Lecture FYS-STK3155/4155, September 16, 2024

FYS-STK3155/4155, September 16

kaggle.com

Linear regression

$$y_i = f(x_i) + \varepsilon_i$$

$\varepsilon_i \sim N(0, \sigma^2)$

$x_i \in (-\infty, \infty)$ $\quad [x_a, x_b]$

$y_i \in (-\infty, \infty)$ $\quad [y_a, y_b]$

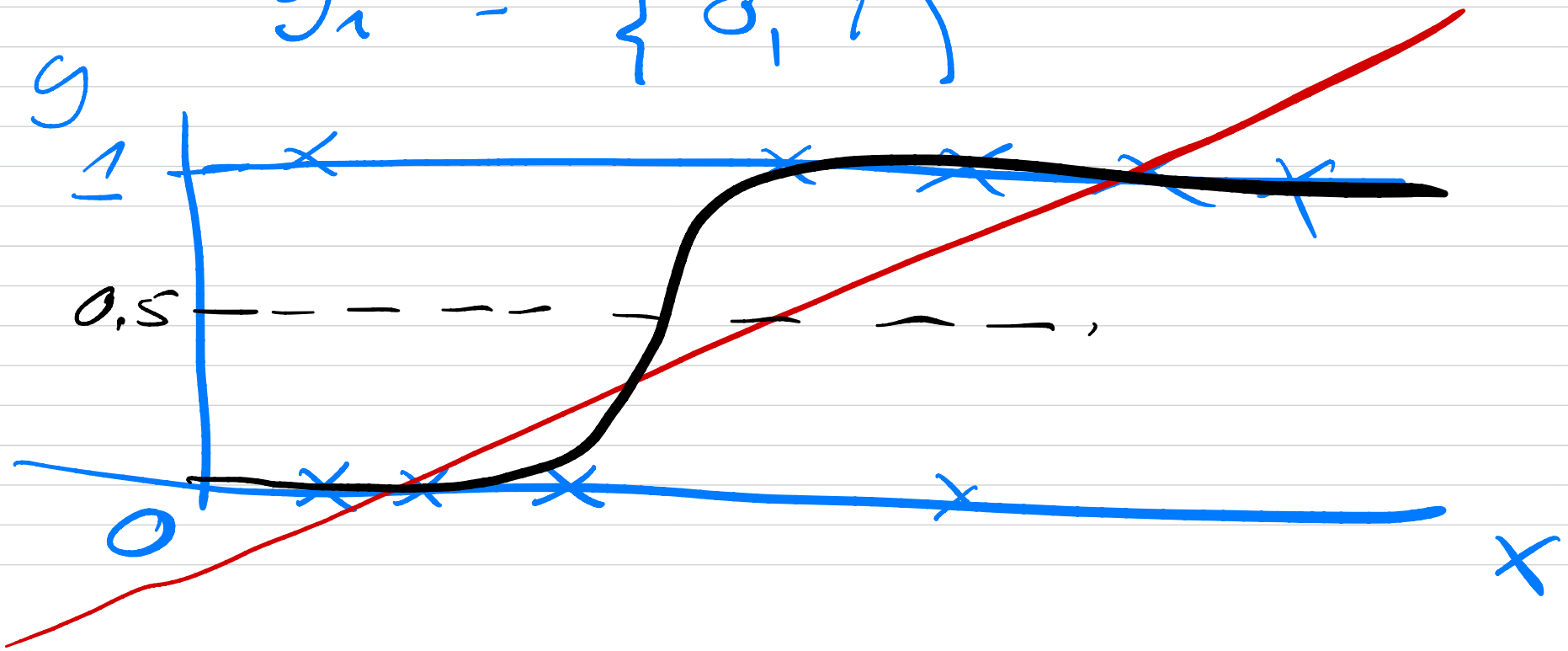
$$y_i = \underbrace{\sum_j x_{ij} \beta_j}_{\tilde{y}_i} + \varepsilon_i$$

First-order polynomial

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

what about y_i given by discrete values ?

$$y_i = \{0, 1\}$$



$y_i = f(x_i) + \varepsilon_i \rightarrow$

$y_i = p(x_i) + \varepsilon_i$

$\int_{x \in D} p(x)\, dx = 1 \quad \left\{ \sum_{i \in D} p(x_i) = 1 \right.$

$x \in [0, 1] \qquad p(1) = 1$

$p(x_i) \leq p(x_j) \qquad x_i \leq x_j$

Example : Sigmoid function

$$p(x) = \frac{e^{\beta x}}{1 + e^{\beta x}}$$

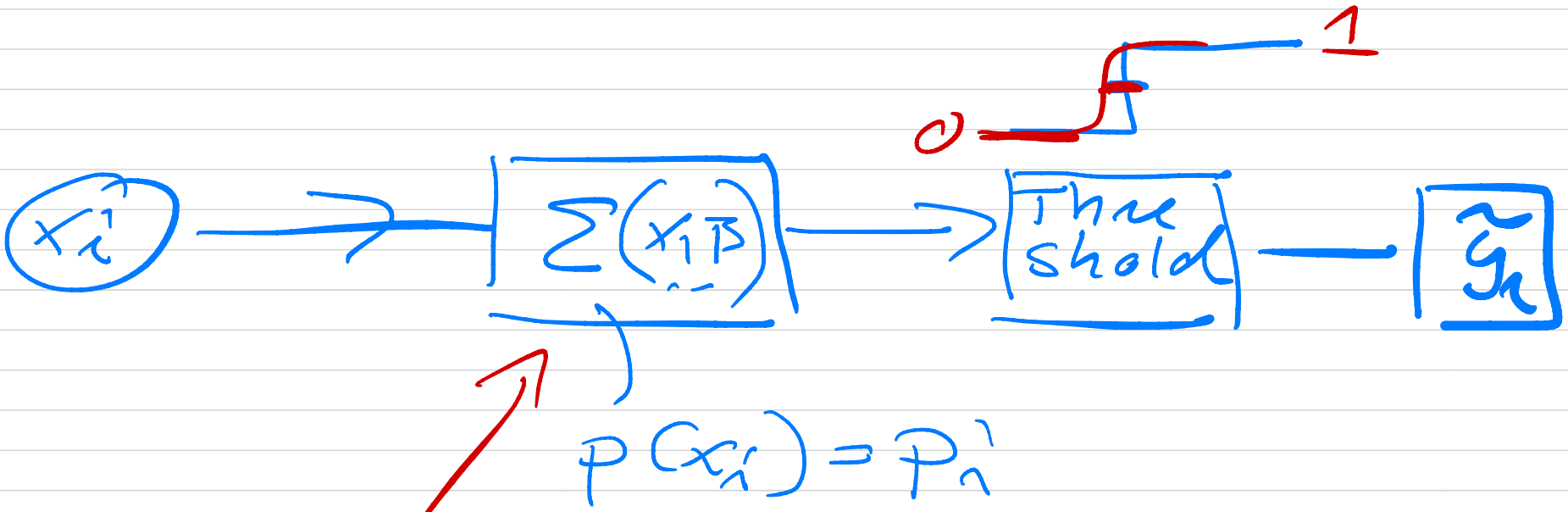
$x \in [0, a]$ $\beta = 1$

$$\int_0^a dx \frac{e^x}{1+e^x} = \ln(1 + e^a)$$

$y = 1$; $p(x) = \frac{e^{\beta x}}{1+e^{\beta x}}$

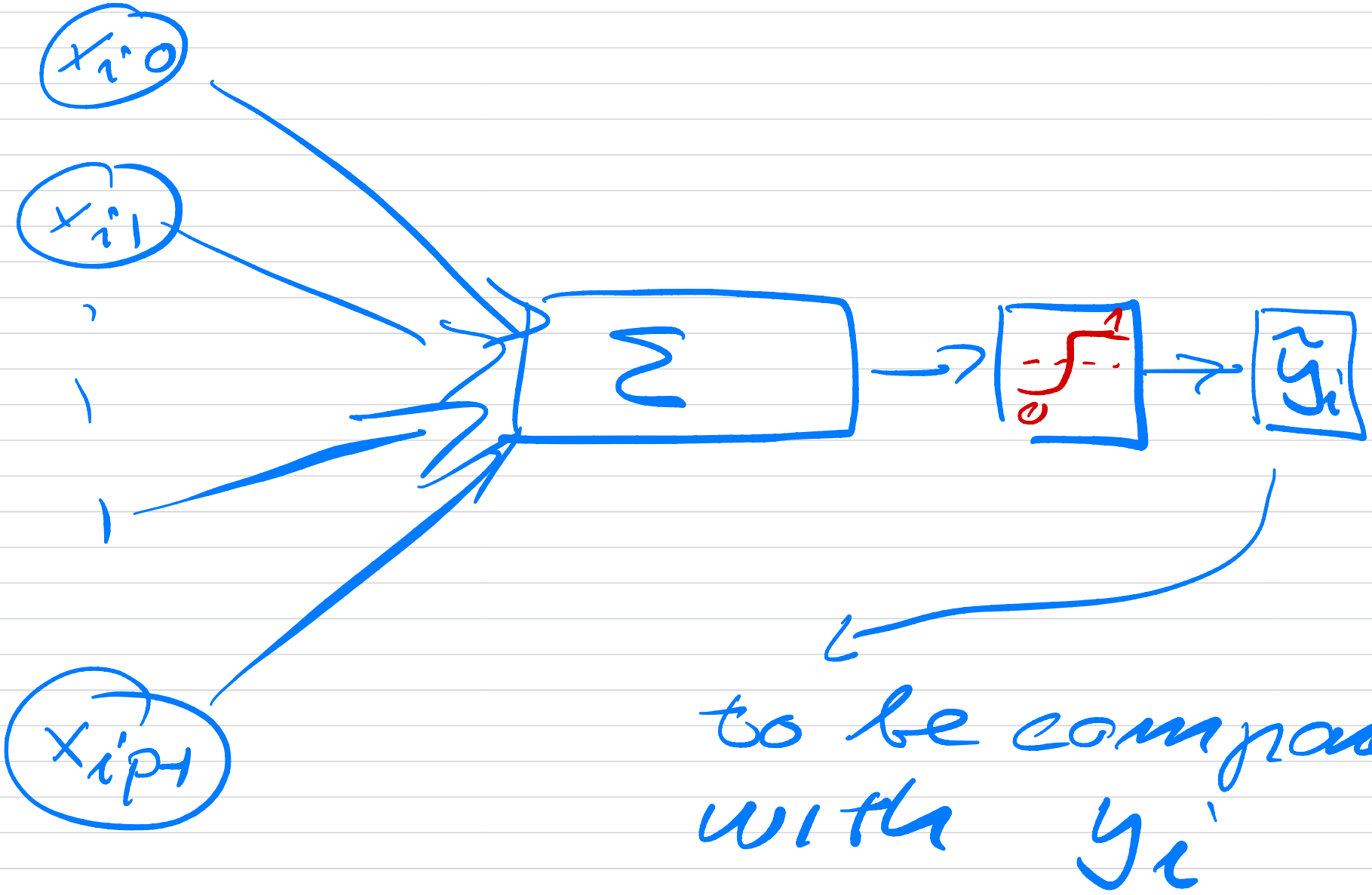
$y = 0$; $1 - p(x) = \frac{1}{1+e^{\beta x}}$

Graphical representation



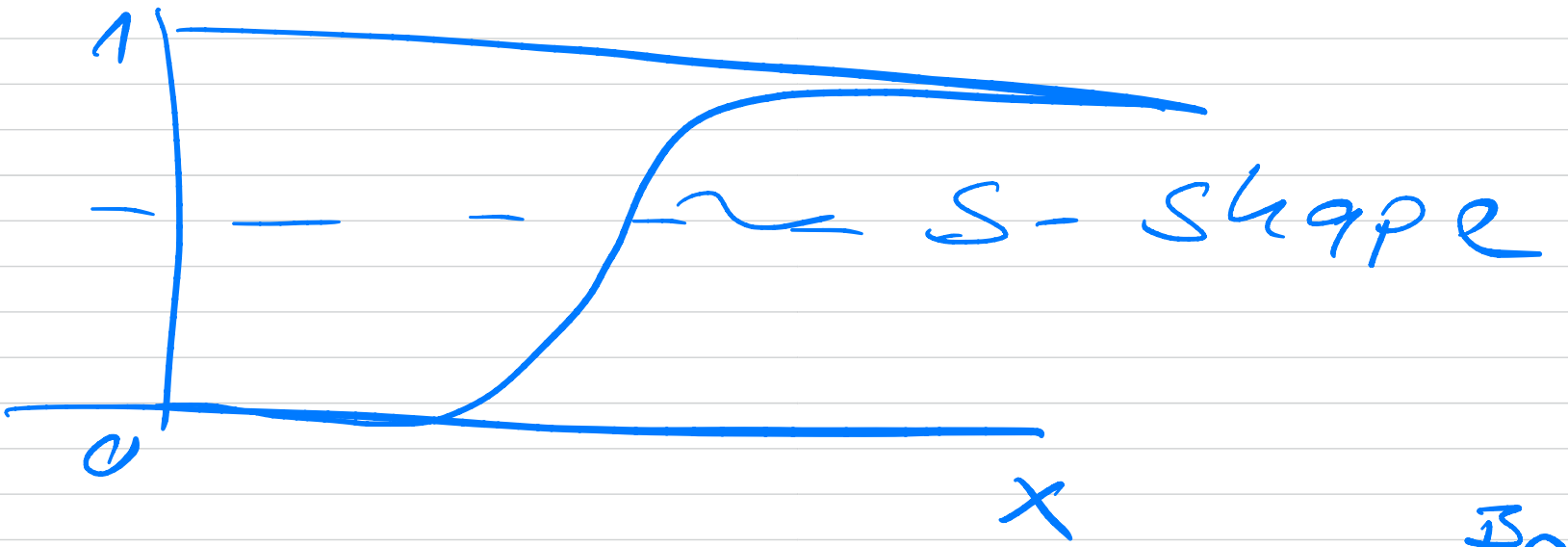
Neural networks activation function

$$X = \begin{bmatrix} x_{00} & x_{01} & \cdots & x_{0\,p-1} \\ x_{10} & & & \\ x_{20} & & & \\ \vdots & & & \\ x_{i0} & x_{i1} & x_{i2} & \cdots x_{i\,p-1} \\ \vdots & & & \\ x_{n-1\,0} & \cdots & \cdots & x_{n-1\,p-1} \end{bmatrix}$$


x_{i0}
x_{i1}
x_{ip}
Σ
$\hat{y}_i$
to be compared with y_i

$$P(x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$P(x_i)$



$$y_i = \underline{1} \quad ; \quad P(x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$$y_i = P_i + \varepsilon_i \qquad = P_i$$

what probability does ε_i' follow?

$$y_i' = 1 = p_i' + \varepsilon_i'$$

$$y_i' = 0 = p_i' + \varepsilon_i'$$

$$\varepsilon_i' = 1 - p_i' \quad \text{when } y_i' = 1$$

$$\varepsilon_i' = -p_i' \quad \text{when } y_i' = 0$$

$$E[\varepsilon_i] = p_i(\underbrace{1-p_i}_{\varepsilon_i}) + (-p_i)(1-p_i)$$

$$\left(E[X] = \sum_{i=0}^{n-1} x_i\, p(x_i)\right)$$

$$= 0$$

$$\mathrm{Var}[\varepsilon_i] = (1-p_i)^2 p_i + (-p_i)^2 (1-p_i)$$

$$= p_i(1-p_i)$$

ε_i follows a binomial distribution.

How do we find β?

y_i are i.i.d.

$$D = \{(x_0, y_0), (x_1, y_1), \ldots, (x_{n-1}, y_{n-1})\}$$

$$P(D|\beta) = \prod_{i=0}^{n-1} p_i^{y_i} (1-p_i)^{1-y_i}$$

$$y_i = \{0, 1\} \qquad p_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\max} \; \underbrace{p(D|\beta)}_{c(\beta)}$$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(\underbrace{-\log p(D|\beta)}_{c(\beta)} \right)$$

$$\Rightarrow c(\beta) = -\sum_{i=0}^{n-1} \left[y_i \log p_i + (1-y_i) \log(1-p_i) \right]$$

$$\log\left[\frac{e^{\beta_0+\beta_1 x_1'}}{1+e^{\beta_0+\beta_1 x_1'}}\right]$$

$$= (\beta_0+\beta_1 x_1') - \log(1+e^{\beta_0+\beta_1 x_1'})$$

$$\Rightarrow c(\beta) = -\sum_{i=0}^{n-1}\left[y_i'(\beta_0+\beta_1 x_1') - \log(1+e^{\beta_0+\beta_1 x_1'})\right]$$

$$\frac{\partial c}{\partial \beta_0} = 0 \quad \wedge \quad \frac{\partial c}{\partial \beta_1} = 0$$

$$\frac{\partial C}{\partial \beta_0} = 0 = -\sum_{i=0}^{n-1} (y_i - p_i)$$

$$p_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$$\frac{\partial C}{\partial \beta_1} = 0 = -\sum_{i=0}^{n-1} x_i (y_i - p_i)$$

in general for the gradient $g = \frac{\partial C}{\partial \beta} =$

$$\frac{\partial c}{\partial p} = -X^T(y - p)$$

$y, p \in \mathbb{R}^n$

$X \in \mathbb{R}^{n \times p}$ $\Rightarrow$

$$g = \frac{\partial c}{\partial p} \in \mathbb{R}^p$$

$\hat{\beta}$ is included in p!

In linear regression

$$\frac{\partial^2 C}{\partial \beta \partial \beta^T} = \frac{2}{n} X^T X$$

Hessian matrix

independent of β

Log Reg

$$\frac{\partial^2 C}{\partial \beta \partial \beta^T} = X^T W X$$ depends on β.

$$W_{ii} = p_i(1-p_i)$$

$$W_{ij}\ (i \neq j) = 0$$

How do we solve the eqs for $\hat{\beta}$?

Newton-Raphson's method.

β - scalar, $\frac{\partial C}{\partial \beta} \to \frac{dC}{d\beta}$

Taylor-expansion around the optimal $\hat{\beta}$

$\hat{\beta} - \beta^{(m)}$

$$C(\hat{\beta}) = C(\beta^{(n)}) + g(\beta^{(n)})\left(\hat{\beta} - \beta^{(n)}\right) + \frac{1}{2} H(\beta^{(n)})\left(\hat{\beta} - \beta^{(n)}\right)^2 + \cdots$$

$$\left(\frac{dC}{d\beta}\Big|_{\beta=\beta^{(n)}}\right)$$

$$\frac{d^2C}{d\beta^2}\Big|_{\beta=\beta^{(n)}}$$

$$h = \hat{\beta} - \beta^{(n)}$$

$$C(\vec{\beta}) \approx C(\beta^{(n)}) + \underbrace{g^{(n)}}_{g(\beta^{(n)})} h + \frac{1}{2} h^2 \underbrace{H^{(n)}}_{H(\beta^{(n)})}$$

$$\frac{dC}{dh} = g^{(n)} + H^{(n)} h = 0$$

$$h = \vec{\beta} - \beta^{(n)} = -\left(H^{(n)}\right)^{-1} g^{(n)}$$

$$\vec{\beta} = \beta^{(n+1)}$$

$$\beta^{(n+1)} = \beta^{(n)} - \left(H^{(n)}\right)^{-1} g^{(n)}$$

$n = 0, 1, \ldots$

(i) start with a guess $\beta^{(0)}$

(ii) iterate over n till

$$\left|\beta^{(n+1)} - \beta^{(n)}\right| \leq \varepsilon \sim 10^{-8}$$

in general, $H(\beta^{(n)})$ is a matrix, $g(\beta^{(n)})$ is a vector

$$g \in \mathbb{R}^p \qquad H \in \mathbb{R}^{p \times p}$$

$$\beta \in \mathbb{R}^p$$

$$\beta^{(n+1)} = \beta^{(n)} - \left(H(\beta^{(n)})\right)^{-1} g(\beta^{(n)})$$

Replace $H(\beta^{(n)})$ with a parameter $\gamma^{(n)}$

$$\beta^{(n+1)} = \beta^{(n)} - \boxed{\gamma^{(n)}} g^{(n)}$$

learning rate

γ

can we find an optimal γ?

Expand $c(p)$ around

$$p^{(n)} - \gamma g^{(n)}$$

$$c(p^{(n)} - \gamma g^{(n)}) = c(p^{(n)})$$

$$- \gamma (g^{(n)})^T (g^{(n)})$$

$$+ \frac{1}{2} \gamma^2 [g^{(n)}]^T H^{(n)} g^{(n)} + \ldots$$

$$g^{(n)} \to g \qquad H^{(n)} \to H$$

Derivative wrt γ

$$\frac{\partial c}{\partial \gamma} = 0 \Rightarrow$$

$$\gamma = \frac{g^T g}{g^T H g} \qquad (\gamma^{(n)})$$

Assume $Hg = \lambda g$

$$\gamma = \frac{g^T g}{g^T \lambda g} = \frac{1}{\lambda} \Rightarrow$$

$$\gamma_{min} = \frac{1}{\lambda_{max}}$$

$$\gamma_{max} = \frac{1}{\lambda_{min}}$$

If you can diagonalize $H^{(n)}$

$$\boxed{\gamma < \frac{2}{\lambda_{max}}}$$