today:  
- finish Gibbs sampling
- variational methods

(Cont for Gibbs sampling)

GOS is MH with time varying proposal  

Suppose we pick \( i \) at time \( t \), then proposal is  

\[
q_t(x^{(t)} | x^{(t-1)}) = p(x_i^t | x_{-i}^{(t-1)}) s(x_i^{(t)}, x_{-i}^{(t-1)})  
\]

\[
\rho(x_i^{(t+1)} | x_i^{(t)})  
\]

Acceptance ratio:

\[
\alpha(x^* | x^{(t-1)}) = \frac{q_t(x^* | x^{(t-1)}) p(x^*)}{q_t(x^{(t-1)} | x^*) p(x^{(t-1)})} \rightarrow \frac{p(x_{-i}^{(t)}) p(x_i^{(t+1)} | x_{-i}^{(t-1)})}{p(x_i^{(t)}) p(x_{-i}^{(t-1)} | x^*)}  
\]

\[
= 1 \quad \text{always accepted?}  
\]

Convergence of GOS:

Let \( A \) be transition kernel of one full cycle of GOS, (i.e. \( n \) steps)  

\[
\rightarrow \text{homogeneous M.C.}  
\]

\( A \) is irreducible and aperiodic because \( A_{ii} > 0 \) for all \( i \).
\( \Rightarrow \quad A \xrightarrow{\text{as} \to \infty} p \)

Also works for random scan. (pick \( i \in \text{Unif}\{1:n\} \) at each step)

Example: G-S, for Ising model.

Ising model \( x_i \in \{0,1\}^2 \)

UGM:

\[
p(x) = \frac{1}{Z(n)} \exp \left( \sum_{i} \eta_i x_i + \sum_{i<j} \sigma_{ij} x_i x_j \right)
\]

For Gibbs sampling,

want to compute \( p(x^*_i | x^*_j) \xrightarrow{\text{Gibbs}} p(x, x^*_i) \)

\[
= \exp \left( \eta_i x^*_i + \sum_{j \in \text{NN}(i)} \sigma_{ij} x^*_i x^*_j + \text{rest} \right)
\]

\( \Rightarrow \) renormalise to get conditional:

\[
p(x_i = 1 | x^*_i) = \frac{\exp(\eta_i + \sum_{j \in \text{NN}(i)} \sigma_{ij} x^*_i)}{1 + \exp(-\eta_i - \sum_{j \in \text{NN}(i)} \sigma_{ij} x^*_i)}
\]

\[
= \sigma \left( \eta_i + \sum_{j \in \text{NN}(i)} \sigma_{ij} x^*_j \right)
\]

\( \text{since } \sigma(x) = \frac{1}{1 + \exp(-x)} \)
Variance:

\[ x_2 \quad \text{issue:} \quad a) \quad \text{block Gibbs sampling} \]

\[ p(x_A | x_{-A}) \]

\[ \sum (\text{block}) \]

\[ UA_c = V \]

b) Rao-Blackwellized Gibbs sampling: marginalize out some variables first

- e.g. say can compute \( p(z_i | y) \) from \( p(x, y, z) \)

- (used in LDA (latent Dirichlet allocation))

Overall

\( \xi \) need to be able to sample from \( p(x_i | x_{-i}) \) to run G-S

(otherwise, need MH)

Gibbs is easy when discrete graph model

\( \xi \)s. DGM with "conjugate distributions"

* good proposal is an art:

look at software: STAN (C++)  (Andrew Gelman)

(BUGS)
diagnostic of mixing:

- monitor mixing by running independent chains

"sticky chain" 

- slow mixing

usually slow mixing comes to difficulty to move between modes

lead to cross

→ annealing methods help this

proposal looks like \( \frac{1}{Z} \exp\left(\frac{-k_B E(x)}{T}\right) \)

example: "Annealed importance sampling"

Variational methods

general idea: say we want to approximate \( Z \)
then express it as solution to optimization problem

\[ \Theta^* = \arg \min_{\Theta} \int_{\Theta} \mathcal{L}(\Theta) \, d\mathcal{OPT} \]

idea: approximate \( \Theta^* \) by approximating \( \mathcal{OPT} \)

linear algebra example: want to compute solution to \( Ax = b \) i.e. \( x = A^{-1}b \)

\[ \min_{\alpha} \|Ax - b\|^2 \]

\textbf{Variational EM:}

recall the EM trick

\[ \text{latent variable model } p(x, z) \]

\[ \log p(z|x) \geq \mathbb{E}_q \left[ \log \frac{p(z|x, z)}{q(z)} \right] = \mathcal{L}(q, \Theta) \]

\[ \log p(z|x) - \mathcal{L}(q, \Theta) = \text{KL}(q \parallel p(.|x, \Theta)) \]

\textbf{E-step:} \( \arg \max_{q \text{ all distributions}} \mathcal{L}(q, \Theta(z|x, \Theta)) = \arg \min_{q} \text{KL}(q \parallel p(z|x, \Theta)) \)

\textbf{a variational approximation for the E-step:}

\[ \text{do } q_{\text{approx}} = \arg \max_{q \text{ simple}} \text{KL}(q \parallel p(.|x)) \]

\[ \text{source of approximation } \rightarrow \text{approximate } p(z|x, \Theta(z|x, \Theta)) \]
approximate M-step: arg max \[ \mathcal{E}_{q \in \mathcal{Q}} \mathcal{L}(q(z | x)) \]

more generally, using \( \text{arg min} \) \( KL(q \parallel p) \) is a variational approach to approximate \( p \)

note: \( I \)-projection; if \( q \) is simple, can compute Eq \( \log q \)

alternative: \( \text{arg min} \) \( KL(p \parallel q) \) \( M \)-projection

"motivation for" EP algorithm \( \rightarrow \) moment matching expectation propagation

\( M \)-projection approximation

(see Figure 10.2 in Bishop)

Mean-field approximation:

Let's suppose that \( p(z) \) is in exponential family

\[ p(z) = \exp(\mathbf{z}^T \mathbf{\theta} - A(\mathbf{\theta})) \]

mean field approximation: \[ q^{MF} = \mathbb{E}_q (z) = \mathbb{E}_z (q(z | x)) \]

(see Figure 10.2 in Bishop)

(See section 10.1 in Bishop)
set of fully factorized distributions

\[ KL(q \parallel p) = E_q \left[ \log \frac{q(z)}{q_i(z)} \right] \]

\[ = -\beta \sum_i E_{q_i} \left[ T(z_i) \right] + \text{const.} + \sum_i E_{q_i} \log q_i(z_i) \]

coordinate descent on \( q_i \) : 

Fix \( q_j \) for \( j \neq i \)

minimize with respect to \( q_i \)

\[ KL(q_i; q_{\cdot i} \parallel p) = -\beta \sum_i E_{q_i} \left[ T(z_i) \right] + \text{const.} + \sum_i q_i(z_i) \log q_i(z_i) \]

add Lagrange multiplier for \( \sum_i q_i(z_i) = 1 \)

\[ \frac{\partial}{\partial q_i(z_i)} = 0 \Rightarrow -\beta i(z_i) + \log q_i(z_i) + 1 - 1 = 0 \]

\[ q_i^*(z_i) \propto \exp \left( \beta i(z_i) \right) \]

general mean field update when target \( p \) is in exp family

\[ q_i^{(t+1)}(z_i) \propto \exp \left( \beta \sum_{j \neq i} T_j(z_i) \right) \]
Ising model example:

\[
T(z) = \frac{(z_i)_{i \in V}}{(z_i z_j)_{i,j \in E}}
\]

\[
\text{Eq.1} \; (z_i) = a_i \; (z_i = 1) \triangleq \mu_i
\]

\[
\text{Eq.2} \; [z_i z_j] = z_i \mu_j \mu_{ij}^{(4)}
\]

\[
\eta^T \text{Eq.1} \; T(z) = \eta_i z_i + \sum_{j \neq i} \eta_j \mu_{ij}^{(4)} [z_i z_j] + \text{const}
\]

\[
= \sum_{j \in \eta(i)} \eta_j \mu_{ij}^{(4)} [z_i z_j] + \text{const}
\]

Result:

\[
q_i (z_i) \triangleq \exp \left( \eta_i z_i + \sum_{j \in \eta(i)} \mu_{ij}^{(4)} [z_i z_j] \right)
\]

\[
\mu_i^{(4i)} = \sigma \left( \eta_i + \sum_{j \in \eta(i)} \mu_{ij}^{(4)} [z_i z_j] \right)
\]

ML update for \( q_i (z_i) \) [with parameter \( \mu_i \)]

Compare with G-S update where \( z_i^{(4i)} = 1 \) with prob \( \sigma \left( \eta_i + \sum_{j \in \eta(i)} \mu_{ij}^{(4)} [z_i z_j] \right) \)

* here min KL(q \| p) \quad KL(q \| p) \text{ is a convex function of } \eta

* but \( q_{\text{ML}} \) is a non-convex constraint set
can get stuck in local minima
KL(q(x) || p(x))

pros & cons of variational methods (optimization based, often faster, often easier to debug)

Sampling
- noisy & hard to debug
- mixing problem for chains
- unbiased estimates

$\mathbb{E}_{q(x)}[f(z)] \neq \mathbb{E}_p[f(z)]$

$\mathbb{E}_{q(x)}[f(z)] = \mathbb{E}_p[f(z)]$

with respect to random sample