Markov models

Source: Hastie et at. (2009), Daumé III. Thanks to D. Hsu.

Sequence models

A **sequence model** (or **time series model**) is a family of probability distributions for (possibly infinite) *sequences* of random variables $\{X_t\}_{t \in \mathcal{T}}$.

Sequence models

A **sequence model** (or **time series model**) is a family of probability distributions for (possibly infinite) *sequences* of random variables $\{X_t\}_{t \in \mathcal{T}}$.

- $\{X_t\}_{t \in \mathcal{T}}$ is a **stochastic process** indexed by the totally-ordered set $\mathcal{T}$ (e.g., $\mathcal{T} = \mathbb{N}$ for discrete time series).

Sequence models

A **sequence model** (or **time series model**) is a family of probability distributions for (possibly infinite) *sequences* of random variables $\{X_t\}_{t \in \mathcal{T}}$.

- $\{X_t\}_{t \in \mathcal{T}}$ is a **stochastic process** indexed by the totally-ordered set $\mathcal{T}$ (e.g., $\mathcal{T} = \mathbb{N}$ for discrete time series).
- Special emphasis is placed on the **linear ordering of $\mathcal{T}$**.

Sequence models

A **sequence model** (or **time series model**) is a family of probability distributions for (possibly infinite) *sequences* of random variables $\{X_t\}_{t \in \mathcal{T}}$.

- $\{X_t\}_{t \in \mathcal{T}}$ is a **stochastic process** indexed by the totally-ordered set $\mathcal{T}$ (e.g., $\mathcal{T} = \mathbb{N}$ for discrete time series).
- Special emphasis is placed on the **linear ordering of $\mathcal{T}$**.

  If $t \in \mathcal{T}$ is the "current time", then X_t is the "current state"; X_τ for $\tau < t$ are "past states"; and X_τ for $\tau > t$ are "future states".

Sequence models

A **sequence model** (or **time series model**) is a family of probability distributions for (possibly infinite) *sequences* of random variables $\{X_t\}_{t \in \mathcal{T}}$.

- $\{X_t\}_{t \in \mathcal{T}}$ is a **stochastic process** indexed by the totally-ordered set $\mathcal{T}$ (e.g., $\mathcal{T} = \mathbb{N}$ for discrete time series).
- Special emphasis is placed on the **linear ordering of $\mathcal{T}$**.

  If $t \in \mathcal{T}$ is the "current time", then X_t is the "current state"; X_τ for $\tau < t$ are "past states"; and X_τ for $\tau > t$ are "future states".

  (May interchange "state" and "observation"—no distinction for now.)

Sequence models

A **sequence model** (or **time series model**) is a family of probability distributions for (possibly infinite) *sequences* of random variables $\{X_t\}_{t \in \mathcal{T}}$.

- $\{X_t\}_{t \in \mathcal{T}}$ is a **stochastic process** indexed by the totally-ordered set $\mathcal{T}$ (e.g., $\mathcal{T} = \mathbb{N}$ for discrete time series).
- Special emphasis is placed on the **linear ordering of $\mathcal{T}$**.

  If $t \in \mathcal{T}$ is the "current time", then X_t is the "current state"; X_τ for $\tau < t$ are "past states"; and X_τ for $\tau > t$ are "future states".

  (May interchange "state" and "observation"—no distinction for now.)

Sequence / time series modeling is an entire subfield in statistics, largely due to the plethora of sequence / time series data in applications:

- Economic / financial data over time
- Climate science
- Genomic sequences
- Speech and natural language
- ...

Markov models

A stochastic process $\{X_t\}_{t\in\mathbb{N}}$ has the **Markov property** if the conditional distribution of the next state X_{t+1} given all previous states $\{X_\tau : \tau \leq t\}$ only depends on the value of the current state X_t.

Markov models

A stochastic process $\{X_t\}_{t \in \mathbb{N}}$ has the **Markov property** if the conditional distribution of the next state X_{t+1} given all previous states $\{X_\tau : \tau \leq t\}$ only depends on the value of the current state X_t.

If the X_t are discrete-valued, then the Markov property means that

$$\Pr(X_{t+1} = x_{t+1} \,|\, X_1 = x_1, \ldots, X_t = x_t) \;=\; \Pr(X_{t+1} = x_{t+1} \,|\, X_t = x_t).$$

$$\cdots \longrightarrow X_{t-1} \longrightarrow \; X_t \; \longrightarrow X_{t+1} \longrightarrow \cdots$$

Markov models

A stochastic process $\{X_t\}_{t\in\mathbb{N}}$ has the **Markov property** if the conditional distribution of the next state X_{t+1} given all previous states $\{X_\tau : \tau \le t\}$ only depends on the value of the current state X_t.

If the X_t are discrete-valued, then the Markov property means that

$$\Pr(X_{t+1} = x_{t+1} \mid X_1 = x_1, \ldots, X_t = x_t) \;=\; \Pr(X_{t+1} = x_{t+1} \mid X_t = x_t).$$

$$\cdots \longrightarrow X_{t-1} \longrightarrow X_t \longrightarrow X_{t+1} \longrightarrow \cdots$$

A stochastic process with the Markov property is called a **Markov chain**.

Markov models

A stochastic process $\{X_t\}_{t\in\mathbb{N}}$ has the **Markov property** if the conditional distribution of the next state X_{t+1} given all previous states $\{X_\tau : \tau \leq t\}$ only depends on the value of the current state X_t.

If the X_t are discrete-valued, then the Markov property means that

$$\Pr(X_{t+1} = x_{t+1} \,|\, X_1 = x_1, \ldots, X_t = x_t) \;=\; \Pr(X_{t+1} = x_{t+1} \,|\, X_t = x_t).$$

$$\cdots \longrightarrow X_{t-1} \longrightarrow X_t \longrightarrow X_{t+1} \longrightarrow \cdots$$

A stochastic process with the Markov property is called a **Markov chain**.

A sequence model for a Markov chain is called a **Markov model**.

Markov chain distributions

To specify a Markov chain (MC):

- Specify the distribution of the initial state X_1.
- Specify a **transition kernel**: $\Pr(X_{t+1} = x' \mid X_t = x)$ for all (x, x').

  (Nothing to do with *kernels* as in SVMs/kernel trick/RKHS.)

Markov chain distributions

To specify a Markov chain (MC):

- Specify the distribution of the initial state X_1.
- Specify a **transition kernel**: $\Pr(X_{t+1} = x' \mid X_t = x)$ for all (x, x').

  (Nothing to do with *kernels* as in SVMs/kernel trick/RKHS.)

We focus on MCs where the **state space** (possible values for each X_t) is finite.

For simplicity, we'll assume the state space is $[d] := \{1, 2, \ldots, d\}$.

Markov chain distributions

To specify a Markov chain (MC):

- Specify the distribution of the initial state X_1.
- Specify a **transition kernel**: $\Pr(X_{t+1} = x' \mid X_t = x)$ for all (x, x').

  (Nothing to do with *kernels* as in SVMs/kernel trick/RKHS.)

We focus on MCs where the **state space** (possible values for each X_t) is finite.

For simplicity, we'll assume the state space is $[d] := \{1, 2, \ldots, d\}$.

- Initial state distribution given by a d-dimensional probability vector $\boldsymbol{\pi}$

$$\pi_i = \Pr(X_1 = i).$$

Markov chain distributions

To specify a Markov chain (MC):

- Specify the distribution of the initial state X_1.
- Specify a **transition kernel**: $\Pr(X_{t+1} = x' \mid X_t = x)$ for all (x, x').

  (Nothing to do with *kernels* as in SVMs/kernel trick/RKHS.)

We focus on MCs where the **state space** (possible values for each X_t) is finite.

For simplicity, we'll assume the state space is $[d] := \{1, 2, \dots, d\}$.

- Initial state distribution given by a d-dimensional probability vector $\boldsymbol{\pi}$

$$\pi_i \;=\; \Pr(X_1 = i).$$

- Transition kernel can be written as a $d \times d$ matrix $\boldsymbol{A}$

$$A_{i,j} \;=\; \Pr(X_{t+1} = j \mid X_t = i)$$

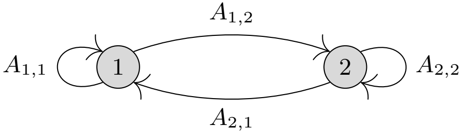
  (rows of $\boldsymbol{A}$ are probability vectors).

  Also called a *transition matrix* or *(right) stochastic matrix*.

Example: a two-state Markov chain

State space: $\{1, 2\}$.
Parameters:

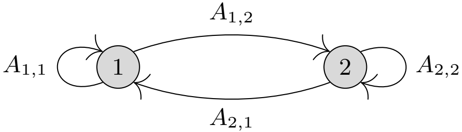
$$\boldsymbol{\pi} = \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{array}{cc} & \begin{array}{cc} \text{state 1} & \text{state 2} \end{array} \\ \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} & \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$



Example: a two-state Markov chain

State space: $\{1, 2\}$.
Parameters:

$$\boldsymbol{\pi} = \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{array}{cc} & \begin{array}{cc} \text{state 1} & \text{state 2} \end{array} \\ \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} & \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$



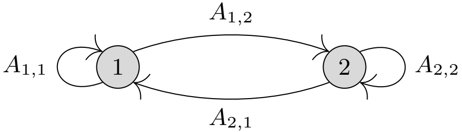
A random state sequence drawn from this MC:

$$(2, 2, 2, 1, 1, 2, 2, 1)$$

Example: a two-state Markov chain

State space: $\{1, 2\}$.
Parameters:

$$\boldsymbol{\pi} = \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{array}{cc} & \begin{array}{cc} \text{state 1} & \text{state 2} \end{array} \\ \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} & \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$



A random state sequence drawn from this MC:

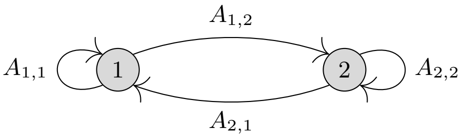
$$(2, 2, 2, 1, 1, 2, 2, 1)$$

What is the probability of this sequence?

Example: a two-state Markov chain

State space: $\{1, 2\}$.
Parameters:

$$\boldsymbol{\pi} = \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{array}{c} \\ \text{state 1} \\ \text{state 2} \end{array} \begin{array}{c} \begin{array}{cc} \text{state 1} & \text{state 2} \end{array} \\ \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$



A random state sequence drawn from this MC:

$$(2, 2, 2, 1, 1, 2, 2, 1)$$

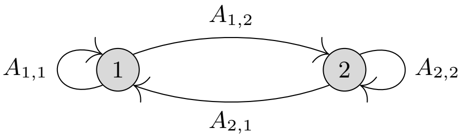
What is the probability of this sequence?

π_2

Example: a two-state Markov chain

State space: $\{1, 2\}$.
Parameters:

$$
\boldsymbol{\pi} = \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{array}{c} \\ \text{state 1} \\ \text{state 2} \end{array} \begin{array}{c} \begin{array}{cc} \text{state 1} & \text{state 2} \end{array} \\ \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.
$$



A random state sequence drawn from this MC:

$$(2, 2, 2, 1, 1, 2, 2, 1)$$

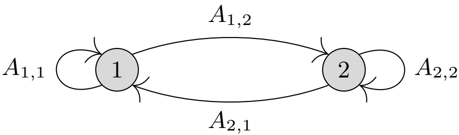
What is the probability of this sequence?

$\pi_2 \times A_{2,2}$

Example: a two-state Markov chain

State space: $\{1, 2\}$.
Parameters:

$$\boldsymbol{\pi} = \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{array}{cc} & \begin{array}{cc} \text{state 1} & \text{state 2} \end{array} \\ \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} & \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$



A random state sequence drawn from this MC:

$$(2, 2, 2, 1, 1, 2, 2, 1)$$

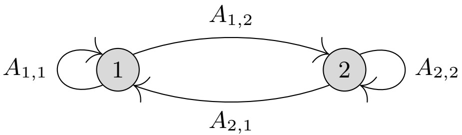
What is the probability of this sequence?

$\pi_2 \times A_{2,2} \times A_{2,2}$

Example: a two-state Markov chain

State space: $\{1, 2\}$.
Parameters:

$$\boldsymbol{\pi} = \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{array}{c} \\ \text{state 1} \\ \text{state 2} \end{array} \begin{array}{c} \begin{array}{cc} \text{state 1} & \text{state 2} \end{array} \\ \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$



A random state sequence drawn from this MC:

$$(2, 2, 2, 1, 1, 2, 2, 1)$$

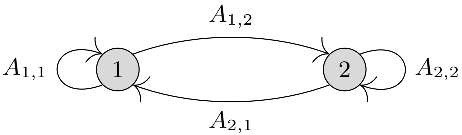
What is the probability of this sequence?

$$\pi_2 \times A_{2,2} \times A_{2,2} \times A_{2,1}$$

Example: a two-state Markov chain

State space: $\{1, 2\}$.
Parameters:

$$\boldsymbol{\pi} = \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{array}{c} \\ \text{state 1} \\ \text{state 2} \end{array} \begin{array}{c} \begin{array}{cc} \text{state 1} & \text{state 2} \end{array} \\ \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$



A random state sequence drawn from this MC:

$$(2, 2, 2, 1, 1, 2, 2, 1)$$

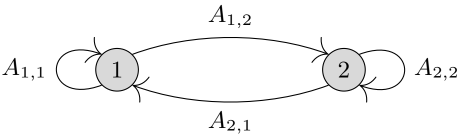
What is the probability of this sequence?

$$\pi_2 \times A_{2,2} \times A_{2,2} \times A_{2,1} \times A_{1,1}$$

Example: a two-state Markov chain

State space: $\{1, 2\}$.
Parameters:

$$\boldsymbol{\pi} = \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{array}{c} \\ \text{state 1} \\ \text{state 2} \end{array} \begin{array}{c} \begin{array}{cc} \text{state 1} & \text{state 2} \end{array} \\ \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$



A random state sequence drawn from this MC:

$$(2, 2, 2, 1, 1, 2, 2, 1)$$

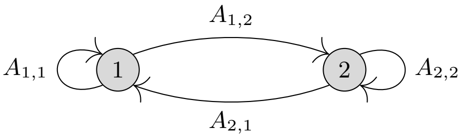
What is the probability of this sequence?

$$\pi_2 \times A_{2,2} \times A_{2,2} \times A_{2,1} \times A_{1,1} \times A_{1,2}$$

Example: a two-state Markov chain

State space: $\{1, 2\}$.
Parameters:

$$\boldsymbol{\pi} = \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{array}{c} \\ \text{state 1} \\ \text{state 2} \end{array} \begin{array}{c} \begin{array}{cc} \text{state 1} & \text{state 2} \end{array} \\ \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$



A random state sequence drawn from this MC:

$$(2, 2, 2, 1, 1, 2, 2, 1)$$

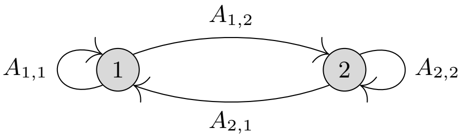
What is the probability of this sequence?

$$\pi_2 \times A_{2,2} \times A_{2,2} \times A_{2,1} \times A_{1,1} \times A_{1,2} \times A_{2,2}$$

Example: a two-state Markov chain

State space: $\{1, 2\}$.
Parameters:

$$\boldsymbol{\pi} = \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{array}{c c} & \begin{array}{cc} \text{state 1} & \text{state 2} \end{array} \\ \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} & \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$



A random state sequence drawn from this MC:

$$(2, 2, 2, 1, 1, 2, 2, 1)$$

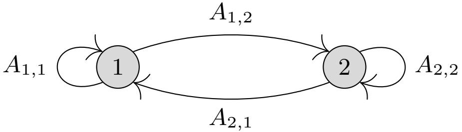
What is the probability of this sequence?

$$\pi_2 \times A_{2,2} \times A_{2,2} \times A_{2,1} \times A_{1,1} \times A_{1,2} \times A_{2,2} \times A_{2,1}$$

Example: a two-state Markov chain

State space: $\{1, 2\}$.
Parameters:

$$\boldsymbol{\pi} = \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{array}{cc} & \begin{array}{cc} \text{state 1} & \text{state 2} \end{array} \\ \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} & \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$



A random state sequence drawn from this MC:

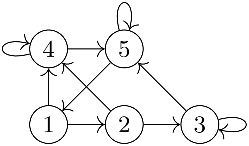
$$(2, 2, 2, 1, 1, 2, 2, 1)$$

What is the probability of this sequence?

$$\pi_2 \times A_{2,2} \times A_{2,2} \times A_{2,1} \times A_{1,1} \times A_{1,2} \times A_{2,2} \times A_{2,1} = 0.00435456$$

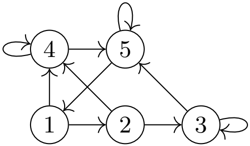
Example: random walk on a directed graph

Consider a directed graph $G = (V, E)$ over $|V| = d$ vertices (self-loops ok).



Example: random walk on a directed graph

Consider a directed graph $G = (V, E)$ over $|V| = d$ vertices (self-loops ok).



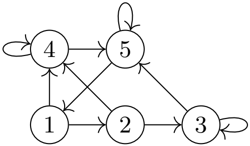
MC for random walk on G:

$$\pi_i \;=\; \mathbb{1}\{\text{start vertex is } i\}, \quad A_{i,j} \;=\; \frac{\mathbb{1}\{(i,j) \in E\}}{\text{out degree}(i)}.$$

$$\begin{array}{cc} & \begin{array}{ccccc} \text{state 1} & \text{state 2} & \text{state 3} & \text{state 4} & \text{state 5} \end{array} \\ \begin{array}{c} \text{state 1} \\ \text{state 2} \\ \text{state 3} \\ \text{state 4} \\ \text{state 5} \end{array} & \begin{pmatrix} 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0 & 0 & 0.5 \end{pmatrix} \end{array}$$

Example: random walk on a directed graph

Consider a directed graph $G = (V, E)$ over $|V| = d$ vertices (self-loops ok).



MC for random walk on G:

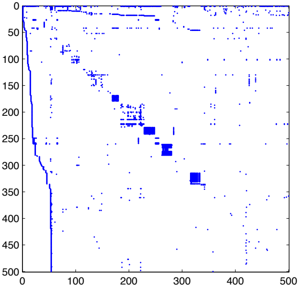
$$\pi_i \;=\; \mathbb{1}\{\text{start vertex is } i\}, \quad A_{i,j} \;=\; \frac{\mathbb{1}\{(i,j) \in E\}}{\text{out degree}(i)}.$$

$$
\begin{array}{rc}
 & \begin{array}{ccccc} \text{state 1} & \text{state 2} & \text{state 3} & \text{state 4} & \text{state 5} \end{array} \\
\begin{array}{r} \text{state 1} \\ \text{state 2} \\ \text{state 3} \\ \text{state 4} \\ \text{state 5} \end{array} &
\begin{pmatrix}
 & * & & * & \\
 & & * & * & \\
 & & * & & * \\
 & & & * & * \\
* & & & & *
\end{pmatrix}
\end{array}
$$

The non-zero pattern of $\boldsymbol{A}$ gives the adjacency structure of G (vertices = states).

Example: PageRank

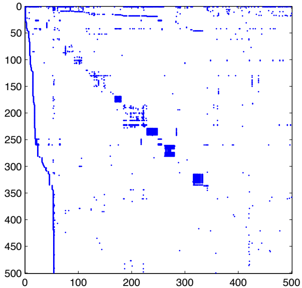
Web graph $G = (V, E)$:
Vertices are webpages, directed edges are hyperlinks between webpages.



Adjacency matrix of the web graph for 500 web pages.

Example: PageRank

Web graph $G = (V, E)$:
Vertices are webpages, directed edges are hyperlinks between webpages.

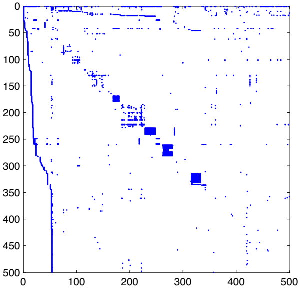


Adjacency matrix of the web graph for 500 web pages.

How popular is webpage i?

Example: PageRank

Web graph $G = (V, E)$:
Vertices are webpages, directed edges are hyperlinks between webpages.



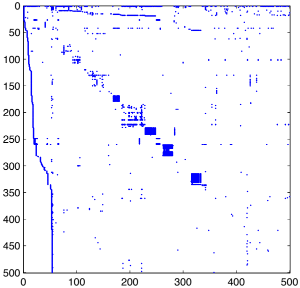
Adjacency matrix of the web graph for 500 web pages.

How popular is webpage i?

Possible answer: probability that **random walk** ends at i after many steps.

Example: PageRank

Web graph $G = (V, E)$:
Vertices are webpages, directed edges are hyperlinks between webpages.



Adjacency matrix of the web graph for 500 web pages.

How popular is webpage i?

Possible answer: probability that **random walk** ends at i after many steps.

$$\Pr(X_t = i) \quad \text{for large } t.$$

Markov chain state distributions

Marginal probabilities

What is the marginal distribution of X_2 in terms of $\boldsymbol{\pi}$ and $\boldsymbol{A}$?

Marginal probabilities

What is the marginal distribution of X_2 in terms of $\boldsymbol{\pi}$ and $\boldsymbol{A}$?

For each $j \in [d]$,

$$\Pr(X_2 = j)$$

Marginal probabilities

What is the marginal distribution of X_2 in terms of $\boldsymbol{\pi}$ and $\boldsymbol{A}$?

For each $j \in [d]$,

$$\Pr(X_2 = j) \quad = \quad \sum_{i=1}^{d} \Pr(X_1 = i,\, X_2 = j)$$

Marginal probabilities

What is the marginal distribution of X_2 in terms of $\boldsymbol{\pi}$ and $\boldsymbol{A}$?

For each $j \in [d]$,

$$\begin{aligned}\Pr(X_2 = j) &= \sum_{i=1}^{d} \Pr(X_1 = i,\, X_2 = j) \\ &= \sum_{i=1}^{d} \Pr(X_1 = i) \cdot \Pr(X_2 = j \mid X_1 = i)\end{aligned}$$

Marginal probabilities

What is the marginal distribution of X_2 in terms of $\boldsymbol{\pi}$ and $\boldsymbol{A}$?

For each $j \in [d]$,

$$\begin{aligned} \Pr(X_2 = j) &= \sum_{i=1}^{d} \Pr(X_1 = i,\, X_2 = j) \\ &= \sum_{i=1}^{d} \Pr(X_1 = i) \cdot \Pr(X_2 = j \mid X_1 = i) \\ &= \sum_{i=1}^{d} \pi_i \cdot A_{i,j} \end{aligned}$$

Marginal probabilities

What is the marginal distribution of X_2 in terms of $\boldsymbol{\pi}$ and $\boldsymbol{A}$?

For each $j \in [d]$,

$$
\begin{aligned}
\Pr(X_2 = j) &= \sum_{i=1}^{d} \Pr(X_1 = i,\, X_2 = j) \\
&= \sum_{i=1}^{d} \Pr(X_1 = i) \cdot \Pr(X_2 = j \mid X_1 = i) \\
&= \sum_{i=1}^{d} \pi_i \cdot A_{i,j} \\
&= j\text{-th entry of } \boldsymbol{\pi}^\top \boldsymbol{A}.
\end{aligned}
$$

Marginal probabilities

What is the marginal distribution of X_3 in terms of $\boldsymbol{\pi}$ and $\boldsymbol{A}$?

Marginal probabilities

What is the marginal distribution of X_3 in terms of $\boldsymbol{\pi}$ and $\boldsymbol{A}$? For each $k \in [d]$,

$$\Pr(X_3 = k) \;=\; \sum_{i=1}^{d} \sum_{j=1}^{d} \Pr(X_1 = i,\, X_2 = j,\, X_3 = k)$$

Marginal probabilities

What is the marginal distribution of X_3 in terms of $\boldsymbol{\pi}$ and $\boldsymbol{A}$? For each $k \in [d]$,

$$\Pr(X_3 = k) \;=\; \sum_{i=1}^{d} \sum_{j=1}^{d} \Pr(X_1 = i,\, X_2 = j,\, X_3 = k)$$

$$= \;\; \sum_{i=1}^{d} \sum_{j=1}^{d} \Pr(X_1 = i) \cdot \Pr(X_2 = j \,|\, X_1 = i) \cdot \Pr(X_3 = k \,|\, X_1 = i,\, X_2 = j)$$

Marginal probabilities

What is the marginal distribution of X_3 in terms of $\boldsymbol{\pi}$ and $\boldsymbol{A}$? For each $k \in [d]$,

$$
\begin{aligned}
\Pr(X_3 = k) &= \sum_{i=1}^{d} \sum_{j=1}^{d} \Pr(X_1 = i,\, X_2 = j,\, X_3 = k) \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} \Pr(X_1 = i) \cdot \Pr(X_2 = j \mid X_1 = i) \cdot \Pr(X_3 = k \mid X_1 = i,\, X_2 = j) \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} \Pr(X_1 = i) \cdot \Pr(X_2 = j \mid X_1 = i) \cdot \Pr(X_3 = k \mid X_2 = j)
\end{aligned}
$$

Marginal probabilities

What is the marginal distribution of X_3 in terms of $\boldsymbol{\pi}$ and $\boldsymbol{A}$? For each $k \in [d]$,

$$
\begin{aligned}
\Pr(X_3 = k) &= \sum_{i=1}^{d} \sum_{j=1}^{d} \Pr(X_1 = i,\, X_2 = j,\, X_3 = k) \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} \Pr(X_1 = i) \cdot \Pr(X_2 = j \,|\, X_1 = i) \cdot \Pr(X_3 = k \,|\, X_1 = i,\, X_2 = j) \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} \Pr(X_1 = i) \cdot \Pr(X_2 = j \,|\, X_1 = i) \cdot \Pr(X_3 = k \,|\, X_2 = j) \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} \pi_i \cdot A_{i,j} \cdot A_{j,k}
\end{aligned}
$$

Marginal probabilities

What is the marginal distribution of X_3 in terms of $\boldsymbol{\pi}$ and $\boldsymbol{A}$? For each $k \in [d]$,

$$\begin{aligned}
\Pr(X_3 = k) &= \sum_{i=1}^{d} \sum_{j=1}^{d} \Pr(X_1 = i,\, X_2 = j,\, X_3 = k) \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} \Pr(X_1 = i) \cdot \Pr(X_2 = j \,|\, X_1 = i) \cdot \Pr(X_3 = k \,|\, X_1 = i,\, X_2 = j) \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} \Pr(X_1 = i) \cdot \Pr(X_2 = j \,|\, X_1 = i) \cdot \Pr(X_3 = k \,|\, X_2 = j) \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} \pi_i \cdot A_{i,j} \cdot A_{j,k} \\
&= k\text{-th entry of } \boldsymbol{\pi}^\top \boldsymbol{A}\boldsymbol{A}.
\end{aligned}$$

Marginal probabilities

What is the marginal distribution of X_3 in terms of $\boldsymbol{\pi}$ and $\boldsymbol{A}$? For each $k \in [d]$,

$$
\begin{aligned}
\Pr(X_3 = k) &= \sum_{i=1}^{d} \sum_{j=1}^{d} \Pr(X_1 = i,\, X_2 = j,\, X_3 = k) \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} \Pr(X_1 = i) \cdot \Pr(X_2 = j \mid X_1 = i) \cdot \Pr(X_3 = k \mid X_1 = i,\, X_2 = j) \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} \Pr(X_1 = i) \cdot \Pr(X_2 = j \mid X_1 = i) \cdot \Pr(X_3 = k \mid X_2 = j) \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} \pi_i \cdot A_{i,j} \cdot A_{j,k} \\
&= k\text{-th entry of } \boldsymbol{\pi}^\top \boldsymbol{A}\boldsymbol{A}.
\end{aligned}
$$

For any $t \in \mathbb{N}$, the marginal distribution of X_t in terms of $\boldsymbol{\pi}$ and $\boldsymbol{A}$ is

$$
\Pr(X_t = k) = k\text{-th entry of } \boldsymbol{\pi}^\top \underbrace{\boldsymbol{A}\boldsymbol{A}\cdots\boldsymbol{A}}_{t-1 \text{ times}}.
$$

Powers of the transition matrix

The (i,j)-th entry of $\boldsymbol{A}^p = \underbrace{\boldsymbol{A}\boldsymbol{A}\cdots\boldsymbol{A}}_{p \text{ times}}$ is the p-step transition matrix

$$\left[\boldsymbol{A}^p\right]_{i,j} \;=\; \Pr(X_{t+p} = j \mid X_t = i).$$

Powers of the transition matrix

The (i, j)-th entry of $\boldsymbol{A}^p = \underbrace{\boldsymbol{A}\boldsymbol{A}\cdots\boldsymbol{A}}_{p \text{ times}}$ is the p-step transition matrix

$$\left[\boldsymbol{A}^p\right]_{i,j} \;=\; \Pr(X_{t+p} = j \mid X_t = i).$$

Example: State space: $\{1, 2\}$. Parameters: $\boldsymbol{\pi} = \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix}$

Powers of the transition matrix

The (i,j)-th entry of $\boldsymbol{A}^p = \underbrace{\boldsymbol{A}\boldsymbol{A}\cdots\boldsymbol{A}}_{p \text{ times}}$ is the p-step transition matrix

$$\left[\boldsymbol{A}^p\right]_{i,j} \;=\; \Pr(X_{t+p} = j \mid X_t = i).$$

Example: State space: $\{1, 2\}$. Parameters: $\boldsymbol{\pi} = \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix}$

$$\boldsymbol{\pi}^\top \boldsymbol{A} \;=\; \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \;=\; \begin{pmatrix} 0.57 & 0.43 \end{pmatrix}$$

Powers of the transition matrix

The (i,j)-th entry of $\boldsymbol{A}^p = \underbrace{\boldsymbol{A}\boldsymbol{A}\cdots\boldsymbol{A}}_{p \text{ times}}$ is the p-step transition matrix

$$\left[\boldsymbol{A}^p\right]_{i,j} \;=\; \Pr(X_{t+p} = j \mid X_t = i).$$

Example: State space: $\{1,2\}$. Parameters: $\boldsymbol{\pi} = \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}$, $\quad \boldsymbol{A} = \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix}$

$$\begin{aligned} \boldsymbol{\pi}^\top \boldsymbol{A} &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} = \begin{pmatrix} 0.57 & 0.43 \end{pmatrix} \\ \boldsymbol{\pi}^\top \boldsymbol{A}^5 &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.46023 & 0.53977 \\ 0.46266 & 0.53734 \end{pmatrix} = \begin{pmatrix} 0.462417 & 0.537583 \end{pmatrix} \end{aligned}$$

Powers of the transition matrix

The (i,j)-th entry of $\boldsymbol{A}^p = \underbrace{\boldsymbol{A}\boldsymbol{A}\cdots\boldsymbol{A}}_{p \text{ times}}$ is the p-step transition matrix

$$\left[\boldsymbol{A}^p\right]_{i,j} \;=\; \Pr(X_{t+p} = j \mid X_t = i).$$

Example: State space: $\{1,2\}$. Parameters: $\boldsymbol{\pi} = \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix}$

$$\begin{aligned}
\boldsymbol{\pi}^\top \boldsymbol{A} &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} = \begin{pmatrix} 0.57 & 0.43 \end{pmatrix} \\
\boldsymbol{\pi}^\top \boldsymbol{A}^5 &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.46023 & 0.53977 \\ 0.46266 & 0.53734 \end{pmatrix} = \begin{pmatrix} 0.462417 & 0.537583 \end{pmatrix} \\
\boldsymbol{\pi}^\top \boldsymbol{A}^{100} &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.461538 & 0.538414 \\ 0.461538 & 0.538414 \end{pmatrix} = \begin{pmatrix} 0.461538 & 0.538414 \end{pmatrix}
\end{aligned}$$

Powers of the transition matrix

The (i,j)-th entry of $\boldsymbol{A}^p = \underbrace{\boldsymbol{A}\boldsymbol{A}\cdots\boldsymbol{A}}_{p \text{ times}}$ is the p-step transition matrix

$$\left[\boldsymbol{A}^p\right]_{i,j} \;=\; \Pr(X_{t+p} = j \mid X_t = i).$$

Example: State space: $\{1,2\}$. Parameters: $\boldsymbol{\pi} = \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}$, $\quad \boldsymbol{A} = \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix}$

$$\begin{aligned}
\boldsymbol{\pi}^\top \boldsymbol{A} &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} = \begin{pmatrix} 0.57 & 0.43 \end{pmatrix} \\
\boldsymbol{\pi}^\top \boldsymbol{A}^5 &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.46023 & 0.53977 \\ 0.46266 & 0.53734 \end{pmatrix} = \begin{pmatrix} 0.462417 & 0.537583 \end{pmatrix} \\
\boldsymbol{\pi}^\top \boldsymbol{A}^{100} &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.461538 & 0.538414 \\ 0.461538 & 0.538414 \end{pmatrix} = \begin{pmatrix} 0.461538 & 0.538414 \end{pmatrix} \\
\boldsymbol{\pi}^\top \boldsymbol{A}^{1000} &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.461538 & 0.538414 \\ 0.461538 & 0.538414 \end{pmatrix} = \begin{pmatrix} 0.461538 & 0.538414 \end{pmatrix}
\end{aligned}$$

Powers of the transition matrix

The (i,j)-th entry of $\boldsymbol{A}^p = \underbrace{\boldsymbol{A}\boldsymbol{A}\cdots\boldsymbol{A}}_{p \text{ times}}$ is the p-step transition matrix

$$\left[\boldsymbol{A}^p\right]_{i,j} \;=\; \Pr(X_{t+p} = j \mid X_t = i).$$

Example: State space: $\{1,2\}$. Parameters: $\boldsymbol{\pi} = \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix}$

$$\begin{aligned}
\boldsymbol{\pi}^\top \boldsymbol{A} &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} = \begin{pmatrix} 0.57 & 0.43 \end{pmatrix} \\
\boldsymbol{\pi}^\top \boldsymbol{A}^5 &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.46023 & 0.53977 \\ 0.46266 & 0.53734 \end{pmatrix} = \begin{pmatrix} 0.462417 & 0.537583 \end{pmatrix} \\
\boldsymbol{\pi}^\top \boldsymbol{A}^{100} &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.461538 & 0.538414 \\ 0.461538 & 0.538414 \end{pmatrix} = \begin{pmatrix} 0.461538 & 0.538414 \end{pmatrix} \\
\boldsymbol{\pi}^\top \boldsymbol{A}^{1000} &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.461538 & 0.538414 \\ 0.461538 & 0.538414 \end{pmatrix} = \begin{pmatrix} 0.461538 & 0.538414 \end{pmatrix}
\end{aligned}$$

Convergence?

Powers of the transition matrix

The (i,j)-th entry of $\boldsymbol{A}^p = \underbrace{\boldsymbol{A}\boldsymbol{A}\cdots\boldsymbol{A}}_{p \text{ times}}$ is the p-step transition matrix

$$\left[\boldsymbol{A}^p\right]_{i,j} = \Pr(X_{t+p} = j \mid X_t = i).$$

Example: State space: $\{1,2\}$. Parameters: $\boldsymbol{\pi} = \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix}$

$$\begin{aligned}
\boldsymbol{\pi}^\top \boldsymbol{A} &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} = \begin{pmatrix} 0.57 & 0.43 \end{pmatrix} \\
\boldsymbol{\pi}^\top \boldsymbol{A}^5 &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.46023 & 0.53977 \\ 0.46266 & 0.53734 \end{pmatrix} = \begin{pmatrix} 0.462417 & 0.537583 \end{pmatrix} \\
\boldsymbol{\pi}^\top \boldsymbol{A}^{100} &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.461538 & 0.538414 \\ 0.461538 & 0.538414 \end{pmatrix} = \begin{pmatrix} 0.461538 & 0.538414 \end{pmatrix} \\
\boldsymbol{\pi}^\top \boldsymbol{A}^{1000} &= \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 0.461538 & 0.538414 \\ 0.461538 & 0.538414 \end{pmatrix} = \begin{pmatrix} 0.461538 & 0.538414 \end{pmatrix}
\end{aligned}$$

Convergence? Doesn't even seem to matter what π is!

Limiting state distribution

Certain "nice" transition matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ have the property that

$$\lim_{p \to \infty} \boldsymbol{A}^p \;=\; \text{stochastic matrix with identical rows}$$

Limiting state distribution

Certain "nice" transition matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ have the property that

$$\lim_{p \to \infty} \boldsymbol{A}^p \ = \ \text{stochastic matrix with identical rows} \ =: \ \begin{pmatrix} \text{——} & \boldsymbol{q}^\top & \text{——} \\ \text{——} & \boldsymbol{q}^\top & \text{——} \\ & \vdots & \\ \text{——} & \boldsymbol{q}^\top & \text{——} \end{pmatrix}$$

Limiting state distribution

Certain "nice" transition matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ have the property that

$$\lim_{p \to \infty} \boldsymbol{A}^p \;=\; \text{stochastic matrix with identical rows} \;=:\; \begin{pmatrix} \text{---} & \boldsymbol{q}^\top & \text{---} \\ \text{---} & \boldsymbol{q}^\top & \text{---} \\ & \vdots & \\ \text{---} & \boldsymbol{q}^\top & \text{---} \end{pmatrix}$$

What can we say about $\boldsymbol{q}$?

Limiting state distribution

Certain "nice" transition matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ have the property that

$$\lim_{p \to \infty} \boldsymbol{A}^p \;=\; \text{stochastic matrix with identical rows} \;=: \begin{pmatrix} \text{---} & \boldsymbol{q}^\top & \text{---} \\ \text{---} & \boldsymbol{q}^\top & \text{---} \\ & \vdots & \\ \text{---} & \boldsymbol{q}^\top & \text{---} \end{pmatrix}$$

What can we say about $\boldsymbol{q}$? For such $\boldsymbol{A}$,

$$\lim_{p \to \infty} \boldsymbol{A}^p$$

LIMITING STATE DISTRIBUTION

Certain "nice" transition matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ have the property that

$$\lim_{p\to\infty} \boldsymbol{A}^p \;=\; \text{stochastic matrix with identical rows} \;=:\; \begin{pmatrix} — & \boldsymbol{q}^\top & — \\ — & \boldsymbol{q}^\top & — \\ & \vdots & \\ — & \boldsymbol{q}^\top & — \end{pmatrix}$$

What can we say about $\boldsymbol{q}$? For such $\boldsymbol{A}$,

$$\lim_{p\to\infty} \boldsymbol{A}^p \;=\; \left(\lim_{p\to\infty} \boldsymbol{A}^{p-1}\right)\boldsymbol{A}$$

Limiting state distribution

Certain "nice" transition matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ have the property that

$$\lim_{p \to \infty} \boldsymbol{A}^p \;=\; \text{stochastic matrix with identical rows} \;=:\; \begin{pmatrix} \text{——} & \boldsymbol{q}^\top & \text{——} \\ \text{——} & \boldsymbol{q}^\top & \text{——} \\ & \vdots & \\ \text{——} & \boldsymbol{q}^\top & \text{——} \end{pmatrix}$$

What can we say about $\boldsymbol{q}$? For such $\boldsymbol{A}$,

$$\lim_{p \to \infty} \boldsymbol{A}^p \;=\; \left(\lim_{p \to \infty} \boldsymbol{A}^{p-1} \right) \boldsymbol{A} \;=\; \begin{pmatrix} \text{——} & \boldsymbol{q}^\top & \text{——} \\ \text{——} & \boldsymbol{q}^\top & \text{——} \\ & \vdots & \\ \text{——} & \boldsymbol{q}^\top & \text{——} \end{pmatrix} \boldsymbol{A}$$

Limiting state distribution

Certain "nice" transition matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ have the property that

$$\lim_{p \to \infty} \boldsymbol{A}^p \;=\; \text{stochastic matrix with identical rows} \;=:\; \begin{pmatrix} \text{---} & \boldsymbol{q}^\top & \text{---} \\ \text{---} & \boldsymbol{q}^\top & \text{---} \\ & \vdots & \\ \text{---} & \boldsymbol{q}^\top & \text{---} \end{pmatrix}$$

What can we say about $\boldsymbol{q}$? For such $\boldsymbol{A}$,

$$\lim_{p \to \infty} \boldsymbol{A}^p \;=\; \left(\lim_{p \to \infty} \boldsymbol{A}^{p-1}\right)\boldsymbol{A} \;=\; \begin{pmatrix} \text{---} & \boldsymbol{q}^\top & \text{---} \\ \text{---} & \boldsymbol{q}^\top & \text{---} \\ & \vdots & \\ \text{---} & \boldsymbol{q}^\top & \text{---} \end{pmatrix}\boldsymbol{A} \;=\; \begin{pmatrix} \text{---} & \boldsymbol{q}^\top & \text{---} \\ \text{---} & \boldsymbol{q}^\top & \text{---} \\ & \vdots & \\ \text{---} & \boldsymbol{q}^\top & \text{---} \end{pmatrix}$$

Limiting state distribution

Certain "nice" transition matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ have the property that

$$\lim_{p \to \infty} \boldsymbol{A}^p \;=\; \text{stochastic matrix with identical rows} \;=:\; \begin{pmatrix} \text{——} & \boldsymbol{q}^\top & \text{——} \\ \text{——} & \boldsymbol{q}^\top & \text{——} \\ & \vdots & \\ \text{——} & \boldsymbol{q}^\top & \text{——} \end{pmatrix}$$

What can we say about $\boldsymbol{q}$? For such $\boldsymbol{A}$,

$$\lim_{p \to \infty} \boldsymbol{A}^p \;=\; \left(\lim_{p \to \infty} \boldsymbol{A}^{p-1}\right)\boldsymbol{A} \;=\; \begin{pmatrix} \text{——} & \boldsymbol{q}^\top & \text{——} \\ \text{——} & \boldsymbol{q}^\top & \text{——} \\ & \vdots & \\ \text{——} & \boldsymbol{q}^\top & \text{——} \end{pmatrix} \boldsymbol{A} \;=\; \begin{pmatrix} \text{——} & \boldsymbol{q}^\top & \text{——} \\ \text{——} & \boldsymbol{q}^\top & \text{——} \\ & \vdots & \\ \text{——} & \boldsymbol{q}^\top & \text{——} \end{pmatrix}$$

i.e.,

$$\boldsymbol{q}^\top \boldsymbol{A} \;=\; \boldsymbol{q}^\top. \tag{$\star$}$$

Limiting state distribution

Certain "nice" transition matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ have the property that

$$\lim_{p \to \infty} \boldsymbol{A}^p \;=\; \text{stochastic matrix with identical rows} \;=: \begin{pmatrix} \text{——} & \boldsymbol{q}^\top & \text{——} \\ \text{——} & \boldsymbol{q}^\top & \text{——} \\ & \vdots & \\ \text{——} & \boldsymbol{q}^\top & \text{——} \end{pmatrix}$$

What can we say about $\boldsymbol{q}$? For such $\boldsymbol{A}$,

$$\lim_{p \to \infty} \boldsymbol{A}^p \;=\; \left(\lim_{p \to \infty} \boldsymbol{A}^{p-1}\right)\boldsymbol{A} \;=\; \begin{pmatrix} \text{——} & \boldsymbol{q}^\top & \text{——} \\ \text{——} & \boldsymbol{q}^\top & \text{——} \\ & \vdots & \\ \text{——} & \boldsymbol{q}^\top & \text{——} \end{pmatrix} \boldsymbol{A} \;=\; \begin{pmatrix} \text{——} & \boldsymbol{q}^\top & \text{——} \\ \text{——} & \boldsymbol{q}^\top & \text{——} \\ & \vdots & \\ \text{——} & \boldsymbol{q}^\top & \text{——} \end{pmatrix}$$

i.e.,

$$\boldsymbol{q}^\top \boldsymbol{A} \;=\; \boldsymbol{q}^\top. \qquad (\star)$$

A solution $\boldsymbol{q}$ to $(\star)$, is called a **stationary distribution**.

Stationary distribution

Suppose a MC has a unique stationary distribution $\boldsymbol{q} = (q_1, q_2, \ldots, q_d)$.

Stationary distribution

Suppose a MC has a unique stationary distribution $\boldsymbol{q} = (q_1, q_2, \ldots, q_d)$.

- For any $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(\left| \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}\{X_t = i\} - q_i \right| > \varepsilon \right) \to 0.$$

Law of Large Numbers for MCs.

Stationary distribution

Suppose a MC has a unique stationary distribution $\boldsymbol{q} = (q_1, q_2, \ldots, q_d)$.

- For any $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{X_t = i\} - q_i\right| > \varepsilon\right) \to 0.$$

  Law of Large Numbers for MCs.

- However, rate of convergence not the same as in the iid case.

Stationary distribution

Suppose a MC has a unique stationary distribution $\boldsymbol{q} = (q_1, q_2, \ldots, q_d)$.

- For any $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{X_t = i\} - q_i\right| > \varepsilon\right) \to 0.$$

  Law of Large Numbers for MCs.

- However, rate of convergence not the same as in the iid case.

  Critically depends on how quickly $\Pr(X_t = \cdot) \to \boldsymbol{q}$ (**mixing rate**).

Stationary distribution

Suppose a MC has a unique stationary distribution $\boldsymbol{q} = (q_1, q_2, \ldots, q_d)$.

- For any $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(\left| \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}\{X_t = i\} - q_i \right| > \varepsilon \right) \to 0.$$

  Law of Large Numbers for MCs.

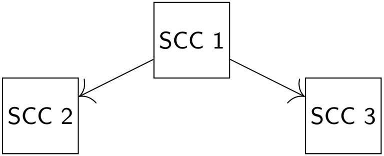
- However, rate of convergence not the same as in the iid case.

  Critically depends on how quickly $\Pr(X_t = \cdot) \to \boldsymbol{q}$ (**mixing rate**).

When does a MC even have a unique stationary distribution?

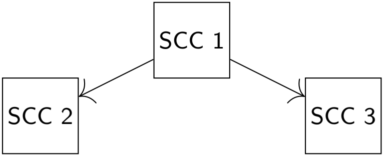
What can go wrong

1. **Directed graph underlying A has more than one strongly connected component "sinks"** $\longrightarrow$ stationary distribution may not be unique.



What can go wrong

1. **Directed graph underlying A has more than one strongly connected component "sinks"** $\longrightarrow$ stationary distribution may not be unique.



Markov chains with only one strongly connected component are called **irreducible**.

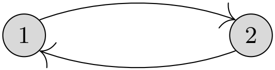
What can go wrong

2. **Oscillation among two or more states** $\longrightarrow$ limit does not exist.

What can go wrong

2. **Oscillation among two or more states** $\longrightarrow$ limit does not exist.

   Example:

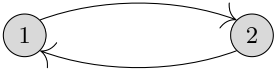
   

   If start at state 1, then never at state 1 on even time steps.

What can go wrong

2. **Oscillation among two or more states** $\longrightarrow$ limit does not exist.

   Example:

   If start at state 1, then never at state 1 on even time steps.

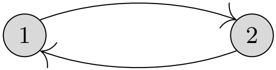
   Markov chains without such oscillation are called **aperiodic**.

   (Formally: there exists p_0 s.t. for all $p \geq p_0$, $[\boldsymbol{A}^p]_{i,i} > 0$ for all $i \in [d]$.)

What can go wrong

2. **Oscillation among two or more states** $\longrightarrow$ limit does not exist.

   Example:

   If start at state 1, then never at state 1 on even time steps.

   Markov chains without such oscillation are called **aperiodic**.

   (Formally: there exists p_0 s.t. for all $p \geq p_0$, $[\boldsymbol{A}^p]_{i,i} > 0$ for all $i \in [d]$.)

   If every state $i \in [d]$ has $A_{i,i} > 0$, then aperiodicity is guaranteed.

Conditions for unique stationary distribution

Theorem: If MC with transition matrix $\boldsymbol{A}$ is *irreducible* and *aperiodic*, then

- There is a unique stationary distribution $\boldsymbol{q}$ (which satisfies $\boldsymbol{q}^\top \boldsymbol{A} = \boldsymbol{q}^\top$).
- $$\lim_{p \to \infty} \boldsymbol{A}^p = \begin{pmatrix} \text{——} & \boldsymbol{q}^\top & \text{——} \\ \text{——} & \boldsymbol{q}^\top & \text{——} \\ & \vdots & \\ \text{——} & \boldsymbol{q}^\top & \text{——} \end{pmatrix}.$$

Computing the stationary distribution

For irreducible and aperiodic MCs, the $\boldsymbol{q}$ that satisfies

$$\boldsymbol{q}^\top \boldsymbol{A} \;=\; \boldsymbol{q}^\top$$

is unique.

Computing the stationary distribution

For irreducible and aperiodic MCs, the $\boldsymbol{q}$ that satisfies

$$\boldsymbol{q}^\top \boldsymbol{A} \;=\; \boldsymbol{q}^\top$$

is unique. Therefore, suffices to find *left eigenvector* of $\boldsymbol{A}$ with eigenvalue 1.

Computing the stationary distribution

For irreducible and aperiodic MCs, the $\boldsymbol{q}$ that satisfies

$$\boldsymbol{q}^\top \boldsymbol{A} \;=\; \boldsymbol{q}^\top$$

is unique. Therefore, suffices to find *left eigenvector* of $\boldsymbol{A}$ with eigenvalue 1.
In fact, $\boldsymbol{A}$ has no other eigenvalue of larger modulus!

Computing the stationary distribution

For irreducible and aperiodic MCs, the $\boldsymbol{q}$ that satisfies

$$\boldsymbol{q}^\top \boldsymbol{A} \;=\; \boldsymbol{q}^\top$$

is unique. Therefore, suffices to find *left eigenvector* of $\boldsymbol{A}$ with eigenvalue 1.
In fact, $\boldsymbol{A}$ has no other eigenvalue of larger modulus!

Direct method: Find any vector in *left null space* of $\boldsymbol{A} - \boldsymbol{I}$

$$\boldsymbol{q}^\top (\boldsymbol{A} - \boldsymbol{I}) \;=\; \boldsymbol{0},$$

and properly normalize it to be a state distribution.

Computing the stationary distribution

For irreducible and aperiodic MCs, the $\boldsymbol{q}$ that satisfies

$$\boldsymbol{q}^\top \boldsymbol{A} \;=\; \boldsymbol{q}^\top$$

is unique. Therefore, suffices to find *left eigenvector* of $\boldsymbol{A}$ with eigenvalue 1.
In fact, $\boldsymbol{A}$ has no other eigenvalue of larger modulus!

Direct method: Find any vector in *left null space* of $\boldsymbol{A} - \boldsymbol{I}$

$$\boldsymbol{q}^\top(\boldsymbol{A} - \boldsymbol{I}) \;=\; \boldsymbol{0},$$

and properly normalize it to be a state distribution.

Power method:

initialize $\boldsymbol{q}$ arbitrarily.
repeat
  $\boldsymbol{q}^\top := \boldsymbol{q}^\top \boldsymbol{A}$.
until bored.
return $\boldsymbol{q}$.

Example: PageRank

Random walk on web graph:

- definitely **not irreducible**,
  (some pages have no links to other pages);
- probably **not aperiodic**.

Example: PageRank

Random walk on web graph:

- definitely **not irreducible**,
  (some pages have no links to other pages);
- probably **not aperiodic**.

Modification:

$$\widetilde{\boldsymbol{A}} := (1-\alpha)\boldsymbol{A} + \frac{\alpha}{d}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}.$$

New MC (with $\widetilde{\boldsymbol{A}}$) is both irreducible and aperiodic.

Example: PageRank

Random walk on web graph:

- definitely **not irreducible**, (some pages have no links to other pages);
- probably **not aperiodic**.

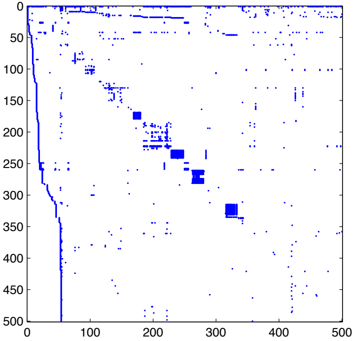
Modification:

$$\widetilde{\boldsymbol{A}} := (1-\alpha)\boldsymbol{A} + \frac{\alpha}{d}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}.$$

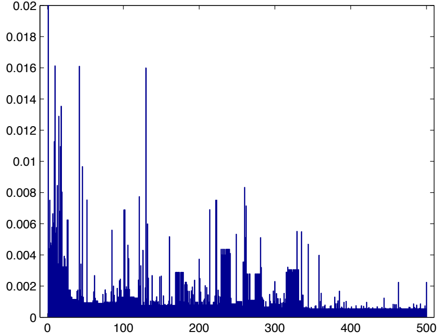
New MC (with $\widetilde{\boldsymbol{A}}$) is both irreducible and aperiodic.

PageRank scores = stationary distribution of this new MC.

Example: PageRank



Adjacency matrix of the web graph for 500 web pages.



PageRank distribution.

(From K. Murphy, "Machine Learning", MIT Press 2012.)

Example: semi-supervised learning

Have some **labeled data** $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ from $\mathcal{X} \times \{\pm 1\}$, and also many **unlabeled data** $x_{m+1}, x_{m+1}, \ldots, x_{m+n}$ from $\mathcal{X}$.

Example: semi-supervised learning

Have some **labeled data** $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ from $\mathcal{X} \times \{\pm 1\}$, and also many **unlabeled data** $x_{m+1}, x_{m+1}, \ldots, x_{m+n}$ from $\mathcal{X}$.

Main question: How to take advantage of U?

Example: semi-supervised learning

Have some **labeled data** $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ from $\mathcal{X} \times \{\pm 1\}$, and also many **unlabeled data** $x_{m+1}, x_{m+1}, \ldots, x_{m+n}$ from $\mathcal{X}$.

Main question: How to take advantage of U?

[Zhu, Ghahramani, and Lafferty, 2003]

- Construct weighted similarity graph $G = (V, W)$ over all data.

Example: semi-supervised learning

Have some **labeled data** $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ from $\mathcal{X} \times \{\pm 1\}$, and also many **unlabeled data** $x_{m+1}, x_{m+1}, \ldots, x_{m+n}$ from $\mathcal{X}$.

Main question: How to take advantage of U?

[Zhu, Ghahramani, and Lafferty, 2003]

- Construct weighted similarity graph $G = (V, W)$ over all data.

  For example:

  - $V = \{1, 2, \ldots, m + n\}$.
  - Weight $W_{i,j} = \exp\left(-\frac{1}{2}\operatorname{dist}(x_i, x_j)^2\right)$.

EXAMPLE: SEMI-SUPERVISED LEARNING

Have some **labeled data** $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ from $\mathcal{X} \times \{\pm 1\}$, and also many **unlabeled data** $x_{m+1}, x_{m+1}, \ldots, x_{m+n}$ from $\mathcal{X}$.

Main question: How to take advantage of U?

[Zhu, Ghahramani, and Lafferty, 2003]

- Construct weighted similarity graph $G = (V, W)$ over all data.

  For example:

  - $V = \{1, 2, \ldots, m + n\}$.
  - Weight $W_{i,j} = \exp\Big(-\frac{1}{2}\operatorname{dist}(x_i, x_j)^2\Big)$.

- Weighted random walk MC:

$$A_{i,j} = \frac{W_{i,j}}{\sum_{k=1}^{m+n} W_{i,k}}.$$

Example: semi-supervised learning

Have some **labeled data** $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ from $\mathcal{X} \times \{\pm 1\}$, and also many **unlabeled data** $x_{m+1}, x_{m+1}, \ldots, x_{m+n}$ from $\mathcal{X}$.

Main question: How to take advantage of U?

[Zhu, Ghahramani, and Lafferty, 2003]

- Construct weighted similarity graph $G = (V, W)$ over all data.

  For example:

  - $V = \{1, 2, \ldots, m + n\}$.
  - Weight $W_{i,j} = \exp\left(-\frac{1}{2}\operatorname{dist}(x_i, x_j)^2\right)$.

- Weighted random walk MC:

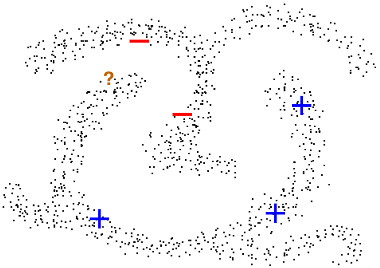
$$A_{i,j} = \frac{W_{i,j}}{\sum_{k=1}^{m+n} W_{i,k}}.$$

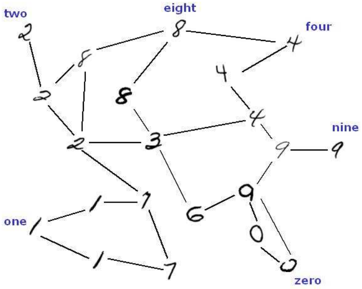
- Start weighted random walk starting from **unlabeled point** x_{m+i}.

  If **first labeled point reached** has label $y \in \{\pm 1\}$, then use $\hat{y}_{m+i} := y$ as the label for x_{m+i}.

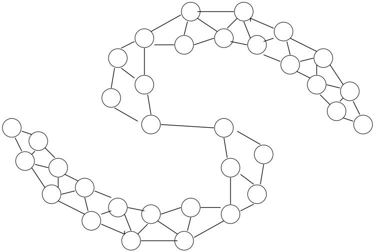
  (Can actually compute, in closed form, the probabilities of $\hat{y}_{m+i} = y$ for each y.)

Example: semi-supervised learning

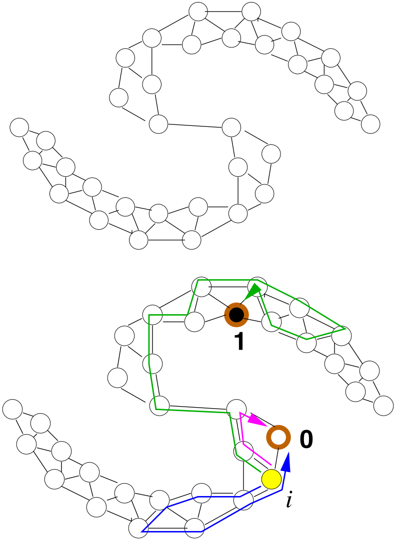




Example: semi-supervised learning



Example: semi-supervised learning



Recap

- Markov property: past and future are conditionally independent given the present.
- Transition matrix: the conditional next-state distributions for each state.
- Random walk on graphs: extremely important process, very well-studied, many applications (including in ML, statistics, etc).
- **Irreducible and aperiodic Markov chains have limiting behavior**: doesn't matter where you start, eventually marginal state distribution is the stationary distribution.

  Some qualities similar to iid processes, some rather different.

  Related to eigenvectors/eigenvalues, computation via power method.
- Forms the basis of PageRank.