EECS 127/227AT Optimization Models in Engineering UC Berkeley Fall 2022

Midterm

1. **Honor Code (0 pts)**

   Please copy the following statement in the space provided below and sign your name.

   As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I will follow the rules and do this exam on my own.

   If you do not copy the honor code and sign your name, you will get a 0 on the exam.

Solution:

2. **Favorites. Any answer, as long as you write it down, will be given full credit. (2 pts)**

   (a) (1 pts) What's your favorite building in Berkeley?

   Solution: Any answer is fine.

   (b) (1 pts) What is a hobby or activity that makes you happy?

   Solution: Any answer is fine.

3. Orthonormal Matrices (6 pts)

Prove the following identities.

(a) (3 pts) Let $\vec{x} \in \mathbb{R}^n$ be a vector, and let $U \in \mathbb{R}^{m \times n}$, where $m \geq n$, be an orthonormal matrix. **Prove the following equality:**

$$\|U\vec{x}\|_2 = \|\vec{x}\|_2 . \tag{1}$$

Solution: We have

$$\|U\vec{x}\|_2^2 = (U\vec{x})^\top (U\vec{x}) = \vec{x}^\top U^\top U\vec{x} = \vec{x}^\top \vec{x} = \|\vec{x}\|_2^2 . \tag{2}$$

Here $U^\top U = I$ as a consequence of U having orthonormal columns. (We've seen this property in many places, such as Discussion 3 Problem 1 (a).)

(b) (3 pts) Let $A \in \mathbb{R}^{n \times n}$ be a square matrix, and suppose $A = QR$ is a QR decomposition of A. **Compute $\|R\|_F$ in terms of $\|A\|_F$.**

Solution: We have

$$\|A\|_F = \|QR\|_F = \|R\|_F \tag{3}$$

since Q is an orthonormal matrix (as seen in lecture, Discussion 1 Problem 2, and Homework 2 Problem 3), and the Frobenius norm is invariant under multiplication by a square orthonormal matrix (as seen in many places such as Homework 3 Problem 3 part (b)).

4. Vector Calculus (7 pts)

Let $\vec{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$. Let $f(\vec{x}) \doteq \|A\vec{x}\|_2^2$.

(a) (4 pts) **Calculate the gradient of $f(\vec{x})$ with respect to $\vec{x}$.** *Show your work.*

Solution: Denote the i^{th} row of A by $\vec{a}_i^\top$. The i^{th} element of $A\vec{x}$ is $\vec{a}_i^\top \vec{x}$, and $f(\vec{x}) = \sum_{i=1}^n (\vec{a}_i^\top \vec{x})^2$. Then

$$\nabla f(\vec{x}) = \sum_{i=1}^n 2(\vec{a}_i^\top \vec{x}) \nabla_{\vec{x}} (\vec{a}_i^\top \vec{x}). \tag{4}$$

For a vector $\vec{c} \in \mathbb{R}^n$, we know that $\nabla_{\vec{x}}(\vec{c}^\top \vec{x}) = \nabla_{\vec{x}}(\sum_{i=1}^n c_i x_i) = \vec{c}$. This means that $\nabla_{\vec{x}}(\vec{a}_i^\top \vec{x}) = \vec{a}_i$. We have

$$\nabla f(\vec{x}) = \sum_{i=1}^n 2(\vec{a}_i^\top \vec{x}) \vec{a}_i. \tag{5}$$

Note that $\vec{a}_i^\top \vec{x}$ are the elements of the vector $\vec{v} = A\vec{x}$, and $\vec{a}_i$ are the columns of $A^\top$. Then $\sum_{i=1}^n (\vec{a}_i^\top \vec{x})\vec{a}_i$ is $A^\top \vec{v} = \sum_{i=1}^n v_i \vec{a}_i = \sum_{i=1}^n (\vec{a}_i^\top \vec{x})\vec{a}_i$, which gives us

$$\nabla f(\vec{x}) = 2A^\top A\vec{x}. \tag{6}$$

For reference, look at Discussion 4 (Q1) and HW 4, Q3 subpart a) (ii).

Alternatively, you can also note that $f(\vec{x}) = x^T A^T A x$, and therefore $\nabla f(\vec{x}) = 2A^\top A\vec{x}$, as done in lecture.

(b) (3 pts) **Calculate the Hessian of $f(\vec{x}) \doteq \|A\vec{x}\|_2^2$ with respect to $\vec{x}$.** *You do not need to show your work.*

Solution: $f(\vec{x}) = x^T A^T A x$. We considered the Hessian of $x^T Q x$ in lecture, for general matrix Q. The Hessian is $\nabla^2 f(\vec{x}) = \nabla 2A^\top A\vec{x} = 2A^\top A$.

For reference, look at Discussion 4 (Q1) and HW 4, Q3 subpart a) (ii).

5. Convexity (9 pts)

(a) (4 pts) Let $f_1, \dots, f_k \colon \mathbb{R}^n \to \mathbb{R}$ be convex functions. **Prove that**

$$\textbf{the set} \qquad S \doteq \{\vec{x} \in \mathbb{R}^n \mid f_i(\vec{x}) \le 0 \quad \forall i = 1, \dots, k\} \qquad \textbf{is convex.} \tag{7}$$

Solution: Fix $\vec{x}, \vec{y} \in S$ and $\lambda \in [0,1]$. We aim to show that $\lambda\vec{x} + (1-\lambda)\vec{y} \in S$.

For $i \in \{1, \dots, k\}$, since f_i is convex we use the definition of convexity (seen in many places such as Homework 6) for f_i to get

$$f_i(\lambda\vec{x} + (1-\lambda)\vec{y}) \le \lambda \underbrace{f_i(\vec{x})}_{\le 0} + (1-\lambda) \underbrace{f_i(\vec{y})}_{\le 0} \tag{8}$$

$$\le 0. \tag{9}$$

Thus $\lambda\vec{x} + (1-\lambda)\vec{y} \in S$. Since this is true for arbitrary $\vec{x}$, $\vec{y}$, and λ, it holds that S is convex.

(b) (5 pts) Let $\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$. **Prove that** $f \colon \mathbb{R}^n \to \mathbb{R}$ **given by** $f(\vec{x}) \doteq \max_{1 \le i \le n} |x_i|$ **is convex.**

Solution: There are many valid solutions, two of which we have listed below

i. **Pointwise Maximum**

The domain of f, $\operatorname{dom}(f) = \mathbb{R}^n$ is convex since as shown in discussion 5 problem 1b, any vector subspace is convex. We can write the absolute value as a max of two affine functions,

$$\begin{aligned} f(x) &= \max_{1 \le i \le n} |x_i| \\ &= \max_{1 \le i \le n} \max\{x_i, -x_i\} \\ &= \max_{1 \le i \le n, j \in \{0,1\}} (-1)^j x_i \end{aligned}$$

Affine functions are convex since the satisfy Jensen's inequality tightly, and since the pointwise maximum of convex functions is convex (as seen in discussion 5 problem 2d), $f(x)$ is convex as well.

ii. **Jensen's Inequality**

The domain of f, $\operatorname{dom}(f) = \mathbb{R}^n$ is convex since as shown in discussion 5 problem 1b, any vector subspace is convex. For any $\vec{x}, \vec{y} \in \mathbb{R}^n$ and all $\theta \in [0,1]$,

$$\begin{aligned} f(\theta\vec{x} + (1-\theta)\vec{y}) &= \max_{1 \le i \le n} |\theta x_i + (1-\theta) y_i| \\ &\le \max_{1 \le i \le n} |\theta x_i| + |(1-\theta) y_i| \\ &\le |\theta| \max_{1 \le i \le n} |x_i| + |1-\theta| \max_{1 \le i \le n} |y_i| \\ &= \theta f(\vec{x}) + (1-\theta) f(\vec{y}) \end{aligned}$$

The first inequality is due to the triangle inequality, and the second inequality is true since if $j = \operatorname{argmax}_{1 \le i \le n} |\theta x_i + (1-\theta) y_i|$, then

$$f(\theta\vec{x} + (1-\theta)\vec{y}) = |\theta\,|x_j| + (1-\theta)\,|y_j||$$

$$
\begin{aligned}
&\leq \theta \left|x_j\right| + (1-\theta) \left|y_j\right| \\
&\leq \theta f(\vec{x}) + (1-\theta) f(\vec{y})
\end{aligned}
$$

6. Low-Rank Approximation (4 pts)

Let $A \in \mathbb{R}^{4\times 3}$ be a matrix whose full SVD is

$$A = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}}_{V^\top}. \tag{10}$$

Give the best rank-2 approximation to A, i.e., the solution to the problem

$$\operatorname*{argmin}_{\substack{B \in \mathbb{R}^{4\times 3} \\ \mathrm{rk}(B) \leq 2}} \|A - B\|_F^2 . \tag{11}$$

No justification is necessary.

NOTE: Please leave your answer in terms of a matrix product.

Solution: By the Eckart-Young theorem, we know that a best rank-2 approximation to A is given by the rank-2 truncated SVD $U_2\Sigma_2V_2^\top$, where $U_2 \in \mathbb{R}^{4\times 2}$ is the first two columns of U, $\Sigma_2 \in \mathbb{R}^{2\times 2}$ is the top left 2×2 sub-block of Σ, and $V_2 \in \mathbb{R}^{3\times 2}$ is the first two columns of V (we've seen this result several times such as in lecture, the note, and Homework 4 Problem 1). Altogether we have

$$A_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{12}$$

as a solution to the problem. An alternate "full SVD"-esque solution is

$$A_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix} \tag{13}$$

which multiplies to the same thing (i.e., our two definitions of A_2 are equivalent, because the second definition is the compact SVD form of the first definition). (We've seen this equivalence in many places, such as Discussion 3 Problem 1.)

7. Power Iteration and SVD (22 pts)

In this problem, we will discuss how to efficiently compute singular values and vectors using an algorithm called "power iteration" that provides eigenvalues and eigenvectors. You do not need any prior knowledge about the algorithm to complete this problem, other than the description below.

The "power iteration" algorithm, denoted by $\textsc{PowIter}$, operates as follows:

- For a symmetric positive semidefinite matrix $B \in \mathbb{S}^n_+$, $\textsc{PowIter}(B) = (\lambda, \vec{v})$, where λ is the largest eigenvalue of B and $\vec{v}$ is a corresponding unit eigenvector.
- For a non-square, non-symmetric, or non-positive semidefinite matrix C, $\textsc{PowIter}(C) = \textsc{Error}$.

(a) (4 pts) Let $A \in \mathbb{R}^{m\times n}$ be known to you. **Explain how to use $\textsc{PowIter}$ to compute a top right singular vector, i.e., the first column $\vec{v}_1$ of V in an SVD of $A = U\Sigma V^\top$, as well as its corresponding singular value σ_1.** *A 1-2 sentence algorithm description or pseudocode will suffice.*

Solution: Let $r \doteq \operatorname{rk}(A)$ and let A have outer product SVD $A = \sum_{i=1}^r \sigma_i \vec{u}_i \vec{v}_i^\top$ (we've seen this from lecture as well as Discussion 3 Problem 1). We know that

$$A^\top A = \left(\sum_{i=1}^r \sigma_i \vec{u}_i \vec{v}_i^\top\right)^\top \left(\sum_{i=1}^r \sigma_i \vec{u}_i \vec{v}_i^\top\right) \tag{14}$$

$$= \sum_{i=1}^r \sum_{j=1}^r \sigma_i \sigma_j \vec{v}_i \vec{u}_i^\top \vec{u}_j \vec{v}_j^\top \tag{15}$$

$$= \sum_{i=1}^r \sum_{j=1}^r \sigma_i^2 \vec{v}_i \vec{v}_i^\top \tag{16}$$

since the $\vec{u}_i$ are orthonormal. This matrix is PSD (as seen in many places such as Discussion 2 Problem 1 or Homework 2 Problem 5), and its eigenvectors $\vec{v}_i$ are the right singular vectors of A by the above derivation, so

$$\textsc{PowIter}(A^\top A) = (\sigma_1^2, \vec{v}_1) \tag{17}$$

where σ_1^2 is the squared top singular value of A. Taking the square root then gets the top singular value of A.

(b) (6 pts) Let $B \in \mathbb{S}^n_+$ be a symmetric positive semidefinite matrix with eigenpairs $(\lambda_1, \vec{w}_1), \ldots, (\lambda_n, \vec{w}_n)$ where $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$. **Prove that the matrix $D \doteq B - \lambda_1 \vec{w}_1 \vec{w}_1^\top$ is a symmetric positive semidefinite matrix with eigenpairs $(0, \vec{w}_1), (\lambda_2, \vec{w}_2), \ldots, (\lambda_n, \vec{w}_n)$.**

Solution: Since B has eigendecomposition

$$B = \sum_{i=1}^n \lambda_i \vec{w}_i \vec{w}_i^\top, \tag{18}$$

(which we discussed in some places such as Homework 2 Problem 4), then

$$D = B - \lambda_1 \vec{w}_1 \vec{w}_1^\top = 0 \cdot \vec{w}_1 \vec{w}_1^\top + \sum_{i=2}^n \lambda_i \vec{w}_i \vec{w}_i^\top = \sum_{i=2}^n \lambda_i \vec{w}_i \vec{w}_i^\top. \tag{19}$$

Thus D has the desired eigenpairs. Moreover, D is a symmetric matrix, since

$$D^\top = (B - \lambda_1 \vec{w}_1 \vec{w}_1)^\top = B^\top - \lambda_1 \vec{w}_1 \vec{w}_1^\top = B - \lambda_1 \vec{w}_1 \vec{w}_1^\top = D, \tag{20}$$

and D is a PSD matrix since its eigenvalues are non-negative (for more details see Homework 2 Problem 5).

(c) (5 pts) Let $A \in \mathbb{R}^{m\times n}$ and $k \leq \min\{m, n\}$ be known to you. **Explain how to use** POWITER **to compute the top k right singular vectors of A, i.e., the first k columns of V in an SVD $A = U\Sigma V^\top$, as well as their corresponding singular values.** *A 1-2 sentence algorithm description or pseudocode will suffice.*

NOTE: You may use the result from part (b), even if you haven't proved it.

Solution: Start with

$$\textsc{PowIter}(A^\top A) = (\sigma_1^2, \vec{v}_1), \tag{21}$$

from which we can get the top singular value σ_1 and top singular vector $\vec{v}_1$. We note that $B_2 = A^\top A - \sigma_1^2 \vec{v}_1\vec{v}_1^\top$ has eigenpairs $(0, \vec{v}_1), (\sigma_2^2, \vec{v}_2), \ldots$ as shown in the previous problem part. The top eigenvalue of B_2 is σ_2^2, and we can perform

$$\textsc{PowIter}(B_2) = \textsc{PowIter}(A^\top A - \sigma_1^2\vec{v}_1\vec{v}_1^\top) = (\sigma_2^2, \vec{v}_2) \tag{22}$$

to get second highest singular value σ_2 and second singular vector $\vec{v}_2$. We can continue doing this operation, and get the k^{th} singular value and vector by performing

$$\textsc{PowIter}(B_k) = \textsc{PowIter}\left(A^\top A - \sum_{i=1}^{k-1}\sigma_i^2\vec{v}_i\vec{v}_i^\top\right) = (\sigma_k^2, \vec{v}_k), \tag{23}$$

since B_k has eigenpairs $(0, \vec{v}_1), (0, \vec{v}_2), \ldots, (0, \vec{v}_{k-1}), (\sigma_k^2, \vec{v}_k), (\sigma_{k+1}^2, \vec{v}_{k+1}), \ldots$.

(d) (4 pts) Suppose that you know how to compute any number of right singular vectors of any matrix using POWITER (regardless of whether or not you completed part (c)). Let $A \in \mathbb{R}^{m\times n}$ and $r \doteq \text{rk}(A)$ be known to you. **Explain how to compute a basis for $\mathcal{R}(A^\top)$.** *A 1-2 sentence solution will suffice.*

Solution:

Solution 1:

We know that the first $r \doteq \text{rk}(A)$ right singular vectors are a basis for $\mathcal{R}(A^\top)$ (as per Discussion 3 Problem 1 and other sources such as lecture), so applying the method of (c) to A generates the appropriate r basis vectors.

Solution 2:

We can take the columns of $A^\top$, or in other words the rows of A, and run Gram-Schmidt on them. This gives an orthonormal basis for $\mathcal{R}(A^\top)$ and some zero vectors, which should be discarded.

(e) (3 pts) Let $A \in \mathbb{R}^{m\times n}$ be unknown to you (so you cannot compute its SVD or even use POWITER). Suppose that you are given a basis for $\mathcal{R}(A^\top)$. **Explain how to compute a basis for $\mathcal{N}(A)$.** *A one sentence solution will suffice.*

Solution: We use Gram-Schmidt to extend our basis for $\mathcal{R}(A^\top)$ to a basis for $\mathbb{R}^n$; the remaining vectors in that extended basis themselves form a basis for $\mathcal{N}(A)$, since $\mathcal{R}(A^\top) \oplus \mathcal{N}(A) = \mathbb{R}^n$ by the FTLA (which was discussed in lecture but can also be proved using properties of the SVD as in Discussion 3 Problem 1 and other sources).

8. Matrix Square Root (9 pts)

Let $A, B \in \mathbb{S}^n_{++}$ be symmetric positive definite matrices.

As B is symmetric, it has an orthonormal eigendecomposition $B = V\Lambda V^\top$. Since B is positive definite, we can define its matrix square root as follows $B^{1/2} = V\Lambda^{1/2}V^\top$, where $\Lambda^{1/2}$ is a diagonal matrix whose entries are the square roots of the corresponding entries of Λ. We denote the inverse of $B^{1/2}$ as $B^{-1/2}$. Finally, define $C \doteq B^{-1/2}AB^{-1/2}$.

Prove that the maximum eigenvalue of C is $\lambda^\star$, where

$$\lambda^\star \doteq \max_{\vec{x}\neq\vec{0}} \frac{\vec{x}^\top A\vec{x}}{\vec{x}^\top B\vec{x}}. \tag{24}$$

Solution: Define $y = B^{1/2}x$ and hence $B^{-1/2}y = x$ (such substitutions but for orthogonal vectors showed up in Homework 2 Q5a and Homework 3 Q4c). As B is positive definite, $x \neq 0 \iff y \neq 0$. So we can rewrite the optimization problem as

$$\begin{aligned}\max_{x\neq 0} \frac{x^\top Ax}{x^\top Bx} &= \max_{y\neq 0} \frac{(B^{-1/2}y)^\top A(B^{-1/2}y)}{(B^{-1/2}y)^\top B(B^{-1/2}y)} \\ &= \max_{y\neq 0} \frac{y^\top B^{-1/2}AB^{-1/2}y}{y^\top y}\end{aligned}$$

Which is the same as finding the maximum eigenvalue and eigenvector of $C = B^{-1/2}AB^{-1/2}$ due to the definition via Rayleigh quotient (see lecture 4).

9. Gradient Descent with A Wide Matrix (31 pts)

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) (3 pts) **Consider the following problem:**

$$X\vec{w} = \vec{y} \tag{25}$$

where $\vec{w} \in \mathbb{R}^d$ is unknown. How many solutions does Equation (25) 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. (definitions of range space and null space of a matrix are seen in many places such as Discussion 3 Problem 2b and 2c. Minimum norm solutions and that systems with wide full-row-rank matrices have infinitely many solutions was also covered in Lecture.)

(b) (5 pts) Consider the minimum-norm problem

$$\vec{w}_\star = \operatorname*{argmin}_{\substack{\vec{w}\in\mathbb{R}^d \\ X\vec{w}=\vec{y}}} \|\vec{w}\|_2^2 . \tag{26}$$

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**

$$\vec{w}_\star = V\begin{bmatrix}\Sigma_1^{-1} \\ 0\end{bmatrix}U^\top\vec{y}. \tag{27}$$

All steps must be shown and justified for full credit.

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

$$\begin{aligned}
\vec{w}_\star &= X^\top(XX^\top)^{-1}\vec{y} && (28)\\
&= V\begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}U^\top\left(U\begin{bmatrix}\Sigma_1 & 0\end{bmatrix}V^\top V\begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}U^\top\right)^{-1}\vec{y}, && \text{(plugged in the SVD of } X\text{)}\\
&= V\begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}U^\top\left(U\begin{bmatrix}\Sigma_1 & 0\end{bmatrix}\begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}U^\top\right)^{-1}\vec{y}, && \text{(by } V^\top V = I\text{)}\\
&= V\begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}U^\top U\left(\begin{bmatrix}\Sigma_1 & 0\end{bmatrix}\begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}\right)^{-1}U^\top\vec{y}, && \text{(by } U^{-1} = U^\top\text{)}\\
&= V\begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}\left(\begin{bmatrix}\Sigma_1 & 0\end{bmatrix}\begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}\right)^{-1}U^\top\vec{y}, && \text{(by } U^\top U = I\text{)}\\
&= V\begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}\left(\Sigma_1^2\right)^{-1}U^\top\vec{y}, && \text{(took the matrix product of } \begin{bmatrix}\Sigma_1 & 0\end{bmatrix}\begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}\text{)}\\
&= V\begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}\Sigma_1^{-2}U^\top\vec{y}, && (\Sigma_1 \text{ is a square matrix and invertible})
\end{aligned}$$

$$= V \begin{bmatrix} \Sigma_1^{-1} \\ 0 \end{bmatrix} U^\top \vec{y}. \tag{29}$$

The concepts of substituting a matrix in an expression with its SVD form, and using the indentities of the orthonormal matrices (the U and V matrices) to simplify terms are seen in many places such as Discussion 1 problem 1b, and Discussion 3 Problem 1a. However, we need to be extra careful when solving this problem since we are now dealing with a wide matrix with full row rank instead of a tall matrix with full column rank which we saw more often in previous exercises. In this problem, the diagonal matrix $\Sigma = \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix}$ is a wide matrix, which is not invertible. Also note that $\Sigma^\top = \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \neq \Sigma$.

(c) (5 pts) 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 $i > 0$:**

$$(I - \eta X^\top X)^i = V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i V^\top. \tag{30}$$

All steps must be shown and justified for full credit.

Solution: We have

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

The concepts of substituting a matrix in an expression with its SVD form, and using the indentities of the orthonormal matrices (the U and V matrices) to simplify terms are seen in many places such as Discussion 1 problem 1b, and Discussion 3 Problem 1a. The repeated application of similar identity as $V^\top V = I$ is also practiced in Discussion 0 Problem 3a. However, again, we need to be extra careful when solving this problem since we are now dealing with a wide matrix with full row rank instead of a tall matrix with full column rank which we saw more often in previous exercises. In this problem, the diagonal matrix $\Sigma = \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix}$ is a wide matrix, so $\Sigma^\top = \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \neq \Sigma$.

(d) (9 pts) 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

$$\min_{\vec{w}\in\mathbb{R}^d} \frac{1}{2}\|X\vec{w} - \vec{y}\|_2^2 \tag{31}$$

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:**

$$(I - \eta X^\top X)^i = V\left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix}\right)^i V^\top. \tag{32}$$

to prove that after k steps, we have

$$\vec{w}_k = \eta \sum_{i=0}^{k-1} V\left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix}\right)^i \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}. \tag{33}$$

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:

$$\begin{aligned}
\vec{w}_{k+1} &= \vec{w}_k - \eta\nabla_{\vec{w}} f(\vec{w}_k) && (34)\\
&= (I - \eta X^\top X)\vec{w}_k + \eta X^\top \vec{y} && (35)\\
\implies \vec{w}_k &= (I - \eta X^\top X)^k \vec{w}_0 + \eta \sum_{i=0}^{k-1} (I - \eta X^\top X)^i X^\top \vec{y} && (36)\\
&= \eta \sum_{i=0}^{k-1} (I - \eta X^\top X)^i X^\top \vec{y}. && (37)
\end{aligned}$$

Using the identity given, we have

$$\begin{aligned}
\vec{w}_k &= \eta \sum_{i=0}^{k-1} (I - \eta X^\top X)^i X^\top \vec{y} && (38)\\
&= \eta \sum_{i=0}^{k-1} V\left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix}\right)^i V^\top \left(V\Sigma^\top U^\top\right)\vec{y} && (39)\\
&= \eta \sum_{i=0}^{k-1} V\left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix}\right)^i \Sigma^\top U^\top \vec{y} && (40)\\
&= \eta \sum_{i=0}^{k-1} V\left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix}\right)^i \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}. && (41)
\end{aligned}$$

The concept of the problem is seen in many places such as Homework 6 Problem 5c, and Homework 7 Problem 4a.

(e) (9 pts) 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

$$\vec{w}_k = \eta \sum_{i=0}^{k-1} V\left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix}\right)^i \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}. \tag{42}$$

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

$$\vec{w}_\star = V \begin{bmatrix} \Sigma_1^{-1} \\ 0 \end{bmatrix} U^\top \vec{y}. \tag{43}$$

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 Equation (33) and simplify, obtaining

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

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. The concept of this problem is seen in many places such as Homework 6 problem 5d, and Homework 7 Problem 1b, 4b.

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.