Lec 4: Exponential Families

9/6/19

Def. Let X be an m-dimensional random vector with sufficient statistics $\phi(x) \in \mathbb{R}^d$. An exponential family density with d-dimensional natural parameter vector η has the form:

$$p(x) = h(x) \exp\{\eta^T \phi(x) - A(\eta)\}$$

with base measure $h(x)$ and log-partition function:

$$A(\eta) = \log \int \exp\{\eta^T \phi(x)\} h(x)\, dx$$

Aside
$\sigma^{-1} = \exp(-\log \sigma)$
$= \exp(-\frac{1}{2}\log \sigma^2)$

Ex. Gaussian

$$p(x) = N(x \mid m, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}(x-m)^2 \sigma^{-2}\right\}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\sigma^{-2} + x\sigma^{-2}m - \frac{1}{2}m^2\sigma^{-2} - \frac{1}{2}\log\sigma^2\right\}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left\{\begin{pmatrix} \sigma^{-2}m \\ -\frac{1}{2}\sigma^{-2} \end{pmatrix}^T \begin{pmatrix} x \\ x^2 \end{pmatrix} - \frac{1}{2}m^2\sigma^{-2} - \frac{1}{2}\log\sigma^2\right\}$$

Where,

$$\eta = \begin{pmatrix} \sigma^{-2}m \\ -\frac{1}{2}\sigma^{-2} \end{pmatrix}, \quad \underbrace{h(x) = \frac{1}{\sqrt{2\pi}}}_{\text{Constant in } X}, \quad A(\eta) = \frac{1}{2}m^2\sigma^{-2} + 1/2\log\sigma^2$$

Mean Parameters

- Alternate d-dimensional set of mean parameters given by:

$$\mu_i = \mathbb{E}[\phi_i(x)] \quad \text{for } i = 1, \ldots, d$$

- $\eta(\mu)$ is an invertible mapping from mean to natural parameters
  - ↳ if minimal!!

Ex. Gaussian Parameters

$$\mu = \begin{pmatrix} \mathbb{E}[x] \\ \mathbb{E}[x^2] \end{pmatrix} = \begin{pmatrix} m \\ \sigma^2 + m^2 \end{pmatrix} \iff \eta(\mu) = \begin{pmatrix} \sigma^{-2} m \\ -\frac{1}{2}\sigma^{-2} \end{pmatrix}$$

- Aside: Gaussian mean params more conveniently written w/ variance:

$$\mu' = (\mathbb{E}[x], \mathbb{E}[x^2] - \mathbb{E}[x]^2)^T = (m, \sigma^2)^T$$

Def. An expfam $p_\eta(x)$ is minimal if both of the following hold:

a) Natural parameters are linearly independent,
b) Suff. stats are linearly independent

Def. An expfam $p_\eta(x)$ is overcomplete if it is not minimal

Ex. Overcomplete Categorical:

- Discrete RV $X \in \{1, \dots, k\}$, mean parameters μ s.t.

$$\{\mu \in \mathbb{R}^k : \mu_1 + \cdots + \mu_k = 1,\ \mu_j \geq 0\ \forall j = 1, \dots, k\}$$

$$p(x) = \prod_{j=1}^{k} \mu_j^{\mathbb{I}(x=j)} = \exp\left(\sum_j \mathbb{I}(x=j) \log \mu_j\right)$$

But,

$$\mu_k = 1 - \sum_{j=1}^{k-1} \mu_j \quad \& \quad \mathbb{I}(x=k) = 1 - \sum_{j=1}^{k-1} \mathbb{I}(x=j)$$

Minimal Parameterization:

$$p(x) = \exp\left\{ \sum_{j=1}^{k-1} \mathbb{I}(x=j) \log \mu_j + \left(1 - \sum_{j=1}^{k-1} \mathbb{I}(x=j)\right) \log\left(1 - \sum_{j=1}^{k-1} \mu_j\right) \right\}$$

EXPFAM PROPERTIES

⇒ Set of natural parameters $\eta \in H$ is convex:

Pf: Show for any $\eta_1, \eta_2 \in H$ & $\lambda \in [0,1]$:

$$\int \exp\left((\lambda \eta_1^T + (1-\lambda)\eta_2^T)\,\phi(x)\right) h(x)\,dx < \infty$$

We have:

$$\int \exp\left\{(\lambda \eta_1^T + (1-\lambda)\eta_2^T)\,\phi(x)\right\} h(x)\,dx \quad (1)$$

$$= \int e^{\lambda \eta_1^T \phi(x)}\, e^{(1-\lambda)\eta_2^T \phi(x)}\, h(x)\,dx$$

$$\leq \left(\int e^{\eta_1^T \phi(x)} h(x)\,dx\right)^{\lambda} \left(\int e^{\eta_2^T \phi(x)} h(x)\,dx\right)^{(1-\lambda)} \quad (2)$$

ASIDE: See gradient properties of $A(\eta)$ for alt. proof

(By Holder's inequality: $\int |f(z)g(z)|\,dz \leq \left(\int |f(z)|^{\frac{1}{\lambda}} dz\right)^{\lambda} \left(\int |g(z)|^{\frac{1}{1-\lambda}} dz\right)^{1-\lambda}$

$< \infty$ (since $\eta_1, \eta_2 \in H$)

□

⇒ Log-partition $A(\eta)$ is convex in η:

Pf: Exponentiating both sides of (1) ≤ (2) yields:

$$e^{A(\lambda \eta_1 + (1-\lambda)\eta_2)} \leq e^{\lambda A(\eta_1) + (1-\lambda) A(\eta_2)}$$

Taking the log yields:

$$A(\lambda \eta_1 + (1-\lambda)\eta_2) \leq \lambda A(\eta_1) + (1-\lambda) A(\eta_2)$$

ASIDE: exp and log are strictly increasing monotonic fun's

□

⇒ Closed under multiplication:

Pf: Let $X_i \overset{iid}{\sim} P_\eta(x)$ for $i = 1, \ldots, N$ and $P_\eta(\cdot)$ is expfam w/ suff. stats $\phi(x)$. The joint density is then,

$$p(x_1,\dots,x_N) = \prod_{i=1}^N p_\eta(x_i) = \prod_i h(x_i)\exp\left\{\eta^T\phi(x_i) - A(\eta)\right\}$$

$$= \left(\prod_i h(x_i)\right)\exp\left\{\eta^T\left(\sum_i \phi(x_i)\right) - nA(\eta)\right\}$$

↑ Add suff. stats. □

Pf: Let $p_{\eta_i}(x)$ be expfam w/ parameters $\eta_1,\dots,\eta_N$ then the product is a density in the expfam:

$$p(x) \propto \prod_{i=1}^N p_{\eta_i}(x) = h(x)^N \exp\left\{\phi(x)^T\left(\sum_i \eta_i\right) - \sum_i A(\eta_i)\right\}$$

w/ Natural parameters $\eta = \sum_i \eta_i$ □

⇒ Differentiating log-partition w.r.t natural parameters yields non-central moments:

$$\underbrace{\frac{\partial A}{\partial \eta_i} = \mathbb{E}\left[\phi_i(x)\right] = \mu_i}_{\text{General inverse mapping between moment-natural params.}}, \quad \frac{\partial^2 A}{\partial \eta_i \partial \eta_j} = \mathbb{E}\left[\phi_i(x)\phi_j(x)\right]$$

Pf: We just show the first derivative result,

$$\frac{\partial A}{\partial \eta_i} = \frac{\partial}{\partial \eta_i}\log\left(\int h(x)\exp\{\langle \eta, \phi(x)\rangle\}dx\right) = e^{-A(\eta)}\int \frac{\partial}{\partial \eta}\exp\{\langle \eta, \phi(x)\rangle\}$$

$$= \int \exp\left\{\langle \eta, \phi(x)\rangle - A(\eta)\right\}\phi_i(x)dx = \mathbb{E}\left[\phi_i(x)\right] = \mu_i$$

→ Note: 2nd derivative result also proves convexity of $A(\eta)$

Since $\frac{\partial^2 A}{\partial^2 \eta_i^2} - \left(\frac{\partial A}{\partial \eta_i}\right)^2 = \mathrm{VAR}(\phi_i(x)) \geq 0$

CONJUGACY:

Ex. Beta-Bernoulli:

- Consider $\theta \sim \text{Beta}(\alpha, \beta)$, $x|\theta \sim \text{Bernoulli}(x|\theta)$ for $x \in \{0,1\}$

$$p(\theta)p(x|\theta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\,\theta^{x}(1-\theta)^{(1-x)}$$

$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \theta^{\alpha-1+x}(1-\theta)^{(\beta+1-x)-1}$$

$$\propto \text{Beta}(\theta \mid \alpha+x, \beta+1-x)$$

$\Rightarrow$ All members of the exponential family have conjugate priors:

Pf: Let $\{x_i\}_{i=1}^N \overset{iid}{\sim} p(\cdot|\eta)$ be from an exponential family w/ PDF:

$$p(x_i|\eta) = h(x)\exp\{\eta^T \phi(x_i) - A(\eta)\}$$

Let $\eta \sim q(\eta|\lambda)$ have prior density,

$$q(\eta|\lambda) = g(\eta)\exp\{\lambda_1^T \eta - \lambda_2 A(\eta) - B(\lambda)\}$$

which is conjugate since,

$$p(\eta|x_1,\dots,x_N) \propto q(\eta|\lambda)\prod_{i=1}^N p(x_i|\eta)$$

$$= g(\eta)\exp\{\lambda_1^T \eta - \lambda_2 A(\eta) - B(\lambda)\} \cdot \prod_i h(x_i)\exp\{\eta^T \phi(x_i) - A(\eta)\}$$

$$\propto g(\eta)\exp\{\eta^T(\lambda_1 + \textstyle\sum_{i=1}^N \phi(x_i)) + (\lambda_2 + N)(-A(\eta))\}$$

$$\propto q\left(\eta \,\middle|\, \begin{pmatrix}\lambda_1 + \sum_i \phi(x_i) \\ \lambda_2 + N\end{pmatrix}\right)$$

ASIDE: Previously showed this is expfam due to closure

Information Theory:

(1) Kullback-Liebler divergence:

$$KL(p\|q) = H_p(q(x)) - H_p(x) = \mathbb{E}\left[\log \frac{p(x)}{q(x)}\right]$$

$H_p(q(x))$: cross entropy; $H_p(x)$: entropy

- Closest expfam $q(x)$ approx. of $p(x)$ in KL sense:

$$q^*(x) = \underset{q}{\operatorname{argmin}}\, KL(p\|q)$$

$$= \operatorname{argmin} H_p(q(x)) + \text{const.}$$

ASIDE: Channel coding interpretation: $KL(p\|q)$ is loss incurred if I code $x\sim p$ as $x\sim q$.

$$H_p(q(x)) = \mathbb{E}_p[-\log q(x)]$$

$$= \mathbb{E}_p[-\log h(x) - \eta^T \phi(x) + A(\eta)]$$

- Take derivative wrt η_i, set to zero, solve:

ASIDE: Recall $\frac{\partial A}{\partial \eta_i} = \mathbb{E}_q[\phi_i(x)]$

$$\frac{\partial H}{\partial \eta_i} = -\mathbb{E}_p[\phi_i(x)] + \mathbb{E}_{q_\eta}[\phi_i(x)] = 0 \Rightarrow$$

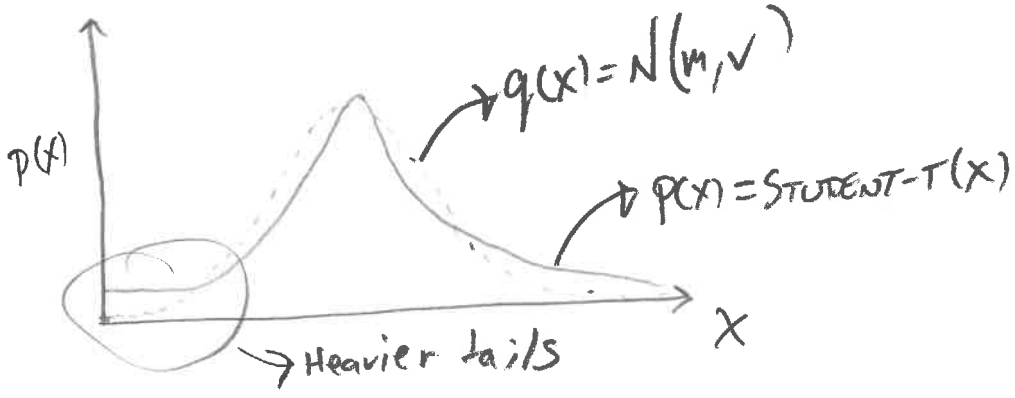
$$\mathbb{E}_{q_\eta}[\phi_i(x)] = \mathbb{E}_p[\phi_i(x)]$$

- This is the well-known "moment-matching" property

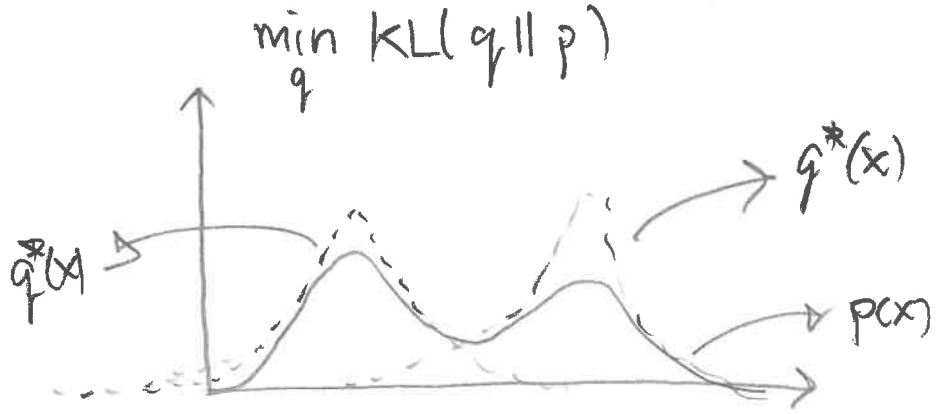
- Note: The sol'n is unique since:

$$\nabla_\eta^2 H = \nabla_\eta^2 A(\eta) = \mathbb{E}_q[\phi(x)\phi(x)^T] \succeq 0 \quad \text{(PSD)}$$

Ex. Gaussian Projection



② Reverse KL projection is non-convex in general:



- $KL(q\|p)$ called "Exclusive KL" because there is high penalty when $q(x)$ high in regions $p(x)$ low
- $\min KL(q\|p)$ tends to capture modes of $p(x)$
- We will see that $KL(q\|p)$ important in variational inference

③ Maximum entropy principle

Thm. Given RV $x \sim p$ and statistics $\{\phi_i(x)\}_{i=1}^d$ w/ known expectations $E_p[\phi_i(x)] = \mu_i$, the dist. $q(x)$ of maximum entropy $q(x) = \max H_q(x)$, subject to moment constraints, is in the expfam w/ suff. Stats. $\{\phi_i(x)\}_{i=1}^d$.

Pf Overview: Can be proven by forming Lagrangian w/ Lagrange multipliers $\vec{\eta}$ for moment constraints. Problem is convex and Sol'n is $q_\eta(x)$

↳ Lagrange multipliers are natural params