Probabilistic Graphical Models

Brown University CSCI 2950-P, Spring 2013
Prof. Erik Sudderth

Lecture 19:
Mean Field Variational Bayesian Learning,
Blocked Gibbs Samplers

Mean Field Free Energy

$$p(x) = \frac{1}{Z} \exp\left\{ - \sum_{(s,t)\in\mathcal{E}} \phi_{st}(x_s, x_t) - \sum_{s\in\mathcal{V}} \phi_s(x_s) \right\}$$

$$q(x) = \prod_{s\in\mathcal{V}} q_s(x_s)$$

$$\phi_{st}(x_s, x_t) = -\log \psi_{st}(x_s, x_t)$$
$$\phi_s(x_s) = -\log \psi_s(x_s)$$

$$D(q \,||\, p) = -H(q) + \sum_x q(x)E(x) + \log Z$$

Mean Field Entropy:

$$H(q) = \sum_{s\in\mathcal{V}} H_s(q_s) = -\sum_{s\in\mathcal{V}} \sum_{x_s} q_s(x_s) \log q_s(x_s)$$

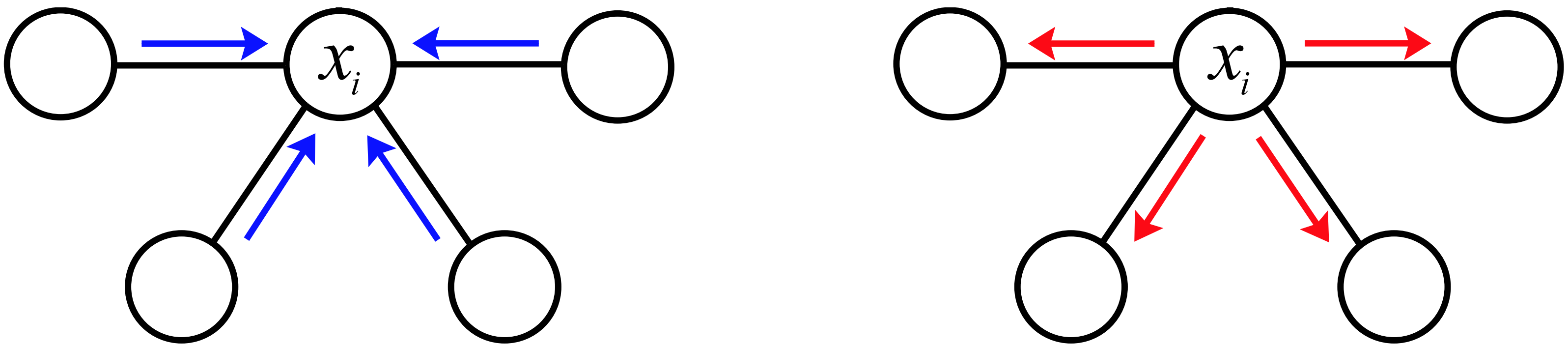
Mean Field Average Energy (expected sufficient statistics):

$$\sum_x q(x)E(x) = \sum_{(s,t)\in\mathcal{E}} \sum_{x_s, x_t} q_s(x_s) q_t(x_t) \phi_{st}(x_s, x_t) + \sum_{s\in\mathcal{V}} \sum_{x_s} q_s(x_s) \phi_s(x_s)$$

Mean Field as Message Passing

- Consider a pairwise undirected graphical model:

$$p(x) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s\in\mathcal{V}} \psi_s(x_s)$$



$$q_i(x_i) \propto \psi_i(x_i) \prod_{j\in\Gamma(i)} m_{ji}(x_i)$$

$$m_{ji}(x_i) \propto \exp\left\{ -\sum_{x_j} \phi_{ij}(x_i, x_j) q_j(x_j) \right\}$$

- For continuous variables, valid with sum replaced by integral
- If marginals place all of their mass on a single state, becomes equivalent to Gibbs sampling update equations

(Mean Field) Variational Bayesian Learning

$$\ln p(x) = \ln \left(\int_\Theta \sum_z p(x, z \mid \theta) p(\theta) \; d\theta \right)$$

$$\ln p(x) \geq \int_\Theta \sum_z q_z(z) q_\theta(\theta) \ln \left(\frac{p(x, z \mid \theta) p(\theta)}{q_z(z) q_\theta(\theta)} \right) \; d\theta$$

$$\ln p(x) \geq \int_\Theta \sum_z q_z(z) q_\theta(\theta) \ln p(x, z, \theta) \; d\theta + H(q_z) + H(q_\theta) \triangleq \mathcal{L}(q_z, q_\theta)$$

- **Initialization:** Randomly select starting distribution $q_\theta^{(0)}$
- **E-Step:** Given parameters, find posterior of hidden data

$$q_z^{(t)} = \arg\max_{q_z} \mathcal{L}(q_z, q_\theta^{(t-1)})$$

- **M-Step:** Given posterior distributions, find likely parameters

$$q_\theta^{(t)} = \arg\max_{q_\theta} \mathcal{L}(q_z^{(t)}, q_\theta)$$

- **Iteration:** Alternate E-step & M-step until convergence

(Mean Field) Variational Bayesian Learning

Temporary notation change: observations y, hidden variables x

$$\ln p(\mathbf{y}\,|\,m) \;\geq\; \int q_{\mathbf{x}}(\mathbf{x})q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\mathbf{y},\mathbf{x},\boldsymbol{\theta}\,|\,m)}{q_{\mathbf{x}}(\mathbf{x})q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}\, d\mathbf{x}\, d\boldsymbol{\theta} \;=\; \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y})$$

Condition (1). *The complete data likelihood is that of an exponential family:* $p(\mathbf{x},\mathbf{y}\,|\,\boldsymbol{\theta}) = f(\mathbf{x},\mathbf{y})\; g(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^T \mathbf{u}(\mathbf{x},\mathbf{y})\right\}$, *where* $\boldsymbol{\phi}(\boldsymbol{\theta})$ *is the vector of natural parameters, and* $\mathbf{u}$ *and* f *and* g *are the functions that define the exponential family.*

Condition (2). *The parameter prior is conjugate to the complete data likelihood:* $p(\boldsymbol{\theta}\,|\,\eta,\boldsymbol{\nu}) = h(\eta,\boldsymbol{\nu})\; g(\boldsymbol{\theta})^{\eta} \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^T \boldsymbol{\nu}\right\}$, *where* η *and* $\boldsymbol{\nu}$ *are hyperparameters.*

EM for MAP estimation	**Variational Bayesian EM**
Goal: maximise $p(\boldsymbol{\theta}\,\|\,\mathbf{y},m)$ w.r.t. $\boldsymbol{\theta}$	**Goal:** lower bound $p(\mathbf{y}\,\|\,m)$
E Step: compute	**VB-E Step:** compute $\bar{\phi}^{(t)} = \mathbb{E}_{q_\theta^{(t)}}[\phi(\theta)]$
$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x}\,\|\,\mathbf{y},\boldsymbol{\theta}^{(t)})$	$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x}\,\|\,\mathbf{y},\overline{\boldsymbol{\phi}}^{(t)})$
M Step:	**VB-M Step:**
$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \int q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x},\mathbf{y},\boldsymbol{\theta})\, d\mathbf{x}$	$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) \propto \exp\left[\int q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x},\mathbf{y},\boldsymbol{\theta})\, d\mathbf{x}\right]$

M-Step: Expected log-likelihood exponentiated to distribution

E-Step: Based on mean of natural parameters, not mode

Exponential Family Variational Learning

Temporary notation change: observations y, hidden variables x

Condition (1). *The complete data likelihood is that of an exponential family:* $p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\theta}) = f(\mathbf{x}, \mathbf{y})\ g(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^T \mathbf{u}(\mathbf{x}, \mathbf{y})\right\}$, *where* $\boldsymbol{\phi}(\boldsymbol{\theta})$ *is the vector of natural parameters, and* $\mathbf{u}$ *and* f *and* g *are the functions that define the exponential family.*

Condition (2). *The parameter prior is conjugate to the complete data likelihood:* $p(\boldsymbol{\theta} \mid \eta, \boldsymbol{\nu}) = h(\eta, \boldsymbol{\nu})\ g(\boldsymbol{\theta})^{\eta} \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^T \boldsymbol{\nu}\right\}$, *where* η *and* $\boldsymbol{\nu}$ *are hyperparameters.*

Theorem. (Conjugate-Exponential Models). *Given an iid data set* $\mathbf{y} = \{\mathbf{y}_1, \ldots \mathbf{y}_n\}$, *if the model satisfies conditions (1) and (2), then at every iteration of the variational Bayesian EM algorithm and at the maxima of* $\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y})$:

(a) $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ *is conjugate with parameters* $\tilde{\eta} = \eta + n$, $\tilde{\boldsymbol{\nu}} = \boldsymbol{\nu} + \sum_{i=1}^{n} \overline{\mathbf{u}}(\mathbf{y}_i)$:

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = h(\tilde{\eta}, \tilde{\boldsymbol{\nu}}) g(\boldsymbol{\theta})^{\tilde{\eta}} \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^T \tilde{\boldsymbol{\nu}}\right\} \tag{9}$$

where $\overline{\mathbf{u}}(\mathbf{y}_i) = \mathrm{E}_{q_{\mathbf{x}_i}}(\mathbf{u}(\mathbf{x}_i, \mathbf{y}_i))$, *using* $\mathrm{E}_{q_{\mathbf{x}_i}}$ *to denote expectation under the variational posterior over the latent variable(s) associated with the ith datum.*

(b) $q_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^{n} q_{\mathbf{x}_i}(\mathbf{x}_i)$ *with*

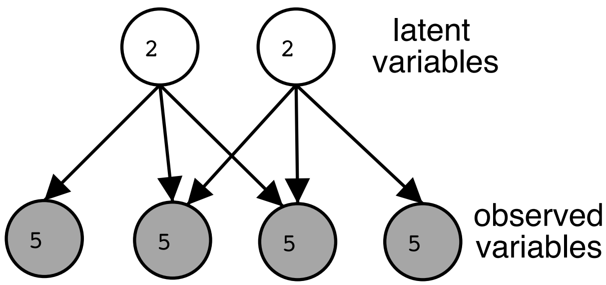
$$q_{\mathbf{x}_i}(\mathbf{x}_i) = p(\mathbf{x}_i \mid \mathbf{y}_i, \overline{\boldsymbol{\phi}}) \propto f(\mathbf{x}_i, \mathbf{y}_i) \exp\left\{\overline{\boldsymbol{\phi}}^T \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)\right\} \tag{10}$$

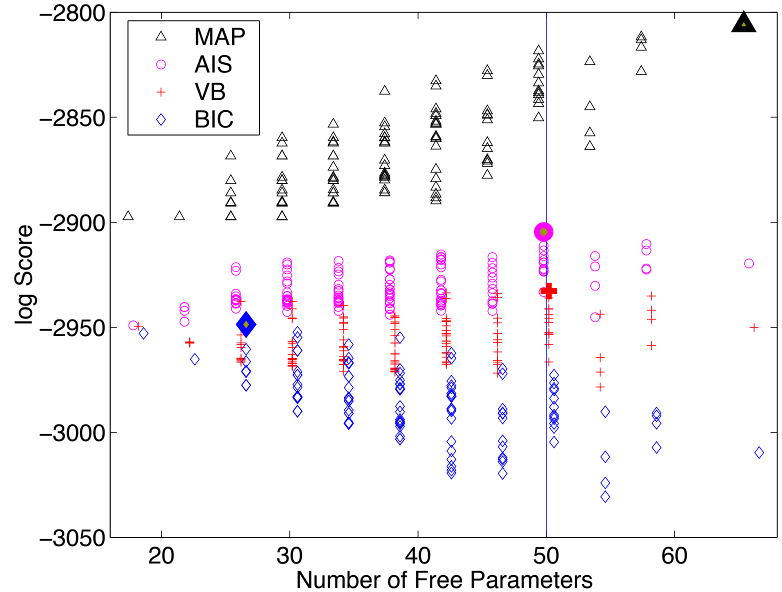
where $\overline{\boldsymbol{\phi}} = \mathrm{E}_{q_{\boldsymbol{\theta}}}(\boldsymbol{\phi}(\boldsymbol{\theta}))$, *the expectation of the natural parameter.*

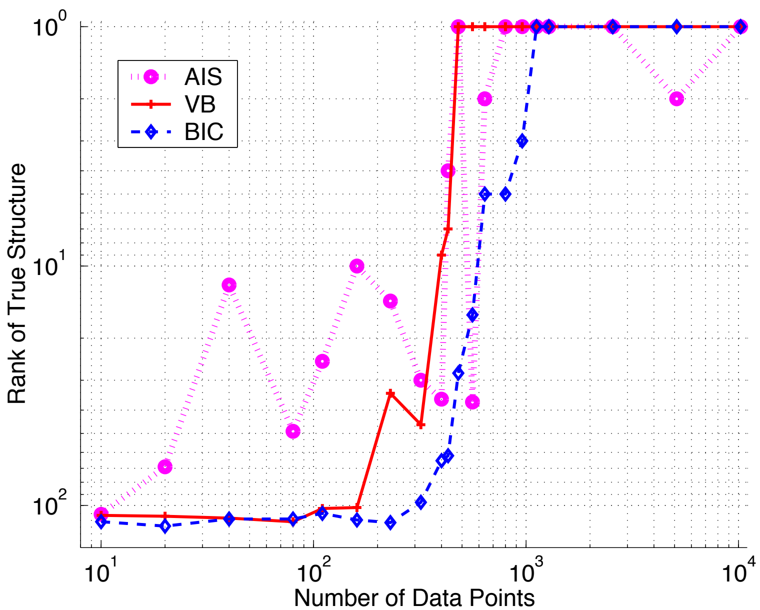
Beal & Ghahramani 2003

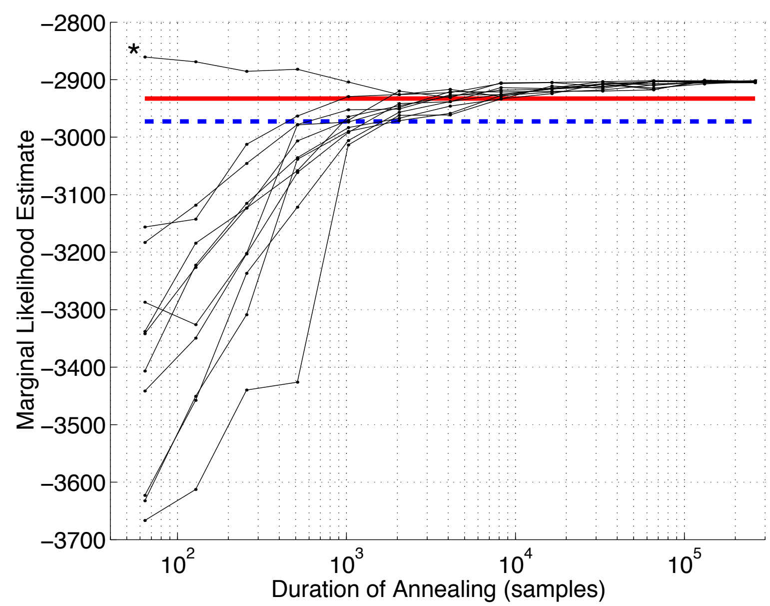
Example: Graph Structure Learning

Consider all possible bipartite graph structures relating 6 discrete variables



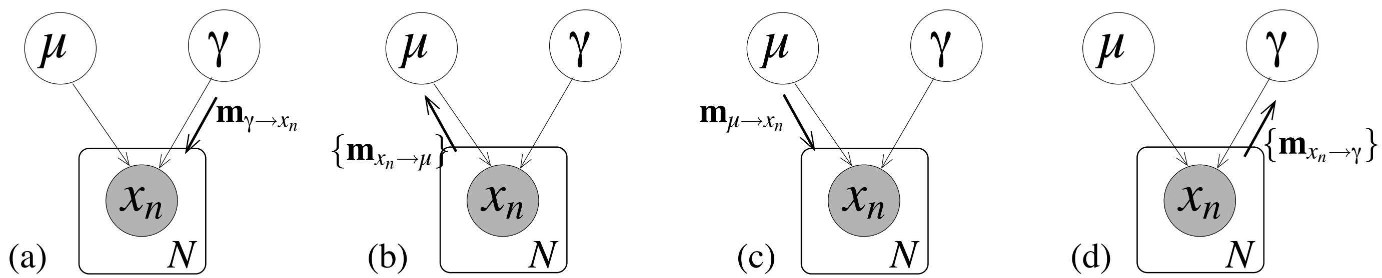






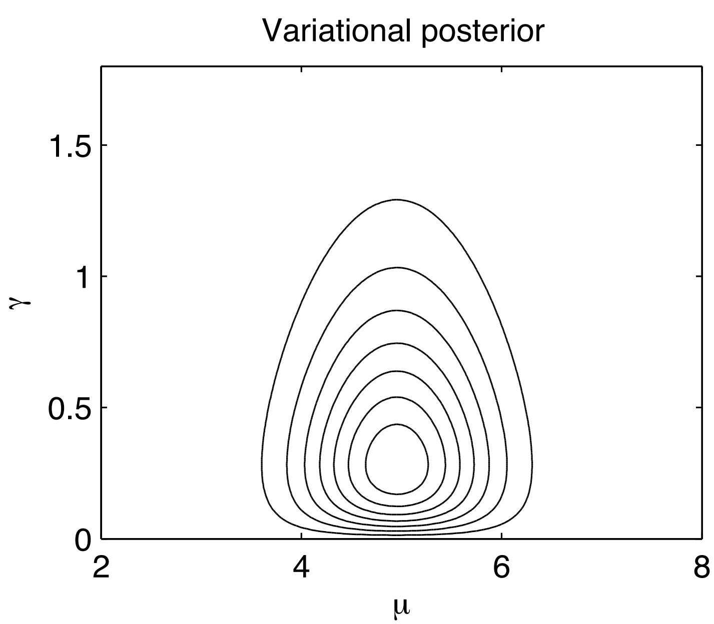
Beal & Ghahramani 2003

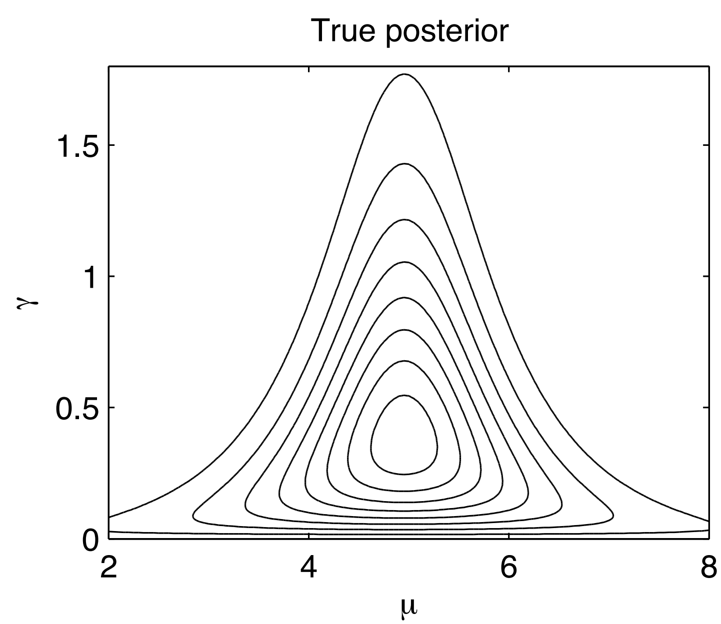
Example: Bayesian Gaussian Learning



$$\ln P(x_n \mid \mu, \gamma^{-1}) = \begin{bmatrix} \gamma\mu \\ -\gamma/2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} x_n \\ x_n^2 \end{bmatrix} + \frac{1}{2}(\ln\gamma - \gamma\mu^2 - \ln 2\pi)$$

$$P(\mu) = \mathcal{N}(0, 1000) \text{ and } P(\gamma) = \mathrm{Gamma}(0.001, 0.001)$$

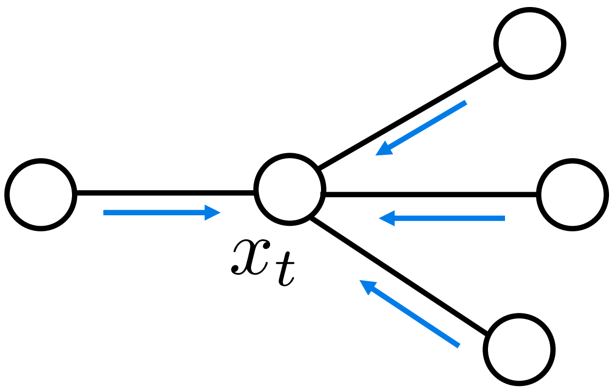




Winn & Bishop 2005

Belief Propagation (Sum-Product)

BELIEFS: Posterior marginals

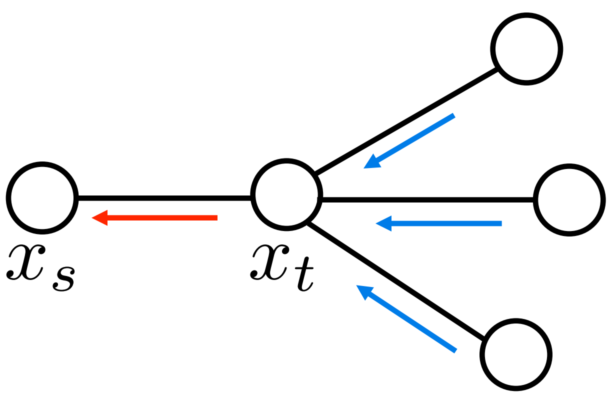


$$q_t(x_t) \propto \psi_t(x_t) \prod_{u \in \Gamma(t)} m_{ut}(x_t)$$

$\Gamma(t) \longrightarrow$ neighborhood of node t (adjacent nodes)

MESSAGES: Sufficient statistics

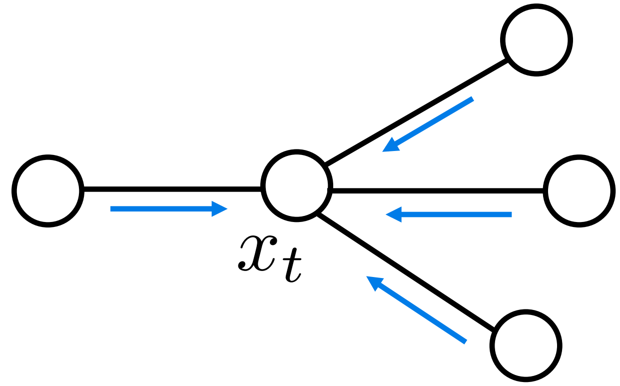
$$m_{ts}(x_s) \propto \sum_{x_t} \psi_{st}(x_s, x_t) \psi_t(x_t) \prod_{u \in \Gamma(t) \setminus s} m_{ut}(x_t)$$



I) Message Product
II) Message Propagation

Mean Field versus Belief Propagation

$$p(x) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s\in\mathcal{V}} \psi_s(x_s)$$



$$q_t(x_t) \propto \psi_t(x_t) \prod_{u\in\Gamma(t)} m_{ut}(x_t)$$

Belief Propagation (Sum-Product) Messages:

$$m_{ts}(x_s) \propto \sum_{x_t} \psi_{st}(x_s, x_t)\psi_t(x_t) \prod_{u\in\Gamma(t)\setminus s} m_{ut}(x_t)$$

$$m_{ts}(x_s) \propto \sum_{x_t} \psi_{st}(x_s, x_t) \frac{q_t(x_t)}{m_{st}(x_t)}$$

Replaces geometric (log-domain) mean by arithmetic mean, and divides by incoming message to avoid "double-counting" information

(Naïve) Mean Field Messages:

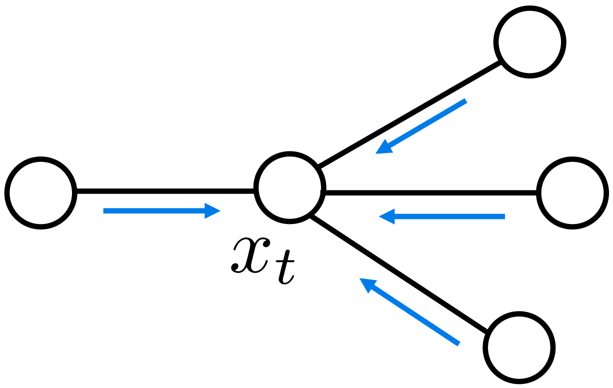
$$m_{ts}(x_s) \propto \exp\left\{ -\sum_{x_t} \phi_{st}(x_s, x_t) q_t(x_t) \right\} \qquad \phi_{st}(x_s, x_t) = -\psi_{st}(x_s, x_t)$$

Mean Field versus Belief Propagation

$$p(x) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s\in\mathcal{V}} \psi_s(x_s)$$

$$\phi_{st}(x_s, x_t) = -\psi_{st}(x_s, x_t)$$

$$q_t(x_t) \propto \psi_t(x_t) \prod_{u\in\Gamma(t)} m_{ut}(x_t)$$



BP: $$m_{ts}(x_s) \propto \sum_{x_t} \psi_{st}(x_s, x_t) \frac{q_t(x_t)}{m_{st}(x_t)}$$

MF: $$m_{ts}(x_s) \propto \exp\left\{ -\sum_{x_t} \phi_{st}(x_s, x_t) q_t(x_t) \right\}$$

Big implications from small changes:

- Belief Propagation: Produces exact marginals for any tree, but for general graphs no guarantees of convergence or accuracy
- Mean Field: Guaranteed to converge for general graphs, always lower-bounds partition function, but approximate even on trees

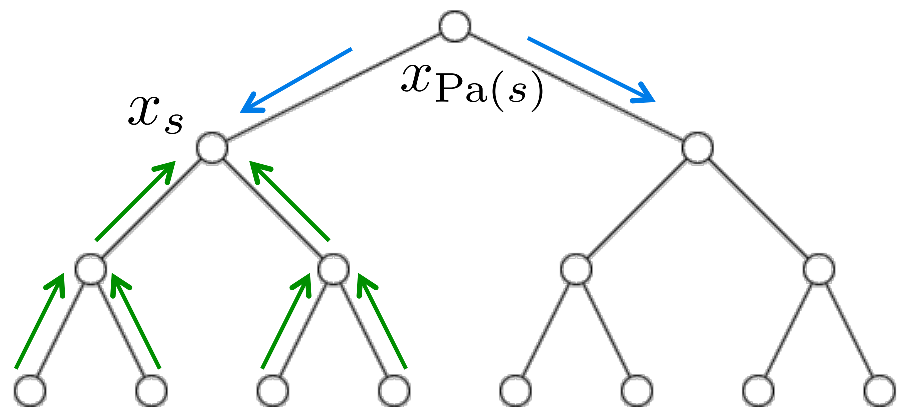
Sum-Product for Blocked Sampling

$$p(x) = \frac{1}{Z} \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s \in \mathcal{V}} \psi_s(x_s)$$

Global Directed Factorization:

- Choose some node as the root of the tree, order by depth
- Define directed factorization from root to leaves:

$$p(x) = p(x_{\text{Root}}) \prod_s p(x_s \mid x_{\text{Pa}(s)})$$



Bottom-Up Message Passing:

- Pass sum-product messages recursively from leaves to root

$$m_{ts}(x_s) \propto \sum_{x_t} \psi_{st}(x_s, x_t) \psi_t(x_t) \prod_{u \in \Gamma(t) \setminus s} m_{ut}(x_t)$$

- Compute marginal of root node:

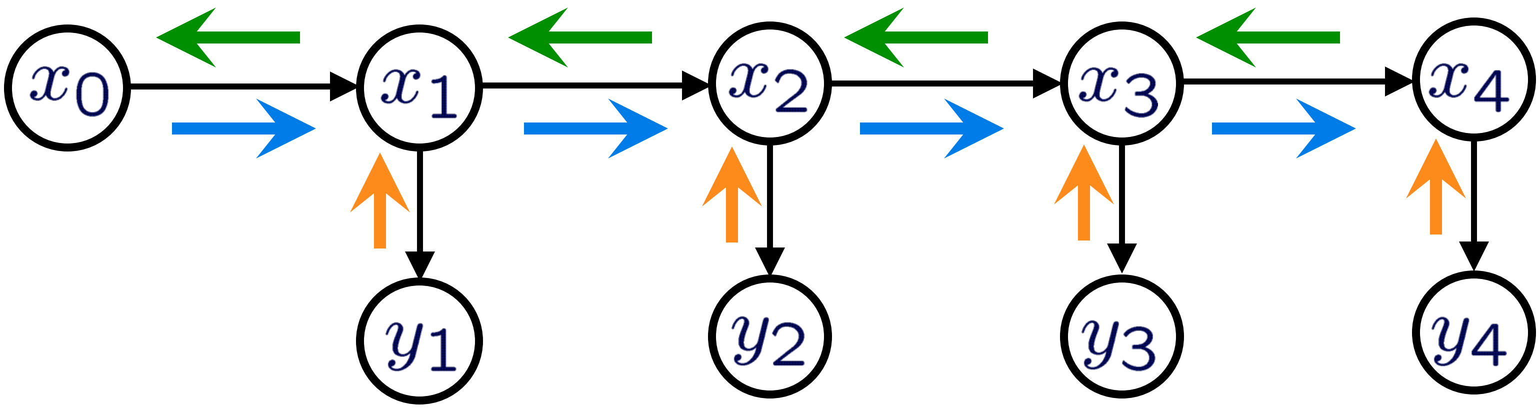
$$q_t(x_t) \propto \psi_t(x_t) \prod_{u \in \Gamma(t)} m_{ut}(x_t)$$

Top-Down Recursive Sampling:

- Sample root from marginal, then sample by depth given parent:

$$p(x_s \mid X_t = \hat{x}_t, t = \text{Pa}(s)) \propto \psi_{ts}(\hat{x}_t, x_s) \psi_s(x_s) \prod_{u \in \Gamma(s) \setminus t} m_{us}(x_s)$$

Monte Carlo Estimation



- Suppose interested in some complex, global function of state:

$$\mathbb{E}[f] = \int f(x) p(x \mid y)\, dx \approx \frac{1}{L} \sum_{\ell=1}^{L} f(x^{(\ell)}) \qquad x^{(\ell)} \sim p(x \mid y)$$

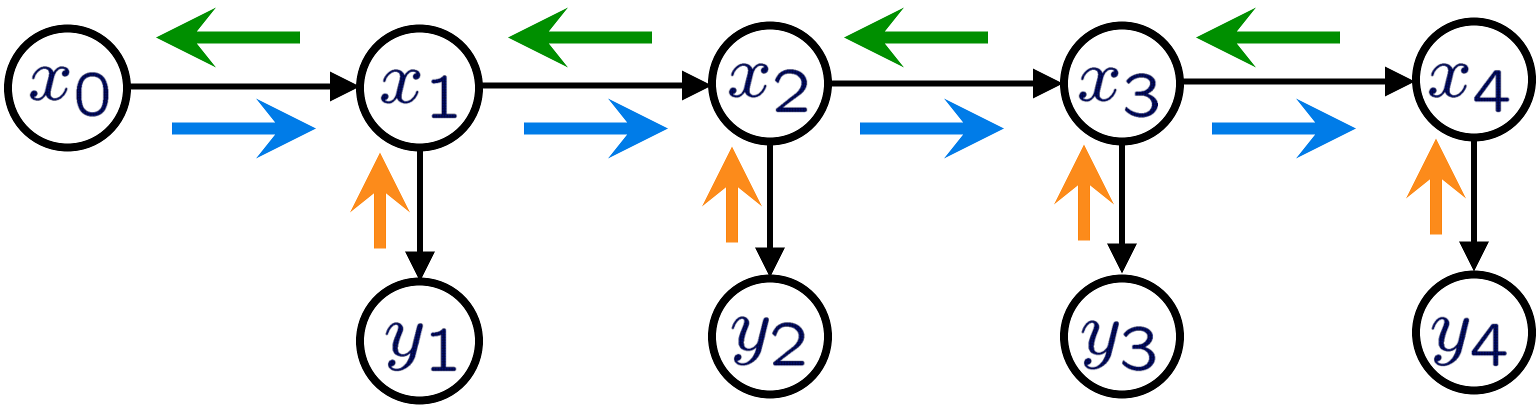
- Can efficiently draw joint samples from posterior marginals:
  - Forward Message Passing:
  - Backwards Sampling:

$$p(x_t \mid y), p(x_t, x_{t+1} \mid y)$$

$$x_T^{(\ell)} \sim p(x_T \mid y)$$
$$x_{T-1}^{(\ell)} \sim p(x_{T-1} \mid x_T^{(\ell)}, y)$$
$$x_{T-2}^{(\ell)} \sim p(x_{T-2} \mid x_{T-1}^{(\ell)}, y)$$

$$(x_1^{(\ell)}, x_2^{(\ell)}, \ldots, x_T^{(\ell)}) \sim p(x \mid y)$$

Monte Carlo Estimation



- Procedure only tractable for a limited class of models:
  - Discrete states: Sum-product belief propagation algorithm
  - Gaussian continuous states: Kalman smoothing algorithm

- Can efficiently draw joint samples from posterior marginals:
  - Forward Message Passing:
  - Backwards Sampling:

$$p(x_t \mid y), p(x_t, x_{t+1} \mid y)$$

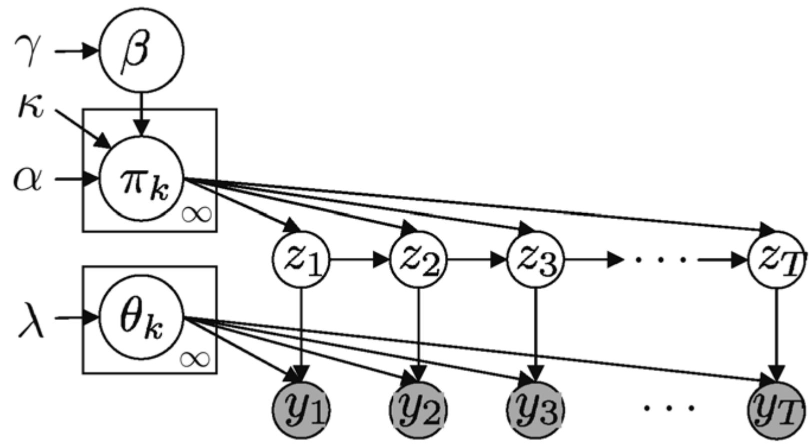
$$x_T^{(\ell)} \sim p(x_T \mid y)$$
$$x_{T-1}^{(\ell)} \sim p(x_{T-1} \mid x_T^{(\ell)}, y)$$
$$x_{T-2}^{(\ell)} \sim p(x_{T-2} \mid x_{T-1}^{(\ell)}, y)$$

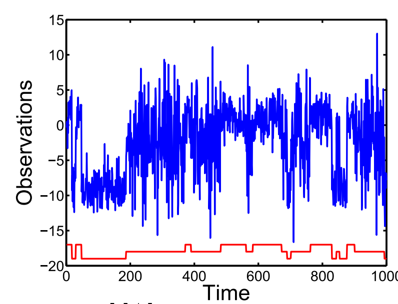
$$(x_1^{(\ell)}, x_2^{(\ell)}, \ldots, x_T^{(\ell)}) \sim p(x \mid y)$$

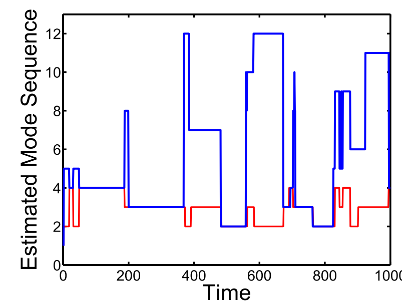
Example: Bayesian HMMs

Given a previous set of state-specific transition probabilities $\boldsymbol{\pi}^{(n-1)}$, the global transition distribution $\beta^{(n-1)}$, and emission parameters $\boldsymbol{\theta}^{(n-1)}$:

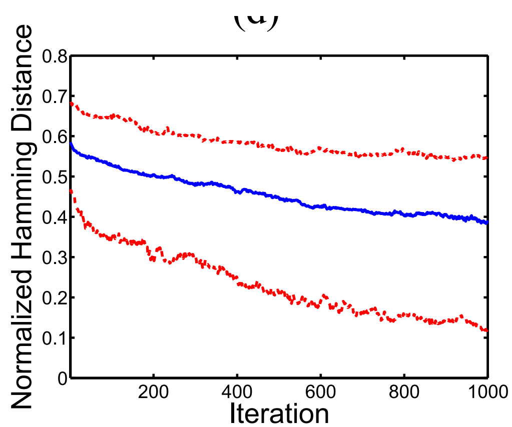
1. Set $\boldsymbol{\pi} = \boldsymbol{\pi}^{(n-1)}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^{(n-1)}$. Working sequentially backwards in time, calculate messages $m_{t,t-1}(k)$:
   (a) For each $k \in \{1, \ldots, L\}$, initialize messages to
   $$m_{T+1,T}(k) = 1$$
   (b) For each $t \in \{T-1, \ldots, 1\}$ and for each $k \in \{1, \ldots, L\}$, compute
   $$m_{t,t-1}(k) = \sum_{j=1}^{L} \pi_k(j) \mathcal{N}(y_t; \mu_j, \Sigma_j) m_{t+1,t}(j)$$
2. Sample state assignments $z_{1:T}$ working sequentially forward in time, starting with $n_{jk} = 0$ and $\mathcal{Y}_k = \emptyset$ for each $(j,k) \in \{1, \ldots, L\}^2$:
   (a) For each $k \in \{1, \ldots, L\}$, compute the probability
   $$f_k(y_t) = \pi_{z_{t-1}}(k) \mathcal{N}(y_t; \mu_k, \Sigma_k) m_{t+1,t}(k)$$
   (b) Sample a state assignment z_t:
   $$z_t \sim \sum_{k=1}^{L} f_k(y_t) \delta(z_t, k)$$
   (c) Increment $n_{z_{t-1}z_t}$ and add y_t to the cached statistics for the new assignment $z_t = k$:
   $$\mathcal{Y}_k \leftarrow \mathcal{Y}_k \oplus y_t$$

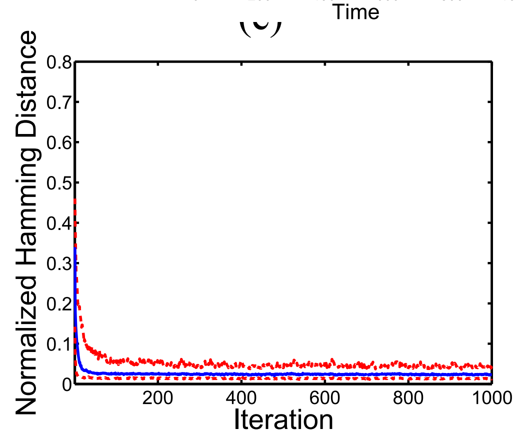






Standard Gibbs





Blocked Gibbs