Lecture 21 - Gibbs sampling; variational methods

Monday, November 25, 2024 12:34 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 $\tilde{p}(x) = p(x_{E^c}, \bar{x}_E)\,\delta(x_E, \bar{x}_E) \propto p(x|\bar{x}_E)$

(UGM)

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

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

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

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

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

$p(x') = \cancel{p(x_{\neg i}')}\,p(x_i'|\cancel{x_{\neg i}'})$

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

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

$= 1$ // always accepts!

convergence of G.S.:

convergence of G.S.:

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

  → homogeneous M.C.

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

  ⟹ $p(x_i \mid x_{\neg i}) > 0 \;\; \forall x_i \;\&\; x_{\neg i}$

  ↳ $A_{kl} > 0$ $\forall k, l$ since can get to any state with n "steps"

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

  ↳ by detailed balance, p is stationary of chain

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

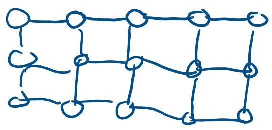
example: G.S. for Ising model

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 $p(x_i \mid x_{\neg i}) \overset{\text{by cond. indep.}}{=} p(x_i \mid x_{\text{neighbors}(i)})$

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

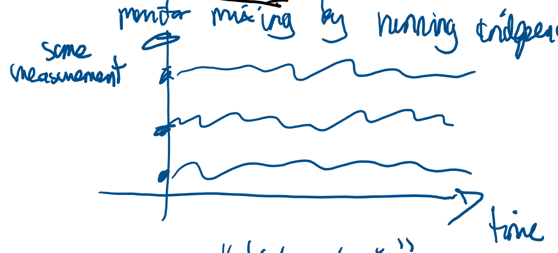
⟹ renormalize to get conditional

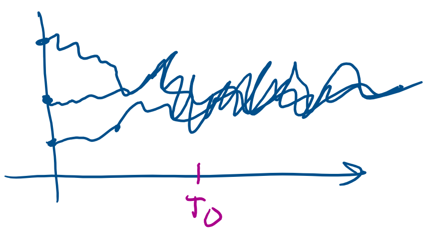
$$p(x_i = 1 \mid x_{\neg i}) = \frac{\exp\big(\eta_i + \sum_{j \in N(i)} \eta_{ij} x_j\big) \cancel{\exp(\text{rest})}}{\big(1 + \exp(\eta_i + \sum_{j \in N(i)} \eta_{ij} x_j)\big) \cancel{\exp(\text{rest})}}$$

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

diagnostic of mixing

monitor mixing by running independent chains





"sticky chain"

→ slow mixing

13h25

variational methods

general idea: say we want to approximate θ^*

then, express it as a solution to opt. problem

$$\left.\theta^* = \arg\min_{\theta\in\Theta} f(\theta)\right] \text{OPT}$$

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

<u>linear algebra example</u>:

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

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

<u>variational EM (motivation for objective)</u>

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

recall EM trick:

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

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

E step: $\arg\max_{q\,\in \text{ all distributions on } Z} \mathcal{L}(q,\theta^{(t)}) \iff \arg\min_q KL(q(\cdot)\,\|\,p(\cdot|x,\theta^{(t)}))$

⊛ a variational approx for E step

do $q^{(t+1)}_{approx} = \arg\min_{q\in \underline{Q_{simple}}} KL(q(\cdot)\,\|\,p(\cdot|x,\theta^{(t)}))$

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

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

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

<u>but</u> <u>lose</u> monotonicity guarantee

on $\log p(x|\theta^{(t)})$ ...

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

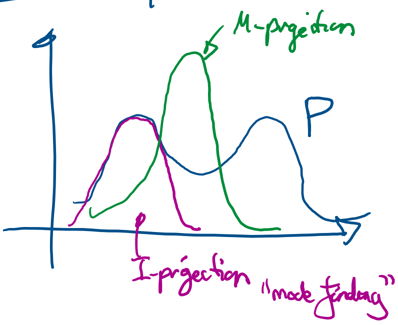
---

more generally, using $\underset{q \in Q}{\arg\min}\, KL(q \| p)$ is a variational approach to approx. p

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

alternative: $\underset{q \in Q}{\arg\min}\, KL(p \| q)$ 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

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

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

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

$$= -\eta^T \mathbb{E}_q[T(z)] + A(\eta) + \sum_z \underbrace{q(z)}_{\prod_j q_j(z_j)} \underbrace{\log q(z)}_{\sum_j \log q_j(z_j)}$$

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

$$\sum_i \sum_{z_i} q_i(z_i) \log q_i(z_i)$$

coordinate descent on q_i's

fix q_j for $j \neq i$ $\quad \underset{\text{w.r.t. } q_i}{\min} \quad KL(q_i q_{\neg i} \| p)$

$$= -\mathbb{E}_{q_i}\left[\eta^T \mathbb{E}_{q_{\neg i}}(T(z)) \right] + cst. + \sum_{z_i} q_i(z_i) \log q_i(z_i)$$

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

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

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

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

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

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

this is fct. of z_i only.
→ sufficient statistics for exp. family for q_i^*

Ising model:

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

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

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

no z_i ⇝ constant

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

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

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

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

parameter for $q_j^{(t)}$

compare with G.S. update

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