Recollect from last time, key abstraction:

\[ \text{Layer } l \]

\[ h^{(l)} = g(h^{(l-1)}, w) \]

\[ \text{loss } L = f(h^{(L)}) \]

Abstractly:

\[ \frac{\partial L}{\partial \text{in}} = \frac{\partial L}{\partial \text{out}} \cdot \frac{\partial \text{out}}{\partial \text{in}} \]

[More]

Concretely:

\[ \frac{\partial L}{\partial h^{(l-1)}} = \frac{\partial L}{\partial h^{(l)}} \cdot \frac{\partial h^{(l)}}{\partial h^{(l-1)}} \]

\[ \frac{\partial L}{\partial w} = \frac{\partial L}{\partial h^{(l)}} \cdot \frac{\partial h^{(l)}}{\partial w} \]
Multi-Layer Perceptions / Fully-Connected Layer / Inner-Product Layer

Another way of drawing the

Forward Pass:
\[ h_i^l = \sum_{j} w_{ij} h_j \]
\[ = \mathbf{W}^T h \]

Let's vectorize some more

\[ h = h \times h \]
\[ \begin{bmatrix} h^T \\ h \end{bmatrix} = \mathbf{W} \begin{bmatrix} h^T \\ h \end{bmatrix} \]

So all of F-PASS
\[ = 1 \text{ matrix-mul} \]
Why did we write this as matrix mult?

1. Because it’s cooler!
2. Gradients, my dear Watson, gradients...

\[ h^0 = W^0 h^{-1} \]

\[ \frac{\partial h^0}{\partial W^0} = W^0 \quad \text{[Now, isn’t that beautiful?]} \]

\[ \frac{\partial h^{-1}}{\partial h} \]

What about \( \partial \mathbf{z} \)?

Simple

\[ h^i = \mathbf{z}_i^T h^{-1} \]

\[ \Rightarrow \frac{\partial h^i}{\partial \mathbf{z}_i} = h^{-1} \]

Now put it all together:

\( a \)

\[ \frac{\partial L}{\partial h^{-1}} = \frac{\partial L}{\partial h^0} \cdot \frac{\partial h^0}{\partial h^{-1}} = \left[ \frac{\partial L}{\partial h^0} \cdot W^0 \right] \text{ Very nice} \]

so B-PASS(a) = 1 matrix mult

\[ \text{[self multiply = gradient]} \]

\[ \text{[right multiply = output]} \]

\( b \)

\[ \frac{\partial L}{\partial z_i} = \frac{\partial L}{\partial h^i} \cdot \frac{\partial h^i}{\partial z_i} = \frac{\partial L}{\partial h^i} \cdot h^{-1} \]

\[ \text{scale by } I \cdot c_i \]
So, \( w^{(e)} = w^{(e-1)} - \eta \frac{\partial L}{\partial w^{(e)}} \)

where \( \frac{\partial L}{\partial w^{(e)}} = \begin{bmatrix} \frac{\partial L}{\partial w^{(e)}_1} \cdot h_1 \cdot e \cdot h_{-1}^T \\ \vdots \\ \frac{\partial L}{\partial w^{(e)}_{e-1}} \cdot h_{e-1} \cdot e \cdot h_{-1} \\ \frac{\partial L}{\partial w^{(e)}_e} \cdot h_e \cdot e \cdot h_{-1} \end{bmatrix} \)

One more intuition:

Row in \( W \) matrix

\( \frac{\partial L}{\partial h_i} = 0 \)

\( \frac{\partial L}{\partial h_i} \rightarrow h_i \)

Row in \( W \) matrix

**3) ReLU Layer**

**F-PASS**

\( h_i^{(e+1)} = \max\{0, h_i^e\} \)

**B-PASS**

\( \frac{\partial h_i^{(e+1)}}{\partial h_i^e} = +1 \) if \( h_i^e \geq 0 \)

\( \frac{\partial h_i^{(e+1)}}{\partial h_i^e} = 0 \) else

\( \Rightarrow \frac{\partial L}{\partial h_i^e} = \frac{\partial L}{\partial h_i^{(e+1)}} \begin{cases} 1, & h_i^e \geq 0 \\ 0, & else \end{cases} \)