1 Gradient Descent Derivation

Example from (https://www.bilibili.com/video/BV1XE411C7mS?p=1)
Parameters to fit through Gradient Decent: $\vec{W} = \{W_a, W_b, W_c\}$
4×3 matrix A stores the data set.$\vec{A_i}$ denotes ith row of the matrix as a vector, and $\vec{A_a}$ denotes the ath column as a vector.
$Y = \{Y_1, Y_2, Y_3, Y_4\}$ is the solution to the data set A.
GOAL : Minimize the following function f by varying $\vec{W}$ through Gradient Descent.

$$f(\vec{W}) := \sum_{\{i=1,2,3,4\}} \left\| \frac{1}{1+e^{-\vec{A_i}\cdot\vec{W}}} - Y_i \right\|^2 \tag{1}$$

Then the gradient in direction Wa is given by:

$$\frac{\partial f(\vec{W})}{\partial W_a} = 2 \sum_{\{i=1,2,3,4\}} \left| \frac{1}{1+e^{-\vec{A_i}\cdot\vec{W}}} - Y_i \right| \frac{A_{ia} \cdot e^{-\vec{A_i}\cdot\vec{W}}}{(1+e^{-\vec{A_i}\cdot\vec{W}})^2} \tag{2}$$

$$= 2 \sum_{\{i=1,2,3,4\}} \left| \frac{1}{1+e^{-\vec{A_i}\cdot\vec{W}}} - Y_i \right| \left[\frac{1}{1+e^{-\vec{A_i}\cdot\vec{W}}} - \frac{1}{(1+e^{-\vec{A_i}\cdot\vec{W}})^2} \right] \cdot A_{ia} \tag{3}$$

$$= 2 \begin{bmatrix} \left| \frac{1}{1+e^{-\vec{A_1}\cdot\vec{W}}} - Y_1 \right| \left[\frac{1}{1+e^{-\vec{A_1}\cdot\vec{W}}} - \frac{1}{(1+e^{-\vec{A_1}\cdot\vec{W}})^2} \right] \\ \vdots \\ \left| \frac{1}{1+e^{-\vec{A_4}\cdot\vec{W}}} - Y_4 \right| \left[\frac{1}{1+e^{-\vec{A_4}\cdot\vec{W}}} - \frac{1}{(1+e^{-\vec{A_4}\cdot\vec{W}})^2} \right] \end{bmatrix} \cdot \begin{bmatrix} A_{1a} \\ \vdots \\ A_{4a} \end{bmatrix} \tag{4}$$

$$= 2 \cdot \overrightarrow{Delta} \cdot \vec{A_a} \tag{5}$$

where $\overrightarrow{Delta}$ is the same 4-row vector as in the example on Bilibili.
Hence,

$$\nabla f(\vec{W}) = \begin{bmatrix} \frac{\partial f(\vec{W})}{\partial W_a} \\ \frac{\partial f(\vec{W})}{\partial W_b} \\ \frac{\partial f(\vec{W})}{\partial W_c} \end{bmatrix} \tag{6}$$

is the gradient vector that has twice the length as the step vector in the example.