Mathematical Foundation of DNN : HW 1

Jeong Min Lee

March 5, 2024

1

In this problem, I followed the notation given by the Petersen and Pedersen[1]. Also, I denote the element of matrix X in ith row and jth column as X_{ij}

a

$$\left(\frac{\partial}{\partial\theta}l_i(\theta)\right)_j = \frac{\partial l_i(\theta)}{\partial\theta_j} \tag{1}$$

$$= (X_i^T\theta - Y)X_{ij} \tag{2}$$

By enumberating the last equation in column vector, one can get the following result.

$$\frac{\partial}{\partial\theta}l_i(\theta) = (X_i^T\theta - Y_i)X_i \tag{3}$$

Note that $(X_i^T\theta - Y_i) \in \mathbb{R}$ and thus, enumeration affects only to the last X_{ij}. (It results in X_i)

b

$$\mathcal{L}(\theta) = \frac{1}{2}\|X\theta - Y\|^2 \tag{4}$$

$$= \frac{1}{2}\sum_i \left(X_i^T\theta - Y_i\right)^2 \tag{5}$$

$$= \sum_i l_i(\theta) \tag{6}$$

From the obseravtion above and the result of the problem(a),

$$\nabla_\theta\mathcal{L}(\theta) = \sum_i \nabla_\theta l_i(\theta) \tag{7}$$

$$= \sum_i (X_i^T\theta - Y_i)X_i \tag{8}$$

$$= \sum_i X_i^T\theta X_i - X_iY_i \tag{9}$$

According to the hint given by the original problem statement, noting that the matrix consisting of the X_i as a column is X^T,

$$\nabla_\theta\mathcal{L}(\theta) = X^T\begin{pmatrix} X_1^T\theta \\ X_2^T\theta \\ \vdots \\ X_N^T\theta \end{pmatrix} - X^TY \tag{10}$$

$$= X^T\begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_N^T \end{pmatrix}\theta - X^TY \tag{11}$$

$$= X^TX\theta - X^TY \tag{12}$$

$$= X^T(X\theta - Y) \tag{13}$$

2

Since $f'(\theta) = \theta$,

$$\theta^{(k+1)} = \theta^{(k)} - \alpha f'(\theta^{(k)}) \tag{14}$$
$$= (1-\alpha)\theta^{(k)} \tag{15}$$

[1]

$$\frac{\theta^{(k+1)}}{\theta^{(k)}} = 1 - \alpha \tag{16}$$
$$\therefore \theta^{(k)} = \theta^{(0)}(1-\alpha)^k \tag{17}$$

If $\alpha > 2$, then $|1 - \alpha| > 1$. This results in $\theta^{(k)} \to \infty$ as $k \to \infty$

3

From problem 1, I showed the following.

$$\nabla f(\theta) = X^T(X\theta - Y) \tag{18}$$

Inserting it to the GD,

$$\theta^{(k+1)} = \theta^{(k)} - \alpha X^T(X\theta^{(k)} - Y) \tag{19}$$
$$= \theta^{(k)} - \alpha X^T X\theta^{(k)} + \alpha X^T Y \tag{20}$$
$$= \theta^{(k)} - \alpha X^T X\theta^{(k)} + \alpha X X^T \theta^\star \tag{21}$$

By substracting $\theta^\star$ on both side of equation 21, the following equation is derived.

$$\theta^{(k+1)} - \theta^\star = (I_p - \alpha X^T X)(\theta^{(k)} - \theta^\star) \tag{22}$$

Note that I_p denotes identity matrix whose dimension is $p \times p$.

4

I implement my own gradient descent algorithm to solve this problem. As mentioned in homework document, I use three learning rate $\alpha = [0.01, 0.3, 4]$. Furthermore, for the robust analysis, I randomly sampled 50 initial point of x for each α. This result agrees to the general notion that GD with small learning rate temps to converge both of local minma, while that with intermediate learning rate converges only to the wide local minimum. The following python code is the code used to solve this problem.

```python
import numpy as np
import matplotlib.pyplot as plt
alpha_lst = [0.01, 0.3,4]
iter_num = 50 # number of sampling
epsilon = 1e-4 # acceptable level
result = dict() # contain the result of experiment
max_step = 1000 # maximum step not to run while loop infintely

for alpha in alpha_lst:
    print(f"--------------------------alpha = {alpha}---------------------------------------")
    result[alpha] = []
    for i in range(iter_num):
        x = 25*np.random.random_sample()-5
        x_init = x
        step = 0
        while fprime(x) > epsilon or step < max_step:
            print(f"step : {step}/{max_step}")
            step +=1
            x = x - alpha * fprime(x)
        result[alpha].append((x,x_init))
```

[1] To resolve the confusion due to the notation between kth element and power of k, I used parenthesis to denote the kth element

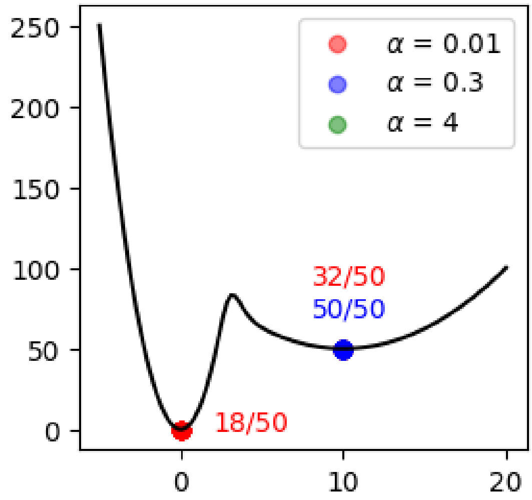

Figure 1: The result of gradient descent of $f(x)$ is described. The proportion of initial points that converges to each local minimum is depicted. The color of given ratio matches to the each α. Note that there is no scatter in this plot corresponding to $\alpha = 4$ since none of them converged.

References

[1] K. B. Petersen and M. S. Pedersen. The matrix cookbook, October 2008. Version 20081110.