today: Gibbs sampling
variational methods

Gibbs sampling alg.

Let $M.H.$ with a clever choice of proposal $q_{\theta}(z'|z)$. 

Example of applications

$$\text{UGM: } \hat{p}(x) = \prod_i \gamma_i(x_i)$$

difficult conditional in $DGM$

$$\hat{p}(x) = \frac{p(x \mid x_E)}{S(x_E, x_E)} \propto p(x \mid x_E)$$

UGM:

cyclic Gibbs sampling alg. nodes $i=1, \ldots, n$

Start at some $x^{(0)}$

For $t=1, \ldots, \ldots$

- Pick $i = (t \mod n) + 1$
- Sample $x^{(t)}_i \sim p(x_i = \cdot \mid x_{-i}^{(t-1)} = x_{-i}^{(t-1)})$

end

- Set $x^{(t)}_j = x^{(t-1)}_j$ for $j \neq i$

Gibbs sampling is M.H. with a time varying proposal

Suppose we pick $i$ at time $t$

Then proposal is $q_{\theta}(z' \mid z^{(t-1)}) = p(x'_i \mid x_i^{(t-1)}) S(x'_E, x_i^{(t-1)})$

Acceptance ratio:

\[ q(z' \mid z^{(t-1)}) = q_{\theta}(z' \mid z^{(t-1)}), \frac{p(z_i^{(t-1)} \mid z^{(t-1)})}{p(z^{(t-1)} \mid z_i^{(t-1)}} \cdot \frac{p(x_E^{(t-1)})}{p(x_E^{(t-1)} \mid x_i^{(t-1)})} \cdot \frac{p(x_i^{(t-1)} \mid x_E^{(t-1)})}{p(x_i^{(t-1)} \mid x_E^{(t-1)})} \cdot \frac{S(x_i^{(t-1)}, x_i^{(t-1)})}{S(x_i^{(t-1)}, x_i^{(t-1)})} \]
Convergence of G-S:

Let $A$ be a Markov transition kernel of one full cycle of G-S (i.e., n steps)

- A homogeneous M.C.

If suppose that $p(z_i) \geq 0 \forall z_i \Rightarrow A$ is irreducible \& aperiodic because $A \mathbb{1} \geq 0$

(since can get to any state with n steps)

$\Rightarrow A \xrightarrow{t \to \infty} \rho$

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

**Example:** G-S for Ising model

Ising model $x_i \in \{0, 1\}$

$$p(z) = \frac{1}{Z(n)} \exp \left( \sum_{i \neq j} \frac{1}{2} J_{ij} x_i x_j \right)$$

UGM:

![Graphical Model]

- Minimal exp family representation

for G-S, want to compute $p(x_i \mid x_{\text{rest}}) = p(x_i \mid \text{neighbors}(i))$

- $p(x_i, x_{\text{rest}})$

- $= \exp \left( m_i x_i + \sum_{j \in \text{rest}} \frac{1}{2} J_{ij} x_i x_j \right)$

- $= \frac{1}{1 + \exp \left( m_i \right)} \exp \left( J_{\text{rest}} \right)$

- $= \sigma \left( m_i + \sum_{j \in \text{rest}} J_{ij} x_j \right)$

**Diagnosis of mixing**

monitor mixing by running independent chains
monitor mixing by running independent chains

usually slow mixing comes from difficulty to move between modes

→ annealing methods help this: proposed looks like \( \frac{1}{Z} \exp \left( -\frac{E(z)}{T} \right) \)

example: "Annealed importance sampling"

Variational methods

general idea: say we want to approximate \( \theta^* \)

then, express it as solution to optimization problem

\( \theta^* = \arg \min_{\theta} f(\theta) \) \[ \text{OPT} \]

idea: approximate \( \theta^* \) via an approximation to \( \text{OPT} \)

linear algebra example
say want solve \( Ax = b \)  \( \text{ie} \ x^* = A^{-1}b \)

\( \arg \min_{x} \|Ax - b\|^2 \)

Variational EM (mutivariate an objective)

recall EM trick: latent variable \( p(z|x) \)

\[ \log p(z|x) \geq \mathbb{E}_{q(z|x)} [ \log p(x|z) ] = \mathbb{E}_{q(z|x)} [q(z|x) \log \frac{p(x,z)}{q(z|x)}] \]
\[ \log p(z|x) \geq \mathbb{E}_q [ \log p(z|x, z) ] \geq \mathbb{E}_q [ S(q; p(x)) ] \]

\[ \log p(z|x) - S(q; \theta) = KL(q(\theta) \| p(\cdot | x, \theta)) \]

E step: argmax \(_{q \in \text{all distributions}} S(q; \theta) \Rightarrow \text{argmin } KL(q(\theta) \| p(z|x, \theta^{(t)})) \]

\( \Rightarrow \) a variational approach for the E-step

\[ q^{(t+1)} = \argmin_{q \in \text{simple}} \text{KL}(q \| p(z|x, \theta^{(t)})) \]

approximate M step: \( \theta^{(t+1)} = \argmax_{\theta \in \Theta} \mathbb{E}_{q^{(t+1)}} [ \log p(z|x, \theta) ] \]

\[ \mathbb{E}_{q^{(t+1)}} [ \log p(z|x, \theta^{(t)}) ] = \mathbb{E}_{q^{(t)}} [ \log p(z|x, \theta^{(t)}) ][\text{stabilized bound on}] \]

\[ \mathbb{E}_{q^{(t+1)}} [ \log p(z|x, \theta^{(t)}) ] \]

more generally, using \( \text{argmin } KL(q \| p) \) is a variational approach to approximate \( p \)

note: I-projection; if \( q \) is simple, \( \mathbb{E}_q [ \log p(x) ] \) can compute

alternative: \( \text{argmin } KL(p \| q) \) M-projection

M "mixture EP algorithm" trying to do moment matching

(see sec 10.2 in Bishop)
Mean-field approximation (section 10.1 in Bishop)

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

\[ p(z_1, \ldots, z_p) = \exp(\mathbf{n}^T T(z) - A(n)) \]

Mean-field approximation: \( q_{MF} = \{ q(z) = \prod_i q_i(z_i) \} \)

Set of fully-factorized dist.

\[
KL(q \| p) = \mathbb{E}_p \left[ \log q(z) \right] - \mathbb{E}_p \left[ T(z) \right] + A(n) + \sum_i z_i \log q_i(z_i)
\]

\[
= -\mathbb{E}_q \left[ \mathbf{n}^T \mathbb{E}_q [T(z)] + A(n) \right] + \sum_i z_i \log q_i(z_i)
\]

\[
\leq \mathbb{E}_q \left[ \frac{\mathbf{n}^T \mathbb{E}_q [T(z)]}{z_i} \right] \log q_i(z_i)
\]

Coordinate descent on \( q_i \)’s

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

Minimize w.r.t. \( q_i \): \( KL(q_i || p) \)

\[
= -\mathbb{E}_q \left[ \frac{\mathbf{n}^T \mathbb{E}_q [T(z)]}{z_i} \right] + \text{const.} + z_i \log q_i(z_i)
\]

Add Lagrange multiplier

\[
\frac{\partial}{\partial q_i(z_i)} + 1 - \lambda \cdot z_i = 0 \quad \Rightarrow \quad \lambda = \frac{1}{z_i}
\]

\[
q_i(z_i) \propto \exp \left[ \frac{\mathbf{n}^T \mathbb{E}_q [T(z_i)]}{z_i} \right]
\]

General mean-field update when \( p \) is in exp-family:

\[
q_i(z_i) \propto \exp \left( \mathbf{n}^T \mathbb{E}_q [T(z)] \right)
\]

Using model example:

\[ T(z) = \begin{cases} \mathbf{z}^T \mathbf{e} & \mathbf{z} \in \mathcal{E}_0, \mathbf{z} \\ \mathbf{z}^T \mathbf{e} & \mathbf{z} \in \mathcal{E}_1, \mathbf{z} \\ \end{cases} \]
\[ q_i^t(z_i) = q_j^t(z_j=1) \alpha \mu_5 \]
\[ q_i^t \delta z_j = z_i \mu_j \]
\[ \eta T \delta z_j = \eta \delta z_j + \sum \frac{z_i \mu_j}{\sum z_i \mu_j} \]
\[ \eta \delta z_j \]
\[ \text{result: } q_i^{t+1}(z_i) \text{ or } \exp(\eta \delta z_i \mu_j^{(t+1)}) \]
\[ M_i^{(t+1)} = \sigma^{-1} m_i + \sum_{j \neq i} \mu_j \mu_j^{(t+1)} \]

*Compare with*

\[ Z_i^{t+1} = 1 \text{ with prob } \sigma(\eta \delta z_i \mu_j^{(t+1)}) \]