General notation remarks. Sometimes the vector sign $\vec{x} = x$ is omitted for clarity and left only when it is necessary to distinguish it from scalars. Upper indices indicate the index of inputs $(x^{(i)},\ y^{(i)})$. Where $x^{(i)}$ is a **vector** of input features and $y^{(i)}$ is a **scalar** of target variable. Lower indices show the dimension $\vec{x} = [x_0, \ldots, x_j, \ldots, x_n]$. Vector $\vec{\theta} = [\theta_0, \ldots, \theta_j, \ldots, \theta_n]$ is a vector of parameters for our model. For our case we have m samples in the dataset and $n+1$ dimensions of the inputs.

Our task is to proove that the Hessian matrix for loss function $J(\theta)$ is positive-semidefinite, meaning

$$z^T H z \geq 0 \quad \text{, for any } \ z$$

However there is another approach to prove a matrix is positive semidefinite, specifically if it can be written in a form $H = A^T A$. To show it lets consider the same product and we can see that squared norm of the resulting vector is always greater than zero.

$$z^T H z = z^T A^T A z = (Az)^T (Az) = \|Az\|^2 \geq 0$$

So lets start from our initial equation for the loss function.

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

$$\text{where } h_\theta(x) = g(\theta^T x), \text{ and } g(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

Let's define two parts inside the sum and differentiate them separately

$$y \cdot log(h_\theta(x)) \tag{3}$$

$$(1 - y) \cdot log(1 - (h_\theta(x))) \tag{4}$$

Taking into account derivative of the logistic function and vector derivative of the dot product

$$\frac{dlog(f(x))}{dx} = \frac{1}{f(x)} f'(x) \tag{5}$$

$$h'_\theta(x) = g(\theta^T x)(1 - g(\theta^T x)) \tag{6}$$

$$\frac{\partial \theta^T x}{\partial \vec{\theta}} = \vec{x} \tag{7}$$

Using those equations for derivative we can now simplify the first part of the gradient sum (1)

$$\nabla_\theta y log(h_\theta(x)) = \frac{\partial y log(h_\theta(x))}{\partial \vec{\theta}} \tag{8}$$
$$= y \cdot \frac{1}{h_\theta(x)} \cdot h'_\theta(x) \tag{9}$$
$$= y \cdot \frac{1}{g(\theta^T x)} \cdot g(\theta^T x)(1 - g(\theta^T x)) \cdot \frac{\partial \theta^T x}{\partial \vec{\theta}} \tag{10}$$
$$= y \cdot (1 - g(\theta^T x)) \cdot \vec{x} \tag{11}$$

Using similar technique we can simplify the second part of the gradient (1)

$$\nabla_\theta (1-y) log(1 - (h_\theta(x))) = (1-y) \cdot \frac{1}{1 - h_\theta(x)} \cdot -h'_\theta(x) \tag{12}$$
$$= (1-y) \cdot \frac{1}{1 - g(\theta^T x)} \cdot g(\theta^T x)(1 - g(\theta^T x)) \cdot -\frac{\partial \theta^T x}{\partial \vec{\theta}} \tag{13}$$
$$= (y-1) \cdot g(\theta^T x)) \cdot \vec{x} \tag{14}$$

Summing the two resulting terms (3) and (4) would provide us the expression under the sum

$$(3) + (4) = y \cdot (1 - g(\theta^T x)) \cdot \vec{x} + (y-1) \cdot g(\theta^T x)) \cdot \vec{x} \tag{15}$$
$$= \vec{x}[y - \cancel{y \cdot g(\theta^T x)} + \cancel{y \cdot g(\theta^T x)} - g(\theta^T x)] \tag{16}$$
$$= \vec{x}[y - g(\theta^T x)] \tag{17}$$

So the final equation for the gradient would be

$$\nabla_\theta J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \vec{x}^{(i)} [y^{(i)} - g(\theta^T x^{(i)})] \tag{18}$$

Now lets rewrite it in the vectorized form. Lets define the matrix of inputs X, where each i-th row represent a separate measurement in n-dimensional space.

$$\underset{m \times n}{X} = \begin{bmatrix} & \vdots & \\ — & x^{(i)} & — \\ & \vdots & \end{bmatrix}$$

The first part of the gradient sum would be

$$\sum_{i=1}^{m} x^{(i)} y^{(i)} = \underbrace{\begin{bmatrix} & | & \\ \dots & x^{(i)} & \dots \\ & | & \end{bmatrix}}_{X^T} \cdot \underbrace{\begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}}_{\vec{y}} = X^T \vec{y}$$

Let's take a closer look at the second part

$$\sum_{i=1}^{m} x^{(i)} g(\theta^T x^{(i)}) = \begin{bmatrix} & | & \\ \dots & x^{(i)} & \dots \\ & | & \end{bmatrix} \cdot \begin{bmatrix} g(\theta^T x^{(1)}) \\ \vdots \\ g(\theta^T x^{(m)}) \end{bmatrix} \tag{19}$$

$$= \begin{bmatrix} & | & \\ \dots & x^{(i)} & \dots \\ & | & \end{bmatrix} \cdot g\left(\begin{bmatrix} \theta^T x^{(1)} \\ \vdots \\ \theta^T x^{(m)} \end{bmatrix} \right) \tag{20}$$

The inner product vector in equation (20) could be rewriten as

$$\begin{bmatrix} \theta^T x^{(1)} \\ \vdots \\ \theta^T x^{(m)} \end{bmatrix} = \begin{bmatrix} x^{(1)} \cdot \vec{\theta} \\ \vdots \\ x^{(m)} \cdot \vec{\theta} \end{bmatrix} = \begin{bmatrix} & \vdots & \\ — & x^{(i)} & — \\ & \vdots & \end{bmatrix} \cdot \underbrace{\begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix}}_{\vec{\theta}} = X\theta$$

So the second part equation (20) could be rewriten as

$$\sum_{i=1}^{m} x^{(i)} g(\theta^T x^{(i)}) = X^T g(X\theta)$$

And the vectorized form of the gradient for loss function

$$\nabla_\theta J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \vec{x}^{(i)} [y^{(i)} - g(\theta^T x^{(i)})] = \frac{1}{m} X^T [\vec{y} - g(X\vec{\theta})] \tag{21}$$

When the gradient is in the vectorized form deriving the Hessian is much simpler. So by the definition of Hessian

$$H = H^T = \frac{\partial^2 J(\theta)}{\partial\theta\partial\theta^T} = \frac{\partial \nabla_\theta J(\theta)}{\partial\vec{\theta}}$$

Using the vector-by-vector differentiating rules we can now simplify the equation

$$\frac{\partial A\vec{y}}{\partial\vec{x}} = A\frac{\partial\vec{y}}{\partial\vec{x}} \qquad \frac{\partial g(\vec{y})}{\partial\vec{x}} = \frac{\partial g(\vec{y})}{\partial\vec{y}} \frac{\partial\vec{y}}{\partial\vec{x}}$$

$$H = \frac{\partial \nabla_\theta J(\theta)}{\partial\vec{\theta}} = \frac{1}{m} \frac{\partial X^T g(X\theta)}{\partial\vec{\theta}} \tag{22}$$

$$= \frac{1}{m} X^T \frac{\partial g(X\theta)}{\partial X\vec{\theta}} \frac{\partial X\theta}{\partial\vec{\theta}} \tag{23}$$

$$= \frac{1}{m} X^T \frac{\partial g(X\theta)}{\partial X\vec{\theta}} X \tag{24}$$

To calculate the inner term $\frac{\partial g(X\theta)}{\partial X\vec{\theta}}$ we can use the extended definition of vector by vector differentiation and taking into account that in our case function $g(\vec{x}) = [g(x_1), \ \dots \ , g(x_n)]$ is a simple function, hence $\frac{\partial g(x)_i}{\partial x_j} = 0$, for $i \neq j$

$$\frac{\partial g(\vec{x})}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial g(x)_1}{\partial x_1} & \dots & \frac{\partial g(x)_1}{\partial x_n} \\ & \dots & \\ \frac{\partial g(x)_n}{\partial x_1} & \dots & \frac{\partial g(x)_n}{\partial x_n} \end{bmatrix} = \begin{bmatrix} g'(x_1) & & \Huge 0 \\ & \ddots & \\ \Huge 0 & & g'(x_n) \end{bmatrix} = D$$

We are close to the end, consider matrix $D' = D/m$ and $\sqrt{D}' \cdot \sqrt{D}' = D'$, then we can rewrite Hessian, taking into account that diagonal matrices is always symmetrical

$$H = \frac{1}{m} \ X^T \ \frac{\partial g(X\theta)}{\partial X\vec{\theta}} \ X = X^T D' X = X^T \sqrt{D}' \cdot \sqrt{D}' X = (\sqrt{D}' X)^T \cdot \sqrt{D}' X = A^T A$$

So the Hessian can be written as a product of $A^T A$ which means it is positive semi-definite. QED.