

today: Gibbs sampling
variational methods

Gibbs sampling alg:

↳ M.H. with a clever choice of proposal $q_t(x'|x)$ [clever: $a(x'|x) = 1$]

example of applications

UGM: $\tilde{p}(x) = \prod_c \gamma_c(x_c)$
 difficult conditional in DGM
 $\tilde{p}(x) = p(x_E | \bar{x}_E) \delta(x_E, \bar{x}_E) \propto p(x | \bar{x}_E)$

UGM:

Cyclic Gibbs sampling alg. nodes $i=1, \dots, n$

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

for $t=1, \dots, \dots$

- pick $i = (t \bmod n) + 1$
- sample $x_i^{(t)} \sim p(x_i = \cdot | x_{\setminus i}^{(t-1)})$ ↳ $i \in \{1, \dots, n\}$
- set $x_j^{(t)} = x_j^{(t-1)}$ for $j \neq i$ true conditional on x_i as proposal

end

⊛ Gibbs Sampling is M.H. with a time varying proposal

suppose we pick i at time t

then proposal is $q_t(x' | x^{(t-1)}) = p(x_i' | x_{\setminus i}^{(t-1)}) \delta(x_{\setminus i}', x_{\setminus i}^{(t-1)})$ for rest to be constant

acceptance ratio:

$$a(x' | x^{(t-1)}) = \frac{q_t(x^{(t-1)} | x') p(x')}{q_t(x' | x^{(t-1)}) p(x^{(t-1)})}$$

$$= \frac{p(x_i^{(t-1)} | x_{\setminus i}^{(t-1)}) \delta(x_{\setminus i}^{(t-1)}, x_{\setminus i}^{(t-1)}) p(x')}{p(x_i' | x_{\setminus i}^{(t-1)}) \delta(x_{\setminus i}', x_{\setminus i}^{(t-1)}) p(x^{(t-1)})}$$

$\rightarrow p(x_i^{(t-1)}) p(x_{\setminus i}^{(t-1)})$
 $\rightarrow p(x_i^{(t-1)}) p(x_{\setminus i}^{(t-1)} | x_i^{(t-1)})$

$= 1$ \rightarrow always accepts?

Convergence of G.S.:

let A be Markov transition kernel of one full cycle of G.S. (i.e. n steps)

\leadsto homog. M.C.

if suppose that $\underline{p(x) > 0 \forall x} \Rightarrow A$ is irreducible & aperiodic because $A_{ij} > 0 \forall i, j$
 (since can get to any state with n steps)

$$\Downarrow$$

$$p(x_i | x_{-i}) > 0 \forall x_i, x_{-i}$$

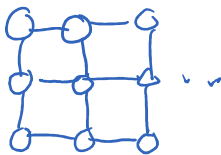
$$\Rightarrow A \xrightarrow[t \rightarrow \infty]{} p$$

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

example: G.S. for Ising model

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

UGM:



$$p(x) = \frac{1}{Z(n)} \exp\left(\sum_i m_i x_i + \sum_{\langle i, j \rangle \in E} m_{ij} x_i x_j\right)$$

[minimal exp family representation]

for G.S.,

want to compute $p(x_i | x_{-i}) \stackrel{\text{by cond. indep}}{=} p(x_i | x_{\text{neighbors}(i)})$

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

$$= \exp\left(m_i x_i + \sum_{j \in N(i)} m_{ij} x_i x_j + \text{rest}\right)$$

\Rightarrow renormalize to get conditional

$$p(x_i = 1 | x_{-i}) = \frac{\exp\left(m_i + \sum_{j \in N(i)} m_{ij} x_j\right) \exp(\text{rest})}{\left(1 + \exp\left(\sum_{j \in N(i)} m_{ij} x_j\right)\right) \exp(\text{rest})}$$

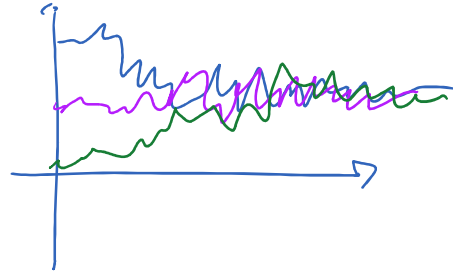
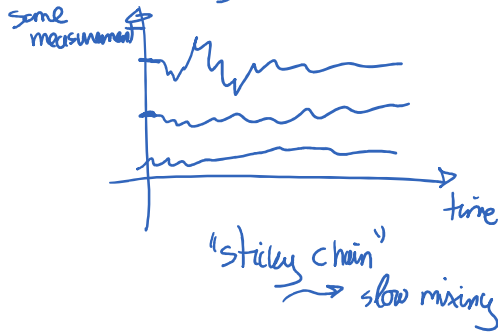
$$= \sigma\left(m_i + \sum_{j \in N(i)} m_{ij} x_j\right)$$

diagnostic of mixing

monitor mixing by running independent chains

Monte Carlo

monitor mixing by running independent chains



usually slow mixing comes from difficulty to move between modes



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

example: "Annealed importance sampling"

"temperature"
↑
high temperature
→ more exploring

15h33

Variational methods

general idea: say we want to approximate θ^*

then, express it as solution to optimization problem

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} f(\theta) \quad] \text{ OPT}$$

idea: approximate θ^* via an approximation to OPT

linear algebra example

say want sol'n $Ax=b$ (ie $x^* = A^{-1}b$)

$$\underset{x}{\operatorname{argmin}} \|Ax - b\|^2$$

Variational EM (motivation and objective)

recall EM trick:

latent variable $p(z, \underline{z})$
↓
underscored

$$\log p(x|\theta) \geq \mathbb{E}_q \left[\log p(x, z|\theta) \right] \triangleq \mathcal{L}(q, \theta)$$

$$\log p(z|\theta) \geq \mathbb{E}_q \left[\log \frac{p(z, z|\theta)}{q(z)} \right] \triangleq \mathcal{J}(q, \theta)$$

$$\log p(z|\theta) - \mathcal{J}(q, \theta) = KL(q(\cdot) \| p(\cdot|z, \theta))$$

$$E \text{ step: } \underset{q \in \text{all distributions on } Z}{\text{argmax}} \mathcal{J}(q, \theta^{(t)}) \Leftrightarrow \underset{q}{\text{argmin}} KL(q \| p(z|z, \theta^{(t)}))$$

⊗ a variational approx. for the E-step

$$\text{do } q_{\text{approx}}^{(t+1)} = \underset{q \in \text{simple}}{\text{argmin}} KL(q \| p(\cdot|z, \theta^{(t)}))$$

↑ source of approximation → to approximate $p(z|z, \theta^{(t)})$

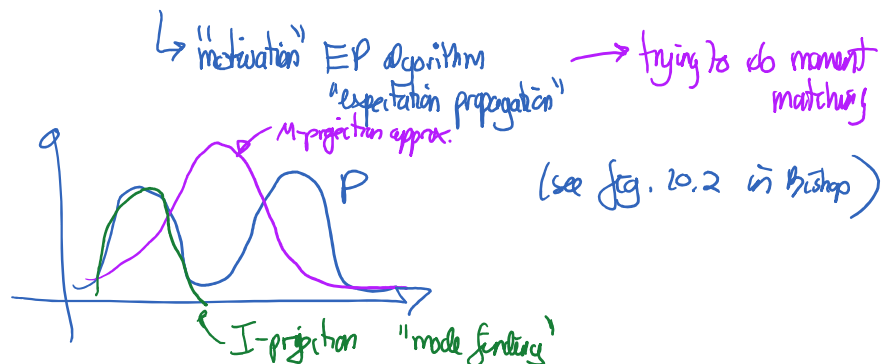
$$\text{approximate M step: } \theta^{(t+1)} = \underset{\theta \in \Theta}{\text{argmax}} \underbrace{\mathbb{E}_{q_{\text{approx}}^{(t+1)}} [\log p(z, z|\theta)]}_{\mathcal{J}(q_{\text{approx}}^{(t+1)}, \theta)}$$

(still a lower bound on $\log p(z|\theta^{(t+1)})$)
 but loose monotonicity guarantees on $\log p(z|\theta^{(t)})$ as θ of θ

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

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

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



Mean-field approximation (section 10.1 in Bishop)

let's suppose that $p(z)$ is in exp.-family

$$z_1, \dots, z_p \quad p(z) = \exp(n^T T(z) - A(n))$$

mean field approximation: $Q_{MF} = \{q(z) = \prod_i q_i(z_i)\}$

set of fully factorized dist.

$$KL(q \| p) = \mathbb{E}_q \left[\log \frac{q(z)}{p(z)} \right]$$

$$\begin{aligned} &= -n^T \mathbb{E}_q [T(z)] + A(n) + \sum_z q(z) \log q(z) \\ &= \sum_z \left(\prod_j q_j(z_j) \right) \left(\sum_j \log q_j(z_j) \right) \\ &= \sum_i \sum_{z_i} q_i(z_i) \log q_i(z_i) \end{aligned}$$

coordinate descent on q_i 's

fix q_j for $j \neq i$

minimize w.r. to q_i $KL(q_i \cdot q_{-i} \| p)$

$$= -\mathbb{E}_{q_i} \left[\underbrace{n^T \mathbb{E}_{q_{-i}}(T(z))}_{\triangleq f_i(z_i)} \right] + \text{cst.} + \sum_{z_i} q_i(z_i) \log q_i(z_i)$$

add Lagrange multiplier

$$\text{for } \sum_{z_i} q_i(z_i) = 1 \quad \rightarrow +\lambda \left(1 - \sum_{z_i} q_i(z_i) \right)$$

$$\frac{\partial}{\partial q_i(z_i)} = 0 \quad -f_i(z_i) + \log q_i(z_i) + 1 - \lambda = 0$$
$$q_i^*(z_i) \propto \exp(f_i(z_i))$$

⊗ general mean-field update when target p is in exp family:

$$q_i^{(t+1)}(z_i) \propto \exp(n^T \mathbb{E}_{q_{-i}^{(t)}} T(z))$$

Ising model example:

$$T(z) = \begin{pmatrix} (z_i)_{i \in V} \\ (z_i z_j)_{i, j \in E} \\ \dots \end{pmatrix} \quad z_i \in \{0, 1\}$$

$$\mathbb{E}q_i(z_j) = q_j'(z_j=1) \triangleq \mu_j$$

$$\mathbb{E}q_i(z_j) = z_j \mu_j$$

$$n^T \mathbb{E}q_i^{(t)}(z) = n_i z_i + \sum_{j \neq i} n_j \underbrace{\mathbb{E}q_i^{(t)}(z_j)}_{\mu_j^{(t)}} + \sum_{j \in N(i)} n_j \underbrace{\mathbb{E}q_i^{(t)}(z_j)}_{z_j \mu_j^{(t)}} + \text{rest (no } z_i)$$

$$\text{result: } q_i^{(t+1)}(z_i) \propto \exp\left(n_i z_i + z_i \sum_{j \in N(i)} n_j \mu_j^{(t)}\right)$$

$$\mu_i^{(t+1)} = \sigma\left(n_i + \sum_{j \in N(i)} n_j \mu_j^{(t)}\right)$$

$\mu_i^{(t+1)}$ update for $q_i(z_i)$
[with parameter μ_i]

Compare with

$$\text{G.S. update where } z_i^{(t+1)} = 1 \text{ with prob } \sigma\left(n_i + \sum_{j \in N(i)} n_j z_j^{(t)}\right)$$