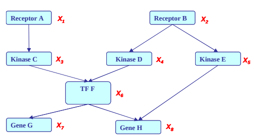
School of Computer Science
Carnegie Mellon

Approximate Inference: Monte Carlo Inference

Probabilistic Graphical Models (10-708)

Lecture 18, Nov 19, 2007



Eric Xing

Reading: J-Chap. 1, KF-Chap. 11

Monte Carlo methods

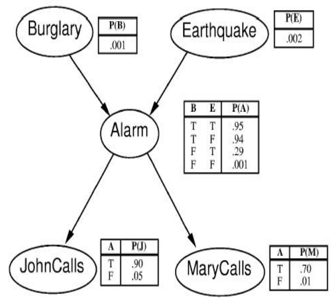
- Draw random samples from the desired distribution
- Yield a stochastic representation of a complex distribution
  - marginals and other expections can be approximated using sample-based averages

$$E[f(x)] = \frac{1}{N}\sum_{t=1}^{N} f(x^{(t)})$$

- **Asymptotically** exact and easy to apply to arbitrary models
- Challenges:
  - how to draw samples from a given dist. (not all distributions can be trivially sampled)?
  - how to make better use of the samples (not all sample are useful, or eqally useful, see an example later)?
  - how to know we've sampled enough?

Example: naive sampling

- Construct samples according to probabilities given in a BN.



Alarm example: (Choose the right sampling sequence)
1) Sampling:P(B)=<0.001, 0.999> suppose it is false, B0. Same for E0. P(A|B0, E0)=<0.001, 0.999> suppose it is false...
2) Frequency counting: In the samples right,
P(J|A0)=P(J,A0)/P(A0)=<1/9, 8/9>.

E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J1
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E1	B0	A1	M1	J1
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0

Example: naive sampling

- Construct samples according to probabilities given in a BN.

Alarm example: (Choose the right sampling sequence)

3) what if we want to compute P(J|A1) ?
we have only one sample ...
P(J|A1)=P(J,A1)/P(A1)=<0, 1>.

4) what if we want to compute P(J|B1) ?
No such sample available!
P(J|A1)=P(J,B1)/P(B1) can not be defined.

For a model with hundreds or more variables, rare events will be very hard to garner evough samples even after a long time or sampling ...

E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J1
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E1	B0	A1	M1	J1
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0

Monte Carlo methods (cond.)

- Direct Sampling
  - We have seen it.
  - Very difficult to populate a high-dimensional state space
- Rejection Sampling
  - Create samples like direct sampling, only count samples which is consistent with given evidences.
- Likelihood weighting, ...
  - Sample variables and calculate evidence weight. Only create the samples which support the evidences.
- Markov chain Monte Carlo (MCMC)
  - Metropolis-Hasting
  - Gibbs

Rejection sampling

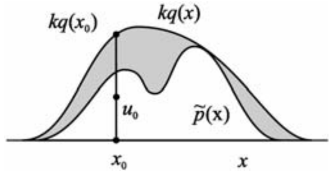
- Suppose we wish to sample from dist. $\Pi(X)=\Pi'(X)/Z$.
  - $\Pi(X)$ is difficult to sample, but $\Pi'(X)$ is easy to evaluate
  - Sample from a simpler dist $Q(X)$
  - Rejection sampling

$$x^* \sim Q(X), \qquad \text{accept } x^* \text{ w.p. } \Pi'(x^*)/kQ(x^*)$$

  - Correctness:

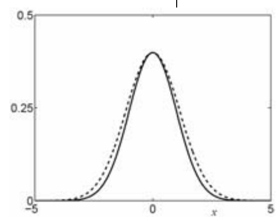
$$p(x) = \frac{[\Pi'(x)/kQ(x)]Q(x)}{\int [\Pi'(x)/kQ(x)]Q(x)dx} = \frac{\Pi'(x)}{\int \Pi'(x)dx} = \Pi(x)$$

  - Pitfall …

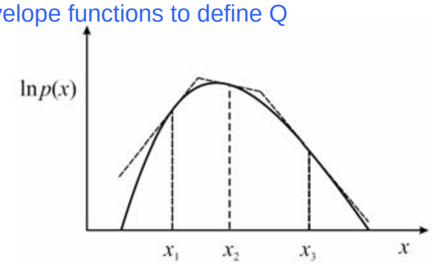


Rejection sampling

- Pitfall:
  - Using $Q=\mathcal{N}(\mu,\sigma_q I)$ to sample $P=\mathcal{N}(\mu,\sigma_p I)$
  - If σ_q exceeds σ_p by 1%, and dimensional=1000,
  - The optimal acceptance rate $k=(\sigma_q/\sigma_p)^d \approx 1/20{,}000$
  - Big waste of samples!

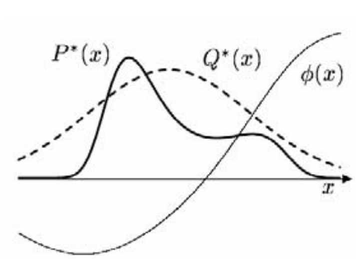


- Adaptive rejection sampling
  - Using envelope functions to define Q



Unnormalized importance sampling

- Suppose sampling from $P(\cdot)$ is hard.
- Suppose we can sample from a "simpler" proposal distribution $Q(\cdot)$ instead.
- If Q dominates P (i.e., $Q(x) > 0$ whenever $P(x) > 0$), we can sample from Q and reweight:



$$\begin{aligned}\langle f(X)\rangle &= \int f(x)P(x)dx \\ &= \int f(x)\frac{P(x)}{Q(x)}Q(x)dx \\ &\approx \frac{1}{M}\sum_m f(x^m)\frac{P(x^m)}{Q(x^m)} \quad \text{where } x^m \sim Q(X) \\ &= \frac{1}{M}\sum_m f(x^m)w^m\end{aligned}$$

Normalized importance sampling

- Suppose we can only evaluate $P'(x) = \alpha P(x)$ (e.g. for an MRF).
- We can get around the nasty normalization constant α as follows:
  - Let $r(X) = \frac{P'(x)}{Q(x)}$ $\quad \Rightarrow \quad \langle r(X) \rangle_Q = \int \frac{P'(x)}{Q(x)} Q(x) dx = \int P'(x) dx = \alpha$
- Now

$$
\begin{aligned}
\langle f(X) \rangle_P &= \int f(x) P(x) dx = \frac{1}{\alpha} \int f(x) \frac{P'(x)}{Q(x)} Q(x) dx \\
&= \frac{\int f(x) r(x) Q(x) dx}{\int r(x) Q(x) dx} \\
&\approx \frac{\sum_m f(x^m) r^m}{\sum_m r^m} \quad \text{where } x^m \sim Q(X) \\
&= \sum_m f(x^m) w^m \quad \text{where } w^m = \frac{r^m}{\sum_m r^m}
\end{aligned}
$$

Normalized vs unnormalized importance sampling

- Unormalized importance sampling is unbiased:

$$E_Q\left[f(X)w(X)\right] =$$

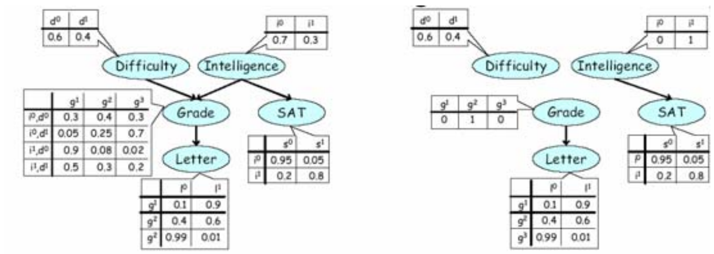
- Normalized importance sampling is biased, eg for M = 1:

$$E_Q\left[\frac{f(x^1)w(x^1)}{w(x^1)}\right] =$$

- However, the variance of the normalized importance sampler is usually lower in practice.
- Also, it is common that we can evaluate P'(x) but not P(x), e.g. P(x|e) = P'(x, e)/P(e) for Bayes net, or P(x) = P'(x)/Z for MRF.

Likelihood weighting

- We now apply normalized importance sampling to a Bayes net.
- The proposal Q is gotten from the mutilated BN where we clamp evidence nodes, and cut their incoming arcs. Call this P_M.



- The unnormalized posterior is P'(x) = P(x, e).
- So for $f(X_i) = \delta(X_i = x_i)$, we get $\hat{P}(X_i = x_i \mid e) = \frac{\sum_m w_m \delta(x_i^m = x_i)}{\sum_m w_m}$ where $w_m = P'(x^m, e) / P_M(x^m)$.

Likelihood weighting algorithm

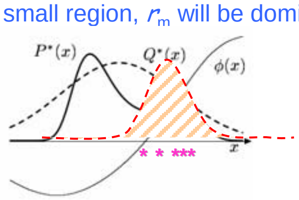
$[x_{1:n}, w]$ = function LW(CPDs, G, E)
let $X_1, \ldots, X_n$ be a topological ordering of G
$w = 1$
$x = (0, \ldots, 0)$
for $i = 1 : n$
    let $u_i = x(Pa_i)$
    if $X_i \notin E$
    then sample x_i from $P(X_i | u_i)$
    else
        $x_i = e(X_i)$
        $w = w * P(x_i | u_i)$

Efficiency of likelihood weighting

- The efficiency of importance sampling depends on how close the proposal Q is to the target P.
- Suppose all the evidence is at the roots. Then Q = P(X|e), and all samples have weight 1.
- Suppose all the evidence is at the leaves. Then Q is the prior, so many samples might get small weight if the evidence is unlikely.
- We can use arc reversal to make some of the evidence nodes be roots instead of leaves, but the resulting network can be much more densely connected.

Weighted resampling

- Problem of importance sampling: depends on how well Q matches P
  - If P(x)f(x) is strongly varying and has a significant proportion of its mass concentrated in a small region, r_m will be dominated by a few samples



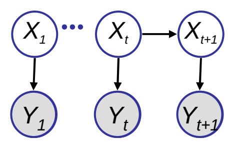
  - Note that if the high-prob mass region of Q falls into the low-prob mass region of P, the variance of $r^m = P(x^m)/Q(x^m)$ can be small even if the samples come from low-prob region of P and potentially erroneous .
- Solution
  - Use heavy tail Q.
  - Weighted resampling

$$w^m = \frac{P(x^m)/Q(x^m)}{\sum_l P(x^l)/Q(x^l)} = \frac{r^m}{\sum_m r^m}$$

Weighted resampling

- Sampling importance resampling (SIR):
  1. Draw N samples from Q: $X_1 \dots X_N$
  2. Constructing weights: $w_1 \dots w_N$, $w^m = \frac{P(x^m)/Q(x^m)}{\sum_l P(x^l)/Q(x^l)} = \frac{r^m}{\sum_m r^m}$
  3. Sub-sample x from $\{X_1 \dots X_N\}$ w.p. $(w_1 \dots w_N)$

- Particular Filtering
  - A special weighted resampler
  - Yield samples from posterior $p(X_t | Y_{1:t})$



Sketch of Particle Filters

- The starting point

$$p(X_t | \mathbf{Y}_{1:t}) = p(X_t | Y_t, \mathbf{Y}_{1:t-1}) = \frac{p(X_t | \mathbf{Y}_{1:t-1}) p(Y_t | X_t)}{\int p(X_t | \mathbf{Y}_{1:t-1}) p(Y_t | X_t) dX_t}$$

  - Thus $p(X_t | Y_{1:t})$ is represented by

$$\left\{ X_t^m \sim p(X_t | \mathbf{Y}_{1:t-1}), \quad w_t^m = \frac{p(Y_t | X_t^m)}{\sum_{m=1}^{M} p(Y_t | X_t^m)} \right\}$$

- A sequential weighted resampler
  - Time update

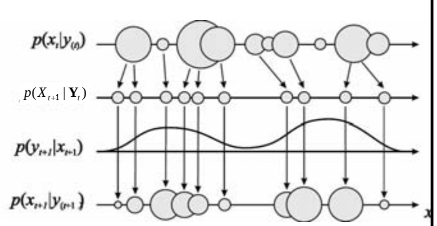
$$p(X_{t+1} | \mathbf{Y}_{1:t}) = \int p(X_{t+1} | X_t) p(X_t | \mathbf{Y}_{1:t}) dX_t$$

$$= \sum_m w_t^m p(X_{t+1} | X_t) \quad \text{(sample from a mixture model)}$$

  - Measurement update

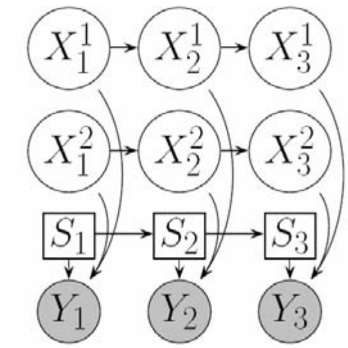
$$p(X_{t+1} | \mathbf{Y}_{1:t+1}) = \frac{p(X_{t+1} | \mathbf{Y}_{1:t}) p(Y_{t+1} | X_{t+1})}{\int p(X_{t+1} | \mathbf{Y}_{1:t}) p(Y_{t+1} | X_{t+1}) dX_{t+1}}$$

$$\Rightarrow \left\{ X_{t+1}^m \sim p(X_{t+1} | \mathbf{Y}_{1:t}), \quad w_{t+1}^m = \frac{p(Y_{t+1} | X_{t+1}^m)}{\sum_{m=1}^{M} p(Y_{t+1} | X_{t+1}^m)} \right\} \quad \text{(reweight)}$$



PF for switching SSM

- Recall that the belief state has O(2t) Gaussian modes

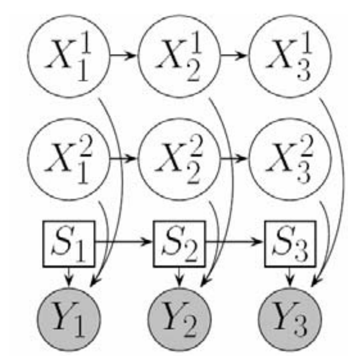


PF for switching SSM

- Key idea: if you knew the discrete states, you can apply the right Kalman filter at each time step.

- So for each old particle *m*, sample $S_t^m \sim P(S_t \mid S_{t-1}^m)$ from the prior, apply the KF (usomg parameters for S_t^m) to the old belief state $(\hat{x}^m_{t-1|t-1}, P^m_{t-1|t-1})$ to get an approximation to $P(X_t \mid y_{1:t}, s^m_{1:t})$

- Useful for online tracking, fault diagnosis, etc.



Rao-Blackwellised sampling

- Sampling in high dimensional spaces causes high variance in the estimate.
- RB idea: sample some variables X_p, and conditional on that, compute expected value of rest X_d analytically:

$$\begin{aligned} E_{p(X|e)}[f(X)] &= \int p(x_p, x_d \mid e) f(x_p, x_d) dx_p dx_d \\ &= \int_{x_p} p(x_p \mid e) \left(\int_{x_d} p(x_d \mid x_p, e) f(x_p, x_d) dx_d \right) dx_p \\ &= \int_{x_p} p(x_p \mid e) E_{p(X_d|x_p,e)}\left[f(x_p, X_d) \right] dx_p \\ &= \frac{1}{M} \sum_m E_{p(X_d|x_p^m,e)}\left[f(x_p^m, X_d) \right], \qquad x_p^m \sim p(x_p \mid e) \end{aligned}$$

- This has lower variance, because of the identity:

$$\mathrm{var}[\tau(X_p, X_d)] = \mathrm{var}[E[\tau(X_p, X_d) \mid X_p]] + E[\mathrm{var}[\tau(X_p, X_d) \mid X_p]]$$

- Hence $\mathrm{var}[E[\tau(X_p, X_d) \mid X_p]] \le \mathrm{var}[\tau(X_p, X_d)]$, so $\tau(X_p, X_d) = E[f(X_p, X_d) \mid X_p]$ is a lower variance estimator.

Markov chain Monte Carlo (MCMC)

- Importance sampling does not scale well to high dimensions.
- Rao-Blackwellisation not always possible.
- MCMC is an alternative.
- Construct a Markov chain whose stationary distribution is the target density $= P(X|e)$.
- Run for T samples (burn-in time) until the chain converges/mixes/reaches stationary distribution.
- Then collect M (correlated) samples x_m .
- Key issues:
  - Designing proposals so that the chain mixes rapidly.
  - Diagnosing convergence.

Markov Chains

- **Definition:**
  - Given an n-dimensional state space
  - Random vector $\mathbf{X} = (x_1,\ldots,x_n)$
  - $\mathbf{x}^{(t)} = \mathbf{x}$ at time-step t
  - $\mathbf{x}^{(t)}$ transitions to $\mathbf{x}^{(t+1)}$ with prob

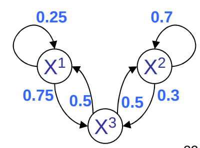
$$P(\mathbf{x}^{(t+1)} \mid \mathbf{x}^{(t)},\ldots,\mathbf{x}^{(1)}) = T(\mathbf{x}^{(t+1)} \mid \mathbf{x}^{(t)}) = T(\mathbf{x}^{(t)} \rightarrow \mathbf{x}^{(t+1)})$$

- **Homogenous**: chain determined by state $\mathbf{x}^{(0)}$, fixed *transition kernel* T (rows sum to 1)
- **Equilibrium**: $\pi(\mathbf{x})$ is a *stationary (equilibrium) distribution* if

$$\pi(\mathbf{x}') = \Sigma_{\mathbf{x}} \pi(\mathbf{x})\ T(\mathbf{x} \rightarrow \mathbf{x}').$$

i.e., is a left eigenvector of the transition matrix $\pi^T T = \pi^T T$.

$$\begin{pmatrix}0.2 & 0.5 & 0.3\end{pmatrix} = \begin{pmatrix}0.2 & 0.5 & 0.3\end{pmatrix}\begin{pmatrix}0.25 & 0 & 0.75\\ 0 & 0.7 & 0.3\\ 0.5 & 0.5 & 0\end{pmatrix}$$



Markov Chains

- An MC is ***irreducible*** if transition graph connected
- An MC is ***aperiodic*** if it is not trapped in cycles
- An MC is ***ergodic*** (regular) if you can get from state x to x' in a finite number of steps.
- **Detailed balance**: prob($\mathbf{x}^{(t)} \rightarrow \mathbf{x}^{(i-1)}$) = prob($\mathbf{x}^{(t-1)} \rightarrow \mathbf{x}^{(t)}$)

$$p(\mathbf{x}^{(t)}) T(\mathbf{x}^{(t-1)} \mid \mathbf{x}^{(t)}) = p(\mathbf{x}^{(t-1)}) T(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)})$$

summing over $\mathbf{x}^{(t-1)}$

$$p(\mathbf{x}^{(t)}) = \sum_{\mathbf{x}^{(t-1)}} p(\mathbf{x}^{(t-1)}) T(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)})$$

- Detailed bal → stationary dist exists

Metropolis-Hastings

- Treat the target distribution as stationary distribution
- Sample from an easier proposal distribution, followed by an acceptance test
- This induces a transition matrix that satisfies detailed balance
  - MH proposes moves according to $Q(x'|x)$ and accepts samples with probability $A(x'|x)$.
  - The induced transition matrix is $T(x \rightarrow x') = Q(x'|x)A(x'|x)$
  - Detailed balance means

$$\pi(x)Q(x'|x)A(x'|x) = \pi(x')Q(x|x')A(x|x')$$

  - Hence the acceptance ratio is

$$A(x'|x) = \min\left(1, \frac{\pi(x')Q(x|x')}{\pi(x)Q(x'|x)}\right)$$

Metropolis-Hastings

1. Initialize $x^{(0)}$
2. While not mixing // burn-in
   - $x = x^{(t)}$
   - t += 1,
   - sample $u \sim$ Unif(0,1)
   - sample $x^* \sim Q(x^*|x)$
     - if $u < A(x^*|x) = \min\left(1, \frac{\pi(x^*)Q(x|x^*)}{\pi(x)Q(x^*|x)}\right)$
       - $x^{(t)} = x^*$ // transition
     - else
       - $x^{(t)} = x$ // stay in current state

$$\left.\begin{array}{l} \\ \\ \\ \end{array}\right\} \text{Function Draw sample } (x(t))$$

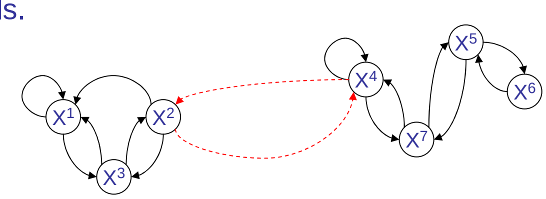
- Reset t=0, for t =1:N
  - x(t+1)) ← Draw sample (x(t))

Mixing time

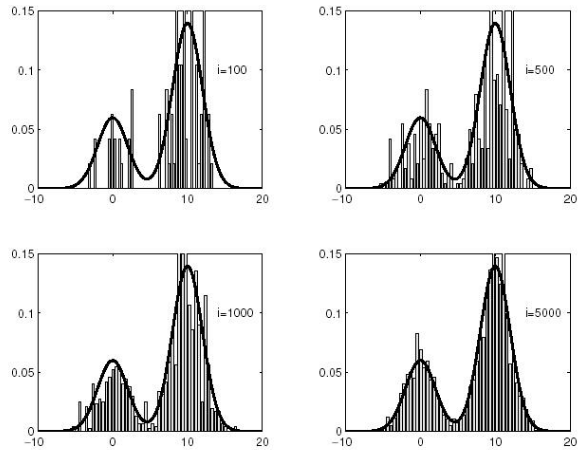
- The ε mixing time T_ε is the minimal number of steps (from any starting distribution) until $D_{var}(P^{(T)}, \pi) \le \varepsilon$, where D_{var} is the variational distance between the two distance:

$$D_{\text{var}}(\mu_1, \mu_2) \overset{\text{def}}{=} \sup_{A \subset S} |\mu_1(A) - \mu_2(A)|$$

- Chains with low bandwidth (conductance) regions of space take a long time to mix.
- This arises for GMs with deterministic or highly skewed potentials.



MCMC example



$q(x^*|x) \sim N(x^{(i)},100)$

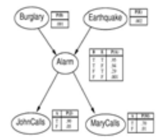
$p(x) \sim 0.3 \exp(-0.2x^2) + 0.7 \exp(-0.2(x-10)^2)$

Summary of MH

- Random walk through state space
- Can simulate multiple chains in parallel
- Much hinges on proposal distribution Q
  - Want to visit state space where $p(X)$ puts mass
  - Want $A(x^*|x)$ high in modes of $p(X)$
  - Chain mixes well
- Convergence diagnosis
  - How can we tell when burn-in is over?
  - Run multiple chains from different starting conditions, wait until they start "behaving similarly".
  - Various heuristics have been proposed.

Gibbs sampling

- Gibbs sampling is an MCMC algorithm that is especially appropriate for inference in graphical models.

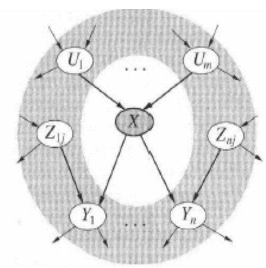


- The procedure
  - we have variable set $\boldsymbol{X}=\{x_1, x_2, x_3, \ldots x_N\}$ for a GM
  - at each step one of the variables X_i is selected (at random or according to some fixed sequences), denote the remaining variables as $\boldsymbol{X}_{-i}$, and its current value as $\boldsymbol{x}_{-i}^{(t-1)}$
    - Using the "alarm network" as an example, say at time t we choose X_E, and we denote the current value assignments of the remaining variables, $\boldsymbol{X}_{-E}$, obtained from previous samples, as $x_{-E}^{(t-1)} = \left\{x_B^{(t-1)}, x_A^{(t-1)}, x_J^{(t-1)}, x_M^{(t-1)}\right\}$
  - the conditonal distribution $p(X_i | \boldsymbol{x}_{-i}^{(t-1)})$ is computed
  - a value $\boldsymbol{x}_i^{(t)}$ is sampled from this distribution
  - the sample $\boldsymbol{x}_i^{(t)}$ replaces the previous sampled value of X_i in $\boldsymbol{X}$.
    - i.e., $x^{(t)} = x_{-E}^{(t-1)} \cup x_E^{(t)}$

Markov Blanket

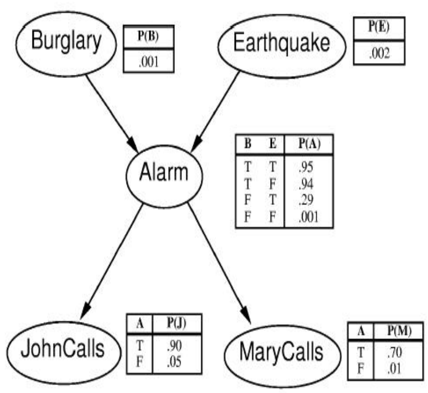
- Markov Blanket in BN
  - A variable is independent from others, given its parents, children and children's parents (d-separation).
- MB in MRF
  - A variable is independent all its non-neighbors, given all its direct neighbors.

  $\Rightarrow p(X_i | X_{-i}) = p(X_i | \mathrm{MB}(X_i))$



- Gibbs sampling
  - Every step, choose one variable and sample it by P(X|MB(X)) based on previous sample.

Gibbs sampling of the alarm network



MB(A)={B, E, J, M}

MB(E)={A, B}

- To calculate P(J|B1,M1)
- Choose (B1,E0,A1,M1,J1) as a start
- **Evidences** are B1, M1, **variables** are A, E, J.
- Choose next variable as A
- Sample A by P(A|MB(A))=P(A|B1, E0, M1, J1) suppose to be false.
- (B1, E0, A0, M1, J1)
- Choose next random variable as E, sample E~P(E|B1,A0)
- ...

Gibbs sampling

- Gibbs sampling is a special case of MH
- The transition matrix updates each node one at a time using the following proposal:

$$Q((x_i,\mathbf{x}_{-i})\to(x_i',\mathbf{x}_{-i})) = p(x_i'|\mathbf{x}_{-i})$$

- This is efficient since for two reasons
  - It leads to samples that is always accepted

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

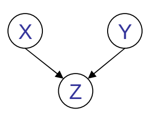
$$= \min\left(1, \frac{p(x_i'|\mathbf{x}_{-i})p(\mathbf{x}_{-i})p(x_i|\mathbf{x}_{-i})}{p(x_i|\mathbf{x}_{-i})p(\mathbf{x}_{-i})p(x_i'|\mathbf{x}_{-i})}\right) = \min(1,1)$$

  Thus $$T((x_i,\mathbf{x}_{-i})\to(x_i',\mathbf{x}_{-i})) = p(x_i'|\mathbf{x}_{-i})$$

  - It is efficient since $p(x_i'|\mathbf{x}_{-i})$ only depends on the values in X_i's Markov blanket

Gibbs sampling

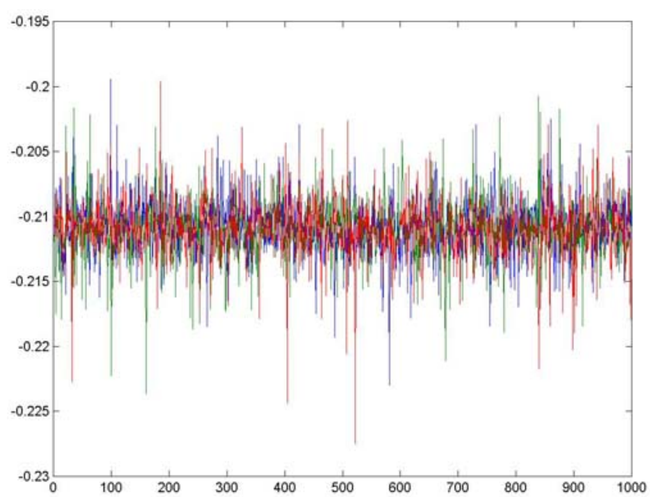
- Scheduling and ordering:
  - Sequential sweeping: in each "epoch" t, touch every r.v. in some order and yield an new sample, $x^{(t)}$, after every r.v. is resampled
  - Randomly pick an r.v. at each time step
- Blocking:
  - Large state space: state vector $\mathbf{X}$ comprised of many components (high dimension)
  - Some components can be correlated and we can sample components (i.e., subsets of r.v.,) one at a time
- Gibbs sampling can fail if there are deterministic constraint



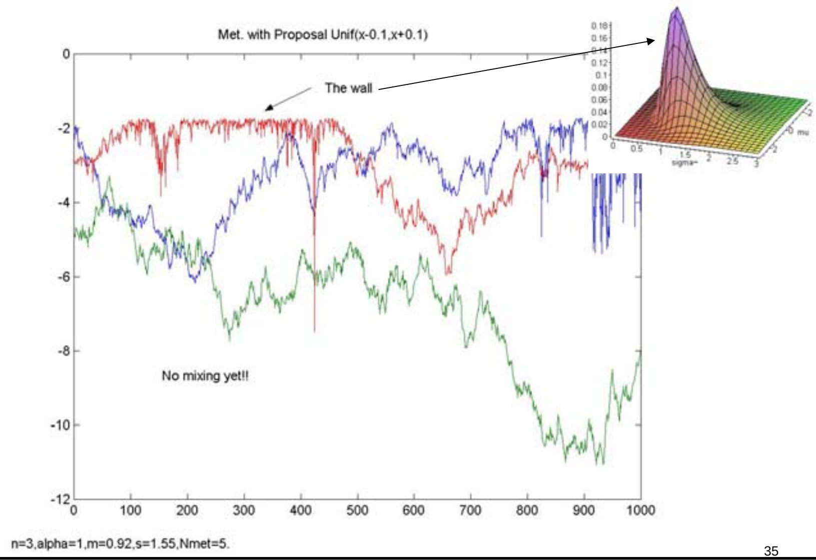
Z is xor

- Suppose we observe $Z = 1$. The posterior has 2 modes: $P(X = 1, Y = 0|Z = 1)$ and $P(X = 0, Y = 1|Z = 1)$. if we start in mode 1, $P(X|Y = 0, Z = 1)$ leaves $X = 1$, so we can't move to mode 2 (Reducible Markov chain).
- If all states have non-zero probability, the MC is guaranteed to be regular.
- Sampling blocks of variables at a time can help improve mixing.

GOOD! Chains



BAD! Chains

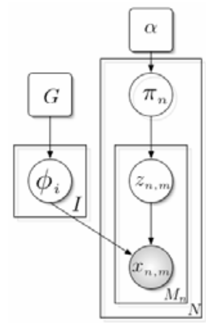


The Art of simulation

- Run several chains
- Start at over-dispersed points
- Monitor the log lik.
- Monitor the serial correlations
- Monitor acceptance ratios
- Re-parameterize (to get approx. indep.)
- Re-block (Gibbs)
- Collapse (int. over other pars.)
- Run with troubled pars. fixed at reasonable vals.

Collapsed Gibbs sampling of M³ model (Tom Griffiths & Mark Steyvers)

- Collapsed Gibbs sampling
  - Integrate out π



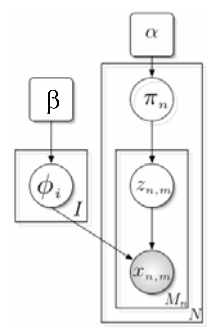
For variables $\mathbf{z} = z_1, z_2, \ldots, z_n$

Draw $z_i^{(t+1)}$ from $P(z_i|\mathbf{z}_{-i}, \boldsymbol{w})$

$\mathbf{z}_{-i} = z_1^{(t+1)}, z_2^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, \ldots, z_n^{(t)}$

Gibbs sampling

- Need full conditional distributions for variables
- Since we only sample z we need



$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto P(w_i|z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i})P(z_i = j|\mathbf{z}_{-i})$$

$$= \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

$n_j^{(w)}$ number of times word w assigned to topic j

$n_j^{(d)}$ number of times topic j used in document d

Gibbs sampling

			iteration
			1
i	w_i	d_i	z_i
1	MATHEMATICS	1	2
2	KNOWLEDGE	1	2
3	RESEARCH	1	1
4	WORK	1	2
5	MATHEMATICS	1	1
6	RESEARCH	1	2
7	WORK	1	2
8	SCIENTIFIC	1	1
9	MATHEMATICS	1	2
10	WORK	1	1
11	SCIENTIFIC	2	1
12	KNOWLEDGE	2	1
.	.	.	.
.	.	.	.
.	.	.	.
50	JOY	5	2

Gibbs sampling

			iteration	
			1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

Gibbs sampling

			iteration	
			1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

			iteration	
			1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

			iteration	
			1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	?
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

			iteration	
			1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	?
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

i	w_i	d_i	iteration 1 z_i	2 z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	1
4	WORK	1	2	?
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Eric Xing

45

Gibbs sampling

i	w_i	d_i	iteration 1 z_i	2 z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	1
4	WORK	1	2	2
5	MATHEMATICS	1	1	?
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Eric Xing

46

Gibbs sampling

			iteration			
			1	2	…	1000
i	w_i	d_i	z_i	z_i		z_i
1	MATHEMATICS	1	2	2		2
2	KNOWLEDGE	1	2	1		2
3	RESEARCH	1	1	1		2
4	WORK	1	2	2		1
5	MATHEMATICS	1	1	2		2
6	RESEARCH	1	2	2		2
7	WORK	1	2	2		2
8	SCIENTIFIC	1	1	1	…	1
9	MATHEMATICS	1	2	2		2
10	WORK	1	1	2		2
11	SCIENTIFIC	2	1	1		2
12	KNOWLEDGE	2	1	2		2
.	.	.	.	.		.
.	.	.	.	.		.
.	.	.	.	.		.
50	JOY	5	2	1		1

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Document tagging

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.