Readings:
K&F: 10.1, 10.2, 10.3 (Particle Based Approximate Inference)

# Approximate Inference by Sampling

Graphical Models – 10708
Ajit Singh
Carnegie Mellon University

November 10th, 2008

# What you've learned so far

- VE & Junction Trees
  - Exact inference
  - Exponential in tree-width
- Belief Propagation, Mean Field
  - Approximate inference for marginals/conditionals
  - Fast, but can get inaccurate estimates
- Sample-based Inference
  - Approximate inference for marginals/conditionals
  - With "enough" samples, will converge to the right answer (or a high accuracy estimate)

*(If you want to be cynical, replace "enough" with "ridiculously many")*

# Goal

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

$$E_P[f] \approx \frac{1}{M}\sum_{m=1}^{M} f(\mathbf{x}[m]) \qquad \mathbf{x}[i] \sim P(\mathbf{X})$$

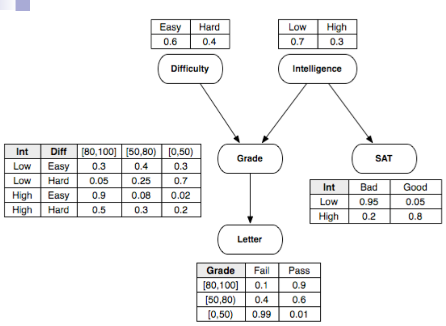
$$P(\mathbf{X}=\mathbf{x}) \approx \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}(\mathbf{x}[m]=\mathbf{x})$$

**Discrete Random Variables:** $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$

**Number of samples from P(X):** $M$



# 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
- Assignment to parents selects P(X|Pa(X))
- End result is one sample from P(X)
- Repeat to get more samples

**D** $\mathbf{x}[m, D] \sim (Easy : 0.6, Hard : 0.4)$ **D = Easy**

**I** $\mathbf{x}[m, I] \sim (Low : 0.7, High : 0.3)$ **I = High**

**G** $\mathbf{x}[m, G|D=d, I=i] \sim ([80,100] : 0.9, [50,80) : 0.08, [0,50) : 0.02)$ **G = [80,100]**

**S** $\mathbf{x}[m, S|I=i] \sim (Bad : 0.2, Good : 0.8)$ **S = Bad**

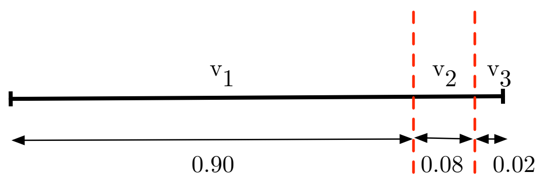
**L** $\mathbf{x}[m, L|G=g] \sim (Fail : 0.1, Pass : 0.9)$ **L = Pass**

# Multinomial Sampling

- Given an assignment to its parents, $X_i$ is a multinomial random variable.

$$\mathbf{x}[m, G | D = d, I = i] \sim (v_1 : 0.9, v_2 : 0.08, v_3 : 0.02)$$

U ~ Unif[0,1]






# Sample-based probability estimates

- Have a set of *M* samples from P(X)
- Can estimate any probability by counting records:

**Marginals:**

$$\hat{P}(D = \text{Easy}, S = \text{Bad}) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}(x[m, D] = \text{Easy}, x[m, S] = \text{Bad})$$

**Conditionals:**

$$\hat{P}(D = \text{Easy} | S = \text{Bad}) = \frac{\sum_{m=1}^{M} \mathbf{1}(\mathbf{x}[m, D] = \text{Easy}, \mathbf{x}[m, S] = \text{Bad})}{\sum_{m=1}^{M} \mathbf{1}(\mathbf{x}[m, S] = \text{Bad})}$$

*Rejection sampling*: once the sample and evidence disagree, throw away the sample.

*Rare events:* If the evidence is unlikely, i.e., P(E = e) small, then the sample size for P(X|E=e) is low

# Sample Complexity

- In many cases the probability estimate is the sum of indicator (Bernoulli) random variables:
  - Forward sampling for marginal probabilities.
  - Rejection sampling for conditional probabilities.
- The indicators are independent and identically distributed

**Additive Chernoff:** $P(P(\mathbf{x}) - \epsilon < \hat{P}(\mathbf{x}) < P(\mathbf{x}) + \epsilon) \leq 2e^{-2M\epsilon^2}$
(absolute error)

**Multiplicative Chernoff:** $P(\hat{P}(\mathbf{x}) < (1+\epsilon)P(\mathbf{x})) \leq 2e^{-M \cdot P(\mathbf{x})\epsilon^2/3}$
(relative error)

Bound the r.h.s. by δ and solve for M.

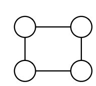
Reducing relative error is hard if P(x) is small.

$P(\mathbf{x})$ can be replaced by any marginal or conditional estimated by the sum of iid Bernoullis



# Importance Sampling

- Limitations of forward and rejection sampling
  - What if the evidence is a rare event ?
    - Either accept low accuracy estimate, or sample a lot more.
  - What if the model has no topological ordering ?
    - Bayesian networks always have a T.O.
    - Tree Markov Random Fields have a T.O.
    - Arbitrary undirected graphical models may not have a T.O.
      - Hard to sample from P(X).
- Importance sampling addresses these issues.



# Importance Sampling

- Want to estimate P(X)
- Basic idea: pick Q(X) such that
  - KL(P||Q) is small.
  - Dominance: Q(x) > 0 whenever P(x) > 0.
  - Sampling from Q is easier than sampling from P.

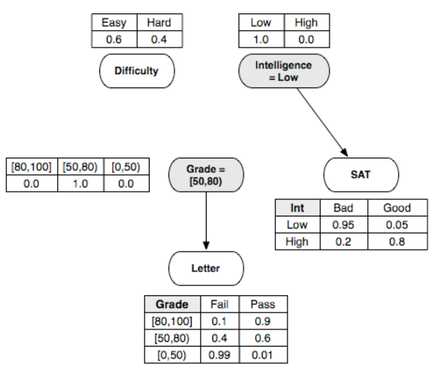
$$E_{P(X)}[f(X)] \approx \frac{1}{M}\sum_{m=1}^{M} f(\mathbf{x}[m])\frac{P(\mathbf{x}[m])}{Q(\mathbf{x}[m])}$$

Assumes it's easy to evaluate P(x)

$$f(\mathbf{X}) = \mathbf{1}(\mathbf{x}[m, D] = \text{Easy}) \implies$$
$$E_{P(\mathbf{X})}[f(\mathbf{X})] = P(D = \text{Easy})$$



# Mutilated Proposal Q(X)




- Fix the evidence distributions.
- Cut edges so that observed nodes have no parents.

**Unlike forward sampling, we do not throw away samples = less work.**

**If Q is good, then the variance of the estimates is lower than forward or rejection sampling.**

**Variance of the estimates reduces at a rate of 1/M.**

# Importance Sampling

- Can be generalized to deal with MRFs, where we can only easily get unnormalized probabilities.
  - Gibbs sampling is more common in undirected models.
  - Importance sampling yields a priori bounds on the sample complexity.

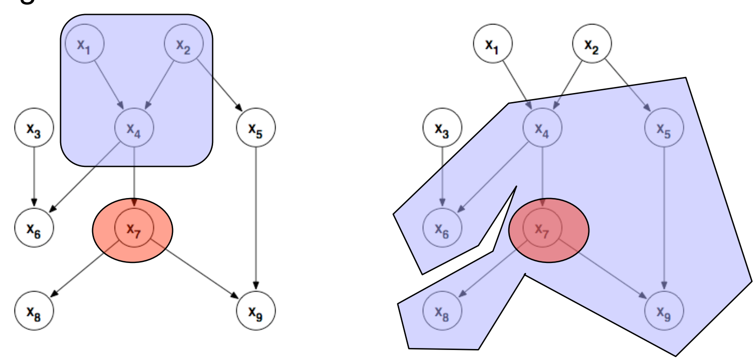


# Limitation of Forward Samplers

- Forward sampling, rejection sampling, and importance sampling are all *forward samplers*
  - Fixing an evidence node only allows it to directly affect its descendents.

# 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.




**Forward Sampler** **Markov Blanket Sampler**



# Gibbs Sampling

- We will focus on Gibbs Sampling
  - The most common Markov Blanket sampler
  - Works for directed and undirected models
  - Exploits independencies in graphical models
  - A common form of Markov Chain Monte Carlo

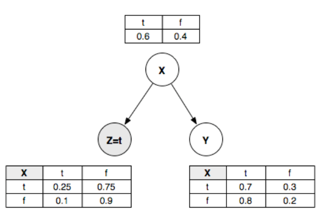
# Gibbs Sampling

1. Let **X** be the non-evidence variables
2. Generate an initial assignment $\xi^{(0)}$
3. For t = 1..MAXITER
   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 $\xi^{(t)}$
      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** - $MB(X_i)$ - $\{X_i\}$

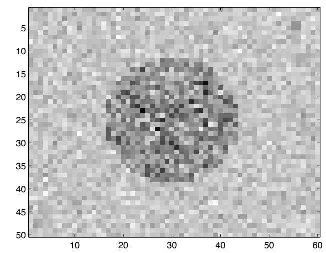


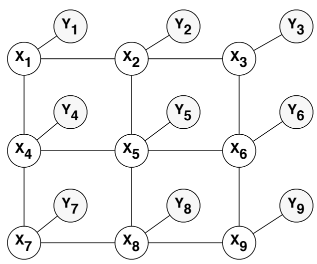

$$\mathbf{P(X|Y = y)} = \frac{P(X, Y = y)}{P(Y = y)}$$

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

**P(Y|X = x) =** CPT Lookup

# (Simple) Image Segmentation




- Noisy grayscale image.
- Label each pixel as on/off.
- Model using a pairwise MRF.

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

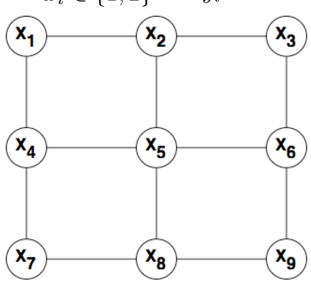
$$\Phi(x_i) = exp\left\{-\frac{(y_i - \mu_{x_i})^2}{2\sigma_{x_i}^2}\right\}$$

$$\Psi(x_i, x_j) = \exp\left\{-\beta(x_i - x_j)^2\right\}$$



# Gibbs Sampling

$x_i \in \{1,2\}$ $y_i \in \mathbb{R}$




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

$$\Phi(x_i) = exp\left\{-\frac{(y_i - \mu_{x_i})^2}{2\sigma_{x_i}^2}\right\}$$

$$\Psi(x_i, x_j) = \exp\left\{-\beta(x_i - x_j)^2\right\}$$

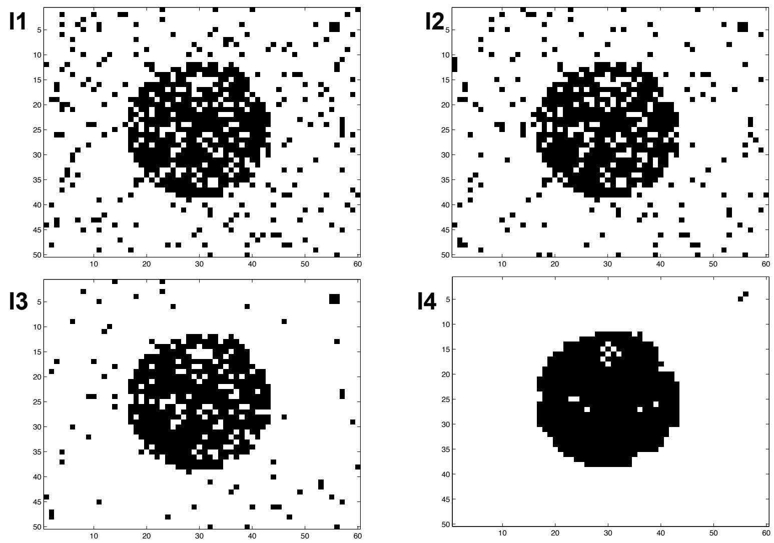
$$P(x_i | x_1, \dots x_{i-1}, x_{i+1}, \dots, x_n) =$$

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

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

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

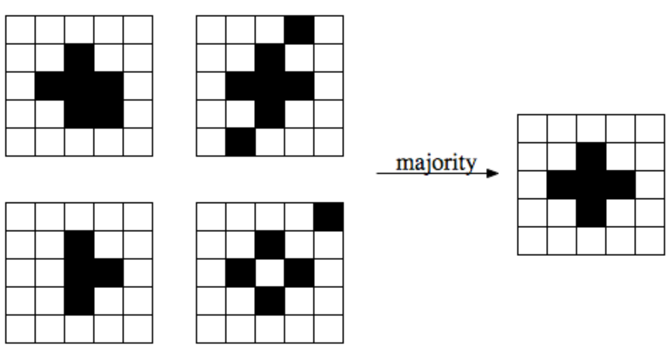
# Gibbs Sampling






# MAP by Sampling

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



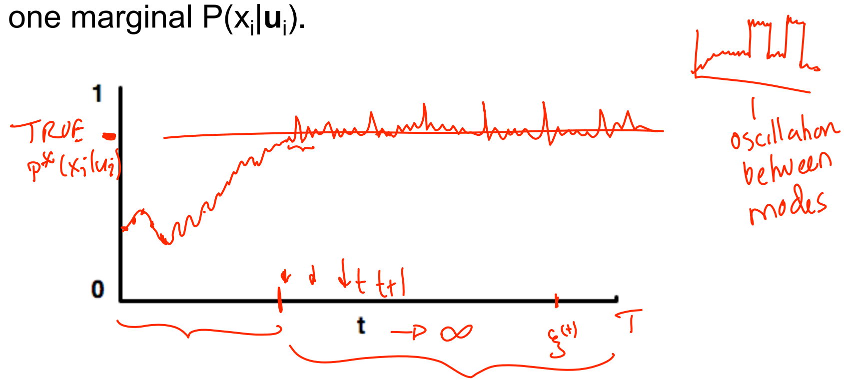
# Markov Chain Interpretation

- The state space consists of assignments to X.
- $P(x_i \mid \mathbf{u}_i)$ 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 !



# Convergence

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




- Burn-in
- Thinning

# 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 $P(X_i \mid \text{rest})$.
    - Markov chain interpretation.
    - Samples are independent when the Markov chain converges.
    - Convergence heuristics, burn-in, thinning.