Lecture 22: 16 April, 2026

Madhavan Mukund

https://www.cmi.ac.in/~madhavan

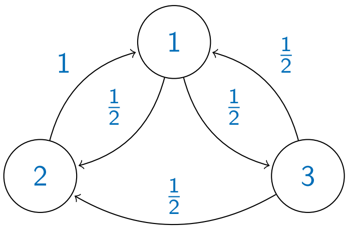
Data Mining and Machine Learning
January–April 2026

Gibbs sampling

- State of a Bayesian network is a valuation of variables $(V_1, V_2, \ldots, V_n)$
- Move probabilistically from $s_j = (x_1, x_2, \ldots, x_n)$ to $s_k = (y_1, y_2, \ldots, y_n)$
- Allow such a move only when s_j, s_k differ at exactly one position
  - $s_j = (x_1, x_2, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)$
  - $s_k = (x_1, x_2, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n)$
- Sampling algorithm
  - Current state is $s_j = (x_1, x_2, \ldots, x_n)$
  - Choose i uniformly in $[1, n]$
  - Resample x_i given current values $(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$
  - Random walk through state space — count number of visits to each state
- Need to compute $P[y_i \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]$
- Why does this work at all?

Markov chains

- Finite set of states, with transition probabilities between states
- For us, a state will be an assignment of values to variables
- A three state Markov Chain



- Represent using a transition matrix — stochastic

$$A = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

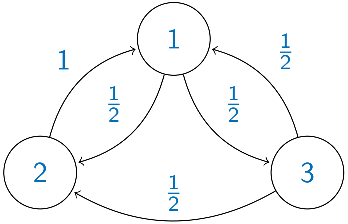
- $P[j]$ is probability of being in state j
- Start in state 1, so initially $P = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

- After one step:

$$P^\top A = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

- After second step:

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

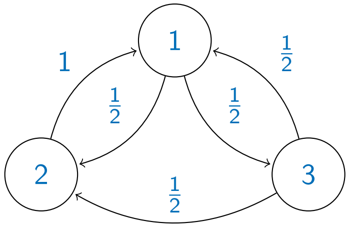


- After k steps, $P[j]$ is probability of being in state j
- Continuing our example,

$$\begin{bmatrix} \frac{3}{4} & \frac{1}{4} & 0 \end{bmatrix} \rightarrow \begin{bmatrix} \frac{1}{4} & \frac{3}{8} & \frac{3}{8} \end{bmatrix} \rightarrow \begin{bmatrix} \frac{9}{16} & \frac{5}{16} & \frac{1}{8} \end{bmatrix}$$

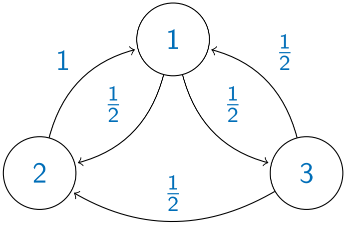
Ergodicity

- Is it the case that $P[j] > 0$ for all j continuously, after some point?
- Markov chain A is ergodic if there is some t_0 such that for every P, for all $t > t_0$, for every j, $(P^\top A^t)[j] > 0$.
  - No matter where we start, after $t > t_0$ steps, every state has a nonzero probability of being visited in step t
- Properties of ergodic Markov chains
  - There is a stationary distribution π, $\pi^\top A = \pi$
    - π is a left eigenvector of A
  - For *any* starting distribution P, $\lim_{t\to\infty} P^\top A^t = \pi$



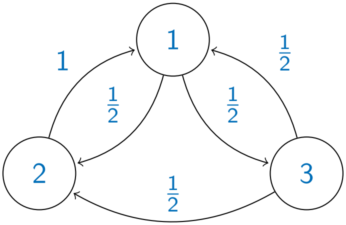
Ergodicity . . .

- How can ergodicity fail?
  - Starting from i, we reach a set of states from which there is no path back to i
  - We have a cycle $i \to j \to k \to i \to j \to k \cdots$, so we can only visit some states periodically
- Sufficient conditions for ergodicity
  - Irreducibility: When viewed as a directed graph, A is strongly connected
  - For all states i, j, there is a path from i to j and a path from j to i
  - Aperiodicity: For any pair of vertices i, j, the gcd of the lengths of all paths from i to j is 1
  - In particular, paths (loops) from i to i do not all have lengths that are multiples of some $k \geq 2$ — prevents bad cycles



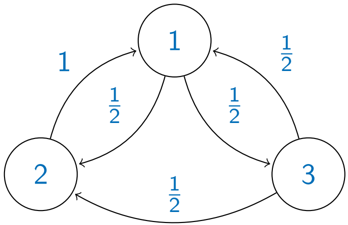
Ergodicity . . .

- Can efficiently approximate $\lim_{t\to\infty} P^\top A^t$ by repeated squaring: $P^\top A^2$, $P^\top A^4$, $P^\top A^8$, . . . , $P^\top A^{2^k}$, . . .
  - Mixing time — how fast this converges to π
- Stationary distribution represents fraction of visits to each state in a long enough execution
- Can we create a Markov chain from a Bayesian network so that the stationary distribution is meaningful?



Approximate inference using Markov chains

- Bayesian network has variables $v_1, v_2, \ldots, v_n$
- Each assignment of values to the variables is a state
- Set up a Markov chain on these states
- Gibbs sampling — random walk through state space, count visits to each state
- Stationary distribution should assign to state s the probability $P(s)$ in the Bayesian network
- How to reverse engineer the transition probabilities to achieve this?



Reversible Markov chains

- Ergodic Markov chain with stationary distribution π
- Transition matrix A, write p_{jk} for $A[j][k]$
  - Probability of transition from state j to state k
- Reversibility : $\pi_j \cdot p_{jk} = \pi_k \cdot p_{kj}$, for all j,k (balance equations)
  - Why reversibility? In steady state, can run the Markov chain forward or backward.
  - Probability of moving forward from state k to state j is same as probability of moving backward from state k to state j
  - Given an evolution $x_1 x_2 \ldots$, for large n, $P[x_n = j \mid x_{n-1} = k] = P[x_{n-1} = j \mid x_n = k]$
- Justification
  - $P[x_{n-1} = j \mid x_n = k] = P[x_n = k \mid x_{n-1} = j] \cdot \dfrac{P[x_{n-1} = j]}{P[x_n = k]}$
  - By reversibility, $p_{kj} = p_{jk}\dfrac{\pi_j}{\pi_k}$, so $P[x_n = j \mid x_{n-1} = k] = P[x_{n-1} = j \mid x_n = k]$

Reversible Markov chains

- Ergodic Markov chain with stationary distribution π
- Transition matrix A, write p_{jk} for $A[j][k]$
  - Probability of transition from state j to state k
- Reversibility : $\pi_j \cdot p_{jk} = \pi_k \cdot p_{kj}$, for all j,k (balance equations)
  - Why reversibility? In steady state, can run the Markov chain forward or backward.
  - Probability of moving forward from state k to state j is same as probability of moving backward from state k to state j
  - Given an evolution $x_1 x_2 \ldots$, for large n, $P[x_n = j \mid x_{n-1} = k] = P[x_{n-1} = j \mid x_n = k]$
- Justification
  - $P[x_{n-1} = j \mid x_n = k] = P[x_n = k \mid x_{n-1} = j] \cdot \dfrac{\pi_j}{\pi_k}$, in steady state
  - By reversibility, $p_{kj} = p_{jk}\dfrac{\pi_j}{\pi_k}$, so $P[x_n = j \mid x_{n-1} = k] = P[x_{n-1} = j \mid x_n = k]$

Reversible Markov chains

- Ergodic Markov chain with stationary distribution π
- Transition matrix A, write p_{jk} for $A[j][k]$
  - Probability of transition from state j to state k
- Reversibility : $\pi_j \cdot p_{jk} = \pi_k \cdot p_{kj}$, for all j,k (balance equations)
  - Why reversibility? In steady state, can run the Markov chain forward or backward.
  - Probability of moving forward from state k to state j is same as probability of moving backward from state k to state j
  - Given an evolution $x_1 x_2 \ldots$, for large n, $P[x_n = j \mid x_{n-1} = k] = P[x_{n-1} = j \mid x_n = k]$
- Justification
  - $P[x_{n-1} = j \mid x_n = k] = p_{jk} \cdot \dfrac{\pi_j}{\pi_k}$, in steady state
  - By reversibility, $p_{kj} = p_{jk} \dfrac{\pi_j}{\pi_k}$, so $P[x_n = j \mid x_{n-1} = k] = P[x_{n-1} = j \mid x_n = k]$

Reversible Markov chains

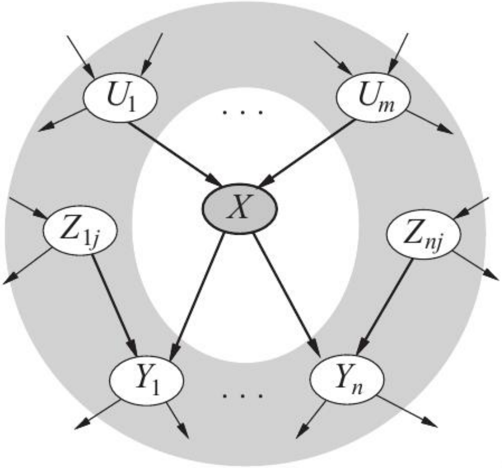
- Ergodic Markov chain
- Suppose $a^\top = (a_1, a_2, \ldots, a_n)$ satisfies reversibility balance equations for all j,k
  - $a_j \cdot p_{jk} = a_k \cdot p_{kj}$
- $\sum_k a_j \cdot p_{jk} = \sum_k a_k \cdot p_{kj}$
- $a_j \sum_k p_{jk} = \sum_k a_k \cdot p_{kj}$
- $a_j \cdot 1 = \sum_k a_k \cdot p_{kj}$
- $a^\top = a^\top A$, so $a^\top$ is the stationary distribution of A

Gibbs sampling

- State of a Bayesian network is a valuation of variables $(V_1, V_2, \ldots, V_n)$
- Move probabilistically from $s_j = (x_1, x_2, \ldots, x_n)$ to $s_k = (y_1, y_2, \ldots, y_n)$
- Allow such a move only when s_j, s_k differ at exactly one position
  - $s_j = (x_1, x_2, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)$
  - $s_k = (x_1, x_2, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n)$
- Sampling algorithm
  - Current state is $s_j = (x_1, x_2, \ldots, x_n)$
  - Choose i uniformly in $[1, n]$
  - Resample x_i given current values $(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$
- Need to compute $P[y_i \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]$

Markov blanket

- Recall $MB(X)$ — Markov blanket of X
  - $Parents(X)$
  - $Children(X)$
  - $Parents\ of\ Children(X)$
- $X \perp \neg MB(X) \mid MB(X)$
- Need to compute
  $P[y_i \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]$
- $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ fix $MB(V_i)$
- Can compute
  $P[y_i \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]$ given conditional probability tables in the network



Gibbs sampling

- Move from $s_j = (x_1, x_2, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)$ to $s_k = (x_1, x_2, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n)$
- Let $\bar{x} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$
- $p_{jk} = \frac{1}{n} P[y_i \mid \bar{x}] = \frac{1}{n} \frac{P(s_k)}{P(\bar{x})}$
- Likewise $p_{kj} = \frac{1}{n} P[x_i \mid \bar{x}] = \frac{1}{n} \frac{P(s_j)}{P(\bar{x})}$
- Therefore, $\frac{p_{jk}}{p_{kj}} = \frac{P(s_k)}{P(s_j)}$, so $P(s_j) \cdot p_{jk} = P(s_k) \cdot p_{kj}$ and this chain is reversible
- By our previous observation about any vector $a^\top$ satisfying balance equations, we must have $(P(s_1), P(s_2), \ldots, P(s_n)) = (\pi_1, \pi_2, \ldots, \pi_n)$ for the current Markov chain

Gibbs sampling

- Move from $s_j = (x_1, x_2, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)$ to $s_k = (x_1, x_2, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n)$
- $\pi_j \cdot p_{jk} = \pi_k \cdot p_{kj}$
- We have created a reversible Markov chain whose stationary distribution provides the true probabilities of the original Bayesian network!
- Gibbs sampling is a special case of the more general Metropolis-Hastings algorithm

Gibbs sampling

- Since we are dealing with steady state probabilities, it is not necessary to change just one variable at a time
  - Generate an entirely new sample state $(y_1, y_2, \ldots, y_n)$
  - First generate y_1, given $x_2, x_3, \ldots, x_n$
  - Then generate y_2, given $y_1, x_3, \ldots, x_n$
  - ...
  - Then generate y_n, given $y_1, y_2, \ldots, y_{n-1}$
- Standard Gibbs sampler — again a reversible Markov chain

Approximate inference using Markov chains

- Bayesian network has variables $V_1, V_2, \ldots, V_n$
- Use Gibbs sampling to set up a reversible Markov chain
- Stationary distribution assigns to each state s its probability $P(s)$ in the Bayesian network
- Run the Markov chain, tabulate frequencies of visits to each state, estimate stationary distribution