Homework 4

Self grades are due at 11 PM on February 16, 2024.

1. Interpreting the Data Matrix

When working in many fields, you'll often find yourself working with a *data matrix* X. Notation can vary — sometimes it has dimensions $\mathbb{R}^{m\times n}$, while others it has dimensions $\mathbb{R}^{n\times d}$, for example — and interpreting its precise meaning can often be confusing. In this problem, we lead you through an example of data matrix interpretation and manipulation.

First, what exactly is a data matrix? As the name suggests, it is a collection of *data points*. Suppose you are collecting data about courses offered in the EECS department in Fall 2022. Each course has certain quantifiable attributes, or *features*, that you are interested in. Possible examples of features are the number of students in the course, the number of GSIs in the course, the number of units the course is worth, the size of the classroom that the course was taught in, the difficulty rating of the course on a numerical (1-5) scale, and so on. Suppose there were a total of 20 courses, and that for each course, we have 10 features. This gives us 20 data points, where each data point is a 10-dimensional vector. We can arrange these data points in a matrix of size 20×10.

Generalizing the above, suppose we have n data points, with each point containing values for d features. Our data matrix X would then be of size $n \times d$, i.e., $X \in \mathbb{R}^{n\times d}$. We can interpret X in the following two (equivalent) ways:

$$X = \begin{bmatrix} \leftarrow \vec{x}_1^\top \rightarrow \\ \leftarrow \vec{x}_2^\top \rightarrow \\ \vdots \\ \leftarrow \vec{x}_n^\top \rightarrow \end{bmatrix} = \begin{bmatrix} \uparrow & \uparrow & \cdots & \uparrow \\ \vec{f}_1 & \vec{f}_2 & \cdots & \vec{f}_d \\ \downarrow & \downarrow & \cdots & \downarrow \end{bmatrix}. \tag{1}$$

Here, $\vec{x}_i \in \mathbb{R}^d,\ i = 1, 2, \ldots, n$, and $\vec{x}_i^\top$ is a row vector that contains values of different features for the i-th data point. Also, $\vec{f}_j \in \mathbb{R}^n,\ j = 1, 2, \ldots, d$, and $\vec{f}_j$ is a column vector that contains values of the the j-th feature for different data points.

In the remainder of this problem, we explore how we can interpret and use X. For subproblems that require answers in Python, assume X is stored as a $n \times d$ NumPy array `X`.

(a) We first introduce the *empirical mean* of each feature. Let $k \geq 1$ be a positive integer, and define $\vec{1}$ to be the vector with 1 in every entry. The empirical mean of a vector $\vec{y} \in \mathbb{R}^k$ is defined as

$$\mu(\vec{y}) \doteq \frac{1}{k}\vec{1}^\top\vec{y} = \frac{1}{k}\sum_{i=1}^{k} y_i. \tag{2}$$

Suppose we want to compute a vector that contains the empirical mean of each feature, i.e., all the $\mu(\vec{f}_j)$'s. What is the length of the vector of empirical means? Which of the following Python commands will generate this vector?

i. `mu = numpy.mean(X, axis = 0)`

ii. `mu = numpy.mean(X, axis = 1)`

Solution: The vector containing the empirical mean of each feature has length d, the number of features. It can be calculated via:

i. `mu = numpy.mean(X, axis = 0)`

(b) The next quantity we will discuss is the *empirical variance*, and through it, the *empirical standard deviation*. The empirical variance of a vector $\vec{y} \in \mathbb{R}^k$ is defined as

$$\sigma^2(\vec{y}) \doteq \frac{1}{k} \left\| \vec{y} - \mu(\vec{y})\vec{1} \right\|_2^2 = \frac{1}{k} \sum_{i=1}^{k} (y_i - \mu(\vec{y}))^2. \tag{3}$$

As the choice of notation would have you expect, the empirical standard deviation is defined as

$$\sigma(\vec{y}) \doteq \sqrt{\sigma^2(\vec{y})}. \tag{4}$$

Suppose we want to compute a vector that contains the empirical standard deviation of each feature, i.e., all the $\sigma(\vec{f}_j)$'s. What is the length of this vector? Which of the following Python commands will generate this vector?

i. `sigma = numpy.std(X, axis = 0)`

ii. `sigma = numpy.std(X, axis = 1)`

Solution: As above, the vector has length d, and can be computed via

i. `sigma = numpy.std(X, axis = 0)`

(c) Suppose we want to modify `X` so that each feature vector is "*centered*", i.e., has zero empirical mean. How would you achieve this using Python code?

Solution: To achieve a "centered" data matrix, we need to subtract the mean of each feature vector from every value in the vector. We can accomplish this in Python code via:

- `X_centered = X - numpy.mean(X, axis = 0)`

(d) Suppose we want to modify `X` so that each feature vector is "*standardized*", i.e., has zero empirical mean with empirical variance equal to 1. How would you achieve this using Python code?

NOTE: This standardization technique is a very common data pre-processing step.

Solution: To achieve a "standardized" data matrix, we need to subtract the mean of each feature vector from every value in the vector, then divide these new values by their corresponding standard deviation. We can accomplish this in Python code via:

- `X_centered = X - numpy.mean(X, axis = 0)`
- `X_std = X_centered / numpy.std(X_centered, axis = 0)`

Note that the standard deviation of feature vectors described by `X` and `X_centered` is the same, so the second line could also read:

- `X_std = X_centered / numpy.std(X, axis = 0)`

(e) The last quantity we will discuss is the *empirical covariance*. For two vectors $\vec{w}, \vec{y} \in \mathbb{R}^k$, the empirical covariance is defined as

$$\sigma(\vec{w}, \vec{y}) \doteq \frac{1}{k} \left(\vec{w} - \mu(\vec{w})\vec{1}\right)^\top \left(\vec{y} - \mu(\vec{y})\vec{1}\right) = \frac{1}{k} \sum_{i=1}^{k} (w_i - \mu(\vec{w}))(y_i - \mu(\vec{y})). \tag{5}$$

What is $\sigma(\vec{y}, \vec{y})$ in terms of the empirical statistics we have previously defined (e.g. mean, variance, and/or standard deviation)?

Solution: We have

$$\sigma(\vec{y}, \vec{y}) = \frac{1}{k}(\vec{y} - \mu(\vec{y})\vec{1})^\top (\vec{y} - \mu(\vec{y})\vec{1}) = \frac{1}{k}\left\|\vec{y} - \mu(\vec{y})\vec{1}\right\|_2^2 = \sigma^2(\vec{y}). \tag{6}$$

In words, this means that the empirical covariance between a vector and itself is just its empirical variance.

(f) For the remainder of this problem, assume that the data matrix is centered, so every feature has zero empirical mean; that is, $\mu(\vec{f}_j) = 0$ for every j.

Let $\Sigma(X) \in \mathbb{R}^{d\times d}$ denote the *empirical covariance matrix* of X. This matrix contains the empirical covariance of each pair of feature vectors $(\vec{f}_i, \vec{f}_j)$. Correspondingly it is defined entry-wise as

$$\Sigma(X)_{i,j} \doteq \sigma(\vec{f}_i, \vec{f}_j). \tag{7}$$

First, show that

$$\Sigma(X) = \frac{X^\top X}{n}. \tag{8}$$

Second, show that

$$\frac{X^\top X}{n} = \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i \vec{x}_i^\top. \tag{9}$$

Therefore, (9) entails that $\Sigma(X)$ can be represented in two ways.

HINT: One way to show two matrices are equal is to show that for all i, j, their (i, j)-th entries are equal.

Solution: We assume that every feature has zero empirical mean (i.e., $\mu(\vec{f}_i) = 0$ for every i), so our expression for each pairwise empirical covariance simplifies to

$$\Sigma(X)_{ij} = \sigma(\vec{f}_i, \vec{f}_j) = \frac{1}{n}\vec{f}_i^\top \vec{f}_j. \tag{10}$$

Now, we have that

$$\begin{aligned}
\frac{1}{n}X^\top X &= \frac{1}{n}\begin{bmatrix}\vec{f}_1 & \cdots & \vec{f}_d\end{bmatrix}^\top \begin{bmatrix}\vec{f}_1 & \cdots & \vec{f}_d\end{bmatrix} \\
&= \frac{1}{n}\begin{bmatrix}\vec{f}_1^\top \\ \vdots \\ \vec{f}_d^\top\end{bmatrix}\begin{bmatrix}\vec{f}_1 & \cdots & \vec{f}_d\end{bmatrix} \\
&= \frac{1}{n}\begin{bmatrix}\vec{f}_1^\top \vec{f}_1 & \cdots & \vec{f}_1^\top \vec{f}_d \\ \vdots & \ddots & \vdots \\ \vec{f}_d^\top \vec{f}_1 & \cdots & \vec{f}_d^\top \vec{f}_d\end{bmatrix} \\
&= \begin{bmatrix}\sigma(\vec{f}_1, \vec{f}_1) & \cdots & \sigma(\vec{f}_1, \vec{f}_d) \\ \vdots & \ddots & \vdots \\ \sigma(\vec{f}_d, \vec{f}_1) & \cdots & \sigma(\vec{f}_d, \vec{f}_d)\end{bmatrix} \\
&= \Sigma(X).
\end{aligned}$$

Next, one can show that both expressions given for $\Sigma(X)$ are the same by algebraic manipulation:

$$\frac{1}{n}\sum_{i=1}^{n} \vec{x}_i \vec{x}_i^\top = \frac{1}{n}\sum_{i=1}^{n} (X^\top \vec{e}_i)(X^\top \vec{e}_i)^\top \tag{11}$$

$$= \frac{1}{n}\sum_{i=1}^{n} X^\top \vec{e}_i \vec{e}_i^\top X \tag{12}$$

$$= \frac{1}{n} X^\top \underbrace{\left(\sum_{i=1}^{n} \vec{e}_i \vec{e}_i^\top\right)}_{=I_n} X \tag{13}$$

$$= \frac{1}{n} X^\top I_n X \tag{14}$$

$$= \frac{1}{n} X^\top X, \tag{15}$$

where $\vec{e}_i$ is the i-th standard basis vector.

(g) In this class, we consider three different interpretations of the term "projection". We define them explicitly here for this problem.

Consider vectors $\vec{a}$ and $\vec{b}$ in $\mathbb{R}^n$. Let $\vec{b}$ be unit norm (i.e., $\left\|\vec{b}\right\|_2^2 = \vec{b}^\top \vec{b} = 1$). We define the following:

i. The **vector projection** of $\vec{a}$ on $\vec{b}$ is given by $(\vec{a}^\top \vec{b})\vec{b}$. The vector projection is a vector in $\mathbb{R}^n$.

ii. The **scalar projection** of $\vec{a}$ on $\vec{b}$ is given by $\vec{a}^\top \vec{b}$. The scalar projection is a scalar but can take both positive and negative values.

iii. The **projection length** of $\vec{a}$ on $\vec{b}$ is given by $\left|\vec{a}^\top \vec{b}\right|$ and is the absolute value of the scalar projection.

Suppose we want to obtain a column vector $\vec{p} \in \mathbb{R}^n$ whose i-th entry is the *scalar* projection of data point $\vec{x}_i$ along the direction given by the unit vector $\vec{w}$. Show that $\vec{p}$ is given by

$$\vec{p} = X\vec{w}. \tag{16}$$

Solution: Because $\vec{w}$ is a unit vector, the scalar projection of any vector $\vec{y} \in \mathbb{R}^d$ onto $\vec{w}$ is $\vec{y}^\top \vec{w}$, as defined above. In the given equation, the i-th entry of $\vec{p}$ is $\vec{x}_i^\top \vec{w}$, precisely the scalar projection of each data point onto $\vec{w}$, as desired.

(h) Performing this kind of projection onto a unit vector $\vec{w}$ is at the heart of the PCA computation, which also requires computing the *variance* of these scalar projections.

Formally, for $i \in \{1, \ldots, n\}$, define $p_i \doteq \vec{x}_i^\top \vec{w}$, and define $\vec{p} \doteq \begin{bmatrix} p_1, \ldots, p_n \end{bmatrix}^\top$. Show that its empirical variance $\sigma^2(\vec{p})$ can be calculated as

$$\sigma^2(\vec{p}) = \frac{1}{n}\vec{w}^\top X^\top X \vec{w} = \vec{w}^\top \Sigma(X) \vec{w}. \tag{17}$$

Recall that X is centered.

Solution: We first calculate an expression for $\mu(\vec{p})$:

$$\mu(\vec{p}) = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{18}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i^\top \vec{w} \tag{19}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} w_j (\vec{x}_i)_j \tag{20}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} w_j(\vec{f}_j)_i && (21)\\
&= \sum_{j=1}^{d} w_j \cdot \frac{1}{n}\sum_{i=1}^{n}(\vec{f}_j)_i && (22)\\
&= \sum_{j=1}^{d} w_j \cdot \underbrace{\mu(\vec{f}_j)}_{=0} && (23)\\
&= 0. && (24)
\end{aligned}
$$

The underlined equality follows since every column of X has zero empirical mean after centering.

We now build an expression for $\sigma^2(\vec{p})$, showing that

$$
\begin{aligned}
\sigma^2(\vec{p}) &= \frac{1}{n}\sum_{i=1}^{n}(p_i - \mu(\vec{p}))^2 && (25)\\
&= \frac{1}{n}\sum_{i=1}^{n}(\vec{x}_i^\top \vec{w} - 0)^2 && (26)\\
&= \frac{1}{n}\sum_{i=1}^{n}(\vec{w}^\top \vec{x}_i)^2 && (27)\\
&= \frac{1}{n}\sum_{i=1}^{n}\vec{w}^\top \vec{x}_i\vec{x}_i^\top \vec{w} && (28)\\
&= \frac{1}{n}\vec{w}^\top\left(\sum_{i=1}^{n}\vec{x}_i\vec{x}_i^\top\right)\vec{w} && (29)\\
&= \frac{1}{n}\vec{w}^\top X^\top X\vec{w} && (30)\\
&= \vec{w}^\top\Sigma(X)\vec{w} && (31)
\end{aligned}
$$

as desired, where the last inequality is due to part (f).

2. PCA and Senate Voting Data

In this problem, we consider a matrix of senate voting data, which we manipulate in Python. The data is contained in a $n \times d$ data matrix X, where each row corresponds to a senator and each column to a bill. Each entry of X is either $1, -1$ or 0, depending on whether the senator voted for the bill, against the bill, or abstained, respectively. Please compute your answers using the attached Jupyter notebook `senator_pca.ipynb`. The sub-parts of this problem can be answered in the notebook itself in the space provided and can be submitted as an attachment to this PDF using the "Download as PDF" feature that Jupyter Notebook supports.

(a) Suppose we want to assign a *score* to each senator based on their voting pattern, and then observe the empirical variance of these scores. To describe this, let us choose a $\vec{a} \in \mathbb{R}^d$ and a scalar $b \in \mathbb{R}$. We define the score for senator i as:

$$f(\vec{x}_i, \vec{a}, b) = \vec{x}_i^\top \vec{a} + b, \quad i = 1, 2, \ldots, n. \tag{32}$$

Note that $\vec{x}_i^\top$ denotes the i^{th} row of X and is a row vector of length d, as in the previous problem.

Let us denote by $\vec{z} = f(X, \vec{a}, b)$ the column vector of length n obtained by stacking the scores for each senator. Then

$$\vec{z} = f(X, \vec{a}, b) = X\vec{a} + b\vec{1} \in \mathbb{R}^n \tag{33}$$

where $\vec{1}$ is a vector with all entries equal to 1. Let us denote the mean value of $\vec{z}$ by $\mu(\vec{z}) = \frac{1}{n}\vec{1}^\top \vec{z}$. Let $\vec{\mu}(X) \in \mathbb{R}^d$ denote the vector containing the mean of each column of X. Then

$$\begin{aligned}
\mu(\vec{z}) &= \frac{1}{n}\sum_{i=1}^{n} z_i && (34)\\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\vec{a}^\top \vec{x}_i + b\right) && (35)\\
&= \vec{a}^\top \left(\frac{1}{n}\sum_{i=1}^{n} \vec{x}_i\right) + b && (36)\\
&= \vec{a}^\top \vec{\mu}(X) + b && (37)
\end{aligned}$$

The empirical variance of the scores can then be obtained as

$$\begin{aligned}
\sigma^2(\vec{z}) &= \frac{1}{n}(\vec{z} - \mu(\vec{z})\vec{1})^\top(\vec{z} - \mu(\vec{z})\vec{1}) && (38)\\
&= \frac{1}{n}((X\vec{a} + b\vec{1}) - (\vec{a}^\top \vec{\mu}(X) + b)\vec{1})^\top((X\vec{a} + b\vec{1}) - (\vec{a}^\top \vec{\mu}(X) + b)\vec{1}) && (39)\\
&= \frac{1}{n}(X\vec{a} + b\vec{1} - (\vec{a}^\top \vec{\mu}(X))\vec{1} - b\vec{1})^\top(X\vec{a} + b\vec{1} - (\vec{a}^\top \vec{\mu}(X))\vec{1} - b\vec{1}) && (40)\\
&= \frac{1}{n}(X\vec{a} - (\vec{a}^\top \vec{\mu}(X))\vec{1})^\top(X\vec{a} - (\vec{a}^\top \vec{\mu}(X))\vec{1}) && (41)\\
&= \frac{1}{n}(X\vec{a} - \vec{1}\vec{\mu}(X)^\top \vec{a})^\top(X\vec{a} - \vec{1}\vec{\mu}(X)^\top \vec{a}) && (42)\\
&= \frac{1}{n}((X - \vec{1}\vec{\mu}(X)^\top)\vec{a})^\top((X - \vec{1}\vec{\mu}(X)^\top)\vec{a}) && (43)\\
&= \frac{1}{n}\vec{a}^\top(X - \vec{1}\vec{\mu}(X)^\top)^\top(X - \vec{1}\vec{\mu}(X)^\top)\vec{a}. && (44)
\end{aligned}$$

Note that this variance is therefore a function of the "centered" data matrix $X - \vec{1}\vec{\mu}(X)^\top$ in which the mean of each column is zero. It also does not depend on b.

For the remainder of this problem, we assume that the data has been pre-centered (i.e., $\vec{\mu}(X) = \vec{0}$); note that this has been pre-computed for you in the code notebook. Assume also that $b = 0$, so that $\mu(\vec{z}) = 0$. Defining $f(X, \vec{a}) \doteq f(X, \vec{a}, 0)$ and replacing $\vec{z}$ with $f(X, \vec{a})$, we can then write the simpler empirical variance formula

$$\sigma^2(f(X, \vec{a})) = \frac{1}{n}\vec{a}^\top X^\top X\vec{a}. \tag{45}$$

Suppose we restrict $\vec{a}$ to have unit-norm. In the provided code, find the maximum empirical variance $\sigma^2(f(X, \vec{a}))$ over all unit-norm $\vec{a}$, and find the $\vec{a}$ that maximizes it.

Solution: The vector $\vec{a}$ can be computed as the first principal component of the data matrix X. For the full solution, see the Jupyter notebook. (Vector $\vec{a}$ is the 542×1 column vector `a_1`, and the maximum variance is just under 150.)

As a side note, computing PCA by finding the direction of maximum variance is only one possible way of conceptualizing the computation; it's also equivalent to minimizing the reconstruction error in a lower-dimensional space. We'll explore this equivalence further in a future homework!

(b) We next consider party affiliation as a predictor for how a senator will vote. Follow the instructions in the notebook to compute the mean voting vector for each party and relate it to the direction of maximum variance.

Solution: For solution, see Jupyter Notebook.

(c) Recall from problem 1 that given a vector $\vec{z} = X\vec{u}$ (i.e., the vector of scalar projections of each row of X along $\vec{u}$), we can compute its empirical variance as

$$\sigma^2(\vec{z}) = \vec{u}^\top \Sigma \vec{u}, \tag{46}$$

where $\Sigma(X) = \frac{X^\top X}{n}$ is the empirical covariance matrix of X. We will show in a future homework problem that the variance along each principal component $\vec{a}_i$ is precisely its corresponding eigenvalue of $\Sigma(X)$, i.e., $\lambda_i\{\Sigma(X)\}$. (For now, just note that this fact should make intuitive sense, since PCA is searching for directions of maximum variance of the data, and these occur along the covariance matrix's eigenvectors.) In the Notebook, compute the sum of the variance along $\vec{a}_1$ and $\vec{a}_2$ and plot the data projected on the $\vec{a}_1$–$\vec{a}_2$ plane.

Solution: For solution, see Jupyter Notebook.

(d) Suppose we want to find the bills that are most and least contentious — i.e., those that have high variability in senators' votes, and those for which voting was almost unanimous. Follow the instructions in the Jupyter notebook to compute a measure of "contentiousness" for each bill, plot the vote counts for exemplar bills, and comment on the voting trends.

Solution: For solution, see Jupyter Notebook.

(e) Suppose we want to infer the political affiliations of two senators whose voting records are known to us. Follow the instructions in the Jupyter notebook to infer the political affiliation of the Green and Grey colored senators using PCA.

Solution: For solution, see Jupyter Notebook.

(f) Finally, we can use the defined score $f(X, \vec{a}, b)$, computed along the first principal component $\vec{a}_1$ to classify the most and least "extreme" senators based on their voting record. Follow the instructions in the Jupyter Notebook to compute these scores and comment on their relationship to partisan affiliation.

Solution: For solution, see Jupyter Notebook.

3. SVD Transformation

In this problem we will interpret the linear map defined by matrix $A \in \mathbb{R}^{n \times n}$ by looking at its singular value decomposition, $A = UDV^\top$. Recall that here $U, D, V \in \mathbb{R}^{n \times n}$ and U, V are orthonormal matrices while D is a diagonal matrix. We will first look at how $V^\top, D$ and U each separately transform the unit circle $C = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$ and then look at their effect as a whole. This problem has an associated jupyter notebook, `svd_transformation.ipynb` that contains parts (b, c, d, e) of the problem. These sub-parts can be answered in the space provided in the notebook itself and can be submitted as an attachment to the solution PDF using the "Download as PDF" feature that Jupyter Notebook supports.

(a) Let $\vec{x} \in \mathbb{R}^n$ and $\vec{z} = V^\top \vec{x}$. Show that $\vec{x} = \sum_{i=1}^{n} z_i \vec{v}_i$, where $\vec{z} = \begin{bmatrix} z_1 & \cdots & z_n \end{bmatrix}^\top$ and $V = \begin{bmatrix} \vec{v}_1 & \cdots & \vec{v}_n \end{bmatrix}$. This shows that $\vec{z}$ is the vector of coordinates that represents $\vec{x}$ in the basis defined by the columns of V.

Solution: Since the columns of V form a basis for $\mathbb{R}^n$, we can represent $\vec{x} \in \mathbb{R}^n$ as $V\vec{z}$ for some $\vec{z}$ in $\mathbb{R}^n$. Then,

$$V^\top \vec{x} = V^\top V \vec{z} \tag{47}$$
$$= \vec{z}. \tag{48}$$

The last equality follows since V is an orthonormal matrix.

For the rest of the problem we restrict ourselves to the case where $A \in \mathbb{R}^{2 \times 2}$ and move to the Jupyter notebook.

Solution: The solutions for the rest of the parts can be found in the Jupyter notebook solution.

4. FTLA, SVD, Pseudoinverse, and Least-Squares

Let $A \in \mathbb{R}^{m\times n}$ be a matrix, and let $\vec{y} \in \mathbb{R}^m$. Let $r \doteq \operatorname{rank}(A)$, and let

$$A = U\Sigma V^\top = \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix} \begin{bmatrix} V_r^\top \\ V_{n-r}^\top \end{bmatrix} = U_r\Sigma_r V_r^\top = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top \tag{49}$$

be an SVD of A. In this problem, we will prove some properties about the SVD, and then apply them to derive a unique solution to an optimization problem which generalizes minimum-norm and least-squares.

(a) Prove that $\mathcal{R}(A) = \mathcal{R}(U_r)$ and that $\mathcal{R}(A^\top) = \mathcal{R}(V_r)$.

Solution: There are many correct ways to show this, and any correct solution merits full credit. The following is only one such correct solution.

Let $\vec{y} \in \mathcal{R}(A)$. Then there exists $\vec{x} \in \mathbb{R}^n$ such that $\vec{y} = A\vec{x}$. We have that

$$\vec{y} = A\vec{x} \tag{50}$$
$$= U_r\Sigma_r V_r^\top \vec{x} \tag{51}$$
$$= U_r(\Sigma_r V_r^\top \vec{x}) \tag{52}$$

so $\vec{y} \in \mathcal{R}(U_r)$. Thus $\mathcal{R}(A) \subseteq \mathcal{R}(U_r)$.

Now let $\vec{y} \in \mathcal{R}(U_r)$. Then there exists $\vec{x} \in \mathbb{R}^r$ such that $\vec{y} = U_r\vec{x}$. Define $\vec{z} \doteq V_r\Sigma_r^{-1}\vec{x}$. Then $\vec{x} = \Sigma_r V_r^\top \vec{z}$, and so

$$\vec{y} = U_r\vec{x} \tag{53}$$
$$= U_r\Sigma_r V_r^\top \vec{z} \tag{54}$$
$$= A\vec{z}. \tag{55}$$

Thus $\vec{y} \in \mathcal{R}(A)$, so $\mathcal{R}(U_r) \subseteq \mathcal{R}(U_r)$. Thus $\mathcal{R}(A) = \mathcal{R}(U_r)$. $\mathcal{R}(A^\top) = \mathcal{R}(V_r)$ is proved similarly.

(b) Use the fundamental theorem of linear algebra to prove that $\mathcal{N}(A) = \mathcal{R}(V_{n-r})$ and $\mathcal{N}(A^\top) = \mathcal{R}(U_{m-r})$.

Solution: There are many correct ways to show this, and any correct solution merits full credit. The following is only one such correct solution.

We know that the columns of V form an orthonormal basis for $\mathbb{R}^n$, and that $\mathcal{R}(V_r) = \mathcal{R}(A^\top)$. Because the columns of $V = \begin{bmatrix} V_r & V_{n-r} \end{bmatrix}$ are orthonormal, the columns of V_r are orthogonal to the columns of V_{n-r}. Since together they span $\mathbb{R}^n$, their individual spans are orthogonal complements; in particular, we have $\mathcal{R}(V_r)^\perp = \mathcal{R}(V_{n-r})$. Finally, by FTLA we have

$$\mathcal{R}(V_{n-r}) = \mathcal{R}(V_r)^\perp = \mathcal{R}(A^\top)^\perp = \mathcal{N}(A). \tag{56}$$

The solution for $\mathcal{R}(U_{m-r}) = \mathcal{N}(A^\top)$ follows from taking transposes, just like the previous part.

Now, we will use the SVD to derive the unique solution to the *least-norm least-squares* problem. That is, we want to use the SVD to find the unique vector in $\mathbb{R}^n$ which solves the least-squares problem with minimum norm. More formally, we wish to solve the least-norm least-squares problem:

$$\min_{\vec{x}\in S} \|\vec{x}\|_2^2 \qquad \text{where} \qquad S \doteq \operatorname*{argmin}_{\vec{x}\in\mathbb{R}^n} \|A\vec{x}-\vec{y}\|_2^2 . \tag{57}$$

In words, we want to find the minimum-norm vector $\vec{x}$ in the set of all least-squares solutions S.[a] This problem generalizes both the traditional least-squares and minimum-norm problems.

To solve the least-norm least-squares problem, we will use the SVD to define a concept called the *pseudoinverse*. The matrix $A^\dagger \in \mathbb{R}^{n\times m}$ is called a *pseudoinverse* (sometimes a *Moore-Penrose pseudoinverse*) of A, where

$$A^\dagger = V\Sigma^\dagger U^\top = \begin{bmatrix} V_r & V_{n-r}\end{bmatrix}\begin{bmatrix}\Sigma_r^{-1} & 0_{r\times(m-r)} \\ 0_{(n-r)\times r} & 0_{(n-r)\times(m-r)}\end{bmatrix}\begin{bmatrix}U_r^\top \\ U_{m-r}^\top\end{bmatrix} = V_r\Sigma_r^{-1}U_r^\top = \sum_{i=1}^{r}\frac{1}{\sigma_i}\vec{v}_i\vec{u}_i^\top . \tag{58}$$

In this problem, we will show that $A^\dagger\vec{y}$ is the unique solution to the least-norm least-squares problem (57).

NOTE: In this problem, we *do not* assume that A has full column rank or full row rank — in fact, A could even be the matrix of all zeros — so the least-squares solution we learn in class does not apply.

[a] Here, the argmin is a set – that is, it is the set of minimizers of the least-squares objective. A simpler example of this notation is that if $f(x) \doteq 0$ for all $x\in\mathbb{R}$, then $\operatorname{argmin}_{x\in\mathbb{R}} f(x) = \mathbb{R}$. More information about this notation is contained in section 1.3 of the course reader.

(c) Show that $\vec{x}\in S$ if and only if $A^\top A\vec{x} = A^\top\vec{y}$ (these are the so-called *normal equations*, which you may remember from our discussion on least-squares).

Solution: There are many correct ways to show this, and any correct solution merits full credit. The following is a small selection of correct solutions.

We have $\vec{x}\in S$ if and only if $A\vec{x}$ is the closest point in $\mathcal{R}(A)$ to $\vec{y}$, or in other words $A\vec{x} = \operatorname{proj}_{\mathcal{R}(A)}(\vec{y})$. This holds if and only if $\vec{y}-A\vec{x}$ is orthogonal to $\mathcal{R}(A)$, or in other words $A^\top(\vec{y}-A\vec{x}) = \vec{0}$, which is equivalent to $A^\top A\vec{x} = A^\top\vec{y}$.

(d) Show that $S = \{A^\dagger\vec{y} + \vec{z} \mid \vec{z}\in\mathcal{N}(A)\}$.

HINT: First, show that $A^\dagger\vec{y}\in S$. Then show that for any two $\vec{x}_1,\vec{x}_2\in S$, that $\vec{x}_1-\vec{x}_2\in\mathcal{N}(A)$. Argue that this implies the conclusion.

Solution: There are many correct ways to show this, and any correct solution merits full credit. The following is only one such correct solution.

As the hint suggests, we argue in two parts.

First, we claim that $A^\dagger\vec{y}\in S$. Indeed, we plug in

$$\begin{aligned}
A^\top AA^\dagger\vec{y} &= (U_r\Sigma_rV_r^\top)^\top(U_r\Sigma_rV_r^\top)(V_r\Sigma_r^{-1}U_r^\top)\vec{y} && (59)\\
&= V_r\Sigma_r^\top U_r^\top U_r\Sigma_rV_r^\top V_r\Sigma_r^{-1}U_r^\top\vec{y} && (60)\\
&= V_r\Sigma_r^\top\Sigma_r\Sigma_r^{-1}U_r^\top\vec{y} && (61)\\
&= V_r\Sigma_r^2\Sigma_r^{-1}U_r^\top\vec{y} && (62)\\
&= V_r\Sigma_rU_r^\top\vec{y} && (63)\\
&= (U_r\Sigma_rV_r^\top)^\top\vec{y} && (64)\\
&= A^\top\vec{y}. && (65)
\end{aligned}$$

Thus $A^\dagger\vec{y}$ fulfills the normal equations, so $A^\dagger\vec{y}\in S$.

Now we claim that for any $\vec{x}_1, \vec{x}_2 \in S$, that $\vec{x}_1 - \vec{x}_2 \in \mathcal{N}(A)$. There are many different solutions for this; we present three here.

i. One solution is to again use the fact that for any $\vec{x} \in S$ we have $A\vec{x} = \text{proj}_{\mathcal{R}(A)}(\vec{y})$. Thus

$$A(\vec{x}_1 - \vec{x}_2) = A\vec{x}_1 - A\vec{x}_2 = \text{proj}_{\mathcal{R}(A)}(\vec{y}) - \text{proj}_{\mathcal{R}(A)}(\vec{y}) = \vec{0}. \tag{66}$$

Thus $\vec{x}_1 - \vec{x}_2 \in \mathcal{N}(A)$.

ii. Another solution is to use the orthogonality principle: namely, the residual of a projection onto a subspace is orthogonal to any vector in the subspace. In our case, this manifests as the following assertion: for any $\vec{v} \in \mathcal{R}(A)$ we have $(\vec{y} - \text{proj}_{\mathcal{R}(A)}(\vec{y}))^\top \vec{v} = 0$.
Since $\vec{x}_1, \vec{x}_2 \in S$, we have $A\vec{x}_1 = A\vec{x}_2 = \text{proj}_{\mathcal{R}(A)}(\vec{y})$, and in particular

$$\|\vec{y} - A\vec{x}_1\|_2^2 = \|\vec{y} - A\vec{x}_2\|_2^2. \tag{67}$$

But we can write

$$\begin{aligned}
\|\vec{y} - A\vec{x}_2\|_2^2 &= \|(\vec{y} - A\vec{x}_1) + A(\vec{x}_1 - \vec{x}_2)\|_2^2 && (68)\\
&= \|\vec{y} - A\vec{x}_1\|_2^2 + 2\underbrace{\underbrace{(\vec{y} - A\vec{x}_1)}_{=\vec{y}-\text{proj}_{\mathcal{R}(A)}(\vec{y})}{}^\top \underbrace{(A\{\vec{x}_1 - \vec{x}_2\})}_{\in\mathcal{R}(A)}}_{=0} + \|A(\vec{x}_1 - \vec{x}_2)\|_2^2 && (69)\\
&= \|\vec{y} - A\vec{x}_1\|_2^2 + \|A(\vec{x}_1 - \vec{x}_2)\|_2^2. && (70)
\end{aligned}$$

Thus we have

$$\|\vec{y} - A\vec{x}_1\|_2^2 = \|\vec{y} - A\vec{x}_2\|_2^2 = \|\vec{y} - A\vec{x}_1\|_2^2 + \|A(\vec{x}_1 - \vec{x}_2)\|_2^2 \tag{71}$$

so $\|A(\vec{x}_1 - \vec{x}_2)\|_2^2 = 0$, which is true if and only if $A(\vec{x}_1 - \vec{x}_2) = \vec{0}$, i.e., $\vec{x}_1 - \vec{x}_2 \in \mathcal{N}(A)$.

iii. A third solution is to note that $\mathcal{N}(A) = \mathcal{N}(A^\top A)$, so we show that $\vec{x}_1 - \vec{x}_2 \in \mathcal{N}(A^\top A)$. We have

$$\begin{aligned}
A^\top A(\vec{x}_1 - \vec{x}_2) &= A^\top A\vec{x}_1 - A^\top A\vec{x}_2 && (72)\\
&= A^\top \vec{y} - A^\top \vec{y} && (73)\\
&= \vec{0}. && (74)
\end{aligned}$$

Thus $\vec{x}_1 - \vec{x}_2 \in \mathcal{N}(A^\top A) = \mathcal{N}(A)$.

Now, since $A^\dagger \vec{y} \in S$, we know that $\vec{z} = A^\dagger \vec{y} - \vec{x} \in \mathcal{N}(A)$ for any $\vec{x} \in S$. Thus, this $\vec{x}$ can be written as $A^\dagger \vec{y} - \vec{z}$ for some $\vec{z} \in \mathcal{N}(A)$, and $\vec{z} \in \mathcal{N}(A)$ if and only if $-\vec{z} \in \mathcal{N}(A)$. Thus, any $\vec{x} \in S$ can be written as $\vec{x} = A^\dagger \vec{y} + \vec{z}$ for some $\vec{z} \in \mathcal{N}(A)$, so $S \subseteq \{A^\dagger \vec{y} + \vec{z} \mid \vec{z} \in \mathcal{N}(A)\}$.

To show the other direction, i.e., $S \supseteq \{A^\dagger \vec{y} + \vec{z} \mid \vec{z} \in \mathcal{N}(A)\}$, note that for any $\vec{z} \in \mathcal{N}(A)$ we can show that

$$\begin{aligned}
A^\top A(A^\dagger \vec{y} + \vec{z}) &= A^\top A A^\dagger \vec{y} + A^\top A\vec{z}\\
&= A^\top \vec{y} + \vec{0}\\
&= A^\top \vec{y},
\end{aligned}$$

so that $A^\dagger \vec{y} + \vec{z} \in S$. Thus $S = \{A^\dagger \vec{y} + \vec{z} \mid \vec{z} \in \mathcal{N}(A)\}$ as desired.

(e) Now show that $\{A^{\dagger}\vec{y}\} = \operatorname{argmin}_{\vec{x}\in S} \|\vec{x}\|_2^2$.

HINT: Pick any $\vec{z} \in \mathcal{N}(A)$. Show that $A^{\dagger}\vec{y}$ is orthogonal to $\vec{z}$. Then show that $\left\|A^{\dagger}\vec{y}\right\|_2^2 \leq \left\|A^{\dagger}\vec{y} + \vec{z}\right\|_2^2$, with equality if and only if $\vec{z} = \vec{0}$. Argue that this implies the conclusion.

Solution: There are many correct ways to show this, and any correct solution merits full credit. The following is only one such correct solution.

Since $\vec{z} \in \mathcal{N}(A)$, and $\mathcal{R}(V_{n-r}) = \mathcal{N}(A)$, we can write $\vec{z} = V_{n-r}\vec{w}$ for some $\vec{w} \in \mathbb{R}^{n-r}$. Then

$$\begin{aligned}
\vec{z}^{\top} A^{\dagger}\vec{y} &= (V_{n-r}\vec{w})^{\top} A^{\dagger}\vec{y} && (75)\\
&= \vec{w}^{\top} V_{n-r}^{\top} A^{\dagger}\vec{y} && (76)\\
&= \vec{w}^{\top} \underbrace{V_{n-r}^{\top} V_r}_{=0} \Sigma_r^{-1} U_r^{\top}\vec{y} && (77)\\
&= \vec{w}^{\top} \cdot 0 \cdot \Sigma_r^{-1} U_r^{\top}\vec{y} && (78)\\
&= 0. && (79)
\end{aligned}$$

Now, expanding the right-hand side of the inequality, we have

$$\begin{aligned}
\left\|A^{\dagger}\vec{y} + \vec{z}\right\|_2^2 &= \left\|A^{\dagger}\vec{y}\right\|_2^2 + \|\vec{z}\|_2^2 + 2\vec{z}^{\top} A^{\dagger}\vec{y}\\
&= \left\|A^{\dagger}\vec{y}\right\|_2^2 + \|\vec{z}\|_2^2\\
&\geq \left\|A^{\dagger}\vec{y}\right\|_2^2,
\end{aligned}$$

with equality if and only if $\vec{z} = \vec{0}$, as desired.

5. Properties of the Frobenius Norm

The Frobenius norm of a matrix P is defined as

$$\|P\|_F = \sqrt{\langle P,\ P\rangle} = \sqrt{\sum_{\substack{1\le i\le m\\1\le j\le n}} P_{ij}^2}, \tag{80}$$

where for two matrices $P, Q \in \mathbb{R}^{m\times n}$, the canonical inner product defined over this space is $\langle P,\ Q\rangle := \operatorname{tr}\left(P^\top Q\right) = \sum_{i,j} P_{ij}Q_{ij}$. The previous definition of the inner product is equivalent to interpreting the matrices P and Q as vectors of length mn and taking the vector inner product of the respective mn-dimensional vectors. The Cauchy-Schwarz inequality for the inner product follows in a straightforward way from the Cauchy-Schwarz inequality for vectors:

$$\langle P,\ Q\rangle = \sum_{\substack{1\le i\le m\\1\le j\le n}} P_{ij}Q_{ij} \le \left(\sum_{\substack{1\le i\le m\\1\le j\le n}} P_{ij}^2\right)^{1/2}\left(\sum_{\substack{1\le i\le m\\1\le j\le n}} Q_{ij}^2\right)^{1/2} = \|P\|_F\,\|Q\|_F\,. \tag{81}$$

(a) Show that the Frobenius norm satisfies all three properties of a norm.

*HINT: The easiest way to do this problem is to **not** look at the individual components P_{ij}, but instead use the inner product formulation $\|P\|_F = \sqrt{\langle P,\ P\rangle}$.*

Solution:

i. $\|P\|_F = 0 \iff P = 0$: $P = 0 \iff \langle P,\ P\rangle = 0 \iff \|P\|_F = 0$

ii. For every $\alpha \in \mathbb{R}$, $\|\alpha P\|_F = |\alpha|\,\|P\|_F$:
$\|\alpha P\|_F = \sqrt{\langle \alpha P,\ \alpha P\rangle} = \sqrt{\alpha^2\langle P,\ P\rangle} = |\alpha|\sqrt{\langle P,\ P\rangle} = |\alpha|\,\|P\|_F$

iii. $\|P+Q\|_F \le \|P\|_F + \|Q\|_F$ for all $P, Q \in \mathbb{R}^{m\times n}$:
Equivalently, we can show

$$\begin{aligned}
\|P+Q\|_F^2 &= \langle P+Q,\ P+Q\rangle && (82)\\
&= \langle P,\ P\rangle + 2\langle P,\ Q\rangle + \langle Q,\ Q\rangle && (83)\\
&= \|P\|_F^2 + \|Q\|_F^2 + 2\langle P,\ Q\rangle && (84)\\
&\le \|P\|_F^2 + \|Q\|_F^2 + 2\|P\|_F\,\|Q\|_F && \text{Cauchy-Schwarz} \quad (85)\\
&= (\|P\|_F + \|Q\|_F)^2 && (86)
\end{aligned}$$

(b) Write the Frobenius norm squared in terms of singular values.

HINT: The cyclic property of traces might be helpful: $\operatorname{tr}(PQR) = \operatorname{tr}(RPQ) = \operatorname{tr}(QRP)$.

Solution:

$$\begin{aligned}
\|P\|_F^2 &= \langle P,\ P\rangle && (87)\\
&= \operatorname{tr}\left(P^\top P\right) && (88)\\
&= \operatorname{tr}\left((U\Sigma V^\top)^\top U\Sigma V^\top\right) && (89)\\
&= \operatorname{tr}\left(V\Sigma^\top U^\top U\Sigma V^\top\right) && U^\top U = I \quad (90)\\
&= \operatorname{tr}\left(V\Sigma^\top \Sigma V^\top\right) && (91)\\
&= \operatorname{tr}\left(V^\top V\Sigma^\top \Sigma\right) && \text{Rotation property of trace} \quad (92)\\
&= \operatorname{tr}\left(\Sigma^\top \Sigma\right) && (93)
\end{aligned}$$

$$= \sum_i \sigma_i^2 \tag{94}$$

(c) Express the Frobenius norm squared in terms of the ℓ_2-norm of the columns of P with $\vec{p}_i$ denoting column i. Concretely, prove $\|P\|_F^2 = \sum_{i=1}^n \|\vec{p}_i\|_2^2$ where $\vec{p}_i$ are the columns of P.

Solution:

$$\|P\|_F^2 = \langle P,\ P\rangle = \operatorname{tr}\left(P^\top P\right) \tag{95}$$

$$P^\top P = \begin{bmatrix} \vec{p}_1^\top \\ \vec{p}_2^\top \\ \vdots \\ \vec{p}_n^\top \end{bmatrix} \begin{bmatrix} \vec{p}_1 & \vec{p}_2 & \cdots & \vec{p}_n \end{bmatrix} = \begin{bmatrix} \vec{p}_1^\top \vec{p}_1 & \vec{p}_1^\top \vec{p}_2 & \cdots & \vec{p}_1^\top \vec{p}_n \\ \vec{p}_2^\top \vec{p}_1 & \vec{p}_2^\top \vec{p}_2 & \cdots & \vec{p}_2^\top \vec{p}_n \\ \vdots & \vdots & \ddots & \vdots \\ \vec{p}_n^\top \vec{p}_1 & \vec{p}_n^\top \vec{p}_2 & \cdots & \vec{p}_n^\top \vec{p}_n \end{bmatrix} \tag{96}$$

$$\operatorname{tr}\left(P^\top P\right) = \sum_i \vec{p}_i^\top \vec{p}_i = \sum_i \|\vec{p}_i\|_2^2 \tag{97}$$

6. Homework Process

With whom did you work on this homework? List the names and SIDs of your group members.

NOTE: If you didn't work with anyone, you can put "none" as your answer.