GIUSEPPE CALAFIORE AND LAURENT EL GHAOUI

OPTIMIZATION MODELS

EXERCISES

CAMBRIDGE

Contents

2. Vectors 4

3. Matrices 7

4. Symmetric matrices 11

5. Singular Value Decomposition 16

6. Linear Equations 21

7. Matrix Algorithms 26

8. Convexity 30

9. Linear, Quadratic and Geometric Models 35

10. Second-Order Cone and Robust Models 40

11. Semidefinite Models 44

12. Introduction to Algorithms 51

13. Learning from Data 57

14. Computational Finance 61

15. Control Problems 71

16. Engineering Design 75

2. Vectors

Exercise 2.1 (Subpaces and dimensions) Consider the set $\mathcal{S}$ of points such that

$$x_1 + 2x_2 + 3x_3 = 0, \;\; 3x_1 + 2x_2 + x_3 = 0.$$

Show that $\mathcal{S}$ is a subspace. Determine its dimension, and find a basis for it.

Exercise 2.2 (Affine sets and projections) Consider the set in $\mathbb{R}^3$ defined by the equation

$$\mathcal{P} = \left\{ x \in \mathbb{R}^3 \; : \; x_1 + 2x_2 + 3x_3 = 1 \right\}.$$

1. Show that the set $\mathcal{P}$ is an affine set of dimension 2. To this end, express it as $x^{(0)} + \text{span}(x^{(1)}, x^{(2)})$, where $x^{(0)} \in \mathcal{P}$, and $x^{(1)}, x^{(2)}$ are linearly independent vectors.

2. Find the minimum Euclidean distance from 0 to the set $\mathcal{P}$, and a point that achieves the minimum distance.

Exercise 2.3 (Angles, lines and projections)

1. Find the projection z of the vector $x = (2, 1)$ on the line that passes through $x_0 = (1, 2)$ and with direction given by vector $u = (1, 1)$.

2. Determine the angle between the following two vectors:

$$x = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \; y = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}.$$

   Are these vectors linearly independent?

Exercise 2.4 (Inner product) Let $x, y \in \mathbb{R}^n$. Under which condition on $\alpha \in \mathbb{R}^n$ does the function

$$f(x, y) = \sum_{k=1}^{n} \alpha_k x_k y_k$$

define an inner product on $\mathbb{R}^n$?

Exercise 2.5 (Orthogonality) Let $x, y \in \mathbb{R}^n$ be two unit-norm vectors, that is, such that $\|x\|_2 = \|y\|_2 = 1$. Show that the vectors $x - y$ and $x + y$ are orthogonal. Use this to find an orthogonal basis for the subspace spanned by x and y.

Exercise 2.6 (Norm inequalities)

1. Show that the following inequalities hold for any vector x:

$$\frac{1}{\sqrt{n}}\|x\|_2 \leq \|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2 \leq n\|x\|_\infty.$$

Hint: use the Cauchy–Schwartz inequality.

2. Show that for any nonzero vector x,

$$\text{card}(x) \geq \frac{\|x\|_1^2}{\|x\|_2^2},$$

where card(x) is the *cardinality* of the vector x, defined as the number of nonzero elements in x. Find vectors x for which the lower bound is attained.

Exercise 2.7 (Hölder inequality) Prove Hölder's inequality (2.4). *Hint:* consider the normalized vectors $u = x/\|x\|_p$, $v = y/\|y\|_q$, and observe that

$$|x^\top y| = \|x\|_p\|y\|_q \cdot |u^\top v| \leq \|x\|_p\|y\|_q \sum_k |u_k v_k|.$$

Then, apply Young's inequality (see Example 8.10) to the products $|u_k v_k| = |u_k||v_k|$.

Exercise 2.8 (Linear functions)

1. For a n-vector x, with $n = 2m - 1$ odd, we define the median of x as the scalar value x_a such that exactly n of the values in x are $\leq x_a$ and n are $\geq x_a$ (i.e., x_a leaves half of the values in x to its left, and half to its right). Now consider the function $f : \mathbb{R}^n \to \mathbb{R}$, with values $f(x) = x_a - \frac{1}{n}\sum_{i=1}^n x_i$. Express f as a scalar product, that is, find $a \in \mathbb{R}^n$ such that $f(x) = a^\top x$ for every x. Find a basis for the set of points x such that $f(x) = 0$.

2. For $\alpha \in \mathbb{R}^2$, we consider the "power law" function $f : \mathbb{R}^2_{++} \to \mathbb{R}$, with values $f(x) = x_1^{\alpha_1} x_2^{\alpha_2}$. Justify the statement: "the coefficients α_i provide the ratio between the relative error in f to a relative error in x_i".

Exercise 2.9 (Bound on a polynomial's derivative) In this exercise, you derive a bound on the largest absolute value of the derivative of a polynomial of a given order, in terms of the size of the coefficients.[1] For $w \in \mathbb{R}^{k+1}$, we define the polynomial p_w, with values

$$p_w(x) \doteq w_1 + w_2 x + \cdots + w_{k+1} x^k.$$

[1] See the discussion on regularization in Section 13.2.3 for an application of this result.

Show that, for any $p \geq 1$

$$\forall\, x \in [-1, 1] \;:\; \left|\frac{\mathrm{d}p_w(x)}{\mathrm{d}x}\right| \leq C(k, p)\|v\|_p,$$

where $v = (w_2, \ldots, w_{k+1}) \in \mathbb{R}^k$, and

$$C(k, p) = \begin{cases} k & p = 1, \\ k^{3/2} & p = 2, \\ \frac{k(k+1)}{2} & p = \infty. \end{cases}$$

Hint: you may use Hölder's inequality (2.4) or the results from Exercise 2.6.

3. Matrices

Exercise 3.1 (Derivatives of composite functions)

1. Let $f : \mathbb{R}^m \to \mathbb{R}^k$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ be two maps. Let $h : \mathbb{R}^n \to \mathbb{R}^k$ be the composite map $h = f \circ g$, with values $h(x) = f(g(x))$ for $x \in \mathbb{R}^n$. Show that the derivatives of h can be expressed via a matrix–matrix product, as $J_h(x) = J_f(g(x)) \cdot J_g(x)$, where $J_h(x)$ is the Jacobian matrix of h at x, i.e., the matrix whose (i, j) element is $\frac{\partial h_i(x)}{\partial x_j}$.

2. Let g be an affine map of the form $g(x) = Ax + b$, for $A \in \mathbb{R}^{m,n}$, $b \in \mathbb{R}^m$. Show that the Jacobian of $h(x) = f(g(x))$ is
$$J_h(x) = J_f(g(x)) \cdot A.$$

3. Let g be an affine map as in the previous point, let $f : \mathbb{R}^n \to \mathbb{R}$ (a scalar-valued function), and let $h(x) = f(g(x))$. Show that
$$\begin{aligned} \nabla_x h(x) &= A^\top \nabla_g f(g(x)), \\ \nabla_x^2 h(x) &= A^\top \nabla_g^2 f(g(x)) A. \end{aligned}$$

Exercise 3.2 (Permutation matrices) A matrix $P \in \mathbb{R}^{n,n}$ is a permutation matrix if its columns are a permutation of the columns of the $n \times n$ identity matrix.

1. For an $n \times n$ matrix A, we consider the products PA and AP. Describe in simple terms what these matrices look like with respect to the original matrix A.

2. Show that P is orthogonal.

Exercise 3.3 (Linear maps) Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a linear map. Show how to compute the (unique) matrix A such that $f(x) = Ax$ for every $x \in \mathbb{R}^n$, in terms of the values of f at appropriate vectors, which you will determine.

Exercise 3.4 (Linear dynamical systems) Linear dynamical systems are a common way to (approximately) model the behavior of physical phenomena, via recurrence equations of the form[2]

$$x(t+1) = Ax(t) + Bu(t), \quad y(t) = Cx(t), \quad t = 0, 1, 2, \ldots,$$

where t is the (discrete) time, $x(t) \in \mathbb{R}^n$ describes the state of the system at time t, $u(t) \in \mathbb{R}^p$ is the input vector, and $y(t) \in \mathbb{R}^m$ is the output vector. Here, matrices A, B, C, are given.

[2] Such models are the focus of Chapter 15.

1. Assuming that the system has initial condition $x(0) = 0$, express the output vector at time T as a linear function of $u(0), \ldots, u(T-1)$; that is, determine a matrix H such that $y(T) = HU(T)$, where
$$U(T) \doteq \begin{bmatrix} u(0) \\ \vdots \\ u(T-1) \end{bmatrix}$$
contains all the inputs up to and including at time $T-1$.

2. What is the interpretation of the range of H?

Exercise 3.5 (Nullspace inclusions and range) Let $A, B \in \mathbb{R}^{m,n}$ be two matrices. Show that the fact that the nullspace of B is contained in that of A implies that the range of $B^\top$ contains that of $A^\top$.

Exercise 3.6 (Rank and nullspace) Consider the image in Figure 3.1, a gray-scale rendering of a painting by Mondrian (1872–1944). We build a 256×256 matrix A of pixels based on this image by ignoring grey zones, assigning $+1$ to horizontal or vertical black lines, $+2$ at the intersections, and zero elsewhere. The horizontal lines occur at row indices $100, 200$ and 230, and the vertical ones at columns indices $50, 230$.

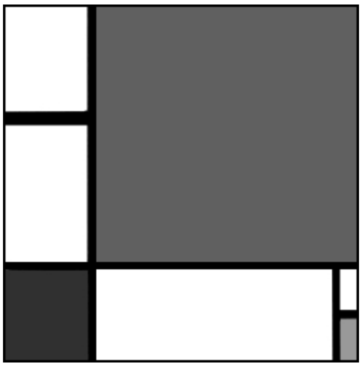

Figure 3.1: A gray-scale rendering of a painting by Mondrian.

1. What is nullspace of the matrix?

2. What is its rank?

Exercise 3.7 (Range and nullspace of $A^\top A$) Prove that, for any matrix $A \in \mathbb{R}^{m,n}$, it holds that
$$\begin{aligned} \mathcal{N}(A^\top A) &= \mathcal{N}(A), \\ \mathcal{R}(A^\top A) &= \mathcal{R}(A^\top). \end{aligned} \qquad (3.1)$$

Hint: use the fundamental theorem of linear algebra.

Exercise 3.8 (Cayley–Hamilton theorem) Let $A \in \mathbb{R}^{n,n}$ and let
$$p(\lambda) \doteq \det(\lambda I_n - A) = \lambda^n + c_{n-1}\lambda^{n-1} + \cdots + c_1\lambda + c_0$$
be the characteristic polynomial of A.

1. Assume A is diagonalizable. Prove that A annihilates its own characteristic polynomial, that is
$$p(A) = A^n + c_{n-1}A^{n-1} + \cdots + c_1 A + c_0 I_n = 0.$$
Hint: use Lemma 3.3.

2. Prove that $p(A) = 0$ holds in general, i.e., also for non-diagonalizable square matrices. *Hint:* use the facts that polynomials are continuous functions, and that diagonalizable matrices are dense in $\mathbb{R}^{n,n}$, i.e., for any $\epsilon > 0$ there exist $\Delta \in \mathbb{R}^{n,n}$ with $\|\Delta\|_F \leq \epsilon$ such that $A + \Delta$ is diagonalizable.

Exercise 3.9 (Frobenius norm and random inputs) Let $A \in \mathbb{R}^{m,n}$ be a matrix. Assume that $u \in \mathbb{R}^n$ is a vector-valued random variable, with zero mean and covariance matrix I_n. That is, $\mathbb{E}\{u\} = 0$, and $\mathbb{E}\{uu^\top\} = I_n$.

1. What is the covariance matrix of the output, $y = Au$?

2. Define the total output variance as $\mathbb{E}\{\|y - \hat{y}\|_2^2\}$, where $\hat{y} = \mathbb{E}\{y\}$ is the output's expected value. Compute the total output variance and comment.

Exercise 3.10 (Adjacency matrices and graphs) For a given undirected graph G with no self-loops and at most one edge between any pair of nodes (i.e., a *simple* graph), as in Figure 3.2, we associate a $n \times n$ matrix A, such that

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between node } i \text{ and node } j, \\ 0 & \text{otherwise.} \end{cases}$$

This matrix is called the *adjacency* matrix of the graph.[3]

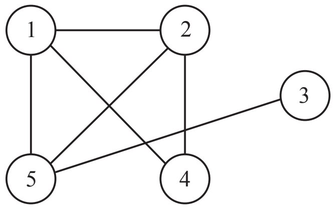

Figure 3.2: An undirected graph with $n = 5$ vertices.

1. Prove the following result: for positive integer k, the matrix A^k has an interesting interpretation: the entry in row i and column j gives the number of *walks* of length k (i.e., a collection of k edges) leading from vertex i to vertex j. *Hint:* prove this by induction on k, and look at the matrix–matrix product $A^{k-1}A$.

2. A *triangle* in a graph is defined as a subgraph composed of three vertices, where each vertex is reachable from each other vertex (i.e., a triangle forms a complete subgraph of order 3). In the graph of Figure 3.2, for example, nodes $\{1, 2, 4\}$ form a triangle. Show that the number of triangles in G is equal to the trace of A^3 divided by 6. *Hint:* For each node in a triangle in an undirected graph, there are two walks of length 3 leading from the node to itself, one corresponding to a clockwise walk, and the other to a counter-clockwise walk.

Exercise 3.11 (Nonnegative and positive matrices) A matrix $A \in \mathbb{R}^{n,n}$ is said to be *non-negative* (resp. *positive*) if $a_{ij} \geq 0$ (resp. $a_{ij} > 0$) for

[3] The graph in Figure 3.2 has adjacency matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

all $i, j = 1, \dots, n$. The notation $A \geq 0$ (resp. $A > 0$) is used to denote non-negative (resp. positive) matrices.

A non-negative matrix is said to be column (resp. row) *stochastic*, if the sum of the elements along each column (resp. row) is equal to one, that is if $\mathbf{1}^\top A = \mathbf{1}^\top$ (resp. $A\mathbf{1} = \mathbf{1}$). Similarly, a vector $x \in \mathbb{R}^n$ is said to be non-negative if $x \geq 0$ (element-wise), and it is said to be a *probability vector*, if it is non-negative and $\mathbf{1}^\top x = 1$. The set of probability vectors in $\mathbb{R}^n$ is thus the set $S = \{x \in \mathbb{R}^n : x \geq 0, \mathbf{1}^\top x = 1\}$, which is called the *probability simplex*. The following points you are requested to prove are part of a body of results known as the Perron–Frobenius theory of non-negative matrices.

1. Prove that a non-negative matrix A maps non-negative vectors into non-negative vectors (i.e., that $Ax \geq 0$ whenever $x \geq 0$), and that a column stochastic matrix $A \geq 0$ maps probability vectors into probability vectors.

2. Prove that if $A > 0$, then its spectral radius $\rho(A)$ is positive. *Hint:* use the Cayley–Hamilton theorem.

3. Show that it holds for any matrix A and vector x that
$$|Ax| \leq |A||x|,$$
where $|A|$ (resp. $|x|$) denotes the matrix (resp. vector) of moduli of the entries of A (resp. x). Then, show that if $A > 0$ and λ_i, v_i is an eigenvalue/eigenvector pair for A, then
$$|\lambda_i||v_i| \leq A|v_i|.$$

4. Prove that if $A > 0$ then $\rho(A)$ is actually an eigenvalue of A (i.e., A has a positive real eigenvalue $\lambda = \rho(A)$, and all other eigenvalues of A have modulus no larger than this “dominant” eigenvalue), and that there exist a corresponding eigenvector $v > 0$. Further, the dominant eigenvalue is simple (i.e., it has unit algebraic multiplicity), but you are not requested to prove this latter fact.

   Hint: For proving this claim you may use the following fixed-point theorem due to Brouwer: *if S is a compact and convex set*[4] *in $\mathbb{R}^n$, and $f : S \to S$ is a continuous map, then there exist an $x \in S$ such that $f(x) = x$.* Apply this result to the continuous map $f(x) \doteq \frac{Ax}{\mathbf{1}^\top Ax}$, with S being the probability simplex (which is indeed convex and compact).

5. Prove that if $A > 0$ and it is column or row stochastic, then its dominant eigenvalue is $\lambda = 1$.

[4] See Section 8.1 for definitions of compact and convex sets.

4. Symmetric matrices

Exercise 4.1 (Eigenvectors of a symmetric 2×2 matrix) Let $p, q \in \mathbb{R}^n$ be two linearly independent vectors, with unit norm ($\|p\|_2 = \|q\|_2 = 1$). Define the symmetric matrix $A \doteq pq^\top + qp^\top$. In your derivations, it may be useful to use the notation $c \doteq p^\top q$.

1. Show that $p + q$ and $p - q$ are eigenvectors of A, and determine the corresponding eigenvalues.
2. Determine the nullspace and rank of A.
3. Find an eigenvalue decomposition of A, in terms of p, q. *Hint:* use the previous two parts.
4. What is the answer to the previous part if p, q are not normalized?

Exercise 4.2 (Quadratic constraints) For each of the following cases, determine the shape of the region generated by the quadratic constraint $x^\top A x \leq 1$.

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

Hint: use the eigenvalue decomposition of A, and discuss depending on the sign of the eigenvalues.

Exercise 4.3 (Drawing an ellipsoid)

1. How would you efficiently draw an ellipsoid in $\mathbb{R}^2$, if the ellipsoid is described by a quadratic inequality of the form

$$\mathcal{E} = \left\{ x^\top A x + 2b^\top x + c \leq 0 \right\},$$

where A is 2×2 and symmetric, positive definite, $b \in \mathbb{R}^2$, and $c \in \mathbb{R}$? Describe your method as precisely as possible.

2. Draw the ellipsoid

$$\mathcal{E} = \left\{ 4x_1^2 + 2x_2^2 + 3x_1 x_2 + 4x_1 + 5x_2 + 3 \leq 1 \right\}.$$

Exercise 4.4 (Minimizing a quadratic function) Consider the *unconstrained* optimization problem

$$p^* = \min_x \frac{1}{2}x^\top Qx - c^\top x$$

where $Q = Q^\top \in \mathbb{R}^{n,n}$, $Q \succeq 0$, and $c \in \mathbb{R}^n$ are given. The goal of this exercise is to determine the optimal value p^* and the set of optimal solutions, $\mathcal{X}^{\text{opt}}$, in terms of c and the eigenvalues and eigenvectors of the (symmetric) matrix Q.

1. Assume that $Q \succ 0$. Show that the optimal set is a singleton, and that p^* is finite. Determine both in terms of Q, c.

2. Assume from now on that Q is not invertible. Assume further that Q is diagonal: $Q = \text{diag}\,(\lambda_1, \dots, \lambda_n)$, with $\lambda_1 \geq \dots \geq \lambda_r > \lambda_{r+1} = \dots = \lambda_n = 0$, where r is the rank of Q ($1 \leq r < n$). Solve the problem in that case (you will have to distinguish between two cases).

3. Now we do not assume that Q is diagonal anymore. Under what conditions (on Q, c) is the optimal value finite? Make sure to express your result in terms of Q and c, as explicitly as possible.

4. Assuming that the optimal value is finite, determine the optimal value and optimal set. Be as specific as you can, and express your results in terms of the pseudo-inverse[5] of Q.

Exercise 4.5 (Interpretation of covariance matrix) As in Example 4.2, we are given m points $x^{(1)}, \dots, x^{(m)}$ in $\mathbb{R}^n$, and denote by Σ the sample covariance matrix:

$$\Sigma \doteq \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \hat{x})(x^{(i)} - \hat{x})^\top,$$

where $\hat{x} \in \mathbb{R}^n$ is the sample average of the points:

$$\hat{x} \doteq \frac{1}{m}\sum_{i=1}^{m} x^{(i)}.$$

We assume that the average and variance of the data projected along a given direction does not change with the direction. In this exercise we will show that the sample covariance matrix is then proportional to the identity.

We formalize this as follows. To a given normalized direction $w \in \mathbb{R}^n$, $\|w\|_2 = 1$, we associate the line with direction w passing through the origin, $\mathcal{L}(w) = \{tw \,:\, t \in \mathbb{R}\}$. We then consider the projection of the points $x^{(i)}$, $i = 1, \dots, m$, on the line $\mathcal{L}(w)$, and look

[5] See Section 5.2.3.

at the associated coordinates of the points on the line. These *projected values* are given by

$$t_i(w) \doteq \arg\min_t \|tw - x^{(i)}\|_2, \quad i = 1, \ldots, m.$$

We assume that for any w, the sample average $\hat{t}(w)$ of the projected values $t_i(w)$, $i = 1, \ldots, m$, and their sample variance $\sigma^2(w)$, are both constant, independent of the direction w. Denote by $\hat{t}$ and σ^2 the (constant) sample average and variance. Justify your answer to the following questions as carefully as you can.

1. Show that $t_i(w) = w^\top x^{(i)}$, $i = 1, \ldots, m$.

2. Show that the sample average $\hat{x}$ of the data points is zero.

3. Show that the sample covariance matrix Σ of the data points is of the form $\sigma^2 I_n$. *Hint:* the largest eigenvalue $\lambda_{\max}$ of the matrix Σ can be written as: $\lambda_{\max} = \max_w \{w^\top \Sigma w \ : \ w^\top w = 1\}$, and a similar expression holds for the smallest eigenvalue.

Exercise 4.6 (Connected graphs and the Laplacian) We are given a graph as a set of vertices in $V = \{1, \ldots, n\}$, with an edge joining any pair of vertices in a set $E \subseteq V \times V$. We assume that the graph is undirected (without arrows), meaning that $(i, j) \in E$ implies $(j, i) \in E$. As in Section 4.1, we define the Laplacian matrix by

$$L_{ij} = \begin{cases} -1 & \text{if } (i,j) \in E, \\ d(i) & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

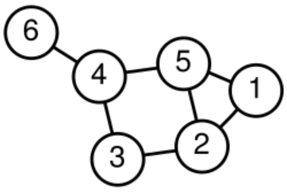

Figure 4.3: Example of an undirected graph.

Here, $d(i)$ is the number of edges adjacent to vertex i. For example, $d(4) = 3$ and $d(6) = 1$ for the graph in Figure 4.3.

1. Form the Laplacian for the graph shown in Figure 4.3.

2. Turning to a generic graph, show that the Laplacian L is symmetric.

3. Show that L is positive-semidefinite, proving the following identity, valid for any $u \in \mathbb{R}^n$:

$$u^\top L u = q(u) \doteq \frac{1}{2} \sum_{(i,j) \in E} (u_i - u_j)^2.$$

   Hint: find the values $q(k)$, $q(e_k \pm e_l)$, for two unit vectors e_k, e_l such that $(k, l) \in E$.

4. Show that 0 is always an eigenvalue of L, and exhibit an eigenvector. *Hint:* consider a matrix square-root[6] of L.

[6] See Section 4.4.4.

5. The graph is said to be connected if there is a path joining any pair of vertices. Show that if the graph is connected, then the zero eigenvalue is simple, that is, the dimension of the nullspace of L is 1. *Hint:* prove that if $u^\top Lu = 0$, then $u_i = u_j$ for every pair $(i,j) \in E$.

Exercise 4.7 (Component-wise product and PSD matrices) Let $A, B \in \mathbb{S}^n$ be two symmetric matrices. Define the component-wise product of A, B, by a matrix $C \in \mathbb{S}^n$ with elements $C_{ij} = A_{ij}B_{ij}$, $1 \le i, j \le n$. Show that C is positive semidefinite, provided both A, B are. *Hint:* prove the result when A is rank-one, and extend to the general case via the eigenvalue decomposition of A.

Exercise 4.8 (A bound on the eigenvalues of a product) Let $A, B \in \mathbb{S}^n$ be such that $A \succ 0$, $B \succ 0$.

1. Show that all eigenvalues of BA are real and positive (despite the fact that BA is not symmetric, in general).

2. Let $A \succ 0$, and let $B^{-1} \doteq \operatorname{diag}\left(\|a_1^\top\|_1, \ldots, \|a_n^\top\|_1\right)$, where $a_i^\top$, $i = 1, \ldots, n$, are the rows of A. Prove that

$$0 < \lambda_i(BA) \le 1, \quad \forall\, i = 1, \ldots, n.$$

3. With all terms defined as in the previous point, prove that

$$\rho(I - \alpha BA) < 1, \quad \forall \alpha \in (0, 2).$$

Exercise 4.9 (Hadamard's inequality) Let $A \in \mathbb{S}^n$ be positive semidefinite. Prove that

$$\det A \le \prod_{i=1}^{n} a_{ii}.$$

Hint: Distinguish the cases $\det A = 0$ and $\det A \neq 0$. In the latter case, consider the normalized matrix $\tilde{A} \doteq DAD$, where $D = \operatorname{diag}\left(a_{11}^{-1/2}, \ldots, a_{nn}^{-1/2}\right)$, and use the geometric–arithmetic mean inequality (see Example 8.9).

Exercise 4.10 (A lower bound on the rank) Let $A \in \mathbb{S}_+^n$ be a symmetric, positive semidefinite matrix.

1. Show that the trace, $\operatorname{trace} A$, and the Frobenius norm, $\|A\|_F$, depend only on its eigenvalues, and express both in terms of the vector of eigenvalues.

2. Show that

$$(\operatorname{trace} A)^2 \le \operatorname{rank}(A)\|A\|_F^2.$$

3. Identify classes of matrices for which the corresponding lower bound on the rank is attained.

Exercise 4.11 (A result related to Gaussian distributions) Let $\Sigma \in \mathbb{S}^n_{++}$ be a symmetric, positive definite matrix. Show that

$$\int_{\mathbb{R}^n} \mathrm{e}^{-\frac{1}{2}x^\top \Sigma^{-1} x} \mathrm{d}x = (2\pi)^{n/2}\sqrt{\det \Sigma}.$$

You may assume known that the result holds true when $n = 1$. The above shows that the function $p : \mathbb{R}^n \to \mathbb{R}$ with (non-negative) values

$$p(x) = \frac{1}{(2\pi)^{n/2} \cdot \sqrt{\det \Sigma}} \mathrm{e}^{-\frac{1}{2}x^\top \Sigma^{-1} x}$$

integrates to one over the whole space. In fact, it is the density function of a probability distribution called the multivariate Gaussian (or normal) distribution, with zero mean and covariance matrix Σ. *Hint:* you may use the fact that for any integrable function f, and invertible $n \times n$ matrix P, we have

$$\int_{x \in \mathbb{R}^n} f(x) \mathrm{d}x = |\det P| \cdot \int_{z \in \mathbb{R}^n} f(Pz) \mathrm{d}z.$$

5. Singular Value Decomposition

Exercise 5.1 (SVD of an orthogonal matrix) Consider the matrix

$$A = \frac{1}{3}\left[\begin{array}{ccc} -1 & 2 & 2 \\ 2 & -1 & 2 \\ 2 & 2 & -1 \end{array}\right].$$

1. Show that A is orthogonal.

2. Find a singular value decomposition of A.

Exercise 5.2 (SVD of a matrix with orthogonal columns) Assume a matrix $A = [a_1, \ldots, a_m]$ has columns $a_i \in \mathbb{R}^n$, $i = 1, \ldots, m$ that are orthogonal to each other: $a_i^\top a_j = 0$ for $1 \leq i \neq j \leq n$. Find an SVD for A, in terms of the a_is. Be as explicit as you can.

Exercise 5.3 (Singular values of augmented matrix) Let $A \in \mathbb{R}^{n,m}$, with $n \geq m$, have singular values $\sigma_1, \ldots, \sigma_m$.

1. Show that the singular values of the $(n+m) \times m$ matrix

   $$\tilde{A} \doteq \left[\begin{array}{c} A \\ I_m \end{array}\right]$$

   are $\tilde{\sigma}_i = \sqrt{1 + \sigma_i^2}$, $i = 1, \ldots, m$.

2. Find an SVD of the matrix $\tilde{A}$.

Exercise 5.4 (SVD of score matrix) An exam with m questions is given to n students. The instructor collects all the grades in a $n \times m$ matrix G, with G_{ij} the grade obtained by student i on question j. We would like to assign a difficulty score to each question, based on the available data.

1. Assume that the grade matrix G is well approximated by a rank-one matrix $sq^\top$, with $s \in \mathbb{R}^n$ and $q \in \mathbb{R}^m$ (you may assume that both s, q have non-negative components). Explain how to use the approximation to assign a difficulty level to each question. What is the interpretation of vector s?

2. How would you compute a rank-one approximation to G? State precisely your answer in terms of the SVD of G.

Exercise 5.5 (Latent semantic indexing) Latent semantic indexing is an SVD-based technique that can be used to discover text documents similar to each other. Assume that we are given a set of m documents $D_1, \ldots, D_m$. Using a "bag-of-words" technique described in

Example 2.1, we can represent each document D_j by an n-vector d_j, where n is the total number of distinct words appearing in the whole set of documents. In this exercise, we assume that the vectors d_j are constructed as follows: $d_j(i) = 1$ if word i appears in document D_j, and 0 otherwise. We refer to the $n \times m$ matrix $M = [d_1, \ldots, d_m]$ as the "raw" term-by-document matrix. We will also use a normalized[7] version of that matrix: $\tilde{M} = [\tilde{d}_1, \ldots, \tilde{d}_m]$, where $\tilde{d}_j = d_j / \|d_j\|_2$, $j = 1, \ldots, m$.

Assume we are given another document, referred to as the "query document," which is not part of the collection. We describe that query document as an n-dimensional vector q, with zeros everywhere, except a 1 at indices corresponding to the terms that appear in the query. We seek to retrieve documents that are "most similar" to the query, in some sense. We denote by $\tilde{q}$ the normalized vector $\tilde{q} = q / \|q\|_2$.

1. A first approach is to select the documents that contain the largest number of terms in common with the query document. Explain how to implement this approach, based on a certain matrix–vector product, which you will determine.

2. Another approach is to find the closest document by selecting the index j such that $\|q - d_j\|_2$ is the smallest. This approach can introduce some biases, if for example the query document is much shorter than the other documents. Hence a measure of similarity based on the normalized vectors, $\|\tilde{q} - \tilde{d}_j\|_2$, has been proposed, under the name of "cosine similarity". Justify the use of this name for that method, and provide a formulation based on a certain matrix–vector product, which you will determine.

3. Assume that the normalized matrix $\tilde{M}$ has an SVD $\tilde{M} = U\Sigma V^\top$, with Σ an $n \times m$ matrix containing the singular values, and the unitary matrices $U = [u_1, \ldots, u_n]$, $V = [v_1, \ldots, v_m]$ of size $n \times n$, $m \times m$ respectively. What could be an interpretation of the vectors u_l, v_l, $l = 1, \ldots, r$? *Hint:* discuss the case when r is very small, and the vectors u_l, v_l, $l = 1, \ldots, r$, are sparse.

4. With real-life text collections, it is often observed that M is effectively close to a low-rank matrix. Assume that a optimal rank-k approximation ($k \ll \min(n, m)$) of $\tilde{M}$, $\tilde{M}_k$, is known. In the latent semantic indexing approach[8] to document similarity, the idea is to first project the documents and the query onto the subspace generated by the singular vectors $u_1, \ldots, u_k$, and then apply the cosine similarity approach to the projected vectors. Find an expression for the measure of similarity.

[7] In practice, other numerical representation of text documents can be used. For example we may use the relative frequencies of words in each document, instead of the ℓ_2-norm normalization employed here.

[8] In practice, it is often observed that this method produces better results than cosine similarity in the original space, as in part 2.

Exercise 5.6 (Fitting a hyperplane to data) We are given m data points $d_1, \ldots, d_m \in \mathbb{R}^n$, and we seek a hyperplane

$$\mathcal{H}(c,b) \doteq \{x \in \mathbb{R}^n : c^\top x = b\},$$

where $c \in \mathbb{R}^n$, $c \neq 0$, and $b \in \mathbb{R}$, that best "fits" the given points, in the sense of a minimum sum of squared distances criterion, see Figure 5.4.

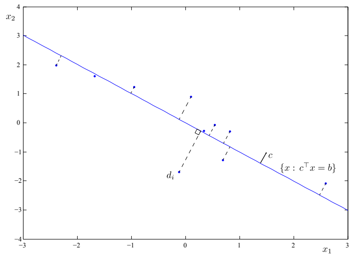

Figure 5.4: Fitting a hyperplane to data.

Formally, we need to solve the optimization problem

$$\min_{c,b} \sum_{i=1}^{m} \mathrm{dist}^2(d_i, \mathcal{H}(c,b)) \;:\; \|c\|_2 = 1,$$

where $\mathrm{dist}(d, \mathcal{H})$ is the Euclidean distance from a point d to $\mathcal{H}$. Here the constraint on c is imposed without loss of generality, in a way that does not favor a particular direction in space.

1. Show that the distance from a given point $d \in \mathbb{R}^n$ to $\mathcal{H}$ is given by

$$\mathrm{dist}(d, \mathcal{H}(c,b)) = |c^\top d - b|.$$

2. Show that the problem can be expressed as

$$\min_{b,c \,:\, \|c\|_2=1} f_0(b,c),$$

where f_0 is a certain quadratic function, which you will determine.

3. Show that the problem can be reduced to

$$\begin{array}{ll} \displaystyle\min_{c} & c^\top(\tilde{D}\tilde{D}^\top)c \\ \text{s.t.:} & \|c\|_2 = 1, \end{array}$$

where $\tilde{D}$ is the matrix of centered data points: the i-th column of $\tilde{D}$ is $d_i - \bar{d}$, where $\bar{d} \doteq (1/m)\sum_{i=1}^{m} d_i$ is the average of the data points. *Hint:* you can exploit the fact that at optimum, the partial derivative of the objective function with respect to b must be zero, a fact justified in Section 8.4.1.

4. Explain how to find the hyperplane via SVD.

Exercise 5.7 (Image deformation) A rigid transformation is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$ that is the composition of a translation and a rotation. Mathematically, we can express a rigid transformation ϕ as $\phi(x) = Rx + r$, where R is an $n \times n$ orthogonal transformation and $r \in \mathbb{R}^n$ a vector.

We are given a set of pairs of points (x_i, y_i) in $\mathbb{R}^n$, $i = 1, \ldots, m$, and wish to find a rigid transformation that best matches them. We can write the problem as

$$\min_{R \in \mathbb{R}^{n,n},\, r \in \mathbb{R}^n} \sum_{i=1}^{m} \|Rx_i + r - y_i\|_2^2 \ : \ R^\top R = I_n, \tag{5.2}$$

where I_n is the $n \times n$ identity matrix.

The problem arises in image processing, to provide ways to deform an image (represented as a set of two-dimensional points) based on the manual selection of a few points and their transformed counterparts.

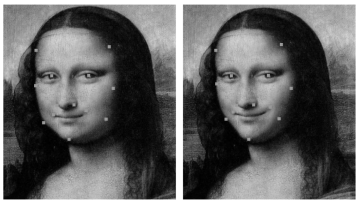

Figure 5.5: Image deformation via rigid transformation. The image on the left is the original image, and that on the right is the deformed image. Dots indicate points for which the deformation is chosen by the user.

1. Assume that R is fixed in problem (5.2). Express an optimal r as a function of R.

2. Show that the corresponding optimal value (now a function of R only) can be written as the original objective function, with $r = 0$ and x_i, y_i replaced with their centered counterparts,

$$\bar{x}_i = x_i - \hat{x}, \ \ \hat{x} = \frac{1}{m}\sum_{j=1}^{m} x_j, \ \ \bar{y}_i = y_i - \hat{y}, \ \ \hat{y} = \frac{1}{m}\sum_{j=1}^{m} y_j.$$

3. Show that the problem can be written as

$$\min_R \|RX - Y\|_F \ : \ R^\top R = I_n,$$

for appropriate matrices X, Y, which you will determine. *Hint:* explain why you can square the objective; then expand.

4. Show that the problem can be further written as

$$\max_R \operatorname{trace} RZ \ : \ R^\top R = I_n,$$

for an appropriate $n \times n$ matrix Z, which you will determine.

5. Show that $R = VU^\top$ is optimal, where $Z = USV^\top$ is the SVD of Z. *Hint:* reduce the problem to the case when Z is diagonal, and use without proof the fact that when Z is diagonal, I_n is optimal for the problem.

6. Show the result you used in the previous question: assume Z is diagonal, and show that $R = I_n$ is optimal for the problem above. *Hint:* show that $R^\top R = I_n$ implies $|R_{ii}| \leq 1$, $i = 1, \ldots, n$, and using that fact, prove that the optimal value is less than or equal to trace Z.

7. How woud you apply this technique to make Mona Lisa smile more? *Hint:* in Figure 5.5, the two-dimensional points x_i are given (as dots) on the left panel, while the corresponding points y_i are shown on the left panel. These points are manually selected. The problem is to find how to transform all the other points in the original image.

6. Linear Equations

Exercise 6.1 (Least squares and total least squares) Find the least-squares line and the total least-squares[9] line for the data points (x_i, y_i), $i = 1, \ldots, 4$, with $x = (-1, 0, 1, 2)$, $y = (0, 0, 1, 1)$. Plot both lines on the same set of axes.

Exercise 6.2 (Geometry of least squares) Consider a least-squares problem

$$p^* = \min_x \|Ax - y\|_2,$$

where $A \in \mathbb{R}^{m,n}$, $y \in \mathbb{R}^m$. We assume that $y \notin \mathcal{R}(A)$, so that $p^* > 0$. Show that, at optimum, the residual vector $r = y - Ax$ is such that $r^\top y > 0$, $A^\top r = 0$. Interpret the result geometrically. *Hint:* use the SVD of A. You can assume that $m \geq n$, and that A is full column rank.

Exercise 6.3 (Lotka's law and least squares) Lotka's law describes the frequency of publication by authors in a given field. It states that $X^a Y = b$, where X is the number of publications, Y the relative frequency of authors with X publications, and a and b are constants (with $b > 0$) that depend on the specific field. Assume that we have data points (X_i, Y_i), $i = 1, \ldots, m$, and seek to estimate the constants a and b.

1. Show how to find the values of a, b according to a linear least-squares criterion. Make sure to define the least-squares problem involved precisely.

2. Is the solution always unique? Formulate a condition on the data points that guarantees unicity.

Exercise 6.4 (Regularization for noisy data) Consider a least-squares problem

$$\min_x \|Ax - y\|_2^2,$$

in which the data matrix $A \in \mathbb{R}^{m,n}$ is noisy. Our specific noise model assumes that each row $a_i^\top \in \mathbb{R}^n$ has the form $a_i = \hat{a}_i + u_i$, where the noise vector $u_i \in \mathbb{R}^n$ has zero mean and covariance matrix $\sigma^2 I_n$, with σ a measure of the size of the noise. Therefore, now the matrix A is a function of the uncertain vector $u = (u_1, \ldots, u_n)$, which we denote by $A(u)$. We will write $\hat{A}$ to denote the matrix with rows $\hat{a}_i^\top$, $i = 1, \ldots, m$. We replace the original problem with

$$\min_x \mathbb{E}_u\{\|A(u)x - y\|_2^2\},$$

[9] See Section 6.7.5.

where $\mathbb{E}_u$ denotes the expected value with respect to the random variable u. Show that this problem can be written as

$$\min_x \|\hat{A}x - y\|_2^2 + \lambda\|x\|_2^2,$$

where $\lambda \geq 0$ is some regularization parameter, which you will determine. That is, regularized least squares can be interpreted as a way to take into account uncertainties in the matrix A, in the expected value sense. *Hint:* compute the expected value of $((\hat{a}_i + u_i)^\top x - y_i)^2$, for a specific row index i.

Exercise 6.5 (Deleting a measurement in least squares) In this exercise, we revisit Section 6.3.5, and assume now that we would like to *delete* a measurement, and update the least-squares solution accordingly.[10]

We are given a full column rank matrix $A \in \mathbb{R}^{m,n}$, with rows $a_i^\top$, $i = 1, \ldots, m$, a vector $y \in \mathbb{R}^m$, and a solution to the least-squares problem

$$x^* = \arg\min_x \sum_{i=1}^{m} (a_i^\top x - y_i)^2 = \arg\min_x \|Ax - y\|_2.$$

Assume now we delete the last measurement, that is, replace (a_m, y_m) by $(0, 0)$. We assume that the matrix obtained after deleting any one of the measurements is still full column rank.

1. Express the solution to the problem after deletion, in terms of the original solution, similar to the formula (6.15). Make sure to explain why any quantities you invert are positive.
2. In the so-called leave-one-out analysis, we would like to efficiently compute all the m solutions corresponding to deleting one of the m measurements. Explain how you would compute those solutions computationally efficiently. Detail the number of operations (flops) needed. You may use the fact that to invert a $n \times n$ matrix costs $O(n^3)$.

Exercise 6.6 The Michaelis–Menten model for enzyme kinetics relates the rate y of an enzymatic reaction to the concentration x of a substrate, as follows:

$$y = \frac{\beta_1 x}{\beta_2 + x},$$

where β_i, $i = 1, 2$, are positive parameters.

1. Show that the model can be expressed as a linear relation between the values $1/y$ and $1/x$.

[10] This is useful in the context of cross-validation methods, as evoked in Section 13.2.2.

2. Use this expression to find an estimate $\hat{\beta}$ of the parameter vector β using linear least squares, based on m measurements (x_i, y_i), $i = 1, \ldots, m$.

3. The above approach has been found to be quite sensitive to errors in input data. Can you experimentally confirm this opinion?

Exercise 6.7 (Least norm estimation on traffic flow networks) You want to estimate the traffic (in San Francisco for example, but we'll start with a smaller example). You know the road network as well as the historical average of flows on each road segment.

1. We call q_i the flow of vehicles on each road segment $i \in I$. Write down the linear equation that corresponds to the conservation of vehicles at each intersection $j \in J$. *Hint:* think about how you might represent the road network in terms of matrices, vectors, etc.

2. The goal of the estimation is to estimate the traffic flow on each of the road segments. The flow estimates should satisfy the conservation of vehicles exactly at each intersection. Among the solutions that satisfy this constraint, we are searching for the estimate that is the closest to the historical average, $\overline{q}$, in the ℓ_2-norm sense. The vector $\overline{q}$ has size I and the i-th element represent the average for the road segment i. Pose the optimization problem.

3. Explain how to solve this problem mathematically. Detail your answer (do not only give a formula but explain where it comes from).

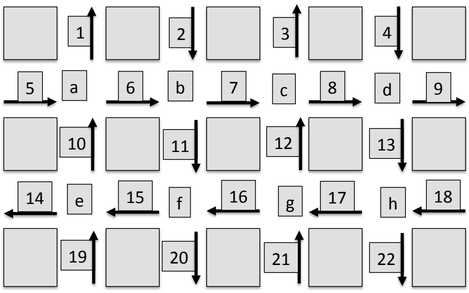

Figure 6.6: Example of the traffic estimation problem. The intersections are labeled a to h. The road segments are labeled 1 to 22. The arrows indicate the direction of traffic.

4. Formulate the problem for the small example of Figure 6.6 and solve it using the historical average given in Table 6.1. What is the flow that you estimate on road segments 1, 3, 6, 15 and 22?

5. Now, assume that besides the historical averages, you are also given some flow measurements on some of the road segments of

the network. You assume that these flow measurements are correct and want your estimate of the flow to match these measurements perfectly (besides matching the conservation of vehicles of course). The right column of Table 6.1 lists the road segments for which we have such flow measurements. Do you estimate a different flow on some of the links? Give the difference in flow you estimate for road segments 1,3, 6, 15 and 22. Also check that your estimate gives you the measured flow on the road segments for which you have measured the flow.

segment	average	measured
1	2047.6	2028
2	2046.0	2008
3	2002.6	2035
4	2036.9	
5	2013.5	2019
6	2021.1	
7	2027.4	
8	2047.1	
9	2020.9	2044
10	2049.2	
11	2015.1	
12	2035.1	
13	2033.3	
14	2027.0	2043
15	2034.9	
16	2033.3	
17	2008.9	
18	2006.4	
19	2050.0	2030
20	2008.6	2025
21	2001.6	
22	2028.1	2045

Table 6.1: Table of flows: historical averages $\bar{q}$ (center column), and some measured flows (right column).

Exercise 6.8 (A matrix least-squares problem) We are given a set of points $p_1, \ldots, p_m \in \mathbb{R}^n$, which are collected in the $n \times m$ matrix $P = [p_1, \ldots, p_m]$. We consider the problem

$$\min_X F(X) \doteq \sum_{i=1}^m \|x_i - p_i\|_2^2 + \frac{\lambda}{2} \sum_{1 \le i,j \le m} \|x_i - x_j\|_2^2,$$

where $\lambda \ge 0$ is a parameter. In the above, the variable is an $n \times m$ matrix $X = [x_1, \ldots, x_m]$, with $x_i \in \mathbb{R}^n$ the i-th column of X, $i = 1, \ldots, m$. The above problem is an attempt at clustering the points p_i; the first term encourages the cluster center x_i to be close to the corresponding point p_i, while the second term encourages the x_is to be close to each other, with a higher grouping effect as λ increases.

1. Show that the problem belongs to the family of ordinary least-squares problems. You do not need to be explicit about the form of the problem.

2. Show that
$$\frac{1}{2} \sum_{1 \le i,j \le m} \|x_i - x_j\|_2^2 = \operatorname{trace} XHX^\top,$$
where $H = mI_m - \mathbf{1}\mathbf{1}^\top$ is an $m \times m$ matrix, with I_m the $m \times m$ identity matrix, and $\mathbf{1}$ the vector of ones in $\mathbb{R}^m$.

3. Show that H is positive semidefinite.

4. Show that the gradient of the function F at a matrix X is the $n \times m$ matrix given by
$$\nabla F(X) = 2(X - P + \lambda XH).$$
Hint: for the second term, find the first-order expansion of the function $\Delta \to \operatorname{trace}((X + \Delta)H(X + \Delta)^\top)$, where $\Delta \in \mathbb{R}^{n,m}$.

5. As mentioned in Remark 6.1, optimality conditions for a least-squares problem are obtained by setting the gradient of the objective to zero. Using the formula (3.10), show that optimal points

are of the form

$$x_i = \frac{1}{m\lambda + 1} p_i + \frac{m\lambda}{m\lambda + 1} \hat{p}, \quad i = 1, \ldots, m,$$

where $\hat{p} = (1/m)(p_1 + \ldots + p_m)$ is the center of the given points.

6. Interpret your results. Do you believe the model considered here is a good one to cluster points?

7. Matrix Algorithms

Exercise 7.1 (Sparse matrix–vector product) Recall from Section 3.4.2 that a matrix is said to be sparse if most of its entries are zero. More formally, assume a $m \times n$ matrix A has sparsity coefficient $\gamma(A) \ll 1$, where $\gamma(A) \doteq d(A)/s(A)$, $d(A)$ is the number of nonzero elements in A, and $s(A)$ is the size of A (in this case, $s(A) = mn$).

1. Evaluate the number of operations (multiplications and additions) that are required to form the matrix–vector product Ax, for any given vector $x \in \mathbb{R}^n$ and generic, non-sparse A. Show that this number is reduced by a factor $\gamma(A)$, if A is sparse.

2. Now assume that A is not sparse, but is a rank-one modification of a sparse matrix. That is, A is of the form $\tilde{A} + uv^\top$, where $\tilde{A} \in \mathbb{R}^{m,n}$ is sparse, and $u \in \mathbb{R}^m$, $v \in \mathbb{R}^m$ are given. Devise a method to compute the matrix–vector product Ax that exploits sparsity.

Exercise 7.2 (A random inner product approximation) Computing the standard inner product between two vectors $a, b \in \mathbb{R}^n$ requires n multiplications and additions. When the dimension n is huge (say, e.g., of the order of 10^{12}, or larger), even computing a simple inner product can be computationally prohibitive.

Let us define a random vector $r \in \mathbb{R}^n$ constructed as follows: choose uniformly at random an index $i \in \{1, \ldots, n\}$, and set $r_i = 1$, and $r_j = 0$ for $j \neq i$. Consider the two scalar random numbers $\tilde{a}, \tilde{b}$ that represent the "random projections" of the original vectors a, b along r:

$$\begin{aligned} \tilde{a} &\doteq r^\top a = a_i, \\ \tilde{b} &\doteq r^\top b = b_i. \end{aligned}$$

Prove that

$$n\mathbb{E}\{\tilde{a}\tilde{b}\} = a^\top b,$$

that is, $n\tilde{a}\tilde{b}$ is an unbiased estimator of the value of the inner product $a^\top b$. Observe that computing $n\tilde{a}\tilde{b}$ requires very little effort, since it is just equal to na_ib_i, where i is the randomly chosen index. Notice, however, that the variance of such an estimator can be large, as it is given by

$$\mathrm{var}\{n\tilde{a}\tilde{b}\} = n\sum_{k=1}^{n} a_i^2 b_i^2 - \left(a^\top b\right)^2$$

(prove also this latter formula). *Hint:* let e_i denote the i-th standard basis vector of $\mathbb{R}^n$; the random vector r has discrete probability distribution $\mathrm{Prob}\{r = e_i\} = 1/n$, $i = 1, \ldots, n$, hence $\mathbb{E}\{r\} = \frac{1}{n}\mathbf{1}$. Further, observe that the products $r_k r_j$ are equal to zero for $k \neq j$ and that the vector $r^2 \doteq [r_1^2, \ldots, r_n^2]^\top$ has the same distribution as r.

Generalizations of this idea to random projections onto k-dimensional subspaces are indeed applied for matrix-product approximation, SVD factorization and PCA on huge-scale problems. The key theoretical tool underlying these results is known as the Johnson–Lindenstrauss lemma.

Exercise 7.3 (Power iteration for SVD with centered, sparse data) In many applications such as principal component analysis (see Section 5.3.2), one needs to find the few largest singular values of a centered data matrix. Specifically, we are given a $n \times m$ matrix $X = [x_1, \ldots, x_m]$ of m data points in $\mathbb{R}^n$, $i = 1, \ldots, m$, and define the centered matrix $\tilde{X}$ to be

$$\tilde{X} = [\tilde{x}_1 \ \cdots \ \tilde{x}_m], \quad \tilde{x}_i \doteq x_i - \bar{x}, \ i = 1, \ldots, m,$$

with $\bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$ the average of the data points. In general, $\tilde{X}$ is dense, even if X itself is sparse. This means that each step of the power iteration method involves two matrix–vector products, with a dense matrix. Explain how to modify the power iteration method in order to exploit sparsity, and avoid dense matrix–vector multiplications.

Exercise 7.4 (Exploiting structure in linear equations) Consider the linear equation in $x \in \mathbb{R}^n$

$$Ax = y,$$

where $A \in \mathbb{R}^{m,n}$, $y \in \mathbb{R}^m$. Answer the following questions to the best of your knowledge.

1. The time required to solve the general system depends on the sizes m, n and the entries of A. Provide a rough estimate of that time as a function of m, n only. You may assume that m, n are of the same order.

2. Assume now that $A = D + uv^\top$, where D is diagonal, invertible, and $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$. How would you exploit this structure to solve the above linear system, and what is a rough estimate of the complexity of your algorithm?

3. What if A is upper-triangular?

Exercise 7.5 (Jacobi method for linear equation) Let $A = (a_{ij}) \in \mathbb{R}^{n,n}$, $b \in \mathbb{R}^n$, with $a_{ii} \neq 0$ for every $i = 1, \ldots, n$. The *Jacobi method* for solving the square linear system

$$Ax = b$$

consists of decomposing A as a sum: $A = D + R$, where $D = \operatorname{diag}(a_{11}, \ldots, a_{nn})$, and R contains the off-diagonal elements of A, and then applying the recursion

$$x^{(k+1)} = D^{-1}(b - Rx^{(k)}), \;\; k = 0, 1, 2, \ldots,$$

with initial point $\hat{x}(0) = D^{-1}b$.

The method is part of a class of methods known as *matrix splitting*, where A is decomposed as a sum of a "simple" invertible matrix and another matrix; the Jacobi method uses a particular splitting of A.

1. Find conditions on D, R that guarantee convergence from an arbitrary initial point. *Hint:* assume that $M \doteq -D^{-1}R$ is diagonalizable.

2. The matrix A is said to be strictly row diagonally dominant if

$$\forall\, i = 1, \ldots, n \; : \; |a_{ii}| > \sum_{j \neq i} |a_{ij}|.$$

Show that when A is strictly row diagonally dominant, the Jacobi method converges.

Exercise 7.6 (Convergence of linear iterations) Consider linear iterations of the form

$$x(k+1) = Fx(k) + c, \quad k = 0, 1, \ldots, \tag{7.3}$$

where $F \in \mathbb{R}^{n,n}$, $c \in \mathbb{R}^n$, and the iterations are initialized with $x(0) = x_0$. We assume that the iterations admit a stationary point, i.e., that there exists $\bar{x} \in \mathbb{R}^n$ such that

$$(I - F)\bar{x} = c. \tag{7.4}$$

In this exercise, we derive conditions under which $x(k)$ tends to a finite limit for $k \to \infty$. We shall use these results in Exercise 7.7, to set up a linear iterative algorithm for solving systems of linear equations.

1. Show that the following expressions hold for all $k = 0, 1, \ldots$:

$$\begin{aligned} x(k+1) - x(k) &= F^k (I - F)(\bar{x} - x_0), && (7.5) \\ x(k) - \bar{x} &= F^k (x_0 - \bar{x}). && (7.6) \end{aligned}$$

2. Prove that, for all x_0, $\lim_{k \to \infty} x(k)$ converges to a finite limit if and only if F^k is convergent (see Theorem 3.5). When $x(k)$ converges, its limit point $\bar{x}$ satisfies (7.4).

Exercise 7.7 (A linear iterative algorithm) In this exercise we introduce some "equivalent" formulations of a system of linear equations

$$Ax = b, \quad A \in \mathbb{R}^{m,n}, \tag{7.7}$$

and then study a linear recursive algorithm for solution of this system.

1. Consider the system of linear equations

$$Ax = AA^{\dagger}b, \tag{7.8}$$

where $A^{\dagger}$ is any pseudoinverse of A (that is, a matrix such that $AA^{\dagger}A = A$). Prove that (7.8) always admits a solution. Show that every solution of equations (7.7) is also a solution for (7.8). Conversely, prove that if $b \in \mathcal{R}(A)$, then every solution to (7.8) is also a solution for (7.7).

2. Let $R \in \mathbb{R}^{n,m}$ be any matrix such that $\mathcal{N}(RA) = \mathcal{N}(A)$. Prove that

$$A^{\dagger} \doteq (RA)^{\dagger}R$$

is indeed a pseudoinverse of A.

3. Consider the system of linear equations

$$RAx = Rb, \tag{7.9}$$

where $R \in \mathbb{R}^{n,m}$ is any matrix such that $\mathcal{N}(RA) = \mathcal{N}(A)$ and $Rb \in \mathcal{R}(RA)$. Prove that, under these hypotheses, the set of solutions of (7.9) coincides with the set of solutions of (7.8), for $A^{\dagger} = (RA)^{\dagger}R$.

4. Under the setup of the previous point, consider the following linear iterations: for $k = 0, 1, \ldots,$

$$x(k+1) = x(k) + \alpha R(b - Ax(k)), \tag{7.10}$$

where $\alpha \neq 0$ is a given scalar. Show that if $\lim_{k\to\infty} x(k) = \bar{x}$, then $\bar{x}$ is a solution for the system of linear equations (7.9). State appropriate conditions under which $x(k)$ is guaranteed to converge.

5. Suppose A is positive definite (i.e., $A \in \mathbb{S}^n$, $A \succ 0$). Discuss how to find a suitable scalar α and matrix $R \in \mathbb{R}^{n,n}$ satisfying the conditions of point 3, and such that the iterations (7.10) converge to a solution of (7.9). *Hint:* use Exercise 4.8.

6. Explain how to apply the recursive algorithm (7.10) for finding a solution to the linear system $\tilde{A}x = \tilde{b}$, where $\tilde{A} \in \mathbb{R}^{m,n}$ with $m \geq n$ and $\operatorname{rank} \tilde{A} = n$. *Hint:* apply the algorithm to the normal equations.

8. Convexity

Exercise 8.1 (Quadratic inequalities) Consider the set defined by the following inequalities:

$$(x_1 \geq x_2 - 1 \text{ and } x_2 \geq 0) \text{ or } (x_1 \leq x_2 - 1 \text{ and } x_2 \leq 0).$$

1. Draw the set. Is it convex?

2. Show that it can be described as a single quadratic inequality of the form $q(x) = x^\top Ax + 2b^\top x + c \leq 0$, for a matrix $A = A^\top \in \mathbb{R}^{2,2}$, $b \in \mathbb{R}^2$ and $c \in \mathbb{R}$ which you will determine.

3. What is the convex hull of this set?

Exercise 8.2 (Closed functions and sets) Show that the indicator function $I_\mathcal{X}$ of a convex set $\mathcal{X}$ is convex. Show that this function is closed whenever $\mathcal{X}$ is a closed set.

Exercise 8.3 (Convexity of functions)

1. For x, y both positive scalars, show that

   $$ye^{x/y} = \max_{\alpha > 0} \alpha(x+y) - y\alpha \cdot \ln \alpha.$$

   Use the above result to prove that the function f defined as

   $$f(x,y) = \begin{cases} ye^{x/y} & \text{if } x > 0,\ y > 0, \\ +\infty & \text{otherwise,} \end{cases}$$

   is convex.

2. Show that for $r \geq 1$, the function $f_r : \mathbb{R}^m_+ \to \mathbb{R}$, with values

   $$f_r(v) = \left(\sum_{j=1}^m v_j^{1/r} \right)^r$$

   is concave. *Hint:* show that the Hessian of $-f$ takes the form $\kappa \text{diag}\,(y) - zz^\top$ for appropriate vectors $y \geq 0$, $z \geq 0$, and scalar $\kappa \geq 0$, and use Schur complements[11] to prove that the Hessian is positive semidefinite.

Exercise 8.4 (Some simple optimization problems) Solve the following optimization problems. Make sure to determine an optimal primal solution.

1. Show that, for given scalars α, β,

   $$f(\alpha,\beta) \doteq \min_{d > 0} \alpha d + \frac{\beta^2}{d} = \begin{cases} -\infty & \text{if } \alpha \leq 0, \\ 2|\beta|\sqrt{\alpha} & \text{otherwise.} \end{cases}$$

[11] See Section 4.4.7.

2. Show that for an arbitrary vector $z \in \mathbb{R}^m$,

$$\|z\|_1 = \min_{d>0} \frac{1}{2} \sum_{i=1}^{m} \left(d_i + \frac{z_i^2}{d_i} \right). \qquad (8.11)$$

3. Show that for an arbitrary vector $z \in \mathbb{R}^m$, we have

$$\|z\|_1^2 = \min_{d} \sum_{i=1}^{m} \frac{z_i^2}{d_i} \ : \ d > 0, \ \sum_{i=1}^{m} d_i = 1.$$

Exercise 8.5 (Minimizing a sum of logarithms) Consider the following problem:

$$\begin{aligned} p^* = \max_{x \in \mathbb{R}^n} \quad & \sum_{i=1}^{n} \alpha_i \ln x_i \\ \text{s.t.:} \quad & x \geq 0, \quad \mathbf{1}^\top x = c, \end{aligned}$$

where $c > 0$ and $\alpha_i > 0$, $i = 1, \ldots, n$. Problems of this form arise, for instance, in maximum-likelihood estimation of the transition probabilities of a discrete-time Markov chain. Determine in closed-form a minimizer, and show that the optimal objective value of this problem is

$$p^* = \alpha \ln(c/\alpha) + \sum_{i=1}^{n} \alpha_i \ln \alpha_i,$$

where $\alpha \doteq \sum_{i=1}^{n} \alpha_i$.

Exercise 8.6 (Monotonicity and locality) Consider the optimization problems (no assumption of convexity here)

$$\begin{aligned} p_1^* &\doteq \min_{x \in \mathcal{X}_1} f_0(x), \\ p_2^* &\doteq \min_{x \in \mathcal{X}_2} f_0(x), \\ p_{13}^* &\doteq \min_{x \in \mathcal{X}_1 \cap \mathcal{X}_3} f_0(x), \\ p_{23}^* &\doteq \min_{x \in \mathcal{X}_2 \cap \mathcal{X}_3} f_0(x), \end{aligned}$$

where $\mathcal{X}_1 \subseteq \mathcal{X}_2$.

1. Prove that $p_1^* \geq p_2^*$ (i.e., enlarging the feasible set cannot worsen the optimal objective).

2. Prove that, if $p_1^* = p_2^*$, then it holds that

$$p_{13}^* = p_1^* \quad \Rightarrow \quad p_{23}^* = p_2^*.$$

3. Assume that all problems above attain unique optimal solutions. Prove that, under such a hypothesis, if $p_1^* = p_2^*$, then it holds that

$$p_{23}^* = p_2^* \quad \Rightarrow \quad p_{13}^* = p_1^*.$$

Exercise 8.7 (Some matrix norms) Let $X = [x_1, \ldots, x_m] \in \mathbb{R}^{n,m}$, and $p \in [1, +\infty]$. We consider the problem

$$\phi_p(X) \doteq \max_u \|X^\top u\|_p \ : \ u^\top u = 1.$$

If the data is centered, that is, $X\mathbf{1} = 0$, the above amounts to finding a direction of largest "deviation" from the origin, where deviation is measured using the l_p-norm.

1. Is ϕ_p a (matrix) norm?

2. Solve the problem for $p = 2$. Find an optimal u.

3. Solve the problem for $p = \infty$. Find an optimal u.

4. Show that
$$\phi_p(X) = \max_v \|Xv\|_2 \ : \ \|v\|_q \leq 1,$$
where $1/p + 1/q = 1$ (hence, $\phi_p(X)$ depends only on $X^\top X$). *Hint:* you can use the fact that the norm dual to the l_p-norm is the l_q-norm and vice versa, in the sense that, for any scalars $p \geq 1$, $q \geq 1$ with $1/p + 1/q = 1$, we have
$$\max_{v:\, \|v\|_q \leq 1} u^\top v = \|u\|_p.$$

Exercise 8.8 (Norms of matrices with non-negative entries) Let $X \in \mathbb{R}_+^{n,m}$ be a matrix with non-negative entries, and $p, r \in [1, +\infty]$, with $p \geq r$. We consider the problem

$$\phi_{p,r}(X) = \max_v \|Xv\|_r \ : \ \|v\|_p \leq 1.$$

1. Show that the function $f_X : \mathbb{R}_+^m \to \mathbb{R}$, with values
$$f_X(u) = \sum_{i=1}^n \left(\sum_{j=1}^m X_{ij} u_j^{1/p} \right)^r$$
is concave when $p \geq r$.

2. Use the previous result to formulate an efficiently solvable convex problem that has $\phi_{p,r}(X)^r$ as optimal value.

Exercise 8.9 (Magnitude least squares) For given n-vectors $a_1, \ldots, a_m$, we consider the problem

$$p^* = \min_x \sum_{i=1}^m \left(|a_i^\top x| - 1 \right)^2.$$

1. Is the problem convex? If so, can you formulate it as an ordinary least-squares problem? An LP? A QP? A QCQP? An SOCP? None of the above? Justify your answers precisely.

2. Show that the optimal value p^* depends only on the matrix $K = A^\top A$, where $A = [a_1, \ldots, a_m]$ is the $n \times m$ matrix of data points (that is, if two different matrices A_1, A_2 satisfy $A_1^\top A_1 = A_2^\top A_2$, then the corresponding optimal values are the same).

Exercise 8.10 (Eigenvalues and optimization) Given an $n \times n$ symmetric matrix Q, define

$$w_1 = \arg \min_{\|x\|_2 = 1} x^\top Q x, \text{ and } \mu_1 = \min_{\|x\|_2 = 1} x^\top Q x,$$

and for $k = 1, 2, \ldots, n-1$:

$$w_{k+1} = \arg \min_{\|x\|_2 = 1} x^\top Q x \quad \text{such that } w_i^\top x = 0, \, i = 1, \ldots, k,$$

$$\mu_{k+1} = \min_{\|x\|_2 = 1} x^\top Q x \quad \text{such that } w_i^\top x = 0, \, i = 1, \ldots, k.$$

Using optimization principles and theory:

1. show that $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$;

2. show that the vectors $w_1, \ldots, w_n$ are linearly independent, and form an orthonormal basis of $\mathbb{R}^n$;

3. show how μ_1 can be interpreted as a Lagrange multiplier, and that μ_1 is the smallest eigenvalue of Q;

4. show how $\mu_2, \ldots, \mu_n$ can also be interpreted as Lagrange multipliers. *Hint:* show that μ_{k+1} is the smallest eigenvalue of $W_k^\top Q W_k$, where $W_k = [w_{k+1}, \ldots, w_n]$.

Exercise 8.11 (Block norm penalty) In this exercise we partition vectors $x \in \mathbb{R}^n$ into p blocks $x = (x_1, \ldots, x_p)$, with $x_i \in \mathbb{R}^{n_i}$, $n_1 + \cdots + n_p = n$. Define the function $\rho : \mathbb{R}^n \to \mathbb{R}$ with values

$$\rho(x) = \sum_{i=1}^{p} \|x_i\|_2.$$

1. Prove that ρ is a norm.

2. Find a simple expression for the "dual norm," $\rho_*(x) \doteq \sup_{z:\, \rho(z) = 1} z^\top x$.

3. What is the dual of the dual norm?

4. For a scalar $\lambda \geq 0$, matrix $A \in \mathbb{R}^{m,n}$ and vector $y \in \mathbb{R}^m$, we consider the optimization problem

$$p^*(\lambda) \doteq \min_x \|Ax - y\|_2 + \lambda \rho(x).$$

Explain the practical effect of a high value of λ on the solution.

5. For the problem above, show that $\lambda > \sigma_{\max}(A_i)$ implies that we can set $x_i = 0$ at optimum. Here, $A_i \in \mathbb{R}^{m,n_i}$ corresponds to the i-th block of columns in A, and $\sigma_{\max}$ refers to the largest singular value.

9. Linear, Quadratic and Geometric Models

Exercise 9.1 (Formulating problems as LPs or QPs) Formulate the problem

$$p_j^* \doteq \min_x f_j(x),$$

for different functions f_j, $j = 1, \dots, 5$, with values given in Table 9.2, as QPs or LPs, or, if you cannot, explain why. In our formulations, we always use $x \in \mathbb{R}^n$ as the variable, and assume that $A \in \mathbb{R}^{m,n}$, $y \in \mathbb{R}^m$, and $k \in \{1, \dots, m\}$ are given. If you obtain an LP or QP formulation, make sure to put the problem in standard form, stating precisely what the variables, objective and constraints are. *Hint:* for the last one, see Example 9.10.

$$\begin{array}{lcl}
f_1(x) & = & \|Ax - y\|_\infty + \|x\|_1 \\
f_2(x) & = & \|Ax - y\|_2^2 + \|x\|_1 \\
f_3(x) & = & \|Ax - y\|_2^2 - \|x\|_1 \\
f_4(x) & = & \|Ax - y\|_2^2 + \|x\|_1^2 \\
f_5(x) & = & \sum_{i=1}^k |Ax - y|_{[i]} + \|x\|_2^2
\end{array}$$

Table 9.2: Table of the values of different functions f. $|z|_{[i]}$ denotes the element in a vector z that has the i-th largest magnitude.

Exercise 9.2 (A slalom problem) A two-dimensional skier must slalom down a slope, by going through n parallel gates of known position (x_i, y_i), and of width c_i, $i = 1, \dots, n$. The initial position (x_0, y_0) is given, as well as the final one, (x_{n+1}, y_{n+1}). Here, the x-axis represents the direction down the slope, from left to right, see Figure 9.7.

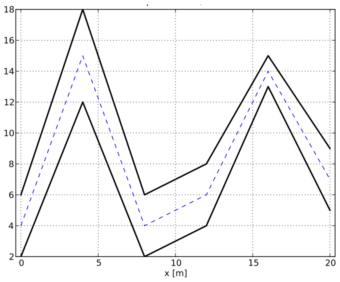

Figure 9.7: Slalom problem with $n = 5$ obstacles. "Uphill" (resp. "downhill") is on the left (resp. right) side. The middle path is dashed, initial and final positions are not shown.

1. Find the path that minimizes the total length of the path. Your answer should come in the form of an optimization problem.

2. Try solving the problem numerically, with the data given in Table 9.3.

i	x_i	y_i	c_i
0	0	4	N/A
1	4	5	3
2	8	4	2
3	12	6	2
4	16	5	1
5	20	7	2
6	24	4	N/A

Table 9.3: Problem data for Exercise 9.2.

Exercise 9.3 (Minimum distance to a line segment) The line segment linking two points $p, q \in \mathbb{R}^n$ (with $p \neq q$) is the set $\mathcal{L} = \{\lambda p + (1 - \lambda)q : 0 \leq \lambda \leq 1\}$.

1. Show that the minimum distance D_* from a point $a \in \mathbb{R}^n$ to the line segment $\mathcal{L}$ can be written as a QP in one variable:

$$\min_\lambda \|\lambda c + d\|_2^2 \ : \ 0 \leq \lambda \leq 1,$$

for appropriate vectors c, d, which you will determine. Explain why we can always assume $a = 0$.

2. Prove that the minimum distance is given by[12]

$$D_*^2 = \begin{cases} q^\top q - \frac{(q^\top(p-q))^2}{\|p-q\|_2^2} & \text{if } p^\top q \leq \min(q^\top q, p^\top p), \\ q^\top q & \text{if } p^\top q > q^\top q, \\ p^\top p & \text{if } p^\top q > p^\top p. \end{cases}$$

3. Interpret the result geometrically.

[12] Notice that the conditions expressing D_*^2 are mutually exclusive, since $|p^\top q| \leq \|p\|_2 \|q\|_2$.

Exercise 9.4 (Univariate LASSO) Consider the problem

$$\min_{x \in \mathbb{R}} f(x) \doteq \frac{1}{2}\|ax - y\|_2^2 + \lambda |x|,$$

where $\lambda \geq 0$, $a \in \mathbb{R}^m$, $y \in \mathbb{R}^m$ are given, and $x \in \mathbb{R}$ is a scalar variable. This is a univariate version of the LASSO problem discussed in Section 9.6.2. Assume that $y \neq 0$ and $a \neq 0$, (since otherwise the optimal solution of this problem is simply $x = 0$). Prove that the optimal solution of this problem is

$$x^* = \begin{cases} 0 & \text{if } |a^\top y| \leq \lambda, \\ x_{\text{ls}} - \text{sgn}(x_{\text{ls}}) \frac{\lambda}{\|a\|_2^2} & \text{if } |a^\top y| > \lambda, \end{cases}$$

where

$$x_{\text{ls}} \doteq \frac{a^\top y}{\|a\|_2^2}$$

corresponds to the solution of the problem for $\lambda = 0$. Verify that this solution can be expressed more compactly as $x^* = \text{sthr}_{\lambda/\|a\|_2^2}(x_{\text{ls}})$, where sthr is the *soft threshold* function defined in (12.65).

Exercise 9.5 (An optimal breakfast) We are given a set of $n = 3$ types of food, each of which has the nutritional characteristics described in Table 9.4. Find the optimal composition (amount of servings per each food) of a breakfast having minimum cost, number of calories between 2000 and 2250, amount of vitamin between 5000 and 10000, and sugar level no larger than 1000, assuming that the maximum number of servings is 10.

Food	Cost	Vitamin	Sugar	Calories
Corn	0.15	107	45	70
Milk	0.25	500	40	121
Bread	0.05	0	60	65

Table 9.4: Food costs and nutritional values per serving.

Exercise 9.6 (An LP with wide matrix) Consider the LP

$$p^* = \min_x c^\top x \ : \ l \leq Ax \leq u,$$

where $A \in \mathbb{R}^{m,n}$, $c \in \mathbb{R}^n$, and $l, u \in \mathbb{R}^m$, with $l \leq u$. We assume that A is wide, and full rank, that is: $m \leq n$, $m = \text{rank}(A)$. We are going to develop a closed-form solution to the LP.

1. Explain why the problem is always feasible.

2. Assume that $c \notin \mathcal{R}(A^\top)$. Using the result of Exercise 6.2, show that $p^* = -\infty$. *Hint:* set $x = x_0 + tr$, where x_0 is feasible, r is such that $Ar = 0$, $c^\top r > 0$, and let $t \to -\infty$.

3. Now assume that there exists $d \in \mathbb{R}^m$ such that $c = A^\top d$. Using the fundamental theorem of linear algebra (see Section 3.2.4), any vector x can be written as $x = A^\top y + z$ for some pair (y, z) with $Az = 0$. Use this fact, and the result of the previous part, to express the problem in terms of the variable y only.

4. Reduce further the problem to one of the form

$$\min_v d^\top v \ : \ l \leq v \leq u.$$

Make sure to justify any change of variable you may need. Write the solution to the above in closed form. Make sure to express the solution steps of the method clearly.

Exercise 9.7 (Median versus average) For a given vector $v \in \mathbb{R}^n$, the average can be found as the solution to the optimization problem

$$\min_{x \in \mathbb{R}} \|v - x\mathbf{1}\|_2^2, \tag{9.12}$$

where $\mathbf{1}$ is the vector of ones in $\mathbb{R}^n$. Similarly, it turns out that the median (any value x such that there is an equal number of values in v above or below x) can be found via

$$\min_{x \in \mathbb{R}} \|v - x\mathbf{1}\|_1. \tag{9.13}$$

We consider a robust version of the average problem (9.12):

$$\min_x \max_{u \,:\, \|u\|_\infty \leq \lambda} \|v + u - x\mathbf{1}\|_2^2, \tag{9.14}$$

in which we assume that the components of v can be independently perturbed by a vector u whose magnitude is bounded by a given number $\lambda \geq 0$.

1. Is the robust problem (9.14) convex? Justify your answer precisely, based on expression (9.14), and without further manipulation.

2. Show that problem (9.14) can be expressed as

$$\min_{x \in \mathbb{R}} \sum_{i=1}^n (|v_i - x| + \lambda)^2.$$

3. Express the problem as a QP. State precisely the variables, and constraints if any.

4. Show that when λ is large, the solution set approaches that of the median problem (9.13).

5. It is often said that the median is a more robust notion of "middle" value than the average, when noise is present in v. Based on the previous part, justify this statement.

Exercise 9.8 (Convexity and concavity of optimal value of an LP) Consider the linear programming problem

$$p^* \doteq \min_x c^\top x \ : \ Ax \leq b,$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m,n}$, $b \in \mathbb{R}^m$. Prove the following statements, or provide a counter-example.

1. The objective function p^* is a concave function of c.
2. The objective function p^* is a convex function of b (you may assume that the problem is feasible).
3. The objective function p^* is a concave function of A.

Exercise 9.9 (Variational formula for the dominant eigenvalue) Recall from Exercise 3.11 that a positive matrix $A > 0$ has a dominant eigenvalue $\lambda = \rho(A) > 0$, and corresponding left eigenvector $w > 0$ and right eigenvector $v > 0$ (i.e., $w^\top A = \lambda w^\top$, $Av = \lambda v$) which belong to the probability simplex $S = \{x \in \mathbb{R}^n \ : \ x \geq 0, \mathbf{1}^\top x = 1\}$. In this exercise, we shall prove that the dominant eigenvalue has an optimization-based characterization, similar in spirit to the "variational" characterization of the eigenvalues of symmetric matrices. Define the function $f : S \to \mathbb{R}_{++}$ with values

$$f(x) \doteq \min_{i=1,\ldots,n} \frac{a_i^\top x}{x_i}, \quad \text{for } x \in S,$$

where $a_i^\top$ is the i-th row of A, and we let $\frac{a_i^\top x}{x_i} \doteq +\infty$ if $x_i = 0$.

1. Prove that, for all $x \in S$ and $A > 0$, it holds that
$$Ax \geq f(x)x \geq 0.$$
2. Prove that
$$f(x) \leq \lambda, \quad \forall x \in S.$$
3. Show that $f(v) = \lambda$, and hence conclude that
$$\lambda = \max_{x \in S} f(x),$$
which is known as the Collatz–Wielandt formula for the dominant eigenvalue of a positive matrix. This formula actually holds more generally for non-negative matrices,[13] but you are not asked to prove this fact.

[13] For a non-negative matrix $A \geq 0$ an extension of the results stated in Exercise 3.11 for positive matrices holds. More precisely, if $A \geq 0$, then $\lambda = \rho(A) \geq 0$ is still an eigenvalue of A, with a corresponding eigenvector $v \geq 0$ (the difference here being that λ could be zero, and not simple, and that v may not be strictly positive). The stronger results of $\lambda > 0$ and simple, and $v > 0$ are recovered under the additional assumption that $A \geq 0$ is *primitive*, that is there exist an integer k such that $A^k > 0$ (Perron–Frobenius theorem).

Exercise 9.10 (LS with uncertain A matrix) Consider a linear least-squares problem where the matrix involved is random. Precisely, the residual vector is of the form $A(\delta)x - b$, where the $m \times n$ A matrix is affected by stochastic uncertainty. In particular, assume that

$$A(\delta) = A_0 + \sum_{i=1}^{p} A_i \delta_i,$$

where δ_i, $i = 1, \ldots, p$ are i.i.d. random variables with zero mean and variance σ_i^2. The standard least-squares objective function $\|A(\delta)x - b\|_2^2$ is now random, since it depends on δ. We seek to determine x such that the expected value (with respect to the random variable δ) of $\|A(\delta)x - b\|_2^2$ is minimized. Is such a problem convex? If yes, to which class does it belong to (LP, LS, QP, etc.)?

10. Second-Order Cone and Robust Models

Exercise 10.1 (Squaring SOCP constraints) When considering a second-order cone constraint, a temptation might be to square it in order to obtain a classical convex quadratic constraint. This might not always work. Consider the constraint

$$x_1 + 2x_2 \geq \|x\|_2,$$

and its squared counterpart:

$$(x_1 + 2x_2)^2 \geq \|x\|_2^2.$$

Is the set defined by the second inequality convex? Discuss.

Exercise 10.2 (A complicated function) We would like to minimize the function $f : \mathbb{R}^3 \to \mathbb{R}$, with values:

$$\begin{aligned} f(x) \quad = \quad & \max \Bigg(x_1 + x_2 - \min \big(\min(x_1 + 2, x_2 + 2x_1 - 5), x_3 - 6 \big), \\ & \frac{(x_1 - x_3)^2 + 2x_2^2}{1 - x_1} \Bigg), \end{aligned}$$

with the constraint $\|x\|_\infty < 1$. Explain precisely how to formulate the problem as an SOCP in standard form.

Exercise 10.3 (A minimum time path problem) Consider Figure 10.8, in which a point in 0 must move to reach point $p = [4 \;\; 2.5]^\top$, crossing three layers of fluids having different densities.

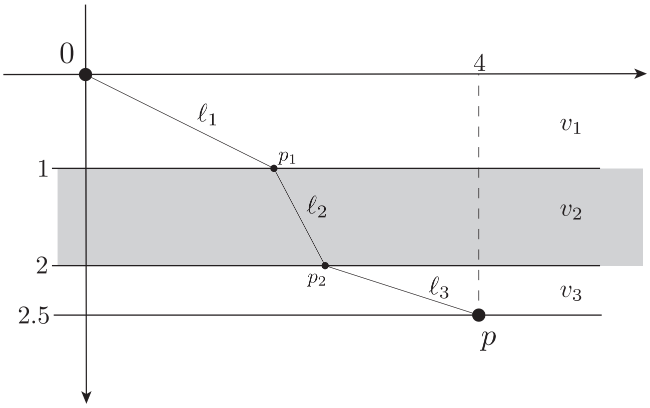

Figure 10.8: A minimum-time path problem.

In the first layer, the point can travel at a maximum speed v_1, while in the second layer and third layers it may travel at lower maximum speeds, respectively $v_2 = v_1/\eta_2$, and $v_3 = v_1/\eta_3$, with $\eta_2, \eta_3 >$

1. Assume $v_1 = 1$, $\eta_2 = 1.5$, $\eta_3 = 1.2$. You have to determine what is the fastest (i.e., minimum time) path from 0 to p. *Hint:* you may use path leg lengths ℓ_1, ℓ_2, ℓ_3 as variables, and observe that, in this problem, equality constraints of the type $\ell_i =$ "something" can be equivalently substituted by inequality constraints $\ell_i \geq$ "something" (explain why).

Exercise 10.4 (k-ellipses) Consider k points $x_1, \ldots, x_k$ in $\mathbb{R}^2$. For a given positive number d, we define the k-ellipse with radius d as the set of points $x \in \mathbb{R}^2$ such that the sum of the distances from x to the points x_i is equal to d.

1. How do k-ellipses look when $k = 1$ or $k = 2$? *Hint:* for $k = 2$, show that you can assume $x_1 = -x_2 = p$, $\|p\|_2 = 1$, and describe the set in an orthonormal basis of $\mathbb{R}^n$ such that p is the first unit vector.

2. Express the problem of computing the *geometric median*, which is the point that minimizes the sum of the distances to the points x_i, $i = 1, \ldots, k$, as an SOCP in standard form.

3. Write a code with input $X = (x_1, \ldots, x_k) \in \mathbb{R}^{2,k}$ and $d > 0$ that plots the corresponding k-ellipse.

Exercise 10.5 (A portfolio design problem) The returns on $n = 4$ assets are described by a Gaussian (normal) random vector $r \in \mathbb{R}^n$, having the following expected value $\hat{r}$ and covariance matrix Σ:

$$\hat{r} = \begin{bmatrix} 0.12 \\ 0.10 \\ 0.07 \\ 0.03 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 0.0064 & 0.0008 & -0.0011 & 0 \\ 0.0008 & 0.0025 & 0 & 0 \\ -0.0011 & 0 & 0.0004 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The last (fourth) asset corresponds to a risk-free investment. An investor wants to design a portfolio mix with weights $x \in \mathbb{R}^n$ (each weight x_i is non-negative, and the sum of the weights is one) so as to obtain the best possible expected return $\hat{r}^\top x$, while guaranteeing that: (i) no single asset weights more than 40%; (ii) the risk-free assets should not weight more than 20%; (iii) no asset should weight less than 5%; (iv) the probability of experiencing a return lower than $q = -3\%$ should be no larger than $\epsilon = 10^{-4}$. What is the maximal achievable expected return, under the above constraints?

Exercise 10.6 (A trust-region problem) A version of the so-called (convex) *trust-region* problem amounts to finding the minimum of a convex quadratic function over a Euclidean ball, that is

$$\begin{array}{ll}\min\limits_{x} & \frac{1}{2}x^\top Hx + c^\top x + d \\ \text{s.t.:} & x^\top x \leq r^2,\end{array}$$

where $H \succ 0$, and $r > 0$ is the given radius of the ball. Prove that the optimal solution to this problem is unique and is given by

$$x(\lambda^*) = -(H + \lambda^* I)^{-1}c,$$

where $\lambda^* = 0$ if $\|H^{-1}c\|_2 \leq r$, or otherwise λ^* is the unique value such that $\|(H + \lambda^* I)^{-1}c\|_2 = r$.

Exercise 10.7 (Univariate square-root LASSO) Consider the problem

$$\min_{x\in\mathbb{R}} f(x) \doteq \|ax - y\|_2 + \lambda|x|,$$

where $\lambda \geq 0$, $a \in \mathbb{R}^m$, $y \in \mathbb{R}^m$ are given, and $x \in \mathbb{R}$ is a scalar variable. This is a univariate version of the square-root LASSO problem introduced in Example 8.23. Assume that $y \neq 0$ and $a \neq 0$, (since otherwise the optimal solution of this problem is simply $x = 0$). Prove that the optimal solution of this problem is

$$x^* = \begin{cases} 0 & \text{if } |a^\top y| \leq \lambda\|y\|_2, \\ x_{\text{ls}} - \text{sgn}(x_{\text{ls}})\frac{\lambda}{\|a\|_2^2}\sqrt{\frac{\|a\|_2^2\|y\|_2^2-(a^\top y)^2}{\|a\|_2^2-\lambda^2}} & \text{if } |a^\top y| > \lambda\|y\|_2,\end{cases}$$

where

$$x_{\text{ls}} \doteq \frac{a^\top y}{\|a\|_2^2}.$$

Exercise 10.8 (Proving convexity via duality) Consider the function $f : \mathbb{R}^n_{++} \to \mathbb{R}$, with values

$$f(x) = 2\max_t \; t - \sum_{i=1}^n \sqrt{x_i + t^2}.$$

1. Explain why the problem that defines f is a convex optimization problem (in the variable t). Formulate it as an SOCP.

2. Is f convex?

3. Show that the function $g : \mathbb{R}^n_{++} \to \mathbb{R}$, with values

$$g(y) = \sum_{i=1}^n \frac{1}{y_i} - \frac{1}{\sum\limits_{i=1}^n y_i}$$

   is convex. *Hint:* for a given $y \in \mathbb{R}^n_{++}$, show that

$$g(y) = \max_{x>0} \; -x^T y - f(x).$$

   Make sure to justify any use of strong duality.

Exercise 10.9 (Robust sphere enclosure) Let B_i, $i = 1, \dots, m$, be m given Euclidean balls in $\mathbb{R}^n$, with centers x_i and radii $\rho_i \geq 0$. We wish to find a ball B of minimum radius that contains all the B_i, $i = 1, \dots, m$. Explain how to cast this problem into a known convex optimization format.

11. Semidefinite Models

Exercise 11.1 (Minimum distance to a line segment revisited) In this exercise, we revisit Exercise 9.3, and approach it using the $\mathcal{S}$-procedure of Section 11.3.3.1.

1. Show that the minimum distance from the line segment $\mathcal{L}$ to the origin is above a given number $R \geq 0$ if and only if

$$\|\lambda(p-q)+q\|_2^2 \geq R^2 \text{ whenever } \lambda(1-\lambda) \geq 0.$$

2. Apply the $\mathcal{S}$-procedure, and prove that the above is in turn equivalent to the LMI in $\tau \geq 0$:

$$\left[\begin{array}{cc} \|p-q\|_2^2+\tau & q^\top(p-q)-\tau/2 \\ q^\top(p-q)-\tau/2 & q^\top q - R^2 \end{array}\right] \succeq 0.$$

3. Using the Schur complement rule,[14] show that the above is consistent with the result given in Exercise 9.3.

Exercise 11.2 (A variation on principal component analysis) Let $X = [x_1, \ldots, x_m] \in \mathbb{R}^{n,m}$. For $p = 1, 2$, we consider the problem

$$\phi_p(X) \doteq \max_u \sum_{i=1}^m |x_i^\top u|^p \; : \; u^\top u = 1. \tag{11.15}$$

If the data is centered, the case $p = 1$ amounts of finding a direction of largest "deviation" from the origin, where deviation is measured using the ℓ_1-norm; arguably, this is less sensitive to outliers than the case $p = 2$, which corresponds to principal component analysis.

1. Find an expression for ϕ_2, in terms of the singular values of X.

2. Show that the problem, for $p = 1$, can be approximated via an SDP, as $\phi_1(X) \leq \psi_1(X)$, where

$$\psi_1(X) \doteq \max_U \sum_{i=1}^m \sqrt{x_i^\top U x_i} \; : \; U \succeq 0, \;\; \mathrm{trace}\, U = 1.$$

   Is ψ_1 a norm?

3. Formulate a dual to the above expression. Does strong duality hold? *Hint:* introduce new variables $z_i = x_i^\top U x_i$, $i = 1, \ldots, m$, and dualize the corresponding constraints.

4. Use the identity (8.11) to approximate, via weak duality, the problem (11.15). How does your bound compare with ψ_1?

[14] See Theorem 4.9.

5. Show that

$$\psi_1(X)^2 = \min_D \operatorname{trace} D \ : \ D \text{ diagonal}, \ D \succ 0, \ \ D \succeq X^\top X.$$

Hint: scale the variables in the dual problem and optimize over the scaling. That is, set $D = \alpha \bar{D}$, with $\lambda_{\max}(X\tilde{D}^{-1}X^\top) = 1$ and $\alpha > 0$, and optimize over α. Then argue that we can replace the equality constraint on $\tilde{D}$ by a convex inequality, and use Schur complements to handle that corresponding inequality.

6. Show that

$$\phi_1(X) = \max_{v \,:\, \|v\|_\infty \leq 1} \|Xv\|_2.$$

Is the maximum always attained with a vector v such that $|v_i| = 1$ for every i? *Hint:* use the fact that

$$\|z\|_1 = \max_{v \,:\, \|v\|_\infty \leq 1} z^\top v.$$

7. A result by Yu. Nesterov[15] shows that for any symmetric matrix $Q \in \mathbb{R}^{m,m}$, the problem

$$p^* = \max_{v \,:\, \|v\|_\infty \leq 1} v^\top Q v$$

can be approximated within $\pi/2$ relative value via SDP. Precisely, $(2/\pi)d^* \leq p^* \leq d^*$, where

$$d^* = \min_D \operatorname{trace} D \ : \ D \text{ diagonal}, \ D \succeq Q. \qquad (11.16)$$

Use this result to show that

$$\sqrt{\frac{2}{\pi}}\psi_1(X) \leq \phi_1(X) \leq \psi_1(X).$$

That is, the SDP approximation is within ≈ 80% of the true value, irrespective of the problem data.

8. Discuss the respective complexity of the problems of computing ϕ_2 and ψ_1 (you can use the fact that, for a given $m \times m$ symmetric matrix Q, the SDP (11.16) can be solved in $O(m^3)$).

Exercise 11.3 (Robust principal component analysis) The following problem is known as robust principal component analysis:[16]

$$p^* \doteq \min_X \|A - X\|_* + \lambda \|X\|_1,$$

where $\|\cdot\|_*$ stands for the nuclear norm,[17] and $\|\cdot\|_1$ here denotes the sum of the absolute values of the elements of a matrix. The interpretation is the following: A is a given data matrix and we would like to

[15] Yu. Nesterov, Quality of semidefinite relaxation for nonconvex quadratic optimization, discussion paper, CORE, 1997.

[16] See Section 13.5.4.

[17] The nuclear norm is the sum of the singular values of the matrix; see Section 11.4.1.4 and Section 5.2.2.

decompose it as a sum of a low rank matrix and a sparse matrix. The nuclear norm and ℓ_1 norm penalties are respective convex heuristics for these two properties. At optimum, X^* will be the sparse component and $A - X^*$ will be the low rank component such that their sum gives A.

1. Find a dual for this problem. *Hint:* we have, for any matrix W:

$$\|W\|_* = \max_Y \operatorname{trace} W^\top Y \ : \ \|Y\|_2 \le 1,$$

   where $\|\cdot\|_2$ is the largest singular value norm.

2. Transform the primal or dual problem into a known programming class (i.e. LP, SOCP, SDP, etc.). Determine the number of variables and constraints. *Hint:* we have

$$\|Y\|_2 \le 1 \Longleftrightarrow I - YY^\top \succeq 0,$$

   where I is the identity matrix.

3. Using the dual, show that when $\lambda > 1$, the optimal solution is the zero matrix. *Hint:* if Y^* is the optimal dual variable, the complementary slackness condition states that $|Y^*_{ij}| < \lambda$ implies $X^*_{ij} = 0$ at optimum.

Exercise 11.4 (Boolean least squares) Consider the following problem, known as *Boolean least squares*:

$$\phi = \min_x \|Ax - b\|_2^2 \ : \ x_i \in \{-1, 1\}, \ \ i = 1, \ldots, n.$$

Here, the variable is $x \in \mathbb{R}^n$, where $A \in \mathbb{R}^{m,n}$ and $b \in \mathbb{R}^m$ are given. This is a basic problem arising, for instance, in digital communications. A brute force solution is to check all 2^n possible values of x, which is usually impractical.

1. Show that the problem is equivalent to

$$\begin{aligned} \phi = \min_{X,x} \quad & \operatorname{trace}(A^\top A X) - 2b^\top Ax + b^\top b \\ \text{s.t.:} \quad & X = xx^\top, \\ & X_{ii} = 1, \quad i = 1, \ldots, n, \end{aligned}$$

   in the variables $X = X^\top \in \mathbb{R}^{n,n}$ and $x \in \mathbb{R}^n$.

2. The constraint $X = xx^\top$, i.e., the set of rank-1 matrices, is not convex, therefore the problem is still hard. However, an efficient approximation can be obtained by relaxing this constraint to $X \succeq xx^\top$, as discussed in Section 11.3.3, obtaining

$$\begin{array}{rll} \phi \geq \phi_{\text{sdp}} = \min\limits_{X} & \operatorname{trace}(A^\top A X) - 2b^\top A x + b^\top b \\ \text{s.t.:} & \begin{bmatrix} X & x \\ x^\top & 1 \end{bmatrix} \succeq 0, \\ & X_{ii} = 1, \quad i = 1, \ldots, n. \end{array}$$

The relaxation produces a lower-bound to the original problem. Once that is done, an approximate solution to the original problem can be obtained by rounding the solution: $x_{\text{sdp}} = \operatorname{sgn}(x^*)$, where x^* is the optimal solution of the semidefinite relaxation.

3. Another approximation method is to relax the non-convex constraints $x_i \in \{-1, 1\}$ to convex interval constraints $-1 \leq x_i \leq 1$ for all i, which can be written $\|x\|_\infty \leq 1$. Therefore, a different lower bound is given by:

$$\phi \geq \phi_{\text{int}} \doteq \min \|Ax - b\|_2^2 \;:\; \|x\|_\infty \leq 1.$$

   Once that problem is solved, we can round the solution by $x_{\text{int}} = \operatorname{sgn}(x^*)$ and compare the original objective value $\|Ax_{\text{int}} - b\|_2^2$.

4. Which one of ϕ_{sdp} and ϕ_{int} produces the closest approximation to ϕ? Justify your answer carefully.

5. Use now 100 independent realizations with normally distributed data, $A \in \mathbb{R}^{10,10}$ (independent entries with mean zero) and $b \in \mathbb{R}^{10}$ (independent entries with mean 1). Plot and compare the histograms of $\|Ax_{\text{sdp}} - b\|_2^2$ of part 2, $\|Ax_{\text{int}} - b\|_2^2$ of part 3, and the objective corresponding to a naïve method $\|Ax_{\text{ls}} - b\|_2^2$, where $x_{\text{ls}} = \operatorname{sgn}\left((A^\top A)^{-1} A^\top b\right)$ is the rounded ordinary least squares solution. Briefly discuss accuracy and computation time (in seconds) of the three methods.

6. Assume that, for some problem instance, the optimal solution (x, X) found via the SDP approximation is such that x belongs to the original non-convex constraint set $\{x : x_i \in \{-1, 1\}, \quad i = 1, \ldots, n\}$. What can you say about the SDP approximation in that case?

Exercise 11.5 (Auto-regressive process model) We consider a process described by the difference equation

$$y(t+2) = \alpha_1(t) y(t+1) + \alpha_2(t) y(t) + \alpha_3(t) u(t), \;\; t = 0, 1, 2, \ldots,$$

where the $u(t) \in \mathbb{R}$ is the input, $y(t) \in \mathbb{R}$ the output, and the coefficient vector $\alpha(t) \in \mathbb{R}^3$ is time-varying. We seek to compute bounds on the vector $\alpha(t)$ that are (a) independent of t, (b) consistent with some given historical data.

The specific problem we consider is: given the values of $u(t)$ and $y(t)$ over a time period $1 \leq t \leq T$, find the smallest ellipsoid $\mathcal{E}$ in $\mathbb{R}^3$ such that, for every t, $1 \leq t \leq T$, the equation above is satisfied for some $\alpha(t) \in \mathcal{E}$.

1. What is a geometrical interpretation of the problem, in the space of αs?

2. Formulate the problem as a semidefinite program. You are free to choose the parameterization, as well as the measure of the size of $\mathcal{E}$ that you find most convenient.

3. Assume we restrict our search to spheres instead of ellipsoids. Show that the problem can be reduced to a linear program.

4. In the previous setting, $\alpha(t)$ is allowed to vary with time arbitrarily fast, which may be unrealistic. Assume that a bound is imposed on the variation of $\alpha(t)$, such as $\|\alpha(t+1) - \alpha(t)\|_2 \leq \beta$, where $\beta > 0$ is given. How would you solve the problem with this added restriction?

Exercise 11.6 (Non-negativity of polynomials) A second-degree polynomial with values $p(x) = y_0 + y_1 x + y_2 x^2$ is non-negative everywhere if and only if

$$\forall\, x \;:\; \begin{bmatrix} x \\ 1 \end{bmatrix}^\top \begin{bmatrix} y_0 & y_1/2 \\ y_1/2 & y_2 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} \geq 0,$$

which in turn can be written as an LMI in $y = (y_0, y_1, y_2)$:

$$\begin{bmatrix} y_0 & y_1/2 \\ y_1/2 & y_2 \end{bmatrix} \succeq 0.$$

In this exercise, you show a more general result, which applies to any polynomial of even degree $2k$ (polynomials of odd degree can't be non-negative everywhere). To simplify, we only examine the case $k = 2$, that is, fourth-degree polynomials; the method employed here can be generalized to $k > 2$.

1. Show that a fourth-degree polynomial p is non-negative everywhere if and only if it is a sum of squares, that is, it can be written as
$$p(x) = \sum_{i=1}^{4} q_i(x)^2,$$

where q_is are polynomials of degree at most two. *Hint:* show that p is non-negative everywhere if and only if it is of the form

$$p(x) = p_0\left((x-a_1)^2 + b_1^2\right)\left((x-a_2)^2 + b_2^2\right),$$

for some appropriate real numbers a_i, b_i, $i = 1, 2$, and some $p_0 \geq 0$.

2. Using the previous part, show that if a fourth-degree polynomial is a sum of squares, then it can be written as

$$p(x) = \begin{bmatrix} 1 & x & x^2 \end{bmatrix} Q \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix} \tag{11.17}$$

for some positive semidefinite matrix Q.

3. Show the converse: if a positive semidefinite matrix Q satisfies condition (11.17) for every x, then p is a sum of squares. *Hint:* use a factorization of Q of the form $Q = AA^\top$, for some appropriate matrix A.

4. Show that a fourth-degree polynomial $p(x) = y_0 + y_1x + y_2x^2 + y_3x^3 + y_4x^4$ is non-negative everywhere if and only if there exists a 3×3 matrix Q such that

$$Q \succeq 0, \;\; y_{l-1} = \sum_{i+j=l+1} Q_{ij}, \;\; l = 1, \ldots, 5.$$

Hint: equate the coefficients of the powers of x in the left and right sides of equation (11.17).

Exercise 11.7 (Sum of top eigenvalues) For $X \in \mathbb{S}^n$, and $i \in \{1, \ldots, n\}$, we denote by $\lambda_i(X)$ the i-th largest eigenvalue of X. For $k \in \{1, \ldots, n\}$, we define the function $f_k : \mathbb{S}^n \to \mathbb{R}$ with values

$$f_k(X) = \sum_{i=1}^{k} \lambda_i(X).$$

This function is an intermediate between the largest eigenvalue (obtained with $k = 1$) and the trace (obtained with $k = n$).

1. Show that for every $t \in \mathbb{R}$, we have $f_k(X) \leq t$ if and only if there exist $Z \in \mathbb{S}^n$ and $s \in \mathbb{R}$ such that

$$t - ks - \text{trace}(Z) \geq 0, \;\; Z \succeq 0, \;\; Z - X + sI \succeq 0.$$

Hint: for the sufficiency part, think about the interlacing property[18] of the eigenvalues.

[18] See Eq. (4.6)

2. Show that f_k is convex. Is it a norm?

3. How would you generalize these results to the function that assigns the sum of the top k singular values to a general rectangular $m \times n$ matrix, with $k \leq \min(m, n)$? *Hint:* for $X \in \mathbb{R}^{m,n}$, consider the symmetric matrix

$$\tilde{X} \doteq \begin{bmatrix} 0 & X \\ X^\top & 0 \end{bmatrix}.$$

12. Introduction to Algorithms

Exercise 12.1 (Successive projections for linear inequalities) Consider a system of linear inequalities $Ax \leq b$, with $A \in \mathbb{R}^{m,n}$, where $a_i^\top$, $i = 1, \ldots, m$, denote the rows of A, which are assumed, without loss of generality, to be nonzero. Each inequality $a_i^\top x \leq b_i$ can be normalized by dividing both terms by $\|a_i\|_2$, hence we shall further assume without loss of generality that $\|a_i\|_2 = 1$, $i = 1, \ldots, m$.

Consider now the case when the polyhedron described by these inequalities, $\mathcal{P} \doteq \{x : Ax \leq b\}$ is nonempty, that is, there exists at least a point $\bar{x} \in \mathcal{P}$. In order to find a feasible point (i.e., a point in $\mathcal{P}$), we propose the following simple algorithm. Let k denote the iteration number and initialize the algorithm with any initial point $x_k = x_0$ at $k = 0$. If $a_i^\top x_k \leq b_i$ holds for all $i = 1, \ldots, m$, then we have found the desired point, hence we return x_k, and finish. If instead there exists i_k such that $a_{i_k}^\top x_k > b_{i_k}$, then we set $s_k \doteq a_{i_k}^\top x_k - b_{i_k}$, we update[19] the current point as

$$x_{k+1} = x_k - s_k a_{i_k},$$

and we iterate the whole process.

1. Give a simple geometric interpretation of this algorithm.
2. Prove that this algorithm either finds a feasible solution in a finite number of iterations, or it produces a sequence of solutions $\{x_k\}$ that converges asymptotically (i.e., for $k \to \infty$) to a feasible solution (if one exists).
3. The problem of finding a feasible solution for linear inequalities can be also put in relation with the minimization of the nonsmooth function $f_0(x) = \max_{i=1,\ldots,m}(a_i^\top x_k - b_i)$. Develop a subgradient-type algorithm for this version of the problem, discuss hypotheses that need be assumed to guarantee convergence, and clarify the relations and similarities with the previous algorithm.

Exercise 12.2 (Conditional gradient method) Consider a constrained minimization problem

$$p^* = \min_{x \in \mathcal{X}} f_0(x), \tag{12.18}$$

where f_0 is convex and smooth and $\mathcal{X} \subseteq \mathbb{R}^n$ is convex and compact. Clearly, a projected gradient or proximal gradient algorithm could be applied to this problem, if the projection onto $\mathcal{X}$ is easy to compute. When this is not the case, the following alternative algorithm has been proposed.[20] Initialize the iterations with some $x_0 \in \mathcal{X}$, and set

[19] This algorithm is a version of the so-called Agmon–Motzkin–Shoenberg *relaxation method* for linear inequalities, which dates back to 1953.

[20] Versions of this algorithm are known as the Franke–Wolfe algorithm, which was developed in 1956 for quadratic f_0, or as the Levitin–Polyak *conditional gradient algorithm* (1966).

$k=0$. Determine the gradient $g_k \doteq \nabla f_0(x_k)$ and solve

$$z_k = \arg\min_{x \in \mathcal{X}} \; g_k^T x.$$

Then update the current point as

$$x_{k+1} = (1-\gamma_k)x_k + \gamma_k z_k,$$

where $\gamma_k \in [0,\, 1]$, and, in particular, we choose

$$\gamma_k = \frac{2}{k+2}, \quad k = 0, 1, \ldots$$

Assume that f_0 has a Lipschitz continuous gradient with Lipschitz constant[21] L, and that $\|x-y\|_2 \le R$ for every $x, y \in \mathcal{X}$. In this exercise, you shall prove that

$$\delta_k \doteq f_0(x_k) - p^* \le \frac{2LR^2}{k+2}, \; k = 1, 2, \ldots \tag{12.19}$$

1. Using the inequality

$$f_0(x) - f_0(x_k) \le \nabla f_0(x_k)^T(x-x_k) + \frac{L}{2}\|x-x_k\|_2^2,$$

which holds for any convex f_0 with Lipschitz continuous gradient,[22] prove that

$$f_0(x_{k+1}) \le f_0(x_k) + \gamma_k \nabla f_0(x_k)^T(z_k - x_k) + \gamma_k^2 \frac{LR^2}{2}.$$

Hint: write the inequality condition above, for $x = x_{k+1}$.

2. Show that the following recursion holds for δ_k:

$$\delta_{k+1} \le (1-\gamma_k)\delta_k + \gamma_k^2 C, \quad k = 0, 1, \ldots,$$

for $C \doteq \frac{LR^2}{2}$. *Hint:* use the optimality condition for z_k, and the convexity inequality $f_0(x^*) \ge f_0(x_k) + \nabla f_0(x_k)^T(x^* - x_k)$.

3. Prove by induction on k the desired result (12.19).

Exercise 12.3 (Bisection method) The bisection method applies to one-dimensional convex problems[23] of the form

$$\min_x \; f(x) \; : \; x_l \le x \le x_u,$$

where $x_l < x_u$ are both finite, and $f : \mathbb{R} \to \mathbb{R}$ is convex. The algorithm is initialized with the upper and lower bounds on x: $\underline{x} = x_l$, $\overline{x} = x_u$, and the initial x is set as the midpoint

$$x = \frac{\underline{x} + \overline{x}}{2}.$$

Then the algorithm updates the bounds as follows: a subgradient g of f at x is evaluated; if $g < 0$, we set $\underline{x} = x$; otherwise,[24] we set $\overline{x} = x$. Then the midpoint x is recomputed, and the process is iterated until convergence.

[21] As defined in Section 12.1.1.

[22] See Lemma 12.1.

[23] See an application in Section 11.4.1.3.

[24] Actually, if $g = 0$ then the algorithm may stop and return x as an optimal solution.

1. Show that the bisection method locates a solution x^* within accuracy ϵ in at most $\log_2(x_u - x_l)/\epsilon - 1$ steps.

2. Propose a variant of the bisection method for solving the unconstrained problem $\min_x f(x)$, for convex f.

3. Write a code to solve the problem with the specific class of functions $f : \mathbb{R} \to \mathbb{R}$, with values

$$f(x) = \sum_{i=1}^{n} \max_{1 \le j \le m} \left(\frac{1}{2} A_{ij} x^2 + B_{ij} x + C_{ij} \right),$$

where A, B, C are given $n \times m$ matrices, with every element of A non-negative.

Exercise 12.4 (KKT conditions) Consider the optimization problem[25]

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & \sum\limits_{i=1}^{n} \left(\frac{1}{2} d_i x_i^2 + r_i x_i \right) \\ \text{s.t.:} & a^\top x = 1, \;\; x_i \in [-1, 1], \quad i = 1, \ldots, n, \end{array}$$

where $a \neq 0$ and $d > 0$.

1. Verify if strong duality holds for this problem, and write down the KKT optimality conditions.

2. Use the KKT conditions and/or the Lagrangian to come up with the fastest algorithm you can to solve this optimization problem.

3. Analyze the running time complexity of your algorithm. Does the empirical performance of your method agree with your analysis?

Exercise 12.5 (Sparse Gaussian graphical models) We consider the following problem in a symmetric $n \times n$ matrix variable X

$$\max_X \log\det X - \operatorname{trace}(SX) - \lambda \|X\|_1 \;:\; X \succ 0,$$

where $S \succeq 0$ is a (given) empirical covariance matrix, $\|X\|_1$ denotes the sum of the absolute values of the elements of the positive definite matrix X, and $\lambda > 0$ encourages the sparsity in the solution X. The problem arises when fitting a multivariate Gaussian graphical model to data.[26] The ℓ_1-norm penalty encourages the random variables in the model to become conditionally independent.

1. Show that the dual of the problem takes the form

$$\min_U -\log\det(S + U) \;:\; |U_{ij}| \le \lambda.$$

[25] Problem due to Suvrit Sra (2013).

[26] See Section 13.5.5.

2. We employ a block-coordinate descent method to solve the dual. Show that if we optimize over one column and row of U at a time, we obtain a sub-problem of the form

$$\min_x \; x^\top Q x \;:\; \|x - x_0\|_\infty \le 1,$$

where $Q \succeq 0$ and $x_0 \in \mathbb{R}^{n-1}$ are given. Make sure to provide the expression of Q, x_0 as functions of the initial data, and the index of the row/column that is to be updated.

3. Show how you can solve the constrained QP problem above using the following methods. Make sure to state precisely the algorithm's steps.

   - Coordinate descent.
   - Dual coordinate ascent.
   - Projected subgradient.
   - Projected subgradient method for the dual.
   - Interior-point method (any flavor will do).

   Compare the performance (e.g., theoretical complexity, running time/convergence time on synthetic data) of these methods.

4. Solve the problem (using block-coordinate descent with five updates of each row/column, each step requiring the solution of the QP above) for a data file of your choice. Experiment with different values of λ, report on the graphical model obtained.

Exercise 12.6 (Polynomial fitting with derivative bounds)
In Section 13.2, we examined the problem of fitting a polynomial of degree d through m data points $(u_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots, m$. Without loss of generality, we assume that the input satisfies $|u_i| \le 1$, $i = 1, \ldots, m$. We parameterize a polynomial of degree d via its coefficients:

$$p_w(u) = w_0 + w_1 u + \cdots + w_d u^d,$$

where $w \in \mathbb{R}^{d+1}$. The problem can be written as

$$\min_w \; \|\Phi^\top w - y\|_2^2,$$

where the matrix Φ has columns $\phi_i = (1, u_i, \ldots, u_i^d)$, $i = 1, \ldots, m$. As detailed in Section 13.2.3, in practice it is desirable to encourage polynomials that are not too rapidly varying over the interval of interest. To that end, we modify the above problem as follows:

$$\min_w \; \|\Phi^\top w - y\|_2^2 + \lambda b(w), \tag{12.20}$$

where $\lambda > 0$ is a regularization parameter, and $b(w)$ is a bound on the size of the derivative of the polynomial over $[-1, 1]$:

$$b(w) = \max_{u\,:\,|u|\leq 1} \left| \frac{d}{du} p_w(u) \right|.$$

1. Is the penalty function b convex? Is it a norm?

2. Explain how to compute a subgradient of b at a point w.

3. Use your result to code a subgradient method for solving problem (12.20).

Exercise 12.7 (Methods for LASSO) Consider the LASSO problem, discussed in Section 9.6.2:

$$\min_x \frac{1}{2}\|Ax - y\|_2^2 + \lambda\|x\|_1,$$

Compare the following algorithms. Try to write your code in a way that minimizes computational requirements; you may find the result in Exercise 9.4 useful.

1. A coordinate-descent method.

2. A subgradient method, as in Section 12.4.1.

3. A fast first-order algorithm, as in Section 12.3.4.

Exercise 12.8 (Non-negative terms that sum to one) Let x_i, $i = 1, \ldots, n$, be given real numbers, which we assume without loss of generality to be ordered as $x_1 \leq x_2 \leq \cdots \leq x_n$, and consider the scalar equation in variable ν that we encountered in Section 12.3.3.3:

$$f(\nu) = 1, \quad \text{where } f(\nu) \doteq \sum_{i=1}^{n} \max(x_i - \nu, 0).$$

1. Show that f is continuous and strictly decreasing for $\nu \leq x_n$.

2. Show that a solution ν^* to this equation exists, it is unique, and it must belong to the interval $[x_1 - 1/n,\, x_n]$.

3. This scalar equation could be easily solved for ν using, e.g., the bisection method. Describe a simpler, "closed-form" method for finding the optimal ν.

Exercise 12.9 (Eliminating linear equality constraints) We consider a problem with linear equality constraints

$$\min_x f_0(x) \;:\; Ax = b,$$

where $A \in \mathbb{R}^{m,n}$, with A full row rank: $\operatorname{rank} A = m \leq n$, and where we assume that the objective function f_0 is decomposable, that is

$$f_0(x) = \sum_{i=1}^{n} h_i(x_i),$$

with each h_i a convex, twice differentiable function. This problem can be addressed via different approaches, as detailed in Section 12.2.6.

1. Use the constraint elimination approach of Section 12.2.6.1, and consider the function $\tilde{f}_0$ defined in Eq. (12.33). Express the Hessian of $\tilde{f}_0$ in terms of that of f_0.

2. Compare the computational effort[27] required to solve the problem using the Newton method via the constraint elimination technique, versus using the feasible update Newton method of Section 12.2.6.3, assuming that $m \ll n$.

[27] See the related Exercise 7.4.

13. Learning from Data

Exercise 13.1 (SVD for text analysis) Assume you are given a data set in the form of an $n \times m$ term-by-document matrix X corresponding to a large collection of news articles. Precisely, the (i, j) entry in X is the frequency of the word i in the document j. We would like to visualize this data set on a two-dimensional plot. Explain how you would do the following (describe your steps carefully in terms of the SVD of an appropriately centered version of X).

1. Plot the different news sources as points in word space, with maximal variance of the points.
2. Plot the different words as points in news-source space, with maximal variance of the points.

Exercise 13.2 (Learning a factor model) We are given a data matrix $X = [x^{(1)}, \ldots, x^{(m)}]$, with $x^{(i)} \in \mathbb{R}^n$, $i = 1, \ldots, m$. We assume that the data is centered: $x^{(1)} + \cdots + x^{(m)} = 0$. An (empirical) estimate of the covariance matrix is[28]

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} x^{(i)} x^{(i)\top}.$$

In practice, one often finds that the above estimate of the covariance matrix is noisy. One way to remove noise is to approximate the covariance matrix as $\Sigma \approx \lambda I + FF^\top$, where F is an $n \times k$ matrix, containing the so-called "factor loadings," with $k \ll n$ the number of factors, and $\lambda \geq 0$ is the "idiosyncratic noise" variance. The stochastic model that corresponds to this setup is

$$x = Ff + \sigma e,$$

where x is the (random) vector of centered observations, (f, e) is a random variable with zero mean and unit covariance matrix, and $\sigma = \sqrt{\lambda}$ is the standard deviation of the idiosyncratic noise component σe. The interpretation of the stochastic model is that the observations are a combination of a small number k of factors, plus a noise part that affects each dimension independently.

To fit F, λ to the data, we seek to solve

$$\min_{F, \lambda \geq 0} \|\Sigma - \lambda I - FF^\top\|_F. \qquad (13.21)$$

1. Assume λ is known and less than λ_k (the k-th largest eigenvalue of the empirical covariance matrix Σ). Express an optimal F as a function of λ, which we denote by $F(\lambda)$. In other words: you are asked to solve for F, with fixed λ.

[28] See Example 4.2.

2. Show that the error $E(\lambda) = \|\Sigma - \lambda I - F(\lambda)F(\lambda)^\top\|_\text{F}$, with $F(\lambda)$ the matrix you found in the previous part, can be written as
$$E(\lambda)^2 = \sum_{i=k+1}^{p} (\lambda_i - \lambda)^2.$$
Find a closed-form expression for the optimal λ that minimizes the error, and summarize your solution to the estimation problem (13.21).

3. Assume that we wish to estimate the risk (as measured by variance) involved in a specific direction in data space. Recall from Example 4.2 that, given a unit-norm n-vector w, the variance along the direction w is $w^\top \Sigma w$. Show that the rank-k approximation to Σ results in an under-estimate of the directional risk, as compared with using Σ. How about the approximation based on the factor model above? Discuss.

Exercise 13.3 (Movement prediction for a time-series) We have a historical data set containing the values of a time-series $r(1), \ldots, r(T)$. Our goal is to predict if the time-series is going up or down. The basic idea is to use a prediction based on the sign of the output of an auto-regressive model that uses n past data values (here, n is fixed). That is, the prediction at time t of the sign of the value $r(t+1) - r(t)$ is of the form
$$\hat{y}_{w,b}(t) = \text{sgn}\left(w_1 r(t) + \cdots + w_n r(t-n+1) + b\right),$$
In the above, $w \in \mathbb{R}^n$ is our classifier coefficient, b is a bias term, and $n \ll T$ determines how far back into the past we use the data to make the prediction.

1. As a first attempt, we would like to solve the problem
$$\min_{w,b} \sum_{t=n}^{T-1} (\hat{y}_{w,b}(t) - y(t))^2,$$
where $y(t) = \text{sgn}(r(t+1) - r(t))$. In other words, we are trying to match, in a least-squares sense, the prediction made by the classifier on the training set, with the observed truth. Can we solve the above with convex optimization? If not, why?

2. Explain how you would set up the problem and train a classifier using convex optimization. Make sure to define precisely the learning procedure, the variables in the resulting optimization problem, and how you would find the optimal variables to make a prediction.

Exercise 13.4 (A variant of PCA) Return to the variant of PCA examined in Exercise 11.2. Using a (possibly synthetic) data set of your choice, compare the classical PCA and the variant examined here, especially in terms of its sensitivity to outliers. Make sure to establish an evaluation protocol that is as rigorous as possible. Discuss your results.

Exercise 13.5 (Squared vs. non-squared penalties) We consider the problems

$$\begin{array}{rcl} P(\lambda) \ : \ p(\lambda) & \doteq & \min_x f(x) + \lambda \|x\|, \\ Q(\mu) \ : \ q(\mu) & \doteq & \min_x f(x) + \frac{1}{2}\mu\|x\|^2, \end{array}$$

where f is a convex function, $\|\cdot\|$ is an arbitrary vector norm, and $\lambda > 0$, $\mu > 0$ are parameters. Assume that for every choice of these parameters, the corresponding problems have a unique solution.

In general, the solutions for the above problems for fixed λ and μ do not coincide. This exercise shows that we can scan the solutions to the first problem, and get the set of solutions to the second, and vice versa.

1. Show that both p, q are concave functions, and $\tilde{q}$ with values $\tilde{q}(\mu) = q(1/\mu)$ is convex, on the domain $\mathbb{R}_+$.
2. Show that
$$p(\lambda) = \min_{\mu > 0} q(\mu) + \frac{\lambda^2}{2\mu}, \ \ q(\mu) = \max_{\lambda > 0} p(\lambda) - \frac{\lambda^2}{2\mu}.$$
For the second expression, you may assume that $\mathbf{dom} f$ has a nonempty interior.
3. Deduce from the first part that the paths of solutions coincide. That is, if we solve the first problem for every $\lambda > 0$, for any $\mu > 0$ the optimal point we thus find will be optimal for the second problem; and vice versa. It will convenient to denote by $x^*(\lambda)$ (resp. $z^*(\mu)$) the (unique) solution to $P(\lambda)$ (resp. $Q(\mu)$).
4. State and prove a similar result concerning a third function
$$r(\kappa) \ : \ r(\kappa) \doteq \min_x f(x) \ : \ \|x\| \leq \kappa.$$
5. What can you say if we remove the uniqueness assumption?

Exercise 13.6 (Cardinality-penalized least squares) We consider the problem

$$\phi(k) \doteq \min_w \|X^\top w - y\|_2^2 + \rho^2 \|w\|_2^2 + \lambda \, \mathbf{card}(w),$$

where $X \in \mathbb{R}^{n,m}$, $y \in \mathbb{R}^m$, $\rho > 0$ is a regularization parameter, and $\lambda \geq 0$ allows us to control the cardinality (number of nonzeros) in the solution. This in turn allows better interpretability of the results. The above problem is hard to solve in general. In this exercise, we denote by $a_i^\top$, $i = 1, \ldots, n$ the i-th row of X, which corresponds to a particular "feature" (that is, dimension of the variable w).

1. First assume that no cardinality penalty is present, that is, $\lambda = 0$. Show that
$$\phi(0) = y^\top \left(I + \frac{1}{\rho^2} \sum_{i=1}^n a_i a_i^\top \right)^{-1} y.$$

2. Now consider the case $\lambda > 0$. Show that
$$\phi(\lambda) = \min_{u \in \{0,1\}^n} y^\top \left(I_m + \frac{1}{\rho^2} \sum_{i=1}^n u_i a_i a_i^\top \right)^{-1} y + \lambda \sum_{i=1}^n u_i.$$

3. A natural relaxation to the problem obtains upon replacing the constraints $u \in \{0,1\}^n$ with interval ones: $u \in [0,1]^n$. Show that the resulting lower bound $\phi(\lambda) \geq \underline{\phi}(\lambda)$ is the optimal value of the convex problem
$$\underline{\phi}(\lambda) \quad = \quad \max_v \; 2y^\top v - v^T v - \sum_{i=1}^n \left(\frac{(a_i^\top v)^2}{\rho^2} - \lambda \right)_+.$$
How would you recover a suboptimal sparsity pattern from a solution v^* to the above problem?

4. Express the above problem as an SOCP.

5. Form a dual to the SOCP, and show that it can be reduced to the expression
$$\underline{\phi}(\lambda) = \|X^\top w - y\|_2^2 + 2\lambda \sum_{i=1}^n B\left(\frac{\rho x_i}{\sqrt{\lambda}} \right),$$
where B is the (convex) *reverse Hüber function:* for $\xi \in \mathbb{R}$,
$$B(\xi) \doteq \frac{1}{2} \min_{0 \leq z \leq 1} \left(z + \frac{\xi^2}{z} \right) = \begin{cases} |\xi| & \text{if } |\xi| \leq 1, \\ \dfrac{\xi^2 + 1}{2} & \text{otherwise.} \end{cases}$$
Again, how would you recover a suboptimal sparsity pattern from a solution w^* to the above problem?

6. A classical way to handle cardinality penalties is to replace them with the ℓ_1-norm. How does the above approach compare with the ℓ_1-norm relaxation one? Discuss.

14. Computational Finance

Exercise 14.1 (Diversification) You have \$12,000 to invest at the beginning of the year, and three different funds from which to choose. The municipal bond fund has a 7% yearly return, the local bank's Certificates of Deposit (CDs) have an 8% return, and a high-risk account has an expected (hoped-for) 12% return. To minimize risk, you decide not to invest any more than \$2,000 in the high-risk account. For tax reasons, you need to invest at least three times as much in the municipal bonds as in the bank CDs. Denote by x, y, z be the amounts (in thousands) invested in bonds, CDs, and high-risk account, respectively. Assuming the year-end yields are as expected, what are the optimal investment amounts for each fund?

I took this out, too simplistic

Exercise 14.2 (Portfolio optimization problems) We consider a single-period optimization problem involving n assets, and a decision vector $x \in \mathbb{R}^n$ which contains our position in each asset. Determine which of the following objectives or constraints can be modeled using convex optimization.

1. The level of risk (measured by portfolio variance) is equal to a given target t (the covariance matrix is assumed to be known).

2. The level of risk (measured by portfolio variance) is below a given target t.

3. The Sharpe ratio (defined as the ratio of portfolio return to portfolio standard deviation) is above a target $t \geq 0$. Here both the expected return vector and the covariance matrix are assumed to be known.

4. Assuming that the return vector follows a known Gaussian distribution, ensure that the probability of the portfolio return being less than a target t is less than 3%.

5. Assume that the return vector $r \in \mathbb{R}^n$ can take three values $r^{(i)}$, $i = 1, 2, 3$. Enforce the following constraint: the smallest portfolio return under the three scenarios is above a target level t.

6. Under similar assumptions as in part 5: the average of the smallest two portfolio returns is above a target level t. *Hint:* use new variables $s_i = x^\top r^{(i)}$, $i = 1, 2, 3$, and consider the function $s \rightarrow s_{[2]} + s_{[3]}$, where for $k = 1, 2, 3$, $s_{[k]}$ denotes the k-th largest element in s.

7. The transaction cost (under a linear transaction cost model, and with initial position $x_{\text{init}} = 0$) is below a certain target.

8. The number of transactions from the initial position $x_{\text{init}} = 0$ to the optimal position x is below a certain target.

9. The absolute value of the difference between the expected portfolio return and a target return t is less than a given small number ϵ (here, the expected return vector $\hat{r}$ is assumed to be known).

10. The expected portfolio return is either above a certain value t_{up}, *or* below another value t_{low}.

Exercise 14.3 (Median risk) We consider a single-period portfolio optimization problem with n assets. We use past samples, consisting of single-period return vectors $r_1, \ldots, r_N$, where $r_t \in \mathbb{R}^n$ contains the returns of the assets from period $t-1$ to period t. We denote by $\hat{r} \doteq (1/N)(r_1 + \cdots + r_N)$ the vector of sample averages; it is an estimate of the expected return, based on the past samples.

As a measure of risk, we use the following quantity. Denote by $\rho_t(x)$ the return at time t (if we had held the position x at that time). Our risk measure is

$$\mathcal{R}_1(x) \doteq \frac{1}{N} \sum_{t=1}^{N} |\rho_t(x) - \hat{\rho}(x)|,$$

where $\hat{\rho}(x)$ is the portfolio's sample average return.

1. Show that $\mathcal{R}_1(x) = \|R^\top x\|_1$, with R an $n \times N$ matrix that you will determine. Is the risk measure $\mathcal{R}_1$ convex?

2. Show how to minimize the risk measure $\mathcal{R}_1$, subject to the condition that the sample average of the portfolio return is greater than a target μ, using linear programming. Make sure to put the problem in standard form, and define precisely the variables and constraints.

3. Comment on the qualitative difference between the resulting portfolio and one that would use the more classical, variance-based risk measure, given by

$$\mathcal{R}_2(x) \doteq \frac{1}{N} \sum_{t=1}^{N} (\rho_t(x) - \hat{\rho}(x))^2.$$

Exercise 14.4 (Portfolio optimization with factor models – 1)

1. Consider the following portfolio optimization problem:

$$\begin{aligned} p^* = \min_x & \quad x^\top \Sigma x \\ \text{s.t.:} & \quad \hat{r}^\top x \geq \mu, \end{aligned}$$

where $\hat{r} \in \mathbb{R}^n$ is the expected return vector, $\Sigma \in \mathbb{S}^n$, $\Sigma \succeq 0$ is the return covariance matrix, and μ is a target level of expected portfolio return. Assume that the random return vector r follows a simplified factor model of the form

$$r = F(f + \hat{f}), \quad \hat{r} \doteq F\hat{f},$$

where $F \in \mathbb{R}^{n,k}$, $k \ll n$, is a factor loading matrix, $\hat{f} \in \mathbb{R}^k$ is given, and $f \in \mathbb{R}^k$ is such that $\mathbb{E}\{f\} = 0$ and $\mathbb{E}\{ff^\top\} = I$. The above optimization problem is a convex quadratic problem that involves n decision variables. Explain how to cast this problem into an equivalent form that involves only k decision variables. Interpret the reduced problem geometrically. Find a closed-form solution to the problem.

2. Consider the following variation on the previous problem:

$$\begin{aligned} p^* = \min_x & \quad x^\top \Sigma x - \gamma \hat{r}^\top x \\ \text{s.t.:} & \quad x \geq 0, \end{aligned}$$

where $\gamma > 0$ is a tradeoff parameter that weights the relevance in the objective of the risk term and of the return term. Due to the presence of the constraint $x \geq 0$, this problem does not admit, in general, a closed-form solution.

Assume that r is specified according to a factor model of the form

$$r = F(f + \hat{f}) + e,$$

where F, f, and $\hat{f}$ are as in the previous point, and e is an idiosyncratic noise term, which is uncorrelated with f (i.e., $\mathbb{E}\{fe^\top\} = 0$) and such that $\mathbb{E}\{e\} = 0$ and $\mathbb{E}\{ee^\top\} = D^2 \doteq \{d_1^2, \ldots, d_n^2\} \succ 0$. Suppose we wish to solve the problem using a logarithmic barrier method of the type discussed in Section 12.3.1. Explain how to exploit the factor structure of the returns to improve the numerical performance of the algorithm. *Hint:* with the addition of suitable slack variables, the Hessian of the objective (plus barrier) can be made diagonal.

Exercise 14.5 (Portfolio optimization with factor models – 2) Consider again the problem and setup of in point 2 of Exercise 14.4. Let

$z \doteq F^\top x$, and verify that the probem can be rewritten as

$$\begin{array}{rll} p^* = & \min\limits_{x \geq 0,\, z} & x^\top D^2 x + z^\top z - \gamma \hat{r}^\top x \\ & \text{s.t.:} & F^\top x = z. \end{array}$$

Consider the Lagrangian

$$\mathcal{L}(x, z, \lambda) = x^\top D^2 x + z^\top z - \gamma \hat{r}^\top x + \lambda^\top (z - F^\top x)$$

and the dual function

$$g(\lambda) \doteq \min_{x \geq 0, z} \mathcal{L}(x, z, \lambda).$$

Strong duality holds, since the primal problem is convex and strictly feasible, thus $p^* = d^* = \max_\lambda g(\lambda)$.

1. Find a closed-form expression for the dual function $g(\lambda)$.

2. Express the primal optimal solution x^* in terms of the dual optimal variable λ^*.

3. Determine a subgradient of $-g(\lambda)$.

Exercise 14.6 (Kelly's betting strategy) A gambler has a starting capital W_0 and repeatedly bets his whole available capital on a game where with probability $p \in [0, 1]$ he wins the stake, and with probability $1 - p$ he loses it. His wealth W_k after k bets is a random variable:

$$W_k = \begin{cases} 2^k W_0 & \text{with probability } p^k, \\ 0 & \text{with probability } 1 - p^k. \end{cases}$$

1. Determine the expected wealth of the gambler after k bets. Determine the probability with which the gambler eventually runs broke at some k.

2. The results of the previous point should have convinced you that the described one is a ruinous gambling strategy. Suppose now that the gambler gets more cautious, and decides to bet, at each step, only a fraction x of his capital. Denoting by w and ℓ the (random) number of times where the gambler wins and loses a bet, respectively, we have that his wealth at time k is given by

$$W_k = (1 + x)^w (1 - x)^\ell W_0,$$

where $x \in [0, 1]$ is the betting fraction, and $w + \ell = k$. Define the exponential rate of growth of the gambler capital as

$$G = \lim_{k \to \infty} \frac{1}{k} \log_2 \frac{W_k}{W_0}.$$

(a) Determine an expression for the exponential rate of growth G as a function of x. Is this function concave?

(b) Find the value of $x \in [0, 1]$ that maximizes the exponential rate of growth G. Betting according to this optimal fraction is known as the optimal Kelly's gambling strategy.[29]

3. Consider a more general situation, in which an investor can invest a fraction of his capital on an investment opportunity that may have different payoffs, with different probabilities. Specifically, if $W_0 x$ dollars are invested, then the wealth after the outcome of the investment is $W = (1 + rx)W_0$, where r denotes the return of the investment, which is assumed to be a discrete random variable taking values $r_1, \ldots, r_m$ with respective probabilities $p_1, \ldots, p_m$ ($p_i \geq 0$, $r_i \geq -1$, for $i = 1, \ldots, m$, and $\sum_i p_i = 1$).

   The exponential rate of growth G introduced in point 2 of this exercise is nothing but the expected value of the log-gain of the investment, that is

$$G = \mathbb{E}\{\log(W/W_0)\} = \mathbb{E}\{\log(1 + rx)\}.$$

   The particular case considered in point 2 corresponds to taking $m = 2$ (two possible investment outcomes), with $r_1 = 1$, $r_2 = -1$, $p_1 = p$, $p_2 = 1 - p$.

   (a) Find an explicit expression for G as a function of $x \in [0, 1]$.

   (b) Devise a simple computational scheme for finding the optimal investment fraction x that maximizes G.

Exercise 14.7 (Multi-period investments) We consider a multi-stage, single-asset investment decision problem over n periods. For any given time period $i = 1, \ldots, n$, we denote by y_i the predicted return, σ_i the associated variance, and u_i the dollar position invested. Assuming our initial position is $u_0 = w$, the investment problem is

$$\phi(w) \doteq \max_u \sum_{i=1}^{n+1} \left(y_i u_i - \lambda \sigma_i^2 u_i^2 - c|u_i - u_{i-1}| \right) \ : \ u_0 = w, \ \ u_{n+1} = 0,$$

where the first term represents profit, the second, risk, and the third, approximate transaction costs. Here, $c > 0$ is the unit transaction cost and $\lambda > 0$ a risk-return trade-off parameter. (We assume $\lambda = 1$ without loss of generality.)

1. Find a dual for this problem.

2. Show that ϕ is concave, and find a subgradient of $-\phi$ at w. If ϕ is differentiable at w, what is its gradient at w?

[29] After J. L. Kelly, who introduced it in 1956.

3. What is the sensitivity issue of ϕ with respect to the initial position w? Precisely, provide a tight upper bound on $|\phi(w+\epsilon) - \phi(w)|$ for arbitrary $\epsilon > 0$, and with y, σ, c fixed. You may assume ϕ is differentiable for any $u \in [w, w+\epsilon]$.

Exercise 14.8 (Personal finance problem) Consider the following personal finance problem. You are to be paid for a consulting job, for a total of $C = \$30,000$, over the next six months. You plan to use this payment to cover some past credit card debt, which amounts to $D = \$7000$. The credit card's APR (annual interest rate) is $r_1 = 15.95\%$. You have the following items to consider:

- At the beginning of each month, you can transfer any portion of the credit card debt to another card with a lower APR of $r_2 = 2.9\%$. This transaction costs $r_3 = 0.2\%$ of the total amount transferred. You cannot borrow any more from either credit cards; only transfer of debt from card 1 to 2 is allowed.

- The employer allows you to choose the schedule of payments: you can distribute the payments over a maximum of six months. For liquidity reasons, the employer limits any month's pay to $4/3 \times (C/6)$.

- You are paid a base salary of $B = \$70,000$ per annum. You cannot use the base salary to pay off the credit card debt; however it affects how much tax you pay (see next).

- The first three months are the last three months of the current fiscal year and the last three months are the first three months of the next fiscal year. So if you choose to be paid a lot in the current fiscal year (first three months of consulting), the tax costs are high; they are lower if you choose to distribute the payments over several periods. The precise tax due depends on your gross annual total income G, which is your base salary, plus any extra income. The marginal tax rate schedule is given in Table 14.5.

- The risk-free rate (interest rate from savings) is zero.

- Time line of events: all events occur at the beginning of each month, i.e. at the beginning of each month, you are paid the chosen amount, and immediately you decide how much of each credit card to pay off, and transfer any debt from card 1 to card 2. Any outstanding debt accumulates interest at the end of the current month.

- Your objective is to maximize the total wealth at the end of the two fiscal years whilst paying off all credit card debt.

Table 14.5: Marginal tax rate schedule.

Total gross income G	Marginal tax rate	Total tax
$\$0 \leq G \leq \$80,000$	10%	$10\% \times G$
$\$80,000 \leq G$	28%	$28\% \times G$ plus $\$8000 = 10\% \times \$80,000$

1. Formulate the decision-making problem as an optimization problem. Make sure to define the variables and constraints precisely. To describe the tax, use the following constraint:

$$T_i = 0.1 \min(G_i, \alpha) + 0.28 \max(G_i - \alpha, 0), \tag{14.22}$$

   where T_i is the total tax paid, G_i is the total gross income in years $i = 1, 2$ and $\alpha = 80,000$ is the tax threshold parameter.

2. Is the problem a linear program? Explain.

3. Under what conditions on α and G_i can the tax constraint (14.22) be replaced by the following set of constraints? Is it the case for our problem? Can you replace (14.22) by (14.23) in your problem? Explain.

$$\begin{aligned} T_i &= 0.1 d_{1,i} + 0.28 d_{2,i}, \qquad (14.23)\\ d_{2,i} &\geq G_i - \alpha,\\ d_{2,i} &\geq 0,\\ d_{1,i} &\geq G_i - d_{2,i},\\ d_{1,i} &\geq d_{2,i} - \alpha. \end{aligned}$$

4. Is the new problem formulation, with (14.23), convex? Justify your answer.

5. Solve the problem using your favorite solver. Write down the optimal schedules for receiving payments and paying off/transferring credit card debt, and the optimal total wealth at the end of two years. What is your total wealth W?

6. Compute an optimal W for $\alpha \in [70k, 90k]$ and plot α vs. W in this range. Can you explain the plot?

Exercise 14.9 (Transaction costs and market impact) We consider the following portfolio optimization problem:

$$\max_x \hat{r}^\top x - \lambda x^\top C x - c \cdot T(x - x^0) \ : \ x \geq 0, \quad x \in \mathcal{X}, \tag{14.24}$$

where C is the empirical covariance matrix, $\lambda > 0$ is a risk parameter, and $\hat{r}$ is the time-average return for each asset for the given period. Here, the constraint set $\mathcal{X}$ is determined by the following conditions.

- No shorting is allowed.
- There is a budget constraint $x_1 + \cdots + x_n = 1$.

In the above, the function T represents transaction costs and market impact, $c \geq 0$ is a parameter that controls the size of these costs, while $x^0 \in \mathbb{R}^n$ is the vector of initial positions. The function T has the form

$$T(x) = \sum_{i=1}^{n} B_M(x),$$

where the function B_M is piece-wise linear for small x, and quadratic for large x; that way we seek to capture the fact that transaction costs are dominant for smaller trades, while market impact kicks in for larger ones. Precisely, we define B_M to be the so-called "reverse Hüber" function with cut-off parameter M: for a scalar z, the function value is

$$B_M(z) \doteq \begin{cases} |z| & \text{if } |z| \leq M, \\ \dfrac{z^2 + M^2}{2M} & \text{otherwise.} \end{cases}$$

The scalar $M > 0$ describes where the transition from a linearly shaped to a quadratically shaped penalty takes place.

1. Show that B_M can be expressed as the solution to an optimization problem:

   $$B_M(z) = \min_{v,w} \; v + w + \frac{w^2}{2M} \;:\; |z| \leq v + w, \;\; v \leq M, \;\; w \geq 0.$$

   Explain why the above representation proves that B_M is convex.

2. Show that, for given $x \in \mathbb{R}^n$:

   $$T(x) = \min_{w,v} \; \mathbf{1}^\top (v + w) + \frac{1}{2M} w^\top w \;:\; \begin{array}{l} v \leq M\mathbf{1}, \;\; w \geq 0, \\ |x - x^0| \leq v + w, \end{array}$$

   where, in the above, v, w are now n-dimensional vector variables, $\mathbf{1}$ is the vector of ones, and the inequalities are component-wise.

3. Formulate the optimization problem (14.24) in convex format. Does the problem fall into one of the categories (LP, QP, SOCP, etc.) seen in Chapter 8?

4. Draw the efficient frontier of the portfolio corresponding to $M = 0.01, 0.05, 0.1, 1, 5$, with $c = 5 \times 10^{-4}$. Comment on the qualitative differences between the optimal portfolio for two different values of $M = 0.01, 1$.

Exercise 14.10 (Optimal portfolio execution) This exercise deals with an optimal portfolio execution problem, where we seek to optimally liquidating a portfolio given as a list of n asset names and initial number of shares in each asset. The problem is stated over a given time horizon T, and shares are to be traded at fixed times $t = 1, \dots, T$. In practice, the dimension of the problem may range from $n = 20$ to $n = 6000$.

The initial list of shares is given by a vector $x_0 \in \mathbb{R}^n$, and the final target is to liquidate our portfolio. The initial position is given by a price vector $p \in \mathbb{R}^n$, and a vector s that gives the side of each asset (1 to indicate long, -1 to indicate short). We denote by $w = p \circ s$ the so-called price weight vector, where $\circ$ denotes the component-wise product[30].

Our decision variable is the *execution schedule*, a $n \times T$ matrix X, with X_{it} the amount of shares (in hundreds, say) of asset i to be sold at time t. We will not account for discretization effects and treat X as a real-valued matrix. For $t = 1, \dots, T$, we denote by $x_t \in \mathbb{R}^n$ the t-th column of X; x_t encapsulates to all the trading that takes place at period t.

In our problem, X is constrained via upper and lower bounds: we express this as $X^l \leq X \leq X^u$, where inequalities are understood component-wise, and X^l, X^u are given $n \times T$ matrices (for example, a no short selling condition is enforced with $X^l = 0$). These upper and lower bounds can be used to make sure we attain our target at time $t = T$: we simply assume that the last columns of X^l, X^u are both equal to the target vector, which is zero in the case we seek to fully liquidate the portfolio.

We may have additional linear equality or inequality constraints. For example we may enforce upper and lower bounds on the trading:

$$0 \leq y_t \doteq x_{t-1} - x_t \leq y_t^u, \quad t = 1, \dots, T,$$

where $Y = [y_1, \dots, y_T] \in \mathbb{R}^{n,T}$ will be referred to as the *trading* matrix, and $Y^u = [y_1^u, \dots, y_T^u]$ is a given (non-negative) $n \times T$ matrix that bounds the elements of Y from above. The lower bound ensures that trading decreases over time; the second constraint can be used to enforce a maximum participation rate, as specified by the user.

We will denote by $\mathcal{X} \subseteq \mathbb{R}^{n,T}$ our feasible set, that is, the set of $n \times T$ matrices $X = [x_1, \dots, x_T]$ that satisfy the constraints above, including the upper and lower bounds on X.

We also want to enforce a *dollar neutral strategy* at each time step. This requires to have the same dollar position both in long and short. This can be expressed with the conditions $w^\top x_t = 0$, $t = 1, \dots, T$, where $w = p \circ s \in \mathbb{R}^m$ contains the price weight of each asset. We

[30] For two n-vectors u, v, the notation $u \circ v$ denotes the vector with components $u_i v_i$, $i = 1, \dots, n$.

can write the dollar-neutral constraint compactly as $X^\top w = 0$.

Our objective function involves three terms, referred to as *impact, risk,* and *alpha* respectively. The *impact function* is modeled as

$$I(X) = \sum_{t=1}^{T} \sum_{i=1}^{n} V_{ti}(X_{ti} - X_{t-1,i})^2,$$

where $V = [v_1, \dots, v_T]$ is a $n \times T$ matrix of non-negative numbers that model the impact of transactions (the matrix V has to be estimated with historical data, but we consider it to be fully known here). In the above, the n-vector of initial conditions $x_0 = (X_{0,i})_{1 \le i \le n}$ is given.

The *risk function* has the form

$$R(X) = \sum_{t=1}^{T} (w \circ x_t)^\top \Sigma (w \circ x_t),$$

where $\circ$ is the component-wise product, $w = p \circ s$ is the price weight vector, and Σ is a positive semidefinite matrix the describes the daily market risk. In this problem, we assume that Σ has a "diagonal-plus-low-rank" structure, corresponding to a factor model. Specifically, $\Sigma = D^2 + FF^\top$, where D is a $n \times n$, diagonal positive definite matrix, and F is a $n \times k$ "factor loading" matrix, with $k \approx 10 - 100$ the number of factors in the model (typically, $k \ll n$). We can write the risk function as

$$R(X) = \sum_{t=1}^{T} x_t^\top (D_w^2 + F_w F_w^\top) x_t,$$

where $D_w \doteq \operatorname{diag}(w) D$ is diagonal, positive definite, and $F_w \doteq \operatorname{diag}(w) F$.

Finally, the alpha function accounts for views on the asset return themselves, and is a linear function of X, which we write as

$$C(X) = \sum_{t=1}^{T} c_t^\top x_t,$$

where $C = [c_1, \dots, c_T] \in \mathbb{R}^{n,T}$ is a given matrix that depends on $\alpha \in \mathbb{R}^n$, which contains our return predictions for the day. Precisely, $c_t = \alpha_t \circ p$, where $p \in \mathbb{R}^n$ is the price vector, and α_t is a vector of predicted returns.

1. Summarize the problem data, and their sizes.

2. Write the portfolio execution problem as a QP. Make sure to define precisely the variables, objective and constraints.

3. Explain how to take advantage of the factor model to speed up computation. *Hint:* look at Exercise 12.9.

15. Control Problems

Exercise 15.1 (Stability and eigenvalues) Prove that the continuous-time LTI system (15.20) is asymptotically stable (or stable, for short) if and only if all the eigenvalues of the A matrix, $\lambda_i(A)$, $i = 1, \ldots, n$, have (strictly) negative real parts.

Prove that the discrete-time LTI system (15.28) is stable if and only if all the eigenvalues of the A matrix, $\lambda_i(A)$, $i = 1, \ldots, n$, have moduli (strictly) smaller than one.

Hint: use the expression $x(t) = e^{At}x_0$ for the free response of the continuous-time system, and the expression $x(k) = A^k x_0$ for the free response of the discrete-time system. You may derive your proof under the assumption that A is diagonalizable.

Exercise 15.2 (Signal norms) A continuous-time *signal* $w(t)$ is a function mapping time $t \in \mathbb{R}$ to values $w(t)$ in either $\mathbb{C}^m$ or $\mathbb{R}^m$. The *energy* content of a signal $w(t)$ is defined as

$$E(w) \doteq \|w\|_2^2 = \int_{-\infty}^{\infty} \|w(t)\|_2^2 \mathrm{d}t,$$

where $\|w\|_2$ is the 2-norm of the signal. The class of finite-energy signal contains signals for which the above 2-norm is finite.

Periodic signals typically have infinite energy. For a signal with period T, we define its *power* content as

$$P(w) \doteq \frac{1}{T} \int_{t_0}^{t_0+T} \|w(t)\|_2^2 \mathrm{d}t.$$

1. Evaluate the energy of the harmonic signal $w(t) = ve^{\jmath\omega t}$, $v \in \mathbb{R}^m$, and of the causal exponential signal $w(t) = ve^{at}$, for $a < 0$, $t \geq 0$ ($w(t) = 0$ for $t < 0$).

2. Evaluate the power of the harmonic signal $w(t) = ve^{\jmath\omega t}$ and of the sinusoidal signal $w(t) = v\sin(\omega t)$.

Exercise 15.3 (Energy upper bound on the system's state evolution) Consider a continuous-time LTI system $\dot{x}(t) = Ax(t)$, $t \geq 0$, with no input (such a system is said to be *autonomous*), and output $y(t) = Cx$. We wish to evaluate the energy contained in the system's output, as measured by the index

$$J(x_0) \doteq \int_0^{\infty} y(t)^\top y(t) \mathrm{d}t = \int_0^{\infty} x(t)^\top Q x(t) \mathrm{d}t,$$

where $Q \doteq C^\top C \succeq 0$.

1. Show that if the system is stable, then $J(x_0) < \infty$, for any given x_0.

2. Show that if the system is stable and there exists a matrix $P \succeq 0$ such that
$$A^\top P + PA + Q \preceq 0,$$
then it holds that $J(x_0) \leq x_0^\top P x_0$. *Hint:* consider the quadratic form $V(x(t)) = x(t)^\top P x(t)$, and evaluate its derivative with respect to time.

3. Explain how to compute a minimal upper bound on the state energy, for the given initial conditions.

Exercise 15.4 (System gain) The *gain* of a system is the maximum energy amplification from the input signal to output. Any input signal $u(t)$ having finite energy is mapped by a stable system to an output signal $y(t)$ which also has finite energy. Parseval's identity relates the energy of a signal $w(t)$ in the time domain to the energy of the same signal in the Fourier domain (see Remark 15.1), that is

$$E(w) \doteq \|w\|_2^2 = \int_{-\infty}^{\infty} \|w(t)\|_2^2 \mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{\infty} \|\hat{W}(\omega)\|_2^2 \mathrm{d}\omega \doteq \|\hat{W}\|_2^2.$$

The *energy gain* of system (15.26) defined as

$$\text{energy gain} \doteq \sup_{u(t):\|u\|_2<\infty, u\neq 0} \frac{\|y\|_2^2}{\|u\|_2^2}.$$

1. Using the above information, prove that, for a stable system,
$$\text{energy gain} \leq \sup_{\omega\geq 0} \|H(\jmath\omega)\|_2^2,$$
where $\|H(\jmath\omega)\|_2$ is the spectral norm of the transfer matrix of system (15.26), evaluated at $s = \jmath\omega$. The (square-root of the) energy gain of the system is also known as the $\mathcal{H}_\infty$-norm, and it is denoted by $\|H\|_\infty$.

   Hint: use Parseval's identity and then suitably bound a certain integral. Notice that equality actually holds in the previous formula, but you are not asked to prove this.

2. Assume that system (15.26) is stable, $x(0) = 0$, and $D = 0$. Prove that if there exists $P \succeq 0$ such that
$$\begin{bmatrix} A^\top P + PA + C^\top C & PB \\ B^\top P & -\gamma^2 I \end{bmatrix} \preceq 0 \qquad (15.25)$$
then it holds that
$$\|H\|_\infty \leq \gamma.$$

Devise a computational scheme that provides you with the lowest possible upper bound γ^* on the energy gain of the system.

Hint: define a quadratic function $V(x) = x^\top Px$, and observe that the derivative in time of V, along the trajectories of system (15.26), is

$$\frac{\mathrm{d}V(x)}{\mathrm{d}t} = x^\top P\dot{x} + \dot{x}^\top Px.$$

Then show that the LMI condition (15.25) is equivalent to the condition that

$$\frac{\mathrm{d}V(x)}{\mathrm{d}t} + \|y\|^2 - \gamma^2\|u\|^2 \leq 0, \quad \forall\, x, u \text{ satisfying (15.26)},$$

and that this implies in turn that $\|H\|_\infty \leq \gamma$.

Exercise 15.5 (Extended superstable matrices) A matrix $A \in \mathbb{R}^{n,n}$ is said to be continuous-time *extended superstable*[31] (which we denote by $A \in E_c$) if there exists $d \in \mathbb{R}^n$ such that

$$\sum_{j \neq i} |a_{ij}| d_j < -a_{ii} d_i,\; d_i > 0, \quad i = 1, \ldots, n.$$

Similarly, a matrix $A \in \mathbb{R}^{n,n}$ is said to be discrete-time extended superstable (which we denote by $A \in E_d$) if there exists $d \in \mathbb{R}^n$ such that

$$\sum_{j=1}^{n} |a_{ij}| d_j < d_i,\; d_i > 0, \quad i = 1, \ldots, n.$$

If $A \in E_c$, then all its eigenvalues have real parts smaller than zero, hence the corresponding continuous-time LTI system $\dot{x} = Ax$ is stable. Similarly, if $A \in E_d$, then all its eigenvalues have moduli smaller than one, hence the corresponding discrete-time LTI system $x(k+1) = Ax(k)$ is stable. Extended superstability thus provides a *sufficient* condition for stability, which has the advantage of being checkable via feasibility of a set of linear inequalities.

1. Given a continuous-time system $\dot{x} = Ax + Bu$, with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, describe your approach for efficiently designing a state-feedback control law of the form $u = -Kx$, such that the controlled system is extended superstable.

2. Given a discrete-time system $x(k+1) = Ax(k) + Bu(k)$, assume that matrix A is affected by interval uncertainty, that is

$$a_{ij} = \hat{a}_{ij} + \delta_{ij}, \quad i, j = 1, \ldots, n,$$

where $\hat{a}_{ij}$ is the given nominal entry, and δ_{ij} is an uncertainty term, which is only known to be bounded in amplitude as $|\delta_{ij}| \leq \rho r_{ij}$, for

[31] See B. T. Polyak, Extended superstability in control theory, *Automation and Remote Control*, 2004.

given $r_{ij} \geq 0$. Define the radius of extended superstability as the largest value ρ^* of $\rho \geq 0$ such that A is extended superstable for all the admissible uncertainties. Describe a computational approach for determining such a ρ^*.

16. Engineering Design

Exercise 16.1 (Network congestion control) A network of $n = 6$ peer-to-peer computers is shown in Figure 16.9. Each computer can upload or download data at a certain rate on the connection links shown in the figure. Let $b^+ \in \mathbb{R}^8$ be the vector containing the packet transmission rates on the links numbered in the figure, and let $b^- \in \mathbb{R}^8$ be the vector containing the packet transmission rates on the reverse links, where it must hold that $b^+ \geq 0$ and $b^- \geq 0$.

Define an arc–node incidence matrix for this network:

$$A \doteq \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix},$$

and let $A_+ \doteq \max(A, 0)$ (the positive part of A), $A_- \doteq \min(A, 0)$ (the negative part of A). Then the total output (upload) rate at the nodes is given by $v_{\text{upl}} = A_+ b^+ - A_- b^-$, and the total input (download) rate at the nodes is given by $v_{\text{dwl}} = A_+ b^- - A_- b^+$. The net outflow at nodes is hence given by

$$v_{\text{net}} = v_{\text{upl}} - v_{\text{dwl}} = Ab^+ - Ab_-,$$

and the flow balance equations require that $[v_{\text{net}}]_i = f_i$, where $f_i = 0$ if computer i is not generating or sinking packets (it just passes on the received packets, i.e., it is acting as a relay station), $f_i > 0$ if computer i is generating packets, or $f_i < 0$ if it is sinking packets at an assigned rate f_i.

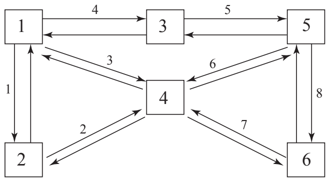

Figure 16.9: A small network.

Each computer can download data at a maximum rate of $\bar{v}_{\text{dwl}} =$ 20 Mbit/s and upload data at a maximum rate of $\bar{v}_{\text{upl}} = 10$ Mbit/s (these limits refer to the total download or upload rates of a computer, through all its connections). The level of congestion of each connection is defined as

$$c_j = \max(0, (b_j^+ + b_j^- - 4)), \quad j = 1, \ldots, 8.$$

Assume that node 1 must transmit packets to node 5 at a rate $f_1 = 9$ Mbit/s, and that node 2 must transmit packets to node 6 at a rate $f_2 = 8$ Mbit/s. Find the rate on all links such that the average congestion level of the network is minimized.

Exercise 16.2 (Design of a water reservoir) We need to design a water reservoir for water and energy storage, as depicted in Figure 16.10.

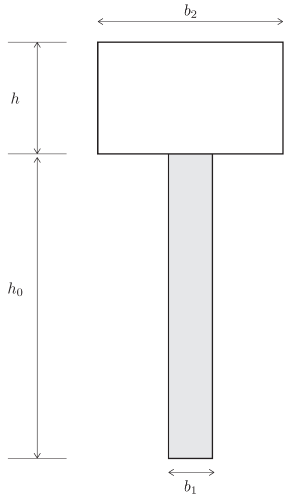

Figure 16.10: A water reservoir on concrete basement.

The concrete basement has a square cross-section of side length b_1 and height h_0, while the reservoir itself has a square cross-section of side length b_2 and height h. Some useful data is reported in Table 16.6.

Table 16.6: Data for reservoir problem.

Quantity	Value	Units	Description
g	9.8	$\mathrm{m/s^2}$	gravity acceleration
E	30×10^9	$\mathrm{N/m^2}$	basement long. elasticity modulus
ρ_w	10×10^3	$\mathrm{N/m^3}$	specific weight of water
ρ_b	25×10^3	$\mathrm{N/m^3}$	specific weight of basement
J	$b_1^4/12$	$\mathrm{m^4}$	basement moment of inertia
N_{cr}	$\pi^2 JE/(2h_0)^2$	N	basement critical load limit

The critical load limit N_{cr} of the basement should withstand at least twice the weight of water. The structural specification $h_0/b_1^2 \leq 35$ should hold. The form factor of the reservoir should be such that $1 \leq b_2/h \leq 2$. The total height of the structure should be no larger than 30 m. The total weight of the structure (basement plus reservoir full of water) should not exceed 9.8×10^5 N. The problem is to find the dimensions b_1, b_2, h_0, h such that the potential energy P_w of the stored water is maximal (assume $P_w = (\rho_w h b_2^2) h_0$). Explain if and how the problem can be modeled as a convex optimization problem and, in the positive case, find the optimal design.

Exercise 16.3 (Wire sizing in circuit design) Interconnects in modern electronic chips can be modeled as conductive surface areas deposed on a substrate. A "wire" can thus be thought as a sequence of rectangular segments, as shown in Figure 16.11.

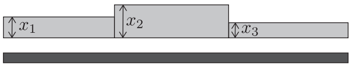

Figure 16.11: A wire is represented as a sequence of rectangular surfaces on a substrate. Lengths ℓ_i are fixed, and the widths x_i of the segments are the decision variables. This example has three wire segments.

We assume that the lengths of these segments are fixed, while the widths x_i need be sized according to the criteria explained next. A common approach is to model the wire as the cascade connection of RC stages, where, for each stage, $S_i = 1/R_i$, C_i are, respectively, the conductance and the capacitance of the i-th segment, see Figure 16.12.

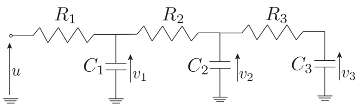

Figure 16.12: RC model of a three-segment wire.

The values of S_i, C_i are proportional to the surface area of the wire segment, hence, since the lengths ℓ_i are assumed known and fixed, they are affine functions of the widths, i.e.,

$$S_i = S_i(x_i) = \sigma_i^{(0)} + \sigma_i x_i, \quad C_i = C_i(x_i) = c_i^{(0)} + c_i x_i,$$

where $\sigma_i^{(0)}, \sigma_i, c_i^{(0)}, c_i$ are given positive constants. For the three-segment wire model illustrated in the figures, one can write the following set of dynamic equations that describe the evolution in time of the node voltages $v_i(t)$, $i = 1, \ldots, 3$:

$$\begin{bmatrix} C_1 & C_2 & C_3 \\ 0 & C_2 & C_3 \\ 0 & 0 & C_3 \end{bmatrix} \dot{v}(t) = -\begin{bmatrix} S_1 & 0 & 0 \\ -S_2 & S_2 & 0 \\ 0 & -S_3 & S_3 \end{bmatrix} v(t) + \begin{bmatrix} S_1 \\ 0 \\ 0 \end{bmatrix} u(t).$$

These equations are actually expressed in a more useful form if we introduce a change of variables

$$v(t) = Qz(t), \quad Q = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix},$$

from which we obtain

$$\mathcal{C}(x)\dot{z}(t) = -\mathcal{S}(x)z(t) + \begin{bmatrix} S_1 \\ 0 \\ 0 \end{bmatrix} u(t),$$

where

$$\mathcal{C}(x) \doteq \begin{bmatrix} C_1 + C_2 + C_3 & C_2 + C_3 & C_3 \\ C_2 + C_3 & C_2 + C_3 & C_3 \\ C_3 & C_3 & C_3 \end{bmatrix}, \quad \mathcal{S}(x) \doteq \mathrm{diag}\,(S_1, S_2, S_3).$$

Clearly, $\mathcal{C}(x)$, $\mathcal{S}(x)$ are symmetric matrices whose entries depend affinely on the decision variable $x = (x_1, x_2, x_3)$. Further, one may observe that $\mathcal{C}(x)$ is nonsingular whenever $x \geq 0$ (as is physically the case in our problem), hence the evolution of $z(t)$ is represented by (we next assume $u(t) = 0$, i.e., we consider only the free-response time evolution of the system)

$$\dot{z}(t) = -\mathcal{C}(x)^{-1}\mathcal{S}(x)z(t).$$

The *dominant time constant* of the circuit is defined as

$$\tau = \frac{1}{\lambda_{\min}(\mathcal{C}(x)^{-1}\mathcal{S}(x))},$$

and it provides a measure of the "speed" of the circuit (the smaller τ, the faster is the response of the circuit).

Describe a computationally efficient method for sizing the wire so as to minimize the total area occupied by the wire, while guaranteeing that the dominant time constant does not exceed an assigned level $\eta > 0$.