EE263 Final Exam

Summer 2026

Name: _____ **Student ID number:** _____

The Stanford Honor Code

The Honor Code is an undertaking of the Stanford academic community, individually and collectively. Its purpose is to uphold a culture of academic honesty.

Students will support this culture of academic honesty by neither giving nor accepting unpermitted academic aid in any work that serves as a component of grading or evaluation, including assignments, examinations, and research.

Instructors will support this culture of academic honesty by providing clear guidance, both in their course syllabi and in response to student questions, on what constitutes permitted and unpermitted aid. Instructors will also not take unusual or unreasonable precautions to prevent academic dishonesty.

Students and instructors will also cultivate an environment conducive to academic integrity. While instructors alone set academic requirements, the Honor Code is a community undertaking that requires students and instructors to work together to ensure conditions that support academic integrity.

I acknowledge and accept the Stanford Honor Code. I affirm that I have neither given nor received unpermitted aid on this examination.

Signature: _____ **Date:** _____

Instructions

- This is a closed-book, closed-notes exam. You may use two double-sided sheets of notes. No calculators, phones, or other electronic devices are allowed.
- You have 150 minutes. The exam has 5 problems worth 100 points in total, followed by a sixth problem that is entirely extra credit. The point value of each problem and of each part is stated. Budget your time for problems 1 through 5 and turn to problem 6 only if you have time left.
- There are 10 extra-credit points in all: 2 at the end of problem 2, part (d), and the 8 of problem 6. Anything you earn there is added to your score on the 100 points above, and skipping them costs you nothing.
- Problem 1 is a set of multiple-choice items and needs no explanation. Problems 2 through 6 ask you to derive, compute, and explain. For those problems, an answer with no reasoning receives little credit, and a clearly reasoned answer with a small arithmetic slip receives most of the credit.
- Write your answers on the separate answer sheet, not in this booklet. Anything written in this booklet will not be graded. The answer sheet has a page for each part, with room sized for the answer we have in mind. If a part runs over, use the extra pages at the end of the answer sheet and write *continued on page ...* in that part's space.
- Wherever we ask for a specific answer, such as a number, a matrix, a line, or a formula in terms of a parameter, the answer sheet gives you a box for it, labelled with what belongs in it. Put that answer in the box and keep the reasoning in the space around it.
- Every part of every problem has a short solution, and every number is chosen so that the arithmetic can be done by hand. If you find yourself in a calculation that looks prohibitively long, or that seems to need a calculator, there is a better way, and you are missing the structure that makes the part easy.
- Keep your answers short. We will deduct points from answers that are technically correct but far more complicated than they need to be.
- We have made every effort to state the problems clearly. If you believe a problem is unclear or ambiguous, you may make a reasonable assumption that resolves the ambiguity and solve the problem under that assumption. State your assumption. If your interpretation is reasonable, it may receive credit even if it differs from the one we intended.
- Good luck!

1. Multiple choice *(20 points).* For each item, write the item number and the letter of your answer in your blue book. Each item has exactly one correct answer. No explanation is required, and none will be read. *(2 points each)*

1. Let $u_1, \cdots, u_4 \in \mathbb{R}^n$ be unit-norm pairwise orthogonal vectors, and define $A = \begin{bmatrix} u_1 & u_2 \end{bmatrix}$, $B = \begin{bmatrix} u_3 & u_4 \end{bmatrix}$. Which of the following statements is correct about the matrix AB^T?

   (a) It has exactly two non-zero eigenvalues
   (b) It can have up to two non-zero eigenvalues
   (c) **All its eigenvalues are zero, but the matrix itself isn't necessarily the zero matrix**
   (d) It is possible for it to have no non-zero eigenvalues, in which case it will be the zero matrix

   Solution. (c). $AB^\mathsf{T} = u_1 u_3^\mathsf{T} + u_2 u_4^\mathsf{T}$, and squaring it gives zero: every term carries an inner product of two distinct u_i. A nilpotent matrix has all its eigenvalues equal to zero. It is not the zero matrix, though, since $AB^\mathsf{T} u_3 = u_1 \neq 0$, so it has rank 2.

2. Let $X \in \mathbb{R}^n$ be a vector of random variables with the covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$. If $\operatorname{rank}(\Sigma) = n - 2$, which of the following statements is the most accurate about these random variables?

   (a) Out of the n random variables, 2 of them have zero variances
   (b) X_n and X_{n-1} are deterministic linear combinations of $X_1, \cdots, X_{n-2}$
   (c) **Some i, j exist, such that if we remove X_i, X_j from X, no information will be lost**
   (d) With probability one, X lies in an affine subspace of $\mathbb{R}^n$ of dimension 2

   Solution. (c). If $\Sigma w = 0$ then $\operatorname{var}(w^\mathsf{T} X) = w^\mathsf{T} \Sigma w = 0$, so $w^\mathsf{T} X$ equals its mean with probability one. Here $\operatorname{null}(\Sigma)$ is two-dimensional; collect a basis in $W \in \mathbb{R}^{n \times 2}$, which has rank 2 and therefore two independent rows i, j. Splitting $W^\mathsf{T}(X - \mathrm{E}\, X) = 0$ on those two coordinates and inverting the resulting 2×2 block writes X_i and X_j as affine functions of the other $n-2$ variables, so dropping them loses nothing. (a) fails because a rank deficiency says nothing about individual variances. (b) names the wrong pair: the number of relations is fixed, which variables are redundant is not. (d) has the dimension the wrong way round, the realizations fill an affine subspace of dimension $n - 2$.

3. Let $A \in \mathbb{R}^{n \times n}$, and suppose $A = P \Sigma P^{-1}$ for some invertible P and some diagonal Σ. Which of the following is true?

   (a) If A is symmetric, then $P^{-1} = P^\mathsf{T}$
   (b) If $P^{-1} = P^\mathsf{T}$, then the eigenvalues of A are distinct
   (c) If the eigenvalues of A are distinct, then $P^{-1} = P^\mathsf{T}$
   (d) **None of the above**

Solution. (d). For (a), the spectral theorem says a symmetric A *can* be diagonalized by an orthogonal matrix, not that every P that diagonalizes it is one: rescaling a column of P leaves $P\Sigma P^{-1}$ unchanged and destroys orthonormality. With $A = \operatorname{diag}(1, 2)$ and $P = \operatorname{diag}(1, 5)$ we have $A = P\Sigma P^{-1}$ and $P^{-1} \neq P^{\mathsf{T}}$.

For (b), take $A = P = I$: orthogonal, with every eigenvalue equal to 1. For (c), distinct eigenvalues buy independent eigenvectors, not orthogonal ones. The matrix $A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$ has eigenvalues 1 and 2 with eigenvectors $\begin{bmatrix} 1 & 0 \end{bmatrix}^{\mathsf{T}}$ and $\begin{bmatrix} 1 & 1 \end{bmatrix}^{\mathsf{T}}$, which are not orthogonal. Orthogonality of the eigenvectors comes from symmetry, and $P^{-1} = P^{\mathsf{T}}$ needs them normalized as well.

4. Let $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p\times n}$, $\mu > 0$, and $C = A^{\mathsf{T}}A + B^{\mathsf{T}}B + \mu I$. Which of the following statements is correct about the matrix C?

   (a) $\lambda_{\max}(C) = \max\left(\sigma^2_{\max}(A), \sigma^2_{\max}(B)\right) + \mu$
   (b) $\lambda_{\max}(C) \leq \max\left(\sigma^2_{\max}(A), \sigma^2_{\max}(B)\right) + \mu$
   (c) $\lambda_{\max}(C) = \sigma^2_{\max}(A) + \sigma^2_{\max}(B) + \mu$
   (d) $\lambda_{\max}(C) \leq \sigma^2_{\max}(A) + \sigma^2_{\max}(B) + \mu$

   Solution. (d). C is symmetric, so for $\|x\| = 1$, $x^{\mathsf{T}}Cx = \|Ax\|^2 + \|Bx\|^2 + \mu$, and maximizing over x gives

   $$\lambda_{\max}(C) = \max_{\|x\|=1} \left(\|Ax\|^2 + \|Bx\|^2\right) + \mu \leq \sigma^2_{\max}(A) + \sigma^2_{\max}(B) + \mu,$$

   since each term is maximized separately by the bound. The inequality is strict unless A and B have a common maximizing direction, which rules out the equality in (c). Both (a) and (b) fail already for $A = B = I$, where $\lambda_{\max}(C) = 2 + \mu$ while the right-hand side is $1 + \mu$.

5. A unit mass sits at rest on a frictionless surface. A force is applied to it over ten unit-time intervals and held constant on each one, with value x_i for $i - 1 < t \leq i$. If y_1 and y_2 are the position and the velocity at $t = 10$, then $y = Ax$ with $A \in \mathbb{R}^{2\times 10}$, where $y_2 = \sum_i x_i$ and $y_1 = \sum_i \left(\frac{21}{2} - i\right) x_i$. To move the mass a unit distance and leave it at rest, we solved

   $$\text{minimize} \quad \|x\| \qquad \text{subject to} \qquad Ax = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

   and the solution came out as $x = \begin{bmatrix} x_0 & 0 & \cdots & 0 & -x_0 \end{bmatrix}^{\mathsf{T}}$: one push in the first interval, an equal pull in the last, and nothing in between. Which norm was minimized?

   (a) $\|x\|_1$
   (b) $\|x\|_2$
   (c) $\|x\|_p$ for some $p > 2$
   (d) $\|x\|_\infty$

Solution. (a). Put the whole force on two intervals $i < j$, with $x_i = a$ and $x_j = -a$: the velocity constraint holds automatically, the position constraint reads $a(j-i) = 1$, and $\|x\|_1 = 2a = 2/(j-i)$, which is smallest for the pair furthest apart, $i = 1$ and $j = 10$, giving $x_0 = 1/9$. Spreading the force over more intervals only costs more. That is the signature of the ℓ_1 norm on an affine set: the solution sits at an extreme point and is sparse.

Every other choice spreads the force out instead. The least ℓ_2 solution $x = A^\mathsf{T}(AA^\mathsf{T})^{-1}y$ is linear in i, a staircase with every entry nonzero, and the least ℓ_∞ solution uses the same magnitude in all ten intervals, $+0.04$ for the first five and -0.04 for the last five. No finite $p > 1$ can produce this shape: optimality makes each x_i a signed power of the ith entry of $A^\mathsf{T}\lambda$, and those entries are affine in i, so at most one of them vanishes. Sparsity needs the kink at zero that only $p = 1$ has.

6. For an analytic function $f(a) = \beta_0 + \beta_1 a + \beta_2 a^2 + \cdots$ and a square matrix A, the matrix $f(A)$ is defined by the same series, $f(A) = \beta_0 I + \beta_1 A + \beta_2 A^2 + \cdots$. Take

$$\cos a = 1 - \frac{a^2}{2!} + \frac{a^4}{4!} - \frac{a^6}{6!} + \cdots,$$

and let $A \in \mathbb{R}^{3\times 3}$ be symmetric with eigenvalues $\pi/3$, $2\pi/3$ and π. Which of the following is true of $\cos A$?

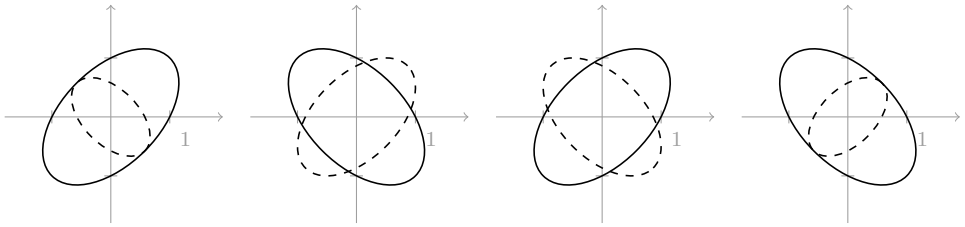
(a) It is positive semidefinite
(b) It is negative semidefinite
(c) **It is indefinite**
(d) It cannot be determined from the given information

Solution. (c). Write $A = T\Lambda T^{-1}$. Every term of the series is $T\Lambda^k T^{-1}$, so $\cos A = T\,\mathrm{diag}(\cos\lambda_1, \cos\lambda_2, \cos\lambda_3)T^{-1}$: the eigenvectors are untouched and each eigenvalue is passed through f. Here the eigenvalues of $\cos A$ are $\cos(\pi/3) = 1/2$, $\cos(2\pi/3) = -1/2$ and $\cos\pi = -1$. They come with both signs, so $\cos A$ is indefinite, and the eigenvalues are all we need, which rules out (d).

Note that A itself is positive definite, since its eigenvalues are positive, and that the largest eigenvalue of $\cos A$ is $1/2$, attained at the *smallest* eigenvalue of A: applying f preserves the ordering of the eigenvalues, or the sign of the matrix, only when f is monotone or of one sign on the spectrum, and cos is neither here.

7. Let $A = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$. In each figure below the solid curve is the set of $x \in \mathbb{R}^2$ with $x^\mathsf{T}Ax = 1$, and the dashed curve is the set of $x \in \mathbb{R}^2$ with $x^\mathsf{T}A^{-1}x = 1$. The tick marks

are at ± 1. Which figure is correct?



(a) **(b)** (c) (d)

Solution. (b). Write

$$v_+ = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}, \qquad v_- = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}.$$

A is symmetric, with eigenvalue $3/2$ along v_+ and $1/2$ along v_-, so A^{-1} has eigenvalues $2/3$ and 2 along those same two directions. For a symmetric positive definite M, the set $\{x \mid x^\mathsf{T} M x \le 1\}$ has semi-axes $1/\sqrt{\lambda_i(M)}$ along the eigenvectors of M: it is *short* where the form is large. So the solid ellipse is long along v_-, where A is small, with semi-axes $\sqrt{2} \approx 1.41$ there and $\sqrt{2/3} \approx 0.82$ along v_+. Inverting turns the small eigenvalue into the large one, so the dashed ellipse is long along v_+: semi-axes $\sqrt{3/2} \approx 1.22$ along v_+ and $1/\sqrt{2} \approx 0.71$ along v_-.

Every panel gets the perpendicularity right, since the two matrices share their eigenvectors, so you have to look at the lengths. Orientation rules out (a) and (c), where the solid curve is long along v_+. That leaves (b) and (d), which differ only in the size of the dashed curve: in (d) it sits inside the solid one, which would say $x^\mathsf{T} A x \le 1$ whenever $x^\mathsf{T} A^{-1} x \le 1$, that is $A^{-1} \succeq A$, impossible here because A has an eigenvalue larger than 1. In (b) the dashed curve pokes out along v_+, where $1.22 > 0.82$, and the two cross.

The sizes are not a coincidence: in two dimensions the second ellipse is always the first turned by $90°$ and scaled by $\sqrt{\det A}$, which here is $\sqrt{3/4} \approx 0.87$, not the 0.58 used in (a) and (d).

8. Let $A \in \mathbb{R}^{10\times 10}$ with $\operatorname{rank}(A) = 7$, and consider

$$\text{minimize} \quad \|A - X\| \qquad \text{subject to} \qquad \operatorname{rank}(X) = k,$$

over $X \in \mathbb{R}^{10\times 10}$, where the constraint asks for rank exactly k, not at most k. Let B be a solution for $k = 6$ and C a solution for $k = 8$. Which of the following is true?

(a) Both B and C exist
(b) **B exists, and C does not exist**
(c) B exists, and C may or may not exist, depending on A
(d) Each of B and C may or may not exist, depending on A

Solution. (b). Write $A = \sum_{i=1}^{7} \sigma_i u_i v_i^\mathsf{T}$ with $\sigma_1 \ge \cdots \ge \sigma_7 > 0$.

For $k = 6$ the best rank 6 or less approximation is the truncated SVD $\hat{A} = \sum_{i=1}^{6} \sigma_i u_i v_i^\mathsf{T}$, at distance σ_7. Since $\sigma_6 > 0$, its rank is exactly 6, so it meets the equality constraint as well, and $B = \hat{A}$ (the only other candidates have rank below 6 and are further away). The equality costs nothing here: the best approximation uses all the rank it is allowed.

For $k = 8$ there is nothing to attain. Any X of rank 8 has $X \neq A$, so $\|A - X\| > 0$, yet you can get arbitrarily close: since $\operatorname{rank}(A) = 7 < 10$, pick unit vectors $u \perp \operatorname{range}(A)$ and $v \perp \operatorname{range}(A^\mathsf{T})$ and set $X_\epsilon = A + \epsilon u v^\mathsf{T}$. Then $Av = 0$, so $\operatorname{range}(X_\epsilon)$ is $\operatorname{range}(A)$ together with u, giving rank exactly 8, and $\|A - X_\epsilon\| = \epsilon$. The infimum is 0 and no rank 8 matrix achieves it, so there is no minimizer.

The set of matrices of rank *at most* k is closed, while the set of rank *exactly* k is not, and A is a limit point of the second set that lies outside it. Nothing changes if $\|\cdot\|$ is the Frobenius norm.

9. A centered data set of $N = 10^6$ points in $\mathbb{R}^{10000}$ is reduced to 20 dimensions in two ways.

   - *One shot:* keep the top 20 principal components of the data.
   - *Two stage:* keep the top 500 principal components, then run PCA again on the resulting 500-dimensional data and keep its top 20 principal components.

   How much variance do the two pipelines capture?

   (a) The two-stage pipeline captures less, because the first stage discards directions that the second one can no longer use
   (b) **Both capture exactly the same variance**
   (c) Both capture the same variance only if the data has rank at most 500
   (d) The two-stage pipeline captures more: dropping the weakest directions first clears away clutter, and the second PCA then works in the 500-dimensional space, where it can pick up structure in the data that was not visible in the original one

   Solution. (b). Let $q_1, \ldots, q_{10000}$ be the principal components of the data, with variances $\lambda_1 \geq \lambda_2 \geq \cdots$. The first stage replaces each point by its coordinates $z_i = (q_1^\mathsf{T} x_i, \ldots, q_{500}^\mathsf{T} x_i)$, and the covariance of the z_i is $Q_{500}^\mathsf{T} S Q_{500} = \operatorname{diag}(\lambda_1, \ldots, \lambda_{500})$: already diagonal, and already sorted. So the principal components of the reduced data are the coordinate vectors $e_1, \ldots, e_{500}$, and keeping its top 20 keeps exactly $q_1, \ldots, q_{20}$, which is what the one-shot pipeline does. Both capture $\lambda_1 + \cdots + \lambda_{20}$.

   The short version: principal subspaces are nested, so the top 20 directions already sit inside the top 500, and a first stage with a larger budget cannot discard anything the second stage wanted.

   Beating the one shot, as (d) claims, is impossible for a simpler reason: the one shot maximizes captured variance over *every* 20-dimensional subspace, and whatever the two-stage pipeline returns is one of those subspaces. The second PCA also gets no new view of the data. Passing to the coordinates z_i preserves lengths and inner products inside the top 500 subspace, so the second PCA is the same eigenvalue problem restricted to that subspace: it

re-sorts directions that were already there instead of finding new ones. Running something nonlinear after PCA is a different story, but PCA after PCA is just PCA.

None of this is a fact about dimension reduction in general, only about PCA with room to spare. Keep 500 *other* directions in the first stage, or whiten the 500 coordinates so that every variance becomes 1 and the ordering is destroyed, and the second PCA no longer reproduces the one-shot answer.

10. Let G be an undirected graph on $n \geq 4$ nodes with positive edge weights, let $L = D - W$ be its Laplacian, and let $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be the eigenvalues of L. Which of the following statements about λ_2 is correct?

    (a) **Adding an edge to G cannot decrease λ_2**
    (b) $\lambda_2 = 0$ if and only if some node of G has degree zero
    (c) $\lambda_2 \geq \min_i d_i$, where d_i is the weighted degree of node i
    (d) Multiplying every edge weight by 2 leaves λ_2 unchanged

    Solution. (a). Since $L\mathbf{1} = 0$, the second smallest eigenvalue is

$$\lambda_2 = \min \left\{ x^{\mathsf{T}} L x \;\middle|\; \mathbf{1}^{\mathsf{T}} x = 0,\ \|x\| = 1 \right\} \quad \text{with} \quad x^{\mathsf{T}} L x = \frac{1}{2} \sum_{i,j} W_{ij} (x_i - x_j)^2,$$

    which is the problem the scalar spectral embedding solves. Putting a new edge of weight $w > 0$ between nodes k and l adds $w(x_k - x_l)^2 \geq 0$ to the objective for *every* x, so the minimum can only go up. It need not go up: if G has two components then $\lambda_2 = 0$, and an edge added inside one of them leaves $\lambda_2 = 0$, which is why the statement says cannot decrease rather than must increase.

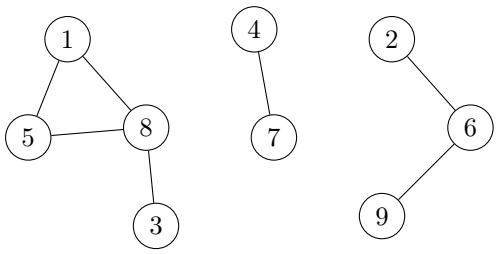
    (b) confuses two things. $\lambda_2 = 0$ exactly when G is disconnected, and the multiplicity of the eigenvalue 0 counts the components. Two disjoint edges have no node of degree zero and still have $\lambda_2 = 0$.

    (c) runs the wrong way. The bound is $\lambda_2 \leq \frac{n}{n-1} \min_i d_i$: take a node v of least degree and test $x = e_v - \frac{1}{n}\mathbf{1}$, which is centered, has $x^{\mathsf{T}} L x = d_v$ and $\|x\|^2 = 1 - \frac{1}{n}$. One weakly attached node caps λ_2 however well connected the rest of the graph is. For the 4-node path $\lambda_2 = 2 - \sqrt{2} \approx 0.59$, below every degree in the graph.

    (d) L is linear in the weights, so scaling them all by 2 doubles every eigenvalue. What scaling leaves alone is the Fiedler *vector*, and with it the spectral embedding and any ordering or bisection read off from it. λ_2 carries the units of the weights because it measures how strongly the graph is held together: it is zero when the graph falls apart, small when the graph is nearly disconnected, and $n\lambda_2$ is the least Dirichlet energy of any centered embedding with RMS value one.

2. Numerical computations *(20 points).* What the four parts of this problem have in common is not a topic but a task: each one asks you to carry a calculation through to actual numbers, and the four are drawn from four different corners of the course. Beyond that they are independent. Nothing in one part is needed in another, so you may do them in any order.

a) *(4 points)* **A graph Laplacian.** The undirected, unweighted graph below has 9 nodes. Let W be its adjacency matrix, D the diagonal matrix of degrees, and $L = D - W$ its Laplacian.



Find a matrix $Z \in \mathbb{R}^{9\times k}$, with k as large as possible, such that

- $LZ = 0$,
- every entry of Z is 0 or 1,
- the columns of Z are nonzero and mutually orthogonal.

Give k and Z, explain why no larger k is possible, and say in one sentence what the columns of Z represent.

Solution. $Lz = 0$ if and only if $z^\mathsf{T} Lz = \frac{1}{2}\sum_{i,j} W_{ij}(z_i - z_j)^2 = 0$, which forces $z_i = z_j$ across every edge, so z is constant on each connected component. The components here are $\{1, 3, 5, 8\}$, $\{2, 6, 9\}$, and $\{4, 7\}$. A vector in $\text{null}(L)$ with entries only 0 and 1 is therefore the indicator vector of a union of components, and two 0/1 vectors are orthogonal exactly when their supports are disjoint. So the columns of Z are indicators of pairwise disjoint, nonempty unions of components: there are only 3 components to go around, hence $k \le 3$. Taking one component per column attains it, with $k = 3$ and

$$Z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

up to a permutation of the columns. Column j is the indicator of the jth connected component: $Z_{ij} = 1$ when node i lies in component j. So Z is a component indicator matrix, and its columns are a basis of $\text{null}(L)$.

Two remarks. First, $Z^{\mathsf{T}}Z = \operatorname{diag}(4,3,2)$, the component sizes, so the columns cannot be made orthonormal while keeping 0/1 entries: rescaling column j by $1/\sqrt{|C_j|}$ is what orthonormalizes them. Second, $k = \dim \operatorname{null}(L) = 3$ here, so by rank-nullity $\operatorname{rank} L = 6$, even though the graph has 7 edges: the triangle 1–5–8 contributes a dependent third edge.

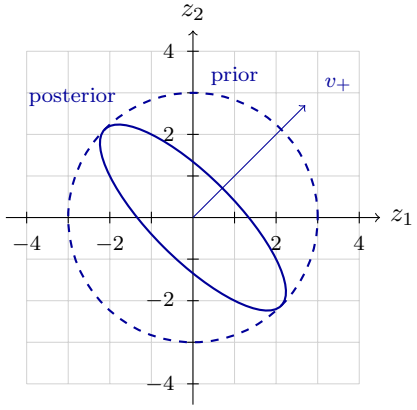
b) *(6 points)* **A noisy sensor.** A single sensor measures the sum of the two components of an unknown $x \in \mathbb{R}^2$:

$$y = Ax + w, \qquad A = \begin{bmatrix} 1 & 1 \end{bmatrix},$$

where $x \sim \mathcal{N}(0, \Sigma_x)$ with $\Sigma_x = 9I$, and the noise $w \sim \mathcal{N}(0, 9/4)$ is independent of x. You observe $y = 9$. Find the MMSE estimate $\hat{x}$ of x and the posterior covariance P. Then sketch the two ellipsoids

$$\{z : z^{\mathsf{T}}\Sigma_x^{-1}z \le 1\} \quad \text{and} \quad \{z : z^{\mathsf{T}}P^{-1}z \le 1\},$$

and give the direction in which the second has shrunk the most relative to the first and the direction in which it has shrunk the least, with the factor in each case. Show that the least shrinkage is in fact no shrinkage at all, and explain why the measurement left that direction untouched.



Solution. Here $A\Sigma_x A^{\mathsf{T}} + \Sigma_w = 18 + 9/4 = 81/4$ and $\Sigma_x A^{\mathsf{T}} = 9\begin{bmatrix} 1 \\ 1 \end{bmatrix}$, so

$$\hat{x} = \Sigma_x A^{\mathsf{T}}(A\Sigma_x A^{\mathsf{T}} + \Sigma_w)^{-1} y = 9\begin{bmatrix} 1 \\ 1 \end{bmatrix} \cdot \frac{4}{81} \cdot 9 = \begin{bmatrix} 4 \\ 4 \end{bmatrix},$$

$$P = \Sigma_x - \Sigma_x A^{\mathsf{T}}(A\Sigma_x A^{\mathsf{T}} + \Sigma_w)^{-1} A\Sigma_x = 9I - 4\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}.$$

The information form is just as quick: $P^{-1} = \Sigma_x^{-1} + A^{\mathsf{T}}\Sigma_w^{-1}A = \frac{1}{9}I + \frac{4}{9}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \frac{1}{9}\begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$, whose determinant is $\frac{1}{81}(25 - 16) = \frac{1}{9}$, giving the same P.

$\Sigma_x = 9I$ has the single eigenvalue 9, so the prior ellipsoid is a circle of radius 3 and every direction is a principal axis. For P, write

$$v_+ = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad v_- = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

The eigenvectors are $v_+/\sqrt{2}$ with eigenvalue $5 - 4 = 1$ and $v_-/\sqrt{2}$ with eigenvalue $5+4 = 9$, so the posterior ellipsoid has semi-axis $\sqrt{1} = 1$ along v_+ and semi-axis $\sqrt{9} = 3$ along v_-.

Comparing the two ellipsoids direction by direction, the semi-axis went from 3 to 1 along v_+, a shrinkage by a factor of 3, and from 3 to 3 along v_-, a factor of 1. So the most shrinkage is along v_+ and the least is along v_-, where there is none: the prior circle is squashed onto an ellipse in one direction only, and the two directions are orthogonal because P is symmetric.

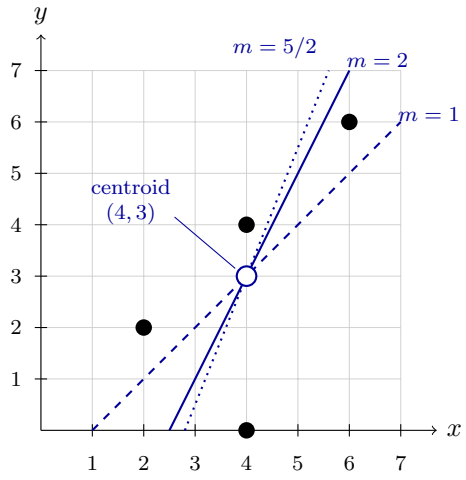
The reason nothing happened along v_- is that the measurement never saw it. Since $Av_- = 0$, the component of x along v_- contributes nothing to y, so no value of y can say anything about it: the posterior variance in that direction is still the prior 9, and $v_-^\mathsf{T}\hat{x} = 0$ whatever y turns out to be. All the information arrives along $v_+ = A^\mathsf{T}$, which is the one direction the sensor responds to.

Sanity check on $\hat{x}$: the sensor reads $y = 9$, and splitting it evenly would give $\begin{bmatrix} 4.5 \\ 4.5 \end{bmatrix}$; the noise shrinks the estimate to $\begin{bmatrix} 4 \\ 4 \end{bmatrix}$.

c) *(6 points)* **Three ways to fit a line.** You are given the four data points

$$(2, 2), \quad (4, 0), \quad (4, 4), \quad (6, 6),$$

plotted below. Fit a line $y = mx + b$ to them three times, determining both the slope m and the intercept b in each case. The first fit minimizes the sum of squared *vertical* distances from the four points to the line, the second minimizes the sum of squared *horizontal* distances, and the third minimizes the sum of squared *perpendicular* distances. Give the three lines, and say in one sentence why they are not all the same.



Solution. Write $p_i = (x_i, y_i)$. The five sums we need are

$$\sum x_i = 16, \quad \sum y_i = 12, \quad \sum x_i^2 = 72, \quad \sum y_i^2 = 56, \quad \sum x_i y_i = 56,$$

so the centroid is $(\bar{x}, \bar{y}) = (4, 3)$.

Vertical distances. With unknowns $\theta = (m, b)$ this is least squares, $y \approx A\theta$ with $A = \begin{bmatrix} x & \mathbf{1} \end{bmatrix}$, so

$$A^{\mathsf{T}}A = \begin{bmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & 4 \end{bmatrix} = \begin{bmatrix} 72 & 16 \\ 16 & 4 \end{bmatrix}, \qquad A^{\mathsf{T}}y = \begin{bmatrix} \sum x_i y_i \\ \sum y_i \end{bmatrix} = \begin{bmatrix} 56 \\ 12 \end{bmatrix}.$$

Since $\det A^{\mathsf{T}}A = 288 - 256 = 32$,

$$\theta = (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}y = \frac{1}{32}\begin{bmatrix} 4 & -16 \\ -16 & 72 \end{bmatrix}\begin{bmatrix} 56 \\ 12 \end{bmatrix} = \frac{1}{32}\begin{bmatrix} 32 \\ -32 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

so the fit is $y = x - 1$.

Horizontal distances. Now x is the dependent variable, so fit $x \approx cy + d$ with $B = \begin{bmatrix} y & \mathbf{1} \end{bmatrix}$:

$$B^{\mathsf{T}}B = \begin{bmatrix} 56 & 12 \\ 12 & 4 \end{bmatrix}, \quad B^{\mathsf{T}}x = \begin{bmatrix} 56 \\ 16 \end{bmatrix}, \quad \det B^{\mathsf{T}}B = 224 - 144 = 80,$$

$$\begin{bmatrix} c \\ d \end{bmatrix} = \frac{1}{80}\begin{bmatrix} 4 & -12 \\ -12 & 56 \end{bmatrix}\begin{bmatrix} 56 \\ 16 \end{bmatrix} = \frac{1}{80}\begin{bmatrix} 32 \\ 224 \end{bmatrix} = \begin{bmatrix} 2/5 \\ 14/5 \end{bmatrix}.$$

So $x = \frac{2}{5}y + \frac{14}{5}$, and solving for y, the fit is $y = \frac{5}{2}x - 7$.

Perpendicular distances. This is PCA. Center the data,

$$\tilde{p}_i: \quad (-2, -1),\ (0, -3),\ (0, 1),\ (2, 3), \qquad C = \sum_i \tilde{p}_i \tilde{p}_i^{\mathsf{T}} = \begin{bmatrix} 8 & 8 \\ 8 & 20 \end{bmatrix},$$

and take the principal direction: $\lambda^2 - 28\lambda + 96 = (\lambda - 24)(\lambda - 4)$, and for $\lambda = 24$, $(8-24)u_1 + 8u_2 = 0$ gives $u_2/u_1 = 2$. The best-fit line passes through the centroid along that direction, so $y - 3 = 2(x - 4)$, that is, $y = 2x - 5$.

The three differ because each charges the error to something different: the first treats x as exact and puts all the error in y, the second does the reverse, and the third treats the two coordinates symmetrically. The slopes are $1 < 2 < \frac{5}{2}$, with the perpendicular fit between the other two, as it always is.

Two things worth noticing. All three lines pass through the centroid $(4, 3)$: for the first two this is the second normal equation, $\sum_i (y_i - mx_i - b) = 0$, and for the third it is where the centering came from. And the reason the three answers differ at all is the single point $(4, 0)$: the other three points lie exactly on $y = x$. That point sits directly below the centroid, so it leaves $C_{xx} = 8$ and $C_{xy} = 8$ untouched and only inflates C_{yy} from 8 to 20. The vertical fit uses C_{xy}/C_{xx} and so keeps slope 1; the horizontal fit uses C_{yy}/C_{xy} and steepens to $\frac{5}{2}$; the perpendicular fit uses both and lands in between.

d) *(4 points)* **A trade-off between two objectives.** With $x \in \mathbb{R}^2$, let

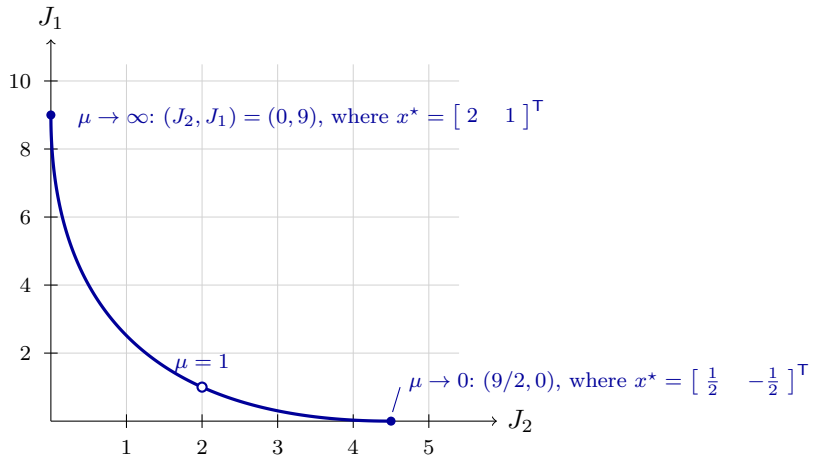
$$A = \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad b = 0, \qquad C = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad d = \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$

and consider the two objectives

$$J_1 = \|Ax - b\|^2, \qquad J_2 = \|Cx - d\|^2.$$

For $\mu > 0$, let $x^\star(\mu)$ be the minimizer of the composite objective $J_1 + \mu J_2$. Find $x^\star(\mu)$ explicitly as a function of μ. Then sketch the optimal trade-off curve in the J_1–J_2 plane, and mark the two points on it that correspond to $\mu \to 0$ and $\mu \to \infty$, giving their coordinates.

Extra credit (2 points): find the equation of the trade-off curve, in the form $J_1 = f(J_2)$.



Solution. Here $J_1 = (x_1 + x_2)^2$ and $J_2 = (x_2 - 1)^2 + (x_1 - 2)^2$. Minimizing $J_1 + \mu J_2 = \|Ax - b\|^2 + \mu\|Cx - d\|^2$ is the least-squares problem with the stacked matrix $\begin{bmatrix} A \\ \sqrt{\mu}\, C \end{bmatrix}$, so its normal equations are

$$\big(A^\mathsf{T}A + \mu C^\mathsf{T}C\big)x = A^\mathsf{T}b + \mu C^\mathsf{T}d.$$

Now C swaps the two coordinates, so $C^\mathsf{T}C = I$ and $C^\mathsf{T}d = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, while $A^\mathsf{T}A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and $A^\mathsf{T}b = 0$. The equations are

$$\begin{bmatrix} 1+\mu & 1 \\ 1 & 1+\mu \end{bmatrix} x = \mu \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \qquad \det = (1+\mu)^2 - 1 = \mu(\mu+2),$$

so for $\mu > 0$

$$x^\star(\mu) = \frac{\mu}{\mu(\mu+2)} \begin{bmatrix} 1+\mu & -1 \\ -1 & 1+\mu \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \frac{1}{\mu+2} \begin{bmatrix} 2\mu+1 \\ \mu-1 \end{bmatrix}.$$

The two ends. As $\mu \to \infty$, $x^\star \to \begin{bmatrix} 2 & 1 \end{bmatrix}^\mathsf{T}$, which solves $Cx = d$ exactly, so $J_2 = 0$ and $J_1 = (2+1)^2 = 9$. As $\mu \to 0$, $x^\star \to \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \end{bmatrix}^\mathsf{T}$, which has $x_1 + x_2 = 0$ and so $J_1 = 0$, with $J_2 = \left(\frac{1}{2} - 2\right)^2 + \left(-\frac{1}{2} - 1\right)^2 = \frac{9}{4} + \frac{9}{4} = \frac{9}{2}$. (At $\mu = 0$ the objective is J_1 alone, which every x on the line $x_1 + x_2 = 0$ minimizes; the limit picks out the one among them that is best for J_2.) So the curve runs from $(J_2, J_1) = (0, 9)$ down to $(9/2, 0)$, as sketched. Points above and to the right of it are achievable and points below and to the left are not.

Extra credit. From the formula above, $x_1 + x_2 = \frac{3\mu}{\mu+2}$ and $x_1 - 2 = x_2 - 1 = \frac{-3}{\mu+2}$, so

$$J_1(\mu) = \frac{9\mu^2}{(\mu+2)^2}, \qquad J_2(\mu) = \frac{18}{(\mu+2)^2}.$$

Hence $\sqrt{J_1} = \frac{3\mu}{\mu+2}$ and $\sqrt{2J_2} = \frac{6}{\mu+2}$, and these two add to 3. The trade-off curve is therefore

$$\sqrt{J_1} + \sqrt{2J_2} = 3, \qquad \text{that is} \qquad J_1 = \left(3 - \sqrt{2J_2}\right)^2 = 9 - 6\sqrt{2J_2} + 2J_2,$$

for $0 \le J_2 \le 9/2$. It is decreasing and convex, and a short computation gives $dJ_1/dJ_2 = -\mu$ at the point $x^\star(\mu)$, as it must be: the level sets of $J_1 + \mu J_2$ are lines of slope $-\mu$ in this plane. That is why the curve leaves the J_1 axis vertically ($\mu \to \infty$) and meets the J_2 axis flat ($\mu \to 0$), and why the slope is -1 at the point $(2, 1)$ where $\mu = 1$.

3. Bringing back a population *(20 points).* A discrete-time linear dynamical system has a *state* $x(t) \in \mathbb{R}^n$ at times $t = 0, 1, 2, \ldots$ and an *input* $u(t) \in \mathbb{R}^m$ that we choose at each time. The state evolves according to

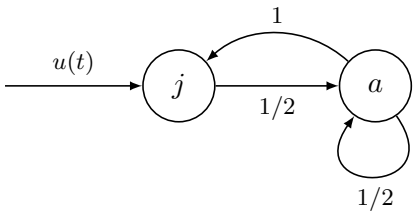
$$x(t+1) = Ax(t) + Bu(t), \qquad A \in \mathbb{R}^{n \times n}, \quad B \in \mathbb{R}^{n \times m},$$

so once $x(0)$ and the inputs are given, the whole trajectory is determined.

A nature reserve holds a threatened bird species, counted once a year just after the young have fledged. Let $j(t)$ be the number of juveniles, meaning birds in their first year, and let $a(t)$ be the number of adults at the count in year t. In our model, from one count to the next each adult leaves one young that survives to the next count, half of the juveniles survive and become adults, and half of the adults survive.

The reserve can also bring birds in from a captive breeding centre. Let $u(t)$ be the number of captive-bred young released just before the count in year $t+1$, so the effect of $u(t)$ shows up in $j(t+1)$, one count later, and from then on those birds face the same odds as the wild ones.

Take the state to be $x(t) = \begin{bmatrix} j(t) \\ a(t) \end{bmatrix}$. The halves in the model will sometimes give counts that are not whole numbers, which is nothing to worry about: rates like half of the adults surviving are averages over a population, not statements about one bird. Treat $j(t)$, $a(t)$ and $u(t)$ as real numbers and do not round anything.



The reserve was founded with 24 adult birds, released just before the year-0 count, so $x(0) = \begin{bmatrix} 0 \\ 24 \end{bmatrix}$. The recovery plan calls for 28 juveniles and 28 adults at the year-3 count, 56 birds in all; part (b) explains why the plan asks for that particular split. Captive-bred birds are the expensive part of the program: a large release in a single year costs more per bird than a small one, since the birds have to come from farther afield and more of them are lost to dispersal and competition on arrival. We model the cost of a release of u birds in one year as u^2. The model ignores crowding, which is reasonable at these numbers.

a) *(4 points)* Write the model in the form $x(t+1) = Ax(t) + Bu(t)$, and give A and B. Then suppose the reserve releases no birds at all, so $u(t) = 0$. Compute $x(1)$, $x(2)$ and $x(3)$, and say whether the reserve reaches the plan on its own.

Solution. The two counts at the next census are $j(t+1) = a(t) + u(t)$ and $a(t+1) = \frac{1}{2}j(t) + \frac{1}{2}a(t)$, so

$$A = \begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The top row says juveniles come from adults, one each, plus the birds released; the bottom row says adults come from the half of last year's juveniles that mature and the half of the adults that survive. With $u = 0$ and $x(0) = \begin{bmatrix} 0 \\ 24 \end{bmatrix}$,

$$x(1) = \begin{bmatrix} 24 \\ 12 \end{bmatrix}, \qquad x(2) = \begin{bmatrix} 12 \\ 18 \end{bmatrix}, \qquad x(3) = \begin{bmatrix} 18 \\ 15 \end{bmatrix},$$

with totals 24, 36, 30, 33. So no: the founders produce a burst of young in the first year, after which the counts swing and settle in the low thirties, well short of 56. Nothing in the model pushes the population toward the plan, which is why the reserve has to bring birds in. (Part (b) explains both the swing and the level it settles at.)

b) *(5 points)* Suppose the program has ended, so $u(t) = 0$ from then on. Find the eigenvalues and eigenvectors of A, say whether A is diagonalizable, and give a closed form for the matrix power A^t. What do these say about the population: does it grow, and what does it look like in the long run? What is special about the state $\begin{bmatrix} 28 & 28 \end{bmatrix}^{\mathsf{T}}$ the plan aims at?

Solution. The characteristic polynomial is $\det(A - \lambda I) = \lambda^2 - \frac{1}{2}\lambda - \frac{1}{2} = (\lambda - 1)\big(\lambda + \frac{1}{2}\big)$, so the eigenvalues are 1 and $-1/2$. For $\lambda = 1$, $(A - I)v = \begin{bmatrix} -1 & 1 \\ 1/2 & -1/2 \end{bmatrix} v = 0$ gives $v = \begin{bmatrix} 1 & 1 \end{bmatrix}^{\mathsf{T}}$; for $\lambda = -1/2$, $(A + \frac{1}{2}I)v = \begin{bmatrix} 1/2 & 1 \\ 1/2 & 1 \end{bmatrix} v = 0$ gives $v = \begin{bmatrix} 2 & -1 \end{bmatrix}^{\mathsf{T}}$. The eigenvalues are distinct, so the eigenvectors are independent and A is diagonalizable: with

$$P = \begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix}, \qquad P^{-1} = \frac{1}{3}\begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix}, \qquad A = P\begin{bmatrix} 1 & 0 \\ 0 & -1/2 \end{bmatrix} P^{-1},$$

we get

$$A^t = P\begin{bmatrix} 1 & 0 \\ 0 & (-1/2)^t \end{bmatrix} P^{-1} = \frac{1}{3}\begin{bmatrix} 1 + 2\big(-\frac{1}{2}\big)^t & 2 - 2\big(-\frac{1}{2}\big)^t \\ 1 - \big(-\frac{1}{2}\big)^t & 2 + \big(-\frac{1}{2}\big)^t \end{bmatrix}.$$

(A quick check: at $t = 1$ this gives A back.)

The population does not grow. The eigenvalue 1 says it sits exactly at replacement: each generation just replaces itself. The eigenvalue $-1/2$ is a transient that halves and flips sign every year, which is the swing seen in part (a): founding with adults only puts a large component along $\begin{bmatrix} 2 & -1 \end{bmatrix}^{\mathsf{T}}$, and it decays away. In the long run $A^t \to \frac{1}{3}\begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}$, so

$$x(t) \longrightarrow \frac{j(0) + 2a(0)}{3}\begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

equal numbers of juveniles and adults, at a level set only by $j(0) + 2a(0)$. An adult counts double a juvenile because only half of juveniles reach adulthood. Here $j(0) + 2a(0) = 48$, so the reserve settles at $\begin{bmatrix} 16 & 16 \end{bmatrix}^{\mathsf{T}}$, 32 birds: it holds, but it never reaches 56.

In fact $j(t) + 2a(t)$ is unchanged by the free dynamics, which is worth noting for the parts that follow: it is 48 at all four counts in part (a).

The plan's state is $\begin{bmatrix} 28 & 28 \end{bmatrix}^{\mathsf{T}} = 28 \begin{bmatrix} 1 & 1 \end{bmatrix}^{\mathsf{T}}$, exactly the eigenvector for $\lambda = 1$. So the day the releases stop, the population stays there with no transient at all: the plan hands over 56 birds in the one split of the two stages that holds. Any other split of 56 would drift.

c) *(5 points)* Starting from $x(0) = \begin{bmatrix} 0 \\ 24 \end{bmatrix}$, you want to meet the plan exactly at the year-3 count:

$$x(3) = \begin{bmatrix} 28 \\ 28 \end{bmatrix}.$$

Show that $x(3) = A^3 x(0) + \mathcal{C}u$, where $u = \begin{bmatrix} u(0) & u(1) & u(2) \end{bmatrix}^{\mathsf{T}}$, and give the matrix $\mathcal{C} \in \mathbb{R}^{2\times 3}$ explicitly. Explain why such a release plan exists, and why there is more than one. Then find the cheapest plan, the one that minimizes $u(0)^2 + u(1)^2 + u(2)^2$, and describe it in words.

Solution. Iterating the recursion,

$$x(1) = Ax(0) + Bu(0), \qquad x(2) = A^2 x(0) + ABu(0) + Bu(1),$$

$$x(3) = A^3 x(0) + A^2 Bu(0) + ABu(1) + Bu(2),$$

so $x(3) = A^3 x(0) + \mathcal{C}u$ with

$$\mathcal{C} = \begin{bmatrix} A^2B & AB & B \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1 \\ 1/4 & 1/2 & 0 \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 2 & 0 & 4 \\ 1 & 2 & 0 \end{bmatrix},$$

using $B = \begin{bmatrix} 1 & 0 \end{bmatrix}^{\mathsf{T}}$, $AB = \begin{bmatrix} 0 & 1/2 \end{bmatrix}^{\mathsf{T}}$ and $A^2B = \begin{bmatrix} 1/2 & 1/4 \end{bmatrix}^{\mathsf{T}}$. Part (a) already gave $A^3x(0) = \begin{bmatrix} 18 & 15 \end{bmatrix}^{\mathsf{T}}$, so the condition on the plan is

$$\mathcal{C}u = \begin{bmatrix} 28 \\ 28 \end{bmatrix} - \begin{bmatrix} 18 \\ 15 \end{bmatrix} = \begin{bmatrix} 10 \\ 13 \end{bmatrix} = d.$$

Existence. The first two columns of $\mathcal{C}$ are independent, so $\operatorname{rank}\mathcal{C} = 2$ and $\operatorname{range}(\mathcal{C}) = \mathbb{R}^2$. Any year-3 count can be reached, including the plan.

More than one plan. $\mathcal{C}$ is 2×3 with rank 2, so $\dim \operatorname{null}(\mathcal{C}) = 1$: solving $2\alpha + 4\gamma = 0$ and $\alpha + 2\beta = 0$ gives $\operatorname{null}(\mathcal{C}) = \operatorname{span}\{\begin{bmatrix} 2 & -1 & -1 \end{bmatrix}^{\mathsf{T}}\}$. Adding any multiple of $\begin{bmatrix} 2 & -1 & -1 \end{bmatrix}^{\mathsf{T}}$ leaves the year-3 count untouched, so the plans that work form a line and we can pick the cheapest point on it.

Cheapest plan. This is a least-norm problem: minimize $\|u\|$ subject to $\mathcal{C}u = d$, with solution $u = \mathcal{C}^{\mathsf{T}}(\mathcal{C}\mathcal{C}^{\mathsf{T}})^{-1}d$. Here

$$\mathcal{C}\mathcal{C}^{\mathsf{T}} = \begin{bmatrix} 5/4 & 1/8 \\ 1/8 & 5/16 \end{bmatrix}, \qquad \det = \frac{3}{8}, \qquad (\mathcal{C}\mathcal{C}^{\mathsf{T}})^{-1} = \frac{8}{3}\begin{bmatrix} 5/16 & -1/8 \\ -1/8 & 5/4 \end{bmatrix},$$

so $(\mathcal{C}\mathcal{C}^{\mathsf{T}})^{-1}\begin{bmatrix}10\\13\end{bmatrix} = \frac{8}{3}\begin{bmatrix}3/2\\15\end{bmatrix} = \begin{bmatrix}4\\40\end{bmatrix}$ and

$$u = \mathcal{C}^{\mathsf{T}}\begin{bmatrix}4\\40\end{bmatrix} = \begin{bmatrix}1/2 & 1/4\\0 & 1/2\\1 & 0\end{bmatrix}\begin{bmatrix}4\\40\end{bmatrix} = \begin{bmatrix}12\\20\\4\end{bmatrix}, \qquad \|u\|^2 = 144 + 400 + 16 = 560.$$

(Pulling the 1/4 out of $\mathcal{C}$ first is easier: with $\mathcal{C} = \frac{1}{4}\mathcal{C}_0$ the same computation is $u = 4\mathcal{C}_0^{\mathsf{T}}(\mathcal{C}_0\mathcal{C}_0^{\mathsf{T}})^{-1}d$ with $\mathcal{C}_0\mathcal{C}_0^{\mathsf{T}} = \begin{bmatrix}20 & 2\\2 & 5\end{bmatrix}$ and $\det = 96$.)

So release twelve birds the first year, twenty the second and four the third, giving the counts

$$x(1) = \begin{bmatrix}36\\12\end{bmatrix}, \qquad x(2) = \begin{bmatrix}32\\24\end{bmatrix}, \qquad x(3) = \begin{bmatrix}28\\28\end{bmatrix}.$$

As a check, $u = \begin{bmatrix}12 & 20 & 4\end{bmatrix}^{\mathsf{T}}$ is orthogonal to $\begin{bmatrix}2 & -1 & -1\end{bmatrix}^{\mathsf{T}}$, which the least-norm solution must be: it has no component in $\text{null}(\mathcal{C})$.

It is worth seeing where the total comes from. By part (b) the free dynamics leave $j+2a$ alone, and each released bird raises it by 1, so every release plan that hits the year-3 target satisfies $u(0) + u(1) + u(2) = 84 - 48 = 36$: thirty-six birds, no matter what. A cost proportional to the number of birds could therefore not tell the plans apart at all. The only thing a plan can choose is the timing, and that is what the quadratic cost prices.

d) *(6 points)* The breeding centre has already published what it will send: 4 birds the first year, 20 the second, 12 the third. That schedule lands at $x(3) = \begin{bmatrix}32 & 26\end{bmatrix}^{\mathsf{T}}$ rather than on the plan. You have to meet the plan exactly, and you want to depart from the published schedule as little as possible. The first year's birds are already committed, so a change there counts double:

$$\begin{aligned}&\text{minimize} \quad 4\big(u(0)-4\big)^2 + \big(u(1)-20\big)^2 + \big(u(2)-12\big)^2\\&\text{subject to} \quad x(3) = \begin{bmatrix}28\\28\end{bmatrix}.\end{aligned}$$

Find the optimal release plan.

One route is to write the optimality conditions for this problem as a single linear equation $Wz = y$, which you could then solve for z. Giving W and y explicitly, saying what the unknown z is, and explaining how solving $Wz = y$ would give you the optimal plan is worth 3 of the 6 points on its own: you do not have to solve that system, which may well be larger than is reasonable to do on paper. Full credit is for the plan itself, by any method you like. If you can find it some other way, do that and skip $Wz = y$ entirely.

Solution. This is norm minimization with an equality constraint. Since $4(u(0)-4)^2 = (2u(0)-8)^2$, the objective is $\|Ru - r\|^2$ with

$$R = \begin{bmatrix}2 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}, \qquad r = \begin{bmatrix}8\\20\\12\end{bmatrix},$$

and the constraint is the one from part (c), $\mathcal{C}u = d$ with $\mathcal{C} = \begin{bmatrix} 1/2 & 0 & 1 \\ 1/4 & 1/2 & 0 \end{bmatrix}$ and $d = \begin{bmatrix} 10 & 13 \end{bmatrix}^{\mathsf{T}}$. Part (c) is the special case $R = I$, $r = 0$ of this problem.

The linear system. With a Lagrange multiplier $\lambda \in \mathbb{R}^2$ for the constraint, the optimality conditions are $R^{\mathsf{T}}Ru - R^{\mathsf{T}}r + \mathcal{C}^{\mathsf{T}}\lambda = 0$ and $\mathcal{C}u = d$, that is $Wz = y$ with

$$W = \begin{bmatrix} R^{\mathsf{T}}R & \mathcal{C}^{\mathsf{T}} \\ \mathcal{C} & 0 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 & 1/2 & 1/4 \\ 0 & 1 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & 1 & 0 \\ 1/2 & 0 & 1 & 0 & 0 \\ 1/4 & 1/2 & 0 & 0 & 0 \end{bmatrix}, \qquad z = \begin{bmatrix} u \\ \lambda \end{bmatrix},$$

$$y = \begin{bmatrix} R^{\mathsf{T}}r \\ d \end{bmatrix} = \begin{bmatrix} 16 \\ 20 \\ 12 \\ 10 \\ 13 \end{bmatrix}.$$

If W is invertible, $z = W^{-1}y$ and the optimal plan is the first three entries of z; the last two are the multipliers, which we do not need.

Solving it without inverting W. Eliminate the constraint using part (c) instead: every release plan that meets the target has the form

$$u = \begin{bmatrix} 12 \\ 20 \\ 4 \end{bmatrix} + \gamma \begin{bmatrix} 2 \\ -1 \\ -1 \end{bmatrix}, \qquad \gamma \in \mathbb{R},$$

so the objective becomes a quadratic in the one remaining variable,

$$4\big(8 + 2\gamma\big)^2 + \gamma^2 + \big(8 + \gamma\big)^2,$$

whose derivative is $16(8 + 2\gamma) + 2\gamma + 2(8 + \gamma) = 36\gamma + 144$. So $\gamma = -4$ and

$$u = \begin{bmatrix} 12 \\ 20 \\ 4 \end{bmatrix} - 4 \begin{bmatrix} 2 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 4 \\ 24 \\ 8 \end{bmatrix},$$

with $\|Ru - r\|^2 = 0 + 16 + 16 = 32$, the smallest possible. Three unknowns and two constraints leave one degree of freedom, so this route replaces a 5×5 system with one derivative.

Compared with the published schedule, the first year is left alone and four birds move from the third year into the second. That is exactly the correction the plan needs: a bird released in year 1 contributes $AB = \begin{bmatrix} 0 & 1/2 \end{bmatrix}^{\mathsf{T}}$ to the year-3 count, since half of that cohort are adults by then, while a bird released in year 2 contributes $B = \begin{bmatrix} 1 & 0 \end{bmatrix}^{\mathsf{T}}$, because those birds are still juveniles at the year-3 count. Moving four birds from the third year to the second therefore trades four juveniles for two adults, turning the published schedule's $\begin{bmatrix} 14 & 11 \end{bmatrix}^{\mathsf{T}}$ into the required $\begin{bmatrix} 10 & 13 \end{bmatrix}^{\mathsf{T}}$. The centre's schedule

sends the right number of birds at the wrong times: too many juveniles, too few adults. The resulting counts are

$$x(1) = \begin{bmatrix} 28 \\ 12 \end{bmatrix}, \qquad x(2) = \begin{bmatrix} 36 \\ 20 \end{bmatrix}, \qquad x(3) = \begin{bmatrix} 28 \\ 28 \end{bmatrix}.$$

The first year coming out exactly at its published value is a coincidence of these numbers, not something the double weight guarantees. As a check on the conditions above, this u gives $R^{\mathsf{T}}Ru - R^{\mathsf{T}}r = \begin{bmatrix} 0 & 4 & -4 \end{bmatrix}^{\mathsf{T}}$, and $\lambda = \begin{bmatrix} 4 & -8 \end{bmatrix}^{\mathsf{T}}$ solves $\mathcal{C}^{\mathsf{T}}\lambda = -\begin{bmatrix} 0 & 4 & -4 \end{bmatrix}^{\mathsf{T}}$. So the vector $z = \begin{bmatrix} 4 & 24 & 8 & 4 & -8 \end{bmatrix}^{\mathsf{T}}$ does satisfy $Wz = y$.

4. The gain of a block matrix *(20 points).* Let $A, B, C, D \in \mathbb{R}^{n\times n}$, and assemble them into the block matrix

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathbb{R}^{2n\times 2n}.$$

Throughout, $\|\cdot\|$ is the matrix norm, that is, the largest singular value, which we read as a gain:

$$\|A\| = \sigma_{\max}(A) = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}.$$

We know the gain of each of the four blocks, and the question is what those four numbers tell us about M itself.

a) *(8 points)* Show that

$$\max\{\|A\|,\ \|B\|,\ \|C\|,\ \|D\|\} \ \leq\ \|M\|,$$

and that

$$\|M\| \ \leq\ \sqrt{\|A\|^2 + \|B\|^2 + \|C\|^2 + \|D\|^2}.$$

Each of the two bounds is worth 4 points.

Solution. Some shorthand keeps the lines short. Abbreviate the four gains as

$$a = \|A\|, \qquad b = \|B\|, \qquad c = \|C\|, \qquad d = \|D\|,$$

and split a vector $z \in \mathbb{R}^{2n}$ into its two halves, $z = \begin{bmatrix} x \\ y \end{bmatrix}$ with $x, y \in \mathbb{R}^n$, so that

$$Mz = \begin{bmatrix} Ax + By \\ Cx + Dy \end{bmatrix}, \qquad \|z\|^2 = \|x\|^2 + \|y\|^2.$$

The lower bound. Each block is what M does between two coordinate subspaces: with

$$E = \begin{bmatrix} I \\ 0 \end{bmatrix}, \qquad F = \begin{bmatrix} 0 \\ I \end{bmatrix} \in \mathbb{R}^{2n\times n},$$

we have $A = E^{\mathsf{T}}ME$, $B = E^{\mathsf{T}}MF$, $C = F^{\mathsf{T}}ME$ and $D = F^{\mathsf{T}}MF$. The columns of E and of F are orthonormal, so $\|E\| = \|F\| = 1$, and a matrix and its transpose have the same norm. Submultiplicativity then gives

$$a = \|E^{\mathsf{T}}ME\| \leq \|E^{\mathsf{T}}\|\,\|M\|\,\|E\| = \|M\|,$$

and the same for b, c and d, so $\max\{a, b, c, d\} \leq \|M\|$.

If you prefer to see it directly: pick a unit x with $\|Ax\| = a$. Then $z = \begin{bmatrix} x \\ 0 \end{bmatrix}$ is a unit vector and $\|Mz\|^2 = \|Ax\|^2 + \|Cx\|^2 \geq a^2$, so $\|M\| \geq a$. Taking $z = \begin{bmatrix} x \\ 0 \end{bmatrix}$ with $\|Cx\| = c$, and $z = \begin{bmatrix} 0 \\ y \end{bmatrix}$ for the other two blocks, covers the rest.

The upper bound. Let $z = \begin{bmatrix} x \\ y \end{bmatrix}$ with $\|z\| = 1$. By the triangle inequality, then the definition of the gains, then Cauchy–Schwarz on the pairs (a, b) and $(\|x\|, \|y\|)$,

$$\|Ax + By\| \le \|Ax\| + \|By\| \le a\|x\| + b\|y\| \le \sqrt{a^2 + b^2}\,\sqrt{\|x\|^2 + \|y\|^2} = \sqrt{a^2 + b^2},$$

and in the same way $\|Cx + Dy\| \le \sqrt{c^2 + d^2}$. Adding the squares,

$$\|Mz\|^2 = \|Ax + By\|^2 + \|Cx + Dy\|^2 \le a^2 + b^2 + c^2 + d^2,$$

for every unit z, which is the claim.

b) *(6 points)* Show that

$$\|M\| \;\le\; \left\| \begin{bmatrix} \|A\| & \|B\| \\ \|C\| & \|D\| \end{bmatrix} \right\|,$$

where the matrix on the right-hand side is 2×2.

Solution. Keep the shorthand of part (a) and call the 2×2 matrix of gains

$$N = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Start from the two inequalities used in part (a), but stop before Cauchy–Schwarz:

$$\|Ax + By\| \le a\|x\| + b\|y\|, \qquad \|Cx + Dy\| \le c\|x\| + d\|y\|.$$

Collect the lengths of the two halves in the single vector $u = \begin{bmatrix} \|x\| & \|y\| \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^2$, which satisfies $\|u\| = \|z\|$. The two right-hand sides are exactly the entries of Nu, and they are nonnegative, so for a unit z,

$$\|Mz\|^2 \le (a\|x\| + b\|y\|)^2 + (c\|x\| + d\|y\|)^2 = \|Nu\|^2 \le \|N\|^2\|u\|^2 = \|N\|^2.$$

Hence $\|M\| \le \|N\|$. The content of the bound is that the $2n \times 2n$ problem is controlled by a 2×2 one in which every block has been replaced by its gain.

c) *(6 points)* Parts (a) and (b) give two upper bounds on $\|M\|$. Decide which of the two is smaller, prove your answer, and say exactly when the two agree. This is worth 3 points.

Then, for the remaining 3 points, give one numerical example: choose n and give explicit numbers for A, B, C, D for which the lower bound of part (a) holds with equality, and exactly one of the two upper bounds holds with equality.

Solution. *Which bound is smaller.* The one in part (b), always. The quantity under the square root in part (a) is the sum of the squared entries of N, so that bound is nothing but $\|N\|_F$, and $\|N\| \le \|N\|_F$ for any matrix. In singular values: if $\sigma_1 \ge \sigma_2$ are those of N, then $\|N\| = \sigma_1$ while $\|N\|_F^2 = \sigma_1^2 + \sigma_2^2$. So the two bounds agree exactly when $\sigma_2 = 0$, that is when N has rank at most one, which for a 2×2 matrix means $ad = bc$; otherwise part (b) is strictly better.

Part (b) is also the best bound available from the four gains alone: with $n = 1$ and nonnegative blocks, M *is* N. And in the rank-one case the two upper bounds are not merely equal but both attained, for instance at $A = B = C = D = I$, where $z = \begin{bmatrix} x \\ x \end{bmatrix} / \sqrt{2}$ is amplified by $2 = \|N\| = \sqrt{a^2 + b^2 + c^2 + d^2}$.

An example where the bounds separate. Which of the two upper bounds can be the tight one is settled by the first half: if the bound of part (a) were exact then, sitting above the bound of part (b), it would force that one to be exact too. So the example must make the bound of part (b) exact and leave the bound of part (a) strict. Take $n = 1$, $B = C = 0$ and $A = 1$, $D = 1/2$:

$$M = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix} = N.$$

Here $a = 1$, $b = c = 0$, $d = 1/2$, so $\max\{a, b, c, d\} = 1$, and $\|M\| = \|N\| = 1$ because M is diagonal, while part (a) offers $\sqrt{1 + 1/4} = \sqrt{5}/2 \approx 1.12$. The lower bound and the bound of part (b) are exact; the bound of part (a) overshoots.

Any block diagonal M does the same in any dimension, since $\|Mz\|^2 = \|Ax\|^2 + \|Dy\|^2 \leq \max\{a, d\}^2 \|z\|^2$ gives $\|M\| = \max\{a, d\}$, against $\sqrt{a^2 + d^2}$ from part (a).

Nothing else works, and the reason is worth seeing. Suppose $\|N\| = \max\{a, b, c, d\}$, and say the largest of the four is a. Since $\|N\|$ is at least the norm of any column of N, and at least the norm of any row because $\|N\| = \|N^\mathsf{T}\|$,

$$\sqrt{a^2 + c^2} \leq \|N\| = a \qquad \text{and} \qquad \sqrt{a^2 + b^2} \leq \|N\| = a,$$

so $b = c = 0$. If instead the largest is b, the same argument forces $a = d = 0$. So the only examples are the ones above: two blocks vanish and the surviving two sit on a diagonal of M. Equivalently: if all four gains are positive, then the row and the column through the largest of them carry something else positive too, so $\|N\| > \max\{a, b, c, d\}$ and the two bounds cannot meet.

The obstruction is in N, not in M: the lower bound alone can be attained with all four blocks nonzero. Take $n = 2$ and let M be the cyclic shift on $\mathbb{R}^4$, $Me_1 = e_2$, $Me_2 = e_3$, $Me_3 = e_4$, $Me_4 = e_1$. Each of its four 2×2 blocks holds a single 1, so $a = b = c = d = 1$, and M is a permutation matrix, hence orthogonal, so $\|M\| = 1$ meets the lower bound exactly. But now $N = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and part (b) promises only $\|M\| \leq 2$: the four gains cannot see that the blocks act along different coordinates.

5. Isolating one effect *(20 points).* A retailer wants to know what a dollar of advertising is worth. Over n weeks you record the sales y_i, the advertising spend d_i, and the conditions of the week that are known to move sales on their own: the season, holidays, the price charged, and so on. Stack the first two into $y, d \in \mathbb{R}^n$, and put week i's conditions in the ith row of a feature matrix $\Phi \in \mathbb{R}^{n \times p}$; assume $p < n$ and Φ full rank. The records satisfy

$$y = \theta d + \Phi\alpha + u, \qquad d = \Phi\gamma + v, \qquad \Phi^\mathsf{T} u = 0, \qquad \Phi^\mathsf{T} v = 0, \qquad v \neq 0.$$

The scalar θ is what you want: the sales brought in by one more dollar of advertising. The vectors $\alpha, \gamma \in \mathbb{R}^p$ are of no interest in themselves, and may be large. The first says how the season and the price move sales by themselves; the second says how the marketing team sets its budget from that same information, since it advertises hardest in the weeks it expects to be strong. That second equation is the difficulty. It makes d and the columns of Φ move together, so a plain fit of y on d credits advertising with the sales the holidays would have brought anyway. The two leftovers carry nothing the features could have explained, which is what the orthogonality conditions say: v is the spending the team's own rule does not account for, and u is the part of sales that neither advertising nor the conditions explain. Write

$$P = \Phi(\Phi^\mathsf{T}\Phi)^{-1}\Phi^\mathsf{T}, \qquad M = I - P.$$

These should look familiar: for any vector $z \in \mathbb{R}^n$, Pz is the least-squares approximation of z by the columns of Φ, and $Mz = z - Pz$ is the residual, the part of z that no combination of features can account for. That is the one idea the whole problem runs on. If we can subtract off everything the features explain, what is left should be linked only by advertising. Part (a) works out what M does to the data; parts (b) and (c) ask what estimate of θ you get when you do that subtraction on one of y and d, and then on both.

a) *(8 points) The tool.* Before using M we should know exactly what it does, so this part is eight short questions about it, one point each.

i. Show that $M^\mathsf{T} = M$ and $M^2 = M$, so that M is an orthogonal projector.

ii. Give range(M), as a subspace built from Φ.

iii. Give null(M), as a subspace built from Φ.

iv. Give rank M, as a number in terms of n and p.

v. Let $\Phi = QR$ be a QR factorization. Express P in terms of Q, with no matrix inverse left in the formula.

vi. Give the eigenvalues of M, and the multiplicity of each in terms of n and p.

vii. Show that $Md = v$.

viii. Show that $My = \theta v + u$.

The last two identities are what the rest of the problem lives on, so they are worth reading: after the subtraction, all that survives of the spend is the team's own v, and all that survives of the sales is θ times that same v, plus u.

Solution. *(i).* Φ is full rank with $p < n$, so $\Phi^\mathsf{T}\Phi$ is invertible and P is defined. Since $\Phi^\mathsf{T}\Phi$ is symmetric, so is its inverse, and $P^\mathsf{T} = \Phi(\Phi^\mathsf{T}\Phi)^{-1}\Phi^\mathsf{T} = P$, while

$$P^2 = \Phi(\Phi^\mathsf{T}\Phi)^{-1}\Phi^\mathsf{T}\Phi(\Phi^\mathsf{T}\Phi)^{-1}\Phi^\mathsf{T} = \Phi(\Phi^\mathsf{T}\Phi)^{-1}\Phi^\mathsf{T} = P.$$

Hence $M^\mathsf{T} = M$ and $M^2 = I - 2P + P^2 = I - P = M$.

(ii), (iii), (iv). P is the orthogonal projector onto range(Φ), so M projects onto its orthogonal complement:

$$\text{range}(M) = \text{range}(\Phi)^\perp = \text{null}(\Phi^\mathsf{T}), \qquad \text{null}(M) = \text{range}(\Phi), \qquad \text{rank}\, M = n - p.$$

(v). With $\Phi = QR$, $Q \in \mathbb{R}^{n\times p}$ has orthonormal columns and $R \in \mathbb{R}^{p\times p}$ is invertible, so $Q^\mathsf{T}Q = I$ and

$$P = QR\big(R^\mathsf{T}Q^\mathsf{T}QR\big)^{-1}R^\mathsf{T}Q^\mathsf{T} = QR(R^\mathsf{T}R)^{-1}R^\mathsf{T}Q^\mathsf{T} = QRR^{-1}R^{-\mathsf{T}}R^\mathsf{T}Q^\mathsf{T} = QQ^\mathsf{T}.$$

So $M = I - QQ^\mathsf{T}$, which is the same projector read in an orthonormal basis for range(Φ).

(vi). Any vector in range(Φ) = range(Q) is annihilated by M and any vector orthogonal to it is left alone, so the eigenvalues of M are 1 with multiplicity $n - p$ and 0 with multiplicity p. (This is what a projector always looks like, and the multiplicities agree with rank $M = n - p$.)

(vii), (viii). $\Phi\gamma \in \text{range}(\Phi) = \text{null}(M)$ gives $M\Phi\gamma = 0$, and $\Phi^\mathsf{T}v = 0$ puts v in range(Φ)$^\perp$ = range(M), so $Mv = v$. Therefore

$$Md = M\Phi\gamma + Mv = v.$$

The same two facts give $M\Phi\alpha = 0$ and $Mu = u$, so

$$My = \theta Md + M\Phi\alpha + Mu = \theta v + u.$$

In words: multiplying by M throws away everything the known conditions can explain. What is left of d is the spending that was unusual for the week, and what is left of y is θ times that spending, plus u.

b) *(4 points) A first idea.* Sales move with the season and the price as well as with advertising, so clean that out of the sales: use My in place of y, and fit θ against the spend by least squares. This is an ordinary least-squares problem with one unknown and the single-column matrix d:

$$\hat\theta_{\text{naive}} = \operatorname*{argmin}_\theta \|My - \theta d\|^2.$$

Give a closed form for $\hat\theta_{\text{naive}}$. Now test the idea: set $u = 0$, so the records follow the model exactly and any error we see is the fault of the method rather than of noise. What we would like is $\hat\theta_{\text{naive}} = \theta$. Show that instead the estimate comes out proportional to θ,

$$\hat\theta_{\text{naive}} = \kappa\,\theta,$$

and find the factor κ. Say what κ is geometrically, and state the one condition under which the idea does return θ.

Solution. The first idea fails, but in a way that says exactly what went wrong. The fit itself is the usual one for a single-column matrix d:

$$\hat\theta_{\text{naive}} = \frac{d^\mathsf{T}My}{d^\mathsf{T}d}.$$

Using $My = \theta v + u$ from part (a) and $d = \Phi\gamma + v$, the orthogonality $\Phi^\mathsf{T} v = \Phi^\mathsf{T} u = 0$ kills the $\Phi\gamma$ term:

$$d^\mathsf{T} My = (\Phi\gamma + v)^\mathsf{T}(\theta v + u) = \theta\|v\|^2 + v^\mathsf{T} u, \qquad \text{so} \qquad \hat{\theta}_{\text{naive}} = \frac{\theta\|v\|^2 + v^\mathsf{T} u}{\|d\|^2}.$$

With $u = 0$ and $v = Md$ this is $\hat{\theta}_{\text{naive}} = \kappa\theta$ with

$$\kappa = \frac{\|Md\|^2}{\|d\|^2}.$$

Geometrically, $\kappa = 1 - \|Pd\|^2/\|d\|^2$ is the fraction of the energy of d that the features cannot explain, that is

$$\kappa = \sin^2 \angle\big(d, \text{range}(\Phi)\big).$$

It lies between 0 and 1, so the estimate is always pulled toward zero: $|\hat{\theta}_{\text{naive}}| \le |\theta|$, with equality if and only if $Pd = 0$, that is $\Phi^\mathsf{T} d = 0$: the team's spending has nothing to do with the season. Note what the bias does *not* depend on: α never appears. Cleaning the features out of y alone is not enough, because they are still sitting inside d.

c) *(8 points) A better idea.* Part (b) cleaned the features out of the sales but left them sitting inside the spend, which is what the factor κ is measuring. So clean both, and fit what is left of the sales against what is left of the spend. This is again a least-squares problem with one unknown, now with the single-column matrix Md:

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \|My - \theta\, Md\|^2.$$

Give a closed form for $\hat{\theta}$, and test it the same way as before: show that when $u = 0$ it recovers θ exactly, for every α and every γ. Then compare it with the approach you would probably have reached for first, which is to fit θ and α together in one least-squares problem and read off the first entry. Show that these give the same answer, that is, that $\hat{\theta}$ is the first component of

$$\operatorname*{argmin}_{\theta,\alpha} \left\| y - \begin{bmatrix} d & \Phi \end{bmatrix} \begin{bmatrix} \theta \\ \alpha \end{bmatrix} \right\|^2.$$

Solution. Again one unknown, now with the single-column matrix Md, and $M^\mathsf{T} M = M^2 = M$, so

$$\hat{\theta} = \frac{(Md)^\mathsf{T}(My)}{(Md)^\mathsf{T}(Md)} = \frac{d^\mathsf{T} My}{d^\mathsf{T} Md}.$$

With $u = 0$ part (a) gives $My = \theta v$ and $Md = v$, so

$$\hat{\theta} = \frac{\theta\|v\|^2}{\|v\|^2} = \theta,$$

whatever α and γ are. Both nuisance vectors have been projected away, which is why neither can bias the answer. (This is also where $v \neq 0$ is needed: $d^\mathsf{T} Md = \|Md\|^2 =$

$\|v\|^2$.) Read the estimate as a comparison of surprises: it lines up the sales that were unusual for the week against the spending that was unusual for the week, and everything the season explains has been taken out of both.

The joint fit. Write $A = \begin{bmatrix} d & \Phi \end{bmatrix}$. Its normal equations $A^{\mathsf{T}}A \begin{bmatrix} \hat{\theta} \\ \hat{\alpha} \end{bmatrix} = A^{\mathsf{T}}y$ split into blocks:

$$d^{\mathsf{T}}d\,\hat{\theta} + d^{\mathsf{T}}\Phi\hat{\alpha} = d^{\mathsf{T}}y, \qquad \Phi^{\mathsf{T}}d\,\hat{\theta} + \Phi^{\mathsf{T}}\Phi\hat{\alpha} = \Phi^{\mathsf{T}}y.$$

The second block gives $\hat{\alpha} = (\Phi^{\mathsf{T}}\Phi)^{-1}\Phi^{\mathsf{T}}(y - \hat{\theta}d)$. Substituting it into the first, and recognizing $\Phi(\Phi^{\mathsf{T}}\Phi)^{-1}\Phi^{\mathsf{T}} = P$,

$$d^{\mathsf{T}}d\,\hat{\theta} + d^{\mathsf{T}}Py - \hat{\theta}\,d^{\mathsf{T}}Pd = d^{\mathsf{T}}y, \qquad \text{that is} \qquad \hat{\theta}\,d^{\mathsf{T}}(I-P)d = d^{\mathsf{T}}(I-P)y,$$

so $\hat{\theta} = d^{\mathsf{T}}My/(d^{\mathsf{T}}Md)$, the same estimate. (Note A is full rank exactly when $d \notin \text{range}(\Phi)$, that is when $Md \neq 0$, which is the assumption $v \neq 0$ again.)

So fitting advertising and the features together in one least-squares problem is the same as cleaning the features out of both y and d and then doing a one-variable fit. Part (b) failed because it cleaned only one of the two. (This equivalence is the Frisch-Waugh-Lovell theorem.)

This problem is **entirely extra credit**. Attempt it only if you have time left over.

6. Extra credit: a huge matrix you can only multiply by *(8 points).* A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is given, with n in the millions. Its eigenvalues satisfy

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|,$$

and you want λ_1 together with an eigenvector belonging to it.

You never get to see the entries of A. What you have is a module that takes any vector $z \in \mathbb{R}^n$ and returns Az, and you may call it many times. Beyond that you can do the cheap things with vectors: add them, scale them, take inner products and norms.

Two limits come with the size, and both matter. Your machine has room for a few vectors of length n and nothing like room for n^2 numbers, so you cannot keep a copy of A, however you might come by it: forming it or factoring it is out of the question. And the module is not free. Every call to it, one matrix-vector product, costs you a dollar. Aim to get the job done cheaply: a good answer for a couple of hundred dollars would be a success.

Give a procedure that produces both λ_1 and an eigenvector for it, and justify it. Your answer should say what you compute and in what order, how the eigenvalue and the eigenvector are read off from what you have computed, why what you get belongs to λ_1 and not to one of the other eigenvalues, which of the assumptions stated above your justification uses and where, and what you have to assume about anything you are free to choose yourself. Since this has to run on a real machine, say something about the sizes of the numbers your procedure carries along, and about the work and the storage each stage costs.

You are not being asked for an exact answer after a fixed amount of arithmetic. A procedure whose two outputs can be brought as close as you like to λ_1 and to an eigenvector for it is a full answer.

Solution. *The procedure.* Take any $z_0 \in \mathbb{R}^n$ with $\|z_0\| = 1$, say a vector of random entries, and repeat: multiply by A, record a number, normalize. For $k = 0, 1, 2, \ldots$

$$w_k = Az_k, \qquad \mu_k = z_k^{\mathsf{T}} w_k, \qquad z_{k+1} = \frac{w_k}{\|w_k\|}.$$

Both answers come out of the same loop: $\mu_k \to \lambda_1$, and z_k itself converges to a unit eigenvector belonging to λ_1, so the vector you are already carrying around *is* the eigenvector. Note that $\mu_k = z_k^{\mathsf{T}} A z_k$ is the Rayleigh quotient of z_k, and it costs nothing extra: it is an inner product with the w_k we already have.

Why it works. A is symmetric, so it has an orthonormal set of eigenvectors $q_1, \ldots, q_n$ with $Aq_i = \lambda_i q_i$, and these are a basis of $\mathbb{R}^n$. Write the starting vector in that basis, $z_0 = \sum_i c_i q_i$ with $c_i = q_i^{\mathsf{T}} z_0$. Each pass multiplies by A and rescales, and rescaling does not change direction, so after k passes the direction is that of

$$A^k z_0 = \sum_i c_i \lambda_i^k q_i = \lambda_1^k \left(c_1 q_1 + \sum_{i \geq 2} c_i \left(\frac{\lambda_i}{\lambda_1} \right)^k q_i \right).$$

Both assumptions in the statement are used here, and it is worth naming them. Symmetry is what supplies the orthonormal eigenbasis, so that the expansion exists and the coefficients are

just inner products. The strict inequality $|\lambda_1| > |\lambda_2|$ is what makes every ratio $|\lambda_i/\lambda_1|$ with $i \geq 2$ strictly less than 1, so each of those terms dies geometrically and the bracket tends to $c_1 q_1$. The scalar λ_1^k in front is exactly what normalizing removes, so $z_k \to \pm q_1$: the iterates line up with the eigenvector belonging to λ_1, which answers the second half of the question. Then, by continuity of the Rayleigh quotient,

$$\mu_k = z_k^{\mathsf{T}} A z_k \;\longrightarrow\; q_1^{\mathsf{T}} A q_1 = \lambda_1.$$

The sign cannot be pinned down, and does not need to be: $-q_1$ is an eigenvector for λ_1 just as much as q_1 is.

The one thing to assume about the free choice is $c_1 = q_1^{\mathsf{T}} z_0 \neq 0$: if the starting vector happens to be orthogonal to q_1, the iteration converges to λ_2 and its eigenvector instead. The bad set is a hyperplane, so a random z_0 satisfies $c_1 \neq 0$ with probability 1.

Why the Rayleigh quotient and not something simpler. Because it gets the sign right. Since $z_k \to \pm q_1$ and the sign alternates when $\lambda_1 < 0$, the quadratic form is insensitive to it, while $\|Az_k\| \to |\lambda_1|$ loses the sign entirely. Reading λ_1 off one component, $(Az_k)_i/(z_k)_i$, is correct in the limit but unusable if that component of q_1 is near zero.

When to stop. Stop when the estimates settle, say $|\mu_k - \mu_{k-1}|$ below a tolerance. Better, the residual certifies the pair you are about to report: for symmetric A, if $\|Az_k - \mu_k z_k\| \leq \epsilon$ with $\|z_k\| = 1$, then A has an eigenvalue within ϵ of μ_k, and z_k is an approximate eigenvector for it. That residual is free as well, again from w_k.

On a real machine. Two points, both about scale. Without the normalization the entries of $A^k z_0$ grow like $|\lambda_1|^k$, or shrink like it if $|\lambda_1| < 1$, and overflow or underflow within a few dozen passes; dividing by $\|w_k\|$ each time keeps every number of order 1, and the λ_1^k we throw away is not needed, since the eigenvalue comes from the Rayleigh quotient rather than from the length. As for cost, a pass is one call to the module, so one dollar, plus a few operations on vectors of length n; the storage is two such vectors. Nothing anywhere is of size n^2, which is the only reason any of this is possible at n in the millions, where A itself would have 10^{12} entries.

How fast, and what it costs. The error in the direction shrinks by a factor of about $|\lambda_2/\lambda_1|$ per pass, and the error in μ_k by about the square of that, because for symmetric A the Rayleigh quotient is accurate to second order in the angle. So the method is quick when λ_1 is well separated and slow when $|\lambda_2|$ is close to $|\lambda_1|$. Against the budget: at a ratio of 0.9, two hundred passes, two hundred dollars, bring the direction error down by a factor of around 10^{-9}, and the eigenvalue error further still. A gap of 0.999 is another matter, and then one reaches for a better method than this one.

Strictness is not decoration. If $\lambda_1 = -\lambda_2$, both terms survive and nothing converges: with $A = \mathrm{diag}(1, -1)$ and $z_0 = \begin{bmatrix} 1 & 1 \end{bmatrix}^{\mathsf{T}}/\sqrt{2}$, every z_k is $\begin{bmatrix} 1 & \pm 1 \end{bmatrix}^{\mathsf{T}}/\sqrt{2}$ and $\mu_k = 0$ forever, which is not an eigenvalue of A at all.

This is the power method, and the estimate is the Rayleigh quotient. It is how the extreme eigenvalues of very large matrices are actually computed, and it needs nothing from A but the ability to multiply by it.