RECITATION 7

Variational inference & convex optimization

Variational inference: We have a latent-variable model with unobserved variables z and observed variables x. Given a family of parametric joint distributions $p(x, z; \theta)$ indexed by θ, and n observations $(x_{1:n})$ from x, we want to compute the maximum likelihood value for θ (in order to obtain an estimate of the joint distribution of θ and x):

$$\max_\theta \sum_i \log p(x_i; \theta) = \max_\theta \sum_i \log \left(\sum_z p(x_i, z; \theta) \right)$$

Example: clustering with discrete assignments.

- $x_i \in \mathbb{R}^p$, $i = 1, \dots n$ data
- $\mu_k \in \mathbb{R}^p$, $k = 1, \dots K$ cluster centers
- $z_i \in [K]$, cluster assignments

$$p(x \mid z; (\mu_k)_k) = \mathcal{N}(\mu_z; Id)$$

EM algorithm

Variational lower bound (VLB):

for some $q \geq 0$, $\sum_z q(z) = 1$

$$\log p(x; \theta) = \log \sum_z p(x, z; \theta) = \log \sum_z q(z) \frac{p(x, z; \theta)}{q(z)}$$

$$= \log \mathbb{E}_q \left[\frac{p(x, z; \theta)}{q(z)} \right] \underset{\text{Jensen's}}{\geq} \mathbb{E}_q \left[\log \frac{p(x, z; \theta)}{q(z)} \right]$$

Thus, $\log p(x; \theta) \geq \underbrace{\mathbb{E}_q[\log p(x, z; \theta)] - \mathbb{E}_q[\log q(z)]}_{= L(\theta, q)}$

Expectation-maximization (E-M) algorithm:

For $t \geq 0$:

$$q_{t+1} \in \arg\max_q L(\theta_t, q) \left[= \overbrace{p(z \mid x; \theta_t)}^{\text{this is the } q \text{ that maximizes } L, \text{ proportional to } p(x, z; \theta_t)} \right]$$

(E-step)

$$\theta_{t+1} \in \arg\max_\theta L(\theta, q_{t+1}) \left[= \arg\max_\theta \mathbb{E}_{q_{t+1}}\left[\log p(x, z; \theta)\right] \right]$$

The gap between $\log p(x; \theta)$ and $L(\theta, q)$ is a KL divergence:

$$\log p(x;\theta) - L(\theta, q) = \log \cancel{p(x;\theta)} - E_q\left[\log \frac{\cancel{p(x;\theta)}\, p(z \mid x;\theta)}{q(z)}\right]$$

$$= E_q\left[\log \frac{q(z)}{p(z \mid x;\theta)}\right] = D_{KL}\left(q, p(\cdot \mid x; \theta)\right)$$

Is EM correct? We can show that any improvement we make on the value of the VLB translates to an improvement on the value of $\log p(x;\theta)$. Namely, for any θ, we have $\log p(x;\theta) - \log p(x;\theta_t) \geq L(\theta, q_{t+1}) - L(\theta_t, q_{t+1})$

Proof: $\log p(x \mid \theta) = \log p(x, z; \theta) - \log p(z \mid x, \theta)$

$$= \sum_z p(z \mid x, \theta_t) \log p(x, z; \theta) - \sum_z p(z \mid x, \theta_t) \log p(z \mid x, \theta)$$

$$= L(\theta, q_{t+1}) + H(\theta \mid \theta_t) \leftarrow \text{we define } H(\cdot \mid \cdot) \text{ like this}$$

Substituting $\theta = \theta_t$: $\log p(x \mid \theta_t) = L(\theta_t, q_{t+1}) + H(\theta_t \mid \theta_t)$

Subtracting we obtain:

$$\log p(x \mid \theta) - \log p(x \mid \theta_t) = L(\theta, q_{t+1}) - L(\theta, q_t) + H(\theta \mid \theta_t) - H(\theta_t \mid \theta_t)$$

By Gibbs' inequality $H(\theta \mid \theta_t) \geq H(\theta_t \mid \theta_t)$, and this concludes the proof.

Review on convexity

(m-strong) Convex function: $f : \Omega \subseteq \mathbb{R}^N \to \mathbb{R}$ if $\forall x, y \in \Omega$, $\forall t \in [0,1]$,

$$f(tx + (1-t)y) \leq t f(x) + (1-t) f(y) - \frac{m}{2} t(1-t) \|x-y\|^2$$

- First order characterization:

$$f(x) \geq f(y) + \langle \nabla f(y), x-y \rangle + \frac{m}{2} \|x-y\|^2$$

- Second order characterization: $\nabla^2 f(x) \geq m I$

"Easy" functions to optimize.

Jensen's ineq: if $z \sim \mu$ prob. measure and f convex,

$$f(\mathbb{E} z) \leq \mathbb{E} f(z) \quad \left(\text{and for } f \text{ concave, } \mathbb{E} f(z) \leq f(\mathbb{E} z)\right)$$

Example: Gibbs inequality: if $p \geq 0$, $q > 0$ are densities of prob. measures, $-\int p(x) \log p(x) \leq -\int p(x) \log q(x)$.

Proof: $\int p(x) \log \frac{q(x)}{p(x)} = -\int q(x) \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)}$

$$= \mathbb{E}_q \Big(\underbrace{- \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)}}_{\text{concave in } \frac{p}{q}} \Big] \underset{\text{Jensen}}{\leq} -\left(\mathbb{E}_q \frac{p}{q}\right) \log \left(\mathbb{E}_q \frac{p}{q}\right) = \left[-x \log x\right]_{x=1} = 0$$

Legendre transform: $f : \Omega \subseteq \mathbb{R}^N \to \mathbb{R}$.

$$f^*(x) = \sup_{y \in \Omega} \{ y^T x - f(y) \}$$

f^* is convex: $f^*(tx + (1-t)y) \leq \cdots \leq t f^*(x) + (1-t) f^*(y)$

Example (try it): $f(x) = x \log x \Rightarrow f^*(y) = e^{y-1}$

- If f is convex, $f^{**} = f$.

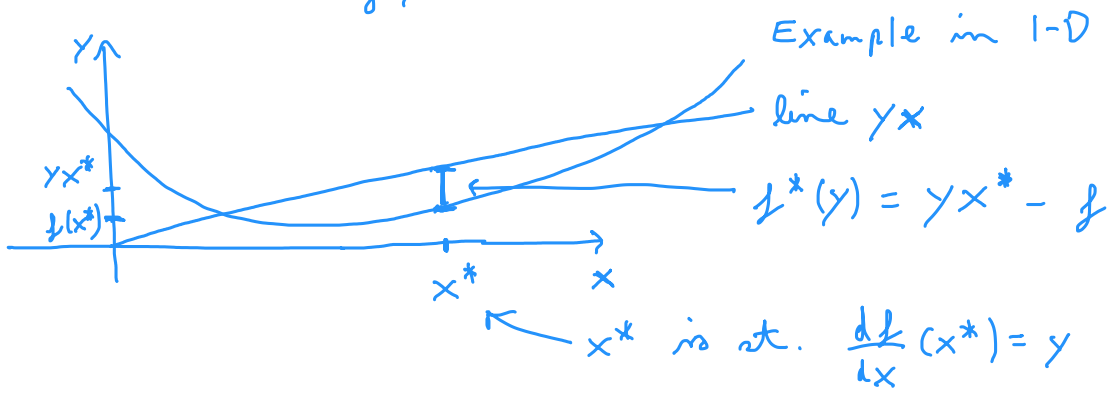
Geometric interpretation of the Legendre transform:

Supp. f diff. and convex. Let

$$y^* = \arg\max \; y^T x - f(x)$$

$$\Rightarrow \quad y - \nabla f(x^*) = 0 \quad \Rightarrow \quad x^* = (\nabla f)^{-1}(y)$$

$$\Rightarrow f^*(y) = y^T (\nabla f)^{-1}(y) - f((\nabla f)^{-1}(y))$$



Another characterization of Legendre transform: the only function f^* such that $(\nabla f^*)(\nabla f(x)) = x$.

<u>Duality</u>

$$p^* = \min_{\substack{f_i(x) \le 0 \\ h_i(x) = 0}} f_0(x) = \min_x \max_{\substack{\lambda_i \ge 0 \\ \mu_i}} \left\{ f_0(x) + \sum_i \lambda_i f_i(x) + \sum_i \mu_i h_i(x) \right\}$$

$$= \min_x \max_{\substack{\lambda_i \ge 0 \\ \mu_i}} \underbrace{L(x, \lambda, \mu)}_{\text{Lagrangian}}$$

Dual problem: $$d^* = \max_{\substack{\lambda_i \ge 0 \\ \mu_i}} \underbrace{\min_x L(x, \lambda, \mu)}_{g(\lambda, \mu)} = \max_{\substack{\lambda_i \ge 0 \\ \mu_i}} g(\lambda, \mu)$$

Weak duality: $d^* \le p^*$.

Proof: Let x^* be the minimizer of the primal problem, $\sum_i \lambda_i f_i(x^*) + \sum_i \mu_i h_i(x^*) \le 0$. Then

$$g(\lambda, \mu) = \min_x L(x, \lambda, \mu) \le \min_x f_0(x) = p^*$$

Strong duality: Suppose f_0, f_i are convex and h_i are affine. Suppose that Slater's condition holds: there exists a feasible point x such that $\forall i$, $f_i(x) < 0$. Then $p^* = d^*$ and $x \in \mathbb{R}^n$ is a solution of the primal problem if and only if there exist $\lambda_i \geq 0, \mu_i$ st:

$$\begin{cases} \nabla f(x) + \sum_i \lambda_i \nabla g_i(x) + \sum_i \mu_i \nabla h_i(x) = 0 \\ \lambda_i g_i(x) = 0 \qquad \forall i = 1, \dots m \end{cases}$$

and x is feasible.

Example: Entropy maximization under moment constraints.

$$\max \quad H(p) = -\sum_{i=1}^{N} p_i \log p_i$$
$$\sum_i p_i = 1$$
$$p_i \geq 0$$
$$\sum_i x_i p_i = m$$

$$\downarrow$$

$$\max_{p} \min_{\substack{\mu, \delta \\ \lambda_i \geq 0}} \left[H(p) + \mu \left(\sum_i p_i - 1 \right) + \delta \left(\sum_i x_i p_i - m \right) + \sum_i \lambda_i p_i \right]$$

$\downarrow$ Dual problem

$$\min_{\substack{\mu, \delta \\ \lambda_i \geq 0}} \underbrace{\max_{p} \left[H(p) + \mu \left(\sum_i p_i - 1 \right) + \delta \left(\sum_i x_i p_i - m \right) + \sum_i \lambda_i p_i \right]}_{g(\mu, \delta, \lambda)}$$

$$g(\mu, \delta, \lambda) = \left[\max_p \left\{ (\mu \mathbb{1} + \delta x + \lambda)^T p - \sum_i p_i \log p_i \right] - \mu - m\delta$$

$$= H^* (\mu \mathbb{1} + \delta x + \lambda) - \mu - m\delta, \qquad H^*(p) = \sum_i e^{p_i - 1}$$

is the legendre transform of H