Lecture 17: March 23, 2015
cs 573: Probabilistic Reasoning
Professor Nevatia
Spring 2015

Review

- HW #5 due Mar 30;
  - Conversion from conditional to canonical form posted on class page
- Exam 1, please return with or without comments
- Previous Lecture
  - Variational approximation examples
  - Intro to sampling or particle based approximation
- Today's objective
  - Importance sampling
  - Markov Chain Monte Carlo (MCMC) methods

Likelihood Weighting

- We can just set evidence variables to the observed values to avoid rejecting samples
- However, this biases the samples. Suppose that evidence is $S=s^1$;
- Apply normal forward sampling, we would start from prior distributions of D and I, so $I=i^1$ in only 30% of the cases; this is clearly not consistent with $S=s^1$.
- In rejection sampling, if we started with $I=i^1$ then $S=s^1$ in 80% of the samples (from CPD); if we started with $I=i^0$ then $S=s^1$ in only 5% samples
- Weight samples by their "likelihood": (i^1,s^1) by .8, (i^0,s^1) by .05
  – This weighting compensation is intuitive, formal justification comes later
  – Generalizes to case where more than one variable is evidence variable (product of probabilities for generating evidence variables given their parents; specifics in Algorithm 12.2, next slide).

$$\hat{P}_{\mathcal{D}}(\boldsymbol{y} \mid \boldsymbol{e}) = \frac{\sum_{m=1}^{M} w[m] \boldsymbol{I}\{\boldsymbol{y}[m] = \boldsymbol{y}\}}{\sum_{m=1}^{M} w[m]}$$

LW Algorithm

Algorithm 12.2 Likelihood-weighted particle generation

Procedure LW-Sample (
$\mathcal{B}$, // Bayesian network over $\mathcal{X}$
$\boldsymbol{Z} = \boldsymbol{z}$ // Event in the network
)

Let $X_1, \ldots, X_n$ be a topological ordering of $\mathcal{X}$
$w \leftarrow 1$
for $i = 1, \ldots, n$
  $\boldsymbol{u}_i \leftarrow \boldsymbol{x}\langle \mathrm{Pa}_{X_i} \rangle$ // Assignment to Pa_{X_i} in $x_1, \ldots, x_{i-1}$
  if $X_i \notin \boldsymbol{Z}$ **then**
    Sample x_i from $P(X_i \mid \boldsymbol{u}_i)$
  else
    $x_i \leftarrow \boldsymbol{z}\langle X_i \rangle$ // Assignment to X_i in $\boldsymbol{z}$
    $w \leftarrow w \cdot P(x_i \mid \boldsymbol{u}_i)$ // Multiply weight by probability of desired value
return $(x_1, \ldots, x_n), w$

Importance Sampling

- Generalization of LW. Also proves that LW is correct
- It may be hard to generate samples from P(**X**), called the *target distribution*; instead, sample from another, simpler but related distribution, Q(**X**), called a *proposal or sampling distribution*.
  - Average the *f* value (to computed expected value of *f* (x))
  - This will not give correct expectation of *f* for P; need to compensate as shown on next slide.
- Require that Q(x) > 0 whenever P(x) > 0; otherwise, Q can be arbitrary .
  - *e.g.* Q (x) can be a uniform distribution
  - However, computational efficiency will be better if Q is similar to P

Unnormalized Importance Sampling

- $E_{P(\mathbf{X})}[f(\mathbf{X})] = \Sigma_x f(x). P(x)$

  $= \Sigma_x Q(x). f(x). (P(x)/ Q(x))$

  $= E_{Q(\mathbf{X})}[f(\mathbf{X}) P(\mathbf{X})/ Q(\mathbf{X}))]$

Note: we can compute P(x) for a given x as the distribution is given

- Given set of samples D = {ξ[1],…, ξ[M]} from Q

$$\hat{\boldsymbol{E}}_{\mathcal{D}}(f) = \frac{1}{M} \sum_{m=1}^{M} f(\boldsymbol{x}[m]) \frac{P(\boldsymbol{x}[m])}{Q(\boldsymbol{x}[m])}.$$

Note: last term is like a weight; still this method is called unnormalized importance sampling

- Can be shown that this estimator is not biased: mean is always the desired value
- Variance of the estimator
  - Formal derivation in the book but depends on the variance of P(x)/Q(x), so it is useful to try to find Q that is similar to P.

Normalized Importance Sampling

- Previous method requires computing P(ξ) (normalized) for each sample.
- Sometimes, this is difficult, e.g. in a MN, we can easily get an unnormalized probability but computing partition function may be expensive
- Define a weight relative to the unnormalized distribution

$$w(\boldsymbol{X}) = \frac{\tilde{P}(\boldsymbol{X})}{Q(\boldsymbol{X})}$$

$$\boldsymbol{E}_{Q(\boldsymbol{X})}[w(\boldsymbol{X})] = \sum_{\boldsymbol{x}} Q(\boldsymbol{x}) \frac{\tilde{P}(\boldsymbol{x})}{Q(\boldsymbol{x})} = \sum_{\boldsymbol{x}} \tilde{P}(\boldsymbol{x}) = Z.$$

Normalized Importance Sampling

$$
\begin{aligned}
\boldsymbol{E}_{P(\mathbf{X})}[f(\boldsymbol{X})] &= \sum_{\boldsymbol{x}} P(\boldsymbol{x}) f(\boldsymbol{x}) \\
&= \sum_{\boldsymbol{x}} Q(\boldsymbol{x}) f(\boldsymbol{x}) \frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})} \\
&= \frac{1}{Z} \sum_{\boldsymbol{x}} Q(\boldsymbol{x}) f(\boldsymbol{x}) \frac{\tilde{P}(\boldsymbol{x})}{Q(\boldsymbol{x})} \\
&= \frac{1}{Z} \boldsymbol{E}_{Q(\mathbf{X})}[f(\boldsymbol{X}) w(\boldsymbol{X})] \\
&= \frac{\boldsymbol{E}_{Q(\mathbf{X})}[f(\boldsymbol{X}) w(\boldsymbol{X})]}{\boldsymbol{E}_{Q(\mathbf{X})}[w(\boldsymbol{X})]}.
\end{aligned}
$$

$$
\hat{\boldsymbol{E}}_{\mathcal{D}}(f) = \frac{\sum_{m=1}^{M} f(\boldsymbol{x}[m]) w(\boldsymbol{x}[m])}{\sum_{m=1}^{M} w(\boldsymbol{x}[m])}.
$$

Similar to likelihood weighting but samples from Q, not necessarily same as P.

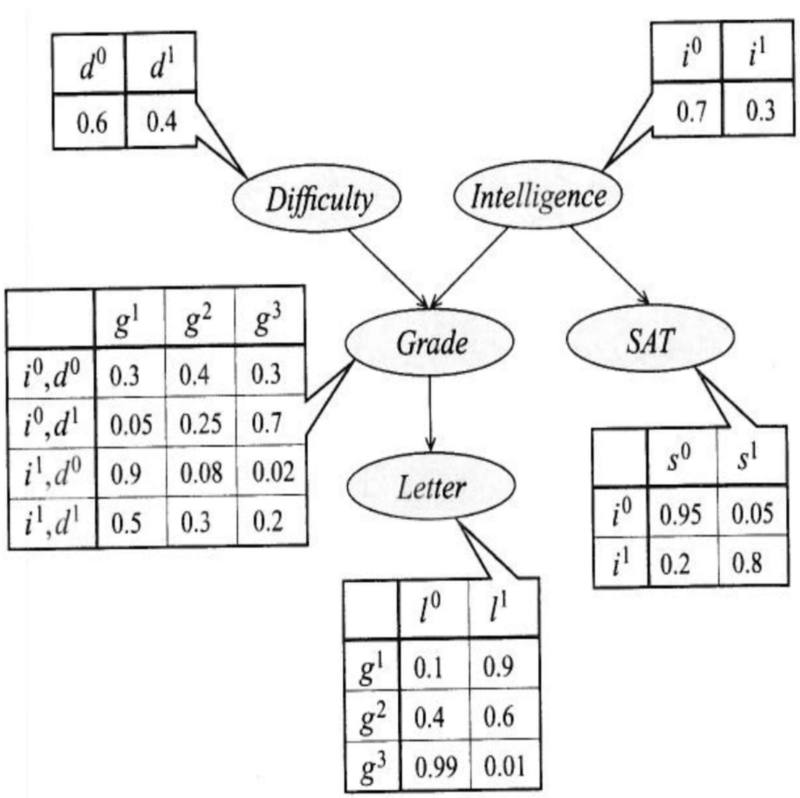
Properties of Normalized IS

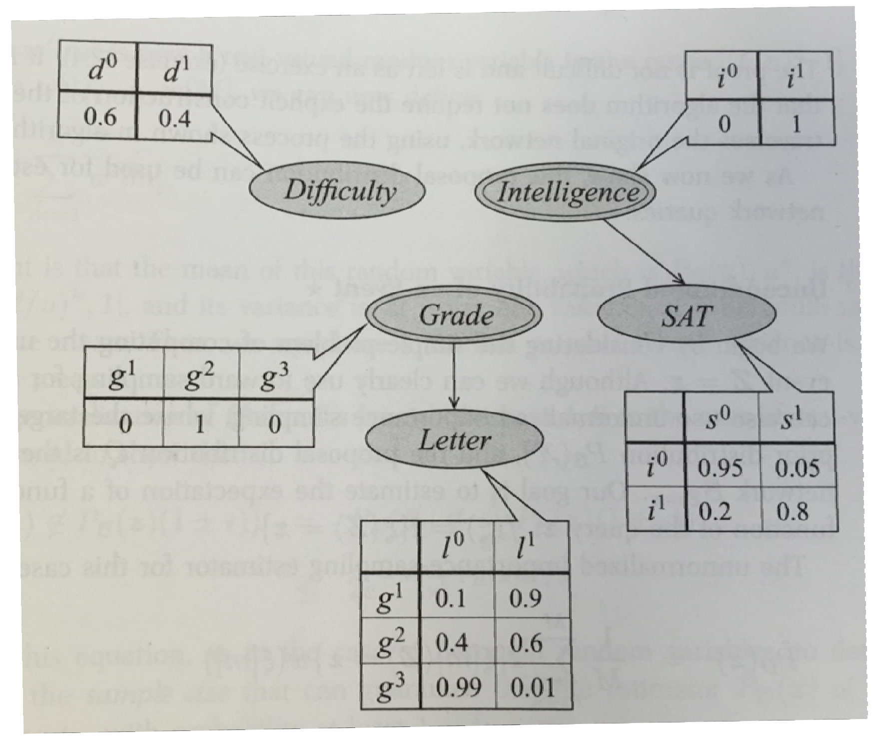
- Estimator has a bias for small values of M, estimates are closer to those of Q than of P
  - Bias goes down as 1/M
- Variance of estimator
  - Analyzed in the book
  - Variance is "typically" lower than for unnormalized sampling though this is not guaranteed
- Normalized sampling may be preferred due to lower variance even though it may have a bias.

Importance Sampling for BN

- For BN, there is a simple construct available for generating a good proposal distribution Q
- Suppose we are interest in the event where **Z** = **z**; e.g. G = g^2
  – This could represent evidence or the specific part of distribution we want to estimate
- In example, influence of Z on its descendants (starting from children) is easy to model: e.g. sample from $P(L|g^2)$.
  – Modeling influence on non-descendants is difficult so one choice is just to ignore this influence
- Modify the network so nodes containing Z_i have no parents and CPD of Z_i is 1 for $Z_i = z_i$, 0 otherwise.
- This defines a *mutilated* network (see next slide). Use this network for the proposal distribution function, Q
  – Property that $Q(x) > 0$ when $P(x) > 0$ is satisfied

Mutilated Network





Importance Sampling for BN

- Don’t actually need to break the network, can incorporate its effect in forward sampling; algorithm 12.2
  - Note: weight modified only when Q and P give different samples
  - Proposition 12.2 shows that the intuitive LW algorithm is equivalent to normalized importance sampling for BNs.

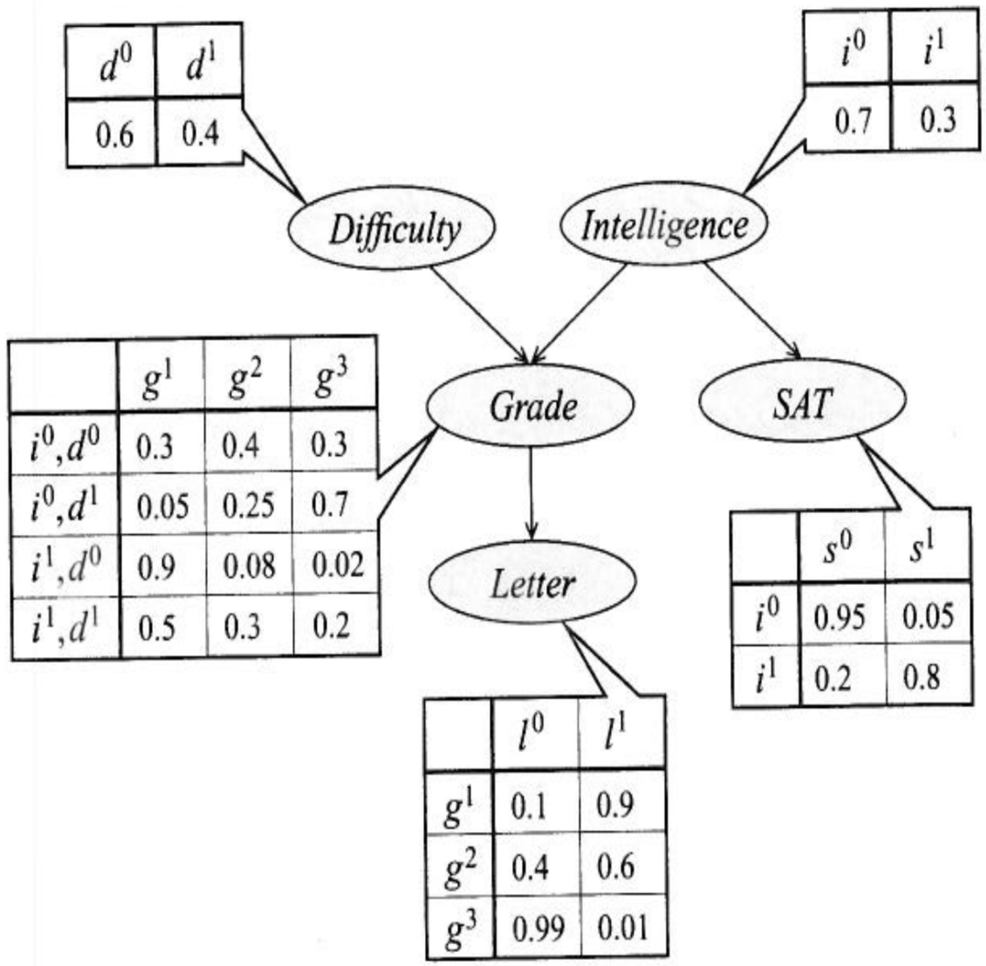
Markov Chain Monte Carlo (MCMC) Methods

- Forward sampling and likelihood weighted sampling methods are defined for BNs only.
  - Also, when evidence is mostly at the leaf nodes, the methods largely sample from the prior distribution
- Another approach is to generate a sequence (chain)
  - Called Markov chain because the next element depends only on the current sample
  - Monte Carlo because of random sampling
- Applies to both directed and undirected models
- Has many practical applications

Gibbs Sampling (Gibbs Chain)

- Reduce all factors according to evidence
- Generate an initial sample
  - Say from forward sampling; reduce factors if evidence variables are given
- Iterate over unobserved variables, one at a time
  - Sample a new value, given the current sample values for all other variables
  - Consider variable X_i ; sample from $P_\Phi (X_i | \boldsymbol{x}_{-i})$; $\boldsymbol{x}_{-i}$ means assignment of all variables in X except variable X_i .
  - Example and algorithm to follow
- At each step, we are sampling from a posterior distribution (though not necessarily the correct one)
- Can be shown that the process converges to the actual posterior distribution

Example 12.4



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

Example 12.4

- Let evidence be s^1 and l^0
- Reduced factors are P(I), P(D), P(G|I,D), P(s^1 |I), P(l^0|G)
- Let first sample, at time 0, by forward sampling, be d^1, i^0, g^2
- Now sample in order of G, I, D (note: need not start from root)
- Sample for G is drawn from P_Φ (G | d^1, i^0, s^1,l^0)

$$
\begin{aligned}
P_\Phi(G \mid d^1, i^0) &= \frac{P(i^0)P(d^1)P(G \mid i^0, d^1)P(l^0 \mid G)P(s^1 \mid i^0)}{\sum_g P(i^0)P(d^1)P(g \mid i^0, d^1)P(l^0 \mid g)P(s^1 \mid i^0)} \\
&= \frac{P(G \mid i^0, d^1)P(l^0 \mid G)}{\sum_g P(g \mid i^0, d^1)P(l^0 \mid g)}.
\end{aligned}
$$

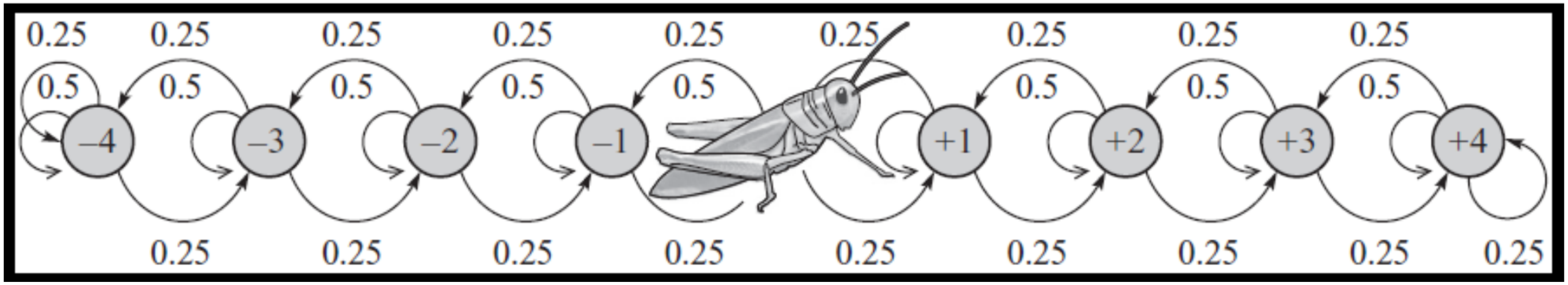
- Multiply all factors containing G and normalize
- Sample from this distribution, let the sample yield G= g^3
- Now sample I from P_Φ (I | d^1, g^3);
- Sample D next and keep iterating
- Distribution from which samples are drawn, get closer and closer to the posterior distribution

Algorithm 12.4 Generating a Gibbs chain trajectory

Procedure Gibbs-Sample (
    $\boldsymbol{X}$ // Set of variables to be sampled
    Φ // Set of factors defining P_Φ
    $P^{(0)}(\boldsymbol{X})$, // Initial state distribution
    T // Number of time steps
)
    Sample $\boldsymbol{x}^{(0)}$ from $P^{(0)}(\boldsymbol{X})$
    for $t = 1, \ldots, T$
        $\boldsymbol{x}^{(t)} \leftarrow \boldsymbol{x}^{(t-1)}$
        for each $X_i \in \boldsymbol{X}$
            Sample $x_i^{(t)}$ from $P_\Phi(X_i \mid \boldsymbol{x}_{-i})$
                // Change X_i in $\boldsymbol{x}^{(t)}$
return $\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(T)}$

Markov Chains

- Defined by a transition function T ($\mathbf{x} \rightarrow \mathbf{x}'$) between a pair of states ($\mathbf{x}$, $\mathbf{x}'$) which defines the probability of going from current state $\mathbf{x}$ to new state $\mathbf{x}'$. A state is given by assignments to variables.
  – Note T will have n^2 entries if $\mathbf{X}$ can take n values
  – Can be viewed as a matrix
- Homogeneous Markov Chain
  – Transition probability does not change over time
- Grasshopper Example
  – State: 9 integers from -4 to +4
  – Initial position: 0
  – At each instance, T(i→i) =.5, T(i→i-1) = .25, T (i→i+1) = .25
  – At two ends, can not jump beyond (stays in the same state)
    - T(4 → 4)= .75
  – Write as a transition matrix



$$P^{(t+1)}(\boldsymbol{X}^{(t+1)} = \boldsymbol{x}') = \sum_{\boldsymbol{x} \in Val(\boldsymbol{X})} P^{(t)}(\boldsymbol{X}^{(t)} = \boldsymbol{x}) \mathcal{T}(\boldsymbol{x} \rightarrow \boldsymbol{x}').$$

At t=0, $P(X^0 = 0) = 1$

At t =1, $P(X^1 = 0) = .5$, $P(X^1 = 1) = .25$, $P(X^1 = -1) = .5$

At t =2, $P(X^2 = 0) = .5 \text{x} .5 + .25 \text{ x} .25 + .25 \text{x} .25 = .375$

$P(X^2 = 1 \text{ or } -1) = .5 \text{ x } .25 + .25 \text{ x } .5 = .25$

$P(X^2 = 2 \text{ or } -2) = ..25 \text{ x } .25 = .0625$

Position probability converges to a nearly uniform distribution with time for this example

MCMC Sampling

- Generate a chain by sampling from the distribution
  - Sample $x^{(t)}$ from distribution $P^{(t)}$
  - Does $P^{(t)}$ converge and if so, to the desired distribution

Algorithm 12.5 Generating a Markov chain trajectory

Procedure MCMC-Sample (
$P^{(0)}(\boldsymbol{X})$, // Initial state distribution
$\mathcal{T}$, // Markov chain transition model
T // Number of time steps
)

Sample $\boldsymbol{x}^{(0)}$ from $P^{(0)}(\boldsymbol{X})$
for $t = 1, \ldots, T$
Sample $\boldsymbol{x}^{(t)}$ from $\mathcal{T}(\boldsymbol{x}^{(t-1)} \rightarrow \boldsymbol{X})$
return $\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(T)}$

Stationary Distribution

- At convergence, we expect:

$$P^{(t)}(\boldsymbol{x}') \approx P^{(t+1)}(\boldsymbol{x}') = \sum_{\boldsymbol{x} \in Val(\boldsymbol{X})} P^{(t)}(\boldsymbol{x}) \mathcal{T}(\boldsymbol{x} \to \boldsymbol{x}')..$$

- Stationary Distribution

A distribution $\pi(\boldsymbol{X})$ *is a* stationary distribution *for a Markov chain* $\mathcal{T}$ *if it satisfies:*

$$\pi(\boldsymbol{X} = \boldsymbol{x}') = \sum_{\boldsymbol{x} \in Val(\boldsymbol{X})} \pi(\boldsymbol{X} = \boldsymbol{x}) \mathcal{T}(\boldsymbol{x} \to \boldsymbol{x}').$$

A stationary distribution is also called an *invariant distribution.*

- In linear algebra formulation: T $\pi(x) = \pi(x)$; *i.e.* the stationary distribution is an eigenvector of the transition matrix with eigenvalue = 1

Next Class

- Read sections 12.3, 6.2, 15.1 of the KF book