DL 2023040165 Zhicheng Jiang

Assignment 1

Question 1 to 3

False, True, False.

Question 4

Proof:
By μ-strong convexity, we have:

$$f(x^*) \geq f(x^k) + \nabla f(x^k)^T(x^* - x^k) + \frac{\mu}{2}\|x^* - x^k\|^2$$

By the definition of gradient descent and the L-Lipschitz continuity of ∇f, we have:

$$\begin{aligned} f(x^*) \leq f(x^{k+1}) \leq f(x^k) - \nabla f(x^k)^T(x^{k+1} - x^k) + \frac{1}{2L}\|\nabla f(x^k)\|^2 \\ = f(x^k) - \frac{L}{2}\|x^{k+1} - x^k\|^2 \end{aligned}$$

Combining the two inequalities, we have:

$$\frac{L}{2}\|x^{k+1} - x^k\|^2 \leq \nabla f(x^k)^T(x^* - x^k) + \frac{1}{2L}\|\nabla f(x^k)\|^2 + \frac{\mu}{2}\|x^* - x^k\|^2$$

Notice that $\nabla f(x^k) = L(x^{k+1} - x^k)$, we have:

$$\frac{L}{2}\|x^{k+1} - x^*\|^2 \leq \frac{L-\mu}{2}\|x^k - x^*\|^2$$

So, the number of steps required to reach ϵ-accuracy is $O(\log_{\frac{1}{1-\frac{\mu}{L}}} \frac{R^2}{\epsilon^2}) = O(\frac{L}{\mu}\log\frac{R}{\epsilon})$.

Question 5

We have: $\nabla f(x) = \begin{cases} 50x, x < 1 \\ 2x + 48, 1 \leq x \leq 2 \\ 50x - 48, x > 2 \end{cases}$

Given $x^0 = 3.3$, we can easily compute that $x^1 = 3.3 - 1/9 * (165 - 48) = -9.7$

We prove by induction that $|x^0| < |x^1| < |x^2| < \cdots < |x^{2k-1}|$, and the even terms are positive and the odd terms are negative.(So that it won't converge)
The base case is true.
Assume that the statement is true for k, then we have:

$$x^{2k-1} < x^1 < 1 \to x^{2k} = x^{2k-1} - \frac{1}{9} * (50x^{2k-1}) + \frac{4}{9} * (x^{2k-1} - x^{2k-2}) = -\frac{37}{9}x^{2k-1} - \frac{4}{9}x^{2k-2} > -x^{2k-1}$$

$$x^{2k} > x^0 > 2 \to x^{2k+1} = x^{2k} - \frac{1}{9} * (50x^{2k} - 48) + \frac{4}{9} * (x^{2k} - x^{2k-1})$$
$$= -\frac{37}{9}x^{2k} - \frac{4}{9}x^{2k-1} + \frac{48}{9} < \frac{-35}{9}x^{2k} + \frac{48}{9} < -x^{2k}$$

Q.E.D.

Question 6

We write $\nabla f(x^*) = \nabla f(x^k) + \nabla^2 f(x^k)(x^* - x^k) + \theta$
By lipchitz continuity of ∇^2 , we have:

$$\|\theta\| = O(1)\|x^* - x^k\|^2$$

Then we know that

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1}(\nabla f(x^*) - \nabla^2 f(x^k)(x^* - x^k) - \theta)$$
$$= x^* - \nabla^2 f(x^k)^{-1}\theta$$

therefore, we have:

$$\|x^{k+1} - x^*\| = \|\nabla^2 f(x^k)^{-1}\theta\| \le O(1)\|\nabla^2 f(x^k)^{-1}\|\|x^* - x^k\|^2$$

By strongly convexity, $\|\nabla^2 f(x^k)^{-1}\| \le \frac{1}{\mu}$, so we have:

$$\|x^{k+1} - x^*\| \le O(1)\frac{1}{\mu}\|x^k - x^*\|^2$$

Q.E.D.

Question 7

First, we prove $\rho_Z(j) = \rho_Z(-j)$ That's because W is i.i.d Guassian, so W and -W have the same distribution.

Then we can know that WX and -WX have the same distribution, by linearlity, Z and -Z have the same distribution.
Then, after the ReLU, the variance of Z reduces by half, so we have:

$$Var(W_1x_1 + W_2x_2 + \cdots + W_{h_l}x_{h_l}) = h_l Var(X^l)Var(W^l)$$

So, to keep it unchange, we need to have:

$$Var(W^l) = \frac{2}{h_l}$$