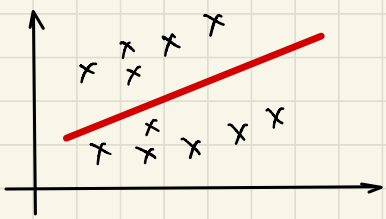
# Generalized linear models and exponential families.

**Motivation:**

- least squares:
  - $\min_w \frac{1}{2} \sum_{i=1}^{N} (y_i - x_i^T w)^2$
  - prob. model



linear model, gaussian Noise

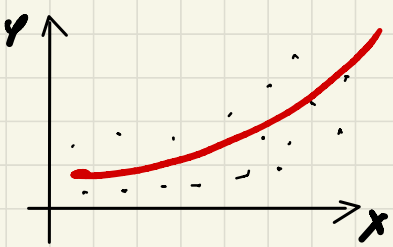
$$y = x^T w + \varepsilon; \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$y \sim \mathcal{N}(x^T w, \sigma^2 I)$$

$\rightarrow$ MLE $\rightarrow \hat{w}$

$\rightarrow$ very simple model

In general we will have more complicated data




- We can do:
  - feature augmentation $(x, x^2, x^3, \ldots)$
  - Richer class of probabilistic model (beyond Gaussian and logistic)

$\rightarrow$ General linear models
Expon. families.

# • Exponential family:

· Logistic regression:

$$\begin{cases} P(y=1|x,w) = \sigma(x^T w) = \dfrac{e^{\eta}}{1+e^{\eta}} \\ P(y=0|x,w) = 1-\sigma(x^T w) = \dfrac{1}{1+e^{\eta}} \end{cases} \quad , \quad \eta = w^T x = x^T w$$

Single form: $P(y|\eta) = \dfrac{e^{\eta y}}{1+e^{\eta}}$ , $\begin{cases} y=1 & \dfrac{e^{\eta}}{1+e^{\eta}} \\ y=0 & \dfrac{1}{1+e^{\eta}} \end{cases}$

$$P(y|\eta) = e^{[\eta y - \log(1+e^{\eta})]}$$

$$\boxed{P(y|\eta) = e^{\eta \Phi(y) - A(\eta)}}$$

exp. fam.

· $y$ is related to $\mu = \mathbb{E}[y]$ through a non-linear relation:

$$\boxed{\eta = \ln \frac{\mu}{1-\mu} \rightarrow \mu = \sigma(\eta)}$$

→ link function

# Exponential family:

$$\mathbb{P}(y|\eta) = h(y)\exp\left[\eta^T \phi(y) - A(\eta)\right]$$

- $y \in \mathbb{R}, \mathbb{R}_+, \{0,1\}, \mathbb{N}$, can be cont. or discr.
- $\eta$ : natural parameter
- $h(y) \geq 0$
- $\phi(y)$ : sufficient statistic
- $A(\eta)$ : cumulat. log partition

$$\int h(y)\exp(\eta^T\phi(y) - A(\eta))\,dy = 1$$

$$\Rightarrow A(\eta) = \log\left[\int h(y)\exp(\eta^T\phi(y))\,dy\right]$$

- $M := \{\eta \,;\, \int h(y)\exp(\eta^T\phi(y))\,dy < \infty\}$

  ↳ natural parameter space

$\Rightarrow$ degrees of freedom $(h, \phi, \eta)$

Ex: $y = \{0,1\}$

- Bernoulli distr. with par. $\nu$

$$\mathbb{P}(y|\nu) = \nu^y(1-\nu)^{1-y} = \left(\frac{\nu}{1-\nu}\right)^y \cdot (1-\nu)$$

$$= e^{\underbrace{\log\left(\frac{\nu}{1-\nu}\right)}_{\text{depends on } y}\, y + \underbrace{\log(1-\nu)}_{\text{indep. of } y}}$$

$$= e^{\left[\eta^T\phi(y) - A(\eta)\right]} \qquad (\nu = \sigma(\eta))$$

$\phi(y) = y$, $\eta = \ln\frac{\nu}{1-\nu}$ ; $h(y) = 1$, $A(\eta) = -\log(1-\nu) = \log(1+e^\eta)$

$\mu$ is the expected value of $y$:

$$\mu = \mathbb{E}(\phi(y))$$

(2 way to parametrize dist.: $\mu$ or $\eta$)

- Gaussian: $\mathcal{N}(\mu, \sigma^2)$

$$\mathbb{P}(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

- $y \in \mathbb{R}$
- We have two parameters $(\mu, \sigma^2)$

$$\mathbb{P}(y \mid \mu, \sigma^2) = e^{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}}$$

$$= e^{\left(\frac{\mu}{\sigma^2}; -\frac{1}{2\sigma^2}\right)\begin{pmatrix} y \\ y^2 \end{pmatrix} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)}$$

$$= e^{\eta^T \phi(y) - A(\eta)}$$

$$\eta = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix}; \quad h(y) = 1 \qquad A(\eta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)$$

$$\phi(y) = \begin{pmatrix} y \\ y^2 \end{pmatrix} \qquad = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log\left(-\frac{\pi}{\eta_2}\right)$$

$$A(\eta) = -\frac{\eta_1^2}{4\eta_2} - \log(\eta_2/\pi)$$

Link fct: $\eta_1 = \mu/\sigma^2 \quad \eta_2 = -1/2\sigma^2$

$$\mu = -\frac{\eta_1}{2\eta_2} \; ; \quad \sigma^2 = -\frac{1}{2\eta_2}$$

- Poisson with mean $\nu$. $y \in \mathbb{N}$

$$\mathbb{P}(y \mid \nu) = \frac{\nu^y}{y!} \cdot e^{-\nu}$$

$$= \frac{1}{y!} \cdot e^{y \ln(\nu) - \nu}$$

$$= h(y) \cdot e^{\eta^T \phi(y) - A(\eta)}$$

$h(y) = \frac{1}{y!}$

$\phi(y) = y$

; $\eta = g(\nu) = \ln(\nu) \Rightarrow \nu = e^{\eta}$

⚠ Cauchy distr. is not a member of the exp. family.

## Basic properties of $A(\eta)$

- $A(\eta)$ is convex
- $\nabla_\eta A(\eta) = \mathbb{E}[\phi(y)]$
- $\nabla_\eta^2 A(\eta) = \mathbb{E}[\phi(y)\phi(y)^T] - \mathbb{E}[\phi(y)]\,\mathbb{E}[\phi(y)]^T$

Ex: The case of logistic regression

$$A(\eta) = \log(1+e^\eta)$$

$$\nabla A(\eta) = \frac{e^\eta}{1+e^\eta} = \sigma(\eta)$$

$$\nabla^2 A(\eta) = \sigma(\eta)(1-\sigma(\eta)) \geq 0$$

$$\nabla A(\eta) = \sigma(\eta) = \mathbb{E}[\phi(y)] = \mathbb{E}[y]$$

$$\nabla^2 A(\eta) = \sigma(\eta)(1-\sigma(\eta)) = \mu(1-\mu) = \mathrm{Var}(y)$$

## Link function

Exists the fct $g$ s.t.: $\eta = g\left(\mathbb{E}[\phi(y)]\right)$

- **Parameter estimation**

. Fixed family $(h, \phi)$ + data $(y_i)_{i=1}^N$ i.i.d

→ Recover $\eta$

- MLE: $\mathbb{P}(y|\eta) = h(y)\, e^{\eta^T \phi(y) - A(\eta)}$

$$\ell(\eta) = -\ln(\mathbb{P}(y|\eta))$$

$$= \sum_{i=1}^N -\ln(\mathbb{P}(y_i|\eta))$$

$$= \sum_{i=1}^N \left[-\ln(h(y_i)) - \eta^T \phi(y_i) + A(\eta)\right]$$

$$\nabla \ell(\eta) = \sum_{i=1}^N -\phi(y_i) + N \nabla A(\eta)$$

$$= -\sum_{i=1}^N \phi(y_i) + N\, \mathbb{E}[\phi(y)]$$

$$\nabla \ell(\eta) = 0 \iff \mathbb{E}[\phi(y)] = \frac{1}{N} \sum_{i=1}^N \phi(y_i)$$

with a link fct $g$:

$$\eta = g(\mathbb{E}[\phi(y)]) \Rightarrow \boxed{\eta = g\left(\frac{1}{N} \sum_{i=1}^N \phi(y_i)\right)}$$

# Generalized linear models

$(x, y)$ iid

$$P(y \mid w, x) = h(y) \cdot e^{x^T w \, \phi(y) - A(x^T w)}$$

for $\eta = x^T w$ : the linear prediction

$S_{Train} = (x_i, y_i)_{i=1}^N$, Goal: Find $w$

How: MLE

## Negative log likelihood:

$$\ell(w) = -\sum_{i=1}^N \ln\left(P(y_i \mid x_i, w)\right)$$

$$= -\left[\sum_{i=1}^N \log(h(y_i)) + x_i^T w \, \phi(y_i) - A(x_i^T w)\right]$$

$\ell$ is convex.

- $$\nabla \ell(w) = \sum_{i=1}^N -x_i \phi(y_i) + A'(x_i^T w) x_i$$

$$= \sum_{i=1}^N -x_i \phi(y_i) + \mathbb{E}[\phi(y_i)] x_i$$

$$= \sum_{i=1}^N -x_i \phi(y_i) + g^{-1}(x_i^T w) x_i$$

$\nabla \ell(w) = 0 \iff g^{-1}(x_i^T w) x_i = \sum_{i=1}^N x_i \phi(y_i)$

$\iff g^{-1}(x_i^T w) x_i - \sum_{i=1}^N x_i \phi(y_i) = 0$ (LS: $g = id$)

$\iff X^T [g^{-1}(Xw) - y] = 0$ (LR: $g^{-1} = \sigma$)

# Recap

- Linear Model: $y = X^T w + \varepsilon \rightsquigarrow$ L.S. estimator
- Logistic regression: $P(y=1 \mid x, w) = \sigma(x^T w)$
- Exponential family: $P(y \mid \eta) = h(y)\, e^{\eta^T \phi(y) - A(\eta)}$
  - $h, \phi$ you can decide
  - $\eta$ : nature par.
  - $A$ : log-partition (cumulat)
    - $A(\eta)$ is convex
    - $\nabla A(\eta) = \mathbb{E}[\phi(y)]$

- G.L.M.

$$P(y \mid x, w) = h(y)\, e^{\left(x^T w\, \phi(y) - A(x^T w)\right)}$$

$\rightsquigarrow$ with MLE find $\hat{w}$