Lecture 18 - MaxEnt duality

November 9, 2021 4:34 PM

today : - Max. Ent. duality
- exponential families

dual problem for max. entropy

Max. ENT. in primal form

$$\min_{q \in \mathbb{R}^K} \; KL(q\|u) = \sum_x q(x) \log \frac{q(x)}{u(x)}$$

$$u(x) \triangleq \frac{1}{|\mathcal{X}|} = \frac{1}{K}$$

$$M: \quad q(x) \geq 0 \;\; \forall x \;\Big\} \; \Delta_{|\mathcal{X}|}, \qquad \sum_x q(x) = 1$$

absorb this constraint in domain of def. of $KL(q\|u)$

$$\sum_x q(x) T_j(x) = \alpha_j \;\; \forall j \quad \text{eq. constraints}$$

$$KL(q\|u) = \begin{cases} +\infty & \text{if } q(x) < 0 \text{ for some } x \\ KL(q\|u) & \text{o.w.} \end{cases}$$

$\mathbb{E}_q[T_j] = \alpha_j$, $\sum_x q(x) = 1$

$$\mathcal{L}(q, \nu, c) = \sum_x q(x) \log \frac{q(x)}{u(x)} + \sum_j \nu_j \left(\alpha_j - \sum_x q(x) T_j(x)\right) + c\left(1 - \sum_x q(x)\right)$$

(q: K-dim, ν: d-dim, c: scalar)

$$\frac{\partial \mathcal{L}}{\partial q(x)} = \log \frac{q(x)}{u(x)} + 1 - \sum_j \nu_j T_j(x) - c \overset{\text{want}}{=} 0$$

$$\nu^T T(x) = \sum_j \nu_j T_j(x)$$

$$\Rightarrow \boxed{q^*_{\nu,c}(x) = u(x) \exp\left(\nu^T T(x) + c - 1\right)}$$

generalized MaxENT: $KL(q\|h_0)$, $h_0(x)$ — exponential family?

dual fct.: plug $q^*_{\nu,c}$ in $\mathcal{L}(\dots)$

$$\tilde{\mathbb{E}}_q[f(x)] \triangleq \sum_x q(x) f(x)$$

$$g(\nu, c) = \mathcal{L}(q^*_{\nu,c}, \nu, c)$$

$$= \tilde{\mathbb{E}}_{q^*}[\nu^T T(x) + c - 1] + \nu^T \alpha - \tilde{\mathbb{E}}_{q^*}[\nu^T T(x)] \; c - \tilde{\mathbb{E}}_{q^*}[\alpha]$$

$$= \nu^T \alpha + c - \tilde{\mathbb{E}}_{q^*}[1]$$

$$\sum_x u(x) \exp(\nu^T T(x)) \exp(c-1) \triangleq Z(\nu)$$

$\max_{\nu,c} g(\nu, c)$ with respect to c

$$\nabla_c = 0 \Rightarrow 1 - Z(\nu) \exp(c-1) \overset{\text{want}}{=} 0$$

$$\Rightarrow \exp(c^* - 1) = \frac{1}{Z(\nu)}$$

plug back c^*: $\max_c g(\nu, c) = \nu^T \alpha + c^* - Z(\nu) \frac{1}{Z(\nu)}$

$$c^* - 1 \triangleq -\log Z(\nu)$$

dual problem $\max_\nu \tilde{g}(\nu)$ $\quad \boxed{\tilde{g}(\nu) \triangleq \nu^T \alpha - \log Z(\nu)}$

link with MLE :

if $\alpha = \frac{1}{n}\sum_i T(x^{(i)}) = \mathbb{E}_{\hat{p}_n}[T(X)]$

then $\tilde{g}(\nu) = \frac{1}{n}\sum_{i=1}^{n}\left[\nu^T T(x^{(i)}) - \log Z(\nu)\right]$

$\log p(x^{(i)}|\nu) + \text{cst.}$

where $p(x|\nu) \triangleq u(x)\exp\left[\nu^T T(x) - \log Z(\nu)\right]$

ie. dual problem is $\max_\nu \tilde{g}(\nu) = \max_\nu \frac{1}{n}\log p(x_{1:n}|\nu)$ ie. MLE

to summarize: ML in exp. family with $T(x)$ as suff. statistics is equivalent to Max ENT with moment constraints on $T(x)$ where $\alpha = \mathbb{E}_{\hat{p}_n}[T(X)]$

they are Lagrangian dual of each others?

MLE in exp. family $\Leftrightarrow$ moment matching in exp. family

note [direct derivation]:

$$\nabla_\nu \log Z(\nu) = \frac{1}{Z(\nu)}\nabla_\nu\left(\sum_x u(x)\exp(\nu^T T(x))\right)$$

$$= \sum_x T(x)\underbrace{\frac{u(x)\exp(\nu^T T(x))}{Z(\nu)}}_{p(x|\nu)}$$

$$\nabla_\nu \log Z(\nu) = \mathbb{E}_{p(x|\nu)}[T(X)] \triangleq \mu(\nu) \quad \text{"model moment"}$$

$$\nabla_\nu \tilde{g}(\nu) = (\alpha - \mu(\nu))$$

$\alpha = \mathbb{E}_{\hat{p}_n}[T(X)]$

$\hat{\mu}_n$ "empirical moment"

want $\nabla_\nu \tilde{g}(\nu) = 0 \Rightarrow \boxed{\mu(\nu^*) = \hat{\mu}_n}$ i.e. moment matching?

→ see lecture 16 in 2017 for "KL Pythagorean theorem"

(see end of old lecture 16 2017 for "KL Pythagorean theorem" and I-projection vs. M-projection for KL + geometry)

17h41

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

base measure ← counting measure (discrete R.V.) → $\sum_x$ pmf
base measure ← Lebesgue measure (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 $(+\Omega_X)$

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

"canonical parameter" · log normalization or log partition fct. or cumulant generating function

If Ω_X is discrete, then $p(x;\eta)$ is a pmf

" is cts. " " " " pdf

(for binary example $d\mu(x)$ is counting $\delta_{\{0\}} + \delta_{\{1\}}$)

* 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 $\left[\sum_x p(x;\eta)\right]$

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

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

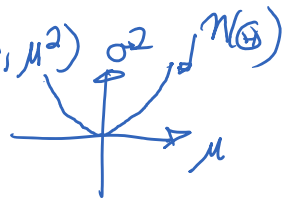
(*) more generally, consider a reparameterization of a subset of the flat family by defining $\eta: \Theta \to \Omega$

Θ = new set of parameters

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

(get "curved exponential family" if $\eta(\Theta)$ is a curved manifold in $\mathbb{R}^2$)

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



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

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

Example: (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 $\pi_j > 0 \ \forall j$

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

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

(think as "0")

we have $\eta_j(\pi) = \log \pi_j$ $\quad \Omega_X \subseteq \mathbb{R}^K$

$T(x) = x$

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

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

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

$\Theta \to$ dimension $K-1$

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

$\Omega_X \to$ " K

we do not have a "minimal exponential family"

note: for any x s.t. $h(x) \neq 0$ we have $\sum_{j=1}^{K} T_j(x) = \sum_j x_j = 1$

affine dependency between components of T

$\Rightarrow$ multiple η's give rise to same dist. ("overparameterization")

↳ not a "minimal" exp. family

(*) for multinoulli, minimal exp. family

$$T(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_{K-1} \end{pmatrix}$$

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

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

$A(\eta)$

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

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

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