

# Exponential Families And Generalized Linear Models

Machine Learning Course - CS-433

Oct 25, 2022

Nicolas Flammarion

**EPFL**

# Motivation

# The LS estimator can be defined in two different ways

## Geometric way:

Minimizing the sum of the squares of the residuals:

$$\hat{w} = \arg \min \frac{1}{2} \sum_{n=1}^N (y_n - x_n^\top w)^2$$

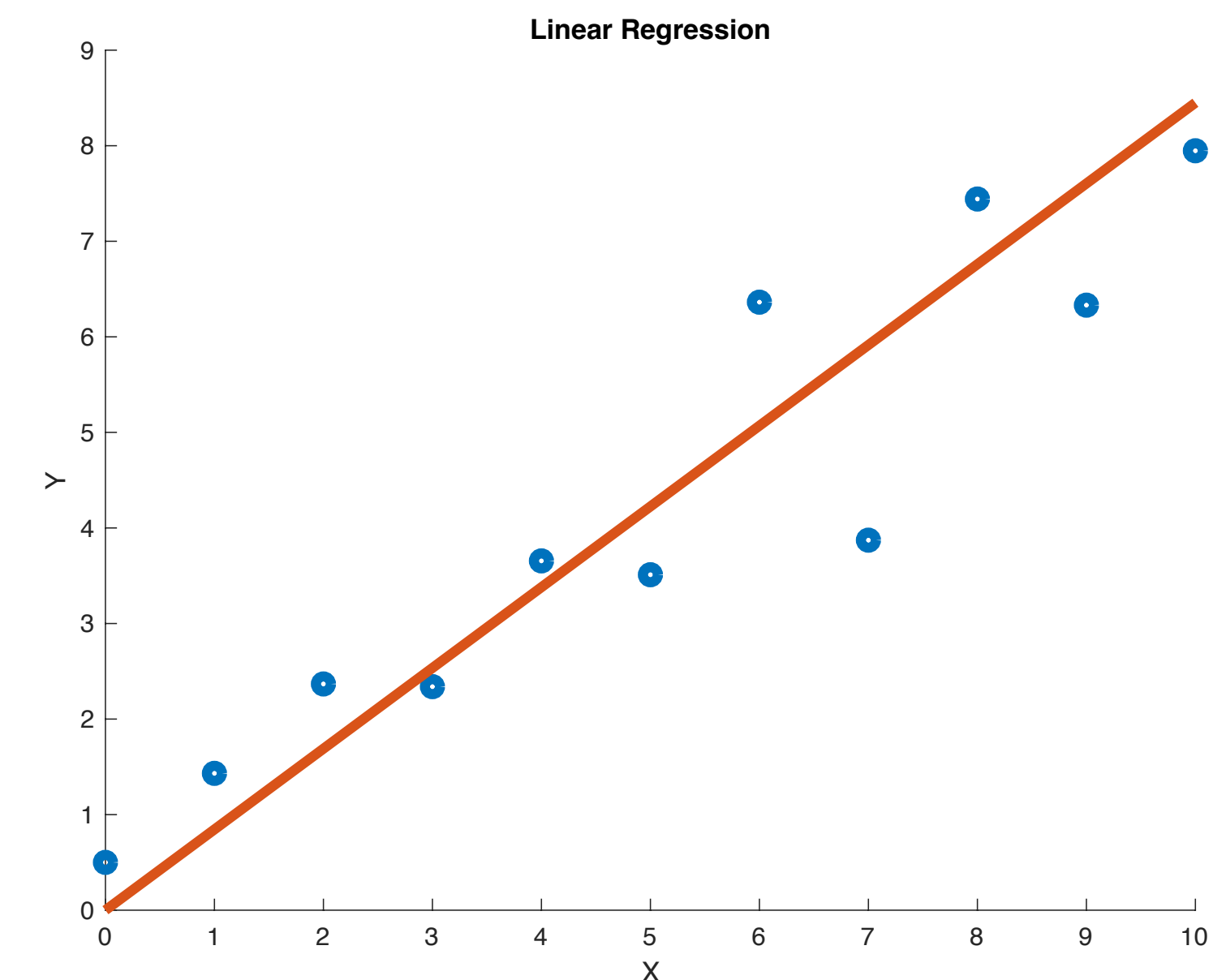
## Probabilistic way:

Assume the data follow a linear Gaussian model:

$$Y = x^\top w + \varepsilon \text{ where } \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\Rightarrow Y \sim \mathcal{N}(x^\top w, \sigma^2)$$

Doing MLE recovers the LS estimator  $\hat{w}$



# How to get non-linear models?

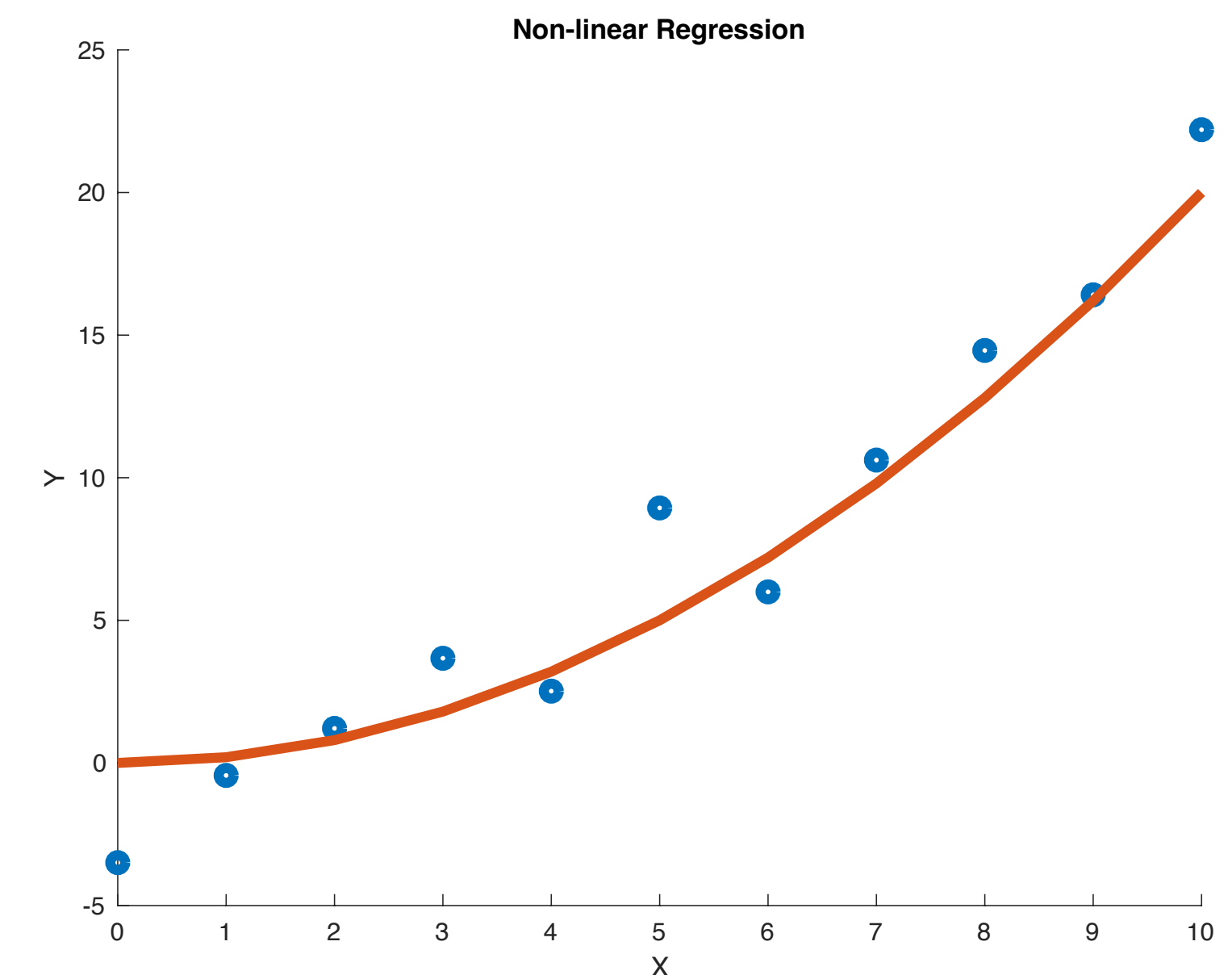
- Features augmentations: add non linear features  $(x, x^2, x^3)$
- Different probabilistic models:
  - LS:  $Y \sim \mathcal{N}(x^\top w, \sigma^2)$

The linear model predicts the mean of a distribution from which the data are sampled

- Logistic regression:  $Y \sim \mathcal{B}(\sigma(x^\top w))$

The linear model predicts an other quantity

- ➡ Generalized linear model
- ➡ Exponential family



# Logistic regression

Logistic regression models the probability of the two classes  $\{0,1\}$  by

$$p(1 | \eta) = \sigma(\eta) \text{ and } p(0 | \eta) = 1 - \sigma(\eta),$$

where  $\eta = x^\top w$ . This can be compactly written as

$$p(y | \eta) = \frac{e^{\eta y}}{1 + e^\eta} = \exp(\eta y - \ln(1 + e^\eta))$$

- The linear model predicts  $\eta$  which is not the mean of the distribution of the observations
- Rather  $\eta$  is related to the mean  $\mu$  through the non-linear relation  $\eta = \ln \frac{\mu}{1 - \mu}$  or  $\mu = \sigma(\eta)$
- The relation between  $\eta$ , the parameter predicted by the linear model and  $\mu$ , the distribution's mean, makes possible to use linear model in this context
  - ➡ It is called the **link function**

# Exponential family: definition

A distribution belongs to the exponential family if it can be written in the form

$$p(y | \eta) = \underbrace{h(y)}_{\geq 0} \exp[\eta^\top \phi(y) - A(\eta)]$$

- $\eta$ : natural or canonical parameter
- $\phi(y)$ : sufficient statistics contains all the relevant information
- $A(\eta)$ : cumulant or log partition, here for normalization but still informative

$$\int p(y | \eta) dy = 1 \implies A(\eta) = \log[\int h(y) \exp(\eta^\top \phi(y))]$$

Degrees of freedom:  $h$ ,  $\phi$  and  $\eta$

# Exponential family: definition

A distribution belongs to the exponential family if it can be written in the form

$$p(y | \eta) = \underbrace{h(y)}_{\geq 0} \exp[\eta^\top \phi(y) - A(\eta)]$$

- $\eta$ : natural or canonical parameter
- $\phi(y)$ : sufficient statistics contains all the relevant information
- $A(\eta)$ : cumulant or log partition, here for normalization but still informative

$$\int p(y | \eta) dy = 1 \implies A(\eta) = \log[\int h(y) \exp(\eta^\top \phi(y)) dy]$$

Natural parameter space  $M = \{\eta : \int h(y) \exp(\eta^\top \phi(y)) dy < \infty\}$

**Why?**



# Bernoulli distributions belong to the exponential family

The Bernoulli distribution is the binary random variable such that for  $\mu \in [0,1]$ :

$$\mathbb{P}(Y = 1) = \mu \quad \text{and} \quad \mathbb{P}(Y = 0) = 1 - \mu$$

Claim: The Bernoulli distribution is a member of the exponential family:

$$\begin{aligned} p(y | \mu) &= \mu^y (1 - \mu)^{1-y} \\ &= \exp\left(\ln \frac{\mu}{1 - \mu} y + \ln(1 - \mu)\right) \\ &= \exp(\eta \phi(y) - A(\eta)) \end{aligned}$$

We can identify:

$$\phi(y) = y, \quad \eta = \ln \frac{\mu}{1 - \mu}, \quad h(y) = 1, \quad \text{and} \quad A(\eta) = -\ln(1 - \mu) = \ln(1 + e^\eta)$$

We have a 1-1 correspondance between  $\mu$  and  $\eta$ :

$$\eta = g(\mu) = \ln \frac{\mu}{1 - \mu} \iff \mu = g^{-1}(\eta) = \frac{e^\eta}{1 + e^\eta}$$

link function

(it links the mean of  $\phi(y)$  to  $\eta$ )

# Gaussian distributions belong to the exponential family

Claim: The Gaussian distribution with mean  $\mu$  and variance  $\sigma^2$  is also a member of the exponential family:

$$\begin{aligned} p(y | \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \\ &= \exp\left[(\mu/\sigma^2, -1/(2\sigma^2))(y, y^2)^\top - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2)\right] \end{aligned}$$

$$\begin{aligned} \phi(y) &= (y, y^2)^\top, \quad \eta = (\mu/\sigma^2, -1/(2\sigma^2))^\top, \quad A(\eta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2} \ln(2\pi\sigma^2), \text{ and } h(y) = 1 \\ &= -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2} \ln(-\eta_2/\pi) \end{aligned}$$

Link function:

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

# Poisson distributions belong to the exponential family

Claim: The Poisson distribution with mean  $\mu$  belongs to the family: for  $y \in \mathbb{N}$

$$\begin{aligned} p(y \mid \mu) &= \frac{\mu^y e^{-\mu}}{y!} \\ &= \frac{1}{y!} e^{y \ln(\mu) - \mu} \\ &= h(y) e^{\eta \phi(y) - A(\eta)} \end{aligned}$$

We can identify:

$$h(y) = 1/y!, \quad \phi(y) = y, \text{ and } \eta = \ln \mu$$

Link function:

$$\eta = g(\mu) = \ln \mu \iff \mu = g^{-1}(\eta) = e^\eta$$

# Basic properties of the cumulant

Claim:

- $A(\eta)$  is convex
- $\nabla A(\eta) = \mathbb{E}[\phi(Y)]$
- $\nabla^2 A(\eta) = \mathbb{E}[\phi(Y)\phi(Y)^\top] - \mathbb{E}[\phi(Y)]\mathbb{E}[\phi(Y)]^\top$

# Convexity of the cumulant

Proof: for  $\eta_1, \eta_2$  two parameters we define  $\eta = \lambda\eta_1 + (1 - \lambda)\eta_2$ . We want to show

$$A(\eta) \leq \lambda A(\eta_1) + (1 - \lambda)A(\eta_2)$$

We have first

$$\begin{aligned} \exp A(\eta) &= \int h(y) \exp(\eta^\top \phi(y)) dy \\ &= \int h(y) \exp((\lambda\eta_1 + (1 - \lambda)\eta_2)^\top \phi(y)) dy \\ &= \int \underbrace{\left[ h(y)^\lambda \exp(\lambda\eta_1^\top \phi(y)) \right]}_{f(y)} \cdot \underbrace{\left[ h(y)^{1-\lambda} \exp((1 - \lambda)\eta_2^\top \phi(y)) \right]}_{g(y)} dy \\ &= \int f(y) g(y) dy \\ &= \|fg\|_1 \end{aligned}$$

# The proof uses Hoelder's inequality

We recall the **Hoelder's inequality**:

$$\|fg\|_1 \leq \|f\|_p \|g\|_q$$

for  $p, q \in [1, +\infty]$  s.t.  $\frac{1}{p} + \frac{1}{q} = 1$ , and  $\|f\|_p = (\int |f(y)|^p dy)^{1/p}$

We apply Hoelder's inequality to  $f$  and  $g$  for  $p = 1/\lambda$  and  $q = 1/(1 - \lambda)$ :

$$\|fg\|_1 \leq \|f\|_p \|g\|_q$$

We check that  $1/p + 1/q = \lambda + (1 - \lambda) = 1$

# Proof

$$\begin{aligned}\|f\|_p &= \left( \int f(y)^p dy \right)^{1/p} \\ &= \left( \int \left( h(y)^\lambda \exp(\lambda \eta_1^\top \phi(y)) \right)^{1/\lambda} dy \right)^\lambda \\ &= \left( \int h(y) \exp(\eta_1^\top \phi(y)) dy \right)^\lambda\end{aligned}$$

$$\begin{aligned}\|g\|_q &= \left( \int g(y)^q dy \right)^{1/q} \\ &= \left( \int \left( h(y)^{1-\lambda} \exp((1-\lambda) \eta_2^\top \phi(y)) \right)^{\frac{1}{1-\lambda}} dy \right)^{1-\lambda} \\ &= \left( \int h(y) \exp(\eta_2^\top \phi(y)) dy \right)^{1-\lambda}\end{aligned}$$

Therefore we have

$$\begin{aligned}\|f\|_p \|g\|_q &= \left( \int h(y) \exp(\eta_1^\top \phi(y)) dy \right)^\lambda \left( \int h(y) \exp(\eta_2^\top \phi(y)) dy \right)^{1-\lambda} \\ &= \exp(\lambda A(\eta_1)) \exp((1-\lambda) A(\eta_2))\end{aligned}$$

# Summary of the proof:

We have

$$\begin{aligned}\exp A(\eta) &= \int h(y) \exp(\eta^\top \phi(y)) dy \\ &= \int h(y) \exp((\lambda \eta_1 + (1 - \lambda) \eta_2)^\top \phi(y)) dy \\ &= \int [h(y)^\lambda \exp(\lambda \eta_1^\top \phi(y))] \cdot [h(y)^{1-\lambda} \exp((1 - \lambda) \eta_2^\top \phi(y))] dy \\ &\leq \left[ \int h(y) \exp(\eta_1^\top \phi(y)) dy \right]^\lambda \cdot \left[ \int h(y) \exp(\eta_2^\top \phi(y)) dy \right]^{1-\lambda} \\ &= \exp(\lambda A(\eta_1)) \exp((1 - \lambda) A(\eta_2))\end{aligned}$$

Taking the log proves the claim:

$$A(\eta) \leq \lambda A(\eta_1) + (1 - \lambda) A(\eta_2)$$



# Derivative of $A(\eta)$ and moments: particular cases

Bernoulli distribution:

$$A'(\eta) = \frac{d}{d\eta} \ln(1 + e^\eta) = \frac{e^\eta}{1 + e^\eta} = \sigma(\eta) = \mu$$

$$A''(\eta) = \frac{d}{d\eta} \sigma(\eta) = \sigma(\eta)(1 - \sigma(\eta)) = \mu(1 - \mu)$$

Gaussian distribution:

$$\frac{\partial}{\partial \eta_1} A(\eta) = \frac{\partial}{\partial \eta_1} \left( -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2} \ln(-\eta_2/\pi) \right) = -\frac{\eta_1}{2\eta_2} = \mu$$

$$\frac{\partial}{\partial \eta_2} A(\eta) = \frac{\partial}{\partial \eta_2} \left( -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2} \ln(-\eta_2/\pi) \right) = \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} = \mu^2 + \sigma^2$$

$$\frac{\partial^2}{\partial \eta_1^2} A(\eta) = \frac{\partial}{\partial \eta_1} \left( -\frac{\eta_1}{2\eta_2} \right) = -\frac{1}{2\eta_2} = \sigma^2$$

# Derivative of $A(\eta)$ and moments: general case

$$\begin{aligned}\nabla A(\eta) &= \nabla \left[ \ln \int h(y) \exp(\eta^\top \phi(y)) dy \right] \\&= \nabla \left[ \int h(y) \exp(\eta^\top \phi(y)) dy \right] \cdot \left( \int h(y) \exp(\eta^\top \phi(y)) dy \right)^{-1} \\&= \nabla \left[ \int h(y) \exp(\eta^\top \phi(y)) dy \right] \cdot \exp(-A(\eta)) \\&= \int \nabla \left[ h(y) \exp(\eta^\top \phi(y)) \right] \cdot \exp(-A(\eta)) \\&= \int h(y) \exp(\eta^\top \phi(y)) \phi(y) dy \cdot \exp(-A(\eta)) \\&= \int h(y) \exp(\eta^\top \phi(y) - A(\eta)) \phi(y) dy \\&= \int \phi(y) p(y | \eta) dy \\&= \mathbb{E}[\phi(Y)]\end{aligned}$$

# Link function

Def: It is the function  $g$  such that:

$$\eta = g(\mathbb{E}[\phi(Y)])$$

Thus the mean parameter  $\mu := \mathbb{E}[\phi(Y)]$  and the natural parameter  $\eta$  are linked through:

$$\eta = g(\mu) \iff \mu = g^{-1}(\eta)$$

Remark:  $g^{-1}(\eta) = \nabla A(\eta)$

Moment parameterization and canonical parametrization

# Applications in ML

# Maximum likelihood estimation

Data  $\{y_n\}_{n=1}^N$  coming from a member of the exponential family with given  $(h, \phi)$

Goal: Estimate the natural parameter  $\eta$

How: MLE for  $p(y | \eta) = h(y)\exp(\eta^\top \phi(y) - A(\eta))$  amounts to minimize

$$\begin{aligned} L(\eta) &= -\frac{1}{N} \ln(p(\mathbf{y} | \eta)) \\ &= \frac{1}{N} \sum_{n=1}^N \left[ -\ln(h(y_n)) - \eta^\top \phi(y_n) + A(\eta) \right] \\ &= -\frac{1}{N} \sum_{n=1}^N \ln(h(y_n)) - \eta^\top \left( \frac{1}{N} \sum_{n=1}^N \phi(y_n) \right) + A(\eta) \end{aligned}$$

➡ The cost function  $L$  is convex since the cumulant  $A$  is convex

# Maximum likelihood parameter estimation

Gradient:

$$\begin{aligned}\nabla L(\eta) &= -\frac{1}{N} \sum_{n=1}^N \phi(y_n) + \nabla A(\eta) \\ &= -\frac{1}{N} \sum_{n=1}^N \phi(y_n) + \mathbb{E}[\phi(Y)]\end{aligned}$$

Stationary point:

$$\mu := \mathbb{E}[\phi(Y)] = \frac{1}{N} \sum_{n=1}^N \phi(y_n)$$

Closed form: assume we have determined the link function  $g(\mu) = \eta$

$$\eta = g\left(\frac{1}{N} \sum_{n=1}^N \phi(y_n)\right)$$

Ex: what does it mean for today's examples (Bernoulli, Poisson and Gaussian)?

# Generalized Linear Models (GLM)

Both linear and logistic regressions focus on the conditional relationship between  $X$  and  $Y$

- LS:  $Y \sim \mathcal{N}(x^\top w, \sigma^2)$
- Logistic regression:  $Y \sim \mathcal{B}(\sigma(x^\top w))$

Common feature of linear and logistic regression:

1. Model the conditional expectation as  $\mu = f(w^\top x)$
2. Endow  $Y$  with a particular probability distribution having  $\mu$  as parameter

The GLM framework extends these ideas to the general exponential family.

# Generalized Linear Models (GLM)

A GLM makes three assumptions regarding the form of  $p(y | x)$ :

- The observed input  $x$  enters into the model via a linear combination  
 $\eta = x^\top w$
- The conditional mean  $\mu$  is represented as a function  $f(\eta)$  of the linear combination  $\eta$
- The observed output  $y$  is assumed to be characterized by an exponential family distribution with conditional mean  $\mu$

The condition probability is thus modeled as:

$$p(y | w, x) = h(y) \exp(\eta \phi(y) - A(\eta)) \quad \text{for } \eta = g \circ f(x^\top w)$$



# Generalized Linear Models (GLM)

Two choice points in the specification of a GLM:

- The choice of the exponential family distribution
  - ➡ Generally constrained by the nature of the data  $Y$
- The choice of the response function  $f$ 
  - ➡ Real degree of freedom!
  - ➡ *Canonical response function:  $f = g^{-1}$* , uniquely associated with the given exponential family distribution

If we decide to use the canonical response function, the choice of the exponential family density completely determines the GLM:

$$p(y | w, x) = h(y) \exp(\eta \phi(y) - A(\eta)) \quad \text{for } \eta = x^\top w$$

# Negative log-likelihood estimation

Data  $\{x_n, y_n\}_{n=1}^N$

Goal: Estimate the parameter  $w$  of the GLM in the case of the canonical response function

How: MLE for  $L(w) = -\frac{1}{N} \sum_{n=1}^N \ln p(y_n | x_n^\top w)$

$$= -\frac{1}{N} \sum_{n=1}^N \ln(h(y_n)) + x_n^\top w \phi(y_n) - A(x_n^\top w)$$

➡  $L$  is convex

$$\begin{aligned} \nabla L(w) &= -\frac{1}{N} \sum_{n=1}^N \phi(y_n) x_n - A'(x_n^\top w) x_n \\ &= -\frac{1}{N} \sum_{n=1}^N \phi(y_n) x_n - \mathbb{E}[\phi(Y_n)] x_n \\ &= -\frac{1}{N} \sum_{n=1}^N \phi(y_n) x_n - g^{-1}(x_n^\top w) x_n \end{aligned}$$

$$\nabla L(w) = 0 \iff \mathbf{X}^\top [g^{-1}(\mathbf{X}w) - \phi(\mathbf{y})] = 0$$

# Summary

- Linear model  $y = x^\top w + \varepsilon \quad \rightarrow \text{LS estimator}$
- Logistic regression:  $p(y = 1 | x, w) = \sigma(x^\top w)$
- Exponential family  $p(y | \eta) = h(y)\exp(\eta^\top \phi(y) - A(\eta))$ 
  - $h, \phi$  degree of freedom
  - $\eta$  natural parameter
  - $A$  log-partition
    - $A(\eta)$  is convex
    - $\nabla_\eta A(\eta) = \mathbb{E}[\phi(y)]$
- GLM:  $p(y | w, x) = h(y)\exp(w^\top x \phi(y) - A(w^\top x))$ 
  - $\rightarrow$  With MLE find  $\hat{w}$