CS 340 Lec. 11: Markov chains, Linear Algebra and PageRank

AD

January 2011

Independence and Conditional Independence of Random Variables

- Consider r.v. $X_1, X_2, ..., X_n$ with a joint p.m.f. $p_{X_1,...,X_n}(x_1, ..., x_n)$ then these variables are called independent if and only if

$$p_{X_1,...,X_n}(x_1, ..., x_n) = \prod_{i=1}^{n} p_{X_i}(x_i).$$

- Consider r.v. $X_1, X_2, ..., X_n$ with a joint p.m.f. $p_{X_1,...,X_n}(x_1, ..., x_n)$ then these variables are called independent upon Y if and only if

$$p_{X_1,...,X_n|Y}(x_1, ..., x_n|y) = \prod_{i=1}^{n} p_{X_i|Y}(x_i|y).$$

- *Example:* $Y \in \{0, 1\}$ indicates spam/non spam and $X_i \in \{0, 1\}$ indicates whether a prespecified word appears in the email.

Markov Chains

- Consider r.v. $X_1, X_2, ..., X_n$ with a joint p.m.f. $p_{X_1,...,X_n}(x_1, ..., x_n)$ then we always have

$$p_{X_1,...,X_n}(x_1, ..., x_n) = p_{X_1}(x_1) \prod_{k=2}^{n} p_{X_k|X_1,...,X_{k-1}}(x_k | x_1, ..., x_{k-1}).$$

- A sequence of r.v. $\{X_k\}_{k\geq 1}$ is said to have the Markov property if and only if

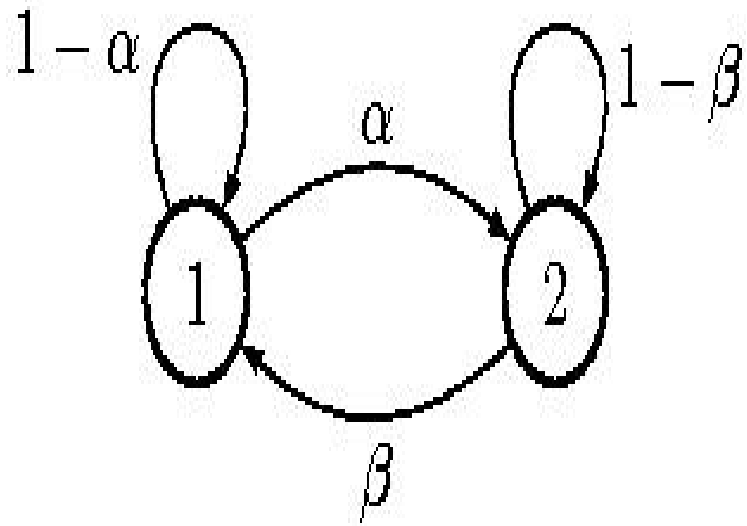
$$p_{X_k|X_1,...,X_{k-1}}(x_k | x_1, ..., x_{k-1}) = p_{X_k|X_{k-1}}(x_k | x_{k-1});$$

  i.e. the conditional distribution of X_k only depends on $(X_1, X_2, ..., X_{k-1})$ through X_{k-1}; in other words X_k and $(X_1, X_2, ..., X_{k-2})$ are conditionally independent given X_{k-1}.
- Markov models are ubiquitous models for time series in Machine learning, EE, Finance etc.

Markov Chains

- Let $X_k \in \{1, 2\}$; e.g. no rain/rain for day k.
- Assume $p_{X_k|X_{k-1}}(x_k|x_{k-1}) = p(x_k|x_{k-1})$, this is an homogeneous Markov chain then we introduce T the transition matrix such that $T_{i,j} = p(j|i)$

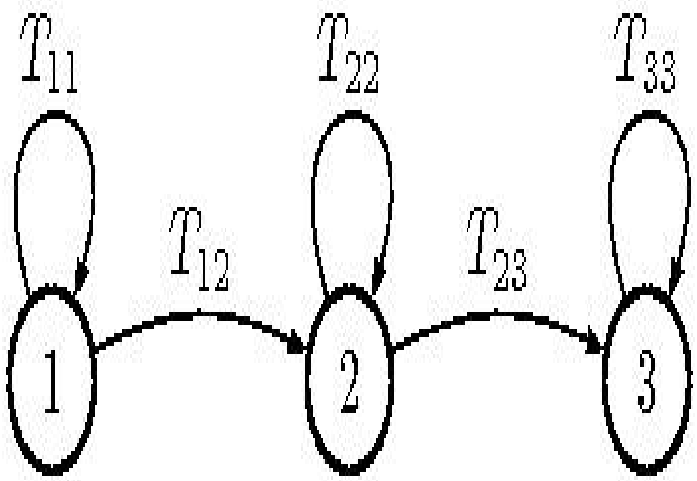
$$T = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}.$$



- T is called a stochastic or Markov transition matrix.

Applications of Markov Chains

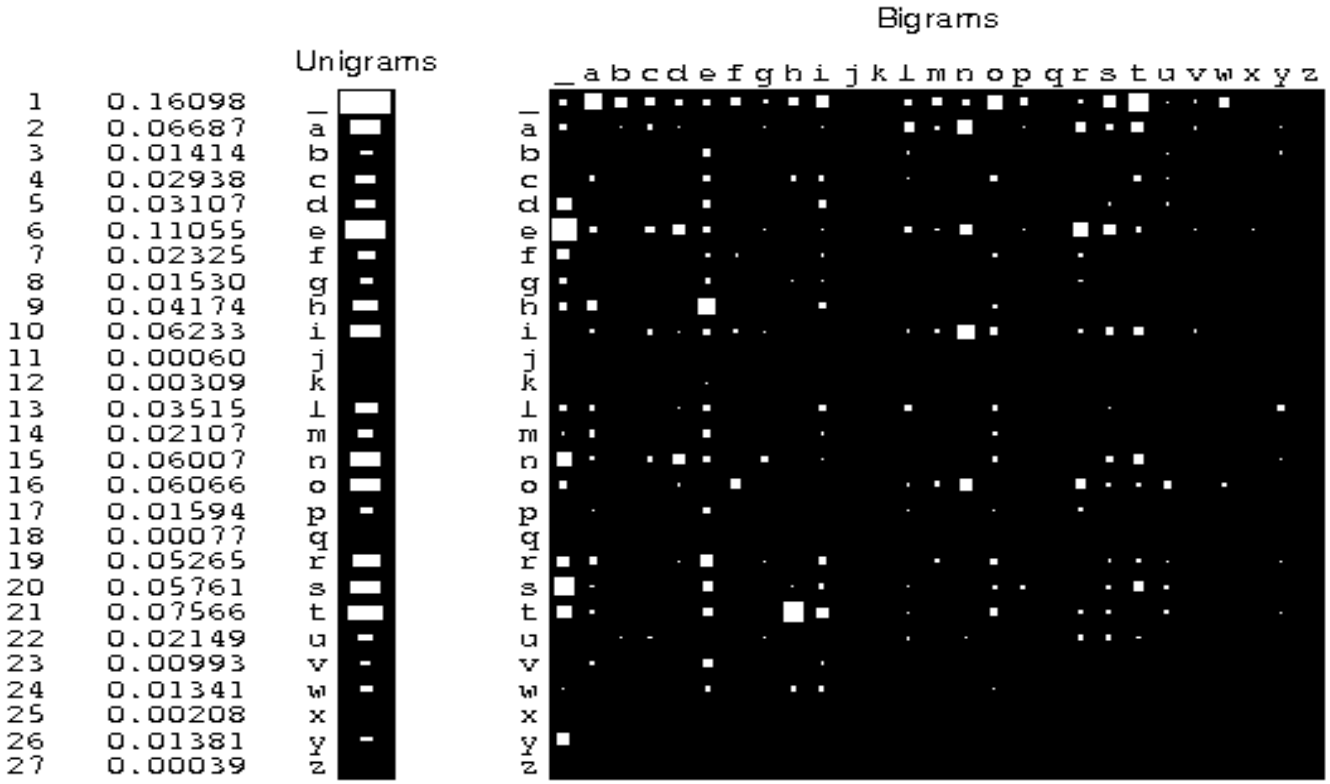
- Left-to-Right Markov chains (used for speech recognition, segmentation etc.)



- DNA Sequencing and Alignment: DNA sequence (ADTTGACATTG....)
- Global optimization via simulated annealing.

Applications of Markov Chains

- Language Modelling: X_k corresponds to a word or a letter in english.



Left: Proba of observing a letter, Right: Proba of observing one letter having just observed another one. The size of the white squares is proportional to the values of the entry. (Based on Darwin's The Origin of Species)

Surfing the Web Randomly

- Let $\mathcal{X} = \{1, 2, ..., n_x\}$ be the Web consisting of n_x webpages ($n_x > 10^{10}$).
- Consider you are surfing the Web randomly and let $X_k \in \mathcal{X}$ be the index of the k^{th} you have visited.
- Whenever you are at a Webpage, you select one of the outbounds links randomly.
- This is a Markov process.

Chapman-Kolmogorov Equation

- **Problem**: We want to compute recursively in time $p_{X_k}(x_k)$ for any $k > 2$ given $p_{X_1}(x_1)$ and $p_{X_k|X_{k-1}}(x_k|x_{k-1}) = p(x_k|x_{k-1})$.
- Chapman-Kolmogorov equation:

$$\begin{aligned} p_{X_k}(x_k) &= \sum_{x_{k-1}\in\mathcal{X}} p_{X_{k-1},X_k}(x_{k-1}, x_k) \\ &= \sum_{x_{k-1}\in\mathcal{X}} p_{X_k|X_{k-1}}(x_k|x_{k-1})\, p_{X_{k-1}}(x_{k-1}) \\ &= \sum_{x_{k-1}\in\mathcal{X}} p(x_k|x_{k-1})\, p_{X_{k-1}}(x_{k-1}) \end{aligned}$$

- Assume X_k takes values in $\mathcal{X} = \{1, 2, ..., n_x\}$ and let

$$\boldsymbol{\pi}_k = \left(p_{X_k}(1) \;\; p_{X_k}(2) \;\; \cdots \;\; p_{X_k}(n_x)\right)^{\mathsf{T}}$$

then Chapman-Kolmogorov can be rewritten as

$$\boldsymbol{\pi}_k^{\mathsf{T}} = \boldsymbol{\pi}_{k-1}^{\mathsf{T}} T \Longleftrightarrow \boldsymbol{\pi}_k = T^{\mathsf{T}} \boldsymbol{\pi}_{k-1}$$

Chapman-Kolmogorov Equation

- Hence, it follows directly that

$$\boldsymbol{\pi}_n = \left(T^{\mathsf{T}}\right)^{n-k} \boldsymbol{\pi}_k$$

  and in particular $\boldsymbol{\pi}_k = \left(T^{\mathsf{T}}\right)^{k-1} \boldsymbol{\pi}_1$.
- One important question is what happens as $k \longrightarrow \infty$. Do we have a "limiting" distribution? i.e. do we have

$$\lim_{k \longrightarrow \infty} \boldsymbol{\pi}_k = \boldsymbol{\pi} \ ?$$

  and, if this limit exists, what is its expression?
- For the two-state example described earlier

$$\lim_{k \longrightarrow \infty} \boldsymbol{\pi}_k \text{ does not exist for } \alpha = \beta = 1$$

  as, if $\boldsymbol{\pi}_1 = (\gamma \quad 1-\gamma)^{\mathsf{T}}$ then $\boldsymbol{\pi}_{2k} = (1-\gamma \quad \gamma)^{\mathsf{T}}$ and $\boldsymbol{\pi}_{2k+1} = \boldsymbol{\pi}_1$.
- For the three-state example, we clearly have for $T_{1,2} > 0$ and $T_{2,3} > 0$

$$\boldsymbol{\pi} = (0 \quad 0 \quad 1)^{\mathsf{T}}$$

Existence of a Limiting Distribution

- If a limit π exists, then it has to satisfies

$$\pi = T^{\mathsf{T}} \pi$$

and it is called the stationary of invariant distribution of the Markov chain. Clearly if π exists then it is an eigenvector of T^{T} associated the eigenvalue $\lambda = 1$.

- **Proposition**: Any stochastic matrix T admits 1 as eigenvalue. Hence T^{T} admits 1 as an eigenvalue.
- *Proof*: We have $\sum_j T_{i,j} = 1$ for any i so for $\mathbf{e} = (1\ 1\ \cdots\ 1)^{\mathsf{T}}$

$$T\ \mathbf{e} = \mathbf{e}$$

and 1 is an eigenvalue of T. As $\det(T - I) = 0 = \det\left((T - I)^{\mathsf{T}}\right) = \det\left(T^{\mathsf{T}} - I\right)$ then 1 is an eigenvalue of T^{T}.

Non-uniqueness of The Eigenvector associated to one

- π is not necessarily unique; e.g. think of two non-communicating sets of states

$$T = \begin{pmatrix} 0.85 & 0.15 & 0.00 & 0.00 \\ 0.50 & 0.50 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.70 & 0.30 \\ 0.00 & 0.00 & 0.15 & 0.85 \end{pmatrix}$$

  then the eigenvalue 1 of T^{T} has two associated eigenvectors

$$\begin{pmatrix} 0.77 & 0.23 & 0.00 & 0.00 \end{pmatrix}^{\mathsf{T}},$$
$$\begin{pmatrix} 0.00 & 0.00 & 0.33 & 0.67 \end{pmatrix}^{\mathsf{T}}.$$

- An even simpler counterexample, think of $T = I$.

The Other Eigenvalues

- **Proposition**: All the eigenvalues $\{\lambda_i\}$ of T, equivalently of T^{T}, satisfy $|\lambda_i| \leq 1$.
- *Proof*: Assume $\mathbf{u}$ is an eigenvector of T associated to λ then

$$T\mathbf{u} = \lambda\mathbf{u} \Leftrightarrow \sum_j T_{i,j} u_j = \lambda u_i.$$

Let select $i_{\max}$ such that $|u_{i_{\max}}|$ is the largest of the components $|u_j|$'s then

$$\begin{aligned} \sum_j T_{i_{\max},j} u_j &= \lambda u_{i_{\max}} \Rightarrow \left| \sum_j T_{i_{\max},j} u_j \right| = |\lambda| \, |u_{i_{\max}}| \\ &\Rightarrow \sum_j T_{i_{\max},j} \, |u_j| \geq |\lambda| \, |u_{i_{\max}}| \end{aligned}$$

However we have by definition of $|u_{i_{\max}}|$

$$\sum_j T_{i_{\max},j} \, |u_j| \leq \sum_j T_{i_{\max},j} \, |u_{i_{\max}}| \leq \left(\sum_j T_{i_{\max},j} \right) |u_{i_{\max}}| = |u_{i_{\max}}| \, .$$

Stationary Distribution as an Eigenvector

- **Perron-Froebenius Theorem.** If there exists $k > 0$ such that

$$\Pr\left(X_k = j \middle| X_1 = i\right) > 0$$

for all i, j; i.e. $\left(T^{\mathsf{T}}\right)^{k-1}$ is a matrix with strictly positive entries then $\boldsymbol{\pi}$ is unique and whatever being $\boldsymbol{\pi}_1$

$$\lim_{k \to \infty} \boldsymbol{\pi}_k = \boldsymbol{\pi}$$

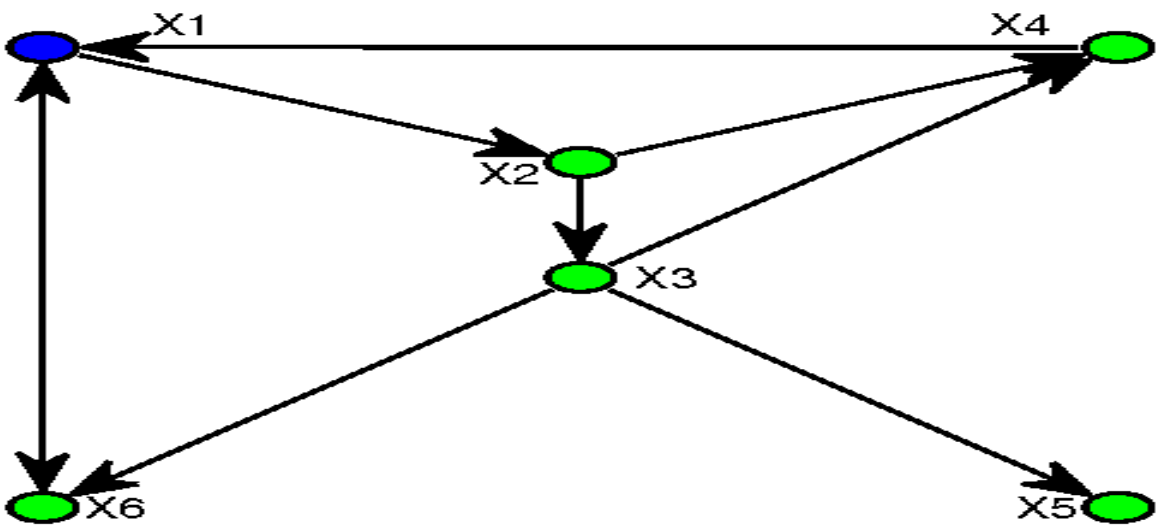
with $\boldsymbol{\pi}(i) > 0$ and $\sum_i \boldsymbol{\pi}(i) = 1$.

- In layman's terms, whatever your initial distribution you will eventually has $\Pr(X_k = i) \approx \boldsymbol{\pi}(i)$ for large k.

Application to Google PageRank

- When one searches for a webpage using a search engine, the system find all the web pages containing the query terms that you specified.
- There are often far too many matches, so the system has to estimate the relevance of each page, it needs to rank them.
- A key idea of Google in the late 90's was to propose a revolutionary approach to ranking known as PageRank.
- There are two equivalent ways to present it
  - it is a system where the importance $\pi(i)$ of each webpage i is made to be proportional to the sum of the importances of all the sites that link to it (with $\pi(i) \geq 0$ and $\sum_i \pi(i) = 1$).
  - if a random surfer was exploring the web, then in the long run he/she will end up on webpage i with proba $\pi(i)$.

A Simplified World Wide Web



- We introduce an adjency matrix G of size $n \times n$ where n is the number of webpages. G is defined by

$$G_{i,j} = \begin{cases} 1 & \text{if outbound link from } i \text{ to } j, \\ 0 & \text{otherwise.} \end{cases}$$

- In this case, a random surfer has a transition matrix

$$T_{i,j} = \begin{cases} G_{i,j} / \left(\sum_j G_{i,j}\right) & \text{if } \sum_j G_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$

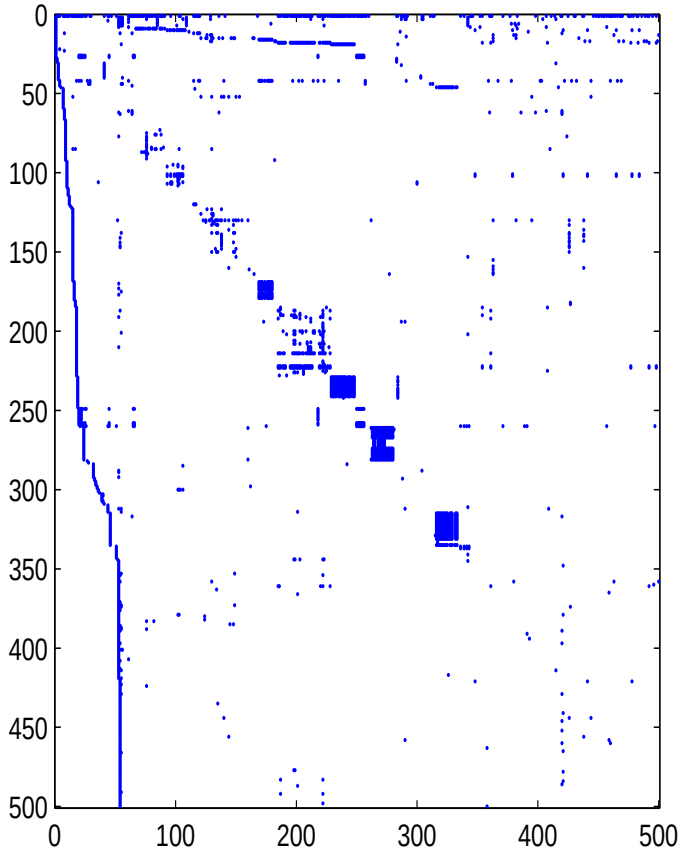
Adding Some Noise

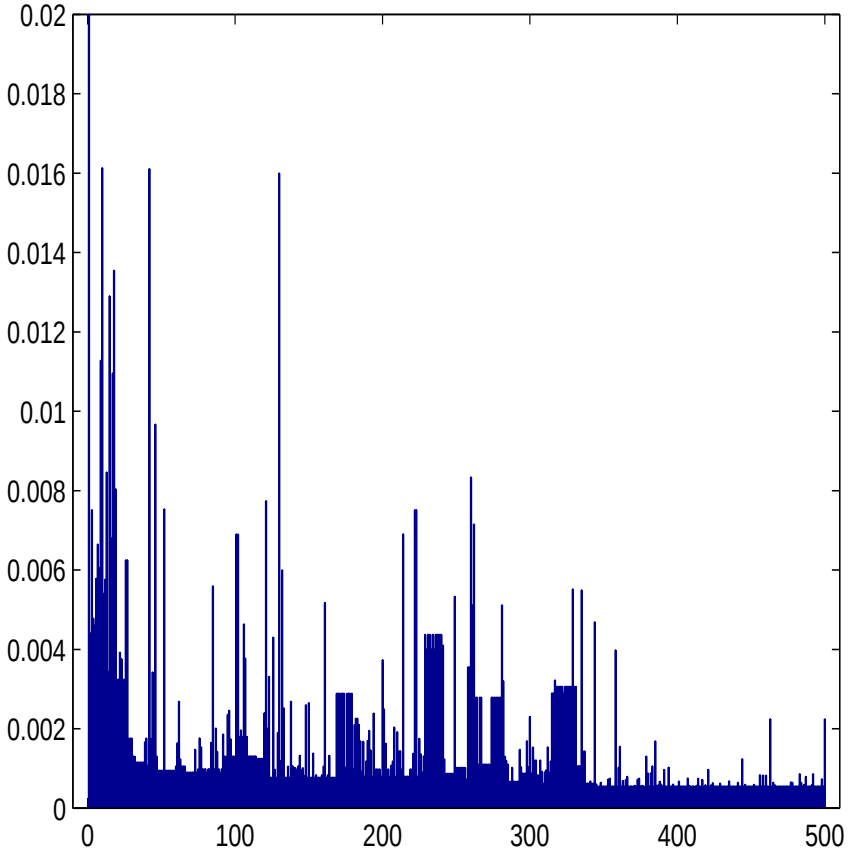
- Clearly, there are two absorbing states '5' and '6': not good!
- To avoid getting trapped, Google considers

$$T_{i,j} = \begin{cases} p \times G_{i,j} / \left(\sum_j G_{i,j}\right) + (1-p) \times 1/n & \text{if } \sum_j G_{i,j} > 0 \\ 1/n & \text{otherwise} \end{cases}$$

  where typically $p = 0.85$.
- In this context, we have $T_{i,j} > 0$ for all i, j so $\lim_{k \to \infty} \boldsymbol{\pi}_k = \boldsymbol{\pi}$ whatever being $\boldsymbol{\pi}_1$.

Example





Left: Adjency matrix, Right: PageRank.

How to Compute the Invariant Distribution

- To compute the invariant distribution, any algorithm to compute eigenvectors can be used. However, we have here $n > 10^9$ so some specific methods have to be developed!
- A simple so-called Monte Carlo approach consists of simulating the Markov chain a very long time and to say

$$\pi(i) \approx \frac{1}{P} \sum_{k=1}^{P} \mathbb{I}(X_k = i)$$

- A law of large numbers hold for this (dependent) process. Such approaches are the basis of Markov chain Monte Carlo methods.

The Power Method

- A powerful method to compute the largest eigenvector consists of the following method.
- Select $\mathbf{v}_0 \in \mathbb{R}^n$ a column vector and iterate for $k \geq 1$

$$\begin{aligned} \mathbf{w}_k &= \left(T^{\mathsf{T}}\right) \mathbf{v}_{k-1}, \\ \mathbf{v}_k &= \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|}. \end{aligned}$$

- For $a_1 = \mathbf{v}_0^{\mathsf{T}} \boldsymbol{\pi} \neq 0$, we have

$$\lim_{k \to \infty} \mathbf{v}_k = \operatorname{sgn}(a_1) \ \boldsymbol{\pi}$$

where $\operatorname{sgn}(x) = 1$ if $x > 0$ and $\operatorname{sgn}(x) = -1$ if $x < 0$.

Outline of the Proof

- Consider the eigenvalues $\{\lambda_i\}$ of T^{T} ordered such that $\lambda_1 = 1 > |\lambda_2| \geq \cdots \geq |\lambda_n|$ with associated orthonormal eigenvectors $\mathbf{u}_1 = \boldsymbol{\pi}, \mathbf{u}_2, ..., \mathbf{u}_n$. We can rewrite

$$\mathbf{v}_0 = a_1 \ \boldsymbol{\pi} + \sum_{i=2}^{n} a_i \ \mathbf{u}_i$$

- Rescaling operation $\mathbf{v}_k = \mathbf{w}_k / \left\|\mathbf{w}_k\right\|$, just ensure that $\left\|\mathbf{v}_k\right\| = 1$ for any k; i.e. we can write alternatively

$$\mathbf{v}_k = \frac{\left(T^{\mathsf{T}}\right)\mathbf{v}_{k-1}}{\left\|\left(T^{\mathsf{T}}\right)\mathbf{v}_{k-1}\right\|} = \frac{\left(T^{\mathsf{T}}\right)^k \mathbf{v}_0}{\left\|\left(T^{\mathsf{T}}\right)^k \mathbf{v}_0\right\|}$$

- We have

$$\begin{aligned}\left(T^{\mathsf{T}}\right)\mathbf{v}_0 &= a_1 \ \left(T^{\mathsf{T}}\right)\boldsymbol{\pi} + \sum_{i=2}^{n} a_i \ \left(T^{\mathsf{T}}\right)\mathbf{u}_i \\ &= a_1\boldsymbol{\pi} + \sum_{i=2}^{n} a_i \ \lambda_i \mathbf{u}_i\end{aligned}$$

Outline of the Proof

- Iterating, we obtain

$$\begin{aligned}\left(T^{\mathsf{T}}\right)^{k}\mathbf{v}_0 &= a_1\boldsymbol{\pi}+\sum_{i=2}^{n} a_i\ \lambda_i^k\mathbf{u}_i \\ &\approx a_1\boldsymbol{\pi} \text{ for large } k\end{aligned}$$

  as $1>|\lambda_2|\geq\cdots\geq|\lambda_n|$. Hence we have

$$\left\|\left(T^{\mathsf{T}}\right)^{k}\mathbf{v}_0\right\|\approx|a_1| \text{ for large } k$$

- It follows that

$$\mathbf{v}_k=\frac{\left(T^{\mathsf{T}}\right)^{k}\mathbf{v}_0}{\left\|\left(T^{\mathsf{T}}\right)^{k}\mathbf{v}_0\right\|}\approx\operatorname{sgn}\left(a_1\right)\ \boldsymbol{\pi} \text{ for large } k$$

  and the convergence is geometric with rate $|\lambda_2|$.

Geometric Rate of Convergence

- Convergence of PageRank is geometric.
- We have for large k

$$\|\mathbf{v}_k - \boldsymbol{\pi}\| \approx \left|\frac{a_2}{a_1}\right| |\lambda_2|^k$$

- $1 - |\lambda_2|$ is known as the spectral gap: the larger the faster the convergence.
- How to estimate λ_2 without knowing π? See assignment.

Outline of the Proof

- Without loss of generality, consider $a_1 > 0$ so that $\text{sgn}(a_1) = 1$ then
- We want to study $\|\mathbf{v}_k - \boldsymbol{\pi}\|$ where

$$\begin{aligned}
\mathbf{v}_k &= \frac{\left(T^{\mathsf{T}}\right)^k \mathbf{v}_0}{\left\|\left(T^{\mathsf{T}}\right)^k \mathbf{v}_0\right\|} \\
&= \frac{a_1\boldsymbol{\pi} + \sum_{i=2}^n a_i\ \lambda_i^k \mathbf{u}_i}{\left\|a_1\boldsymbol{\pi} + \sum_{i=2}^n a_i\ \lambda_i^k \mathbf{u}_i\right\|} \\
&\approx \boldsymbol{\pi} + \frac{a_2}{|a_1|}\ \lambda_2^k \mathbf{u}_2
\end{aligned}$$

so

$$\|\mathbf{v}_k - \boldsymbol{\pi}\| \approx \left\|\frac{a_2}{a_1}\ \lambda_2^k \mathbf{u}_2\right\| = \left|\frac{a_2}{a_1}\right| |\lambda_2|^k$$