# Deep Learning: Homework 3

Qi Luo
A02274095

2/13/2019

## Problem 1

1. For one-dimensional case, the gradient is the slope of a tangent which is a line and the slope can be represented as a single number. And we only can go left or right, and the function increase in one direction and decrease in the other.

2. Advantage is online learning can constanly updating its weights, thus its error calculation uses different weights for each input. Disadvantage is it is very noisy when the weights are updated.

3. The accuracy I got is: 75.11% based on learning rate is 3 and mini batch size is 10.

## Problem 2

1. $\delta^L = \nabla_a C \odot \sigma'(z^L)$ which we can represnt it as $\delta_j^L = \nabla_a C_j \sigma'(z_j^L)$ for the j-th element. Since $\Sigma'(z^L)$ is a square matrix whose diagonal entries are the values $\sigma'(z_j^L)$ and whose off-diagonal entries are zero, the matix multilication for $\delta_j^L = \sigma'(z_j^L)\nabla_a C_j$. Therefore, $\delta^L = \nabla_a C \odot \sigma'(z^L)$ can be written as $\delta^L = \sum'(z^L)\nabla_a C$.

2. Similar to the previous queation,$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$ can represent as $\delta_j^l = ((w_j^{l+1})^T \delta^{l+1})\sigma'(z_j^l)$ for the j-th element. Since $\Sigma'(z^L)$ is a square matrix whose diagonal entries are the values $\sigma'(z_j^L)$ and whose off-diagonal entries are zero, the matix multilication for $\delta^l = \Sigma'(z^l)(w^{l+1})^T \delta^{l+1}$ can be represent as $\delta_j^l = \sigma'(z_j^l)(w_j^{l+1})^T \delta^{l+1}$. Therefore, $\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$ can be rewritten as $\delta^l = \Sigma'(z^l)(w^{l+1})^T \delta^{l+1}$.

3. $\delta^l = \Sigma'(z^l)(w^{l+1})^T \delta^{l+1}$
$= \Sigma'(z^l)(w^{l+1})^T \Sigma'(z^{l+1})(w^{l+2})^T \delta^{l+2}$
$= \Sigma'(z^l)(w^{l+1})^T \ldots \Sigma'(z^{L-1})(w^L)^T \delta^L$
$= \Sigma'(z^l)(w^{l+1})^T \ldots \Sigma'(z^{L-1})(w^L)^T \Sigma'(z^L)\nabla_a C.$

4. $\sigma(z) = z$ and $\sigma'(z) = 1$ Therefore, the backprop algorithm is:
1) $\delta^L = \nabla_{a^L} C \odot \sigma'(z^L) = \nabla_{a^L} C$
2) $\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l) = (w^{l+1})^T \delta^{l+1} = (w^{l+1})^T (w^{l+2})^T \ldots (w^L)^T \nabla_{a^L} C$
3) and 4) will not change.

# Problem 3

1. Since $f$ is strictly convex, by defination,
$\forall x_1 \neq x_2 \in x, \forall t \in (0,1) : f(tx_1 + (1-t)x_2) < tf(x_1) + (1-t)f(x_2)$.
Assume $x_1, x_2$ are two global minimize, which $x_1 < x_2$ and $f(x_1) < f(x_2)$, then $tf(x_1) + (1-t)f(x_2) < tf(x_2) + (1-t)f(x_2) = f(x_2)$.
Therefore, $f(tx_1 + (1-t)x_2) < f(x_2)$. However, $x_2$ is assumed to be a global minimizer which is contradict with the condition. So, in order to satisfy the condition is $x_1 = x_2$, which means at most have one global minimizer in strictly convex function.

2. Since a convex function its Hessian is PSD matrix and the sum of two PSD matrix is also PSD, two convex function its Hessian sum is also a PSD matrix. Also, the Hessian of a function is a PSD matrix which function is convex. Therefore, the sum of two convex function is also convex.

3. The Hessian of $f$ is $A$. Therefore, if $A$ is PSD matrix then $f$ is convex; if $A$ is PD matrix, then $f$ is strirtly convex.

4. Assume that function $f(x) = x^3$ is convex, then $\forall x_1 \neq x_2 \in x, \forall t \in (0,1) :$
$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$.
When $x_1 = -3, x_2 = 1$, and $t = 0.7$,
$f(tx_1 + (1-t)x_2) = f(0.7 * (-3) + (1-0.7) * 1) = f(-1.8) = -5.832$
$tf(x_1) + (1-t)f(x_2) = 0.7 * (-3)^3 + (1-0.7) * (1)^3 = -18.6$
Then, $f(tx_1 + (1-t)x_2) > tf(x_1) + (1-t)f(x_2)$ which is contradict with the definition of convexity. Therefore, $f$ is not convex.(Proof by counterexample)

5. The Hessian of $f(x) = x^3$ is $6x$, when $x < 0$ the funtion is not convex.

6. The function $-f(x) = -ln(x)$, and its Hessian is $\frac{1}{x^2}$ which is always positive for all $x > 0$. Therefore, $-f(x)$ is convex. Hence, $f$ is concave.

7. The Hessian for $f(x) = ax + b$ is 0, since 0 is nonpositive and nonnegative, the function is both convex and concave. This function do not have a global minimum or maximum. I don't think there is any other function that are twice continuously differentiable and both convex and concave that do not have the form of an affine function. The reason is because we need the Hessian of the function is 0 and also twice continuously differentiable, except the affine function, there is no other function can satisfy all the requriements.