

# Stochastic matrices

Richard Yi Da Xu

University of Technology, Sydney

June 5, 2017

# Stochastic matrices

- ▶ **Right stochastic matrix** (or row stochastic matrix) is a real square matrix, with **each row** summing to 1.

$$\begin{bmatrix} K_{1 \rightarrow 1} & \dots & K_{1 \rightarrow n} \\ \dots & \dots & \dots \\ K_{d \rightarrow 1} & \dots & K_{d \rightarrow n} \\ \dots & \dots & \dots \\ K_{n \rightarrow 1} & \dots & K_{n \rightarrow n} \end{bmatrix}$$

- ▶ **Left stochastic matrix** (or column stochastic matrix) is a real square matrix, with **each column** summing to 1

$$\begin{bmatrix} K_{1 \rightarrow 1} & \dots & K_{n \rightarrow 1} \\ \dots & \dots & \dots \\ K_{1 \rightarrow d} & \dots & K_{n \rightarrow d} \\ \dots & \dots & \dots \\ K_{1 \rightarrow n} & \dots & K_{n \rightarrow n} \end{bmatrix}$$

- ▶ **doubly stochastic matrices**: is a real square matrix, where both **each column** and **each row** summing to 1.

# Product of two stochastic matrix is still stochastic

- ▶ Each entry in the product  $AB$  is a dot product of a row from  $A$  and a column from  $B$ .

$$(AB)_{ij} = \sum_{k=1}^n A_{ik} B_{kj}$$

- ▶ We need to prove, for a single row of product  $(AB)_{i,:}$ :

$$\sum_{j=1}^n (AB)_{ij} = \sum_{j=1}^n \sum_{k=1}^n A_{ik} B_{kj} = \sum_{k=1}^n (A_{ik} \sum_{j=1}^n B_{kj})$$

- ▶ Because  $B$  is stochastic,  $\sum_{j=1}^n B_{kj} = 1$
- ▶ Because  $A$  is stochastic,  $\sum_{k=1}^n A_{ik} = 1$

# Perron-Frobenius Theorem:

If  $K$  is a **positive**, left stochastic matrix, then:

- ▶ 1 is an eigenvalue of multiplicity one.
- ▶ 1 is the largest eigenvalue: all the other eigenvalues have absolute value smaller than 1.
- ▶ the eigenvectors corresponding to the eigenvalue 1 have either only positive entries or only negative entries.
- ▶ Note that  $K$  is a **positive** means,  $K_{ij} \geq 0 \ \forall i, j$ . It's NOT **positive definite matrix**

# Power Method Convergence Theorem

- ▶ Let  $K$  be a positive, left (i.e., column) stochastic  $n \times n$  matrix.
- ▶  $\pi^*$  its **probabilistic eigenvector** corresponding to the eigenvalue 1.

$$\begin{bmatrix} K_{1 \rightarrow 1} & \dots & K_{n \rightarrow 1} \\ \vdots & \ddots & \vdots \\ K_{1 \rightarrow n} & \dots & K_{n \rightarrow n} \end{bmatrix} \begin{bmatrix} \pi_1^* \\ \vdots \\ \pi_n^* \end{bmatrix} = \begin{bmatrix} \pi_1^* \\ \vdots \\ \pi_n^* \end{bmatrix}$$

- ▶ Let  $\pi^{(1)}$  be the column vector with all entries equal to some arbitrary stochastic vector.
- ▶ Then sequence  $\{\pi^{(1)}, K\pi^{(1)}, K^2\pi^{(1)}, \dots, K^t\pi^{(1)}, \dots, K^\infty\pi^{(1)}\}$  converges to the vector  $\pi^*$

$$\lim_{t \rightarrow \infty} K^t = K^\infty \implies \lim_{t \rightarrow \infty} K^t \pi^{(1)} = \pi^*$$

- ▶ **Exercise** Generate some random matrix in MATLAB and to show an example of the above.

# Extend to continuous case

- ▶ in the **discrete** case:

$$\begin{bmatrix} K_{1 \rightarrow 1} & K_{2 \rightarrow 1} & \dots & K_{n \rightarrow 1} \\ \vdots & \vdots & \vdots & \vdots \\ K_{1 \rightarrow d} & K_{2 \rightarrow d} & \dots & K_{n \rightarrow d} \\ \vdots & \vdots & \vdots & \vdots \\ K_{1 \rightarrow n} & K_{2 \rightarrow n} & \dots & K_{n \rightarrow n} \end{bmatrix} \begin{bmatrix} \pi_1^* \\ \vdots \\ \pi_d^* \\ \vdots \\ \pi_n^* \end{bmatrix} = \begin{bmatrix} \pi_1^* \\ \vdots \\ \pi_d^* \\ \vdots \\ \pi_n^* \end{bmatrix} \implies \pi_d^* = \sum_{i=1}^n \pi_i^* K_{i \rightarrow d}$$

- ▶ in the **continuous** case, let  $\pi(x)$  be the target distribution:

$$\pi(x^{(n+1)}) = \int_{x_n} \pi(x^{(n)}) K(x^{(n)} \rightarrow x^{(n+1)})$$

- ▶ A transition kernel  $K$  contains element-wise entries:  
 $\{K(x^{(n)} \rightarrow x^{(n+1)})\} \quad \forall x^{(n)}, x^{(n+1)}$
- ▶ Sometimes we prefer to write  $(x^{(n)})$  as  $x$  and  $(x^{(n+1)})$  as  $x^*$ .
- ▶  $K(x \rightarrow x^*)$  is the probability a process at state  $x$  moves to state  $x^*$  in a **one step**
- ▶  $K^n(x \rightarrow x^*)$  is the probability a process at state  $x$  moves to state  $x^*$  in **n steps**

# Power Method Convergence in continuous case

- ▶ One may have first sample  $x^{(1)}$  distributed from an arbitrary distribution:

$$x^{(1)} \sim \pi^{(1)}$$

- ▶ by applying  $K$  function, to obtain  $x^{(2)}$  given  $x^{(1)}$  with probability:

$$\pi^{(2)}(x^{(2)}) = \int_{x^{(1)}} \pi^{(1)}(x^{(1)}) K(x^{(1)} \rightarrow x^{(2)}) dx^{(1)}$$

- ▶ by applying  $K$  function again, to obtain  $x^{(3)}$  with probability:

$$\begin{aligned}\pi^{(3)}(x^{(3)}) &= \int_{x^{(1)}} \int_{x^{(2)}} \pi^{(1)}(x^{(1)}) K(x^{(1)} \rightarrow x^{(2)}) K(x^{(2)} \rightarrow x^{(3)}) dx^{(1)} dx^{(2)} \\ &= \int_{x^{(1)}} \pi^{(1)}(x^{(1)}) \underbrace{\int_{x^{(2)}} K(x^{(1)} \rightarrow x^{(2)}) K(x^{(2)} \rightarrow x^{(3)}) dx^{(2)}}_{K^2(x^{(1)} \rightarrow x^{(3)})} dx^{(1)} \\ &= \int_{x^{(1)}} \pi^{(1)}(x^{(1)}) \underbrace{K^2(x^{(1)} \rightarrow x^{(3)})}_{\rightarrow \text{converge closer to } \pi(x^{(3)})} dx^{(1)}\end{aligned}$$

- ▶ This says,

$$\lim_{t \rightarrow \infty} \pi^{(t)}(x^{(t)}) \rightarrow \pi(x^{(t)})$$

# Burn in samples

- ▶ We know,

$$\lim_{t \rightarrow \infty} \pi^{(t)}(x^{(t)}) \rightarrow \pi(x^{(t)})$$

- ▶ But, in practice,

$$\lim_{t \rightarrow B} \pi^{(t)}(x^{(t)}) \rightarrow \pi(x^{(t)})$$

- ▶  $\{x^{(1)}, \dots, x^{(B)}\}$  are the **burn-in** samples, which we discard.



# What is MCMC research is all about

- ▶ **equilibrium equation:**

$$\pi(x^*) = \int_x \pi(x) K(x \rightarrow x^*) dx$$

- ▶ In machine learning, we always know the expression of stationary distribution  $\pi(x)$ ,
- ▶ Our task is therefore, **find an appropriate**  $K(x \rightarrow x^*)$  to generate samples in a Markov fashion.

# Detailed Balance

- ▶ At equilibrium, that stationary distribution satisfies:

$$\pi(x^*) = \int_x \pi(x) K(x \rightarrow x^*) dx \quad \text{equilibrium equation}$$

- ▶ Proving **equilibrium equation** may be difficult in some cases, therefore, we instead prove detail balance:
- ▶ **detailed balance** condition holds when:

$$\pi(x) K(x \rightarrow x^*) = \pi(x^*) K(x^* \rightarrow x)$$

- ▶ **detailed balance** implies **equilibrium equation**:

$$\begin{aligned} \int_x \pi(x) K(x \rightarrow x^*) dx &= \int_x \pi(x^*) K(x^* \rightarrow x) dx \\ &= \pi(x^*) \int_x K(x^* \rightarrow x) dx \\ &= \pi(x^*) \quad \text{equilibrium equation} \end{aligned}$$

- ▶ the reverse is not always true.

# Extend target distribution with auxiliary variables

- ▶ At equilibrium, that stationary distribution satisfies:

$$\pi(x^*) = \int_x \pi(x) K(x \rightarrow x^*) dx$$

- ▶ under many scenarios, we may have an extended joint density  $(x, u)$ :

$$\pi(x|u)\pi(u)K(u, x \rightarrow u^*, x^*) = \pi(x^*|u^*)\pi(u^*)K(x^*, u^* \rightarrow x, u)$$

- ▶  $u$  is auxiliary variables help sampling
- ▶ one needs to ensure that:

$$\int_u \pi(x, u) du = \pi(x)$$

# Alternative Use of Stochastic Matrix

- ▶ Before dive deep into MCMC algorithms, let's have a look at alternative use of stochastic matrix
- ▶ PageRank algorithm is different to MCMC, in PageRank algorithm:  $K$  is known
- ▶ PageRank algorithm then computes  $\pi$  which is the **invariant distribution**, tells the importance of each web page.

# PageRank algorithm

- ▶ Imagine we have the following four web pages and their links
- ▶ we can then compute the probability of navigating from  $i^{\text{th}}$  page (discrete state) to  $j^{\text{th}}$  page (discrete state)

- ▶ Page 1 links to pages  $\{2, 3\}$

$$\implies K_{1 \rightarrow 1} = 0, K_{1 \rightarrow 2} = \frac{1}{2}, K_{1 \rightarrow 3} = \frac{1}{2}, K_{1 \rightarrow 4} = 0$$

- ▶ Page 2 has links to pages  $\{1, 3, 4\}$

$$\implies K_{2 \rightarrow 1} = \frac{1}{3}, K_{2 \rightarrow 2} = 0, K_{2 \rightarrow 3} = \frac{1}{3}, K_{2 \rightarrow 4} = \frac{1}{3}$$

- ▶ Page 3 has links to pages  $\{1, 3\}$

$$\implies K_{3 \rightarrow 1} = \frac{1}{2}, K_{3 \rightarrow 2} = 0, K_{3 \rightarrow 3} = \frac{1}{2}, K_{3 \rightarrow 4} = 0$$

- ▶ Page 4 has links to pages  $\{2, 3\}$

$$\implies K_{4 \rightarrow 1} = 0, K_{4 \rightarrow 2} = \frac{1}{2}, K_{4 \rightarrow 3} = \frac{1}{2}, K_{4 \rightarrow 4} = 0$$

# Stochastic matrix $K$

- ▶ From the preceding example, **Left stochastic matrix** is:

$$\begin{bmatrix} K_{1 \rightarrow 1} & K_{2 \rightarrow 1} & K_{3 \rightarrow 1} & K_{4 \rightarrow 1} \\ K_{1 \rightarrow 2} & K_{2 \rightarrow 2} & K_{3 \rightarrow 2} & K_{4 \rightarrow 2} \\ K_{1 \rightarrow 3} & K_{2 \rightarrow 3} & K_{3 \rightarrow 3} & K_{4 \rightarrow 3} \\ K_{1 \rightarrow 4} & K_{2 \rightarrow 4} & K_{3 \rightarrow 4} & K_{4 \rightarrow 4} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0 \\ 0 & \frac{1}{3} & 0 & 0 \end{bmatrix}$$

- ▶ From Power Method Convergence Theorem, we know:
  - ▶ sequence  $\{\pi^{(1)}, K\pi^{(1)}, K^2\pi^{(1)}, \dots, K^t\pi^{(1)}, \dots, K^\infty\pi^{(1)}\}$  converges to the vector  $\pi^*$

$$\lim_{t \rightarrow \infty} K^t \pi^{(1)} = \pi^*$$

where  $\pi^*$  is a **probabilistic eigenvector** of  $K$  corresponding to the eigenvalue 1.

- ▶ **Exercise** What is the usefulness of  $\pi^*$  in the setting of web pages?

# Usefulness of $\pi^*$ in the setting of web pages

The **answer** to usefulness of  $\pi^*$  in the setting of web pages is:

- ▶ Shows how **important** each webpage is
- ▶ i.e., regardless of the probabilities of the initial webpage visit:  $\pi^{(1)}$ ,
- ▶  $\pi^{(1)} \rightarrow \pi^*$ , where  $\pi^*(i)$  is the target distribution i.e, the probability that the visit will end up at a web page  $i$ .
- ▶ Note that this is a **reverse problem** of MCMC

# Dangling nodes

- What happens when you have the following  $K$ :

$$\begin{bmatrix} K_{1 \rightarrow 1} & K_{2 \rightarrow 1} & K_{3 \rightarrow 1} & K_{4 \rightarrow 1} \\ K_{1 \rightarrow 2} & K_{2 \rightarrow 2} & K_{3 \rightarrow 2} & K_{4 \rightarrow 2} \\ K_{1 \rightarrow 3} & K_{2 \rightarrow 3} & K_{3 \rightarrow 3} & K_{4 \rightarrow 3} \\ K_{1 \rightarrow 4} & K_{2 \rightarrow 4} & K_{3 \rightarrow 4} & K_{4 \rightarrow 4} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 \end{bmatrix}$$

- Note that 4<sup>th</sup> has no out-going node
- **Exercise** check eigenvector correspond to eigenvalue of 1
- What is the eigenvector correspond to eigenvalue of 1, if we change  $K$  into:

$$\begin{bmatrix} K_{1 \rightarrow 1} & K_{2 \rightarrow 1} & K_{3 \rightarrow 1} & K_{4 \rightarrow 1} \\ K_{1 \rightarrow 2} & K_{2 \rightarrow 2} & K_{3 \rightarrow 2} & K_{4 \rightarrow 2} \\ K_{1 \rightarrow 3} & K_{2 \rightarrow 3} & K_{3 \rightarrow 3} & K_{4 \rightarrow 3} \\ K_{1 \rightarrow 4} & K_{2 \rightarrow 4} & K_{3 \rightarrow 4} & K_{4 \rightarrow 4} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 & \mathbf{1} \end{bmatrix}$$

**Exercise** give reason to why this is so?

- **Exercise** How can we solve this?



# Dangling nodes: what may be the solution?

$$\begin{bmatrix} K_{1 \rightarrow 1} & K_{2 \rightarrow 1} & K_{3 \rightarrow 1} & K_{4 \rightarrow 1} \\ K_{1 \rightarrow 2} & K_{2 \rightarrow 2} & K_{3 \rightarrow 2} & K_{4 \rightarrow 2} \\ K_{1 \rightarrow 3} & K_{2 \rightarrow 3} & K_{3 \rightarrow 3} & K_{4 \rightarrow 3} \\ K_{1 \rightarrow 4} & K_{2 \rightarrow 4} & K_{3 \rightarrow 4} & K_{4 \rightarrow 4} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 \end{bmatrix}$$

- ▶ One simply solution is:

$$\begin{bmatrix} K_{1 \rightarrow 1} & K_{2 \rightarrow 1} & K_{3 \rightarrow 1} & K_{4 \rightarrow 1} \\ K_{1 \rightarrow 2} & K_{2 \rightarrow 2} & K_{3 \rightarrow 2} & K_{4 \rightarrow 2} \\ K_{1 \rightarrow 3} & K_{2 \rightarrow 3} & K_{3 \rightarrow 3} & K_{4 \rightarrow 3} \\ K_{1 \rightarrow 4} & K_{2 \rightarrow 4} & K_{3 \rightarrow 4} & K_{4 \rightarrow 4} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 \end{bmatrix}$$

- ▶ in words, it means any page doesn't have out-link, we assume it has equal probability of visiting entire web.
- ▶ Of course, **data mining** researchers may argue certain web page (having certain properties) may attract higher weights etc.

# Disconnected sub-graphs

- ▶ What happens when you have the following  $K$ :

$$\begin{bmatrix} K_{1 \rightarrow 1} & K_{2 \rightarrow 1} & K_{3 \rightarrow 1} & K_{4 \rightarrow 1} \\ K_{1 \rightarrow 2} & K_{2 \rightarrow 2} & K_{3 \rightarrow 2} & K_{4 \rightarrow 2} \\ K_{1 \rightarrow 3} & K_{2 \rightarrow 3} & K_{3 \rightarrow 3} & K_{4 \rightarrow 3} \\ K_{1 \rightarrow 4} & K_{2 \rightarrow 4} & K_{3 \rightarrow 4} & K_{4 \rightarrow 4} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & \frac{1}{2} \end{bmatrix}$$

- ▶ node  $\{1, 2\}$  and  $\{3, 4\}$  each form a sub-graph.
- ▶ **Exercise** check eigenvector correspond to eigenvalue of 1, also multiplicity of eigenvalue 1
- ▶ **Exercise** How can we solve this?

# Disconnected sub-graphs: what may be the solution?

$$\begin{bmatrix} K_{1 \rightarrow 1} & K_{2 \rightarrow 1} & K_{3 \rightarrow 1} & K_{4 \rightarrow 1} \\ K_{1 \rightarrow 2} & K_{2 \rightarrow 2} & K_{3 \rightarrow 2} & K_{4 \rightarrow 2} \\ K_{1 \rightarrow 3} & K_{2 \rightarrow 3} & K_{3 \rightarrow 3} & K_{4 \rightarrow 3} \\ K_{1 \rightarrow 4} & K_{2 \rightarrow 4} & K_{3 \rightarrow 4} & K_{4 \rightarrow 4} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & \frac{1}{2} \end{bmatrix}$$

- ▶ One solution is to use a convex combination between  $K$  and a square matrix having identical elements  $\frac{1}{n}$ :

$$K' = (1 - p)K + p \left( \frac{1}{n} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \right)$$

- ▶ in words, it means most of the time  $1 - p$ , a surfer will follow links to navigate a page
- ▶ but with probability  $p$ , it will arbitrarily close the current page and go to the new one
- ▶ **Exercise** Prove  $K$  remains a left stochastic matrix

# How to compute the **one hundred billion** dimension eigenvector?

- ▶ starting from the vector (not probabilistic eigenvector),  $x$ :

$$x = [1 \quad 1 \quad \dots \quad 1]^T$$

- ▶ generate the sequence:  $\{x, Kx, K^2x \dots K^tx\}$  until convergence then its is the eigenvectors of  $K$  correspond to eigenvalue of 1, up to a normalisation constant  $c$
- ▶ This is solved using **power method**

# Power method

- ▶ **power method** is used to finding an eigenvector of a square matrix corresponding to the **largest** eigenvalue (in terms of absolute value)
- ▶ for stochastic matrix  $K$  has eigenvalues:

$$1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_n|$$

- ▶ the initial vector:  $x$  as a linear combination of eigenvectors of  $K$ :

$$x = c_1 v_1 + c_2 v_2 + \dots c_n v_n$$

Then,

$$\begin{aligned} Kx &= K(c_1 v_1 + c_2 v_2 + \dots c_n v_n) \\ &= c_1 \underbrace{\lambda_1}_{=1} v_1 + c_2 \lambda_2 v_2 + \dots c_n \lambda_n v_n \quad \text{definition of eigen value/vector} \\ &= c_1 v_1 + c_2 \lambda_2 v_2 + \dots c_n \lambda_n v_n \\ \implies K^2 x &= c_1 v_1 + c_2 \lambda_2^2 v_2 + \dots c_n \lambda_n^2 v_n \\ \implies K^t x &= c_1 v_1 + c_t \lambda_2^t v_2 + \dots c_n \lambda_n^t v_n \end{aligned}$$

$$\lambda_j^k \rightarrow 0 \text{ when } j \geq 2 \implies K^t x \rightarrow c_1 v_1$$

- ▶ The second largest eigen value determines the convergence
- ▶ **Homework** Perform the following simulations:
  - ▶ generate lots of  $K$  and choose one which has **large** second eigen values in absolute value
  - ▶ also generate a  $K$  which has **small** second eigen values in absolute value
  - ▶ in both cases, try to compute the sequence  $\{x, Kx, K^2x \dots K^tx\}$ , using an arbitrary vector  $x$
- ▶ **Homework** Generate  $K$  from known eigen value/vectors are called Inverse Eigenvalue Problems. Use IEP to generate stochastic matrices above
- ▶ Try something like, “Doubly Stochastic Matrices with Prescribed Positive Spectrum”

# Reversible Jump MCMC

- ▶ the problem setting is:

$$\Pr(\mathbb{M}_k) = \pi_k$$

$$\sum_k \pi_k = 1$$

$$\theta | \mathbb{M}_k \sim p_k(\theta)$$

$$\mathcal{L}_k(y_1, \dots, y_n | \theta) = \prod_{i=1}^n l_k(y_i | \theta)$$

- ▶  $p_k(\theta | \mathcal{D}) =$

# Reversible Jump MCMC (2)

- ▶ the problem setting is:

$$\Pr(\mathbb{M}_k) = \pi_k$$

$$\sum_k \pi_k = 1$$

$$\theta | \mathbb{M}_k \sim p_k(\theta)$$

$$\mathcal{L}_k(y_1, \dots, y_n | \theta) = \prod_{i=1}^n l_k(y_i | \theta)$$

- ▶  $p_k(\theta | \mathcal{D}) =$