Probabilistic Graphical Models

Markov Chain Monte Carlo

Previously

Monte Carlo Sampling

Sample M i.i.d. samples from a distribution $P(X = x)$ and estimate expectation of function $f(X)$

$$\widehat{E_P}[f(X)] \quad = \frac{1}{M}\sum_m f(x_m)$$

If you cannot sample from $P(X = x)$:

Rejection sampling: Sample x from proposal distribution $\mathrm{k}Q$ and accept samples proportionally to $\frac{P(x)}{kQ(x)}$

Importance sampling: Sample from proposal distribution Q and weigh sample by $w(x) = \frac{P(x)}{Q(x)}$

For BNs

Monte Carlo Sampling

Forward Sampling: Sample M i.i.d. samples from a distribution $P(X = x)$ and estimate expectation of function $f(X)$

$$\widehat{E_P}[f(X)] = \frac{1}{M}\sum_m f(x_m)$$

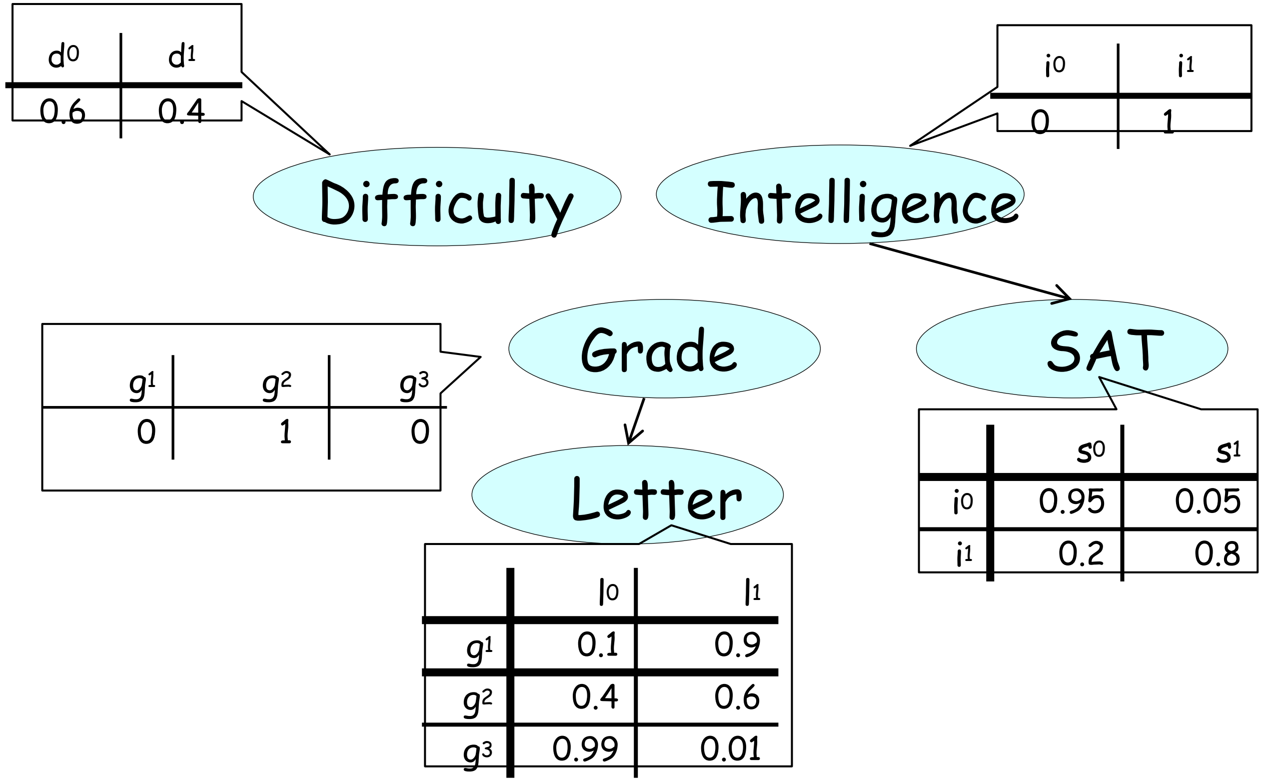
If you cannot sample from $P(X = x)$ (e.g., you want to sample from $P(X = x|E = e)$

Rejection sampling: Forward sample and accept samples where $E = e$

Importance sampling: Sample from proposal distribution (mutilated network B_e) and weigh sample by $w(\xi)$

$$w(\xi) = \frac{P(\xi)}{B_e(\xi)}$$

Importance Sampling for BNs



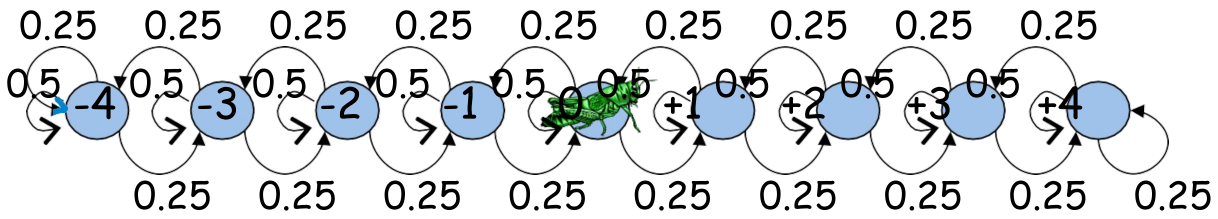
When your evidence is in the leaves, essentially all your sampling is done from the prior distribution

Idea: Sequence of non-i.i.d. samples that will start from the prior and hang out in the posterior.

Mutilated Graph $\mathcal{B}_{Z=z}$, proposal distribution $P_{\mathcal{B}_{Z=z}}$

Weight of a sample $w(\xi) = \frac{P_{\mathcal{B}}(\xi)}{P_{\mathcal{B}_{Z=z}}(\xi)}$.

Markov Chain

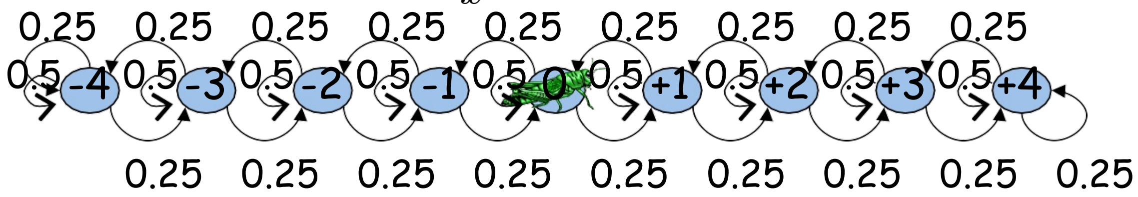


- A Markov chain defines a probabilistic transition model $T(x \rightarrow x')$ over states x:
  - for all x: $$\sum_{x'} T(x \rightarrow x') = 1$$

Daphne Koller

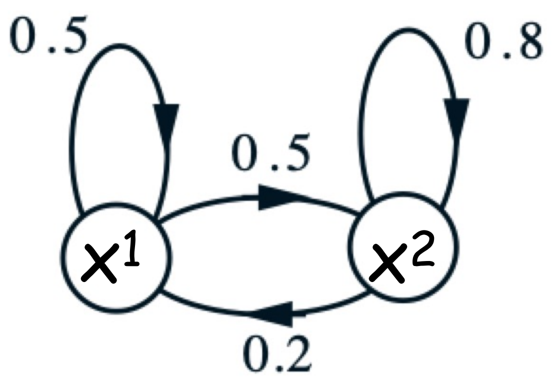
Temporal Dynamics

$$P^{(t+1)}(X^{(t+1)} = x') = \sum_x P^{(t)}(X^{(t)} = x) T(x \to x')$$



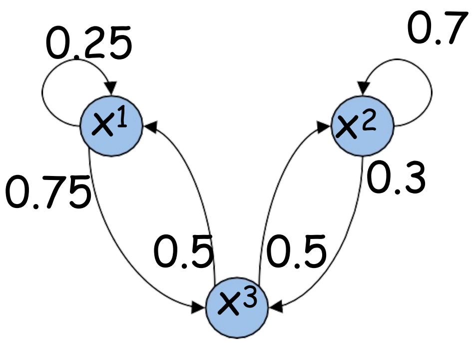
	-2	-1	0	+1	+2
$P^{(0)}$	0	0	1	0	0
$P^{(1)}$	0	.25	.5	.25	0
$P^{(2)}$	$.25^2 = .0625$	$2\times(.5\times.25) = .25$	$.5^2+2\times.25^2 = .375$	$2\times(.5\times.25) = .25$	$.25^2 = .0625$

Example



		$n=0$	$n=1$	$n=2$	$n=100$	$n=101$
Start at x^1	$P(X=x^1)$					
	$P(X=x^2)$					
Start at x^2	$P(X=x^1)$					
	$P(X=x^2)$					

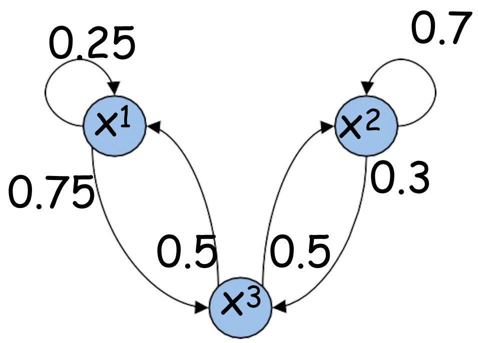
Stationary Distribution



$$P^{(t)}(x') \approx P^{(t+1)}(x') = \sum_x P^{(t)}(x)T(x \to x')$$

$$\pi(x') = \sum_x \pi(x)T(x \to x')$$

Stationary Distribution



$$P^{(t)}(x') \approx P^{(t+1)}(x') = \sum_x P^{(t)}(x)T(x \rightarrow x')$$

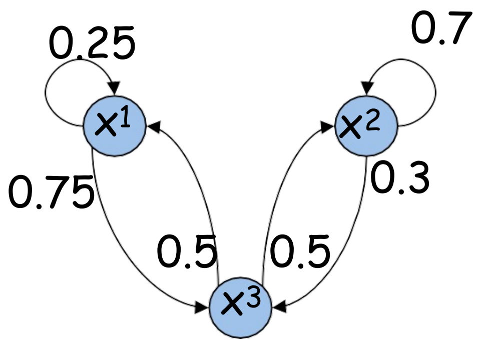
$$\pi(x') = \sum_x \pi(x)T(x \rightarrow x')$$

$$\begin{aligned}\pi(x^1) &= 0.25\pi(x^1) + 0.5\pi(x^3) \\ \pi(x^2) &= 0.7\pi(x^2) + 0.5\pi(x^3) \\ \pi(x^3) &= 0.75\pi(x^1) + 0.3\pi(x^2)\end{aligned}$$

$$\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$$

Daphne Koller

Stationary Distribution



$$P^{(t)}(x') \approx P^{(t+1)}(x') = \sum_x P^{(t)}(x) T(x \to x')$$

$$\pi(x') = \sum_x \pi(x) T(x \to x')$$

$$\begin{aligned} \pi(x^1) &= 0.25\pi(x^1) + 0.5\pi(x^3) \\ \pi(x^2) &= 0.7\pi(x^2) + 0.5\pi(x^3) \\ \pi(x^3) &= 0.75\pi(x^1) + 0.3\pi(x^2) \end{aligned} \qquad \begin{aligned} \pi(x^1) &= 0.2 \\ \pi(x^2) &= 0.5 \\ \pi(x^3) &= 0.3 \end{aligned}$$

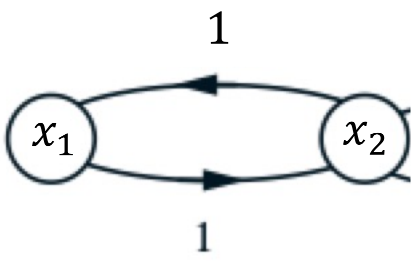
$$\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$$

Daphne Koller

Properties of Markov Chains

Not all Markov Chains have a unique stationary distribution

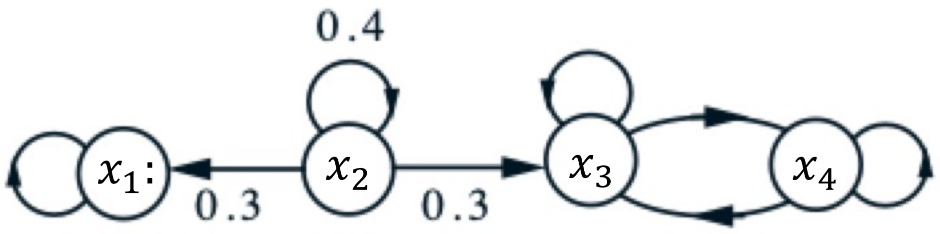
Periodic MC: No stationary distribution



$P^0(x_1) = 0$, then

$P^n(x_1) = 0$ if n even
$P^n(x_1) = 1$ if n odd

Reducible MC: No unique stationary distribution, depends on initial state



Start at x_1: $\pi(x_1) = 1$

Start at x_3: $\pi(x_1) = 0$

Start at x_2: $\pi(x_1) = \frac{1}{2}, for\ n \to \infty$

Regular Markov Chains

- A Markov chain is regular if there exists k such that, for every x, x', the probability of getting from x to x' in exactly k steps is > 0

- Theorem: A regular Markov chain converges to a unique stationary distribution regardless of start state

Regular Markov Chains

- A Markov chain is regular if there exists k such that, for every x, x', the probability of getting from x to x' in exactly k steps is > 0
- Sufficient conditions for regularity:
  - Every two states are connected
  - For every state, there is a self-transition

Using a Markov Chain

- Goal: compute $P(x \in S)$
  - but P is too hard to sample from directly
- Construct a Markov chain T whose unique stationary distribution is P
- Sample $x^{(0)}$ from some $P^{(0)}$
- For t = 0, 1, 2, …
  - Generate $x^{(t+1)}$ from $T(x^t \rightarrow x')$

Using a Markov Chain

- We only want to use samples that are sampled from a distribution close to P
- At early iterations, $P^{(t)}$ is usually far from P
- Start collecting samples only after the chain has run long enough to “mix”

Mixing

- How do you know if a chain has mixed or not?
  - In general, you can never “prove” a chain has mixed
  - But in many cases you can show that it has NOT

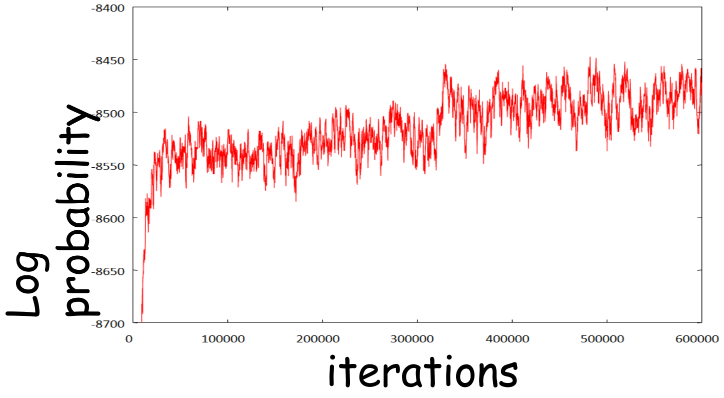
- How do you know a chain has not mixed?
  - Compare chain statistics in different windows within a single run of the chain
  - and across different runs initialized differently

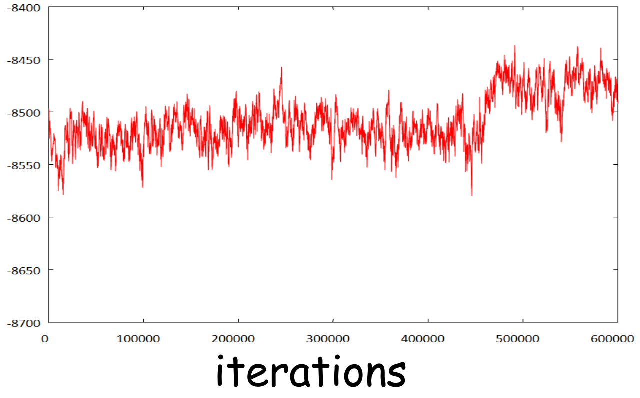
Mixing

Initialized from an arbitrary state

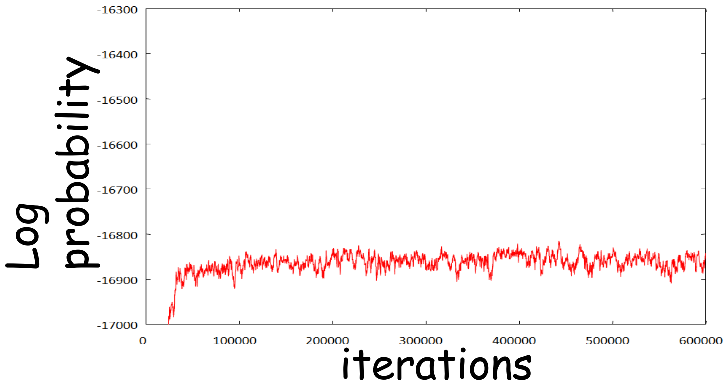
Initialized from a high-probability state

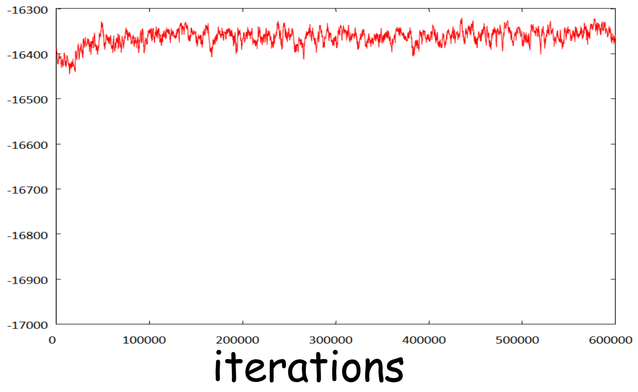
Mixing?





Maybe

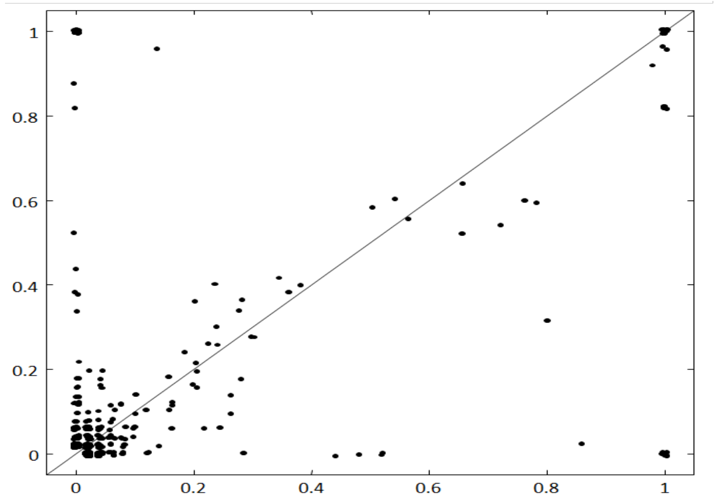


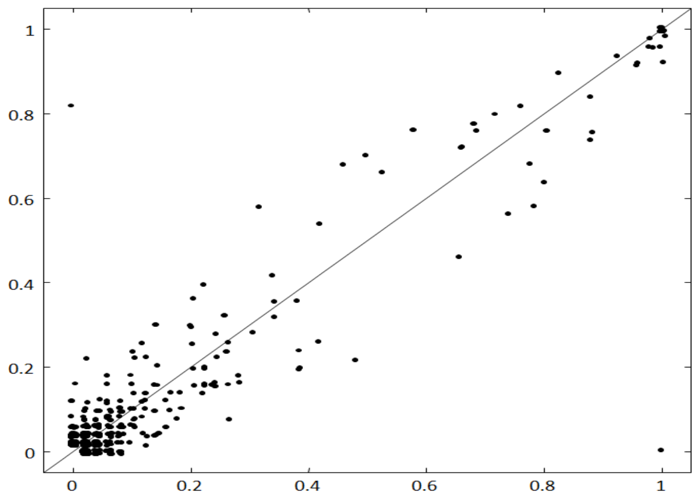


NO

Mixing

- Each dot is a statistic (e.g., P($x \in S$))
- x-position is its estimated value from chain 1
- y-position is its estimated value from chain 2





Mixing? NO Maybe

Using the Samples

- Once the chain mixes, all samples $x^{(t)}$ are from the stationary distribution π
    - So we can (and should) use all $x^{(t)}$ for $t > T_{mix}$
    - However, nearby samples are correlated!
  - So we shouldn't overestimate the quality of our estimate by simply counting samples
- The faster a chain mixes, the less correlated (more useful) the samples

MCMC Algorithm Summary I

- For $c = 1, \ldots, C$
  - Sample $x^{(0)}$ from $P^{(0)}$
- Repeat until mixing
  - For $c = 1, \ldots, C$
  - Generate $x^{(c,t+1)}$ from $\mathrm{T}(x^{(c,t+1)} \rightarrow x')$
  - Compare window statistics in different chains to determine mixing
  - $t := t + 1$

MCMC Algorithm Summary II

- Repeat until sufficient samples
  - $D := \emptyset$
  - For c=1,…,C
  - Generate $x^{(c,t+1)}$ from $\mathrm{T}(x^{(c,t+1)} \rightarrow x')$
    - $D := D \cup x^{(c,t+1)}$
  - t := t+1
- Let $D = \{x[1], \dots, x[M]\}$
- Estimate $\boldsymbol{E}_P[f] \approx \frac{1}{M} \sum_{m=1}^{M} f(x[m])$

Summary

- Pros:
  - Very general purpose
  - Often easy to implement
  - Good theoretical guarantees as $t \to \infty$
- Cons:
  - Lots of tunable parameters / design choices
  - Can be quite slow to converge
  - Difficult to tell whether it's working

MCMC for PGMs: Gibbs sampling

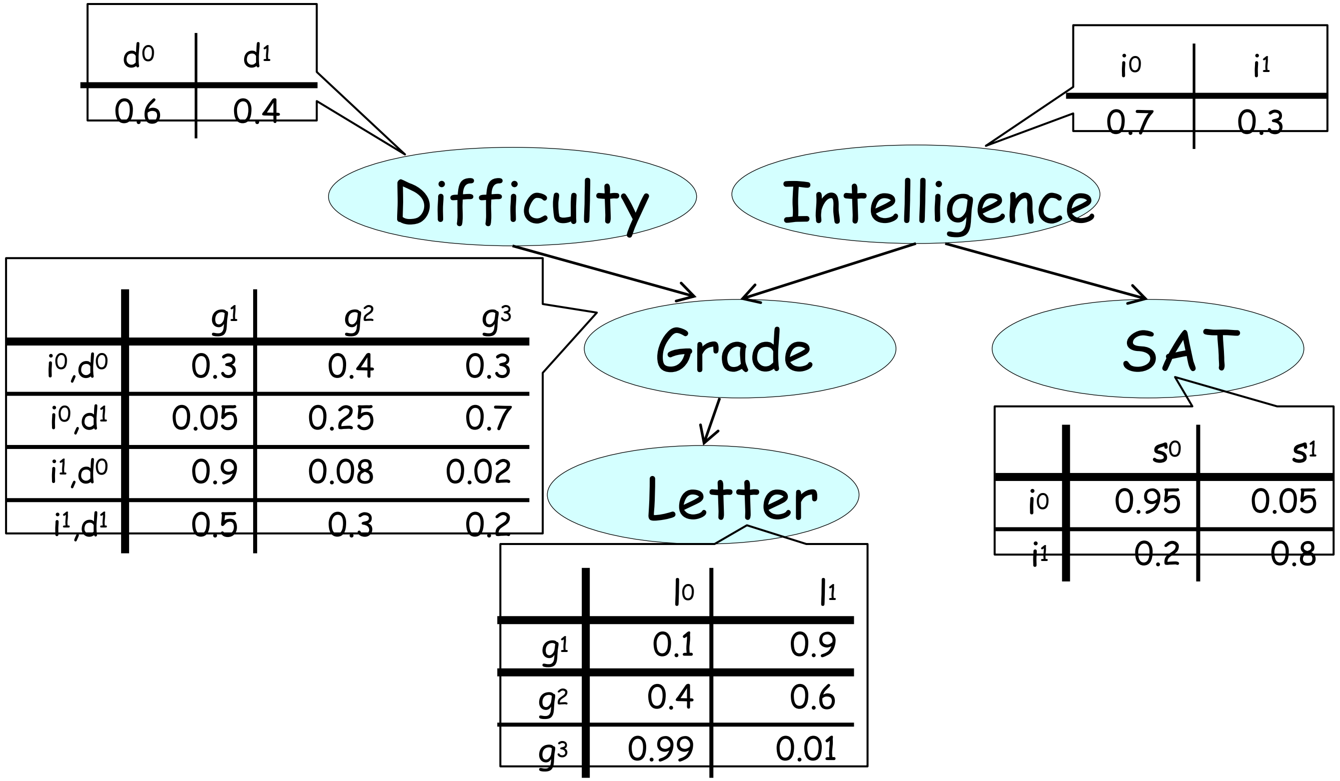
- Target distribution $P_{\Phi}(X_1, \ldots, X_n)$
- Markov chain state space: complete assignments $\boldsymbol{x}$ to $\boldsymbol{X} = \{X_1, \ldots, X_n\}$
- Transition model given starting state **x**:
  - For i=1,…,n
    - Sample $x_{i} \sim P_{\Phi}(X_i | \boldsymbol{x_{-i}})$ (all except x_i)
    - Set $x' = x$
- Example: X_1, X_2, X_3, X_4

MCMC for PGMs: Gibbs sampling

Transition model given starting state **x**:

- For i=1,…,n
  - Sample $x_i \sim P_\Phi\ (X_i | \boldsymbol{x}_{-\boldsymbol{i}})$ (all except x_i)
  - Set $\boldsymbol{x}' = \boldsymbol{x}$
- Example: X_1, X_2, X_3, X_4 :
  - Start from a random state, e.g. (0,0,0,0)
  - Sample $x_1 \sim P(X_1 | x_2 = 0, x_3 = 0, x_4 = 0)$
  - Sample $x_2 \sim P(X_2 | x_1, x_3 = 0, x_4 = 0)$
  - Sample $x_3 \sim P(X_3 | x_1, x_2, x_4 = 0)$
  - Sample $x_4 \sim P(X_4 | x_1, x_2, x_3)$

Example



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

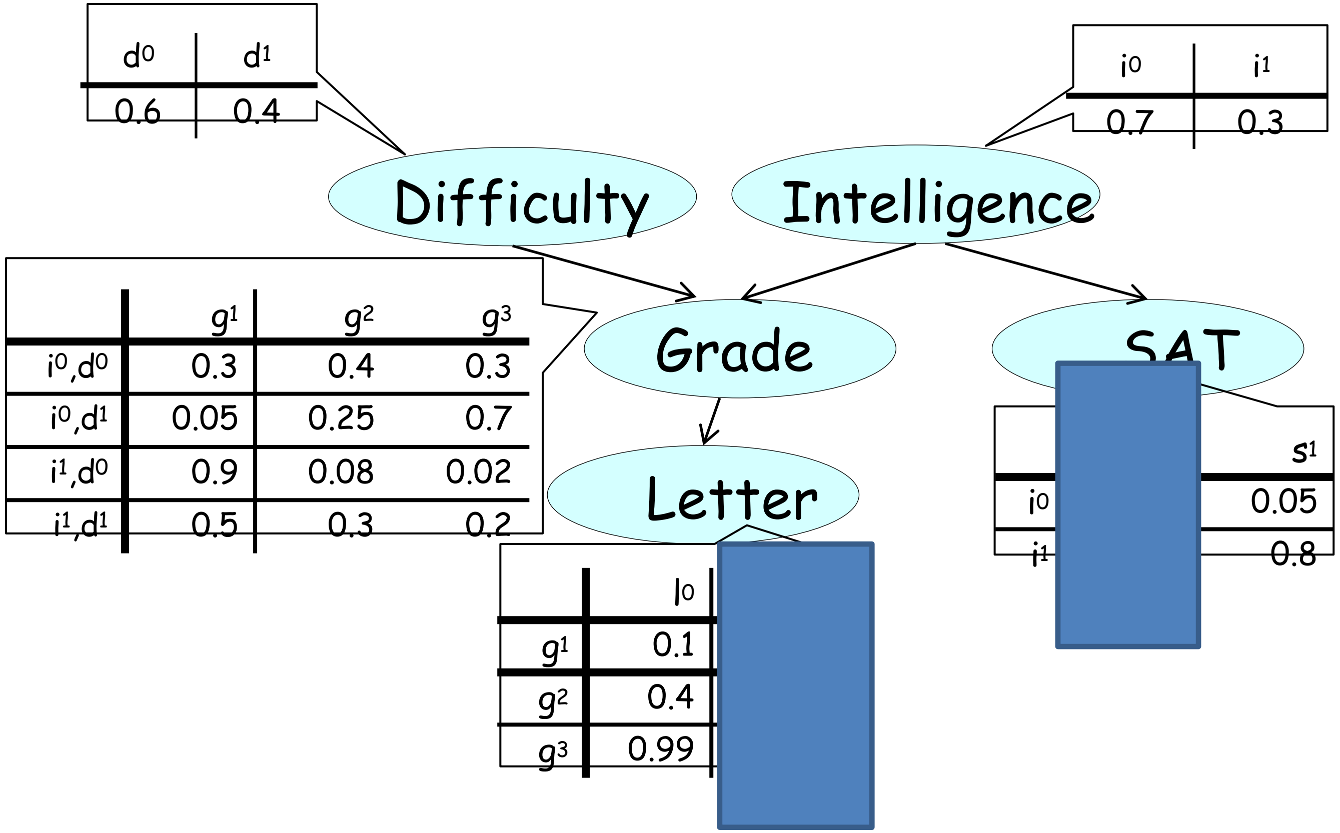
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Assume you want to sample from

$$L = l^0, S = s^1$$

Example



Assume you want to sample from

$$L = l^0, S = s^1$$

Step 1: Reduce factors according to evidence.

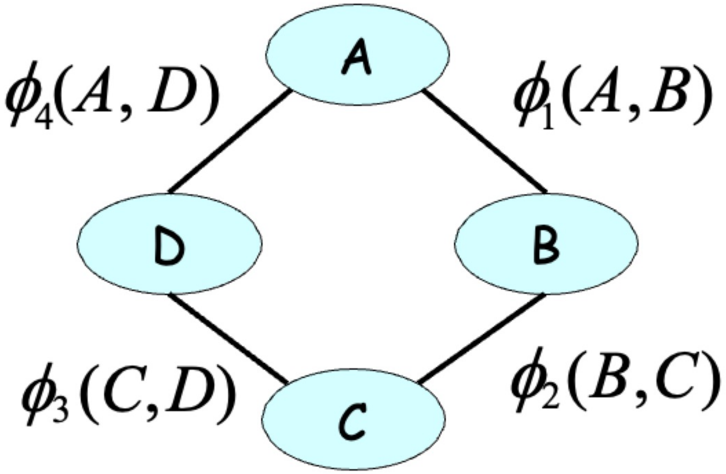
Step 2: Gibbs sampling

Sampling from. $P_\Phi(X_i|\boldsymbol{x_{-i}})$

For every step of the Gibbs sampler (every step of the MCMC), you want to sample from

$$P_\Phi(X_i \mid \boldsymbol{x}_{-i}) = \frac{P_\Phi(X_i, \boldsymbol{x}_{-i})}{P_\Phi(\boldsymbol{x}_{-i})} = \frac{\tilde{P}_\Phi(X_i, \boldsymbol{x}_{-i})}{\tilde{P}_\Phi(\boldsymbol{x}_{-i})}$$

Another Example



$$P_\Phi(A = a|b, c, \mathrm{d}) = \frac{\tilde{P}_\Phi(a,b,c,d)}{\sum_{A'} \tilde{P}_\Phi(A',b,c,d)} =$$

$$= \frac{\phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)}{\sum_{A'} \phi_1(A',b)\phi_2(b,c)\phi_3(c,d)\phi_4(A',d)} =$$

$$= \frac{\phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)}{\sum_{A'} \phi_1(A',b)\phi_2(b,c)\phi_3(c,d)\phi_4(A',d)}$$

Sampling from. $P_\Phi(X_i|\boldsymbol{x}_{-i})$

For every step of the Gibbs sampler (every step of the MCMC), you want to sample from

Multiply all

$$P_\Phi(X_i \mid \boldsymbol{x}_{-i}) = \frac{P_\Phi(X_i, \boldsymbol{x}_{-i})}{P_\Phi(\boldsymbol{x}_{-i})} = \frac{\tilde{P}_\Phi(X_i, \boldsymbol{x}_{-i})}{\tilde{P}_\Phi(\boldsymbol{x}_{-i})}$$

Sum over x_{-i}

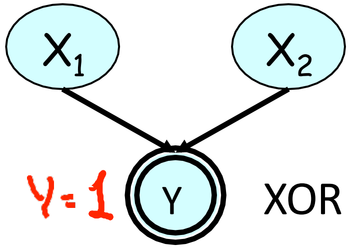
Reminder: Markov Boundary

For every X_i, $P(X_i|V \setminus X_i) = P(X_i|MB(X_i))$

For DAGs: Parents, Children, Spouses

For UGMs: Neighbors

Gibbs Chain and Regularity



X_1	X_2	Y	Prob
0	0	0	0.25
0	1	1	0.25
1	0	1	0.25
1	1	0	0.25

- If all factors are positive, Gibbs chain is regular
- However, mixing can still be very slow

Summary: Gibbs Sampling

- Converts the hard problem of inference to a sequence of “easy” sampling steps
- Pros:
  - Probably the simplest Markov chain for PGMs
  - Computationally efficient to sample
- Cons:
  - Only applies if we can sample from product of factors
  - Often slow to mix, esp. when probabilities are very high
    - How can you move away from the current space?

Reversible Chains

Detailed Balance Equation:

$$\pi(x)T(x \rightarrow x') = \pi(x')T(x' \rightarrow x)$$

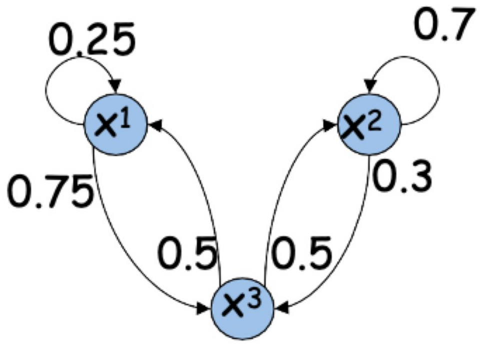
Definition: A Markov Chain is reversible if it satisfies the detailed balance equation for a unique distribution π

Reversible Chains

Detailed Balance Equation:

$$\pi(x)T(x \to x') = \pi(x')T(x' \to x)$$

Definition: A Markov Chain is reversible if it satisfies the detailed balance equation for a unique distribution π



$$\begin{aligned} \pi(x^1) &= 0.2 \\ \pi(x^2) &= 0.5 \\ \pi(x^3) &= 0.3 \end{aligned}$$

$$\pi(x^1)T(x^1 \to x^2) = \pi(x^2)T(x^2 \to x^1)$$
$$\pi(x^2)T(x^2 \to x^3) = \pi(x^3)T(x^3 \to x^2)$$
$$\pi(x^3)T(x^1 \to x^3) = \pi(x^1)T(x^3 \to x^1)$$

Metropolis Hastings Chain

Proposal distribution $Q(x \rightarrow x')$

Acceptance probability: $A(x \rightarrow x')$

- At each state x, sample x' from $Q(x \rightarrow x')$
- Accept proposal with probability $\mathrm{A}(x \rightarrow x')$
  - If proposal accepted, move to x'
  - Otherwise stay at x

$$T(x \rightarrow x') = Q(x \rightarrow x')A(x \rightarrow x'), if\ x \neq x'$$

$$T(x \rightarrow x) = Q(x \rightarrow x) + \sum_{x \neq x'} Q(x \rightarrow x')[1 - A(x \rightarrow x')]$$

Acceptance Probability

Construct A such that detailed balance holds

$$\pi(\boldsymbol{x})T(\boldsymbol{x} \to \boldsymbol{x}') = \pi(\boldsymbol{x}')T(\boldsymbol{x}' \to \boldsymbol{x})$$

$$\pi(\boldsymbol{x})Q(\boldsymbol{x} \to \boldsymbol{x}')A(\boldsymbol{x} \to \boldsymbol{x}') = \pi(\boldsymbol{x}')Q(\boldsymbol{x}' \to \boldsymbol{x})A(\boldsymbol{x}' \to \boldsymbol{x})$$

$$\frac{A(\boldsymbol{x} \to \boldsymbol{x}')}{A(\boldsymbol{x}' \to \boldsymbol{x})} = \frac{\pi(\boldsymbol{x}')Q(\boldsymbol{x}' \to \boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x} \to \boldsymbol{x}')}$$

Acceptance Probability

Construct A such that detailed balance holds

$$\pi(\boldsymbol{x})T(\boldsymbol{x} \to \boldsymbol{x}') = \pi(\boldsymbol{x}')T(\boldsymbol{x}' \to \boldsymbol{x})$$

$$\pi(\boldsymbol{x})Q(\boldsymbol{x} \to \boldsymbol{x}')A(\boldsymbol{x} \to \boldsymbol{x}') = \pi(\boldsymbol{x}')Q(\boldsymbol{x}' \to \boldsymbol{x})A(\boldsymbol{x}' \to \boldsymbol{x})$$

$A(x \to x') = \rho$
$A(x' \to x) = 1$

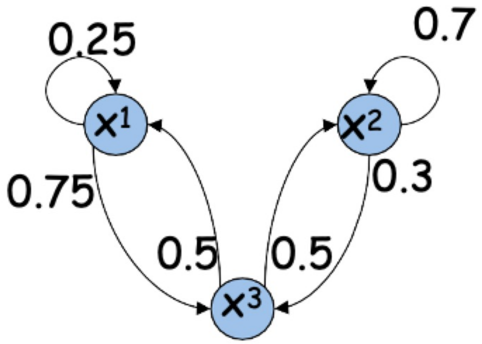
$$\frac{A(\boldsymbol{x} \to \boldsymbol{x}')}{A(\boldsymbol{x}' \to \boldsymbol{x})} = \frac{\pi(\boldsymbol{x}')Q(\boldsymbol{x}' \to \boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x} \to \boldsymbol{x}')}$$

$$A(\boldsymbol{x} \to \boldsymbol{x}') = \min\left[1, \frac{\pi(\boldsymbol{x}')Q(\boldsymbol{x}' \to \boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x} \to \boldsymbol{x}')}\right]$$

Example: Acceptance Probability

If $Q = T$, but you want to sample from a different stationary distribution $\pi'(x^1) = 0.6, \pi'(x^2) = 0.3, \pi'(x^3) = 0.1$

Find the Acceptance Probability



$$\pi(x^1) = 0.2$$
$$\pi(x^2) = 0.5$$
$$\pi(x^3) = 0.3$$

$$\pi(x^1)T(x^1 \to x^2) = \pi(x^2)T(x^2 \to x^1)$$
$$\pi(x^2)T(x^2 \to x^3) = \pi(x^3)T(x^3 \to x^2)$$
$$\pi(x^3)T(x^1 \to x^3) = \pi(x^1)T(x^3 \to x^1)$$

Proposal Distiribution

$$A(\boldsymbol{x} \to \boldsymbol{x}') = \min\left[1, \frac{\pi(\boldsymbol{x}')Q(\boldsymbol{x}' \to \boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x} \to \boldsymbol{x}')}\right]$$

- Q must be reversible:
  - $Q(x \to x') > 0 \Rightarrow Q(x' \to x) > 0$
- Opposing forces
  - Q should try to spread out, to improve mixing
  - But then acceptance probability often low

Relationship to Gibbs Sampling

Gibbs Sampling is a special case of MH

- The GS proposal distribution is

$$Q(x_i', \mathbf{x}_{-i} \mid x_i, \mathbf{x}_{-i}) = P(x_i' \mid \mathbf{x}_{-i})$$

($\mathbf{x}_{-i}$ denotes all variables except $\mathbf{x_i}$)

- Applying Metropolis-Hastings with this proposal, we obtain:

$$A(x_i', \mathbf{x}_{-i} \mid x_i, \mathbf{x}_{-i}) = \min\left(1, \frac{P(x_i', \mathbf{x}_{-i})Q(x_i, \mathbf{x}_{-i} \mid x_i', \mathbf{x}_{-i})}{P(x_i, \mathbf{x}_{-i})Q\left(x_i', \mathbf{x}_{-i} \mid x_i, \mathbf{x}_{-i}\right)}\right)$$

$$= \min\left(1, \frac{P(x_i', \mathbf{x}_{-i})P(x_i \mid \mathbf{x}_{-i})}{P(x_i, \mathbf{x}_{-i})P\left(x_i' \mid \mathbf{x}_{-i}\right)}\right) = \min\left(1, \frac{P(x_i' \mid \mathbf{x}_{-i})P(\mathbf{x}_{-i})P(x_i \mid \mathbf{x}_{-i})}{P(x_i \mid \mathbf{x}_{-i})P(\mathbf{x}_{-i})P\left(x_i' \mid \mathbf{x}_{-i}\right)}\right)$$

$$= \min(1,1) = 1$$

GS is simply MH with a proposal that is always accepted!

Summary

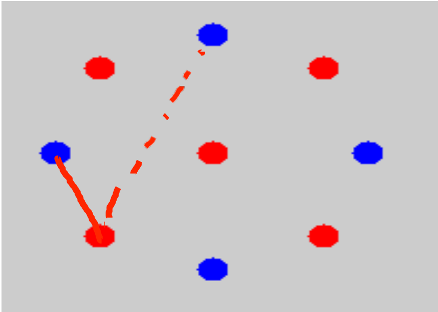
- MH is a general framework for building Markov chains with a particular stationary distribution
  - Requires a proposal distribution
  - Acceptance computed via detailed balance
- Tremendous flexibility in designing proposal distributions that explore the space quickly
  - But proposal distribution makes a big difference
  - and finding a good one is not always easy

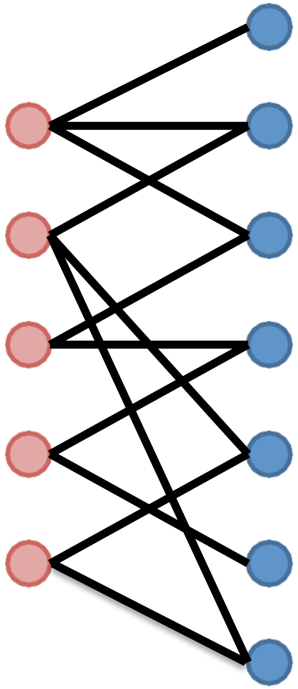
Gibbs Sampler is a special case of MH

MCMC for Matching

X_i = j if i matched to j

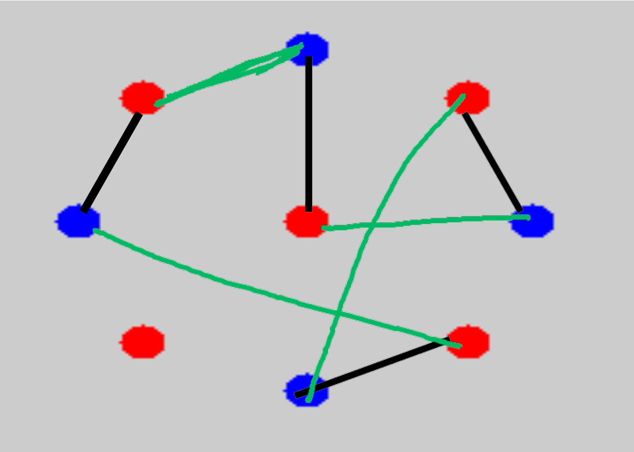
$$P(X_1 = v_1, \ldots, X_4 = v_4) \propto \begin{cases} \exp(-\sum_i \text{dist}(i, v_i)) & \text{if every } X_i \text{ has different value} \\ 0 & \text{otherwise} \end{cases}$$





Daphne Koller

MH for Matching: Augmenting Path



1) randomly pick one variable X_i
2) sample X_i, pretending that all values are available
3) pick the variable whose assignment was taken (conflict), and return to step 2

- When step 2 creates no conflict, modify assignment to flip augmenting path

Example Results