Lecture 22 - scribbles

Friday, November 24, 2017
13:30

today: - finish variational methods
- ML for graphs models

structured mean field:

context: suppose target p is UGM with graph $G=(V,E)$

(see https://metacademy.org/graphs/concepts/structured_mean_field for pointers)

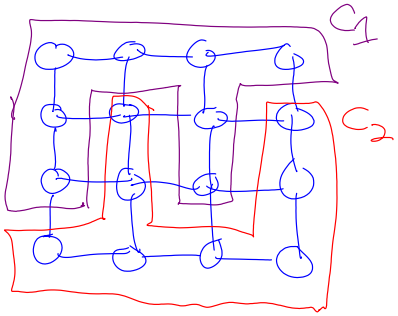
general idea: use $q(z) = \prod_{j=1}^{k} q_j(z_{C_j})$

where $C_1, \dots, C_k$ partition V

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

to do this, you choose C_j

s.t. G restricted to C_j is a low treewidth UGM



⊛ suppose $q(z) = \prod_j q_j(z_{C_j})$

where q_j is arbitrary distribution over z_{C_j}

it's easy to show that

$q_j^*(z_{C_j})$ is UGM with structure $G|C_j$ (graph G restricted on nodes in C_j)

i.e. when do $\min_{q \in \mathcal{Q}} KL(q \| p)$

⊛ we can derive mean field update as before using coordinate descent on $KL(q\|p)$

to get $q_j^{(t+1)}(z_{C_j}) \propto \exp\left(\eta^T \mathbb{E}_{q^{(t)}_{\setminus C_j}}(T(z))\right)$

this will be UGM $G|C_j$

Ising model example:

$$\eta^T T(z) = \sum_i \eta_i z_i + \sum_{\{i,j\}\in E} \eta_{ij} z_i z_j \qquad z_i \in \{0,1\}$$

$$\mathbb{E}_{q^{(t)}_{\setminus C_k}}[\eta^T T(z)] = \sum_{i\in C_k}\Big[\eta_i z_i + z_i \Big(\sum_{j\in N(i)\setminus C_k} \eta_{ij}\, \mathbb{E}_{q^{(t)}_{\setminus C_k}}(z_j)\Big)$$

(effect of rest of graph)

$$+ \sum_{\substack{\{i,j\}\in E\\ i,j\in C_k}} \eta_{ij} z_i z_j + \underbrace{\text{rest}}_{q}\Big]$$

does not depend on z_i for $i \in C_k$

Convex optimization viewpoint on variational inference Wainwright & Jordan FTML 2008 paper

p in exponential family

$$KL(q\|p) = \mathbb{E}_q\left[\log \frac{q(x)}{p(x)}\right] = -\eta^T \underbrace{\mathbb{E}_q[T(x)]}_{\triangleq \mu(q)} + A(\eta) - (-\mathbb{E}_q[\log q(x)])$$

$$= -\eta^T \mu(q) + A(\eta) - H(q)$$

$KL(q\|p) \geq 0$ (recall that $\nabla_\eta A(\eta) = \mu(p_\eta)$)

$$\Rightarrow \quad A(\eta) \geq \eta^T \mu(q) + H(q) \quad \text{for any } q$$

Wainwright showed

$$A(\eta) = \sup_{q \in \text{all dist.}} \eta^T \mu(q) + H(q) = \sup_{\mu \in \mathcal{M}} \eta^T \mu + H(\eta(\mu))$$

$\mu^{-1}(\mu)$

since $\mu(q) \triangleq \mathbb{E}_q[T(x)]$

$\mathcal{M} = \{ \mu : \exists \text{ a dist. } q \text{ s.t. } \mathbb{E}_q[T(x)] = \mu \}$ (convex set)

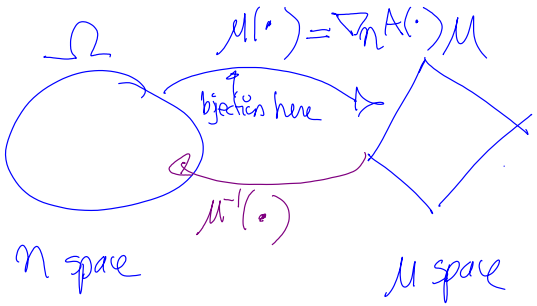
$\triangleq$ Marginal polytope

(e.g. for Ising model : $T(z) = \begin{pmatrix} (z_i)_{i \in V} \\ (z_i z_j)_{\{i,j\} \in E} \end{pmatrix}$)

for a minimal exponential family, Wainwright et al. showed

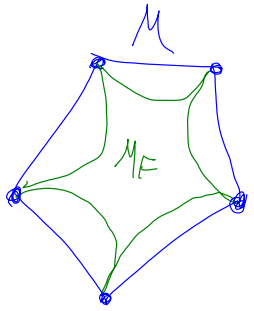
that there is a one-to-one correspondence between $\mu \in \text{int}(\mathcal{M})$

and $\eta \in \text{int}(\Omega)$



$$A(\eta) = \sup_{\mu \in \mathcal{M}} \eta^T \mu + H(\eta(\mu))$$

intractable for Ising model



idea : approximate $\mathcal{M}$ & H to get tractable approximation

mean field approximate is to restrict $\mathcal{M}$ to $\mathbb{E}_q[T(x)]$ where q is fully factorized

recall that for Ising model

here for mean field :

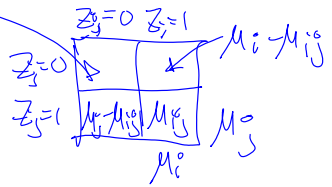
$(\mu_i)_{i \in V}$

$\mu_{ij} \triangleq p(z_i=1, z_j=1) \quad \{i,j\} \in E$

$\mu_{ij} = \mu_i \cdot \mu_j$

consistency constraints: $\mu_i = \mu_{ij} + p(z_i=1, z_j=0)$

$1 + \mu_{ij} - \mu_j - \mu_i$



⊛ outer bound approximation: → approximation of entropy as if it was a tree

$\sup_{\mu \in \mathcal{L}} \eta^T \mu + H_{\text{Bethe}}(\mu)$

$\mathcal{L}$ local consistency polytope $\mathcal{M} \subseteq \mathcal{L}$

which says e.g. that $\sum_k p(z_i, z_j=k) = p(z_i)$

↓

recall that for a tree, we had $p(z) = \prod_i p(z_i) \prod_{\{i,j\} \in E} \frac{p(z_i, z_j)}{p(z_i)p(z_j)}$

other approaches: TRW approach → gives a convex upper bound on $A(\eta)$ optim.

aside: $f^*(\mu) \triangleq \sup_{\eta} \eta^T \mu - f(\eta)$ Fenchel-conjugate

when f is convex, $f^{**}(\eta) = f(\eta)$

it turns out that $A^*(\mu) = \begin{cases} -H(\eta(\mu)) & \text{for } \mu \in \text{int}(\mathcal{M}) \\ +\infty & \text{for } \mu \notin \mathcal{M} \end{cases}$

- Once $\eta(\text{data}:\eta) = \sum \eta^T T(z_i) - \eta A(\eta)$

- $\log p(\text{data}; \eta) = \sum_i \eta^T T(x_i) - nA(\eta)$

Estimation of parameters for PGM:

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

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. $\Rightarrow \hat{\theta}_j^{ML}$ = proportion of observations e.g. $\frac{\#(x_j = k, x_{\pi_j} = \text{something})}{\#(x_{\pi_j} = \text{something})}$

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

$\Rightarrow$ use EM

PGM -

UGM:

example for exponential family:

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

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

gradient ascent on log-likelihood:

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

$\hat{\mu}_c$

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

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

to compute this, need inference

eg. 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 — variational, sampling