University of New Orleans

Department of Computer Science

# FALL 2019: CSCI 6522

# Homework # 1

## Machine Learning - II

Submitted to:

**Professor Dr. Tamjidul Hoque**

By

Name: **Astha Sharma**

Student ID: **2564173**

**Part A:**

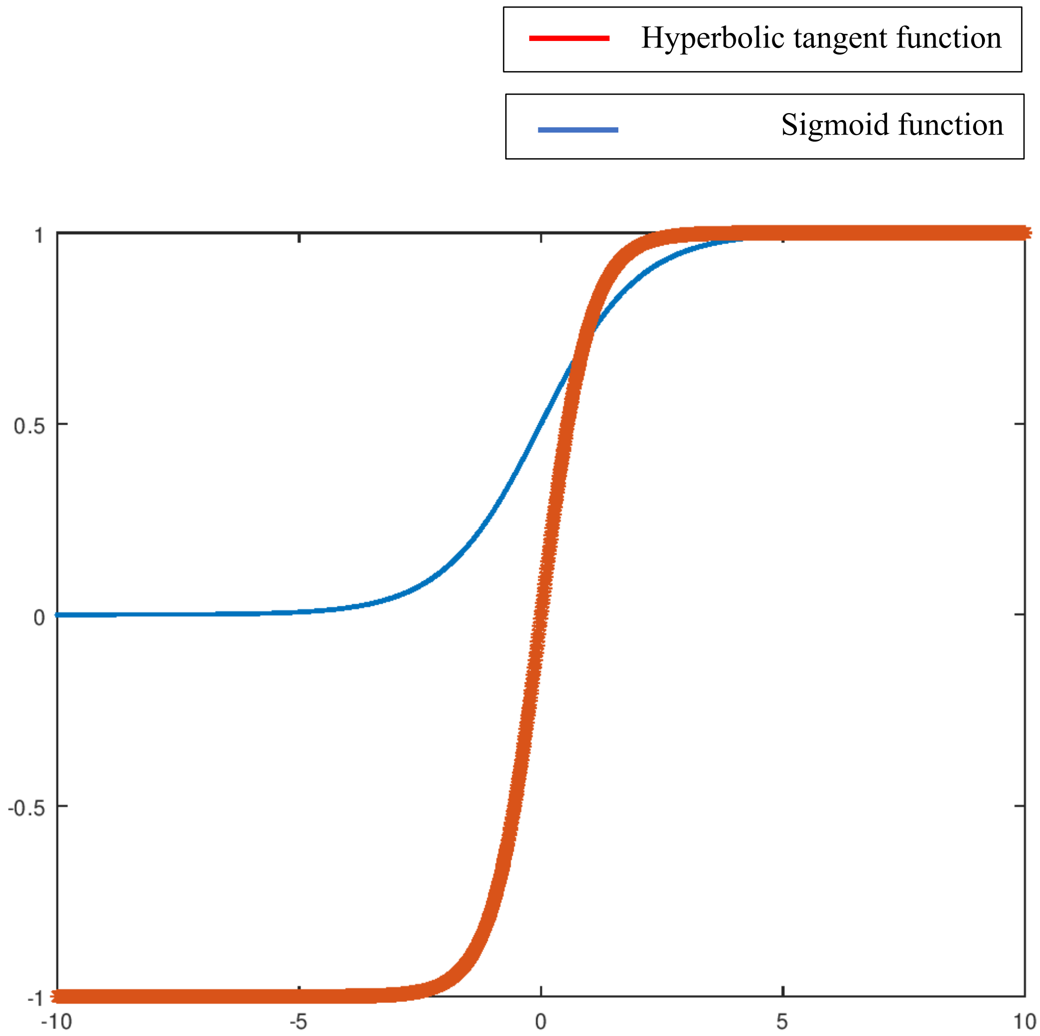


Fig: Sigmoid function vs Hyperbolic tangent function

In the above graph, the sigmoid function $f_{sig}(x) = \frac{1}{1+e^{-x}}$ is denoted by the blue curve and ranges from 0 to 1 with midpoint as 0.5. The output is often interpreted as probabilities.

And the hyperbolic tanh function $f_{tanh}(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$ is denoted by the red curve which ranges from -1 to 1 with a midpoint of 0. It is considered as the rescaling of the sigmoid function because of its greater range.

**Part (B):**

Here we have a classification problem for determining if a cancer cell is malignant or benign based on the feature set like size and age of the tumor.

So, the predictor model can be written as:

$$P(G|X)$$

where, $X = \{x_1, x_2\}$ and $\hat{G} \in \{Benign, Malignant\}$, and $x_1$= Size of the tumor and $x_2$= Age of the tumor.

And we know, the hyperbolic tangent function can be written as follows:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Also, we have the linear function as: $X_0\beta_O + X_1\beta_1 + \cdots + X_P\beta_P = X^T\beta$

Now, the predictive model can be written using hyperbolic function as:

$$P(G|X) = f_{tanh}(X^T\beta)$$

We can establish the classification rule as:

$\hat{G} = 1, if\ [f_{tanh}(X^T\beta) \geq 0]$, i.e. $(X^T\beta) \geq 0$ → malignant class (1)

$\hat{G} = 2, if\ [f_{tanh}(X^T\beta) < 0]$, i.e. $(X^T\beta) <0$ → benign class (-1)

$y_i = Bernoulli\ (\eta_i)$, where $\eta_i = f_{tanh}(X^T\beta) = p(x_i; \beta)$ is the hyperbolic tanh function

Therefore, from Bernoulli equation, we can write:

$$\boldsymbol{P(y_i) = \eta_i^{y_i}(-\eta_i)^{1-y_i}}$$

So, the likelihood is given as:

$$L(\beta) = \prod_{i=1}^{N} P(y_i) = \prod_{i=1}^{N} \eta_i^{y_i}(-\eta_i)^{1-y_i}$$

Now the Bernoulli distribution should give:

~ **1** -> if correctly classified

~ **-1** -> else case

So, checking for all possible cases.

**Case 1: Classification of correctly classified malignant**

$y_i = 1 \; and \; \eta_i = 0.95 \; (assumption \; for \; large \; positive \; number)$

Now,

$P(y_i) = \eta^{y_i}(-\eta_i)^{1-y_i} = 0.95^1(-0.95)^{1-1} =$ 0.95; which is close to 1

**Case 2: Classification of correctly classified benign**

$y_i = 0 \; and \; \eta_i = -0.95$

Now,

$P(y_i) = \eta^{y_i}(-\eta_i)^{1-y_i} = (-0.95)^0(0.95)^{1-0} =$ 0.95; which is close to 1

**Case 3: Classification of incorrectly classified malignant** $y_i = 1 and \; \eta_i = -0.95$

Now,
$P(y_i) = \eta^{y_i}(-\eta_i)^{1-y_i} = (-0.95)^1(0.95)^{1-1} =$ -0.95; which is close to -1

**Case 4: Classification of incorrectly classified benign**

$y_i = 0 \; and \; \eta_i = 0.95$

Now,
$P(y_i) = \eta^{y_i}(-\eta_i)^{1-y_i} = (0.95)^0(-0.95)^{1-0} =$ -0.95; which is close to -1

These four cases prove the Bernoulli distribution works for classification using Hyperbolic Tangent Function.

Therefore, moving forward, the Log-likelihood can be written as:

$$l(\beta) \quad = \log L(\beta) = \sum_i^N \{y_i \log(\eta_i) + (1 - y_i) \log(-\eta_i)\}$$

$$= \sum_{i=1}^N \{y_i \log p(x_i; \beta) + (1 - y_i) \log(-p(x_i; \beta))\}$$

$$= \sum_{i=1}^N \left\{y_i \log \frac{e^{X^T\beta} - e^{-X^T\beta}}{e^{X^T\beta} + e^{-X^T\beta}} + (1 - y_i) \log - \frac{e^{X^T\beta} - e^{-X^T\beta}}{e^{X^T\beta} + e^{-X^T\beta}}\right\}$$

$$= \sum_{i=1}^N \left\{y_i \log \frac{e^{X^T\beta} - e^{-X^T\beta}}{e^{X^T\beta} + e^{-X^T\beta}} + (1 - y_i) \log \frac{e^{-X^T\beta} - e^{X^T\beta}}{e^{X^T\beta} + e^{-X^T\beta}}\right\}$$

In order to maximize the log likelihood of correct classification, first we need to find the gradient of $\beta$ that maximizes it.

$$\frac{\partial l(\beta)}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} \left\{\sum_{i=1}^N \left\{\underbrace{y_i \log \frac{e^{X^T\beta} - e^{-X^T\beta}}{e^{X^T\beta} + e^{-X^T\beta}}}_{Part\ A} + (1 - y_i) \underbrace{\log \frac{e^{-X^T\beta} - e^{X^T\beta}}{e^{X^T\beta} + e^{-X^T\beta}}}_{Part\ B}\right\}\right\}$$

Calculating the gradient in part:

i) Part A:

$$\frac{\partial}{\partial \beta_j}\left\{\log \frac{e^{X^T\beta} - e^{-X^T\beta}}{e^{X^T\beta} + e^{-X^T\beta}}\right\}$$

$$= \frac{\partial}{\partial \beta_j}\{\log(e^{X^T\beta} - e^{-X^T\beta}) - \log(e^{X^T\beta} + e^{-X^T\beta})\}$$

$$= \frac{\partial\left(\log\left(e^{X^T\beta} - e^{-X^T\beta}\right)\right)}{\partial\, e^{X^T\beta} - e^{-X^T\beta}} * \frac{\partial\left(e^{X^T\beta} - e^{-X^T\beta}\right)}{\partial(\beta_j)} - \frac{\partial\left(\log\left(e^{X^T\beta} + e^{-X^T\beta}\right)\right)}{\partial\, e^{X^T\beta} + e^{-X^T\beta}} * \frac{\partial\left(e^{X^T\beta} + e^{-X^T\beta}\right)}{\partial(\beta_j)}$$

$$= \frac{\partial\left(\log\left(e^{X^T\beta} - e^{-X^T\beta}\right)\right)}{\partial\, e^{X^T\beta} - e^{-X^T\beta}} * \left\{\frac{\partial\left(e^{X^T\beta}\right)}{\partial(\beta_j)} - \frac{\partial\left(e^{-X^T\beta}\right)}{\partial(\beta_j)}\right\} - \frac{\partial\left(\log\left(e^{X^T\beta} + e^{-X^T\beta}\right)\right)}{\partial\, e^{X^T\beta} + e^{-X^T\beta}} * \left\{\frac{\partial\left(e^{X^T\beta}\right)}{\partial(\beta_j)} + \frac{\partial\left(e^{-X^T\beta}\right)}{\partial(\beta_j)}\right\}$$

$$= \frac{1}{e^{X^T\beta} - e^{-X^T\beta}} * \left\{x_j\left(e^{X^T\beta}\right) + x_j\left(e^{-X^T\beta}\right)\right\} - \frac{1}{e^{X^T\beta} + e^{-X^T\beta}} * \left\{x_j\left(e^{X^T\beta}\right) - x_j\left(e^{-X^T\beta}\right)\right\}$$

$$= x_j \frac{e^{X^T\beta} + e^{-X^T\beta}}{e^{X^T\beta} - e^{-X^T\beta}} - x_j \frac{e^{X^T\beta} - e^{-X^T\beta}}{e^{X^T\beta} + e^{-X^T\beta}}$$

$$= \boldsymbol{\frac{x_j}{\eta_i} - x_j \eta_i}$$

ii) Part B:

$$\frac{\partial}{\partial \beta_j}\left\{\log \frac{e^{-X^T\beta} - e^{X^T\beta}}{e^{X^T\beta} + e^{-X^T\beta}}\right\}$$

$$= \frac{\partial}{\partial \beta_j}\{\log(e^{-X^T\beta} - e^{X^T\beta}) - \log(e^{X^T\beta} + e^{-X^T\beta})\}$$

$$=\frac{\partial\left(\log\left(e^{-X^T\beta}-e^{X^T\beta}\right)\right)}{\partial\, e^{-X^T\beta}-e^{X^T\beta}} * \frac{\partial\left(e^{-X^T\beta}-e^{X^T\beta}\right)}{\partial(\beta_j)} - \frac{\partial\left(\log\left(e^{X^T\beta}+e^{-X^T\beta}\right)\right)}{\partial\, e^{X^T\beta}+e^{-X^T\beta}} * \frac{\partial\left(e^{X^T\beta}+e^{-X^T\beta}\right)}{\partial(\beta_j)}$$

$$=\frac{\partial\left(\log\left(e^{-X^T\beta}-e^{X^T\beta}\right)\right)}{\partial\, e^{-X^T\beta}-e^{X^T\beta}} * \left\{\frac{\partial\left(e^{-X^T\beta}\right)}{\partial(\beta_j)} - \frac{\partial\left(e^{X^T\beta}\right)}{\partial(\beta_j)}\right\} - \frac{\partial\left(\log\left(e^{X^T\beta}+e^{-X^T\beta}\right)\right)}{\partial\, e^{X^T\beta}+e^{-X^T\beta}} * \left\{\frac{\partial\left(e^{X^T\beta}\right)}{\partial(\beta_j)} + \frac{\partial\left(e^{-X^T\beta}\right)}{\partial(\beta_j)}\right\}$$

$$=\frac{1}{e^{-X^T\beta}-e^{X^T\beta}} * \left\{-x_j\left(e^{-X^T\beta}\right) - x_j\left(e^{X^T\beta}\right)\right\} - \frac{1}{e^{X^T\beta}+e^{-X^T\beta}} * \left\{x_j\left(e^{X^T\beta}\right) - x_j\left(e^{-X^T\beta}\right)\right\}$$

$$=x_j\,\frac{e^{X^T\beta}+e^{-X^T\beta}}{e^{X^T\beta}-e^{-X^T\beta}} - x_j\,\frac{e^{X^T\beta}-e^{-X^T\beta}}{e^{X^T\beta}+e^{-X^T\beta}}$$

$$=\boldsymbol{\frac{x_j}{\eta_i} - x_j\eta_i}$$

Now, we can replace the equation with these values for each part, as:

$$\begin{aligned}
\frac{\partial l(\beta)}{\partial\beta_j} &= \sum_{i=1}^{N}\left\{y_i\left(\boldsymbol{\frac{x_{i,j}}{\eta_i} - x_{i,j}\eta_i}\right) + (1-y_i)\left(\boldsymbol{\frac{x_{i,j}}{\eta_i} - x_{i,j}\eta_i}\right)\right\} \\
&= \sum_{i=1}^{N}\left\{y_i\left(\boldsymbol{\frac{x_{i,j}-x_{i,j}{\eta_i}^2}{\eta_i}}\right) + (1-y_i)\left(\boldsymbol{\frac{x_{i,j}-x_{i,j}{\eta_i}^2}{\eta_i}}\right)\right\} \\
&= \sum_{i=1}^{N}\left\{\boldsymbol{\frac{x_{i,j}-x_{i,j}{\eta_i}^2}{\eta_i}}\right\} \\
&= \sum_{i=1}^{N}\boldsymbol{x_{i,j}}\left\{\boldsymbol{\frac{1-{\eta_i}^2}{\eta_i}}\right\} \\
&= \sum_{i=1}^{N}\boldsymbol{X^T}\left\{\boldsymbol{\frac{1}{\eta} - \eta}\right\}
\end{aligned}$$

We can apply the gradient ascent to get the value(s) for which the log-likelihood is ***maximized***.

$$\beta_j(t+1) = \beta_j(t) + \alpha\frac{\partial l(\beta)}{\partial\beta_j}$$

However, Newton's method is more efficient, and we prefer to use it. For Newton's method, we will further need to compute the second-derivatives or, Hessian Matrix (**H**).

Therefore, continuing for second derivative:

$$\mathbf{H} = \frac{\partial}{\partial\beta_j}\left(\frac{\partial l(\beta)}{\partial\beta_j}\right) = \frac{\partial}{\partial\beta_j}\left(\sum_{i=1}^{N}\boldsymbol{x_{i,j}}\left\{\boldsymbol{\frac{1-{\eta_i}^2}{\eta_i}}\right\}\right) = \frac{\partial}{\partial\beta_j}\left(\sum_{i=1}^{N}\boldsymbol{x_{i,j}}\left\{\boldsymbol{\frac{1}{\eta_i} - \eta_i}\right\}\right)$$

Simplifying for $\frac{\mathbf{1}}{\boldsymbol{\eta_i}} - \boldsymbol{\eta_i}$**:**

$$= \frac{e^{X^T\beta} + e^{-X^T\beta}}{e^{X^T\beta} - e^{-X^T\beta}} - \frac{e^{X^T\beta} - e^{-X^T\beta}}{e^{X^T\beta} + e^{-X^T\beta}}$$

$$= \frac{(e^{X^T\beta} + e^{-X^T\beta})^2 - (e^{X^T\beta} - e^{-X^T\beta})^2}{(e^{X^T\beta} - e^{-X^T\beta})(e^{X^T\beta} + e^{-X^T\beta})}$$

$$= \frac{(e^{X^T\beta})^2 + 2 * e^{X^T\beta} * e^{-X^T\beta} + (e^{-X^T\beta})^2 \; - \; (e^{X^T\beta})^2 + 2 * e^{X^T\beta} * e^{-X^T\beta} - (e^{-X^T\beta})^2}{(e^{X^T\beta} - e^{-X^T\beta})(e^{X^T\beta} + e^{-X^T\beta})}$$

$$= \frac{4}{(e^{X^T\beta} - e^{-X^T\beta})(e^{X^T\beta} + e^{-X^T\beta})}$$

$$= \frac{4}{(e^{2X^T\beta} - e^{-2X^T\beta})}$$

So, getting back to the second derivative,

$$\mathbf{H} \quad = \quad \frac{\partial}{\partial \beta_j}\left(\sum_{i=1}^{N} \boldsymbol{x_{i,j}} \left\{\frac{\mathbf{1}}{\boldsymbol{\eta_i}} - \boldsymbol{\eta_i}\right\}\right)$$

$$= \quad \frac{\partial}{\partial \beta_j}\left(\sum_{i=1}^{N} \boldsymbol{x_{i,j}} \left\{\frac{4}{(e^{2x_i^T\beta} - e^{-2x_i^T\beta})}\right\}\right)$$

$$= \quad \frac{\partial}{\partial \beta_j} \sum_{i=1}^{N} \mathbf{4}\, \boldsymbol{x_{i,j}} \; \left(e^{2x_i^T\beta} - e^{-2x_i^T\beta}\right)^{\mathbf{-1}}$$

$$= \quad \sum_{i-1}^{N} \left\{-4x_j \left(e^{2X_i^T\beta} - e^{-2X_i^T\beta}\right)^{-2} \left\{\frac{\boldsymbol{\partial}}{\boldsymbol{\partial\beta_j}} e^{2X_i^T\beta} - \frac{\boldsymbol{\partial}}{\boldsymbol{\partial\beta_j}} e^{-2X_i^T\beta}\right\}\right\}$$

$$= \quad \sum_{i-1}^{N} \left\{-\frac{4x_j}{\left(e^{2X_i^T\beta} - e^{-2X_i^T\beta}\right)^2} \left\{2x_j e^{2X_i^T\beta} - (-2x_j) e^{-2X_i^T\beta}\right\}\right\}$$

$$= \quad \sum_{i-1}^{N} \left\{-\frac{4x_j}{\left(e^{2X_i^T\beta} - e^{-2X_i^T\beta}\right)^2} \left\{2x_j e^{2X_i^T\beta} + 2x_j e^{-2X_i^T\beta}\right\}\right\}$$

$$= \quad \sum_{i-1}^{N} \left\{-\frac{8\, x_j x_j \left\{e^{2X_i^T\beta} + e^{-2X_i^T\beta}\right\}}{\left(e^{2X_i^T\beta} - e^{-2X_i^T\beta}\right)^2}\right\}$$

$$= \quad \sum_{i-1}^{N}\left\{-8\, x_j x_j * \frac{\left\{\left(e^{X_i^T\beta}\right)^2 + \left(e^{-X_i^T\beta}\right)^2\right\}}{\left\{\left(e^{X_i^T\beta}\right)^2 - \left(e^{-X_i^T\beta}\right)^2\right\}^2}\right\}$$

$$= \quad \sum_{i-1}^{N}\left\{-8\, x_j x_j * \frac{\left\{e^{X_i^T\beta} + e^{-X_i^T\beta}\right\}^2 - 2e^{X_i^T\beta}.e^{-X_i^T\beta}}{\left\{\left(e^{X_i^T\beta} - e^{-X_i^T\beta}\right)\left(e^{X_i^T\beta} + e^{-X_i^T\beta}\right)\right\}^2}\right\}$$

$$= \quad \sum_{i-1}^{N}\left\{-8\, x_j x_j * \left(\frac{\left\{e^{X_i^T\beta} + e^{-X_i^T\beta}\right\}^2}{\left\{\left(e^{X_i^T\beta} - e^{-X_i^T\beta}\right)\left(e^{X_i^T\beta} + e^{-X_i^T\beta}\right)\right\}^2} - \frac{2}{\left\{\left(e^{X_i^T\beta} - e^{-X_i^T\beta}\right)\left(e^{X_i^T\beta} + e^{-X_i^T\beta}\right)\right\}^2}\right)\right\}$$

$$=. \quad \sum_{i-1}^{N}\left\{-8\, x_j x_j * \left(\frac{1}{\left\{\left(e^{X_i^T\beta} - e^{-X_i^T\beta}\right)\right\}^2} - \frac{2}{\left\{\left(e^{X_i^T\beta} - e^{-X_i^T\beta}\right)\left(e^{X_i^T\beta} + e^{-X_i^T\beta}\right)\right\}^2}\right)\right\}$$

1-$\eta^2$ can also be written as:

$$\mathbf{1\text{-}\eta^2} = 1 - \left(\frac{e^{X_i^T\beta} - e^{-X_i^T\beta}}{e^{X_i^T\beta} + e^{-X_i^T\beta}}\right)^2$$

$$= \frac{\left(e^{X_i^T\beta} + e^{-X_i^T\beta}\right)^2 - \left(e^{X_i^T\beta} - e^{-X_i^T\beta}\right)^2}{\left(e^{X_i^T\beta} + e^{-X_i^T\beta}\right)^2}$$

$$= \frac{4e^{X_i^T\beta}e^{-X_i^T\beta}}{\left(e^{X_i^T\beta} + e^{-X_i^T\beta}\right)^2}$$

$$= \frac{\mathbf{4}}{\left(e^{X_i^T\beta} + e^{-X_i^T\beta}\right)^{\mathbf{2}}}$$

**And,** 1-$\frac{1}{\eta^2}$ can also be written as:

$$1\text{-}\frac{1}{\eta^2} = 1 - \left(\frac{e^{X_i^T\beta} + e^{-X_i^T\beta}}{e^{X_i^T\beta} - e^{-X_i^T\beta}}\right)^2$$

$$= \frac{\left(e^{X_i^T\beta} - e^{-X_i^T\beta}\right)^2 - \left(e^{X_i^T\beta} + e^{-X_i^T\beta}\right)^2}{\left(e^{X_i^T\beta} - e^{-X_i^T\beta}\right)^2}$$

$$= \frac{\mathbf{4}}{\left(e^{X_i^T\beta} - e^{-X_i^T\beta}\right)^{\mathbf{2}}}$$

We have,

$$\frac{\partial}{\partial \boldsymbol{\beta_j}} \frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta_j}} = \sum_{i-1}^{N} \left\{ -8\, x_j x_j * \left( \frac{1}{\left\{\left(e^{X_i^T\beta} - e^{-X_i^T\beta}\right)\right\}^{\mathbf{2}}} - \frac{\mathbf{2}}{\left\{\left(e^{X_i^T\beta} - e^{-X_i^T\beta}\right)\left(e^{X_i^T\beta} + e^{-X_i^T\beta}\right)\right\}^{\mathbf{2}}} \right) \right\}$$

$$= \sum_{i-1}^{N} \left\{ -8\, x_j x_j * \left( \frac{1}{4} \frac{4}{\left\{\left(e^{X_i^T\beta} - e^{-X_i^T\beta}\right)\right\}^{\mathbf{2}}} - \frac{\mathbf{1}}{\mathbf{16}} \frac{\mathbf{2*16}}{\left\{\left(e^{X_i^T\beta} - e^{-X_i^T\beta}\right)\left(e^{X_i^T\beta} + e^{-X_i^T\beta}\right)\right\}^{\mathbf{2}}} \right) \right\}$$

Now, substituting the corresponding values, we can write:

$$= \sum_{i-1}^{N} \left\{ -8\, x_j x_j * \left( \frac{1}{4}\left(1 - \frac{1}{\eta^2}\right) - \frac{1}{8}(1 - \frac{1}{\eta^2})(1 - \eta^2) \right) \right\}$$

$$= \sum_{i-1}^{N} \left\{ -\, x_j x_j * \left( 2\left(1 - \frac{1}{\eta^2}\right) - (1 - \frac{1}{\eta^2})(1 - \eta^2) \right) \right\}$$

$$= \sum_{i-1}^{N} \left\{ -\, x_j x_j * \left(1 - \frac{1}{\eta^2}\right)(2 - 1 + \eta^2) \right\}$$

$$= \sum_{i-1}^{N} \left\{ -\, x_j x_j * \left(1 - \frac{1}{\eta^2}\right)(1 + \eta^2) \right\}$$

$$= \sum_{i-1}^{N} \left\{ -\, x_j x_j * (1 + \eta^2 - \frac{1}{\eta^2} - 1) \right\}$$

$$= \sum_{i-1}^{N} \left\{ -\, x_j x_j * (\eta^2 - \frac{1}{\eta^2}) \right\}$$

$$= \sum_{i-1}^{N} \left\{ x_j x_j * (\frac{1}{\eta^2} - \eta^2) \right\}$$

$$\frac{\partial}{\partial \boldsymbol{\beta_j}} \frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta_j}} = \boldsymbol{X^T X} * \left( \frac{\mathbf{1}}{\boldsymbol{\eta}^{\mathbf{2}}} - \boldsymbol{\eta}^{\mathbf{2}} \right)$$

**Finally, for the Newton Raphson method, we have:**

$$\beta_{t+1} = \beta_t - \left( \frac{\partial}{\partial \beta_j} \frac{\partial l(\beta)}{\partial \beta_j} \right)^{-1} \frac{\partial l(\beta)}{\partial \beta_j}$$

or, $\beta_{t+1} = \beta_t - \left( X^T X * \left( \frac{1}{\eta^2} - \eta^2 \right) \right)^{-1} X^T \left( \frac{1}{\eta} - \eta \right)$

or, $\beta_{t+1} = \beta_t - \frac{X^T\left(\frac{1}{\eta} - \eta\right)}{X^T X * \left(\frac{1}{\eta^2} - \eta^2\right)}$

or, $\beta_{t+1} = \beta_t - \frac{X^T\left(\frac{1}{\eta} - \eta\right)}{X^T X * \left(\frac{1}{\eta} - \eta\right)\left(\frac{1}{\eta} + \eta\right)}$

or, $\beta_{t+1} = \beta_t - \frac{X^T}{X^T X\left(\frac{1}{\eta} + \eta\right)}$

$$\boxed{\boldsymbol{\beta_{t+1} = \beta_t - \left(X^T X\left(\frac{1}{\eta} + \eta\right)\right)^{-1} X^T}}$$