MATH 6490. NONLINEAR OPTIMIZATION IN MACHINE LEARNING.
ASSIGNMENT 2.

There are 3 problems, each problem is worth 5 points, total is 15 points.

1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function, show that f is m–strongly convex in the sense of (2.4) in the Lecture Notes, i.e, for any $x, y \in \mathbb{R}^n$

$$f((1-\alpha)x + \alpha y) \leq (1-\alpha)f(x) + \alpha f(y) - \frac{1}{2}m\alpha(1-\alpha)\|x-y\|_2^2 \ , \tag{0.1}$$

if and only if $\nabla^2 f(x) \succeq mI$ for all x, i.e., the eigenvalues $\lambda_1, ..., \lambda_n$ of $\nabla^2 f(x)$ satisfy $\min\limits_{1\leq i\leq n} \lambda_i \geq m$. You can make use of the following fact: Let A be a symmetric non–negative definite $n \times n$ matrix. Then the eigenvalues $\lambda_1, ..., \lambda_n$ of A satisfy $\min\limits_{1\leq i\leq n} \lambda_i \geq m$ if and only if for any $u \in \mathbb{R}^n$ we have $u^T A u \geq m\|u\|^2$.

(Hint: Suppose that the function f is strongly–m convex with the standard inequality (0.1) above for m–convexity. Set $z_\alpha = (1-\alpha)x + \alpha y$. Then by (0.1) we have

$$\alpha f(y) - \alpha f(x) \geq f(z_\alpha) - f(x) + \frac{1}{2}m\alpha(1-\alpha)\|x-y\|^2 \ .$$

Making use of Taylor's expansion

$$\alpha f(y) - \alpha f(x) \geq (\nabla f(x))^T(z_\alpha - x) + O(\|z_\alpha - x\|^2) + \frac{1}{2}m\alpha(1-\alpha)\|x-y\|^2 \ .$$

Since $z_\alpha \to x$ as $\alpha \to 0$, and $z_\alpha - x = \alpha(y-x)$, we can set $\alpha \to 0$ to get from above that for any $x, y \in \mathbb{R}^n$

$$f(y) \geq f(x) + (\nabla f(x))^T(y-x) + \frac{m}{2}\|y-x\|^2 \ . \tag{0.2}$$

Set $u \in \mathbb{R}^n$ and $\alpha > 0$, then we consider the Taylor's expansion

$$f(x+\alpha u) = f(x) + \alpha \nabla f(x)^T u + \frac{1}{2}\alpha^2 u^T \nabla^2 f(x + t\alpha u)u \tag{0.3}$$

for some $0 \leq t \leq 1$. We apply (0.2) with $y = x + \alpha u$ so that

$$f(x+\alpha u) \geq f(x) + \alpha(\nabla f(x))^T u + \frac{m}{2}\alpha^2\|u\|^2 \ . \tag{0.4}$$

Comparing (0.3) and (0.4) we see that for arbitrary choice of $u \in \mathbb{R}^n$ we have

$$u^T \nabla^2 f(x + t\alpha u)u \geq m\|u\|^2 \ .$$

This implies that $\nabla^2 f(x) \succeq mI$ as claimed. This shows the "only if" part.

For the "if" part, we assume that $\nabla^2 f(x) \succeq mI$. Then for any $z \in \mathbb{R}^n$ we have that $(x-z)^T \nabla^2 f(z + t(x-z))(x-z) \geq m\|x-z\|^2$. Thus

$$\begin{aligned} f(x) &= f(z) + (\nabla f(z))^T(x-z) + \frac{1}{2}(x-z)^T\nabla^2 f(z+t(x-z))(x-z) \\ &\geq f(z) + (\nabla f(z))^T(x-z) + \frac{m}{2}\|x-z\|^2 \ . \end{aligned} \tag{0.5}$$

Similarly

$$
\begin{aligned}
f(y) &= f(z) + (\nabla f(z))^T(y-z) + \frac{1}{2}(y-z)^T\nabla^2 f(z+t(y-z))(y-z) \\
&\geq f(z) + (\nabla f(z))^T(y-z) + \frac{m}{2}\|y-z\|^2 \ .
\end{aligned}
\tag{0.6}
$$

We consider $(1-\alpha)$(0.5)$+\alpha$(0.6) and we set $z=(1-\alpha)x+\alpha y$. This gives

$$
\begin{aligned}
&(1-\alpha)f(x)+\alpha f(y) \\
&\geq (\alpha+(1-\alpha))f(z) + (\nabla f(z))^T((1-\alpha)(x-z)+\alpha(y-z)) + \frac{m}{2}\left((1-\alpha)\|x-z\|^2+\alpha\|y-z\|^2\right) \\
&= f(z) + (\nabla f(z))^T((1-\alpha)(x-z)+\alpha(y-z)) + \frac{m}{2}\left((1-\alpha)\|x-z\|^2+\alpha\|y-z\|^2\right) \ .
\end{aligned}
\tag{0.7}
$$

Since $x-z=\alpha(x-y)$ and $y-z=(1-\alpha)(y-x)$, we see that $((1-\alpha)(x-z)+\alpha(y-z))=0$. Moreover, this means that

$$
(1-\alpha)\|x-z\|^2+\alpha\|y-z\|^2 = \left[(1-\alpha)\alpha^2+\alpha(1-\alpha)^2\right]\|x-y\|^2 = \alpha(1-\alpha)\|x-y\|^2 \ .
$$

From these we see that (0.7) is the same as saying

$$
(1-\alpha)f(x)+\alpha f(y) \geq f((1-\alpha)x+\alpha y) + \frac{m}{2}\alpha(1-\alpha)\|x-y\|^2 \ ,
$$

which is (0.1).)

2. Consider the fully connected-neural network via the recursive relation

$$
a^{[1]}=x, z^{[l]}=W^{[l]}a^{[l]}+b^{[l]}, a^{[l]}=\sigma(z^{[l-1]}) \text{ for } l=2,3,...,L \ ,
$$

in which l stands for the number of layers in the neural network, and $W=(W^{[1]},...,W^{[L]})$ and $b=(b^{[1]},...,b^{[L]})$ are the weight matrices and bias vectors. We can view the neural network function as a chain with $a^{[1]}=x$ and $a^{[L]}=g(x;\omega)$. Let the training data be one point (x,y). Then the loss function is given by

$$
C=C(W,b)=\frac{1}{2}\|y-a^{[L]}\|^2 \ .
$$

Set the *error* in the j–th neuron at layer l to be $\delta^{[l]}=(\delta_j^{[l]})_j$ (viewed as a column vector) with $\delta_j^{[l]}=\dfrac{\partial C}{\partial z_j^{[l]}}$. Prove the back-propagation relation

$$
\delta^{[l]}=\sigma'(z^{[l]})\circ(W^{[l+1]})^T\delta^{[l+1]} \ , \quad \text{for } 2\leq l\leq L-1 \ ,
$$

where for $x,y\in\mathbb{R}^n$, we let $x\circ y\in\mathbb{R}^n$ to be the Hadamard product defined by $(x\circ y)_i=x_iy_i$.

(Hint: Let the l-th layer contain n_l neurons. We can directly calculate that

$$
\delta_j^{[l]}=\frac{\partial C}{\partial z_j^{[l]}}=\sum_{k=1}^{n_{l+1}}\frac{\partial C}{\partial z_k^{[l+1]}}\frac{\partial z_k^{[l+1]}}{\partial z_j^{[l]}}=\sum_{k=1}^{n_{l+1}}\delta_k^{[l+1]}\frac{\partial z_k^{[l+1]}}{\partial z_j^{[l]}} \ .
$$

Since we have

$$z^{[l+1]} = W^{[l+1]}\sigma(z^{[l]}) + b^{[l+1]} \ ,$$

we see that we can further have

$$\frac{\partial z_k^{[l+1]}}{\partial z_j^{[l]}} = \frac{\partial [W^{[l+1]}\sigma(z^{[l]})]_k}{\partial z_j^{[l]}} = \left(W^{[l+1]} \frac{\partial \sigma(z^{[l]})}{\partial z_j^{[l]}} \right)_k \ ,$$

Notice that $\dfrac{\partial \sigma(z^{[l]})}{\partial z_j^{[l]}} = (0, ..., 0, \sigma'(z_j^{[l]}), 0, ..., 0)^T$ where the $\sigma'(z_j^{[l]})$ term lies on the j-th position, we have $W^{[l+1]} \dfrac{\partial \sigma(z^{[l]})}{\partial z_j^{[l]}} = \sigma'(z_j^{[l]}) W_j^{[l+1]}$, where $W^{[l+1]} = (W_j^{[l+1]})$. This gives us $\left(W^{[l+1]} \dfrac{\partial \sigma(z^{[l]})}{\partial z_j^{[l]}} \right)_k = \sigma'(z_j^{[l]}) W_{jk}^{[l+1]}$, where W_{jk} is the element at the j-th column and k-th row of W. So we finally get

$$\delta_j^{[l]} = \sum_{k=1}^{n_{l+1}} \delta_k^{[l+1]} \sigma'(z_j^{[l]}) W_{jk}^{[l+1]} = \sigma'(z_j^{[l]}) [(W^{[l+1]})^T \delta^{[l+1]}]_j \ ,$$

which leads to what we wanted to prove.)

3. Consider solving an optimization problem $\min_{x \in \mathbb{R}^n} f(x)$ via a line search method of the form $x^{k+1} = x^k + \alpha d^k$ where the stepsize $\alpha > 0$ and the descent direction $d^k \in \mathbb{R}^n$. Assume one can obtain an inequality of the form

$$f(x^{k+1}) \leq f(x^k) - C\|\nabla f(x^k)\|^2$$

for some constant $C > 0$. Assume that the objective function f is bounded from below and its gradient ∇f is L–Lipschitz. Suppose there is a subsequence $x_{n_k} \to \bar{x}$ as $k \to \infty$, show that $\nabla f(\bar{x}) = 0$. In particular, if the function f is convex, this implies that $\bar{x}$ is a solution to the optimization problem.

(Hint: Rewrite the inequality into $\|\nabla f(x^k)\|^2 \leq \dfrac{f(x^k) - f(x^{k+1})}{C}$ and use the fact that $\{f(x^k)\}_{k \geq 1}$ is a monotonically decreasing sequence that is bounded from below. This indicates that $\lim_{k \to \infty} [f(x^k) - f(x^{k+1})] = 0$, which implies that $\lim_{k \to \infty} \|\nabla f(x^k)\| = 0$. As we have $x^{n_k} \to \bar{x}$ as $k \to \infty$, by the continuity of $\nabla f(x)$ we have $\nabla f(\bar{x}) = \lim_{k \to \infty} \nabla f(x^{n_k}) = 0$, as claimed.)