Lecture 18 - Exp. family and estimation in PGM

Friday, November 11, 2022 2:52 PM

today: exp. families
estimation in PGM

Exponential family

a (flat / canonical) exponential family on $\mathcal{X}$
is a parametric family of dist. on $\mathcal{X}$ defined by two quantities

I) $\underbrace{h(x)}_{\text{reference density}}\underbrace{d\mu(x)}_{\text{base measure}} \longrightarrow$ reference measure

base measure: counting measure (discrete R.V.) $\rightarrow \sum_x$ pmf; Lebesgue " (cts. R.V.) $\rightarrow \int_x$ pdf

II) $T: \mathcal{X} \rightarrow \mathbb{R}^p$ called "sufficient statistics" vector
aka feature vector

members of the family will have pmf/pdf

$$p(x;\eta)\,d\mu(x) = \exp(\underbrace{\eta}_{\text{"canonical" parameter}}{}^T \boxed{T(x)} - \underbrace{A(\eta)}_{\text{log-normalization or cumulant generating function or log partition fct.}})\boxed{h(x)\,d\mu(x)}$$

defining pieces $(+\Omega_X)$

if Ω_X is discrete; then $p(x;\eta)$ is a pmf
" " cts. " " " pdf

* want $1 = \int_{\mathcal{X}} p(x;\eta)\,d\mu(x) = \int_{\mathcal{X}} \exp(\eta^T T(x))\, e^{-A(\eta)}\, h(x)\,d\mu(x)$

discrete: $\left[\sum_{x \in \Omega_X} p(x;\eta)\right]$

$$\Rightarrow \boxed{A(\eta) \triangleq \log\Big(\underbrace{\int_{\mathcal{X}} \exp(\eta^T T(x))\,h(x)\,d\mu(x)}_{Z(\eta)}\Big)}$$

$\boxed{\text{domain } \Omega \triangleq \{\eta \in \mathbb{R}^p \mid A(\eta) < \infty\}}$
set of valid canonical parameters

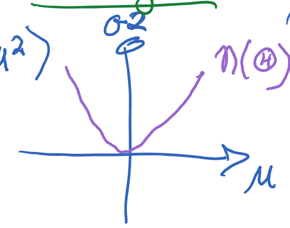
note: $A(\eta)$ is convex in $\eta \Rightarrow \Omega$ is convex

⊕ more generally, consider a reparameterization of a subset of the flat family
by defining $\eta: \underbrace{\Theta}_{\text{new set of parameters}} \rightarrow \Omega$

$$p(x;\theta) \triangleq p(x;\eta(\theta)) \quad \text{for } \theta \in \Theta$$

(get "curved exponential family" if $\eta(\Theta)$ is a curved manifold in Ω)

↳ e.g. could consider Gaussians where $N(\mu, \mu^2)$



* note: any single dist. $p(x)$ can be put in an exponential family by using $h(x) = p(x)$

* two examples of family not an exp. family:
- $\text{unif}(0, \theta)$
- mixture of Gaussians (latent variable model)

Example 1: (multinoulli)

$X \sim \text{Mult}(\pi)$ $\quad \mathcal{X} = \{0,1\}^K$

$\Omega_X = \Delta_K \cap \mathcal{X}$ (one hot encodings)

parameter $\pi \in \Delta_K$; suppose $\pi_i > 0 \ \forall i$

$$p(x; \underset{\text{think as "}\theta\text{"}}{\pi}) = \prod_{j=1}^{K} \pi_j^{x_j} = \exp\left(\sum_{j=1}^{K} x_j \log \pi_j\right)$$

$$= \exp\left(\sum_{j=1}^{K} x_j \log \pi_j - 0\right)$$

we have $\eta_j(\pi) = \log \pi_j$

$T(x) = x$ $\qquad \Omega_X \subseteq \mathbb{R}^K$

$d\mu(x)$ = counting measure on $\mathcal{X}$

$h(x) = \mathbb{1}\{x \in \Omega_X\} = \mathbb{1}\{x \in \Delta_K \cap \mathcal{X}\}$

$\Theta = \text{int}(\Delta_K)$ $\quad A(\eta(\pi)) = 0 \ \forall \pi \in \Theta$

$\Theta \rightarrow$ dimension $K-1$

$\eta(\Theta) \rightarrow$ " "

$\Omega_X \rightarrow$ " K

we do not have a "minimal exp family"

note: here, for any x s.t. $h(x) \neq 0$

here $\underbrace{\sum_{j=1}^{K} T_j(x)}_{\text{affine linear dep. between the components of } T} = \sum_j x_j = 1$

$\Rightarrow$ multiple η's give rise to same dist.

$\rightarrow$ "overparameterization"

↳ not a "minimal" exp. family

(*) for a Multinoulli, min. exp family

$$T(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_{K-1} \end{pmatrix} \qquad \left["x_K = 1 - \sum_{j=1}^{K-1} x_j" \right]$$

$$Z(\eta) = \sum_{x \in \Omega_X} \exp(\eta^T T(x)) = \sum_{j=1}^{K-1} e^{\eta_j} + 1$$

$$\boxed{p(x_j;\eta) = \exp\Big(\sum_{j=1}^{K-1} \eta_j x_j - \underbrace{\log\Big(\sum_{j=1}^{K-1} e^{\eta_j} + 1\Big)}_{A(\eta)} \Big)}$$

recall: $\nabla_\eta A(\eta) = \mathbb{E}_{p(x;\eta)}[T(x)]$ (valid for $\eta \in \text{int}(\Omega)$)

for multinoulli, $\frac{\partial A}{\partial \eta_j} = \frac{1}{Z(\eta)} e^{\eta_j} = p("x=j"|\eta)$

$= \mathbb{E}_{p(x;\eta)}[T_j(x)]$ as required //

moment matching can be different than MLE in exp. family

gamma dist $\Gamma(\alpha,\beta)$ has $T(x) = \begin{bmatrix} \log x \\ x \end{bmatrix}$

so moment matching with $\tilde{T}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$ will yield different estimate than MLE

16h07

example 2: 1d Gaussian

$X \sim N(\mu, \sigma^2)$ $\mathcal{X} = \mathbb{R}$ $\theta = (\mu, \sigma^2)$ "moment parameterization"

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

$$= \exp\Big(-\frac{x^2}{2} \underbrace{\left[\frac{1}{\sigma^2}\right]}_{\eta_2} + x \underbrace{\left[\frac{\mu}{\sigma^2}\right]}_{\eta_1} - \overbrace{\left[\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right]}^{A(\eta)} \Big)$$

$$T(x) = \begin{bmatrix} x \\ -\frac{x^2}{2} \end{bmatrix} \qquad \eta(\theta) = \begin{bmatrix} \mu/\sigma^2 \\ 1/\sigma^2 \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}$$

$\eta_2 = \frac{1}{\sigma^2}$ = precision > 0

$\eta_1 = \eta_2 \cdot \mu$

$h(x) = 1$ (but some people use $h(x) = \frac{1}{\sqrt{2\pi}}$ for Gaussian)

σ^2 $\qquad$ $h(x)=1$ $\qquad$ (but some people use $h(x)=\frac{1}{\sqrt{2\pi}}$ for Gaussian)

$$\eta_1 = \eta_2 \cdot \mu$$

$$\underbrace{\Omega = \{(\eta_1,\eta_2) : \eta_2 > 0, \eta_1 \in \mathbb{R}\}}$$

canonical parameter

[we'll see later : multivariate Gaussian $T(x) = \begin{bmatrix} x \\ -\frac{xx^T}{2} \end{bmatrix}$ $\begin{cases} \eta = \Lambda\mu = \Sigma^{-1}\mu \\ \Lambda = \Sigma^{-1} \end{cases}$]

<u>example 3</u>: discrete UGM

let $p \in \mathcal{L}(G)$ $\quad$ G is undirected

with $\psi_C(x) > 0 \ \forall C, x_C$

$$p(x) = \frac{1}{Z}\prod_{C\in\mathcal{C}} \psi_C(x_C) = \exp\left(\sum_C \log \psi_C(x_C) - \log Z\right)$$

$$= \exp\left(\sum_{C\in\mathcal{C}} \sum_{y_C\in\mathcal{X}_C} \underbrace{\mathbb{1}\{y_C = x_C\}}_{T_{C,y_C}(x)} \underbrace{\log \psi_C(y_C)}_{\eta_{C,y_C}} - \log Z\right)$$

$$T(x) = \begin{pmatrix} \vdots \\ \mathbb{1}\{x_C = y_C\} \end{pmatrix} \leftarrow \begin{matrix} y_C \in \mathcal{X}_C \\ C \in \mathcal{C} \end{matrix} \qquad \mathcal{X}_C = \{(y_i)_{i\in C} : y_i \in \mathcal{X}_i\}$$

$$\eta(\theta) = \begin{pmatrix} \vdots \\ \log \psi_C(y_C) \\ \vdots \end{pmatrix} \leftarrow \begin{matrix} y_C \in \mathcal{X}_C \\ C \in \mathcal{C} \end{matrix} \qquad \text{[not a minimal representation]}$$

<u>notes</u>: a) $\text{Mult}(\pi)$ is a special case where have <u>complete graph</u> (1 big clique)

b) <u>feature perspective</u>: instead of using <u>all</u> possible indicators $\mathbb{1}\{y_C = x_C\}$ you could use a <u>subset</u> for a task

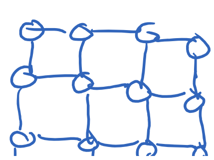
for example: suppose x is a sentence, x_i is a word

feature on x_i & x_{i+1} e.g. $\mathbb{1}\{x_i \text{ is a verb}, x_{i+1} \text{ is a noun}\}$

$\rightsquigarrow$ much smaller set of parameters

c) <u>binary Ising model</u>

$x_i \in \{0,1\}$ $\qquad$ $|C| \leq 2$



suppose use nodes & pairs (edges) as cliques

$\Rightarrow$ dimension of $T(x)$ $\quad 2|V| + 4|E|$ $\rightsquigarrow$ "overparameterized exp family"

$\sum_{y_c} T_{c, y_c}(x) = 1$ for any C

$\Rightarrow$ not a min. exp family

$*$ a minimal representation

is $T(x) = \begin{pmatrix} (x_i)_{i \in V} \\ (x_i x_j)_{\{i,j\} \in E} \end{pmatrix} \begin{matrix} \eta_i \\ \eta_{ij} \end{matrix}$ $\rightarrow$ dim: $|V| + |E|$

$x_i x_j \rightsquigarrow \mathbb{1}\{x_i = 1, x_j = 1\}$

$\mathbb{E}\,T(x) = \begin{pmatrix} (\mu_i)_{i \in V} \\ (\mu_{ij})_{\{i,j\} \in E} \end{pmatrix}$

$x_j \backslash x_i$	0	1	
0	$1 - \mu_i - \mu_j + \mu_{ij}$	$\mu_i - \mu_{ij}$	
1	$\mu_j - \mu_{ij}$	μ_{ij}	μ_j
		μ_i	

$\rightarrow$ 3 numbers are sufficient to parameterize

properties of A: (for generic flat exp. family)

- $\nabla_\eta A(\eta) = \mathbb{E}_{p(x;\eta)}[T(x)] \triangleq \mu(\eta)$ "moment vector" (for $\eta \in \operatorname{int}(\Omega)$)
- $\left(\dfrac{\partial^2 A(\eta)}{\partial \eta_i \partial \eta_j}\right)_{ij} = \mathbb{E}_{p(x;\eta)}\left[[T(x) - \mu(\eta)][T(x) - \mu(\eta)]^T\right] = \operatorname{Cov}(T(X))$

(proof as exercise)

"cumulant generating fct."

Estimation of parameters DGM/UGM

DGM (fully observed)

parametric family $\mathcal{P}_\Theta = \{ p_\theta(x) = \prod_i p(x_i | x_{\pi_i}, \theta_i) :$

$\theta = (\theta_1, \dots, \theta_{|V|})$

$\in \Theta = \Theta_1 \times \Theta_2 \cdots \times \Theta_{|V|} \}$

independent parameterization i.e. no tying of parameters

$\Rightarrow$ MLE decouples in $|V|$ independent MLE problems

$\{x^{(i)}\}_{i=1}^n \quad p(\text{data} | \theta) = \prod_{i=1}^n p(x^{(i)} | \theta) = \prod_{i=1}^n \prod_{j=1}^{|V|} p(x_j^{(i)} | x_{\pi_j}^{(i)}; \theta_j)$

$\log(\quad) = \sum_{j=1}^{|V|} \underbrace{\left(\sum_{i=1}^n \log p(x_j^{(i)} | x_{\pi_j}^{(i)}; \theta_j) \right)}_{\ell_j(\theta_j)}$

$f_j(\theta_j)$

example: for discrete R.V. (multinoulli conditionals) $\Rightarrow \theta_j^{MLE}$ = proportion of observations

$$\hat{p}(x_j = k \mid x_{\pi_j} = \text{stuff}) = \frac{\#(x_j = k, x_{\pi_j} = \text{stuff})}{\#(x_{\pi_j} = \text{stuff})}$$

(fully observed DGM is relatively easy)
often closed form?

⊗ if have latent variables (ie. unobserved variables)

$\Longrightarrow$ use E.M. (like HMM)

UGM :

example for exp. family

$$p(x;\eta) = \exp\left(\sum_c \langle \eta_c, T_c(x_c) \rangle - A(\eta)\right)$$

$\rightarrow$ unlike in a DGM, $\log p(x;\eta)$ does not separate as $\sum_c f_c(\eta_c)$

gradient ascent on log-likelihood

$$\frac{1}{n}\sum_{i=1}^{n} \log p(x^{(i)} \mid \eta) = \sum_c \eta_c^T \underbrace{\left[\frac{1}{n}\sum_{i=1}^{n} T_c(x_c^{(i)})\right]}_{\hat{\mu}_c} - \frac{\cancel{n} A(\eta)}{\cancel{n}}$$

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

$\hookrightarrow \mathbb{E}_{p(x;\eta)}[T_c(X_c)]$

to compute this, need inference

e.g. Ising model $T_{ij}(x_i, x_j) = x_i \cdot x_j$

$$\mathbb{E}[T_{ij}] = \mu_{ij} = p(x_i = 1, x_j = 1 \mid \eta)$$

here need approximate inference $\leftarrow$ sampling [e.g. Gibbs sampling]
variational [mean field]