Normal Equation.

La función de costo se escribe de la siguiente manera:

$$J = \frac{1}{2m}\left(\begin{bmatrix}\theta_0 & \theta_1 & \cdots & \theta_n\end{bmatrix}\begin{bmatrix}1 & 1 & \cdots & 1\\ x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(m)}\\ \vdots & \vdots & \cdots & \vdots\\ x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(m)}\end{bmatrix} - \begin{bmatrix}y^{(1)} & \cdots & y^{(m)}\end{bmatrix}\right)$$
$$\cdot\left(\begin{bmatrix}1 & x_1^{(1)} & \cdots & x_n^{(1)}\\ 1 & x_1^{(2)} & \cdots & x_n^{(2)}\\ \vdots & \vdots & \cdots & \vdots\\ 1 & x_1^{(m)} & \cdots & x_n^{(m)}\end{bmatrix}\begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix} - \begin{bmatrix}y^{(1)}\\ y^{(2)}\\ \cdots\\ y^{(m)}\end{bmatrix}\right)$$

Si se definen las siguientes cantidades matriciales:

$$\Theta = \begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix} \qquad \tilde{X} = \begin{bmatrix}1 & 1 & \cdots & 1\\ x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(m)}\\ \vdots & \vdots & \cdots & \vdots\\ x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(m)}\end{bmatrix} \qquad y = \begin{bmatrix}y^{(1)}\\ y^{(2)}\\ \vdots\\ y^{(m)}\end{bmatrix} \tag{1}$$

La función de coste se expresa así:

$$J(\Theta) = \frac{1}{2m}(\Theta^T\tilde{X} - y^T)(\Theta^T\tilde{X} - y^T)^T$$

Aplicando la transpuesta y expandiendo los productos:

$$\begin{aligned} J(\Theta) &= \frac{1}{2m}(\Theta^T\tilde{X} - y^T)([\Theta^T\tilde{X}]^T - y)\\ &= \frac{1}{2m}(\Theta^T\tilde{X} - y^T)(\tilde{X}^T\Theta - y)\\ &= \frac{1}{2m}(\Theta^T\tilde{X}\tilde{X}^T\Theta - \Theta^T\tilde{X}y - y^T\tilde{X}^T\Theta + y^Ty) \end{aligned}$$

Con esta expresión podemos calcular el gradiente de J:

$$\boldsymbol{\nabla}_\Theta J = \frac{1}{2m}\boldsymbol{\nabla}_\Theta(\Theta^T\tilde{X}\tilde{X}^T\Theta - \Theta^T\tilde{X}y - y^T\tilde{X}^T\Theta + y^Ty)$$

El último término no contribuye al gradiente pues es un producto de vectores constantes, además el segundo y tercer término son iguales: $\tilde{X}^T\Theta$ es un vector de dimensiones $(m \times n)(n \times 1) = (m \times 1)$ y y^T es un vector de $(1 \times m)$ de modo que $y^T\tilde{X}^T\Theta$ es de dimensiones $(1 \times m)(m \times 1) = (1 \times 1)$, al ser un escalar este es igual a su transpuesta, por lo que $y^T\tilde{X}^T\Theta = (y^T\tilde{X}^T\Theta)^T = \Theta^T\tilde{X}y$.

$$\boldsymbol{\nabla}_\Theta J = \frac{1}{2m}\boldsymbol{\nabla}_\Theta(\Theta^T\tilde{X}\tilde{X}^T\Theta - 2\Theta^T\tilde{X}y) \tag{2}$$

Para efectuar la derivada traemos dos resultados importantes, para una matriz simétrica A y un vector $\mathbf{b}$:

$$\begin{aligned} \boldsymbol{\nabla}_\mathbf{x}(\mathbf{x}^T A\mathbf{x}) &= 2A\mathbf{x} \\ \boldsymbol{\nabla}_\mathbf{x}(\mathbf{b}^T\mathbf{x}) = \boldsymbol{\nabla}_\mathbf{x}(\mathbf{x}^T\mathbf{b}) &= \mathbf{b} \end{aligned} \tag{3}$$

Al notar que $S = \tilde{X}\tilde{X}^T = (\tilde{X}\tilde{X}^T)^T = \tilde{X}\tilde{X}^T = S^T$, y haciendo uso de la ecuación (3) en la expresión (2):

$$\begin{aligned} \boldsymbol{\nabla}_\Theta J &= \frac{1}{2m}\boldsymbol{\nabla}_\Theta(\Theta^T\tilde{X}\tilde{X}^T\Theta - 2\Theta^T\tilde{X}y) \\ &= \frac{1}{2m}[\boldsymbol{\nabla}_\Theta(\Theta^T S\Theta) - 2\boldsymbol{\nabla}_\Theta(\Theta^T\tilde{X}y)] \\ &= \frac{1}{2m}[2S\Theta - 2\tilde{X}y] \\ &= \frac{1}{m}[\tilde{X}\tilde{X}^T\Theta - \tilde{X}y] \end{aligned}$$

El mínimo de J se obtiene igualando a cero el gradiente.

$$\begin{aligned} \boldsymbol{\nabla}_\Theta J = 0 \longrightarrow & \frac{1}{m}[\tilde{X}\tilde{X}^T\Theta - \tilde{X}y] = 0 \\ & \tilde{X}\tilde{X}^T\Theta - \tilde{X}y = 0 \\ & \Theta = (\tilde{X}\tilde{X}^T)^{-1}\tilde{X}y \end{aligned} \tag{4}$$