Exponential Family

A family of distribution $\{P_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$ is said to be a k-parameter exponential family on $\mathbb{R}^q$, if there exist real valued functions:

- $\eta_1, \eta_2, \cdots, \eta_k$ and B of θ,
- T_1, $T_2, \cdots, T_k$, and h of $x \in \mathbb{R}^q$ such that the density function (pmf or pdf) of P_θ can be written as

$$p_\theta(x) = \exp[\sum_{i=1}^{k} \eta_i(\theta) T_i(x) - B(\theta)] h(x)$$

$$\mathbb{P}_\theta(x) = P(\theta, x)$$

$X \sim$ r.v

θ = Parameter.

* there is many way's θ, and x can interact.
* What the exponential Family does for you is that it restricts the way these thing's x, θ interact with each other.
* in Exponential family, θ, x interact is

$$P_\theta(x) = \exp(\underbrace{\theta \times X}_{\text{Product}})\, f(x)\, g(\theta)$$

when $\theta \in \mathbb{R}$, $x \in \mathbb{R}$

for $\theta \in \mathbb{R}^k$, $x \in \mathbb{R}^q$, we cannot Just do $\theta \times X$, we can do is

$$\theta \times X \longrightarrow \begin{array}{cc} \eta & T \\ \begin{bmatrix} \eta_1(\theta) \\ \eta_2(\theta) \\ \vdots \\ \eta_k(\theta) \end{bmatrix} \times & \begin{bmatrix} T_1(x) \\ T_2(x) \\ \vdots \\ T_k(x) \end{bmatrix} \end{array}$$

$$\longrightarrow \langle \eta, T \rangle = \sum_{j=1}^{k} \eta_j(\theta)\, T_j(x)$$

$$\mathbb{P}_\theta(x) = \exp\left[\sum_{j=1}^{k} \eta_j(\theta) T_j(x)\right] c(\theta) h(x)$$

$$= \exp\left[\sum_{j=1}^{k} \eta_j(\theta) T_j(x)\right] \cdot \exp\left(-\log\left(\frac{1}{c(\theta)}\right)\right) h(x)$$

$$= \exp\left[\sum_{j=1}^{k} \eta_j(\theta) T_j(x) - \log\left(\frac{1}{c(\theta)}\right)\right] h(x)$$

$$\log\left(\frac{1}{c(\theta)}\right) = B(\theta)$$

$$= \exp\left(\sum_{j=1}^{k} \eta_j(\theta) T_i(x) - B(\theta)\right) h(x)$$

$B(\theta)$:- in many way's normalizing constant.

Normal distribution example

- Consider $X \sim \mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$. The density is

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

  which forms a two-parameter exponential family with

$$\eta_1 = \frac{\mu}{\sigma^2},\ \eta_2 = -\frac{1}{2\sigma^2},\ T_1(x) = x,\ T_2(x) = x^2,$$

$$B(\theta) = \frac{\mu^2}{2\sigma^2} + \log(\sigma\sqrt{2\pi}),\ h(x) = 1.$$

- When σ^2 is known, it becomes a one-parameter exponential family on $\mathbb{R}$:

$$\eta = \frac{\mu}{\sigma^2},\ T(x) = x,\ B(\theta) = \frac{\mu^2}{2\sigma^2},\ h(x) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}.$$

$X \sim N(\mu, \sigma^2)$ $\qquad \theta = (\mu, \sigma^2) \in \mathbb{R}^2$

$x \in \mathbb{R}$

$$\mathbb{P}_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\frac{1}{\sqrt{2\pi}\sigma} = e^{\log \frac{1}{\sqrt{2\pi}\sigma}} = \exp(-\log(\sigma\sqrt{2\pi}))$$

$$\Rightarrow \mathbb{P}_\theta(x) = \exp\left(-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi})\right)$$

$$= \exp\left[\underbrace{x^2}_{T_2(x)} \cdot \underbrace{-\frac{1}{2\sigma^2}}_{\eta_2(\theta)} + \underbrace{x}_{T_1(x)} \cdot \underbrace{\frac{\mu}{\sigma^2}}_{\eta_1(\theta)} - \underbrace{\left(\frac{\mu^2}{2\sigma^2} + \log(\sigma\sqrt{2\pi})\right)}_{B(\theta)}\right]$$

$h(x) = 1$

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

$$B(\theta) = \frac{\mu^2}{2\sigma^2} + \log(\sigma\sqrt{2\pi}) \quad ; \quad h(x) = 1$$

$$P_\theta(x) = \exp\left(\eta(\theta) \cdot T(x) - B(\theta)\right) h(x)$$

When σ^2 is known

$$\Rightarrow \mathbb{P}_\theta(x) = \exp\left(x^2 \cdot \underbrace{-\frac{1}{2\sigma^2}}_{\text{this is not function of } \theta} + x \cdot \frac{\mu}{2\sigma^2} - \left(\frac{\mu^2}{2\sigma^2} + \log(\sigma\sqrt{2\pi})\right)\right)$$

$$\mathbb{P}_\theta(x) = \exp\left[x \cdot \frac{\mu}{2\sigma^2} - \left(\frac{\mu^2}{2\sigma^2} + \log(\sigma\sqrt{2\pi})\right)\right] \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

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

$T(x) = x$; $\eta(\theta) = \dfrac{\mu}{2\sigma^2}$

$B(\theta) = \dfrac{\mu^2}{2\sigma^2}$; $h(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left(\dfrac{-x^2}{2\sigma^2}\right)$

Examples of discrete distributions

The following distributions form discrete exponential families of distributions with pmf

- Bernoulli(p): $p^x(1-p)^{1-x}$, $x \in \{0, 1\}$
- Poisson(λ): $\dfrac{\lambda^x}{x!}e^{-\lambda}$, $x = 0, 1, \ldots$.

for discrete r.v

$X \sim \text{Bernoulli}(p)$

$$f_X(x) = \mathbb{P}(X=x) = p^x (1-p)^{1-x}$$

$$\Rightarrow \mathbb{P}_p(x) = \exp\left(x \cdot \log p + (1-x) \log(1-p) \right)$$

$$= \exp\left[x \cdot \log p - x \cdot \log(1-p) - (-\log(1-p)) \right] \cdot 1$$

$$= \exp\left[x \cdot \log \frac{p}{1-p} - \log\left(\frac{1}{1-p}\right) \right] \cdot 1$$

$$T(x) = x \; ; \; \eta(\theta) = \frac{p}{1-p}$$

$$B(\theta) = \log\left(\frac{1}{1-p}\right) \; ; \; h(x) = 1$$

$X \sim \text{Poisson}(\lambda)$

$$\mathbb{P}_\lambda(x) = \frac{e^{-\lambda} \lambda^x}{x!} \qquad x = 0, 1, 2, \ldots$$

$$\mathbb{P}_\lambda(x) = \exp(-\lambda - x \log \lambda) \frac{1}{x!}$$

$$\mathbb{P}_\lambda(x) = \exp\left(x \cdot -\log\lambda - \lambda\right) \frac{1}{x!}$$

$T(x) = x \qquad \eta(\theta) = -\log\lambda$

$h(x) = \frac{1}{x!} \qquad B(\theta) = \lambda$

Examples of Continuous distributions

The following distributions form continuous exponential families of distributions with pdf:

- Gamma(a, b): $\dfrac{1}{\Gamma(a) b^a} x^{a-1} e^{-\frac{x}{b}}$;
  - above: a: shape parameter, b: scale parameter
  - reparametrize: $\mu = ab$: mean parameter

$$\frac{1}{\Gamma(a)} \left(\frac{a}{\mu}\right)^a x^{a-1} e^{-\frac{ax}{\mu}}.$$

- Inverse Gamma(α, β): $\dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$.
- Inverse Gaussian(μ, σ^2): $\sqrt{\dfrac{\sigma^2}{2\pi x^3}} e^{\frac{-\sigma^2 (x-\mu)^2}{2\mu^2 x}}$.

Others: Chi-square, Beta, Binomial, Negative binomial distributions.

Components of GLM

1. Random component:

$$Y \sim \text{some exponential family distribution}$$

2. Link: between the random and covariates:

$$g\Big(\mu(X)\Big) = X^\top \beta$$

where g called link function and $\mu(X) = \mathbb{E}(Y|X)$.

One-parameter canonical exponential family

- Canonical exponential family for $k = 1$, $y \in \mathbb{R}$

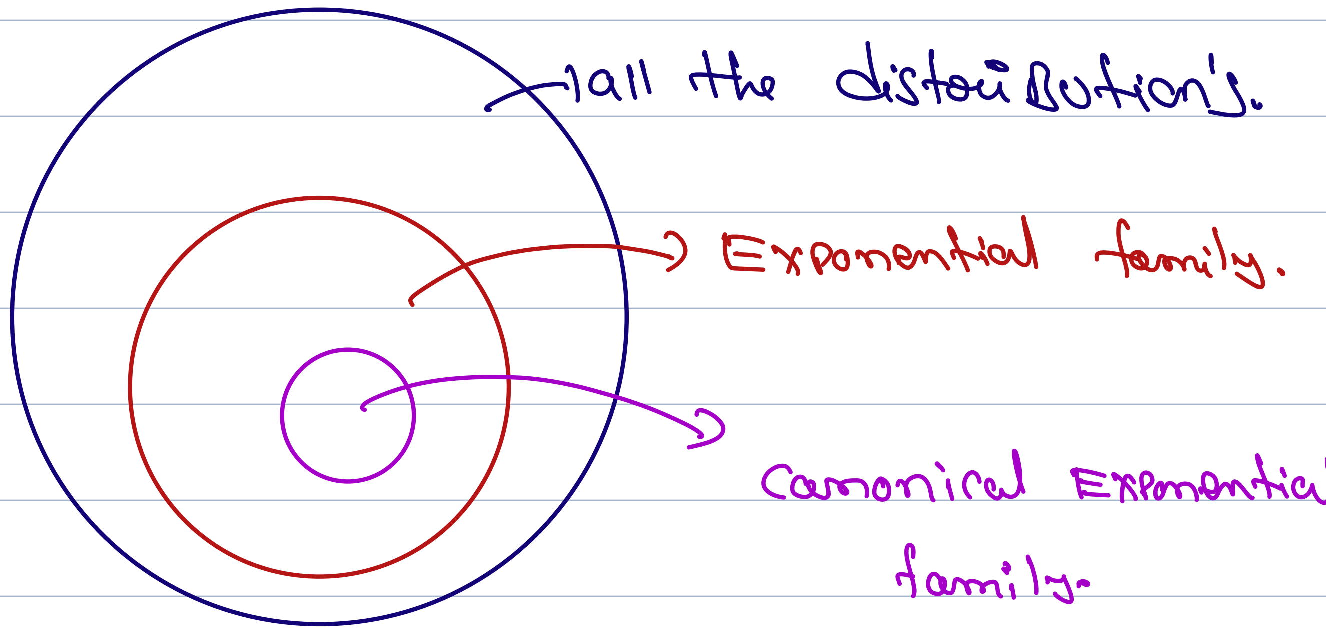
$$f_\theta(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

for some *known* functions $b(\cdot)$ and $c(\cdot, \cdot)$.

- If ϕ is known, this is a one-parameter exponential family with θ being the canonical parameter .
- If ϕ is unknown, this may/may not be a two-parameter exponential family. ϕ is called dispersion parameter.
- In this class, we always assume that ϕ is *known*.

in the canonical exponential family, what I have, is that I have $x = \theta$, and we have some normalization factor ϕ

ϕ is known



$$P_\theta(y) = \exp\left(\frac{y \cdot \theta - b(\theta)}{\phi} + c(y, \phi) \right)$$

if ϕ is known

then $h(y) = \exp(c(y, \phi))$

Normal distribution example

- Consider the following Normal density function with known variance σ^2,

$$
\begin{aligned}
f_\theta(y) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \\
&= \exp\left\{ \frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right) \right\},
\end{aligned}
$$

- Therefore $\theta = \mu, \phi = \sigma^2, \; , b(\theta) = \frac{\theta^2}{2}$, and

$$
c(y, \phi) = -\frac{1}{2}(\frac{y^2}{\phi} + \log(2\pi\phi)).
$$

Normal distribution with known variance

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

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

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

$\phi = 2\sigma^2$ known

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

$c(y,\phi) = -\frac{1}{2}\left(\frac{x^2}{\phi} + \log(2\pi\phi)\right)$

$b(\theta)$ what's gonna make the difference b/w Gaussian's, Bernoullis, Gamma, beta etc.. $b(\theta)$ contain's everything that idiosyncratic to the particular distribution.

$b(\theta)$:= Cummulant Generating function.

= log of MGF

Other distributions

Table 1: Exponential Family

	Normal	Poisson	Bernoulli
Notation	$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{P}(\mu)$	$\mathcal{B}(p)$
Range of y	$(-\infty, \infty)$	$[0, -\infty)$	$\{0, 1\}$
ϕ	σ^2	1	1
$b(\theta)$	$\frac{\theta^2}{2}$	e^θ	$\log(1 + e^\theta)$
$c(y, \phi)$	$-\frac{1}{2}(\frac{y^2}{\phi} + \log(2\pi\phi))$	$-\log y!$	1

Likelihood

Let $\ell(\theta) = \log f_\theta(Y)$ denote the log-likelihood function.
The mean $\mathbb{E}(Y)$ and the variance $\mathrm{var}(Y)$ can be derived from the following identities

- First identity
$$\mathbb{E}(\frac{\partial \ell}{\partial \theta}) = 0,$$
- Second identity
$$\mathbb{E}(\frac{\partial^2 \ell}{\partial \theta^2}) + \mathbb{E}(\frac{\partial \ell}{\partial \theta})^2 = 0.$$

Obtained from $\displaystyle\int f_\theta(y) dy \equiv 1\,.$

1st Identity:

$$l(\theta) = \log f_\theta(Y)$$

$$\frac{\partial}{\partial\theta} \log f_\theta(Y) = \frac{1}{f_\theta(Y)} \frac{\partial}{\partial\theta} f_\theta(Y)$$

$$\Rightarrow \mathbb{E}\left[\frac{\partial l}{\partial\theta}\right] = \int \frac{1}{\cancel{f_\theta(y)}} \frac{\partial}{\partial\theta} f_\theta(y) \cdot \cancel{f_\theta(y)} \cdot dy$$

$$= \int \frac{\partial}{\partial\theta} f_\theta(y)\, dy$$

$$= \frac{\partial}{\partial\theta} \int f_\theta(y)\, dy$$

$$= \frac{\partial}{\partial\theta} 1 = 0$$

$$\Rightarrow \mathbb{E}\left[\frac{\partial l}{\partial\theta}\right] = 0$$

$$\frac{\partial^2}{\partial\theta^2} \log f_\theta(x)$$

$$= \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta} \log f_\theta(x)\right]$$

$$= \frac{\partial}{\partial\theta}\left[\frac{1}{f_\theta(x)} \frac{\partial}{\partial\theta} f_\theta(x)\right]$$

$$= \frac{f_\theta(x)\left[\frac{\partial}{\partial\theta} f_\theta(x)\right]^2 - \left(\frac{\partial}{\partial\theta} f_\theta(x)\right)^2}{\left(f_\theta(x)\right)^2}$$

$$= \frac{1}{f_\theta(x)}\left[\frac{\partial}{\partial\theta} f_\theta(x)\right]^2 - \frac{1}{\left(f_\theta(x)\right)^2}\left[\frac{\partial}{\partial\theta} f_\theta(x)\right]^2$$

$$\mathbb{E}\left[\frac{\partial^2 l}{\partial\theta^2}\right] = \int \frac{\partial^2 l}{\partial\theta^2} \cdot f_\theta(x)\, dx$$

$$= \int \frac{\partial}{\partial\theta} f_\theta(x)\, dx - \int \frac{1}{f_\theta(x)}\left[\frac{\partial}{\partial\theta} f_\theta(x)\right]^2 dx$$

$$= \underbrace{\frac{\partial}{\partial\theta} \underbrace{\int f_\theta(x)\,dx}_{\text{Pdf}=1}}_{\frac{\partial}{\partial\theta}1=0} - \int \frac{1}{f_\theta(x)} \left[\frac{\partial}{\partial\theta} f_\theta(x)\right]^2 dx$$

$$= 0 - \int \frac{1}{f_\theta(x)} \left[\frac{\partial}{\partial\theta} f_\theta(x)\right]^2 dx$$

$$\boxed{\Rightarrow \mathbb{E}\left[\frac{\partial^2 l}{\partial\theta^2}\right] = -\int \frac{1}{f_\theta(x)} \cdot \left[\frac{\partial}{\partial\theta} f_\theta(x)\right]^2 dx}$$

$$\left(\frac{\partial l}{\partial\theta}\right)^2 = \left(\frac{\partial \log f_\theta(x)}{\partial\theta}\right]^2 = \left(\frac{1}{f_\theta(x)} \frac{\partial}{\partial\theta} f_\theta(x)\right)^2$$

$$\boxed{\Rightarrow \mathbb{E}\left[\left(\frac{\partial l}{\partial\theta}\right)^2\right] = \int \frac{1}{f_\theta(x)} \left(\frac{\partial}{\partial\theta} f_\theta(x)\right)^2 dx}$$

Identity 2

$$\boxed{\Rightarrow \mathbb{E}\left[\frac{\partial^2 l}{\partial\theta^2}\right] + \mathbb{E}\left[\left(\frac{\partial l}{\partial\theta}\right)^2\right] = 0}$$

Expected value

Note that

$$\ell(\theta) = \frac{Y\theta - b(\theta}{\phi} + c(Y;\phi),$$

Therefore

$$\frac{\partial \ell}{\partial \theta} = \frac{Y - b'(\theta)}{\phi}$$

It yields

$$0 = \mathbb{E}(\frac{\partial \ell}{\partial \theta}) = \frac{\mathbb{E}(Y) - b'(\theta))}{\phi},$$

which leads to

$$\mathbb{E}(Y) = \mu = b'(\theta).$$

Canonical Exponential Distribution

$$f_\theta(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

$$l(\theta) = \log f_\theta(y) = \frac{y\theta - b(\theta)}{\phi} + c(y,\phi)$$

1st Identity:

$$\mathbb{E}\left[\frac{\partial l}{\partial \theta}\right] = 0 \quad \Rightarrow \quad \frac{\partial l}{\partial \theta} = \frac{y - b'(\theta)}{\phi}$$

$$\mathbb{E}\left[\frac{\partial l}{\partial \theta}\right] = \mathbb{E}\left[\frac{y - b'(\theta)}{\phi}\right] = 0$$

$$\Rightarrow \frac{\mathbb{E}[y] - b'(\theta)}{\phi} = 0$$

$$\Rightarrow \quad \mathbb{E}[y] = b'(\theta)$$

Variance

On the other hand we have we have

$$\frac{\partial^2 \ell}{\partial \theta^2} + (\frac{\partial \ell}{\partial \theta})^2 = -\frac{b''(\theta)}{\phi} + \Big(\frac{Y - b'(\theta)}{\phi}\Big)^2$$

and from the previous result,

$$\frac{Y - b'(\theta)}{\phi} = \frac{Y - \mathbb{E}(Y)}{\phi}$$

Together, with the second identity, this yields

$$0 = -\frac{b''(\theta)}{\phi} + \frac{\operatorname{var}(Y)}{\phi^2},$$

which leads to

$$\operatorname{var}(Y) = V(Y) = b''(\theta)\phi.$$

Identity 2

$$\mathbb{E}\left\{ \frac{\partial^2 \ell}{\partial \theta^2} + \left[\frac{\partial \ell}{\partial \theta}\right]^2 \right\} = 0$$

$$\frac{\partial^2 \ell}{\partial \theta^2} = \frac{-b''(\theta)}{\phi} \qquad \left(\frac{\partial \ell}{\partial \theta}\right)^2 = \left(\frac{Y - b'(\theta)}{\phi}\right)^2$$

$$\mathbb{E}\left[\frac{-b''(\theta)}{\phi} + \left(\frac{Y - b'(\theta)}{\phi}\right)^2 \right] = 0$$

$$\Rightarrow \qquad \mathbb{E}\left[\left(\frac{Y - b'(\theta)}{\phi}\right)^2\right] = \frac{b''(\theta)}{\phi}$$

$$\mathbb{E}\left[(Y - b'(\theta))^2\right] = b''(\theta) \cdot \phi$$

$$\Rightarrow \qquad \mathrm{var}(Y) = b''(\theta) \cdot \phi$$

Example: Poisson distribution

Example: Consider a Poisson likelihood,

$$f(y) = \frac{\mu^y}{y!}e^{-\mu} = e^{y\log\mu - \mu - \log(y!)},$$

Thus,

$$\theta = \log\mu, \quad b(\theta) = \mu, \quad c(y,\phi) = -\log(y!),$$

$$\phi = 1,$$

$$\mu = e^{\theta},$$

$$b(\theta) = e^{\theta},$$

$$b''(\theta) = e^{\theta} = \mu,$$

$$f_{\mu}(y) = \frac{e^{-\mu} \mu^{y}}{y!}$$

$$= \exp\left(-\mu + y \log \mu - \log y!\right)$$

$$= \exp\left(\frac{y \cdot \log \mu - e^{\log \mu}}{1} - \log y!\right)$$

$\Rightarrow \quad \theta = \log \mu \quad \Rightarrow \mu = e^{\theta}$

$b(\theta) = e^{\theta}$

$c(y, \phi) = -\log(y!)$

$\phi = 1$

$b(\theta) = e^{\theta}$

$b'(\theta) = e^{\theta}$

$b''(\theta) = e^{\theta} =$

$$E[Y] = b'(\theta) = e^{\theta} = \mu$$

$$\mathrm{Var}(Y) = b''(\theta)\phi = e^{\theta} \cdot 1 = \mu$$

$\mathbb{E}[Y \mid X] = \mu(X) = (X^T \beta$ for Linear models$)$

in GLM

we are takin

$$g(\mu(X)) = X^T \beta$$

Link function

- β is the parameter of interest, and needs to appear somehow in the likelihood function to use maximum likelihood.
- A link function g relates the linear predictor $X^\top \beta$ to the mean parameter μ,
$$X^\top \beta = g(\mu).$$
- g is required to be monotone increasing and differentiable
$$\mu = g^{-1}(X^\top \beta).$$

$$\mathbb{E}[Y|X] = \mu(X)$$

$$g(\mu(X)) = X^T\beta$$

$\Rightarrow$ we want the link function $g(\cdot)$ to be countinuously differentiable

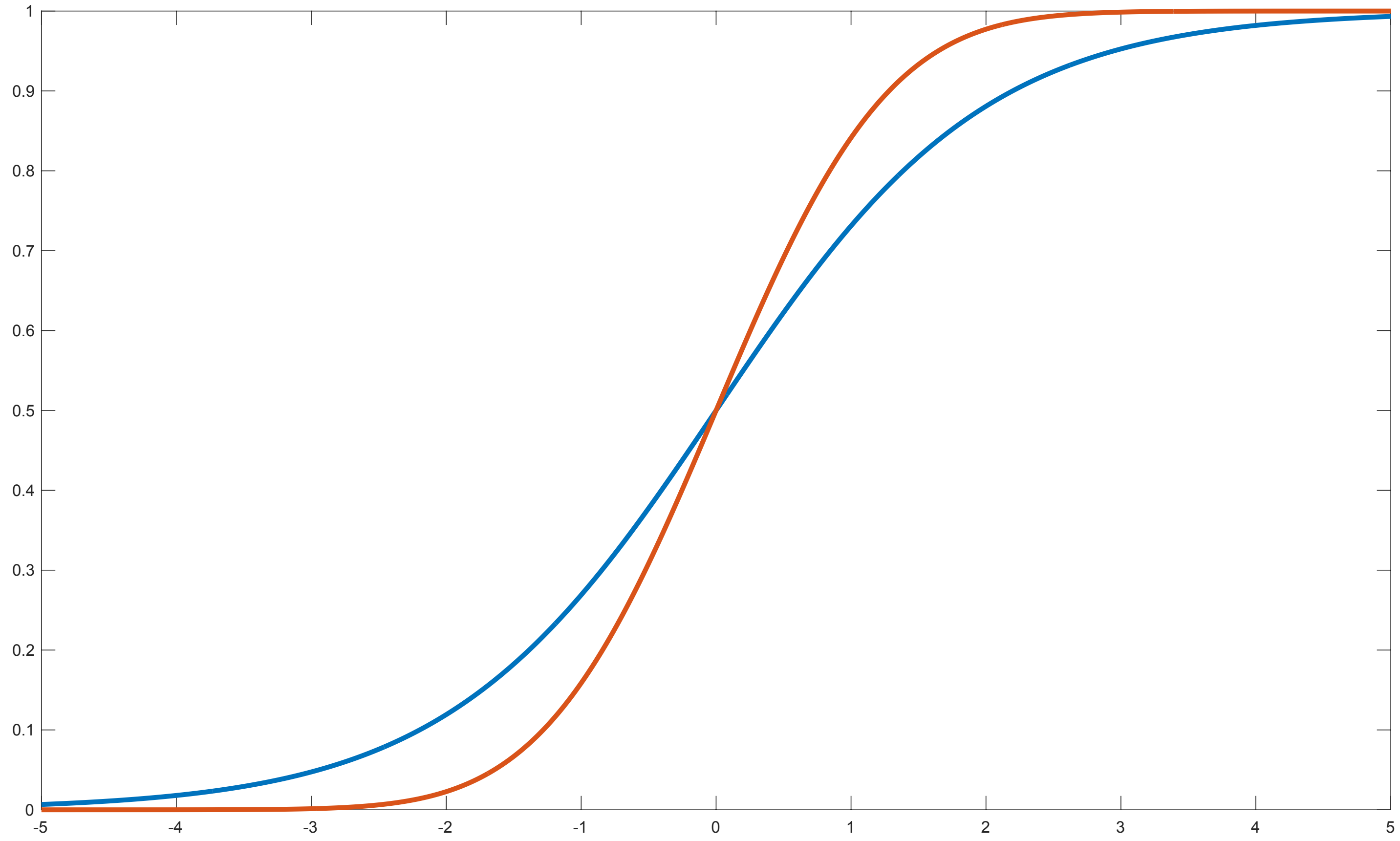
$\Rightarrow$ and $g(\cdot)$ to be strictly increasing.

$g(\mu(x))$ span's $\mathbb{R}$

Examples of link functions

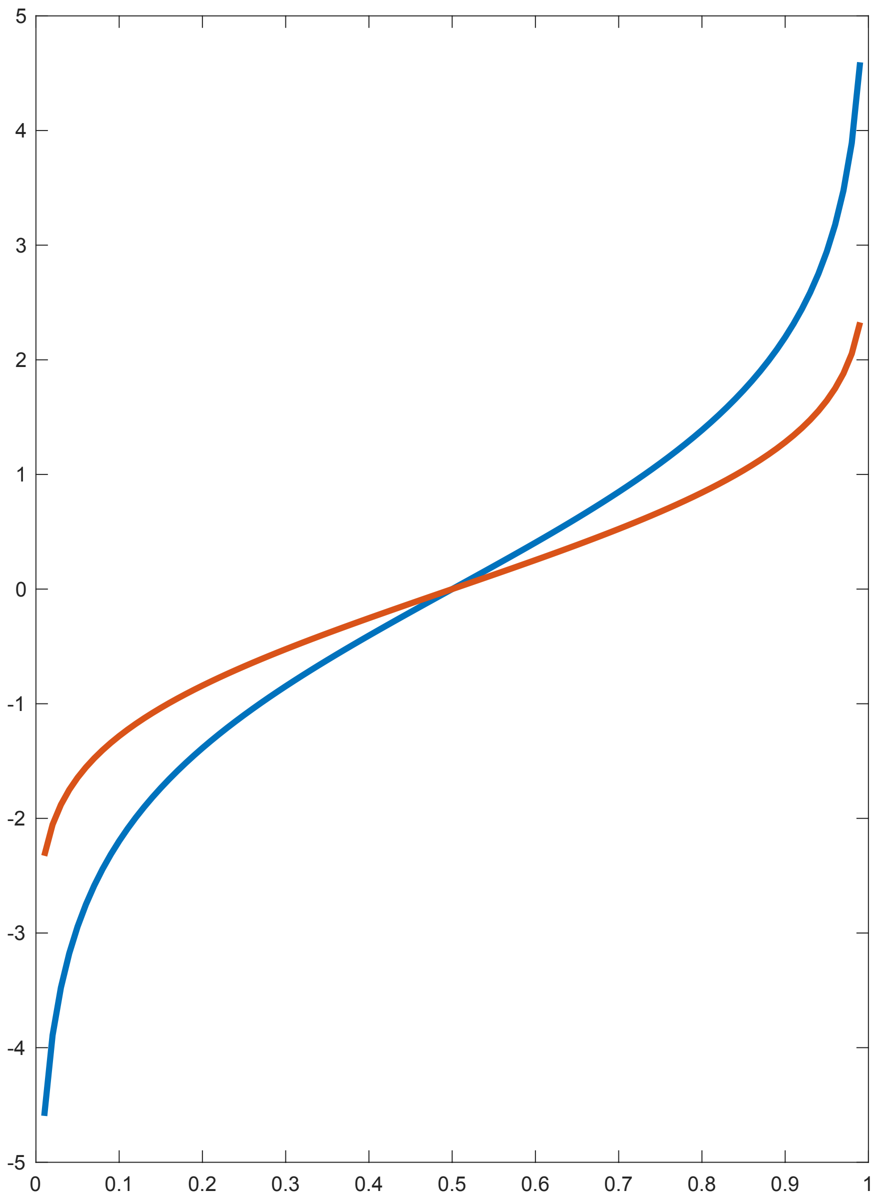
- For LM, $g(\cdot) =$ identity.
- Poisson data. Suppose $Y|X \sim \mathrm{Poisson}(\mu(X))$.
  - $\mu(X) > 0$;
  - $\log(\mu(X)) = X^\top \beta$;
  - In general, a link function for the count data should map $(0, +\infty)$ to $\mathbb{R}$.
  - The log link is a natural one.
- Bernoulli/Binomial data.
  - $0 < \mu < 1$;
  - g should map $(0, 1)$ to $\mathbb{R}$:
  - 3 choices:
    1. logit: $\log\left(\frac{\mu(X)}{1-\mu(X)}\right) = X^\top \beta$;
    2. probit: $\Phi^{-1}(\mu(X)) = X^\top \beta$ where $\Phi(\cdot)$ is the normal cdf;
    3. complementary log-log: $\log(-\log(1 - \mu(X))) = X^\top \beta$
  - The logit link is the natural choice.

Examples of link functions for Bernoulli response (1)



- in blue: $f_1(x) = \dfrac{e^x}{1 + e^x}$
- in red: $f_2(x) = \Phi(x)$ (Gaussian CDF)

Examples of link functions for Bernoulli response (2)



- in blue: $g_1(x) = f_1^{-1}(x) = \log \dfrac{x}{1-x}$ (logit link)
- in red: $g_2(x) = f_2^{-1}(x) = \Phi^{-1}(x)$ (probit link)

Canonical Link

- The function g that links the mean μ to the canonical parameter θ is called Canonical Link:

$$g(\mu) = \theta$$

- Since $\mu = b'(\theta)$, the canonical link is given by

$$g(\mu) = (b')^{-1}(\mu)\,.$$

- If $\phi > 0$, the canonical link function is strictly increasing. `Why?`

Example: the Bernoulli distribution

- We can check that

$$b(\theta) = \log(1 + e^{\theta})$$

- Hence we solve

$$b'(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)} = \mu \qquad \Leftrightarrow \qquad \theta = \log\left(\frac{\mu}{1-\mu}\right)$$

- The canonical link for the Bernoulli distribution is the logit link.

Other examples

	$b(\theta)$	$g(\mu)$
Normal	$\theta^2/2$	μ
Poisson	$\exp(\theta)$	$\log \mu$
Bernoulli	$\log(1 + e^\theta)$	$\log \frac{\mu}{1-\mu}$
Gamma	$-\log(-\theta)$	$-\frac{1}{\mu}$

Model and notation

- Let $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ be independent random pairs such that the conditional distribution of Y_i given $X_i = x_i$ has density in the canonical exponential family:

$$f_{\theta_i}(y_i) = \exp\Big\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\Big\}.$$

- $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, $\mathbb{X} = (X_1^\top, \ldots, X_n^\top)^\top$
- Here the mean μ_i is related to the canonical parameter θ_i via

$$\mu_i = b'(\theta_i)$$

- and μ_i depends linearly on the covariates through a link function g:

$$g(\mu_i) = X_i^\top \beta\,.$$

Back to β

- Given a link function g, note the following relationship between β and θ:

$$\begin{aligned} \theta_i &= (b')^{-1}(\mu_i) \\ &= (b')^{-1}(g^{-1}(X_i^\top \beta)) \equiv h(X_i^\top \beta), \end{aligned}$$

  where h is defined as

$$h = (b')^{-1} \circ g^{-1} = (g \circ b')^{-1}.$$

- Remark: if g is the canonical link function, h is identity.

Log-likelihood

- The log-likelihood is given by

$$\begin{aligned}\ell_n(\beta; \mathbf{Y}, \mathbb{X}) &= \sum_i \frac{Y_i\theta_i - b(\theta_i)}{\phi} \\ &= \sum_i \frac{Y_i h(X_i^\top \beta) - b(h(X_i^\top \beta))}{\phi}\end{aligned}$$

  up to a constant term.
- Note that when we use the canonical link function, we obtain the simpler expression

$$\ell_n(\beta, \phi; \mathbf{Y}, \mathbb{X}) = \sum_i \frac{Y_i X_i^\top \beta - b(X_i^\top \beta)}{\phi}$$

Strict concavity

- The log-likelihood $\ell(\theta)$ is strictly concave using the canonical function when $\phi > 0$. `Why?`
- As a consequence the maximum likelihood estimator is unique.
- On the other hand, if another parameterization is used, the likelihood function may not be strictly concave leading to several local maxima.