Lec 6: Missing Data & EM Algorithm

Yanjun Han

Oct. 17, 2023

Convex conjugate

Defn (convex conjugate). Let f be convex on $\mathbb{R}^d$. The convex conjugate f^* of f is defined as

$$f^*(t) = \max_{x \in \mathbb{R}^d} \langle t, x \rangle - f(x).$$

Examples. 1. if $f(x) = \frac{1}{2}\|x\|_2^2$, then

$$\begin{aligned} f^*(t) &= \max_{x \in \mathbb{R}^d} \langle t, x \rangle - \tfrac{1}{2}\|x\|_2^2 \\ &= \max_{x \in \mathbb{R}^d} -\tfrac{1}{2}\|x - t\|_2^2 + \tfrac{1}{2}\|t\|_2^2 \\ &= \tfrac{1}{2}\|t\|_2^2 \end{aligned}$$

2. if $f(x) = \sum_{i=1}^d x_i \log x_i$, then

$$f^*(t) = \max_x \left(t_1 x_1 + t_2 x_2 + \cdots + t_d x_d - \sum_{i=1}^d x_i \log x_i \right).$$

F.O.C. for maximization:

$$t_i - (1 + \log x_i^*) = 0 \iff x_i^* = e^{t_i - 1}$$

$$\Rightarrow f^*(t) = \sum_i (t_i x_i^* - x_i^* \log x_i^*) = \sum_i x_i^* = \frac{1}{e}\left(e^{t_1} + \cdots + e^{t_d}\right).$$

Properties. Assume that f is convex & continuously differentiable.

1. $t \mapsto f^*(t)$ is convex.

Pf. $t \mapsto \langle t, x \rangle - f(x)$ is affine.

$\Rightarrow t \mapsto \max_x \langle t, x \rangle - f(x) = f^*(t)$ is convex.

2. The maximizer x^* satisfies $t = \nabla f(x^*)$.

Pf. F.O.C.: $0 = \nabla_x \left(\langle t, x \rangle - f(x) \right)\big|_{x = x^*} = t - \nabla f(x^*)$.

3 (Fenchel-Young inequality) $f(x) + f^*(t) \geq \langle x, t \rangle$.

Equality holds iff $t = \nabla f(x)$

Pf. $f^*(t) = \max_{x_0} \langle x_0, t \rangle - f(x_0) \geq \langle x, t \rangle - f(x)$,

equality holds iff $x_0 = x$ is the maximizer, i.e. $t = \nabla f(x)$.

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

(In other words, $f(x) = \max_{t \in \mathbb{R}^d} \langle t, x \rangle - f^*(t)$.)

Pf. Fenchel-Young $\Rightarrow f(x) \geq \langle t, x \rangle - f^*(t) \quad \forall t$

$\Rightarrow f(x) \geq \max_{t \in \mathbb{R}^d} \langle t, x \rangle - f^*(t)$.

Equality condition in Fenchel-Young.

$$f(x) = \langle \nabla f(x), x \rangle - f^*(\nabla f(x))$$
$$\leq \max_{t \in \mathbb{R}^d} \langle t, x \rangle - f^*(t).$$

Therefore,

$$f(x) = \max_{t \in \mathbb{R}^d} \langle t, x \rangle - f^*(t) = f^{**}(x). \qquad \square$$

Application in exponential families. $p_\theta(y) = \exp(\langle \theta, T(y) \rangle - A(\theta))\, h(y)$.

Since $\nabla^2 A(\theta) = \mathrm{Cov}_\theta(T(y)) \succeq 0 \Rightarrow \theta \mapsto A(\theta)$ convex.

Therefore, $A(\theta) = \max_{\mu \in \mathbb{R}^d} \langle \theta, \mu \rangle - A^*(\mu)$. (Property 4)

Maximizer $\mu^*(\theta)$: $\mu^*(\theta) = \nabla_\theta A(\theta) = \mathbb{E}_\theta[T(y)]$.

↑ property 2

Missing data in exponential families

$(x_1, y_1), \cdots, (x_n, y_n) \overset{\text{i.i.d}}{\sim} p_\theta(x, y) = \exp(\langle \theta, T(x,y) \rangle - A(\theta))\, h(x,y)$.

$(x_1, \cdots, x_n)$: unobserved variables $\quad (y_1, \cdots, y_n)$: observed variables

Goal: find the MLE for θ.

EM algorithm: derivation.

incomplete log-likelihood for y:

$$\ell_\theta(y_1, \cdots, y_n) = \frac{1}{n} \sum_{i=1}^n \log p_\theta(y_i)$$
$$= \frac{1}{n} \sum_{i=1}^n \log \int_{\mathcal{X}} p_\theta(x, y_i)\, dx$$
$$= \frac{1}{n} \sum_{i=1}^n \log \int_{\mathcal{X}} \exp(\langle \theta, T(x, y_i) \rangle - A(\theta))\, h(x, y_i)\, dx$$
$$= \frac{1}{n} \sum_{i=1}^n \left(A_{y_i}(\theta) - A(\theta) \right),$$

where

$$A_{y_i}(\theta) = \log \int_{\mathcal{X}} \exp(\langle \theta, T(x, y_i) \rangle)\, h(x, y_i)\, dx$$

is convex in θ.

MLE: $\max_\theta \ell_\theta(y_1, \cdots, y_n) = \max_\theta \frac{1}{n} \sum_{i=1}^n (A_{y_i}(\theta) - A(\theta))$

$= \max_\theta \frac{1}{n} \sum_{i=1}^n \left(\max_{\mu_i} \langle \mu_i, \theta \rangle - A^*_{y_i}(\mu_i) - A(\theta) \right)$

$= \max_\theta \max_{\mu_1, \cdots, \mu_n} \underbrace{\frac{1}{n} \sum_{i=1}^n \left(\langle \mu_i, \theta \rangle - A^*_{y_i}(\mu_i) - A(\theta) \right)}_{=: f(\theta, \mu_1, \cdots, \mu_n)}$

Key intuition: 1. for fixed θ, $(\mu_1, \cdots, \mu_n) \mapsto f(\theta, \mu_1, \cdots, \mu_n)$ is concave

2. for fixed $(\mu_1, \cdots, \mu_n)$, $\theta \mapsto f(\theta, \mu_1, \cdots, \mu_n)$ is concave.

(Warning: $(\theta, \mu_1, \cdots, \mu_n) \mapsto f(\theta, \mu_1, \cdots, \mu_n)$ is NOT "jointly" concave!)

EM algorithm. Initialize any $\theta^{(0)} \in \mathbb{R}^d$. For $t = 0, 1, 2, \cdots$:

- E-step: $(\mu_1^{(t+1)}, \mu_2^{(t+1)}, \cdots, \mu_n^{(t+1)}) = \operatorname{argmax}_{(\mu_1, \cdots, \mu_n)} f(\theta^{(t)}, \mu_1, \cdots, \mu_n)$

$\Rightarrow \mu_i^{(t+1)} = \nabla_\theta A_{y_i}(\theta^{(t)}) = \mathbb{E}_{X \sim P_{\theta^{(t)}}(x|y_i)}[T(X, y_i)]$

("expectation" step)

- M-step: $\theta^{(t+1)} = \operatorname{argmax}_\theta f(\theta, \mu_1^{(t+1)}, \cdots, \mu_n^{(t+1)})$

$= \operatorname{argmax}_\theta \langle \frac{1}{n} \sum_{i=1}^n \mu_i^{(t+1)}, \theta \rangle - A(\theta)$

$\Rightarrow \nabla_\theta A(\theta^{(t+1)}) = \frac{1}{n} \sum_{i=1}^n \mu_i^{(t+1)}$

("maximization" step)

EM Intuition:

1. E-step: for each sample i with missing data x_i, think of a fake $\tilde{x}_i \sim p_\theta(x_i | y_i)$ and compute sufficient statistic $\mu_i = \mathbb{E}[T(\tilde{x}_i, y_i)]$.
2. M-step: aggregate the sufficient statistics "as if" there were no missing data problem: $\frac{1}{n} \sum_{i=1}^n \mu_i = \nabla A(\theta)$
3. Iterate the above process.

Example: Gaussian mixture model

For $i=1,\cdots,n$: $\mathbb{P}(x_i=j)=\pi_j$, $j=1,2,\cdots,m$,

$$y_i \mid x_i=j \sim N(\gamma_j, 1).$$

Unknown parameters: $(\pi_1,\cdots,\pi_m,\gamma_1,\cdots,\gamma_m)$

Unobserved variables: $(x_1,\cdots,x_n)$ observed variables: $(y_1,\cdots,y_n)$

Joint likelihood: $p_\theta(x,y)=\prod_{j=1}^m\left(\frac{\pi_j}{\sqrt{2\pi}}\exp\left(-\frac{(y-\gamma_j)^2}{2}\right)\right)^{1(x=j)}$

$$\propto \exp\left(\sum_{j=1}^m 1(x=j)\left(\log\pi_j+y\gamma_j-\frac{\gamma_j^2}{2}\right)-\frac{y^2}{2}\right)$$

Reparametrization: $\pi_j=\frac{e^{\alpha_j}}{\sum_{k=1}^m e^{\alpha_k}}$, $\tilde{\alpha}_j=\alpha_j-\frac{\gamma_j^2}{2}$.

Then
$$\begin{aligned}p_\theta(x,y) &\propto \exp\left(\sum_{j=1}^m 1(x=j)\left(\log\frac{e^{\alpha_j}}{\sum_k e^{\alpha_k}}+y\gamma_j-\frac{\gamma_j^2}{2}\right)\right)e^{-y^2/2}\\ &= \exp\left(\sum_{j=1}^m 1(x=j)\left(\tilde{\alpha}_j+y\gamma_j\right)-\log\left(\sum_{k=1}^m e^{\alpha_k}\right)\right)e^{-y^2/2}\\ &= \exp\left(\langle\theta,T(x,y)\rangle-A(\theta)\right)e^{-y^2/2},\end{aligned}$$

where
$$\theta=(\tilde{\alpha}_1,\cdots,\tilde{\alpha}_m,\gamma_1,\cdots,\gamma_m),$$
$$A(\theta)=\log\sum_{k=1}^m e^{\alpha_k}=\log\sum_{k=1}^m e^{\tilde{\alpha}_k+\frac{\gamma_k^2}{2}},$$
$$T(x,y)=(1(x=1),\cdots,1(x=m),y\cdot 1(x=1),\cdots,y\cdot 1(x=m))$$

Auxiliary variable $\mu_i=(p_{i,1},\cdots,p_{i,m},\mu_{i,1},\cdots,\mu_{i,m})$

E-step: $p_\theta(x=j\mid y)=\frac{p_\theta(x=j,y)}{\sum_{k=1}^m p_\theta(x=k,y)}=\frac{\pi_j\exp(-\frac{(y-\gamma_j)^2}{2})}{\sum_{k=1}^m\pi_k\exp(-\frac{(y-\gamma_k)^2}{2})}$,

$$\Rightarrow\begin{cases}p_{i,j}^{(t+1)}=\mathbb{E}_{\theta^{(t)}}[1(x=j)\mid y_i]=\dfrac{\pi_j^{(t)}\exp(-\frac{(y_i-\gamma_j^{(t)})^2}{2})}{\sum_{k=1}^m\pi_k^{(t)}\exp(-\frac{(y_i-\gamma_k^{(t)})^2}{2})}\\ \mu_{i,j}^{(t+1)}=\mathbb{E}_{\theta^{(t)}}[y_i 1(x=j)\mid y_i]=y_i\,p_{i,j}^{(t+1)}.\end{cases}$$

(In vector form: $\mu_i^{(t+1)}=\mathbb{E}_{x\sim p_{\theta^{(t)}}(x\mid y_i)}[T(x,y_i)]$)

M-step.

$$\frac{1}{n}\sum_{i=1}^{n}\mu_i^{(t+1)} = \nabla A(\theta^{(t+1)})$$

$$= \frac{1}{\sum_{k=1}^{m} e^{\widetilde{\alpha}_k + \frac{V_k^2}{2}}}\left(e^{\widetilde{\alpha}_1+\frac{V_1^2}{2}}, \cdots, e^{\widetilde{\alpha}_m+\frac{V_m^2}{2}}, V_1 e^{\widetilde{\alpha}_1+\frac{V_1^2}{2}}, \cdots, V_m e^{\widetilde{\alpha}_m+\frac{V_m^2}{2}}\right)$$

(all quantities here have superscript $(t+1)$)

$$= (\pi_1^{(t+1)}, \cdots, \pi_m^{(t+1)}, \pi_1^{(t+1)}V_1^{(t+1)}, \cdots, \pi_m^{(t+1)}V_m^{(t+1)})$$

$$\Rightarrow \begin{cases} \pi_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} p_{i,j}^{(t+1)} \\ V_j^{(t+1)} = \frac{1}{n\pi_j^{(t+1)}}\sum_{i=1}^{n} y_i p_{i,j}^{(t+1)} \end{cases}, \quad j = 1, 2, \cdots, m.$$

Generalization: Evidence Lower Bound (Optional)

Def. For probability distributions P, Q over $\mathcal{X}$, the Kullback-Leibler (KL) divergence is

$$D_{KL}(P \| Q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \quad (\text{could be } +\infty)$$

Thm. $D_{KL}(P\|Q) \geq 0$.

Pf. Since $\log t \geq 1 - \frac{1}{t}$ for all $t > 0$,

$$D_{KL}(P\|Q) = \sum_x p(x)\log\frac{p(x)}{q(x)} \geq \sum_x p(x)\left(1 - \frac{q(x)}{p(x)}\right)$$

$$= \sum_x p(x) - \sum_x q(x) = 1 - 1 = 0 \quad \square$$

Evidence lower bound (ELBO).

$$\log p_\theta(y^n) \geq \underbrace{\mathbb{E}_{x^n\sim q(\cdot)}\left[\log\frac{p_\theta(x^n, y^n)}{q(x^n)}\right]}_{\text{ELBO}}, \quad \forall q.$$

Pf. $\text{LHS} - \text{RHS} = D_{KL}(q(x^n) \| p_\theta(x^n|y^n)) \geq 0.$

Choice of q.

1. $q(x^n) = p_{\theta'}(x^n \mid y^n)$ for some $\theta' \in \Theta$.

$$\max_\theta \log p_\theta(y^n) = \max_\theta \max_{\theta'} \underbrace{\mathbb{E}_{x^n \sim p_{\theta'}(\cdot \mid y^n)}\left[\log \frac{p_\theta(x^n, y^n)}{p_{\theta'}(x^n \mid y^n)}\right]}_{=: L(\theta, \theta')}$$

- for fixed θ: $\operatorname{argmax}_{\theta'} L(\theta, \theta') = \theta$.
- for fixed θ':

$$\operatorname*{argmax}_\theta L(\theta, \theta') = \operatorname*{argmax}_\theta \mathbb{E}_{x^n \sim p_{\theta'}(\cdot \mid y^n)}[\log p_\theta(x^n, y^n)]$$

$$= \operatorname*{argmax}_\theta \left\langle \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{x_i \sim p_{\theta'}(\cdot \mid y_i)}[T(x_i, y_i)], \theta \right\rangle - A(\theta)$$

(in exponential families)

- recovers EM iteration $\theta^{(t+1)} = \operatorname{argmax}_\theta L(\theta, \theta^{(t)})$
  - E-step: evaluation of $L(\theta, \theta^{(t)})$ (or equivalently, the posterior mean)
  - M-step: maximization over θ (the MLE on "fake data")

2. $q(x^n) = q_\phi(x^n \mid y^n)$ indexed by ϕ.

$$\max_\theta \log p_\theta(y^n) \geq \max_\theta \max_\phi \mathbb{E}_{x^n \sim q_\phi(\cdot \mid y^n)}\left[\log \frac{p_\theta(x^n, y^n)}{q_\phi(x^n \mid y^n)}\right]$$

$$= \max_\theta \max_\phi \underbrace{-D_{KL}(q_\phi(x^n \mid y^n) \,\|\, p_\theta(x^n))}_{\text{typically admits closed-form expression by taking } p_\theta = N(0, I),\ q_\phi(\cdot \mid y^n) = N(\mu_\phi(y^n), \sigma_\phi^2(y^n))} + \mathbb{E}_{x^n \sim q_\phi(\cdot \mid y^n)}[\log p_\theta(y^n \mid x^n)]$$

Variational auto encoder: jointly maximize over (θ, ϕ) using SGD

- Gradient of $\mathbb{E}_{z \sim q_\phi}[f(z)]$ w.r.t. ϕ computed via

$$\nabla_\phi \mathbb{E}_{q_\phi}[f(z)] = \mathbb{E}_{q_\phi}[f(z) \nabla_\phi \log q_\phi(z)] \approx \frac{1}{L} \sum_{i=1}^{L} f(z^{(i)}) \nabla_\phi \log q_\phi(z^{(i)})$$

or $\nabla_\phi \mathbb{E}_{q_\phi}[f(z)] \approx \nabla_\phi \left(\frac{1}{L} \sum_{i=1}^{L} f(g_\phi(\varepsilon^{(i)})) \right)$ for $g_\phi(\varepsilon) \sim q_\phi$.