Lecture 21 - Gibbs sampling; variational methods

Tuesday, November 21, 2023 2:30 PM

today: • Gibbs sampling
• variational methods

Gibbs sampling alg.

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

examples of applications

UGM; $\tilde{p}(x) = \prod_C \psi_c(x_c)$

difficult conditional in DGM $\tilde{p}(x) = p(x_{E^c}, \bar{x}_E)\,\delta(x_E, \bar{x}_E)$

or $p(x|\bar{x}_E)$

(UGM)

cyclic Gibbs sampling alg: nodes $i = 1, \dots, n$

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

for $t = 1, \dots,$

• pick $i = (t \bmod n) + 1$

• sample $x_i^{(t)} \sim \underbrace{p(X_i = \cdot \mid X_{\neg i} = x_{\neg i}^{(t-1)})}_{\text{true conditional on } x_i \text{ as a proposal}}$ $\quad \neg i \to 1{:}n \setminus \{i\}$

• set $x_j^{(t)} = x_j^{(t-1)}$ for $j \neq i$

end

(*) Gibbs sampling is M.H. with a time varying proposal

suppose we pick i at time t

then the proposal is $q_t(x'|x^{(t-1)}) = p(x_i'|x_{\neg i}^{(t-1)})\,\underbrace{\delta(x'_{\neg i}, x_{\neg i}^{(t-1)})}_{\text{for } x_{\neg i} \text{ to stay constant}}$

M.H. acceptance ratio:

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

$q_t(x^{(t-1)}|x') = p(x_i^{(t-1)}|\cancel{x'_{\neg i}})\,\delta(x'_{\neg i}, x_{\neg i}^{(t-1)})$

$p(x') = p(\cancel{x'_{\neg i}})\,p(x'_i|\cancel{x'_{\neg i}})$

$p(x^{(t-1)}) = p(\cancel{x_{\neg i}^{(t-1)}})\,p(x_i^{(t-1)}|\cancel{x_{\neg i}^{(t-1)}})$

$$\underbrace{q_t(x'|x^{(t-1)})} \, p(\underbrace{x^{(t-1)}}) \longrightarrow p(\cancel{x_{\neg i}^{(t-1)}})\, p(x_i^{(t-1)} | \cancel{x_{\neg i}^{(t-1)}}) \quad //$$

$$\hookrightarrow p(x_i' | \cancel{x_{\neg i}^{(t-1)}})\, \delta(x'_{\neg i}, x_{\neg i}^{(t-1)})$$

always accepts!

Convergence of G.S.:

- let A be a Markov transition kernel of one full cycle of Gibbs sampling (i.e. n steps)

$\rightarrow$ homogeneous M.C.

if suppose that $p(x) > 0 \ \forall x$ $\Rightarrow A$ is irreducible & aperiodic because $A_{ii} > 0$

$\Rightarrow p(x_i | x_{\neg i}) > 0 \quad \forall x_i \,\&\, x_{\neg i}$

$A_{kl} > 0 \ \forall k, l$

since can get to any state with n "flips"

$\Rightarrow A^t \pi_0 \xrightarrow{t \to \infty} p$

$\hookrightarrow$ by detailed balance, p is stationary dist. of chain

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

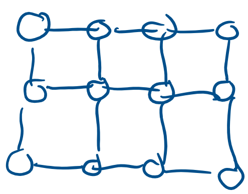
example: G.S. for Ising model

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

$$p(x) = \frac{1}{Z(\eta)} \exp\Big(\sum_i \eta_i x_i + \sum_{\{i,j\} \in E} \eta_{ij} x_i x_j\Big)$$

[minimal exp. family representation]

UGM:



for G.S.

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

$$\propto \exp\Big(\eta_i x_i + \sum_{j \in N(i)} \eta_{ij} x_i x_j + \text{rest}\Big)$$

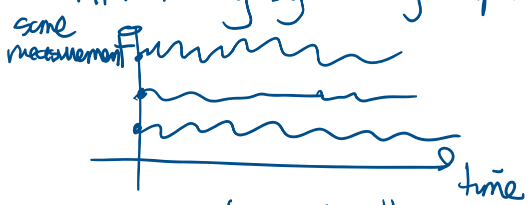
$\Rightarrow$ renormalize to get conditional

$$p(x_i = 1 | x_{\neg i}) = \frac{\exp(\eta_i + \sum_{j \in N(i)} \eta_{ij} x_j)\, \cancel{\exp(\text{rest})}}{\big(1 + \exp(\eta_i + \sum_{j \in N(i)} \eta_{ij} x_j)\big)\, \cancel{\exp(\text{rest})}}$$

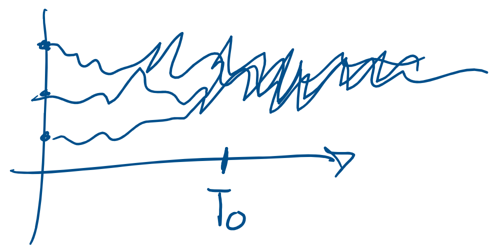
$$\boxed{p(x_i^{(t)}=1 \mid x_{\neg i}^{(t-1)}) = \sigma\Big(\eta_i + \sum_{j\in M(i)} \eta_{ij}\, x_j^{(t-1)}\Big)}$$

diagnostic of mixing

monitor mixing by running independent chains



"sticky chain" → slow mixing



15h16

back at 15h40

variational methods

general idea: say we want to approximate θ^*

then, express it as a solution to opt. problem

$$\theta^* = \arg\min_{\theta\in\Theta} f(\theta) \quad \Big] \text{ OPT}$$

$\boxed{\text{idea: approximate } \theta^* \text{ via an approximation of OPT}}$

linear algebra example:

say want sol'n to $Ax=b$ (i.e. $x^*=A^{-1}b$)

$$x^* = \arg\min_x \|Ax-b\|^2$$

variational EM (motivation for objective)

recall EM trick   latent $p(x,\underset{\text{unobserved}}{z}\mid\theta)$

$$\log p(x\mid\theta) \geq \mathbb{E}_q\Big[\log \frac{p(x,z\mid\theta)}{q(z)}\Big] \triangleq \mathcal{L}(q,\theta)$$

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

$$\log p(x|\theta) - \mathcal{L}(q,\theta) = KL\left(q(\cdot) \,\|\, p(z|x,\theta)\right)$$

E step: $\underset{q \in \text{ all distributions on } z}{\operatorname{argmax}} \mathcal{L}(q,\theta^{(t)}) \Leftrightarrow \underset{q}{\operatorname{argmin}} KL\left(q(\cdot) \,\|\, p(\cdot|x,\theta^{(t)})\right)$

(*) a variational approx. for E step:

do $q^{(t)}_{approx} = \underset{q \in \mathcal{Q}_{simple}}{\operatorname{argmin}} KL\left(q \,\|\, p(\cdot|x,\theta^{(t)})\right)$

source of approx $\rightarrow$ to approximate $p(z|x,\theta^{(t)})$

approximate M step:

$$\theta^{(t+1)} = \underset{\theta \in \Theta}{\operatorname{argmax}} \underbrace{\mathbb{E}_{q^{(t+1)}_{approx}}\left[\log p(x,z|\theta)\right]}_{\mathcal{L}(q^{(t+1)}_{approx},\theta)}$$

still a lower bound on $\log p(x|\theta)$

but lose monoticity guarantee on $\log p(x|\theta^{(t)})$ as fct. of t

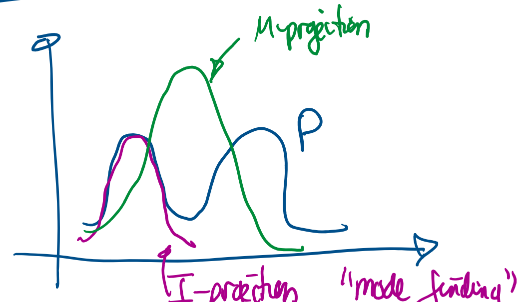
---

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

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

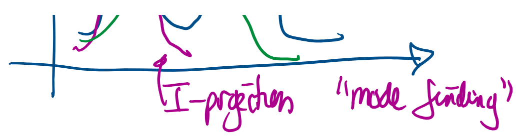
alternative: $\underset{q \in Q}{\operatorname{argmin}} KL(p\|q)$ M-projection $\rightarrow$ "moment matching"

"motivation" for EP algorithm "expectation propagation"



(see fig. 10.2 in Bishop)

expectation propagation



(see fig. 10.2 in Bishop)

Mean-field approximation (section 10.1 in Bishop)

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

$z_1, \dots, z_d \qquad p(z) = \exp(\eta^T T(z) - A(\eta))$

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

set of fully factorized dist.

$$KL(q \| p) = \mathbb{E}_q[\log q/p]$$
$$= -\eta^T \mathbb{E}_q[T(z)] + A(\eta) + \sum_z \underbrace{q(z)}_{\left(\prod_j q_j(z_j)\right)} \underbrace{\log q(z)}_{\sum_j \log q_j(z_j)}$$
$$\sum_i \underbrace{\sum_z q_{\neg i}(z_{\neg i}) q_i(z_i) \log q_i(z_i)}_{1 \cdot \sum_{z_i} q_i(z_i)}$$
$$\sum_i \sum_{z_i} q_i(z_i) \log q_i(z_i)$$

coordinate descent on q_i's

fix q_j for $j \neq i$; $\quad \min_{\text{w.r.to } q_i} KL(q_i \cdot q_{\neg i} \| p)$

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

(like MaxEnt) add lagrange multiplier for $\sum_{z_i} q_i(z_i) = 1 \rightarrow \lambda(1 - \sum_{z_i} q_i(z_i))$

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

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

$$\boxed{q_i^{(t+1)}(z_i) \propto \exp\left(\eta^T \mathbb{E}_{q_{\neg i}^{(t)}}[T(z)]\right)}$$

Ising model

$T(z) \Rightarrow \begin{pmatrix} z_i \\ (z_i z_j)_{\{i,j\} \in E} \end{pmatrix}$ $\qquad z_i \in \{0,1\}$

$\mathbb{E}_{q_{\neg i}^{(t)}}(z_j) = q_j^{(t)}(z_j = 1) \triangleq \mu_j^{(t)}$

$\mathbb{E}_{q_{\neg i}^{(t)}}(z_i z_j) = z_i \mu_j^{(t)}$

no z_i $\Rightarrow$ constant

$$\eta^T \mathbb{E}_{q_{\neg i}^{(t)}}[T(z)] = \eta_i z_i + \boxed{\sum_{j \neq i} \eta_j \underbrace{\mathbb{E}_{q_{\neg i}^{(t)}}[z_j]}_{\mu_j^{(t)}}} + \sum_{j \in N(i)} \eta_{ij} \underbrace{\mathbb{E}_{q_{\neg i}^{(t)}}[z_i z_j]}_{z_i \mu_j^{(t)}} + \underbrace{\text{rest}}_{\text{no } z_i}$$

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

parameter for $q_j^{(t)}$

$$\boxed{\mu_i^{(t+1)} = \sigma\left(\eta_i + \sum_{j \in N(i)} \eta_{ij} \mu_j^{(t)}\right)}$$

MF update for $q_i(z_i)$
[with parameter μ_i]

compare with G.S. update

where $z_i^{(t)} = 1$ with prob. $\sigma\left(\eta_i + \sum_{j \in N(i)} \eta_{ij} z_j^{(t)}\right)$