SOLUTIONS HW 3

Problem 1

(a) Note that f has a unique global minimum at $x^* = 0$, $\nabla f(x) = 4x^3$, and $\nabla^2 f(x) = 12x^2$. Then for $x_k \neq 0$:

$$x_{k+1} = x_k - \frac{\alpha(4x_k)^3}{12x_k^2} = (1 - \frac{\alpha}{3})x_k.$$

Therefore, as long as $|1 - \frac{\alpha}{3}| < 1$, x_k converges to $x^* = 0$ as $k \to \infty$. The range of α can be found using $|1 - \frac{\alpha}{3}| < 1 \Rightarrow 0 < \alpha < 6$. Note that for $\alpha = 3$, the method converges in one step.

For this range of α and any $x_0 \in \mathbb{R}$, we can show

$$x_k = (1 - \frac{\alpha}{3})^k x_0,$$

hence x_k converges to 0 geometrically, i.e., the method converges "linearly".

(b)

$$\nabla f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \qquad \nabla^2 f(x) = \frac{4e^{2x}}{(e^{2x} + 1)^2}.$$

Substituting values in formula for Newton's method for $\alpha = 1$, we get the desired expression.

Example code:

```
import numpy as np
alpha = 1
x = 1
n = 5
iterates = np.zeros(n)
for i in range(n):
    x = x - (np.exp(4*x)-1)/(4*np.exp(2*x))
    iterates[i] = x
print(iterates)
```

For initialization $x_0 = 1$, iterates are:

$$\begin{bmatrix} -8.13430204e-01 & 4.09402317e-01 & -4.73049165e-02 & 7.06028036e-05 & -2.34633642e-13 \end{bmatrix}.$$

For initialization $x_0 = 1.1$, iterates are:

$$\begin{bmatrix} -1.12855259e+00 & 1.23413113e+00 & -1.69516598e+00 & 5.71536010e+00 & -2.30213565e+04 \end{bmatrix}.$$

The iterates converges to $x^* = 0$ with $x_0 = 1$ and diverges for $x_0 = 1.1$. Newton's method converges as long as the initial estimate is sufficiently close to x^*.

(c) Since the cost function is quadratic and strongly convex, the Newton's method converges in one step as we discussed in Lecture.

Problem 2

(a) For layer M, starting with the hint, and following steps similar to those in the lecture note:

$$\begin{aligned}\frac{\partial J}{\partial W_{ij}^{(M)}} &= \sum_{n=1}^{N} \frac{\partial J}{\partial y_i^{(M)}[n]} \frac{\partial y_i^{(M)}[n]}{\partial W_{ij}^{(M)}} \\ &= \sum_{n=1}^{N} \frac{\partial J}{\partial y_i^{(M)}[n]} \frac{\partial y_i^{(M)}[n]}{\partial x_i^{(M)}[n]} \frac{\partial x_i^{(M)}[n]}{\partial W_{ij}^{(M)}} \\ &= \sum_{n=1}^{N} \frac{\partial J}{\partial y_i^{(M)}[n]} \sigma'(x_i^{(M)}[n]) y_j^{(M-1)}[n]\end{aligned}$$

and

$$\frac{\partial J}{\partial b_i^{(M)}} = \sum_{n=1}^{N} \frac{\partial J}{\partial y_i^{(M)}[n]} \sigma'(x_i^{(M)}[n])$$

For layers $M-1, \cdots, 1$, going backwards, we can again follow the steps in the lecture notes (with the additional summation over the datapoints) to obtain:

$$\frac{\partial J}{\partial W_{ij}^{(m)}} = \sum_{n=1}^{N} \frac{\partial J}{\partial y_i^{(m)}[n]} \sigma'(x_i^{(m)}[n]) y_j^{(m-1)}[n]$$

$$\frac{\partial J}{\partial b_i^{(m)}} = \sum_{n=1}^{N} \frac{\partial J}{\partial y_i^{(m)}[n]} \sigma'(x_i^{(m)}[n])$$

with

$$\frac{\partial J}{\partial y_i^{(m)}[n]} = \sum_{k} \frac{\partial J}{\partial y_k^{(m+1)}[n]} \sigma'(x_k^{(m+1)}[n]) W_{ki}^{(m+1)},$$

where the summation is over the number of outputs of layer $m+1$.

(b) Note from part (a) that running the back-propagation algorithm directly on J results in the summation of independent terms obtained by running the algorithm on the loss corresponding to the individual data-points. Given a single computing machine, there is no computational advantage of running the algorithm directly on J.

The advantage of SGD: It is computationally fast as only a subset of the training set is processed at a time. For larger datasets, it requires less computation resources.

Problem 3

(a) True. Suppose $\mathcal{A} = \{x \in \mathbb{R}^n : Ax = b\}$. If $x_1, x_2 \in \mathcal{A}$, for any $\lambda \in [0, 1]$, we have $A(\lambda x_1 + (1-\lambda)x_2) = \lambda A x_1 + (1-\lambda)Ax_2 = b$. Therefore, $\lambda x_1 + (1-\lambda)x_2 \in \mathcal{A}$. $\mathcal{A}$ is a convex set.

(b) False. If $f(x) = x^2$ and $r = 1$, then the set $\{x \in R : x^2 = 1\} = \{1, -1\}$, which is clearly not convex.

(c) True. It suffices to show $f(x) = x^\mathsf{T} Q x$ is a convex function. Since Q is a positive semidefinite matrix, i.e., $Q \geq 0$, we have $\nabla^2 f(x) = 2Q \geq 0$, which is positive semidefinite as well. Therefore, $f(x)$ is a convex function. Then the considered set is a convex set by using the results in Lecture 3 (page 8).

(d) True. Since $f(x)$ is μ strongly convex, we have:

$$f(y) \geq f(x) + \nabla f(x)^\mathsf{T}(y - x) + \frac{\mu}{2}\|y - x\|^2.$$

Minimizing the both sides of the above inequality with respect to y:

$$\min\{\text{LHS}\} = \min\{f(y)\} = f(x^*)$$
$$\frac{\partial \text{RHS}}{\partial y} = \nabla f(x) + \mu(y - x) \Rightarrow y^* = x - \frac{1}{\mu}\nabla f(x) \Rightarrow \min\{\text{RHS}\} = f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2$$

Overall, we have $f(x^*) \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2 \Rightarrow f(x) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x)\|^2$.