EECS 127/227AT Optimization Models in Engineering

Spring 2019 Midterm 1

1. **(1 Point) What is one of your favorite things to do outside of school?**

2. **(1 Point) What is one of things you learned in 127/227AT that you enjoyed?**

Do not turn this page until the proctor tells you to do so. You may work on the questions above.

PRINT your name and student ID: ____________________

3. (13 points) Singular value decomposition

The compact form of the singular value decomposition of a matrix $A \in \mathbb{R}^{3\times 3}$ is given as

$$A = \begin{bmatrix} \frac{2}{3} & \frac{1}{\sqrt{2}} \\ \frac{2}{3} & -\frac{1}{\sqrt{2}} \\ \frac{1}{3} & 0 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \end{bmatrix}.$$

(a) (2 points) What is the rank of A? Justify.

Solution: A has two nonzero singular values, so its rank is 2.

(b) (3 points) What is the dimension of the column space (range) of A? Write a basis for the column space (range) of A.

Solution: Given the SVD decomposition $A = U\Sigma V^\top$, $\mathcal{R}(A)$ is spanned by the columns of U corresponding to the nonzero singular values. Therefore,

$$\left\{ \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \right\}$$

is a basis for $\mathcal{R}(A)$, and $\dim(\mathcal{R}(A)) = 2$.

(c) (4 points) What is the dimension of the null space of $A^\top$? Write a basis for the null space of $A^\top$. **Solution:** By Fundamental Theorem of Linear Algebra, $\mathcal{N}(A^\top)$ is orthogonal to $\mathcal{R}(A)$, and it complements $\mathcal{R}(A)$ to fill up $\mathbb{R}^3$. Therefore, $\mathcal{N}(A^\top)$ is spanned by the columns of U corresponding to the zero singular values. In other words, $\mathcal{N}(A^\top) = \operatorname{span}\left(\begin{bmatrix} a & b & c \end{bmatrix}^\top\right)$, where $\begin{bmatrix} a & b & c \end{bmatrix}^\top$ is the unshown column of U. The fact that U is an orthonormal matrix implies that $2a + 2b + c = 0$ and $a - b = 0$. Consequently,

$$\left\{ \begin{bmatrix} 1 \\ 1 \\ -4 \end{bmatrix} \right\}$$

is a basis for $\mathcal{N}(A^\top)$ and $\dim(\mathcal{N}(A^\top)) = 1$.

(d) (4 points) Let $\mathcal{B}_2$ denote the unit-norm ball in ℓ_2 norm: $\mathcal{B}_2 = \{z \in \mathbb{R}^3 : \|z\|_2 \leq 1\}$. Compute the minimum value of $x^\top Ay$, where x and y are two vectors in $\mathcal{B}_2$; that is, find $\min_{x,y\in\mathcal{B}_2} x^\top Ay$.

Solution: First note that $\min_{x,y\in\mathcal{B}_2} x^\top Ay = -\max_{x,y\in\mathcal{B}_2} x^\top Ay$. Then,

$$\max_{x,y\in\mathcal{B}_2} x^\top Ay = \max_{x,y\in\mathcal{B}_2} x^\top U\Sigma V^\top y = \max_{x,y\in\mathcal{B}_2} x^\top \Sigma y = \sigma_{\max}(A) = 3,$$

where the second equality follows from the fact that U and V are orthonormal matrices. As a result, $\min_{x,y\in\mathcal{B}_2} x^\top Ay = -3$.

PRINT your name and student ID: ____________________

4. (12 points) Symmetric and skew-symmetric matrices

A square matrix $A \in \mathbb{R}^{n\times n}$ is called skew-symmetric if all its diagonal elements are zero and $A_{ij} = -A_{ji}$ for all $i, j \in \{1, \dots, n\}$. In other words, A is skew-symmetric if and only if $A^\top = -A$.

(a) (3 points) Let $A \in \mathbb{R}^{n\times n}$ be a skew-symmetric matrix, and let $B \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Show that $\langle A, B\rangle = 0$, where inner product of A and B is defined as

$$\langle A, B\rangle = \mathrm{Trace}(A^\top B).$$

Note that this implies the space of symmetric matrices is orthogonal to the space of skew-symmetric matrices.

Solution:

Solution 1: Note that the diagonal entries of skew-symmetric matrix A must be 0.

$$\langle A, B\rangle = \sum_i \sum_j A_{ij}B_{ij} = \sum_i \sum_{j>i} A_{ij}B_{ij} + \sum_i \sum_{j<i} A_{ij}B_{ij} = \sum_i \sum_{j>i} A_{ij}\left(B_{ij} - B_{ij}\right) = 0.$$

Solution 2: B can be written as $B = UU^\top$ for some $U \in \mathbb{R}^{n\times n}$. Then,

$$\langle A, B\rangle = \langle A, UU^\top\rangle = \langle AU, U\rangle = \langle -A^\top U, U\rangle = -\langle U, AU\rangle,$$

which can hold only when $\langle AU, U\rangle = 0 \iff \langle A, B\rangle = 0$.

(b) (2 points) The set of all matrices in $\mathbb{R}^{2\times 2}$ forms a vector space. All elements in this space can be written as a linear combination of

$$\left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\},$$

and because these matrices are linearly independent, they provide a basis for $\mathbb{R}^{2\times 2}$.

Similarly, the set of all skew-symmetric matrices in $\mathbb{R}^{2\times 2}$ forms a vector space. Write a basis for the space of skew-symmetric matrices in $\mathbb{R}^{2\times 2}$.

Solution: $\left\{ \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & 0 \end{bmatrix} \right\}$

(c) (4 points) Consider the matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathbb{R}^{2\times 2}$. Find a symmetric matrix $A_{\text{sym}} \in \mathbb{R}^{2\times 2}$ and a skew-symmetric matrix $A_{\text{skew}} \in \mathbb{R}^{2\times 2}$ such that

$$A = A_{\text{sym}} + A_{\text{skew}}.$$

Solution:

$$\left\langle \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & 0 \end{bmatrix} \right\rangle \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & 0 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 0 & b-c \\ c-b & 0 \end{bmatrix}.$$

Solution: From part (a), we know that the space of symmetric matrices and the space of skew-symmetric matrices are orthogonal. Then, the projection of the given matrix onto the space of symmetric matrices will be equal to the projection of

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} - \frac{1}{2}\begin{bmatrix} 0 & b-c \\ c-b & 0 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 2a & b+c \\ b+c & 2d \end{bmatrix}.$$

However, since this matrix is symmetric, its projection onto the space of symmetric matrices will be itself.

(d) (3 points) Consider the function $f : \mathbb{R}^2 \mapsto \mathbb{R}$, which is defined as $f(x) = \frac{1}{2}x^\top \begin{bmatrix} a & b \\ c & d \end{bmatrix} x$. Find the Hessian of the function. Show your calculations.

Solution: For any $A \in \mathbb{R}^{n\times n}$, the Hessian of $f(x) = \frac{1}{2}x^\top Ax$ is given as $\frac{1}{2}(A + A^\top)$. Then, the Hessian of the given function is

$$\frac{1}{2}\begin{bmatrix} 2a & b+c \\ b+c & 2d \end{bmatrix}.$$

PRINT your name and student ID: ____________________

5. (12 points) Online Least Squares

Consider n sensors located in different areas of California to measure the temperature of the air (as a scalar).

Let $x_{i,t}$ denote the measurement from sensor i at time t, for for $i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, T$. Assume $T < n$.

We represent $x_t \in \mathbb{R}^n$ as a column vector of all measurements at time t. Thus,

$$x_t = \begin{bmatrix} x_{1,t} \\ x_{2,t} \\ \vdots \\ x_{n,t} \end{bmatrix}.$$

Let $X_t \in \mathbb{R}^{n \times t}$ denote the matrix with columns $x_1, x_2 \ldots x_t$. Thus we have,

$$X_t = \begin{bmatrix} x_1 & x_2 & \ldots & x_t \end{bmatrix}.$$

We additionally consider scalars $y_1, y_2, \ldots, y_n$ where $y_i \in \mathbb{R}$. Here y_i represents wind chill at the region corresponding to sensor i, as predicted by meteorological department with the help of weather satellites. Note that y_i does not depend on time t. Let $y \in \mathbb{R}^n$ denote the column vector containing the wind chill at all sensors. Thus,

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Define $q_t \in \mathbb{R}^n$ iteratively as follows. First,

$$q_1 = \frac{x_1}{\|x_1\|_2}.$$

For $t = 2, 3, \ldots, T$:

$$s_t = x_t - \sum_{j=1}^{t-1} \langle x_t, q_j \rangle q_j$$

$$q_t = \frac{s_t}{\|s_t\|_2}.$$

Let $Q_t \in \mathbb{R}^{n \times t}$ denote the matrix with columns $q_1, q_2, \ldots, q_t$. Thus we have,

$$Q_t = \begin{bmatrix} q_1 & q_2 & \ldots & q_t \end{bmatrix}.$$

Assume that for all $t = 1, 2, \ldots, T$, the matrix X_t is full column rank, i.e the columns of X_t are linearly independent. Further for this problem, assume that the inner product of x and y, $\langle x, y \rangle$, is given by $x^\top y$.

For each time $i = 1, 2, \dots t$, we are interested in fitting a linear model to predict y from X_t, the sensor measurements up to time t. Consider the following two problems at time t:

$$
\begin{aligned}
w_t^* &= \underset{w}{\operatorname{argmin}} \|y - X_t w\|_2 \\
v_t^* &= \underset{v}{\operatorname{argmin}} \|y - Q_t v\|_2 .
\end{aligned}
$$

(a) (1 point) Write an expression for w_t^* in terms of X_t and y.

Solution: Taking gradient of the squared objective and setting it to zero we obtain:

$$w_t^* = (X_t^\top X_t)^{-1} X_t^\top y.$$

(b) (1 point) Write an expression for v_t^* in terms of Q_t and y.

Solution: Taking gradient of the squared objective and setting it to zero we obtain:

$$v_t^* = (Q_t^\top Q_t)^{-1} Q_t^\top y.$$

(c) (3 points) Show that for each t,

$$X_t w_t^* = Q_t v_t^*.$$

Note that this implies that the matrices Q and X have the information for fitting a linear model for y.

Solution: Method 1:

$X_t w_t^*$ is the projection of y onto the column space of X. Similarly $Q_t v_t^*$ is the projection of y onto the column space of Q. Since X and Q have the same column space, the result holds.

Method 2:

Note that the QR decomposition of X_t yields

$$X_t = Q_t R_t,$$

since Q_t is an orthonormal basis for column space of X_t.

Substituting X_t in terms of Q_t and R_t and using the fact that R_t is invertible we get,

$$
\begin{aligned}
w_t^* &= ((Q_t R_t)^\top Q_t R_t)^{-1} (Q_t R_t)^\top y \\
&= (R_t^\top Q_t^\top Q_t R_t)^{-1} R_t^\top Q_t^\top y \\
&= (R_t)^{-1} (Q_t^\top Q_t)^{-1} (R_t^\top)^{-1} R_t^\top Q_t^\top y \\
&= (R_t)^{-1} (Q_t^\top Q_t)^{-1} Q_t^\top y \\
&= (R_t)^{-1} v_t^*.
\end{aligned}
$$

Substituting X_t in terms of Q_t and R_t and w_t^* from the expression above we get,

$$
\begin{aligned}
X_t w_t^* &= Q_t R_t (R_t)^{-1} v_t^* \\
&= Q_t v_t^*.
\end{aligned}
$$

(d) (3 points) For each t, show that we can express w_t^* in terms of X_t, Q_t and v_t^* as,

$$w_t^* = (X_t^\top X_t)^{-1} X_t^\top Q_t v_t^*.$$

Hint: It might be useful to start by justifying that y can be expressed as $y = X_t w_t^* + e_t$, where e_t is orthogonal to columns of X_t. Then use the fact that $X_t w_t^* = Q_t v_t^*$.

Solution: Since the least squares problem finds the projection of y onto the column space of X_t, the error term $e_t = y - X_t w_t^*$ is orthogonal to the column space of X_t and thus e_t is orthogonal to columns of X_t. Now,

$$\begin{aligned} w_t &= (X_t^\top X_t)^{-1} X_t^\top y \\ &= (X_t^\top X_t)^{-1} X_t^\top (X_t w_t^* + e_t) \\ &= (X_t^\top X_t)^{-1} X_t^\top (X_t w_t^*) + (X_t^\top X_t)^{-1} (X_t^\top e_t) \\ &= (X_t^\top X_t)^{-1} X_t^\top (X_t w_t^*) \\ &= (X_t^\top X_t)^{-1} X_t^\top Q_t v_t^*. \end{aligned}$$

In the second-last inequality we used the fact that e_t is orthogonal to columns of X_t and in the last inequality we used the fact that $X_t w_t^* = Q_t v_t^*$.

(e) (4 points) Show that v_t^* can be obtained in terms of y and $q_i, i = 1, 2, \ldots, t$ as:

$$v_t^* = \begin{bmatrix} (v_t^*)_1 \\ (v_t^*)_2 \\ \vdots \\ (v_t^*)_t \end{bmatrix},$$

where,

$$(v_t^*)_i = q_i^\top y, \quad i = 1, 2, \ldots, t.$$

For $t = 2, 3, \ldots, T$, use the above result to obtain v_t^* using only v_{t-1}^*, q_t and y.

Solution:

$$\begin{aligned} v_t^* &= (Q_t^\top Q_t)^{-1} Q_t^\top y \\ &= Q_t^\top y. \end{aligned}$$

The last equality uses the fact that Q_t is an orthogonal matrix. Now the ith entry of the column vector $Q_t y$ can be expressed as $q_t^\top y$. For $t = 2, 3, \ldots, T$ we can express v_t^* as:

$$v_t^* = \begin{bmatrix} v_{t-1}^* \\ q_t^\top y \end{bmatrix}.$$

PRINT your name and student ID: ______________________________

6. (9 points) PCA

Let $\{x_1, x_2, \ldots, x_m\}$ be a set of points in $\mathbb{R}^3$. Assume that their empirical mean $\hat{x}$ and empirical covariance matrix Σ are given as:

$$\hat{x} = \frac{1}{m}\sum_{i=1}^{m} x_i = 0,$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m} x_i x_i^\top = \begin{bmatrix} 0.8 & 0.6 & 0 \\ -0.36 & 0.48 & 0.8 \\ -0.48 & 0.64 & -0.6 \end{bmatrix} \begin{bmatrix} 9 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} 0.8 & -0.36 & -0.48 \\ 0.6 & 0.48 & 0.64 \\ 0 & 0.8 & -0.6 \end{bmatrix}.$$

Note that Σ is a symmetric matrix, and it is given with its singular value decomposition, which is equivalent to its eigen-decomposition.

(a) (3 points) Let $w \in \mathbb{R}^3$ be a vector. Let $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m\}$ be projections of the points onto w. Write a vector w that maximizes the variance of the projected points $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m\}$.

Solution: The question asks for the first principal component for the data, which is the singular vector of the matrix Σ that corresponds to its largest singular value. Because the largest and the second largest singular values of Σ are equal, **any vector in**

$$\operatorname{span}\left\{ \begin{bmatrix} 0.8 \\ -0.36 \\ -0.48 \end{bmatrix}, \begin{bmatrix} 0.6 \\ 0.48 \\ 0.64 \end{bmatrix} \right\}$$

maximizes the variance of the projected points.

(b) (6 points) Let $A \in \mathbb{R}^{3\times 3}$ be a matrix such that $A^\top A = \Sigma$, and consider the set $S \subset \mathbb{R}^3$ defined as $S = \{Au : u \in \mathbb{R}^3, \|u\|_2 \le 1\}$. Assume that the points in S are projected onto the hyperplane $\mathcal{H}(w) = \{z \in \mathbb{R}^3 : w^\top z = 0\}$ for some $w \in \mathbb{R}^3$. Find the vector w for which the projection of S onto $\mathcal{H}(w)$ is a circular disc.

Solution:

A is given to be the square-root matrix of Σ. If A is the symmetric square-root of Σ, then the SVD of A is

$$A = \begin{bmatrix} 0.8 & 0.6 & 0 \\ -0.36 & 0.48 & 0.8 \\ -0.48 & 0.64 & -0.6 \end{bmatrix} \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 0.8 & -0.36 & -0.48 \\ 0.6 & 0.48 & 0.64 \\ 0 & 0.8 & -0.6 \end{bmatrix}.$$

In general, however, A may not be symmetric, and its SVD could be

$$A = U \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 0.8 & -0.36 & -0.48 \\ 0.6 & 0.48 & 0.64 \\ 0 & 0.8 & -0.6 \end{bmatrix}$$

with some orthonormal matrix $U \in \mathbb{R}^{3\times 3}$.

Let $\hat{x}$ denote the projection of a point $x \in S$ onto a vector $w \in \mathbb{R}^3$ with unit norm:

$$\hat{x} = w^\top x.$$

Note that there is a vector $u \in \mathbb{R}^3$ such that $Au = x$ and $\|u\|_2 \leq 1$. Then,

$$\begin{aligned}
\hat{x} = w^\top x = w^\top A u &= w^\top U \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 0.8 & -0.36 & -0.48 \\ 0.6 & 0.48 & 0.64 \\ 0 & 0.8 & -0.6 \end{bmatrix} u \\
&= \tilde{w}^\top \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix} \tilde{u}
\end{aligned}$$

where $\tilde{w} = U^\top w$ and $\tilde{u}$ is a unit vector.

If $\tilde{w}$ is $\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$ or $\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$, then $|\hat{x}| \leq 3$. In addition, there exist points with $\hat{x} = 3$ and $\hat{x} = -3$. Consequently, for every vector $\tilde{w} \in \operatorname{span}\left\{\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}\right\}$ with $\|\tilde{w}\|_2 = 1$, the projection of x onto corresponding $w = U\tilde{w}$ changes value between -3 and 3. On the other hand, if we project a point x onto the last column of U, it can take value between -2 and 2. Therefore, we want to project onto the plane spanned by the first two columns of U, and the normal vector for this plane is given by the last column of U. This shows that **the normal vector of the hyperplane $\mathcal{H}(w)$ needs to be the last column of U.**

If A is the symmetric square-root of Σ, then the normal vector of $\mathcal{H}(w)$ needs to be $\begin{bmatrix} 0 \\ 0.8 \\ -0.6 \end{bmatrix}$.

Print your name and student ID: ______________________________

7. (12 points) Errors in the measurement apparatus

This question is about solving the following optimization problem:

$$Q^*_\lambda = \underset{Q}{\operatorname{argmin}} \|Qw - y\|_2^2 + \lambda \|X - Q\|_F^2 . \tag{1}$$

$w \in \mathbb{R}^m, y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times m}, \lambda \in \mathbb{R}, \lambda > 0$ **are all known and constant in Eq.** (1)**.** You do not have to carefully read the rest of the setup to solve the question, but it might give you context.

We perform a series of experiments where we illuminate an object with patterned light and collect the reflected light after it was incident on the object, to help us understand the properties of the object. This is the key idea behind tomography.

The patterned light that is used for illumination is measured and recorded in the "measurement" matrix $X \in \mathbb{R}^{n\times m}$. The ith row of this matrix, $x_i^\top$, represents the illumination measurement for the ith experiment, with a total of n experiments. The intensity of the reflected light is measured as a scalar observation, y_i, for the ith experiment.

Thus we have n pairs $(x_i, y_i) \quad i = 1, 2, \dots, n$, corresponding to the n experiments, where $x_i \in \mathbb{R}^m$ is a vector and $y_i \in \mathbb{R}$ is a scalar.

If our observations are accurate, we expect that (x_i, y_i) should satisfy the equation,

$$x_i^\top w = y_i, \quad i = 1, 2, \dots, n.$$

where vector $w \in \mathbb{R}^m$ represents a **known** image.

To be precise, let $X \in \mathbb{R}^{n\times m}$ denote the matrix with rows as $x_i^\top$,

$$X = \begin{bmatrix} \leftarrow & x_1^\top & \rightarrow \\ \leftarrow & x_2^\top & \rightarrow \\ \leftarrow & \vdots & \rightarrow \\ \leftarrow & x_n^\top & \rightarrow \end{bmatrix}.$$

Let $y \in \mathbb{R}^n$ denote the column vector with entries y_i,

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

We expect the following equation to hold:

$$Xw = y,$$

where w is a **known** vector.

Unfortunately, after completing the experiment we find that our apparatus made errors while measuring the illumination on the image, i.e. there there are small errors in the recorded values of X. However the observations, y_i, are accurate. We would like to recover the true value of the illumination, represented by $X_{true} = Q$. For this we use the two pieces of information that we have:

- $Qw \approx y$
- $Q \approx X$.

We mathematize this by writing the objective function Eq. (1).

PRINT your name and student ID: ______________________________

(a) (3 points) Consider $Z \in \mathbb{R}^{m \times m}$ defined as $Z = ww^\top + \lambda I$. Show that Z is invertible.
Solution:

$$\begin{aligned} x^\top Z x &= x^\top w w^\top x + \lambda x^\top x \\ &= (x^\top w)^2 + \lambda \|x\|_2^2 \\ &> 0, \quad \text{if } x \neq 0. \end{aligned}$$

(b) (8 points) Find an expression for Q_λ^* as defined in Eq. (1) in terms of λ, X, w and y. Assume that we can find the minimum value by setting gradient with respect to Q of the objective function to zero. Justify any algebraic manipulations you make.
Hint: The following identities might be useful:

$$\begin{aligned} \|X\|_F^2 &= \mathrm{Trace}(X^\top X), \\ \nabla_Q \mathrm{Trace}(Q^\top Q B) &= QB^\top + QB, \quad \text{for B square}, \\ \nabla_Q \mathrm{Trace}(AQ) &= A^\top. \end{aligned}$$

Solution:

$$\begin{aligned} f(Q) &= \|Qw - y\|_2^2 + \lambda \|X - Q\|_F^2 \\ &= (Qw - y)^\top (Qw - y) + \lambda \mathrm{Trace}((X - Q)^\top (X - Q)) \\ &= w^\top Q^\top Q w - w^\top Q^\top y - y^\top Q w + y^\top y \\ &\quad + \lambda(\mathrm{Trace}(X^\top X) - \mathrm{Trace}(X^\top Q) - \mathrm{Trace}(Q^\top X) + \mathrm{Trace}(X^\top X)) \\ &= w^\top Q^\top Q w - 2y^\top Q w + \lambda \mathrm{Trace}(Q^\top Q) - 2\lambda \mathrm{Trace}(X^\top Q) + y^\top y + \lambda \mathrm{Trace}(X^\top X) \\ &= \mathrm{Trace}(w^\top Q^\top Q w) - 2\mathrm{Trace}(y^\top Q w) + \lambda \mathrm{Trace}(Q^\top Q) - 2\lambda \mathrm{Trace}(X^\top Q) + y^\top y + \lambda \mathrm{Trace}(X^\top X) \\ &= \mathrm{Trace}(Q^\top Q w w^\top) - 2\mathrm{Trace}(w y^\top Q) + \lambda \mathrm{Trace}(Q^\top Q) - 2\lambda \mathrm{Trace}(X^\top Q) + y^\top y + \lambda \mathrm{Trace}(X^\top X). \end{aligned}$$

Thus,

$$\begin{aligned} \nabla_Q f(Q) &= Qww^\top + Qww^\top - 2yw^\top + \lambda(Q + Q) - 2\lambda X + 0 + 0 \\ &= 2Qww^\top - 2yw^\top + 2\lambda Q - 2\lambda X \\ &= 2(Q(ww^\top + \lambda I) - (yw^\top + \lambda X)). \end{aligned}$$

Setting the gradient to 0, we obtain:

$$\begin{aligned} Q(ww^\top + \lambda I) &= (yw^\top + \lambda X) \\ \implies Q &= (yw^\top + \lambda X)(ww^\top + \lambda I)^{-1}. \end{aligned}$$

Thus,

$$Q_\lambda^* = (yw^\top + \lambda X)(ww^\top + \lambda I)^{-1}.$$

(c) (1 point) Find $\lim_{\lambda \to \infty} Q_\lambda^*$.

Solution:

$$\begin{aligned}
\lim_{\lambda \to \infty} Q_\lambda^* &= \lim_{\lambda \to \infty} (yw^\top + \lambda X)(ww^\top + \lambda I)^{-1} \\
&= \lambda X (\lambda I)^{-1} \\
&= X.
\end{aligned}$$