CS 228T: Probabilistic graphical models – theoretical foundations
Lecture 2

Kevin Murphy

13 April 2012

Figures from *Machine Learning: a Probabilistic Perspective* by Kevin Murphy.

Outline

Changes to the course

- Grading: 9 problem sets, all mandatory, each worth 11% of final grade. No late days (except for documented medical emergencies). No final.
- Reading material is always posted to `https://sites.google.com/site/cs228tspring2012`.
- Links to Courseware require that you login first. All requests from Stanford students to join will now be automatically accepted. If you have already joined, you may need to rejoin.
- The bookstore will reprint my book within 24 hours.

Outline

Basic idea

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s) \tag{1}$$

- $\overline{x} = \frac{1}{S}\sum_{s=1}^{S} x_s \to \mathbb{E}[X]$
- $\frac{1}{S}\sum_{s=1}^{S}(x_s - \overline{x})^2 \to \mathrm{var}[X]$
- $\frac{1}{S}\#\{x_s \leq c\} \to P(X \leq c)$
- $\mathrm{median}\{x_1, \ldots, x_S\} \to \mathrm{median}(X)$

Monte Carlo integration for estimating π

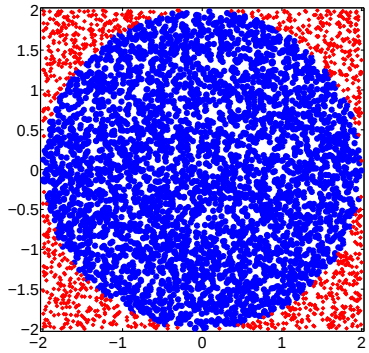
Area of circle with radius r:

$$I = \int_{-r}^{r} \int_{-r}^{r} \mathbb{I}(x^2 + y^2 \leq r^2) dx dy \tag{2}$$

Let $f(x, y) = \mathbb{I}(x^2 + y^2 \leq r^2)$.

$$I = (2r)(2r) \int \int f(x, y) p(x) p(y) dx dy \tag{3}$$

$$\approx 4r^2 \frac{1}{S} \sum_{s=1}^{S} f(x_s, y_s) \tag{4}$$



Posterior inference for $p(\theta_1 > \theta_2|\mathcal{D})$

Supose you are about to buy something from Amazon.com, and there are two sellers offering it for the same price. Seller 1 has 90 positive reviews and 10 negative reviews. Seller 2 has 2 positive reviews and 0 negative reviews. Who should you buy from?

Let θ_1 and θ_2 be the unknown reliabilities of the two sellers. Since we don't know much about them, we'll endow them both with uniform priors, $\theta_i \sim \mathrm{Beta}(1, 1)$. The posteriors are $p(\theta_1|\mathcal{D}_1) = \mathrm{Beta}(91, 11)$ and $p(\theta_2|\mathcal{D}_2) = \mathrm{Beta}(3, 1)$. We want to compute $p(\theta_1 > \theta_2|\mathcal{D})$.

Numerical integration for $p(\theta_1 > \theta_2|\mathcal{D})$

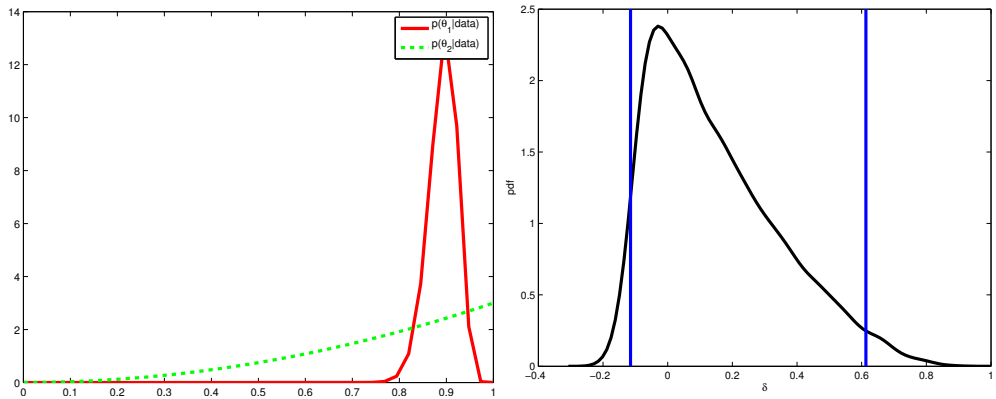
let us define $\delta = \theta_1 - \theta_2$ as the difference in the rates. (Alternatively we might want to work in terms of the log-odds ratio.) We can compute the desired quantity using numerical integration:

$$p(\delta > 0|\mathcal{D}) = \int_0^1 \int_0^1 \mathbb{I}(\theta_1 > \theta_2)\mathrm{Beta}(\theta_1|y_1 + 1, N_1 - y_1 + 1) \mathrm{Beta}(\theta_2|y_2 + 1, N_2 - y_2 + 1) d\theta_1 d\theta_2 \quad (5)$$

We find $p(\delta > 0|\mathcal{D}) = 0.710$, which means you are better off buying from seller 1!

MC integration for $p(\theta_1 > \theta_2|\mathcal{D})$

Sample $\theta_i \sim \mathrm{Beta}(y_i + 1, N_i - y_1 + 1)$. An MC approximation to $p(\delta > 0|\mathcal{D})$ is obtained by counting the fraction of samples where $\theta_1 > \theta_2$; this turns out to be 0.718, which is very close to the exact value.



Two key questions

1. How many samples do we need?

2. How do we generate the samples?

How many samples do we need?

If we denote the exact mean by $\mu = \mathbb{E}[f(X)]$, and the MC approximation by $\hat{\mu}$, one can show that, with *independent samples*,

$$(\hat{\mu} - \mu) \to \mathcal{N}(0, \frac{\sigma^2}{S}) \tag{6}$$

where

$$\sigma^2 = \mathrm{var}[f(X)] = \mathbb{E}\left[f(X)^2\right] - \mathbb{E}[f(X)]^2 \tag{7}$$

This is a consequence of the central-limit theorem.
Of course, σ^2 is unknown in the above expression, but it can also be estimated by MC:

$$\hat{\sigma}^2 = \frac{1}{S}\sum_{s=1}^{S}(f(x_s) - \hat{\mu})^2 \tag{8}$$

Then we have

$$P\left\{\mu - 1.96\frac{\hat{\sigma}}{\sqrt{S}} \leq \hat{\mu} \leq \mu + 1.96\frac{\hat{\sigma}}{\sqrt{S}}\right\} \approx 0.95 \tag{9}$$

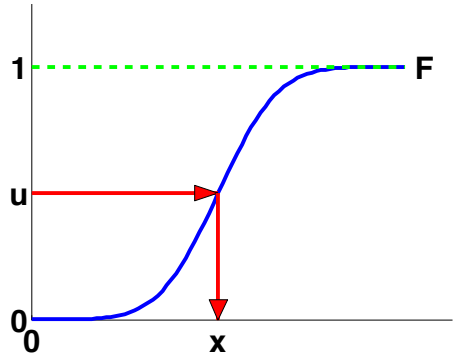
How many samples do we need?

The term $\sqrt{\frac{\hat{\sigma}^2}{S}}$ is called the (numerical or empirical) **standard error**, and is an estimate of our uncertainty about our estimate of μ.

If we want to report an answer which is accurate to within $\pm\epsilon$ with probability at least 95%, we need to use a number of samples S which satisfies $1.96\sqrt{\hat{\sigma}^2/S} \leq \epsilon$. We can approximate the 1.96 factor by 2, yielding $S \geq \frac{4\hat{\sigma}^2}{\epsilon^2}$.

When samples are correlated, we will need many more (see later).

How to generate samples?



- We assume we can sample $u \sim U(0, 1)$ using a pseudo random number generator.
- We can sample from standard univariate distributions with known cdf F as shown above.
- For more complex distributions, we can draw independent samples using rejection sampling, importance sampling, etc.
- For very complex distributions, we can use Markov Chain Monte Carlo (MCMC)

Outline

Basic idea

Goal: approximate

$$I = \mathbb{E}[f] = \int f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \tag{10}$$

Sample $\mathbf{x}^s \sim q()$ from a proposal. Then

$$\mathbb{E}[f] = \int f(\mathbf{x}) \frac{p(\mathbf{x})}{q(\mathbf{x})} q(\mathbf{x}) d\mathbf{x} \approx \frac{1}{S} \sum_{s=1}^{S} w_s f(\mathbf{x}^s) = \hat{I} \tag{11}$$

where $w_s \triangleq \frac{p(\mathbf{x}^s)}{q(\mathbf{x}^s)}$ are the **importance weights**.

Want to make q as close as possible to p to reduce variance of the weights.

Handling unnormalized distributions

Let $p(\mathbf{x}) = \tilde{p}(\mathbf{x})/Z_p$, $q(\mathbf{x}) = \tilde{q}(\mathbf{x})/Z_q$.
First we evaluate

$$\mathbb{E}[f] = \frac{Z_q}{Z_p} \int f(\mathbf{x}) \frac{\tilde{p}(\mathbf{x})}{\tilde{q}(\mathbf{x})} q(\mathbf{x}) d\mathbf{x} \approx \frac{Z_q}{Z_p} \frac{1}{S} \sum_{s=1}^{S} \tilde{w}_s f(\mathbf{x}^s) \tag{12}$$

where $\tilde{w}_s \triangleq \frac{\tilde{p}(\mathbf{x}^s)}{\tilde{q}(\mathbf{x}^s)}$ is the unnormalized importance weight. Hence

$$\hat{I} = \frac{\frac{1}{S} \sum_s \tilde{w}_s f(\mathbf{x}^s)}{\frac{1}{S} \sum_s \tilde{w}_s} = \sum_{s=1}^{S} w_s f(\mathbf{x}^s) \tag{13}$$

where

$$w_s \triangleq \frac{\tilde{w}_s}{\sum_{s'} \tilde{w}_{s'}} \tag{14}$$

are the normalized importance weights.

Likelihood weighting for a Bayes net

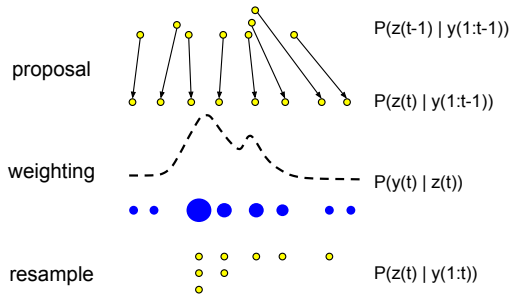
- Forward sampling from prior is easy.
- To sample from posterior, sample each unclamped node and weight clamped nodes by their probability given their parents.
- Equivalent to this proposal:

$$q(\mathbf{x}) = \prod_{t \notin E} p(x_t | \mathbf{x}_{\mathrm{pa}(t)}) \prod_{t \in E} \delta_{x_t^*}(x_t) \tag{15}$$

- Weight is

$$w(\mathbf{x}) = \frac{p(\mathbf{x})}{q(\mathbf{x})} = \prod_{t \notin E} \frac{p(x_t | \mathbf{x}_{\mathrm{pa}(t)})}{p(x_t | \mathbf{x}_{\mathrm{pa}(t)})} \prod_{t \in E} \frac{p(x_t | \mathbf{x}_{\mathrm{pa}(t)})}{1} = \prod_{t \in E} p(x_t | \mathbf{x}_{\mathrm{pa}(t)})(1$$

Particle filtering



PF is a form of importance sampling over trajectories, combined with a resampling step. See lecture 1.

Quiz

1. Suppose we are trying to estimate an integral of the form

$$\mathbb{E}[g(X)] = \int f(x)g(x)dx \tag{17}$$

   where $f(x)$ is a probability density function, and where $g(x)$. What is the standard Monte Carlo estimator, a_n?
2. Now suppose f is large only in places where g is small, and vice versa. What will happen to the standard Monte Carlo estimator?
3. Now suppose that we want to estimate the expectation using importance sampling, where we sample from a different distribution h, and weight the terms accordingly in our estimator. What does our new estimator look like?

Quiz

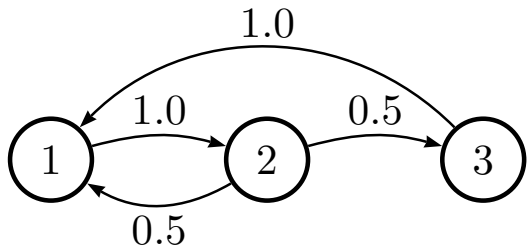
1. Recall that we want to estimate

$$\mathbb{E}[g(X)] = \int f(x)g(x)dx \tag{18}$$

   using importance sampling, where f is the pdf. Intuitively, what is the optimal proposaldistribution h to sample from?
2. If we do this, what does our estimator become? What is the variance, in terms of the number of samples?
3. What is the catch with using this proposal?
4. What are some alternative approaches that this analysis suggests?

Outline

What is a stationary distribution?

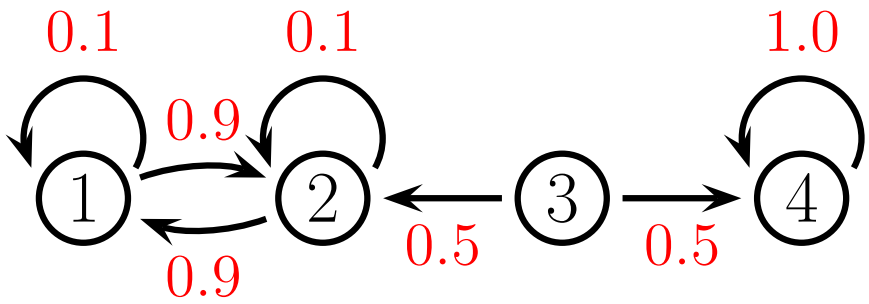


Let $A_{ij} = p(X_t = j | X_{t-1} = i)$ be the one-step transition matrix, and let $\pi_t(j) = p(X_t = j)$ be the probability of being in state j at time t. Stationary distribution satisfies

$$\boldsymbol{\pi} = \boldsymbol{\pi} \mathbf{A} \tag{19}$$

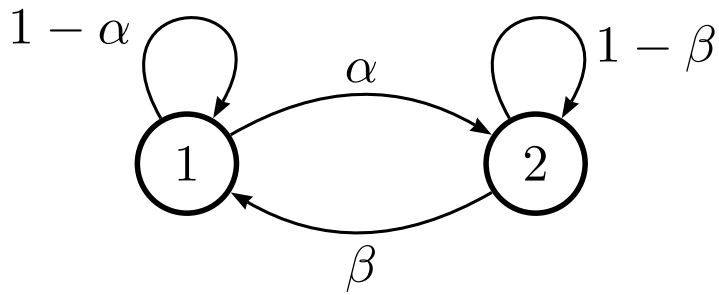
$\boldsymbol{\pi} = (0.4, 0.4, 0.2)$

When does a unique stationary distribution exist?



- If we start in 4, $\boldsymbol{\pi} = (0, 0, 0, 1)$ is one possible stationary distribution.
- If we start in 1 or 2, $\boldsymbol{\pi} = (0.5, 0.5, 0, 0)$ is one possible stationary distribution.
- If we start in 3, could end up in either.
- A necessary condition for a unique $\boldsymbol{\pi}$ is that the transition diagram is singly connected, i.e., chain is **irreducible**.

When does a unique stationary distribution exist?



- This is irreducible provided $\alpha, \beta > 0$.
- Suppose $\alpha = \beta = 0.9$. It is clear by symmetry that this chain will spend 50% of its time in each state. Thus $\boldsymbol{\pi} = (0.5, 0.5)$.
- If $\alpha = \beta = 1$, the system never forgets where it started. If started in an odd state, it will be in an odd state for every odd time step.
- So we also require the chain to be **aperiodic**.
- A sufficient condition is to add a self loop, cf Page-Rank.

When does a unique stationary distribution exist?

Theorem

Every irreducible (singly connected), aperiodic finite state Markov chain has a unique stationary distribution.

For the non finite state space case, we need an additional property called **ergodicity**.

Regular Markov chains

A finite-state Markov chain is **regular** if $A_{ij}^n > 0$ for some integer n and all i, j, i.e., it is possible to get from any state to any other state in n steps. Consequently, after n steps, the chain could be in any state, no matter where it started.

Theorem

If the MC is regular, it has a unique stationary distribution.

A sufficient condition to ensure regularity is the graph is singly connected and every state has a self-loop.

Detailed balance

Detailed balance:

$$\pi_i A_{ij} = \pi_j A_{ji} \tag{20}$$

Theorem

If a Markov chain with transition matrix $\mathbf{A}$ *is regular and satisfies detailed balance wrt distribution* $\boldsymbol{\pi}$*, then* $\boldsymbol{\pi}$ *is a stationary distribution of the chain.*

Proof.

To see this, note that

$$\sum_i \pi_i A_{ij} = \sum_i \pi_j A_{ji} = \pi_j \sum_i A_{ji} = \pi_j \tag{21}$$

and hence $\boldsymbol{\pi} = \mathbf{A}\boldsymbol{\pi}$. □

Quiz

Consider the following two statements:

1. A Markov chain with transition operator T satisfies detailed balance with respect to a distribution p, *i.e.*,

$$p(x)\,T(x \to x') = p(x')\,T(x' \to x)$$

   holds for all possible outcomes x, x'.
2. The distribution p is a stationary distribution for T.

Assume here that the Markov chains mentioned are regular and have finite state space, so they can get from any state to any other state in a finite number of steps with positive probability.
Which of the following is true?

(a) (i) implies (ii), but not vice versa
(b) (ii) implies (i), but not vice versa
(c) (i) if and only if (ii)
(d) (i) and (ii) are unrelated

Outline

Basic idea

Stochastic coordinate descent

- $x_1^{s+1} \sim p(x_1|x_2^s, x_3^s)$
- $x_2^{s+1} \sim p(x_2|x_1^{s+1}, x_3^s)$
- $x_3^{s+1} \sim p(x_3|x_1^{s+1}, x_2^{s+1})$

$p(x_i|\mathbf{x}_{-i})$ is i's full conditional, depends on its Markov blanket.

Quiz

Consider the Bayesian network $X \to Y \to Z$. If the current sample is (x_0, y_0, z_0), and the first substep of Gibbs sampling is to sample y, with what probability will the first subsample be (x_0, y_1, z_0)?

(a) $p(y_1 \mid x_0)$

(b) $p(y_1 \mid x_0, z_0)$

(c) $p(y_1 \mid z_0)$

(d) $p(y_1)$

Example: Gibbs for Ising model

$x_t \in \{-1, +1\}$

$$p(\mathbf{x}) \propto \prod_{s,t} \exp(J x_s x_t) \prod_t \psi_t(x_t) \tag{22}$$

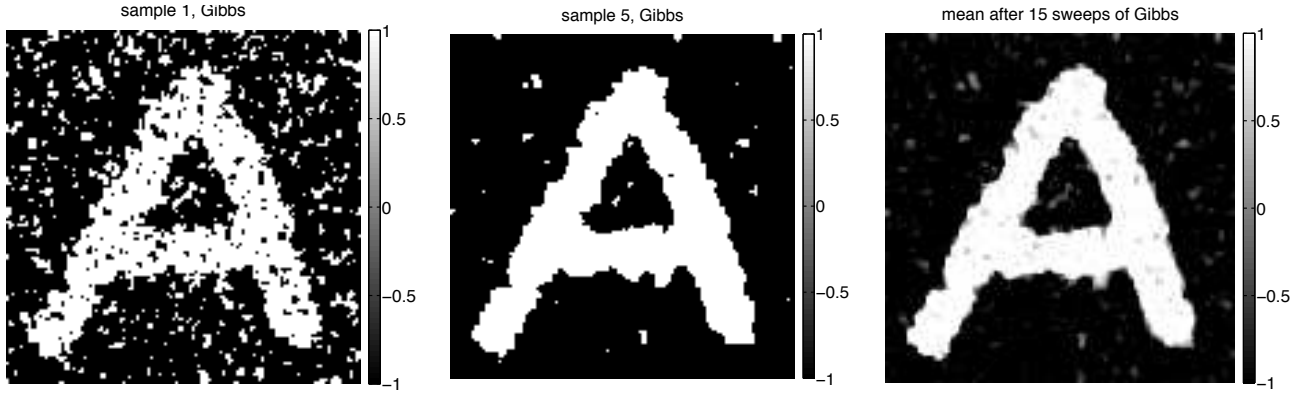
The full conditional becomes

$$p(x_t = +1 | \mathbf{x}_{-t}, \mathbf{y}, \boldsymbol{\theta}) = \frac{\exp[J\eta_t]\psi_t(+1)}{\exp[J\eta_t]\psi_t(+1) + \exp[-J\eta_t]\psi_t(-1)} \tag{23}$$

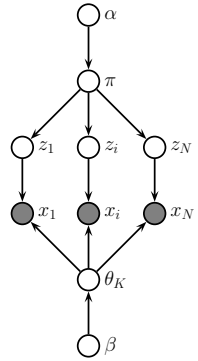
$$= \mathrm{sigm}\left(2J\eta_t - \log \frac{\psi_t(+1)}{\psi_t(-1)}\right) \tag{24}$$

where $\eta_t \triangleq \sum_{s \in \mathrm{nbr}(t)} x_t$.

Example: Gibbs for Ising model



Example: Gibbs for fitting a GMM



Semi-conjugate prior

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}) \tag{25}$$

$$p(z_i|\boldsymbol{\pi}) \sim \mathrm{Cat}(\boldsymbol{\pi}) \tag{26}$$

$$p(\mathbf{x}_i|z_i = k, \boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{27}$$

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\mathbf{m}_0, \mathbf{V}_0) \tag{28}$$

$$\boldsymbol{\Sigma}_k \sim \mathrm{IW}(\mathbf{S}_0, \nu_0) \tag{29}$$

For the discrete indicators, we have

$$p(z_i = k|\mathbf{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \quad \propto \quad \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{30}$$

For the mixing weights, we have

$$p(\boldsymbol{\pi}|\mathbf{z}) \quad = \quad \mathrm{Dir}(\{\alpha_k + \sum_{i=1}^{N} \mathbb{I}(z_i = k)\}_{k=1}^{K}) \tag{31}$$

For the means, we have

$$\begin{aligned} p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k, \mathbf{z}, \mathbf{x}) &= \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_k, \mathbf{V}_k) && (32)\\ \mathbf{V}_k^{-1} &= \mathbf{V}_0^{-1} + N_k \boldsymbol{\Sigma}_k^{-1} && (33)\\ \mathbf{m}_k &= \mathbf{V}_k(\boldsymbol{\Sigma}_k^{-1} N_k \bar{\mathbf{x}}_k + \mathbf{V}_0^{-1}\mathbf{m}_0) && (34)\\ N_k &\triangleq \sum_{i=1}^{N} \mathbb{I}(z_i = k) && (35)\\ \bar{\mathbf{x}}_k &\triangleq \frac{\sum_{i=1}^{N} \mathbb{I}(z_i = k)\mathbf{x}_i}{N_k} && (36) \end{aligned}$$

For the covariances, we have

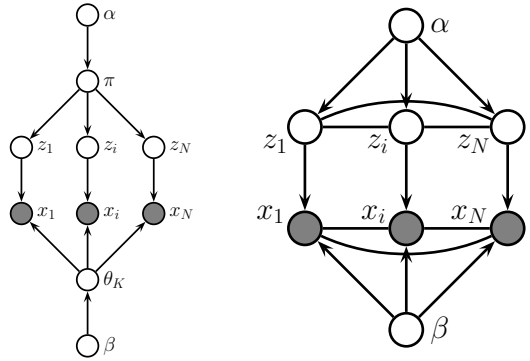
$$p(\boldsymbol{\Sigma}_k|\boldsymbol{\mu}_k, \mathbf{z}, \mathbf{x}) \quad = \quad \mathrm{IW}(\boldsymbol{\Sigma}_k|\mathbf{S}_k, \nu_k) \tag{37}$$

Outline

Basic idea

- In some cases, we can analytically integrate out some hidden variables and just sample what's left over.
- This is called collapsing or Rao-Blackwellisation.
- The variance of the samples will be lower, so we need fewer samples (although it may take longer to create each one).

Example: collapsed Gibbs for fitting a GMM



$$
\begin{aligned}
p(z_i = k|\mathbf{z}_{-i}, \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &\propto p(z_i = k|\mathbf{z}_{-i}, \boldsymbol{\alpha}) p(\mathbf{x}|z_i = k, \mathbf{z}_{-i}, \boldsymbol{\beta}) \\
&\propto p(z_i = k|\mathbf{z}_{-i}, \boldsymbol{\alpha}) p(\mathbf{x}_i|\mathbf{x}_{-i}, z_i = k, \mathbf{z}_{-i}, \boldsymbol{\beta}) \\
&\quad p(\mathbf{x}_{-i}|\mathbf{z}_{-i}, \cancel{z_i = k}, \boldsymbol{\beta}) \\
&\propto p(z_i = k|\mathbf{z}_{-i}, \boldsymbol{\alpha}) p(\mathbf{x}_i|\mathbf{x}_{-i}, z_i = k, \mathbf{z}_{-i}, \boldsymbol{\beta})
\end{aligned}
$$

Aside: computing marginal likelihood for beta-Bernoulli model

Since we know $p(\theta|\mathcal{D}) = \mathrm{Beta}(\theta|a', b')$, where $a' = a + N_1$ and $b' = b + N_0$, we know the normalization constant of the posterior is $B(a', b')$. Hence

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \tag{40}$$

$$= \frac{1}{p(\mathcal{D})}\left[\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\right]\left[\binom{N}{N_1}\theta^{N_1}(1-\theta)^{N_0}\right] \tag{41}$$

$$= \binom{N}{N_1}\frac{1}{p(\mathcal{D})}\frac{1}{B(a,b)}\left[\theta^{a+N_1-1}(1-\theta)^{b+N_0-1}\right] \tag{42}$$

So

$$\frac{1}{B(a+N_1, b+N_0)} = \binom{N}{N_1}\frac{1}{p(\mathcal{D})}\frac{1}{B(a,b)} \tag{43}$$

$$p(\mathcal{D}) = \binom{N}{N_1}\frac{B(a+N_1, b+N_0)}{B(a,b)} \tag{44}$$

Aside: computing marginal likelihood for Dirichlet-multinoulli model

$$p(\mathcal{D}) = \frac{B(\mathbf{N} + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})} \tag{45}$$

where

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)} \tag{46}$$

Hence we can rewrite the above result in the following form, which is what is usually presented in the literature:

$$p(\mathcal{D}) = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N + \sum_k \alpha_k)} \prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)} \tag{47}$$

Collapsed Gibbs for fitting a GMM: full conditional

$$p(z_1, \ldots, z_N|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \alpha/K)}{\Gamma(\alpha/K)} \tag{48}$$

Hence

$$p(z_i = k|\mathbf{z}_{-i}, \alpha) = \frac{p(\mathbf{z}_{1:N}|\alpha)}{p(\mathbf{z}_{-i}|\alpha)} = \frac{\frac{1}{N+\alpha}}{\frac{1}{N+\alpha-1}} \times \frac{\Gamma(N_k + \alpha/K)}{\Gamma(N_{k,-i} + \alpha/K)} \tag{49}$$

$$= \frac{N_{k,-i} + \alpha/K}{N + \alpha - 1} \tag{50}$$

where $N_{k,-i} \triangleq \sum_{n \neq i} \mathbb{I}(z_n = k) = N_k - 1$, and where we exploited the fact that $\Gamma(x+1) = x\Gamma(x)$.

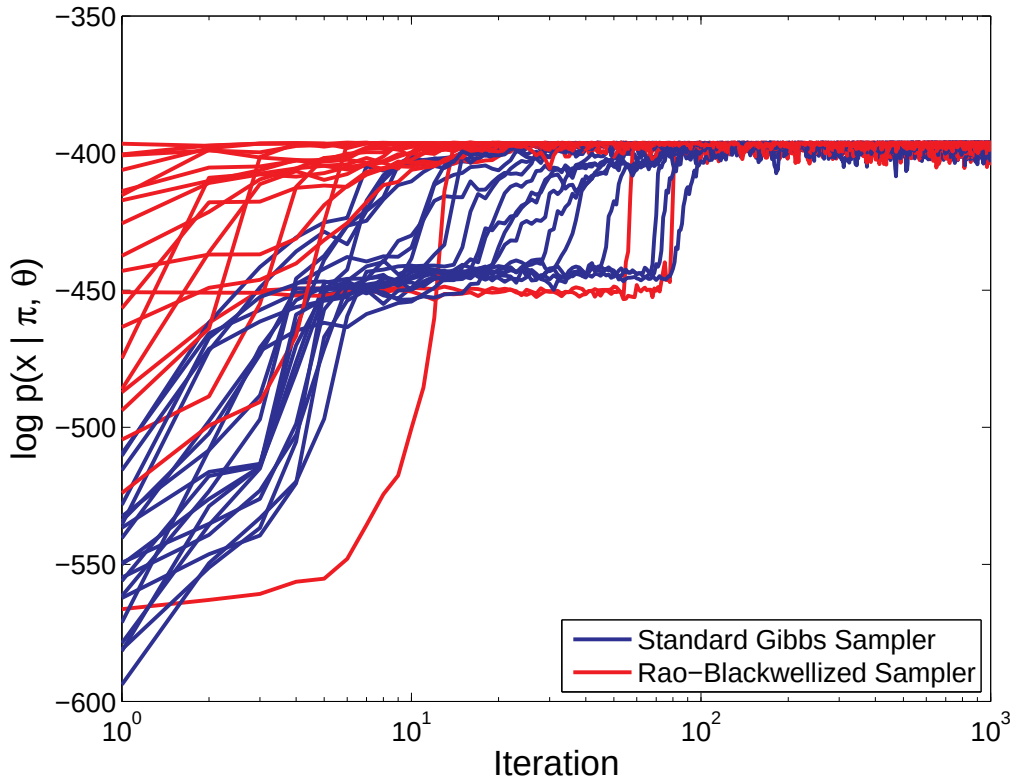
Collapsed Gibbs for fitting a GMM: code

Algorithm 24.1: Collapsed Gibbs sampler for a mixture model

for *each* $i = 1 : N$ *in random order* **do**
  Remove $\mathbf{x}_i$'s sufficient statistics from old cluster z_i ;
  for *each* $k = 1 : K$ **do**
    Compute $p_k(\mathbf{x}_i) \triangleq p(\mathbf{x}_i | \{\mathbf{x}_j : z_j = k, j \neq i\})$;
  Compute $p(z_i = k | \mathbf{z}_{-i}, \mathcal{D}) \propto (N_{k,-i} + \alpha/K) p_k(\mathbf{x}_i)$;
  Sample $z_i \sim p(z_i | \cdot)$;
  Add $\mathbf{x}_i$'s sufficient statistics to new cluster z_i

Collapsed Gibbs for fitting a GMM: comparison

Red = collapsed, K=5 GMM on N=300 points in D=2 dim.



Quiz

Suppose you have a bipartite MRF with two sets of variables, $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$. Assume that n is large, e.g. $n = 1000$. Each X_i is connected to each Y_j, none of the Y_j's are connected to each other, and the X_i's are internally connected using a tree structure. Assume that the edges in the tree structure connecting the X_i's induce very strong correlations between the X_i's that they connect. If you are applying collapsed Gibbs sampling to this MRF, which variables should you use as the sampled variables?

(a) The X_i's.

(b) The Y_j's.

(c) All the variables.

(d) Either the X_i's or the Y_j's, it doesn't matter.

Outline

Basic idea

Propose a move according to $q(\mathbf{x}'|\mathbf{x})$. Accept with prob

$$r = \min(1, \alpha) \tag{51}$$

$$\alpha = \frac{p^*(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}{p^*(\mathbf{x})q(\mathbf{x}'|\mathbf{x})} = \frac{p^*(\mathbf{x}')/q(\mathbf{x}'|\mathbf{x})}{p^*(\mathbf{x})/q(\mathbf{x}|\mathbf{x}')} \tag{52}$$

MC Converges to $p^*(\mathbf{x})$.

Gibbs sampling is a special case of MH

Gibbs is MH with a sequence of proposals of this form:

$$q(\mathbf{x}'|\mathbf{x}) = p(x_i'|\mathbf{x}_{-i})\mathbb{I}(\mathbf{x}'_{-i} = \mathbf{x}_{-i}) \tag{53}$$

Acceptance rate is

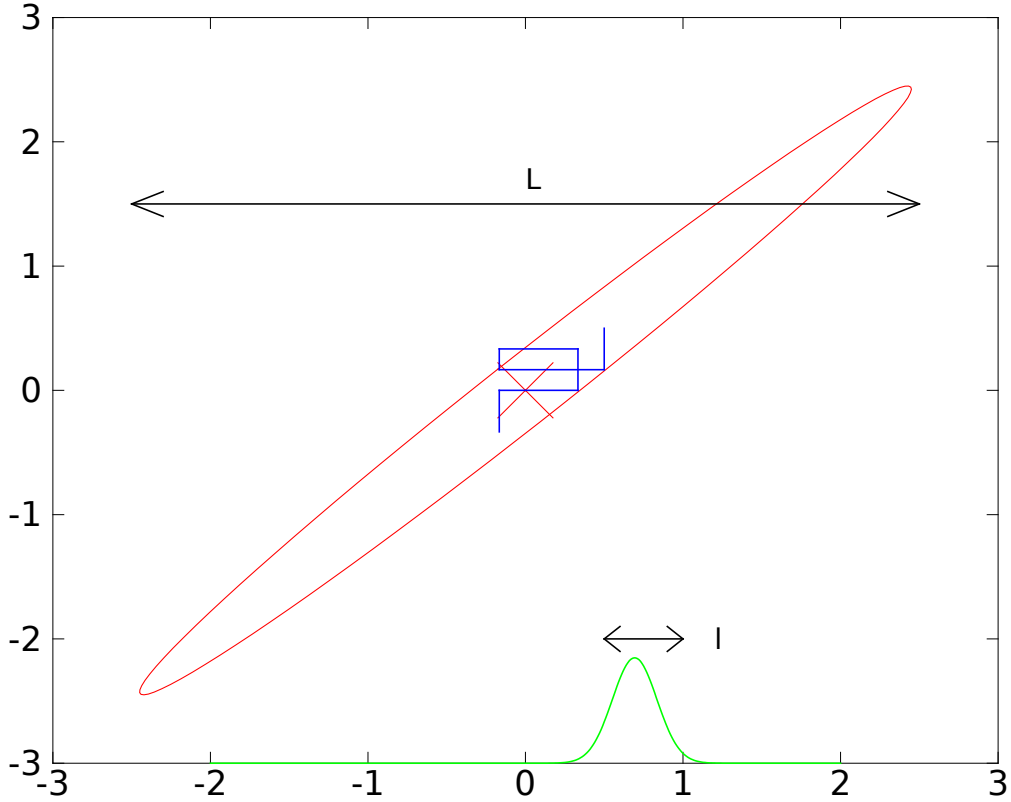
$$\alpha = \frac{p(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}{p(\mathbf{x})q(\mathbf{x}'|\mathbf{x})} = \frac{p(x_i'|\mathbf{x}'_{-i})p(\mathbf{x}'_{-i})p(x_i|\mathbf{x}'_{-i})}{p(x_i|\mathbf{x}_{-i})p(\mathbf{x}_{-i})p(x_i'|\mathbf{x}_{-i})} \tag{54}$$

$$= \frac{p(x_i'|\mathbf{x}_{-i})p(\mathbf{x}_{-i})p(x_i|\mathbf{x}_{-i})}{p(x_i|\mathbf{x}_{-i})p(\mathbf{x}_{-i})p(x_i'|\mathbf{x}_{-i})} = 1 \tag{55}$$

where we exploited the fact that $\mathbf{x}'_{-i} = \mathbf{x}_{-i}$, and that $q(\mathbf{x}'|\mathbf{x}) = p(x_i'|\mathbf{x}_{-i})$.

Gibbs can be slow

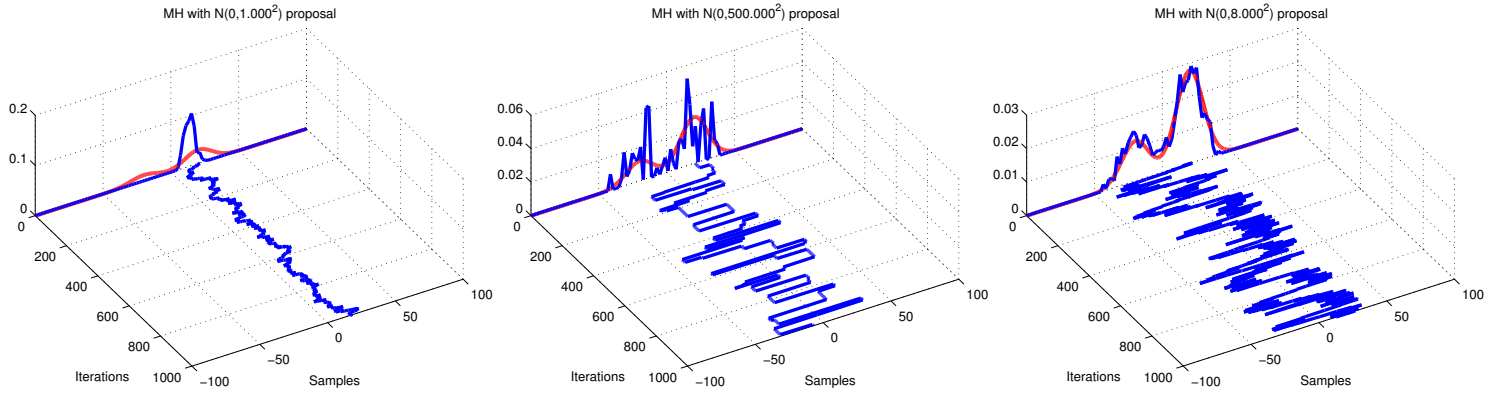
Vanilla gibbs does single site updating.



Blocked gibbs and collapsed Gibbs can be better.

MH proposals

Random walk Metropolis.

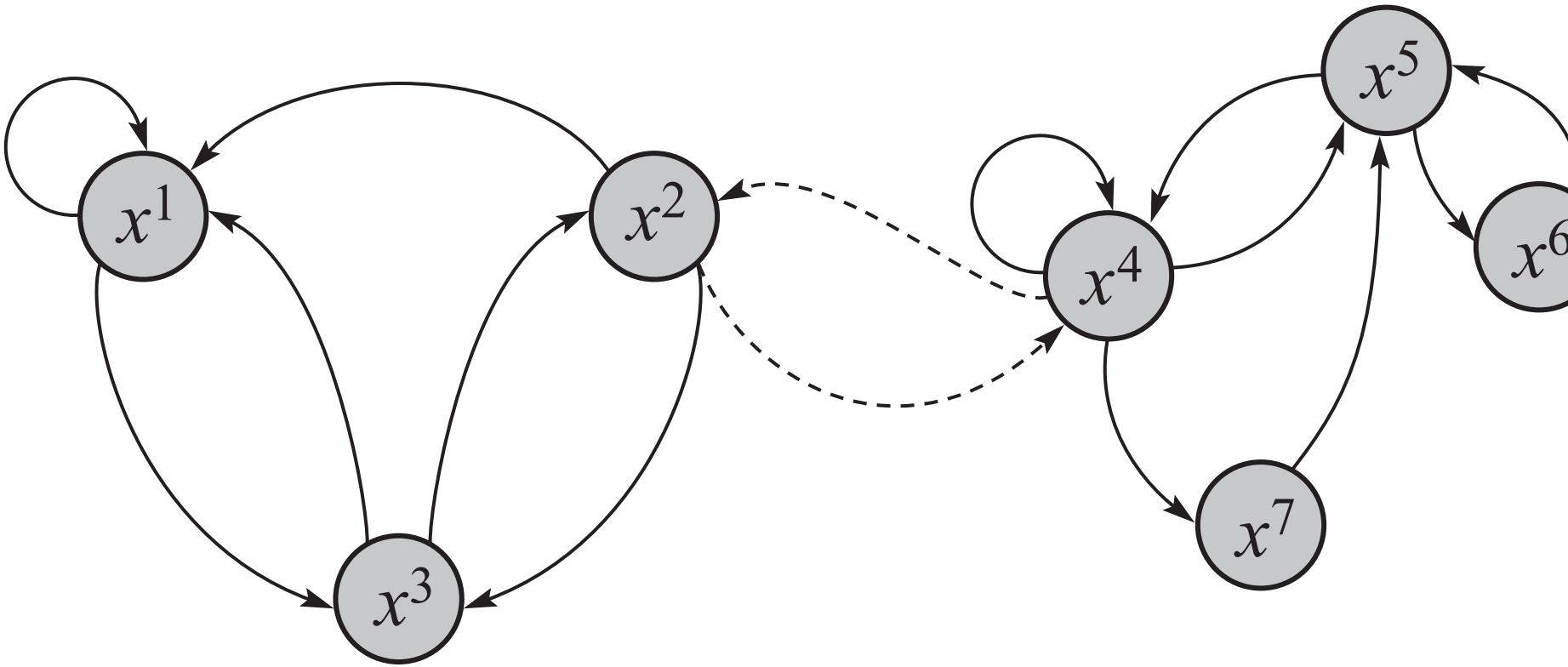


- Adaptive MCMC
- Mixture of proposals
- Data driven proposals

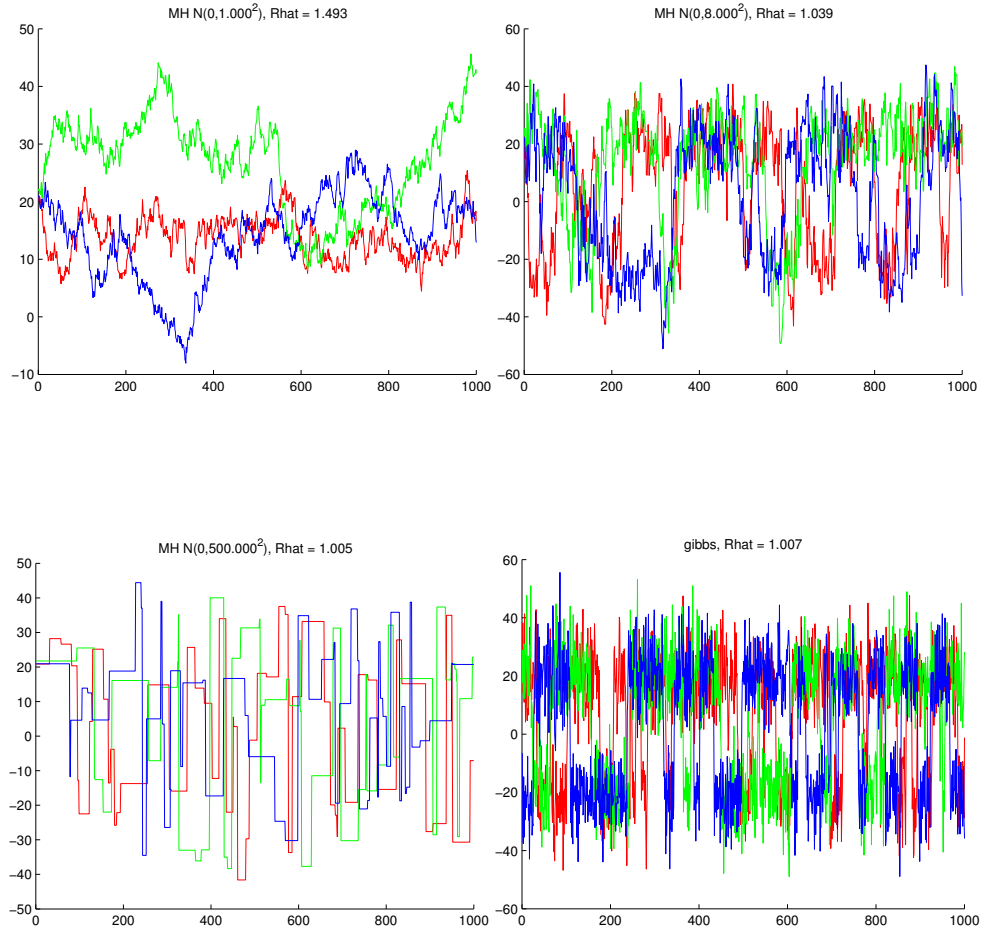
Outline

Theory

Mixing rate is the speed with which we forget the initial distribution. This is slow if the chain has low conductance.

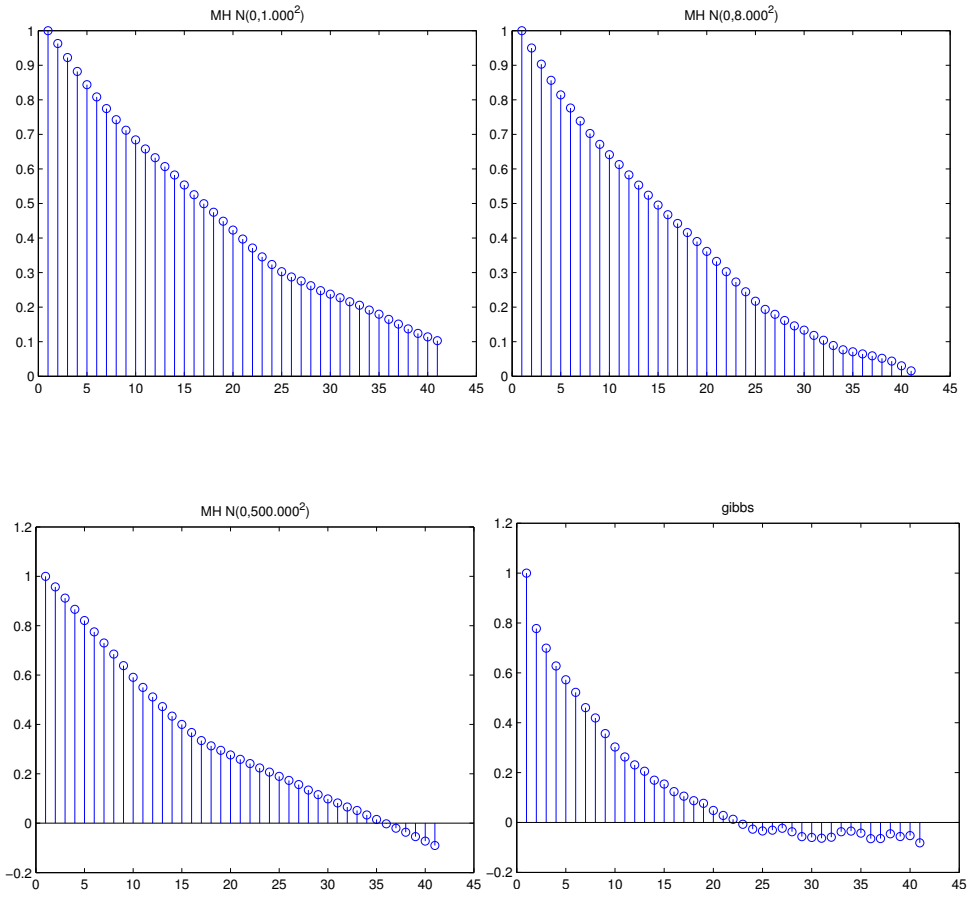


Traceplots



Quantify using EPSR.

Autocorrelation



Effective sample size is lower if samples are autocorrrelated. To save space we can use thinning.

Outline

Basic idea

If the original variables are denoted by $\mathbf{x}$, and the auxiliary variables by $\mathbf{z}$, we require that $\sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x})$, and that $p(\mathbf{x}, \mathbf{z})$ is easier to sample from than just $p(\mathbf{x})$.

Swendsen Wang

Consider an Ising model of the following form:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_e f_e(\mathbf{x}_e) \tag{56}$$

where $\mathbf{x}_e = (x_i, x_j)$ for edge $e = (i, j)$, $x_i \in \{+1, -1\}$, and the edge factor f_e is defined by $\begin{pmatrix} e^J & e^{-J} \\ e^{-J} & e^J \end{pmatrix}$, where J is the edge strength. These are called **bond variables**, and will be denoted by $\mathbf{z}$. We then define an extended model $p(\mathbf{x}, \mathbf{z})$ of the form

$$p(\mathbf{x}, \mathbf{z}) = \frac{1}{Z'} \prod_e g_e(\mathbf{x}_e, z_e) \tag{57}$$

where $z_e \in \{0, 1\}$, and we define the new factor as follows:

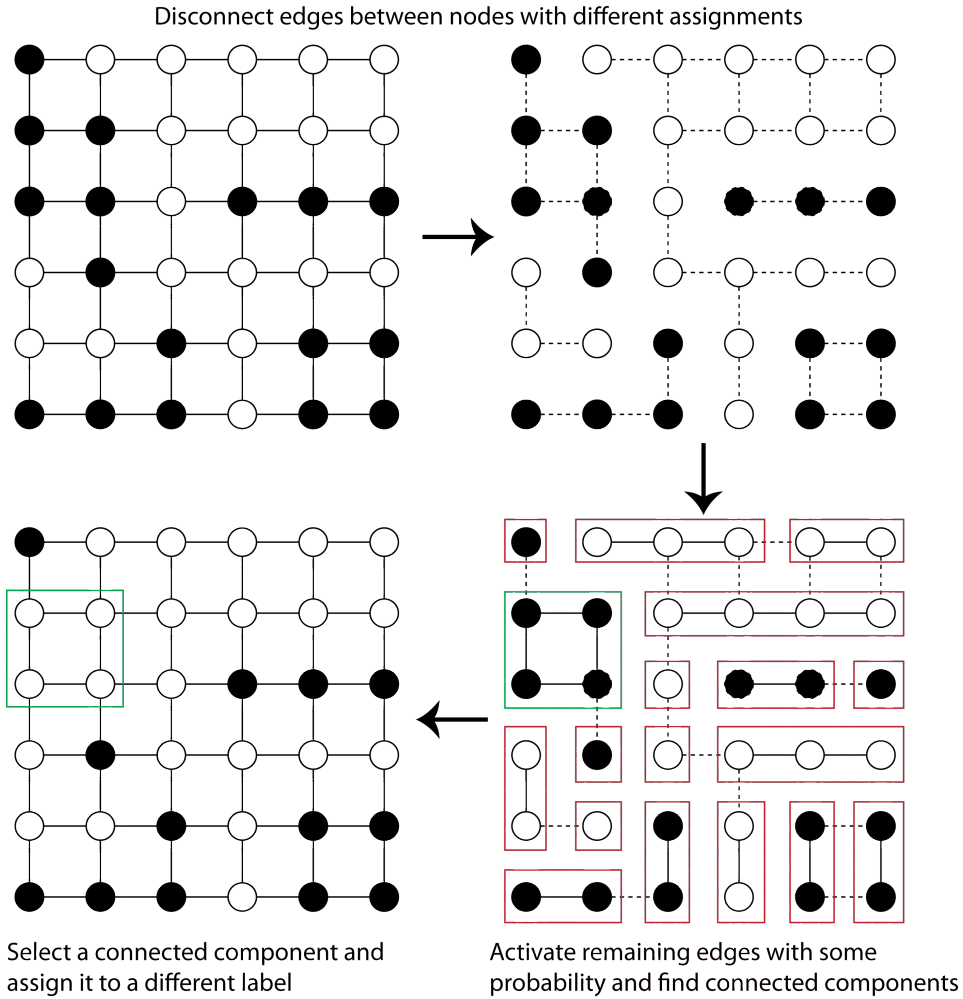
$g_e(\mathbf{x}_e, z_e = 0) = \begin{pmatrix} e^{-J} & e^{-J} \\ e^{-J} & e^{-J} \end{pmatrix}$, and

$g_e(\mathbf{x}_e, z_e = 1) = \begin{pmatrix} e^J - e^{-J} & 0 \\ 0 & e^J - e^{-J} \end{pmatrix}$. It is clear that $\sum_{z_e=0}^{1} g_e(\mathbf{x}_e, z_e) = f_e(\mathbf{x}_e)$, and hence that $\sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x})$.

Swendsen Wang

- $p(\mathbf{z}|\mathbf{x})$ factorizes
- To compute $p(z_e|\mathbf{x}_e)$: if the nodes on either end of the edge are in the same state ($x_i = x_j$), we set the bond z_e to 1 with probability $p = 1 - e^{-2J}$, otherwise we set it to 0.
- To sample $p(\mathbf{x}|\mathbf{z})$: note that all nodes which are connected by a set of bonds must have the same state, since the off-diagonal terms in the $g_e(\mathbf{x}_e, z_e = 1)$ factor are 0.
- Pick a connected component at random and force all variables in this component to adopt the same state (chosen at random).

Swendsen Wang



Other auxiliary variable methods

- Slice sampling
- Hamiltonian MCMC
- Data augmentation for logistic regression
- etc.

Quiz

When running an MCMC method, is it better to have to sample more variables or fewer variables?