P1

(a)

$$l_i(\theta) = \frac{1}{2}(X_i^T\theta - Y_i)^2 = \frac{1}{2}\left(\sum_j X_{ij}\theta_j - Y_i\right)^2$$

$$\frac{\partial}{\partial\theta_k} l_i(\theta) = \frac{\partial}{\partial\theta_k}\left(\frac{1}{2}\left(\sum_j X_{ij}\theta_j - Y_i\right)^2\right)$$

$$= \left(\sum_j X_{ij}\theta_j - Y_i\right) X_{ik} = (X_i^T\theta - Y_i) X_{ik}$$

$$\therefore \nabla_\theta l_i(\theta) = (X_i^T\theta - Y_i)(X_{i1}\ X_{i2} \cdots X_{ip})^T = (X_i^T\theta - Y_i)X_i \quad ※$$

(b)

$$f(\theta) = \frac{1}{2}\|X\theta - Y\|^2 = \frac{1}{2}\sum_{i=1}^N (X\theta - Y)_i^2 = \frac{1}{2}\sum_{i=1}^N\left(\sum_{j=1}^p X_{ij}\theta_j - Y_i\right)^2 = \sum_{i=1}^N\sum_{j=1}^p \frac{1}{2}(X_{ij}\theta_j - Y_i)^2$$

$$\therefore \frac{\partial}{\partial\theta_k} f(\theta) = \sum_{i=1}^N\left(\sum_{j=1}^p X_{ij}\theta_j - Y_i\right)X_{ik} = \sum_{i=1}^N (X\theta - Y)_i X_{ik}$$

$$\therefore \nabla_\theta f(\theta) = ((X\theta - Y)^T X)^T = X^T(X\theta - Y) \quad ※$$

P2

$$\theta_{k+1} = \theta_k - \alpha f'(\theta_k) = \theta_k - \alpha\theta_k = (1-\alpha)\theta_k$$

$$= (1-\alpha)^{k+1}\theta_0$$

if $|1-\alpha| > 1$, it diverges $\therefore \alpha > 2$ makes GD diverge.

P3

$$\theta_{k+1} - \theta_* = \theta_k - \alpha X^T(X\theta_k - Y) - \theta_*$$

$$= \theta_k - \alpha X^T(X\theta_k - X\theta_*) - \theta_* \quad (\because X\theta_* = Y)$$

$$= (I - \alpha X^TX)(\theta_k - \theta_*)$$

$I - \alpha X^TX$ is invertible

$$\theta_k - \theta_* = (I - \alpha X^TX)^{-1}(\theta_{k+1} - \theta_*) = (I + \alpha X^TX + (\alpha X^TX)^2 + \cdots)(\theta_{k+1} - \theta_*)$$

Let v be eigenvector corresponding with ρ.

$$v^T(\theta_k - \theta_*) = (v^T + \alpha(X^TXv)^T + \alpha^2(X^TXX^TXv^T) + \cdots)(\theta_{k+1} - \theta_*)$$

$$= (v^T + \rho\alpha v^T + \rho^2\alpha^2 v^T + \cdots)(\theta_{k+1} - \theta_*)$$

$$= v^T(I + \rho\alpha I + \rho^2\alpha^2 I + \cdots)(\theta_{k+1} - \theta_*)$$

$$= v^T \frac{1}{1-\rho\alpha}(\theta_{k+1} - \theta_*)$$

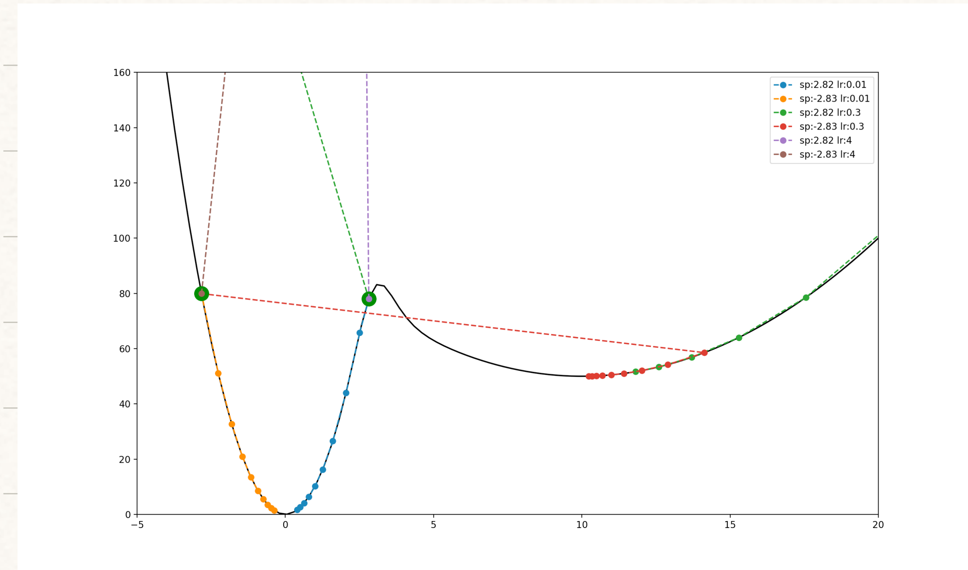
$$V^T(\theta_k - \theta_*) = \frac{1}{1-\rho\alpha} V^T(\theta_{k+1} - \theta_*)$$

$$|1-\rho\alpha| \, \|V^T(\theta_k - \theta_*)\| = \|V^T(\theta_{k+1} - \theta_*)\|$$

Using result of problem 2, $\rho\alpha > 2 \iff \alpha > \frac{2}{\rho}$ makes $\{\|V^T(\theta_{k+1} - \theta_*)\|\}$ diverge.

$\|V^T\| < \infty$, thus $\alpha > \frac{2}{\rho}$ makes $\{\theta_k\}$ diverge.

P4



You can see with learning rate (lr) 0.01, it approaches to sharp minima as iteration goes on. Meanwhile, with lr 0.3, they converges to wide minima. On the other hand, with lr 4, gradient descent diverges, thus aren't shown in the figure even after their first iteration.

P5

```
❯ python3 -u "/Users/yongjae/Desktop/SNU/Mathematical-Foundations-of-DNN/hw/week1/conv1D.py"
5.098441152369599
```