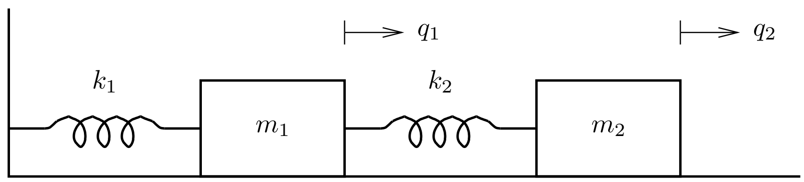
Homework 4 Solutions

EE 263 Stanford University

Summer 2017

1. **Controlling a system using the initial conditions.** Consider the mechanical system shown below:



Here q_i give the displacements of the masses, m_i are the values of the masses, and k_i are the spring stiffnesses, respectively. The dynamics of this system are

$$\dot{x} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{k_1+k_2}{m_1} & \frac{k_2}{m_1} & 0 & 0 \\ \frac{k_2}{m_2} & -\frac{k_2}{m_2} & 0 & 0 \end{bmatrix} x$$

where the state is given by

$$x = \begin{bmatrix} q_1 \\ q_2 \\ \dot{q}_1 \\ \dot{q}_2 \end{bmatrix}.$$

Immediately before $t = 0$, you are able to apply a strong impulsive force α_i to mass i, which results in initial condition

$$x(0) = \begin{bmatrix} 0 \\ 0 \\ \alpha_1/m_1 \\ \alpha_2/m_2 \end{bmatrix}.$$

(i.e., each mass starts with zero position and a velocity determined by the impulsive forces.) This problem concerns selection of the impulsive forces α_1 and α_2. For parts a–c below, the parameter values are

$$m_1 = m_2 = 1, \quad k_1 = k_2 = 1.$$

Consider the following specifications:

a) $q_2(10) = 2$

b) $q_1(10) = 1$, $q_2(10) = 2$

c) $q_1(10) = 1$, $q_2(10) = 2$, $\dot{q}_1(10) = 0$, $\dot{q}_2(10) = 0$

d) $q_2(10) = 2$ when the parameters have the values used above (*i.e.*, $m_1 = m_2 = 1$, $k_1 = k_2 = 1$), and *also*, $q_2(10) = 2$ when the parameters have the values $m_1 = 1$, $m_2 = 1.3$, $k_1 = k_2 = 1$.

Determine whether each of these specifications is feasible or not (*i.e.*, whether there exist $\alpha_1, \alpha_2 \in \mathbb{R}$ that make the specification hold). If the specification is feasible, find the particular α_1, α_2 that satisfy the specification and minimize $\alpha_1^2 + \alpha_2^2$. If the specification is infeasible, find the particular α_1, α_2 that come closest, in a least-squares sense, to satisfying the specification. (For example, if you cannot find α_1, α_2 that satisfy $q_1(10) = 1$, $q_2(10) = 2$, then find α_i that minimize $(q_1(10) - 1)^2 + (q_2(10) - 2)^2$.) Be sure to be very clear about which alternative holds for each specification.

Solution. The dynamics of the system is given by $\dot{x} = Ax$ where (for $m_1 = m_2 = 1$ and $k_1 = k_2 = 1$)

```
>> A=[0 0 1 0;0 0 0 1;-2 1 0 0;1 -1 0 0]
A =
0      0      1      0
0      0      0      1
-2      1      0      0
1     -1      0      0
>>
```

$x(10)$ is related to $x(0)$ through $x(10) = \Phi x(0)$ where $\Phi = e^{10A}$ is

```
>> phi=expm(10*A)
phi =
-0.3694    0.8431   -0.2494    0.0515
0.8431    0.4737    0.0515   -0.1979
0.5503   -0.3009   -0.3694    0.8431
-0.3009    0.2494    0.8431    0.4737
>>
```

The values of $q_1(0)$ and $q_2(0)$ are taken as zero and therefore

$$x(10) = \begin{bmatrix} -0.2494 & 0.0515 \\ 0.0515 & -0.1979 \\ -0.3694 & 0.8431 \\ 0.8431 & 0.4737 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix},$$

in other words, the first two columns of Φ are irrelevant. Take

$$\Psi = \begin{bmatrix} -0.2494 & 0.0515 \\ 0.0515 & -0.1979 \\ -0.3694 & 0.8431 \\ 0.8431 & 0.4737 \end{bmatrix}$$

so that $x(10) = \Psi\alpha$ where $\alpha = [\alpha_1\ \alpha_2]^\mathsf{T}$. We define the matrix `psi` in the matlab environment:

```
>> psi=phi(:,3:4)
psi =
-0.2494    0.0515
0.0515   -0.1979
-0.3694    0.8431
0.8431    0.4737
>>
```

a) We have $q_2(10) = e_2^\mathsf{T} x(10)$ where e_2 is the 2nd unit vector in $\mathbb{R}^4$. Therefore, α_1 and α_2 should satisfy the linear equation $2 = e_2^\mathsf{T}\Psi\alpha$. There are two variables α_1 and α_2 but only one equation. Therefore the choice of α_1 and α_2 is not unique and we pick the minimum norm solution. In matlab:

   ```
   >> alpha=pinv(psi(2,:))*2
   alpha =
   2.4614
   -9.4647
   >>
   ```

   So here we can meet the spec, and even have one extra degree of fredom, which we use to minimize the norm of α.

b) The requirements in this part are more stringent than in the previous one. Here we have the additional requirement $q_1(10) = 1$ and we get two linear equations in two unknowns, *i.e.*, $[e_1^\mathsf{T}\ e_2^\mathsf{T}]x(10) = \Psi\alpha$ which has a unique solution sincethe resulting matrix is invertible. In matlab:

   ```
   >> alpha=psi(1:2,:)\[1;2]
   alpha =
   -6.4408
   -11.7798
   >>
   ```

   In this case there is only one α that meets the specs; there is no extra freedom.

c) Here we have two more requirements $\dot{q}_1(10) = 0$ and $\dot{q}_2(10) = 0$ and therefore we get an overdetermined system of linear equations (four equations in two unknowns) $x(10) = \Psi\alpha$. We solve for α in a least-squares sense, which will show us if we are lucky and can find an exact solution. Using matlab:

   ```
   >> alpha=psi\[1;2;0;0]
   alpha =
   -0.1361
   -0.3435
   >>
   ```

It can be checked that this α does not meet the spec, but it comes closest in the sense that $\|x(10) - [1\ 2\ 0\ 0]^\mathsf{T}\|$ is minimized.

d) With the new set of parameters $m_1 = 1$, $m_2 = 1.3$, $k_1 = k_2 = 1$ we get a new system $\dot{x} = \tilde{A}x$ with

```
>> tilde_A=[0 0 1 0;0 0 0 1;-2 1 0 0;1/1.3 -1/1.3 0 0]
tilde_A =
0          0     1.0000          0
0          0          0     1.0000
-2.0000     1.0000          0          0
0.7692    -0.7692          0          0
>>
```

Similarly we define $\tilde{\Phi}$ and $\tilde{\Psi}$:

```
>> tilde_phi=expm(10*tilde_A)
tilde_phi =
-0.6221     0.8272   -0.2229   -0.5395
0.6363     0.3961   -0.4150   -0.8869
0.0308     0.1921   -0.6221    0.8272
0.1478     0.2672    0.6363    0.3961
>> tilde_psi=tilde_phi(:,3:4)
tilde_psi =
-0.2229   -0.5395
-0.4150   -0.8869
-0.6221    0.8272
0.6363    0.3961
>>
```

The requirements $q_2(10) = 2$ for $m_1 = m_2 = 1$, $k_1 = k_2 = 1$ and $q_2(10) = 2$ for $m_1 = 1$, $m_2 = 1.3$, $k_1 = k_2 = 1$ can be written as

$$2 = e_2^\mathsf{T}\Psi \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}, \quad 2 = e_2^\mathsf{T}\tilde{\Psi} \begin{bmatrix} \alpha_1 \\ \alpha_2/1.3 \end{bmatrix} = e_2^\mathsf{T}\tilde{\Psi} \begin{bmatrix} 1 & 0 \\ 0 & 1/1.3 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix},$$

or in matrix form

$$\begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} e_2^\mathsf{T}\Psi \\ e_2^\mathsf{T}\tilde{\Psi} \begin{bmatrix} 1 & 0 \\ 0 & 1/1.3 \end{bmatrix} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}.$$

Using matlab:

```
>> alpha=[psi(2,:);tilde_psi(2,:)*diag([1;1/1.3])]\[2;2]
alpha =
8.2608
-7.9566
>>
```

Here we can meet the spec, but there is only one solution.

2. **Analysis of a power control algorithm.** In this problem we consider again the power control method described in homework problem 2.1 Please refer to this problem for the setup and background. In that problem, you expressed the power control method as a discrete-time linear dynamical system, and simulated it for a specific set of parameters, with several values of initial power levels, and two target SINRs. You found that for the target SINR value $\gamma = 3$, the powers converged to values for which each SINR exceeded γ, no matter what the initial power was, whereas for the larger target SINR value $\gamma = 5$, the powers appeared to diverge, and the SINRs did not appear to converge. You are going to analyze this, now that you know alot more about linear systems.

   a) *Explain the simulations.* Explain your simulation results from the problem 1(b) for the given values of G, α, σ, and the two SINR threshold levels $\gamma = 3$ and $\gamma = 5$.

   b) *Critical SINR threshold level.* Let us consider fixed values of G, α, and σ. It turns out that the power control algorithm works provided the SINR threshold γ is less than some critical value γ_{crit} (which might depend on G, α, σ), and doesn't work for $\gamma > \gamma_{\text{crit}}$. ('Works' means that no matter what the initial powers are, they converge to values for which each SINR exceeds γ.) Find an expression for γ_{crit} in terms of $G \in \mathbb{R}^{n\times n}$, α, and σ. Give the simplest expression you can. Of course you must explain how you came up with your expression.

Solution.

a) In the homework we found that the powers propagate according to a linear system. The power update rule for a single transmitter can be found by manipulating the definitions given in the problem.

$$\begin{aligned} p_i(t+1) &= \frac{\alpha\gamma p_i(t)}{S_i(t)} = \frac{\alpha\gamma p_i(t) q_i(t)}{s_i(t)} = \frac{\alpha\gamma p_i(t)\left[\sigma + \sum_{j\neq i} G_{ij}p_j(t)\right]}{G_{ii}p_i(t)} \\ &= \frac{\alpha\gamma\left[\sigma + \sum_{j\neq i} G_{ij}p_j(t)\right]}{G_{ii}} \end{aligned}$$

In matrix form the equations represent a linear dynamical system with constant input, $p(t+1) = Ap(t) + b$.

$$\underbrace{\begin{bmatrix} p_1(t+1) \\ p_2(t+1) \\ p_3(t+1) \\ \vdots \\ p_n(t+1) \end{bmatrix}}_{p(t+1)} = \alpha\gamma \underbrace{\begin{bmatrix} 0 & \frac{G_{12}}{G_{11}} & \frac{G_{13}}{G_{11}} & \cdots & \frac{G_{1n}}{G_{11}} \\ \frac{G_{21}}{G_{22}} & 0 & \frac{G_{23}}{G_{22}} & \cdots & \frac{G_{2n}}{G_{22}} \\ \frac{G_{31}}{G_{33}} & \frac{G_{32}}{G_{33}} & 0 & \cdots & \frac{G_{3n}}{G_{33}} \\ \vdots & & & \ddots & \\ \frac{G_{n1}}{G_{nn}} & \frac{G_{n2}}{G_{nn}} & \frac{G_{n3}}{G_{nn}} & \cdots & 0 \end{bmatrix}}_{P} \underbrace{\begin{bmatrix} p_1(t) \\ p_2(t) \\ p_3(t) \\ \vdots \\ p_n(t) \end{bmatrix}}_{p(t)} + \underbrace{\begin{bmatrix} \frac{\alpha\gamma\sigma}{G_{11}} \\ \frac{\alpha\gamma\sigma}{G_{22}} \\ \frac{\alpha\gamma\sigma}{G_{33}} \\ \vdots \\ \frac{\alpha\gamma\sigma}{G_{nn}} \end{bmatrix}}_{b}.$$

where $A = \alpha\gamma P$. This is a discrete LDS, and is stable if and only if $|\lambda_i| < 1$ for all $i = 1, \ldots, n$, where λ_i are the eigenvalues of A. When $\gamma = 3$ the eigenvalues of A are 0.6085, -0.3600, and -0.2485, so the system is stable; for all initial conditions, the powers converge to their equilibrium values.

Also, the SINR at each receiver i, given by S_i, converges to the same constant value $\alpha\gamma$, which is enough for a successful signal reception. This can be shown by observing that at equilibrium $p_i(t+1) = p_i(t) = \bar{p}_i$, and the power update equation gives

$$\bar{p}_i = \bar{p}_i(\alpha\gamma/S_i(t)).$$

After cancellation, we obtain the constant value for each SINR, $S_i = \alpha\gamma$.

When $\gamma = 5$, the eigenvalues of A are 1.0141, -0.6000, and -0.4141. This system is unstable because of the first eigenvalue, so this means there are initial conditions from which the powers diverge.

```
>> inv(v)*b
-0.0670
-0.0000
-0.0182
```

b) The critical SINR threshold level is a function of dominant system eigenvalue. We will assume that matrix P is diagonizable and that its eigenvalues are ordered by their magnitude when forming Λ matrix. Using the property that scaling of any matrix scales its eigenvalues by the same constant, we can derive:

$$\begin{aligned} A &= \alpha\gamma P = \alpha\gamma T\Lambda T^{-1} \\ &= T\,\mathrm{diag}(\alpha\gamma\lambda_1, \ldots, \alpha\gamma\lambda_n)T^{-1} \end{aligned}$$

For a marginally stable system we need to have $|\alpha\gamma\lambda_1| \leq 1$. Manipulating equation $\alpha\gamma_{\text{crit}}|\lambda_1| = 1$, we obtain the critical SINR threshold level,

$$\gamma_{\text{crit}} = \frac{1}{\alpha|\lambda_1|}.$$

3. Harmonic oscillator. The system $\dot{x} = \begin{bmatrix} 0 & \omega \\ -\omega & 0 \end{bmatrix} x$ is called a *harmonic oscillator.*

a) Find the eigenvalues, resolvent, and state transition matrix for the harmonic oscillator. Express $x(t)$ in terms of $x(0)$.

b) Sketch the vector field of the harmonic oscillator.

c) The state trajectories describe circular orbits, *i.e.*, $\|x(t)\|$ is constant. Verify this fact using the solution from part (a).

d) You may remember that circular motion (in a plane) is characterized by the velocity vector being orthogonal to the position vector. Verify that this holds for any trajectory of the harmonic oscillator. Use only the differential equation; do not use the explicit solution you found in part (a).

Solution.

a) We have

$$(sI - A)^{-1} = \frac{1}{s^2 + \omega^2} \begin{bmatrix} s & \omega \\ -\omega & s \end{bmatrix}.$$

From this result it follows that the eigenvalues of A are given by $\{\pm j\omega\}$. The inverse Laplace transform gives

$$\Phi(t) = \begin{bmatrix} \cos\omega t & \sin\omega t \\ -\sin\omega t & \cos\omega t \end{bmatrix}$$

and we have $x(t) = \Phi(t)x(0)$.

b) Here is the vector field:

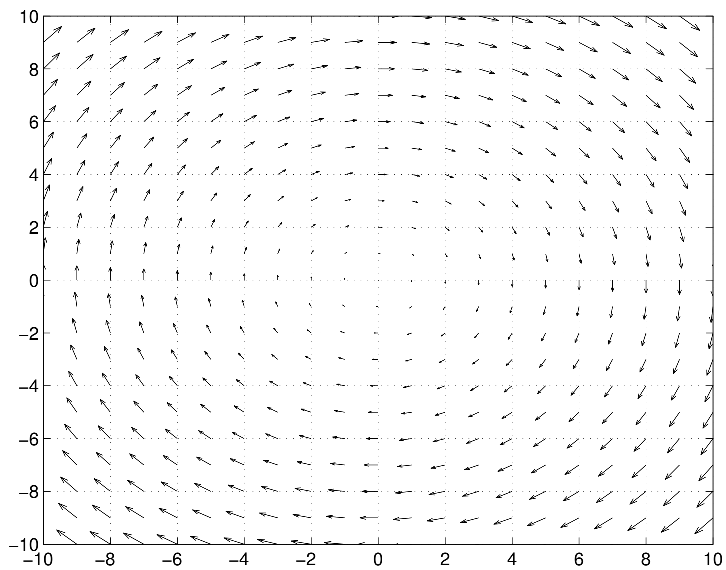


Figure 1: Vector field of harmonic oscillator

c) First we note from basic trigonometric relations that $\Phi^\mathsf{T}(t)\Phi(t) = I$. From this we conclude that $\Phi(t)$ is *orthogonal*. Now it follows that $x^\mathsf{T}(t)x(t) = x^\mathsf{T}(0)\Phi^\mathsf{T}(t)\Phi(t)x(0) = x^\mathsf{T}(0)x(0)$, i.e. $\|x(t)\| = \|x(0)\|$.

d) Using previous relations we can write

$$\dot{x}^\mathsf{T} x = x^\mathsf{T} \begin{bmatrix} 0 & -\omega \\ \omega & 0 \end{bmatrix} x = \begin{bmatrix} -\omega x_2 & \omega x_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

This shows that the velocity vector is always orthogonal to the position vector, as claimed.

4. **Norm expressions for quadratic forms.** Let $f(x) = x^\mathsf{T} Ax$ (with $A = A^\mathsf{T} \in \mathbb{R}^{n\times n}$) be a quadratic form.

a) Show that f is positive semidefinite (*i.e.*, $A \geq 0$) if and only if it can be expressed as $f(x) = \|Fx\|^2$ for some matrix $F \in \mathbb{R}^{k\times n}$. Explain how to find such an F (when $A \geq 0$). What is the size of the smallest such F (*i.e.*, how small can k be)?

b) Show that f can be expressed as a difference of squared norms, in the form $f(x) = \|Fx\|^2 - \|Gx\|^2$, for some appropriate matrices F and G. How small can the sizes of F and G be?

Solution.

a) We know that the norm expression $f(x) = \|Fx\|^2$ is a positive semidefinite quadratic form simply because $f(x) \geq 0$ for all x and $f(x) = x^\mathsf{T} Ax$ with $A = F^\mathsf{T} F \geq 0$. In this problem we will show the converse, *i.e.*, any positive semidefinite quadratic form $f(x) = x^\mathsf{T} Ax$ can be written as a norm expression $f(x) = \|Fx\|^2$. Suppose the eigenvalue decomposition of $A \geq 0$ is $Q\Lambda Q^\mathsf{T}$, with $Q^\mathsf{T} Q = I$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ where λ_i are the eigenvalues of A. Since $\lambda_i \geq 0$ (because $A \geq 0$) then $\Lambda^{1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$ is a real matrix. Let $F = \Lambda^{1/2} Q^\mathsf{T} \in \mathbb{R}^{n\times n}$. Then we have $\|Fx\|^2 = x^\mathsf{T} F^\mathsf{T} Fx = Q\Lambda^{1/2}\Lambda^{1/2}Q^\mathsf{T} = x^\mathsf{T} Ax = f(x)$. To get smallest F suppose that $\mathrm{rank}(A) = r$. Therefore, $A \in \mathbb{R}^{n\times n}$ has exactly r nonzero eigenvalues $\lambda_1, \ldots, \lambda_r$. Suppose $\Lambda_+ = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$. Hence, the eigenvalue decomposition of A can be written as

$$A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} \Lambda_+ & 0_{r\times(n-r)} \\ 0_{(n-r)\times r} & 0_{(n-r)\times(n-r)} \end{bmatrix} \begin{bmatrix} Q_1^\mathsf{T} \\ Q_2^\mathsf{T} \end{bmatrix}$$

and as a result $A = Q_1 \Lambda_+ Q_1^\mathsf{T}$ where $Q_1 \in \mathbb{R}^{n\times r}$. Now we can take $F = \Lambda_+^{1/2} Q_1^\mathsf{T} \in \mathbb{R}^{r\times n}$. Therefore, k can be as small as r, *i.e.*, $k = \mathrm{rank}(r)$. Note that k cannot be any smaller than $\mathrm{rank}(A)$ because $A = F^\mathsf{T} F$ implies that $\mathrm{rank}(A) \leq k$.

b) In general, a quadratic form need not to be positive semidefinite. In this problem we show that any quadratic form can be decomposed into its "positive" and "negative" parts. In other words, we can write $f(x)$ as the difference of two norm expressions, *i.e.*, $f(x) = \|Fx\|^2 - \|Gx\|^2$. Suppose A has n_1 positive eigenvalues $\lambda_1, \ldots, \lambda_{n_1}$, n_2 negative eigenvalues $\lambda_{n_1+1}, \ldots, \lambda_{n_1+n_2}$, and therefore $n - n_1 - n_2$ zero eigenvalues. Let

$$\Lambda_+ = \mathrm{diag}(\lambda_1, \ldots, \lambda_{n_1}), \quad \Lambda_- = \mathrm{diag}(-\lambda_{n_1+1}, \ldots, -\lambda_{n_1+n_2}).$$

The eigenvalue decomposition of A can be written as

$$A = \begin{bmatrix} Q_1 & Q_2 & Q_3 \end{bmatrix} \begin{bmatrix} \Lambda_+ & 0_{n_1\times n_2} & 0_{n_1\times(n-n_1-n_2)} \\ 0_{n_2\times n_1} & -\Lambda_- & 0_{n_2\times(n-n_1-n_2)} \\ 0_{(n-n_1-n_2)\times n_1} & 0_{(n-n_1-n_2)\times n_2} & 0_{(n-n_1-n_2)\times(n-n_1-n_2)} \end{bmatrix} \begin{bmatrix} Q_1^\mathsf{T} \\ Q_2^\mathsf{T} \\ Q_3^\mathsf{T} \end{bmatrix}$$

so $A = Q_1 \Lambda_+ Q_1^\mathsf{T} - Q_2^\mathsf{T} \Lambda_- Q_2$. Now simply take $F = \Lambda_+^{1/2} Q_1^\mathsf{T} \in \mathbb{R}^{n_1\times n}$ and $G = \Lambda_-^{1/2} Q_2^\mathsf{T} \in \mathbb{R}^{n_2\times n}$. It is easy to verify that $A = F^\mathsf{T} F - G^\mathsf{T} G$ and therefore $x^\mathsf{T} Ax = \|Fx\|^2 - \|Gx\|^2$. In fact, this method gives the smallest sizes for F and G.

5. **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 $-Q \geq -P$.

   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 λ_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 $Q = U\Lambda U^\mathsf{T}$ is an eigenvalue decomposition of Q. Since $Q > 0$ then $\Lambda > 0$ and therefore $Q^{-1/2} = Q^{-T/2} = U\Lambda^{-1/2}U^\mathsf{T}$ 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 $Q = -2$.

6. Frobenius norm of a matrix. The Frobenius norm of a matrix $A \in \mathbb{R}^{n\times n}$ is defined as $\|A\|_{\mathrm{F}} = \sqrt{\operatorname{trace} A^{\mathsf{T}}A}$. (Recall trace is the trace of a matrix, *i.e.*, the sum of the diagonal entries.)

a) Show that

$$\|A\|_{\mathrm{F}} = \left(\sum_{i,j} |A_{ij}|^2\right)^{1/2}.$$

Thus the Frobenius norm is simply the Euclidean norm of the matrix when it is considered as an element of $\mathbb{R}^{n^2}$. Note also that it is much easier to compute the Frobenius norm of a matrix than the (spectral) norm (*i.e.*, maximum singular value).

b) Show that if U and V are orthogonal, then $\|UA\|_{\mathrm{F}} = \|AV\|_{\mathrm{F}} = \|A\|_{\mathrm{F}}$. Thus the Frobenius norm is not changed by a pre- or post- orthogonal transformation.

c) Show that $\|A\|_{\mathrm{F}} = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$, where $\sigma_1, \ldots, \sigma_r$ are the singular values of A. Then show that $\sigma_{\max}(A) \leq \|A\|_{\mathrm{F}} \leq \sqrt{r}\sigma_{\max}(A)$. In particular, $\|Ax\| \leq \|A\|_{\mathrm{F}}\|x\|$ for all x.

Solution.

a) Simply by definition

$$\|A\|_{\mathrm{F}}^2 = \operatorname{trace} A^{\mathsf{T}}A = \sum_i [A^{\mathsf{T}}A]_{ii} = \sum_i \left(\sum_j A^{\mathsf{T}}_{ij}A_{ji}\right) = \sum_{i,j} A_{ij}^2.$$

b) First note that $\|UA\|_{\mathrm{F}} = \|A\|_{\mathrm{F}}$ because

$$\|UA\|_{\mathrm{F}}^2 = \operatorname{trace}(UA)^{\mathsf{T}}(UA) = \operatorname{trace} A^{\mathsf{T}}U^{\mathsf{T}}UA = \operatorname{trace} A^{\mathsf{T}}A = \|A\|_{\mathrm{F}}^2.$$

and $\|AV\|_{\mathrm{F}} = \|A\|_{\mathrm{F}}$ since

$$\|AV\|_{\mathrm{F}}^2 = \operatorname{trace}(AV)^{\mathsf{T}}(AV) = \operatorname{trace}(AV)(AV)^{\mathsf{T}} = \operatorname{trace} AVV^{\mathsf{T}}A^{\mathsf{T}} = \operatorname{trace} AA^{\mathsf{T}} = \operatorname{trace} A^{\mathsf{T}}A = \|A\|_{\mathrm{F}}^2$$

where we have used the fact that $\operatorname{trace} XY = \operatorname{trace} YX$.

c) We start with the full SVD of $A = U\Sigma V^\mathsf{T}$. By the previous problem,

$$\|A\|_\mathrm{F} = \|U^\mathsf{T}\Sigma V\|_\mathrm{F} = \|\Sigma V\|_\mathrm{F} = \|\Sigma\|_\mathrm{F} = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}.$$

Since $\sigma_2^2, \ldots, \sigma_r^2 \geq 0$, we have $\sigma_1 \leq \sqrt{\sigma_1^2 + \cdots + \sigma_r^2} = \|A\|_\mathrm{F}$. Next, since $\sigma_2, \ldots, \sigma_r \leq \sigma_1$, we have $\|A\|_\mathrm{F} = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2} \leq \sqrt{\sigma_1^2 + \cdots + \sigma_1^2} = \sqrt{r}\sigma_1$.

7. Drawing a graph. We consider the problem of drawing (in two dimensions) a graph with n vertices (or nodes) and m undirected edges (or links). This just means assigning an x- and a y- coordinate to each node. We let $x \in \mathbb{R}^n$ be the vector of x- coordinates of the nodes, and $y \in \mathbb{R}^n$ be the vector of y- coordinates of the nodes. When we draw the graph, we draw node i at the location $(x_i, y_i) \in \mathbb{R}^2$. The problem, of course, is to make the drawn graph look good. One goal is that neighboring nodes on the graph (*i.e.*, ones connected by an edge) should not be too far apart as drawn. To take this into account, we will choose the x- and y-coordinates so as to minimize the objective

$$J = \sum_{i<j,\ i\sim j} \left((x_i - x_j)^2 + (y_i - y_j)^2\right),$$

where $i \sim j$ means that nodes i and j are connected by an edge. The objective J is precisely the sum of the squares of the lengths (in $\mathbb{R}^2$) of the drawn edges of the graph. We have to introduce some other constraints into our problem to get a sensible solution. First of all, the objective J is not affected if we shift all the coordinates by some fixed amount (since J only depends on differences of the x- and y-coordinates). So we can assume that

$$\sum_{i=1}^n x_i = 0, \qquad \sum_{i=1}^n y_i = 0,$$

i.e., the sum (or mean value) of the x- and y-coordinates is zero. These two equations 'center' our drawn graph. Another problem is that we can minimize J by putting all the nodes at $x_i = 0$, $y_i = 0$, which results in $J = 0$. To force the nodes to spread out, we impose the constraints

$$\sum_{i=1}^n x_i^2 = 1, \qquad \sum_{i=1}^n y_i^2 = 1, \qquad \sum_{i=1}^n x_i y_i = 0.$$

The first two say that the variance of the x- and y- coordinates is one; the last says that the x- and y- coordinates are uncorrelated. (You don't have to know what variance or uncorrelated mean; these are just names for the equations given above.) The three equations above enforce 'spreading' of the drawn graph. Even with these constraints, the coordinates that minimize J are not unique. For example, if x and y are any set of coordinates, and $Q \in \mathbb{R}^{2\times 2}$ is any orthogonal matrix, then the coordinates given by

$$\begin{bmatrix} \tilde{x}_i \\ \tilde{y}_i \end{bmatrix} = Q \begin{bmatrix} x_i \\ y_i \end{bmatrix}, \qquad i = 1, \ldots, n$$

satisfy the centering and spreading constraints, and have the same value of J. This means that if you have a proposed set of coordinates for the nodes, then by rotating or reflecting them, you get another set of coordinates that is just as good, according to our objective. We'll just live with this ambiguity. Here's the question:

a) Explain how to solve this problem, *i.e.*, how to find x and y that minimize J subject to the centering and spreading constraints, given the graph topology. You can use any method or ideas we've encountered in the course. Be clear as to whether your approach solves the problem exactly (*i.e.*, finds a set of coordinates with J as small as it can possibly be), or whether it's just a good heuristic (*i.e.*, a choice of coordinates that achieves a reasonably small value of J, but perhaps not the absolute best). In describing your method, you may not refer to any programming commands or operators; your description must be entirely in mathematical terms.

b) Implement your method, and carry it out for the graph given in `dg_data.json`. This JSON file contains the *node adjacency matrix* of the graph, denoted A, and defined as $A_{ij} = 1$ if nodes i and j are connected by an edge, and $A_{ij} = 0$ otherwise. (The graph is undirected, so A is symmetric. Also, we do not have self-loops, so $A_{ii} = 0$, for $i = 1, \dots, n$.) Give the value of J achieved by your choice of x and y, and verify that your x and y satisfy the centering and spreading conditions, at least approximately. If your method is iterative, plot the value of J versus iteration. Draw the corresponding graph by plotting nodes as small circles and edges as lines. For comparison, the JSON file also contains the vectors `x_circ` and `y_circ`. These coordinates were obtained using a standard technique for drawing a graph, by placing the nodes in order on a circle. The radius of the circle has been chosen so that `x_circ` and `y_circ` satisfy the centering and spread constraints. Draw this graph on a separate plot.

Hint. You are welcome to use the results described below, without proving them. Let $A \in \mathbb{R}^{n\times n}$ be symmetric, with eigenvalue decomposition $A = \sum_{i=1}^{n} \lambda_i q_i q_i^\mathsf{T}$, with $\lambda_1 \geq \cdots \geq \lambda_n$, and $\{q_1, \dots, q_n\}$ orthonormal. You know that a solution of the problem

$$\begin{array}{ll} \text{minimize} & x^\mathsf{T} A x \\ \text{subject to} & x^\mathsf{T} x = 1, \end{array}$$

where the variable is $x \in \mathbb{R}^n$, is $x = q_n$. The related maximization problem is

$$\begin{array}{ll} \text{maximize} & x^\mathsf{T} A x \\ \text{subject to} & x^\mathsf{T} x = 1 \end{array}$$

with variable $x \in \mathbb{R}^n$. A solution to this problem is $x = q_1$. Now consider the following generalization of the first problem:

$$\begin{array}{ll} \text{minimize} & \mathrm{trace}(X^\mathsf{T} A X) \\ \text{subject to} & X^\mathsf{T} X = I_k \end{array}$$

where the variable is $X \in \mathbb{R}^{n\times k}$, and I_k denotes the $k \times k$ identity matrix, and we assume $k \leq n$. The constraint means that the columns of X, say, $x_1, \dots, x_k$, are orthonormal; the objective can be written in terms of the columns of X as $\mathrm{trace}(X^\mathsf{T} A X) = \sum_{i=1}^{k} x_i^\mathsf{T} A x_i$. A solution of this problem is $X = [q_{n-k+1} \cdots q_n]$. Note that when $k = 1$, this reduces to the first problem above. The related maximization problem is

$$
\begin{array}{ll}
\text{maximize} & \operatorname{trace}(X^\mathsf{T} A X) \\
\text{subject to} & X^\mathsf{T} X = I_k
\end{array}
$$

with variable $X \in \mathbb{R}^{n \times k}$. A solution of this problem is $X = [q_1 \cdots q_k]$.

Solution. We first note that the objective function of this problem is just a sum the same quadratic form of x and y:

$$J = x^\mathsf{T} L x + y^\mathsf{T} L y,$$

where

$$x^\mathsf{T} L x = \sum_{i<j,\ i\sim j} (x_i - x_j)^2 = \sum_{i<j,\ i\sim j} (x_i^2 - 2x_i x_j + x_j^2) = \sum_{i,j,\ i\sim j} (x_i^2 - x_i x_j), \tag{1}$$

and similarly for y. We can express L as follows. If node i has d_i edges emanating from it (*i.e.*, it has degree d_i), then the coefficient corresponding to x_i^2 in the above sum is d_i. Now, if nodes i and j are connected, then the sum contains a $-x_i x_j$ component. Thus the elements of L are:

$$L_{ij} = \begin{cases} d_i & i = j \\ -1 & i \sim j \\ 0 & \text{otherwise.} \end{cases}$$

Now, in terms of A, d_i is just the sum of the elements of A along its ith row or column. Furthermore, if $i \sim j$ then $A_{ij} = 1$, otherwise $A_{ij} = 0$. We can thus express the coefficients of L in terms of the coefficients of A as:

$$L_{ij} = \begin{cases} \sum_{i=1}^m A_{ij} & i = j \\ -A_{ij} & \text{otherwise.} \end{cases}$$

The matrix L is called the Laplacian of the graph, and shows up in many different problems involving graphs. We can therefore write down the problem as

$$
\begin{array}{ll}
\text{minimize} & x^\mathsf{T} L x + y^\mathsf{T} L y \\
\text{subject to} & \mathbf{1}^\mathsf{T} x = 0, \quad \mathbf{1}^\mathsf{T} y = 0 \\
 & \|x\|_2 = 1, \quad \|y\|_2 = 1, \quad x^\mathsf{T} y = 0.
\end{array} \tag{2}
$$

First we note that L is positive semidefinite: $x^\mathsf{T} L x$ is a sum of squared terms, hence nonnegative. Therefore all the eigenvalues of L are nonnegative. Second, we have $L\mathbf{1} = 0$, since the sum of the rows of L are all zero. This means that $\mathbf{1}$ is an eigenvector of L with eigenvalue 0 (*i.e.*, it's in the nullspace of L). Of course, 0 must the smallest eigenvalue, since all eigenvalues are nonnegative, *i.e.*, we have $\lambda_1 = 0$. Now, since $L\mathbf{1} = 0$, we have $v^\mathsf{T} L v = 0$, where we take v to be $v = n^{-1/2}\mathbf{1}$. (v is the normalized eigenvector of L corresponding to eigenvalue 0.) Thus we can write the objective function as

$$J = x^\mathsf{T} L x + y^\mathsf{T} L y + v^\mathsf{T} L v,$$

or, using matrix notation,

$$J = \operatorname{trace}\left([x\ y\ v]^\mathsf{T} L [x\ y\ v]\right).$$

We can also gather all the equality constraints in problem (**??**) into a single compact matrix equality constraint:

$$[x\ y\ v]^\mathsf{T}[x\ y\ v] = I,$$

where I is the 3×3 identity matrix. Thus our problem is

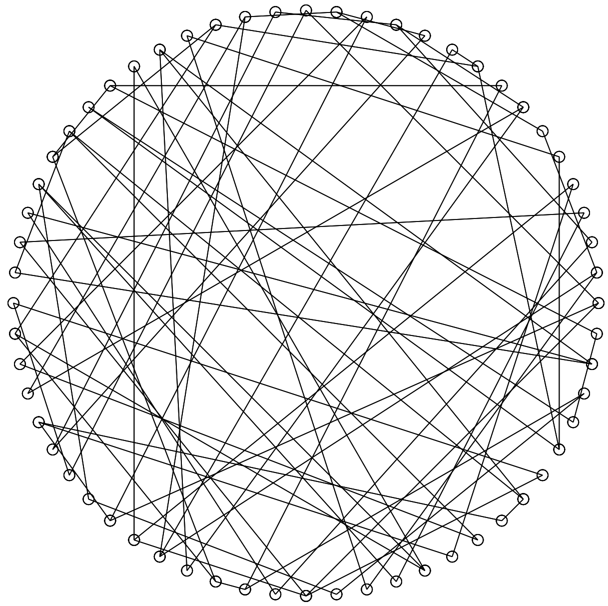
$$\begin{array}{ll} \text{minimize} & \operatorname{trace}\left([x\ y\ v]^\mathsf{T} L [x\ y\ v]\right) \\ \text{subject to} & [x\ y\ v]^\mathsf{T}[x\ y\ v] = I. \end{array} \tag{3}$$

here the variables are x and y; v is given by $v = n^{-1/2}\mathbf{1}$. We know how to solve a problem very similar to this problem. For $C = C^\mathsf{T} \in \mathbb{R}^{n \times n}$, a solution $Q \in \mathbb{R}^{n \times r}$ of

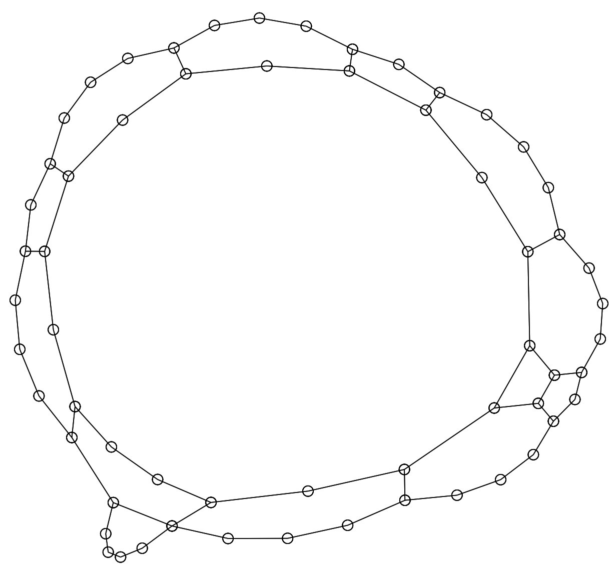
$$\begin{array}{ll} \text{minimize} & \operatorname{trace}\left(Q^\mathsf{T} C Q\right) \\ \text{subject to} & Q^\mathsf{T} Q = I, \end{array} \tag{4}$$

is to take the r columns of the (orthonormal) eigenvectors of C corresponding to its r smallest eigenvalues. In fact $v = n^{-1/2}\mathbf{1}$ is already fixed as a normalized eigenvector of L corresponding to its smallest eigenvalue, $\lambda_n = 0$. So to get the optimal values of x and y, we just take them to be the eigenvectors of L corresponding to the eigenvalues λ_{n-1} and λ_{n-2}: We take $x = v_{n-1}$ and $y = v_{n-2}$. A very simple solution to a problem that looks pretty complicated. Some people came up with this solution, without a clear argument as to why it's the solution, especially, why it is that the constraints are satisfied. You pretty much have to mention the nullspace of L, (the span of the vector $\mathbf{1}$), to show you really understand it. Believe it or not, several people invented their own iterative methods for solving the problem, and got the correct answer. All they were doing was re-inventing famous methods for computing the eigenvectors corresponding to the smallest eigenvalues of a symmetric matrix. By the way, there's a really interesting field, called *spectral graph theory*, that studies the relationship between graphs and the eigenvalues of associated matrices (such as the Laplacian). It's not only very interesting, but extremely useful in practice too. It's a critical element in web search (like Google's Pagerank), and also partitioning large graphs (such as in circuit design), and lots of other problems, too (like image segmentation). We also mention that the problem can be solved in terms of the singular value decomposition. Number the edges of the graph $1, \ldots, m$, and assign an arbitrary orientation to each one. Define the *incidence matrix* $B \in \mathbb{R}^{n \times n}$ as $B_{ij} = 1$ if edge j points in to node i, $B_{ij} = -1$ if edge j comes out of node i, and zero otherwise. Then we have the formula $L = BB^\mathsf{T}$. The matrix B has rank exactly $n-1$, by a famous theorem of graph theory, since the graph is connected. (Of course, you could just check this numerically.) The solution to our problem is then to *take x and y to be the right singular vectors associated with the two smallest (positive) singular values.* The figure below shows the graph generated by placing all nodes on a circle, whose radius is such that x and y

satisfy the problem constraints. In this case we have $J = 5.328$.



The following figure shows a graph which minimizes J by the proposed method. This graph achieves $J = 0.107$.



8. **Two representations of an ellipsoid.** In the lectures, we saw two different ways of representing an ellipsoid, centered at 0, with non-zero volume. The first uses a quadratic form:

$$\mathcal{E}_1 = \left\{ x \,\middle|\, x^\mathsf{T} S x \leq 1 \right\},$$

with $S^\mathsf{T} = S > 0$. The second is as the image of a unit ball under a linear mapping:

$$\mathcal{E}_2 = \{ y \,|\, y = Ax, \|x\| \leq 1 \},$$

with $\det(A) \neq 0$.

a) Given S, explain how to find an A so that $\mathcal{E}_1 = \mathcal{E}_2$.

b) Given A, explain how to find an S so that $\mathcal{E}_1 = \mathcal{E}_2$.

c) What about uniqueness? Given S, explain how to find *all* A that yield $\mathcal{E}_1 = \mathcal{E}_2$. Given A, explain how to find *all* S that yield $\mathcal{E}_1 = \mathcal{E}_2$.

Solution. First we will show that

$$\mathcal{E}_2 = \left\{ y \mid y^\mathsf{T}(AA^\mathsf{T})^{-1} y \leq 1 \right\} \tag{5}$$

$$\begin{aligned}
\mathcal{E}_2 &= \{y \mid y = Ax, \|x\| \leq 1\} \\
&= \left\{y \mid A^{-1}y = x, \|x\| \leq 1\right\} \text{ since } A \text{ is invertible square matrix} \\
&= \left\{y \mid \|A^{-1}y\| \leq 1\right\} \\
&= \left\{y \mid y^\mathsf{T} A^{-T} A^{-1} y \leq 1\right\} = \left\{y \mid y^\mathsf{T}(AA^\mathsf{T})^{-1} y \leq 1\right\}
\end{aligned}$$

Since S is symmetric positive definite, the eigenvalues of S are all positive and we can choose n orthonormal eigenvectors. So $S = Q\Lambda Q^\mathsf{T}$ where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) > 0$ and Q is orthogonal. Let $\Lambda^{\frac{1}{2}} = \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$. If we let $A = Q(\Lambda^{\frac{1}{2}})^{-1} = Q\Lambda^{-\frac{1}{2}}$,

$$\begin{aligned}
(AA^\mathsf{T})^{-1} &= (Q\Lambda^{-\frac{1}{2}}\Lambda^{-\frac{1}{2}}Q^\mathsf{T})^{-1} \\
&= (Q\Lambda^{-1}Q^\mathsf{T})^{-1} = Q\Lambda Q^\mathsf{T} \\
&= S
\end{aligned}$$

Therefore, by (**??**) $A = Q\Lambda^{-\frac{1}{2}}$ yields $\mathcal{E}_1 = \mathcal{E}_2$. By (**??**), $S = (AA^\mathsf{T})^{-1}$ yields $\mathcal{E}_1 = \mathcal{E}_2$. *Uniqueness:* We show that

$$\mathcal{E}_S = \mathcal{E}_T \Leftrightarrow S = T \tag{6}$$

where $\mathcal{E}_S = \{x \mid x^\mathsf{T}Sx \leq 1\}$, $\mathcal{E}_T = \{x \mid x^\mathsf{T}Tx \leq 1\}$, $S^\mathsf{T} = S > 0$ and $T^\mathsf{T} = T > 0$. It is clear that if $S = T$, then $\mathcal{E}_S = \mathcal{E}_T$. Now we show that $\mathcal{E}_S = \mathcal{E}_T \Rightarrow x^\mathsf{T}Sx = x^\mathsf{T}Tx,\ \forall x \in \mathbb{R}^n$. Without loss of generality let's assume $\exists x_0 \in \mathbb{R}^n$ such that $x_0^\mathsf{T}Sx_0 > x_0^\mathsf{T}Tx_0 = \alpha \neq 0$. If we let $x_1 = x_0/\sqrt{\alpha}$, then $x_1^\mathsf{T}Tx_1 = 1$, but $x_1^\mathsf{T}Sx_1 > 1$, thus $x_1 \in \mathcal{E}_T$ but $x_1 \notin \mathcal{E}_S$, and therefore $\mathcal{E}_S \neq \mathcal{E}_T$. Finally, $\mathcal{E}_S = \mathcal{E}_T \Rightarrow x^\mathsf{T}Sx = x^\mathsf{T}Tx,\ \forall x \in \mathbb{R}^n \Rightarrow S = T$ by the uniqueness of the symmetric part in a quadratic form. Hence S is unique. *Given S, find all A that yield $\mathcal{E}_1 = \mathcal{E}_2$.*

The answer is

$$A = Q\Lambda^{-\frac{1}{2}}V^\mathsf{T}$$

where $V \in \mathbb{R}^{n\times n}$ is any orthogonal matrix and

$$S^\mathsf{T} = S = Q\Lambda Q^\mathsf{T} > 0$$

where Q is orthogonal and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) > 0$. Let the singular value decomposition of

A be

$$A = U\Sigma V^\mathsf{T}$$

where $U, V \in \mathbb{R}^{n\times n}$ are orthogonal and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n) > 0$ (since $\det(A) \neq 0$.) By (**??**),

$$\begin{aligned}
\mathcal{E}_2 &= \left\{ y \mid y^\mathsf{T}(AA^\mathsf{T})^{-1}y \leq 1 \right\} \\
&= \left\{ y \mid y^\mathsf{T}(U\Sigma^2U^\mathsf{T})^{-1}y \leq 1 \right\} \\
&= \left\{ y \mid y^\mathsf{T}U\Sigma^{-2}U^\mathsf{T}y \leq 1 \right\}
\end{aligned}$$

Thus, if $\mathcal{E}_1 = \mathcal{E}_2$, then $S = U\Sigma^{-2}U^\mathsf{T}$ by (**??**). Therefore $U = Q$ and $\Sigma = \Lambda^{-\frac{1}{2}}$, and V can be any orthogonal matrix. You can also see why A's are different only by right-side multiplication by an orthogonal matrix by the following argument. By (**??**) and (**??**), $AA^\mathsf{T} = S^{-1}$. Let

$$A = \left[\begin{array}{cccc} \tilde{a}_1 & \tilde{a}_2 & \ldots & \tilde{a}_n \end{array}\right]^\mathsf{T}$$

Then we have,

$$\begin{aligned}
\|\tilde{a}_i\|^2 &= (S^{-1})_{ii}, \\
\tilde{a}_i^\mathsf{T}\tilde{a}_j &= (S^{-1})_{ij},
\end{aligned}$$

and

$$\cos\theta_{ij} = \frac{(S^{-1})_{ij}}{\sqrt{(S^{-1})_{ii}(S^{-1})_{jj}}}.$$

This means that the row vectors of any A satisfying $AA^\mathsf{T} = S^{-1}$ have the same length and the same angle between any two of them. So the rows of A can vary only by the application of an identical rotation or reflection to all of them. These are the transformations preserving length and angle, and correspond to orthogonal matrices. Since we are considering row vectors, the orthogonal matrix should be multiplied on the right.

9. **Worst-case analysis of impact.** We consider a (time-invariant) linear dynamical system

$$\dot{x} = Ax + Bu, \qquad x(0) = x_{\mathrm{init}},$$

with state $x(t) \in \mathbb{R}^n$, and input $u(t) \in \mathbb{R}^m$. We are interested in the state trajectory over the time interval $[0, T]$. In this problem the input u represents an *impact* on the system, so it has the form

$$u(t) = g\delta(t - T_{\mathrm{imp}}),$$

where $g \in \mathbb{R}^m$ is a vector that gives the direction and magnitude of the impact, and T_{imp} is the time of the impact. We assume that $0 \leq T_{\mathrm{imp}} \leq T_-$. ($T_{\mathrm{imp}} = T_-$ means that the impact occurs right at the end of the period of interest, and does affect $x(T)$.) We let $x_{\mathrm{nom}}(T)$ denote the state, at time $t = T$, of the linear system $\dot{x}_{\mathrm{nom}} = Ax_{\mathrm{nom}}$, $x_{\mathrm{nom}}(0) = x_{\mathrm{init}}$. The vector $x_{\mathrm{nom}}(T)$ is what the final state $x(T)$ of the system above would have been at time $t = T$, had the impact not occurred (*i.e.*, with $u = 0$). We are interested in the deviation D between $x(T)$ and $x_{\rm nom}(T)$, as measured by the norm:

$$D = \|x(T) - x_{\rm nom}(T)\|.$$

D measures how far the impact has shifted the state at time T. We would like to know how large D can be, over all possible impact directions and magnitudes no more than one (*i.e.*, $\|g\| \le 1$), and over all possible impact times between 0 and T_-. In other words, we would like to know the maximum possible state deviation, at time T, due to an impact of magnitude no more than one. We'll call the choices of $T_{\rm imp}$ and g that maximize D the *worst-case impact time* and *worst-case impact vector*, respectively.

a) Explain how to find the worst-case impact time, and the worst-case impact vector, given the problem data A, B, $x_{\rm init}$, and T. Your explanation should be as short and clear as possible. You can use any of the concepts we have encountered in the class. Your approach can include a simple numerical search (such as plotting a function of one variable to find its maximum), if needed. If either the worst-case impact time or the worst-case impact vector do not depend on some of the problem data (*i.e.*, A, B, $x_{\rm init}$, and T) say so.

b) Get the data from `worst_case_impact_data.json`, which defines A, B, $x_{\rm init}$, and T, and carry out the procedure described in part (a). Be sure to give us the worst-case impact time (with absolute precision of 0.01), the worst-case impact vector, and the corresponding value of D.

Solution.

a) We have $x_{\rm nom}(T) = e^{AT} x_{\rm init}$. The state of the system right before the impact is given by

$$x(T_{\rm imp-}) = e^{AT_{\rm imp}} x_{\rm init},$$

and the state right after the impact is

$$x(T_{\rm imp+}) = e^{AT_{\rm imp}} x_{\rm init} + Bg,$$

so the final state is given by

$$x(T) = e^{A(T-T_{\rm imp})} \left(e^{AT_{\rm imp}} x_{\rm init} + Bg\right) = e^{AT} x_{\rm init} + e^{A(T-T_{\rm imp})} Bg.$$

The deviation is then

$$D = \|x(T) - x_{\rm nom}(T)\| = \|e^{A(T-T_{\rm imp})} Bg\|.$$

Thus, the deviation does not depend at all on $x_{\rm init}$, the initial state. Now suppose that $T_{\rm imp}$ is fixed. How do we choose g, subject to $\|g\| \le 1$, to maximize D? That's simple: we choose g to be v_1, the right singular vector of the matrix

$$e^{A(T-T_{\rm imp})} B,$$

associated with its largest singular value (*i.e.*, its norm). The associated maximum deviation is then $D = \|e^{A(T-T_{\rm imp})} B\|$. There is no analytical formula or method for

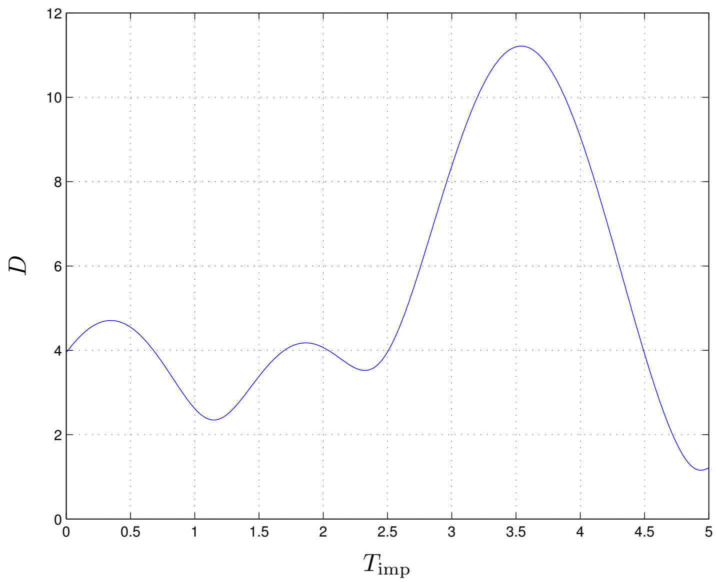


Figure 2: Final state deviation $D = \|x(T) - x_{\text{nom}}(T)\|$ versus impact time T_{imp}. The worst-case impact time is the time that maximizes D.

finding the worst-case inpact time. Instead, we have to just compute and plot $D = \|e^{A(T-T_{\text{imp}})}B\|$ over the interval of interest, and find the value of T_{imp} that gives the largest norm. We then find the worst-case impact vector as the right singular vector associated with the largest singular value.

b) The solution procedure described above was carried out for the given problem instance. In figure **??** we plot D versus impact time over the time interval $T_{\text{imp}} \in [0, 5]$. (The corresponding matlab code is given below). The worst-case impact time occurs at $T_{\text{imp}} = 3.54$, and the associated maximum state deviation is $D_{\max} = 11.213$. The worst-case impact vector is the corresponding right singular vector, $v_{\max} = [-0.0525\ 0.8926\ 0.4477]^{\mathsf{T}}$.

```
wc_impact_data;  % system data for worst-case impact problem
D = []; % compute deviation versus Timp relation curve
for Timp=0:0.01:T
D = [D svds(expm(A*(T-Timp))*B,1)];
end
Timp=0:0.01:T;
[Dmax indmax] = max(D);
Tmax = Timp(indmax);
[umax,S,vmax] = svds(expm(A*(T-Tmax))*B,1);
plot(Timp,D); xlabel('t'); ylabel('D'); grid;
print -depsc wc_impact.eps
```