Probabilistic Graphical Models

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

Lecture 18:
Collapsed Gibbs Sampling, Mean Field Variational Methods, Variational Bayesian Learning

Gibbs Sampling

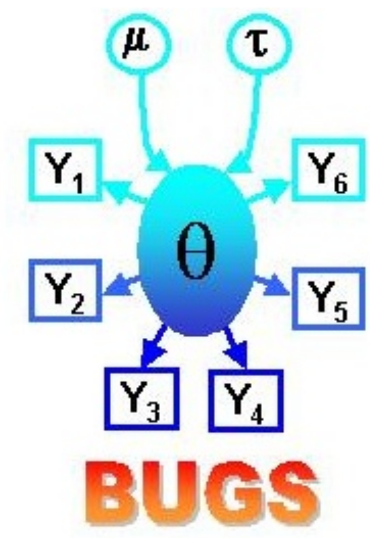
Gibbs sampling benefits from few free choices and convenient features of conditional distributions:

- Conditionals with a few discrete settings can be explicitly normalized:

$$
\begin{aligned}
P(x_i|\mathbf{x}_{j\neq i}) &\propto P(x_i, \mathbf{x}_{j\neq i}) \\
&= \frac{P(x_i, \mathbf{x}_{j\neq i})}{\sum_{x_i'} P(x_i', \mathbf{x}_{j\neq i})} \quad \leftarrow \text{this sum is small and easy}
\end{aligned}
$$

- Continuous conditionals only univariate
  $\Rightarrow$ amenable to standard sampling methods.

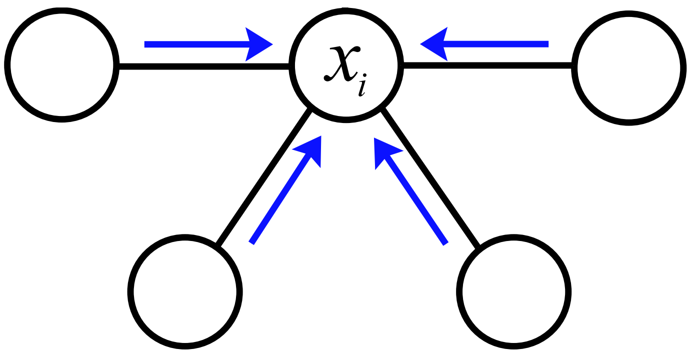
  - Inverse CDF sampling
  - Rejection sampling
  - Slice sampling
  - …

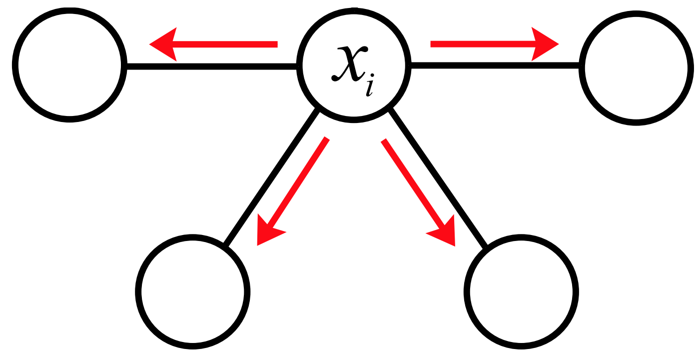


Gibbs Sampling as Message Passing

- Consider a pairwise undirected graphical model:

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





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

$$m_{ij}(x_j) \propto \psi_{ij}(\hat{x}_i, x_j)$$

$$\hat{x}_i \sim q_i(x_i)$$

Draw single sample from marginal

Use sample to extract a “slice” of pairwise potential

- Valid for discrete and continuous variables, although sampling step may be harder for continuous models
- General factor graphs have similar form

Rao-Blackwellized Estimation

- Basic Monte Carlo estimation for joint distribution of *x, z*:

$$(x^{(\ell)}, z^{(\ell)}) \sim p(x,z) \qquad \ell = 1, 2, \ldots, L$$

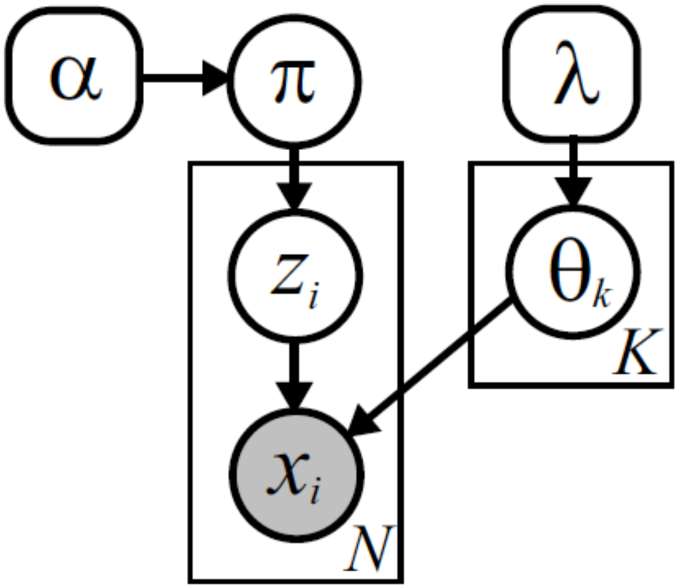
$$\mathbb{E}_p[f(x,z)] = \int_{\mathcal{Z}} \int_{\mathcal{X}} f(x,z) p(x,z)\, dx\, dz \approx \frac{1}{L} \sum_{\ell=1}^{L} f(x^{(\ell)}, z^{(\ell)}) = \mathbb{E}_{\tilde{p}}[f(x,z)]$$

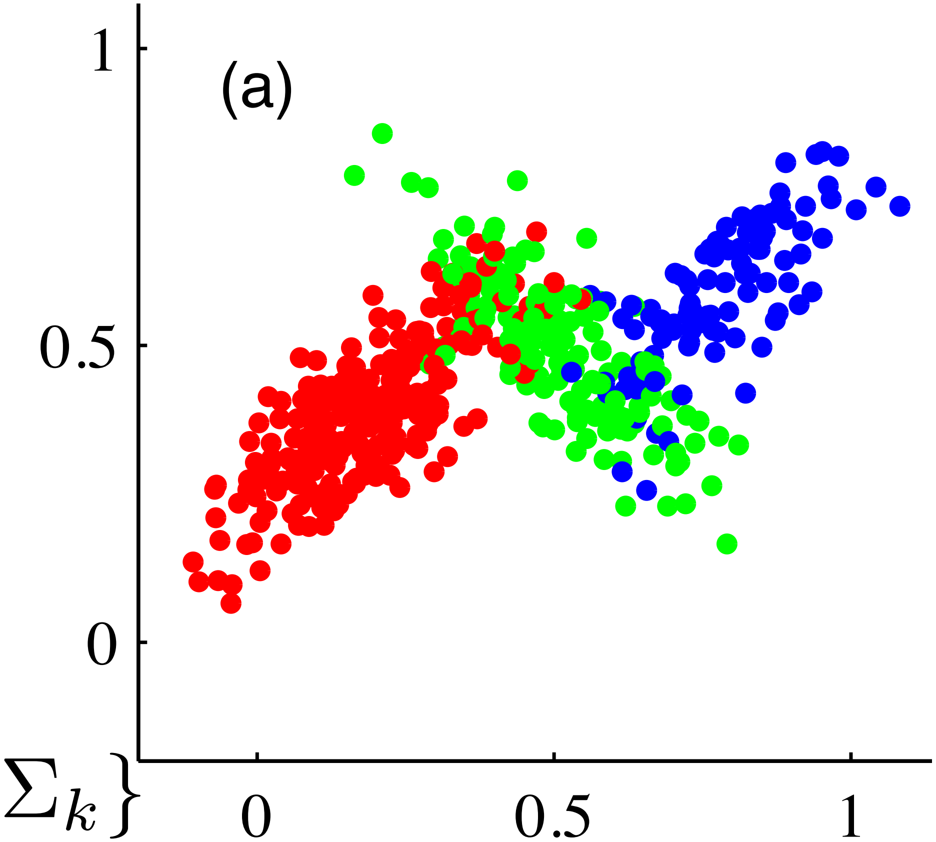
- But suppose that the conditional distribution $p(x \mid z)$ is tractable:

$$\begin{aligned}
\mathbb{E}_p[f(x,z)] &= \int_{\mathcal{Z}} \int_{\mathcal{X}} f(x,z) p(x \mid z)\, p(z)\, dx\, dz \\
&= \int_{\mathcal{Z}} \left[\int_{\mathcal{X}} f(x,z) p(x \mid z)\ dx \right] p(z)\, dz \\
&\approx \frac{1}{L} \sum_{\ell=1}^{L} \int_{\mathcal{X}} f(x, z^{(\ell)}) p(x \mid z^{(\ell)})\ dx = \mathbb{E}_{\tilde{p}}[\mathbb{E}_p[f(x,z) \mid z]]
\end{aligned}$$

- Rao-Blackwell: Collapsed estimator *always* has lower variance

Probabilistic Mixture Models





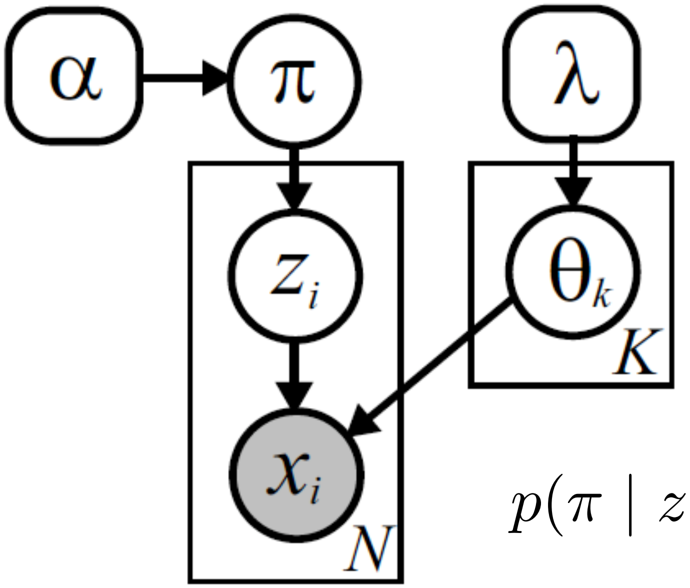
$$\pi \sim \mathrm{Dir}(\alpha)$$

$$\theta_k \sim H(\lambda) \qquad \theta_k = \{\mu_k, \Sigma_k\}$$

$$p(z_i \mid \pi) = \mathrm{Cat}(z_i \mid \pi)$$

$$p(x_i \mid z_i, \mu, \Sigma) = \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

A Collapsed Monte Carlo Estimator



$$(z^{(\ell)}, \theta^{(\ell)}, \pi^{(\ell)}) \sim p(z, \pi, \theta \mid x)$$
$$\ell = 1, 2, \ldots, L$$

Approximate joint samples from Gibbs:

$$p(z_i = k \mid x, \pi, \theta) \propto \pi_k f(x_i \mid \theta_k)$$

$$p(\pi \mid z, x, \theta) = \mathrm{Dir}(\pi \mid N_1 + \alpha/K, \ldots, N_K + \alpha/K)$$

- A conventional estimator of the probability that a pair of observations comes from the same cluster:

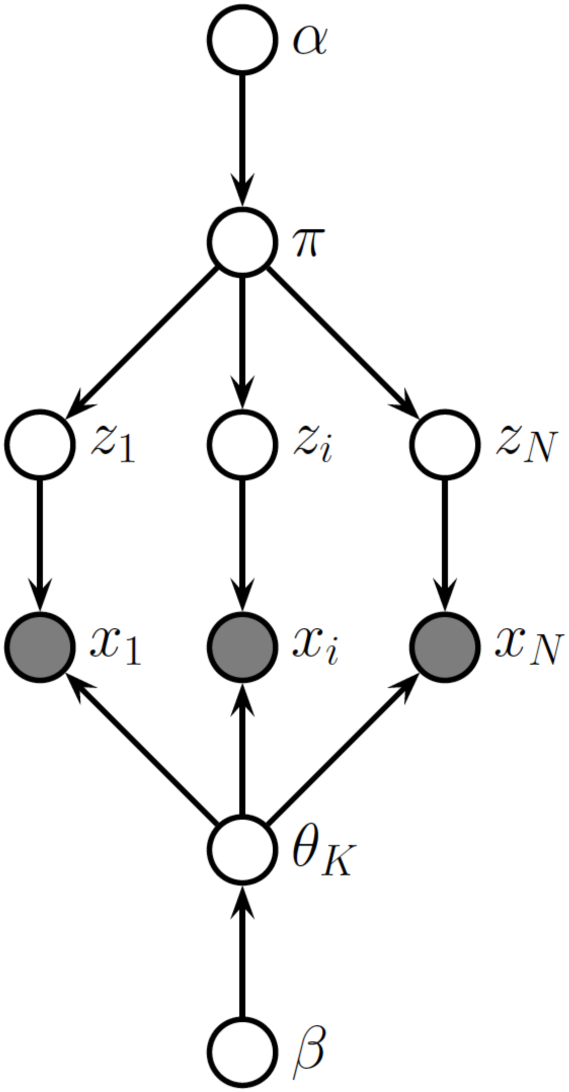
$$p(z_i = z_j) \approx \frac{1}{L} \sum_{\ell=1}^{L} \delta(z_i^{(\ell)}, z_j^{(\ell)})$$

Note choice of statistic which avoids "label switching"

- A provably superior, collapsed estimator of the probability that a pair of observations comes from the same cluster:

$$p(z_i = z_j) \approx \frac{1}{L} \sum_{\ell=1}^{L} \sum_{k=1}^{K} q_{ik}^{(\ell)} q_{jk}^{(\ell)} \qquad q_{ik}^{(\ell)} = p(z_i = k \mid x, \pi^{(\ell)}, \theta^{(\ell)})$$

Collapsed Sampling Algorithms

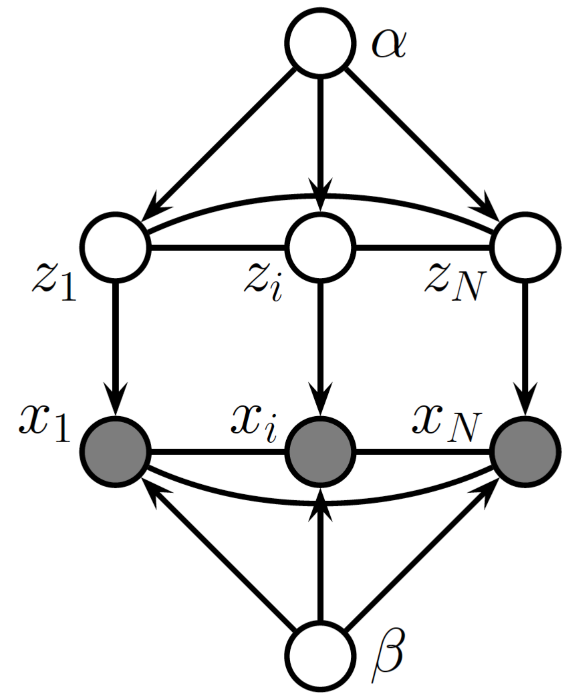


$$\pi \sim \mathrm{Dir}(\alpha)$$

$$z_i \sim \mathrm{Cat}(\pi)$$

$$x_i \sim F(\theta_{z_i})$$

$$\theta_k \sim G(\beta)$$



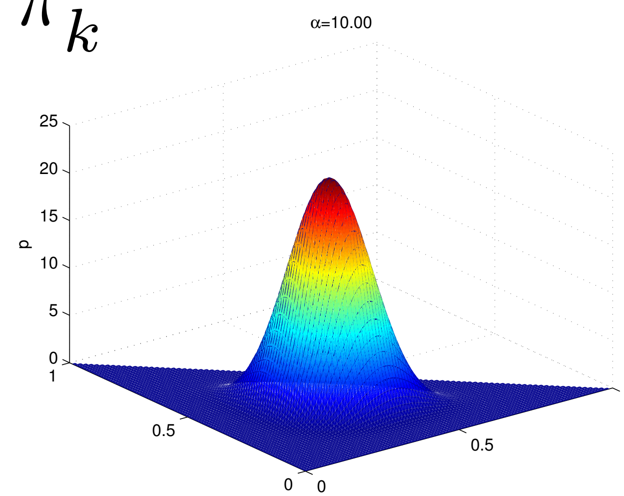
Conjugate priors allow exact marginalization of parameters, to make an equivalent model with fewer variables

Bayesian Learning of Probabilities

Multinoulli Distribution: Single roll of a (possibly biased) die

$$\mathrm{Cat}(z \mid \pi) = \prod_{k=1}^{K} \pi_k^{z_k} \qquad \mathcal{Z} = \{0,1\}^K, \sum_{k=1}^{K} z_k = 1$$

$$p(z_1, \ldots, z_N \mid \pi) = \prod_{k=1}^{K} \pi_k^{N_k}$$



Dirichlet Prior Distribution:

$$p(\pi) = \mathrm{Dir}(\pi \mid \alpha) \propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

Posterior Distribution:

$$p(\pi \mid z) \propto \prod_{k=1}^{K} \pi_k^{N_k + \alpha_k - 1} \propto \mathrm{Dir}(\pi \mid N_1 + \alpha_1, \ldots, N_K + \alpha_K)$$

- *This is a conjugate prior, because posterior is in same family*

Bayesian Learning of Probabilities

Posterior Predictive Distribution: For the next observation,

$$p(\bar{z} = k \mid z_1, \ldots, z_N) = \int_{\Pi} \pi_k p(\pi \mid z_1, \ldots, z_N) \, d\pi$$

$$= \frac{N_k + \alpha_k}{N + \alpha_0} = \mathbb{E}[\pi_k \mid z_1, \ldots, z_N]$$

Dirichlet Prior Distribution:

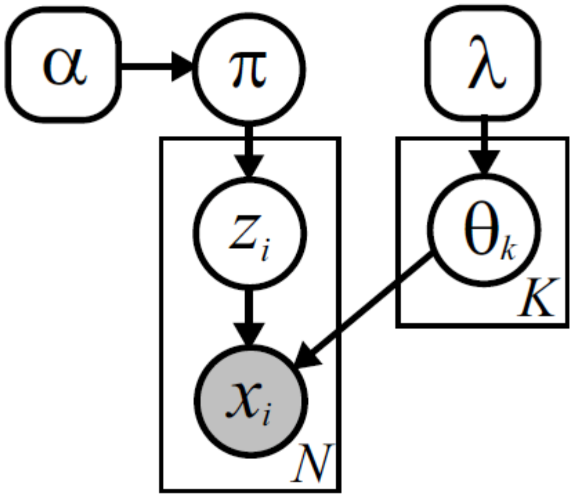
$$p(\pi) = \mathrm{Dir}(\pi \mid \alpha) \propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

Posterior Distribution:

$$p(\pi \mid z) \propto \prod_{k=1}^{K} \pi_k^{N_k + \alpha_k - 1} \propto \mathrm{Dir}(\pi \mid N_1 + \alpha_1, \ldots, N_K + \alpha_K)$$

- *This is a conjugate prior, because posterior is in same family*

A Collapsed Gibbs Sampler



- Collapsed mixture model representation:

$$p(z \mid x) \propto p(z)p(x \mid z)$$

$$\propto \int_{\Pi} p(z \mid \pi)p(\pi \mid \alpha)\, d\pi \int_{\Theta} p(x \mid z, \theta)p(\theta \mid \lambda)\, d\theta$$

- Apply standard Gibbs sampling updates:

$$p(z_i \mid z_{\setminus i}, x) \propto p(z_i \mid z_{\setminus i})p(x \mid z_i, z_{\setminus i})$$

- Conditional prior:

$$N_k^{\setminus i} = \sum_{j=1, j\neq i}^{N} \delta(z_j, k) \qquad p(z_i = k \mid z_{\setminus i}) = \frac{N_k^{\setminus i} + \alpha/K}{N - 1 + \alpha}$$

- Conditional likelihood:

$$X_k^{\setminus i} \triangleq \{x_j \mid z_j = k, j \neq i\} \qquad p(x \mid z) \propto p(x_i \mid z, x_{\setminus i})$$

$$p(x_i \mid z_i = k, z_{\setminus i}, x_{\setminus i}) = \int_{\Theta_k} p(x_i \mid \theta_k)p(\theta_k \mid X_k^{\setminus i})\, d\theta_k$$

Conjugate analysis of "other" data assigned to this cluster

Mixture Sampler Pseudocode

Given previous cluster assignments $z^{(t-1)}$, sequentially sample new assignments as follows:

1. Sample a random permutation $\tau(\cdot)$ of the integers $\{1, \ldots, N\}$.
2. Set $z = z^{(t-1)}$. For each $i \in \{\tau(1), \ldots, \tau(N)\}$, sequentially resample z_i as follows:
   (a) For each of the K clusters, determine the predictive likelihood
   $$f_k(x_i) = p(x_i \mid \{x_j \mid z_j = k, j \neq i\}, \lambda)$$
   This likelihood can be computed from cached sufficient statistics via Prop. 2.1.4.
   (b) Sample a new cluster assignment z_i from the following multinomial distribution:
   $$z_i \sim \frac{1}{Z_i} \sum_{k=1}^{K} (N_k^{-i} + \alpha/K) f_k(x_i) \delta(z_i, k) \qquad Z_i = \sum_{k=1}^{K} (N_k^{-i} + \alpha/K) f_k(x_i)$$
   N_k^{-i} is the number of other observations assigned to cluster k (see eq. (2.162)).
   (c) Update cached sufficient statistics to reflect the assignment of x_i to cluster z_i.
3. Set $z^{(t)} = z$. Optionally, mixture parameters may be sampled via steps 2–3 of Alg. 2.1.

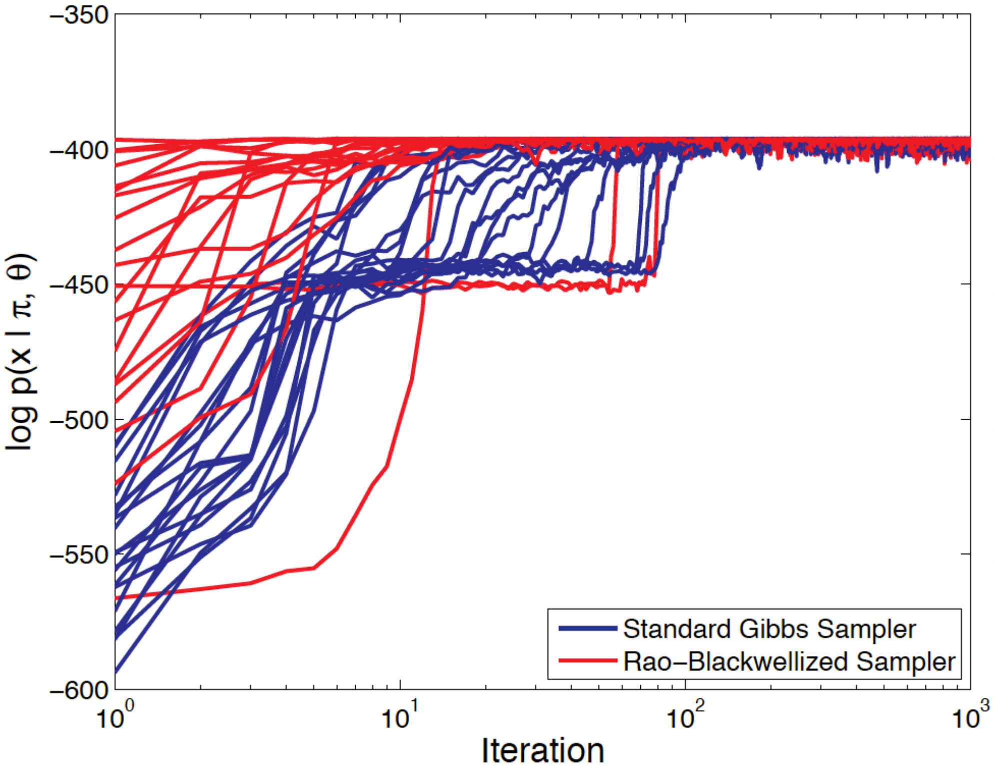
$$p(\theta \mid x^{(1)}, \ldots, x^{(L)}, \lambda) = p(\theta \mid \bar{\lambda}) \tag{2.31}$$

$$\bar{\lambda}_0 = \lambda_0 + L \qquad \bar{\lambda}_a = \frac{\lambda_0 \lambda_a + \sum_{\ell=1}^{L} \phi_a(x^{(\ell)})}{\lambda_0 + L} \qquad a \in \mathcal{A} \tag{2.32}$$

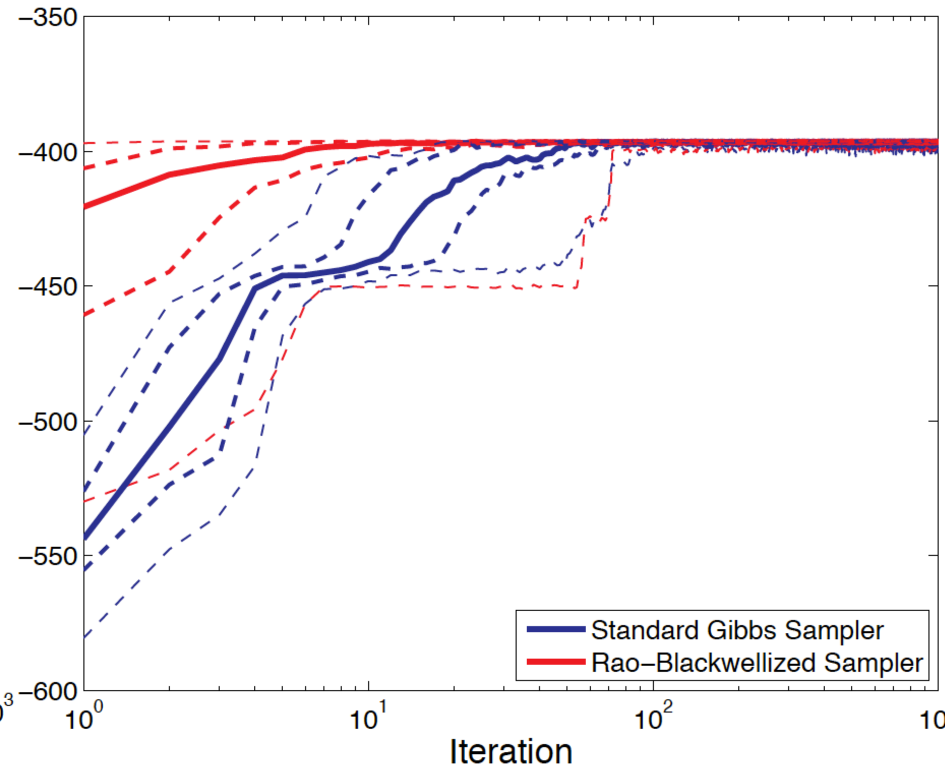
Integrating over Θ, the log–likelihood of the observations can then be compactly written using the normalization constant of eq. (2.29):

$$\log p(x^{(1)}, \ldots, x^{(L)} \mid \lambda) = \Omega(\bar{\lambda}) - \Omega(\lambda) + \sum_{\ell=1}^{L} \log \nu(x^{(\ell)}) \tag{2.33}$$

Gibbs: Representation and Mixing



Multiple Initializations



Quantiles of 100 Chains

Standard Gibbs: Alternatively sample assignments, parameters

Collapsed Gibbs: Marginalize parameters, sample assignments

Variational Approximate Inference

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

- Choose a family of approximating distributions which is tractable. The simplest example:

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

- Define a distance to measure the quality of different approximations. Two possibilities:

$$D(p \,||\, q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \qquad D(q \,||\, p) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

- Find the approximation minimizing this distance

Fully Factored Approximations

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

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

$$D(p \,||\, q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$= \sum_{i\in\mathcal{V}} H(p_i) - H(p) + \sum_{i\in\mathcal{V}} D(p_i \,||\, q_i)$$

Marginal Entropies

Joint Entropy

- Trivially minimized by setting $q_i(x_i) = p_i(x_i)$
- Doesn't provide a computational method…

Variational Approximations

$$D(q(x) \,||\, p(x \mid y)) = \sum_x q(x) \log \frac{q(x)}{p(x \mid y)}$$

$$\begin{aligned}
\log p(y) &= \log \sum_x p(x, y) \\
&= \log \sum_x q(x) \frac{p(x, y)}{q(x)} && \text{(Multiply by one)} \\
&\geq \underbrace{\sum_x q(x) \log \frac{p(x, y)}{q(x)}}_{= -D(q(x) \,||\, p(x \mid y)) + \log p(y)} && \text{(Jensen's inequality)}
\end{aligned}$$

- Minimizing KLD maximizes lower bound on data likelihood
- Generalize EM by restricting to *tractable families*

Free Energies

$$p(x \mid y) = \frac{1}{Z}\exp\{-E(x)\}$$

$$\begin{aligned} D(q \,||\, p) &= \sum_x q(x)\log q(x) - \sum_x q(x)\log p(x \mid y) \\ &= \underbrace{\underbrace{-H(q)}_{\text{Negative Entropy}} + \underbrace{\sum_x q(x)E(x)}_{\text{Average Energy}}}_{\text{Gibbs Free Energy}} + \underbrace{\log Z}_{\text{Normalization}} \end{aligned}$$

- Free energies equivalent to KL divergence, up to a fixed normalization constant that can be ignored
- Variational inference equivalent to “energy minimization”

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):

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

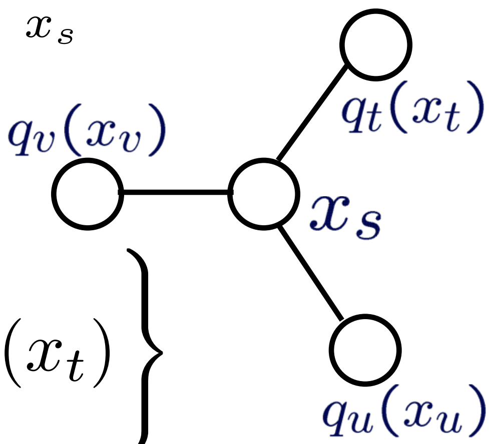
Mean Field Equations

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

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

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

- Add Lagrange multipliers to enforce $\sum_{x_s} q_s(x_s) = 1$
- Taking derivatives and simplifying, we find a set of fixed point equations:



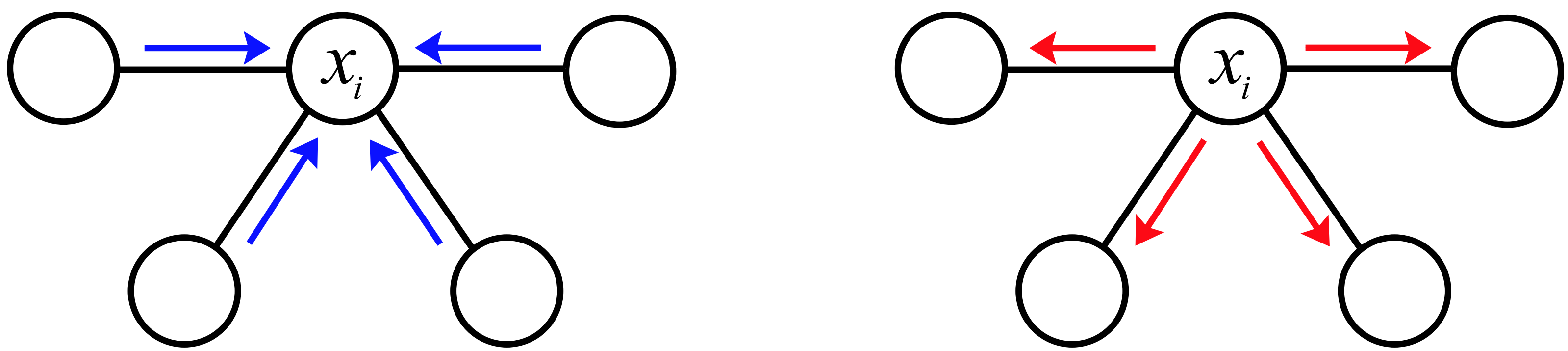
$$q_s(x_s) \propto \psi_s(x_s) \prod_{t \in \Gamma(s)} \exp\left\{ -\sum_{x_t} \phi_{st}(x_s, x_t) q_t(x_t) \right\}$$

- Updating one marginal at a time gives convergent coordinate descent

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