ECE421 Homework 6

Name: Mark Qi

Student ID: 1006764645

Question 1,

$x=2, y=1$

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \times \frac{e^{x}}{e^{x}}$$

$$\sinh x = \frac{e^{x}-e^{-x}}{2} \qquad = \frac{e^{2x}-e^{0}}{e^{2x}+e^{0}} = \frac{e^{2x}-1}{e^{2x}+1}$$

$$\cosh x = \frac{e^{x}+e^{-x}}{2}$$

Forward Propagation

Compute

$S^{(1)} = (S_1^{(1)}, S_2^{(1)})$

$x^{(1)} = (x_1^{(1)}, x_2^{(1)})$

$S^{(2)}$

$x^{(2)}$

$S^{(3)}$

$x^{(3)}$

$$W^{(1)} = \begin{bmatrix} W_{0,1}^{(1)} & W_{0,2}^{(1)} \\ W_{1,1}^{(1)} & W_{1,2} \end{bmatrix} = \begin{bmatrix} 0.1 & 0.2 \\ 0.3 & 0.4 \end{bmatrix}$$

$$W^{(2)} = \begin{bmatrix} W^{(2)}_{0,1} \\ W^{(2)}_{1,1} \\ W^{(2)}_{2,1} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 1 \\ -3 \end{bmatrix} \qquad W^{(3)} = \begin{bmatrix} W^{(3)}_{0,1} \\ W^{(3)}_{1,1} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$x^{(0)} = \begin{bmatrix} 1 \\ x \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$s^{(1)} = W^{(1)T} x^{(0)} = \underbrace{\begin{bmatrix} 0.1 & 0.3 \\ 0.2 & 0.4 \end{bmatrix}}^{2\times2} \overbrace{\begin{bmatrix} 1 \\ 2 \end{bmatrix}}^{2\times1} = \begin{bmatrix} 0.7 \\ 1.0 \end{bmatrix}$$

$$x^{(1)} = \begin{bmatrix} 1 \\ \tanh(0.7) \\ \tanh(1.0) \end{bmatrix} = \begin{bmatrix} 1 \\ 0.60 \\ 0.76 \end{bmatrix}$$

$$s^{(2)} = W^{(2)T} x^{(1)} = \overset{1\times3}{\begin{bmatrix} 0.2 & 1 & -3 \end{bmatrix}} \overset{3\times1}{\begin{bmatrix} 1 \\ 0.60 \\ 0.76 \end{bmatrix}} = -1.48$$

$$x^{(2)} = \begin{bmatrix} 1 \\ \tanh(-1.48) \end{bmatrix} = \begin{bmatrix} 1 \\ -0.90 \end{bmatrix}$$

$$s^{(3)} = W^{(3)T} x^{(2)} = \underset{1\times2}{\begin{bmatrix} 1 & 2 \end{bmatrix}} \underset{2\times1}{\begin{bmatrix} 1 \\ -0.90 \end{bmatrix}} = -0.80$$

$$x^{(3)} = \theta(s^{(3)}) = \tanh(s^{(3)}) = \tanh(-0.80) = -0.66$$

Backward Propagation

$d^{(l)}$: the number of nodes for layer L minus 1

$[v]_1^{d^{(l)}}$: given vector $v = (v_0, \ldots, v_{d^{(l)}})$, discard v_0

$$\delta^{(l)} = \frac{\partial e}{\partial s^{(l)}} = \frac{\partial e}{\partial x^{(l)}} \times \frac{\partial x^{(l)}}{\partial s^{(l)}}$$

$$= \theta'(s^{(l)}) \odot [W^{(l+1)} \delta^{(l+1)}]_1^{d^{(l)}}, \quad 1 \leq l \leq L-1$$

$$\delta^{(L)} = \frac{\partial e}{\partial s^{(L)}} = \frac{\partial (x^{(L)} - y)^2}{\partial s^{(L)}} = \frac{\partial (x^{(L)} - y)^2}{\partial x^{(L)}} \times \frac{\partial x^{(L)}}{\partial s^{(L)}}$$

$$= 2(x^{(L)} - y) \times \theta'(s^{(L)})$$

$$\frac{\partial}{\partial s} \tanh s = 1 - \tanh^2 s$$

$$\delta^{(L)} = \begin{cases} 2(x^{(L)} - y)(1 - x^{(L)^2}) & \theta(s) = \tanh(s) \\ 2(x^{(L)} - y) & \theta(s) = s \end{cases}$$

Starting From $L=3$, Calculate $\delta^{(3)}$

$$\delta^{(3)} = \frac{\partial e}{s^{(3)}} = 2(x^{(3)} - y)(1 - x^{(3)^2}) = -1.855$$

$x^{(3)} = -0.66$, $y = 1$

Different story for intermediate layers

$$\delta^{(l)} = \frac{\partial e}{\partial s^{(l)}} = \theta'(s^{(l)}) \odot \left[W^{(l+1)} \delta^{(l+1)}\right]_1^{d^{(l)}}$$

For tanh hidden layers only

$$\theta'(s^{(l)}) = \left[1 - x^{(l)} \odot x^{(l)}\right]_1^{d^{(l)}} = \begin{bmatrix} 1 - (x_1^{(l)})^2 \\ \vdots \\ 1 - (x_{d^{(l)}}^{(l)})^2 \end{bmatrix}$$

Layer $l=2$, Calculate $\delta^{(2)}$

$$\delta^{(2)} = \theta'(s^{(2)}) \odot \left[W^{(2+1)} \delta^{(2+1)}\right]_1^{d^{(2)}}$$

$$= \left[1 - \begin{bmatrix} 1 \\ -0.9 \end{bmatrix} \odot \begin{bmatrix} 1 \\ -0.9 \end{bmatrix}\right]_1^{d^{(2)}} \odot \left[\begin{bmatrix} 1 \\ 2 \end{bmatrix} (-1.855)\right]_1^{d^{(2)}=2}$$

$$= \left[1 - \begin{bmatrix} 1 \\ 0.81 \end{bmatrix}\right]_1^{d^{(2)}} \odot \left[\begin{matrix} -1.855 \\ -3.71 \end{matrix}\right]_1^{d^{(2)}}$$

$$= -0.705$$

Layer $l=1$, Calculate $\delta^{(1)}$

$$\delta^{(1)} = \theta'(s^{(1)}) \odot \left[W^{(1+1)} \delta^{(1+1)}\right]_1^{d^{(1)}}$$

$$= \left[1 - \begin{bmatrix} 1 \\ 0.60 \\ 0.76 \end{bmatrix} \odot \begin{bmatrix} 1 \\ 0.60 \\ 0.76 \end{bmatrix}\right]_1^{d^{(1)}} \odot \left[\begin{bmatrix} 0.2 \\ 1 \\ -3 \end{bmatrix} (-0.705)\right]_1^{d^{(1)}}$$

$$= 1 - \begin{bmatrix} 1 \\ 0.36 \\ 0.5776 \end{bmatrix}^{d^{(1)}}_{1} \odot \begin{bmatrix} -0.141 \\ -0.705 \\ 2.115 \end{bmatrix}^{d^{(1)}}_{1}$$

$$= \begin{bmatrix} 0.64 \\ 0.4224 \end{bmatrix} \odot \begin{bmatrix} -0.705 \\ 2.115 \end{bmatrix} = \begin{bmatrix} -0.4512 \\ 0.89676 \end{bmatrix}$$

$$\frac{\partial e}{\partial W^{(1)}} = x^{(0)} \delta^{(1)T} = \overset{2\times 1}{\begin{bmatrix} 1 \\ 2 \end{bmatrix}} \overset{1\times 2}{\begin{bmatrix} -0.45 & 0.89 \end{bmatrix}} = \begin{bmatrix} -0.45 & 0.89 \\ -0.90 & 1.78 \end{bmatrix}$$

$$\frac{\partial e}{\partial W^{(2)}} = x^{(1)} \delta^{(2)T} = \begin{bmatrix} 1 \\ 0.60 \\ 0.76 \end{bmatrix} (-0.705) = \begin{bmatrix} -0.705 \\ -0.423 \\ -0.535 \end{bmatrix}$$

$$\frac{\partial e}{\partial W^{(3)}} = x^{(2)} \delta^{(3)T} = \begin{bmatrix} 1 \\ -0.90 \end{bmatrix} (-1.855) = \begin{bmatrix} -1.855 \\ 1.6695 \end{bmatrix}$$

Vanishing Gradient

Question 2,

Input (x, y)

$x = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}) \in \mathbb{R}^3$

r and w are shared weights

$\theta(\cdot)$ is some arbitrary activation function

$e(\Omega) = (\hat{y} - y)^2$

(a) $\frac{de}{dr} = \frac{de}{d\hat{y}} \times \frac{d\hat{y}}{dx_3^{(2)}} \times \frac{dx_3^{(2)}}{dr}$

$\frac{d}{dr} x_3^{(2)} = \frac{d}{dr}(x_3^{(1)} + r x_2^{(2)}) = \frac{d}{dr}(x_3^{(1)} + r(x_2^{(1)} + r x_1^{(1)}))$

$= x_2^{(1)} + 2r x_1^{(1)}$

$\frac{de}{dr} = 2(\hat{y} - y)(1)(x_2^{(1)} + 2r x_1^{(1)})$

$= 2(\hat{y} - y)(1)(x_2^{(1)} + r x_1^{(1)} + r x_1^{(1)})$

$= 2\triangle(x_2^{(2)} + r x_1^{(1)})$

Product Rule

$\frac{d(f(r)g(r))}{dr} = f(r)\frac{dg(r)}{dr} + g(r)\frac{df(r)}{dr}$

$\frac{\partial}{\partial r}(r x_2^{(2)}) = r x_1^{(1)} + x_2^{(2)}$

(b)

$$\frac{de}{dx_2^{(2)}} = \frac{de}{d\hat{y}} \times \frac{d\hat{y}}{dx_3^{(2)}} \times \frac{dx_3^{(2)}}{dx_2^{(2)}}$$

$$= 2(\hat{y}-y)(1)(r)$$

$$= 2r\Delta$$

$$\frac{de}{dx_1^{(1)}} = \frac{de}{dx_2^{(2)}} \times \frac{dx_2^{(2)}}{dx_1^{(1)}} = (2r\Delta)(r) = 2r^2\Delta$$

$$\frac{de}{dx_2^{(1)}} = \frac{de}{dx_2^{(2)}} \times \frac{dx_2^{(2)}}{dx_2^{(1)}} = (2r\Delta)(1) = 2r\Delta$$

$$\frac{de}{dx_3^{(1)}} = \frac{de}{d\hat{y}} \times \frac{d\hat{y}}{dx_3^{(2)}} \times \frac{dx_3^{(2)}}{dx_3^{(1)}} = 2(\hat{y}-y)(1) = 2\Delta$$

(c)

$$\frac{de}{dw} = \frac{de}{dx_1^{(1)}} \times \frac{dx_1^{(1)}}{dw} + \frac{de}{dx_2^{(1)}} \times \frac{dx_2^{(1)}}{dw} + \frac{de}{dx_3^{(1)}} \times \frac{dx_3^{(1)}}{dw}$$

$$= (2r^2\Delta)\,\theta'(wx_1^{(0)})\,x_1^{(0)} + 2r\Delta\,\theta'(wx_2^{(0)})\,x_2^{(0)} +$$

$$2\Delta\,\theta'(wx_3^{(0)})\,x_3^{(0)}$$

(d)

$$\frac{de}{dx_1^{(0)}} = \frac{de}{dx_1^{(1)}} \times \frac{dx_1^{(1)}}{dx_1^{(0)}} = (2r^2\Delta)\,\theta'(wx_1^{(0)})\,w$$

$$\frac{de}{dx_2^{(0)}} = \frac{de}{dx_2^{(1)}} \times \frac{dx_2^{(1)}}{dx_2^{(0)}} = (2r\Delta)\,\theta'(wx_2^{(0)})\,w$$

$$\frac{de}{dx_3^{(0)}} = \frac{de}{dx_3^{(1)}} \times \frac{dx_3^{(1)}}{dx_3^{(0)}} = (2v)\theta'(wx_3^{(0)})w$$

(e)

$x^{(0)} = (1, -1, 1)$ and $y = 1$

$w = v = 1$

$\theta(s) = \max(0, s)$

$$e(\Omega) = (\overset{1}{\theta(wx_3^{(0)})} + v\overset{0}{\theta(wx_2^{(0)})} + v^2\overset{1}{\theta(wx_1^{(0)})} - y)^2$$

$$= (2-1)^2$$
$$= 1^2$$
$$= 1$$

$$\hat{y} = x_3^{(2)} = x_3^{(1)} + v x_2^{(2)} = \theta(wx_3^{(0)}) + v(x_2^{(1)} + v x_1^{(1)})$$
$$= \theta(wx_3^{(0)}) + v(\theta(wx_2^{(0)}) + v\theta(wx_1^{(0)}))$$
$$= \theta(wx_3^{(0)}) + v\theta(wx_2^{(0)}) + v^2\theta(wx_1^{(0)})$$

$$\frac{\partial e}{\partial v} = 2(\theta(wx_3^{(0)}) + v\theta(wx_2^{(0)}) + v^2\theta(wx_1^{(0)}) - y)$$
$$[\theta(wx_2^{(0)}) + 2v\theta(wx_1^{(0)})]$$
$$= (2)(2-1)(0+2) = (2)(1)(2) = 4$$

$$\frac{\partial e}{\partial w} = 2(\theta(wx_3^{(0)}) + v\theta(wx_2^{(0)}) + v^2\theta(wx_1^{(0)}) - y)$$
$$[\underset{1}{\theta'(wx_3^{(0)})}\underset{1}{x_3^{(0)}} + v\underset{0}{\theta'(wx_2^{(0)})}x_2^{(0)} + \underset{1^2}{v^2}\underset{1}{\theta'(wx_1^{(0)})}\underset{1}{x_1^{(0)}}]$$

$$= 2(2-1)(1+1) = (2)(1)(2) = 4$$

(d)

1. Pick (x_n, y_n) from D at random
2. Forward Pass, Calculate e
3. Backward Pass

$$\frac{\partial e}{\partial v}, \quad \frac{\partial e}{\partial w}$$

4. $$v = v - \epsilon\left(\frac{\partial e}{\partial v} + 2v\lambda\right)$$

$$w = w - \epsilon\left(\frac{\partial e}{\partial w} + 2w\lambda\right)$$