We want a drone to visit \( m \) positions \( p_1, \ldots, p_m \in \mathbb{R}^2 \) at times \( s_1, \ldots, s_m \in \mathbb{R}_+ \). Define the matrix \( P \in \mathbb{R}^{2 \times m} \) to have columns \( p_j \), where \( j = 1, \ldots, m \). The following Julia code (copy and paste works) defines these:

\[
P = \begin{bmatrix} -1 & -0.5 & 0 & 0.5 & 1 & 0 \\ 2 & 0 & 1 & 0 & 2 & 0 \end{bmatrix}; \quad s = [3; 7; 10; 13; 17; 20];
\]

a) The dynamics of the drone are

\[
\ddot{q} = u,
\]

where \( q : \mathbb{R} \to \mathbb{R}^2 \) is the position of the drone as a function of time, and \( u : \mathbb{R} \to \mathbb{R}^2 \) is an input force as a function of time. Write this as a linear dynamical system of the form

\[
\dot{x} = Ax + Bu, \quad y = Cx,
\]

where \( y : \mathbb{R} \to \mathbb{R}^4 \) is the position and velocity of the drone. In particular, define \( y \) to satisfy \( y_1 = q_1 \), \( y_2 = q_2 \), \( y_3 = \dot{q}_1 \), \( y_4 = \dot{q}_2 \). Is the system stable? Is the system controllable?

b) We will discretize this system with a sample interval \( h > 0 \). Assume the force is piecewise constant on sample intervals. Construct the exact discretization

\[
x_d(k + 1) = A_dx_d(k) + B_du_d(k), \quad \text{and} \quad y_d(k) = C_dx_d(k),
\]

where \( x_d : \mathbb{Z}_+ \to \mathbb{R}^n \) satisfies \( x_d(k) = x(kh) \), and likewise for \( y_d \) and \( u_d \). Give \( A_d \), \( B_d \) and \( C_d \).

c) We operate the drone on the time interval \([0, T]\) where \( T = s_m \). Define \( n = T/h \). Assume the drone starts at the origin with velocity zero. Explain how to choose forces to steer the drone through the desired positions at the desired times, i.e. \( y_{1,2}(s_i) = p_i \), while minimizing

\[
J_1 = \sum_{k=0}^{n-1} \|u_d(k)\|^2_2.
\]

The drone need not be stationary when passing through the points \( p_1, \ldots, p_m \).

d) With \( h = 0.1 \), use your method to compute the optimal \( u_d \). Report the optimal value of \( J_1 \) that you obtained. Plot the components of \( u \) with respect to time. Plot the trajectory of the drone with axes \( q_1 \) and \( q_2 \), so that the plot shows the path followed by the drone. Mark on your plot the points \( p_i \) where the drone has deliveries.

e) Suppose we want to find forces minimizing \( J_1 \) so that the drone is also stationary (velocity zero) when it is at position \( i \), in order to make a clean drop. Explain how to do this. Report the optimal cost. Plot the trajectory and control inputs as before. (Even if you cannot solve this part, you may complete (f) without the condition that the drone is stationary.)
f) Suppose that, in addition to being stationary, we want to avoid the drone jerking too much, in order to avoid damaging its payload. We penalize the discrete jerk. Define

\[ J_2 = \sum_{k=1}^{n-1} \| u_d(k) - u_d(k-1) \|^2. \]

Explain how to find forces to steer the drone to be at the positions \( p_1, \ldots, p_m \) with velocity zero at the desired times \( s_1, \ldots, s_m \), while minimizing \( J_1 + \mu J_2 \), where \( \mu > 0 \) is given.

For \( \mu = 100 \), report \( J_1 \) and \( J_2 \). Plot the trajectory and controls as before.

Make a trade-off curve with \( J_2 \) on the horizontal axis, and \( J_1 \) on the vertical. Briefly interpret the endpoints.

**Solution.**

a) The dynamics of the drone are

\[
\begin{align*}
\dot{x} &= Ax + Bu \\
y &= Cx
\end{align*}
\]

where \( x = \begin{bmatrix} q_1 & q_2 & \dot{q}_1 & \dot{q}_2 \end{bmatrix}^\top \),

\[
A = \begin{bmatrix} 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\
0 & 1 \\
1 & 0 \\
0 & 0 \end{bmatrix}, \quad \text{and} \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \end{bmatrix}.
\]

The system is not stable. The eigenvalues of \( A \) are each 0, and \( \Re(0) = 0 \nless 0 \).

The system is controllable. The rank of \( \begin{bmatrix} B & BA & B^2A & B^3A \end{bmatrix} \) is 4.

b) The discretization is

\[
A_d = e^{hA} = I + Ah + \frac{1}{2}A^2h^2 + \cdots 
\]

\[ = I + Ah \]

\[ = \begin{bmatrix} 1 & 0 & h & 0 \\
0 & 1 & 0 & h \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \end{bmatrix}, \]

since \( A^2 = 0 \) and so \( A^p = 0 \) for all \( p \geq 2 \). Consequently \( e^{sA} = I + sA \) for all \( s \in \mathbb{R} \).

Next,

\[
B_d = \int_0^h e^{\tau A} B d\tau = \int_0^h (I + A\tau) B d\tau = \int_0^h \begin{bmatrix} \tau & 0 \\
0 & \tau \\
1 & 0 \\
0 & 1 \end{bmatrix} d\tau = \begin{bmatrix} h^2/2 & 0 \\
0 & h^2/2 \\
h & 0 \\
0 & h \end{bmatrix}.
\]

Lastly, \( C_d = C \).
c) Recall that we have
\[
\begin{bmatrix}
y(0) \\
y(1) \\
y(2) \\
\vdots \\
y(200)
\end{bmatrix} =
\begin{bmatrix}
D_d & D_d & D_d & D_d & \cdots \\
C_dB_d & C_dB_d & C_dB_d & \cdots \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
C_dA_d^{-1}B_d & C_dA_d^{-2}B & \cdots & C_dB_d & D_d \\
\end{bmatrix}
\begin{bmatrix}
u(0) \\
u(1) \\
u(2) \\
\vdots \\
u(200)
\end{bmatrix}
+ \begin{bmatrix}
C_d \\
C_dA_d \\
\vdots \\
C_dA_d^t \\
\end{bmatrix}
\begin{bmatrix}
x(0) \\
\end{bmatrix}
\]
where \(D_d = 0\) and \(x(0) = 0\) here. Define the block matrix in the above displayed equation \(F \in \mathbb{R}^{4(n+1)\times 2(n+1)}\). Denote the \(i\)th block row by \(F_i\). Define \(Q = \begin{bmatrix}1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}\) and define \(A^{(c)}_{\text{act}} \in \mathbb{R}^{12\times 402}\) by
\[
A^{(c)}_{\text{act}} = \begin{bmatrix}
QF_{10s_1+1} \\
QF_{10s_2+1} \\
\vdots \\
QF_{10s_6+1}
\end{bmatrix}.
\]
(It also works to define a different \(\tilde{C}_d = \begin{bmatrix}1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}\) and define \(A^{(c)}_{\text{act}}\) without \(Q\).)
Define \(\tilde{u} \in \mathbb{R}^{402}\) by \([u(0) \cdots u(200)]^\top\). Define \(\bar{p} \in \mathbb{R}^{12}\) by \(\bar{p} = [p_1 \ p_2 \ \cdots \ p_6]^\top\).
Formulate minimizing the size of the forces while passing through the points as a least norm problem:
\[
\begin{align*}
\text{minimize} & \quad \|\tilde{u}\|^2 \\
\text{subject to} & \quad A^{(c)}_{\text{act}}\tilde{u} = \bar{p}.
\end{align*}
\] (Note: \(J_1\) doesn’t include the square of \(u(200)\), but since this term does not affect \(y(200)\), since \(D_d = 0\), any solution of Problem (??) will have \(u_d(200) = 0\).)
An optimal control sequence \(\tilde{u}^*\) is given by \(\tilde{u}^* = (A^{(c)}_{\text{act}})^\dagger \bar{p}\).
See code below.
The optimal cost is 48.47. Here is the trajectory:
d) Define $A_{\text{act}}^{(d)} \in \mathbb{R}^{24 \times 402}$ by

$$A_{\text{act}}^{(c)} = \begin{bmatrix} F_{10s_1+1} \\ F_{10s_2+1} \\ \vdots \\ F_{10s_6+1} \end{bmatrix}$$

Define $\bar{p}_{\text{stop}} \in \mathbb{R}^{24}$ by $\bar{p}_{\text{stop}} = \begin{bmatrix} p_1 & 0_2 & p_2 & 0_2 & \cdots & p_m & 0_2 \end{bmatrix}^\top$ where $0_2 \in \mathbb{R}^2$ is the 2-vector of zeros. Again, formulate a least norm problem:

$$\begin{align*}
\text{minimize} & \quad \|\bar{u}\|^2 \\
\text{subject to} & \quad A_{\text{act}}^{(d)} \bar{u} = \bar{p}_{\text{stop}}.
\end{align*}$$

(2)

An optimal control sequence $\bar{u}^*$ is given by $\bar{u}^* = (A_{\text{act}}^{(d)})^\dagger \bar{p}_{\text{stop}}$. 
The optimal cost is 71.56. Here is the trajectory:

Here are the forces:

e) Define $W \in \mathbb{R}^{2(n-1) \times 2(n+1)}$ by

$$W = \begin{bmatrix}
-1 & 0 & 1 & 0 & \cdots \\
0 & -1 & 0 & 1 & 0 & \cdots \\
0 & 0 & -1 & 0 & 1 & 0 & \cdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
\cdots & \cdots & \cdots & \cdots & -1 & 0 & 1 & 0 \\
\cdots & \cdots & \cdots & \cdots & 1 & -1 & 0 & 1
\end{bmatrix}$$

We interpret $W$ as the differencing matrix for the controls.
(Note: it was common to define $W \in \mathbb{R}^{2n \times 2(n-1)}$ and to penalize the deviation between $u(200)$ and $u(199)$. We accept such solutions.)

Then $J_2 = ||W \bar{u}||^2$ so

$$J_1 + \mu J_2 = \bar{u}^T I \bar{u} + \mu \bar{u}^T W^T W \bar{u} = \bar{u}^T (I + \mu W^T W) \bar{u}.$$ 

We want to

$$\text{minimize } \bar{u}^T (I + \mu W^T W) \bar{u}$$

subject to $A^{(d)}_{\text{act}} \bar{u} = \bar{p}_{\text{stop}}.$

An optimal control sequence $\bar{u}^*$ is (part of) a solution of the KKT system

$$\begin{bmatrix} I + \mu W^T W & (A^{(d)}_{\text{act}})^T \\ A^{(d)}_{\text{act}} & 0 \end{bmatrix} \begin{bmatrix} \bar{u}^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} 0 \\ \bar{p}_{\text{stop}} \end{bmatrix}.$$

See code below.

The optimal solution for $\mu = 100$ gives $J_1 = 74.85$ and $J_2 = 0.90$.

(For those penalizing the difference of the last two controls, the optimal solution for $\mu = 100$ gives $J_1 = 76.51$ and $J_2 = 0.90$.)

Here is the trajectory (penalizing the last two controls does not visually change things):

[Diagram of trajectory with waypoints]

Here are the forces (penalizing the difference of the last two controls gives $u(200) =$
We solve the KKT conditions for each $\mu$ in a wide range (e.g., $10^{-8}$ to $10^8$). Here is the tradeoff curve (it looks visually similar if penalizing last two controls):

The top left corresponds to $\mu \to 0$, and the bottom right corresponds to $\mu \to \infty$. Roughly speaking, the top left is a low energy control sequence that has more jerk and the bottom right is a low jerk sequence that uses more energy. (We accept many such statements. We also accept flipping the order of the axes: $J_2$ vertical and $J_1$ horizontal.)

Here is Julia code solving this problem.
15.2200. **Properties of symmetric matrices.** In this problem $P$ and $Q$ are symmetric matrices. For each statement below, either give a proof or a specific counterexample.

a) If $P \geq 0$ then $P + Q \geq Q$.

b) If $P \geq Q$ then $-P \leq -Q$.

c) If $P > 0$ then $P^{-1} > 0$.

d) If $P \geq Q > 0$ then $P^{-1} \leq Q^{-1}$.

e) If $P \geq Q$ then $P^2 \geq Q^2$.

**Hint:** you might find it useful for part (d) to prove $Z \geq I$ implies $Z^{-1} \leq I$.

**Solution.**

a) By definition, $A \geq B$ if and only if $A - B \geq 0$. So, if $P \geq 0$, then $P + Q - Q \geq 0$ and therefore $P + Q \geq Q$.

b) If $P \geq Q$ then $P - Q \geq 0$, and by definition $-(P - Q) \leq 0$ or $-P + Q \leq 0$ so finally $-Q \leq -P$.

c) If $P > 0$ then all eigenvalues of $P$ are strictly positive and $P^{-1}$ exists. If $\lambda_1, \ldots, \lambda_n > 0$ are the eigenvalues of $P$ then the eigenvalues of $P^{-1}$ are $1/\lambda_1, \ldots, 1/\lambda_n$. Since $\lambda_i > 0$ then $1/\lambda_i > 0$ so the eigenvalues of $P^{-1}$ are all positive and therefore $P^{-1} > 0$.

d) First we prove the hint, i.e., if $Z \geq I$ then $Z^{-1} \leq I$. Suppose the eigenvalues of $Z \in \mathbb{R}^{n \times n}$ are $\lambda_1, \ldots, \lambda_n$. Then the eigenvalues of $Z - I$ are $\lambda_1 - 1, \ldots, \lambda_n - 1$ because if $v_i$ is the eigenvector associated with $\lambda_i$ then

$$(Z - I)v_i = Zv_i - v_i = \lambda_i v_i - v_i = (\lambda_i - 1)v_i$$

which means that $\lambda_i - 1$ is an eigenvalue of $Z - I$. Since $Z \geq I$ or $Z - I \geq 0$ then all eigenvalues of $Z - I$ are nonnegative or $\lambda_i \geq 1$. The eigenvalues of $Z^{-1}$ are $1/\lambda_1, \ldots, 1/\lambda_n$ and from $\lambda_i \geq 1$ we conclude that $1/\lambda_i \leq 1$ or the eigenvalues of $Z^{-1}$ are all less than or equal to 1. The eigenvalues of $Z^{-1} - I$ are $1/\lambda_1 - 1, \ldots, 1/\lambda_n - 1$ and therefore are all nonpositive. Hence $Z^{-1} - I \leq 0$ or $Z^{-1} \leq I$ and we are done. Now we prove that $P \geq Q > 0$ implies that $P^{-1} \leq Q^{-1}$ or $P^{-1} - Q^{-1} \leq 0$. Suppose that $Q = U\Lambda U^T$ is an eigenvalue decomposition of $Q$. Since $Q > 0$ then $\Lambda > 0$ and therefore $Q^{-1/2} = Q^{-T/2} = U\Lambda^{-1/2}U^T$ exists. By congruence, $P - Q \geq 0$ implies that

$$Q^{-T/2}(P - Q)Q^{-1/2} \geq 0$$

or

$$Q^{-T/2}PQ^{-1/2} - Q^{-T/2}QQ^{-1/2} \geq 0$$

and therefore

$$Q^{-T/2}PQ^{-1/2} - I \geq 0.$$
Now according to the hint (take \( Z = Q^{-T/2}PQ^{-1/2} \)) we have
\[
(Q^{-T/2}PQ^{-1/2})^{-1} - I \leq 0
\]
or
\[
Q^{1/2}P^{-1}Q^{T/2} - I \leq 0.
\]
Again by congruence this implies
\[
Q^{-1/2}(Q^{1/2}P^{-1}Q^{T/2} - I)Q^{-T/2} \leq 0
\]
or
\[
P^{-1} - Q^{-1/2}Q^{-T/2} \leq 0
\]
and finally
\[
P^{-1} - Q^{-1} \leq 0.
\]
e) The statement is false. A simple counterexample is \( P = -1 \) and \( Q = -2 \).

15.2301. Matrix norms and singular values.

a) Eigenvalues and singular values of a symmetric matrix. Suppose \( A \in \mathbb{R}^{n \times n} \) with \( A = A^\top \).
Let \( \lambda_1, \ldots, \lambda_n \) be the eigenvalues of \( A \), and assume that the eigenvalues are ordered such that \( |\lambda_1| \geq \cdots \geq |\lambda_n| \). Let \( \sigma_1, \ldots, \sigma_n \) be the singular values of \( A \); by definition the singular values are ordered such that \( \sigma_1 \geq \cdots \geq \sigma_n \geq 0 \). How are the eigenvalues and singular values of \( A \) related?

b) Suppose \( X \in \mathbb{R}^{n \times n} \). Is \( \sigma_{\max}(X) \geq \max_{1 \leq i \leq n} \sqrt{\sum_{1 \leq j \leq n} |X_{ij}|^2} \)? Prove or give a counterexample.

c) Suppose \( X \in \mathbb{R}^{n \times n} \). Is \( \sigma_{\min}(X) \geq \min_{1 \leq i \leq n} \sqrt{\sum_{1 \leq j \leq n} |X_{ij}|^2} \)? Prove or give a counterexample.

d) Suppose \( X \in \mathbb{R}^{n \times n} \). Is \( \sigma_{\max}(XY) \leq \sigma_{\max}(X)\sigma_{\max}(Y) \)? Prove or give a counterexample.

e) Suppose \( X, Y \in \mathbb{R}^{n \times n} \). Is \( \sigma_{\min}(XY) \geq \sigma_{\min}(X)\sigma_{\min}(Y) \)? Prove or give a counterexample.

f) Suppose \( X, Y \in \mathbb{R}^{n \times n} \). Is \( \sigma_{\min}(X + Y) \geq \sigma_{\min}(X) - \sigma_{\max}(Y) \)? Prove or give a counterexample.

g) Recall that the Frobenius norm of a matrix \( A \in \mathbb{R}^{m \times n} \) is defined to be
\[
\|A\|_F = \sqrt{\text{trace}(A^\top A)}.
\]
Show that
\[
\|A\|_F = \left( \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2 \right)^{1/2}.
\]
Thus, the Frobenius norm is simply the Euclidean norm of a matrix, when we think of the matrix as an element of \( \mathbb{R}^{mn} \). Additionally, note that the Frobenius norm is much easier to compute than the spectral norm.
h) Show that if $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal, then
\[
\|UA\|_F = \|AV\|_F = \|A\|_F.
\]
Thus, multiplication by orthogonal matrices on the left or right does not change the Frobenius norm.

i) Show that
\[
\|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2},
\]
where $\sigma_1, \ldots, \sigma_r$ are the nonzero singular values of $A$. Use this result to deduce that
\[
\sigma_{\text{max}}(A) \leq \|A\|_F \leq \sqrt{r} \sigma_{\text{max}}(A).
\]
In particular, we have that $\|Ax\| \leq \|A\|_F \|x\|$ for all $x \in \mathbb{R}^n$.

Solution.

a) Let the eigenvalue expansion of $A$ be
\[
A = \sum_{i=1}^{n} \lambda_i q_i q_i^T = \sum_{i=1}^{n} |\lambda_i|(\text{sign}(\lambda_i)q_i)q_i^T.
\]
Since $A$ is symmetric, we can assume that $q_1, \ldots, q_n$ form an orthonormal set. This implies that $\text{sign}(\lambda_1)q_1, \ldots, \text{sign}(\lambda_n)q_n$ form an orthonormal set. Therefore,
\[
A = \sum_{i=1}^{n} |\lambda_i|(\text{sign}(\lambda_i)q_i)q_i^T
\]
is the singular-value expansion of $A$. This means that $|\lambda_1|, \ldots, |\lambda_n|$ are the singular values of $A$, $\text{sign}(\lambda_1)q_1, \ldots, \text{sign}(\lambda_n)q_n$ are corresponding left singular vectors, and $q_1, \ldots, q_n$ are corresponding right singular vectors.

b) True. To prove it, let $e_i$ denote the $i$th standard unit vector in $\mathbb{R}^n$. Then
\[
\sigma_{\text{max}}(X) = \sigma_{\text{max}}(X^T) = \max_{\|u\|=1} \|X^T u\| \geq \max_{u \in \{e_1, \ldots, e_n\}} \|X^T u\| = \max_{1 \leq i \leq n} \sqrt{\sum_{1 \leq j \leq n} |X_{ij}|^2},
\]
since $\|X^T e_i\| = \sqrt{\sum_{1 \leq j \leq n} |X_{ij}|^2}$.

c) False. As a counterexample, take $X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, then we have
\[
\min_{1 \leq i \leq n} \sqrt{\sum_{1 \leq j \leq n} |X_{ij}|^2} = \sqrt{2} > \sigma_{\text{min}}(X) = 0.
\]

d) True.
\[
\sigma_{\text{max}}(XY) = \max_{\|x\|=1} \|X(Yx)\| \leq \sigma_{\text{max}}(X) \max_{\|x\|=1} \|Yx\| = \sigma_{\text{max}}(X) \sigma_{\text{max}}(Y).
\]
e) True.

\[ \sigma_{\min}(XY) = \min_{\|x\|=1} \|X(Yx)\| \geq \sigma_{\min}(X) \min_{\|x\|=1} \|Yx\| = \sigma_{\min}(X)\sigma_{\min}(Y). \]

f) True. First we start with a little trick:

\[ \|Xx\| = \|(X + Y)x - Yx\| \leq \|(X + Y)x\| + \|Yx\|. \]

Rearranging this inequality yields

\[ \|(X + Y)x\| \geq \|Xx\| - \|Yx\| \geq \sigma_{\min}(X)\|x\| - \sigma_{\max}(Y)\|x\|. \]

If \( x \neq 0 \), then

\[ \frac{\|(X + Y)x\|}{\|x\|} \geq \sigma_{\min}(X) - \sigma_{\max}(Y). \]

Since this is true for all \( x \neq 0 \), it is true for the minimizing \( x \):

\[ \min_{x \neq 0} \frac{\|(X + Y)x\|}{\|x\|} = \sigma_{\min}(X + Y) \geq \sigma_{\min}(X) - \sigma_{\max}(Y), \]

and we are done.

g) Expanding the definition of the Frobenius norm, we have that

\[ \|A\|_F = \sqrt{\text{trace}(A^T A)} = \sqrt{\sum_{j=1}^{n} (A^T A)_{jj}} = \sqrt{\sum_{j=1}^{n} \sum_{i=1}^{m} (A^T)_{ij} a_{ij}} = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2}. \]

h) For any orthogonal matrix \( U \in \mathbb{R}^{m \times m} \), we have that

\[ \|UA\|_F = \sqrt{\text{trace}((UA)^T (UA))} = \sqrt{\text{trace}(A^T U^T UA)} = \sqrt{\text{trace}(A^T A)} = \|A\|_F, \]

where \( U^T U = I \) because \( U \) is orthogonal. Similarly, for any orthogonal matrix \( V \in \mathbb{R}^{n \times n} \), we have that

\[ \|AV\|_F = \sqrt{\text{trace}((AV)^T (AV))} = \sqrt{\text{trace}(V^T A^T AV)} = \sqrt{\text{trace}(A^T AV V^T)} = \sqrt{\text{trace}(A^T A)} = \|A\|_F, \]

where we have used the fact that \( \text{trace}(XY) = \text{trace}(YX) \) whenever \( XY \) and \( YX \) are both defined, and \( VV^T = I \) because \( V \) is orthogonal.
Let $A = U \Sigma V^T$ be the (full) singular value decomposition of $A$. Then, we have that

$$\|A\|_F = \|U \Sigma V^T\|_F = \|\Sigma V^T\|_F = \|\Sigma\|_F$$

because $U$ and $V$ are orthogonal. Since $\Sigma$ is a diagonal matrix whose nonzero diagonal elements are $\sigma_1, \cdots, \sigma_r$, we have that

$$\|A\|_F = \|\Sigma\|_F = \sqrt{\sum_{i=1}^r \sigma_i^2}$$

Since $\sigma_1 \geq \cdots \geq \sigma_r > 0$, this expression implies that

$$\sigma_{\text{max}}(A) \leq \|A\|_F \leq \sqrt{r} \sigma_{\text{max}}(A).$$

In particular, we have that

$$\|Ax\| \leq \|A\| \|x\| \leq \|A\|_F \|x\|$$

for all $x \in \mathbb{R}^n$.

**16.2980. Smoothing.** We have a discrete-time signal given by $x \in \mathbb{R}^n$. We get to measure $y \in \mathbb{R}^n$, given by

$$y_i = \sum_{k=-h}^h c_k x_{i+k} + w_i \quad \text{for } i = 1, \ldots, n$$

where $w_i$ is noise. Here we use the convention that $x_i = 0$ for $i < 1$ or $i > n$. That is, $y$ is $c$ convolved with $x$ plus noise. In applications, very often the effect of convolution with $c$ is to smooth or blur $x$, and we would like to undo this.

The file `regl_data.json` contains $c$, $w$ and $x$.

a) In Julia, construct the $n \times n$ matrix such that $y = Ax + w$. Plot the singular values $\sigma_k$ against $k$.

b) Plot the first 6 right singular vectors of $A$ (i.e. plot $V_{ij}$ against $i$ for $j = 1, \ldots, 6$.) Explain what you see.

c) Find and plot the least-squares estimate of $x$ given $y_{\text{meas}}$, computing $y_{\text{meas}}$ using $c$, $x$ and $w$ given in `regl_data.json`. Explain what happens.

d) Many of the singular values of $A$ are very small; this means that the measurement in the directions of the corresponding right singular vectors is being swamped by the noise.

If we believe these components are small, we can remove them from our estimate of $x$ altogether by truncating the SVD of $A$ and using the truncated SVD to compute the estimate. This is called the truncated SVD regularization of least-squares.
Suppose we decided only to keep the first $r$ components. Then truncate by letting $	ilde{V}$ and $	ilde{U}$ be the first $r$ columns of $V$ and $U$, and letting $	ilde{\Sigma}$ be the top-left $r \times r$ submatrix of $\Sigma$. Then we can construct an estimator that ignores the noise components by

$$A_{est} = \tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^T$$

and set

$$x_{est} = A_{est}y_{meas}$$

For values of $r$ in 5, 10, 15, 30, 50, compute and plot the corresponding estimates of $x$. Explain what you see.

e) For each $r$ between 1 and 35, compute the norm of the error

$$\|x - x_{est}\|$$

Plot this against $r$. Explain what you see.

f) Pick the ‘best’ $r$ and plot the corresponding estimate.

g) Another approach is to use Tychonov regularization. Find and plot the vector $x_{reg} \in \mathbb{R}^n$ that minimizes the function

$$\|Ax - y\|^2 + \mu\|x\|^2,$$

where $\mu > 0$ is the regularization parameter. Pick a value of $\mu$ that gives a good estimate, in your opinion.

h) The regularized solution is a linear function of $y$, so it can be expressed as $x_{reg} = By$ where $B \in \mathbb{R}^{n \times n}$. Express the SVD of $B$ in terms of the SVD of $A$. To be more specific, let

$$B = \sum_{i=1}^{n} \tilde{\sigma}_i \tilde{u}_i \tilde{v}_i^T$$

denote the SVD of $B$. Express $\tilde{\sigma}_i$, $\tilde{u}_i$, $\tilde{v}_i$, for $i = 1, \ldots, n$, in terms of $\sigma_i$, $u_i$, $v_i$, $i = 1, \ldots, n$ (and, possibly, $\mu$). Recall the convention that $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_n$.

i) Find the norm of $B$. Give your answer in terms of the SVD of $A$ (and $\mu$).

j) Find the worst-case relative inversion error, defined as

$$\max_{y \neq 0} \frac{\|ABy - y\|}{\|y\|}.$$

give your answer in terms of the SVD of $A$ (and $\mu$).

**Solution.**

a) $A$ can be constructed from $c$ easily. The singular values of $A$ is as follows.
Note that they decay very fast, and some are extremely small. In particular, this shows that this estimation problem will be extremely sensitive to noise and numerical errors in the measurement, since the condition number is large.

b) The first six singular vectors are below.

The matrix $A$ is a smoother, or low-pass filter. So we should expect that it has a gain that
depends on the ‘frequency’ of the input. The plot show that this intuition is correct; the input is broken down into different frequency components by the matrix of right singular vectors $V$, each is scaled by the corresponding singular value, and then the output is constructed. The left singular vectors look similar. Since its a low-pass filter, the larger singular values correspond to low frequencies.

c) Below are the plots of $x$, $y_{\text{meas}}$, and $x_{ls}$.

Our estimate is extremely badly corrupted by the noise, because of the high condition number of $A$; the estimation ellipsoid is long and thin, and we have extremely poor estimates of the components of $x$ corresponding to small singular values of $A$. These are the high frequency components. They are multiplied by the small singular values, and so are swamped by the noise.

d) The plots below show the estimates obtained with different values of $r$. 
The regularization ignores components of the frequency corresponding to right singular vectors $v_k$ when $k > r$. These are the higher-frequency and more noise components. Removing them is equivalent to making the assumption that that component is actually zero, rather than using the measured data.

The estimate is poor for $r$ very small, because much of the signal is being ignored by the regularization. It is also poor at large $r$, when very noisy measurements of the high-frequency components are used.

e) The error is plotted below.
The graph shows the error phenomena described in part (d). Using too few singular values means we lose too much information, using too many means we use information which is very badly corrupted by noise.

f) From the above graph, we find the number of singular values which results in the minimum error is 28.

The corresponding estimate is below.

g) One of good values of $\mu$ can be chosen as $\mu = 0.05$ by computing errors for various values of $\mu$. Below is the estimate corresponding to $\mu = 0.05$. 
h) The regularized least-squares solution is given by \( x_{\text{rls}}(\mu) = (A^T A + \mu I)^{-1} A^T y \), and thus

\[
B = (A^T A + \mu I)^{-1} A^T
\]

\[
= \left( (U\Sigma V^T)^T (U\Sigma V^T + \mu I) \right)^{-1} (U\Sigma V^T)^T
\]

\[
= (V\Sigma U^T U\Sigma V^T + \mu I)^{-1} V\Sigma U^T
\]

\[
= (V (\Sigma^2 + \mu I) V^T)^{-1} V\Sigma U^T
\]

\[
= (V (\Sigma^2 + \mu I)^{-1} V^T) V\Sigma U^T
\]

\[
= V (\Sigma^2 + \mu I)^{-1} \Sigma U^T
\]

\[
= V \text{diag} \left( \frac{\sigma_i}{\sigma_i^2 + \mu} \right) U^T.
\]

This is almost the SVD of \( B \), except for one detail: the numbers \( \frac{\sigma_i}{\sigma_i^2 + \mu} \) aren’t necessarily ordered from largest to smallest. Thus we have

\[
\tilde{\sigma}_i = \frac{\sigma_i}{\sigma_i^2 + \mu} \quad \tilde{u}_i = v_i, \quad \tilde{v}_i = u_i,
\]

where the notation \( x_{[i]} \) means the \( i \)th largest element of \( x \). (We accepted all sorts of descriptions of this!) One common misconception was that the numbers \( \frac{\sigma_i}{\sigma_i^2 + \mu} \) were simply in reverse order, so all that had to be done was to reverse the ordering. That isn’t true; just sketch the function \( \sigma/(\sigma^2 + \mu) \) as a function of \( \mu \) to see that it is not always decreasing. (It increases first, then decreases.)
i) The norm of $B$ is its largest singular value, i.e.,
\[ \| B \| = \max_i \frac{\sigma_i}{\sigma_i^2 + \mu}. \]

j) The worst-case relative inversion error is the matrix norm of $AB - I$:
\[
AB - I = U \Sigma V^T (\Sigma + \mu \Sigma^{-1})^{-1} U^T - I \\
= U \Sigma (\Sigma + \mu \Sigma^{-1})^{-1} U^T - I \\
= U \left( (I + \mu \Sigma^{-2})^{-1} - I \right) U^T \\
= U \text{diag} \left( \frac{1}{1 + \mu/\sigma_i^2} - 1 \right) U^T \\
= -U \text{diag} \left( \frac{\mu}{\sigma_i^2 + \mu} \right) U^T
\]

This is the SVD of $AB - I$ (to within reordering). Its largest singular value, i.e., the norm of $AB - I$, is given by
\[ \| AB - I \| = \frac{\mu}{\sigma_n^2 + \mu}. \]

Note that we had to absorb the negative sign in the lefthand orthogonal matrix; one common error was to keep the negative sign in the norm. Obviously that couldn’t be right because norms are always nonnegative!

18.2890. **Linear dynamical systems for portfolio management.** We consider a portfolio of $n$ financial assets (like stocks) and cash, which we manage over $T$ time steps of unit length (e.g. one month). We call $x_t \in \mathbb{R}^{n+1}$ for $t = 1, \ldots, T$ our state vector. The first $n$ elements are our positions in each of the assets, in dollars, and the last element is the dollar amount of cash we hold. Every element of $x$ can be either positive (for long positions) and negative (for short or borrowing). For $t = 1, \ldots, T - 1$, the transition from $x_t$ to $x_{t+1}$ is composed of two steps.

- First, the portfolio positions change value because of market returns. Let $\mu \in \mathbb{R}^n_{++}$ be the vector of returns, where $\mathbb{R}^n_{++}$ is the set of all vectors of length $n$ with strictly positive entries. We define the post-return portfolio $\tilde{x}_t$ to be
\[
(\tilde{x}_t)_i = \begin{cases} 
\mu_i(x_t)_i & i = 1, \ldots, n, \\
(x_t)_i & i = n + 1.
\end{cases}
\]

(Intuitively, cash is unchanged, and the asset positions are multiplied by the corresponding element of the vector of returns.) For simplicity we assume that the vector of returns does not change in time.

- Then we trade. We can exchange any amount of cash for the corresponding amount of any of the assets. Note that the only valid trades are cash for asset. If you wish to trade some amount of an asset with the same amount of another asset, you have to perform two trades: trade the first asset with cash, and then trade cash with the second
asset. (Think carefully about this definition of trade when you formulate the transaction costs.) For example, if we buy \(c > 0\) dollars of the first asset and sell \(d > 0\) dollars of the second asset the state evolves as

\[
x_{t+1} = \tilde{x}_t + \begin{bmatrix} c \\ -d \\ 0 \\ \vdots \\ 0 \\ -(c - d) \end{bmatrix}.
\]

Finally, we define the portfolio value \(v_t \in \mathbb{R}\) for \(t = 1, \ldots, T\) to be

\[
v_t = 1^T x_t.
\]

a) Formulate the problem as a linear dynamical system of the form

\[
x_{t+1} = Ax_t + Bu_t, \quad t = 1, \ldots, T - 1.
\]

The control vector \(u_t\) should have dimension \(n\).

b) Assume that our trades incur quadratic transaction costs with parameter \(\rho > 0\). For example, if at time \(t\) we buy \(c > 0\) dollars of the first asset, and we sell \(d > 0\) dollars of the second asset (the example above), then the transaction costs for the transition \(x_t\) to \(x_{t+1}\) are

\[
\rho(c^2 + d^2).
\]

(Be careful, they are not \(\rho(c + d)^2\).) Explain how to solve the problem of maximizing the final value of the portfolio \(v_T\) minus the total transaction costs. (The sequence of controls \(u_1, \ldots, u_{T-1}\) that achieves the maximum should be a function of \(A, B,\) and \(\rho\).) Use methods from EE263.

c) Apply your method to the following data.

\[
T = 12; \\
x_{-1} = [1000, 1000, 0, 1000, 0, 0]; \\
mu = [1.001, 1.003, 1.004, 1.006, 1.007]; \\
rho = 0.0001;
\]

What is the final value \(v_T\)? What are the total transaction costs? Plot the trajectories of the portfolio positions \(x_t\). (On the same plot you should draw \(n + 1\) lines, one for each of the assets and cash, with time on the \(x\)-axis.)

d) Now assume that we aim to liquidate an initial portfolio, which means that at time \(T\) we want to have zero positions in any of the \(n\) assets and only hold cash. We thus impose the constraint \((x_T)_i = 0\), for \(i = 1, \ldots, n\). Explain how to solve the problem of maximizing the final portfolio value (in this case, all cash) minus the transaction costs with this additional constraint. Use methods from EE263.
e) Apply your method to the data given above. What is the final value $v_T$? What are the total transaction costs? Plot the trajectories of the portfolio positions $x_t$. (On the same plot you should draw $n+1$ lines, one for each of the assets and cash, with time on the $x$-axis.)

**Solution.**

a) We define the vector

$$
\tilde{\mu} = \begin{bmatrix} \mu \\ 1 \end{bmatrix},
$$

then

$$
\tilde{x}_t = \text{diag}(\tilde{\mu})x_t,
$$

and we call

$$
A = \text{diag}(\tilde{\mu}).
$$

We represent the trades by a vector $u_t \in \mathbb{R}^n$ for $t = 1, \ldots, T-1$. Each element is the dollar amount of each of the assets that we exchange for cash. Then the matrix $B$ is

$$
B = \begin{bmatrix}
1 & 0 & \cdots & 0 \\
0 & 1 & 0 & & \\
\vdots & & \ddots & \vdots \\
0 & & 1 & \cdots & -1
\end{bmatrix}.
$$

We thus have that

$$
x_{t+1} = Ax_t + Bu_t, \quad t = 1, \ldots, T - 1.
$$

b) The final state is given by

$$
x_T = A^{T-1}x_1 + \sum_{t=1}^{T-1} A^{T-t-1}B u_t
$$

or equivalently

$$
x_T = A^{T-1}x_1 + B^T U
$$

with

$$
B^T = \begin{bmatrix} A^{T-2}B & A^{T-3}B & \cdots & B \end{bmatrix}, \quad U = \begin{bmatrix} u_1 \\
u_2 \\
\vdots \\
u_{T-1}
\end{bmatrix}.
$$

Thus, the final value of the portfolio is given by

$$
v_T = 1^T A^{T-1}x_1 + 1^T B^T U.
$$

The quadratic transaction costs are

$$
\rho ||U||^2.
$$
The maximization problem is

$$\text{maximize} \quad v_T - \rho \|U\|^2$$

since the first term of $v_T$ is a constant, this is equivalent to

$$\text{maximize} \quad 1^T B^T U - \rho \|U\|^2$$

the solution is (by the first order condition, since the function is strictly concave)

$$U = \frac{1}{2\rho} B 1.$$

c) The final value and total transaction costs are

$$v_T = 3335.06, \quad \rho \|U\|^2 = 111.24, \quad v_T - \rho \|U\|^2 = 3223.81.$$  

The yellow line is cash, the others are the asset positions.

d) The optimization problem is

$$\text{maximize} \quad v_T - \rho \|U\|^2$$

s.t. \quad $(x_T)_i = 0$ \quad $i = 1, \ldots, n.$

Let $\bar{x}_t \in \mathbb{R}^n$ be the first $n$ elements of each vector $x_t$. Then we have

$$\bar{x}_{t+1} = \text{diag}(\mu) \bar{x}_t + I u_t,$$

$$\bar{x}_T = \text{diag}(\mu)^{T-1} \bar{x}_1 + \sum_{t=1}^{T-1} \text{diag}(\mu)^{T-t-1} u_t$$
or
\[ \bar{x}_T = \text{diag}(\mu)^{T-1} \bar{x}_1 + C^T U \]
with
\[ C = \begin{bmatrix} \text{diag}(\mu)^{T-2} \\ \text{diag}(\mu)^{T-3} \\ \vdots \\ I \end{bmatrix}. \]

Thus the optimization problem is
\[ \begin{align*}
\text{maximize} & \quad v_T - \rho \|U\|_2^2 \\
\text{s.t.} & \quad C^T U = -\text{diag}(\mu)^{T-1} \bar{x}_1.
\end{align*} \]

We could solve this problem by introducing a Lagrange multiplier, but instead we transform the problem by subtracting the optimal solution to (b)
\[ U' = U - \frac{1}{2\rho} B1 \]

The objective function is
\[ 1^T B^T (U' + \frac{1}{2\rho} B1) - \rho \|U' + \frac{1}{2\rho} B1\|_2^2. \]

Expanding the squared norm and ignoring constant terms we obtain that the problem is equivalent to
\[ \begin{align*}
\text{maximize} & \quad -\rho \|U'\|_2^2 \\
\text{s.t.} & \quad C^T U' = y
\end{align*} \]

where
\[ y = -\text{diag}(\mu)^{T-1} \bar{x}_1 - \frac{1}{2\rho} C^T B1. \]

Note that \( C \) is full rank. The solution (least-norm underdetermined system) is
\[ U' = C(C^T C)^{-1} y \]
and
\[ U = C(C^T C)^{-1} y + \frac{1}{2\rho} B1. \]

e) The final value and total transaction costs are
\[ v_T = 3120.65, \quad \rho \|U\|_2^2 = 58.72, \quad v_T - \rho \|U\|_2^2 = 3061.94. \]
The yellow line is cash, the others are the asset positions.