Lecture 21 - scribbles

Tuesday, November 21, 2017
14:17

today: · finish Gibbs sampling
· variational methods

(ctd for Gibbs sampling)

GS is MH with time varying proposal

suppose we pick i at time t

force rest to be constant

then proposal is $q_t(x' \mid x^{(t-1)}) = p(x_i' \mid x_{-i}^{(t-1)})\, \delta(x_{-i}', x_{-i}^{(t-1)})$

acceptance ratio:

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

$p(x_i^{(t-1)} \mid x_{-i}')$

$\rightarrow p(x_{-i}')\, p(x_i' \mid x_{-i}')$

$p(x_i' \mid x_{-i}^{(t-1)})\, \delta(x_{-i}', x_{-i}^{(t-1)}) \rightarrow p(x_{-i}^{(t-1)})\, p(x_i^{(t-1)} \mid x_{-i}^{(t-1)})$

$= 1$ → always accepts?

convergence of G-S.:

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

→ homogeneous M.C.

A is irreducible and aperiodic because $A_{ii} > 0 \ \forall i$

(since can get to any state with n steps)

(supposing $p(x_i | x_{\neg i}) > 0$ $\forall x \in \{0,1\}$)

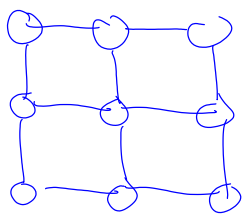
$\Rightarrow A^t \pi_0 \xrightarrow{t \to \infty} p$

⊗ 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:



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

for Gibbs sampling,

want to compute $p(x_i | x_{\neg i}) \propto p(x_i, \underbrace{x_{\neg i}}_{\text{fixed}})$

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

$\Rightarrow$ renormalize to get conditional:

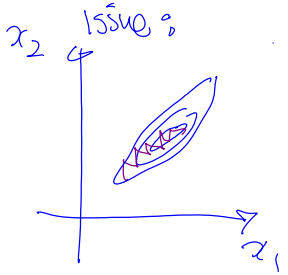
$$p(x_i = 1 | x_{\neg i}) = \frac{\exp(\eta_i + \sum_{j \in N(i)} \eta_{ij} x_j)}{1 + \exp(\dots)}$$

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

σ sigmoid $\sigma(z) \triangleq \frac{1}{1}$

$\frac{1}{1+\exp(-z)}$

Variants:



a) block Gibbs sampling

$p(x_A \mid x_{-A})$

set (block)

$\bigcup A_i = V$

b) Rao-Blackwellized Gibbs sampling: marginalize out some variables first

e.g. say can compute $p(x,y)$ from $p(x,y,z)$

(used in LDA (latent Dirichlet allocation))

overall

- need to be able to sample from $p(x_i \mid x_{\neq i})$ to run G.-S.

(otherwise, need M.H.)

Gibbs is easy when — discrete graph. model

— cts. DGM with "conjugate distributions"

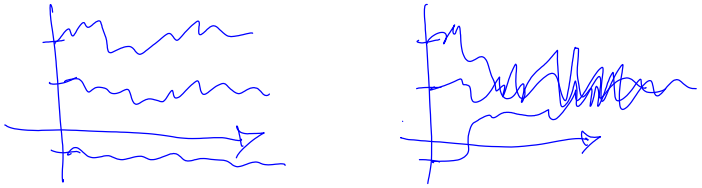
* good proposal is an art.

look at software: STAN (C++) (Andrew Gelman)

(BUGS)

diagnostic of mixing:

monitor mixing by running independent chains



"sticky chain"
→ slow mixing

usually slow mixing comes to difficulty to move between modes

→ annealing methods help this

proposal looks like $\frac{1}{Z}\exp\left(-\frac{k_B}{T}E(x)\right)$

physics analogy
energy
temperature
↳ high temperature ⇒ more exploration

example: "Annealed importance sampling"

Variational methods

general idea: say we want to approximate θ^*

then, express it as soln to optimization problem

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

idea: approximate θ^* by approximating OPT

linear algebra example: want to compute soln to $Ax = b$ i.e. $x = A^{-1}b$

$$\min_x \|Ax - b\|^2$$

variational EM:

recall the EM trick latent variable model $p(x, \underset{\text{unobserved}}{z})$

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

$$\log p(x|\theta) = \underbrace{KL(q(\cdot) \,\|\, p(\cdot|x,\theta))}_{\log p(x|\theta) - f(q,\theta)}$$

E-step: $\underset{q \in \text{all distributions}}{\operatorname{argmax}} \; f(q, \theta^{(t)}) \iff \underset{q}{\operatorname{argmin}} \; KL(q \,\|\, p(z|x,\theta^{(t)}))$

"a" variational approximation for the E-step:

do $q^{t+1}_{\text{approx}} = \underset{q \in \underline{Q_{\text{simple}}}}{\operatorname{argmin}} \; KL(q(\cdot) \,\|\, p(\cdot|x))$

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

(still get lower bound on $\log p(x|\theta^{(t)})$ but no more monotonicity guarantees)

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

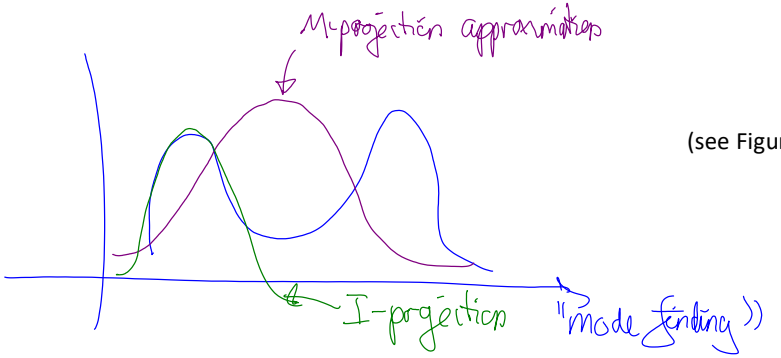
---

more generally, using $\arg\min_{q \in \mathcal{Q}} KL(q \| p)$ is a variational approach to approximate p

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

alternative : $\arg\min_{q \in \mathcal{Q}} KL(p \| q)$ M-projection

"motivation for" EP algorithm → moment matching
expectation propagation



(see Figure 10.2 in Bishop)

Mean-field approximation:

(section 10.1 in Bishop)

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

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

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

set of fully factorized distributions

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

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

$$\sum_i \left(\sum_{z_i} q(z_i)\log q(z_i)\right)$$

coordinate descent on q_is:

Fix q_j for $j \neq i$

minimize with respect to q_i $\quad KL(q_i q_{\neg i} \| p)$

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

add lagrange multiplier for $\sum_{z_i} q_i(z_i) = 1$ $\quad\rightsquigarrow\quad + \lambda\left(1 - \sum_{z_i} q_i(z_i)\right)$

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

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

general mean field update ~~when target p is in exp family~~

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

Ising model example:

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

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

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

$$\eta^\top \mathbb{E}_{q_{\neg i}^{(t)}} T(z) = \eta_i z_i + \sum_{j\neq i} \eta_j \underbrace{\mathbb{E}_{q_{\neg i}^{(t)}}[z_j]}_{\mu_j^{(t)}}$$

$$+ \sum_{j\in N(i)} \eta_{ij} \underbrace{\mathbb{E}_{q_{\neg i}^{(t)}}[z_i z_j]}_{z_i \mu_j^{(t)}} + \begin{matrix}\text{cst} \\ (\text{no } z_i)\end{matrix}$$

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

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

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

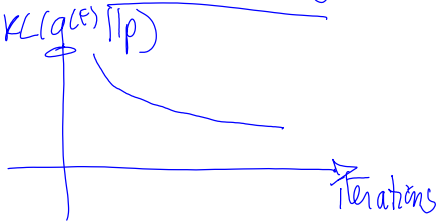
Compare with Gibbs update where $z_i^{(t+1)} = 1$ with prob. of $\sigma\left(\eta_i + \sum_{j\in N(i)} \eta_{ij} z_j^{(t)}\right)$

(*) here $\min_{q\in Q} KL(q\|p)$ $\quad KL(\cdot\|p)$ is a convex function of q

but Q_{MF} is a non-convex constraint set

⇒ can get stuck in local minima

but can monitor progress by



pros & cons of variational methods vs. sampling

⊕ optimization based
⇒ often faster & easier to debug

⊖ biased estimate

$\mathbb{E}_{q^{(\infty)}}[f(z)] \neq \mathbb{E}_p[f(z)]$

⊖ noisy ⇒ hard to debug
mixing problem for chains

⊕ unbiased estimate

$\mathbb{E}\left[\mathbb{E}_{q^{(\infty)}}[f(z)]\right] = \mathbb{E}_p[f(z)]$

$\mathbb{E}$ with respect to random sample