Lecture 23:
Markov Chains

November 17th, 2015
My typical day (when I was a student)

9:00am

Work

\[ f(x) = \sum_{S \subseteq [n]} \hat{f}(S) \chi_S(x) \]
My typical day (when I was a student)

9:01am

Work

\[ f(x) = \sum_{S \subseteq [n]} \hat{f}(S) \chi_S(x) \]

40%

Surf

60%
My typical day (when I was a student)

9:02am

\[ f(x) = \sum_{S \subseteq [n]} \hat{f}(S) \chi_S(x) \]

- Work: 40%
- Surf: 60%
- Email: 30%

Diagram:
- Email: 30% to Work
- Surf: 60% to Work
- Surf: 60% to Email
My typical day (when I was a student)

\[ f(x) = \sum_{S \subseteq [n]} \hat{f}(S) \chi_S(x) \]

9:03am

- Email: 50%
- Surf: 50%

Work: 40%
- Surf: 60%
- Email: 10%

Email: 50%
- 30% to Surf

Surf: 60%
- 60% to Work
My typical day (when I was a student)

9:00am

Email

Surf

Work

\[ f(x) = \sum_{S} (S)\chi_S(x) \]

40%

50%

60%

30%

50%
My typical day (when I was a student)

9:01am

Email: 50%

Work:

\[ f(x) = \sum_{S \subseteq [n]} \hat{f}(S) \chi_S(x) \]

Surf: 60%

Email: 30%

Email: 50%

Surf: 60%

Email: 50%

Email: 10%

Email: 40%
My typical day (when I was a student)

9:02am

\[ f(x) = \sum_{S \subseteq [n]} \hat{f}(S) \chi_S(x) \]

- Work: 40%
- Email: 50%
- Surf: 60%

Email: 50%

Surf: 60%
My typical day (when I was a student)

9:03am

Work

\[ f(x) = \sum_{S \subseteq [n]} \hat{f}(S) \chi_S(x) \]

40%

Email

50%

50%

Surf

10%

60%

30%

60%

30%
My typical day (when I was a student)

9:04am

$$f(x) = \sum_{S \subseteq [n]} \hat{f}(S) \chi_S(x)$$

Work

40%

Email

50%

50%

Surf

60%

30%

60%
My typical day (when I was a student)

9:05am

- **Work**: 40%
- **Email**: 50%
- **Surf**: 50%

Math formula:

\[ f(x) = \bigoplus_{S \in S} \chi_S(x) \]

Graph:

- From 9:05am to Work: 40%
- From Work to Email: 50%
- From Work to Surf: 60%
- From Email to Work: 30%
- From Email to Email: 50%
- From Surf to Work: 60%
- From Surf to Surf: 60%
And now

Prepare 15-251 slides

100%
Markov Model
Markov Model

Andrey Markov (1856 - 1922)
Russian mathematician.
Famous for his work on random processes.

Pr[X ≥ c · E[X]] ≤ 1/c is Markov’s Inequality.)
Markov Model

Andrey Markov (1856 - 1922)
Russian mathematician.
Famous for his work on random processes.

\[ \Pr[X \geq c \cdot E[X]] \leq 1/c \] is Markov’s Inequality.

A model for the evolution of a random system.

The future is independent of the past, given the present.
Cool things about the Markov model

- It is a very general and natural model.
  Extraordinary number of applications in many different disciplines:
  computer science, mathematics, biology, physics, chemistry, economics, psychology, music, baseball,...

- The model is simple and neat.

- A beautiful mathematical theory behind it.
  Starts simple, goes deep.
The plan

Motivating examples and applications

Basic mathematical representation and properties

Applications
The future is independent of the past, given the present.
Some Examples of Markov Models
Example: Drunkard Walk

Salvador Dali (1922)
The Drunkard

Home
Example: Diffusion Process
Example: Weather

A very(!!) simplified model for the weather.

Probabilities on a daily basis:

\[ \text{Pr[\text{sunny to rainy}]} = 0.1 \]
\[ \text{Pr[\text{sunny to sunny}]} = 0.9 \]
\[ \text{Pr[\text{rainy to rainy}]} = 0.5 \]
\[ \text{Pr[\text{rainy to sunny}]} = 0.5 \]

Encode more information about current state for a more accurate model.
Example: Life Insurance

Goal of insurance company:

figure out how much to charge the clients.

Find a model for how long a client will live.

Probabilistic model of health on a monthly basis:

Pr[healthy to sick] = 0.3
Pr[sick to healthy] = 0.8
Pr[sick to death] = 0.1
Pr[healthy to death] = 0.01
Pr[healthy to healthy] = 0.69
Pr[sick to sick] = 0.1
Pr[death to death] = 1
Example: Life Insurance

Goal of insurance company:
figure out how much to charge the clients.

Find a model for how long a client will live.

Probabilistic model of health on a monthly basis:
Some Applications of Markov Models
Application: Algorithmic Music Composition

Nicholas Vasillo

Megalithic Copier #2: Markov Chains (2011)

written in Pure Data
Application: Image Segmentation
Random text generated by a computer (putting random words together):

“While at a conference a few weeks back, I spent an interesting evening with a grain of salt.”

Google: Mark V Shaney
Speech recognition software programs use Markov models to listen to the sound of your voice and convert it into text.
1997: Web search was horrible

Sorts webpages by number of occurrences of keyword(s).
Application: Google PageRank

Founders of Google

Larry Page  Sergey Brin

$20 Billionaires
Application: Google PageRank

Jon Kleinberg

Nevanlinna Prize
How does Google order the webpages displayed after a search?

2 important factors:

- Relevance of the page.

- Reputation of the page.

  The number and reputation of links pointing to that page.

Reputation is measured using PageRank.

PageRank is calculated using a Markov Chain.
The answer to life the universe and everything = 42
The plan

Motivating examples and applications

Basic mathematical representation and properties

Applications
The Setting

There is a system with $n$ possible states/values.

At each time step, the state changes probabilistically.
The Setting

There is a system with \( n \) possible states/values.

At each time step, the state changes probabilistically.

Memoryless
The next state only depends on the current state.

Evolution of the system: random walk on the graph.
There is a system with $n$ possible states/values.

At each time step, the state changes probabilistically.

**Memoryless**

The next state only depends on the current state.

**Evolution of the system:** random walk on the graph.
A **Markov Chain** is a directed graph with $V = \{1, 2, \ldots, n\}$ such that:

- Each edge is labeled with a value in $(0, 1]$ (a positive probability).
- At each vertex, the probabilities on outgoing edges sum to 1.

(- We usually assume the graph is strongly connected. i.e. there is a path from $i$ to $j$ for any $i$ and $j$.)

The vertices of the graph are called **states**.

The edges are called **transitions**.

The label of an edge is a **transition probability**.
Example: Markov Chain for a Lecture

This is not strongly connected.
Given some Markov Chain with $n$ states:

For each $t = 0, 1, 2, 3, \ldots$ we have a random variable:

$$X_t = \text{the state we are in after } t \text{ steps.}$$

Define $\pi_t[i] = \Pr[X_t = i]$.  

$$\pi_t = [p_1 \ p_2 \ \cdots \ p_n]$$

$$\sum_{i} p_i = 1$$

We write $X_t \sim \pi_t$.  \ $(X_t \text{ has distribution } \pi_t)$

Note that someone has to provide $\pi_0$.  \ Once this is known, we get the distributions $\pi_1, \pi_2, \ldots$
Let’s say we start at state $1$, i.e., $X_0 \sim \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} = \pi_0$

$X_0 = 1 \quad X_0 \sim \pi_0$
Let’s say we start at state 1, i.e., \( X_0 \sim [1 \ 2 \ 3 \ 4] = \pi_0 \)

\[ X_0 = 1 \quad X_0 \sim \pi_0 \]
\[ X_1 = 4 \quad X_1 \sim \pi_1 \]
Let's say we start at state 1, i.e., $X_0 \sim \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix} = \pi_0$

- $X_0 = 1$
- $X_1 = 4$
- $X_2 = 3$

$X_0 \sim \pi_0$  
$X_1 \sim \pi_1$  
$X_2 \sim \pi_2$
Let's say we start at state 1, i.e., $X_0 \sim [1 \ 0 \ 0 \ 0] = \pi_0$

$X_0 = 1 \quad X_1 = 4 \quad X_2 = 3 \quad X_3 = 4$

$X_0 \sim \pi_0 \quad X_1 \sim \pi_1 \quad X_2 \sim \pi_2 \quad X_3 \sim \pi_3$
Let's say we start at state 1, i.e., \( X_0 \sim \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} = \pi_0 \)
Let’s say we start at state 1, i.e., \( X_0 \sim \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \pi_0 \)

\[
\begin{align*}
X_0 &= 1 & X_0 &\sim \pi_0 \\
X_1 &= 4 & X_1 &\sim \pi_1 \\
X_2 &= 3 & X_2 &\sim \pi_2 \\
X_3 &= 4 & X_3 &\sim \pi_3 \\
X_4 &= 2 & X_4 &\sim \pi_4 \\
X_5 &= 3 & X_5 &\sim \pi_5 
\end{align*}
\]
Let’s say we start at state 1, i.e., $X_0 \sim [1 \ 0 \ 0 \ 0] = \pi_0$
Let’s say we start at state 1, i.e., \( X_0 \sim [1 \ 0 \ 0 \ 0] = \pi_0 \)

\[
\Pr[1 \rightarrow 2 \text{ in one step}] = \Pr[X_1 = 2 \mid X_0 = 1] = \Pr[X_t = 2 \mid X_{t-1} = 1]
\]
Let's say we start at state 1, i.e., $X_0 \sim [1 \ 0 \ 0 \ 0] = \pi_0$.

### Notation

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1/4</td>
<td>3/4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>1/2</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pr[$X_1 = 2 | X_0 = 1$] = $\frac{1}{2}$

Pr[$X_1 = 3 | X_0 = 1$] = 0

Pr[$X_1 = 4 | X_0 = 1$] = $\frac{1}{2}$

Pr[$X_1 = 1 | X_0 = 1$] = 0

$\forall t$ Pr[$X_t = 2 | X_{t-1} = 4$] = $\frac{1}{4}$

$\forall t$ Pr[$X_t = 3 | X_{t-1} = 2$] = 1
A Markov Chain with \( n \) states can be characterized by the \( n \times n \) transition matrix \( K \):

\[
\forall i, j \in \{1, 2, \ldots, n\} \quad K[i, j] = \Pr[X_t = j \mid X_{t-1} = i] = \Pr[i \rightarrow j \text{ in one step}]
\]

Note: rows of \( K \) sum to 1.
Some Fundamental and Natural Questions

What is the probability of being in state \( i \) after \( t \) steps (given some initial state)?

\[
\pi_t[i] = ?
\]

What is the expected time of reaching state \( i \) when starting at state \( j \)?

What is the expected time of having visited every state (given some initial state)?

How do you answer such questions?
Mathematical representation of the evolution

Suppose we start at state 1 and let the system evolve.

How can we mathematically represent the evolution?

\[ \pi_0 = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 0 & 0 & 0 \end{bmatrix} \]

What is \( \pi_1 \)? By inspection, \( \pi_1 = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \).
Given $\pi_1 = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$, what is $\pi_2$?
Mathematical representation of the evolution

\[ \pi_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix} \]

What is \( \pi_1 \)?

\[ \pi_1[j] = \Pr[X_1 = j] \]

(law of total probability)

\[
= \sum_{i=1}^{4} \Pr[X_1 = j \mid X_0 = i] \Pr[X_0 = i]
\]

This is true for any \( j \).

\[
= \sum_{i=1}^{4} K[i, j] \cdot \pi_0[i] = (\pi_0 \cdot K)[j]
\]
Mathematical representation of the evolution

The probability of states after 1 step:

\[
\pi_0 \begin{bmatrix}
0 & \frac{1}{2} & 0 & \frac{1}{2} \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & \frac{1}{4} & \frac{3}{4} & 0 \\
\end{bmatrix} = \begin{bmatrix}
0 & \frac{1}{2} & 0 & \frac{1}{2} \\
\end{bmatrix}
\]

the new state (probabilistic)
Mathematical representation of the evolution

The probability of states after 2 steps:

\[
\pi_1 = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}, \quad \pi_2 = \begin{bmatrix} 0 & \frac{1}{8} & \frac{7}{8} & 0 \end{bmatrix}
\]

\[K = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 2 & 0 & 0 & 1 & 0 \\ 3 & 0 & 0 & 0 & 1 \\ 4 & 0 & \frac{1}{4} & \frac{3}{4} & 0 \end{bmatrix}\]

the new state (probabilistic)
Mathematical representation of the evolution

\[ \pi_1 = \pi_0 \cdot K \]

\[ \pi_2 = \pi_1 \cdot K \]

So

\[ \pi_2 = (\pi_0 \cdot K) \cdot K \]

\[ = \pi_0 \cdot K^2 \]
Mathematical representation of the evolution

In general:

If the initial probabilistic state is \( \begin{bmatrix} p_1 & p_2 & \cdots & p_n \end{bmatrix} = \pi_0 \)

\( p_i \) = probability of being in state \( i \),

\( p_1 + p_2 + \cdots + p_n = 1 \),

after \( t \) steps, the probabilistic state is:

\[
\begin{bmatrix} p_1 & p_2 & \cdots & p_n \end{bmatrix} \begin{bmatrix} \text{Transition Matrix} \end{bmatrix}^t = \pi_t
\]
What happens in the long run?

i.e., can we say anything about $\pi_t$ for large $t$?

Suppose the Markov chain is “aperiodic”.

Then, as the system evolves, the probabilistic state $\text{converges}$ to a limiting probabilistic state.

As $t \to \infty$, for any $\pi_0 = [p_1 \ p_2 \ \cdots \ p_n]$:

$$
[p_1 \ p_2 \ \cdots \ p_n]^{\text{Transition Matrix}}_t \rightarrow \pi
$$
Remarkable Property of Markov Chains

In other words:

\[ \pi_t \to \pi \quad \text{as} \quad t \to \infty. \]

Note:

\[
\begin{bmatrix}
\pi \\
\text{Transition Matrix}
\end{bmatrix}
= \pi
\]

stationary/invariant distribution

This \( \pi \) is unique.
Remarkable Property of Markov Chains

Stationary distribution is \([ \begin{bmatrix} \frac{5}{6} & \frac{1}{6} \end{bmatrix} \].

\[
\begin{bmatrix} \frac{5}{6} & \frac{1}{6} \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix} = \begin{bmatrix} \frac{5}{6} & \frac{1}{6} \end{bmatrix}
\]

*In the long run, it is sunny \( \frac{5}{6} \) of the time, it is rainy \( \frac{1}{6} \) of the time.*
How did I find the stationary distribution?

\[
\begin{bmatrix}
0.9 & 0.1 \\
0.5 & 0.5 \\
\end{bmatrix}^2 = \begin{bmatrix}
0.86 & 0.14 \\
0.7 & 0.3 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
0.9 & 0.1 \\
0.5 & 0.5 \\
\end{bmatrix}^4 = \begin{bmatrix}
0.8376 & 0.1624 \\
0.812 & 0.188 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
0.9 & 0.1 \\
0.5 & 0.5 \\
\end{bmatrix}^8 = \begin{bmatrix}
0.833443 & 0.166557 \\
0.832787 & 0.167213 \\
\end{bmatrix}
\]

**Exercise:** Why do the rows converge to \( \pi \)?
We needed the Markov chain to be “aperiodic”.

What is a “periodic” Markov chain?

There is still a stationary distribution.

But it is not a limiting distribution.
Summary so far

Markov Chains can be characterized by the transition matrix $K$.

$$K[i, j] = \Pr[X_t = j \mid X_{t-1} = i] = \Pr[i \to j \text{ in one step}]$$

What is the probability of being in state $i$ after $t$ steps?

$$\pi_t = \pi_0 \cdot K^t \quad \pi_t[i] = (\pi_0 \cdot K^t)[i]$$

There is a unique invariant distribution $\pi$:

$$\pi : \quad \pi = \pi \cdot K$$

For aperiodic Markov Chains: $\pi_t \to \pi$ as $t \to \infty$. 
The plan

Motivating examples and applications

Basic mathematical representation and properties

Applications
How are Markov Chains applied?

2 common types of applications:

1. Build a Markov chain as a statistical model of a real-world process.

   Use the Markov chain to simulate the process.

   e.g. text generation, music composition.

2. Use a measure associated with a Markov chain to approximate a quantity of interest.

   e.g. Google PageRank, image segmentation
How are Markov Chains applied?

2 common types of applications:

1. Build a Markov chain as a statistical model of a real-world process.

   Use the Markov chain to simulate the process.

   e.g. text generation, music composition.

2. Use a measure associated with a Markov chain to approximate a quantity of interest.

   e.g. Google PageRank, image segmentation
Generate a superficially real-looking text given a sample document.

**Idea:**
From the sample document, create a Markov chain. Use a random walk on the Markov chain to generate text.

**Example:**
Collect speeches of Obama, create a Markov chain. Use a random walk to generate new speeches.
1. For each word in the document, create a node/state.

2. Put an edge \textit{word1} \textsuperscript{--->} \textit{word2} if there is a sentence in which \textit{word2} comes after \textit{word1}.

3. Edge probabilities reflect frequency of the pair of words.

The Markov Chain:

- \textit{like} \quad \text{a} \quad \text{the} \quad \text{to}

- \text{like a} \quad 3 \text{ times}
- \text{like the} \quad 4 \text{ times}
- \text{like to} \quad 2 \text{ times}
“I jumped up. I don't know what's going on so I am coming down with a road to opportunity. I believe we can agree on or do about the major challenges facing our country.”
Another use:

Build a Markov chain based on speeches of Obama.
Build a Markov chain based on speeches of Bush.

Given a new quote, can predict if it is by Obama or Bush.

(by testing which Markov model the quote fits best)
Simple version

Given an image that contains an object, figure out:
which pixels correspond to the object,
which pixels correspond to the background.

i.e., label each pixel “object” or “background”

(user labels a small number of pixels with known labels)
The Markov Chain:

1. Each pixel is a node/state.
2. There is an edge between adjacent pixels.
3. Edge probabilities reflect similarity between pixels.

Which one is more likely: random walker first visits "background" or "object"?

Salvador Dali (1922)
The Drunkard
Image Segmentation
Google PageRank

PageRank is a measure of reputation:
The number and reputation of links pointing to you.

The Markov Chain:
Google PageRank

PageRank is a measure of reputation:
The number and reputation of links pointing to you.

The Markov Chain:

1. Every webpage is a node/state.

2. Each hyperlink is an edge:
   
   if webpage A has a link to webpage B, \( A \rightarrow B \)

3a. If A has \( m \) outgoing edges, each gets label \( 1/m \).

3b. If A has no outgoing edges, put edge \( A \rightarrow B \) \( \forall B \)
   (jump to a random page)
A little tweak:
Random surfer jumps to a random page with 15% prob.

Stationary distribution:
probability of being in state A in the long run

PageRank of webpage A
= The stationary probability of A
Google: 

“PageRank continues to be the heart of our software.”
How are Markov Chains applied?

2 common types of applications:

1. Build a Markov chain as a statistical model of a real-world process.

   Use the Markov chain to simulate the process.

   e.g. text generation, music composition.

2. Use a measure associated with a Markov chain to approximate a quantity of interest.

   e.g. Google PageRank, image segmentation
The plan

Motivating examples and applications

Basic mathematical representation and properties

Applications