# MCMC

CS B553
Spring 2013

## Announcements

- A4 posted soon
- Another quiz soon

## Final project

- Choose a topic of interest to you, related to probabilistic approaches (graphical models)
  - Option 1: Choose a research paper that applies probabilistic approaches to a problem. Re-implement, improve, and/or validate their results.
  - Option 2: Apply probabilistic approaches to some new problem of interest to you.
- Work alone or in partnerships
- Deliverables
  - Project proposal: ~~Friday March 22~~ Monday March 25
  - Interim report: Monday April 8
  - Project presentation: Week of April 22
  - Project report and code: Wednesday May 1

## Markov Chain Monte Carlo (MCMC)

- General class of techniques that produce a *sequence* of samples
- Main idea: Save effort by using information from *past samples* in producing *future samples*
  - Initial samples are from a proposal distribution Q
  - Subsequent sampling is biased towards P
  - Eventually the samples are drawn from a distribution that is closer and closer to P

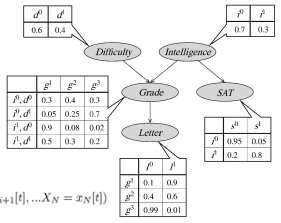
## Example: Gibbs sampling

- Generate initial sample *x[0]*
- For each sample *t=1…T*
  - Let *x[t] = x[t-1]*
  - For each unobserved variable $X_i$,
    - Sample a value for $X_i$ given values for all other variables in *x[t];* i.e. sample from:

$$P(X_i|X_1 = x_1[t], \ldots X_{i-1} = x_{i-1}[t], X_{i+1} = x_{i+1}[t], \ldots X_N = x_N[t]) = P(X_i|\mathbf{X}_{-i} = \mathbf{x}_{-i}[t])$$

where $\mathbf{X}_{-\mathbf{i}} = \mathbf{X} - \{X_i\}$

- Put this sampled value in $x_i[t]$

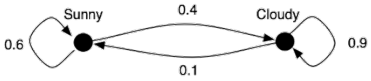



## Properties of Gibbs sampling

- Gibbs can be applied to Markov or Bayes networks
  - Unlike forward sampling and importance sampling, which can in general only be applied to Bayes nets
- Gibbs sampling will converge to sampling from the correct distribution, ***eventually***
  - Under weak assumptions (that the clique potential functions are positive)
  - But may require a long time to converge
  - Why does this happen?

## Markov chains

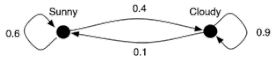
- Stochastic process model
  - Due to Andrey Markov (1906)
  - e.g.,




- The Markov assumption:
  - The probability of transitioning to each new state depends *only* on the current state (and not on the prior states)
  - More formally,

$$P(Q_{t+1} = q_{t+1} | Q_t = q_t, Q_{t-1} = q_{t-1}, ..., Q_0 = q_0) = P(Q_{t+1} = q_{t+1} | Q_t = q_t)$$

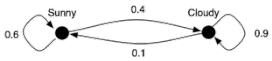
## Markov chains




- What is $P(Q_t=q)$?
- This can be written more compactly with matrix notation.
  - How?

8

## Markov chains




- Suppose there's an 80% chance of sun on day 0. What is the probability of sun on day 3?

$$B = \begin{bmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{bmatrix} \qquad w = \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix}$$

$$(B^T)^3 w = \begin{bmatrix} 0.275 \\ 0.725 \end{bmatrix}$$

$P(X_3 = \text{sun})$

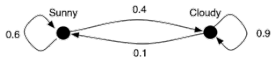
$P(X_3 = \text{cloudy})$

What's the probability of sun on day #1 million?

What's the probability of sun on day #1 million and 1?

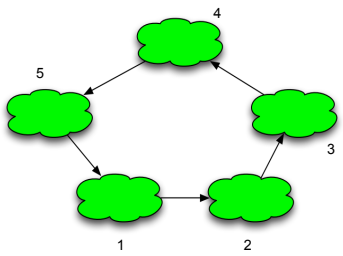
9

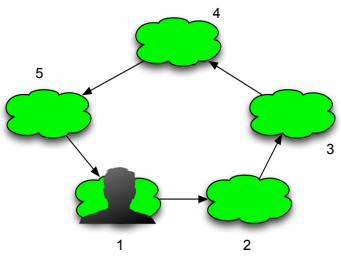
## Stationary distributions

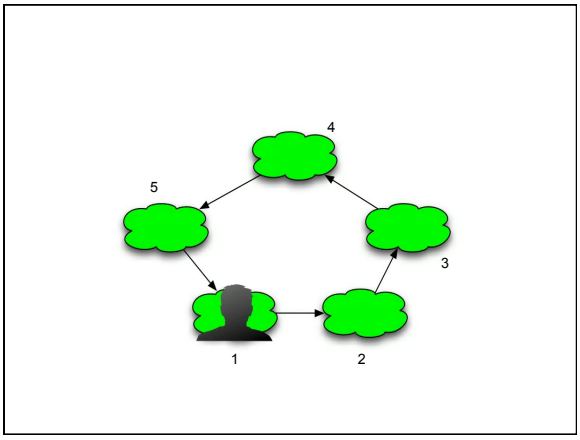


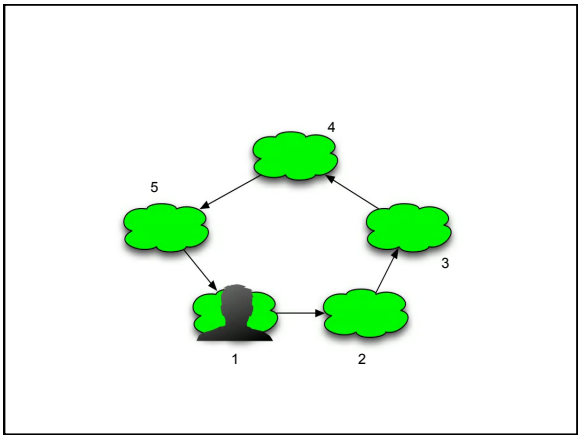

- The *stationary distribution* of a Markov Chain is the distribution over states after a large number of steps
  - Intuitively, after a while, it doesn't matter which time step you're on or which state you started in

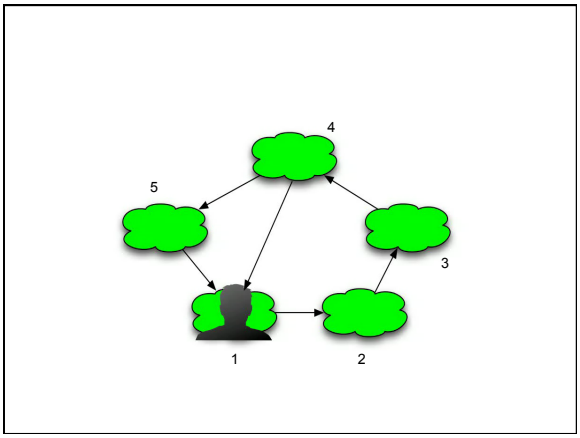
10

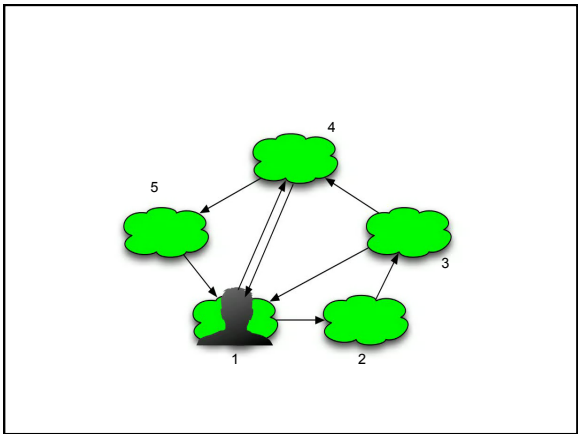


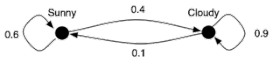


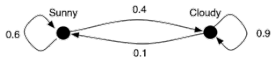

## Stationary distributions




- The *stationary distribution* of a Markov Chain is the distribution over states after a large number of steps
  - Intuitively, after a while, it doesn't matter which time step you're on or which state you started in
- Does a stationary distribution always exist?



## Stationary distributions

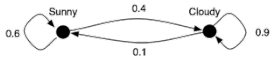



- For an *ergodic* chain, a *stationary distribution* exists
  - ergodic: all states are recurrent and aperiodic
  - stationary distribution: for large *t*, the probability of being in state *i* at time *t* depends *only* on the transition probabilities
  - the stationary distribution $\pi$ is the vector satisfying

$$B^T\pi = \pi$$

How do we compute $\pi$?

## Stationary distribution of Markov chain




- What is the stationary distribution of this chain?

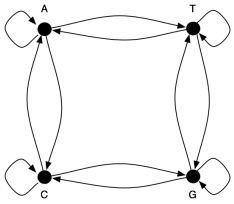
```
>> % e.g. in Matlab:

>> [v d]=eigs([0.6 0.4; 0.1 0.9]',1)

v =
  -0.24253562503633
  -0.97014250014533

d =
     1

>> v/sum(v)

ans =

   0.20000000000000
   0.80000000000000
```

$$\pi = \begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix}$$



## Application: bioinformatics

- Markov chains are used to model biological sequences
  - e.g. peptide

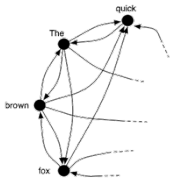





## Application: language modeling

- A sentence is just a sequence of words
  - which we can model as a sequence of states.
- Sentence generation can be modeled as a Markov chain

The quick brown fox jumps over the lazy dog.






## Automatic sentence generation

- Random walks on the Markov chain produce sentences!
  - e.g. using a model trained on an essay of Jean Baudrillard, "The Precession of Simulacra"

> If we were to revive the fable is useless. Perhaps only the allegory of simulation is unendurable--more cruel than Artaud's Theatre of Cruelty, which was the first to practice deterrence, abstraction, disconnection, deterritorialisation, etc.; and if it were our own past. We are witnessing the end of the negative form. But nothing separates one pole from the very swing of voting "rights" to electoral "duties" that the disinvestment of the revolutionary and total strike collapses at the real and its object, as Castaneda does, etc., and to escape the spectre raised by simulation--namely that truth, reference and objective causes have ceased to exist.

By "Mark V. Shaney" and Rob Pike, 1989



## Automatic sentence generation

- Random walks on the Markov chain produce sentences!
  - e.g. using a model trained on poetry

> He was a light, slow, and there is a small Saturn -- away from a high flame lying in the life within it, a new dune, we are formations of caterpillars, we are formations of craziness to innocent, and as it moves it is complete different than the rising face, the cold water, even we can't see infinity is an ocean of downy treasure the welldeep pleasure of caterpillars, we are formations of the world, and what it with the ecstasy of the day is an iceberg we find ourselves on a caress mingled with sleep kill me its lights bands of subjective experience, and wonder why I had dirt a star-crystal-flower plants, made the dragon. Its neck was a novel entitled "Kaleidoscope Vision," which is hat crinkle were like fresh glass domain key - you become someone mentioned them and build in. We see the white my own rising and thunder clapping in the singularity of it, evaporating into a tree, like a long before shade.

By "Mark V. Shaney" and Justin McHale, 1994



## Automatic sentence generation

- Random walks on the Markov chain produce sentences!
  - e.g. using a model trained on postings from alt.singles

> When I meet someone on a professional basis, I want them to shave their arms. While at a conference a few weeks back, I spent an interesting evening with a grain of salt. I wouldn't dare take them seriously! This brings me back to the brash people who dare others to do so or not. I love a good flame argument, probably more than anyone....

By "Mark V. Shaney" and Rob Pike et al, 1984

## Practical uses?

- Generating spam
- SCIgen: Generating scientific papers?!
  - "Rooter: A Methodology for the Typical Unification of Access Points and Redundancy," *11th World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI),* 2005.
  - "I/O Automata No Longer Considered Harmful," *3rd International Symposium of Interactive Media Design*, 2005.
  - Cooperative, Compact Algorithms for Randomized Algorithms, *Applied Mathematics and Computation* (accepted but eventually rescinded)

See http://pdos.csail.mit.edu/scigen/

25

## PageRank

- Basic idea: assign a score to every web page
  - The score represents importance or quality of the page
- The structure of the web can help
  - High-quality pages are linked to by high quality pages

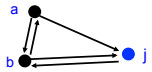
26

## Computing page quality score

- How do we compute the score for a page?
  - Sum up the votes cast by its incoming links

$$r_j = \sum_{i|i\to j} \frac{r_i}{d_i}$$

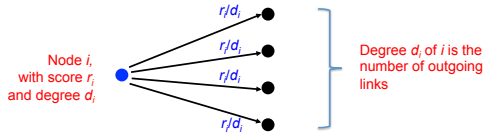
Sum over each of *i*'s Incoming links




27

## Page quality score

- Let's say page *i* has a quality score $r_i$
  - A link from *i* to a page is a "vote" for that page
  - If *i* is high-quality, its opinion should get more weight (be allowed to cast more votes)
  - So, let *i* have $r_i$ votes, to split equally among its outgoing links

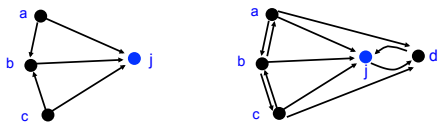



## Computing page quality score

- How do we compute the score for all pages?
- A simple model
  - A user (frog) is randomly clicking on links on webpages (lilypads)
  - At any given moment, there's some probability that the random user is on page p
  - If p is important, many important pages will link to it, so the probability that the random user is at p will be high
  - If p is unimportant, few important pages will link to it, so the probability of being at the page is low
- This defines a Markov chain, and we can compute the stationary distribution!
  - Pages with higher stationary probability are more important

## Page quality score – problems?

- Unfortunately, this doesn't always work.



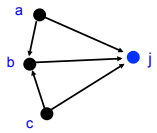
## The PageRank model

- Problem: some nodes have in-degree 0, so their score is 0 and their votes do not count
  - Simple fix: add a small constant to every rank score

Constant (e.g. 0.8)

$$r_j = \frac{p}{n} + (1-p)\sum_{i|i\to j}\frac{r_i}{d_i}$$

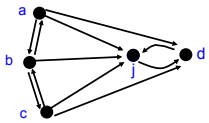
Total number of pages on web




31

## What does the model mean?

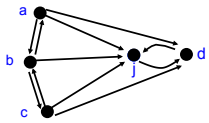
- Say you're a user on the web, visiting a page
  - with probability *1-p*, follow a random link on the page;
  - otherwise (probability *p*), visit a random webpage
  - continue doing this forever




32

## What does the model mean?

- Say you're a user on the web, visiting a page
  - with probability *1-p*, follow a random link on the page;
  - otherwise (probability *p*), visit a random webpage
  - continue doing this forever
- A page's score is equal to the probability that the user is visiting that page at any moment in time




33

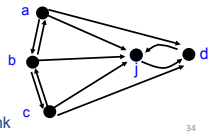
## What does the model mean?

- Say you're a user on the web, visiting a page
  - with probability *1-p*, follow a random link on the page;
  - otherwise (probability *p*), visit a random webpage
  - continue doing this forever
- A page's score is equal to the probability that user is visiting that page at any moment in time

$$r_j = \boxed{\frac{p}{n}} + \boxed{(1-p)\sum_{i|i\to j}\frac{r_i}{d_i}}$$

Probability of visiting page j randomly

Probability of visiting page j by following a link




34

## An *enormous* eigenvector problem!

$$B_{ij} = \frac{p}{n} + (1-p)\frac{1}{d_i}A_{ij} \qquad \vec{r} = B^T\vec{r}$$

- How to solve an eigen problem this huge?
  - There's a simple way of finding the principal eigenvector
  - For any nondegenerate vector v, the principal eigenvector of B is,

$$B^T B^T B^T B^T \ldots B^T v = \lim_{n\to\infty} B^n v$$

  - B is typically very sparse, so it's possible to compute this explicitly.

35