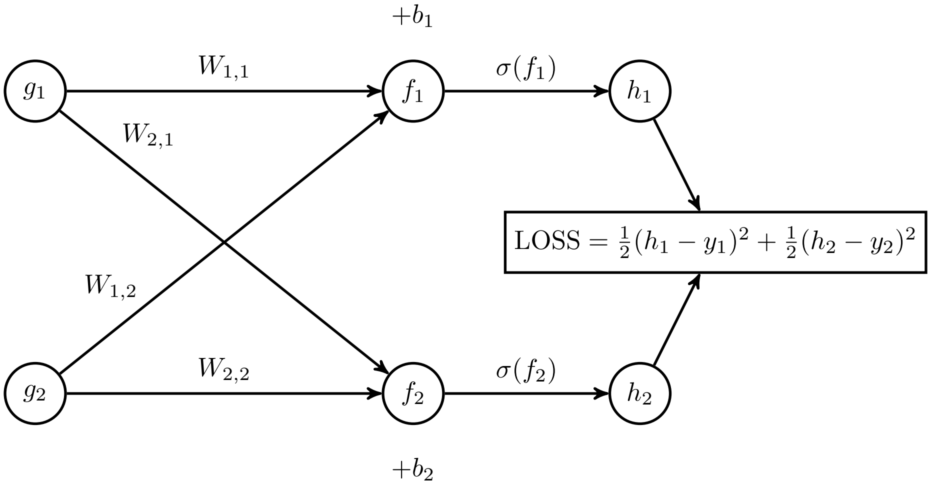

$*$ is standard numpy point-wise multiplication, @ is numpy matrix multiplication (see `https://numpy.org/doc/stable/reference/generated/numpy.matmul.html`)

Chain rule (`https://en.wikipedia.org/wiki/Chain_rule`):

$$\frac{\partial f(g(x))}{\partial x} = \frac{\partial f(g(x))}{\partial g(x)} \frac{\partial g(x)}{\partial x}$$

The main things that we use here are that things below have common prefixes

$$\frac{\partial c(d(e(a(x,y))))}{\partial x} = \frac{\partial c(d(e(a(x,y))))}{\partial d(e(a(x,y)))} \frac{\partial d(e(a(x,y)))}{\partial e(a(x,y))} \frac{\partial e(a(x,y))}{\partial a(x,y)} \frac{\partial a(x,y)}{\partial x}$$

$$\frac{\partial c(d(e(a(x,y))))}{\partial y} = \frac{\partial c(d(e(a(x,y))))}{\partial d(e(a(x,y)))} \frac{\partial d(e(a(x,y)))}{\partial e(a(x,y))} \frac{\partial e(a(x,y))}{\partial a(x,y)} \frac{\partial a(x,y)}{\partial y}$$

and that

$$\frac{\partial(c(x) + d(x))}{\partial x} = \frac{\partial c(x)}{\partial x} + \frac{\partial d(x)}{\partial x}$$

We want to compute derivatives with respect to each thing present in the computation graph presented above:

$$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

$$W = \begin{bmatrix} W_{1,1} & W_{1,2} \\ W_{2,1} & W_{2,2} \end{bmatrix}$$

$$g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$$

$$f = W@g + b = \begin{bmatrix} W_{1,1}g_1 + W_{1,2}g_2 + b_1 \\ W_{2,1}g_1 + W_{2,2}g_2 + b_2 \end{bmatrix}$$

$$h = \sigma(f) = \begin{bmatrix} \sigma(f_1) \\ \sigma(f_2) \end{bmatrix}$$

Because

$$\frac{\partial \mathrm{LOSS}}{\partial h_i} = h_i - y_i$$

therefore

$$\frac{\partial \mathrm{LOSS}}{\partial h} = h - y = \begin{bmatrix} h_1 - y_1 \\ h_2 - y_2 \end{bmatrix}$$

Because

$$\frac{\partial h_i}{\partial f_i} = \sigma(f_i)(1 - \sigma(f_i)) = h_i(1 - h_i)$$

therefore

$$\frac{\partial h}{\partial f} = h * (1 - h) = \begin{bmatrix} h_1(1 - h_1) \\ h_2(1 - h_2) \end{bmatrix}$$

where $*$ is a point-wise multiplication as in numpy. Using chain rule we have that

$$\frac{\partial \mathrm{LOSS}}{\partial f} = \frac{\partial \mathrm{LOSS}}{\partial h} * \frac{\partial h}{\partial f}$$

Because

$$\frac{\partial f_k}{\partial W_{i,j}} = \begin{cases} g_j & \text{if } k = i \\ 0 & otherwise \end{cases}$$

therefore

$$\frac{\partial \mathrm{LOSS}}{\partial W_{i,j}} = \frac{\partial \mathrm{LOSS}}{\partial f_i} g_j$$

$$\frac{\partial \mathrm{LOSS}}{\partial W} = \frac{\partial \mathrm{LOSS}}{\partial f} @g.T$$

where $D = \frac{\partial \mathrm{LOSS}}{\partial W}$ is such that $D_{i,j} = \frac{\partial \mathrm{LOSS}}{\partial W_{i,j}}$

Note that

$$\frac{\partial \mathrm{LOSS}}{\partial b} = \frac{\partial \mathrm{LOSS}}{\partial f}$$

as

$$\frac{\partial f_i}{\partial b_j} = \begin{cases} 1 & \text{if } j = i \\ 0 & otherwise \end{cases}$$

As

$$\frac{\partial f_i}{\partial g_j} = W_{i,j}$$

We can observe that indeed

$$\frac{\partial \text{LOSS}}{\partial g} = W.T@\frac{\partial \text{LOSS}}{\partial f} = \begin{bmatrix} W_{1,1}\frac{\partial \text{LOSS}}{\partial f_1} + W_{2,1}\frac{\partial \text{LOSS}}{\partial f_2} \\ W_{1,2}\frac{\partial \text{LOSS}}{\partial f_1} + W_{2,2}\frac{\partial \text{LOSS}}{\partial f_2} \end{bmatrix}$$

So the most important parts are:

$$\frac{\partial \text{LOSS}}{\partial h} = h - y$$

$$\frac{\partial h}{\partial f} = h * (1 - h)$$

$$\frac{\partial \text{LOSS}}{\partial f} = \frac{\partial \text{LOSS}}{\partial h} * \frac{\partial h}{\partial f}$$

$$\frac{\partial \text{LOSS}}{\partial b} = \frac{\partial \text{LOSS}}{\partial f}$$

$$\frac{\partial \text{LOSS}}{\partial W} = \frac{\partial \text{LOSS}}{\partial f}@g.T$$

$$\frac{\partial \text{LOSS}}{\partial g} = W.T@\frac{\partial \text{LOSS}}{\partial f}$$