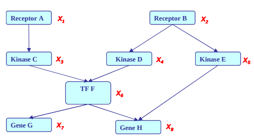
School of Computer Science
Carnegie Mellon

Approximate Inference: Mean Field Methods

Probabilistic Graphical Models (10-708)

Lecture 17, Nov 12, 2007



Eric Xing

Reading: KF-Chap. 12

- Questions????
  - Kalman Filters
  - Complex models
  - LBP-Bethe Minimization

Approximate Inference

Variational Methods

- For a distribution $p(X|\theta)$ associated with a complex graph, computing the marginal (or conditional) probability of arbitrary random variable(s) is intractable
- Variational methods
  - formulating probabilistic inference as an optimization problem:

$$e.g. \quad f^* = \arg\max_{f \in \mathcal{S}} \{ F(f) \}$$

f: a (tractable) probability distribution
or, solutions to certain probabilistic queries

Exponential Family

- Exponential representation of graphical models:

$$P(\mathbf{X}) = \frac{1}{Z}\prod_{c\in C}\psi_c(\mathbf{X}_c) \quad \Rightarrow \quad p(\mathbf{X}\,|\,\theta) = \exp\left\{\sum_\alpha \theta_\alpha \phi_\alpha(\mathbf{X}_{D_\alpha}) - A(\theta)\right\}$$

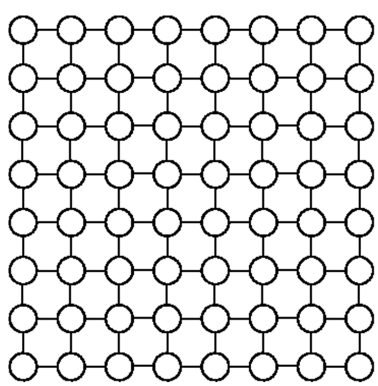
- Includes discrete models, Gaussian, Poisson, exponential, and many others

$$E(\mathbf{X}) = -\sum_\alpha \theta_\alpha \phi_\alpha(\mathbf{X}_{D_\alpha}) \text{ is referred to as the } \textit{energy} \text{ of state } \mathbf{x}$$

$$\Rightarrow \quad p(\mathbf{X}\,|\,\theta) = \exp\{-E(\mathbf{X}) - A(\theta)\}$$

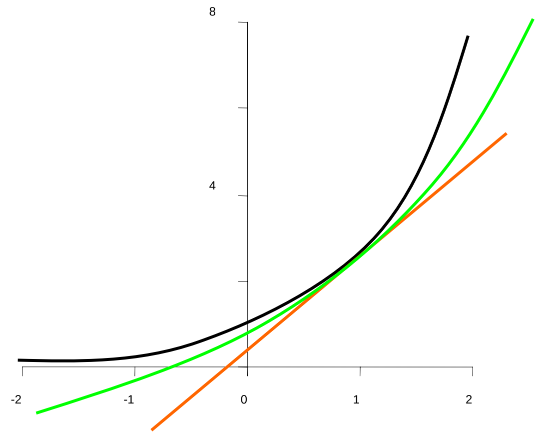
$$= \exp\{-E(\mathbf{X}_H, \mathbf{x}_E) - A(\theta, \mathbf{x}_E)\}$$

Example: the Boltzmann distribution on atomic lattice



$$p(X) = \frac{1}{Z}\exp\left\{\sum_{i<j}\theta_{ij}X_iX_j + \sum_i \theta_{i0}X_i\right\}$$

Lower bounds of exponential functions



$$\exp(x) \geq \exp(\mu)(x - \mu + 1)$$

$$\exp(x) \geq \frac{1}{6}\exp(\mu)\left((x-\mu)^3 + 3(x-\mu)^2 + 6(x-\mu+1)\right)$$

Lower bounding likelihood

Representing $q(\mathbf{X}_H)$ by $\exp\{-E'(\mathbf{X}_H)\}$:

Lemma: Every marginal distribution $q(X_H)$ defines a lower bound of likelihood:

$$p(\mathbf{x}_E) \geq \int d\mathbf{x}_H \exp\{-E'(\mathbf{x}_H)\}$$
$$\left(1 - A(\mathbf{x}_E) - \left(E(\mathbf{x}_H, \mathbf{x}_E) - E'(\mathbf{x}_H)\right)\right),$$

where x_E denotes observed variables (evidence).

Upgradeable to higher order bound [Leisink and Kappen, 2000]

Lower bounding likelihood

Representing $q(\mathbf{X}_H)$ by $\exp\{-E'(\mathbf{X}_H)\}$:

Lemma: Every marginal distribution $q(\mathrm{X}_H)$ defines a lower bound of likelihood:

$$p(\mathbf{x}_E) \geq C - \langle E(\mathbf{X}_H, \mathbf{x}_E) \rangle_{q(\mathbf{x}_H)} - \int d\mathbf{x}_H q(\mathbf{x}_H) \log q(\mathbf{x}_H)$$
$$= C - \langle E \rangle_q + H_q,$$

where x_E denotes observed variables (evidence).

$\langle E \rangle_q$: expected energy

$\langle E \rangle_q - H_q$: Gibbs free energy

H_q : entropy

Eric Xing

KL and variational (Gibbs) free energy

- Kullback-Leibler Distance:

$$KL(q \| p) \equiv \sum_z q(z) \ln \frac{q(z)}{p(z)}$$

- "Boltzmann's Law" (definition of "energy"):

$$p(z) = \frac{1}{C} \exp[-E(z)]$$

$$KL(q \| p) \equiv \underbrace{\sum_z q(z) E(z) + \sum_z q(z) \ln q(z)}_{\text{Gibbs Free Energy } G(q);\ \text{minimized when } q(Z) = p(Z)} + \ln C$$

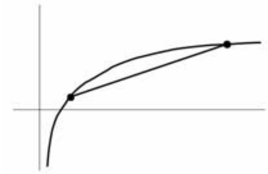
Gibbs Free Energy $G(q)$;
minimized when $q(Z) = p(Z)$

Eric Xing

KL and Log Likelihood

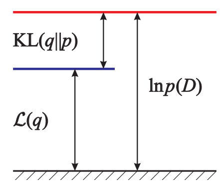
- Jensen's inequality

$$\begin{aligned}\ell(\theta;x) &= \log p(x\mid\theta)\\ &= \log\sum_z p(x,z\mid\theta)\\ &= \log\sum_z q(z\mid x)\frac{p(x,z\mid\theta)}{q(z\mid x)}\\ &\geq \sum_z q(z\mid x)\log\frac{p(x,z\mid\theta)}{q(z\mid x)}\end{aligned}$$



$$\Rightarrow\quad \ell(\theta;x)\geq\langle\ell_c(\theta;x,z)\rangle_q + H_q = \mathcal{L}(q)$$

- KL and Lower bound of likelihood



$$\begin{aligned}\ell(\theta;x) &= \log p(x\mid\theta) = \log\frac{p(x,z\mid\theta)}{p(z\mid x,\theta)} = \sum_z q(z)\log\frac{p(x,z\mid\theta)}{p(z\mid x,\theta)}\\ &= \sum_z q(z)\log\frac{p(x,z\mid\theta)}{q(z)}\frac{q(z)}{p(z\mid x,\theta)}\\ &= \sum_z q(z)\log\frac{p(x,z\mid\theta)}{q(z)} + \sum_z q(z)\log\frac{q(z)}{p(z\mid x,\theta)}\end{aligned}$$

$$\Rightarrow\quad \ell(\theta;x) = \mathcal{L}(q) + KL(q\,\|\,p)$$

- Setting $q()=p(z|x)$ closes the gap (c.f. EM)

A variational representation of probability distributions

$$\begin{aligned} q &= \arg\max_{q\in Q}\ \left\{\ -\langle E\rangle_q + H_q\ \right\}\\ &= \arg\min_{q\in Q}\ \left\{\ \langle E\rangle_q - H_q\ \right\}\end{aligned}$$

where Q is the equivalent sets of realizable distributions, *e.g.*, all valid parameterizations of exponential family distributions, marginal polytopes [winright *et al.* 2003].

Difficulty: H_q is intractable for general q

"solution": approximate H_q
and/or,
relax or tighten Q

Bethe Free Energy/LBP

- But we do not optimize $q(\mathbf{X})$ explicitly, focus on the set of beliefs
  - $e.g.,\quad b = \{b_{i,j} = \tau(x_i, x_j),\quad b_i = \tau(x_i)\}$
- Relax the optimization problem
  - approximate objective: $H_{Betha} = H(b_{i,j}, b_i)$
  - relaxed feasible set: $\mathcal{M}_o = \{\ \tau \geq 0 \mid \sum_{x_i} \tau(x_i) = 1, \sum_{x_i} \tau(x_i, x_j) = \tau(x_j)\ \}$

$$b^* = \arg\min_{b \in \mathcal{M}_o} \left\{ \ \langle E \rangle_b - F(b) \ \right\}$$

- The loopy BP algorithm:
  - a fixed point iteration procedure that tries to solve b^*

Eric Xing 13

Mean field methods

- Optimize $q(\mathbf{X}_H)$ in the space of tractable families
  - $i.e.,$ subgraph of G_p over which exact computation of H_q is feasible
- Tightening the optimization space
  - exact objective: H_q
  - tightened feasible set: $\mathcal{Q} \rightarrow \mathcal{T} \quad (\mathcal{T} \subseteq \mathcal{Q})$

$$q^* = \arg\min_{q \in \mathcal{T}} \langle E \rangle_q - H_q$$

Eric Xing 14

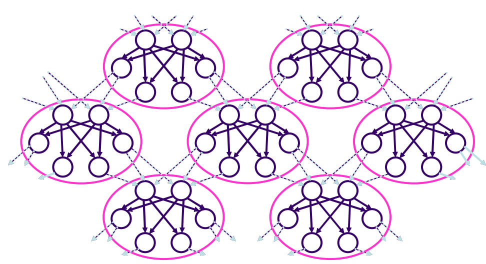
Mean Field Approximation

Cluster-based approx. to the Gibbs free energy

(Wiegerinck 2001, Xing *et al* 03,04)

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

Clusters: $G[\{q_c(X_c)\}]$



Mean field approx. to Gibbs free energy

- Given a disjoint clustering, {C_1, … , C_I}, of all variables
- Let $$q(\mathbf{X}) = \prod_i q_i(\mathbf{X}_{C_i}),$$
- Mean-field free energy

$$G_{\rm MF} = \sum_i \sum_{\mathbf{x}_{C_i}} \prod_i q_i(\mathbf{x}_{C_i}) E(\mathbf{x}_{C_i}) + \sum_i \sum_{\mathbf{x}_{C_i}} q_i(\mathbf{x}_{C_i}) \ln q_i(\mathbf{x}_{C_i})$$

e.g., $$G_{\rm MF} = \sum_{i<j} \sum_{x_i x_j} q(x_i) q(x_j) \phi(x_i x_j) + \sum_i \sum_{x_i} q(x_i) \phi(x_i) + \sum_i \sum_{x_i} q(x_i) \ln q(x_i)$$ (naïve mean field)

  - Will **never** equal to the exact Gibbs free energy no matter what clustering is used, but it does **always** define a lower bound of the likelihood
- Optimize each $q_i(x_c)$'s.
  - Variational calculus …
  - Do inference in each $q_i(x_c)$ using any tractable algorithm

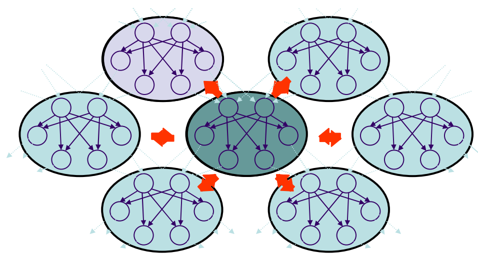
The Generalized Mean Field theorem

Theorem: The optimum GMF approximation to the cluster marginal is isomorphic to the cluster posterior of the original distribution given internal evidence and its generalized mean fields:

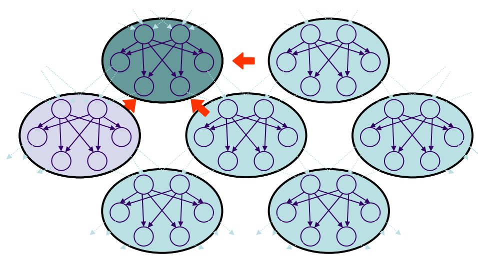
$$q_i^*(\mathbf{X}_{H,C_i}) = p(\mathbf{X}_{H,C_i} \mid \mathbf{x}_{E,C_i}, \langle \mathbf{X}_{H,MB_i} \rangle_{q_{j \neq i}})$$

GMF algorithm: Iterate over each q_i

A generalized mean field algorithm [xing *et al*. UAI 2003]



A generalized mean field algorithm [xing *et al*. UAI 2003]



Convergence theorem

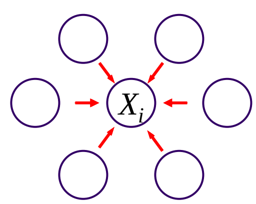
Theorem: The GMF algorithm is guaranteed to converge to a local optimum, and provides a lower bound for the likelihood of evidence (or partition function) the model.

The naive mean field approximation

- Approximate $p(\mathbf{X})$ by fully factorized $q(\mathbf{X})=\mathrm{P}_i q_i(X_i)$
- For Boltzmann distribution $p(X)=\exp\{\sum_{i<j} q_{ij}X_iX_j+q_{io}X_i\}/Z$:

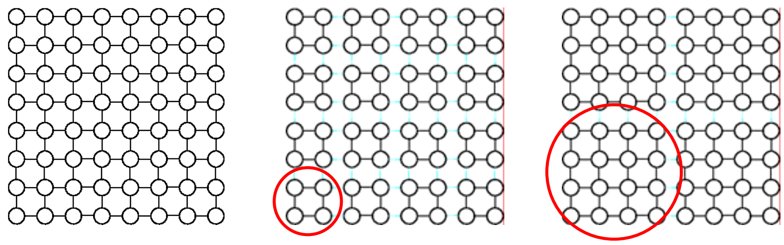
mean field equation:

$$q_i(X_i) = \exp\left\{\theta_{i0}X_i + \sum_{j\in\mathcal{N}_i}\theta_{ij}X_i\left\langle X_j\right\rangle_{q_j} + A_i\right\}$$
$$= p(X_i \mid \{\langle X_j\rangle_{q_j} : j\in\mathcal{N}_i\})$$



- $\left\langle X_j\right\rangle_{q_j}$ resembles a "message" sent from node j to i
- $\{\langle X_j\rangle_{q_j} : j\in\mathcal{N}_i\}$ forms the "mean field" applied to X_i from its neighborhood

Generalized MF approximation to Ising models

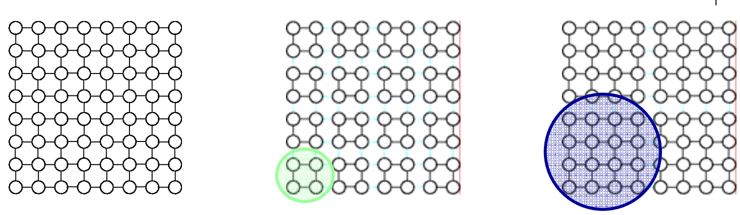


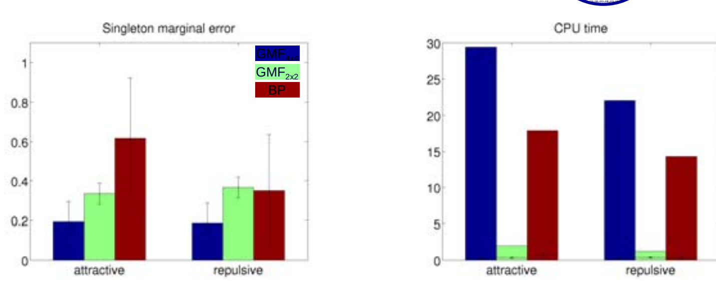
Cluster marginal of a square block C_k:

$$q(X_{C_k}) \propto \exp\left\{ \sum_{i,j\in C_k} \theta_{ij} X_i X_j + \sum_{i\in C_k} \theta_{i0} X_i + \sum_{\substack{i\in C_k,\, j\in MB_k,\\ k'\in MBC_k}} \theta_{ij} X_i \langle X_j \rangle_{q(X_{C_{k'}})} \right\}$$

Virtually a reparameterized Ising model of small size.

GMF approximation to Ising models





Attractive coupling: positively weighted
Repulsive coupling: negatively weighted

Automatic Variational Inference

fHMM | Mean field approx. | Structured variational approx.

- Currently for each new model we have to
  - derive the variational update equations
  - write application-specific code to find the solution
- Each can be time consuming and error prone
- Can we build a general-purpose inference engine which automates these procedures?

Cluster-based MF (e.g., GMF)

- a general, iterative message passing algorithm
- clustering completely defines approximation
  - preserves dependencies
  - flexible performance/cost trade-off
  - clustering automatable
- recovers model-specific structured VI algorithms, including:
  - fHMM, LDA
  - variational Bayesian learning algorithms
- easily provides new structured VI approximations to complex models

Example 1: Bayesian Gaussian Model

- Likelihood function

$$p(D|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

mean

precision (inverse variance)

- Conjugate priors

$$\begin{aligned} p(\mu|\mu_0,\lambda_0) &= \mathcal{N}(\mu|\mu_0,\lambda_0^{-1}) \\ p(\tau|a_0,b_0) &= \mathcal{G}(\tau|a_0,b_0) \end{aligned}$$

- Factorized variational distribution

$$q(\mu,\tau) = q(\mu)q(\tau)$$

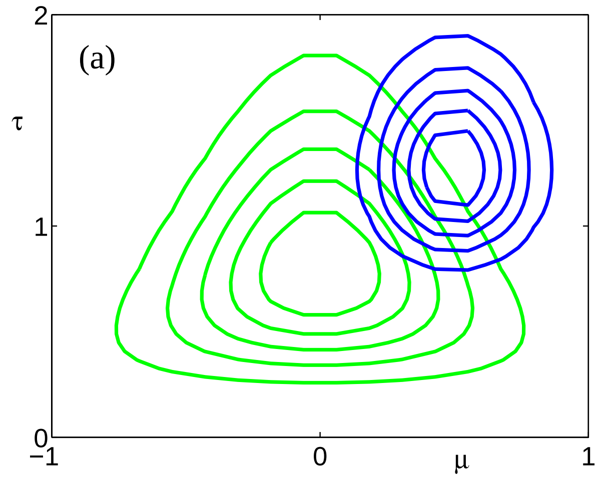
Variational Posterior Distribution

$$\begin{aligned} q(\mu) &= \mathcal{N}(\mu|\mu_N,\lambda_N^{-1}) \\ q(\tau) &= \mathcal{G}(\tau|a_N,b_N) \end{aligned}$$

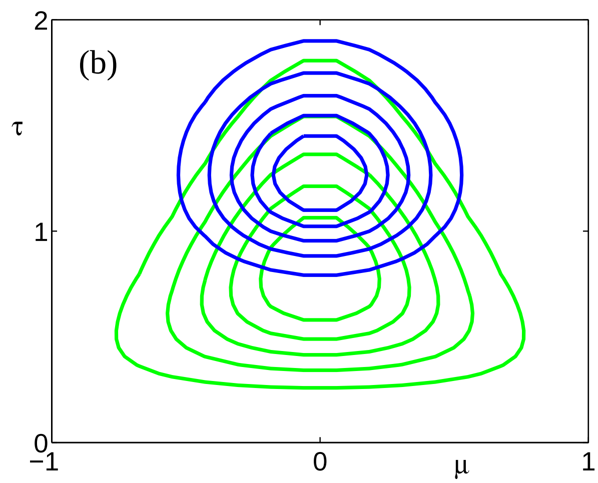
where

$$\begin{aligned} \mu_N &= \frac{\lambda_0\mu_0 + \langle\tau\rangle N\overline{x}}{\lambda_0 + N\langle\tau\rangle} \\ \lambda_N &= \lambda_0 + N\langle\tau\rangle \\ a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2}\left\langle\sum_n (x_n-\mu)^2\right\rangle_\mu \end{aligned}$$

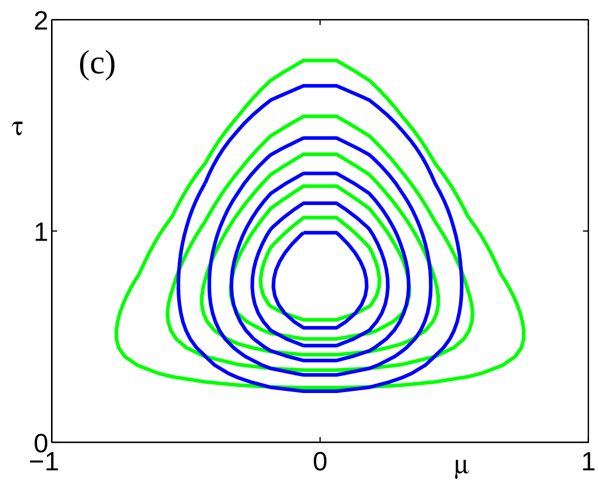
Initial Configuration



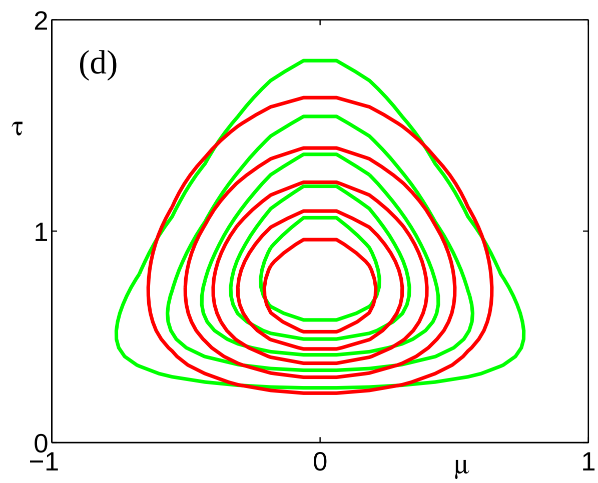
After Updating $q(\mu)$



After Updating $q(\mu)$

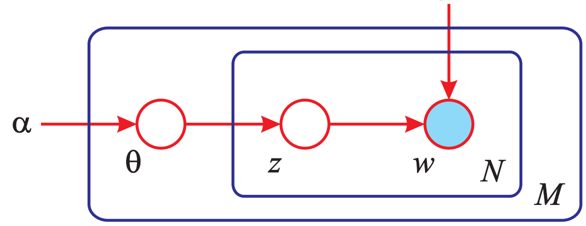


Converged Solution



Example 2: Latent Dirichlet Allocation

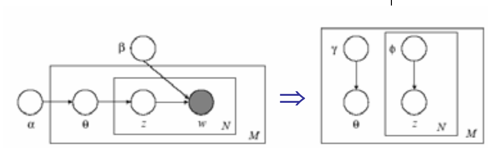
- Blei, Jordan and Ng (2003)
- Generative model of documents (but broadly applicable e.g. collaborative filtering, image retrieval, bioinformatics)
- Generative model:
  - choose $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
  - choose topic $\mathbf{z}_n \sim \mathrm{Mult}(\boldsymbol{\theta})$
  - choose word $\mathbf{w}_n \sim p(\mathbf{w}_n | \mathbf{z}_n, \boldsymbol{\beta})$



Latent Dirichlet Allocation

- Variational approximation

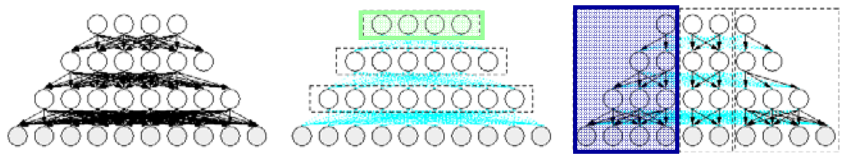
$$q(\theta, z) = q_\theta(\theta) q_z(z) = \mathrm{Dir}\big(\theta \mid \gamma = f(\alpha, \langle z \rangle)\big) \times \mathrm{Multi}\big(z \mid \phi = f(\beta_w, \langle \ln \theta \rangle)\big)$$

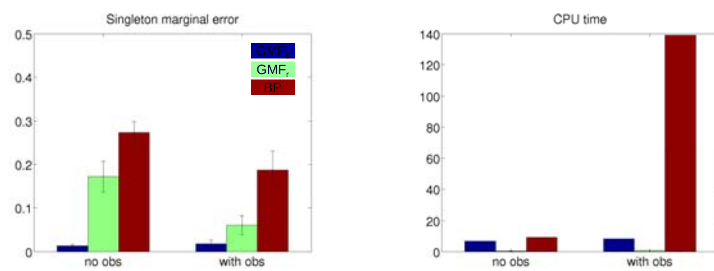


$$\phi_{ni} \propto \beta_{iw_n} \exp\{E_q[\log(\theta_i) \mid \gamma]\}$$
$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}.$$

- Data set:
  - 15,000 documents
  - 90,000 terms
  - 2.1 million words
- Model:
  - 100 factors
  - 9 million parameters
- MCMC could be totally infeasible for this problem

Example 3: Sigmoid belief network





Example 4: Factorial HMM