Lecture 18 - Exp. family and estimation in PGM

Thursday, November 14, 2024 2:37 PM

today: · exp. family
· estimation for 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) $h(x)\,d\mu(x)$ → reference measure

reference density, base measure

→ counting measure (discrete R.V.) → $\sum_x$ pmf

→ Lebesgue " (cts. R.V.) → $\int_x$ pdf

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

members of the family will have pmf/pdf defining pieces ($+\Omega_X$)

$$p(x;\eta)\,d\mu(x) = \exp\left(\eta^T T(x) - A(\eta)\right) h(x)\,d\mu(x)$$

η: "canonical parameter"

$A(\eta)$: log-normalization or log partition fct. or cumulant generating function

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

$\mathbb{E}[f] = \int_x f(x)\,d\mu(x)$ ($p(x)\,dx$)

or $\sum_x f(x)\,\mu(x)$ if μ is discrete

(*) 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: $\sum_x p(x;\eta)$]

$$\Rightarrow \boxed{A(\eta) \triangleq \log\left(\int_{\mathcal{X}} \exp(\eta^T T(x))\, h(x)\,d\mu(x)\right)}$$

$Z(\eta)$

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

clash of notation: $\Omega \neq \Omega_X \subseteq \mathcal{X}$, $\Omega \subseteq \mathbb{R}^p$

(*) more generally, consider a reparameterization of a subset of the flat exponential family by defining $\eta: \underset{\text{new set}}{(H)} \to \Omega$, $H \subseteq \mathbb{R}^p$

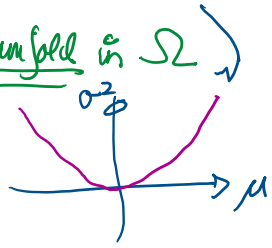
by defining $\eta : \Theta \to \mathcal{H}$

η: new set of parameters

$p(x;\theta) \triangleq p(x;\eta(\theta))$ for $\theta \in \Theta$

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

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



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

two examples of family not an exp family : • 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 $\pi_j > 0 \ \forall j$

$$p(x;\pi) = \prod_{j=1}^{k} \pi_j^{x_j} = \exp\left(\sum_{j=1}^{k} x_j \log \pi_j\right)$$

$x \in \mathcal{X}$ "think as θ"

$$= \exp\left(\sum_j x_j \log \pi_j - 0\right)$$

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

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

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

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

$A(\eta) = 0$??

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

$\mathcal{H} \to$ dim $k-1$

$\eta(\Theta) \to$ " "

$\Omega \subseteq \mathbb{R}^k \to$ dim k

we do not have a "minimal exp family"

we do not have a "minimal exp. family"

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

$$\sum_{j=1}^{K} T_j(x) = \sum_j x_j = 1$$

affine linear dep. between components of T

$\Rightarrow$ multiple η's give rise to same distribution $\rightarrow$ "over parameterization"

not a minimal exp. family

e.g. $(\eta + \mathbb{1}\alpha)^T T(x) = \eta^T T(x) + \alpha \underbrace{T(x)^T \mathbb{1}}_{1}$ ($\mathbb{1}$: vector $\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$)

(*) for a multinoulli; a min. exp. family

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

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

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

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

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

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

moment matching can be different than MLE in exp. family (with wrong moments!)

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

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

15h35

example 2: 1d Gaussian

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

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

$$= \exp\left(-\frac{x^2}{2}\left[\frac{1}{\sigma^2}\right] + x\left[\frac{\mu}{\sigma^2}\right] - \left[\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right]\right)$$

(η_2, η_1, $A(\eta)$)

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

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

$\eta_1 = \eta_2 \cdot \mu$

$\Omega = \{(\eta_1, \eta_2) : \eta_2 > 0, \eta_1 \in \mathbb{R}\}$ $h(x) = 1$ (but some people use $h(x) = \frac{1}{\sqrt{2\pi}}$)

[we'll see later: multivariate Gaussian $T(x) = \begin{bmatrix} x \\ -\frac{xx^T}{2} \end{bmatrix}$ for Gaussian "η_1" $= \Lambda\mu = \Sigma^{-1}\mu$ "η_2" $= \Lambda = \Sigma^{-1}$

Example 3: discrete UGM

Let $p \in \mathcal{G}(G)$ G is undirected with $\psi_C(x_C) > 0$ $\forall C, x_C$

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

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

$$T(x) = \left(\vdots \; \mathbb{1}\{x_C = y_C\} \; \vdots\right)_{\substack{y_C \in \mathcal{X}_C \\ C \in \mathcal{C}}} \qquad \mathcal{X}_C = \{(y_i)_{i \in C} : y_i \in \mathcal{X}_i\}$$

$$\eta(\theta) = \left(\vdots \; \log \psi_C(y_C) \; \vdots\right)_{\substack{y_C \in \mathcal{X}_C \\ C \in \mathcal{C}}}$$

[not a minimal representation]

notes: a) mult(π) is a special case where have ~~complete graph~~ (1 big clique)

b) ~~feature perspective~~: instead of using all possible indicators $\mathbb{1}\{y_C = x_C\}$ you could use a subset or a function of them for a task

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

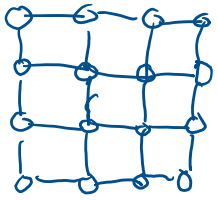
feature on x_i & x_{i+1} e.g. $\mathbb{1}\{x_i$ is a verb
x_{i-1} " " a noun$\}$

→ much smaller set of parameters

"parameter sharing"

c) binary Ising model

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



suppose use nodes & pairs (edges) as cliques

⇒ dimension of $T(x)$ $2|V| + 4|E|$ → "overparameterized exp. family"

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

⇒ not a min. exp fam

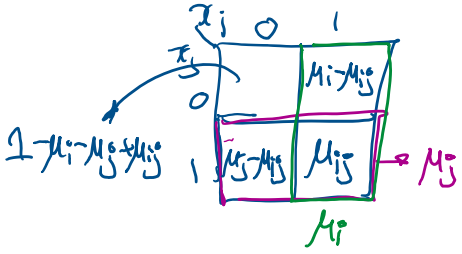
* a minimal representation

$T(x) = \begin{pmatrix} (x_i)_{i \in V} \\ (x_i x_j)_{\{i,j\} \in E} \end{pmatrix}$ $\begin{matrix} \} n_i \\ \} n_{ij} \end{matrix}$ → dim $|V| + |E|$

$(x_i x_j) \to \mathbb{1}\{x_i = 1, x_j = 1\}$

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

$\mu_{ij} \to p(x_i = 1, x_j = 1)$



properties of A (for generic $\mathcal{S}$ at exponential family)

- $\nabla_\eta A(\eta) = \mathbb{E}_{p(x;\eta)}[T(x)] \triangleq \mu(\eta)$ "moment vector" (for $\eta \in \text{int}(\Omega)$)
- $\left(\frac{\partial^2 A(\eta)}{\partial \eta_i \partial \eta_j}\right)_{(i,j) \in (1:p)^2} = \mathbb{E}_{p(x;\eta)}\left[(T(x) - \mu(\eta))(T(x) - \mu(\eta))^T\right] = \text{cov}(T(x))$

(proof as exercise)

"cumulant generating fct."

Estimation of parameters DGM/UGM

"independent ..."

Estimation of parameters DGM/UGM

DGM (fully observed)

parametric family $\mathcal{P}_\Theta = \{p_\theta(x) = \prod_i p(x_i \mid x_{\pi_i}, \theta_i)\}$; $(\theta_1, \dots, \theta_{|V|}) \in \Theta = \Theta_1 \times \Theta_2 \times \dots \Theta_{|V|}$

independent parameterization ⟹ i.e. no tying of parameters

⟹ MLE decouples in $|V|$ independent MLE problems:

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

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

example: for discrete R.V. (multinomial conditionals) ⟹ $\hat{\theta}_j^{MLE}$ = 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 in closed form?

⊗ if have latent variable (i.e. unobserved variables)

⟹ use E.M. (like HMMs)

UGM

example for exp. family

$$p(x; \eta) = \exp\left(\sum_C \langle \eta_C, T_C(x_C) \rangle - A(\eta) \right)$$

⟶ unlike in a DGM, $\log p(x, \eta)$ does not separate $\sum_C f_C(\eta_C)$

gradient ascent on log-likelihood

$$\frac{1}{n} \sum_{i=1}^n \log p(x^{(i)} \mid \eta) = \sum_C \eta_C^T \left[\frac{1}{n} \sum_{j=1}^n T_C(x^{(i)}) \right] - A(\eta)$$

$\left[\frac{1}{n} \sum_{j=1}^n T_C(x^{(i)}) \right] = \mu_C$

$\nabla_{\eta_C}[\ ] = \hat{\mu} - \mu_C(\eta)$

$\mu_C(\eta) = \mathbb{E}_{p(x;\eta)}[T_C(X_C)]$

to compute this, need inference

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

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

here need approximate inference

- sampling [Gibbs sampling]
- variational [mean field]