Exercises for Chapter 6: Deep Feedforward Networks

ranjita.naik.edu@gmail.com

Exercise 1

Optional exercise to brush up gradients.

Notation

$\mathbf{X}$	Bold-face, capital letters refer to matrices
$\mathbf{w}$	Bold-face, lower case letters refer to vectors
b	Lower case letters refer to scalars

For the XOR problem described in the book - choosing $f(x, \boldsymbol{\theta})$ to be linear model

$$f(\mathbf{x}; \mathbf{w}, b) = \mathbf{x}^T\mathbf{w} + b$$

and using mean squared error as cost function

$$J(\boldsymbol{\theta}) = \frac{1}{4}\sum_{x=\mathbf{X}}(f^*(\mathbf{x}) - f(\mathbf{x}; \boldsymbol{\theta}))^2$$

minimize $J(\boldsymbol{\theta})$ in closed form with respect to $\mathbf{w}$ and b to show that $\mathbf{w} = \mathbf{0}$ and $\mathrm{b} = \frac{1}{2}$

Table 1: Training Data

x_1	x_2	$f^*(\mathbf{x})$
0	0	0
0	1	1
1	0	1
1	1	0

Solution

Solution 1 Using matrix differentiation rules

- $\frac{d\mathbf{X}\mathbf{w}}{d\mathbf{w}} = \mathbf{X}^T$
- $\frac{d\mathbf{w}^T\mathbf{X}\mathbf{w}}{d\mathbf{w}} = 2\mathbf{X}^T d\mathbf{w}$

Let $\mathbf{y}$ be the vector of true labels, $w_0 = \text{b}$ and $x_0 = 1$

$J(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})$
$J(\mathbf{w}) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} - \mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}$

Note that $(\mathbf{y}^T\mathbf{X}\mathbf{w})^T = (\mathbf{w}^T\mathbf{X}^T\mathbf{y})$ is scalar and equals its own tranpose, hence $(\mathbf{y}^T\mathbf{X}\mathbf{w}) = (\mathbf{w}^T\mathbf{X}^T\mathbf{y})$.

$J(\mathbf{w}) = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}$

The convex function $J(\mathbf{w})$ is minimized when the gradient vector is $\mathbf{0}$.

$\frac{d\mathbf{J}}{d\mathbf{w}} = 0 - 2(\mathbf{y}^T\mathbf{X})^T + 2(\mathbf{X}^T\mathbf{X})^T\mathbf{w}$
$-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{0}$
$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$
$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \qquad \mathbf{X}^T = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X}) = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 2 \end{bmatrix} \qquad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} \frac{3}{4} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \begin{bmatrix} \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{bmatrix} \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ 0 \\ 0 \end{bmatrix}$$

Solution 2 For those unfamiliar with matrix differentiation, here is the solution in scalar form

$$J(\boldsymbol{\theta}) = \frac{1}{4}\sum_{x=\mathbf{X}}((w_1x_1 + w_2x_2 + b) - f^*(\mathbf{x}))^2$$

$$J(\boldsymbol{\theta}) = \frac{1}{4}\sum_{x=\mathbf{X}}((w_1x_1 + w_2x_2 + b) - f^*(\mathbf{x}))^2 \tag{1}$$

$J(\boldsymbol{\theta})$ is convex, and hence, global minimum can be found by setting the partial derivative of $J(\boldsymbol{\theta})$ w.r.t each of the parameters to 0 and solving for the parameters.

Derivative of any finite sum of functions is the sum of the derivatives of those functions.

$$\frac{\partial J}{\partial w_1} = \frac{1}{4}\sum_{x=\mathbf{X}} 2((w_1x_1 + w_2x_2 + b) - f^*(\mathbf{x}))\ x_1$$

$$\begin{aligned}\frac{\partial J}{\partial w_1} = \frac{1}{2}[&((w_1(0) + w_2(0) + b) - 0)(0) + ((w_1(0) + w_2(1) + b) - 1)(0) + \\ &((w_1(1) + w_2(0) + b) - 1)(1) + ((w_1(1) + w_2(1) + b) - 0)(1)\ ]\end{aligned}$$

$$\frac{\partial J}{\partial w_1} = \frac{1}{2}(2w_1 + w_2 + 2b - 1)$$

$$2w_1 + w_2 + 2b - 1 = 0 \tag{2}$$

$$\frac{\partial J}{\partial w_2} = \frac{1}{4}\sum_{x=\mathbf{X}} 2((w_1x_1 + w_2x_2 + b) - f^*(\mathbf{x}))\ x_2$$

$$\frac{\partial J}{\partial w_2} = \frac{1}{2}(2w_2 + w_1 + 2b - 1)$$

$$2w_2 + w_1 + 2b - 1 = 0 \tag{3}$$

$$\frac{\partial J}{\partial b} = \frac{1}{4}\sum_{x=\mathbf{X}} 2((w_1x_1 + w_2x_2 + b) - f^*(\mathbf{x}))$$

$$\frac{\partial J}{\partial b} = \frac{1}{2}(2w_2 + 2w_1 + 4b - 2)$$

$$2b + w_1 + w_2 - 1 = 0 \tag{4}$$

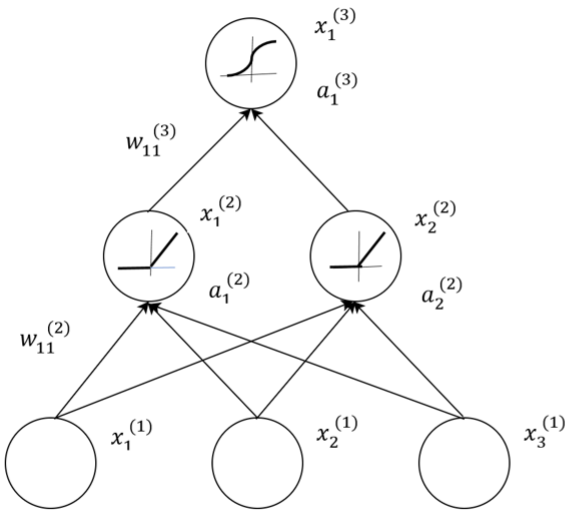
Substituting (4) *in* (2)
$2w_1 - w_2 - w_1 + w_2 = 0$
$w_1 = 0$

Substituting (4) *in* (3)
$w_1 + 2w_2 - w_1 - w_2 = 0$
$w_2 = 0$

Solving for b
$2b - 1 = 0$
$b = \frac{1}{2}$

Exercise 2

When sigmoid output unit is used in combination with squared error loss, learning can cease when the sigmoid saturates.For the following neural network, show this behavior analytically by working out the partial derivative of $w_{11}^{(3)}$ with respect to the squared error loss E and true label y.



	$1 \le l \le L$	layers
$w_{ij}^{(l)} =$	$0 \le i \le d^{(l-1)}$	inputs
	$1 \le j \le d^{l}$	outputs
$x_0^{(l)}$	= +1	

Output Layer $d^{(l-1)}$

$$x_j^{(l)} = \sigma\left(a_j^{(l)}\right)$$

$$a_j^{(l)} = \sum_i^{d^{(l-1)}} w_{ij}^{(l)}\, x_j^{(l-1)}$$

Hidden Layer

$$x_j^{(l)} = max\left(0, a_j^{(l)}\right)$$

$$a_j^{(l)} = \sum_i^{d^{(l-1)}} w_{ij}^{(l)}\, x_j^{(l-1)}$$

$$E = \frac{1}{2}(x_1^{(3)} - y)^2$$

Solution

Using the chain rule

$$\frac{\partial E}{\partial w_{11}^{(3)}} = \frac{\partial a_1^{(3)}}{\partial w_{11}^{(3)}} \frac{\partial x_1^{(3)}}{\partial a_1^{(3)}} \frac{\partial E}{\partial x_1^{(3)}}$$

$$\frac{\partial a_1^{(3)}}{\partial w_{11}^{(3)}} = x_1^{(2)}$$
$$\frac{\partial x_1^{(3)}}{\partial a_1^{(3)}} = x_1^{(3)}(1 - x_1^{(3)})$$
$$\frac{\partial E}{\partial x_1^{(3)}} = (x_1^{(3)} - y)$$

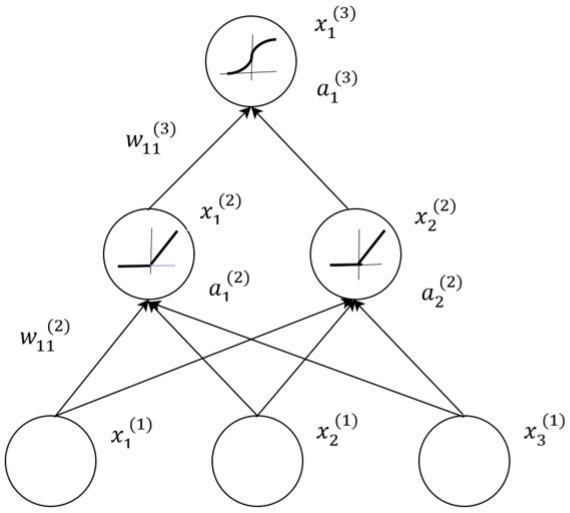
$$\frac{\partial E}{\partial w_{11}^{(3)}} = x_1^{(2)}\ x_1^{(3)}\ (1 - x_1^{(3)})\ (x_1^{(3)} - y) \tag{5}$$

$$w_{11}^{(3)} = w_{11}^{(3)} - \eta \frac{\partial E}{\partial w_{11}^{(3)}} \tag{6}$$

As shown in eq. (5), when the sigmoid saturates, on either side (0 or 1), derivative approaches zero due to $x_1^{(3)}\ (1 - x_1^{(3)})$ and hence the weight updae for $w_{11}^{(3)}$ ceases.

Exercise 3

For the following neural network,show analytically that sigmoid output unit when used in combination with cross entropy cost doesnt suffer from ceasing of learning issue when the sigmoid saturates.



$1 \le l \le L$ layers

$w_{ij}^{(l)} =$ $0 \le i \le d^{(l-1)}$ inputs

$1 \le j \le d^{l}$ outputs

$x_0^{(l)} = +1$

Output Layer $d^{(l-1)}$

$$x_j^{(l)} = \sigma\left(a_j^{(l)}\right)$$

$$a_j^{(l)} = \sum_i^{d^{(l-1)}} w_{ij}^{(l)}\, x_j^{(l-1)}$$

Hidden Layer

$$x_j^{(l)} = max\left(0, a_j^{(l)}\right)$$

$$a_j^{(l)} = \sum_i^{d^{(l-1)}} w_{ij}^{(l)}\, x_j^{(l-1)}$$

$$E = -(y \log x_1^{(3)} + (1 - y) \log(1 - x_1^{(3)}))$$

Solution

Using the chain rule

$$\frac{\partial E}{\partial w_{11}^{(3)}} = \frac{\partial a_1^{(3)}}{\partial w_{11}^{(3)}} \frac{\partial x_1^{(3)}}{\partial a_1^{(3)}} \frac{\partial E}{\partial x_1^{(3)}}$$

$$\frac{\partial a_1^{(3)}}{\partial w_{11}^{(3)}} = x_1^{(2)}$$

$$\frac{\partial x_1^{(3)}}{\partial a_1^{(3)}} = x_1^{(3)}(1 - x_1^{(3)})$$

$$\frac{\partial E}{\partial x_1^{(3)}} = -\left(\frac{y}{x_1^{(3)}} + \frac{(1-y)(-1)}{(1 - x_1^{(3)})}\right)$$

$$\frac{\partial E}{\partial x_1^{(3)}} = \left(\frac{(x_1^{(3)} - y)}{x_1^{(3)}(1 - x_1^{(3)})}\right)$$

$$\frac{\partial E}{\partial w_{11}^{(3)}} = x_1^{(2)}\ x_1^{(3)}(1 - x_1^{(3)})\ \left(\frac{(x_1^{(3)} - y)}{x_1^{(3)}(1 - x_1^{(3)})}\right)$$

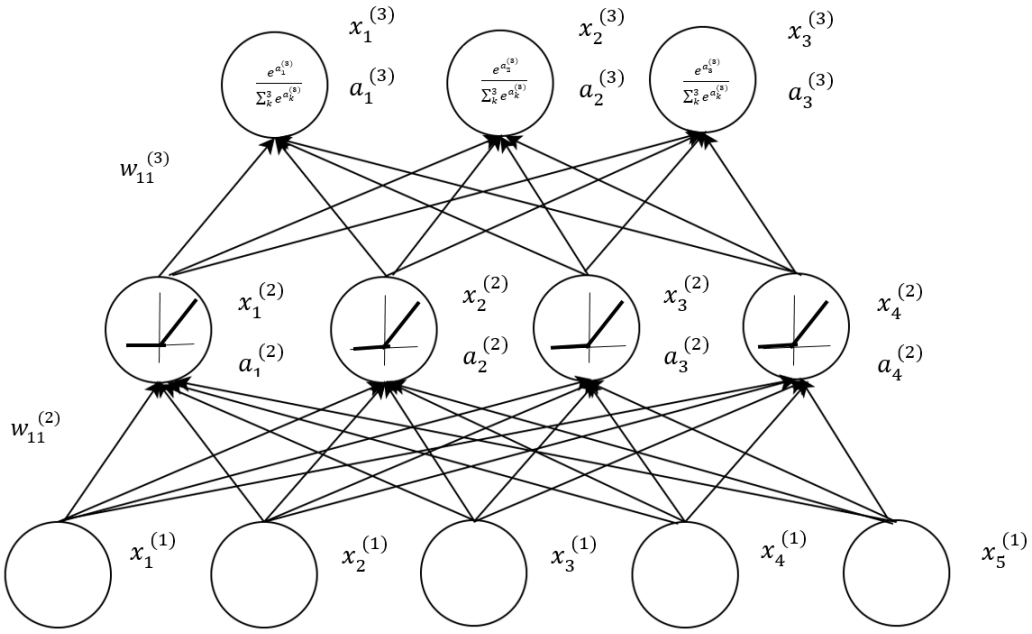
$$\frac{\partial E}{\partial w_{11}^{(3)}} = x_1^{(2)}\ (x_1^{(3)} - y) \tag{7}$$

$$w_{11}^{(3)} = w_{11}^{(3)} - \eta \frac{\partial E}{\partial w_{11}^{(3)}} \tag{8}$$

Note that the term $x_1^{(3)}\ (1 - x_1^{(3)})$ in eq. (5), which was making the derivative zero on sigmoidal saturation doesnt appear in eq. (7).

Exercise 4

As explained in section 6.2.2.3, softmax output unit can saturate when the difference between the input values become extreme. Learning can cease when softmax is used with squared error loss function. Show this behavior analytically by computing the partial derivative of $w_{11}^{(3)}$ with respect to the squared loss E and true label y.



	$1 \le l \le L$	layers
$w_{ij}^{(l)} =$	$0 \le i \le d^{(l-1)}$	inputs
	$1 \le j \le d^{l}$	outputs
$x_0^{(l)} = +1$		

Output Layer

$$x_j^{(l)} = \frac{e^{a_j^{(l)}}}{\sum_k^l e^{a_k^{(l)}}}$$

$$a_j^{(l)} = \sum_i^{d^{(l-1)}} w_{ij}^{(l)} x_j^{(l-1)}$$

Hidden Layer

$$x_j^{(l)} = max\left(0, a_j^{(l)}\right)$$

$$a_j^{(l)} = \sum_i^{d^{(l-1)}} w_{ij}^{(l)} x_j^{(l-1)}$$

$$E = \sum_k^3 (x_k^{(3)} - y_k)^2$$

Solution

$$\frac{\partial E}{\partial w_{11}^{(3)}} = \frac{\partial a_1^{(3)}}{\partial w_{11}^{(3)}} \frac{\partial E}{\partial a_1^{(3)}}$$

$$\frac{\partial E}{\partial a_i^{(3)}} = \sum_k^3 \frac{\partial x_k^{(3)}}{\partial a_i^{(3)}} \frac{\partial E}{\partial x_k^{(3)}}$$

$$\frac{\partial E}{\partial x_i^{(3)}} = 2(x_i^{(3)} - y_i)$$

$$\frac{\partial x_i^{(3)}}{\partial a_k^{(3)}} = \begin{cases} \left(\frac{\exp a_k^{(3)}}{\sum_k^3 \exp a_k^{(3)}} - \left(\frac{\exp a_k^{(3)}}{\sum_k^3 \exp a_k^{(3)}} \right)^2 \right) & = \; x_i^{(3)} \; (1 - x_i^{(3)}) \qquad i = k \\ -\frac{\exp a_k^{(3)} \exp a_i^{(3)}}{\left(\sum_k^3 \exp a_k^{(3)} \right)^2} & = \; x_i^{(3)} \; x_k^{(3)} \qquad i \neq k \end{cases}$$

$$\frac{\partial E}{\partial a_i^{(3)}} = \sum_k^3 \frac{\partial x_k^{(3)}}{\partial a_i^{(3)}} \frac{\partial E}{\partial x_k^{(3)}}$$

$$\frac{\partial E}{\partial a_i^{(3)}} = \frac{\partial x_i^{(3)}}{\partial a_i^{(3)}} \frac{\partial E}{\partial x_i^{(3)}} + \sum_{i \neq k}^3 \frac{\partial x_k^{(3)}}{\partial a_i^{(3)}} \frac{\partial E}{\partial x_k^{(3)}}$$

$$\frac{\partial E}{\partial a_i^{(3)}} = 2(x_i^{(3)} - y_i) \; x_i^{(3)} \; (1 - x_i^{(3)}) - 2 \sum_{i \neq k}^3 (x_k^{(3)} - y_k) x_i^{(3)} x_k^{(3)}$$

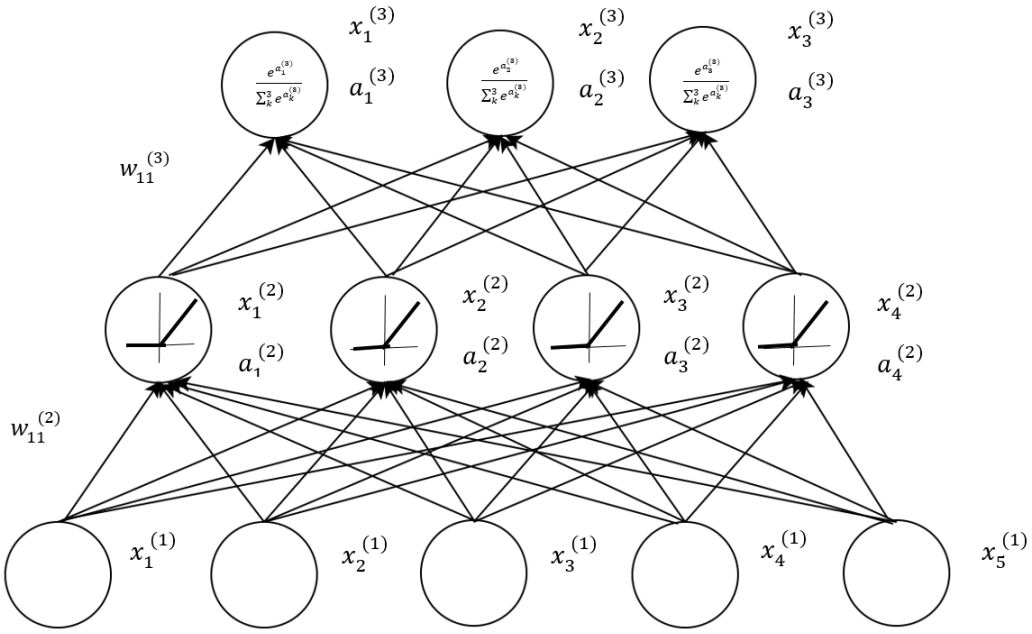
$$\frac{\partial E}{\partial w_{11}^{(3)}} = 2 \; x_1^{(2)} \left(\underbrace{(x_i^{(3)} - y_i) \; x_i^{(3)} \; (1 - x_i^{(3)})}_{\text{(a)}} - \underbrace{\sum_{i \neq k}^3 (x_k^{(3)} - y_k) \; x_i^{(3)} \; x_k^{(3)}}_{\text{(b)}} \right) \tag{9}$$

Case 1: $a_i^{(3)} >> a_{k \neq i}^{(3)}$
Term (a) can vanish due to $(1 - x_i^{(3)})$ approaching zero and term (b) can vanish due to $x_k^{(3)}$ approaching zero.

Case 2: $a_i^{(3)} << a_{k \neq i}^{(3)}$
Both terms (a) and (b) can vanish due to $x_i^{(3)}$ approaching zero.

Exercise 5

For the neural network in the diagram below, show analytically that softmax output unit when used in combination with cross entropy cost doesnt suffer from ceasing of the learning issue when difference between the softmax input values become extreme. Show this behavior analytically by computing the partial derivative of $w_{11}^{(3)}$ with respect to the squared error loss E and true label y.



	$1 \le l \le L$	layers
$w_{ij}^{(l)} =$	$0 \le i \le d^{(l-1)}$	inputs
	$1 \le j \le d^{l}$	outputs
$x_0^{(l)} = +1$		

Output Layer

$$x_j^{(l)} = \frac{e^{a_j^{(l)}}}{\sum_k^l e^{a_k^{(l)}}}$$

$$a_j^{(l)} = \sum_i^{d^{(l-1)}} w_{ij}^{(l)} x_j^{(l-1)}$$

Hidden Layer

$$x_j^{(l)} = max\left(0, a_j^{(l)}\right)$$

$$a_j^{(l)} = \sum_i^{d^{(l-1)}} w_{ij}^{(l)} x_j^{(l-1)}$$

$$E = \sum_i^3 y_i \log x_i^{(3)}$$

Solution

$$\frac{\partial E}{\partial w_{11}^{(3)}} = \frac{\partial a_1^{(3)}}{\partial w_{11}^{(3)}} \frac{\partial E}{\partial a_1^{(3)}}$$

$$\frac{\partial E}{\partial a_i^{(3)}} = \sum_k^3 \frac{\partial x_k^{(3)}}{\partial a_i^{(3)}} \frac{\partial E}{\partial x_k^{(3)}}$$

$$\frac{\partial E}{\partial x_i^{(3)}} = -\frac{y_i}{x_i^{(3)}}$$

$$\frac{\partial x_i^{(3)}}{\partial a_k^{(3)}} = \begin{cases} \left(\frac{\exp a_k^{(3)}}{\sum_k^3 \exp a_k^{(3)}} - \left(\frac{\exp a_k^{(3)}}{\sum_k^3 \exp a_k^{(3)}} \right)^2 \right) & = x_i^{(3)} \ (1 - x_i^{(3)}) \qquad i = k \\ -\frac{\exp a_k^{(3)} \exp a_i^{(3)}}{\left(\sum_k^3 \exp a_k^{(3)} \right)^2} & = x_i^{(3)} \ x_k^{(3)} \qquad i \neq k \end{cases}$$

$$\frac{\partial E}{\partial a_i^{(3)}} = \sum_k^3 \frac{\partial x_k^{(3)}}{\partial a_i^{(3)}} \frac{\partial E}{\partial x_k^{(3)}}$$

$$\frac{\partial E}{\partial a_i^{(3)}} = \frac{\partial x_i^{(3)}}{\partial a_i^{(3)}} \frac{\partial E}{\partial x_i^{(3)}} + \sum_{i \neq k}^3 \frac{\partial x_k^{(3)}}{\partial a_i^{(3)}} \frac{\partial E}{\partial x_k^{(3)}}$$

$$\frac{\partial E}{\partial a_i^{(3)}} = -y_i \ (1 - x_i^{(3)}) - \sum_{i \neq k}^3 y_k x_i^{(3)}$$

$$\frac{\partial E}{\partial a_i^{(3)}} = -y_i - x_i^{(3)} \sum_k y_k$$

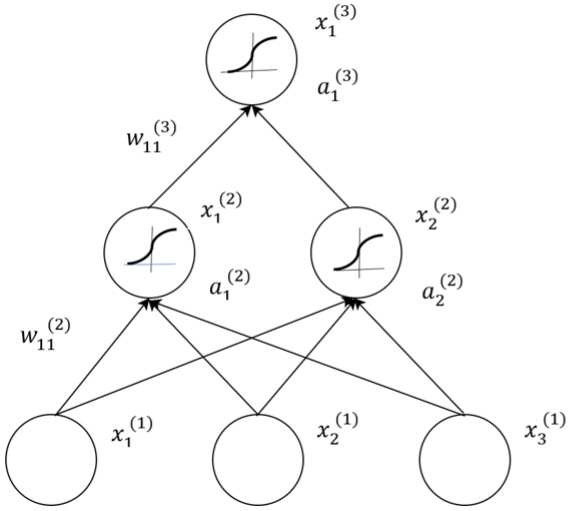
$$\frac{\partial E}{\partial a_i^{(3)}} = x_i^{(3)} - y_i$$

$$\frac{\partial E}{\partial w_{11}^{(3)}} = x_1^{(2)} (x_i^{(3)} - y_i) \tag{10}$$

Hence derivative doesnt vanish when the difference between the softmax input units is extreme.

Exercise 6

For the neural network in the diagram below, show analytically why the usage of sigmoidal units in the hidden layers is discouraged, by working out the gradients of the lower layer weights with respect to the cross entropy loss function.



	$1 \le l \le L$	layers
$w_{ij}^{(l)} =$	$0 \le i \le d^{(l-1)}$	inputs
	$1 \le j \le d^{l}$	outputs
$x_0^{(l)} = +1$		

$$x_j^{(l)} = \sigma\left(a_j^{(l)}\right)$$

$$a_j^{(l)} = \sum_i^{d^{(l-1)}} w_{ij}^{(l)} x_j^{(l-1)}$$

$$E = -\left(y \log x_1^{(3)} + (1-y)\log(1-x_1^{(3)})\right)$$

Solution

$$\frac{\partial E}{\partial a_1^{(3)}} = \frac{\partial x_1^{(3)}}{\partial a_1^{(3)}} \frac{\partial E}{\partial x_1^{(3)}}$$

$$\frac{\partial E}{\partial x_1^{(3)}} = -\left(\frac{y}{x_1^{(3)}} + \frac{(1-y)(-1)}{(1-x_1^{(3)})}\right)$$

$$\frac{\partial E}{\partial x_1^{(3)}} = \left(\frac{x_1^{(3)} - y}{x_1^{(3)}(1-x_1^{(3)})}\right)$$

$$\frac{\partial x_1^{(3)}}{\partial a_1^{(3)}} = x_1^{(3)}(x_1^{(3)} - 1)$$

$$\frac{\partial E}{\partial a_1^{(3)}} = x_1^{(3)}(x_1^{(3)} - 1)\left(\frac{x_1^{(3)} - y}{x_1^{(3)}(1-x_1^{(3)})}\right)$$

$$\frac{\partial E}{\partial a_1^{(2)}} = \frac{\partial a_1^{(3)}}{\partial a_1^{(2)}} \frac{\partial E}{\partial a_1^{(3)}}$$
$$\frac{\partial a_1^{(3)}}{\partial a_1^{(2)}} = \frac{\partial x_1^{(2)}}{\partial a_1^{(2)}} \frac{\partial a_1^{(3)}}{\partial x_1^{(2)}}$$
$$\frac{\partial a_1^{(3)}}{\partial a_1^{(2)}} = x_1^{(2)}(x_1^{(2)} - 1)w_{11}^{(3)}$$
$$\frac{\partial E}{\partial w_{11}^{(2)}} = \frac{\partial a_1^{(2)}}{\partial w_{11}^{(2)}} \frac{\partial a_1^{(3)}}{\partial a_1^{(2)}} \frac{\partial E}{\partial a_1^{(3)}}$$

$$\frac{\partial E}{\partial w_{11}^{(2)}} = x_1^{(1)}\ x_1^{(2)}\ (x_1^{(2)} - 1)\ w_{11}^{(3)}\ (x_1^{(3)} - y) \tag{11}$$

$$w_{11}^{(2)} = w_{11}^{(2)} - \eta \frac{\partial E}{\partial w_{11}^{(2)}} \tag{12}$$

When the sigmoidal unit in the hidden layer saturates, gradient vanishes due to $x_1^{(2)}\ (x_1^{(2)} - 1)$ term in eq. (11) and hence the weight update for $w_{11}^{(2)}$ ceases.