VARIATIONAL INFERENCE

9/9/19

VARIATIONAL LOWER BOUND

⇒ Consider observed RV $X = x$ w/ marginal likelihood:

$$\log p(x) = \log \int p(x,\theta)\, d\theta$$
$$= \log \int q(\theta) \frac{p(x,\theta)}{q(\theta)}\, d\theta$$
$$= \log \mathbb{E}_q\left[\frac{p(x,\theta)}{q(\theta)} \right]$$
$$\geq \mathbb{E}_q\left[\log \frac{p(x,\theta)}{q(\theta)} \right] \quad \text{(Jensen's inequality)}$$
$$= -KL(q \,\|\, p(x,\theta))$$

TALK: FRI, 1-2pm, ENR2 S395. "PROB. REASONING IN COMPLEX SYSTEMS: ALG. & APP"

⇒ Variational optimization finds tightest lower bound:

$$\log p(x) \geq \max_{q \in Q} -KL(q(\theta) \,\|\, p(x,\theta))$$

for some class of dist'ns Q

- Bound is tight iff $q(\theta) = p(\theta | x)$
- KEY IDEA: VI turns statistical inference into optimization

⇒ Different VI methods amount to different Q

- $\boxed{\text{Mean field}} = Q^{MF}$ marginal posterior independence
- Belief Propogation $= Q^{BETHE}$ "Tree-like" posterior
- Expectation Propatation $= Q^{EP}$ "Tree-Like" in moments

Today's Focus

⇒ ASIDE: Info-Theoretic interpretation suggests we should minimize

$$\min_q KL(p(\theta|x) \,\|\, q(\theta)) = \mathbb{E}_{p(\theta|x)}\left[\log \frac{p(\theta|x)}{q(\theta)} \right]$$

Can't compute this

Don't have this ☹

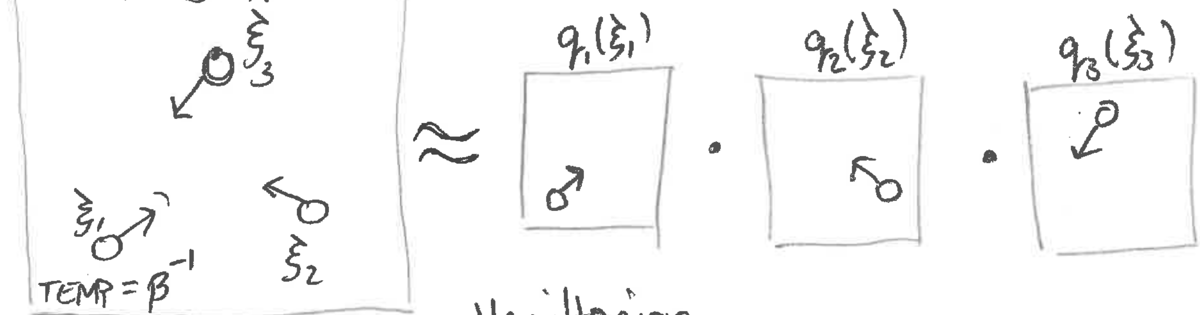
MEAN FIELD VARIATIONAL

⇒ "Mean field" assumes d-dimensional θ, marginally independent.

$$q(\theta) = \prod_{i=1}^{d} q_i(\theta_i)$$

⇒ Originates from many-body problem in stat. mechanics:

$p(\xi)$ $\qquad$ ξ_i = "microstates" (e.g. spin, velocity, ...)



Hamiltonian

$$p(\xi) = \underbrace{\frac{1}{Z} e^{-\beta H(\xi)}}_{\substack{\text{GIBBS' DIST'N} \\ \text{BOLTZMANN DIST'N}}} \approx \prod_i \frac{1}{Z_i} e^{-\beta h_i(\xi_i)} \triangleq \prod_i q_i(\xi_i)$$

⇒ Mean field variational lower bound:

$$\log p(x) \geq \max_q \mathbb{E}_q[\log p(\theta, x)] + \sum_i H(q_i) \triangleq \mathcal{L}(q)$$

- Sometimes called "Evidence Lower Bound" (ELBO)

⇒ How do we optimize $\mathcal{L}(q)$? Coordinate ascent.

- Take derivative wrt each $q_i(\theta_i)$, set to 0, solve:

$$q_i^*(\theta_i) \propto \exp\left\{ \mathbb{E}_{q_{\neg i}}\left[\log \underbrace{p(\theta_i \mid \theta_{\neg i}, x)}_{\text{KNOWN AS "COMPLETE CONDITIONAL"}}\right]\right\} \qquad (*)$$

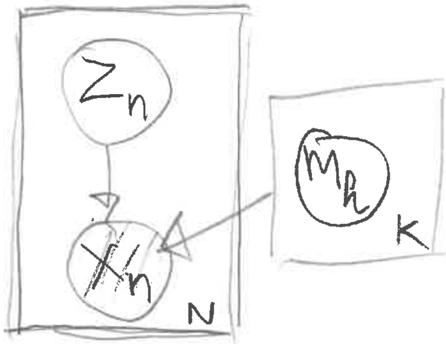
where $\neg i = \{1, \ldots, d\} \setminus i$

COORDINATE ASCENT VI:

$\Rightarrow$ Monotonically increases $\mathcal{L}(q)$

$\Rightarrow$ Fixed-points of (*) $\Leftrightarrow$ Local optima of $\mathcal{L}(q)$

Ex. GAUSSIAN MIXTURE MODEL



$$m_k \sim N(0, \sigma^2) \quad \text{for } k = 1, \dots, K$$
$$Z_n \sim \text{CAT}(1/K, \dots, 1/K) \quad \text{for } n = 1, \dots, N$$
$$X_n | m_{z_n} \sim N(m_{z_n}, 1) \quad \dots$$

- Variational distributions:

$$q_n(z_n | \pi_n) = \text{CAT}(z_n | \pi_n) \quad , \quad q_k(m_k | \mu_k, s_k) = N(m_k | \mu_k, s_k)$$

where $\{\pi_n, \mu_k, s_k\}$ are variational parameters

(1) Update cluster assignments:

$$q_n(z_n | \pi_n) \propto \exp\Big\{ \underbrace{\log p(z_n)}_{\text{PRIOR} = 1/K} + \underbrace{\mathbb{E}_{q(m|\mu,s)}\left[\log p(x_n | z_n, m)\right]}_{\text{EXPECTED LOG-LIKELIHOOD}} \Big\} \quad (**)$$

- Compute expected log-likelihood:

$$\mathbb{E}_{q(m)}\left[\log p(x_n | z_n, m)\right] = \mathbb{E}\left[\sum_k \mathbb{1}(z_n = k) \log N(x_n | m_k, 1)\right]$$
$$= \sum_k \mathbb{1}(z_n = k)\, \mathbb{E}\left[-\tfrac{1}{2}\log 2\pi 1 - \tfrac{1}{2}x_n^2 + x_n m_k - \tfrac{1}{2} m_k^2\right]$$
$$= \sum_k \mathbb{1}(z_n = k)\left(\underbrace{\mathbb{E}_{q(m_k)}[m_k]}\, x_n - \tfrac{1}{2}\,\mathbb{E}_{q_k(m_k)}[m_k^2]\right) + \text{Const.}$$

- Plug back into (**)

$$q_n(z_n = k) = \pi_n(k) \propto \exp\Big\{ \mu_k x_n - \tfrac{1}{2}(s_k + \mu_k^2) \Big\}$$

(2) Update component means $q(m_k | \mu_k, s_k)$ (we will skip that)

CONDITIONALLY CONJUGATE MODELS

$\Rightarrow$ In the GMM example our update was:

$$q_n(z_n) \propto p(z_n) \exp\{\mathbb{E}[\log p(x_n \mid z_n, m)]\} \propto \mathrm{CAT}(\cdot)$$

Which is same family as prior $p(z_n)$

- We call this conditional conjugacy

$\Rightarrow$ In general this holds when:

1) Complete conditionals are expfam
2) $q(\theta) \propto p(\theta) \exp\{\mathbb{E}_z[\log p(x, z \mid \theta)]\}$
$$= p(\theta) f(x \mid \theta)$$
is conjugate $\Rightarrow$ Lies in same family as $p(\theta)$

$\Rightarrow$ Conditional conjugacy $\Rightarrow$ Closed-form updates

VARIATIONAL OPTIMIZATION

$\Rightarrow$ How did we get fixed-point condition?

$$q_i(\theta_i) \propto \exp\{\mathbb{E}[\log p(\theta_i \mid \theta_{\neg i}, x)]\}$$

$\Rightarrow$ Form Lagrangian of variational problem:

$$\max_q \mathcal{L}(q) \triangleq \mathbb{E}_q[\log p(x, \theta)] + \sum_i H(q_i(\theta_i))$$
$$\text{s.t.} \int q_i(\theta_i) = 1 \quad \forall i = 1, \ldots, N$$

Lagrangian:

$$\mathcal{L}(q) + \sum_i \lambda_i \left(1 - \int q_i(\theta_i) d\theta_i\right) \triangleq J(q, \lambda)$$

$\Rightarrow$ Calculus of variations defines derivative of functional

$$J(f) = \int L(f(x))\, dx$$

AS:

$$\frac{\partial J}{\partial f(x)} = \frac{\partial L}{\partial f}$$

NOTE: This is a simplification of a functional and the Euler-Lagrange equations.

$\Rightarrow$ Take functional derivative wrt $q_j(\theta_j)$:

$$\frac{\partial}{\partial q_j(\theta_j)} \mathcal{L}(q) + \frac{\partial}{\partial q_j(\theta_j)} \lambda_j \left(1 - \int q_j(\theta_j) d\theta_j\right)$$

$$= \frac{\partial}{\partial q_j(\theta_j)} \mathbb{E}_{q_j}\left[\mathbb{E}_{q_{\neg j}}[\log p(\theta_j \mid \theta_{\neg j}, x)]\right]$$

$$+ \frac{\partial}{\partial q_j(\theta_j)} H(q_j(\theta_j)) - \lambda_j$$

$$= \mathbb{E}_{q_{\neg j}}[\log p(\theta_j \mid \theta_{\neg j}, x)] - \log q_j(\theta_j) - 1 - \lambda_j = 0 \Rightarrow$$

$$q_j^*(\theta_j) = \exp\left\{\mathbb{E}_{q_{\neg j}}[\log p(\theta_j \mid \theta_{\neg j}, x)]\right\} \div Z(\lambda_j)$$

Bethe Variational Problem:

$\Rightarrow$ Key Idea: Optimize over dist'ns $q \in \mathcal{Q}^{\text{BETHE}}$ consistent w/ tree-structured MRFs.

$\Rightarrow$ Any tree-structured dist'n $q(\theta)$ can be written as:

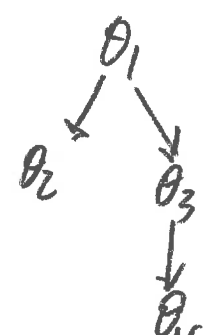
$$q(\theta) = \prod_{s \in \mathcal{V}} q_s(\theta_s) \prod_{(s,t) \in \mathcal{E}} \frac{q_{st}(\theta_s, \theta_t)}{q_s(\theta_s)\, q_t(\theta_t)} \qquad (*)$$

where $q_s(\theta_s)$ are marginals:

$$q_s(\theta_s) = \int q(\theta)\, d\theta_{\setminus s}$$

Ex.

$$q(\theta) = \left(q_1(\theta_1)\, \cancel{q_2(\theta_2)}\, \cancel{q_3(\theta_3)}\, \cancel{q_4(\theta_4)}\right) \left(\frac{q_{12}(\theta_1, \theta_2)}{q_1(\theta_1)\, \cancel{q_2(\theta_2)}}\right) \left(\frac{q_{13}(\theta_1, \theta_3)}{q_1(\theta_1)\, \cancel{q_3(\theta_3)}}\right) \cdot \left(\frac{q_{34}(\theta_3, \theta_4)}{q_3(\theta_3)\, \cancel{q_4(\theta_4)}}\right)$$



$$= q_1(\theta_1)\, q_{2|1}(\theta_2 | \theta_1)\, q_{3|1}(\theta_3 | \theta_1)\, q_{4|3}(\theta_4 | \theta_3)$$

$\Rightarrow$ Dist'ns of the form $(*)$ necessarily satisfy Local Marginal Consistency:

$$q_s(\theta_s) = \int q_{st}(\theta_s, \theta_t)\, d\theta_t \quad \forall t \in \Gamma(s) \leftarrow \text{neighbors of } s$$

$\Rightarrow$ Bethe Variational Problem:

$$\max_q \; \mathbb{E}_q[\log p(\theta, x)] + H^{\text{BETHE}}(q)$$

s.t.

$$\int q_{st}(\theta_s, \theta_t)\, d\theta_t = q_s(\theta_s) \quad \forall t \in \Gamma(s) \quad \text{(Local Consistency)}$$

$$\int q_s(\theta_s)\, d\theta_s = 1 \quad \forall s \in \mathcal{V} \quad \text{(Normalization)}$$