Lec 2: Properties of Exponential Family

Yanjun Han

Sept 10, 2024

Def (Exponential family).

Let $\mathcal{Y}$ be the observation space. A class of probability distributions $(P_\theta)_{\theta \in \Theta}$ is called an exponential family iff

$$p_\theta(y) = \exp(\langle \theta, T(y) \rangle - A(\theta))\, h(y), \quad \forall \theta \in \Theta,\ y \in \mathcal{Y}.$$

Notations.
- $T(y) = (T_1(y), \cdots, T_d(y))$: sufficient statistic
- $\langle x, y \rangle = x_1 y_1 + \cdots + x_d y_d$ denotes the inner product
- $A(\theta)$: log-partition function
- $h(y)$: base measure

Intuition of exp. family :

$$p_\theta(y) = \exp(\langle \theta, T(y) \rangle) \times \text{function of } \theta \times \text{function of } y$$

" θ and y interact ONLY through $\langle \theta, T(y) \rangle$ in the exponent"

Examples.

1. Gaussian location family : $Y \sim N(\mu, 1)$.

$$p_\mu(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2}\right)$$
$$= \frac{1}{\sqrt{2\pi}} \exp\left(\mu y - \frac{\mu^2}{2} - \frac{y^2}{2}\right)$$

Correspondence to exp. family : $\theta = \mu$

$T(y) = y$

$A(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2}$

$h(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$

2. Gaussian location and scale family: $y \sim N(\mu, \sigma^2)$

$$p_{\mu,\sigma^2}(y) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$
$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2\sigma^2} + \frac{\mu y}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log\sigma\right)$$

Correspondence to exp. family:

$$\theta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$$
$$T(y) = (y, y^2)$$
$$A(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2)$$
$$h(y) = \frac{1}{\sqrt{2\pi}}$$

$\boxed{\text{Reparametrization may be necessary for an exp. family}}$

3. Bernoulli model: $y \sim \mathrm{Bern}(p)$, $p \in [0,1]$

$$p(y) = \begin{cases} p & \text{if } y=1 \\ 1-p & \text{if } y=0 \end{cases} = p^{\mathbb{1}(y=1)}(1-p)^{1-\mathbb{1}(y=1)}$$
$$= \exp\left(\mathbb{1}(y=1)\cdot\log\frac{p}{1-p} + \log(1-p)\right)$$

Correspondence to exp. family:

$$\theta = \log\frac{p}{1-p} \in (-\infty, +\infty)$$
$$T(y) = \mathbb{1}(y=1)$$
$$A(\theta) = -\log(1-p) = \log(1+e^\theta)$$
$$h(y) = 1.$$

4. Poisson model: $y \sim \mathrm{Poi}(\lambda)$, $\lambda > 0$

$$p_\lambda(y) = e^{-\lambda}\frac{\lambda^y}{y!}$$
$$= \exp(y\log\lambda - \lambda)\frac{1}{y!}$$

$$\begin{cases} \theta = \log\lambda \in (-\infty, +\infty) \\ T(y) = y \\ A(\theta) = \lambda = e^\theta \\ h(y) = \frac{1}{y!} \end{cases}$$

5. Multinomial model : $y \sim (p_1, \cdots, p_K)$: $p_i \geq 0$, $p_1 + \cdots + p_K = 1$.

$$p(y) = \begin{cases} p_1 & \text{if } y=1 \\ p_2 & \text{if } y=2 \\ \vdots & \\ p_K & \text{if } y=k \end{cases} = p_1^{1(y=1)} p_2^{1(y=2)} \cdots p_K^{1(y=k)}$$

$$= \left(\frac{e^{\theta_1}}{e^{\theta_1}+\cdots+e^{\theta_K}}\right)^{1(y=1)} \cdots \left(\frac{e^{\theta_K}}{e^{\theta_1}+\cdots+e^{\theta_K}}\right)^{1(y=k)}$$

(reparametrization : $p_j = \dfrac{e^{\theta_j}}{e^{\theta_1}+\cdots+e^{\theta_K}}$)

$$= \exp\left(\theta_1 1(y=1) + \cdots + \theta_K 1(y=k) - \log(e^{\theta_1}+\cdots+e^{\theta_K})\right)$$

Correspondence to exp. family : $\theta = (\theta_1, \cdots, \theta_K)$

$T(y) = (1(y=1), 1(y=2), \cdots, 1(y=k))$

$A(\theta) = \log(e^{\theta_1} + \cdots + e^{\theta_K})$

$h(y) = 1$.

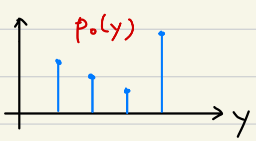
Modeling idea behind exponential family.

- $y \in \mathcal{Y}$: response variable, either discrete or continuous
- $p_0(y)$: a "base probability distribution" trying to fit $(y_1, \cdots, y_n)$
- $p_\theta(y)$: an "exponential tilting" of $p_0(y)$:

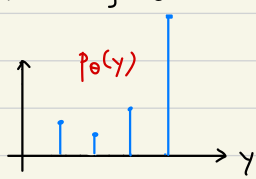
$$p_\theta(y) = \exp(\langle \theta, T(y) \rangle - A(\theta)) h(y)$$

$$\Rightarrow \frac{p_\theta(y)}{p_0(y)} = \exp(\langle \theta, T(y) \rangle - A(\theta) + A(0))$$

$$= \exp(\langle \theta, T(y) \rangle) \times \text{function of } \theta$$

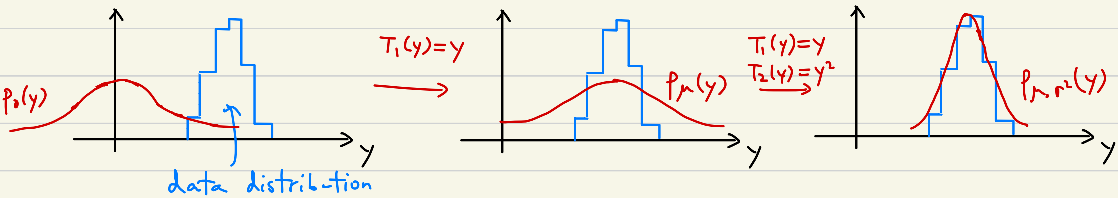


exp tilting



Exponential tilting maintains the support $\mathcal{Y}$, but changes the shape of the distribution.

- $T(y)$: the "quantity of interest" to be fit



- $A(\theta)$: normalization factor

y discrete: $\quad A(\theta) = \log \sum_{y \in \mathcal{Y}} \exp(\langle \theta, T(y) \rangle) h(y)$

y continuous: $\quad A(\theta) = \log \int_{\mathcal{Y}} \exp(\langle \theta, T(y) \rangle) h(y)\, dy$

- response with covariate: $(x_1, y_1), \cdots, (x_n, y_n)$

  modeling via exp. family: $\quad Y_i \sim p_{\theta_i}(y_i)$ with $\theta_i = \beta^T x$

  this is called "generalized linear model" (more in Lec 4)

  (GLM)

Examples of GLM:

1. Gaussian family. $\mathbb{E}[Y|X] = \beta^T x$ (linear regression)
2. Bernoulli family: $\log \frac{\mathbb{E}[Y|X]}{1 - \mathbb{E}[Y|X]} = \beta^T x$ (logistic regression)
3. Poisson family: $\log \mathbb{E}[Y|X] = \beta^T x$ (Poisson regression)

Properties of exp. family.

1. Mean. $\mu_\theta = \mathbb{E}_{p_\theta}[T(y)] = \sum_{y \in \mathcal{Y}} T(y) \cdot p_\theta(y)$

$$= \sum_{y \in \mathcal{Y}} T(y) \cdot \exp(\langle \theta, T(y) \rangle - A(\theta)) h(y)$$

$$\nabla A(\theta) = \nabla_\theta \log \left(\sum_{y \in \mathcal{Y}} \exp(\langle \theta, T(y) \rangle) h(y) \right)$$

$$= \frac{\sum_{y \in \mathcal{Y}} \nabla_\theta \exp(\langle \theta, T(y) \rangle) h(y)}{\sum_{y \in \mathcal{Y}} \exp(\langle \theta, T(y) \rangle) h(y)}$$

$$= e^{-A(\theta)} \sum_{y \in \mathcal{Y}} T(y) \exp(\langle \theta, T(y) \rangle) h(y)$$

$$= \sum_{y \in \mathcal{Y}} T(y) \exp(\langle \theta, T(y) \rangle - A(\theta)) h(y)$$

$$\boxed{\mathbb{E}_\theta[T(y)] = \nabla A(\theta)}$$

2. Covariance.

$$\nabla^2 A(\theta) = \nabla(\nabla A(\theta)) = \nabla_\theta \left(\frac{\sum_{y \in \mathcal{Y}} \nabla_\theta \exp(\langle \theta, T(y) \rangle) h(y)}{\sum_{y \in \mathcal{Y}} \exp(\langle \theta, T(y) \rangle) h(y)} \right)$$

$$= \frac{\sum_{y \in \mathcal{Y}} \nabla_\theta^2 \exp(\langle \theta, T(y) \rangle) h(y)}{\sum_{y \in \mathcal{Y}} \exp(\langle \theta, T(y) \rangle) h(y)}$$

$$- \left(\frac{\sum_{y \in \mathcal{Y}} \nabla_\theta \exp(\langle \theta, T(y) \rangle) h(y)}{\sum_{y \in \mathcal{Y}} \exp(\langle \theta, T(y) \rangle) h(y)} \right) \left(\frac{\sum_{y \in \mathcal{Y}} \nabla_\theta \exp(\langle \theta, T(y) \rangle) h(y)}{\sum_{y \in \mathcal{Y}} \exp(\langle \theta, T(y) \rangle) h(y)} \right)^T$$

$$\left(\nabla \frac{f}{g} = \frac{(\nabla f) g - f (\nabla g)}{g^2} \right)$$

$$= \sum_{y \in \mathcal{Y}} T(y) T(y)^T \underbrace{\exp(\langle \theta, T(y) \rangle - A(\theta)) h(y)}_{p_\theta(y)} - (\mathbb{E}_\theta T(y))(\mathbb{E}_\theta T(y))^T$$

$$\boxed{\mathrm{Cov}_\theta(T(y)) = \nabla^2 A(\theta)}$$

Corollary: $\nabla^2 A(\theta) \succeq 0$, so $A(\theta)$ convex in θ.

Note: the above corollary implies that

the correspondence $\theta \mapsto \mu_\theta = \mathbb{E}_\theta[T(y)]$ is one-to-one

Therefore, exp. families have two parametrizations:

parametrize by θ : natural parametrization

parametrize by μ_θ : mean parametrization

3. Repeated sampling.

$$y_1, y_2, \cdots, y_n \overset{\text{i.i.d.}}{\sim} p_\theta(y)$$

$$\Rightarrow p_\theta(y_1, \cdots, y_n) = p_\theta(y_1)\, p_\theta(y_2) \cdots p_\theta(y_n)$$

$$= \exp\left(\langle \theta, \sum_{i=1}^n T(y_i)\rangle - nA(\theta)\right) h(y_1) h(y_2) \cdots h(y_n)$$

$(y_1, \cdots, y_n)$ belongs to a new exponential family:

$$\begin{cases} \theta^{(n)} = n\theta \\ T^{(n)}(y_1, \cdots, y_n) = \frac{1}{n}\sum_{i=1}^n T(y_i) \\ A^{(n)}(\theta^{(n)}) = nA(\theta^{(n)}/n) \\ h^{(n)}(y_1, \cdots, y_n) = \prod_{i=1}^n h(y_i) \end{cases}$$

Note: As $\frac{1}{n}\sum_{i=1}^n T(y_i)$ is the sufficient statistic, for estimation/inference of θ, one may discard $(y_1, \cdots, y_n)$ and only keep $\frac{1}{n}\sum_{i=1}^n T(y_i)$

4. Conditioning.

If $\{p_\theta(y)\}_{\theta\in\Theta}$ is an exp. family, then for $y_0 \subseteq y$, the conditional probability distributions $\{p_\theta(y \mid y_0)\}_{\theta\in\Theta}$ is also an exp. family.

5. Conjugate prior.

A natural prior associated with exp. family:

$$\theta \sim \pi_{\xi,\tau}(\theta) = \exp(\langle \theta, \xi \rangle - \tau A(\theta))\, \underbrace{b(\xi,\tau)}_{\text{normalization factor}}$$

The posterior distribution of θ given y.

$$\pi_{\xi,\tau}(\theta \mid y) = \frac{\pi_{\xi,\tau}(\theta)\, p_\theta(y)}{p(y)}$$

marginal distribution of y: $p(y) = \int_\Theta \pi_{\xi,\tau}(\theta)\, p_\theta(y)\, d\theta$

$$= \frac{1}{p(y)} \exp(\langle \theta, \xi + T(y) \rangle - (\tau+1) A(\theta))\, h(y)\, b(\xi,\tau)$$

$$=: \exp(\langle \theta, \xi + T(y) \rangle - (\tau+1) A(\theta))\, b_y(\xi,\tau)$$

$$= \pi_{\xi+T(y),\tau+1}(\theta)$$

Under the conjugate prior, the posterior takes the same form as the prior, with

$$(\xi, \tau) \longmapsto (\xi + T(y), \tau+1)$$