Graphical Models: Approximate Inference

Course Webpage:
http://www.cs.cornell.edu/Courses/cs4786/2016sp/
Belief Propagation or Message Passing

- Each node passes messages to its parents and children
- Guaranteed to work on a tree
- To get marginal of node $X_v$ (given evidence)
  \[
  \sum_{\text{Parents}(X_v)} \left( \prod_{\text{messages to } X_v} \right) \times P(X_v|\text{Parent}(X_v))
  \]
- For general graphs, belief propagation need not work
- Inference for general graphs can be computationally hard

Can we perform inference approximately?
What is Approximate Inference?

- Obtain \( \hat{P}(X_v|\text{Observation}) \) that is close to \( P(X_v|\text{Observation}) \)
  - Additive approximation:
    \[
    |\hat{P}(X_v|\text{Observation}) - P(X_v|\text{Observation})| \leq \epsilon
    \]
  - Multiplicative approximation:
    \[
    (1 - \epsilon) \leq \frac{\hat{P}(X_v|\text{Observation})}{P(X_v|\text{Observation})} \leq (1 + \epsilon)
    \]
Two approaches:

- Inference via sampling:
  generate instances from the model, compute marginals

- Use exact inference but move to a close enough simplified model
Inference Via Sampling

- Law of large numbers: empirical distribution using large samples approximates the true distribution

- Some approaches:
  - Rejection sampling: sample all the variables, retain only ones that match evidence
  - Importance sampling: Sample from a different distribution but then apply correction while computing empirical marginals
  - Gibbs sampling: iteratively sample from distributions closer and closer to the true one
Example:

\[ |\hat{P}(X_v = 1) - P(X_v = 1)| \approx \frac{1}{\sqrt{\text{# of samples}}} \]
Algorithm:

Topologically sort variables (parents first children later)
For $t = 1$ to $n$
    For $i = 1$ to $N$
        Sample $x_i^t \sim P(X_i|X_1 = x_1^t, \ldots, X_{i-1} = x_{i-1}^t)$
    End For
End For
Throw away $x^t$'s that do not match observations
Example:

What about $|\hat{P}(X_v = 1|\text{observation}) - P(X_v = 1|\text{observation})|$?
Example: Likelihood weighting
**Importance Sampling**

**Likelihood weighting:**

Topologically sort variables (parents first children later)

For $t = 1$ to $n$
  
  Set $w_t = 1$
  
  For $i = 1$ to $N$
    
    If $X_i$ is observed, set $w_t \leftarrow w_t \cdot P(X_i = x_i | X_1 = x^t_1, \ldots, X_{i-1} = x^t_{i-1})$
    
    Else, sample $x^t_i \sim P(X_i | X_1 = x^t_1, \ldots, X_{i-1} = x^t_{i-1})$
  
  End For

End For

To compute $P(Variable|Observation)$ set,

$$P(Variable = value|Observation) = \frac{\sum_{t=1}^{n} w_t 1\{Variable = value\}}{\sum_{t=1}^{n} w_t}$$
More generally importance sampling is given by:

- Draw $x_1, \ldots, x_n \sim Q$

Notice that

$$
\mathbb{E}_{X \sim P}[f(X)] = \mathbb{E}_{X \sim Q}
\left[
\frac{P(X)}{Q(X)} f(X)
\right]
$$

Hence,

$$
\hat{P}(X = x) \approx \frac{1}{n} \sum_{t=1}^{n} 1\{x_t = x\} \frac{P(X = x)}{Q(X = x)}
$$

Idea draw samples from $Q$ but re-weight them
Gibbs Sampling

- Fix values of observed variables $v$ to the observations ($x_v^1 = x_v$)
- Randomly initialize all other variables $u$ by randomly sampling $x_u^1$
- For $t = 2$ to $n$
  - For $i = 1$ to $N$
    - If $X_i$ is observed set
      $$x_i^t = x_i^{t-1}$$
    - Else sample $x_i^{t+1}$ from
      $$x_i^{t+1} \sim P(X_i|X_1 = x_1^{t+1}, \ldots, X_{i-1} = x_{i-1}^{t+1}, X_{i+1} = x_{i+1}^{t}, \ldots, X_N = x_N^t)$$
  - End For
- End For

Take $(x_1^n, \ldots, x_N^n)$ as one sample and repeat
Gibbs sampling belongs to a class of methods called Markov Chain Monte Carlo methods.

We start by sampling from some simple distribution.

Set up a Markov chain whose stationary distribution is the target distribution.

That is, based on previous sample (state) we transit to the next state, and then to the next state and so on.

If the transition probabilities are set up right, after multiple transitions, our sample looks like one from target distribution.