Probabilistic Graphical Models

10-708

Markov Chain Monte Carlo and Belief Propagation

Eric Xing

Lecture 17, Nov 9, 2005

Reading: MJ-Chap. 21, KF-Chap. 9

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.

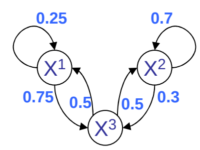
Markov Chains

- **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

    $P(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)},\ldots,\mathbf{x}^{(1)}) = T(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)}) = T(\mathbf{x}^{(t-1)} \rightarrow \mathbf{x}^{(t)})$
- **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 $\pi^T T = \pi^T T$.



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

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 $T(x \to x') = Q(x'|x)A(x'|x)$
  - 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 // burin-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

   Function
   Draw sample (x(t))

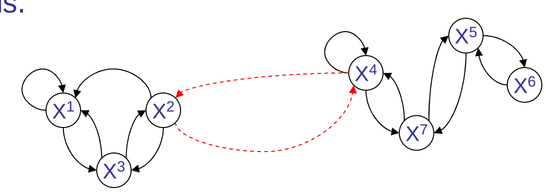
- Reset t=0, for t =1:N
  - x(t+1)) ← Draw sample (x(t))

Mixing time

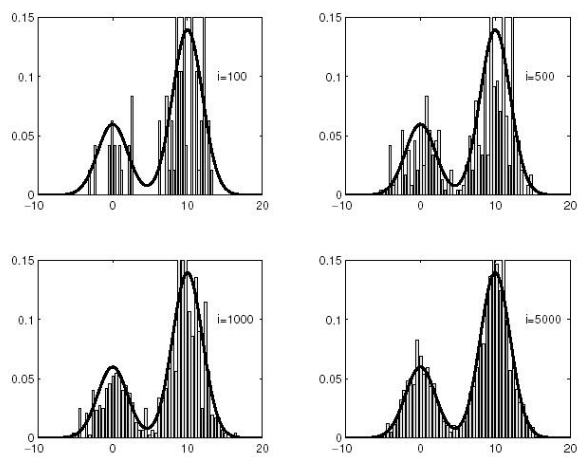
- The ε mixing time T_ε is the minimal number of steps (from any starting distribution) until $D_{var}(P^{(T)}, \pi) \le \varepsilon$, where D_{var} is the variational distance between the two distance:

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

- 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



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

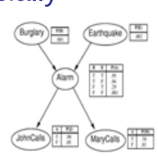
$p(x) \sim 0.3 \exp(-0.2x^2) + 0.7 \exp(-0.2(x-10)^2)$

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.

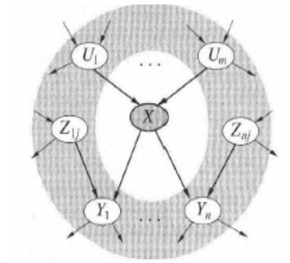


- The procedue
  - we have variable set $\mathbf{X}=\{x_1, x_2, x_3, \ldots x_N\}$ for a GM
  - at each step one of the variables X_i is selected (at random or according to some fixed sequences), denote the remaining variables as $\mathbf{X}_{-i}$, and its current value as $\mathbf{x}_{-i}^{(t-1)}$
    - 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, $\mathbf{X}_{-E}$, 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 $p(X_i| \mathbf{x}_{-i}^{(t-1)})$ is computed
  - a value $\mathbf{x}_i^{(t)}$ is sampled from this distribution
  - the sample $\mathbf{x}_i^{(t)}$ replaces the previous sampled value of X_i in $\mathbf{X}$.
    - i.e., $x^{(t)} = x_{-E}^{(t-1)} \cup x_E^{(t)}$

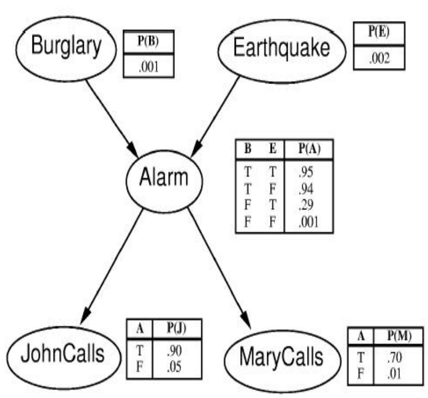
Markov Blanket

- 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.

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

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

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

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

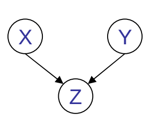
$$= \min\left(1, \frac{p(x'_i \mid \mathbf{x}_{-i})p(x_i \mid \mathbf{x}_{-i})}{p(x_i \mid \mathbf{x}_{-i})p(x'_i \mid \mathbf{x}_{-i})}\right) = \min(1,1)$$

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

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

Gibbs sampling

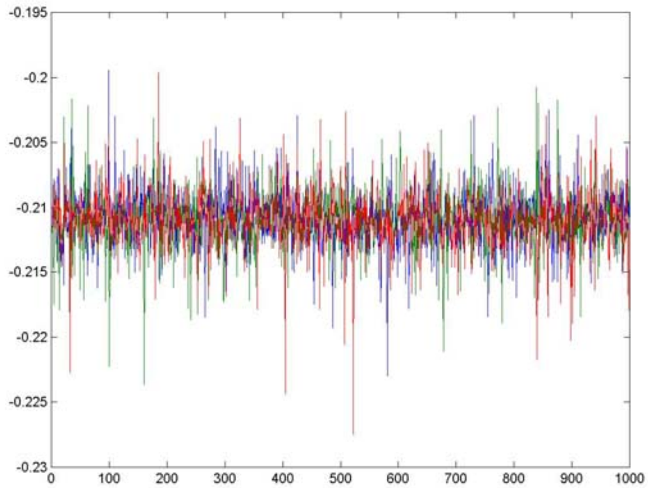
- Scheduling and ordering:
  - Sequential sweeping: in each "epoch" *t,* touch every r.v. in some order and yield an new sample, $x^{(t)}$, after every r.v. is resampled
  - Randomly pick an r.v. at each time step
- Blocking:
  - Large state space: state vector $\mathbf{X}$ comprised of many components (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



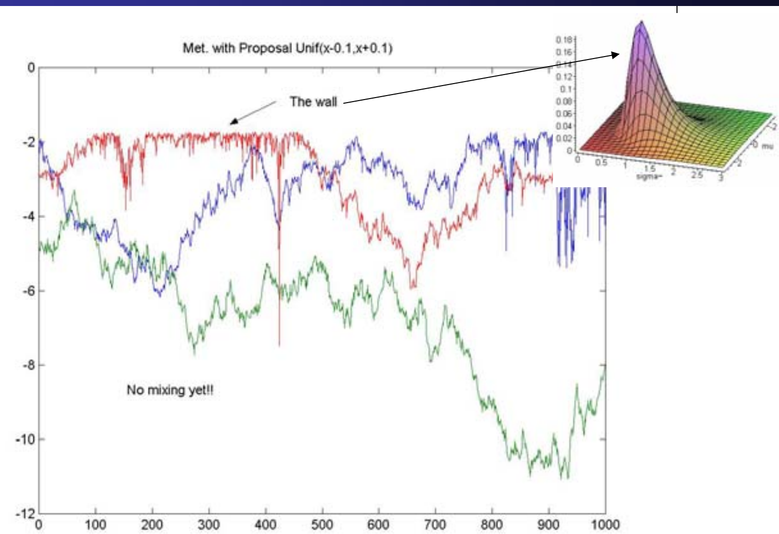
Z is xor

- Suppose we observe $Z = 1$. The posterior has 2 modes: $P(X = 1, Y = 0 \mid Z = 1)$ and $P(X = 0, Y = 1 \mid Z = 1)$. if we start in mode 1, $P(X \mid Y = 0, Z = 1)$ 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



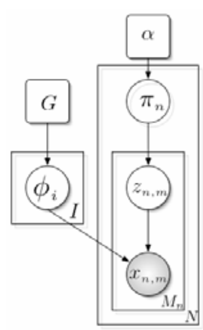
n=3,alpha=1,m=0.92,s=1.55,Nmet=5.

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



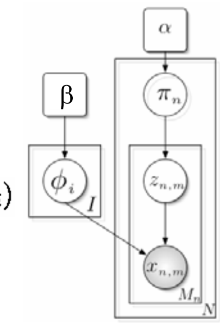
For variables $\mathbf{z} = z_1, z_2, \ldots, z_n$

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

i	w_i	d_i	iteration 1 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	

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

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	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	

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

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	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	

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

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

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

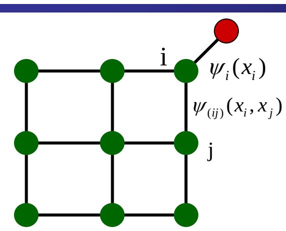
Document tagging

"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.

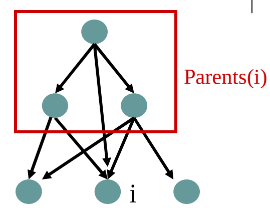
Toward variational inference

From GM to factored graphs



Undirected graph
(Markov random field)

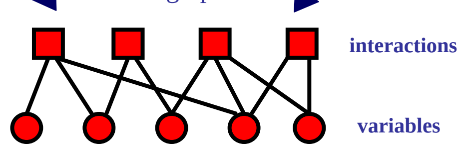
$$P(x) = \frac{1}{Z}\prod_i \psi_i(x_i) \prod_{(ij)} \psi_{(ij)}(x_i, x_j)$$



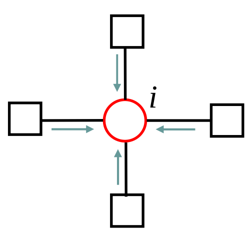
Directed graph
(Bayesian network)

$$P(x) = \prod_i P(x_i \mid x_{\text{parents}(i)})$$

factor graphs

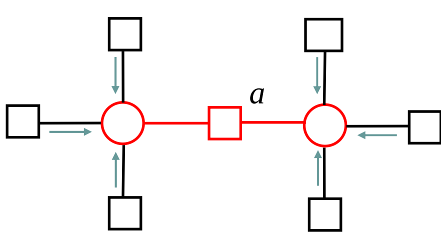


Beliefs and messages in FG



$$b_i(x_i) \propto f_i(x_i) \prod_{a \in N(i)} m_{a \to i}(x_i)$$

"beliefs" "messages"



$$b_a(X_a) \propto f_a(X_a) \prod_{i \in N(a)} \prod_{c \in N(i) \backslash a} m_{c \to i}(x_i)$$

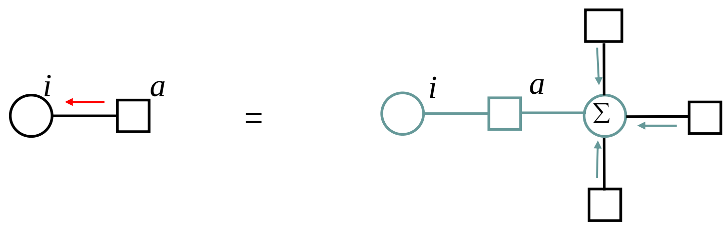
The "belief" is the BP approximation of the marginal probability.

BP Message-update Rules

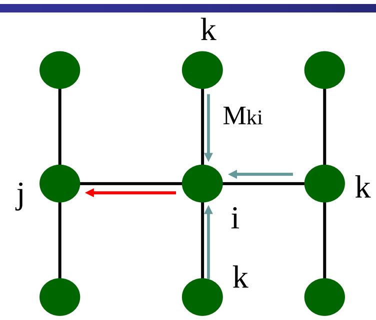
Using $b_{a\to i}(x_i) = \sum_{X_a \setminus x_i} b_a(X_a)$, we get

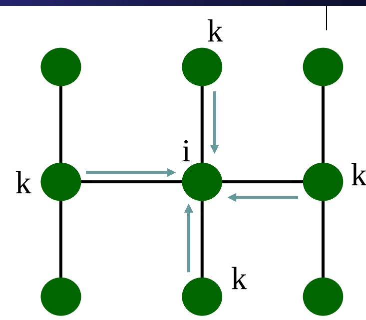
$$\boxed{m_{a\to i}(x_i) = \sum_{X_a \setminus x_i} f_a(X_a) \prod_{j\in N(a)\setminus i} \prod_{b\in N(j)\setminus a} m_{b\to j}(x_j)}$$

(A sum product algorithm)



Belief Propagation on trees





- BP Message-update Rules

$$M_{i\to j}(x_j) \propto \sum_{x_i} \psi_{ij}(x_i, x_j)\psi_i(x_i)\prod_k M_{k\to i}(x_i) \qquad b_i(x_i) \propto \psi_i(x_i)\prod_k M_k(x_k)$$

Compatibilities (interactions) external evidence

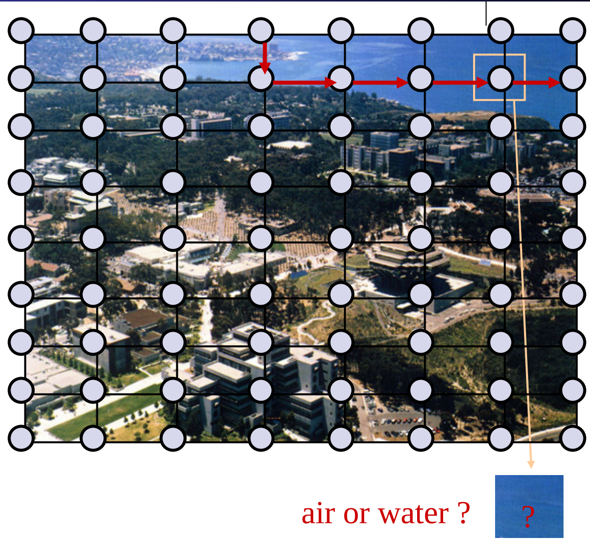
- BP on trees always converges to exact marginals (cf. Junction tree algorithm)

Loopy Belief Propagation

- If BP is used on graphs with loops, messages may circulate indefinitely
- Empirically, a good approximation is still achievable
  - Stop after fixed # of iterations
  - Stop when no significant change in beliefs
  - If solution is not oscillatory but converges, it usually is a good approximation
- Example: Ising models

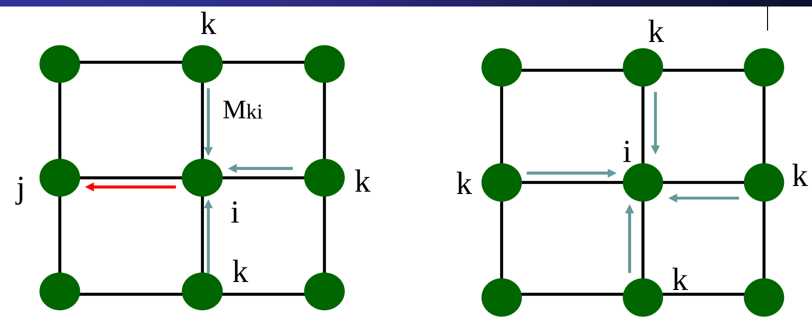
An Ising model on 2-D image

- Nodes encode hidden information (patch-identity).
- They receive local information from the image (brightness, color).
- Information is propagated though the graph over its edges.
- Edges encode ‘compatibility’ between nodes.



air or water ?

Belief Propagation on loopy graphs



- BP Message-update Rules

$$M_{i\to j}(x_j) \propto \underbrace{\sum_{x_i} \psi_{ij}(x_i,x_j)}_{\text{Compatibilities (interactions)}} \underbrace{\psi_i(x_i)}_{\text{external evidence}} \prod_k M_{k\to i}(x_i) \qquad b_i(x_i) \propto \psi_i(x_i)\prod_k M_k(x_k)$$

- May not converge or converge to a wrong solution

Variational (Gibbs) Free Energy

- Kullback-Leibler Distance:

$$D(b \,\|\, p) \equiv \sum_X b(X) \ln \frac{b(X)}{p(X)}$$

- “Boltzmann’s Law” (definition of “energy”):

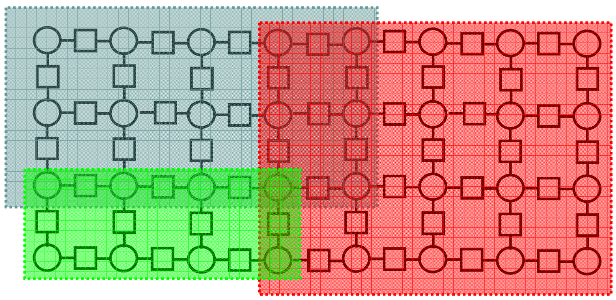
$$p(X) = \frac{1}{Z} \exp[-E(X)]$$

$$D(b \,\|\, p) \equiv \underbrace{\overbrace{\sum_X b(X) E(X)}^{U(b)} + \overbrace{\sum_X b(X) \ln b(X)}^{-H(b)}}_{\text{Gibbs Free Energy } G(b);\ \text{minimized when } b(X)=p(X)} + \ln Z$$

Region-based Approximations to the Gibbs Free Energy (Kikuchi, 1951)

Exact: $G[b(X)]$ *(intractable)*

Regions: $G[\{b_r(X_r)\}]$



Bethe Approximation to Gibbs Free Energy

- Bethe free energy

$$G_{Bethe} = \sum_a \sum_{X_a} b_a(X_a) \ln\left(\frac{b_a(X_a)}{f_a(X_a)}\right) + \sum_i (1 - d_i) \sum_{x_i} b_i(x_i) \ln b_i(x_i)$$

- Equal to the exact Gibbs free energy when the factor graph is a tree because in that case,

$$b(X) = \prod_a b_a(X_a) \prod_i b_i(x_i)^{1-d_i}$$

Minimizing the Bethe Free Energy

$$L = G_{Bethe} + \sum_i \gamma_i \{ \sum_{x_i} b_i(x_i) - 1 \}$$

$$+ \sum_a \sum_{i \in N(a)} \sum_{x_i} \lambda_{ai}(x_i) \left\{ \sum_{X_a \backslash x_i} b_a(X_a) - b_i(x_i) \right\}$$

$$\frac{\partial L}{\partial b_i(x_i)} = 0 \quad \Longrightarrow \quad b_i(x_i) \propto \exp\left(\frac{1}{d_i - 1} \sum_{a \in N(i)} \lambda_{ai}(x_i) \right)$$

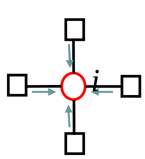
$$\frac{\partial L}{\partial b_a(X_a)} = 0 \quad \Longrightarrow \quad b_a(X_a) \propto \exp\left(-E_a(X_a) + \sum_{i \in N(a)} \lambda_{ai}(x_i) \right)$$

Bethe = BP

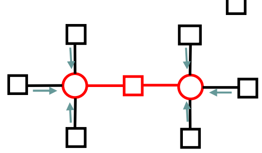
- Identify

$$\lambda_{ai}(x_i) = \ln \prod_{b \in N(i) \neq a} m_{b \to i}(x_i)$$

- to obtain BP equations:



$$b_i(x_i) \propto \prod_{a \in N(i)} m_{a \to i}(x_i)$$

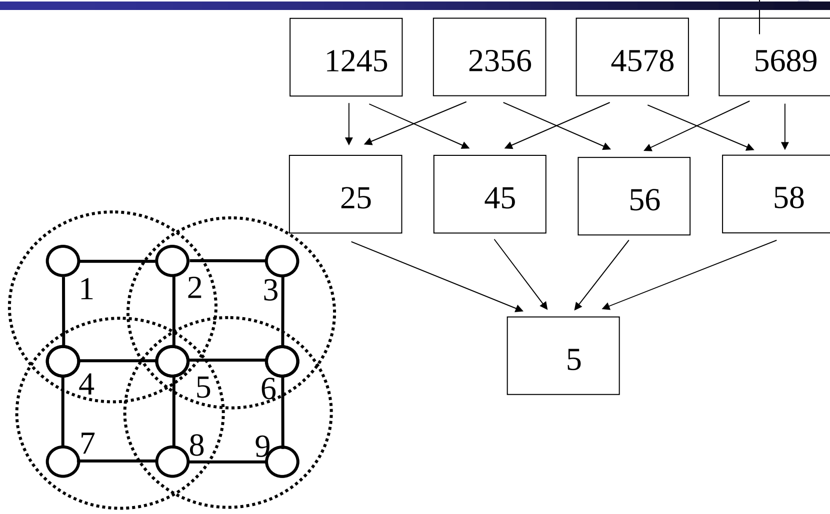


$$b_a(X_a) \propto f_a(X_a) \prod_{i \in N(a)} \prod_{b \in N(i) \backslash a} m_{b \to i}(x_i)$$

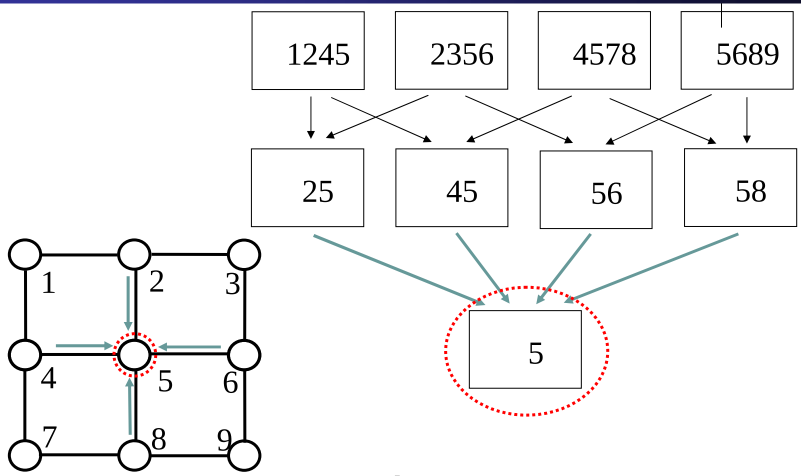
Generalized Belief Propagation

- Belief in a region is the product of:
  - Local information (factors in region)
  - Messages from parent regions
  - Messages into descendant regions from parents who are not descendants.
- Message-update rules obtained by enforcing marginalization constraints.
- Kikuchi free energy

Generalized Belief Propagation

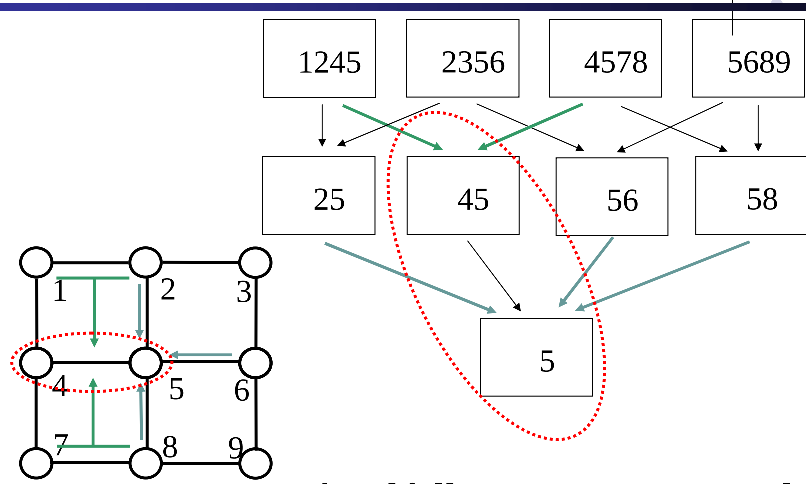


Generalized Belief Propagation



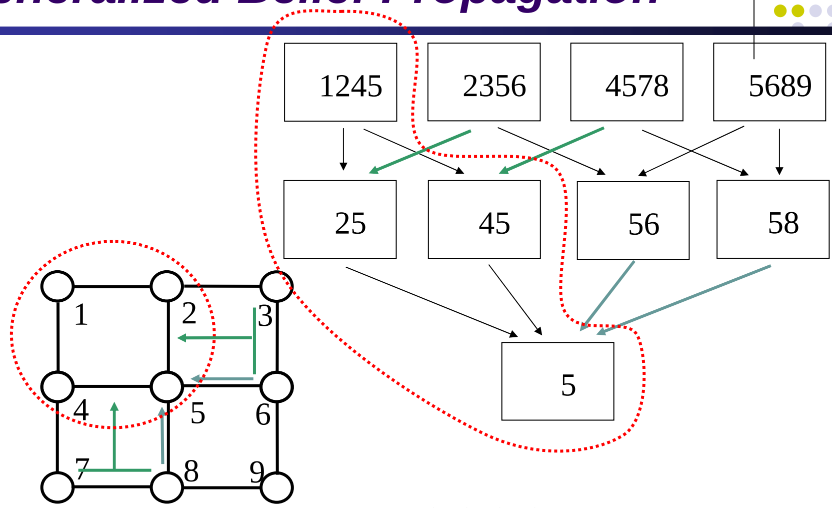
$$b_5 \propto m_{2\to5}m_{4\to5}m_{6\to5}m_{8\to5}$$

Generalized Belief Propagation



$$b_{45} \propto [\,f_{45}\,][m_{12\to45}m_{78\to45}m_{2\to5}m_{6\to5}m_{8\to5}]$$

Generalized Belief Propagation



$$b_{1245} \propto [f_{12} f_{14} f_{25} f_{45}][m_{36\to25} m_{78\to45} m_{6\to5} m_{8\to5}]$$