# Lecture 22 - scribbles

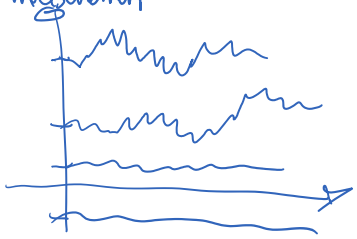
Friday, November 23, 2018 13:28

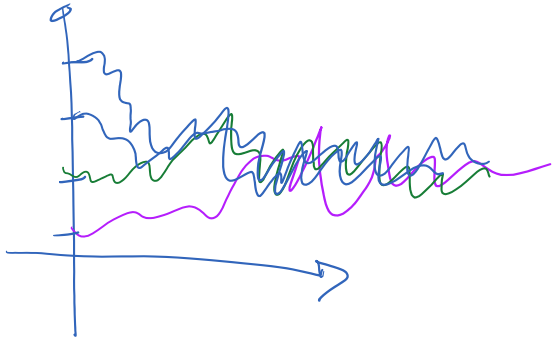
today: variational methods
mean field eq.
estimation is POM

## (sampling cted): diagnostic of mixing

monitor mixing by running independent chains

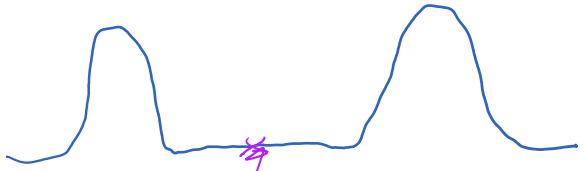
same measurement





"sticky chain"
↝ slow mixing

usually slow mixing comes from difficulty to move between modes



hard to cross

→ annealing methods help this

physics analogy
energy

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

"temperature"

high temperature ⇒ more exploration

example: "Annealed importance sampling"

14h47

## variational methods

general idea: say we want to approximate $\theta^*$

then, express it as solution to optimization problem

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

idea: approximate $\theta^*$ by approximating OPT

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

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

## variational EM (motivation)

recall EM trick: latent variable $p(x,z)$ (unobserved)

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

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

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

(*) a variational approximation for the E-step

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

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

( still get a lower on $\log p(x|\theta^{(t)})$ but no monotonicity guarantees for approximate EM )

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

---

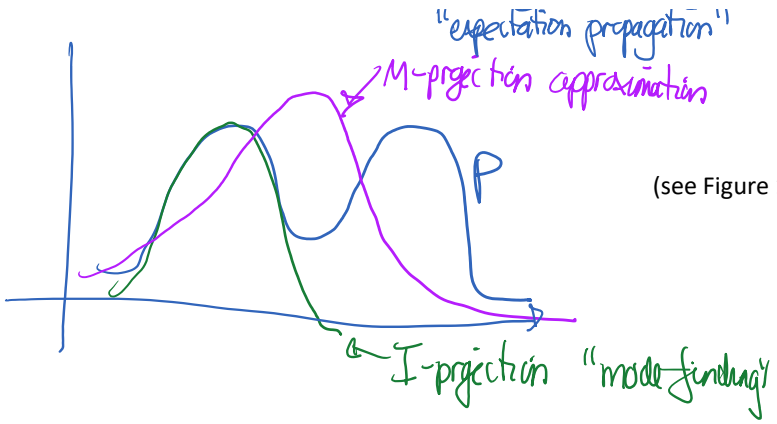
more generally, using $\underset{q \in Q}{\operatorname{argmin}} KL(q \| p)$ is a variational approach to approximate $p$

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

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

"motivation" EP algorithm "expectation propagation" → moment matching

↘ M-projection approximation




(see Figure 10.2 in Bishop)

## Mean-field approximation: (section 10.1 in Bishop)

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

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

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

set of fully factorized dist.

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

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

$$\sum_i \left[ \sum_z \prod_j q_j(z_j) \log q(z_i) \right]$$

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

coordinate descent on $q_i$'s:

fix $q_j$ for $j \neq i$

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

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

add Lagrange multiplier for $\sum_{z_i} q_i(z_i) = 1$ $\longrightarrow$ $+\lambda\left(1 - \sum_{z_i} q_i(z_i)\right)$

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

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

⊛ genral mean field update when target $p$ is in exp. family

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

14h40

Ising model example

$z_i \in \{0,1\}$

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

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

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

$\mu_j$ update

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

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

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

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

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

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

Compare with

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

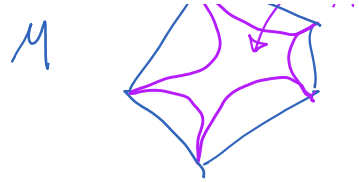
(*) here $\min_{q \in Q_{MF}} KL(q \| p)$

- $KL(\cdot \| p)$ is a convex of $q$
- but $Q_{MF}$ is a non-convex constraint set $\rightarrow \mu_{ij} = \mu_i \mu_j$

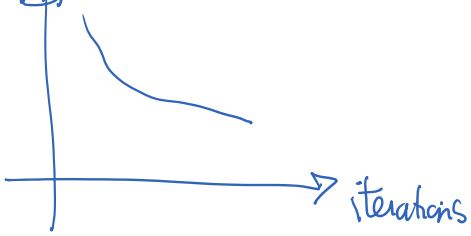
$\Rightarrow$ can get stuck in local minima

[See lecture 22 last year for "marginal polytope" def. [set of feasible mean parameters]]


lecture 22 Fall 2017 link



but can monitor progress by $KL(q^{(t)} \| p)$




pros & cons of variational methods vs sampling

⊕ optimization based ⇒ often faster & easier to debug

⊖ noisy ⇒ hard to debug

mixing problem for chains

⊖ biased estimate

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

⊕ unbiased estimate

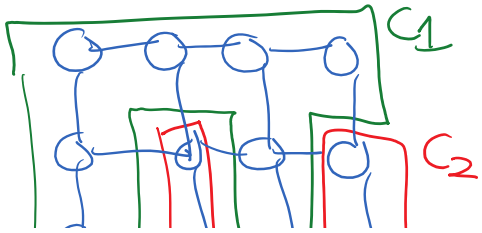
$\mathbb{E}[\mathbb{E}_{q^{(\infty)}}[f(z)]] = \mathbb{E}_p[f(z)]$

↳ with respect to random sample

lecture 22 Fall 2017 link

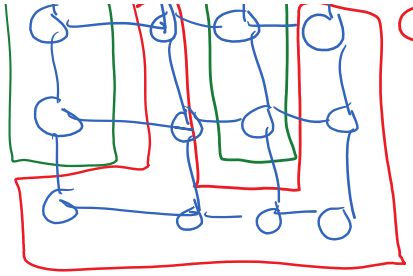
Structured mean field: (see lect. 22 last year)

idea $q(z) = \prod_{j=1}^{k} q_j(z_{C_j})$




where $C_1, \dots, C_k$ partition of $V$

and $q_j$s are tractable distributions (for example tree UGM)




(for example tree GSM)

## Estimation of parameters for PGM

DGM: parametric family $P_\Theta = \{ p_\theta(x) = \prod_i p(x_i \mid x_{\pi_i}, \theta_i) \mid \theta = (\theta_1, \dots, \theta_{|V|}) \in \Theta = \Theta_1 \times \dots \times \Theta_{|V|} \}$

independent parameterization

ie. no tying of parameters

$\Rightarrow$ MLE decouples in $|V|$ independent problems

$\{x^{(i)}\}_{i=1}^n$ $\quad p(\text{data} \mid \theta) = \prod_{i=1}^n p(x^{(i)} \mid \theta) = \prod_{i=1}^n \prod_{j=1}^{|V|} p(x_j^{(i)} \mid x_{\pi_j}^{(i)}, \theta_j)$

$$\log[\quad] = \sum_{j=1}^{|V|} \underbrace{\left( \sum_{i=1}^n \log p(x_j^{(i)} \mid x_{\pi_j}^{(i)}, \theta_j) \right)}_{f_j(\theta_j)}$$

example: for discrete R.V. $\Rightarrow \theta_j^{ML}$ = proportion of observations

$$\frac{\#(x_j = k,\ x_{\pi_j} = \text{something})}{\#(x_{\pi_j} = \text{something})}$$

$\#(x_j = \text{something})$

(*) if have latent variable (i.e. unobserved variable)

$\Rightarrow$ use E.M.

UGM:

example for exp family

$\rightarrow \exp(\eta_c^T T_c(x_c)) = \psi_c(x_c)$

$$p(x|\eta) = \exp\left(\sum_c \eta_c^T T_c(x_c) - A(\eta)\right)$$

gradient ascent on log-likelihood:

$$\frac{1}{n}\sum_{i=1}^{n} \log p(x^{(i)}|\eta) = \sum_c \eta_c^T \left(\underbrace{\frac{1}{n}\sum_{i=1}^{n} T_c(x_c^{(i)})}_{\hat{\mu}_c}\right) - \cancel{\frac{1}{n}} A(\eta)$$

$$\nabla_{\eta_c}(\quad) = \hat{\mu}_c - \mu_c(\eta)$$

$\mu_c(\eta) \rightarrow \mathbb{E}_{p(x|\eta)}[T_c(x_c)]$

to compute this, need inference

e.g. Ising model $T_{ij}(x_i, x_j) = x_i x_j$

$\mathbb{E}[T_{ij}] = \mu_{ij} = p(x_i=1, x_j=1|\eta)$

perhaps use approximate inference $\longleftarrow$ variational

15h26

sampling