Deep Learning 2023040163 Zhao Han Hong

Homework 1

1 True or False Questions

Problem 1

False.

Problem 2

True.

Problem 3

False.

2 Q & A

Problem 4

We first prove the descent lemma mentioned in the class.

Descent Lemma $f(y) \leq f(x) + \nabla f(y-x)^T + \frac{L}{2}\left|\left|y-x\right|\right|^2$.

Proof By L smoothness, we have

$$\begin{aligned}
f(y) &= f(x) + \int_0^1 \nabla f(x+t(y-x))^T(y-x)dt \\
&\leq f(x) + \nabla f(x)^T(y-x) + \int_0^1 [\nabla f(x+t(y-x)) - \nabla f]^T(y-x)dt \\
&\leq f(x) + \nabla f(x)^T(y-x) + \int_0^1 \left|\left|\nabla f(x+t(y-x)) - \nabla f\right|\right| \cdot \left|\left|y-x\right|\right| dt \\
&\leq f(x) + \nabla f(x)^T(y-x) + \int_0^1 L\left|\left|t(y-x)\right|\right| \cdot \left|\left|y-x\right|\right| dt \\
&= f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}L\left|\left|y-x\right|\right|^2,
\end{aligned}$$

so we are done.

Now, we can prove the convergence. First, use our lemma and obtain

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

Next, we use the μ-convergence to get

$$\begin{aligned} f(x^\star) &\geq f(x^k) + \nabla f(x^k)^T(x^\star - x^k) + \frac{\mu}{2}\left|\left|x^k - x^\star\right|\right|^2 \\ &\geq f(x^k) + L(x^k - x^{k+1})^T(x^\star - x^k) + \frac{\mu}{2}\left|\left|x^k - x^\star\right|\right|^2. \end{aligned}$$

Putting these two together and note that $f(x^{k+1}) \geq f(x^\star)$, we then have

$$\frac{\mu}{2}\left|\left|x^k - x^\star\right|\right|^2 \leq -\frac{L}{2}\left|\left|x^k - x^{k+1}\right|\right|^2 - L(x^k - x^{k+1})^T(x^\star - x^k).$$

But notice that

$$\begin{aligned} & -\frac{L}{2}\left|\left|x^k - x^{k+1}\right|\right|^2 - L(x^k - x^{k+1})^T(x^\star - x^k) \\ =& \left(-\frac{L}{2}\right)(x^k - x^{k+1})^T(-x^{k+1} + 2x^\star - x^k) \\ =& \frac{L}{2}((x^\star - x^k) - (x^\star - x^{k+1}))^T((x^\star - x^k) + (x^\star - x^{k+1})) \\ =& \frac{L}{2}\left(\left|\left|x^\star - x^k\right|\right|^2 - \left|\left|x^\star - x^{k+1}\right|\right|^2\right), \end{aligned}$$

so we have

$$\frac{\mu}{2}\left|\left|x^k - x^\star\right|\right|^2 \leq \frac{L}{2}\left(\left|\left|x^\star - x^k\right|\right|^2 - \left|\left|x^\star - x^{k+1}\right|\right|^2\right).$$

This implies that

$$\left|\left|x^k - x^\star\right|\right|^2 \leq \left(1 - \frac{\mu}{L}\right)^k \left|\left|x^0 - x^\star\right|\right|^2,$$

so the iteration time

$$k = \frac{\ln\left(\left(\frac{R}{\epsilon}\right)^2\right)}{-\ln\left(1 - \frac{\mu}{L}\right)} = \mathcal{O}\left(\frac{L}{\mu}\left(2\ln\frac{R}{\epsilon}\right)\right) = \mathcal{O}\left(\frac{L}{\mu}\ln\frac{R}{\epsilon}\right)$$

is enough.

Problem 5

We use the original function, i.e.

$$f(x) = \begin{cases} 25x^2 & \text{if } x \le 1 \\ x^2 + 48x - 24 & \text{if } 1 < x \le 2 \\ 25x^2 - 48x + 72 & \text{otherwise} \end{cases}.$$

The sequence is uniquely determined after the first term x^0 is given. Now we state that **for even terms,** $x > 2$**; for odd terms,** $x < 1$. We first assume that and only have to verify it afterward. To facilitate our discussion, let $a_0 = b_0 = 3.3$, $x^{2k-1} = b_k, x^{2k} = a_k$. Then, a calculation yields

$$\begin{cases} b_{n+1} = \frac{48}{9} - \frac{37}{9}a_n - \frac{4}{9}b_n \\ a_{n+1} = -\frac{37}{9}b_{n+1} - \frac{4}{9}a_n = -\frac{1776}{81} + \frac{1333}{81}a_n + \frac{148}{81}b_n \end{cases}.$$

Let $c_n = a_n - 1.48, d_n = b_n + 0.52$, then

$$\begin{pmatrix} c_{n+1} \\ d_{n+1} \end{pmatrix} = \begin{pmatrix} \frac{1333}{81} & \frac{148}{81} \\ -\frac{37}{9} & -\frac{4}{9} \end{pmatrix} \begin{pmatrix} c_n \\ d_n \end{pmatrix} = \begin{pmatrix} 1 & -4 \\ -9 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{81} & 0 \\ 0 & 16 \end{pmatrix} \frac{1}{-35} \begin{pmatrix} 1 & 4 \\ 9 & 1 \end{pmatrix} \begin{pmatrix} c_n \\ d_n \end{pmatrix}.$$

Thus,

$$\begin{pmatrix} c_n \\ d_n \end{pmatrix} = \begin{pmatrix} 1 & -4 \\ -9 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{81^n} & 0 \\ 0 & 16^n \end{pmatrix} \frac{1}{-35} \begin{pmatrix} 1 & 4 \\ 9 & 1 \end{pmatrix} \begin{pmatrix} \frac{91}{50} \\ \frac{191}{50} \end{pmatrix} = \begin{pmatrix} -\frac{171}{350}\frac{1}{81^n} + \frac{404}{175}16^n \\ \frac{1539}{350}\frac{1}{81^n} - \frac{101}{175}16^n \end{pmatrix},$$

so we have solved the whole sequence. We can immediately notice that $a_n > a_0 > 2$ and $b_n \le b_1 < 1$ for $n \ge 1$, so our assumption holds. Also, due to the factor 16, we know that the sequence is not going to converge.

Problem 6

Assume that $\nabla^2 f(x^k)$ is M-Lipchitz, then $\nabla f(x^k)$ is M-smooth. This leads to

$$\left|\left|\nabla f(x^\star) - \nabla f(x^k) - \nabla^2 f(x^k)(x^\star - x^k)\right|\right| \le \frac{M}{2}\left|\left|x^k - x^\star\right|\right|^2.$$

But $\nabla f(x^\star) = 0$, so we can estimate the distance between x^{k+1} and $x^\star$:

$$\begin{aligned} \left|\left|x^{k+1} - x^\star\right|\right| &= \left|\left|x^k - x^\star - \left(\nabla^2 f(x^k)\right)^{-1}\nabla f(x^k)\right|\right| \\ &= \left|\left|\left(\nabla^2 f(x^k)\right)^{-1}\left(\nabla^2 f(x^k)(x^k - x^\star) - \nabla f(x^k)\right)\right|\right| \\ &\le \frac{M}{2}\left|\left|\nabla^2 f(x^k)\right|\right|^{-1}\left|\left|x^k - x^\star\right|\right|^2 \le \frac{M}{2\mu}\left|\left|x^k - x^\star\right|\right|^2, \end{aligned}$$

so we are done.

Problem 7

We first demonstrate that $Z^l (l \geq 1)$ has a symmetric probability distribution. We prove this by showing that the probability distribution of the random variable $u = W_{ij}^l X_j^l = wv$ is symmetric, where j is arbitrary. We first write

$$p_u(a) = \int p_w \left(\frac{a}{t}\right) p_v(t) dt,$$

where the integration is over the t at which both of the two probabilities are nonzero. Now,

$$p_u(-a) = \int p_w \left(\frac{-a}{t}\right) p_v(t) dt = \int p_w \left(\frac{a}{t}\right) p_v(t) dt = p_u(a)$$

by to the symmetry of p_w. Then, we are done.

After that, we then know that ReLU will reduce the variance by half, namely,

$$\begin{aligned}
Var(Z_i^l) &= Var\left(\sum_j W_{ij}^l \text{ReLU}(Z_j^{l-1})\right) \\
&= \sum_j E\left((W_{ij}^l)^2 \text{ReLU}(Z_j^{l-1})^2\right) - \left(\sum_j E(W_{ij}^l \text{ReLU}(Z_j^{l-1}))\right)^2 \\
&= \sum_j Var(W_{ij}^l) E\left((\text{ReLU}(Z_j^{l-1}))^2\right) \\
&= \sum_j Var(W_{ij}^l) \cdot \frac{1}{2} E((Z_j^{l-1})^2) = \frac{1}{2} Var(W^l) Var(Z^{l-1}).
\end{aligned}$$

(Here Z^{l-1} is the total variance for the $(l-1)$-layer neurons.) Now, since

$$Var(Z^l) = h_l Var(Z_i^l),$$

we immediately obtain that $Var(W^l) = \frac{2}{h_l}$, finishing the proof.