Quiz03

1. Convergence of Gradient Descent for Ridge Regression

Let $A \in \mathbb{R}^{m \times n}$, $\vec{y} \in \mathbb{R}^m$, and $\lambda > 0$. Consider a slight variation of the ridge regression problem where the least squares loss is normalized by the number of data points:

$$\min_{\vec{x} \in \mathbb{R}^n} f_\lambda(\vec{x}) \qquad \text{where} \qquad f_\lambda(\vec{x}) \doteq \frac{1}{2}\left\{\frac{1}{m}\|A\vec{x} - \vec{y}\|_2^2 + \lambda \|\vec{x}\|_2^2\right\}. \tag{1}$$

The unique solution to the problem in Equation (1) is

$$\vec{x}_\lambda^\star = \left(A^\top A + \lambda m I\right)^{-1} A^\top \vec{y}. \tag{2}$$

(a) Show that the GD update

$$\vec{x}_{t+1} = \vec{x}_t - \eta\left(\frac{1}{m}A^\top(A\vec{x}_t - \vec{y}) + \lambda \vec{x}_t\right) \tag{3}$$

can be rearranged into the form

$$\vec{x}_{t+1} - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)(\vec{x}_t - \vec{x}_\lambda^\star). \tag{4}$$

Use this to show that

$$\vec{x}_t - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)^t (\vec{x}_0 - \vec{x}_\lambda^\star). \tag{5}$$

for every positive integer t.

Solution: We have

$$\begin{aligned}
\vec{x}_{t+1} &= \vec{x}_t - \eta\left(\frac{1}{m}A^\top(A\vec{x}_t - \vec{y}) + \lambda \vec{x}_t\right) && (6)\\
&= \vec{x}_t - \eta \cdot \frac{A^\top A}{m}\vec{x}_t + \eta \cdot \frac{1}{m}A^\top \vec{y} + \eta\lambda \vec{x}_t && (7)\\
&= \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)\vec{x}_t + \eta \cdot \frac{1}{m}A^\top \vec{y} && (8)\\
\implies \vec{x}_{t+1} - \vec{x}_\lambda^\star &= \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)\vec{x}_t + \eta \cdot \left(\frac{A^\top A}{m} + \lambda I\right)\vec{x}_\lambda^\star - \vec{x}_\lambda^\star && (9)\\
&= \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)(\vec{x}_t - \vec{x}_\lambda^\star). && (10)
\end{aligned}$$

Iterating this relation obtains the second equality.

(b) We now discuss the insight that the SVD can give us regarding the convergence of GD. Let $A = U\Sigma V^\top$ be a full SVD of A. Let $\vec{z}_t = V^\top \vec{x}_t$ and $\vec{z}_\lambda^\star = V^\top \vec{x}_\lambda^\star$. Show that

$$\vec{z}_t - \vec{z}_\lambda^\star = \left(I - \eta\left(\frac{\Sigma^\top \Sigma}{m} + \lambda I\right)\right)^t (\vec{z}_0 - \vec{z}_\lambda^\star), \tag{11}$$

and, moreover, show that for each $i \in \{1, \ldots, n\}$, we have

$$(\vec{z}_t)_i - (\vec{z}_\lambda^\star)_i = \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right)^t ((\vec{z}_0)_i - (\vec{z}_\lambda^\star)_i) \tag{12}$$

where $\sigma_i\{A\}$ is the i^{th} largest singular value of A.

Solution: If $A = U\Sigma V^\top$ then

$$A^\top A = V\Sigma^\top U^\top U \Sigma V^\top = V\Sigma^\top \Sigma V^\top. \tag{13}$$

Thus we have

$$\vec{x}_t - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)^t (\vec{x}_0 - \vec{x}_\lambda^\star) \tag{14}$$

$$= \left(I - \eta\left(\frac{V\Sigma^\top \Sigma V^\top}{m} + \lambda I\right)\right)^t (\vec{x}_0 - \vec{x}_\lambda^\star) \tag{15}$$

$$= \left(I - \eta\left(V\left(\frac{\Sigma^\top \Sigma}{m} + \lambda I\right)V^\top\right)\right)^t (\vec{x}_0 - \vec{x}_\lambda^\star) \tag{16}$$

$$= \left(I - \eta V\left(\frac{\Sigma^\top \Sigma}{m} + \lambda I\right)V^\top\right)^t (\vec{x}_0 - \vec{x}_\lambda^\star) \tag{17}$$

$$= \left(V\left(I - \eta\left(\frac{\Sigma^\top \Sigma}{m} + \lambda I\right)\right)V^\top\right)^t (\vec{x}_0 - \vec{x}_\lambda^\star) \tag{18}$$

$$= V\left(I - \eta\left(\frac{\Sigma^\top \Sigma}{m} + \lambda I\right)\right)^t V^\top(\vec{x}_0 - \vec{x}_\lambda^\star) \tag{19}$$

$$\implies V^\top(\vec{x}_t - \vec{x}_\lambda^\star) = \left(I - \eta\left(\frac{\Sigma^\top \Sigma}{m} + \lambda I\right)\right)^t V^\top(\vec{x}_0 - \vec{x}_\lambda^\star) \tag{20}$$

$$\implies \vec{z}_t - \vec{z}_\lambda^\star = \left(I - \eta\left(\frac{\Sigma^\top \Sigma}{m} + \lambda I\right)\right)^t (\vec{z}_0 - \vec{z}_\lambda^\star). \tag{21}$$

Now note that the quantity $I - \eta(\frac{\Sigma^\top \Sigma}{m} + \lambda I)$ is a diagonal matrix. Thus we have

$$\vec{z}_t - \vec{z}_\lambda^\star = \left(I - \eta\left(\frac{\Sigma^\top \Sigma}{m} + \lambda I\right)\right)^t (\vec{z}_0 - \vec{z}_\lambda^\star) \tag{22}$$

$$\implies (\vec{z}_t - \vec{z}_\lambda^\star)_i = \left(I - \eta\left(\frac{\Sigma^\top \Sigma}{m} + \lambda I\right)\right)_i^t (\vec{z}_0 - \vec{z}_\lambda^\star)_i \tag{23}$$

$$\implies (\vec{z}_t)_i - (\vec{z}_\lambda^\star)_i = \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right)^t ((\vec{z}_0)_i - (\vec{z}_\lambda^\star)_i) \tag{24}$$

(c) Show that $\lim_{t\to\infty} \vec{z}_t = \vec{z}_\lambda^\star$ for all initializations $\vec{x}_0 = V\vec{z}_0$ if and only if

$$\max_{i\in\{1,\dots,n\}} \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right| < 1. \tag{25}$$

Use this to show that GD converges for all initializations $\vec{x}_0$ if and only if

$$\eta \in \left(0, \frac{2m}{\sigma_{\max}\{A\}^2 + \lambda m}\right) \tag{26}$$

where $\sigma_{\max}\{A\} = \sigma_1\{A\}$ is the largest singular value of A.

Solution: We have

$$\lim_{t\to\infty} \vec{z}_t = \vec{z}_\lambda^\star, \qquad \forall \vec{x}_0 \tag{27}$$

$$\iff \lim_{t\to\infty} (\vec{z}_t)_i = (\vec{z}_\lambda^\star)_i, \qquad \forall i \quad \forall \vec{x}_0 \tag{28}$$

$$\iff \lim_{t\to\infty} \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right)^t = 0, \qquad \forall i \tag{29}$$

$$\iff \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right| < 1, \qquad \forall i \tag{30}$$

$$\iff \max_{i\in\{1,\dots,n\}} \left|1-\eta\left(\frac{\sigma_i\{A\}^2}{m}+\lambda\right)\right| < 1. \tag{31}$$

This proves the first part of the question. The second part of the question follows by noting that

$$\max_{i\in\{1,\dots,n\}} \left|1-\eta\left(\frac{\sigma_i\{A\}^2}{m}+\lambda\right)\right| < 1 \tag{32}$$

$$\iff \max_{i\in\{1,\dots,n\}} \left(1-\eta\left(\frac{\sigma_i\{A\}^2}{m}+\lambda\right)\right) < 1 \tag{33}$$

$$\text{and} \quad \min_{i\in\{1,\dots,n\}} \left(1-\eta\left(\frac{\sigma_i\{A\}^2}{m}+\lambda\right)\right) > -1 \tag{34}$$

$$\iff 1-\eta\left(\frac{\sigma_{\min}\{A\}^2}{m}+\lambda\right) < 1 \tag{35}$$

$$\text{and} \quad 1-\eta\left(\frac{\sigma_{\max}\{A\}^2}{m}+\lambda\right) > -1. \tag{36}$$

Now the first equation is always satisfied for $\eta > 0$ and $\lambda > 0$ because $\frac{\sigma_{\min}\{A\}^2}{m} + \lambda > 0$ so $1-\eta(\frac{\sigma_{\min}\{A\}^2}{m}+\lambda) < 1$. The second equation is satisfied when $\eta < \frac{2m}{\sigma_{\max}\{A\}^2+\lambda m}$. Since $\lim_{t\to\infty} \vec{x}_t = \vec{x}^\star_\lambda$ if and only if $\lim_{t\to\infty} \vec{z}_t = \vec{z}^\star_\lambda$, we have that gradient descent converges for all initializations $\vec{x}_0$ if and only if $0 < \eta < \frac{2m}{\sigma_{\max}\{A\}^2+\lambda m}$.