Q1-1: Are these statements true or false?
(A) Stochastic gradient descent has fewer amount of computation per gradient update than standard gradient descent.
(B) Large-batch methods often have a worse generalization ability compared to small-batch methods.

1. True, True
2. True, False
3. False, True
4. False, False

Q1-1: Are these statements true or false?
(A) Stochastic gradient descent has fewer amount of computation per gradient update than standard gradient descent.
(B) Large-batch methods often have a worse generalization ability compared to small-batch methods.

1. True, True ←
2. True, False
3. False, True
4. False, False

(A) Since stochastic GD uses single instance per iteration while standard GD uses full batch training data per iteration, stochastic GD has fewer amount of computation per iteration.
(B) Small-batch methods is less susceptible to local minimum, thus having a better generalization ability.

Q1-2: Assume $net^{(d)} = w_0 + \sum_{i=1}^{n} w_i x_i^{(d)}$, $o^{(d)} = \text{Sigmoid}(net^{(d)}) = \frac{1}{1+\exp(-net^{(d)})}$, and $E^{(d)} = -y^{(d)} \ln\left(o^{(d)}\right) - \left(1 - y^{(d)}\right) \ln\left(1 - o^{(d)}\right)$ for a data $(x^{(d)}, y^{(d)})$, please calculate $\frac{\partial E^{(d)}}{\partial w_i}$ by using the chain rule.

1. $-\left(y^{(d)} - o^{(d)}\right)x_i^{(d)}$
2. $-\left(y^{(d)} - o^{(d)}\right)o^{(d)}(1 - o^{(d)})x_i^{(d)}$
3. $-(\frac{y^{(d)}}{o^{(d)}} - \frac{1-y^{(d)}}{1-o^{(d)}})x_i^{(d)}$
4. $-(y^{(d)} - 2y^{(d)}o^{(d)} + o^{(d)})x_i^{(d)}$

Q1-2: Assume $net^{(d)} = w_0 + \sum_{i=1}^{n} w_i x_i^{(d)}$, $o^{(d)} = \text{Sigmoid}(net^{(d)}) = \frac{1}{1+\exp(-net^{(d)})}$, and $E^{(d)} = -y^{(d)} \ln\left(o^{(d)}\right) - \left(1 - y^{(d)}\right) \ln\left(1 - o^{(d)}\right)$ for a data $(x^{(d)}, y^{(d)})$, please calculate $\frac{\partial E^{(d)}}{\partial w_i}$ by using the chain rule.

1. $-\left(y^{(d)} - o^{(d)}\right)x_i^{(d)}$
2. $-\left(y^{(d)} - o^{(d)}\right)o^{(d)}(1 - o^{(d)})x_i^{(d)}$
3. $-(\frac{y^{(d)}}{o^{(d)}} - \frac{1-y^{(d)}}{1-o^{(d)}})x_i^{(d)}$
4. $-(y^{(d)} - 2y^{(d)}o^{(d)} + o^{(d)})x_i^{(d)}$

$$\frac{\partial E^{(d)}}{\partial w_i} = \frac{\partial E^{(d)}}{\partial o^{(d)}} \frac{\partial o^{(d)}}{\partial net^{(d)}} \frac{\partial net^{(d)}}{\partial w_i}$$
$$= \left(-\frac{y^{(d)}}{o^{(d)}} + \frac{1 - y^{(d)}}{1 - o^{(d)}}\right) o^{(d)}(1 - o^{(d)})x_i^{(d)}$$
$$= \left(-y^{(d)}(1 - o^{(d)}) + (1 - y^{(d)})o^{(d)}\right)x_i^{(d)}$$
$$= -(y^{(d)} - o^{(d)})\, x_i^{(d)}.$$

Q2-1: In backpropagation, every weight is changed by $\Delta w_{ji} = -\eta \frac{\partial E}{\partial net_j} \frac{\partial net_j}{\partial w_{ji}} = \eta \delta_j o_i$, where $\delta_j = -\frac{\partial E}{\partial net_j} = -\frac{\partial E}{\partial o_j} \frac{\partial o_j}{\partial net_j}$. If j is a tanh output unit with $o_j = Tanh(net_j) = \frac{1-\exp(-2net_j)}{1+\exp(-2net_j)}$, and $E = \frac{1}{2}(y_j \ - o_j \ )^2$. Please calculate the δ_j here. Hint: $Tanh(z) = 2Sigmoid(2z) - 1$, so $Tanh'(z) = 1 - \left(Tanh(z)\right)^2$

1. $(y_j - o_j)(1 - o_j)o_j$
2. $(y_j - o_j)(1 - o_j^2)$
3. $y_j - o_j$
4. $\frac{y_j(1-o_j^2)}{o_j} - (1 - y_j)(1 + o_j)$

Q2-1: In backpropagation, every weight is changed by $\Delta w_{ji} = -\eta \frac{\partial E}{\partial net_j} \frac{\partial net_j}{\partial w_{ji}} = \eta \delta_j o_i$, where $\delta_j = -\frac{\partial E}{\partial net_j} = -\frac{\partial E}{\partial o_j} \frac{\partial o_j}{\partial net_j}$. If j is a tanh output unit with $o_j = Tanh(net_j) = \frac{1-\exp(-2net_j)}{1+\exp(-2net_j)}$, and $E = \frac{1}{2}(y_j \ - o_j \ )^2$. Please calculate the δ_j here. Hint: $Tanh(z) = 2Sigmoid(2z) - 1$, so $Tanh'(z) = 1 - \left(Tanh(z)\right)^2$

1. $(y_j - o_j)(1 - o_j)o_j$
2. $(y_j - o_j)(1 - o_j^2)$ ⬅
3. $y_j - o_j$
4. $\frac{y_j(1-o_j^2)}{o_j} - (1 - y_j)(1 + o_j)$

$$-\frac{\partial E}{\partial net_j}$$
$$= (y_j - o_j)\frac{\partial o_j}{\partial net_j}$$
$$= (y_j - o_j)(1 - o_j^2)$$
with tanh activation.

Q2-2: In backpropagation, every weight is changed by $\Delta w_{ji} = -\eta \frac{\partial E}{\partial net_j} \frac{\partial net_j}{\partial w_{ji}} = \eta \delta_j o_i$, where $\delta_j = -\frac{\partial E}{\partial net_j} = -\sum_k \frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial o_j} \frac{\partial o_j}{\partial net_j}$. If j is a hidden unit with the activation $o_j = Tanh(net_j) = \frac{1-\exp(-2net_j)}{1+\exp(-2net_j)}$, and $E = \frac{1}{2}(y_j - o_j)^2$. Please calculate the δ_j here. Hint: $Tanh(z) = 2Sigmoid(2z) - 1$, so $Tanh'(z) = 1 - \left(Tanh(z)\right)^2$

1. $\sum_k \delta_k w_{kj}$
2. $o_j(1 - o_j) \sum_k \delta_k w_{kj}$
3. $(1 - o_j^2) \sum_k \delta_k w_{kj}$
4. $(y_j - o_j)(1 - o_j^2)$

Q2-2: In backpropagation, every weight is changed by $\Delta w_{ji} = -\eta \frac{\partial E}{\partial net_j} \frac{\partial net_j}{\partial w_{ji}} = \eta \delta_j o_i$, where $\delta_j = -\frac{\partial E}{\partial net_j} = -\sum_k \frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial o_j} \frac{\partial o_j}{\partial net_j}$. If j is a hidden unit with the activation $o_j = Tanh(net_j) = \frac{1-\exp(-2net_j)}{1+\exp(-2net_j)}$, and $E = \frac{1}{2}(y_j - o_j)^2$. Please calculate the δ_j here. Hint:

$Tanh(z) = 2Sigmoid(2z) - 1$, so $Tanh'(z) = 1 - \left(Tanh(z)\right)^2$

1. $\sum_k \delta_k w_{kj}$
2. $o_j(1 - o_j) \sum_k \delta_k w_{kj}$
3. $(1 - o_j^2) \sum_k \delta_k w_{kj}$ ⟵
4. $(y_j - o_j)(1 - o_j^2)$

$$-\frac{\partial E}{\partial net_j}$$
$$= -\frac{\partial E}{\partial o_j} \frac{\partial o_j}{\partial net_j}$$
$$= -\sum_k \frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial o_j} \frac{\partial o_j}{\partial net_j}$$
$$= \sum_k \delta_k w_{kj} \, (1 - o_j^2)$$

with tanh activation.

Q3-1: Are these statements true or false?
(A) Backpropagation is based on the chain rule.
(B) Backpropagation contains only forward passes.

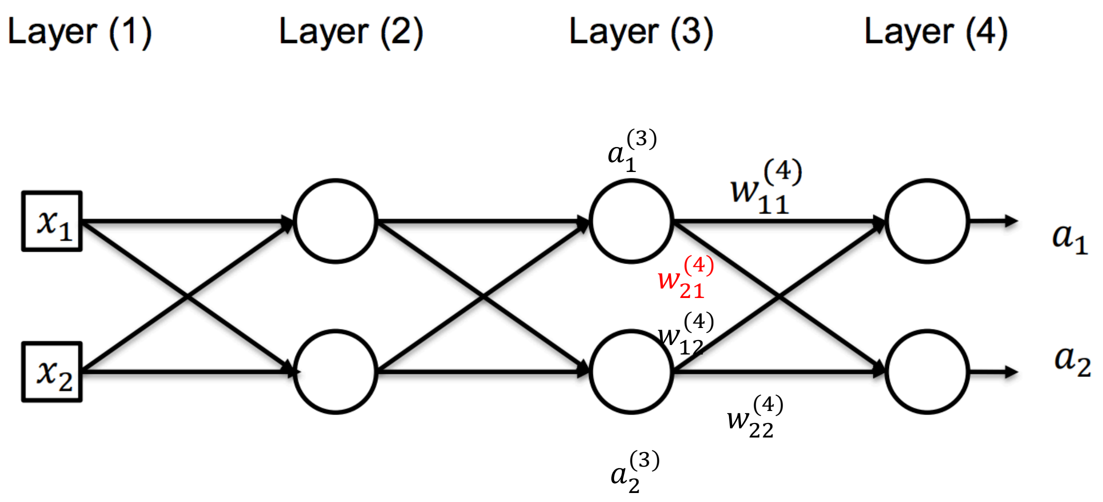
1. True, True
2. True, False
3. False, True
4. False, False

Q3-1: Are these statements true or false?
(A) Backpropagation is based on the chain rule.
(B) Backpropagation contains only forward passes.

1. True, True
2. True, False ←
3. False, True
4. False, False

(A) We use chain rule to calculate the partial derivatives of composite functions like neural network.
(B) It contains both forward and backward passes.

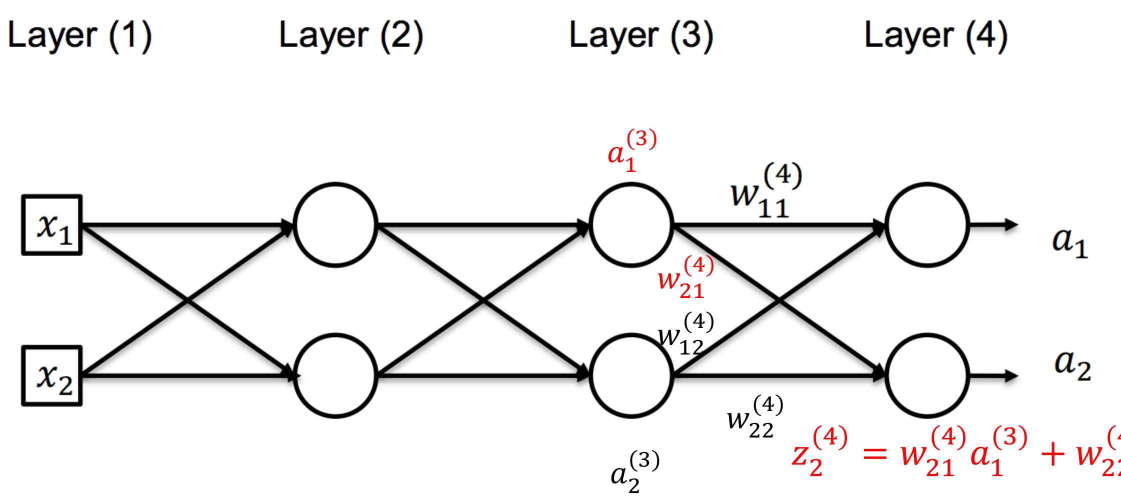
Q3-2: Please calculate the $\frac{\partial E_x}{\partial w_{21}^{(4)}}$ with sigmoid activation.



$E_x = \|y - a\|^2$

1. $2(a_1 - y_1)a_1(1 - a_1)a_1^{(3)}$
2. $2(a_2 - y_2)a_2(1 - a_2)a_1^{(3)}$
3. $2(a_2 - y_2)a_2(1 - a_2)a_2^{(3)}$
4. $2(a_1 - y_1)a_1(1 - a_1)a_2^{(3)}$

Q3-2: Please calculate the $\frac{\partial E_x}{\partial w_{21}^{(4)}}$ with sigmoid activation.



$E_x = \|y - a\|^2$

$z_2^{(4)} = w_{21}^{(4)} a_1^{(3)} + w_{22}^{(4)} a_2^{(3)}$

1. $2(a_1 - y_1)a_1(1 - a_1)a_1^{(3)}$
2. $2(a_2 - y_2)a_2(1 - a_2)a_1^{(3)}$ ←
3. $2(a_2 - y_2)a_2(1 - a_2)a_2^{(3)}$
4. $2(a_1 - y_1)a_1(1 - a_1)a_2^{(3)}$

$w_{21}^{(4)} a_1^{(3)}$, $w_{22}^{(4)} a_2^{(3)}$ → $+$ → $z_2^{(4)}$ $\xrightarrow{g(z_2^{(4)})}$ a_2 $\xrightarrow{\| y - a \|^2}$ E_x

$\frac{\partial a_2}{\partial z_2^{(4)}} = g'(z_2^{(4)})$

$\frac{\partial E_x}{\partial a_2} = 2(a_2 - y_2)$

$$\frac{\partial E_x}{\partial w_{21}^{(4)}} = 2(a_2 - y_2) g'\left(z_2^{(4)}\right) a_1^{(3)}$$

$$= 2(a_2 - y_2) g\left(z_2^{(4)}\right)\left(1 - g\left(z_2^{(4)}\right)\right) a_1^{(3)}$$

$$= 2(a_2 - y_2) a_2 (1 - a_2) a_1^{(3)}$$