# 09-Exponential Family and GLIMs

## Exponential Family

**Def** (exponential family): A distribution belonging to the exponential family has the following form

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x}) \exp\left\{ \boldsymbol{\eta}^T \mathbf{T}(\boldsymbol{x}) - A(\boldsymbol{\eta}) \right\}$$

where $$A(\boldsymbol{\eta}) = \log \int h(\boldsymbol{x}) \exp\left\{ \boldsymbol{\eta}^T \mathbf{T}(\boldsymbol{x}) \right\} d\boldsymbol{x}$$

is a log-normalizing factor. In other words,

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \frac{1}{Z} h(\boldsymbol{x}) \exp\left\{ \boldsymbol{\eta}^T \mathbf{T}(\boldsymbol{x}) \right\}$$

where $$Z = \int h(\boldsymbol{x}) \exp\left\{ \boldsymbol{\eta}^T \mathbf{T}(\boldsymbol{x}) \right\} d\boldsymbol{x}$$ [assuming convergent]

Here, $\boldsymbol{T}(\boldsymbol{X})$ are the sufficient statistics.

$\boldsymbol{\eta}$ are the parameters, and are called the natural parameters.

Basically, exponential family is log-linear in the sufficient statistics of $\boldsymbol{x}$, with some residual effect $h(\boldsymbol{x})$.

## A variety of common distributions are exponential family

- Gaussian

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

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

Thus $$\mathbf{T}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

$$\boldsymbol{\eta} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$$

Let $\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$. Then $\mu = -\frac{\eta_1}{2\eta_2}$, $\sigma^2 = -\frac{1}{2\eta_2}$

$$A(\boldsymbol{\eta}) = \frac{\mu^2}{2\sigma^2} + \log\sigma = -\frac{\eta_1^2}{4\eta_2} + \log\sqrt{-\frac{1}{2\eta_2}}$$

$$h(x) = \frac{1}{\sqrt{2\pi}}$$

- Bernoulli

$$p(x; \pi) = \pi^x (1-\pi)^{1-x}$$

$$= \exp\{x \log \pi + (1-x)\log(1-\pi)\}$$
$$= \exp\{[\log \pi - \log(1-\pi)]\cdot x + \log(1-\pi)\}$$

Thus, $T(x) = x$

$$\eta = \log \frac{\pi}{1-\pi} \quad \Rightarrow \pi = \frac{e^\eta}{1+e^\eta} = \frac{1}{1+e^{-\eta}}$$
$$A(\eta) = -\log(1-\pi) = \log(1+e^\eta)$$
$$h(x) = 1$$

For the Gaussian distribution, $\mu = \mathbb{E}[x]$, $\sigma^2 = \mathbb{E}[X^2]$
For the Bernoulli distribution, $\pi = \mathbb{E}[x]$

They are known as moment parameters, and have some correspondence with natural parameters.

$\psi: \mu \mapsto \eta \qquad \psi^{-1}: \eta \mapsto \mu$

**A few useful properties of exponential family**

- $A(\boldsymbol{\eta})$ is convex in natural parameters $\boldsymbol{\eta}$.
- $\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \mathbb{E}[\mathbf{T}(\boldsymbol{x})] \qquad \frac{\partial^2 A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^\top} = \mathrm{Var}[\mathbf{T}(\boldsymbol{x})]$
- Maximum likelihood estimation <=> Moment matching

$$\text{log-likelihood } l(\boldsymbol{\eta}) = \log \prod_{m=1}^{M}\left[h(\boldsymbol{x}^{(m)}) \exp\{\boldsymbol{\eta}^\top \mathbf{T}(\boldsymbol{x}^{(m)}) - A(\boldsymbol{\eta})\}\right]$$
$$= \sum_{m=1}^{M}\left[\log h(\boldsymbol{x}^{(m)})\right] + \boldsymbol{\eta}^\top \sum_{m=1}^{M} \mathbf{T}(\boldsymbol{x}^{(m)}) - M A(\boldsymbol{\eta})$$

Closed-form solution

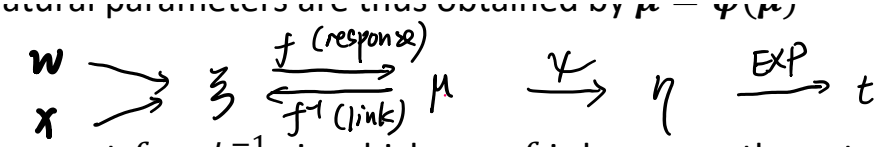
$$\frac{\partial l(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \sum_{m=1}^{M} \mathbf{T}(\boldsymbol{x}^{(m)}) - M \frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \overset{\text{set}}{=} \mathbf{0}$$
$$\Rightarrow \quad \frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{T}(\boldsymbol{x}^{(m)})$$
$$\Rightarrow \quad \hat{\boldsymbol{\mu}}_{MLE} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{T}(\boldsymbol{x}^{(m)})$$

**Generalized Linear Models**

In generalized linear models (GLM, GLIM), we assume

- $\boldsymbol{x}$ enters the model by linear (or affine) transformation, yielding $\xi = \boldsymbol{w}^\top \boldsymbol{x}$
- A response function $f$ transforms $\xi$ to the mean parameters $\mu$
- Natural parameters are thus obtained by $\boldsymbol{\mu} = \boldsymbol{\psi}(\boldsymbol{\mu})$

- Natural parameters are thus obtained by $\boldsymbol{\mu} = \boldsymbol{\psi}(\boldsymbol{\mu})$




Usually, we set $f = \psi^{-1}$, in which case $f$ is known as the natural response. The goal of prediction is to estimate the mean of $t$.

- For Gaussian distribution,

$$f(\boldsymbol{x}) = \mathbb{E}[t|\boldsymbol{x}] = \mu = \eta = \boldsymbol{w}^T\boldsymbol{x}$$

- For Bernoulli distribution,

$$f(\boldsymbol{x}) = \mathbb{E}[t|\boldsymbol{x}] = \mu = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}}$$

**MLE for GLIM with natural response**

Log-likelihood

$$\ell(\boldsymbol{w}, \boldsymbol{t}) = \sum_{m=1}^{M} \log h(t^{(m)}) + \sum_{m=1}^{M} \left( (\boldsymbol{w}^T\boldsymbol{x}^{(m)})\, t^{(m)} - A(\eta) \right)$$

$$\frac{\partial \ell}{\partial \boldsymbol{w}} = \sum_{m=1}^{M} \left[ \boldsymbol{x}^{(m)} y^{(m)} - \frac{\partial A(\eta^{(m)})}{\partial \eta^{(m)}} \cdot \frac{\partial \eta^{(m)}}{\partial \boldsymbol{w}} \right]$$

$$= \sum_{m=1}^{M} \left( y^{(m)} - \mu^{(m)} \right) \boldsymbol{x}^{(m)}$$

$$= \boldsymbol{X}^T(\boldsymbol{t} - \boldsymbol{y})$$

**Summary**

- We showed that exponential family is a useful family of distributions with nice properties.
- With further assumptions of generalized linear models and the natural response function, we may derive the formulas for linear regression, logistic regression, and softmax regression (HW). And they share the same formula of the derivative in MLE.
- It should be emphasized that GLIM and natural response that lead to the sigmoid or softmax is still an assumption. Other assumptions may also be acceptable and used.