Explanatory Notes for 6.390

Shaunticlair Ruiz (Current TA)

Fall 2022

Relevant Derivatives

If you aren't interesting in understanding matrix derivatives, here we provide the general format of each of the derivatives we care about.

Notation 1

Here, we give useful **derivatives** for **neural network gradient descent**.

Loss is not given, so we can't compute it, as before:

$$\frac{\overbrace{\partial \mathcal{L}}^{(n^L \times 1)}}{\partial A^L}$$

We get the same result for each of these terms as we did before, except in matrix form.

$$\frac{\overbrace{\partial Z^\ell}^{(m^\ell \times 1)}}{\partial W^\ell} = A^{\ell-1}$$

$$\frac{\overbrace{\partial Z^\ell}^{(m^\ell \times n^\ell)}}{\partial A^{\ell-1}} = W^\ell$$

The last one is actually pretty different from before:

$$\frac{\overbrace{\partial a^\ell}^{(n^\ell \times n^\ell)}}{\partial z^\ell} = \begin{bmatrix} f'(z_1^\ell) & 0 & 0 & \cdots & 0 \\ 0 & f'(z_2^\ell) & 0 & \cdots & 0 \\ 0 & 0 & f'(z_3^\ell) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & f'(z_r^\ell) \end{bmatrix}$$

Where r is the length of Z^ℓ.

In short, we only have the z_i derivative on the i^{th} diagonal.

Why this is will be explained in the matrix derivative notes.

Example: Suppose you have the activation $f(z) = z^2$.

Your pre-activation might be

$$z^\ell = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \tag{1}$$

The output would be

$$a^\ell = f(z^\ell) = \begin{bmatrix} 1 \\ 2^2 \\ 3^2 \end{bmatrix} \tag{2}$$

But the derivative would be:

$$f(z) = 2z \tag{3}$$

Which, gives our matrix derivative as:

$$\frac{\partial a^\ell}{\partial z^\ell} = \begin{bmatrix} 2\cdot 1 & 0 & 0 \\ 0 & 2\cdot 2 & 0 \\ 0 & 0 & 2\cdot 3 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 6 \end{bmatrix}$$

If you want to be able to **derive** some of the derivatives, without reading the matrix derivative section, just use this formula for vector derivatives:

If you have time, do read - you won't understand what you're doing otherwise!

$$\frac{\partial w}{\partial v} = \overbrace{\begin{bmatrix} \frac{\partial w_1}{\partial v_1} & \frac{\partial w_2}{\partial v_1} & \cdots & \frac{\partial w_n}{\partial v_1} \\ \frac{\partial w_1}{\partial v_2} & \frac{\partial w_2}{\partial v_2} & \cdots & \frac{\partial w_n}{\partial v_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial w_1}{\partial v_m} & \frac{\partial w_2}{\partial v_m} & \cdots & \frac{\partial w_n}{\partial v_m} \end{bmatrix}}^{\text{Column } j \text{ matches } w_j} \Bigg\} \text{Row } i \text{ matches } v_i \tag{4}$$

We can use this for scalars as well: we just treat them as a vector of length 1.

With some cleverness, you can derive the Scalar/Matrix and Matrix/Scalar derivatives as well.

Part of what the next section covers.