Readings:
K&F: 10.1, 10.2, 10.3

Approximate Inference by Sampling

Graphical Models – 10708
Ajit Singh
Carnegie Mellon University

November 3rd, 2006

Approximate inference overview

- There are many many many many approximate inference algorithms for PGMs
- We will focus on three representative ones:
  - sampling - *today*
  - variational inference - *continues next class*
  - loopy belief propagation and generalized belief propagation

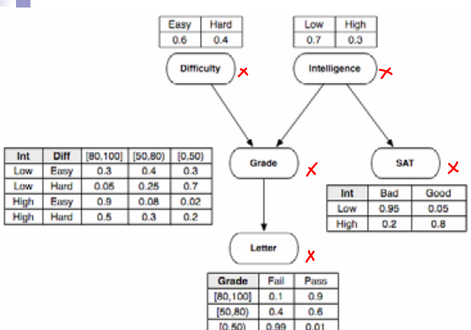
Goal

- Often we want expectations given samples x[1] … x[m] from a distribution P.

$$E_P[f] \approx \frac{1}{M}\sum_{m=1}^{M} f(x[m])$$

$$P(\mathbf{Y}=\mathbf{y}) \approx \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}(\mathbf{y}[m]=\mathbf{y})$$

Forward Sampling



Easy	Hard
0.6	0.4

Low	High
0.7	0.3

Int	Diff	[80,100]	[50,80)	[0,50)
Low	Easy	0.3	0.4	0.3
Low	Hard	0.05	0.25	0.7
High	Easy	0.9	0.08	0.02
High	Hard	0.5	0.3	0.2

Int	Bad	Good
Low	0.95	0.05
High	0.2	0.8

Grade	Fail	Pass
[80,100]	0.1	0.9
[50,80)	0.4	0.6
[0,50)	0.99	0.01

Sample nodes in topological order

D I G S L

= = = = =

Easy High [80,100] Bad Pass

$(\overset{v_1}{0.9}, \overset{v_2}{0.08}, \overset{v_3}{0.02})$

$r \sim \text{Unif}[0,1]$

0.9 0.98 1.0

v_1 v_2 v_3

Forward Sampling

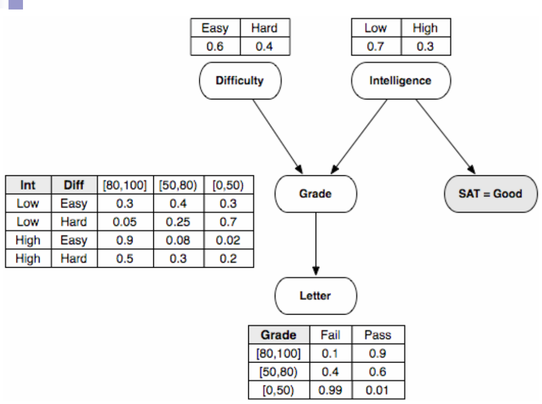
- $P(Y = y) = \#(Y = y) / N$
- $P(Y = y \mid E = e) = \#(Y = y, E = e) / \#(E = e)$
  - Rejection sampling: throw away samples that do not match the evidence.
- Sample efficiency
  - How often do we expect to see a record with E = e ?

Want sample w/ $Y=y$

Happens wp $P(Y=y)$

In expectation this happens $\frac{1}{P(Y=y)}$

Idea



Easy	Hard
0.6	0.4

Low	High
0.7	0.3

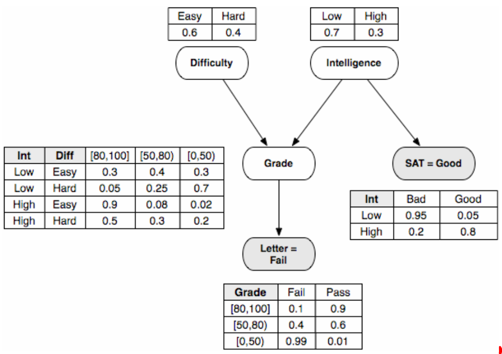
Int	Diff	[80,100]	[50,80)	[0,50)
Low	Easy	0.3	0.4	0.3
Low	Hard	0.05	0.25	0.7
High	Easy	0.9	0.08	0.02
High	Hard	0.5	0.3	0.2

Grade	Fail	Pass
[80,100]	0.1	0.9
[50,80)	0.4	0.6
[0,50)	0.99	0.01

What is we just fix the value of evidence nodes ?

What is expected number of records with (Intelligence = Low) ?

Likelihood Weighting



Easy	Hard
0.6	0.4

Low	High
0.7	0.3

Int	Diff	[80,100]	[50,80)	[0,50)
Low	Easy	0.3	0.4	0.3
Low	Hard	0.05	0.25	0.7
High	Easy	0.9	0.08	0.02
High	Hard	0.5	0.3	0.2

Int	Bad	Good
Low	0.95	0.05
High	0.2	0.8

Grade	Fail	Pass
[80,100]	0.1	0.9
[50,80)	0.4	0.6
[0,50)	0.99	0.01

$\xi[m] =$
D = Easy
I = Low
G = [50,80)

$w[m] = 0.05 * 0.4$

$$P(D=\text{Easy} \mid e) = \frac{\sum_{m=1}^{M} w[m]\, \mathbb{1}(D=\text{Easy})}{\sum_{m=1}^{M} w[m]}$$

Importance Sampling

- What if you cannot easily sample ?
  - Posterior distribution on a Bayesian network
    - P(Y = y | E = e) where the evidence itself is a rare event.
  - Sampling from a Markov network with cycles is always hard
    - See homework 4
- Pick some distribution Q(X) that is easier to sample from. Proposal
  - Assume that if P(x) > 0 then Q(x) > 0
  - Hopefully D(P||Q) is small

KL-Divergence

Importance Sampling

- Unnormalized Importance Sampling

$$E_{P(X)}[f(X)] = \sum_{x} f(x) P(x)$$

True BN / distr

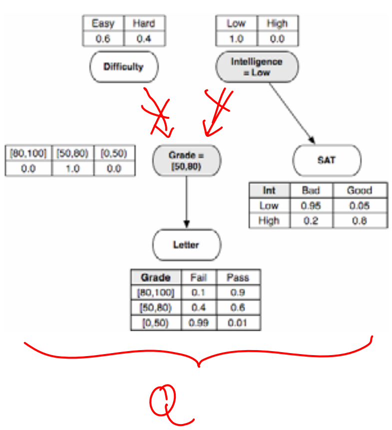
$$= \sum_{x} Q(x) f(x) \frac{P(x)}{Q(x)}$$

likelihood weighting $w(X)$

$$= E_{Q(X)}\left[f(X) \frac{P(X)}{Q(X)} \right]$$

$$\approx \frac{1}{M} \sum_{m=1}^{M} f(x[m]) \frac{P(x[m])}{Q(x[m])}$$

Mutilated BN Proposal



- Generating a proposal distribution for a Bayesian network
- Evidenced nodes have no parents.
- Each evidence node $Z_i = z_i$ has distribution $P(Z_i = z_i) = 1$
- Equivalent to likelihood weighting

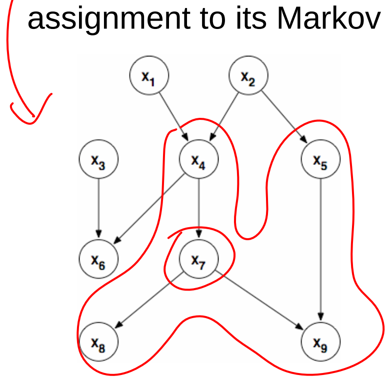
Forward Sampling Approaches

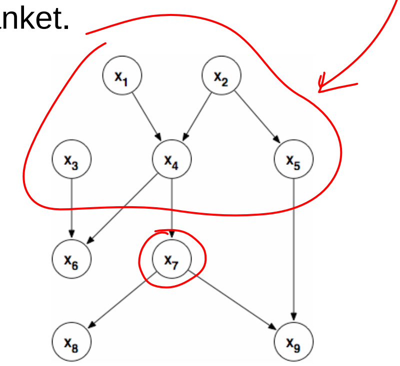
- Forward sampling, rejection sampling, and likelihood weighting are all *forward samplers*
  - Requires a topological ordering. This limits us to
    - Bayesian networks
    - Tree Markov networks
  - Unnormalized importance sampling can be done on cyclic Markov networks, but it is expensive
    - See homework 4
- Limitation
  - Fixing an evidence node only allows it to directly affect its descendents.

Scratch space

Markov Blanket Approaches

- *Forward Samplers*: Compute weight of X_i given assignment to ancestors in topological ordering
- *Markov Blanket Samplers*: Compute weight of X_i given assignment to its Markov Blanket.



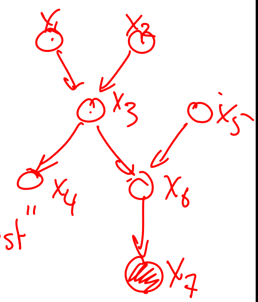


Markov Blanket Samplers

- Works on any type of graphical model covered in the course thus far.

Gibbs Sampling

1. Let **X** be the non-evidence variables
2. Generate an initial assignment $\xi^{(0)}$ } asmt to $X_1 \ldots X_6$ here
3. For t = 1..T } T= max iterations
   1. $\xi^{(t)} = \xi^{(t-1)}$
   2. For each X_i in **X**
      1. $\mathbf{u}_i$ = Value of variables **X** - $\{X_i\}$ in sample $\xi^{(t)}$] "rest"
      2. Compute $P(X_i \mid \mathbf{u}_i)$
      3. Sample $x_i^{(t)}$ from $P(X_i \mid \mathbf{u}_i)$
      4. Set the value of $X_i = x_i^{(t)}$ in $\xi^{(t)}$
4. Samples are taken from $\xi^{(0)} \ldots \xi^{(T)}$

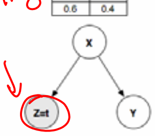


Computing $P(X_i \mid \mathbf{u}_i)$

- The major task in designing a Gibbs sampler is deriving $P(X_i \mid \mathbf{u}_i)$
- Use conditional independence
  - $X_i \perp X_j \mid MB(X_i)$ for all X_j in **X** - $MB(X_i)$ - $\{X_i\}$

no sampling

t	f
0.6	0.4



X	t	f
t	0.25	0.75
f	0.1	0.9

X	t	f
t	0.7	0.3
f	0.8	0.2

P(X|Y = y) = $\frac{P(X, Y=y)}{P(Y=y)}$ if z=f P(.)=0

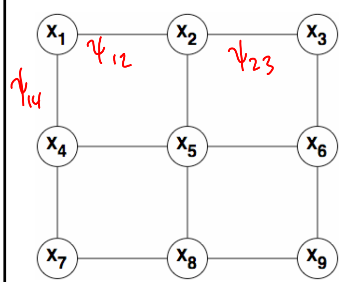
$\xi^{(t\to\infty)}$

$$= \frac{\sum_Z P(X, Y=y, Z=z)}{\sum_X \sum_Z P(X=x, Y=y, Z=z)}$$

P(Y|X = x) =

$\xi^0 \ldots \xi^T \approx P(X, y \mid Z=t)$ lookup in

Pairwise Markov Random Field



$x_i \in \{1,2\}$ $\forall i$

$$P(x_i | x_1, \dots x_{i-1}, x_{i+1}, \dots, x_n) =$$

$$= \frac{P(x_1 \dots x_n)}{P(x_1 \dots x_{i-1}, x_{i+1}, \dots, x_n)}$$

$$= \frac{\cancel{\frac{1}{Z}} \prod_i \Phi(x_i) \prod_{(j,k)\in E} \psi(x_j, x_k)}{\sum_{x_i} \cancel{\frac{1}{Z}} \prod_i \Phi(x_i) \prod_{(j,k)\in E} \psi(x_j, x_k)}$$

Cancel terms

$$\propto \Phi(x_i) \prod_{j \in N(i)} \psi(x_i, x_j)$$

$$P(x) = \frac{1}{Z} \prod_i \Phi(x_i) \prod_{(j,k)\in E} \Psi(x_j, x_k)$$

See scratch slide

Markov Chain Interpretation

- The state space consists of assignments to X. $(X_1 \dots X_6)$
- $P(x_i \mid \mathbf{u}_i)$ are the transition probability (neighboring states differ only in one variable)
- Given the transition matrix you could compute the exact stationary distribution
  - Typically impossible to store the transition matrix.
- Gibbs does not need to store the transition matrix !

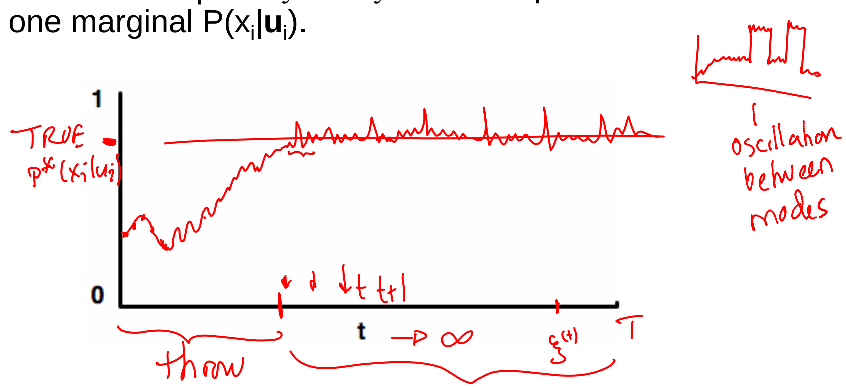
Scratch space

$$P(x_i \mid x_1 \dots x_{i-1}, x_{i+1}, \dots x_n) \propto \Phi(x_i) \prod_{j \in N(i)} \psi(x_i, x_j)$$

HW4 Q6

Convergence

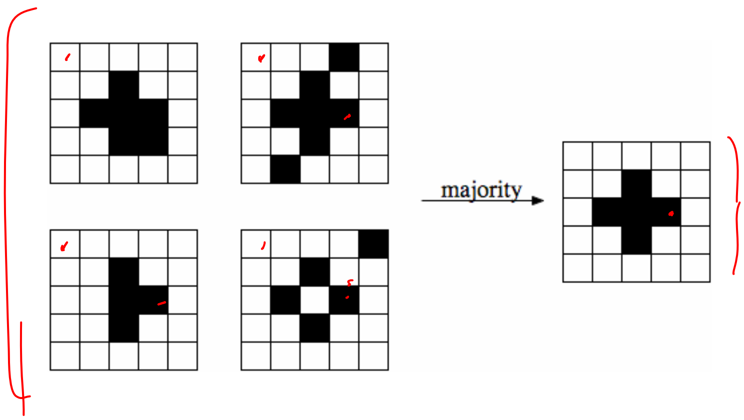
- Not all samples $\xi^{(0)} \dots \xi^{(T)}$ are independent. Consider one marginal $P(x_i|\mathbf{u}_i)$.



- Burn-in
- Thinning

MAP by Sampling

- Generate a few samples from the posterior
- For each X_i the MAP is the majority assignment



What you need to know

- Forward sampling approaches
  - Forward Sampling / Rejection Sampling
    - Generate samples from P(X) or P(X|e)
  - Likelihood Weighting / Importance Sampling
    - Sampling where the evidence is rare
    - Fixing variables lowers variance of samples when compared to rejection sampling.
  - Useful on Bayesian networks & tree Markov networks
- Markov blanket approaches
  - Gibbs Sampling
    - Works on any graphical model where we can sample from P(X_i | rest).
    - Markov chain interpretation.
    - Samples are independent when the Markov chain converges.
    - Convergence heuristics, burn-in, thinning.