Lecture 21 - Gibbs sampling; variational methods

Tuesday, November 22, 2022 2:58 PM

today: · Gibbs sampling
· variational methods

<u>Gibbs sampling alg.</u>

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

example of applications:

- UGM: $\tilde{p}(x) = \prod_c \psi_c(x_c)$
- difficult conditional in DGM: $\tilde{p}(x) = p(x_{E^c}, \bar{x}_E)\, \delta(x_E, \bar{x}_E) \propto p(x|\bar{x}_E)$

(UGM)

<u>cyclic Gibbs sampling alg.</u>: nodes $i=1,\dots,n$

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

for $t=1,\dots,$
- pick $i = (t \bmod n)+1$
- sample $x_i^{(t)} \sim \underbrace{p(X_i = \cdot \mid X_{\neg i} = x_{\neg i}^{(t-1)})}_{\text{true conditional on } X_i \text{ as proposal}}$, $\neg i \to 1{:}n \setminus \{i\}$
- 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:

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

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

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

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

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

$= 1$ → always accepts!

$\hookrightarrow p(x_i^* \cancel{| x_{\neg i}^{(t-1)}}) \, \delta(x_{\neg i}, x_{\neg i}^{(t-1)})$

always accepts?

convergence of G.S.:

- let A be a Markov transition kernel of $\underline{\underline{\text{one}}}$ full cycle of Gibbs Sampling (i.e. n steps)

$\rightarrow$ homogeneous M.C.

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

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

$A_{ij} > 0 \;\; \forall i \,\&\, j$

since can get to any state with n "flips"

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

by detailed balance, p is stationary dist. of chain...

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

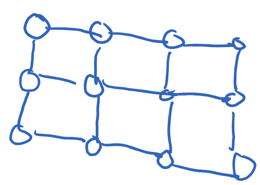
example: G.-S. for Ising model

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:

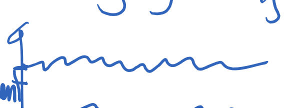
$$p(x_i^{(t)} = 1 | x_{\neg i}^{(t-1)}) = \frac{\exp\Big(\eta_i + \sum_{j \in N(i)} \eta_{ij} x_j^{(t-1)}\Big) \cancel{\exp(\text{rest})}}{\Big(1 + \exp\big(\eta_i + \sum_{j \in N(i)} \eta_{ij} x_j^{(t-1)}\big)\Big) \cancel{\exp(\text{rest})}}$$

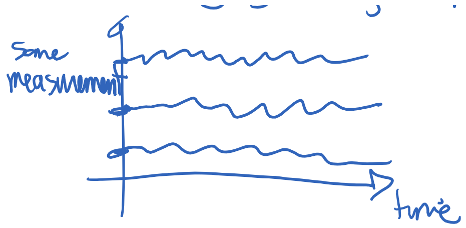
$$\boxed{p(x_i^{(t)} = 1 | x_{\neg i}^{(t-1)}) = \sigma\Big(\eta_i + \sum_{j \in N(i)} \eta_{ij} x_j^{(t-1)}\Big)}$$

diagnostic of mixing

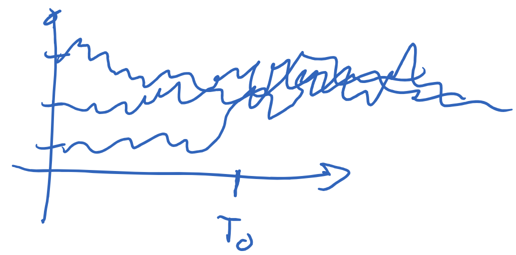
monitor mixing by running independent chains

some measurement





"sticky chain" → slow mixing



16/10/2

variational methods

general idea: say we want to approximate θ^*

then, express it as solution to opt. problem

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

Idea: approximate θ^* via an approximation to OPT

linear algebra example:

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

$$x^* = \underset{x}{\operatorname{arg\,min}} \; \|Ax - b\|^2$$

variational EM (motivation for objective)

recall EM trick

latent variable $p(x, \overset{\text{unobserved}}{z} \mid \theta)$

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

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

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

⊛ a variational approx. for E step:

$$\text{do } q^{(t+1)}_{\text{approx}} = \underset{q \in \underbrace{\mathcal{Q}_{\text{simple}}}_{\text{source of approx}}}{\operatorname{argmax}} \; KL\big(q \,\|\, p(\cdot \mid x, \theta^{(t)})\big)$$

→ to approximate $p(z \mid x, \theta^{(t)})$

..

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

still a lower bound on $\log p(x|\theta)$
but lose monotonicity guarantee on $\log p(x|\theta^{(t)})$ as fct. of t

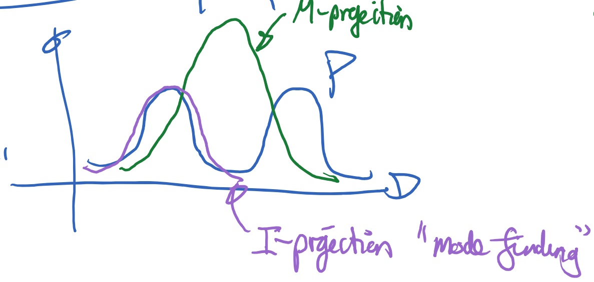
---

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, $\mathbb{E}_q [\log p/q]$ can compute

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

"motivation" for EP algorithm "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

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

mean field approximation $\quad Q_{MF} = \{ q(z) = \prod_i q_i(z_i) \}$
set of fully factorized dist.

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

$$= -\eta^T \mathbb{E}_q [T(z)] + A(\eta) + \sum_z q(z) \log q(z)$$

$$\underbrace{\left(\prod_j q_j(z_j)\right) \sum_j \log q_j(z_j)}$$

$$\sum_i \underbrace{\sum_{z} q_{\neg i}(z_{\neg i})}_{1 \cdot \sum_{z_i}} q_i(z_i) \log q_i(z_i) = \sum_i \sum_{z_i} q_i(z_i) \log q_i(z_i)$$

Coordinate descent on q_i's

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

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

(like MaxENT) add Lagrange multiplier for $\sum_{z_i} q_i(z_i) = 1 \rightsquigarrow \lambda(1 - \sum_{z_i} q_i(z_i))$

$$\partial / \partial q_i(z_i) = 0 \Rightarrow -f_i(z_i) + \log q_i(z_i) + 1 - \lambda = 0$$

$$\Rightarrow q_i^*(z_i) \propto \exp(f_i(z_i))$$

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

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

Ising model:

$T(z) \Rightarrow (z_i)_{i \in V}$ $\quad z_i \in \{0,1\}$

$(z_i z_j)_{\{i,j\} \in E}$

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

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

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

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

parameter for $q_j^{(t)}$

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

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

Compare with G.-S. update where $z_i^{(t+1)} = 1$ with prob. $\sigma(\eta_i + \sum_{j \in N(i)} \eta_{ij} z_j^{(t)})$