Discussion 7

1. Gradient Descent with A Wide Matrix (Fall 2022 Midterm)

Consider a matrix $X \in \mathbb{R}^{n \times d}$ with $n < d$ and a vector $\vec{y} \in \mathbb{R}^n$, both of which are known and given to you. Suppose X has full row rank.

(a) Consider the following problem:

$$X\vec{w} = \vec{y} \tag{1}$$

where $\vec{w} \in \mathbb{R}^d$ is unknown. How many solutions does 1 have? *Justify your answer.*

Solution: Since $\vec{y}$ is in the range of X, this implies that there exists $\vec{w}_0$ such that $\vec{y} = X\vec{w}_0$. Now let $\vec{s}$ be any non-zero vector in the null space of X (which exists since $\dim(\mathcal{N}(X)) = d - n > 0$), and consider an arbitrary vector $\vec{w}_{\text{new}} = \vec{w}_0 + t\vec{s}$, where $t \in \mathbb{R}$. Since $X\vec{w}_{\text{new}} = X\vec{w}_0 = \vec{y}$, we conclude that there are infinitely many solutions.

(b) Consider the minimum-norm problem

$$\vec{w}_\star = \underset{\substack{\vec{w} \in \mathbb{R}^d \\ X\vec{w} = \vec{y}}}{\text{argmin}} \|\vec{w}\|_2^2 . \tag{2}$$

We know that the optimal solution to this problem is $\vec{w}_\star = X^\top (XX^\top)^{-1}\vec{y}$. Now let $X = U\Sigma V^\top = U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} V^\top$ be the SVD of X, where $\Sigma_1 \in \mathbb{R}^{n \times n}$. Recall that this is possible because $n < d$ and X is full row rank. Prove that $\vec{w}_\star$ is given by

$$\vec{w}_\star = V \begin{bmatrix} \Sigma_1^{-1} \\ 0 \end{bmatrix} U^\top \vec{y}. \tag{3}$$

Solution: By plugging in the SVD of X in the expression of $\vec{w}_\star$, we have

$$\begin{aligned}
\vec{w}_\star &= X^\top (XX^\top)^{-1}\vec{y} && (4)\\
&= V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \left(U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} V^\top V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \right)^{-1} \vec{y}, && \text{(plugged in the SVD of } X)\\
&= V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \left(U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \right)^{-1} \vec{y}, && \text{(by } V^\top V = I)\\
&= V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top U \left(\begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \right)^{-1} U^\top \vec{y}, && \text{(by } U^{-1} = U^\top)\\
&= V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \left(\begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \right)^{-1} U^\top \vec{y}, && \text{(by } U^\top U = I)\\
&= V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \left(\Sigma_1^2\right)^{-1} U^\top \vec{y}, && \text{(took the matrix product of } \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix})\\
&= V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \Sigma_1^{-2} U^\top \vec{y}, && (\Sigma_1 \text{ is a square matrix and invertible})\\
&= V \begin{bmatrix} \Sigma_1^{-1} \\ 0 \end{bmatrix} U^\top \vec{y}. && (5)
\end{aligned}$$

(c) Let $\eta > 0$, and I be the identity matrix of appropriate dimension. Using the SVD $X = U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} V^\top$, prove the following identity for all positive integers $i > 0$:

$$(I - \eta X^\top X)^i = V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i V^\top. \tag{6}$$

Solution: We have

$$\begin{aligned}
(I - \eta X^\top X)^i &= \left(I - \eta (U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} V^\top)^\top (U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} V^\top) \right)^i, && \text{(plugged in the SVD of } X) \\
&= \left(I - \eta V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} V^\top \right)^i, && \text{(took the transpose of } U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} V^\top) \\
&= \left(I - \eta V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} V^\top \right)^i, && \text{(by } U^\top U = I) \\
&= \left(I - \eta V \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} V^\top \right)^i, && \text{(took the matrix product of } \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix}) \\
&= \left(V V^\top - \eta V \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} V^\top \right)^i, && \text{(by } I = VV^\top) \\
&= \left(V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right) V^\top \right)^i, && \text{(combine the diagonal matrices)} \\
&= V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i V^\top, && \text{(by applying } V^\top V = I \text{ repeatedly)}
\end{aligned}$$

(d) Recall that $X \in \mathbb{R}^{n \times d}$, and that we can write the SVD of X as $X = U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} V^\top$. We will use gradient descent to solve the minimization problem

$$\min_{\vec{w} \in \mathbb{R}^d} \frac{1}{2} \|X\vec{w} - \vec{y}\|_2^2 \tag{7}$$

with step-size $\eta > 0$. Let $\vec{w}_0 = \vec{0}$ be the initial state, and $\vec{w}_k$ be the k^{th} iterate of gradient descent. Use the identity:

$$(I - \eta X^\top X)^i = V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i V^\top. \tag{8}$$

to prove that after k steps, we have

$$\vec{w}_k = \eta \sum_{i=0}^{k-1} V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}. \tag{9}$$

HINT: Remember to set $\vec{w}_0 = \vec{0}$.

Solution: With $\nabla_{\vec{w}} f(\vec{w}) = X^\top (X\vec{w} - y)$, the gradient updates are of the form:

$$\begin{aligned}
\vec{w}_{k+1} &= \vec{w}_k - \eta \nabla_{\vec{w}} f(\vec{w}_k) && (10) \\
&= (I - \eta X^\top X)\vec{w}_k + \eta X^\top \vec{y} && (11)
\end{aligned}$$

$$\implies \vec{w}_k = (I - \eta X^\top X)^k \vec{w}_0 + \eta \sum_{i=0}^{k-1} (I - \eta X^\top X)^i X^\top \vec{y} \quad (12)$$

$$= \eta \sum_{i=0}^{k-1} (I - \eta X^\top X)^i X^\top \vec{y}. \quad (13)$$

Using the identity given, we have

$$\vec{w}_k = \eta \sum_{i=0}^{k-1} (I - \eta X^\top X)^i X^\top \vec{y} \quad (14)$$

$$= \eta \sum_{i=0}^{k-1} V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i V^\top \left(V \Sigma^\top U^\top \right) \vec{y} \quad (15)$$

$$= \eta \sum_{i=0}^{k-1} V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i \Sigma^\top U^\top \vec{y} \quad (16)$$

$$= \eta \sum_{i=0}^{k-1} V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}. \quad (17)$$

(e) Now let $0 < \eta < \frac{1}{\sigma_1^2}$, where σ_1 denotes the maximum singular value of $X = U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} V^\top$. Let $\vec{w}_k$ be given as

$$\vec{w}_k = \eta \sum_{i=0}^{k-1} V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}. \quad (18)$$

and let $\vec{w}_\star$ be the minimum norm solution given as

$$\vec{w}_\star = V \begin{bmatrix} \Sigma_1^{-1} \\ 0 \end{bmatrix} U^\top \vec{y}. \quad (19)$$

Prove that $\lim_{k \to \infty} \vec{w}_k = \vec{w}_\star$.

HINT: You may use the following result without proof. When all eigenvalues of $A \in \mathbb{R}^{n \times n}$ have magnitude < 1, we have the identity $(I - A)^{-1} = I + A + A^2 + \dots$.

Solution: We start with 9 and simplify, obtaining

$$\vec{w}_k = \eta \sum_{i=0}^{k-1} V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}$$

$$= \eta \sum_{i=0}^{k-1} V \begin{bmatrix} I - \eta \Sigma_1^2 & 0 \\ 0 & I \end{bmatrix}^i \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}$$

$$= \eta \sum_{i=0}^{k-1} V \begin{bmatrix} (I - \eta \Sigma_1^2)^i & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}$$

$$= \eta \sum_{i=0}^{k-1} V \begin{bmatrix} (I - \eta \Sigma_1^2)^i \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}$$

$$= \eta V \left\{ \sum_{i=0}^{k-1} \begin{bmatrix} (I - \eta \Sigma_1^2)^i \Sigma_1 \\ 0 \end{bmatrix} \right\} U^\top \vec{y}$$

$$= \eta V \begin{bmatrix} \sum_{i=0}^{k-1} (I - \eta \Sigma_1^2)^i \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}.$$

Taking limits, we have

$$\begin{aligned}
\lim_{k\to\infty} \vec{w}_k &= \eta V \begin{bmatrix} \sum_{i=0}^{\infty}(I-\eta\Sigma_1^2)^i\Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y} \\
&= \eta V \begin{bmatrix} (I-(I-\eta\Sigma_1^2))^{-1}\Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}, && \text{(applied the identity in the hint on } I-\eta\Sigma_1^2\text{)} \\
&= \eta V \begin{bmatrix} (\eta\Sigma_1^2)^{-1}\Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}, && (\Sigma_1^2 \text{ is a square matrix and invertible}) \\
&= \eta V \begin{bmatrix} \frac{1}{\eta}\Sigma_1^{-2}\Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y} \\
&= V \begin{bmatrix} \Sigma_1^{-1} \\ 0 \end{bmatrix} U^\top \vec{y}
\end{aligned}$$

as desired. Here the infinite sum is evaluated as in the hint because the eigenvalues of $I-\eta\Sigma_1^2$ are all in the interval $(0,1) \subseteq (-1,1)$. Indeed, the eigenvalues of $I-\eta\Sigma_1^2$ are $1-\eta\sigma_i^2$, where σ_i are the entries of Σ_1 and thus the nonzero singular values of X. Since $\sigma_i > 0$, we know $1-\eta\sigma_i^2 < 1$. Now, since $\eta < \frac{1}{\sigma_1^2}$, we have $1-\eta\sigma_i^2 > 1-\frac{\sigma_i^2}{\sigma_1^2} \geq 0$. Thus the eigenvalues of $I-\eta\Sigma_1^2$ are contained in $(-1,1)$ and the hint applies.

A common error, is to apply the hint directly on $\left(I-\eta\begin{bmatrix}\Sigma_1^2 & 0 \\ 0 & 0\end{bmatrix}\right)$. Note that the eigenvalues of

$$I-\eta\begin{bmatrix}\Sigma_1^2 & 0 \\ 0 & 0\end{bmatrix} = \begin{bmatrix} I-\eta\Sigma_1^2 & 0 \\ 0 & I\end{bmatrix}$$

are in the interval $(0,1]$, which breaks the condition we made on the A matrix described in the hint, all eigenvalues of A having magnitude strictly < 1.

2. Convexity and Composition of Functions

Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$. Define the composition of f with g as $h = f \circ g : \mathbb{R}^n \to \mathbb{R}$ such that $h(\vec{x}) = f(g(\vec{x}))$.

(a) Show that if f is convex and non decreasing and g is convex, then h is convex.

Solution:

$$\begin{aligned}
h(\lambda\vec{x}+(1-\lambda)\vec{y}) &= f(g(\lambda\vec{x}+(1-\lambda)\vec{y})) && (20)\\
&\leq f(\lambda g(\vec{x})+(1-\lambda)(g(\vec{y}))) && (g \text{ convex and } f \text{ nondecreasing}) \quad (21)\\
&\leq \lambda f(g(\vec{x}))+(1-\lambda)f(g(\vec{y})) && (f \text{ convex}) \quad (22)\\
&= \lambda h(\vec{x})+(1-\lambda)h(\vec{y}) && (23)
\end{aligned}$$

So h is convex.

(b) Show that there exists f non decreasing and g convex, such that $h = f \circ g$ is not convex.

Solution: Take $n = 1$, $f(x) = \log(x)$ and $g(x) = x$. Then $h(x) = \log(x)$ is not convex.

(c) Show that there exists f convex and g convex such that $h = f \circ g$ is not convex.

Solution: Take $n = 1$, $f(x) = -x$ and $g(x) = x^2$, then $h(x) = -x^2$ is not convex.