# Markov Models

*Presented by:* Wei Fan

School of the Gifted Young
University of Science and Technology of China

2020.12

# Focus

## I. Theories

**I.1. Introduction**

**I.2. Classification of States**

**I.3. Stationary Distribution**

## II. Applications - MCMC

## III. Examples

**III.1. Language Modeling**

**III.2. Google's PageRank Algorithm for web page ranking**

# Theories - Introduction

**For a Stochastic Process, if we assume $X_t$ captures all the information that predicting the future, then we may write the joint distribution as follow:**

$$P(X_{1:T}) = P(X_1)P(X_2|X_1)\cdots P(X_T|X_{T-1}) = P(X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})$$

**This is called a Markov Chain or Markov Model, it has following properties:**

- **Markov Property: $\forall\ t_1 < t_2 < \cdots < t_n < t$, $x_i, x \in S$, we have:**

$$P(X(t) = x|X(t_1) = x_1, X(t_2) = x_2 \cdots X(t_n) = x_n) = P(X(t) = x|X(t_n) = x_n)$$

- **Homogeneous: $\forall\ t_0 < t$, $x, x_0 \in S$, we have:**

$$P(X(t) = x|X(t_0) = x_0) \text{ is only dependent with } t - t_0$$

**For Markov Chain is discrete and has finite state space, we may define:**

- **Transition matrix: Denoted as P, with:**

$$P_{ij} = P(X_{t+1} = j | X_t = i)$$

  **Then the distribution of $X_n$ is uniquely determined by $X_0$ and $P$.**

- **n-step transition matrix: Denoted as $P^{(n)}$, with**

$$P_{ij}^{(n)} = P(X_{t+n} = j | X_t = i)$$

- **We have following equality: $P^{(n)} = P^n$**

**State transition diagram: Represent these Markov chains with a directed graph.**

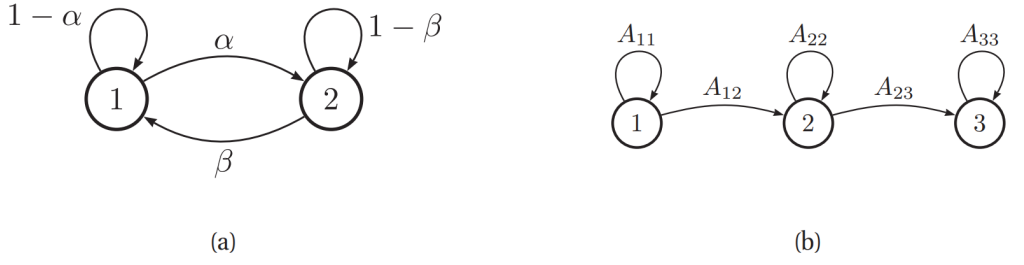
**For example, the following 2-state chain**

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

**is illustrated in the left Figure below, and the following 3-state chain**

$$P = \begin{pmatrix} A_{11} & A_{12} & 0 \\ 0 & A_{22} & A_{23} \\ 0 & 0 & 1 \end{pmatrix}$$

**is illustrated in the right Figure below.**




(a) (b)

**Figure 17.1** State transition diagrams for some simple Markov chains. Left: a 2-state chain. Right: a 3-state left-to-right chain.

# Theories - Classification of States

- **For a Markov chain, we may divide its states into different classes: each class is disjoint and the states in each class is interconnected.**
- **A Markov chain is irreducible if it contains only one class (i.e. we may get from any state to any other state)**
- **A class C is closed if** $P_{jk}^n = 0 \quad \forall j \in C,\ k \notin C,\ n \geq 0$ **(i.e. the chain will not exit class C once entrance).**
- **We may define the period of state i as:** $d(i) = \mathbf{gcd}\{n : P_{ii}^n > 0\}$**. We say a state i is aperiodic if** $d(i) = 1$**.**
- **If state i and j are in the same class, then** $d(i) = d(j)$**.**

**Consider a Markov chain $\{X_n : n \geq 0\}$, we may define:**

$$N_k(n) = \#\{j : X_j = k, 1 \leq j \leq n\}$$

**that is, the number of visits to k before time n. We also define:**

$$f_{jk} = P(N_k(\infty) > 0 | X_0 = j)$$

$$g_{jk} = P(N_k(\infty) = \infty | X_0 = j)$$

**$f_{jk}$ is the probability that the visit to k will occur starting at j, $g_{jk}$ is the probability that the visit to k will occur infinitely starting at j.**

**We may also define:**

$$f_{jk}^n = P(X_n = k, X_j \neq k, j = 0, q, \cdots, n-1 | X_0 = j)$$

**Proposition 1.**

$$f_{jk} = \sum_{n=1}^{\infty} f_{jk}^n, \qquad P_{jk}^n = \sum_{m=1}^{n} f_{jk}^m P_{jk}^{n-m}$$

**Proposition 2.**

$$g_{jk} = f_{jk} g_{kk}, \qquad g_{kk} = \lim_{n \to \infty} (f_{kk})^n$$

**Therefore for each state k, either of the following is true:**

$$g_{kk} = 1 \iff f_{kk} = 1$$

$$g_{kk} = 0 \iff f_{kk} < 1$$

- **We call a state j recurrent if $f_{jj} = 1$.**
- **We call a state j transient if $f_{jj} < 1$.**

**We define $T_{jj}$ to be the time of first return to j, then we have:**

$$\mu_{jj} = ET_{jj} = \begin{cases} \infty & \text{if j is transient} \\ \sum_{n=0}^{\infty} n f_{jj}^n & \text{if j is recurrent} \end{cases}$$

**Then if j is recurrent, we may further classify as:**

- **j is non-null recurrent if $\mu_{jj} < \infty$**
- **j is null recurrent if $\mu_{jj} = \infty$**
- **j is ergodic if it's aperiodic and non-null recurrent.**

**We may think of Markov chain to be a renewal process/delayed renewal process, then we have following conclusion:**

**Proposition 3.**

$$\lim_{n\to\infty} P_{ij}^n = \frac{1}{\mu_{jj}}$$

## Example: Simple random walk

**Consider Stochastic process $\{S_n : n \geq 0\}$ with $S_0 = 0$, $S_n = \sum_{i=1}^{n} X_i$, where $\{X_n : n \geq 0\}$ iid with**

$$P(X_n = 1) = p, \quad P(X_n = -1) = q$$

**Then $S_n$ is an irreducible Markov chain, and we may find out that**

$$P_{00}^{2n} = \binom{2n}{n} p^n q^n \sim \frac{(4pq)^n}{\sqrt{2\pi n}}$$

**Therefore we have**

$$\sum_{n=1}^{\infty} P_{00}^{2n} < \infty \iff p \neq \frac{1}{2}$$

**which shows state 0 is transient iff $p = \frac{1}{2}$**

# Theories - Stationary Distribution

- **Stationary distribution:** $\{\pi_i : i \geq 0\}$ **that satisfies:**

$$\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij} \quad \forall j$$

**Equivalently we may write it as** $\pi = \pi P$**.**

- **Global balance equations:** $\pi_i \sum_{j \neq i} P_{ij} = \sum_{j \neq i} \pi_j P_{ji}$**.**

- **Time reversible: If the chain satisfies detailed balance equations** $\pi_i P_{ij} = \pi_j P_{ji}$ **. Clearly such** $\pi$ **is the stationary distribution.**

**Theorem 1.** *An irreducible aperiodic Markov chain must satisfy one of the following:*
*(i) Every state is transient or null recurrent (i.e.*$P_{ij}^n \to 0,\ \forall i, j$*), and the stationary distribution does not exist.*
*(ii) Every state is non-null recurrent, and*

$$\pi_j = \lim_{n\to\infty} P_{ij}^n \quad \forall i, j$$

*is the unique stationary distribution.*

**Remark 1.** *We can express* $\pi_j$ *as:*

$$\pi_j = \lim_{n\to\infty} \frac{N_j(n)}{n} = \frac{1}{\mu_{jj}}$$

# Focus

## I. Theories

**I.1. Introduction**

**I.2. Classification of States**

**I.3. Stationary Distribution**

## II. Applications - MCMC

## III. Examples

**III.1. Language Modeling**

**III.2. Google's PageRank Algorithm for web page ranking**

# Applications - MCMC

- **Goal: Generate a sample with certain distribution with the stationary distribution of Markov Chain.**

**Theorem 2.** *Given a certain pmf* $\{\pi_i : i \in S\}$*, there exists a time reversible Markov chain* $\{X_n : n \geq 0\}$ *with stationary distribution* $\{\pi_i : i \in S\}$*.*

*Proof.* We suppose that $S = \mathbb{N}$, and there exists a Markov chain $\{Y_n : n \geq 0\}$ with stationary distribution $\mathbf{Q} = (Q_{ij})$ such that:

$$Q_{ij} = 0 \iff Q_{ji} = 0$$

We define $\alpha_{ij}$ as follow

$$\alpha_{ij} = \begin{cases} 1 & Q_{ij} = 0 \\ \min\left\{\frac{\pi_j Q_{ji}}{\pi_i Q_{ij}}, 1\right\} & Q_{ij} \neq 0 \end{cases}$$

Notice that $\pi_i Q_{ij} \alpha_{ij} = \pi_j Q_{ji} \alpha_{ji}$, let

$$P_{ij} = Q_{ij}\alpha_{ij}, \qquad P_{ii} = Q_{ii} + \sum_{j \neq i} Q_{ij}(a - \alpha_{ij})$$

It's easy to check here that $\mathbf{P}$ is a transition matrix with stationary distribution $\{\pi_i : i \geq 0\}$. ∎

**The way used to proof the Theorem is called Hastings-Metropolis Algorithm, we may apply the algorithm to generate random sample with certain distribution.**

## Algorithm - Hastings-Metropolis Sampling

1. **Assign a random value to $X_0$.**
2. **Suppose now we have $X_k = i$.**
3. **Generate a random variable with pmf $\{Q_{ij}, j \geq 0\}$, denoted as j.**
4. **If $\pi_j Q_{ji}/\pi_i Q_{ij} \geq 1$, then we renew the stage to$X_{k+1} = j$ and back to step 2, else we go to step 5.**
5. **We pick a random sample $U \sim U(0,1)$, if $U \leq \pi_j Q_{ji}/\pi_i Q_{ij}$, then we renew the state to $X_{k+1} = j$ and back to step 2, else we don't renew and back to step 2.**

**Now we're going to introduce another way of sampling. We consider a random vector $\mathbf{Z} = (Z_1, Z_2, \cdots Z_n)$ and assume:**

- **The distribution of Z satisfies:**

$$\pi_{\mathbf{z}} = P(\mathbf{Z} = \mathbf{z}) = cg(\mathbf{z}) > 0$$

- $\forall 1 \leq i \leq n$ **and** $z_j : 1 \leq j \leq n$ **we have:**

$$P(Z_i = z_i | Z_j = z_j \text{ for } j \neq i)$$

  **exists and is known.**

**We may apply Gibbs Sampling method to generate a random vector satisfies those conitions.**

## Algorithm - Gibbs Sampling

1. **Assign a random value $\mathbf{y}_0$ to $\mathbf{Y}_0$**
2. **Suppose now we have $\mathbf{Y}_k = \mathbf{y} = (y_1, y_2, \cdots, y_n)$**
3. **Randomly pick a number i in set $\{1, 2, \cdots, n\}$, then we generate a random number wrt to conditional probability $P(Z_i = \cdot | Z_j = y_j, j \neq i)$**
4. **We renew the state to $\mathbf{Y}_{k+1} = (y_1, \cdots, y_{i-1}, z, y_{i+1}, \cdots, y_n)$, back to step 2.**

**Theorem 3.** *The Markov chain generated above is aperiodice and irreducible, and its stationary distribution is the distribution of* $\boldsymbol{Z}$.

*Proof.* For such Markov chain $\{Y_k : k \geq 0\}$, we calculate the transtition matrix $\mathbf{Q}$:

- If $\mathbf{x}$ and $\mathbf{y}$ are different in at least 2 elements: $Q_{x,y} = 0$.
- If $\mathbf{x}$ and $\mathbf{y}$ are only different in element i, we have:

$$Q_{x,y} = \frac{1}{n} P(Z_i = y_i | Z_j = x_j, \forall j \neq i) = \frac{cg(\mathbf{y})}{nP(Z_j = x_j, \forall j \neq i)}$$

- If $\mathbf{x}$=$\mathbf{y}$, then:

$$\begin{aligned} Q_{x,x} &= 1 - \sum_{y \neq x} Q_{x,y} = 1 - \frac{1}{n} \sum_{i=1}^{n} [1 - P(Z_i = x_i | Z_j = x_j, \forall j \neq i)] \\ &= \frac{cg(\mathbf{x})}{n} \sum_{i=1}^{n} \frac{1}{P(Z_j = x_j, \forall j \neq i)} > 0 \end{aligned}$$

This shows that the chain is aperiodic and irreducible. Furthermore: we may see directly that $\pi_x Q_{x,y} = \pi_y Q_{y,x}$ holds $\forall\, \mathbf{x}, \mathbf{y}$. Since that $P_{x,y} = \alpha_{x,y} Q_{x,y}$, we conclude that $\mathbf{P} = \mathbf{Q}$. ∎

# Focus

## I. Theories

**I.1. Introduction**

**I.2. Classification of States**

**I.3. Stationary Distribution**

## II. Applications - MCMC

## III. Examples

**III.1. Language Modeling**

**III.2. Google's PageRank Algorithm for web page ranking**

# Examples - Language Modeling

- Language Models: **Probability distributions over sequences of words.**
- **The state space is defined to be all words in English**
- **The marginal probabilities $P(X_t = k)$ is called unigram statistics.**
- **If we apply a first-order Markov model $P(X_t = k|X_{t-1} = j)$ is called bigram model. And more generally, we have trigram model and n-gram models**

**Language models can be used for following:**

- **Sentence completion**
- **Data compression**
- **Text classification**
- **Automatic essay writing**

## MLE for Markov language models

- **Suppose now we have a set of sequences $\mathcal{D} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)$, with $\mathbf{x}_i = (x_{i1}, \cdots, x_{i,T_i})$ is a sequence of length $T_i$.**

- **Log likelihood:**

$$\log p(\mathcal{D}|\theta) = \sum_j N_j^1 \log \pi_j + \sum_j \sum_k N_{jk} \log P_{jk}$$

**where**

$$N_j = \sum_{i=1}^{N} \mathbf{1}(x_{i1} = j), \quad N_{jk} = \sum_{i=1}^{N} \sum_{t=1}^{T_i - 1} \mathbf{1}(x_{i,t} = j, x_{i,t+1} = k)$$

- **MLE:**

$$\hat{\pi}_j = \frac{N_j^1}{\sum_j N_j^1}, \quad \hat{P}_{jk} = \frac{N_{jk}}{\sum_k P_{jk}}$$

- **In reality, we may apply "add-one smoothing".**

## Empirical Bayes version of deleted interpolation

- **Deleted interpolation: Transition matrix as a convex combination of the bigram frequencies $f_{jk} = N_{jk}/N_j$ and unigram frequencies $f_k = N_k/N$:**

$$P_{jk} = (1 - \lambda) f_{jk} + \lambda f_k$$

- **Prior: Dirichlet Prior**

$$\mathbf{P}_j \sim Dir(\alpha_0 m_1, \cdots, \alpha_0 m_k) = Dir(\boldsymbol{\alpha})$$

- **Posterior: $\mathbf{P}_j \sim Dir(\boldsymbol{\alpha} + \mathbf{N}_j)$, where $\mathbf{N}_j = (N_{j1}, N_{j2}, \cdots, N_{jk})$.**

- **Posterior predictive density:**

$$P(X_{t+1} = k | X_t = j, \mathcal{D}) = \frac{f_{jk} N_j + \alpha_0 m_k}{N_j + \alpha_0} = (1 - \lambda_j) f_{jk} + \lambda_j m_k$$

**where** $\lambda_j = \frac{\alpha_0}{\lambda_j + \alpha_0}$

# Examples - Google's PageRank Algorithm for web page ranking

- **History: Originally is considered as a model of user behaviour, where a user click on links randomly. The factor $1-p$ is the probability that he does not click the link, but jump to a random website. Usually we suppose $p = 0.85$.**

- **Goal: Measure the authoritativeness of each webpage.**

- **Intuition: Use the frequency of clicks to measure authoritativeness.**

- **Method: The model we described above can be transformed to Markov model, and the frequency of clicks are equal to its stationary distribution. All is left is to find the transition matrix M.**

- **The stationary distribution $\{\pi_i : i \geq 0\}$ of the Markov chain satisfies:**

$$\pi_j = \sum_i \pi_i P_{ij}$$

  **where $P_{ij}$ is the probability of following a link from i to j.**

- **In the simplest setting, we define such probability to be same for all links jump from i (including the one jumps to itself), allowing the chain to be aperiodic and irreducible.**

- **In such setting, we take into account both the importance of the linking pages and the outgoing links they have.**

First we have a look at the example below:

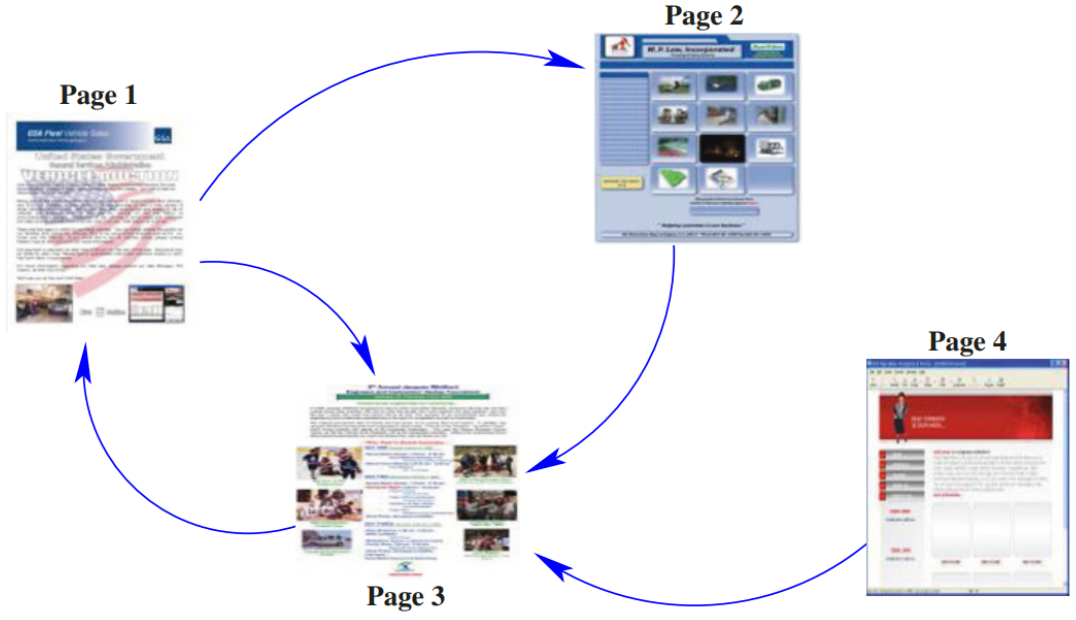



**FIGURE 14.46.** *PageRank algorithm: example of a small network*

It can be calculated that the PageRank distribution is:

$$\boldsymbol{\pi} = (1.49, 0.78, 1.58, 0.15)$$

## Computing the PageRank Vector

- **Let $L_{ij} = 1$ iff $j$ points to page $i$, and zero otherwise. Let $c_j = \sum_{i=1}^{N} L_{ij}$ equal the number of pages pointed to by page $j$ (number of outlinks).**

- **Then the Google PageRanks $\pi_i$ are defined by**

$$\pi_i = (1-p) + p \sum_{j=1}^{N} \left( \frac{L_{ij}}{c_j} \right) \pi_j$$

**where $p$ is a positive constant (apparently set to 0.85), which ensure the each page gets the PageRank for at least $1-p$.**

- **In matrix notation the equality is**

$$\boldsymbol{\pi} = (1-p)\mathbf{e} + p \cdot \mathbf{L}\mathbf{D}_c^{-1}\boldsymbol{\pi}$$

**where e is the vector of N ones, and $\mathbf{D}_c = diag(\mathbf{c})$ is the diagonal matrix with elements $c_j^{-1}$.**

- **Introducing the normalization $\mathbf{e}^T\boldsymbol{\pi} = N$, we can write as**

$$\begin{aligned} \boldsymbol{\pi} &= \left[(1-p)\mathbf{e}\mathbf{e}^T/N + p\mathbf{L}\mathbf{D}_c^{-1}\right]\boldsymbol{\pi} \\ &= \mathbf{A}\mathbf{p} \end{aligned}$$

- **Since matrix A have positive entries in each column summing to 1, it has a unique eigenvector with eigenvalue 1 (i.e. the stationary distribution).**

- **Ways to calculate $\pi$: Power method or Monte Carlo approximation.**

## Web spam

- **PageRank is not foolproof.**

- **JCPenney: planted many links to its homepage on 1000s irrevelent web pages, thus increasing its ranking on Google's search engine.**

- **Even though each of the source pages has low rank, the effect add up.**