1.

(a)
$$\frac{\partial}{\partial \theta_j} l_i(\theta) = \frac{\partial}{\partial \theta_j}\left[\frac{1}{2}(X_i^T\theta - Y_i)^2\right]$$
$$= \frac{\partial}{\partial \theta_j}\left[\frac{1}{2}(X_{i1}\theta_1 + \cdots + X_{ip}\theta_p - Y_i)^2\right]$$
$$= \frac{1}{2} \cdot 2(X_{i1}\theta_1 + \cdots + X_{ip}\theta_p - Y_i) \cdot X_{ij}$$
$$= (X_i^T\theta - Y_i) X_{ij}$$

$$\nabla_\theta l_i(\theta) = \begin{bmatrix} \frac{\partial}{\partial\theta_1} l_i(\theta) \\ \vdots \\ \frac{\partial}{\partial\theta_p} l_i(\theta) \end{bmatrix} = \begin{bmatrix} (X_i^T\theta - Y_i)X_{i1} \\ \vdots \\ (X_i^T\theta - Y_i)X_{ip} \end{bmatrix} = (X_i^T\theta - Y_i) X_i$$

$\square$

(b)
$$\nabla_\theta L(\theta) = \nabla_\theta\left[\frac{1}{2}\|X\theta - Y\|^2\right]$$
$$= \nabla_\theta\left[\frac{1}{2}\sum_{i=1}^{N}(X_i\theta - Y_i)^2\right]$$
$$= \sum_{i=1}^{N} \nabla_\theta\, l_i(\theta)$$
$$= \sum_{i=1}^{N} (X_i^T\theta - Y_i) X_i$$
$$= \sum_{i=1}^{N} X_i (X_i^T\theta - Y_i)$$
$$= [X_1, \cdots, X_N] \begin{bmatrix} X_1^T\theta - Y_1 \\ \vdots \\ X_N^T\theta - Y_N \end{bmatrix}$$
$$= X^T(X\theta - Y)$$

$\square$

2.

$$f(\theta) = \frac{\theta^2}{2}$$

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

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

$$\therefore \theta^n = (1-\alpha)^n \theta^0 , \quad k = 0, 1, 2, \cdots$$

$$\alpha > 2 \Rightarrow |1-\alpha| > 1$$

$\therefore$ If $\theta^0 \neq 0$, $\alpha > 2$,

then θ^n diverges as $n \to \infty$. $\square$

3.

$$\nabla f(\theta) = X^T(X\theta - Y)$$

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

$$= \theta^k - \alpha X^T(X\theta^k - Y)$$

$$= (I - \alpha X^TX)\theta^k + \alpha X^TY$$

$$\theta^{k+1} - (X^TX)^{-1}X^TY = (I - \alpha X^TX)\theta^k + \alpha X^TY - (X^TX)^{-1}X^TY$$

$$= (I - \alpha X^TX)(\theta^k - (X^TX)^{-1}X^TY)$$

By letting $(X^TX)^{-1}X^TY = \theta^*$,

$$\theta^{k+1} - \theta^* = (I - \alpha X^TX)(\theta^k - \theta^*)$$

$$\theta^n - \theta^* = (I - \alpha X^TX)^n(\theta^0 - \theta^*)$$

X^TX is symmetric psd and so diagonalizable.

$X^TX = Q^T\Lambda Q$, $Q \in \mathbb{R}^{p\times p}$: orthogonal, $Q = \begin{bmatrix} q_1 \\ \vdots \\ q_p \end{bmatrix}$

$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$: eigenvalues of X^TX

$$(I - \alpha X^TX)^n = (Q^TIQ - \alpha Q^T\Lambda Q)^n$$

$$= [Q^T(I - \alpha\Lambda)Q]^n$$

$$= Q^T \begin{bmatrix} (1-\alpha\lambda_1)^n & & 0 \\ & \ddots & \\ 0 & & (1-\alpha\lambda_p)^n \end{bmatrix} Q = \sum_{i=1}^{p} q_i^T(1-\alpha\lambda_i)^n q_i$$

Since $\alpha > \frac{2}{\rho(X^TX)} = \frac{2}{\lambda_1} \Rightarrow 1 - \alpha\lambda_1 < -1$

$(1-\alpha\lambda_1)^n$ diverges as $n \to \infty$

$\therefore$ $\theta^n = (I - \alpha X^TX)^n(\theta^0 - \theta^*) + \theta^*$ diverges

if first element of $\theta^0 - \theta^*$ is non zero.

$\therefore$ $\langle \theta^n \rangle$ diverges for most starting points θ_0.

$\square$