(1) $(x_1, y_1), (x_2, y_2), (x_3, y_3) - \dots - (x_m, y_m)$

II $h_\theta(x) = \theta_0 + \theta_1 x$

$h_\theta(x) = \theta_0 + \theta_1 x_2 + \theta_2 (x_2)$ $\quad x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$

~~$\| y - h_\theta(x) \|$~~ $\quad \text{Loss } H_\theta = \text{Cost}(h_\theta(x), y)$

$$= \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

$$\min_{\theta_1, \theta_2} \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 ; \quad \min_{\theta_1, \theta_2} J(\theta)$$

$h_\theta(x) = \theta_0 + \theta_1 x$

~~$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 (x_2)$~~

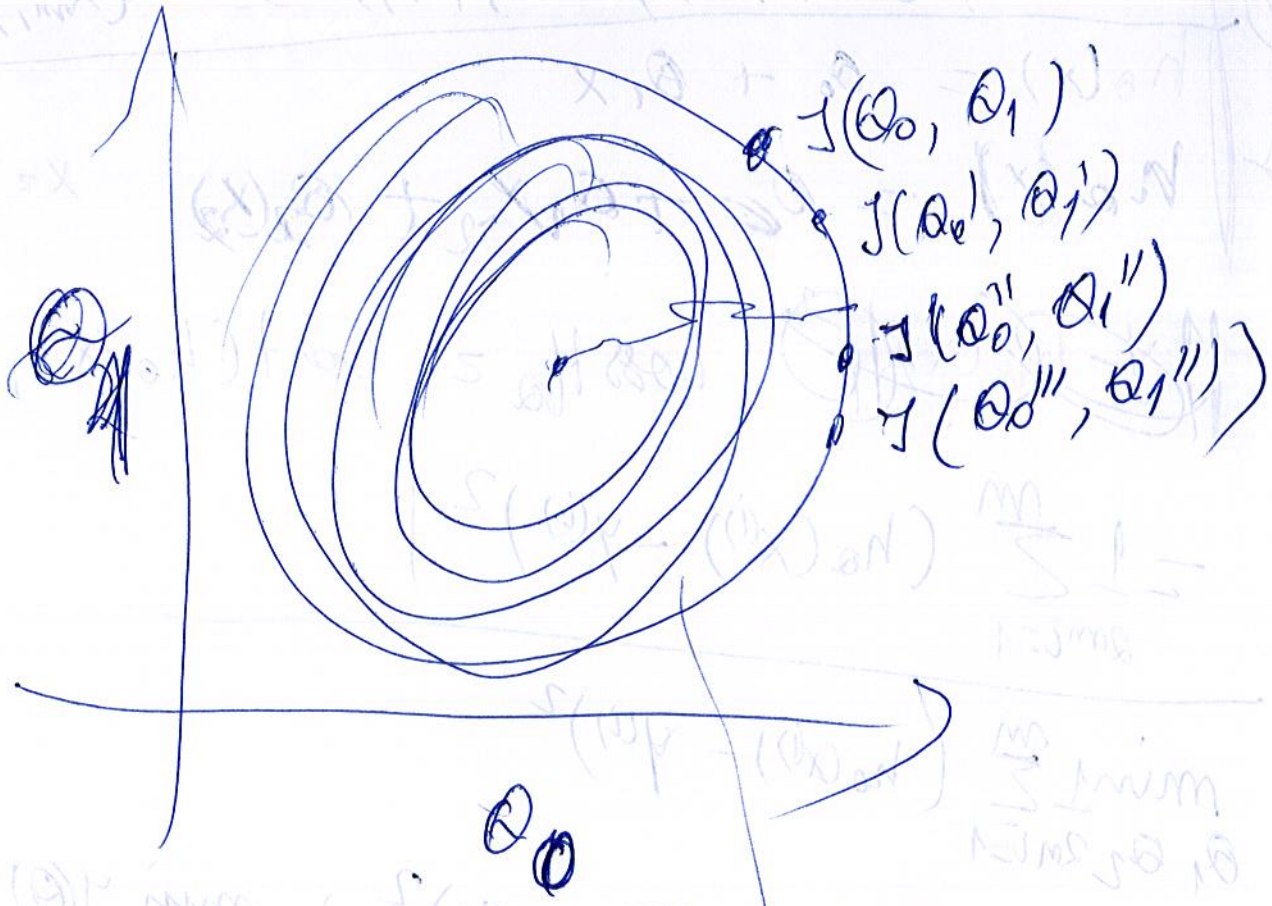
$\theta_0 \cdot 1$

$\theta_1 x$

$\theta_2 x_2 =$

$\downarrow \quad y = b + ax$

(2)

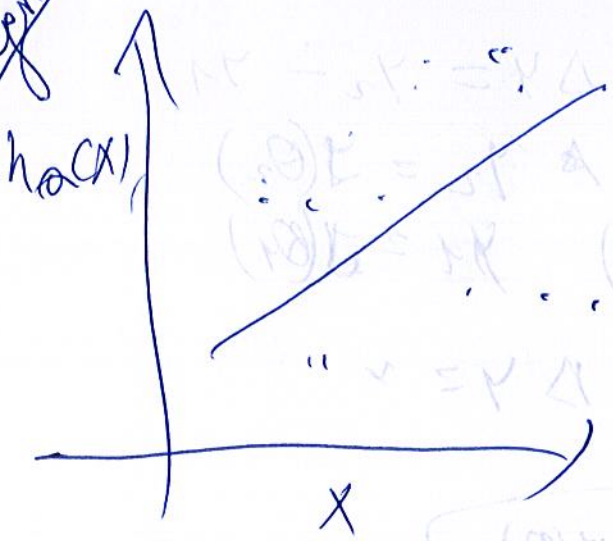


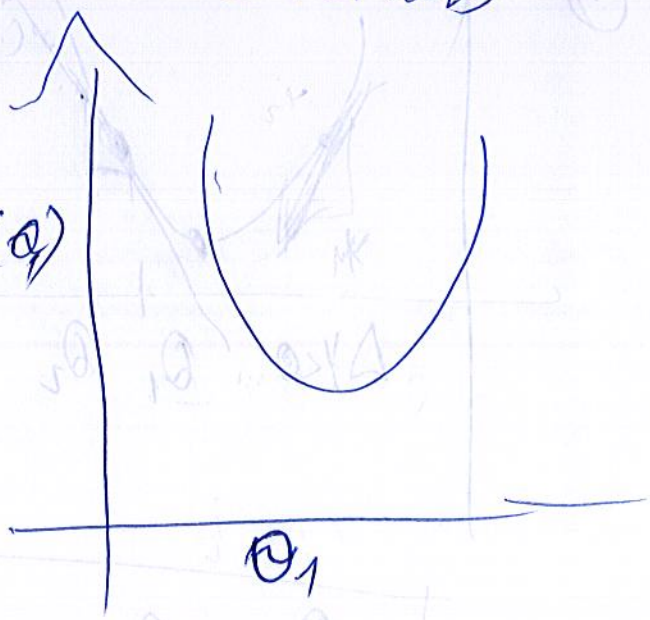
We pick up here θ_0, θ_1

such that

$$\min_{\theta_0, \theta_1} J(\theta)$$

(4) Linear Regression $h_\theta(x) = \theta_0 + \theta_1(x)$ One variable





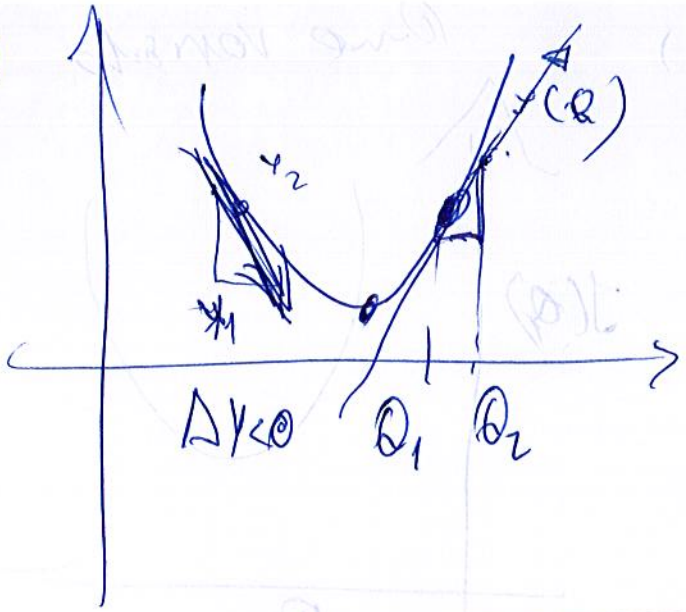
$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

$$\boxed{J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x)^{(i)} - y^{(i)} \right)^2}$$

linear regression

Predict continuous valued output

③



each

$\Delta y = y_2 - y_1$

$y_2 = J(\theta_2)$

$y_1 = J(\theta_1)$

$\Delta y = -$

$$\theta_i = \theta_i - \alpha \frac{\partial}{\partial \theta_i} J(\theta)$$

Direction of motion is opposite of $\vec{\Delta y}$, so opposite of $\frac{\partial}{\partial \theta_i} J(\theta)$

Logistic regression (classification)

$y \in \{0, 1\}$

$0 \leq h_\theta(x) \leq 1$

$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$

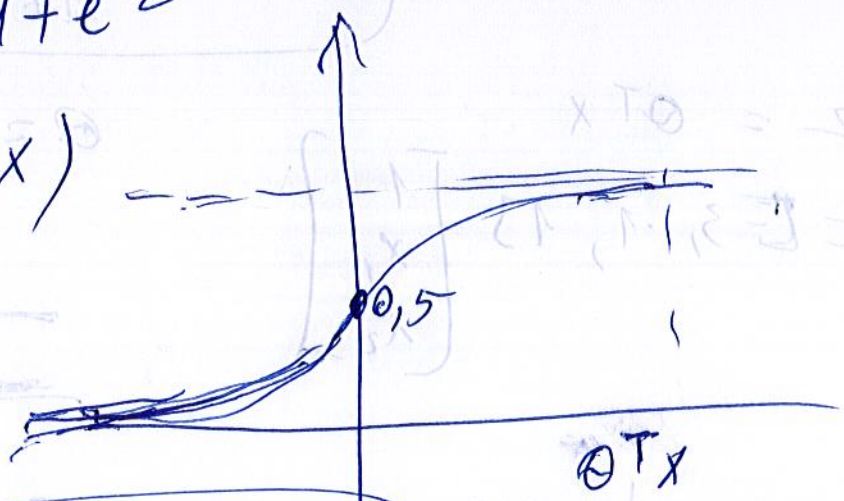
Regression = Predict continuous valued output

$h_\theta(x) = \frac{1}{1 + e^{-h_\theta(x)}}$

$z = \theta^T x$

(5) $g(z) = \frac{1}{1+e^{-z}}$ $\quad z = \theta^T x$

$h_\theta(x) = g(\theta^T x)$



$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

$h_\theta(z)$

$y = 1$ if $h_\theta(x) \geq 0,5$ $\quad \theta^T x \geq 0$

$y = 0$ $\quad h_\theta(x) < 0,5$ $\quad \theta^T x < 0$

predict $y = 1$ $\Leftrightarrow$ $\theta^T x \geq 0$ (a)

predict $y = 0$ $\Leftrightarrow$ $\theta^T x < 0$ (b)

a) $g(z) \geq 1 \Leftrightarrow \theta^T x \geq 0$ $\quad$ example: $3 \cdot 1 + 2 + x_1 - 7x_2 + 8x_3 \geq 0$

Decision boundary

$h_\theta(x) = g(z) = g(3 \cdot 1 + 2x_1 - 7x_2 + 8x_3)$

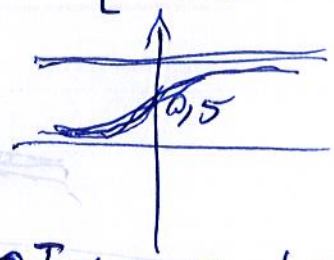
⑥ example 2

$$g(z) = \frac{1}{1+e^{-z}}$$

$z = \theta^T x$

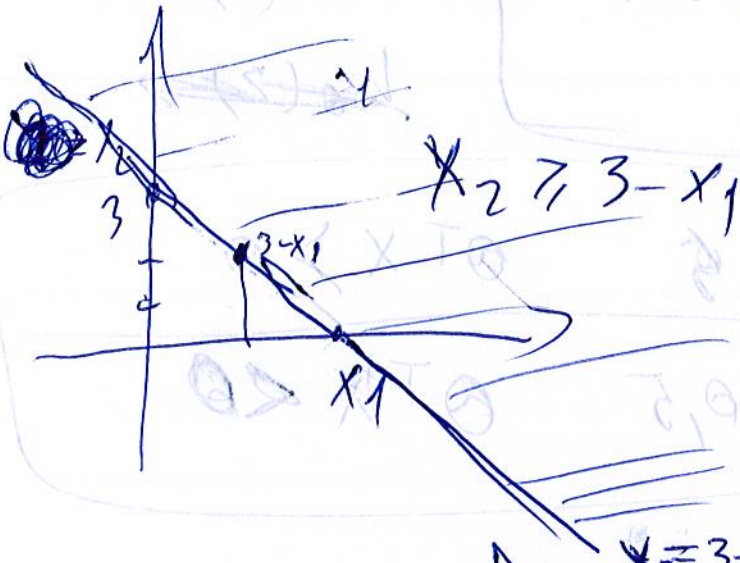
$\theta^T x = [-3, 1, 1] \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$

$\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$



$y = 1 \Leftrightarrow h_\theta(x) \geq 0,5 \Leftrightarrow \theta^T x \geq 0 \Leftrightarrow$

$\Leftrightarrow -3 + x_1 + x_2 \geq 0 \quad x_1 + x_2 \geq 3$



$x_2 \geq 3 - x_1$

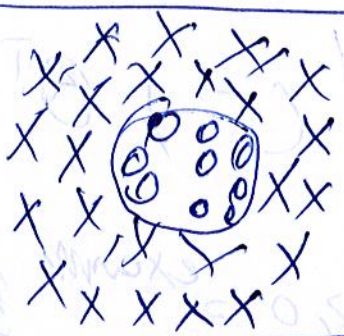
$x_2 = 3 - x_1$

$x_1 = 3 \quad x = 0$

$x_1 = 0 \quad x = 3$

Decision bound

Ex



$h_\theta(x) = g(\underbrace{\theta_0}_{=-1} + \underbrace{\theta_1}_{=0} x_1 + \underbrace{\theta_2}_{=0} x_2 + \underbrace{\theta_3}_{=1} x_1^2 + \underbrace{\theta_4}_{=1} x_2^2)$

$-1 + x_1^2 + x_2^2 \geq 0 \Leftrightarrow \boxed{x_1^2 + x_2^2 \geq 1}$

circle

Linear regression: $h_\theta(x) = [\theta_0, \dots \theta_m] \begin{bmatrix} t_0 \\ t_2 \\ \vdots \\ t_m \end{bmatrix}$

$\theta_i \in R$

$t \in \{x_i, x_i^2, x_2 x_1, x_3^2 x_2^2 \dots\}$

t is monomial

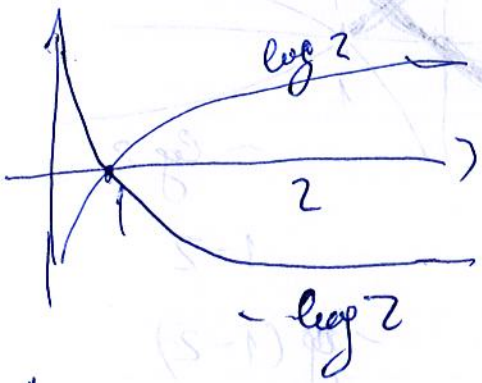
$$J(\theta) = \frac{1}{m} \left[\sum_{i=1}^{m} \frac{1}{2} (h_\theta(x^i) - y^i)^2 \right] = \sum_{i=1}^{m} cost(h_\theta x^{(i)}, y^{(i)})$$

$$cost(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{2} (h_\theta(x^i) - y^{(i)})^2$$

Logistic Regression i.e. classification

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)), & y = 1 \\ -\log(1 - h_\theta(x)), & y = 0 \end{cases}$$

$$Cost(h_\theta(x), y) = -y \log(h_\theta(x)) - (1-y) \log(1 - h_\theta(x))$$

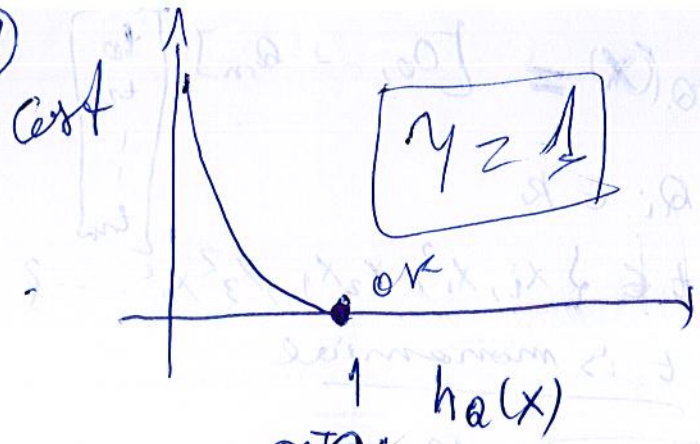


$0 \le h_\theta(x) \le 1$, because

$$h_\theta(x) = \frac{1}{1 - e^{-\theta^T x}}$$

So $h_\theta(x)$ still $= g(\theta^T x)$

where g is sigmoid function
or Kellk



$cost(h_\theta(x), y) = -\log(h_\theta(x))$

$h_\theta(x) = 1 \quad cost(\,) = 0$

$h_\theta(x) \to 0 \quad cost(\,) \to \infty$

but $y = 1$

CASE $\to y = 1$

a) $y = 1$, $h_\theta(x) = 1 \Rightarrow cost(h_\theta(x), y) = 0$ is OK

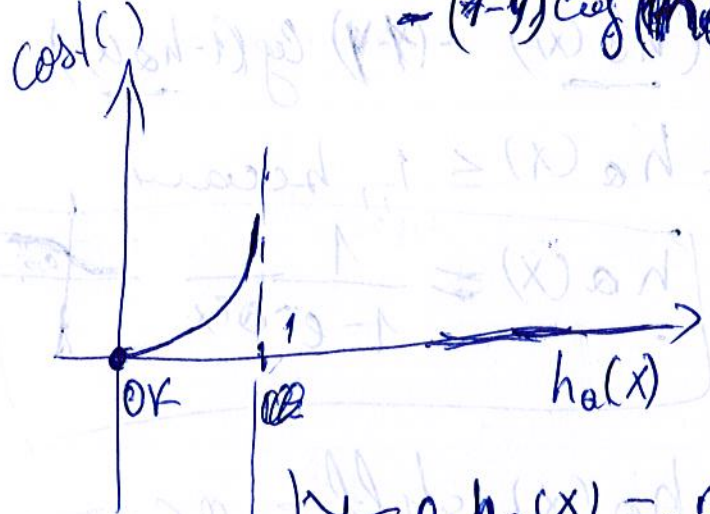
b) $y = 1$, $h_\theta(x) \to 0 \Rightarrow cost(h_\theta(x), y) \to \infty$

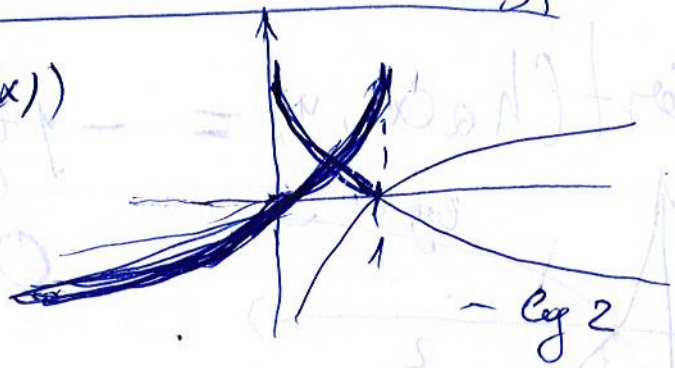
Not ok so we penalize learning algorithm by a very large cost

a) $P(y = 1 \mid x; \theta) = 1$ OK

b) $P(y = 1 \mid x; \theta) = 0$ Penalize he large cost

$-(1-y)\log(1 - h_\theta(x))$





$y = 0, h_\theta(x) = 0 \Rightarrow cost(h_\theta(x), y) = 0$ OK

$y = 0 \; h_\theta(x) = 1 \to cost(h_\theta(x), y) \to +\infty$

Penalize by large cost

⑨ Logistic Regression Cost function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{cost}(h_\theta(x^{(i)}), y^{(i)})$$ → this is more general than logistic regression

Let $x = x^{(i)}$, $y = y^{(i)}$, $y \in \{0,1\}$ (i.e. logistic regression)

then

$$\text{Cost}(h_\theta(x), y) = -y \log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))$$

$$= -[y \log(h_\theta(x)) + (1-y)\log(1-h_\theta(x))]$$

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\left(y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right)\right]$$

↑ Logestic Regression Cost Function ↑

Gradient Descent (general notes)

G.D. is used to find those value of $\theta \in R^{n+1}$ for which $J(\theta)$ achieves minimal value.

Let $\theta \in R^{1+1}$ $\quad \theta = \begin{bmatrix}\theta_0 \\ \theta_1\end{bmatrix} = \begin{bmatrix}1 \\ \theta_1\end{bmatrix}$

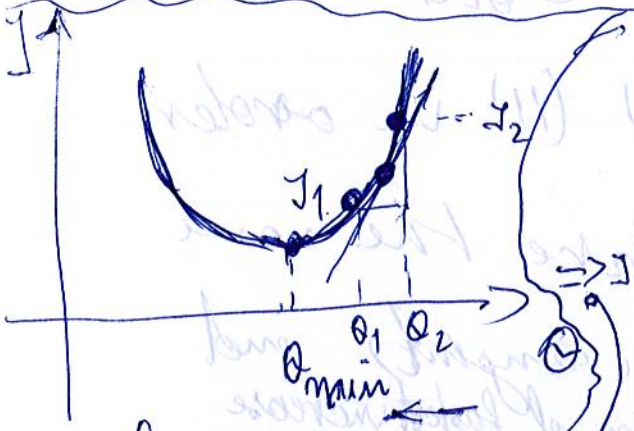


figure a

case (a)

$\Delta J = J_2 - J_1 > 0$

$\Delta\theta = \theta_2 - \theta_1 > 0$ $\Big\}$ $\frac{\Delta J}{\Delta\theta} > 0 \Rightarrow$

$\Rightarrow J' = \frac{dJ}{d\theta} > 0$

$\frac{dJ}{d\theta} = 0 \Leftrightarrow \theta = \theta_{min}$

$\frac{dJ}{d\theta}$ { • is positive • decreases as value when we move from right toward θ_{min}

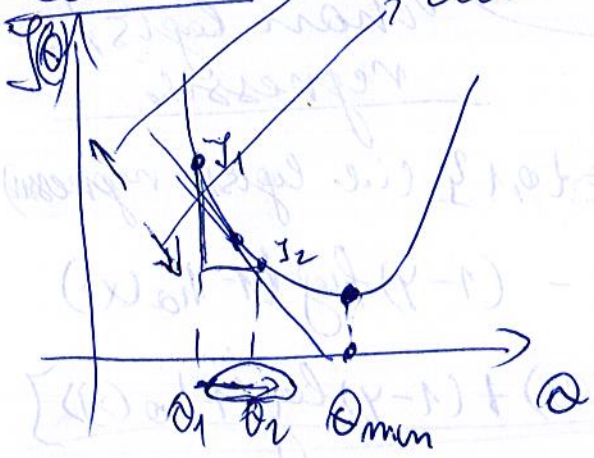
P.S $J'(\theta_1) = \lim_{\theta_2 \to \theta_1} \frac{J(\theta_2) - J(\theta_1)}{\theta_2 - \theta_1}$

i.e $\frac{dJ}{d\theta_1} = \lim_{\Delta\theta \to 0} \frac{\Delta J}{\Delta\theta}$

i.e in order to move toward θ_{min} we should remove from actual value of θ a positive value

10/ Thus we can decide $\Theta := \Theta - \alpha \frac{dJ}{d\Theta}$ (1)
(i.e. we decrease actual Θ) $\alpha > 0$

Case 1



$\Delta J = J_2 - J_1 < 0$
$\Delta\Theta = \Theta_2 - \Theta_1 > 0$ $\Rightarrow$

$\Rightarrow \frac{\Delta J}{\Delta \Theta} < 0 \Rightarrow \frac{dJ}{d\Theta} < 0$

$\frac{\Delta J}{\Delta \Theta} = 0 \Leftrightarrow \Theta = \Theta_{min}$

$\frac{dJ}{d\Theta}$
- is negative
- decreases as absolute value when we move toward Θ_{min} from left.

(for this function we know that this local min, Θ_{min}, is general minimum as function is convex.

Hence, to go toward Θ_{min}, we add a positive value to actual value of Θ to achieve Θ_{min}. (so we increase actual Θ)

$\frac{dJ}{d\Theta} < 0 \Rightarrow -\Theta \frac{dJ}{d\Theta} > 0$

Thus $\Theta := \Theta - \alpha \frac{dJ}{d\Theta}$ $(\alpha > 0)$ (II) in order to come closer to Θ_{min}

We see that (I) and (II) are the same

derivate $\frac{\partial J}{\partial \Theta}$ → shows the slope (tangent) and the direction of fastest increase

More generally $\Theta \in R^{n+1}$

$\nabla J(\Theta) = \left(\frac{\partial J}{\partial \Theta_0}, \frac{\partial J}{\partial \Theta_1}, \frac{\partial J}{\partial \Theta_2}, \frac{\partial J}{\partial \Theta_3}, \dots, \frac{\partial J}{\partial \Theta_n}\right)$ → Gradient
- slope
- the direction of fastest increase

(11) # HOW TO CALCULATE $\left(\frac{\partial J}{\partial \theta_j}\right)$?

We need $\frac{\partial J}{\partial \theta_j}$ while running θ_j ($\forall j = 0 \dots n$) in order to (ρ^0) converge to $\theta_{min} = [\theta_{0min}, \dots \theta_{nmin}]$

I.C We use $\frac{\partial J}{\partial \theta_j}$, ~~for example~~, during G. Descent. or any other algorithm with similar purpose

I way → calculate directly $\frac{\partial J}{\partial \theta_j}$ as you do in calculus

Case a <-> Linear regression where

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \dots + \theta_n x_n$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

Let denote $U = h_\theta(x^{(i)}) - y^{(i)}$ | i.e U is function of θ and x (y is const)

$U(\theta, x) = h_\theta(x^{(i)}) - y^{(i)}$

- $(U^2)' = 2U \cdot U'$ | if we derivate by θ $\frac{\partial (U^2)}{\partial \theta} = 2U \cdot \frac{\partial U}{\partial \theta}$

in fact we are interested to derivate by θ the let denote $U(\theta) := h_\theta(x^{(i)}) - y^{(i)}$

- $(\sum U^2)' = 2 \sum U^2 \cdot U'$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} U^2(\theta, x^{(i)}) \quad ; \quad cost^{(i)} = \frac{1}{2} U^2(\theta, x^{(i)})$$

$x^{(i)} = (x_0^{(i)}, x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)})$ | $U(\theta) \overset{\text{denote}}{=} U(\theta, x^{(i)}) = h_\theta(x^{(i)}) - y^{(i)}$

(12) Let's derivate initially only $cost^{(i)} = \frac{1}{2} U^2(\Theta, x^{(i)})$

$(\forall i = 1 \dots m)\ x^{(i)} \in \mathbb{R}^{n+1} \qquad x_0^{(i)} = 1 \ \forall i$

Let's omit superscript (i) for a moment i.e.

let $x = x^{(i)} = (x_0, x_1, x_2, \dots x_n) \qquad y = y^{(i)}$

$$\frac{\partial cost}{\partial \Theta_j} = \frac{\partial U^2(\Theta, x)}{\partial \Theta_j} \cdot \frac{1}{2} = \frac{1}{2} 2U \cdot \frac{\partial U}{\partial \Theta_j} = U \cdot \frac{\partial (h_\Theta(x) - y)}{\partial \Theta_j} =$$

$$= U \cdot \frac{\partial h_\Theta(x)}{\partial \Theta_j}$$ (because y is constant as we derivate by Θ_j)

$$= U \cdot \frac{\partial (\Theta_0 x_0 + \Theta_1 x_1 + \Theta_2 x_2 + \dots + \Theta_n x_n)}{\partial \Theta_j} = U \cdot (0 + 0 + \dots + x_j + 0 + 0)$$

$$= U \cdot x_j = (h_\Theta(x) - y) \cdot x_j$$

($U = h_\Theta(x) - y$)

i.e $\frac{\partial cost}{\partial \Theta_j} = (h_\Theta(x) - y) x_j$ → this is true for any $x^{(i)} = x$ ⇒

$$\Rightarrow \boxed{\frac{\partial cost^{(i)}}{\partial \Theta_j} = h_\Theta(x^{(i)}) - y^{(i)}) x_j^{(i)}} \Rightarrow$$

→ that how we calculate in this scenario

$$\boxed{\frac{\partial J}{\partial \Theta_j} = \frac{1}{m} \sum_{i=1}^{m} (h_\Theta(x^{(i)}) - y^{(i)}) x_j^{(i)}} \Rightarrow \frac{\partial J}{\partial \Theta_j}$$ for Linear Regression

where $h_\Theta(x) = \Theta_0 x_0 + \Theta_1 x_1 + \dots + \Theta_n x_n$

(P.S. We use the formula above during gradient Descent

$$\Theta_j = \Theta_j - \alpha \frac{\partial J}{\partial \Theta_j} \qquad \Theta \in \mathbb{R}^{n+1} : \begin{bmatrix} \Theta_0 \\ \Theta_1 \\ \vdots \\ \Theta_n \end{bmatrix} = \begin{bmatrix} \Theta_0 \\ \Theta_1 \\ \vdots \\ \Theta_n \end{bmatrix} - \alpha \begin{bmatrix} \frac{\partial J}{\partial \Theta_0} \\ \frac{\partial J}{\partial \Theta_1} \\ \vdots \\ \frac{\partial J}{\partial \Theta_n} \end{bmatrix}$$

$\begin{bmatrix} \frac{\partial J}{\partial \Theta_0} \\ \frac{\partial J}{\partial \Theta_n} \end{bmatrix} \overset{def}{=} \nabla J$ → gradient of cost $J(\Theta)$.

(13)

$$\begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix} = \begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix} - \alpha \cdot \begin{bmatrix}\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\,x_0^{(i)}\\ \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\,x_1^{(i)}\\ \vdots\\ \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\,x_n^{(i)}\end{bmatrix} \qquad x_0^{(i)} = 1 \ \forall i$$

θ_{new} | $\theta \in \mathbb{R}^{n+1}$ | ∇

Note For parallel update, keep θ unchanged then $\theta = \theta_{new}$, because we use θ in $h_\theta(x)$.

$j=0$

$$\sum_{i=1}^{m}\left[(h_\theta(x^{(i)}) - y^{(i)})\cdot x_0^{(i)}\right] = (\theta_0 x_0^{(1)} + \dots + \theta_n x_n^{(1)} - y^{(1)})\,x_0^{(1)} + (\theta_0 x_0^{(2)} + \dots + \theta_n x^{(2)} - y^{(2)})\,x_0^{(2)} + \dots + (\theta_0 x_0^{(m)} + \dots + \theta_n x^{(m)} - y^{(m)})\,x_0^{(m)}$$

$$= (\theta^T \cdot x^{(1)} - y^{(1)})\,x_0^{(1)} + (\theta^T x^{(2)} - y^{(2)})\,x_0^{(2)} + \dots + (\theta^T \cdot x^{(m)} - y^{(m)})\,x_0^{(m)}$$

$$[\theta_0 \ \dots \ \theta_n]\,[x^{(1)}, x^{(2)}, \dots, x^{(m)}] = \theta^T X$$

(θ: n, X: $n \times m$ (m times))

$$X := [x^{(1)}\ x^{(2)}\ \dots\ x^{(m)}] \qquad y^T = [y^{(1)} \ \dots \ y^{(m)}]$$

$$\theta^T X - y^T = [\theta^T x^{(1)} - y^{(1)},\ \theta^T x^{(2)} - y^{(2)},\ \dots\ \theta^T x^{(m)} - y^{(m)}]$$

$$\times \begin{bmatrix}x_0^{(1)}\\ x_0^{(2)}\\ \vdots\\ x_0^{(m)}\end{bmatrix}$$

$j=0$

$j=0$: $(\theta^T x - y^T) \cdot x_0^{(i)}$ $\quad (X^T)^{(i)}$

$$\boxed{\frac{1}{m}(\theta^T X - y^T)\,X^T = \nabla} \qquad \theta = \theta - \alpha \nabla$$

$\Theta = \Theta - \alpha \cdot \nabla J$; $\boxed{\nabla J = \frac{1}{m}(\Theta^T X - y^T)\, X^T}$

Where $\nabla J \overset{Def}{=} \begin{bmatrix} \frac{\partial J}{\partial \Theta_0} \\ \frac{\partial J}{\partial \Theta_1} \\ \vdots \\ \frac{\partial J}{\partial \Theta_n} \end{bmatrix}$; While to calculate ∇J we accomplish $\frac{1}{m}(\Theta^T X - y^T)\, X^T$

We use this formula to calculate ∇J

Wher $\Theta^T = [\Theta_0, \dots \Theta_n]$; $y^T = [y^{(1)} \dots y^{(m)}]$

$$X = \begin{bmatrix} x_0^{(1)} & x_0^{(2)} & \dots & x_0^{(m)} \\ x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(m)} \\ \vdots & \vdots & & \vdots \\ x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(m)} \end{bmatrix} \quad X \in R^{(n+1) \times m}$$

(columns $x^{(1)}$, $x^{(2)}$, …, $x^{(m)}$; n rows, m columns)

$$X^T = \begin{bmatrix} x_0^{(1)} & x_1^{(1)} & \dots & x_n^{(1)} \\ x_0^{(2)} & x_1^{(2)} & \dots & x_n^{(2)} \\ x_0^{(3)} & x_1^{(3)} & \dots & x_n^{(3)} \\ \vdots & & & \vdots \\ x_0^{(m)} & x_1^{(m)}\; x_2^{(m)} & \dots & x_n^{(m)} \end{bmatrix} \quad X^T \in R^{m \times (n+1)}$$

(rows $x^{(1)}$, $x^{(2)}$, …, $x^{(m)}$; m rows, n columns)

i.e. $\Theta = \Theta - \alpha \cdot \nabla J$ where $\nabla J = \frac{1}{m} \cdot (\Theta^T X - y^T)\, X^T$

(15) **Note:** If we define: [illegible]

but as $\Theta = \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix}$ and $y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in R^m$

and $x^{(i)} \in R^{n+1}$ $\quad x^{(i)} = \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}$

This is transpose version of what I used on previous pages

$$X = \begin{bmatrix} -(x^{(1)})^T- \\ -(x^{(2)})^T- \\ \vdots \\ -(x^{(m)})^T- \end{bmatrix} \Rightarrow X^T = \begin{bmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \cdots & x^{(m)} \\ | & | & & | \end{bmatrix}$$

$X \in R^{m \times (n+1)}$ $\quad X^T \in R^{(n+1) \times m}$

Then the previous equality $\nabla J = \frac{1}{m}(\Theta^T X - y^T) X^T$ will look different (X here transposed version of the one in previous pages)

$j=0$ $\sum_{i=1}^{m} (\Theta^T x^{(i)} - y^{(i)}) x_0^{(i)}$

$j=1$ $\sum_{i=1}^{m} (\Theta^T x^{(i)} - y^{(i)}) x_1^{(i)}$

$j=2$ $\sum_{i=1}^{m} (\Theta^T x^{(i)} - y^{(i)}) x_2^{(i)}$

$$X = \begin{bmatrix} x_0^{(1)} & x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ x_0^{(2)} & x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ \vdots & & & & \\ x_0^{(m)} & x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \quad \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} = \Theta$$

$$X \cdot \Theta = \begin{bmatrix} x^{(1)} \cdot \Theta \\ x^{(2)} \cdot \Theta \\ \vdots \\ x^{(m)} \cdot \Theta \end{bmatrix} \qquad x^{(i)} \Theta = \sum_{j=1}^{n} x_j^{(i)} \theta_j$$

$= X \cdot \Theta \in R^m$

i-th column of $X \cdot \Theta = \Theta^T x^{(i)}$ | $x_0^{(i)}$ where $i = 1 \ldots m$ is x_0^1 to x_0^m the 0-column of X or 0-row of X^T

1) Let j be fixed then $x_j^{(0\ldots m)} = x_j^{(i)}$ is the j-th column of X.
or the j-th row of X^T.

$$\sum (\theta^T x^{(i)} - y^{(i)})\, x_j^{(i)} \quad = \; ?$$

$$X\cdot\theta - y \Rightarrow \begin{bmatrix} x^{(1)}\theta \\ x^{(2)}\theta \\ \vdots \\ x^{m}\theta \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} = \begin{bmatrix} x^{(1)}\theta - y^{(1)} \\ x^{(2)}\theta - y^{(2)} \\ \vdots \\ x^{m}\theta - y^{(m)} \end{bmatrix} = B$$

2) $\theta^T x^{(i)} - y^{(i)}$ is the i-th row of B

1) ∧ 2)

$$j = 0 \ldots n \qquad \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})\, x_j^{(i)}$$

$$X^T \cdot \begin{bmatrix} x^{(1)}\theta - y^{(1)} \\ x^{2}\theta - y^{(2)} \\ \vdots \\ x^{n}\theta - y^{(m)} \end{bmatrix} = B = X\theta - y$$

$$X^T = \begin{bmatrix} x_0^{(1)} & x_0^{(2)} & x_0^{(3)} & \cdots & x_0^{(m)} \\ x_1^{(1)} & x_1^{(2)} & x_1^{(3)} & \cdots & x_1^{(m)} \\ \vdots & & & & \\ x_n^{(1)} & x_n^{(2)} & x_n^{(3)} & \cdots & x_n^{(m)} \end{bmatrix} \qquad \begin{matrix} \sum_{i=0}^{m} (x^{(i)}\theta - y^{(i)})\, x_0^{(i)} \\ \sum_{i=0}^{m} (x^{(i)}\theta - y^{(i)})\, x_1^{(i)} \\ \\ \sum_{i=0}^{m} (x^{(i)}\theta - y^{(i)})\, x_n^{(m)} \end{matrix}$$

$x^{(1)} \quad x^{(3)} \quad x^{(m)}$

$$\frac{1}{m}\begin{bmatrix} \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})\, x_0^{(i)} \\ \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})\, x_j^{(i)} \\ \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})\, x_n^{(i)} \end{bmatrix} = \nabla J = \frac{1}{m} X^T B = \frac{1}{m} X^T (X\theta - y) \Rightarrow$$

$$= \boxed{\frac{1}{m} X^T \cdot (h_\theta(x) - y) = \nabla J}$$

$$\nabla J = \frac{1}{m} X^T (X\theta - y)$$

(1.7)

(1) $\nabla J = \frac{1}{m} X^T (X \cdot \theta - y)$

if $x^{(i)} = \begin{bmatrix} x_0^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}$ and $X = \begin{bmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(m)})^T \end{bmatrix}$

(2) $\nabla J = \frac{1}{m} (\theta^T X' - y^T) X'^T$

if $X' = X^T \Leftrightarrow X = (X')^T$

Check if our calculation was Ok.

$$\frac{1}{m} \left(X^T (X \cdot \theta - y) \right)^T = \frac{1}{m} (X \cdot \theta - y)^T \cdot (X^T)^T = \frac{1}{m} (X \cdot \theta - y)^T \cdot X =$$

$$= \frac{1}{m} (\theta^T X^T - y^T) \cdot X = \frac{1}{m} (\theta^T X' - y^T) \cdot X'^T \equiv (2).$$

P.S. If we write $h_\theta(X) = \begin{bmatrix} h_\theta(x^{(1)}) \\ \vdots \\ h_\theta(x^{(m)}) \end{bmatrix}$, then

$$\nabla J = \boxed{\frac{1}{m} X^T (X \cdot \theta - y)} = \boxed{\frac{1}{m} X^T (h_\theta(X) - y)}$$

Note: $\nabla J = \frac{\partial J}{\partial \theta} \frac{\partial J(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial J(\theta)}{\partial \theta_n} \end{bmatrix} = \left(\frac{\partial J(\theta)}{\partial \theta_1}, \dots, \frac{\partial J(\theta)}{\partial \theta_n} \right)$

$$\boxed{\theta := \theta - \alpha \nabla J} \equiv \boxed{\theta := \theta - \frac{\alpha}{m} X^T (h_\theta(X) - y)}$$

(18)

SUMMARY

How to calculate $\frac{\partial J}{\partial \theta_j}$ for Linear Regression of type $(h_\theta(x) = \sum_{j=0}^{n} \theta_j x_j)$? $x_j = 1$

I way → Like in calculus exercises; we derivate and find our

$$\frac{\partial J}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} \quad (1)$$

$$\nabla J = \begin{bmatrix} \frac{\partial J}{\partial \theta_0} \\ \frac{\partial J}{\partial \theta_1} \\ \vdots \\ \frac{\partial J}{\partial \theta_n} \end{bmatrix} \quad (2)$$

Note:
∇J → is gradient of cost function

Gradient shows the direction of fastest increase of function $J(\theta)$ (This is proven in calculus) at a certain point.

so we have point (current point) $\theta = \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix} \in \mathbb{R}^{n+1}$

and we derivate at this point and we get formula (1) which we apply to concrete values of θ, x, y and we get a concrete value which we use to update θ with the purpose of achieving (getting to) θ_{min}.

The (2) using (1) and matrix calculus, we can write and calculate in compact way as following:

$$\nabla J = \frac{1}{m} X^T (X \cdot \theta - y)$$

$\theta \in \mathbb{R}^{n+1}$, $y \in \mathbb{R}^m$
$X \in \mathbb{R}^{m \times (n+1)}$ } Then we calculate $\theta = \theta - \alpha \nabla J$

(19) case b **Logistic Regression** → $\frac{\partial J(\theta)}{\partial \theta_j} = ?$

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} \quad (1)$$

$$\boxed{g(u) = \frac{1}{1+e^{-u}}} \quad (2) \text{ sigmoid function}$$

$$= \frac{e^u}{e^u+1} \quad (3)$$

$g'(u) = (1-g(u)) \cdot g(u)$ (7)

$\frac{\partial g(h_\theta(x))}{\partial x} = (1-g(u))\,g(u)\cdot u'(x)$ (8)

$\theta \in R^{n+1}$, $x \in R^{(n+1)}$, $x^{(i)} \in R^{(n+1)}$

$$\boxed{J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right]} \quad (4)$$

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} = g(\theta^T x) \qquad \boxed{h_\theta(x) = g(\theta^T x)} \quad (5)$$

$$\boxed{J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(g(\theta^T x^{(i)})) + (1-y^{(i)})\log(1-g(\theta^T x^{(i)}))\right]} \quad (6)$$

$\log(z) = \ln(z) = \log_e(z)$

$$cost^{(i)} = -\left[y^{(i)}\log(g(\theta^T x^{(i)})) + (1-y^{(i)})\log(1-g(\theta^T x^{(i)}))\right]$$

simply

$$cost^{(i)} = -\left[y\log(g(\theta^T x)) + (1-y)\log(1-g(\theta^T x))\right]$$

$$cost_a^{(i)} = y\log g(\theta^T x) \quad ; \quad cost_b^{(i)} = (1-y)\log(1-g(\theta^T x))$$

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} cost^{(i)} = -\frac{1}{m}\sum_{i=1}^{m}(cost_a^{(i)} + cost_b^{(i)})$$

$$\frac{\partial cost_a^{(i)}}{\partial \theta_j} = \frac{\partial(y\log(g(\theta^T x)))}{\partial \theta_j} = \frac{y}{g(\theta^T x)} \cdot \frac{\partial g(\theta^T x)}{\partial \theta_j} \overset{(7)}{=} \frac{y}{g(\theta^T x)}(1-g(\theta^T x))g(\theta^T x)$$

$$\times \frac{\partial(\theta^T x)}{\partial \theta_j} = y(1-g(\theta^T x))\cdot x_j = yx_j - yg(\theta^T x)x_j$$

$$= yx_j - yh_\theta(x)x_j \quad (9)$$

$g(z) = h_\theta(x) = g(z)$

$z = \theta^T x$

$$\frac{\partial\, cost b^{(i)}}{\partial \theta_j} = \frac{\partial\,(1-y)\log(1-g(\theta^T x))}{\partial \theta_j} = \frac{1-y}{1-g(\theta^T x)} \frac{\partial\,(1-g(\theta^T x))}{\partial \theta_j} =$$

$$= \frac{1-y}{1-g(\theta^T x)} \cdot -\frac{\partial g(\theta^T x)}{\partial \theta_j} = \frac{y-1}{1-g(\theta^T x)} \cdot (1-g(\theta^T x))\, g(\theta^T x) \cdot \frac{\partial(\theta^T x)}{\partial \theta_j} =$$

$$= (y-1)\, g(\theta^T x) \cdot x_j = (y-1)\, h_\theta(x) \cdot x_j =$$

$$= y\, h_\theta(x) \cdot x_j - h_\theta(x) \cdot x_j \quad (10)$$

$$(9) + (10) = \frac{\partial(cost a^{(i)} + cost b^{(i)})}{\partial \theta_j} = y \cdot x_j - h_\theta(x) x_j = cost^{(i)}$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\left[\sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)}))\, x_j^{(i)}\right] \Rightarrow \frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m} \cdot \sum_{i}^{m} cost^{(i)} \cdot x_j^{(i)}$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} \left[(h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}\right]$$

see on previous pages how we go to matrix form.

$$\nabla J(\theta) = \frac{\partial(J(\theta))}{\partial \theta} = \frac{1}{m} X^T [h_\theta(X) - y] \qquad X =$$

because $\nabla J(\theta) = \dfrac{\partial J(\theta)}{\partial \theta} =$

$$\begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial J(\theta)}{\partial \theta_n} \end{bmatrix} = \begin{bmatrix} \frac{1}{m}\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_1^{(i)} \\ \vdots \\ \frac{1}{m}\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_n^{(i)} \end{bmatrix}$$

P.S. $g(z) = \frac{1}{1+e^{-z}} \xrightarrow{f=1+e^{-z}} g'(z) = -\frac{f'}{f^2} = \frac{-(1+e^{-z})'}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})^2}$

$$= \frac{e^{-z}}{(1+e^{-z})} \cdot \frac{1}{(1+e^{-z})} = \frac{e^{-z}}{1+e^{-z}} \cdot g(z) = \frac{1-1+e^{-z}}{1+e^{-z}}\, g(z) =$$

$$= 1 - \frac{1}{1+e^{-z}} \cdot g(z) = (1-g(z))\, g(z); \text{ if } z = z(x), \text{ then } \frac{\partial g}{\partial x} = g'(z(x)) = (1-g(z))\, g(z) \cdot z'(x)$$

$g(z) = \frac{1}{1+e^{-z}}$

$$\left(\frac{1}{1+e^{-z}}\right)' = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{e^{-z}}{1+e^{-z}} \cdot \frac{1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}} \cdot g(z)$$

$$= \frac{1+e^{-z}-1}{1+e^{-z}} \cdot g(z)$$

$$1 - g(z) = \frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}} \qquad (1-g(z))$$

$$1 - \frac{1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}} = e^{-z} \cdot \frac{1}{1+e^{-z}}$$

$$1 - g(z) = e^{-z} \cdot g(z)$$

$$\frac{e^{-z}}{1+e^{-z}} = \frac{1+e^{-z}}{1+e^{-z}} = \frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}} = 1 - \frac{1}{1+e^{-z}}$$

$$\frac{1}{g(\theta^T x)} \frac{\partial g(\theta^T x)}{\partial \theta_j} = \frac{1}{g(\theta^T x)} (1 - g(\theta^T x)) g(\theta^T x) \frac{\partial \theta^T x}{\partial \theta_j}$$

$$(1 - g(\theta^T x)) \cdot x_j$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = 0$$

$$\frac{1}{m} X^T(X\cdot\theta - y) = 0 \iff X^T X\theta - X^T y = 0$$

$$X^T X\theta = X^T y \iff \boxed{\theta = (X^T X)^{-1} X^T \cdot y}$$

Normal equation

$$\frac{\partial J}{\partial \theta^{(3)}} = (\tilde{y}^{(i)} - y^{(i)})\, a^{(3)}$$

$$\frac{\partial J}{\partial \theta^{(2)}} = (\tilde{y}^{(i)} - y^{(i)})\, \theta^{(3)} \cdot g'(z^{(3)}) \cdot a^{(2)}$$

$$\frac{\partial J}{\partial \theta^{(1)}} = (\tilde{y}^{(i)} - y^{(i)})\, \theta^{3}$$

Summary

$$\frac{\partial J}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m} (h_\theta(x) - y) x_j$$

→ For both linear regression and logistic regression $\frac{\partial J}{\partial \theta_j}$ looks the same, but the difference is on $h_\theta(x)$.

Linear regression → $h_\theta(x) = \theta^T x$ (or $x\theta$)

Logistic regression →

- $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$ (or $\frac{1}{1+e^{x\theta}}$)
- $h_\theta(x) = \sigma(\theta^T x)$
- $h_\theta(x) = \sigma(z)$, where $z = \theta^T x$ ($z = \theta_1^T x + b$, $b = \theta_0$)
- $h_\theta(x) = f(z)$, where $f() = \sigma()$