm{T}_{\mathrm{mix}} \leq \frac{-2}{\log \alpha} \approx 8.5 \text{ for } \alpha = .85.$$

Undirected graph:

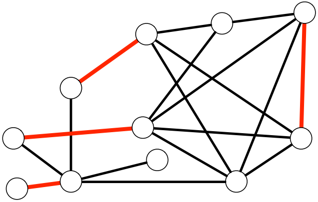


Matching = subset of vertex disjoint edges.

Let $\Omega =$ collection of all matchings of G (of all sizes).

Example Chain: Random Matching

Undirected graph:



Matching = subset of vertex disjoint edges.

Let $\Omega =$ collection of all matchings of G (of all sizes).

Can we generate a matching uniformly at random from Ω?
in time polynomial in $n = |V|$?

Markov Chain for Matchings

Consider an undirected graph $G = (V, E)$.

From a matching X_t the transition $X_t \to X_{t+1}$
is defined as follows:

1. Choose an edge $e = (v, w)$ uniformly at random from E.
2. If $e \in X_t$ then set $X_{t+1} = X_t \setminus \{e\}$.
3. If v and w are unmatched in X_t then set $X_{t+1} = X_t \bigcup \{e\}$.
4. Otherwise, set $X_{t+1} = X_t$.

Symmetric and ergodic and thus π is uniform over Ω.

Markov Chain for Matchings

Consider an undirected graph $G = (V, E)$.

From a matching X_t the transition $X_t \to X_{t+1}$
is defined as follows:

1. Choose an edge $e = (v, w)$ uniformly at random from E.
2. If $e \in X_t$ then set $X_{t+1} = X_t \setminus \{e\}$.
3. If v and w are unmatched in X_t then set $X_{t+1} = X_t \bigcup \{e\}$.
4. Otherwise, set $X_{t+1} = X_t$.

Symmetric and ergodic and thus π is uniform over Ω.

How fast does it reach π?

Next class: we'll see that it's close after poly(n) steps for every G.