Explanatory Notes for 6.390

Shaunticlair Ruiz (Current TA)

Fall 2022

Activation Derivatives

We weren't able to **simplify** our expressions above, partly because we didn't know which **loss** or **activation** function we were going to use.

So, here, we will look at the **common** choices for these functions, and **catalog** what their derivatives look like.

- **Step function** step(z):

$$\frac{d}{dz}\text{step}(z) = 0 \tag{1}$$

  This is part of why we don't use this function: it has no gradient. We can show this by looking piecewise:

$$\text{step}(z) = \begin{cases} 1 & \text{if } z \geqslant 0 \\ 0 & \text{if } z < 0 \end{cases} \tag{2}$$

  And take the derivative of each piece:

$$\frac{d}{dz}\text{ReLU}(z) = 0 = \begin{cases} 0 & \text{if } z \geqslant 0 \\ 0 & \text{if } z < 0 \end{cases} \tag{3}$$

- **Rectified Linear Unit** ReLU(z):

$$\frac{d}{dz}\text{ReLU}(z) = \text{step}(z) \tag{4}$$

  This one might be a bit surprising at first, but it makes sense if you **also** break it up into cases:

$$\text{ReLU}(z) = \max(0, z) = \begin{cases} z & \text{if } z \geqslant 0 \\ 0 & \text{if } z < 0 \end{cases} \tag{5}$$

  And take the derivative of each piece:

$$\frac{d}{dz}\text{ReLU}(z) = \text{step}(z) = \begin{cases} 1 & \text{if } z \geqslant 0 \\ 0 & \text{if } z < 0 \end{cases} \tag{6}$$

- **Sigmoid** function $\sigma(z)$:

$$\frac{d}{dz}\sigma(z) = \sigma(z)(1 - \sigma(z)) = \frac{e^{-z}}{(1 + e^{-z})^2} \tag{7}$$

  This derivative is useful for simplifying NLL, and has a nice form.

  We can just compute the derivative with the single-variable chain rule.

As a reminder, the function looks like:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{8}$$

- **Identity** ("linear") function $f(z) = z$:

$$\frac{d}{dz} z = 1 \tag{9}$$

This one follows from the definition of the derivative.

We cannot rely on a linear activation function for our **hidden** layers, because a linear neural network is no more **expressive** than one layer.

But, we use it for **regression**.

- **Softmax** function $\text{softmax}(z)$:

This function has a difficult derivative we won't go over here.

If you're curious, here's a link.

- **Hyperbolic tangent** function $\tanh(z)$:

$$\frac{d}{dz} \tanh(z) = 1 - \tanh(z)^2 \tag{10}$$

This strange little expression is the "hyperbolic secant" squared. We won't bother further with it.

Notation 1

For our various **activation** functions, we have the **derivatives**:

Step:

$$\frac{d}{dz}\text{step}(z) = 0$$

ReLU:

$$\frac{d}{dz}\text{ReLU}(z) = \text{step}(z)$$

Sigmoid:

$$\frac{d}{dz}\sigma(z) = \sigma(z)(1 - \sigma(z))$$

Identity/Linear:

$$\frac{d}{dz}z = 1$$

Loss derivatives

Now, we look at the loss derivatives.

- **Square loss** function $\mathcal{L}_{sq} = (a - y)^2$:

$$\frac{d}{da}\mathcal{L}_{sq} = 2(a - y) \tag{11}$$

  Follows from chain rule+power rule, used for regression.

- **Linear loss** function $\mathcal{L}_{sq} = |a - y|$:

$$\frac{d}{da}\mathcal{L}_{lin} = \text{sign}(a - y) \tag{12}$$

  This one can also be handled piecewise, like $\text{step}(z)$ and $\text{ReLU}(z)$:

$$|u| = \begin{cases} u & \text{if } z \geqslant 0 \\ -u & \text{if } z < 0 \end{cases} \tag{13}$$

  We take the piecewise derivative:

$$\frac{d}{du}|u| = \text{sign}(u) = \begin{cases} 1 & \text{if } z \geqslant 0 \\ -1 & \text{if } z < 0 \end{cases} \tag{14}$$

- **NLL** (Negative-Log Likelihood) function $\mathcal{L}_{NLL} = -(y \log(a) + (1-y)\log(1-a))$

$$\frac{d}{da}\mathcal{L}_{NLL} = -\left(\frac{y}{a} - \frac{1-y}{1-a}\right) \tag{15}$$

- **NLLM** (Negative-Log Likelihood Multiclass) function $\mathcal{L}_{NLL} = -\sum_j y_j \log(a_j)$

  Similar to softmax, we will omit this derivative.

Notation 2

For our various **loss** functions, we have the **derivatives**:

Square:

$$\frac{d}{da}\mathcal{L}_{sq} = 2(a-y) \tag{16}$$

Linear (Absolute):

$$\frac{d}{da}\mathcal{L}_{lin} = \text{sign}(a-y) \tag{17}$$

NLL (Negative-Log Likelihood):

$$\frac{d}{da}\mathcal{L}_{NLL} = -\left(\frac{y}{a} - \frac{1-y}{1-a}\right) \tag{18}$$