

Exponential Families And Generalized Linear Models

Machine Learning Course - CS-433

Oct 26, 2021

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:

$$\hat{w} = \arg \min \frac{1}{2} \sum_{i=1}^n (y_i - x_i^\top w)^2$$

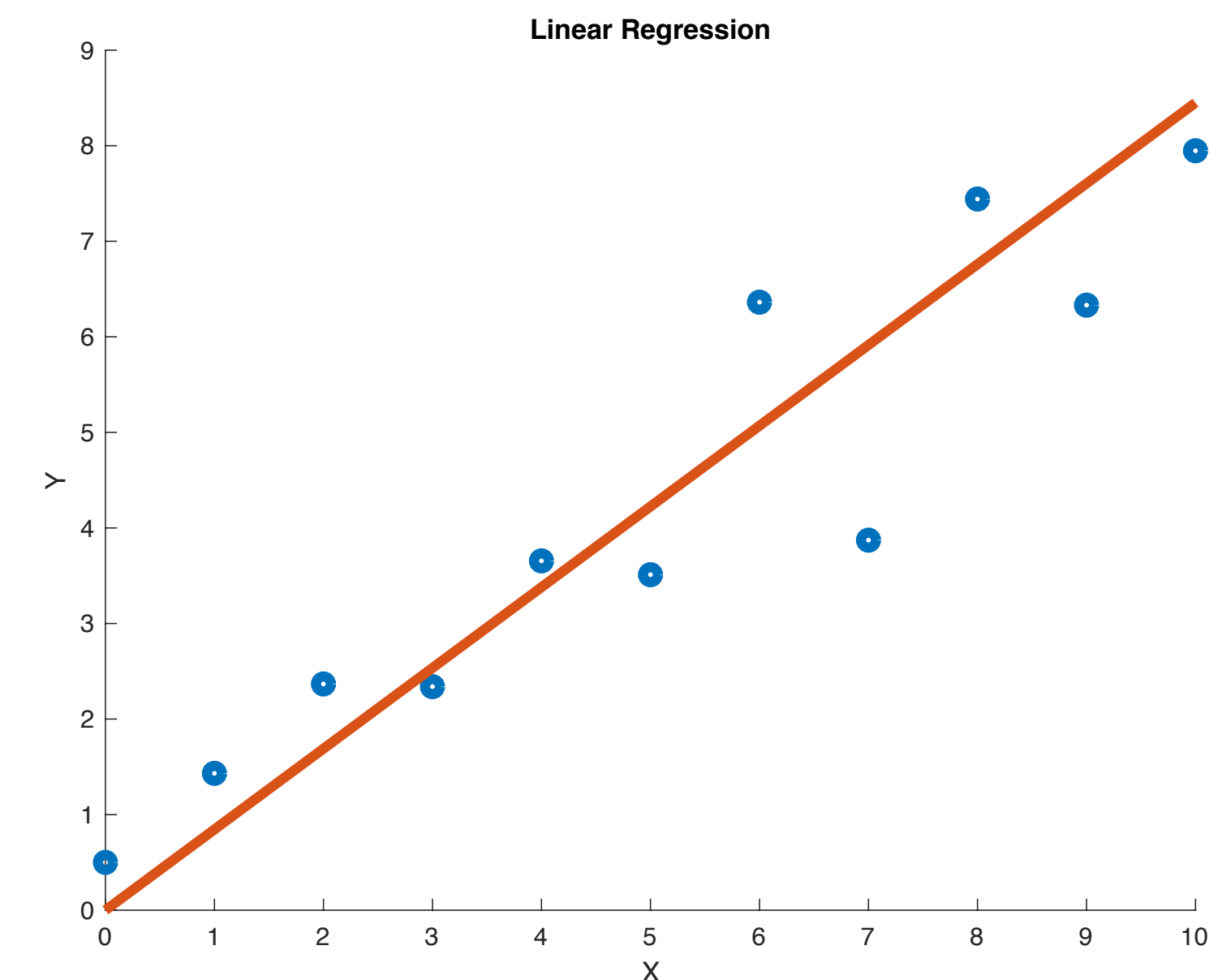
Probabilistic way:

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?

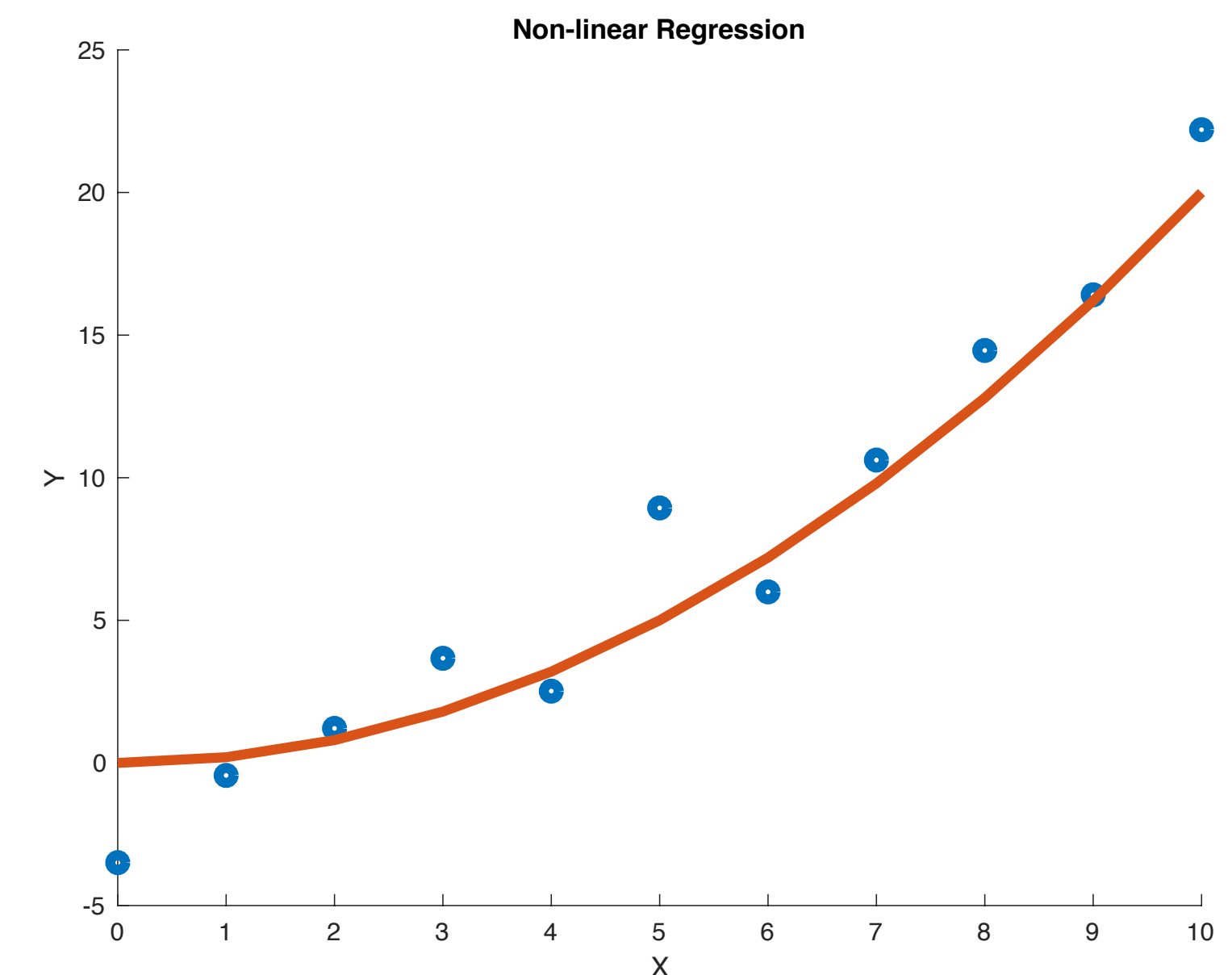
- Features augmentations: add non linear features (x, x^2, x^3)
- Different probabilistic models:
 - LS: $Y \sim \mathcal{N}(x^\top w, \sigma^2)$

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 η which is not the mean of the distribution of the observations
- Rather η is related to the mean μ through the non-linear relation $\eta = \ln \frac{\mu}{1 - \mu}$ or $\mu = \sigma(\eta)$
- The relation between η , the parameter predicted by the linear model and μ , 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

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

- η : 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 , ϕ and η

Exponential family: definition

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

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

- η : 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 $\mu \geq 0$:

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

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 μ et η :

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

link function

(it links the mean of $\phi(y)$ to η)

Gaussian distributions belong to the exponential family

Claim: The Gaussian distribution with mean μ and variance σ^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 μ 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 η_1, η_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

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

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 η are linked through:

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

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

Moment parameterization and canonical parametrization

Applications in ML

Maximum likelihood estimation

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

Goal: Estimate the natural parameter η

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

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

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

Maximum likelihood parameter estimation

Gradient:

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

Stationary point:

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

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

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

Ex: what does it mean for today 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: $Y \sim \mathcal{B}(\sigma(x^\top w))$

Commun 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 μ as parameter

The GLM frameworks extends these to the general exponential family by modeling the conditional probability as

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

Generalized Linear Models (GLM)

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

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 $\mu := \mathbb{E}[\phi(Y) | X]$ is represented as a function $g^{-1}(\eta)$ of the linear combination η
- The observed output y is assumed to be characterized by an exponential family distribution with conditional mean μ

Negative log-likelihood estimation

Data $\{x_i, y_i\}_{i=1}^n$

Goal: Estimate the parameter w of the GLM

How: MLE for $L(w) = - \sum_{i=1}^n \ln p(y_i | x_i^\top w)$
 $= - \sum_{i=1}^n \ln(h(y_i)) + x_i^\top w \phi(y_i) - A(x_i^\top w)$

➡ L is convex

$$\begin{aligned}\nabla L(w) &= - \sum_{i=1}^n \phi(y_i) x_i - A'(x_i^\top w) x_i \\ &= - \sum_{i=1}^n \phi(y_i) x_i - \mathbb{E}[\phi(Y_i)] x_i \\ &= - \sum_{i=1}^n \phi(y_i) x_i - g^{-1}(x_i^\top w) x_i\end{aligned}$$

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