MathDNN Homework 1

Department of Computer Science and Engineering
2021-16988 Jaewan Park

Problem 1

(a) Let $X_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$ and $\theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}$, then for $j = 1, \cdots, p$,

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} l_i(\theta) &= \frac{\partial}{\partial \theta_j} \left[\frac{1}{2} (X_i^\intercal \theta - Y_i)^2 \right] \\
&= \frac{\partial}{\partial \theta_j} \left[\frac{1}{2} (x_{i1}\theta_1 + \cdots + x_{ip}\theta_p - Y_i)^2 \right] \\
&= (x_{i1}\theta_1 + \cdots + x_{ip}\theta_p - Y_i) x_{ij} \\
&= (X_i^\intercal \theta - Y_i) x_{ij}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\nabla_\theta l_i(\theta) &= \begin{bmatrix} \frac{\partial}{\partial \theta_1} l_i(\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_p} l_i(\theta) \end{bmatrix} \\
&= \begin{bmatrix} (X_i^\intercal \theta - Y_i) x_{i1} \\ \vdots \\ (X_i^\intercal \theta - Y_i) x_{ip} \end{bmatrix} = (X_i^\intercal \theta - Y_i) \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix} \\
&= (X_i^\intercal \theta - Y_i) X_i.
\end{aligned}$$

(b) Let $X_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$ and $\theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}$, then for $j = 1, \cdots, p$,

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} \mathcal{L}(\theta) &= \frac{\partial}{\partial \theta_j} \left[\frac{1}{2} \left\| X\theta - Y \right\|^2 \right] \\
&= \frac{\partial}{\partial \theta_j} \left[\frac{1}{2} \left\| \begin{matrix} X_1^\intercal \theta - Y_1 \\ \vdots \\ X_N^\intercal \theta - Y_N \end{matrix} \right\|^2 \right] = \frac{\partial}{\partial \theta_j} \left[\sum_{i=1}^N \frac{1}{2} (X_i^\intercal \theta - Y_i)^2 \right] = \frac{\partial}{\partial \theta_j} \left[\sum_{i=1}^N l_i(\theta) \right] \\
&= \sum_{i=1}^N \frac{\partial}{\partial \theta_j} l_i(\theta) = \sum_{i=1}^N (X_i^\intercal \theta - Y_i) x_{ij}.
\end{aligned}$$

Therefore

$$
\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) &= \begin{bmatrix} \sum_{i=1}^N (X_i^\intercal \theta - Y_i) x_{i1} \\ \vdots \\ \sum_{i=1}^N (X_i^\intercal \theta - Y_i) x_{ip} \end{bmatrix} \\
&= \begin{bmatrix} (X_1^\intercal \theta - Y_1) x_{11} + (X_2^\intercal \theta - Y_2) x_{21} + \cdots + (X_N^\intercal \theta - Y_N) x_{N1} \\ \vdots \\ (X_1^\intercal \theta - Y_1) x_{1p} + (X_2^\intercal \theta - Y_2) x_{2p} + \cdots + (X_N^\intercal \theta - Y_N) x_{Np} \end{bmatrix} \\
&= \sum_{i=1}^N X_i (X_i^\intercal \theta - Y_i) = \sum_{i=1}^N (X^\intercal)_{:,i} (X\theta - Y)_i \\
&= X^\intercal (X\theta - Y).
\end{aligned}
$$

Problem 2

Since $f(\theta) = \theta^2/2$, we can get $f'(\theta) = \theta$ and the iteration equation can be written as

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

Thus the nth iteration gives

$$\theta^n = (1-\alpha)\theta^{n-1} = \cdots = (1-\alpha)^n \theta^0.$$

Therefore if $\alpha > 2$ and $\theta^0 \neq 0$, θ^n diverges since it is a geometric sequence where its ratio is less than -1.

Problem 3

Since $f(\theta) = \dfrac{1}{2} \|X\theta - Y\|^2$, we can get $\nabla f(\theta) = X^\intercal (X\theta - Y)$ and the iteration equation can be written as

$$\theta^{k+1} = \theta^k - \alpha X^\intercal \big(X\theta^k - Y\big).$$

Assume $X^\intercal X$ is invertible and let $\theta^* = (X^\intercal X)^{-1} X^\intercal Y$, then

$$
\begin{aligned}
\theta^{k+1} - \theta^* &= \theta^k - \alpha X^\intercal \big(X\theta^k - Y\big) - (X^\intercal X)^{-1} X^\intercal Y \\
&= \theta^k - \alpha X^\intercal X \theta^k + \alpha X^\intercal Y - (X^\intercal X)^{-1} X^\intercal Y \\
&= (I - \alpha X^\intercal X)\theta^k + (\alpha X^\intercal X - I)(X^\intercal X)^{-1} X^\intercal Y = (I - \alpha X^\intercal X)\theta^k + (\alpha X^\intercal X - I)\theta^* \\
&= (I - \alpha X^\intercal X)\big(\theta^k - \theta^*\big)
\end{aligned}
$$

where $I \in \mathbb{R}^{p \times p}$ is an identity matrix. Thus the nth iteration gives

$$\theta^n - \theta^* = (I - \alpha X^\intercal X)\big(\theta^{n-1} - \theta^*\big) = \cdots = (I - \alpha X^\intercal X)^n \big(\theta^0 - \theta^*\big).$$

Since $X^\intercal X$ is a symmetric matrix and a positive matrix, it has p real positive eigenvalues and is diagonalizable by them. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ the eigenvalues and M the matrix with the eigenvectors as its columns, then

$$M^{-1} X^\intercal X M = \mathrm{diag}(\lambda_1, \cdots, \lambda_p), \quad X^\intercal X \sim \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$$

and gradually

$$\begin{aligned} I - \alpha X^\intercal X &= I - \alpha M \mathrm{diag}(\lambda_1, \cdots, \lambda_p) M^{-1} \\ &= M I M^{-1} - M(\alpha\, \mathrm{diag}(\lambda_1, \cdots, \lambda_p)) M^{-1} \\ &= M(I - \alpha\, \mathrm{diag}(\lambda_1, \cdots, \lambda_p)) M^{1} = M \mathrm{diag}(1 - \alpha\lambda_1, \cdots, 1 - \alpha\lambda_p) M^{-1} \end{aligned}$$

so $I - \alpha X^\intercal X \sim \mathrm{diag}(1 - \alpha\lambda_1, \cdots, 1 - \alpha\lambda_p)$. Therefore

$$\begin{aligned} (I - \alpha X^\intercal X)^n &= M \mathrm{diag}(1 - \alpha\lambda_1, \cdots, 1 - \alpha\lambda_p)^n M^{-1} \\ &= M \mathrm{diag}((1 - \alpha\lambda_1)^n, \cdots, (1 - \alpha\lambda_p)^n) M^{-1}. \end{aligned}$$

If $\alpha > 2/\rho(X^\intercal X)$, we can obtain $1 - \alpha\lambda_1 < -1$ since $\rho\big(X^T X\big) = \lambda_1 \geq 0$. Therefore $(1 - \alpha\lambda_1)^n$ diverges, and consequently $(I - \alpha X^\intercal X)^n$, $\theta^n - \theta^*$, and θ^n also diverge.