Problem 1 Take the neural network from Lecture 4. Assume the loss is $e = \frac{(t-y)^2}{2}$. This is

$$\begin{aligned} s &= W_1 x + b_1 \\ h &= \text{sigmoid}(s) \\ z &= W_2 h + b_2 \\ y &= \text{sigmoid}(z) \end{aligned}$$

What is $\frac{\partial e}{\partial W_1}, \frac{\partial e}{\partial W_2}, \frac{\partial e}{\partial b_1}, \frac{\partial e}{\partial b_2}$? If we modify one equation so that $z = W_2 h + b_2 + s$, what are the gradients now?

1. $\frac{\partial e}{\partial w_2} = \frac{\partial e}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial w_2}$

$= -\sigma(z)(1-\sigma(z))h$

$\frac{\partial e}{\partial b_2} = \frac{\partial e}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial b_2}$

$= -\sigma(z)(1-\sigma(z))$

$\frac{\partial e}{\partial w_1} = \frac{\partial e}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial h} \cdot \frac{\partial h}{\partial s} \cdot \frac{\partial s}{w_1}$

$= -\sigma(z)(1-\sigma(z)) \cdot W_2 \cdot \sigma(s)(1-\sigma(s)) \cdot x$

$\frac{\partial e}{\partial b_1} = \frac{\partial e}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial h} \cdot \frac{\partial h}{\partial s} \cdot \frac{\partial s}{\partial b_1}$

$= -\sigma(z)(1-\sigma(z))w_2 \cdot \sigma(s)(1-\sigma(s))$

If $z = w_2 h + b_2 + s$.

$\frac{\partial e}{\partial w_2} = \frac{\partial e}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial w_2}$

$= -\sigma(z)(1-\sigma(z))h$

$\frac{\partial e}{\partial b_2} = \frac{\partial e}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial b_2}$

$= -\sigma(z)(1-\sigma(z))$

$\frac{\partial e}{\partial w_1} = \frac{\partial e}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \left(\frac{\partial z}{\partial h} \cdot \frac{\partial h}{\partial s} \cdot \frac{\partial s}{\partial w_1} + \frac{\partial z}{\partial s} \cdot \frac{\partial s}{\partial w_1}\right)$

$= -\sigma(z)(1-\sigma(z))(W_2 \cdot \sigma(s)(1-\sigma(s)) \cdot x + x)$

$\frac{\partial e}{\partial b_1} = \frac{\partial e}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \left(\frac{\partial z}{\partial h} \cdot \frac{\partial h}{\partial s} \cdot \frac{\partial s}{\partial b_1} + \frac{\partial z}{\partial s} \cdot \frac{\partial s}{\partial b_1}\right)$

$= -\sigma(z)(1-\sigma(z))(w_2 \cdot \sigma(s)(1-\sigma(s)) + 1)$

Problem 2 Drop Out at inference time. At training time, we randomly pick weights in the network and zero them out with probability p. So, each weight in the network, we can zero it out with a probability p and keep it with a probability $1-p$. Express each weight as a random variable times an appropriate independent Bernoulli random variable. What is the expected value of this random variable. This is what is used at training time.

$E[w'] = E[w] \cdot (1-p)$

Which means the drop-out rate is P during training time.

Problem 3 Imagine that you have a MLP network where each Linear layer is followed by a Batch Norm layer. Do you need the bias term in each Linear layer? Prove that it is unnecessary. Thus, prove that if you fit a model where the layer is specified with $nn.Linear(\ldots, bias = False)$ no information is lost, the bias adds nothing and you can specify this. This is one effect of using Batch Norm.

The bias term in each linear is unnecessary

Consider a linear layer : $y = wx + b$.

After apply batch nomalization, $y' = \gamma \cdot (x - \mu) / \sigma + \beta$

$$= \frac{\gamma}{\sigma} x + (\beta - \gamma \cdot \mu / \sigma)$$

Comparing two models, b can be replaced by the $(\beta - \gamma \cdot \frac{\mu}{\sigma})$

because γ and β are introduced to prevent similar data be homogeneous

therefore, γ and β can be seem as two bias.