Optimization Models Solution Manual

Optimization Models in Engineering (University of California, Berkeley)

GIUSEPPE CALAFIORE AND LAURENT EL GHAOUI

OPTIMIZATION MODELS

SOLUTIONS MANUAL

CAMBRIDGE

Ver. 0.1 – Oct. 2014

DISCLAIMER

This is the first draft of the Solution Manual for exercises in the book "Optimization Models" by Calafiore & El Ghaoui.

This draft is under construction. It is still incomplete and it is very likely to contain errors.

This material is offered "as is," non commercially, for personal use of instructors.

Comments and corrections are very welcome.

Contents

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.

Solution 2.1 The set $\mathcal{S}$ is a subspace, as can be checked directly: if $x, y \in \mathcal{S}$, then for every $\lambda, \mu \in \mathbb{R}$, we have $\lambda x + \mu y \in \mathcal{S}$. To find the dimension, we solve the equation and find that any solution to the equations is of the form $x_1 = -1/2x_2$, $x_3 = -1/3x_2$, where x_2 is free. Hence the dimension of $\mathcal{S}$ is 1, and a basis for S is the vector $(-1/2, 1, -1/3)$.

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.

Solution 2.2

1. We can express any vector $x \in \mathcal{P}$ as $x = (x_1, x_2, 1/3 - x_1/3 - 2x_2/3)$, where x_1, x_2 are arbitrary. Thus

   $$x = x^{(0)} + x_1 x^{(1)} + x_2 x^{(2)},$$

   where

   $$x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ \frac{1}{3} \end{bmatrix}, \ x^{(1)} = \begin{bmatrix} 1 \\ 0 \\ -\frac{1}{3} \end{bmatrix}, \ x^{(2)} = \begin{bmatrix} 0 \\ 1 \\ -\frac{2}{3} \end{bmatrix}.$$

   Since $x^{(1)}$ and $x^{(2)}$ are linearly independent, $\mathcal{P}$ is of dimension 2.

2. The set $\mathcal{P}$ is defined by a single linear equation $a^\top x = b$, with $a^\top = [1 \ 2 \ 3]$ and $b = 1$, i.e., $\mathcal{P}$ is a hyperplane. The minimum Euclidean distance from o to $\mathcal{P}$ is the ℓ_2 norm of the projection of 0 onto $\mathcal{P}$, which can be determined as discussed in Section 2.3.2.2. That is, the projection x^* of 0 onto $\mathcal{P}$ is such that $x^* \in \mathcal{P}$ and x^*

is orthogonal to the subspace generating $\mathcal{P}$ (which coincides with the span of a), that is $x^* = \alpha a$. Hence, it must be that $a^\top x^* = 1$, thus $\alpha \|a\|_2^2 = 1$, and $\alpha = 1/\|a\|_2^2$. We thus have that

$$x^* = \frac{a}{\|a\|_2^2},$$

and the distance we are seeking is $\|x^*\|_2 = 1/\|a\|_2 = 1/\sqrt{14}$.

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:

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

   Are these vectors linearly independent?

Solution 2.3

1. We can observe directly that $u^\top (x - x_0) = 0$, hence the projection of x is the same as that of x_0, which is $z = x_0$ itself.

   Alternatively, as seen in Section (2.3.2.1), the projection is

$$z = x_0 + \frac{u^\top (x - x_0)}{u^\top u} u$$

   which gives $z = x_0$.

   Another method consists in solving

$$\begin{aligned} \min_t \|x_0 + tu - x\|_2^2 &= \min_t t^2 u^T u - 2tu^\top (x - x_0) + \|x - x_0\|_2^2 \\ &= \min_t (u^\top u)(t - t_0)^2 + \text{constant}, \end{aligned}$$

   where $t_0 = (x - x_0)^\top u/(u^T u)$. This leads to the optimal $t^* = t_0$, and provides the same result as before.

2. The angle cosine is given by

$$\cos\theta = \frac{x^\top y}{\|x\|_2 \|y\|_2} = \frac{10}{14},$$

   which gives $\theta \approx 41°$.

   The vectors are linearly independent, since $\lambda x + \mu y = 0$ for $\lambda, \mu \in \mathbb{R}$ implies that $\lambda = \mu = 0$. Another way to prove this is to observe that the angle is not $0°$ nor $180°$.

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 $\mathbb{R}^n$?

Solution 2.4 The axioms of 2.2 are all satisfied for any $\alpha \in \mathbb{R}^n$, except the conditions

$$\begin{aligned} &f(x,x) \geq 0; \\ &f(x,x) = 0 \text{ if and only if } x = 0. \end{aligned}$$

These properties hold if and only if $\alpha_k > 0$, $k = 1, \ldots, n$. Indeed, if the latter is true, then the above two conditions hold. Conversely, if if there exist k such that $\alpha_k \leq 0$, setting $x = e_k$ (the k-th unit vector in $\mathbb{R}^n$) produces $f(e_k, e_k) \leq 0$; this contradicts one of the two above conditions.

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.

Solution 2.5 When x, y are both unit-norm, we have

$$(x-y)^\top (x+y) = x^\top x - y^\top y - y^\top x + x^\top y = x^\top x - y^\top y = 0,$$

as claimed.

We can express any vector $z \in \text{span}(x,y)$ as $z = \lambda x + \mu y$, for some $\lambda, \mu \in \mathbb{R}$. We have $z = \alpha u + \beta v$, where

$$\alpha = \frac{\lambda + \mu}{2}, \quad \beta = \frac{\lambda - \mu}{2}.$$

Hence $z \in \text{span}(u,v)$. The converse is also true for similar reasons. Thus, (u,v) is an orthogonal basis for $\text{span}(x,y)$. We finish by normalizing u, v, replacing them with $(u/\|u\|_2, v/\|v\|_2)$. The desired orthogonal basis is thus given by $((x-y)/\|x-y\|_2, (x+y)/\|x+y\|_2)$.

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 non-zero vector x,

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

where $\mathrm{card}(x)$ is the *cardinality* of the vector x, defined as the number of non-zero elements in x. Find vectors x for which the lower bound is attained.

Solution 2.6

1. We have

$$\|x\|_2^2 = \sum_{i=1}^n x_i^2 \leq n \cdot \max_i x_i^2 = n \cdot \|x\|_\infty^2.$$

Also, $\|x\|_\infty \leq \sqrt{x_1^2 + \ldots + x_n^2} = \|x\|_2$.

The inequality $\|x\|_2 \leq \|x\|_1$ is obtained after squaring both sides, and checking that

$$\sum_{i=1}^n x_i^2 \leq \sum_{i=1}^n x_i^2 + \sum_{i \neq j} |x_i x_j| = \left(\sum_{i=1}^n |x_i| \right)^2 = \|x\|_1^2.$$

Finally, the condition $\|x\|_1 \leq \sqrt{n}\|x\|_2$ is due to the Cauchy-Schwartz inequality

$$|z^\top y| \leq \|y\|_2 \cdot \|z\|_2,$$

applied to the two vectors $y = (1, \ldots, 1)$ and $z = |x| = (|x_1|, \ldots, |x_n|)$.

2. Let us apply the Cauchy-Schwartz inequality with $z = |x|$ again, and with y a vector with $y_i = 1$ if $x_i \neq 0$, and $y_i = 0$ otherwise. We have $\|y\|_2 = \sqrt{k}$, with $k = \mathrm{card}(x)$. Hence

$$|z^\top y| = \|x\|_1 \leq \|y\|_2 \cdot \|z\|_2 = \sqrt{k} \cdot \|x\|_2,$$

which proves the result. The bound is attained for vectors with k non-zero elements, all with the same magnitude.

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 $|u_k v_k| = |u_k||v_k|$.

Solution 2.7 The inequality is trivial if one of the vectors x, y is zero. We henceforth assume that none is, which allows us to define the normalized vectors u, v. We need to show that

$$\sum_k |u_k v_k| \leq 1.$$

Using the hint given, we apply Young's inequality, which states that for any given numbers $a, b \geq 0$ and $p, q > 0$ such that

$$\frac{1}{p} + \frac{1}{q} = 1,$$

it holds that

$$ab \leq \frac{1}{p}a^p + \frac{1}{q}b^q.$$

We thus have, with $a = |u_k|$ and $b = |v_k|$, and summing over k:

$$\begin{aligned} \sum_k |u_k v_k| &\leq \frac{1}{p}\sum_k |u_k|^p + \frac{1}{q}\sum_k |v_k|^q \\ &= \frac{1}{p}\|u\|_p^p + \frac{1}{q}\|v\|_q^q \\ &= \frac{1}{p} + \frac{1}{q} = 1, \end{aligned}$$

where we have used the fact that $\|u\|_p = \|v\|_q = 1$.

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".

Solution 2.8 (Linear functions) TBD

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

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

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

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.

Solution 2.9 (Bound on a polynomial's derivative) We have, with $z = (1, 2, \ldots, k)$, and using Hölder's inequality:

$$\begin{aligned} \left|\frac{\mathrm{d}p_w(x)}{\mathrm{d}x}\right| &= \left|w_2 + 2w_3 x + \ldots + k w_{k+1} x^{k-1}\right| \\ &\leq |w_2| + 2|w_3| + \ldots + k|w_{k+1}| \\ &= |v^\top z| \\ &\leq \|v\|_p \cdot \|z\|_q. \end{aligned}$$

When $p = 1$, we have

$$\|z\|_q = \|z\|_\infty = k.$$

When $p = 2$, we have

$$\|z\|_q = \|z\|_2 = \sqrt{1 + 4 + \ldots + k^2} \leq \sqrt{k \cdot k^2} = k^{3/2}.$$

When $p = \infty$, we have

$$\|z\|_q = \|z\|_1 = 1 + 2 + \ldots + k = \frac{k(k+1)}{2}.$$

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

$$\begin{aligned} \nabla_x h(x) &= A^\top \nabla_g f(g(x)) \\ \nabla_x^2 h(x) &= A^\top \nabla_g^2 f(g(x)) A. \end{aligned}$$

Solution 3.1

1. We have, by the composition rule for derivatives:

$$\begin{aligned} [J_h(x)]_{i,j} &= \frac{\partial h_i(x)}{\partial x_j} = \sum_{l=1}^{m} \frac{\partial f_i}{\partial g_l}(x) \frac{\partial g_l}{\partial x_j}(x) \\ &= \sum_{l=1}^{m} [J_f(g(x))]_{i,l} [J_g(x)]_{l,j}, \end{aligned}$$

which proves the result.

2. Since $g_i(x) = \sum_{k=1}^n a_{ik} x_k + b_i$, $i = 1, \ldots, m$, we have that the (i, j)-th element of the Jacobian of g is

$$[J_g(x)]_{ij} = \frac{\partial g_i(x)}{\partial x_j} = a_{ij},$$

hence $J_g(x) = A$, and the desired result follows from applying point 1. of this exercise.

3. For a scalar-valued function, the gradient coincides with the transpose of the Jacobian, hence the expression for the gradient of h w.r.t. x follows by applying the previous point. For the Hessian,

we have instead

$$\begin{aligned}
[\nabla_x^2 h(x)]_{ij} &= \frac{\partial^2 h(x)}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_j}\frac{\partial h(x)}{\partial x_i} = \frac{\partial}{\partial x_j} a_i^\top \nabla_g f(g(x)) \\
&= \frac{\partial}{\partial x_j}\sum_{k=1}^m a_{ik}\frac{\partial f(g(x))}{\partial g_k} = \sum_{k=1}^m a_{ik}\frac{\partial}{\partial x_j}\frac{\partial f(g(x))}{\partial g_k} \\
&= \sum_{k=1}^m a_{ik}\sum_{p=1}^m \frac{\partial}{\partial g_p}\left(\frac{\partial f(g(x))}{\partial g_k}\right)\frac{\partial g_p(x)}{\partial x_j} \\
&= \sum_{k=1}^m a_{ik}\sum_{p=1}^m \frac{\partial^2 f(g(x))}{\partial g_p \partial g_k}\frac{\partial g_p(x)}{\partial x_j} \\
&= \sum_{k=1}^m a_{ik}\sum_{p=1}^m \frac{\partial^2 f(g(x))}{\partial g_p \partial g_k} a_{pj} \\
&= a_i^\top \nabla_g^2 f(g(x)) a_j,
\end{aligned}$$

which proves the statement.

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 a $n \times n$ 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.

Solution 3.2

1. Given the matrix A, the product PA is the matrix obtained by permuting the rows of A; AP corresponds to permuting the columns.
2. Every pair of columns (p_k, p_l) of P is of the form (e_k, e_l), where e_k, e_l are the k-th and the l-th standard basis vectors in $\mathbb{R}^n$. Thus, $\|p_k\|_2 = 1$, and $p_k^\top p_l = 0$ if $k \neq l$, as claimed.

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.

Solution 3.3 For $i = 1, \ldots, n$, let e_i be the i-th unit vector in $\mathbb{R}^n$. We have

$$f(e_i) = Ae_i = a_i,$$

where a_i is the i-th column of A. Hence we can compute the matrix A column-wise, by evaluating f at the points $e_1, \ldots, e_n$.

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]

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

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.

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?

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

Solution 3.4

1. We have

$$\begin{aligned} x(1) &= Bu(0) \\ x(2) &= Ax(1) + Bu(1) = ABu(0) + Bu(1) \\ x(3) &= Ax(2) + Bu(2) = A^2Bu(0) + ABu(1) + Bu(2). \end{aligned}$$

We now prove by induction that, for $T \geq 1$:

$$x(T) = \sum_{k=0}^{T-1} A^k Bu(T-1-k) = \begin{bmatrix} A^{T-1}B & \ldots & AB & B \end{bmatrix} U(T).$$

The formula is correct for $T = 1$. Let $T \geq 2$. Assume the formula is correct for $T-1$; we have

$$\begin{aligned} x(T) = Ax(T-1) + Bu(T-1) &= A\left(\sum_{k=0}^{T-2} A^k Bu(T-2-k)\right) + Bu(T-1) \\ &= \sum_{k=0}^{T-2} A^{k+1} Bu(T-2-k) + Bu(T-1) \\ &= \sum_{k=1}^{T-1} A^k Bu(T-1-k) + Bu(T-1) \\ &= \sum_{k=0}^{T-1} A^k Bu(T-1-k), \end{aligned}$$

as claimed. Finally, we have $y(T) = HU(T)$, with

$$H = C \cdot \left[\begin{array}{cccc} A^{T-1}B & \ldots & AB & B \end{array} \right].$$

2. The range of H is the set of output vectors that are attainable at time T by the system by proper choice of the sequence of inputs, starting from the initial state $x(0) = 0$.

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 $A^\top$.

Solution 3.5 Assume that the nullspace of B is contained in that of A. This means that

$$Bx = 0 \Longrightarrow Ax = 0.$$

Let $z \in \mathcal{R}(A^\top)$: there exist $y \in \mathbb{R}^m$ such that $z = A^\top y$. We have thus, for any element $x \in \mathcal{N}(A)$, $z^\top x = y^\top Ax = 0$. Hence, z is orthogonal to the nullspace of A, so it is orthogonal to the nullspace of B. We have obtained $\mathcal{R}(A^\top) \subseteq \mathcal{N}(B)^\perp = \mathcal{R}(B^\top)$, as claimed. Here, we have used the fundamental theorem of linear algebra (3.1).

Exercise 3.6 (Rank and nullspace) Consider the image in Figure 3.6, 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 $50, 230$.

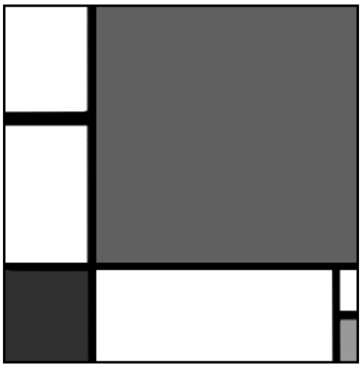

Figure 3.1: A gray-scale rendering of a painting by Mondrian.

1. What is nullspace of the matrix?

2. What is its rank?

Solution 3.6

1. Denote by e_i the i-th unit vector in $\mathbb{R}^{256}$, by $z_1 \in \mathbb{R}^{256}$ the vector with all first 50 components equal to one, by $z_2 \in \mathbb{R}^{256}$ the vector with all last 26 components equal to one, and by $z_3 \in \mathbb{R}^{256}$ the vector with all last 56 components equal to one. Finally, $\mathbf{1}$ denotes the vector of all ones in $\mathbb{R}^{256}$. We can express the matrix as

$$M = e_{100}z_1^\top + e_{200}\mathbf{1}^\top + e_{230}z_3^\top + \mathbf{1}e_{50}^\top + z_3 e_{230}^\top.$$

The condition $Mx = 0$, for some vector $x \in \mathbb{R}^n$, translates as

$$(z_1^\top x)e_{100} + (\mathbf{1}^\top x)e_{200} + (z_3^\top x)e_{230} + (e_{50}^\top x)\mathbf{1} + (e_{230}^\top x)z_3 = 0.$$

Since the vectors $(e_{100}, e_{200}, e_{230}, \mathbf{1}, z_3)$ are linearly independent, we obtain that the five coefficients in the above must be zero:

$$0 = z_1^\top x = \mathbf{1}^\top x = z_3^\top x = e_{50}^\top x = e_{230}^\top x.$$

It is easy to check that the corresponding subspace of $\mathbb{R}^{256}$ is of dimension $256 - 5 = 251$. Indeed, two elements of x are zero ($x_{50} = x_{230} = 0$), the remaining ones satisfy three independent equality constraints. From these we can express (say) x_1, x_{201}, x_{51} from the remaining variables, which then are free of any constraints. We can eliminate a total of five variables from the above five conditions, so the nullspace is of dimension 251.

2. The rank is 5.

Exercise 3.7 (Range and nullspace of $A^\top A$) Prove that, for any matrix $A \in \mathbb{R}^{m,n}$, it holds that

$$\begin{aligned} \mathcal{N}(A^\top A) &= \mathcal{N}(A) \\ \mathcal{R}(A^\top A) &= \mathcal{R}(A^\top). \end{aligned} \tag{3.1}$$

Hint: use the fundamental theorem of linear algebra.

Solution 3.7 First, suppose $x \in \mathcal{N}(A)$, then $Ax = 0$ and obviously $A^\top Ax = 0$. Conversely, suppose $x \in \mathcal{N}(A^\top A)$, we show by contradiction that it must be $x \in \mathcal{N}(A)$, hence proving the first claim. Indeed, suppose $x \in \mathcal{N}(A^\top A)$ but $x \notin \mathcal{N}(A)$. Define then $v = Ax \neq 0$. Such a v is by definition in the range of A, and $A^\top v = A^\top Ax = 0$, so v is also in the nullspace of $A^\top$, which is impossible since, by the fundamental theorem of linear algebra, $\mathcal{R}(A) \perp \mathcal{N}(A^\top)$. Next,

$$\mathcal{R}(A^\top) = \mathcal{N}(A)^\perp = \mathcal{N}(A^\top A)^\perp = \mathcal{R}(A^\top A),$$

which proves (3.1).

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.

Solution 3.8

1. The result is immediate from Lemma 3.3: if $A = U\Lambda U^{-1}$ is a diagonal factorization of A, then $p(\Lambda) = 0$, since by definition eigenvalues are roots of the characteristic polynomial, hence
$$p(A) = Up(\Lambda)U^{-1} = 0.$$

2. The map $p : \mathbb{R}^{n,n} \to \mathbb{R}^{n,n}$ with values $p(A) = A^n + c_{n-1}A^{n-1} + \cdots + c_1 A + c_0 I_n$ is continuous on $\mathbb{R}^{n,n}$. This map is identically zero on the dense subset of $\mathbb{R}^{n,n}$ formed by diagonalizable matrices (proved in the previous point of the exercise), hence by continuity it must be zero everywhere in $\mathbb{R}^{n,n}$.

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.

Solution 3.9

1. The mean of the output is zero: $\hat{y} = \mathbb{E}y = A\mathbb{E}u = 0$. Hence the covariance matrix is given by
$$\begin{aligned} \mathbb{E}(yy^\top) &= \mathbb{E}(Auu^\top A^\top) \\ &= A\mathbb{E}(uu^\top)A^\top \\ &= AA^\top. \end{aligned}$$

2. The total variance is
$$\begin{aligned} \mathbb{E}(y^\top y) &= \operatorname{trace} \mathbb{E}(yy^\top) \\ &= \operatorname{trace}(AA^\top) \\ &= \|A\|_F^2. \end{aligned}$$

The total output variance is the square of the Frobenius norm of the matrix. Hence the Frobenius norm captures the response of the matrix to a class of random inputs (zero mean, and unit covariance matrix).

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].

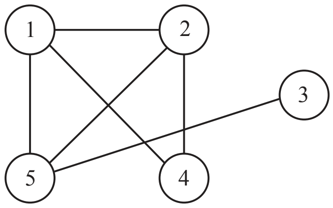

Figure 3.2: An undirected graph with $n = 5$ vertices.

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.

Solution 3.10

1. We can prove the result by induction on k. For $k = 1$, the result follows from the very definition of A. Let $L_k(i, j)$ denote the number of paths of length k between nodes i and j, and assume that the result we wish to prove is true for some given $h \geq 1$, so that $L_h(i, j) = [A^h]_{i,j}$. We next prove that it must also hold that $L_{h+1}(i, j) = [A^{k+1}]_{i,j}$, thus proving by inductive argument that $L_k(i, j) = [A^k]_{i,j}$ for all $k \geq 1$.

   Indeed, to go from a node i to a node j with a walk of length $h + 1$, one needs first reach, with a walk of length h, a node l linked to j by an edge. Thus:

$$L_{h+1}(i, j) = \sum_{l \in V(j)} L_h(i, l),$$

[3] The graph in Figure 3.2 has adjacency matrix

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

where $V(j)$ is the neighbor set of j, which is the set of nodes connected to the j-th node, that is, nodes l such that $A_{l,j} \neq 0$. Thus:

$$L_{h+1}(i,j) = \sum_{l=1}^{n} L_h(i,l) A_{l,j}.$$

But we assumed that $L_h(i,j) = [A^h]_{i,j}$, hence the previous equation can be written as

$$L_{h+1}(i,j) = \sum_{l=1}^{n} [A^h]_{i,l} A_{l,j}.$$

In the above we recognize the (i,j)-th element of the product $A^h A = A^{h+1}$, which proves that $L_{h+1}(i,j) = [A^{h+1}]_{i,j}$, and hence concludes the inductive proof.

2. Following the hint, we observe that 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. Therefore, each triangle in the graph produces 6 walks of length 3 (two walks for each vertex composing the triangle). From the previous result, the number of walks of length 3 from node i to itself is given by $[A^3]_{i,i}$, hence the total number of of walks of length 3 from each node to itself is $\sum_{i=1}^{n} [A^3]_{i,i} = \text{trace}(A^3)$, and therefore the number of triangles is $\text{trace}(A^3)/6$.

Exercise 3.11 (Nonnegative and positive matrices) A matrix $A \in \mathbb{R}^{n,n}$ is said to be *nonnegative* (resp. *positive*) if $a_{ij} \geq 0$ (resp. $a_{ij} > 0$) for all $i,j = 1,\ldots,n$. The notation $A \geq 0$ (resp. $A > 0$) is used to denote nonnegative (resp. positive) matrices.

A nonnegative 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 nonnegative if $x \geq 0$ (element-wise), and it is said to be a *probability vector*, if it is nonnegative 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 nonnegative matrices.

1. Prove that a nonnegative matrix A maps nonnegative vectors into nonnegative 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 $\lambda = 1$.

Solution 3.11 (Nonnegative and positive matrices)

1. Let $A \geq 0$, $x \geq 0$, $y = Ax$, and denote with $a_i^\top$ the i-th row of A. Then obviously

$$y_i = a_i^\top x = \sum_{j=1}^{n} a_{ij} x_j \geq 0, \quad i = 1, \ldots, n,$$

which shows that a nonnegative matrix maps nonnegative vectors into nonnegative vectors. Further, if x is a probability vector and A is stochastic, then

$$\mathbf{1}^\top y = \mathbf{1}^\top Ax = \mathbf{1}^\top x = 1,$$

which shows that y is also a probability vector.

2. Suppose by contradiction that $A > 0$ and $\rho(A) = 0$. This would imply that A has an eigenvalue of maximum modulus in $\lambda = 0$,

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

thus, all eigenvalues of A are actually zero. This means that the characteristic polynomial of A is $p_A(s) = s^n$ and, by the Cayley-Hamilton theorem, it must hold that $A^n = 0$, which is impossible since A^n is the n-fold product of positive matrices, hence it must be positive.

3. By the triangle inequality, we have that, for $i = 1, \ldots, n$,

$$\begin{aligned} |a_i^\top x| &\leq \sum_{j=1}^n |a_{ij} x_j| = \sum_{j=1}^n |a_{ij}||x_j| \\ &= |a_i^\top||x|, \end{aligned}$$

whih proves the first part. If $A > 0$ the above relation reads $|Ax| \leq A|x|$ which, for $x = v_i$, becomes

$$A|v_i| \geq |Av_i| = |\lambda_i v_i| = |\lambda_i||v_i|.$$

4. Let S be the probability simplex, and $f(x) \doteq \frac{Ax}{\mathbf{1}^\top Ax}$. From Brouwer's fixed-point theorem there exist $v \in S$ such that $f(v) = v$, that is such that

$$Av = (\mathbf{1}^\top Av)v = \lambda v, \quad \lambda \doteq \mathbf{1}^\top Av.$$

Moreover, since $A > 0$, it holds that $\lambda > 0$ and $v > 0$; thus A has a positive eigenvalue and a corresponding positive eigenvector. We next apply the same result to $A^\top$, obtaining that there exist $w \in S$ such that

$$A^\top w = (\mathbf{1}^\top A^\top w)w = \mu w, \quad \mu \doteq \mathbf{1}^\top A^\top w,$$

where again $\mu > 0$ and $w > 0$. Now, $v^\top w > 0$, and

$$\lambda v^\top w = v^\top A^\top w = \mu v^\top w,$$

which implies that $\mu = \lambda$, whence $A^\top w = \lambda w$.

Next, consider any eigenvalue/eigenvector pair λ_i, v_i for A, and apply the result of point 3. in this exercise, to obtain that

$$|\lambda_i||v_i| \leq A|v_i|, \quad i = 1, \ldots, n.$$

Multiply both sides on the left by $w^\top$ to get

$$w^\top |\lambda_i||v_i| \leq w^\top A|v_i| = \lambda w^\top |v_i|,$$

from which we obtain that

$$|\lambda_i| \leq \lambda, \quad i = 1, \ldots, n,$$

which proves that λ (which is real and positive, as shown above) is indeed a maximum modulus eigenvalue of A (thus, $\lambda = \rho(A)$), and the corresponding eigenvector v is positive.

5. By definition, if A is column stochastic then $\mathbf{1}^\top A = \mathbf{1}^\top$, which means that $\lambda = 1$ is an eigenvalue of A. Next, recall from Section 3.6.3.1 that the spectral radius of A is no larger than its induced ℓ_1 norm:

$$\rho(A) \leq \|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| = 1,$$

hence $\lambda = 1$ is indeed the dominant eigenvalue. An analogous argument applies to row stochastic matrices.

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?

Solution 4.1

1. We have
$$Ap = (cp + q), \;\; Aq = p + cq,$$
from which we obtain
$$A(p - q) = (c - 1)(p - q), \;\; A(p + q) = (c + 1)(p + q).$$
Thus $u_\pm := p \pm q$ is an (un-normalized) eigenvector of A, with eigenvalue $c \pm 1$.

2. The condition on $x \in \mathbb{R}^n$: $Ax = 0$, holds if and only if
$$0 = (q^\top x)p + (p^\top x)q = 0.$$
Since p, q are linearly independent, the above is equivalent to $p^\top x = q^\top x = 0$. The nullspace is the set of vectors orthogonal to p and q. The range is the span of p, q. The rank is thus 2.

3. Since the rank is 2, there is a total of two non-zero eigenvalues. Note that, since p, q are normalized, c is the cosine angle between p, q; $|c| < 1$ since p, q are independent. We have found two linearly independent eigenvectors $u_\pm = p \pm q$ that do not belong to the nullspace (since $|c| < 1$). We can complete this set with eigenvectors corresponding to the eigenvalue zero; simply choose an orthonormal basis for the nullspace.

   Then, the eigenvalue decomposition is
$$A = (c - 1)v_- v_-^\top + (c + 1)v_+ v_+^\top,$$

where $v_{\pm}$ are the normalized vectors $v_{\pm} = u_{\pm}/\|u_{\pm}\|_2$. We have

$$v_{\pm} = \frac{1}{\sqrt{2(1 \pm c)}}(p \pm q),$$

so that the eigenvalue decomposition amounts to the trivial identity

$$A = \frac{1}{2}\left((p+q)(p+q)^{\top} - (p-q)(p-q)^{\top}\right).$$

4. We can always scale the matrix: with $\bar{p} = p/\|p\|_2$, $\bar{q} = q/\|q\|_2$, we have

$$A = \|p\|_2\|q\|_2\left(\bar{p}\bar{q}^{\top} + \bar{q}\bar{p}^{\top}\right).$$

The eigenvalues are scaled accordingly: with $c = (p/\|p\|_2)^{\top}(q/\|q\|_2)$,

$$\lambda_{\pm} = \|p\|_2\|q\|_2(c \pm 1) = p^{\top}q \pm \|p\|_2\|q\|_2.$$

Note that, since p, q are independent, none of these eigenvalues is zero; one is positive, the other negative. This is due to the Cauchy-Schwartz inequality, which says that $|p^{\top}q| \leq \|p\|_2\|q\|_2$, with equality if and only if p, q are linearly dependent.

The corresponding (un-normalized) eigenvectors are $\bar{p} \pm \bar{q}$. The eigenvalue decomposition obtained before leads to

$$A = \frac{\|p\|_2\|q\|_2}{2}\left((\bar{p}+\bar{q})(\bar{p}+\bar{q})^{\top} - (\bar{p}-\bar{q})(\bar{p}-\bar{q})^{\top}\right).$$

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

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.

Solution 4.2 The region determined by $x^{\top}Ax \leq 1$ is best understood in terms of an eigenvalue decomposition of A: $A = U\Lambda U^{\top}$, with $\Lambda = \mathrm{diag}\,(\lambda_1, \lambda_2)$ are the two eigenvalues of A, with by convention $\lambda_1 \geq \lambda_2$. We have

$$x^{\top}Ax \leq 1 \iff \lambda_1\bar{x}_1^2 + \lambda_2\bar{x}_2^2 \leq 1,$$

where $\bar{x} = U^\top x$ (so that $x = U\bar{x}$). Several cases can occur.

If $\lambda_i > 0$, $i = 1, 2$, in the rotated space where $\bar{x}$ lives, the above set takes the shape of an ellipsoid centered at zero, with semi-axes lengths given by $1/\sqrt{\lambda_i}$, and principal directions $\bar{x}^{(i)} = e_i$ (the i-th unit vector), $i = 1, 2$. The principal directions in the original space are given by $x^{(i)} = Ue_i = u_i$, where u_i is the i-th column of U, $i = 1, 2$. Those vectors are nothing else than the eigenvectors of A.

If $\lambda_i < 0$, $i = 1, 2$, the set is the whole space $\mathbb{R}^2$.

If $\lambda_1 > 0$, $\lambda_2 < 0$, the set is an hyperboloid centered at zero, with principal directions given by the eigenvectors.

If one of the eigenvalues is zero, the set is a cylinder, again centered at zero. If $\lambda_1 > \lambda_2 = 0$, then the set is characterized by $|\bar{x}_1| \leq 1/\sqrt{\lambda_1}$. If $0 = \lambda_1 > \lambda_2$, the set is the outside of the set characterized by $|\bar{x}_2| \leq 1/\sqrt{-\lambda_2}$.

1. The eigenvalues are characterized by

$$0 = \det(\lambda I - A) = (\lambda - 2)^2 - 1,$$

which gives $\lambda = 3, 1$. We then solve for $u \in \mathbb{R}^2$ in $Au = \lambda u$, which gives two equations that are redundant by construction. We obtain from the first one $u_1 = u_2$ (if $\lambda = 3$) and $u_1 = -u_2$ (if $\lambda = 1$). Two normalized eigenvectors for $\lambda = 3, 1$ respectively, are

$$u_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad u_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

We obtain

$$A = \frac{3}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix}^\top + \frac{1}{2} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}^\top.$$

The region determined by $x^\top Ax \leq 1$ is

$$3\bar{x}_1^2 + \bar{x}_2^2 \leq 1,$$

with $\bar{x}_1 = (x_1 - x_2)/\sqrt{2}$, $\bar{x}_2 = (x_1 + x_2)\sqrt{2}$. This corresponds to an ellipsoid centered at zero with semi-axis length of $3, 1$, with length $1/\sqrt{3}$ associated with the direction $(1, 1)$ and the length 1 associated with the direction $(1, -1)$; the ellipsoid is rotated anti-clockwise by a 45° degree angle.

2. The set is characterized by

$$1 \geq x^\top Ax = x_1(x_1 - x_2) + x_2(-x_1 + x_2) = (x_1 - x_2)^2,$$

which translates as the slab $-1 \leq x_1 - x_2 \leq 1$. The slab is parallel to the direction $(1, 1)$, and is between the two lines $x_2 = x_1 \pm 1$.

3. The set is characterized by $-x_1^2 - x_2^2 \leq 1$, which is the whole space $\mathbb{R}^2$.

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

$$\mathcal{E} = \left\{ 4x_1^2 + 2x_2^2 + 3x_1x_2 + 4x_1 + 5x_2 + 3 \leq 1 \right\}.$$

Solution 4.3

1. Since A is positive-definite, it admits a Cholesky decomposition, if the form $A = R^\top R$, with R upper-triangular and invertible. In terms of the new variable $\bar{x} = Rx$, we have $x \in \mathcal{E}$ if and only if

$$1 \geq \bar{x}^\top \bar{x} - 2\bar{x}_0^\top \bar{x} + c = \|\bar{x} - \bar{x}_0\|_2^2 + c - \|\bar{x}_0\|_2^2.$$

where $\bar{x}_0 = -(R^\top)^{-1}b$. The set is empty when $1 + \|\bar{x}_0\|_2^2 < c$. Otherwise, the inequality writes $\|\bar{x} - \bar{x}_0\|_2 \leq \rho$, with $\rho^2 = 1 + \|\bar{x}_0\|_2^2 - c$. In the $\bar{x}$-space, the set is a circle with center x_0 and radius ρ. The set in the x-space is then obtained by the linear transformation $x = R^{-1}\bar{x}$. The resulting set is an ellipsoid is centered at $x_0 = R^{-1}\bar{x}_0 = -A^{-1}b$, and principal axes and directions given by the singular value decomposition of R (we don't need to detail those to draw the ellipsoid).

To produce a set of N points on the ellipsoid, encoded in a $2 \times N$ matrix X, the method starts by generating a set of N points on a unit circle, encoded in a $2 \times N$ matrix Z; the set of points in x-space is then obtained by scaling by ρ, adding $\bar{x}_0$ defined above to each column, and multiplying the resulting matrix by the matrix R^{-1}. That is:

$$X = R^{-1} \left(\rho \cdot Z + \bar{x}_0 \mathbf{1}^\top \right),$$

with $\mathbf{1}$ the vector of ones in $\mathbb{R}^N$.

2. We have now

$$A = \begin{bmatrix} 4 & 3/2 \\ 3/2 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ 5/2 \end{bmatrix}, \quad c = 2.$$

A Cholesky decomposition of A is $A = R^T R$, with

$$R = \begin{bmatrix} 2 & 3/4 \\ 0 & 1.19 \end{bmatrix}.$$

We have $\bar{x}_0 = -(R^\top)^{-1}b = (-1, -1.4596)$, and $\rho = 1.4596$. The result is shown in Fig 4.3.

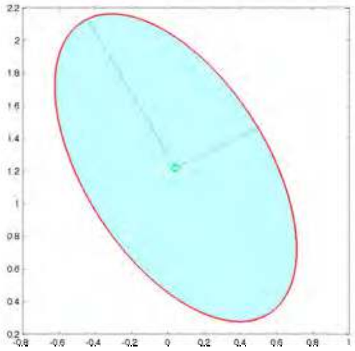

Figure 4.3: An ellipsoid in two dimensions.

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

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

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: $Q = \text{diag}(\lambda_1, \ldots, \lambda_n)$, with $\lambda_1 \geq \ldots \geq \lambda_r > \lambda_{r+1} = \ldots = \lambda_n = 0$, 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.

Solution 4.4

1. When $Q \succ 0$, it admits a Cholesky decomposition $Q = R^\top R$, with R upper-triangular and invertible. We can define the new variable $\bar{x} = Rx$, which leads to the problem

$$\min_{\bar{x}} \frac{1}{2} \bar{x}^\top \bar{x} - \bar{c}^\top \bar{x},$$

where $\bar{c} = (R^{-1})^\top c$. We can express the objective in the above problem as

$$\frac{1}{2} \|\bar{x} - \bar{c}\|_2^2 - \|\bar{c}\|_2^2,$$

[5] See Section 5.2.3.

from which it is clear that the unique minimizer is $\bar{x} = \bar{c}$. In terms of the x-variable, the unique solution is $x = R^{-1}\bar{c} = Q^{-1}b$.

The same result can be obtained by invoking the fact that the minimizers of a convex differentiable function f without any constraints, are characterized by the optimality condition [6] $\nabla f(x) = 0$. In our case, we have $\nabla f(x) = Qx - c$.

2. The objective function writes

$$f(x) = \sum_{i=1}^{r} \left(\frac{1}{2} \lambda_i x_i^2 - c_i x_i \right) + \sum_{i=r+1}^{n} c_i x_i .$$

If any element c_i, $i = r+1, \ldots, n$, is non-zero, the optimal value is $-\infty$. Otherwise, that is, when c is in the range of Q, the optimal value is obtained with $x_i = c_i / \lambda_i$, $i = 1, \ldots, r$, and the other variables $x_{r+1}, \ldots, x_n$ free. That value is

$$p^* = -\frac{1}{2} \sum_{i=1}^{r} \frac{c_i^2}{\lambda_i} .$$

3. We use the eigenvalue decomposition of Q: $Q = U \Lambda U^\top$, with $\Lambda = \mathbf{diag}\,(\lambda_1, \ldots, \lambda_n)$. The problem is more conveniently formulated in terms of the new variable $\bar{x} = U^\top x$:

$$p^* = \min_{\bar{x}} \frac{1}{2} \bar{x}^\top \Lambda \bar{x} - \bar{c}^\top \bar{x},$$

with $\bar{c} \doteq Uc$.

Assuming as before $\lambda_1 \geq \ldots \geq \lambda_r > \lambda_{r+1} = \ldots = \lambda_n = 0$, where r is the rank of Q ($1 \leq r < n$), we are lead to the similar conclusions. In particular, the optimal value is finite if and only if the last $n - r$ components of $\bar{c}$ are zero. This means that c must be in the range of Q, in order for the value to be finite.

4. Let us assume that the optimal value is finite. In that case, the optimal set is the set of points $U\bar{x}$, where $\bar{x}_{r+1}, \ldots, \bar{x}_n$ are free, and $x_i = \bar{c}_i / \lambda_i$, $i = 1, \ldots, r$. The solution corresponding to $\bar{x}_{r+1} = \ldots = \bar{x}_n = 0$ is nothing else than $Q^\dagger c$, where $Q^\dagger = U \Lambda^\dagger U^\top$ is the pseudo-inverse of Q, defined via

$$\Lambda^\dagger = \mathbf{diag}\,(1/\lambda_1, \ldots, 1/\lambda_r, 0, \ldots, 0) .$$

The optimal value is then

$$p^* = -\frac{1}{2} \sum_{i=1}^{r} \frac{\bar{c}_i^2}{\lambda_i} = -\frac{1}{2} c^\top Q^\dagger c,$$

and the optimal set is of the form

$$x = Q^\dagger c + (I - QQ^\dagger) v,$$

with v a free variable.

[6] See Section 8.4.1.

Exercise 4.5 (Interpretation of covariance matrix) As in Example 4.2, we are given m data points $x^{(1)}, \ldots, x^{(m)}$ 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 $x^{(i)}$, $i = 1, \ldots, m$, on the line $\mathcal{L}(w)$, and look 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.

Solution 4.5

1. The result is obtained by noting that, when $\|w\|_2 = 1$, we have for any $x \in \mathbb{R}^n$:

$$\|tw - x\|_2^2 = (t - t^*)^2 + \text{ constant},$$

where $t^* = x^\top w$. Minimizing the above over t gives $t = t^*$.

2. We have that
$$\sum_{i=1}^{m} t_i(w) = \hat{x}^\top w$$
is a constant on the unit sphere, independent of w. This is only possible if $\hat{x} = 0$. Indeed, taking $w = e_i$ (the i-th unit vector in $\mathbb{R}^n$), we see that all the elements of $\bar{x}$ must be equal, so that $\bar{x} = \beta\mathbf{1}$, with $\beta \in \mathbb{R}$ and $\mathbf{1}$ the vector of ones in $\mathbb{R}$. If $\beta \neq 0$, we obtain that the linear fuction $w \to \mathbf{1}^\top w$ is constant over the unit sphere, a contradiction. Hence $\beta = 0$, and thus $\hat{x} = 0$.

3. The sample variance of the projected values is given by
$$\sigma^2(w) = \frac{1}{m}\sum_{i=1}^{m} t_i(w)^2 = \frac{1}{m}\sum_{i=1}^{m}(w^\top x^{(i)})^2 = w^\top \Sigma w,$$
where we have used that $\hat{x} = 0$, which implies
$$\Sigma = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}(x^{(i)})^\top.$$
Since $\sigma^2(w)$ is a constant on the unit sphere, the function $w \to w^\top \Sigma w$ is also constant on the unit sphere. Thus, the largest and smallest eigenvalues coincide, and Σ is proportional to the identity: $\Sigma = \alpha I$, with $\alpha \in \mathbb{R}$. The proportion factor is σ^2, since $\sigma^2 = w^\top \Sigma w = \alpha w^\top w = \alpha$.

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
$$L_{ij} = \begin{cases} -1 & \text{if } (i,j) \in E, \\ d(i) & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$
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 4.4.

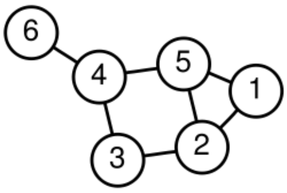

Figure 4.4: Example of an undirected graph.

1. Form the Laplacian for the graph shown in Figure 4.4.

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[7] of L.

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 $(i,j) \in E$.

Solution 4.6

1. The Laplacian for the graph is

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

2. L is symmetric, by definition of the edge set E.

3. Let us prove that the expression is valid by looking at the values of the quadratic form at specific points. Let

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

Clearly, q is of the form $u^\top Qu$ for some symmetric matrix Q. Note that $q(u) \geq 0$ for every u, hence Q is positive semi-definite. Let us show that $Q = L$, which will imply that L is positive semi-definite.

Expanding, we obtain

$$q(u) = \sum_{(i,j)\in E} d(i)u_i^2 - \sum_{i\neq j,\,(i,j)\in E} u_i u_j.$$

We have for every $k \in \{1, \ldots, n\}$:

$$Q_{kk} = q(e_k) = \sum_{j\,:\,(k,j)\in E} 1 = d(k),$$

We next find Q_{kl} for $k \neq l$, thanks to the formula

$$Q_{kl} = \frac{1}{2}(q(e_k + e_l) - q(e_k - e_l)).$$

[7] See Section 4.4.4.

If $(k,l) \in E$, $k \neq l$, we have

$$q(e_k + e_l) = d(k) + d(l) - 1, \;\; q(e_k - e_l) = d(k) + d(l) + 1,$$

which implies that $Q_{kl} = -1$. Otherwise, both terms are zero, and $Q_{kl} = 0$. This proves the desired result.

4. Since L is PSD, we can write $L = R^\top R$, for some matrix R. Let $\mathbf{1}$ be the vector of ones in $\mathbb{R}^n$. We have

$$q(\mathbf{1}) = \|R\mathbf{1}\|_2^2 = 0,$$

which implies $R\mathbf{1} = 0$, and in turn, $L\mathbf{1} = 0$. Hence 0 is always an eigenvalue of L, associated for the eigenvector $\mathbf{1}$.

5. If u is a vector such that $q(u) = 0$, then from the expression of q, $u_k = u_l$ for every pair of vertices $(k,l) \in E$. Now let $(k,l) \in V \times V$, $(k,l) \notin E$ be a pair of vertices that are not directly connected. Since the graph is connected, there is a path of vertices connecting vertex k and l, that is, a sequence of indices such that $i_0 = k \leq i_1 \leq \ldots \leq i_m = l$, for some path length $m > 1$, and with $(i_s, i_{s+1}) \in E$, $s = 0, \ldots, m-1$. Since $u_{i_s} = u_{i_{s+1}}$, we obtain by induction on s that $u_k = u_{i_0} = u_{i_m} = u_l$, which shows that u is proportional to the vector of ones, as claimed.

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 $C_{ij} = A_{ij}B_{ij}$, $1 \leq i,j \leq n$. 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.

Solution 4.7 If A is rank-one, we can express it as $A = vv^\top$ for some $v \in \mathbb{R}^n$. We then have, for an arbitrary vector $z \in \mathbb{R}^n$:

$$z^\top Cz = \sum_{i,j} z_i z_j A_{ij} B_{ij} = \sum_{i,j} z_i z_j v_i v_j B_{ij} = y^\top By,$$

where y is the n-vector with i-th component $v_i z_i$, $i = 1, \ldots, n$. The proof follows from the positive semi-definiteness of B.

When A is of arbitrary rank $r \leq n$, we can write

$$A = \sum_{k=1}^r v^{(k)} (v^{(k)})^\top,$$

for some vectors $v^{(k)}$, $k = 1, \ldots, r$. For an arbitrary vector $z \in \mathbb{R}^n$:

$$z^\top Cz = \sum_{k=1}^r \sum_{i,j} z_i z_j v_i^{(k)} v_j^{(k)} B_{ij} = \sum_{k=1}^r (y^{(k)})^\top By^{(k)},$$

where, for $k = 1, \ldots, r$, $y^{(k)}$ is the n-vector with i-th component $v_i^{(k)} z_i$, $i = 1, \ldots, n$. The proof again follows from the positive semi-definiteness of B.

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

3. With all terms defined as in the previous point, prove that
$$\rho(I - \alpha BA) < 1, \quad \forall \alpha \in (0, 2).$$

Solution 4.8 (A bound on the eigenvalues of a product)

1. This point is an immediate consequence of Corollary 4.5.

2. From the first point, we have that
$$0 < \lambda_{\max}(BA) = \max_{i=1,\ldots,n} |\lambda_i(BA)| = \rho(BA) \leq \|BA\|_\infty,$$
where the last inequality is proved in Section 3.6.3.1. Further, since B is positive and diagonal,
$$\|BA\|_\infty = \max_i [B|A|\mathbf{1}]_i = \max_i B_{ii} |a_i^\top| \mathbf{1} = 1,$$
whence
$$0 < \lambda_{\max}(BA) \leq 1,$$
as desired.

3. We have that $\rho(I - \alpha BA) < 1$ iff $|1 - \alpha\lambda_i(BA)| < 1$ for all i, iff
$$0 < \alpha\lambda_i(BA) < 2, \quad \text{for all } i.$$
Since $0 < \lambda_i(BA) \leq 1$ for all i, the above condition requires $\alpha > 0$ and
$$\alpha \max_i \lambda_i(BA) < 2,$$
which is satisfied if $\alpha < 2$, since $\max_i \lambda_i(BA) \leq 1$.

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

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

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).

Solution 4.9 (Hadamard's inequality) First observe that $A \succeq 0$ implies $a_{ii} \geq 0$ for all i, hence $\prod_i a_{ii} \geq 0$. Thus, the inequality is satisfied if $\det A = 0$. Suppose next $\det A \neq 0$, that is, A is invertible (hence positive definite), thus $a_{ii} > 0$ for all i. Let then $D = \operatorname{diag}\left(a_{11}^{-1/2}, \ldots, a_{nn}^{-1/2}\right)$ and observe that

$$\tilde{A} \doteq DAD$$

has all diagonal entries equal to one, and $\det \tilde{A} \leq 1$ if an only if $\det A \leq \prod_i a_{ii}$. Therefore, it suffices to prove that $\det \tilde{A} \leq 1$. Indeed,

$$\det \tilde{A} = \prod_i \lambda_i(\tilde{A}) \leq \left(\frac{1}{n}\sum_i \lambda_i(\tilde{A})\right)^n = \left(\frac{1}{n} \operatorname{trace} \tilde{A}\right)^n = 1,$$

where the first inequality follows from the geometric/arithmetic mean inequality.

Exercise 4.10 (A lower bound on the rank) Let $A \in \mathbb{S}_+^n$ be a symmetric, positive semi-definite 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
$$(\operatorname{trace} A)^2 \leq \operatorname{rank}(A)\|A\|_F^2.$$

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

Solution 4.10

1. We begin with an eigenvalue decomposition of $A = U\Lambda U^\top$, where $\Lambda = \operatorname{diag}(\lambda_, \ldots, \lambda_n)$ contains the (non-negative) eigenvalues, and the columns of the matrix U are eigenvectors. Since $UU^\top = I$, we have
$$\operatorname{trace} A = \operatorname{trace}(U\Lambda U^\top) = \operatorname{trace}(\Lambda U^\top U) = \operatorname{trace} \Lambda = \sum_{i=1}^{n} \lambda_i.$$

Since $\lambda_i \geq 0$, $i = 1, \ldots, n$, we obtain that the trace is simply the ℓ_1-norm of the vector of eigenvalues $\lambda \doteq (\lambda_1, \ldots, \lambda_n)$:

$$\text{trace}\, A = \|\lambda\|_1.$$

Likewise, the Frobenius norm depends only on the eigenvalues:

$$\begin{aligned} \|A\|_F^2 &= \text{trace}(A^\top A) \\ &= \text{trace}(U\Lambda U^\top U\Lambda U^\top) \\ &= \text{trace}(U\Lambda^2 U^\top) \\ &= \text{trace}(\Lambda^2) \\ &= \sum_{i=1}^n \lambda_i^2. \end{aligned}$$

We obtain $\|A\|_F = \|\lambda\|_2$.

2. The rank of A is simply the number of eigenvalues that are non-zero. Without loss of generality, we assume that the eigenvalues are ordered in decreasing fashion:

$$\lambda_1 \geq \ldots \geq \lambda_r > \lambda_{r+1} = \ldots = \lambda_n = 0.$$

Define the vector $\mu = (\lambda_1, \ldots, \lambda_r) \in \mathbb{R}^r$. We have

$$\text{trace}\, A = \|\mu\|_1, \;\; \|A\|_F = \|\mu\|_2.$$

The inequality follows from the Cauchy-Schwartz inequality: with $\mathbf{1}_r$ the vector of ones in $\mathbb{R}^r$, we have

$$\|\mu\|_1 = \mu^\top \mathbf{1}_r \leq \|\mathbf{1}_r\|_2 \cdot \|\mu\|_2 = \sqrt{r}\|\mu\|_2,$$

which proves the desired result.

3. The lower bound is attained when $\|\mu\|_1 = \sqrt{r}\|\mu\|_2$. According to the Cauchy-Schwartz inequality result[8], this implies that $\mu, \mathbf{1}_r$ are collinear, which means that $\mu = \alpha \mathbf{1}_r$ for some $\alpha \in \mathbb{R}_+$. The matrices for which the lower bound on the rank is attained are of the form

$$A = \alpha B, \;\; B \doteq \sum_{k=1}^r u^{(k)} (u^{(k)})^\top,$$

where $\alpha \geq 0$, $u^{(k)}$, $k = 1, \ldots, r$ are r orthonormal vectors in $\mathbb{R}^n$. Up to a non-negative scaling factor α, such matrices are orthogonal projections on subspaces (with basis given by the vectors $u^{(k)}$).

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

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

[8] See Section 2.2.

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

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

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

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

Solution 4.11 Let $\Sigma = R^\top R$ be Cholesky decomposition of $\Sigma \succ 0$, so that $\Sigma = R^\top R$, with R a $n \times n$ upper-triangular, invertible matrix. We have $\Sigma^{-1} = R^{-1}(R^{-1})^\top$, and thus

$$\begin{aligned} q(x) &\doteq x^\top \Sigma^{-1} x \\ &= x^\top R^{-1} (R^{-1})^\top x \\ &= z^\top z, \end{aligned}$$

where $R^\top z = x$. The transformation $x \to z$ is a valid change of variable, and the infinitesimal volume $\mathrm{d}x$ is expressed as $\gamma \mathrm{d}z$, with $\gamma = |\det R^\top| = |\det R| = \sqrt{\det \Sigma}$.

We have, using the hint,

$$\begin{aligned} \int_{\mathbb{R}^n} e^{-\frac{1}{2}x^\top \Sigma^{-1} x} \mathrm{d}x &= \int_{\mathbb{R}^n} e^{-\frac{1}{2}\|(R^{-1})^\top x\|_2^2} \mathrm{d}x \\ &= \gamma \cdot \int_{\mathbb{R}^n} e^{-\frac{1}{2}\|z\|_2^2} \mathrm{d}z \\ &= \gamma \cdot \Pi_{i=1}^n \int_{z_i \in \mathbb{R}} e^{-\frac{1}{2}z_i^2} \mathrm{d}z_i \\ &= \gamma \cdot \left(\int_{\eta \in \mathbb{R}} e^{-\frac{1}{2}\eta^2} \mathrm{d}\eta \right)^n . \end{aligned}$$

The result follows from the fact that, as implied from the case $n = 1$, $\Sigma = 1$:

$$\int_{\eta \in \mathbb{R}} e^{-\frac{1}{2}\eta^2} \mathrm{d}\eta = \sqrt{2\pi}.$$

5. Singular Value Decomposition

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

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

1. Show that A is orthogonal.

2. Find a singular value decomposition of A.

Solution 5.1

1. The columns are easily shown to be mutually orthogonal, and have unit Euclidean norms.

2. Since A is orthogonal, a valid SVD of A is $U\Sigma V^\top$, with $U = A$, $\Sigma = V = I$. As usual, there is no unicity: we can also choose $U = I$, $V = A^\top$.

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_i's. Be as explicit as you can.

Solution 5.2 Assume first that none of the a_i's is zero.Let $\sigma_i = \|a_i\|_2$, $u_i = a_i/\sigma_i$, $i = 1, \ldots, n$. By assumption the matrix $U \doteq (u_1, \ldots, u_n)$ is orthogonal. In addition, we have

$$A = U\Sigma,$$

where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$. The above is an SVD of A, with right singular vector matrix $V = I$.

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

$$\tilde{A} \doteq \begin{bmatrix} A \\ I_m \end{bmatrix}$$

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

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

Solution 5.3

1. We have $\tilde{A}^\top \tilde{A} = A^\top A + I$. Since the eigenvalues of $A^\top A$ are σ_i^2, $i = 1, \ldots, m$, those of the shifted matrix $\tilde{A}^\top \tilde{A}$ are $\sigma_i^2 + 1$, $i = 1, \ldots, m$. Hence, the singular values of $\tilde{A}$ satisfy the desired relation with those of A.

2. Let $A = USV^\top$ be an SVD of A, with the familiar notation. The SVD of $\tilde{A} = \tilde{U}\tilde{S}\tilde{V}^\top$ has $\tilde{S} = \mathrm{diag}(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_m)$; matrix $\tilde{U}$ contain the eigenvectors of $\tilde{A}\tilde{A}^\top$, while $\tilde{V}$ contains those of $\tilde{A}^\top \tilde{A}$. Since $\tilde{A}^\top \tilde{A} = A^\top A + I$, we can choose $\tilde{V} = V$.

   The condition for $(\tilde{u}, \tilde{v})$ to be a pair of left- and right-singular vectors of $\tilde{A}$ are that both vectors must be unit-norm, and
$$\tilde{A}\tilde{v} = \tilde{\sigma}\tilde{u}, \;\; \tilde{A}^\top \tilde{u} = \tilde{\sigma}\tilde{v}.$$

   We have seen that we can choose $\tilde{v} = v$ to be an eigenvector of $A^\top A$ (that is, a right singular vector of A). Further, decomposing $\tilde{u} = (\tilde{u}^1, \tilde{u}^2)$, with $\tilde{u}^1 \in \mathbb{R}^n$, we obtain
$$Av = \tilde{\sigma}\tilde{u}^1, \;\; v = \tilde{\sigma}\tilde{u}^2, \;\; A^\top \tilde{u}^1 + \tilde{u}^2 = \tilde{\sigma} v.$$

   Solving for the second equation: $\tilde{u}^2 = v/\tilde{\sigma}$, we obtain from the third $A^\top \tilde{u}^1 = (\tilde{\sigma} - 1/\tilde{\sigma})v$. Multiplying by A, and with the first equation, we then obtain
$$AA^\top \tilde{u}^1 = \tilde{\sigma}(\tilde{\sigma} - 1/\tilde{\sigma})\tilde{u}^1 = \sigma u^1.$$

   This shows that we can set $\tilde{u}^1$ to be proportional to a left singular vector u of A, and $\tilde{u}^2 = v/\tilde{\sigma}$ proportional to v. We have
$$\tilde{u} = \begin{bmatrix} \alpha u \\ \frac{1}{\tilde{\sigma}} v \end{bmatrix},$$
   where α must be chosen so that the above has unit Euclidean norm, that is:
$$\alpha = \frac{\sigma}{\sqrt{\sigma^2 + 1}}.$$
   We have obtained that a generic pair of left- and right singular vectors $(\tilde{u}, \tilde{v})$ of $\tilde{A}$ corresponding to the singular value $\sqrt{\sigma^2 + 1}$, can be constructed from a generic pair of left- and right singular vectors (u, v) of A corresponding to the singular value σ, with the choice
$$\tilde{u} = \begin{bmatrix} \frac{\sigma}{\sqrt{\sigma^2+1}} u \\ \frac{1}{\sqrt{\sigma^2+1}} v \end{bmatrix}, \;\; \tilde{v} = v.$$

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.

Solution 5.4

1. We have $G_{ij} \approx s_i q_j$, $i = 1, \ldots, n$, $j = 1, \ldots, m$. In this rank-one model, each student i is characterized by an ability level s_i, and each question j has a difficulty level j.

2. We solve for the rank-one approximation problem

$$\min_{q \geq 0, s \geq 0} \|G - sq^\top\|_F$$

   Ignoring the non-negativity constraints, we obtain s, q by the SVD of G. If σ_1 is the largest singular value, and (u_1, v_1) corresponding left- and right singular vectors, we set $s = \sqrt{\sigma_1} u_1$, $q = \sqrt{\sigma_1} v_1$.

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 is described by an n-vector d_j, where n is the total number of distinct words appearing in the whole corpus. 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[9] 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 a 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 $\tilde{q} = q / \|q\|_2$.

[9] 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.

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 a $n \times m$ 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[10] to document similarity, the idea is to first project the documents and the query onto the sub-space generated by the singular vectors $u_1, \ldots, u_k$, and then apply cosine similarity approach to the projected vectors. Find an expression for the measure of similarity.

Solution 5.5

1. If $d \in \mathbb{R}^n$ is a specific document, with d_i indicating the presence or absence of word i in the document, then $q^\top d$ is the number of common terms between the document and the query.

   For $j = 1, \ldots, m$, the number of common terms between the query q and the j-th document D_j is given by $q^\top M e_j$, with e_j the j-th unit vector in $\mathbb{R}^m$. Hence the m-vector $M^\top q$ gives the number of co-occurring terms between the query and each document. We select the closest documents by selecting indices j that achieve the maximum, that is

$$j \in \arg\max_{1 \le j \le m} (M^\top q)_j.$$

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

2. With $\tilde{d}$ the normalized vector corresponding to a generic document, we have

$$\begin{aligned} \|\tilde{q}-\tilde{d}\|_2^2 &= \|\tilde{q}\|_2^2+\|\tilde{d}\|_2^2-2\tilde{q}^\top \tilde{d} \\ &= \|\tilde{q}\|_2^2+1-2\cos(\theta), \end{aligned}$$

where θ is the angle between the two vectors[11] d, q. This explains the name of the method, which seeks the document D_j such that the angle between q and d_j is the smallest.

The cosine similarity approach is based on the matrix-vector product $\tilde{M}^\top \tilde{q}$: it involves ordering that vector by decreasing magnitude, and selecting the index j with the largest:

$$j \in \arg\max_{1\le j\le m} (\tilde{M}^\top \tilde{q})_j.$$

3. We can write

$$M=\sum_{l=1}^{r} \sigma_l u_l v_l^\top,$$

which can be interpreted as follows. Each u_l corresponds to a n-vector assigning a (positive or negative) weight to each term in the dictionary. Hence u_l corresponds to a synthetic document. The vectors v_l assign a weight to each document, hence can be understood as a "concept", if we accept the idea that a concept is a weighted list of words. Indeed, if v_l is sparse, the indices where it is non-zero correspond to a short list of terms that can be understood as a concept. The term-by-document matrix M is a linear combination of concept and synthetic document pairs.

4. For a generic document D_j in the text collection, we have

$$d_j = Me_j = \sum_{l=1}^{r} \sigma_l (v_l^\top e_j) u_l.$$

Thus, the coordinates of d_j in the orthonormal basis $(u_1,\ldots,u_n)$ are $(\sigma_l(v_l^\top e_j))_{1\le j\le n}$, with the convention $\sigma_{r+1}=\ldots=\sigma_n=0$. The projection onto the subspace spanned by the first k vectors of the basis is the k-vector $\hat{d}_j$ with the same k first coordinates as d:

$$\hat{d}_j = \sum_{l=1}^{k} \sigma_l (v_l^\top e_j) u_l.$$

For a query document q, we can write

$$q=\sum_{l=1}^{n} (q^\top u_l) u_l.$$

[11] See Section 2.2.2.2.

The projection $\hat{q}$ onto the subspace spanned by the $(u_1, \ldots, u_k)$ is

$$\hat{q} = \sum_{l=1}^{k} (q^\top u_l) u_l.$$

The Latent Semantic Indexing approach is based on evaluating the angles between the projected query $\hat{q}$ and the projected documents $\hat{d}_j$, $j = 1, \ldots, m$.

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 an 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.

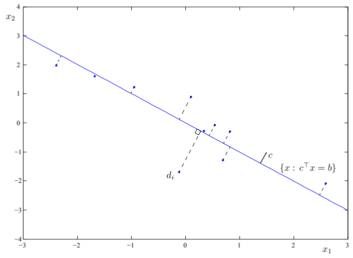

Figure 5.5: Fitting an hyperplane to data.

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

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

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 Chapter 8.4.1.

4. Explain how to find the hyperplane via SVD.

Solution 5.6

1. We have from (2.6) that

$$\mathrm{dist}(d,\mathcal{H}) = |c^\top d - b| = \left| [d^\top \; -1] \begin{bmatrix} c \\ b \end{bmatrix} \right|.$$

2. Letting $D \in \mathbb{R}^{n,m}$ be the matrix having d_i as columns, $D = [d_1 \cdots d_m]$, the problem objective becomes

$$\begin{aligned} f_0 &= \sum_{i=1}^m \left([d_i^\top \; -1] \begin{bmatrix} c \\ b \end{bmatrix} \right)^2 \\ &= \left\| [D^\top \; -\mathbf{1}] \begin{bmatrix} c \\ b \end{bmatrix} \right\|_2^2 \\ &= \begin{bmatrix} c \\ b \end{bmatrix}^\top \begin{bmatrix} DD^\top & -m\bar{d} \\ -m\bar{d}^\top & m \end{bmatrix} \begin{bmatrix} c \\ b \end{bmatrix} \\ &= c^\top (DD^\top) c - 2mbc^\top \bar{d} + mb^2, \end{aligned}$$

where

$$\bar{d} = \frac{1}{m} D\mathbf{1} = \frac{1}{m} \sum_{i=1}^m d_i$$

is the barycenter (average) of the given points.

3. Since the problem is unconstrained in variable b, we can partially minimize f_0 with respect to b by simply imposing that

$$\frac{\partial f_0}{\partial b} = -2mc^\top \bar{d} + 2mb = 0,$$

resulting in

$$b = c^\top \bar{d},$$

which, substituted back in f_0 gives

$$\tilde{f}_0(c) = c^\top H c, \quad H \doteq DD^\top - m\bar{d}\bar{d}^\top = D\left(I_m - \frac{1}{m}\mathbf{1}\mathbf{1}^\top \right) D^\top.$$

Notice that matrix

$$E \doteq I_m - \frac{1}{m}\mathbf{1}\mathbf{1}^\top$$

is symmetric and *idempotent*, that is $EE = E$, therefore $H = \tilde{D}\tilde{D}^\top$, where

$$\tilde{D} = DE = D\left(I_m - \frac{1}{m}\mathbf{1}\mathbf{1}^\top \right) = D - \bar{d}\mathbf{1}^\top$$

represent the matrix of *centered* data points, i.e., the i-th column of $\tilde{D}$ is $d_i - \bar{d}$, where $\bar{d}$ is the average of the data points.

4. From Theorem 4.3, the optimal objective value of our problem

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

is the minimum eigenvalue of $H = \tilde{D}\tilde{D}^\top$, which coincides with the smallest singular value of $\tilde{D}$. The optimal value is

$$f_0^* = \lambda_{\min}(\tilde{D}\tilde{D}^\top) = \sigma_n^2,$$

with σ_n the smallest singular value of $\tilde{D}$. The optimal value is achieved for $c = u_n$, being u_n the left singular vector of $\tilde{D}$ corresponding to the smallest singular value σ_n.

To summarize, the problem of finding the hyperplane that best fits given points $d_1, \ldots, d_m$ can be solved as follows: we construct the centered data points matrix $\tilde{D} \in \mathbb{R}^{n,m}$, and find its minimum singular value σ_n and the corresponding left singular vector u_n. Then, the best fitting hyperplane is

$$\mathcal{H} = \{x : u_n^\top x = u_n^\top \bar{d}\},$$

where $\bar{d}$ is the barycenter of the data points.

An interesting interpretation of this result is that we found a direction in data space, u_n, along which the centered data have the *least variation*, in the mean-square sense, meaning that the mean-square deviation of the centered data from the flat $\mathcal{H}$, along the direction u_n, is minimal and equal to σ_n^2.

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.

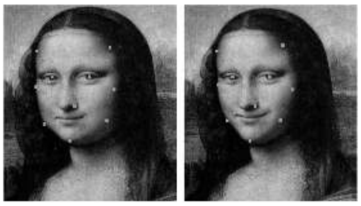

Figure 5.6: Image deformation via rigid transformation. The image on the left is the original image, and that on the right is the deformed image. Yellow dots indicate points for which the deformation is chosen by the user.

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) writes 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 $\operatorname{trace} Z$.

7. How woud you apply this technique to make Mona Lisa smile more? *Hint:* in Figure 5.6, the two-dimensional points x_i are given (as yellow 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.

Solution 5.7

1. For fixed R, denoting by $z_i = y_i - Rx_i$, $i = 1, \ldots, m$, we obtain the problem in variable r

$$\min_r \sum_{i=1}^{m} \|r - z_i\|_2^2$$

The solution to this unconstrained least-squares problem can be obtained by setting the derivative to zero. This yields a (unique) optimal point r^*

$$r^* = \hat{z} = \frac{1}{m}\sum_{i=1}^{m} z_i = \hat{y} - R\hat{x}$$

2. At optimum, the objective is

$$\sum_{i=1}^{m} \|z_i - \hat{z}\|_2^2 = \sum_{i=1}^{m} \|y_i - \hat{y} - R(x_i - \hat{x})\|_2^2 = \sum_{i=1}^{m} \|\bar{y}_i - R\bar{x}_i\|_2^2,$$

as claimed.

3. We now consider the problem

$$\min_{R \in \mathbb{R}^{n,n}} \sum_{i=1}^{m} \|R\bar{x}_i - \bar{y}_i\|_2^2 \;:\; R^\top R = I_n,$$

We can express the objective as stated, with

$$X = [\bar{x}_1, \ldots, \bar{x}_m] \in \mathbb{R}^{n,m}, \;\; \bar{Y} = [\bar{y}_1, \ldots, \bar{y}_m] \in \mathbb{R}^{n,m}.$$

Indeed, the columns of the matrix $RX - Y$ are $R\bar{x}_i - \bar{y}_i$, $i = 1, \ldots, m$.

4. We can safely square the objective, since it is always non-negative. We have, for any R with $R^\top R = I_n$:

$$\begin{aligned}
\|RX - Y\|_F^2 &= \operatorname{trace}(RX - Y)^\top (RX - Y) \\
&= \operatorname{trace} X^\top R^\top RX - 2\operatorname{trace} Y^\top RX + \operatorname{trace} Y^\top Y \\
&= c - 2\operatorname{trace} Y^\top RX,
\end{aligned}$$

where $c := \operatorname{trace} X^\top X + \operatorname{trace} Y^\top Y$ is a constant. We note that $\operatorname{trace} Y^\top RX = \operatorname{trace} RXY^\top$. Hence, the problem can be expressed as claimed, with $Z = XY^\top \in \mathbb{R}^{n,n}$.

5. Let $Z = USV^\top$ be the SVD of Z, with U, V $n \times n$ and orthogonal, and S diagonal. Then the objective of the problem writes

$$\operatorname{trace} RZ = \operatorname{trace} R(USV^\top) = \operatorname{trace}(V^\top RU)S = \operatorname{trace} MS,$$

where M is the new variable $M = V^\top RU$. We note that $R^\top R = I_n$ translates as $M^\top M = I_n$. We have reduced the problem to the case when Z is replaced with the diagonal matrix S:

$$\max_{M} \operatorname{trace} MS \;:\; M^\top M = I_n.$$

Using the fact that $M = I$ is optimal in that case, we obtain $V^\top RU = I_n$ at optimum, which leads to the desired result.

6. The fact that $M = I_n$ is optimal for the above problem stems from the fact that since M is orthogonal, $|M_{ii}| \leq 1$, $i = 1, \dots, n$. Indeed, for every $i = 1, \dots, m$, defining e_i to be the i-th unit vector in $\mathbb{R}^n$:

$$1 = (M^\top M)_{ii} = e_i^\top M^\top M e_i = \|Me_i\|_2^2 = \sum_{k=1}^{n} M_{ki}^2 \geq M_{ii}^2.$$

Thus, with $p \leq m$ the rank of Z, and $\sigma_1, \dots, \sigma_p$ its singular values:

$$\operatorname{trace} MS = \sum_{i=1}^{p} M_{ii}\sigma_i \leq \sum_{i=1}^{p} \sigma_i.$$

The result follows from the fact that the upper bound is attained when $M = I_n$.

7. For the image deformation problem, we find the matrix R and vector r based on the input-output pairs (x_i, y_i), $i = 1, \dots, m$. We then apply the transformation to every point $x \in \mathbb{R}^2$ on the original image (left panel in Fig. 5.6), to obtain the transformed points $y = Rx + r$, and form a new image (right panel in Fig. 5.6).

6. Linear Equations

Exercise 6.1 (Least-squares and total least-squares) Find the least squares line and the total least-squares[12] 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.

Solution 6.1 A perfect line would have

$$y_i = w_1 x_i + w_2, \quad i = 1, 2, 3, 4,$$

for some scalars w_1, w_2.

To find the least-squares line, we solve

$$\min_x \|X^\top w - y\|_2,$$

where $w = (w_1, w_2)$, and

$$X = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}.$$

Since not all coordinates of x are equal, X is full row rank, and hence the LS solution is

$$w_{\text{LS}} = (XX^\top)^{-1} Xy = \begin{bmatrix} 0.4 \\ 0.3 \end{bmatrix}.$$

To find the TLS solution, we have the formula

$$w_{\text{TLS}} = ((XX^\top - \sigma_{\min}^2 I)^{-1} Xy,$$

with $\sigma_{\min}$ the smallest singular value of $[X^\top, y]$. We obtain

$$w_{\text{TLS}} = \begin{bmatrix} 0.4207 \\ 0.3217 \end{bmatrix}.$$

As expected, both coefficients are slightly larger in magnitude.

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 geometrically the result. *Hint:* use the SVD of A. You can assume that $m \geq n$, and that A is full column rank.

[12] See Section 6.7.5.

Solution 6.2 Assuming $m \geq n$ and using the SVD of the full column-rank matrix A:

$$A = U\tilde{\Sigma}V^\top, \ \ \tilde{\Sigma} = \begin{bmatrix} \Sigma \\ 0 \end{bmatrix}, \ \ \Sigma = \mathbf{diag}\left(\sigma_1, \ldots, \sigma_n\right),$$

we can easily reduce the problem to the case when A is diagonal. Precisely, we set $\bar{y} = U^\top y = (\bar{y}_1, \bar{y}_2)$, with $\bar{y}_1 \in \mathbb{R}^n$, and $\bar{y}_2 \neq 0$. The optimal point is $x^* = V\bar{x}^*$, with $\bar{x}^* = \Sigma^{-1}\bar{y}_1$, and the corresponding optimal residual $r^* = y - Ax^*$ satisfies

$$U^\top r^* = U^\top(y - Ax^*) = \bar{y} - \tilde{\Sigma}\bar{x}^* = \begin{bmatrix} 0 \\ \bar{y}_2 \end{bmatrix}.$$

We do have $A^\top r^* = 0$, since

$$VA^\top r^* = \tilde{\Sigma}^\top U^\top r^* = \begin{bmatrix} \Sigma & 0 \end{bmatrix} \begin{bmatrix} 0 \\ y_2 \end{bmatrix} = 0,$$

and

$$y^\top r^* = \bar{y}^\top (U^\top r^*) = \|\bar{y}_2\|_2^2 > 0.$$

The geometrical interpretation is apparent from Figure 6.7. The residual vector $r = y - Ax$ forms an acute angle with the point y. In addition, it is orthogonal to the range of A.

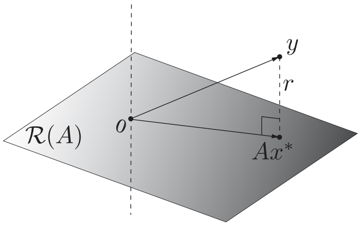

Figure 6.7: Projection onto the range of A.

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 $X^aY = b$, 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 precisely the least-squares problem involved.

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

Solution 6.3

1. If the model was exactly true, we would have $b = x_i^a y_i$, $i = 1, \ldots, m$, or taking logarithms:

$$\log b = a \log x_i + \log y_i, \ \ i = 1, \ldots, m.$$

This leads to a least-squares problem with variables $\theta = (a, \log b)$:

$$\min_\theta \|A\theta - z\|_2,$$

where $z = -(\log y_1, \ldots, \log y_m)$ is the response vector, and

$$A = \begin{bmatrix} \log x_1 & -1 \\ \vdots & \vdots \\ \log x_m & -1 \end{bmatrix}.$$

2. The answer is unique if and only if A is full column rank. This is always the case except when all the x_i's are equal.

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

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

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.

Solution 6.4 Consider a fixed row index i, and drop the dependence of the vectors on that index. Define $r = y - \hat{a}^\top x$. We consider the expected value

$$\begin{aligned} \mathbb{E}_u((\hat{a} + u)^\top x - y)^2 &= \mathbb{E}\left((u^\top x)^2 - 2r(u^\top x) + r^2\right) \\ &= \mathbb{E}\left((u^\top x)^2 + r^2\right) \\ &= \sigma^2\|x\|_2^2 + r^2, \end{aligned}$$

where we have used $\mathbb{E}_u u = 0$ in the second line, and in the third, the expression

$$\mathbb{E}_u (u^\top x)^2 = \mathbb{E}(x^\top u u^\top x) = \mathbb{E}\,\mathrm{trace}(uu^\top)(xx^\top) = \mathrm{trace}\, xx^\top = x^\top x.$$

Summing we obtain the desired result, with regularization parameter $\lambda = \sigma^2$.

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[13].

We are given a full column rank matrix $A \in \mathbb{R}^{m,n}$, with rows $a_i^\top$, $i = 1, \ldots, m$, and 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 $O(n^3)$.

Solution 6.5

1. We have
$$A = \begin{bmatrix} A_- \\ a^\top \end{bmatrix}, \quad y = \begin{bmatrix} y_- \\ \eta \end{bmatrix}$$
where A_-, y_- corresponds to the least-squares problem with the last measurement deleted, $a = a_m$ and $\eta = y_m$ corresponds to the last measurement. With $H \doteq A^\top A$, $H_- \doteq A_-^\top A_-$, the current solution expresses $x^* = H^{-1} A^\top y$, while the solution with the last measurement deleted is $x_-^* = H_-^{-1} A_-^\top y_-$. We would like to express x_-^* in a computationally efficient way, as a modification of x.

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

We have

$$H_- = A_-^\top A_- = AA^\top - aa^\top, \quad A_-^\top y_- = A^\top y - a\eta.$$

Using the rank-one perturbation formula (3.10), we obtain

$$H_-^{-1} = H^{-1} + \frac{1}{\gamma} H^{-1} a a^\top H^{-1}, \quad \gamma \doteq 1 - a^\top H^{-1} a.$$

Note that $\gamma > 0$, since the matrix H_- is full rank.

We then have

$$\begin{aligned} x_-^* = H_-^{-1} A_-^\top y_- &= \left(H^{-1} + \frac{1}{\gamma} H^{-1} a a^\top H^{-1} \right) \left(A^\top y - a\eta \right) \\ &= x^* + \frac{(a^\top x^* - \eta)}{\gamma} H^{-1} a. \end{aligned}$$

Compare with the similar formula (6.15), which was obtained in the context of *adding* a measurement.

2. In a leave-one-out approach, we can first find $K \doteq H^{-1}$, where $H = A^\top A$ corresponds to all the measurements. Then we set $P = KA^\top$, and $x^* = Py$. We then obtain the solution corresponding to the i-th measurement deleted as

$$x_i^* = x^* + \frac{(a_i^\top x^* - y_i)}{\gamma_i} p_i, \quad i = 1, \ldots, m,$$

where $p_i = H^{-1} a_i$ is the i-th column of P, and $\gamma_i = 1 - a^\top p_i$, $i = 1, \ldots, m$.

Let us compare the complexity of the above method, with a more brute-force approach that involves solving each least-squares problem with one measurement deleted. The complexity of solving a least-squares problem with n variables and m measurements is $C(m, n) = O(mn^2 + n^3)$, so solving m such problems is $mC(m, n)$. The above method, in contrast, requires:

- forming $n \times n$ matrix H ($O(n^2 m)$) and its inverse K ($O(n^3)$);
- forming $n \times m$ matrix $P = KA^\top$ ($O(n^2 m)$);
- forming x^* and x_i^*, $i = 1, \ldots, m$ ($O(nm)$).

In total, the complexity is $C(m, n) = O(mn^2 + n^3)$, which is the same as *one* least-squares problem.

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 $1/x$.

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), $i = 1, \dots, m$.

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

Solution 6.6 (Enzyme kinetics model) 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 parameters.

1. We have
$$\frac{1}{y} = w_1 + \frac{w_2}{x}, \quad w_1 = \frac{1}{\beta_1}, \quad w_2 = \frac{\beta_2}{\beta_1}.$$

2. The least-squares problem takes the form
$$\min_w \|X^\top w - z\|_2,$$
where $z = (1/y_1, \dots, 1/y_m)$, and
$$X = \begin{bmatrix} 1/x_1 & \dots & 1/x_m \\ 1 & \dots & 1 \end{bmatrix}.$$
Once w^* is found, we can set $\beta^* = (1/w_1^*, w_2^*/w_1^*)$.

3. According to section (6.5.4), the sensitivity of the solution to an LS problem depends on the condition number of the matrix $A = X^\top$. We can experiment with random data and observe that the condition number can be very high. Fishy.

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 segment. 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, $\bar{q}$, in the l_2-norm sense. The vector $\bar{q}$ 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).

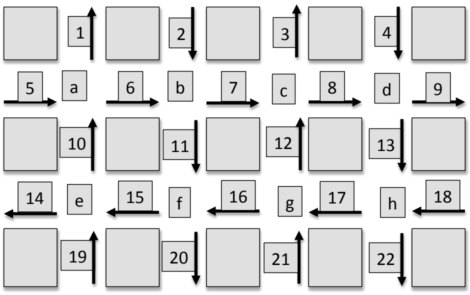

Figure 6.8: Example of traffic estimation problem. The intersections are labeled a to h. The road segments are labeled 1 to 22. The arrows indicate the direction of traffic.

4. Formulate the problem for the small example of Figure 6.8 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 you 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).

Solution 6.7

1. At each intersection, the sum of all incoming flows must be equal to the sum of all outgoing flows. We construct the matrix $A \in \mathbb{R}^{J \times I}$ such that the element on the jth line and ith column is

- 0 if link i does not arrive or leave intersection j;

- 1 if link i arrives at intersection j;
- -1 if link i leaves intersection j.

With such a construction, the conservation of flow is written $Aq = 0$.

2. The problem can be formulated as follows:

$$\min_{q \in \mathbb{R}^I} \quad ||q - \overline{q}||_2 \text{ subject to } Aq = 0.$$

3. To solve this problem we introduce the variable $x = q - \overline{q}$. With this change of variable, we solve the following optimization problem:

$$\min_{x \in \mathbb{R}^I} \quad ||x||_2 \text{ subject to } Ax = b,$$

where $b = -A\overline{q}$. Our estimation problem is a least norm problem. The matrix A is full row rank, that is, the matrix AA^T is invertible. We can solve the minimum norm problem in closed-form $x^* = A^T(AA^T)^{-1}b$. We get the flow estimate as $q^* = x^* + \overline{q}$.

4. We first need a function that constructs the matrix A that will represent the incidence matrix of links at intersections.

```
function A = construct_network
A = zeros(8,22);
for i=1:4
A(i,i) = (-1)^i;
A(i,4+i) = 1;
A(i,5+i) = -1;
A(i,9+i) = -(-1)^i;
end
for i=5:8
A(i,9+i-4) = (-1)^i;
A(i,13+i-4) = 1;
A(i,14+i-4) = -1;
A(i,18+i-4) = -(-1)^i;
end
```

Then, we solve the estimation problem as a least norm problem. The function below loads the historical flows, constructs the incidence matrix and estimates the flow that is closest to the historical averages while satisfying the conservation of vehicles.

```
function x_hat = estimate_historic
```

```
A = construct_network;
hist = load('historical.mat');
hist = hist.historical;

b = zeros(8,1);

z_hat = A\(b-A*hist);
x_hat = z_hat + hist;
```

5. The additional flow measurements are incorporated into the constraints. Let B be the matrix of size N times I where N represents the number of flow measurements. This matrix has a one on row n and column i if the nth measurement concerns link i. It has zeros everywhere else. We call q_m the vector of measurements. The optimization problem reads:

$$\min_{q \in \mathbb{R}^I} \quad ||q - \overline{q}||_2 \text{ subject to } Aq = 0, \;\; Bq = q_m.$$

With the same change of variable as before ($x = q - \overline{q}$), defining $\tilde{A}$ the matrix obtained by stacking A and B vertically, $\tilde{A} = [A^T \; B^T]^T$ and $\tilde{b}$ the vector obtained by stacking b and $q_m - B\overline{q}$, $\tilde{b} = [b^T \; (q_m - B\overline{q})^T]^T$, we solve again a least norm problem:

$$\min_{x \in \mathbb{R}^I} \quad ||x||_2 \text{ subject to } \tilde{A}x = \tilde{b}.$$

In the data provided, the measurements are such that $\tilde{A}$ is still full row rank, and we can solve the problem with the same method as before. In the general case, $\tilde{A}$ may not be full row rank anymore. In this case, two situations may arise:

- If $\tilde{b}$ is not in the range of $\tilde{A}$, then the problem is not feasible (the optimal value is $+\infty$).
- Otherwise, any solution of the linear equation $\tilde{A}x = \tilde{b}$ can be written $x = x^* + z$. Here, z is any element of the nullspace of $\tilde{A}$. Also, x^* is *the* element in the orthogonal of the nullspace which is solution to the linear equation *i.e.* $\tilde{A}x^* = \tilde{b}$ and for all z in the nullspace of $\tilde{A}$, $z^T x^* = 0$.

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 $\lambda \geq 0$ is a parameter. In the above, the variable is a $n \times m$ 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_i's 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 problem. You do not need to be explicit about the form of the problem.

2. Show that
$$\frac{1}{2} \sum_{1 \leq i,j \leq m} \|x_i - x_j\|_2^2 = \operatorname{trace} XHX^\top,$$
where $H = mI_m - \mathbf{1}\mathbf{1}^\top$ is a $m \times m$ 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 semi-definite.

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?

Solution 6.8

1. The objective function is a sum of squares. Inside the squares are linear or affine functions of the variables. There are no constraints. Hence the problem is an ordinary least-squares problem.

2. We have

$$\begin{aligned}\frac{1}{2}\sum_{i,j}\|x_i - x_j\|_2^2 &= \frac{1}{2}\sum_{i,j}\left(\|x_i\|_2^2 + \|x_j\|_2^2 - 2x_i^\top x_j\right)\\ &= m\sum_i \|x_i\|_2^2 - \left(\sum_i x_i\right)^\top\left(\sum_i x_i\right)\\ &= m\,\mathrm{trace}(X^\top X) - (X\mathbf{1})^\top(X\mathbf{1})\\ &= m\,\mathrm{trace}(XX^\top) - \mathrm{trace}(X\mathbf{1})(X\mathbf{1})^\top\\ &= \mathrm{trace}(X(mI_m - \mathbf{1}\mathbf{1}^\top)X^\top),\end{aligned}$$

as claimed.

3. H is positive semi-definite, since for any vector $z \in \mathbb{R}^m$, we have

$$\frac{1}{m}z^\top Hz = z^Tz - \frac{1}{m}(\sum_i z_i)^2 = \sum_i(z_i - \hat{z})^2 \geq 0,$$

where $\hat{z} = (1/m)(z_1 + \ldots + z_m)$.

4. The objective function writes

$$F(X) = \|X - P\|_F^2 + \lambda\,\mathrm{trace}(XHX^\top).$$

The gradient of the first term is $X - P$. For the second term, we start by looking at the following first-order expansion: for any $\Delta \in \mathbb{R}^{n,m}$, we have

$$\begin{aligned}&\mathrm{trace}((X+\Delta)H(X+\Delta)^\top) - \mathrm{trace}(XHX^\top)\\ &\quad= 2\,\mathrm{trace}(\Delta HX^\top) + \text{h.o.t.}\\ &\quad= 2\,\mathrm{trace}(HX^\top\Delta) + \text{h.o.t.}\\ &\quad= 2\,\mathrm{trace}((XH)^\top\Delta) + \text{h.o.t.},\end{aligned}$$

which shows that the gradient of the second term is $2\lambda XH$. The formula for $\nabla F(X)$ follows.

5. Optimality conditions are $(1/2)\nabla F(X) = X - P + \lambda XH = 0$, or $X(I + \lambda H) = P$. Since H is positive semi-definite, the matrix $I + \lambda H$ is invertible whenever $\lambda \geq 0$. In fact, we have

$$\begin{aligned}(I+\lambda H)^{-1} &= \left((1+m\lambda)I - \lambda\mathbf{1}\mathbf{1}^\top\right)^{-1}\\ &= \frac{1}{1+m\lambda}\left(I - \theta\mathbf{1}\mathbf{1}^\top\right)^{-1} \text{ (with } \theta \doteq \lambda/(1+m\lambda))\\ &= \frac{1}{1+m\lambda}\left(I + \frac{\theta}{1-\theta\mathbf{1}^\top\mathbf{1}}\mathbf{1}\mathbf{1}^\top\right) \text{ (from formula 3.10)}\\ &= \frac{1}{1+m\lambda}I + \theta\mathbf{1}\mathbf{1}^\top.\end{aligned}$$

The formula for the optimal points then follows from the identity $P\mathbf{1} = m\hat{p}$.

6. The model will simply shrink the points towards their averages. This is not a desirable effect; a good clustering algorithm would group points in several different clusters, rather than sending them all towards the global mean $\hat{p}$.

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 $\gamma(A) \doteq d(A)/s(A)$, where $d(A)$ 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.

Solution 7.1

1. We have to form m scalar products between the m rows of A and the n-vector x. The total number of flops is then $s(A) = mn$. When A is sparse, the number is $d(A)$.

2. We have $Ax = y + z$, with $y = \tilde{A}x$ and $z = (v^\top x)u$. The cost of forming $v^\top x$ is n flops; forming $(v^\top x)u$ is m flops. Thus, the total cost required is $d(\tilde{A}) + m + n$, which is much less than $s(A)$.

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:

$$\tilde{a} \doteq r^\top a = a_i, \quad \tilde{b} \doteq r^\top b = b_i.$$

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

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

(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 $\text{Prob}\{r = e_i\} = 1/n$, $i = 1, \dots, n$, 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 $r^2 \doteq [r_1^2, \dots, r_n^2]^\top$ 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.

Solution 7.2 We have that

$$\tilde{a}\tilde{b} = r^\top a r^\top b = a^\top r r^\top b,$$

hence

$$\mathbb{E}\{\tilde{a}\tilde{b}\} = a^\top \mathbb{E}\{rr^\top\} b = a^\top \left(\sum_{i=1}^{n} \frac{1}{n} e_i e_i^\top\right) b = \frac{1}{n} a^\top b,$$

which proves the first statement. Next, observe that, since $r_i r_j = 0$ with probability one for $i \neq j$, then

$$\begin{aligned} \tilde{a}\tilde{b} &= a^\top r r^\top b = \sum_{i=1}^{n} r_i^2 a_i b_i, \quad \text{w.p. } 1 \\ &= \sum_{i=1}^{n} r_i a_i b_i, \quad \text{w.p. } 1, \end{aligned}$$

and

$$\begin{aligned} (\tilde{a}\tilde{b})^2 &= \left(\sum_{i=1}^{n} r_i a_i b_i\right)^2 = \sum_{i=1}^{n} r_i^2 a_i^2 b_i^2, \quad \text{w.p. } 1 \\ &= \sum_{i=1}^{n} r_i a_i^2 b_i^2, \quad \text{w.p. } 1. \end{aligned}$$

Thus,

$$\begin{aligned} \text{var}\{\tilde{a}\tilde{b}\} &= \mathbb{E}\{(\tilde{a}\tilde{b} - \mathbb{E}\{\tilde{a}\tilde{b}\})^2\} = \mathbb{E}\{(\tilde{a}\tilde{b} - \frac{1}{n} a^\top b)^2\} \\ &= \mathbb{E}\{(\tilde{a}\tilde{b})^2\} - \frac{2}{n}(a^\top b)\mathbb{E}\{\tilde{a}\tilde{b}\}\} + \frac{1}{n^2}(a^\top b)^2 \\ &= \sum_{i=1}^{n} \mathbb{E}\{r_i\} a_i^2 b_i^2 - \frac{2}{n}(a^\top b)\frac{1}{n}(a^\top b) + \frac{1}{n^2}(a^\top b)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} a_i^2 b_i^2 - \frac{1}{n^2}(a^\top b)^2, \end{aligned}$$

which proves the second statement (recalling that $\mathrm{var}\{\alpha x\} = \alpha^2 \mathrm{var}\{x\}$).

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 barycenter 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.

Solution 7.3 We have

$$\tilde{X} = X - \bar{x}\mathbf{1}^\top.$$

The power iteration updates take the form

$$\begin{aligned} u(k+1) &= \frac{\tilde{X}v(k)}{\|\tilde{X}v(k)\|_2}, \\ v(k+1) &= \frac{\tilde{X}^\top u(k+1)}{\|\tilde{X}^\top u(k+1)\|_2}. \end{aligned}$$

Consider the problem of a matrix-vector multiplication $u = \tilde{X}v$: we can write it as

$$y = \mathbf{1}^\top v, \;\; u = Xv - \bar{x}y,$$

where the opetations only involve a simple scalar product and a matrix-vector product Xv that involves the sparse matrix X. A similar result holds with the transpose operation $v = \tilde{X}^\top u$, which can be written as

$$z = \bar{x}^\top u, \;\; v = Xu - \mathbf{1}z.$$

The normalizations involved are also simple. We write the power iteration method as

$$\begin{aligned} y(k) &= \mathbf{1}^\top v(k), \\ \tilde{u}(k+1) &= Xv(k) - \bar{x}y(k), \\ u(k+1) &= \frac{\tilde{u}_{(}k+1)}{\|\tilde{u}(k+1))\|_2}, \\ z(k+1) &= \bar{x}^\top u(k+1), \\ \tilde{v}(k+1) &= Xu(k+1) - \mathbf{1}^\top y(k), \\ v(k+1) &= \frac{\tilde{v}(k+1)}{\|\tilde{v}(k+1)\|_2}. \end{aligned}$$

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?

Solution 7.4

1. The count is $O(n^3)$. This can be seen from the flop count for the QR algorithm[14].

2. We can write the system as

   $$Dx + uz = y, \quad z = v^\top x.$$

   Assuming D is invertible, we can solve for x:

   $$x = D^{-1}(y - uz).$$

   Then

   $$z = v^\top x = v^\top D^{-1}(y - uz) \Longrightarrow z = \frac{v^\top D^{-1} y}{1 + v^\top D^{-1} u}.$$

   The flop count in this case is $O(n)$.

3. When A is upper-triangular, we can use the backward substitution algorithm[15]. The flop count is then $O(n^2)$, as explained in Remark 7.1.

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

$$Ax = b$$

[14] See section 7.3.1.

[15] See Section 6.

consists in decomposing A as a sum: $A = D + R$, where $D = \text{diag}(a_{11}, \ldots, a_{nn})$, 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.

Solution 7.5

1. Assume that $M = -D^{-1}R$ is diagonalizable: $M = V^{-1}EV$, with E a diagonal matrix containing the eigenvalues $\lambda_1, \ldots, \lambda_n$ on the diagonal, and V a matrix with columns equal to corresponding eigenvectors. Defining $z_0 = D^{-1}b$, the recursion writes

   $$x_{k+1} = Mx_k + z_0, \ \ k = 0, 1, 2, \ldots$$

   This can be written in terms of the vectors $\bar{z} = Vz_0$, and $\bar{x}_k = Vx_k$, as

   $$\bar{x}_k = E\bar{x}_k + \bar{z}, \ \ k = 0, 1, 2, \ldots$$

   The above is a set of n independent scalar recursions of the form

   $$\xi_{k+1} = \lambda_i \xi_k + \bar{z}_i, \ \ k = 0, 1, 2, \ldots,$$

   the convergence of which is guaranteed for any initial point when $|\lambda_i| < 1$. Our condition thus writes $\rho(D^{-1}R) < 1$, where ρ is the modulus of the largest eigenvalue of its matrix argument.

2. Let λ, v be an eigenvalue-eigenvector pair for M: $\lambda Dv = Rv$, with $v \neq 0$. Let i be such that $|v_i| = \max_j |v_j| = \|v\|_\infty \neq 0$. We have

   $$|\lambda||a_{ii}||v_i| = \left|\sum_{j \neq i} a_{ij} v_j\right| \leq \sum_{j \neq i} |a_{ij}||v_j| \leq \|v\|_\infty \sum_{j \neq i} |a_{ij}|.$$

Dividing by $|a_{ii}||v_i|$ (a non-zero quantity since $|a_{ii}| > 0$):

$$|\lambda| \leq \frac{\sum_{j \neq i} |a_{ij}|}{|a_{ii}|} < 1,$$

which proves that any eigenvalue of M has modulus strictly less than one.

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 exist $\bar{x} \in \mathbb{R}^n$ 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).

Solution 7.6 (Convergence of linear iterations)

1. Applying (7.4) iteratively we obtain that

$$x(k) = F^k x_0 + \left(\sum_{i=0}^{k-1} F^i\right) c, \quad k = 1, 2, \ldots$$

hence, for $\bar{x}$ satisfying (7.4),

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

which proves the first expression. Also, since $\bar{x}$ is a stationary point, it holds for all $k \geq 1$ that

$$\bar{x} = F^k \bar{x} + \left(\sum_{i=0}^{k-1} F^i \right) c,$$

thus

$$x(k) - \bar{x} = F^k x_0 + \left(\sum_{i=0}^{k-1} F^i \right) c - F^k \bar{x} - \left(\sum_{i=0}^{k-1} F^i \right) c = F^k (x_0 - \bar{x}),$$

which proves the second expression.

2. Suppose first that, for any x_0, $x(k)$ converges[16] for $k \to \infty$. Then,

$$0 = \lim_{k\to\infty} x(k+1) - x(k) = \lim_{k\to\infty} F^k (I - F)(\bar{x} - x_0),$$

thus $\lim_{k\to\infty} F^k (I - F)(\bar{x} - x_0) = 0$ for all x_0, which happens if and only if $\lim_{k\to\infty} F^k (I - F) = 0$, which implies (by Theorem 3.5) that F^k is convergent.

Conversely, suppose that F^k is convergent. Then, $\lim_{k\to\infty} F^k = \bar{F}$, and

$$\lim_{k\to\infty} x(k) - \bar{x} = \lim_{k\to\infty} F^k (x_0 - \bar{x}) = \bar{F}(x_0 - \bar{x}),$$

thus $x(k)$ converges to $\hat{x} = \bar{x} + \bar{F}(x_0 - \bar{x})$. This point must a fortiori satisfy the stationarity conditions, that is $\hat{x} = F\hat{x} + c$. This property can also be verified directly, since

$$\begin{aligned} F\hat{x} + c &= F\bar{x} + F\bar{F}(x_0 - \bar{x}) + c = \bar{x} + F\bar{F}(x_0 - \bar{x}) \\ &= \bar{x} + \bar{F}(x_0 - \bar{x}) = \hat{x}, \end{aligned}$$

where we have used the fact[17] that $F\bar{F} = \bar{F}$.

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).

[16] The point of convergence, $\bar{x}(x_0)$ may depend on x_0. In any case, since, in the limit, $x(k+1) = x(k) = \bar{x}(x_0)$, we see that the convergence point must satisfy the stationarity equations (7.4).

[17] Any convergent matrix F can be factored as $F = U \mathrm{diag}\,(I, \tilde{F})\, U^{-1}$, where $U = [U_1\, \tilde{U}]$ and U_1 contains by columns the eigenvectors of F associated to the eigenvalue $\lambda = 1$ (if F has such eigenvalue). Moreover, $\rho(\tilde{F}) < 1$. It follows that the limit matrix is $\bar{F} = U \mathrm{diag}\,(I, 0)\, U^{-1}$, and one can verify directly that $F\bar{F} = \bar{F}F = \bar{F}$.

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 rank $\tilde{A} = n$. *Hint:* apply the algorithm to the Normal equations.

Solution 7.7 (A linear iterative algorithm)

1. We recall[18] preliminary that $AA^{\dagger}$ is an orthogonal projector onto $\mathcal{R}(A)$, therefore, $AA^{\dagger}b \in \mathcal{R}(A)$, which means that (7.8) always admits a solution.

   Suppose next that x satisfies (7.7). Then, multiplying this equation by $A^{\dagger}$ on the left, we have

$$A^{\dagger}Ax = A^{\dagger}b,$$

and multiplying it once again on the left by A we obtain

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

which proves that (7.8) is satisfied by x, since $AA^{\dagger}A = A$.

[18] See Section 5.2.3

Finally, let $b \in \mathcal{R}(A)$ and suppose (7.8) is satisfied for some x. We write $b = AA^{\dagger}z$ for some z, hence

$$Ax = AA^{\dagger}b = AA^{\dagger}AA^{\dagger}z = AA^{\dagger}z = b,$$

which shows that (7.7) is satisfied at x.

2. Recall that, for any matrix X, $\mathcal{R}(X) = \mathcal{R}(XX^{\dagger})$, and $\mathcal{N}(X)^{\perp} = \mathcal{R}(X^{\dagger}X)$. Then, by the fundamental theorem of linear algebra, we have that

$$\begin{aligned}\mathbb{R}^n &= \mathcal{N}(RA) \oplus \mathcal{N}(RA)^{\perp} = \mathcal{N}(RA) \oplus \mathcal{R}((RA)^{\dagger}(RA)) \\ &= \mathcal{N}(A) \oplus \mathcal{R}((RA)^{\dagger}(RA)).\end{aligned}$$

Therefore, any $x \in \mathbb{R}^n$ can be written as $x = x_1 + x_2$, with $x_1 \in \mathcal{R}((RA)^{\dagger}(RA))$ and $x_2 \in \mathcal{N}(A)$. Hence, we can write $x_1 = (RA)^{\dagger}(RA)z$ for some z, and

$$\begin{aligned}A(RA)^{\dagger}(RA)x &= A(RA)^{\dagger}(RA)x_1 = A(RA)^{\dagger}(RA)(RA)^{\dagger}(RA)z \\ &= A(RA)^{\dagger}(RA)z = Ax_1 = Ax,\end{aligned}$$

which shows that $A((RA)^{\dagger}R)A = A$, thus $(RA)^{\dagger}R$ is a pseudoinverse of A.

3. Any solution of (7.9) is of the form

$$x = \bar{x} + z, \quad \bar{x} = (RA)^{\dagger}Rb,\ z \in \mathcal{N}(RA) = \mathcal{N}(A),$$

where $\bar{x}$ is a solution, since it is assumed that $Rb \in \mathcal{R}(RA)$. Then,

$$Ax = A\bar{x} + Az = A\bar{x} = A(RA)^{\dagger}Rb = AA^{\dagger}b,$$

which shows that x satisfies (7.8). Conversely, any solution of (7.8) is of the form

$$x = \bar{x} + z, \quad \bar{x} = A^{\dagger}b,\ z \in \mathcal{N}(A),$$

Then,

$$RAx = RAA^{\dagger}b + RAz = RAA^{\dagger}b = (RA)(RA)^{\dagger}Rb = Rb,$$

where the last passage follows from the fact that $(RA)(RA)^{\dagger}$ is a projector onto $\mathcal{R}(RA)$, but $Rb \in \mathcal{R}(RA)$ by assumption, hence the projection leaves this vector unchanged.

4. The solution to this point is easily obtained by applying the statements of Exercise 7.6, considering that the iterations have the form (7.3), for $F = I - \alpha RA$ and $c = \alpha Rb$. The iterations converge if and only if F^k is convergent. A sufficient condition for this is given by $\rho(I - \alpha RA) < 1$.

5. Let
$$R \doteq \operatorname{diag}\left(\frac{1}{\|a_1^\top\|_1}, \ldots, \frac{1}{\|a_n^\top\|_1}\right),$$
where $a_i^\top$ are the rows of A. Since $A \succ 0$ and $R \succ 0$, this choice of R satisfies the conditions of point 3. Next, applying the results in points 2. and 3. of Exercise 4.8, we have that $\rho(I - \alpha RA) < 1$ for any $\alpha \in (0, 2)$ hence, for such values of α, $(I - \alpha RA)^k$ converges to zero and the iterations (7.10) converge to the (unique) solution of (7.9).

6. All solutions of $\tilde{A}x = \tilde{b}$ are also solutions to the Normal equations $Ax = b$, with $A = \tilde{A}^\top \tilde{A}$, $b = \tilde{A}^\top \tilde{b}$. If $\tilde{A} \in \mathbb{R}^{m,n}$ with $m \geq n$ and rank $\tilde{A} = n$, then $A \in \mathbb{S}^n$ is positive definite, and we can apply the results of the previous point to the system of 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 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?

Solution 8.1

1. The set is not convex: $(0, 1)$ and $(-2, 0)$ both belong to the set, but the midpoint $(-1, 1/2)$ does not.
2. Define $q(x) = x_2(x_2 - x_1 - 1)$; we have $q(x) \leq 0$ if and only if $(x_2 - x_1 - 1) \leq 0$ and $x_2 \geq 0$, or $(x_2 - x_1 - 1) \geq 0$ and $x_2 \leq 0$. We can write
$$q(x) = x^\top Ax + 2b^\top x + c,$$
where
$$A = \begin{bmatrix} 0 & -1/2 \\ -1/2 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ -1/2 \end{bmatrix}, \quad c = 0.$$
3. The convex hull of the set is the whole space, $\mathbb{R}^2$.

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.

Solution 8.2 (Closed functions and sets) The indicator function of a convex set $\mathcal{X}$ is an extended-valued function defined as

$$I_\mathcal{X}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{X} \\ +\infty & \text{otherwise.} \end{cases}$$

Clearly, the domain of this function is $\mathcal{X}$, which is a convex set. Hence, for any $x, y \in \mathcal{X}$ and $\lambda \in [0, 1]$, we have that

$$I_\mathcal{X}(\lambda x + (1 - \lambda)y) = 0 \leq \lambda I_\mathcal{X}(x) + (1 - \lambda)I_\mathcal{X}(y) = 0, \qquad (8.11)$$

which shows that $I_\mathcal{X}$ is convex.

Suppose next that $\mathcal{X}$ is a closed set. Let $\alpha \in \mathbb{R}$. If $\alpha \geq 0$, the sublevel set $S_\alpha = \{x \in \mathbb{R}^n : I_\mathcal{X}(x) \leq \alpha\}$ coincides with $\mathcal{X}$, which is a closed set. If $\alpha < 0$, then $S_\alpha = \emptyset$, which is also closed. Thus, all sublevel sets of $I_\mathcal{X}$ are closed, which proves that $I_\mathcal{X}$ is a closed function.

Exercise 8.3 (Convexity of functions)

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

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

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 $\kappa \mathrm{diag}\,(y) - zz^\top$ for appropriate vectors $y \geq 0$, $z \geq 0$, and scalar $\kappa \geq 0$, and use Schur complements[19] to prove that the Hessian is positive semi-definite.

Solution 8.3

1. For x, y both positive scalars, the function $g : \alpha \to \alpha(x+y) - y\alpha \log \alpha$, with domain $\mathbb{R}_{++}$, is convex, and differentiable. Its minimum is obtained by setting the derivative to zero, which leads to

$$x + y = y(\log \alpha + 1) \Longrightarrow \alpha = e^{x/y}.$$

The minimizer is in the interior of the domain. Plugging its value in the expression of the function we obtain the desired result. That result proves that f is convex, since it is the point-wise maximum (over $\alpha > 0$) of affine functions.

2. Let $V \doteq \sum_j v_j^{1/r}$, so that $f(v) = V^r$. We have

$$\frac{\partial V}{\partial v_i}(v) = \frac{1}{r} v_i^{\frac{1}{r}-1}, \quad i = 1, \ldots, n,$$

and
$$\frac{\partial f}{\partial v_i}(v) = rV^{r-1} \frac{\partial V}{\partial v_i}(v) = V^{r-1} v_i^{\frac{1}{r}-1}, \quad i = 1, \ldots, n.$$

[19] See Section 4.4.7.

Then, for $j \neq i$:

$$\frac{\partial^2 f}{\partial v_i \partial v_j}(v) = (r-1)V^{r-2}\frac{\partial V}{\partial v_j}(v)v_i^{\frac{1}{r}-1} = \frac{r-1}{r}V^{r-2}v_i^{\frac{1}{r}-1}v_j^{\frac{1}{r}-1},$$

while when $i = j$, there is an additional term:

$$\begin{aligned}\frac{\partial^2 f}{\partial v_i^2}(v) &= \frac{r-1}{r}V^{r-2}v_i^{\frac{2}{r}-2} + (\frac{1}{r}-1)V^{r-1}v_i^{\frac{1}{r}-2} \\ &= \frac{r-1}{r}V^{r-2}\left(v_i^{\frac{2}{r}-2} - V \cdot v_i^{\frac{1}{r}-2}\right)\end{aligned}$$

We can write the Hessian as

$$\nabla^2 f(v) = (r-1)V^{r-2}\left(zz^\top - V \cdot \operatorname{diag}(y)\right),$$

where $z = (v_1^{\frac{1}{r}-1}, \ldots, v_n^{\frac{1}{r}-1})$, and $y = (v_1^{\frac{1}{r}-2}, \ldots, v_n^{\frac{1}{r}-2})$.

Due to Schur complements, and from $\operatorname{diag}(y) \succ 0$, the condition $\nabla^2 f(v) \preceq 0$ is equivalent to

$$V \geq z^\top \operatorname{diag}(y)^{-1} z = \sum_{i=1}^{n} \frac{z_i^2}{y_i} = \sum_{i=1}^{n} v_i^{\frac{1}{r}} = V,$$

which holds.

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,

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

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

$$\|z\|_1 = \min_{d>0} \frac{1}{2}\sum_{i=1}^{m}\left(d_i + \frac{z_i^2}{d_i}\right). \tag{8.12}$$

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

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

Solution 8.4

1. The result obtains with $d \to +\infty$ when $\alpha \leq 0$. When $\alpha > 0$, we take the derivative with respect to d, and obtain a unique minimizer in the domain $d^* = \beta/\alpha > 0$; the result then follows.

2. The result is a direct consequence of the previous part.

3. We can apply strong duality to the problem, since it is convex and strictly feasible. Dualizing the equality constraint, and using the first part:

$$\begin{aligned} p^* &= \min_{d>0} \max_{\nu} \sum_{i=1}^{m} \frac{z_i^2}{d_i} + \nu(\sum_{i=1}^{m} d_i - 1) \\ &= \max_{\nu} \min_{d>0} \sum_{i=1}^{m} \frac{z_i^2}{d_i} + \nu(\sum_{i=1}^{m} d_i - 1) \\ &= \max_{\nu>0} -\nu + 2\sqrt{\nu}\left(\sum_{i=1}^{n} |z_i|\right) \\ &= \|z\|_1^2. \end{aligned}$$

The result shows that the function g_β is concave, as the point-wise minimum of affine functions.

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

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

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

Solution 8.5 Let us consider the equivalent problem

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

Since the objective is strictly decreasing over $x \geq 0$ and $\mathbf{1}^\top x$ is non-decreasing over $x \geq 0$, we can replace the equality constraint by an inequality one, thus we consider the problem

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

The partial Lagrangian for this problem is

$$\begin{aligned}\mathcal{L}(x,\mu) &= \sum_{i=1}^{n} \alpha_i \ln 1/x_i + \mu(\mathbf{1}^\top x - c) \\ &= \sum_{i=1}^{n} (\alpha_i \ln 1/x_i + \mu x_i) - \mu c,\end{aligned}$$

and, for $\mu \geq 0$,

$$\begin{aligned}g(\mu) &= \min_{x \geq 0} \mathcal{L}(x,\mu) = -\mu c + \sum_{i=1}^{n} \min_{x_i \geq 0} (\alpha_i \ln 1/x_i + \mu x_i) \\ &= -\mu c + \sum_{i=1}^{n} (\alpha_i \ln(\mu/\alpha_i) + \alpha_i) \\ &= -\mu c + \ln \mu \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \alpha_i(1 - \ln \alpha_i),\end{aligned}$$

the minimium with respect to $x_i \geq 0$ in the previous expression being attained at the unique point $x_i = \alpha_i/\mu \geq 0$. The dual is thus $d^* = \max_{\mu \geq 0} g(\mu)$, and strong duality holds, since the primal problem is strictly feasible. The optimal dual solution is easily obtained as

$$\mu^* = \frac{\sum_{i=1}^{n} \alpha_i}{c},$$

from which we obtain the optimal primal solution as

$$x_i^* = \frac{\alpha_i}{\mu^*} = c\frac{\alpha_i}{\sum_{i=1}^{n} \alpha_i}, \quad i = 1, \ldots, n.$$

The expression for the optimal objective value follows by substituting this optimal solution back into the objective.

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 hypothesis, if $p_1^* = p_2^*$, then it holds that

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

Solution 8.6 (Monotonicity and locality)

1. The first point is obvious, since any optimal point of the first problem is feasible for the second problem.

2. Clearly, $\mathcal{X}_1 \subseteq \mathcal{X}_2$ implies that $\mathcal{X}_1 \cap \mathcal{X}_3 \subseteq \mathcal{X}_2 \cap \mathcal{X}_3$, therefore, by the first point, it must be $p_2^* = p_1^* = p_{13}^* \geq p_{23}^*$. But $\mathcal{X}_2 \cap \mathcal{X}_3 \subseteq \mathcal{X}_2$ implies that $p_{23}^* \geq p_2^* = p_{13}^*$ (the last equality follows since we are assuming that $p_1^* = p_2^*$ and $p_{13}^* = p_1^*$). It must therefore be $p_{23}^* = p_{13}^*$, as desired.

3. Call x_1^*, x_{13}^*, x_2^*, x_{23}^*, the unique points attaining p_1^*, p_{13}^*, p_2^*, and p_{23}^*, respectively. Since $p_1^* = p_2^*$, uniqueness of the optimal solutions implies that it must be $x_1^* = x_2^*$ (for if $x_1^* \neq x_2^*$, then p_2^* would be attained at two optimal solutions). Also, $p_{23}^* = p_2^*$ implies that $x_{23}^* = x_2^*$, hence $x_1^* = x_2^* = x_{23}^*$. This means in particular that x_1^* belongs to $\mathcal{X}_1 \cap \mathcal{X}_3$, thus

$$p_1^* = f_0(x_1^*) \geq f_0(x_{13}^*) = p_{13}^*.$$

On the other hand, $\mathcal{X}_1 \cap \mathcal{X}_3 \subseteq \mathcal{X}_2 \cap \mathcal{X}_3$ implies that $p_{13}^* \geq p_{23}^* = p_2^* = p_1^*$, hence it must hold that $p_1^* = p_{13}^*$, as desired.

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 of 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

$$\max_{v:\, \|v\|_q \leq 1} u^\top v = \|u\|_p.$$

Solution 8.7

1. For every u, the function $X \to \|X^\top u\|_p$ is convex, hence ϕ_p is. Further, ϕ_p is positively homogeneous, since $\phi_p(\alpha X) = \alpha \phi_p(X)$. Together with convexity this ensures that ϕ_p satisfies the triangle inequality. Finally, if $\phi_p(X) = 0$ then for every u, $u \neq 0$, $\|X^\top u\|_p = 0$, which implies $X^\top u = 0$ for every u; thus, $X = 0$.

2. When $p = 2$, we have

$$\begin{aligned} \phi_2(X)^2 &= \max_{u\,:\,u^\top u=1} u^\top X^\top X u \\ &= \lambda_{\max}(XX^\top) \\ &= \sigma_{\max}(X)^2, \end{aligned}$$

where $\sigma_{\max}$ denotes the largest singular value. The above is due to the variational characterization of eigenvalues, see Theorem 4.3. We obtain that $\phi_2(X) = \sigma_{\max}(X)$. An optimal u is given by a right singular vector corresponding to the largest singular value (see Corollary 5.1).

3. For $p = +\infty$, we have

$$\begin{aligned} \phi_p(X) &= \max_{u\,:\,u^\top u=1} \max_{1 \leq i \leq m} |x_i^\top u| \\ &= \max_{1 \leq i \leq m} \max_{u\,:\,u^\top u=1} |x_i^\top u| \\ &= \max_{1 \leq i \leq m} \|x_i\|_2. \end{aligned}$$

4. We have, using the hint given:

$$\begin{aligned} \phi_p(X) &= \max_{u\,:\,u^\top u=1} \max_{v\,:\,\|v\|_q \leq 1} v^\top X^\top u \\ &= \max_{v\,:\,\|v\|_q \leq 1} \max_{u\,:\,u^\top u=1} v^\top X^\top u \\ &= \max_{v\,:\,\|v\|_q \leq 1} \|Xv\|_2, \end{aligned}$$

as claimed.

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.

Solution 8.8

1. It suffices to prove the result for $n = 1$. Using the change of variable $v_j = \sqrt{X_{1j}} u_j$, $j = 1, \dots, m$, we can further assume $X_{1j} = 1$ for every $j = 1, \dots, m$. We are led to prove convexity of the function $f : \mathbb{R}^m_+ \to \mathbb{R}$, with values
$$f(v) = -\left(\sum_{j=1}^{m} v_j^{1/p} \right)^r .$$
The convexity condition is obvious if $p \geq r = 1$. Let us assume $r > 1$.

   The function f is twice differentiable, so we can use the Hessian condition. Define
$$S = \sum_{j=1}^{m} v_j^{1/p}.$$
We have
$$\frac{\partial f}{\partial v_i}(u) = -\frac{r}{p} S^{r-1} v_i^{1/p-1}, \quad i = 1, \dots, m.$$
Further, denoting by δ_{ij} the Kronecker symbol (equal to 1 if $i = j$, 0 otherwise)
$$\frac{\partial^2 f}{\partial u_i^2} = \frac{r(p-1)}{p^2} S^{r-1} v_i^{1/p-1} \delta_{ij} - \frac{r(r-1)}{p^2} S^{r-2} v_i^{1/p-1} v_j^{1/p-1}, \quad i, j = 1, \dots, m.$$
We can write the Hessian as
$$\nabla^2 f(u) = \frac{r(p-1)}{p^2} S^{r-2} \left(\operatorname{diag}(y) - zz^\top \right),$$
where $y_i = \frac{S(p-1)}{r-1} v_i^{1/p-2}$, $z_i = v_i^{1/p-1}$, $i = 1, \dots, m$.

   Noting that $y_i = 0$ implies $z_i = 0$, we can reduce the problem to the case when $y > 0$ (component-wise). Precisely, we can always re-order the variables so that the zero elements in y, z come first. We then have
$$\operatorname{diag}(y) - zz^\top = \begin{pmatrix} 0 & 0 \\ 0 & \operatorname{diag}(\tilde{y}) - \tilde{z}\tilde{z}^\top \end{pmatrix},$$

where $\tilde{y}, \tilde{z}$ contains the non-zero components of y, z. This proves that it suffices to show the result when y, z are positive componentwise.

Using Schur complements, the condition $\mathbf{diag}(y) \succeq zz^\top$ becomes equivalent to

$$1 \geq \sum_{i=1}^{m} \frac{z_i^2}{y_i} = \frac{r-1}{S(p-1)} \left(\sum_{i=1}^{m} v_i^{1/p} \right) = \frac{r-1}{p-1},$$

The above holds, due to $p \geq r > 1$.

2. With the change of variable $u_j = v_j^p$, $j = 1, \ldots, m$, we obtain the convex formulation

$$\max_u \; f_X(u) \;:\; u \geq 0, \; \sum_{i=1}^{m} u_j \leq 1.$$

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).

Solution 8.9

1. The problem is not convex. Indeed, assume that $m = n = 1$, and set $a_1 = 1$. The objective function is the function f with values

$$f(x) = (|x| - 1)^2,$$

which is not convex, since

$$f(0) = 1 > \frac{1}{2}(f(1) + f(-1)) = 0.$$

2. The fundamental theorem of linear algebra[20] states that any vector $x \in \mathbb{R}^n$ can be written as $x = Ay + z$, for some y, z with $A^\top z = 0$. We observe that the variable z does not play any role in the objective; in other words, we can restrict our search for x to those

[20] See Section 3.2.4.

that lie in the span of A. That is, the objective can be written as a function of y only:

$$\sum_{i=1}^{m} \left(|a_i^\top (Ay)| - 1 \right)^2 = \sum_{i=1}^{m} \left(|k_i^\top y| - 1 \right)^2,$$

where $k_i \doteq A^\top a_i$ is the i-th column of $A^\top A$.

Exercise 8.10 (Eigenvalues and optimization)

Given an $n \times n$ symmetric matrix Q, define

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

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

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

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

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 $W_k^\top QW_k$, where $W_k = [w_{k+1}, \ldots, w_n]$.

Solution 8.10

1. The result follows from the basic fact[21] that the optimization problems yielding optimal values μ_k all have the same objective, while their feasible sets are nested (included in each other).
2. By construction the n vectors w_i, $i = 1, \ldots, n$ are unit-norm and mutually orthogonal, hence they form an orthonormal basis.
3. We have from the variational Theorem 4.3 that μ_1 is the smallest eigenvalue of Q, and that this optimal value is attained for w_1 equal to the corresponding normalized eigenvector of Q. The Lagrangian of problem $\min_{x:x^\top x=1} x^\top Qx$ is

$$\mathcal{L}(x, \mu) = x^\top Qx + \mu(1 - x^\top x) = x^\top (Q - \mu I)x + \mu$$

[21] See point 1. in Exercise 8.6.

and the dual function is[22]

$$g(\mu) = \inf_x \mathcal{L}(x,\mu) = \begin{cases} \mu & \text{if } Q - \mu I \succeq 0 \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem amounts thus to maximizing $g(\mu) = \mu$ subject to $Q - \mu I \succeq 0$, which means that $\mu \leq \min_{i=1,\ldots,n} \lambda_i(Q)$ (here, $\lambda_i(Q)$ are the eigenvalues of Q, arranged in the usual decreasing order), hence the optimal μ is $\mu^* = \lambda_n(Q) = \mu_1$. This shows that strong duality holds between the primal and dual problems, and that μ_1 can be interpreted as the optimal dual multiplier.

4. For $k \in \{1,\ldots,n-1\}$, consider the problem defining w_{k+1}, μ_{k+1}:

$$\mu_{k+1} = \min_{x\,:\,\|x\|_2=1} x^\top Qx \;:\; (w_i^\top x)^2 = 0,\;\; 1 \leq i \leq k.$$

Here, x is constrained to belong to the subspace $\mathcal{V}$ orthogonal to $w_1,\ldots,w_k$ (that is, since $w_1,\ldots,w_n$ form an orthonormal set, $\mathcal{V} = \text{span}(w_{k+1},\ldots,w_n)$) and, by the fundamental theorem of linear algebra, this subspace has dimension $n-k$. From the minimax principle stated in Corollary 4.1, it follows that $\mu_{k+1} = \lambda_{n-k}(Q)$, i.e., it is equal to the $(n-k)$-largest eigenvalue of Q. Thus, μ_{k+1} is the $(k+1)$-smallest eigenvalue of Q, and this optimal value is attained for $x = w_{k+1}$. Since $x \in \mathcal{V}$, we let $x = W_k z$, and rewrite the problem in the variable z

$$\mu_{k+1} = \min_{z\,:\,\|z\|_2=1} z^\top W_k Q W_k z,$$

where $W_k Q W_k = \text{diag}\,(\mu_{k+1},\ldots,\mu_n)$. Applying the result of the previous point, we have that strong duality holds for this problem, and that μ_{k+1} is the optimal Lagrange multiplier.

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 $x_i \in \mathbb{R}^{n_i}$, $n_1 + \ldots + n_p = n$. 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

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

[22] Use the result for the minimum of a quadratic function, given in Section 9.1.

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.

Solution 8.11

1. The function ρ must pass three tests:

   (a) positive homogeneity: $\rho(\alpha x) = |\alpha|\rho(x)$ for every $\alpha \in \mathbb{R}$;

   (b) convexity;

   (c) positive definiteness: $\rho(x) = 0$ if and only if $x = 0$.

   Together, the first two conditions imply the triangle inequality:

   $$\forall\, x, y \in \mathbb{R}^n \ : \ \rho(x+y) \le \rho(x) + \rho(y).$$

   Clearly ρ satisfies the conditions.

2. The dual norm has values, for $y \in \mathbb{R}^n$:

   $$\begin{aligned} \rho^*(y) &= \max_x \ y^T x \ : \ \rho(x) \le 1 \\ &= \max_x \min_{\lambda \ge 0} \ y^T x + \lambda \left(1 - \sum_{i=1}^p \|x_i\|_2\right) \\ &= \min_{\lambda \ge 0} \max_x \ y^T x + \lambda \left(1 - \sum_{i=1}^p \|x_i\|_2\right) \\ &= \min_{\lambda \ge 0} \lambda \ : \ \lambda \ge \|y_i\|_2, \ \ i = 1, \ldots, p \\ &= \max_{1 \le i \le p} \ \|y_i\|_2, \end{aligned}$$

   with $y = (y_1, \ldots, y_p)$, $y_i \in \mathbb{R}^{n_i}$.

   In the third line, we have used Slater's condition for convex programs to obtain strong duality; in the fourth,the fact that, for any vector y,

   $$\max_x \ y^T x - \lambda \|x\|_2 = \begin{cases} 0 & \text{if } \|y\|_2 \le \lambda \\ +\infty & \text{otherwise.} \end{cases}$$

3. The dual of the dual norm is always itself; indeed, ρ is closed (its

epigraph is closed). This can be checked directly:

$$\begin{aligned} \rho^{**}(y) &= \max_x \; y^T x \;:\; \rho^*(y) \leq 1 \\ &= \max_x \; y^T x \;:\; \|y_i\|_2 \leq 1, \;\; i = 1, \ldots, p \\ &= \max_x \; \sum_{i=1}^{p} y_i^T x_i \;:\; \|y_i\|_2 \leq 1 \;\; i = 1, \ldots, p \\ &= \sum_{i=1}^{n} \|x_i\|_2, \end{aligned}$$

where we have used the fact that the problem in the third line is decomposable.

4. The practical effect is that when λ is high enough, we will see many *blocks* of x equal to zero. This is not merely the same effect as if we'd use the ℓ_1-norm: the latter would just encourage elements (possibly scattered across different blocks) to be zero.

5. The dual of the problem is as follows.

$$\begin{aligned} p^*(\lambda) &= \min_x \max_{u,v} \; u^T(y - Ax) + v^T x \;:\; \|u\|_2 \leq 1, \;\; \rho^*(v) \leq \lambda \\ &= \max_{u,v} \; u^T y \;:\; A^T u = v, \;\; \|u\|_2 \leq 1, \;\; \rho^*(v) \leq \lambda \\ &= \max_u \; u^T y \;:\; \|u\|_2 \leq 1, \;\; \rho^*(A^T u) \leq \lambda \\ &= \max_u \; u^T y \;:\; \|u\|_2 \leq 1, \;\; \|A_i^T u\|_2 \leq \lambda, \;\; i = 1, \ldots, p. \end{aligned}$$

In the second line, we have used Sion's minimax theorem[23].

Let $i \in \{1, \ldots, p\}$. If $\sigma_{\max}(A_i) < \lambda$, then

$$\forall\, u \text{ with } \|u\|_2 \leq 1 \;:\; \|A_i^T u\|_2 < \lambda.$$

This implies that the constraint $\|A_i^T u\|_2 \leq \lambda$ cannot be active, at any optimum in the dual. Hence, p^* is the value of an optimization problem where we have set A_i to zero, or alternatively, x_i to zero[24]. This means that we can set $x_i = 0$ at optimum.

[23] Theorem 8.8.

[24] See the related discussion of removing inactive constraints in a convex problem (Proposition 8.1).

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, $j = 1, \ldots, 5$, 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 $k \in \{1, \ldots, m\}$ 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.

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

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.

Solution 9.1

1. For p_1^*, we gave the LP formulation

$$p_1^* = \min_{x,t,z} t + \sum_{i=1}^n z_i \ : \ \begin{array}{l} z_i \geq x_i \geq -z_i, \ \ i = 1, \ldots, n \\ t \geq (Ax-y)_i \geq -t, \ \ i = 1, \ldots, m. \end{array}$$

2. Likewise, for p_2^*, we obtain the QP

$$p_2^* = \min_{x,t,z} t^2 + \sum_{i=1}^n z_i \ : \ \begin{array}{l} z_i \geq x_i \geq -z_i, \ \ i = 1, \ldots, n \\ t \geq (Ax-y)_i \geq -t, \ \ i = 1, \ldots, m. \end{array}$$

3. For p_3^*, the problem is not convex.

4. For p_4^*, we have the QP

$$p_4^* = \min_{x,t,z} t^2 + \left(\sum_{i=1}^n z_i \right)^2 \ : \ \begin{array}{l} z_i \geq x_i \geq -z_i, \ \ i = 1, \ldots, n \\ t \geq (Ax-y)_i \geq -t, \ \ i = 1, \ldots, m. \end{array}$$

5. For p_5^*, we use the fact that the sum of the k largest elements of a vecctor $z \in \mathbb{R}^m$ can be expressed as[25]

$$\begin{aligned} \sum_{i=1}^k z_{[i]} &= \min_s ks + \sum_{i=1}^m \max(0, z_i - s) \\ &= \min_s ks + \sum_{i=1}^m u_i \ : \ u_i \geq 0, \ \ u_i \geq z_i - s, \ \ i = 1, \ldots, m. \end{aligned}$$

We obtain the QP

$$p_5^* = \min_{x,s,z,u} ks + \sum_{i=1}^m u_i + x^\top x \ : \ \begin{array}{l} z_i \geq x_i \geq -z_i, \ \ i = 1, \ldots, n \\ u_i \geq z_i - s, \ \ u_i \geq 0, \ \ i = 1, \ldots, m, \\ z_i \geq (Ax-y)_i \geq -z_i, \ \ i = 1, \ldots, m. \end{array}$$

[25] See Section 9.10.

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, $i = 1, \ldots, n$. 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.

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.

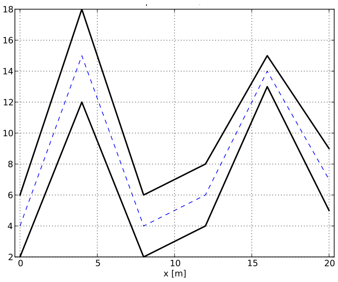

Figure 9.9: Slalom problem with $n = 5$ obstacles. "Uphill" (resp. "downhill") is on the left (resp. right) side. Middle path is dashed, initial and final positions are not shown.

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.

Solution 9.2

1. Assume that (x_i, z_i) is the crossing point of gate i, the path length minimization problem is thus

$$\begin{array}{ll} \min\limits_{z} & \sum\limits_{i=1}^{n+1} \left\| \begin{pmatrix} x_i \\ z_i \end{pmatrix} - \begin{pmatrix} x_{i-1} \\ z_{i-1} \end{pmatrix} \right\|_2 \\ \text{subject to} & y_i - c_i/2 \leq z_i \leq y_i + c_i/2, \text{ for } i = 1, \ldots, n \\ & z_0 = y_0, z_{n+1} = y_{n+1}, \end{array}$$

which is equivalent to

$$\begin{array}{ll} \min\limits_{z} & \sum\limits_{i=1}^{n+1} t_i \\ \text{subject to} & y_i - c_i/2 \leq z_i \leq y_i + c_i/2, \text{ for } i = 0, \ldots, n+1 \\ & \left\| \begin{pmatrix} x_i \\ z_i \end{pmatrix} - \begin{pmatrix} x_{i-1} \\ z_{i-1} \end{pmatrix} \right\|_2 \leq t_i, \text{ for } i = 1, \ldots, n+1. \end{array}$$

with the convention $c_0 = c_{n+1} = 0$. Hence, the problem is an SOCP.

2. A CVX code for the problem is

```
x = [0 4 8 12 16 20 24]';
y = [4 5 4 6 5 7 4]';
c = [0 3 2 2 1 2 0]';
n = 5;
cvx_begin
    variable z(n+2);
    variable t(n+1);
    minimize sum(t)
    subject to
        z <= y + c/2
        z >= y - c/2
        for i=1:n+1
            norm([x(i+1); z(i+1)] - [x(i); z(i)]) <= t(i)
        end
cvx_end
z = [4.0000 4.3749 4.7498 5.1248 5.5000 6.000 5.000];
```

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[26]

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

3. Interpret the result geometrically.

Solution 9.3

1. We can write

   $$\|\lambda p + (1-\lambda)q - a\|_2 = \|\lambda c + d\|_2,$$

   where $c \doteq p - q \neq 0$, $d = q - a$. We can always translate p, q to $p - a$, $q - a$, hence we can assume without loss of generality that $a = 0$.

2. We have

   $$D_* = \|c\|_2 \cdot \min_{0 \leq \lambda \leq 1} \|\lambda \tilde{c} + \tilde{d}\|_2$$

   with $\tilde{c} \doteq c/\|c\|_2$, $\tilde{d} = d/\|c\|_2$. The problem writes, after squaring,

   $$D_*^2 = \|c\|_2^2 \cdot \min_{\lambda \in [0,1]} \lambda^2 + 2\lambda \tilde{c}^\top \tilde{d} + \tilde{d}^\top \tilde{d}.$$

   The optimizer is either $\lambda_0 \doteq -\tilde{c}^\top \tilde{d}$ (if $\lambda_0 \in [0,1]$), or 0 (if $\lambda_0 < 0$), or 1 (if $\lambda_0 > 1$). Hence

   $$D_*^2 = \|c\|_2^2 \cdot \begin{cases} \tilde{d}^\top \tilde{d} - (\tilde{c}^\top \tilde{d})^2 & \text{if } -1 \leq \tilde{c}^\top \tilde{d} \leq 0, \\ \tilde{d}^\top \tilde{d} & \text{if } \tilde{c}^\top \tilde{d} > 0, \\ (\tilde{c} + \tilde{d})^\top (\tilde{c} + \tilde{d}) & \text{if } \tilde{c}^\top \tilde{d} < -1, \end{cases}$$

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

or, equivalently,

$$D_*^2 = \begin{cases} d^\top d - (c^\top d)^2/\|c\|_2^2 & \text{if } -\|c\|_2^2 \le c^\top d \le 0, \\ d^\top d & \text{if } c^\top d > 0, \\ (c+d)^\top(c+d) & \text{if } c^\top d < -\|c\|_2^2. \end{cases}$$

Further, for $c = p - q$, $d = q$, we have that $c^\top d = p^\top q - q^\top q$, $\|c\|_2^2 = p^\top p + q^\top q - 2p^\top q$, hence

$$D_*^2 = \begin{cases} q^\top q - \frac{(q^\top(p-q))^2}{\|p-q\|_2^2} & \text{if } p^\top q \le \min(q^\top q, p^\top p), \\ q^\top q & \text{if } p^\top q > q^\top q, \\ p^\top p & \text{if } p^\top q > p^\top p, \end{cases} \tag{9.13}$$

which is the expression we wished to prove.

3. Consider the line through p and q

$$L = \{x : x = q + \mu\tilde{c}\}, \quad \tilde{c} = \frac{p-q}{\|p-q\|_2}.$$

The projection of $a = 0$ onto this line is given by (see Section 2.3.2.1)

$$z^* = q + \mu^*\tilde{c}, \quad \mu^* = -\tilde{c}^\top q,$$

and the optimal squared distance is $\|z^*\|_2^2 = (\tilde{c}^\top q)^2$. If $\mu^* \in [0, \|p-q\|_2]$, then z^* is also the projection of $a = 0$ onto the segment S. The condition $\mu^* \in [0, \|p-q\|_2]$ is indeed equivalent to $p^\top q \le \min(q^\top q, p^\top p)$, which is the first condition in (9.13). When z^* is outside the segment, that is, when it falls to the left of q (for $\mu^* < 0$) or to the right of p (for $\mu^* > \|p-q\|_2$), then the optimal projection onto the segment is either q or p, respectively, see Figure 9.10.

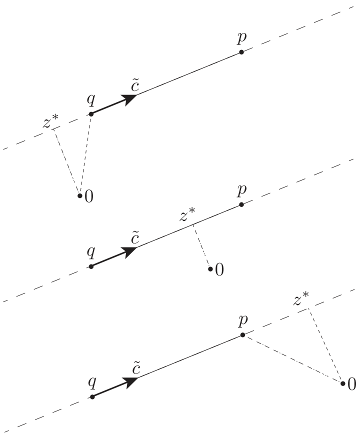

Figure 9.10: The three cases that may arise when projecting a point onto a segment.

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 $\lambda \ge 0$, $a \in \mathbb{R}^m$, $y \in \mathbb{R}^m$ are given, and $x \in \mathbb{R}$ is a scalar variable.

Assume that $y \ne 0$ and $a \ne 0$, (since otherwise the optimal solution of this problem is simply $x = 0$). Prove that the optimal solution of this problem is

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

where

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

corresponds to the solution of the problem for $\lambda = 0$. Verify that this solution can be expressed more compactly as $x^* = \mathrm{sthr}_{\lambda/\|a\|_2^2}(x_{\rm ls})$, where sthr is the *soft threshold* function defined in (12.64).

Solution 9.4 (Univariate LASSO) The subdifferential of the objective is

$$\partial f(x) = \frac{1}{2}\partial\|ax - y\|_2^2 + \lambda\partial|x|,$$

where

$$\partial\|ax - y\|_2^2 = x\|a\|_2^2 - a^\top y$$

and

$$\partial|x| = \begin{cases} \mathrm{sgn}(x) & \text{if } x \neq 0 \\ \{v : \ |v| \leq 1\} & \text{if } x = 0. \end{cases}$$

We first check under what conditions 0 is contained in the subdifferential of f at $x = 0$, that is

$$x^* = 0 \text{ is optimal} \quad \Leftrightarrow \quad 0 \in \partial f(0) = \left\{-a^\top y + \lambda v, \ |v| \leq 1\right\}.$$

Since the term λv may take any value in the interval $[-\lambda, \lambda]$, it follows that the above condition is satisfied if and only if $|a^\top y| \leq \lambda$.

Let next $|a^\top y| > \lambda$. Then, $x^* \neq 0$ and x is optimal if and only if

$$x\|a\|_2^2 - a^\top y + \lambda\,\mathrm{sgn}(x) = 0,$$

i.e., for

$$x = \frac{a^\top y}{\|a\|_2^2} - \frac{\lambda}{\|a\|_2^2}\mathrm{sgn}(x) = x_{\rm ls} - \frac{\lambda}{\|a\|_2^2}\mathrm{sgn}(x).$$

But since $|a^\top y| > \lambda$ the sign of x is not affected by the second term in the above expression, that is $\mathrm{sgn}(x) = \mathrm{sgn}(x_{\rm ls})$, and therefore the optimal solution is given by

$$x^* = x_{\rm ls} - \frac{\lambda}{\|a\|_2^2}\mathrm{sgn}(x_{\rm ls}),$$

which concludes the proof.

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.

Solution 9.5 (An optimal breakfast) This is a simple LP problem, which is soved via the following CVX code.

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.

```
% data
D = [.15 107 45 70; .25 500 40 121; 0.05 0 60 65];
cvx_begin
variable x(3)
cost = D(:,1)'*x;
vitamin = D(:,2)'*x;
sugar = D(:,3)'*x;
calories = D(:,4)'*x;
%
minimize ( cost )
subject to
calories <= 2250; calories >= 2000;
vitamin >= 5000; vitamin  <= 10000;
sugar <= 1000;
x >= 0;
x <= 10;
cvx_end
```

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 exist $d \in \mathbb{R}^m$ 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. MAake sure to express the solution steps of the method clearly.

Solution 9.6

1. The problem is always feasible because the range of A is the whole space $\mathbb{R}^m$.

2. We first prove the existence of r such that $c^\top r > 0$, $Ar = 0$ when $c \notin \mathbf{Range}(A)$. There are two possible techniques.

   A first one is based entirely on linear algebra. Due to the fundamental theorem of linear algebra, c can be written as $c = A^\top d + r$, with $Ar = 0$. Our assumption $c \notin \mathbf{Range}(A)$ implies $r \neq 0$. Since A is full row rank, $AA^\top \succ 0$, hence we can compute y, r:

   $$d = (AA^\top)^{-1}Ac, \;\; r = c - A^\top d = Pc, \text{ with } P \doteq I - A^\top(AA^\top)^{-1}A.$$

   Using the SVD of A, or by direct analysis, it is easy to see that the symmetric matrix P satisfies $P^2 = P$. (P is actually the projection operator on $\mathbf{Range}(A)^\perp = \mathbf{Null}(A)$.) We obtain $c^\top r = c^\top Pc = c^\top P^2 c = r^\top r > 0$.

   Another method for proving the existence of r is LP duality. Consider the LP

   $$\phi \doteq \max_r \, c^\top r \; : \; Ar = 0.$$

   The problem is feasible, hence, due to LP strong duality, ϕ has the same value as the dual:

   $$\phi = \min_y \max_r \, (c - A^\top y)r = \begin{cases} 0 & \text{if } c = A^\top y, \\ +\infty & \text{otherwise.} \end{cases}$$

   Since $c \notin \mathbf{Range}(A)$, we must have $\phi = +\infty$, and there exist a r such that $c^\top r > 0$.

   To finish the proof, we let $x(t) = x_0 + tr$, with x_0 feasible, and $t \in \mathbb{R}$. For every t, $x(t)$ is feasible, and $c^\top x(t) = c^\top x_0 + tc^\top r$. Letting $t \to -\infty$, we obtain $p^* = -\infty$.

3. Now assume that $c = A^\top d$ for some $d \in \mathbb{R}^m$. The problem writes

   $$p^* = \min_x \, d^\top Ax \; : \; l \leq Ax \leq u.$$

   We see that the variable x appears only via Ax. Letting $v = Ax$, we obtain the desired result. Note that once an optimal v is found, any x such that $Ax = v$ is optimal; the existence of x is guaranteed since A is full row rank. Of course, there will be possibly many choices for x. The minimum-norm solution is $x^* = A^\top(AA^\top)^{-1}v$.

4. Let $d = (AA^\top)^{-1}Ac$, the unique solution to a least-squares problem with a full rank matrix:

$$d = \arg\min_{\xi} \|A^\top \xi - c\|_2.$$

The condition $c \notin \textbf{Range}(A)$ is equivalent to $\|A^\top d - c\|_2 > 0$. When this is the case, we proceed by solving

$$p^* = \min_{v} d^\top v \ : \ l \leq v \leq u, \tag{9.14}$$

which can be solved in time linear in m: it simply requires finding the sign of d_i and choose v_i accordingly. Precisely, one choice for v is

$$v_i^* = \begin{cases} l_i & \text{if } d_i > 0, \\ u_i & \text{if } d_i < 0, \\ 0 & \text{if } d_i = 0. \end{cases} \tag{9.15}$$

The above takes into account the case when l_i or u_i are infinite; a zero in f indicates degrees of freedom. The above is the least-norm solution (when both l, u are finite).

The algorithm is as follows.

(a) Compute $d = (AA^\top)^{-1}A^\top c = \arg\min_\xi \|A^\top \xi - c\|$.

(b) Set v^* according to (9.15).

(c) Set $x^* = A^\top(AA^\top)^{-1}v^*$.

Note that the algorithm can be made more efficient if a full SVD of A is available. The following is entirely expressed in terms of the pseudo-inverse of A, $A^\dagger$.

(a) Form the SVD of A: $A = U(S, 0)V^\top$ with S diagonal positive-definite, U, V unitary of size m, n respectively.

(b) Compute $f = U(S^{-1}, 0)V^\top c = A^\dagger c$.

(c) Set v^* according to (9.15).

(d) Set $x^* = V(S^{-1}, 0)^\top U^\top v^* = (A^\dagger)^\top v^*$.

A similar approach can be taken based on the QR decomposition of A.

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

$$\min_{x \in \mathbb{R}} \|v - x\mathbf{1}\|_2^2, \tag{9.16}$$

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

$$\min_{x\in\mathbb{R}} \|v - x\mathbf{1}\|_1. \tag{9.17}$$

We consider a robust version of the average problem (9.16):

$$\min_{x} \max_{u\,:\,\|u\|_\infty\le\lambda} \|v + u - x\mathbf{1}\|_2^2, \tag{9.18}$$

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 $\lambda \ge 0$.

1. Is the robust problem (9.18) convex? Justify your answer precisely, based on expression (9.18), and without further manipulation.

2. Show that problem (9.18) can be expressed as

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

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 (9.17).

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.

Solution 9.7

1. The robust problem is convex, since the objective function is the pointwise maximum (over u) of convex functions, $x \to \|v + u - x\mathbf{1}\|_2^2$.

2. For a given vector $z \in \mathbb{R}^n$, we have

$$\begin{aligned} \max_{u\,:\,\|u\|_\infty\le\lambda} \|z+u\|_2^2 &= \sum_{i=1}^n \max_{\eta\,:\,|\eta|\le\lambda} (z_i+\eta)^2 \\ &= \sum_{i=1}^n (|z_i| + \lambda)^2, \end{aligned}$$

the last line resulting from

$$\forall\, \eta,\ \ |\eta| \le \lambda\ :\ |z_i + \eta| \le |z_i| + \lambda,$$

with upper bound attained with $\eta = \lambda \mathrm{sgn}(z_i)$.

3. A QP formulation is

$$\min_{x,t} \sum_{i=1}^{n} (t_i + \lambda)^2 \ : \ t_i \geq \pm(v_i - x), \ \ i = 1, \ldots, n.$$

4. The objective function takes the form

$$\sum_{i=1}^{n} (|v_i - x| + \lambda)^2 = \lambda^2 + 2\lambda \|v - x\mathbf{1}\|_1 + \|v - x\mathbf{1}\|_2^2,$$

   The corresponding optimization problem has the same minimizers as the problem

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

   When λ is large, the minimizer will tend to minimize the first term only, which implies the desired result.

5. The median problem can be interpreted as a robust version of the average problem, when the uncertainty is large.

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.

Solution 9.8

1. True, since p^* is the point-wise minimum of linear (hence, concave) functions (indexed by x), $f_x : c \to c^T x$. That is:

$$p^* = \min_{x \in \mathcal{X}} \ f_x(c),$$

   where $\mathcal{X} = \{x \ : \ Ax \leq b\}$. (When the latter set is empty, the problem is infeasible, and p^* is the function with constant value $+\infty$.)

2. True. Consider the function

$$f(x,b) = \begin{cases} c^T x & \text{if } Ax \le b, \\ +\infty & \text{otherwise} \end{cases}$$

We have

$$p^*(b) = \min_x f(x,b),$$

therefore if f is convex (with respect to the *pair* (x,b)), then $p^*(b)$ is. The former fact comes from the epigraph characterization of convexity: the condition $f(x,b) \le t$, for $t \in \mathbb{R}$, is equivalent to

$$c^T x \le t, \quad Ax \le b,$$

which are jointly convex conditions on (x,b,t).

3. False. Consider the problem with $n = 1, m = 1$, $b = 1$, $c = 1$, and with $A < 0$. We have $p^* = -1/A$ if $A < 0$, $-\infty$ otherwise. This function is not concave on its domain.

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 $S = \{x \in \mathbb{R}^n : x \ge 0, \mathbf{1}^\top x = 1\}$. 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

$$Ax \ge f(x)x \ge 0.$$

2. Prove that

$$f(x) \le \lambda, \quad \forall x \in S.$$

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 nonnegative matrices[27], but you are not asked to prove this fact.

[27] For a nonnegative matrix $A \ge 0$ an extension of the results stated in Exercise 3.11 for positive matrices holds. More precisely, if $A \ge 0$, then $\lambda = \rho(A) \ge 0$ is still an eigenvalue of A, with a corresponding eigenvector $v \ge 0$ (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 $A \ge 0$ is *primitive*, that is there exist an integer k such that $A^k > 0$ (Perron-Frobenius theorem).

Solution 9.9 (Variational formula for the dominant eigenvalue)

1. Since $a_i^\top > 0$ and $0 \neq x \geq 0$, we have that $a_i^\top x > 0$ for all $x \in S$, hence $f(x) > 0$, and $f(x)x \geq 0$. Also, for all i such that $x_i \neq 0$,

$$a_i^\top x = \frac{a_i^\top x}{x_i} x_i \geq \min_j \left(\frac{a_j^\top x}{x_j} \right) x_i = f(x)x_i,$$

and for $x_i = 0$ clearly $a_i^\top x \geq f(x)x_i = 0$, thus $a_i^\top x \geq f(x)x_i$ holds for all i, hence $Ax \geq f(x)x$, which, together with the previous point, proves the first claim

$$Ax \geq f(x)x \geq 0.$$

2. Multiply the previous inequality on the left by $w^\top$, to obtain

$$\lambda w^\top x = w^\top Ax \geq f(x) w^\top x,$$

which, since $w^\top x > 0$, implies that $f(x) \leq \lambda$ for all $x \in S$, as desired.

3. $v > 0$ belongs to S, and

$$f(v) = \min_{i=1,\ldots,n} \frac{[Av]_i}{[v]_i} = \min_{i=1,\ldots,n} \frac{\lambda [v]_i}{[v]_i} = \lambda.$$

So, $f(x) \leq \lambda$ for all $x \in S$, and $f(v) = \lambda$, which proves that the maximum of $f(x)$ over S is attained at $x = v$, and it is equal to λ.

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.)?

Solution 9.10 (LS with uncertain A matrix) We have that

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

and, since δ_i are i.i.d. and have zero mean,

$$\begin{aligned}
\mathbb{E}\{A(\delta)x\} &= A_0 x \\
\operatorname{var}\{A(\delta)x\} &= \mathbb{E}\{x^\top A(\delta)^\top A(\delta)x\} \\
&= \mathbb{E}\{x^\top A_0^\top A_0 x + 2\sum_{i=1}^{p} x^\top A_0^\top A_i x\delta_i + \sum_{i=1}^{p}\sum_{j=1}^{p} x^\top A_i^\top A_j x\delta_i\delta_j\} \\
&= x^\top A_0^\top A_0 x + \sum_{i=1}^{p} \sigma_i^2 x^\top A_i^\top A_i x \\
&= x^\top (A_0^\top A_0 + \sum_{i=1}^{p} \sigma_i^2 A_i^\top A_i)x.
\end{aligned}$$

Now, the expected objective is

$$\begin{aligned}
\mathbb{E}\{\|A(\delta)x - b\|_2^2\} &= \mathbb{E}\{x^\top A(\delta)^\top A(\delta)x + 2b^\top A(\delta)x + b^\top b\} \\
&= \mathbb{E}\{x^\top A(\delta)^\top A(\delta)x\} + 2b^\top A_0 x + b^\top b \\
&= x^\top (A_0^\top A_0 + \sum_{i=1}^{p} \sigma_i^2 A_i^\top A_i)x + 2b^\top A_0 x + b^\top b.
\end{aligned}$$

Minimizing $\mathbb{E}\{\|A(\delta)x - b\|_2^2\}$ is thus a convex quadratic optimization problem.

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.

Solution 10.1 (Squaring SOCP constraints) It is indeed possible to square a SOC constraint

$$\|Ax + b\|_2 \leq c^\top x + d,$$

provided that one takes care of adding the implicit constraint that $c^\top x + d \geq 0$. The SOC constraint is thus equivalent to

$$\begin{array}{rcl} \|Ax + b\|_2^2 & \leq & (c^\top x + d)^2 \\ 0 & \leq & c^\top x + d. \end{array}$$

In the exercise, the set defined by

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

can be expressed as

$$x_2(4x_1 - 3x_2) \leq 0$$

and it is depicted in Figure 10.11.

This set is clearly nonconvex. Instead, the intersection with the above set with the halfspace $x_1 + 2x_2 \geq 0$ is convex and identical to the set defined by the original SOC. Notice however that certain solvers for convex optimization (such as CVX) may not accept a constraint of the form $\|Ax + b\|_2^2 - (c^\top x + d)^2 \leq 0$ since, as shown above, the difference of two convex quadratic functions may well be nonconvex.

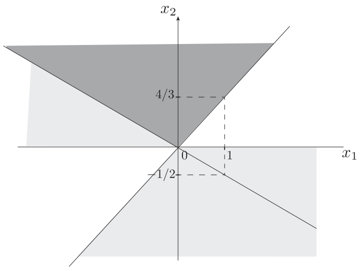

Figure 10.11: Squared SOC region.

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

$$\begin{array}{rcl} f(x) & = & \max\big(x_1 + x_2 - \min\big(\min(x_1 + 2, x_2 + 2x_1 - 5), x_3 - 6\big), \\ & & \dfrac{(x_1 - x_3)^2 + 2x_2^2}{1 - x_1}\big), \end{array}$$

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

Solution 10.2 (A complicated function) We first write the problem in epigraphic form as $\min_{x,t} t$ subject to $f(x) \le t$. Since f is the max of two component functions $f(x) = \max(f_1(x), f_2(x))$, the constraint is equivalent to $f_i(x) \le t$, for $i = 1, 2$. Further,

$$\begin{aligned}
f_1(x) \le t \quad &\Leftrightarrow \quad x_1 + x_2 - t \le \min(\min(x_1 + 2, x_2 + 2x_1 - 5), x_3 - 6) \\
&\Leftrightarrow \quad x_1 + x_2 - t \le \min(x_1 + 2, x_2 + 2x_1 - 5),\ x_1 + x_2 - t \le x_3 - 6 \\
&\Leftrightarrow \quad x_1 + x_2 - t \le x_1 + 2,\ x_1 + x_2 - t \le x_2 + 2x_1 - 5,\ x_1 + x_2 - t \le x_3 - 6; \\
f_2(x) \le t \quad &\Leftrightarrow \quad \frac{(x_1 - x_3)^2 + 2x_2^2}{1 - x_1} \le t \\
\text{(since } |x_1| < 1) \quad &\Leftrightarrow \quad (x_1 - x_3)^2 + 2x_2^2 \le t(1 - x_1)
\end{aligned}$$

The latter constraint is a hyperbolic one, and can be written as

$$\|Ax\|_2^2 \le t(1 - x_1), \quad A = \begin{bmatrix} 1 & 0 & -1 \\ 0 & \sqrt{2} & 0 \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},$$

which is equivalent, for $1 - x_1 > 0$, to the SOC constraint (see Section 10.1.1.1)

$$\left\| \begin{bmatrix} 2Ax \\ 1 - x_2 - t \end{bmatrix} \right\|_2 \le 1 - x_1 + t.$$

Exercise 10.3 (A minimum time path problem) Consider Figure 10.12, in which a point in 0 must move to reach point $p = [4 \ \ 2.5]^\top$, crossing three layers of fluids having different densities.

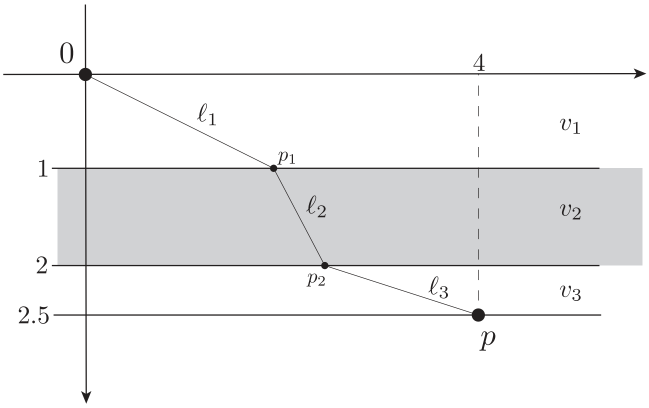

Figure 10.12: A minimum-time path problem.

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 a 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).

Solution 10.3 (A minimum time path problem) Let us denote with x_1, x_2, x_3 the horizontal coordinates of points p_1, p_2 and p, respectively, and with y_1, y_2, y_3 the corresponding vertical coordinates (which are given: $y_1 = 1$, $y_2 = 2$, $y_3 = 2.5$). Define as h_1, h_2, h_3 the lengths of the horizontal projections of the three legs, that is

$$h_1 = x_1, \quad h_2 = x_2 - x_1, \quad h_3 = x_3 - x_2.$$

We shall write the problem using ℓ_1, ℓ_2, ℓ_3, and h_1, h_2, h_3 as variables. Note that by Pythagoras theorem it must hold that

$$\begin{aligned}
\ell_1 &= \sqrt{h_1^2 + y_1^2} = \|(h_1, 1)\|_2 \\
\ell_2 &= \sqrt{h_2^2 + (y_2 - y_1)^2} = \|(h_2, 1)\|_2 \\
\ell_3 &= \sqrt{h_3^2 + (y_3 - y_2)^2} = \|(h_3, 0.5)\|_2.
\end{aligned}$$

The total travel time is

$$T = \frac{\ell_1}{v_1} + \frac{\ell_2}{v_2} + \frac{\ell_2}{v_2}.$$

We hence set up our optimization problem as follows:

$$\begin{aligned}
\min_{h,\ell} \quad & \tfrac{\ell_1}{v_1} + \tfrac{\ell_2}{v_2} + \tfrac{\ell_2}{v_2} \\
\text{s.t.:} \quad & \ell_1 = \|(h_1, 1)\|_2 \\
& \ell_2 = \|(h_2, 1)\|_2 \\
& \ell_3 = \|(h_3, 0.5)\|_2 \\
& h_1 + h_2 + h_3 = 4,
\end{aligned}$$

where the last constraint imposes that the horizontal coordinate of point p is equal to 4, as indicated by the figure. This formulation of the problem is not convex, due to the presence of nonlinear *equality* constraints. However, since we are minimizing a positive linear combination of the ℓ_i, we can substitute $=$ with $\geq$ in the constraints, obtaining

$$\begin{aligned}
\min_{h,\ell} \quad & \tfrac{\ell_1}{v_1} + \tfrac{\ell_2}{v_2} + \tfrac{\ell_2}{v_2} \\
\text{s.t.:} \quad & \ell_1 \geq \|(h_1, 1)\|_2 \\
& \ell_2 \geq \|(h_2, 1)\|_2 \\
& \ell_3 \geq \|(h_3, 0.5)\|_2 \\
& h_1 + h_2 + h_3 = 4.
\end{aligned}$$

This problem is a SOCP, and equality will hold at optimum. Hence, this problem is equivalent to the original one.

The following CVX code solves the problem with the given numerical data.

```
% minimum-time path exercise
p=[4 2.5]'; % target point
tk=[1 1 .5]; % layers thickness
v1 = 1;
nu2=1.5;
nu3=1.2;
v = [v1 v1/nu2 v1/nu3]';
%
cvx_begin
variable leg(3) % lenghts of paths
variable h(3) % orizontal lengths of path legs
time = leg'*(1./v);
minimize (time)
subject to
leg(1) >= norm([h(1) tk(1)]);
leg(2) >= norm([h(2) tk(2)]);
leg(3) >= norm([h(3) tk(3)]);
sum(h) == 4;
cvx_end
%% plot results
px = cumsum([0 h']);
py = cumsum([0 tk]);
plot(px,py,'k');
hold on
plot(px,py,'ko');
hold off
grid on
```

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 like 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 a 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.

Solution 10.4

1. For $k = 1$, we obtain a circle of radius d and center x_1. For $k = 2$, the set is an ellipse[28].

Indeed, without loss of generality we can assume $x_1 = -x_2 = p$, so that the set has now 0 as center of symmetry. Let $x \in \mathcal{E}$, with

$$\mathcal{E} \doteq \{x \ : \ \|x - p\|_2 + \|x + p\|_2 \leq d\}$$

We can always assume without loss of generality that $\|p\|_2 = 1$, and $d > 2$ (otherwise $\mathcal{E} = \emptyset$). We now change basis: we set p to be the first unit vector, and find $Q = [u_2, \ldots, u_n]$ such that $[p, Q]$ is unitary; in particular $Q^\top Q = I_{n-1}$. Then, any $x \in \mathbb{R}^n$ can be written as $x = tp + Qq$, with $q \in \mathbb{R}^{n-1}$, such that $q^\top p = 0$, and $t = p^\top x$.

The condition $x \in \mathcal{E}$ writes

$$\sqrt{(t-1)^2 + r^2} + \sqrt{(t+1)^2 + r^2} \leq d, \tag{10.19}$$

where $r \doteq q^\top Q^\top Q q = q^\top q$. The above is equivalent to $d \geq \sqrt{(t+1)^2 + r^2}$, and

$$\begin{aligned} (t-1)^2 + r^2 &\leq \left(d - \sqrt{(t+1)^2 + r^2}\right)^2 \\ &= d^2 + (t+1)^2 + r^2 - 2d\sqrt{(t+1)^2 + r^2}, \end{aligned}$$

which becomes

$$2d\sqrt{(t+1)^2 + r^2} \leq d^2 + (t+1)^2 - (t-1)^2 = d^2 + 4t$$

The above in turn implies

$$4d^2((t+1)^2 + r^2) \leq (d^2 + 4t)^2,$$

or, after some algebra, and upon replacing $r = q^\top q$:

$$q^\top q + (1 - \frac{4}{d^2})t^2 \leq \frac{d^2}{4} - 1.$$

Note that the above condition does imply that $d \geq \sqrt{(t+1)^2 + r^2}$, so it is equivalent to the original condition (10.19).

The above is an ellipse in (t, q) space, hence the set $\mathcal{E}$ is also an ellipse (rotated via the unitary transformation $[p, Q]$).

[28] For $n = 2$, you may remember from High School the fact that an ellipse can be described as the set of points such that the sum of the distances to a pair of points, called the foci, is a constant.

2. The problem

$$\min_x \sum_{i=1}^{k} \|x - x_i\|_2$$

can be written as the SOCP

$$\min_{x,t} \sum_{i=1}^{k} t_i \ : \ t_i \geq \|x - x_i\|_2, \ \ i = 1, \ldots, k.$$

3. We first determine the center of the k-ellipse, which by symmetry is the average point $\hat{x} = (x_1 + \ldots + x_k)/k$. Then for every $\theta \in [0, 2\pi]$, we solve the problem of maximizing the distance away from the center in the direction determined by the angle θ, with the constraint to be inside the k-ellipse:

$$\max_r \ r \ : \ \sum_{i=1}^{k} \|\hat{x} + ru(\theta) - x_i\|_2 \leq d.$$

where $u(\theta) = (\cos(\theta, \sin(\theta)))$. The above is an SOCP. In practice, we would discretize the variable θ and obtain a finite set of points on the 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 nonnegative, and the sum of the weights is one) so 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?

Solution 10.5 (A portfolio design problem) All constraints in this problem are simple linear constraints, except for the one involving the short-fall probability requirement. First observe that

$$\begin{aligned} \text{Prob}\{r^\top x < q\} \leq \epsilon \quad &\Leftrightarrow \quad \text{Prob}\{r^\top x \geq q\} \geq 1 - \epsilon \\ &\Leftrightarrow \quad \text{Prob}\{-r^\top x \leq -q\} \geq 1 - \epsilon. \end{aligned}$$

Following the developments in Example 10.4 in the book, we see that the latter constraint is equivalent to a SOC constraint of the form

$$\Phi^{-1}(1-\epsilon)\|\Sigma^{1/2}x\|_2 \leq \hat{r}^\top x - q,$$

where Φ^{-1} is the inverse standard Normal cumulative distribution function. The following CVX code then solves the problem. The maximal achievable expected return is 0.0964 (i.e., 9.6%). Notice that, in Matlab, the command `norminv(p,mu,sig)` returns the inverse cumulative distribution function for the Normal distribution with mean `mu` and standard deviation `sig`, evaluated at the values in `p`.

```
% Portfolio exercise
r = [0.12 0.10 0.07 0.03]';
S = [0.0064  0.0008  -0.0011 0; 0.0008  0.0025  0  0;...
    -0.0011  0  0.0004  0; 0 0 0 0];
Ss=sqrtm(S);
n=length(r);
ep = 1e-4; % bound of risk of loss q
q = -0.03; % loss
%
cvx_begin
cvx_solver sdpt3
variable x(n)
maximize ( r'*x )
subject to
x >= 0 ;
sum(x) == 1;
x >= 0.05;
x <= 0.4;
x(4) <= 0.2;
norminv(1-ep,0,1)*norm(Ss*x) <= r'*x - q;
cvx_end
```

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 an Euclidean ball, that is

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

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 it 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 $\|(H + \lambda^* I)^{-1}c\|_2 = r$.

Solution 10.6 (A trust-region problem) The Lagrangian of this problem can be written as

$$\begin{aligned} \mathcal{L}(x, \lambda) &= \frac{1}{2}x^\top Hx + c^\top x + d + \frac{\lambda}{2}(x^\top x - r^2) \\ &= \frac{1}{2}x^\top (H + \lambda I)x + c^\top x + d - \frac{\lambda}{2}r^2. \end{aligned}$$

The Lagrangian is strongly convex, hence it has a unique minimizer

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

Since strong duality holds, the optimal solution of the problem is $x^*(\lambda^*)$, where λ^* is the optimal solution of the dual problem

$$\max_{\lambda \geq 0} \; g(\lambda),$$

with

$$g(\lambda) \doteq -\frac{1}{2}c^\top (H + \lambda I)^{-1}c + d - \frac{\lambda}{2}r^2.$$

Let $H = U\Lambda U^\top$ be a spectral factorization of H, where U is orthogonal, and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \succ 0$. We can write $(H + \lambda I) = U(\Lambda + \lambda I)U^\top$, hence

$$g(\lambda) = d - \frac{\lambda}{2}r^2 - \frac{1}{2}\sum_{i=1}^{n} \frac{\tilde{c}_i^2}{\lambda_i + \lambda},$$

where $\tilde{c}_i$ are the entries of $\tilde{c} \doteq U^\top c$, and it holds that

$$\begin{aligned} g'(\lambda) \doteq \frac{\mathrm{d}g}{\mathrm{d}\lambda} &= -\frac{1}{2}r^2 + \frac{1}{2}\sum_{i=1}^{n} \frac{\tilde{c}_i^2}{(\lambda_i + \lambda)^2} \\ &= -\frac{1}{2}r^2 + \frac{1}{2}\|(H + \lambda I)^{-1}c\|_2^2. \end{aligned}$$

Notice that the term $\|(H + \lambda I)^{-1}c\|_2$ is strictly decreasing over $\lambda \geq 0$. Hence, if $g'(0) \leq 0$ then $g(\lambda) < 0$ for all $\lambda > 0$, which means that $g(\lambda)$ is decreasing for positive λ, hence the maximum is at the boundary point where $\lambda = 0$. If otherwise $g'(0) > 0$, then the concave function g has a maximum over $\lambda > 0$, at the point where the derivative g' is zero, that is where

$$\|(H + \lambda I)^{-1}c\|_2^2 = r^2,$$

which is what we needed to prove. The value of λ satisfying this equation can be found numerically via any univariate search technique. For instance, one may use Newton method, starting with some initial $\lambda > 0$, and iteratively updating

$$\lambda \leftarrow \lambda - \frac{g'(\lambda)}{g''(\lambda)},$$

where $g''(\lambda)$ is the second derivative

$$g''(\lambda) \doteq \frac{\mathrm{d}^2 g}{\mathrm{d}\lambda^2} = -\sum_{i=1}^{n} \frac{\tilde{c}_i^2}{(\lambda_i + \lambda)^3}.$$

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

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

Solution 10.7 (Univariate square-root LASSO) The problem is convex but nonsmooth, hence we write the optimality conditions in terms of the subdifferential of the objective:

$$0 \in \partial f(x) = \partial \|ax - y\|_2 + \lambda \partial |x|,$$

where

$$\partial \|ax - y\|_2 = \begin{cases} \frac{a^\top (ax - y)}{\|ax - y\|_2} & \text{if } ax - y \neq 0 \\ \{a^\top g : \ \|g\|_2 \leq 1\} & \text{if } ax - y = 0, \end{cases}$$

and

$$\partial |x| = \begin{cases} \text{sgn}(x) & \text{if } x \neq 0 \\ \{v : \ |v| \leq 1\} & \text{if } x = 0. \end{cases}$$

We first check under what conditions 0 is contained in the subdifferential of f at $x = 0$, that is

$$x^* = 0 \text{ is optimal} \quad \Leftrightarrow \quad 0 \in \partial f(0) = \left\{ \frac{a^\top y}{\|y\|_2} + \lambda v, \ |v| \leq 1 \right\}.$$

Since the term λv may take any value in the interval $[-\lambda, \lambda]$, it follows that the above condition is satisfied if and only if $|a^\top y|/\|y\|_2 \leq \lambda$, which proves the part regarding optimality of $x = 0$. Also, since by the Cauchy-Schwartz inequality it holds that

$$|a^\top y| \leq \|a\|_2 \|y\|_2,$$

it is clear that $\|a\|_2 \leq \lambda$ implies $|a^\top y| \leq \lambda \|y\|_2$, hence the optimal solution is certainly zero when $\|a\|_2 \leq \lambda$.

Consider next the case when the optimal solution is nonzero, i.e., when $|a^\top y| > \lambda \|y\|_2$, thus $\|a\|_2 > \lambda$. We initially assume for simplicity that a and y are not collinear, so that $ax - y \neq 0$ for all x; later we show that the derived solution is still valid if this assumption is lifted. With this assumption, and since $x \neq 0$, we have that

$$x \text{ is optimal} \quad \Leftrightarrow \quad 0 = \partial f(x) = \frac{a^\top (ax - y)}{\|ax - y\|_2} + \lambda \, \mathrm{sgn}(x),$$

that is, since $\|ax - y\|_2 \neq 0$, for

$$a^\top (ax - y) = -\lambda \|ax - y\|_2 \mathrm{sgn}(x). \tag{10.20}$$

All solution to this equation are also solutions of the squared equation

$$(a^\top ax - a^\top y)^2 = \lambda^2 \|ax - y\|_2^2, \tag{10.21}$$

which is a quadratic equation in x, equivalent to:

$$\|a\|_2^2(\|a\|_2^2 - \lambda^2)x^2 - 2a^\top y(\|a\|_2^2 - \lambda^2)x + (a^\top y)^2 - \lambda^2 \|y\|_2^2 = 0.$$

The roots of this equation are in

$$x_\pm = x_{\mathrm{ls}} \pm \sqrt{x_{\mathrm{ls}}^2 - \frac{(a^\top y)^2 - \lambda^2 \|y\|_2^2}{\|a\|_2^2(\|a\|_2^2 - \lambda^2)}}.$$

Observe that the term under the square root is nonnegative, since

$$\begin{aligned} \delta \doteq x_{\mathrm{ls}}^2 - \frac{(a^\top y)^2 - \lambda^2 \|y\|_2^2}{\|a\|_2^2(\|a\|_2^2 - \lambda^2)} &= \frac{(a^\top y)^2}{\|a\|^4} - \frac{(a^\top y)^2 - \lambda^2 \|y\|_2^2}{\|a\|_2^2(\|a\|_2^2 - \lambda^2)} \\ &= \frac{\lambda^2}{\|a\|_2^2} \cdot \frac{\|a\|_2^2 \|y\|_2^2 - (a^\top y)^2}{\|a\|_2^2(\|a\|_2^2 - \lambda^2)}, \end{aligned}$$

where, under the current conditions, $\|a\|_2^2 - \lambda^2 > 0$, and $\|a\|_2^2 \|y\|_2^2 - (a^\top y)^2 \geq 0$, by the Cauchy-Schwartz inequality. Further, $\delta \geq 0$ is smaller in magnitude than x_{ls}^2, since the condition $|a^\top y| > \lambda \|y\|_2$ implies that $x_{\mathrm{ls}}^2 - \delta > 0$. It follows that the sign of $x_\pm = x_{\mathrm{ls}} \pm \sqrt{\delta}$ is the same sign of x_{ls} (since adding $\pm\sqrt{\delta}$ to x_{ls} cannot change its sign).

Then, plugging $x \leftarrow x_{\pm}$ into equation (10.20), we have the left-hand side

$$\|a\|_2^2 x_{\pm} - a^\top y = \|a\|_2^2 (x_{\rm ls} \pm \sqrt{\delta}) - a^\top y = \pm\sqrt{\delta}$$

and the right-hand side

$$-\lambda \|ax_{\pm} - y\|_2 \mathrm{sgn}(x_{\pm}) = -\lambda \|ax_{\pm} - y\|_2 \mathrm{sgn}(x_{\rm ls}).$$

Thus, sign consistency is obtained by choosing the solution with "+" when $x_{\rm ls}$ is negative, and with "-" when $x_{\rm ls}$ is positive. In conclusion, the unique solution to eq. (10.20) is given by

$$x^* = x_{\rm ls} - \mathrm{sgn}(x_{\rm ls}) \frac{\lambda}{\|a\|_2^2} \sqrt{\frac{\|a\|_2^2 \|y\|_2^2 - (a^\top y)^2}{\|a\|_2^2 - \lambda^2}}, \tag{10.22}$$

which is the expression we wished to prove.

It only remains to be proved that the above expression is still valid also when y and a are collinear. In this case, since $\|a\|_2^2 \|y\|_2^2 = (a^\top y)^2$, eq. (10.22) gives $x^* = x_{\rm ls}$, and we have that $ax^* - y = 0$. Let us check that this solution is indeed optimal. The subdifferential of f at $x^* \neq 0$ such that $ax^* - y = 0$ is

$$\partial f(x^*) = \{a^\top g + \lambda \,\mathrm{sgn}(x^*),\ \|g\|_2 \leq 1\},$$

and we see that $0 \in \partial f(x^*)$ if $\|a\|_2 \geq \lambda$, which is indeed the condition under which the expression (10.22) for x^* holds.

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 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

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

   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.

Solution 10.8

1. The objective function of the problem defining f is concave in t, hence that problem is convex. We can write it as the SOCP
$$f(x) = 2 \max_{t,z} \ t - \sum_{i=1}^{n} z_i \ : \ z_i \geq \left\| \begin{bmatrix} \sqrt{x_i} \\ t \end{bmatrix} \right\|_2, \ \ i = 1, \ldots, n.$$

2. Yes, since f is the pointwise maximum (over t) of the convex functions with domain $\mathbb{R}_+^x$
$$f_t \ : \ x \to t - \sum_{i=1}^{n} \sqrt{x_i + t^2}.$$

3. We have, for a given $y \in \mathbb{R}_{++}^n$:
$$\begin{aligned} &\max_{x>0} \ -x^T y - f(x) \\ &= \max_{x>0} \min_{t} \ -x^\top y - 2t + 2\sum_{i=1}^{n} \sqrt{x_i + t^2} \\ &= \min_{t} \ -2t + \max_{x>0} \sum_{i=1}^{n} \left(2\sqrt{x_i + t^2} - x_i y_i \right), \\ &= \min_{t} \ -2t + \sum_{i=1}^{n} \max_{x_i>0} \left(2\sqrt{x_i + t^2} - x_i y_i \right), \end{aligned}$$
where we have used strong duality for the prblem defining f, as allowed by the Slater's condition. Next, we observe that for every scalars $y > 0$, $t > 0$, we have
$$\max_{x>0} 2\sqrt{x + t^2} - yx = 2\frac{1}{y} - y(\frac{1}{y^2} - t^2) = \frac{1}{y} + yt^2.$$
Hence,
$$\begin{aligned} &\max_{x>0} \ -x^T y - f(x) \\ &= \min_{t} \ -2t + \sum_{i=1}^{n} \left(\frac{1}{y_i} + y_i t^2 \right) \\ &= \min_{t} \ -2t + t^2 \left(\sum_{i=1}^{n} y_i \right) + \sum_{i=1}^{n} \frac{1}{y_i} \\ &= \min_{t} \sum_{i=1}^{n} \frac{1}{y_i} - \frac{1}{\sum_{i=1}^{n} y_i} \\ &= g(y). \end{aligned}$$
This shows that g is convex, as the pointwise maximum of affine functions. Note that the convexity of g is not immediately obvious.

Exercise 10.9 (Robust sphere enclosure) Let B_i, $i = 1, \ldots, m$, 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, $i = 1, \ldots, m$. Explain how to cast this problem into a known convex optimization format.

Solution 10.9 (Robust sphere enclosure) Let $c \in \mathbb{R}^n$ and $r \geq 0$ denote the center and radius of the enclosing ball B, respectively. We express the given balls B_i as

$$B_i = \{x : x = x_i + \delta_i, \ \|\delta_i\|_2 \leq \rho_i\}, \quad i = 1, \ldots, m.$$

We have that $B_i \subseteq B$ if and only if

$$\max_{x \in B_i} \|x - c\|_2 \leq r.$$

But

$$\max_{x \in B_i} \|x - c\|_2 = \max_{\|\delta_i\|_2 \leq \rho_i} \|x_i - c + \delta_i\|_2 = \|x_i - c\|_2 + \rho_i.$$

The problem is then cast as the following SOCP

$$\begin{array}{ll} \min\limits_{c,r} & r \\ \text{s.t.:} & \|x_i - c\|_2 + \rho_i \leq r, \quad i = 1, \ldots, m. \end{array}$$

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 $\tau \geq 0$
$$\begin{bmatrix} \|p-q\|_2^2+\tau & q^\top(p-q)-\tau/2 \\ q^\top(p-q)-\tau/2 & q^\top q - R^2 \end{bmatrix} \succeq 0.$$

3. Using the Schur complement rule[29], show that the above is consistent with the result given in Exercise 9.3.

[29] See Theorem 4.9.

Solution 11.1

1. The minimum distance is bounded below by a number $R \geq 0$ if and only if
$$\|\lambda(p-q)+q\|_2^2 \geq R^2 \text{ whenever } \lambda \in [0,1].$$
The desired result follows from the equivalence between $\lambda \in [0,1]$, and the quadratic inequality $\lambda(1-\lambda) \geq 0$.

2. The $\mathcal{S}$-procedure states that the above condition is equivalent to the existence of a scalar $\tau \geq 0$ such that
$$\forall \lambda \ : \ \|\lambda(p-q)+q\|_2^2 \geq R^2 + \tau\lambda(1-\lambda).$$
The above can be written as
$$\forall \lambda \ : \ \begin{bmatrix} \lambda \\ 1 \end{bmatrix}^\top \begin{bmatrix} \|p-q\|_2^2+\tau & q^\top(p-q)-\tau/2 \\ q^\top(p-q)-\tau/2 & q^\top q - R^2 \end{bmatrix} \begin{bmatrix} \lambda \\ 1 \end{bmatrix} \geq 0,$$
which in turn is equivalent to the desired result.

3. The squared minimal distance is the largest R such that the linear matrix inequality (in R^2, τ) holds:
$$D_*^2 = \max_{R^2,\tau} R^2 \ : \ \begin{bmatrix} \|p-q\|_2^2+\tau & q^\top(p-q)-\tau/2 \\ q^\top(p-q)-\tau/2 & q^\top q - R^2 \end{bmatrix} \succeq 0.$$

Using Schur complements:

$$\begin{aligned} D_*^2 &= \max_{\tau \geq 0} q^\top q - \frac{(q^\top(p-q) - \tau/2)^2}{\tau + \|p-q\|_2^2} \\ &= q^\top q - \min_{\tau \geq 0} f(\tau, \alpha, \beta), \end{aligned}$$

where $\alpha = q^\top(p-q)$, $\beta = \|p-q\|_2$, and

$$f(\tau, \alpha, \beta) = \frac{(\alpha - \tau/2)^2}{\tau + \beta^2}.$$

Setting the gradient to zero, we obtain that

$$\frac{\tau/2 - \alpha}{\tau + \beta^2} = 1 \text{ or } \tau = 2\alpha.$$

The first equality is the same as $\tau = -2(\alpha + \beta^2)$. We have obtained that the optimal τ is one of the three values: 0, 2α, or $-2(\alpha + \beta^2)$.

If $\alpha \geq 0$, that is, $p^\top q \geq q^\top q$, then $-2(\alpha + \beta^2) \leq 0$ cannot be optimal. If we set $\tau^* = 2\alpha \geq 0$ we attain the value 0 for f, which is the global minimum in that case; the corresponding squared minimum distance is then $D_*^2 = q^\top q$.

Now assume that $\alpha < 0$, that is, $p^\top q < q^\top q$, then the unconstrained minimizer is $-2(\alpha + \beta^2)$. Two cases can occur.

If $\alpha \leq -\beta^2 (\leq 0)$, that is, $p^\top p \geq p^\top q$, the unconstrained minimizer $-2(\alpha + \beta^2)$ is feasible for the problem, and thus, optimal. The corresponding optimal value for f is $-2\alpha - \beta^2$, and the squared minimum distance is then

$$D_*^2 = q^\top q + 2q^\top(p-q) + \|p-q\|_2^2 = p^\top p.$$

Otherwise, that is, if $0 \geq \alpha \geq -\beta^2$, that is, then the unconstrained minimizer is not feasible, and $\tau^* = 0$, resulting in the optimal value for f α^2/β^2, and a minimal distance

$$D_*^2 = q^\top q - \frac{(q^\top(p-q))^2}{\|p-q\|_2^2}.$$

We have obtained

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

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

$$\phi_p(X) \doteq \max_u \sum_{i=1}^m |x_i^\top u|^p \;:\; u^\top u = 1. \tag{11.23}$$

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
$$\psi_1(X) \doteq \max_U \sum_{i=1}^m \sqrt{x_i^\top U x_i} \;:\; U \succeq 0, \;\; \operatorname{trace} U = 1.$$
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 (8.12) to approximate, via weak duality, the problem (11.23). How does your bound compare with ψ_1?

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[30] 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

$$d^* = \min_D \operatorname{trace} D \;:\; D \text{ diagonal}, \, D \succeq Q. \tag{11.24}$$

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 $\approx 80\%$ of the true value, irrespective of 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 (11.24) can be solved in $O(m^3)$).

Solution 11.2

1. When $p = 2$,

$$\begin{aligned} \phi_2(X) &= \max_u \sum_{i=1}^m (x_i^\top u)^2 \;:\; u^\top u = 1 \\ &= \max_u \|X^\top u\|_2^2 \;:\; u^\top u = 1 \\ &= \lambda_{\max}(XX^\top) \\ &= \sigma_{\max}(X)^2. \end{aligned}$$

2. We have

$$\phi_1(X) = \max_U \sum_{i=1}^m \sqrt{x_i^\top U x_i} \;:\; U \succeq 0, \;\; \operatorname{trace} U = 1, \;\; \operatorname{rank}(U) = 1,$$

where we have used the variable $U = uu^\top$. The expression obtains upon dropping the rank constraint on the positive semi-definite matrix U. Since we relax (that is, drop) constraints in a maximization problem, we obtain an upper bound.

The function ψ_1 is convex, since $x \rightarrow \|U^{1/2} x\|_2$ is, for every $U \succeq 0$. It is positively homogeneous, and positive-definite (that is, $\psi_1(X) = 0$ implies $X = 0$). Hence, it is a norm.

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

3. We can express $\psi_1(X)$ as

$$\psi_1(X) = \max_{U,z} \sum_{i=1}^m \sqrt{z_i} \;:\; \begin{array}{l} U \succeq 0,\ \ \mathrm{trace}\, U = 1,\ \ z \geq 0, \\ z_i = x_i^\top U x_i,\ \ i = 1, \ldots, n. \end{array}$$

The above problem is convex and strictly feasible, hence strong duality holds: denoting by $\mathcal{U}$ the set of $n \times n$ positive semi-definite matrices with unit trace,

$$\begin{aligned} \psi_1(X) &= \max_{U \in \mathcal{U}, z \geq 0} \min_y \sum_{i=1}^m \left(\sqrt{z_i} + y_i (x_i^\top U x_i - z_i) \right) \\ &= \min_y \max_{U \in \mathcal{U}, z \geq 0} \sum_{i=1}^m \left(\sqrt{z_i} + y_i (x_i^\top U x_i - z_i) \right) \\ &= \min_{y \geq 0} \max_{U \in \mathcal{U}} \sum_{i=1}^m \frac{1}{4y_i} + \mathrm{trace}\, U \left(\sum_{i=1}^m y_i x_i x_i^\top \right) \\ &= \min_{y \geq 0} \sum_{i=1}^m \frac{1}{4y_i} + \lambda_{\max} \left(\sum_{i=1}^m y_i x_i x_i^\top \right), \end{aligned}$$

where we have used the fact that, for a given $\eta \in \mathbb{R}$:

$$\max_{\xi \geq 0} \sqrt{\xi} - \eta \xi = \begin{cases} \frac{1}{4\eta} & \text{if } \eta \geq 0, \\ +\infty & \text{otherwise,} \end{cases}$$

and that for any symmetric matrix Y:

$$\max_{U \in \mathcal{U}} \mathrm{trace}(YU) = \lambda_{\max}(Y).$$

4. Use the identity (8.12), we obtain

$$\begin{aligned} \phi_1(X) &= \max_{u \,:\, u^\top u = 1} \min_{d > 0} \frac{1}{2} \left(\sum_{i=1}^m d_i + \frac{(x_i^\top u)^2}{d_i} \right) \\ &\geq \min_{d > 0} \max_{u \,:\, u^\top u = 1} \frac{1}{2} \left(\sum_{i=1}^m d_i + \frac{(x_i^\top u)^2}{d_i} \right) \\ &= \min_{d > 0} \frac{1}{2} \sum_{i=1}^m d_i + \frac{1}{2} \max_{u \,:\, u^\top u = 1} u^\top \left(\sum_{i=1}^m \frac{1}{d_i} x_i x_i^\top \right) u \\ &= \psi_1(X), \end{aligned}$$

where we have used the the change of variable $d_i = 1/(2y_i)$, $i = 1, \ldots, m$, to prove the last equality.

5. We have, denoting by $\mathcal{D}$ the set of $m \times m$ positive-definite diagonal

matrices:

$$\begin{aligned}\psi_1(X) &= \frac{1}{2}\min_{D\in\mathcal{D}} \operatorname{trace} D + \lambda_{\max}(XD^{-1}X^\top) \\ &= \frac{1}{2}\min_{\alpha,\tilde{D}} \alpha \operatorname{trace} D + \frac{1}{\alpha} \ : \ \tilde{D}\in\mathcal{D},\, \lambda_{\max}(XD^{-1}X^\top) = 1 \\ &= \min_{\tilde{D}} \sqrt{\operatorname{trace}\tilde{D}} \ : \ \tilde{D}\in\mathcal{D},\ \ \lambda_{\max}(X\tilde{D}^{-1}X^\top) = 1 \\ &= \min_{\tilde{D}} \sqrt{\operatorname{trace}\tilde{D}} \ : \ \tilde{D}\in\mathcal{D},\ \ \lambda_{\max}(X\tilde{D}^{-1}X^\top) \leq 1,\end{aligned}$$

where in the last line we have exploited the fact that we can always scale $\tilde{D}$ in the last problem, and improve the objective while making the inequality an equality. Now that inequality is equivalent to

$$I \succeq X\tilde{D}^{-1}X^\top,$$

which, since $\tilde{D} \succ 0$, is equivalent, via Schur complements, to

$$\left[\begin{array}{cc} I & X \\ X^\top & \tilde{D}\end{array}\right] \succeq 0,$$

and, using Schur complements again, to $\tilde{D} \succeq X^\top X$.

6. Using the hint:

$$\begin{aligned}\psi_1(X) &= \max_{u\,:\,u^\top u=1}\ \max_{v\ \|v\|_\infty\leq 1} \sum_{i=1}^m v_i(x_i^\top u) \\ &= \max_{v\ \|v\|_\infty\leq 1}\ \max_{u\,:\,u^\top u=1} (Xv)^\top u \\ &= \max_{v\ \|v\|_\infty\leq 1} \|Xv\|_2.\end{aligned}$$

The maximum is always attained as claimed, since we are maximizing a convex function over a polytope, so that the maximum is always attained at the vertices.

7. We have, according to part 6,

$$\phi_1(X)^2 = \max_{v\,:\,\|v\|_\infty\leq 1} v^\top Q v,$$

where $Q = X^\top X$. The corresponding bound from Nesterov is

$$\phi_1(X)^2 \geq \min_D \operatorname{trace} D \ : \ D \text{ diagonal}, D \succeq X^\top X,$$

which is precisely the expression we have found in part 5.

8. For $p = 2$ the complexity is in $O(m^3 + nm^2)$ (to account for the cost of forming $X^\top X$), the same as $\psi_1(X)$.

Exercise 11.3 (Robust Principal Component Analysis) The following problem is known as Robust Principal Component Analysis[31]:

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

where $\|\cdot\|_*$ stands for the nuclear norm[32], 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 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:

   $$\|W\|_* = \max_Y \text{trace}\, W^\top Y \;:\; \|Y\|_2 \leq 1,$$

   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

   $$\|Y\|_2 \leq 1 \iff I - YY^\top \succeq 0,$$

   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.

Solution 11.3

1. We can write, thanks to the hint:

   $$p^* = \min_X \max_{Y,Z} \text{trace}\, Y^\top (A - X) + \text{trace}(Z^\top X) \;:\; \|Y\| \leq 1, \;\; \|Z\|_\infty \leq \lambda,$$

   where $\|\cdot\|_\infty$ is the largest magnitude of the entries of the matrix argument. Applying Sion's minimax theorem[33], we obtain

   $$\begin{aligned} p^* &= \max_{Y,Z} \left(\min_X \text{trace}\, Y^\top (A - X) + \text{trace}(Z^\top X) \right) \;:\; \|Y\| \leq 1, \;\; \|Z\|_\infty \leq \lambda \\ &= \max_Y \text{trace}\, Y^\top A \;:\; \|Y\| \leq 1, \;\; \|Y\|_\infty \leq \lambda. \end{aligned}$$

[31] See Section 13.5.4.

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

[33] See Theorem 8.8.

2. The condition $\|Y\|_\infty \leq \lambda$ is equivalent to $n(n+1)$ ordinary inequalities:
$$-\lambda \leq Y_{ij} \leq \lambda, \;\; 1 \leq i \leq j \leq n.$$
The condition $\|Y\| \leq 1$ is equivalent to $I - YY^\top \succeq 0$, which, using Shcur complements, writes as the linear matrix inequality in Y:
$$\begin{bmatrix} I & Y \\ Y^\top & I \end{bmatrix} \succeq 0.$$
Hence the problem can be written as an SDP:
$$\max_Y \; \operatorname{trace} Y^\top A \;:\; \begin{bmatrix} I & Y \\ Y^\top & I \end{bmatrix} \succeq 0, \quad -\lambda \leq Y_{ij} \leq \lambda, \;\; 1 \leq i \leq j \leq n.$$

3. The complementarity condition can be written as
$$X^*_{ij}(\lambda - |Y^*_{ij}|) = 0, \;\; 1 \leq i, j \leq n.$$
Hence if $\|Y^*\|_\infty < \lambda$ then $X^* = 0$. Now the condition is in turn satisfied if
$$\forall\, Y, \;\; \|Y\| \leq 1 \;:\; \|Y\|_\infty < \lambda.$$
Thus, $X^* = 0$ if $\max_{i,j} \phi_{ij} < \lambda$, where
$$\phi_{i,j} = \max_Y \; Y_{ij} \;:\; \|Y\| \leq 1.$$
Note that $Y_{ij} = \operatorname{trace} A^\top Y$, where $A = e_j e_i^\top$, where e_k denotes the k-th unit vector in $\mathbb{R}^n$. Note that the singular values of A are all zero, except the largest, which is 1.

Applying the hint to part 1, we obtain $\phi_{ij} = \|A\|_* = 1$. This concludes the proof.

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

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

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: $x_{\rm sdp} = \mathrm{sgn}(x^*)$, 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:

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

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

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

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 $\|Ax_{\rm sdp} - b\|_2^2$ of part 2, $\|Ax_{\rm int} - b\|_2^2$ of part 3, and the objective corresponding to a naïve method $\|Ax_{\rm ls} - b\|_2^2$, where $x_{\rm ls} = \mathrm{sgn}\left((A^\top A)^{-1} A^\top b\right)$ 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 $\{x : x_i \in \{-1, 1\}, \; i = 1, \ldots, n\}$. What can you say about the SDP approximation in that case?

Solution 11.4

1. Expanding the square and replacing Boolean constraints with $x_i^2 = 1$ we get

$$\phi = \min_x \; x^\top A^\top A x - 2b^\top A x + b^\top b \;:\; x_i^2 = 1, \;\; i = 1, \ldots, n.$$

Using the property trace UV = trace VU and introducing a new variable X, we obtain the desired expression:

$$\phi = \min_{X,x} \text{trace}(A^\top A X) - 2b^\top A x + b^\top b \;:\; X = xx^\top, \;\; X_{ii} = 1, \;\; i = 1, \ldots, n.$$

2. Since

$$\begin{bmatrix} X & x \\ x^\top & 1 \end{bmatrix} \succeq 0$$

we obtain in particular $X_{i,i} \geq x_i^2 \geq 0$, $i = 1, \ldots, n$. Therefore ϕ_{INT} is a relaxation of ϕ_{SDP}. Hence, $\phi_{SDP} \geq \phi_{INT}$.

3. A CVX code is below.

```
for iter = 1:100
m = 10;
n = 10;
A = randn(m,n);
b = 1*ones(m,1) + randn(m,1);

cvx_begin SDP
variable X(n,n) symmetric
variable x(n)
minimize(trace(A’*A*X)-2*b’*A*x+b’*b)
subject to
[X,x;x’,1]>=0;
for i=1:n
    X(i,i) == 1;
end
cvx_end
xs = sign(x);

cvx_quiet(true);
cvx_begin
variable x(n)
minimize(norm(A*x-b,2))
subject to
norm(x,Inf)<=1;
cvx_end
xi = sign(x);
```

```
xn = sign(pinv(A)*b);

sdp(iter) = norm(A*xs-b);
naive(iter) = norm(A*xn-b);
linf(iter) = norm(A*xi-b);

end
```

As seen in Fig. 11.13, on average, SDP out-performs INT and Least Squares in estimation and approximation quality. The computation time is shortest for Least Squares, while INT and SDP are similar.

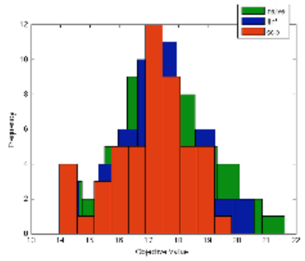

Figure 11.13: *Histogram comparing the performance of algorithms to approximate Boolean Least Squares.*

4. In that case, the optimal solution of the original problem is the same as the optimal solution of the relaxation.

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

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

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 α's?

2. Formulate the problem as a semidefinite program. You are free to choose the parametrization, 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 impose a bound 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?

Solution 11.5

1. For each t, denote by $\mathcal{L}(t)$ the line in α-space:

   $$\mathcal{L}(t) = \left\{\alpha \in \mathbb{R}^3 \ : \ y(t+2) = \alpha_1 y(t+1) + \alpha_2 y(t) + \alpha_3 u(t)\right\}.$$

   Geometrically, we are seeking the ellipsoid with minimal "size" that intersects every one of the above lines.

2. Let us parametrize the ellipsoid as

   $$\mathcal{E} = \{c + Ru \ : \ \|u\|_2 \le 1\},$$

   where $c \in \mathbb{R}^3$ and $R \in \mathbb{R}^{3\times 3}$, $R \neq 0$. For given $a \in \mathbb{R}^3$, $a \neq 0$, $b \in \mathbb{R}$, the ellipsoid above intersects the line $\mathcal{L} \doteq \{\alpha \ : \ a^\top \alpha = b\}$ if and only if

   $$1 \ge p^* \doteq \min_u \ \|u\|_2 \ : \ a^\top(c + Ru) = b.$$

   The above is a simple minimum-norm problem, with optimal value

   $$p^* = \frac{|b - a^\top c|}{\|R^\top a\|_2}.$$

   The intersection condition reduces to

   $$a^\top R R^\top a \ge (b - a^\top c)^2,$$

   Note that the condition can be written as a linear matrix inequality in $P \doteq RR^\top$ and c:

   $$\begin{bmatrix} a^\top P a & b - a^\top c \\ b - a^\top c & 1 \end{bmatrix} \succeq 0.$$

   Applying this result to our problem, and choosing the trace of P as measure of size (this corresponds to the sum of the squared semi-axis lengths of $\mathcal{E}$), we obtain the problem

   $$\min_{P,c} \ \mathrm{trace}\, P \ : \ \begin{bmatrix} a_t^\top P a_t & b_t - a_t^\top c \\ b_t - a_t^\top c & 1 \end{bmatrix} \succeq 0, \ \ t = 1, \ldots, T-2,$$

   where, for $t = 1, \ldots, T-2$:

   $$a_t = \begin{bmatrix} y(t+1) \\ y(t) \\ u(t) \end{bmatrix}, \ \ b_t = y(t+2).$$

   The above is an SDP.

3. When the ellipsoid must be a sphere, P is restricted to be of the form $r^2 I_3$, with $r \geq 0$. In that case, the problem reduces to

$$\min_{r,c} r \ : \ \|a_t\|_2 r \geq |b_t - a_t^\top c|, \ \ t = 1, \ldots, T-2,$$

which is a linear program, as claimed.

4. The issue is to formulate the conditions, keeping the variables α_t. For every α, the condition $\alpha \in \mathcal{E}$, that is, $\alpha = c + Ru$ holds for some u, $\|u\|_2 \leq 1$, if and only if $P \succeq (\alpha - c)(\alpha - c)^\top$. Using this, we express the problem as

$$\min_{P,c,(\alpha_t)_{t=1}^{T-1}} \operatorname{trace} P \ : \ \begin{array}{l} \|\alpha_{t+1} - \alpha_t\|_2 \leq \beta, \ a_t^\top \alpha_t = b_t, \\ \begin{bmatrix} P & \alpha_t - c \\ (\alpha_t - c)^\top & 1 \end{bmatrix} \succeq 0, \ \ t = 1, \ldots, T-2. \end{array}$$

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_i's 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

$$p(x) = \begin{bmatrix} 1 & x & x^2 \end{bmatrix} Q \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}. \tag{11.25}$$

for some positive-semidefinite matrix Q.

3. Show the converse: if a positive semi-definite matrix Q satisfies condition (11.25) 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 $p(x) = y_0 + y_1 x + y_2 x^2 + y_3 x^3 + y_4 x^4$ is non-negative everywhere if and only if there exist a 5×5 matrix Q such that

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

Hint: equate the coefficients of the powers of x in the left and right sides of equation (11.25).

Solution 11.6

1. A fourth-degree polynomial that is everywhere non-negative must have no real root; since the polynomial has real coefficients, the roots must come in complex conjugate pairs. Hence the roots must be of the form $a_1 \pm \jmath b_i$, $i = 1, 2$. We obtain the expression for $p(x)$ given in the hint; the leading coefficient must be non-negative, as seen from the case $x \to +\infty$. The result ensues by expanding the expression.

2. Since every polynomial q_i is of degree 2 at most, it can be written as

$$q_i(x) = \begin{bmatrix} 1 & x & x^2 \end{bmatrix} q_i,$$

for some appropriate vector $q_i \in \mathbb{R}^3$. We obtain

$$\begin{aligned} p(x) &= \sum_{i=1}^{4} q_i(x)^2 \\ &= \sum_{i=1}^{4} (\begin{bmatrix} 1 & x & x^2 \end{bmatrix} q_i)^2 \\ &= \begin{bmatrix} 1 & x & x^2 \end{bmatrix} Q \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}, \end{aligned}$$

where

$$Q = \sum_{i=1}^{4} q_i q_i^\top \succeq 0.$$

3. Assume that Q satisfies (11.25), and $Q \succeq 0$. We can write $Q = AA^\top$ for some 3×3 matrix A. Denote by q_i, $i = 1,2,3$ the i-th column of Q: we have $A = [q_1, q_2, q_3]$, and

$$Q = AA^\top = \sum_{i=1}^{4} q_i q_i^\top.$$

This shows that

$$p(x) = \left[\begin{array}{ccc} 1 & x & x^2 \end{array}\right] Q \left[\begin{array}{c} 1 \\ x \\ x^2 \end{array}\right] = \sum_{i=1}^{4} q_i(x)^2,$$

where $q_i(x) = \left[\begin{array}{ccc} 1 & x & x^2 \end{array}\right] q_i$, $i = 1,2,3$.

4. Let $p(x) = y_0 + y_1 x + y_2 x^2 + y_3 x^3 + y_4 x^4$. By identifying the coefficients of terms x^l, $l = 0,\ldots,4$, in the condition

$$\forall\, x \;:\; p(x) = \left[\begin{array}{ccc} 1 & x & x^2 \end{array}\right] Q \left[\begin{array}{c} 1 \\ x \\ x^2 \end{array}\right],$$

we obtain

$$y_{l-1} = \sum_{i+j=l} Q_{ij}, \quad l = 1,\ldots,5.$$

The result follows.

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 $f_k(X) \le t$ if and only if there exist $Z \in \mathbb{S}^n$ and $s \in \mathbb{R}$ such that

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

Hint: for the sufficiency part, think about the interlacing property[34].

[34] 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 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

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

Solution 11.7

1. First, the sufficiency part: assume that the stated conditions hold with some matrix $Z \in \mathbb{S}^n$ and scalar s. We then have $X \preceq Z + sI$. From the interlacing property of eigenvalues, this implies that

$$\lambda_i \leq \lambda_i(Z + sI), \quad i = 1, \ldots, n,$$

where $\lambda_i(\cdot)$ is the i-th largest eigenvalue of its matrix argument. Summing, we obtain

$$f_k(X) \leq \sum_{i=1}^k \lambda_i(Z + sI) = ks + \sum_{i=1}^k \lambda_i(Z) \leq ks + \mathrm{trace}(Z) \leq t,$$

as needed.

For the necessary part, assume that $f_k(X) \leq t$. Let $U \in \mathbb{R}^{n,n}$ be a unitary matrix such that $U^\top X U = \Lambda$ is diagonal, with the eigenvalues ordered in decreasing fashion: $\lambda_1 \geq \ldots \lambda_n$. Let $Z = UYU^\top$, with

$$Y = \mathrm{diag}\left(\lambda_1 - s, \ldots, \lambda_{k-1} - s, 0, \ldots, 0\right), \quad s = \lambda_k.$$

We have

$$t - ks - \mathrm{trace}(Y) \geq 0, \quad Y \succeq 0, \quad Y - \Lambda + sI \succeq 0.$$

Using congruence transformations[35], we obtain that the pair (Z, s) satisfies the conditions.

2. The condition $f_k(X) \leq t$ is a linear matrix inequality in (t, s, Z, X); thus the epigraph, as the projection (onto the (t, X)-space) of a convex set, is itself convex. Thus, the function f_k is convex. The function is also a norm: it satisfies the scaling property ($f_k(\alpha X) = \alpha f_k(X)$ for every $\alpha \geq 0$). Since it is convex, it satisfies the triangle inquality. Indeed, for every $X, Y \in \mathbb{S}^n$,

$$f_k(\frac{1}{2}(X + Y)) \leq \frac{1}{2}(f_k(X) + f_k(Y)).$$

Using the scaling property, and multiplying by $\alpha = 2$, we get the triangle inequality.

[35] See Section 4.4.3.

3. For $X \in \mathbb{R}^{m,n}$, let us examine the eigenvalues of the symmetric matrix $\tilde{X}$ given in the hint. A vector $(u, v) \in \mathbb{R}^{m+n}$ is an eigenvector of $\tilde{X}$ associated with the eigenvalue λ if and only if

$$\lambda u = Xv, \;\; \lambda v = X^\top u,$$

which implies that

$$XX^\top u = \lambda^2 u, \;\; X^\top X v = \lambda^2 v.$$

Thus, λ^2 is an eigenvalue of $XX^\top$, and $|\lambda|$ a singular value of X. Conversely, if σ is a singular value of X, then there exist u, v such that

$$\sigma u = Xv, \;\; \sigma v = X^\top u.$$

Changing u in $-u$, we see that

$$-\sigma(-u) = Xv, \;\; -\sigma v = X^\top(-u).$$

This means that the eigenvalues of the augmented matrix $\tilde{X}$ are $\pm\sigma_i(X)$, with σ_i the singular values of X. The sum of the top k singular values is thus the same as the sum of the top k eigenvalues of the augmented matrix. We obtain that the sum of the top k largest singular values of X is bounded above by t if and only if there exist $Z \in \mathbb{S}^{m+n}$ and $s \in \mathbb{R}$ such that

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

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 exist 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 found the desired point, hence we return x_k, and finish. If instead there exist 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[36] 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.

Solution 12.1 (Successive projections for linear inequalities)

1. The iterate $x_k - s_k a_{i_k}$ simply takes point x_k (who violates the i_k-th constraint), and projects it back onto the hyperplane $H_k \doteq \{x : a_{i_k}^\top x_k - b_{i_k} = 0\}$. Indeed, $s_k = a_{i_k}^\top x_k - b_{i_k} > 0$ is the distance from x_k to the hyperplane, and a_{i_k} is the unit norm vector orthogonal to the hyperplane, hence x_{k+1} is the projection of x_k onto the hyperplane. The algorithm then simply checks if the current point violates some constraint and, in the positive case, chooses any constraint that is violated and projects x_k on it. A particular choice

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

would be to select the constraint with the maximal violation (i.e., the index i_k such that s_k is maximal), in which case the algorithm becomes closely connected to the variation mentioned in point 3. of the exercise.

2. If at some finite iteration k the current point x_k satisfies all inequalities, then the algorithm exits with a feasible solution. Otherwise, suppose the algorithm runs indefinitely, and let $\bar{x} \in \mathcal{P}$ (such a feasible point is supposed to exist). We have that

$$\begin{aligned} \|x_{k+1} - \bar{x}\|_2^2 &= \|(x_k - \bar{x}) - s_k a_{i_k}\|_2^2 \\ &= \|x_k - \bar{x}\|_2^2 + s_k^2 \|a_{i_k}\|_2^2 - 2 s_k a_{i_k}^\top (x_k - \bar{x}) \\ &= \|x_k - \bar{x}\|_2^2 + s_k^2 - 2 s_k \left((a_{i_k}^\top x_k - b_{i_k}) - (a_{i_k}^\top \bar{x} - b_{i_k}) \right) \\ &\leq \|x_k - \bar{x}\|_2^2 + s_k^2 - 2 s_k (a_{i_k}^\top x_k - b_{i_k}), \end{aligned}$$

where the last expression follows from the fact that, since $\bar{x}$ is feasible by definition, it holds that $a_{i_k}^\top \bar{x} - b_{i_k} \leq 0$. Finally, noting that $s_k \doteq a_{i_k}^\top x_k - b_{i_k}$, from the last expression we obtain

$$\|x_{k+1} - \bar{x}\|_2^2 \leq \|x_k - \bar{x}\|_2^2 - s_k^2,$$

which shows that the distance of the current point from any fixed feasible point decreases at each iteration. Applying this inequality recursively from 0 onward, we obtain that

$$\|x_k - \bar{x}\|_2^2 \leq \|x_0 - \bar{x}\|_2^2 - \sum_{i=0}^{k} s_k^2.$$

Now, for $k \to \infty$, the summation must tend to a constant (for otherwise the right-hand side of the above expression would become negative and unbounded below), and this implies that s_k must tend to zero. Since s_k represents the distance from x_k to a violated constraint, we conclude that the algorithm will produce a sequence of solutions x_k that tends to a feasible solution as $k \to \infty$.

3. Let us assume that the finite minimum value p^* of f_0 is attained at some point x^* (this is for instance guaranteed if the polyhedron $\mathcal{P}$ is assumed to be bounded). A subgradient of f_0 at x_k is given by $g_k = a_{i_k}$, where i_k is the value of index i that attains the maximum in $\max_{i=1,\ldots,m}(a_i^\top x_k - b_i)$, i.e., such that $f_0(x_k) = a_{i_k}^\top x_k - b_{i_k}$. We can thus minimize f_0 by applying the projected subgradient method described in Section 12.4.1 (only, there is nothing to project here, since the problem is unconstrained), which takes the form

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

where now s_k must be a non-summable and diminishing sequence (e.g., $s_k = 1/(k+1)$). Denoting with $f^*_{0,k}$ the lowest objective value obtained up to iteration k of this algorithms, we have that $f^*_{0,k} \to p^*$. In practice, if at some k we find $f^*_{0,k} \leq 0$, then we exit and return a feasible solution x_k. If, otherwise, the sequence of $f^*_{0,k}$ seemingly converges to a positive p^*, then we conclude that the problem is infeasible ($\mathcal{P}$ is empty).

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

$$p^* = \min_{x \in \mathcal{X}} f_0(x), \tag{12.26}$$

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[37].

Initialize the iterations with some $x_0 \in \mathcal{X}$, and set $k = 0$. 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[38] L, and that $\|x - y\|_2 \leq R$ for every $x, y \in \mathcal{X}$. In this exercise, you shall prove that

$$\delta_k \doteq f_0(x_k) - p^* \leq \frac{2LR^2}{k+2}, \quad k = 1, 2, \ldots \tag{12.27}$$

1. Using the inequality

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

which holds for any convex f_0 with Lipschitz continuous gradient[39], prove that

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

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

[37] 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).

[38] As defined in Section 12.1.1.

[39] See Lemma 12.1.

2. Show that the following recursion holds for δ_k:

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

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

3. Prove by induction on k the desired result (12.27).

Solution 12.2 (Conditional gradient method)

1. Using inequality (12.5), we have that

$$\begin{aligned} f_0(x_{k+1}) &\leq f_0(x_k) + \nabla^\top f_0(x_k)(x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 \\ &= f_0(x_k) + \gamma_k \nabla^\top f_0(x_k)(z_k - x_k) + \gamma_k \frac{L}{2}\|z_k - x_k\|_2^2 \\ &\leq f_0(x_k) + \gamma_k \nabla^\top f_0(x_k)(z_k - x_k) + \gamma_k^2 \frac{LR^2}{2} \\ &\quad \left(\begin{array}{c} \text{since, } \forall x \in X, \\ \nabla^\top f_0(x_k) z_k \leq \nabla^\top f_0(x_k) x \end{array}\right) \\ &\leq f_0(x_k) + \gamma_k \nabla^\top f_0(x_k)(x^* - x_k) + \gamma_k^2 \frac{LR^2}{2}, \end{aligned}$$

which proves the first point in the exercise.

2. Continuing the previous chain of inequalities, and considering that, due to convexity inequality, it holds $\forall y$ that

$$f_0(y) \geq f_0(x) + \nabla^\top f_0(x)(y - x),$$

we have that

$$f_0(x_{k+1}) \leq f_0(x_k) + \gamma_k(f_0(x^*) - f_0(x_k)) + \gamma_k^2 \frac{LR^2}{2}.$$

Subtracting $p^* = f_0(x^*)$ from both sides of this inequality, we obtain

$$f_0(x_{k+1}) - p^* \leq (1 - \gamma_k)(f_0(x_k) - p^*) + \gamma_k^2 \frac{LR^2}{2},$$

which we rewrite more compactly as

$$\delta_{k+1} \leq (1-\gamma_k)\delta_k + \gamma_k^2 C, \quad k = 0, 1, \ldots, \tag{12.28}$$

where $\delta_k \doteq (f_0(x_k) - p^*)$, $C \doteq \frac{LR^2}{2}$.

3. We next prove[40] that

$$\delta_k \leq \frac{4C}{k+2}, \quad \text{for } k = 1, 2, \ldots. \tag{12.29}$$

[40] This approach was proposed by M. Jaggi, Ph.D. Thesis, ETH, Zürich, 2011.

using induction on k. First, evaluating (12.28) for $k = 0$, and recalling that $\gamma_k = \frac{2}{k+2}$, we have that

$$\delta_1 \leq C \leq \frac{4}{3}C,$$

which shows that (12.29) is true for $k = 1$. Assume next (12.29) is true for generic k, and evaluate δ_{k+1}:

$$\begin{aligned}
\delta_{k+1} &\leq (1-\gamma_k)\delta_k + \gamma_k^2 C \\
\text{(by the inductive hyp.)} &\leq (1-\gamma_k)\frac{4C}{k+2} + \gamma_k^2 C \\
&= \left(1 - \frac{2}{k+2}\right)\frac{4C}{k+2} + \frac{4C}{k+2}\frac{1}{k+2} \\
&= \frac{4C}{k+2}\left(1 - \frac{1}{k+2}\right) \\
&= \frac{4C}{k+2}\frac{(k+2)-1}{(k+2)} \leq \frac{4C}{k+2}\frac{(k+2)-1+1}{(k+2)+1} \\
&= \frac{4C}{k+2}\frac{k+2}{k+3} = \frac{4C}{k+3},
\end{aligned}$$

which concludes the inductive proof.

Exercise 12.3 (Bisection method) The bisection method applies to one-dimensional convex problems[41] of the form

$$\min_x f(x) \; : \; x_l \leq x \leq x_u$$

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 sub-gradient g of f at x is evaluated; if $g < 0$, we set $\underline{x} = x$; otherwise[42], we set $\overline{x} = x$. Then the midpoint x is recomputed, and the process is iterated until convergence.

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

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

[41] See an application in Section 11.4.1.3.

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

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

Solution 12.3 (A bisection method)

1. Let $g(x)$ be a subgradient of f at x. By the subgradient inequality it holds for all y that

   $$f(y) \geq f(x) + g(x)(y - x).$$

   If $g(x) < 0$, then all points y such that $y < x$ are such that $f(y) > f(x)$, thus it is useless to look for minimizers on the left of the current point, and the algorithm indeed sets $\underline{x} = x$. If otherwise $g(x) > 0$, then all points y such that $y > x$ are such that $f(y) > f(x)$, thus it is useless to look for minimizers on the right of the current point, and the algorithm indeed sets $\overline{x} = x$.

   Let ℓ_k denote the length of the search interval at iteration k of the algorithm, being $\ell_0 = x_u - x_l$ the initial length, let x_k denote the candidate solution at iteration k, and let x^* be the optimal solution. At each iteration $k = 0, 1, \ldots$, it holds that $|x_k - x^*| \leq \frac{1}{2}\ell_k$. Further, the length of the search interval is halved at each iteration, i.e., $\ell_{k+1} = \frac{1}{2}\ell_k$, therefore, $\ell_k = \frac{1}{2^k}\ell_0$, and

   $$|x_k - x^*| \leq \frac{1}{2^{k+1}}\ell_0.$$

   For given accuracy $\epsilon > 0$ it thus holds that

   $$|x_k - x^*| \leq \epsilon$$

   whenever

   $$k \geq \log_2 \frac{\ell_0}{\epsilon} - 1.$$

2. Initialize the algorithm with some $x \in \mathbb{R}$, and evaluate a subgradient $g = g(x)$. If $g = 0$ we terminate the algorithm, since x is optimal. If $g < 0$ we set $x_l = x$ and increase $x \leftarrow 2^{\mathrm{sgn}(x)}x$; if $g > 0$ we set $x_u = x$ and decrease $x \leftarrow 2^{-\mathrm{sgn}(x)}x$. We iterate until both x_l and x_u are set, or otherwise declare that f is unbounded below. Once x_l and x_u are obtained, we switch to the standard bisection algorithm described in the previous point.

3. This function is the sum of maxima of convex functions (since $A_{ij} \geq 0$), hence it is convex. A subgradient g of f at x can be found using the max rule (see Section 8.2.3.1) as

   $$g = \sum_{i=1}^{n} A_{ij_*}x + B_{ij_*},$$

where j_* is any index attaining the maximum in

$$\max_{1 \leq j \leq m} \frac{1}{2} A_{ij} x^2 + B_{ij} x + C_{ij}.$$

The problem can then be solved via the approach described in point 2. of this exercise.

Exercise 12.4 (KKT conditions) Consider the optimization problem[43]

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} \sum_{i=1}^{n} \left(\tfrac{1}{2} d_i x_i^2 + r_i x_i \right) \\ &\text{s.t.: } a^\top x = 1, \;\; x_i \in [-1, 1], \;\; i = 1, \ldots, n, \end{aligned}$$

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?

Solution 12.4 (KKT conditions)

1. Write the interval constraints on x_i as $x_i^2 \leq 1$, and define $D = \operatorname{diag}(d_1, \ldots, d_n)$, $\Lambda = \operatorname{diag}(\lambda)$, $\lambda = (\lambda_1, \ldots, \lambda_n)$, $r = (r_1, \ldots, r_n)$. The problem is a standard convex quadratic problem of the form

$$\begin{aligned} \min_{x} \quad & \tfrac{1}{2} x^\top D x + r^\top x \\ \text{s.t.:} \quad & a^\top x = 1 \\ & x_i^2 \leq 1 \qquad i = 1, \ldots, n. \end{aligned}$$

Strong duality holds for this problem, due to satisfaction of the Slater's conditions (see Proposition 8.7). The Lagrangian of the problem is

$$\mathcal{L}(x, \lambda, \mu) = \frac{1}{2} x^\top (D + \Lambda) x + (r + \mu a)^\top x - \left(\mu + \frac{1}{2} \sum_{i=1}^{n} \lambda_i \right)$$

The KKT necessary and sufficient conditions for optimality of x^* and of λ^*, μ^* require primal feasibility, dual feasiblity (i.e., $\lambda^* \geq 0$), complementary slackness $\lambda_i^*((x_i^*)^2 - 1) = 0$, and Lagrangian stationarity:

$$\nabla_x \mathcal{L}(x, \lambda^*, \mu^*) = (D + \Lambda^*) x + (r + \mu^* a) = 0.$$

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

That is, since D is strictly positive,

$$x^* = -(D + \Lambda^*)^{-1}(r + \mu^* a),$$

i.e., since D and Λ are diagonal,

$$x_i^* = -\frac{r_i + \mu^* a_i}{d_i + \lambda_i^*}, \quad i = 1, \ldots, n. \tag{12.30}$$

Notice that the problem is essentially separable in the variables x_i, except for the coupling constraint $a^\top x = 1$.

2. The dual function is

$$\begin{aligned} g(\lambda, \mu) &= \inf_x \mathcal{L}(x, \lambda, \mu) \\ &= -\frac{1}{2}(r + \mu a)^\top (D + \Lambda)^{-1}(r + \mu a) - \left(\mu + \frac{1}{2}\sum_{i=1}^n \lambda_i\right) \\ &= -\mu - \frac{1}{2}\sum_{i=1}^n \frac{(r_i + \mu a_i)^2}{d_i + \lambda_i} + \lambda_i, \end{aligned}$$

and the dual problem amounts to maximising $g(\lambda, \mu)$ w.r.t. μ and $\lambda \geq 0$. Let us find the maximum of $g(\lambda, \mu)$ w.r.t. λ, for fixed μ. The problem is separable, and it amounts to finding

$$\max_{\lambda_i \geq 0} -\frac{(r_i + \mu a_i)^2}{d_i + \lambda_i} - \lambda_i.$$

The unconstrained optimal point is at $\lambda_i = |r_i + \mu a_i| - d_i$. When this quantity is nonnegative, it is also the optimal solution of the constrained problem, otherwise the optimal solution is at the boundary, i.e., $\lambda_i = 0$. We thus have that

$$\lambda_i^*(\mu) = \begin{cases} |r_i + \mu a_i| - d_i & \text{if } |r_i + \mu a_i| - d_i \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{12.31}$$

Substituting (12.31) into (12.30), we can express $x_i^*(\mu)$ as a function of the scalar dual variable μ only:

$$x_i^*(\mu) = \begin{cases} -\text{sgn}(r_i + \mu a_i) & \text{if } |r_i + \mu a_i| - d_i \geq 0 \\ -(r_i + \mu a_i)/d_i & \text{otherwise.} \end{cases} \quad i = 1, \ldots, n. \tag{12.32}$$

The optimal μ is the value which makes the constraint $\sum_i a_i x_i^*(\mu) = 1$ satisfied. We can find such value by bisection over μ. Since the gradient of $g(\lambda^*, \mu)$ with respect to μ is simply given by $a^\top x_i^*(\mu) - 1$ (see Section 8.5.9), we increase μ when the gradient is positive, and decrease it when it is negative. In practice, we initialize two values $\mu_l < 0$, $\mu_r > 0$ for which we know a priori that $\mu^* \in [\mu_l, \mu_r]$, and execute the following bisection algorithm[44]

[44] See also Exercise 12.3.

(a) Set $\mu = (\mu_r + \mu_l)/2$

(b) Evaluate $h = a^\top x^*(\mu) - 1$

(c) If $h > 0$, let $\mu_l = \mu$, else let $\mu_r = \mu$

(d) If $|\mu_r - \mu_l| \leq \epsilon$, or $h = 0$, exit and return μ, else goto (a).

This algorithm returns a value of μ that is within ϵ from the true optimum μ^* in a number of iterations upper bounded by $\lceil \log_2(\ell/\epsilon) - 1 \rceil$, where ℓ is the length of the initial localization interval.

3. The numerical simulations confirm precisely this iteration bound. The following Matlab script implements the proposed bisection algorithm and compares it with the solution obtained via CVX.

```
% Exercise on KKT conditions
n=10;
% some problem data
d=rand(n,1);
r=randn(n,1);
a = randn(n,1);

%% solve with CVX
cvx_begin quiet
Dsqrt=diag(sqrt(d));
variable xc(n)
dx = Dsqrt*xc;
minimize (0.5*(dx’*dx)+r’*xc)
subject to
a’*xc == 1;
abs(xc) <= 1;
cvx_end

%% solve by bisection
mu_l=-100;
mu_r=100;
ell = mu_r-mu_l;
ep=1e-6;
k=1;
while 1
    mu = (mu_l+mu_r)/2;
    y = r+mu*a;
    yind = find(abs(y)-d >= 0);
    x = -y./d;
    x(yind) = -sign(y(yind));
    h=a’*x-1;
    if h>0
        mu_l=mu;
    else
        mu_r=mu;
    end;
    if abs(mu_r-mu_l)<=ep | h==0
        break;
    end;
    k=k+1;
end;
```

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 an 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[45]. 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

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

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 subproblem of the form

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

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 following methods. Make sure to state precisely the algorithm's steps.

   - Coordinate descent.
   - Dual coordinate ascent.
   - Projected subgradient.

[45] See Section 13.5.5.

- 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 5 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.

Solution 12.5 (Sparse Gaussian graphical models) TBD

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 $|u_i| \leq 1$, $i = 1, \ldots, m$. We parametrize a polynomial of degree d via its coefficients:

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

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:

$$\min_w \|\Phi^\top w - y\|_2^2 + \lambda b(w), \tag{12.33}$$

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 (12.33).

Solution 12.6 (Polynomial fitting with derivative bounds)

1. We have that

$$\frac{d}{du}p_w(u) = w_1 + 2w_2 u + \cdots + d w_d u^{d-1} = \tilde{w}^\top C\tilde{\phi}(u),$$

where

$$C = \mathrm{diag}\,(1, 2, \ldots, d)\,, \quad \tilde{w}^\top = [w_1 \, \cdots \, w_d], \quad \tilde{\phi}(u)^\top = [1\, u \, \cdots \, u^{d-1}].$$

Thus,

$$\left|\frac{d}{du}p_w(u)\right| = \left|\tilde{w}^\top C\tilde{\phi}(u)\right| = \max(\tilde{w}^\top C\tilde{\phi}(u), -\tilde{w}^\top C\tilde{\phi}(u)),$$

hence

$$b(w) = b(\tilde{w}) = \max_{u:|u|\leq 1}\left|\frac{d}{du}p_w(u)\right| = \max_{u:|u|\leq 1}\max(\tilde{w}^\top C\tilde{\phi}(u), -\tilde{w}^\top C\tilde{\phi}(u))$$

is the max of linear functions of w, hence it is convex.

$b(w)$ is not a norm on $\mathbb{R}^{d+1}$, since it does not depend on w_0. However, it is indeed a norm on $\mathbb{R}^d$, since it satisfied the three axioms defining a norm (see Section 2.2.1.2). Specifically, if $\tilde{w} = 0 \in \mathbb{R}^d$, then clearly $b(\tilde{w}) = 0$. Conversely, if $b(\tilde{w}) = 0$ then by definition

$$\max_{|u|\leq 1}|\tilde{w}^\top C\tilde{\phi}(u)| = 0,$$

that is the polynomial in u, $\tilde{w}^\top C\tilde{\phi}(u)$ is zero for all u: $|u| \leq 1$, which implies that this polynomial must be the zero polynomial, that is $\tilde{w} = 0$. Further, it is easily verified that $b(\alpha\tilde{w}) = |\alpha| b(\tilde{w})$. Finally, the triangle inequality holds, since for any $\tilde{w}^{(1)} \in \mathbb{R}^d$, $\tilde{w}^{(2)} \in \mathbb{R}^d$ we have

$$\begin{aligned} b(\tilde{w}^{(1)} + \tilde{w}^{(2)}) &= \max_{|u|\leq 1}|(\tilde{w}^{(1)} + \tilde{w}^{(2)})^\top C\tilde{\phi}(u)| \\ &= \max_{|u|\leq 1}|\tilde{w}^{(1)\top} C\tilde{\phi}(u) + \tilde{w}^{(1)\top} C\tilde{\phi}(u)| \\ &\leq \max_{|u|\leq 1}|\tilde{w}^{(1)\top} C\tilde{\phi}(u)| + |\tilde{w}^{(1)\top} C\tilde{\phi}(u)| \\ &\leq \max_{|u|\leq 1}|\tilde{w}^{(1)\top} C\tilde{\phi}(u)| + \max_{|v|\leq 1}|\tilde{w}^{(2)\top} C\tilde{\phi}(v)| \\ &= b(w^{(1)}) + b(w^{(2)}). \end{aligned}$$

2. Let us find a value u^* of u that achieves the optimum in

$$b(\tilde{w}) = \max_{|u|\leq 1}|p'_w(u)| = \max_{|u|\leq 1}\max(p'_w(u), -p'_w(u)),$$

where $p'_w(u) \doteq \tilde{w}^\top C\tilde{\phi}(u)$. To this end, consider the derivative of $p'_w(u)$ w.r.t. u

$$p''_w(u) = \tilde{w}^\top C\tilde{\phi}'(u) = 2w_2 + 6w_3 u + \cdots + d(d-1)w_d u^{d-2}.$$

$p''_w(u)$ is a polynomial of degree at most $d-2$. We define a set of points $\mathcal{R}$ containing all the real roots of $p''_w(u) = 0$ lying in the interval $[-1, 1]$, plus the extremes of this interval (thus, $\mathcal{R}$ contains at most d real numbers). The optimum we seek is achieved at one of the values in $\mathcal{R}$, hence

$$u^* = \arg\min_{u\in\mathcal{R}} |p'_w(u)|,$$

which is readily computed by evaluating $|p'_w(u)|$ at all the points in $\mathcal{R}$. Then, applying the rule of max for the subgradients[46], we have that a subgradient for b at w is given by

$$g_b(w) = (0, \operatorname{sgn}(\tilde{w}^\top C\tilde{\phi}(u^*))C\tilde{\phi}(u^*)). \tag{12.34}$$

3. A subgradient for the objective function in (12.33) at w is given by

$$g(w) = 2\Phi(\Phi^\top w - y) + \lambda g_b(w),$$

where $g_b(w)$ is given by (12.34). A subgradient algorithm for solving problem (12.33) is thus specified as follows[47]: initialize with $k = 0$ and some $w^{(0)} \in \mathbb{R}^{d+1}$, and iterate until convergence

$$\begin{array}{rcl} g_k & = & g(w^{(k)}) \\ s_k & = & \gamma/(k+1) \\ w^{(k+1)} & = & w^{(k)} - s_k g_k \\ k & \leftarrow & k+1, \end{array}$$

where $\gamma > 0$ is an algorithm parameter, e.g., $\gamma = 0.1$.

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 sub-gradient method, as in Section 12.4.1.

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

Solution 12.7 (Methods for LASSO) Exercise seems doable and OK.

[46] See Section 8.2.3.1.

[47] See Section 12.4.1.

Exercise 12.8 (Nonnegative 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.

Solution 12.8 (Nonnegative terms that sum to one) Let us denote with s_k the sum of the k-th largest elements in $x = (x_1, \ldots, x_n)$, that is, since x is ordered, $s_k = \sum_{i=n-k+1}^{n} x_i$, thus $s_1 = x_n$, $s_n = \sum_{i=1}^{n} x_i$.

1. If $\nu \geq x_n$, then $x_i - \nu \leq 0$ for all i, hence $f(\nu) = 0$. If $\nu \leq x_1$, then $x_i - \nu \geq 0$ for all i, hence $f(\nu) = s_n - n\nu$. Therefore, f is linear and decreasing at rate n for $\nu \leq x_1$, and it is identically zero for $\nu \geq x_n$. Further, if $\nu \in (x_k, x_{k+1}]$, $k = 1, \ldots, n-1$, then we readily see that $f(\nu) = s_{n-k} - (n-k)\nu$, hence it is linear and decreasing with rate $(n-k)$ in each of these intervals. The limits on the left and on the right at each break-point x_k coincide (since $s_{n-(k-1)} - (n-(k-1))x_k = s_{n-(k-1)} - (n-k)x_k + x_k = s_{n-k} - (n-k)x_k$), thus f is piece-wise linear, continuous, and strictly decreasing for $\nu \leq x_n$.

2. From the previous point we have that

$$f(x_1 - 1/n) = s_n - n(x_1 - 1/n) = s_n - nx_1 + 1,$$

thus

$$\frac{f(x_1 - 1/n)}{n} = \left(\frac{s_n}{n} - x_1\right) + \frac{1}{n}.$$

Since the average of the points s_n/n is no smaller than the minimum x_1 of the points, we have that $\frac{f(x_1-1/n)}{n} \geq 1/n$ and hence $f(x_1 - 1/n) \geq 1$. Since we have also that $f(x_n) = 0$, and since f is continuous and strictly decreasing, we conclude that there exist a unique point ν in the interval $[x_1 - 1/n,\, x_n]$ for which $f(\nu) = 1$.

3. Evaluate the function at the following n points:

$$f_0 \doteq f(x_1 - 1/n),\; f_1 \doteq f(x_1),\; \ldots,\; f_{n-1} \doteq f(x_{n-1}).$$

We have $f_0 = s_n - nx_1 + 1$, and $f_k = s_{n-k} - (n-k)x_k$, for $k = 1, \ldots, n-1$. Since f is decreasing, we have that $f_0 > f_1 \geq f_2 \geq \cdots \geq f_{n-1} \geq 0$. Then, find the largest index $k \in \{0, \ldots, n-1\}$ such that $f_k \geq 1$, and call it k^*. The optimal ν is

$$\nu^* = \frac{s_{n-k^*} - 1}{n - k^*}.$$

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: 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 12.33. Express the Hessian of $\tilde{f}_0$ in terms of that of f_0.

2. Compare the computational effort[48] 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 $m \ll n$.

Solution 12.9 (Eliminating linear equality constraints)

1. Since f_0 is separable, we have that

$$\nabla f_0(x) = \sum_{i=1}^{n} \frac{\partial h_i}{\partial x_i}(x_i) = \begin{bmatrix} \frac{\partial h_1}{\partial x_1}(x_1) \\ \vdots \\ \frac{\partial h_n}{\partial x_n}(x_n) \end{bmatrix},$$

and

$$\nabla^2 f_0(x) = \text{diag}\left(\frac{\partial^2 h_1}{\partial x_1^2{}^2}(x_1), \ldots, \frac{\partial^2 h_n}{\partial x_n^2{}^2}(x_n) \right).$$

Let $N \in \mathbb{R}^{n,n-m}$ contain by columns a basis for the nullspace of A, then all x satisfying the equality constraints are written as

$$x = \bar{x} + Nz$$

[48] See the related Exercise 7.4.

where $\bar{x} \in \mathbb{R}^n$ is some fixed solution of $Ax = b$, and $z \in \mathbb{R}^{n-m}$ is a new free variable. The minimization problem becomes now unconstrained in the variable z:

$$\min_z \tilde{f}_0(z),$$

where $\tilde{f}_0(z) = f_0(\bar{x} + Nz)$. The Hessian of $\tilde{f}_0$ is obtained using the result in point 3. of Exercise 3.1, i.e.,

$$\nabla^2 \tilde{f}_0(z) = N^\top \nabla^2 f_0(x(z)) N.$$

We observe that the Hessian of $\tilde{f}_0$ is in general a full symmetric matrix, while the Hessian of f_0 is diagonal.

2. The standard Newton method applied at $\tilde{f}_0(z)$ is based on the basic iteration

$$z_{k+1} = z_k + s_k h_k,$$

where h_k solves the full symmetric system of dimension $n - m$

$$[\nabla^2 \tilde{f}_0(z_k)] h_k = -\nabla \tilde{f}_0(z_k).$$

The feasible update Newton method iterates instead directly on x

$$x_{k+1} = x_k + s_k h_k,$$

where h_k solves the augmented system of dimension $m + n$

$$\begin{bmatrix} \nabla^2 f_0(x_k) & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} h_k \\ \lambda \end{bmatrix} = \begin{bmatrix} -\nabla f_0(x_k) \\ 0 \end{bmatrix},$$

where $\nabla^2 f_0(x_k)$ is diagonal. The solution of this latter system can be expressed as

$$h_k = -[\nabla^2 f_0(x_k)]^{-1}(\nabla f_0(x_k) + A^\top \lambda),$$

where

$$\lambda = -(A[\nabla^2 f_0(x_k)]^{-1} A^\top)^{-1} A [\nabla^2 f_0(x_k)]^{-1} \nabla f_0(x_k)$$

This approach can be computationally more convenient than the first one when $m \ll n$, since it only requires inversion of the $m \times m$ matrix $A[\nabla^2 f_0(x_k)]^{-1} A^\top$ (the Hessian is readily inverted, since it is diagonal).

13. Learning from Data

Exercise 13.1 (SVD for text analysis) Assume you are given a data set in the form of a $n \times m$ 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 to 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.

Solution 13.1

1. Denote by d_j the j-th column of A, which corresponds to particular news source. Thus $A = [d_1, \dots, d_n]$. The SVD of A can be written as
$$A = \sum_{l=1}^{r} \sigma_l u_l v_l^\top,$$
where r is the rank of A ($r \leq \min(m, n)$), and the singular values are ordered, so that $\sigma_1 \geq \dots \geq \sigma_r$.

   Since the j-th news source is $d_j = Ae_j \in \mathbb{R}^m$ with e_j the j-th unit vector in $\mathbb{R}^n$, we have
$$d_j = Ae_j = \sum_{l=1}^{r} \sigma_l u_l v_l^\top e_j = \sum_{l=1}^{r} \alpha_l u_l,$$
where $\alpha_l \doteq \sigma_l v_l^\top e_j$, $l = 1, \dots, r$. Since the u_l's form an orthonormal basis, the numbers α_l are simply the components of d_j along u_l. That is, $\alpha_l = d_j^\top u_l$ for every l. We can use this for plotting *any* news source $d \in \mathbb{R}^m$, even if it is not any one of the columns of A, using the numbers $(d^\top u_1, d^\top u_2)$.

   The particular projection on (u_1, u_2) is of maximal variance, in the sense that u_1 is a direction that captures the largest variance, if we assume the data to be centered ($d_1 + \dots + d_n = 0$). Indeed, if we consider the *columns* of A to be data points then the associated covariance matrix is the $m \times m$ matrix:
$$\Sigma = \frac{1}{n} \sum_{j=1}^{n} d_j d_j^\top = \frac{1}{n} A A^\top.$$

We can write $n\Sigma = AA^\top = US^2U^\top$. The direction of maximal variance is the solution to

$$\max_{u\,:\,\|u\|_2 \leq 1} \quad u^\top \Sigma u,$$

a solution of which is u_1, the first column of U. Similarly u_2 is the direction of maximal variance once we have projected the data on the hyperplane orthogonal to u_2.

2. This second question simply amounts to replace A with its transpose. In fact, if $A = USV^\top$ is the SVD of A, then that of $A^\top$ is $A^\top = VS^\top U^\top$, with $S^\top$ still diagonal, so the roles of U, V are simply reversed.

To *summarize:* to plot a news source vector $d \in \mathbb{R}^m$ in word space, we simply project that vector on the first two columns of U, u_1, u_2, that is, we use the coordinates $(d^\top u_1, d^\top u_2)$. To plot a word vector $w \in \mathbb{R}^n$ in news source space, we use $(w^\top v_1, w^\top v_2)$.

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: $x^{(1)} + \ldots + x^{(m)} = 0$. An (empirical) estimate of the covariance matrix is[49]

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

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 a $n \times k$ 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 is 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 data, we seek to solve

$$\min_{F, \lambda \geq 0} \|\Sigma - \lambda I - FF^\top\|_F. \tag{13.35}$$

[49] See Example 4.2.

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 $F(\lambda)$. In other words: you are asked to solve for F, with fixed λ.

2. Show that the error $E(\lambda) = \|\Sigma - \lambda I - F(\lambda)F(\lambda)^\top\|_F$, 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 (13.35).

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 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.

Solution 13.2

1. Let $\Sigma = U\Lambda U^\top$ be an eigenvalue decomposition of Σ, with U unitary, and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. The eigenvalue decomposition of $\Sigma - \lambda I$ is then $\Sigma - \lambda I = U(\Lambda - \lambda I)U^\top$. The best rank-$k$ approximation to that matrix is then
$$\hat{\Sigma}_k \doteq \sum_{i=1}^{k} (\lambda_i - \lambda) u_i u_i^\top.$$
Since $\lambda \le \lambda_k$, we see that the above is positive semi-definite. It can be written as $F(\lambda)F(\lambda)^\top$, with
$$F(\lambda) = \begin{bmatrix} \sqrt{\lambda_1 - \lambda}\, u_1 & \cdots & \sqrt{\lambda_k - \lambda}\, u_k \end{bmatrix}.$$

2. From the previous part, it follows that
$$\begin{aligned} E(\lambda)^2 &= \|\Sigma - \lambda I - F(\lambda)F(\lambda)^\top\|_F^2 \\ &= \left\| \sum_{i=k+1}^{n} (\lambda_i - \lambda) u_i u_i^\top \right\|_F^2 \\ &= \sum_{i=k+1}^{p} (\lambda_i - \lambda)^2, \end{aligned}$$

as claimed. For the optimal λ, the first order conditions give us

$$\lambda^* = \frac{1}{n-k} \sum_{i=k+1}^{p} \lambda_i.$$

We do check that $\lambda^* \leq \lambda_k$; hence λ^* is indeed optimal for the problem (13.35).

The solution approach to that problem is:

(a) Form the sample covariance matrix.

(b) Set λ^* to be the average of the $n-k$ smallest eigenvalues.

(c) Find the corresponding factor loading matrix $F_* \doteq F(\lambda^*)$.

3. For any direction $w \in \mathbb{R}^n$, with $\hat{\Sigma}_k$ the rank-k estimate:

$$w^\top (\Sigma - \hat{\Sigma}_k) w = w^\top \left(\sum_{i=k+1}^{n} \lambda_i u_i u_i^\top \right) w \geq 0.$$

Using the other estimate, denoting by (λ^*, F_*) the optimal values found before:

$$\begin{aligned} w^\top (\Sigma - (\lambda^* I + F_* F_*^\top)) w &= w^\top \left(\sum_{i=k+1}^{n} (\lambda_i - \lambda^*) u_i u_i^\top \right) w \\ &= \sum_{i=k+1}^{n} (\lambda_i - \lambda^*)(w^\top u_i)^2. \end{aligned}$$

The sign of the error depends on w. For example, with $w = e_{k+1}$, the error is non-negative, and the factor model under-estimates the risk along w; for $w = u_n$, the opposite is true.

Exercise 13.3 (Movement prediction for a time-series) We have a historical data set containing the values of a times 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

$$\hat{y}_{w,b}(t) = \text{sgn}\left(w_1 r(t) + \ldots + w_n r(t-n+1) + b \right),$$

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 $y(t) = \mathrm{sgn}(r(t+1) - r(t))$. 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.

Solution 13.3

1. The problem is not convex in (w, b).

2. We can use a support vector machine for example. For $t \in \{1, \ldots, T-1\}$, we set $y_t \in \{-1, 1\}$ to be the sign of $y(t) = \mathrm{sgn}(r(t+1) - r(t))$, and we define the "feature vector" at time t to be

$$x_t = (r(t), \ldots, r(t-n+1)), \ \ t = n, \ldots, T-1.$$

We try to find w, b so that, on the available data, the sign predictions match. If that is so, then w, b must satisfy

$$y(t)(w^\top x_t + b) \geq 0, \ \ t = n, \ldots, T-1.$$

The next steps are entirely similar to those taken in Section 13.3.1. Once an optimal pair (w, b) is found, our prediction at time T is

$$\hat{y}(T) = \mathrm{sgn}\left(w_1 r(t) + \ldots + w_n r(t-n+1) + b\right).$$

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.

Solution 13.4 To be completed later. This exercise should be correct.

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

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

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

$$p(\lambda) = \min_{\mu>0} q(\mu) + \frac{\lambda^2}{2\mu}, \quad q(\mu) = \max_{\lambda>0} p(\lambda) - \frac{\lambda^2}{2\mu},$$

   For the second expression, you may assume that $\mathbf{dom} f$ has non-empty interior.

3. Deduce from the first part that the path 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?

Solution 13.5

1. p, q are concave, since they are point-wise minima of affine functions of their argument. We have

$$\tilde{q}(\mu) = \min_x g(x,\mu) \text{ where } g(x,\mu) \doteq f(x) + \frac{1}{2\mu}\|x\|^2.$$

   The function g is jointly convex with respect to (x, μ) on the convex domain $\mu > 0$, $x \in \mathbb{R}^n$. Indeed, thanks to the Schur complement Lemma, the condition $g(x,\mu) \leq t$ for $\mu > 0$ can be written as a linear matrix inequality in (x, μ, t):

$$\begin{pmatrix} 2\mu & x^\top \\ x & tI \end{pmatrix} \succeq 0.$$

2. The first expression comes from the fact that, for every $\gamma, \lambda \in \mathbb{R}_+$:

$$\min_{\mu>0} \frac{\mu\gamma^2}{2} + \frac{\lambda^2}{2\mu} = \gamma\lambda,$$

with $\mu = \lambda/\gamma$ at optimum. Hence:

$$\begin{aligned} \min_{\mu>0} q(\mu) + \frac{\lambda^2}{2\mu} &= \min_{x,\mu>0} f(x) + \frac{1}{2}\mu\|x\|^2 + \frac{\lambda^2}{2\mu} \\ &= \min_x f(x) + \lambda\|x\| \\ &= p(\lambda). \end{aligned}$$

At optimum, we have $\mu\|x\| = \lambda$.

For the second expression, we use the fact that for every $\gamma \geq 0$, $\mu > 0$:

$$\max_{\lambda\geq 0} \lambda\gamma - \frac{\lambda^2}{2\mu} = \frac{\mu}{2}\gamma^2,$$

with $\lambda \ = \gamma\mu$ at optimum. We have then

$$\begin{aligned} \max_{\lambda\geq 0} p(\lambda) - \frac{\lambda^2}{2\mu} &= \max_{\lambda\geq 0} \min_x f(x) + \lambda\|x\| - \frac{\lambda^2}{2\mu} \\ &= \min_x \max_{\lambda\geq 0} f(x) + \lambda\|x\| - \frac{\lambda^2}{2\mu} \\ &= \min_x f(x) + \frac{1}{2}\mu\|x\|^2 \\ &= q(\mu). \end{aligned} \tag{13.36}$$

In the third line, we use strong duality. This is allowed if $\mathbf{dom} f$ has non-empty interior. Again, at optimum we have $\mu\|x\| = \lambda$.

3. We first prove that for any $\mu > 0$, there exists a value of λ such that the (unique) solution to $P(\lambda)$, $x^\star(\lambda)$, is optimal for $Q(\mu)$. We now show that one such value is given by any maximizer $\lambda^\star(\mu)$ for the problem

$$\max_{\lambda\geq 0} p(\lambda) - \frac{\lambda^2}{2\mu}.$$

Indeed, defining $y^\star = x^\star(\lambda^\star(\mu))$, at optimum of the above problem we have

$$\mu\|y^\star\| = \lambda^\star(\mu).$$

Thus:

$$\begin{aligned} q(\mu) &= p(\lambda^\star(\mu)) - \frac{\lambda^\star(\mu)^2}{2\mu} \\ &= f(y^\star) + \lambda^\star(\mu)\|y^\star\| - \frac{\lambda^\star(\mu)^2}{2\mu} \\ &= f(y^\star) + \frac{\mu}{2}\|y^\star\|^2. \end{aligned}$$

This shows that $y^\star$ is optimal for $Q(\mu)$.

Likewise, we now show that any $\lambda \geq 0$, there exists a value of μ such that the (unique) solution to $Q(\mu)$, $z^\star(\mu)$, is optimal for $Q(\mu)$. We now show that one such value is given by any maximizer $\mu^\star(\lambda) > 0$ for the problem

$$\min_{\mu > 0} q(\mu) + \frac{\lambda^2}{2\mu}.$$

Indeed, defining $y^\star = z^\star(\mu^\star(\lambda))$, at optimum of the above problem we have

$$\mu^\star(\lambda)\|y^\star\| = \lambda.$$

Thus:

$$\begin{aligned} p(\lambda) &= q(\mu^\star(\lambda)) + \frac{\lambda^2}{2\mu^\star(\lambda)} \\ &= f(y^\star) + \frac{\mu^\star(\lambda)}{2}\|y^\star\|^2 - \frac{\lambda^2}{2\mu^\star(\lambda)} \\ &= f(y^\star) + \lambda\|y^\star\|. \end{aligned}$$

This shows that $y^\star$ is optimal for $P(\lambda)$.

4. Similar statements can be made, as strong duality holds. Precisely, we have for $\kappa > 0$

$$\begin{aligned} r(\kappa) &= \min_x \max_{\lambda \geq 0} f(x) + \lambda(\|x\| - \kappa) \\ &= \max_{\lambda \geq 0} \min_x f(x) + \lambda(\|x\| - \kappa) \\ &= \max_{\lambda \geq 0} p(\lambda) - \lambda\kappa. \end{aligned}$$

Here, strong duality can be invoked, since the problem $R(\kappa)$ is strictly feasible. Likewise, since

$$r(\kappa) = \min_x f(x) \ : \ \frac{1}{2}\|x\|^2 \leq \frac{1}{2}\kappa^2,$$

applying strong duality we obtain

$$\begin{aligned} r(\kappa) &= \min_x \max_{\mu \geq 0} f(x) + \frac{\mu}{2}(\|x\|^2 - \kappa^2) \\ &= \max_{\mu \geq 0} \min_x f(x) + \frac{\mu}{2}(\|x\|^2 - \kappa^2) \\ &= \max_{\mu \geq 0} q(\mu) - \frac{\mu\kappa^2}{2}. \end{aligned}$$

5. If we remove the uniqueness assumption, a similar statement can be made. The solution to $Q(\mu)$ is unique for any $\mu > 0$, but that

of $P(\lambda)$ may not be. However, we can show that for any $\lambda > 0$, there exist a $\mu > 0$ such that one of the optimal points for $P(\lambda)$ is optimal for $Q(\mu)$.

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

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

where $X \in \mathbb{R}^{n,m}$, $y \in \mathbb{R}^m$, $\rho > 0$ is a regularization parameter, and $\lambda \geq 0$ allows to control the cardinality (number of non-zeros) in the solution. This in turn allows better interpretability of the results. The above problem is hard to solve in general. In the sequel, 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 sub-optimal 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 sub-optimal 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.

Solution 13.6

1. We first find an expression for the regularized LS problem, which corresponds to $\lambda = 0$:

$$\begin{aligned} \phi(0) &= \min_x \|X^\top w - b\|_2^2 + \rho^2 \|w\|_2^2 \\ &= \max_\tau \tau \ : \ \forall\, w, \ \ \tau \le \|X^\top w - b\|_2^2 + \rho^2 w^\top w. \end{aligned}$$

The constraint holds for every w if and only if

$$\forall\, w \ : \ \left(\begin{array}{c} w \\ -1 \end{array}\right)^\top \left(\begin{array}{cc} XX^\top + \rho^2 I_n & Xy \\ (Xy)^\top & y^\top y - \tau \end{array}\right) \left(\begin{array}{c} w \\ -1 \end{array}\right) \ge 0.$$

In turn the above is equivalent to

$$\left(\begin{array}{cc} XX^\top + \rho^2 I_n & Xy \\ y^\top X^\top & y^\top y - \tau \end{array}\right) \succeq 0.$$

Using Schur complements, and a matrix inversion formula, we obtain

$$\phi(0) = y^\top y - y^\top X^\top (XX^\top + \rho^2 I_n)^{-1} Xy = y^\top (\frac{1}{\rho^2} X^\top X + I_m)^{-1} y.$$

2. We have

$$\phi(\lambda) = \min_{u,z} \|X^\top D(u) z - y\|_2^2 + \rho^2 z^\top D(u) z + \lambda \mathbf{1}^\top u \ : \ u \in \{0,1\}^n,$$

where $D(u) = \mathbf{diag}\,(u)$, and using $D(u)^2 = D(u)$. In fact, the problem above is equivalent to

$$\min_{u,x} \|X^\top D(u) z - y\|_2^2 + \rho^2 z^\top z + \lambda \mathbf{1}^\top u \ : \ u \in \{0,1\}^n,$$

since the components of z corresponding to the zero components of u do not appear in the first term in the objective, and are driven down to zero due to the second term. We can use the expression we found above for the regularized LS problem, in order to eliminate z. This simply entails replacing X with $D(u)X$ in the expression of $\phi(0)$ above. We obtain the Boolean formulation:

$$\phi(\lambda) = \min_{u \in \{0,1\}^n} \lambda \mathbf{1}^\top u + y^\top \mathcal{A}(u)^{-1} y,$$

where

$$\mathcal{A}(u) = \frac{1}{\rho^2} X^\top D(u)^2 X + I_m = \frac{1}{\rho^2} X^\top D(u) X + I_m = I_m + \frac{1}{\rho^2} \sum_{i=1}^n u_i a_i a_i^\top,$$

with $a_i^\top \in \mathbb{R}^m$ is the i-th row of X.

Alternatively the problem reads

$$\phi(\lambda) = \min_{u \in \{0,1\}^n} \max_v \lambda \mathbf{1}^\top u + 2y^\top v - v^\top \mathcal{A}(u) v,$$

with $v = \mathcal{A}(u)^{-1} y$ at optimum.

3. A natural relaxation involves replacing the Boolean constraints with interval ones. This amounts to exchange the min and the max in the above expression. We obtain the convex *lower* bound $\phi(\lambda) \geq \underline{\phi}(\lambda)$, where

$$\begin{aligned} \underline{\phi}(\lambda) &= \max_v \min_{u \in \{0,1\}^n} \lambda \mathbf{1}^\top u + 2y^\top v - v^\top \left(I_m + \frac{1}{\rho^2} \sum_{i=1}^n u_i a_i a_i^\top \right) v \\ &= \max_v 2y^\top v - v^\top v - \sum_{i=1}^n \left(\frac{(a_i^\top v)^2}{\rho^2} - \lambda \right)_+ . \end{aligned}$$

A sub-optimal sparsity pattern can be inferred from the solution to the above problem, setting $u_i = 0$ if $\lambda \rho^2 \geq (a_i^\top v)^2$, and $u_i = 1$ otherwise.

4. We can express the problem as an SOCP with rotated second-order cone constraints. Using the fact that, for every vector z and scalars r, t:

$$z^\top z \leq rt, \;\; r \geq 0, \;\; t \geq 0 \Longleftrightarrow \left\| \begin{pmatrix} 2z \\ r - t \end{pmatrix} \right\|_2 \leq r + t,$$

we obtain

$$\begin{aligned} \underline{\phi}(\lambda) &= \max_{v,p} 2y^\top v - v^\top v - p^\top \mathbf{1} \; : \; p \geq 0, \;\; p_i \geq \frac{1}{\rho^2} (a_i^\top v)^2 - \lambda, \;\; i = 1, \ldots, n \\ &= \max_{v,p} 2y^\top v - v^\top v - p^\top \mathbf{1} \; : \; p \geq 0, \;\; \left\| \begin{pmatrix} 2\rho^{-1} a_i^\top v \\ p_i + \lambda - 1 \end{pmatrix} \right\|_2 \leq p_i + \lambda + 1, \;\; i = 1, \ldots, n. \end{aligned}$$

5. We proceed with the dual to the above SOCP:

$$\begin{aligned} \underline{\phi}(\lambda) &= \max_{v,p \geq 0} \min_{\alpha,\beta,\gamma} \; 2y^\top v - v^\top v - p^\top \mathbf{1} + \sum_{i=1}^n \Big(\gamma_i (p_i + \lambda - 1) - 2\rho^{-1} \alpha_i a_i^\top v - \beta_i (p_i + \lambda + 1) \Big) \\ &\qquad\qquad \text{s.t.} \quad \|(\alpha_i, \beta_i)\|_2 \leq \gamma_i, \;\; i = 1, \ldots, n \\ &= \min_{\alpha,\beta,\gamma} \max_{v,p \geq 0} 2(y - \rho^{-1} X^\top \alpha)^\top v - v^\top v + p^\top ((\gamma - \beta - \mathbf{1}) + \mathbf{1}^\top (\lambda(\gamma - \beta) - (\gamma + \beta)) \\ &\qquad \text{s.t.} \quad \|(\alpha_i, \beta_i)\|_2 \leq \gamma_i, \;\; i = 1, \ldots, n. \end{aligned}$$

Setting $w = \rho^{-1}\alpha$, $s \doteq \gamma + \beta$, $z \doteq \gamma - \beta$, and eliminating $v = b - X^\top w$, we obtain

$$\begin{array}{rcll}
\underline{\phi}(\lambda) & = & \min\limits_{w,s,z} & \max\limits_{p \geq 0} \|X^\top w - y\|_2^2 + p^\top(z - \mathbf{1}) + \mathbf{1}^\top(\lambda z + s) \\
& & \text{s.t.} & \|(2\rho w_i, s_i - z_i)\|_2 \leq s_i + z_i, \;\; i = 1, \ldots, n \\
& = & \min\limits_{w,s,z} & \|X^\top w - b\|_2^2 + \mathbf{1}^\top(\lambda z + s) \\
& & \text{s.t.} & 0 \leq z_i \leq 1, \;\; s_i \geq 0, \;\; \rho^2 x_i^2 \leq s_i z_i, \;\; i = 1, \ldots, n \\
& = & \min\limits_{w,z} & \|X^\top w - b\|_2^2 + \sum\limits_{i=1}^n \left(\lambda z_i + \dfrac{\rho^2 w_i^2}{z_i} \right) \;:\; 0 \leq z \leq \mathbf{1}.
\end{array}$$

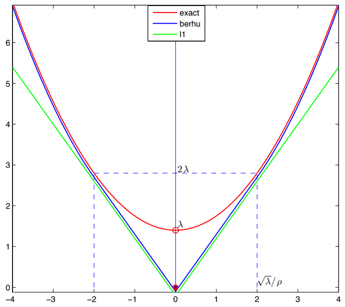

Figure 13.14: The relaxation consists in replacing the non-convex penalty (red) with its convex envelope (blue), which corresponds to the reverse Huber function. The classical ℓ_1 penalty (green) is a lower bound only if we choose the penalty parameter to be $2\rho\sqrt{\lambda}$.

We obtain that our relaxation is in the form

$$\underline{\phi}(\lambda) = \|X^\top w - b\|_2^2 + 2\lambda \sum_{i=1}^n B(\frac{\rho x_i}{\sqrt{\lambda}}).$$

We can extract a sparsity pattern from the above, setting $v = y - X^\top w$, and $u_i = 0$ if $\lambda\rho^2 \geq (a_i^\top v)^2$, and $u_i = 1$ otherwise.

We observe that we have the following bound:

$$\rho^2\xi^2 + \lambda\mathbf{1}(\xi) \geq 2\lambda B(\frac{\rho\xi}{\sqrt{\lambda}}) \geq 2\rho\sqrt{\lambda}|\xi|,$$

where $\mathbf{1}(\cdot)$ is zero when its argument is zero, and 1 otherwise. This is illustrated in Fig. 13.14.

Hence

$$\phi(\lambda) \geq \underline{\phi}(\lambda) \geq \psi(\lambda) \doteq \min_x \|X^\top w - b\|_2^2 + 2\rho\sqrt{\lambda}\|x\|_1,$$

which is the classical approximation. Note that the penalty in the classical approximation is not the same as that on the l_0 norm. Our approach provides a natural guideline for how to threshold the result when solving for the relaxation, namely that we should zero out elements of x with magnitude less than $\sqrt{\lambda}/\rho$.

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}.

Solution 14.1

1. The constraint reads $x^\top Cx = t$. This is not a convex constraint, since it is an *equality* involving a non-linear function.

2. The constraint reads $x^\top Cx \leq t$. This is a convex constraint, since C is positive semi-definite, hence the function $f(x) = x^\top Cx - t$ is convex.

3. The Sharpe ratio is $\hat{r}^\top x / \sqrt{x^\top Cx}$. The constraint writes

$$\hat{r}^\top x \geq t\sqrt{x^\top Cx}.$$

   This is a second-order cone constraint. Indeed, since C is positive semi-definite, there exist $R \in \mathbb{R}^{n,n}$ such that $C = R^\top R$. Then the constraint above can be written as: the Euclidean norm of a vector affine in x is less than some affine function of x:

$$\hat{r}^\top x \geq t\|Rx\|_2.$$

   Since $t \geq 0$, this constraint is convex.

4. As seen in Prop. 10.1, the constraint can be expressed as

$$\hat{r}^\top x \geq \kappa\sqrt{x^\top Cx}$$

   where $\kappa > 0$ is a function of ϵ. Similar to the previous part, this constraint is a (convex) second-order cone constraint.

5. This is convex, as it can be represented as three affine constraint $x^\top r^{(i)} \geq t$, $i = 1, 2, 3$.

6. Here we introduce new variables s_i, $i = 1, 2, 3$ and add the affine equality constraints $s_i = x^\top r^{(i)}$, $i = 1, 2, 3$. It remains to examine if the constraint on $s \in \mathbb{R}^3$

$$\frac{1}{2}(s_{[2]} + s_{[3]}) \geq t$$

is convex, where for $k = 1, 2, 3$, $s_{[k]}$ denotes the k-th largest element in s. The above constraint is indeed convex, since the function on the LHS is concave. To see this, we write

$$s_{[2]} + s_{[3]} = \min(s_1 + s_2, s_2 + s_3, s_3 + s_1).$$

Hence the function $s \to s_{[2]} + s_{[3]}$ is the point-wise minimum of linear functions, thus it is concave.

7. This is the convex constraint $\|x\|_1 \leq t$.

8. This is a non-convex constraint.

9. This is the convex constraint $|\hat{r}^\top x - t| \leq \epsilon$.

10. Here we introduce new variable s, and add the linear equality constraint $s = \hat{r}^\top x$. The constraint can be written as the quadratic constraint on s, x

$$(t_{\rm up} - s)(s - t_{\rm low}) \leq 0, \;\; s = \hat{r}^\top x.$$

To check that this is not convex, we observe that the first constraint can be written as the non-convex constraint $q(s) \geq 0$, where $q(s) = s^2 +$ (a linear function of s). (In x-space, the constraint above represents the *outside* of a slab.)

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 with $\hat{r} \doteq (1/N)(r_1 + \ldots + r_N)$ 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 a $n \times N$ 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

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

Solution 14.2 (Median risk)

1. We have

$$\rho_t(x) = r_t^\top x, \;\; t = 1, \ldots, N, \;\; \hat{\rho}(x) = \hat{r}^\top x.$$

   Hence,

$$\mathcal{R}_1(x) = \frac{1}{N}\sum_{t=1}^{N} |(r_t - \hat{r})^\top x| = \|R^\top x\|_1,$$

   where R is the $n \times N$ matrix with t-th colum equal to $(1/N)(r_t - \hat{r})$, $t = 1, \ldots, N$. The risk measure is a convex function of x, as the composition of the ℓ_1-norm with an affine transformation $x \to R^\top x$.

2. The problem writes

$$\min_x \; \|R^\top x\|_1 \; : \; \hat{r}^\top x \geq \mu.$$

   We can put the above problem in standard LP format:

$$\min_{x,z} \sum_{t=1}^{N} z_t \; : \; \begin{array}{l} -Nz_t \leq (r_t - \hat{r})^\top x \leq Nz - t, \;\; t = 1, \ldots, N, \\ \hat{r}^\top x \geq \mu. \end{array}$$

3. In practice the resulting portfolio would have most of the deviations $\rho_t(x) - \hat{\rho}(x)$ very small in magnitude, and tolerate a few that are potentially very large. In constrast, the variance-based measure would tend to produce a deviation vector that is more evenly distributed.

Exercise 14.4 (Portfolio optimization with factor models – 1)

1. Consider the following portfolio optimization problem

$$\begin{array}{rl} p^* = \min\limits_x & x^\top \Sigma x \\ \text{s.t.:} & \hat{r}^\top x \geq \mu, \end{array}$$

   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

$$\begin{array}{rl} p^* = \min\limits_x & x^\top \Sigma x - \gamma \hat{r}^\top x \\ \text{s.t.:} & x \geq 0, \end{array}$$

where $\gamma > 0$ is a tradeoff parameters that weigths 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.

Solution 14.3 (Portfolio optimization with factor model – 1)

1. Since $r = \hat{r} + Ff$, and $\mathbb{E}\{f\} = 0$, $\mathbb{E}\{f^\top f\} = I$, we have that $\Sigma = \text{var}\{r\} = FF^\top$. Thus the original optimization problem is

$$p^* = \min_x \quad x^\top FF^\top x \quad \text{s.t.: } \hat{f}^\top F^\top x \geq \mu.$$

By defining a new variable $z = F^\top x$, this problem reduces to

$$p^* = \min_z \quad z^\top z \quad \text{s.t.: } \hat{f}^\top z \geq \mu.$$

This problem has $k \ll n$ variables. Moreover, the problem has the geometric interpretation of finding the projection of the origin onto the halfspace $\{\hat{f}^\top z \geq \mu\}$, which has the unique optimal solution

$$z^* = \frac{\mu}{\|\hat{f}\|_2^2}\hat{f}.$$

All optimal solutions to the original problem solve $F^\top x = z^*$, and are thus given by

$$x^* = (F^\top)^\dagger z^* + v, \quad v \in \mathcal{N}(F^\top).$$

If F is full rank, then we may write $(F^\top)^\dagger = F(F^\top F)^{-1}$, and

$$x^* = \frac{\mu}{\|\hat{f}\|_2^2} F(F^\top F)^{-1}\hat{f} + v, \quad v \in \mathcal{N}(F^\top).$$

2. For the proposed factor model, we have that

$$\Sigma = \mathrm{var}\{r\} = D^2 + FF^\top,$$

and the problem objective becomes

$$f_0(x) = x^\top D^2 x + x^\top F F^\top x - \gamma \hat{r}^\top x.$$

The composite objective in a barrier method for this problem would be of the form

$$t f_0(x) + \phi(x), \quad \phi(x) = -\sum_i \ln x_i,$$

where the Hessian of ϕ is $\nabla^2 \phi(x) = \mathrm{diag}\left(x_1^{-2}, \ldots, x_n^{-2}\right)$. However, the Hessian of f_0 is Σ, which is not diagonal, in general.

Let us introduce slack variables $z \doteq F^\top x$. The problem now becomes

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

We can solve this problem again via a barrier method, using feasible updates[50] in the Newton iterations to deal with the equality constraint $F^\top x = z$. The advantage is that the Hessian of the objective function is now

$$\nabla^2 f_0(x,z) = \mathrm{diag}\left(D^2, I\right),$$

[50] See Section 12.2.6.3.

which is diagonal, thus readily invertible. The Newton step Δ is then found by solving the KKT system of the form

$$\begin{bmatrix} H & \tilde{F} \\ \tilde{F}^\top & 0 \end{bmatrix} \begin{bmatrix} \Delta \\ \eta \end{bmatrix} = \begin{bmatrix} -g \\ 0 \end{bmatrix},$$

where H and ξ are, respectively, the Hessian (which is diagonal) and the gradient of the composite objective function at (x, z), and $\tilde{F}^\top \doteq [F^\top \ -I] \in \mathbb{R}^{k,n+k}$. From this system, we obtain η by solving

$$(\tilde{F}^\top H^{-1} \tilde{F})\eta = -\tilde{F}^\top H^{-1} \xi,$$

where the matrix on the right-hand side has dimension $k \times k$ (hence, much smaller than $n \times n$), and then obtain Δ as

$$\Delta = -H^{-1}(\xi + \tilde{F}\eta).$$

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

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

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 $-g(\lambda)$.

Solution 14.4 (Portfolio optimization with factor model – 2)

1. The Lagrangian is strongly convex in (x, z), and it is decomposable as

$$\mathcal{L}(x, z, \lambda) = \sum_{i=1}^{k} \left(z_i^2 + \lambda_i z_i \right) + \sum_{i=1}^{n} \left(d_i^2 x_i^2 - (\gamma \hat{r}_i + [F\lambda]_i) x_i \right),$$

thus

$$\begin{aligned} g(\lambda) &= \min_{x\geq 0,z} \mathcal{L}(x,z,\lambda) \\ &= \min_{x\geq 0,z} \sum_{i=1}^{k} \left(z_i^2 + \lambda_i z_i\right) + \sum_{i=1}^{n} \left(d_i^2 x_i^2 - (\gamma\hat{r}_i + [F\lambda]_i)x_i\right) \\ &= \min_{z} \sum_{i=1}^{k} \left(z_i^2 + \lambda_i z_i\right) + \min_{x\geq 0} \sum_{i=1}^{n} \left(d_i^2 x_i^2 - (\gamma\hat{r}_i + [F\lambda]_i)x_i\right) \\ &= \sum_{i=1}^{k} \min_{z_i} \left(z_i^2 + \lambda_i z_i\right) + \sum_{i=1}^{n} \min_{x_i\geq 0} \left(d_i^2 x_i^2 - (\gamma\hat{r}_i + [F\lambda]_i)x_i\right). \end{aligned}$$

The minimum of $z_i^2 + \lambda_i z_i$ is equal to $-\lambda_i^2/4$ (attained at $z_i = -\lambda_i/2$). Also, the unique minimizer of $\min_{\xi\geq 0} \alpha\xi^2 - \beta\xi$ is $\xi = [\beta/(2\alpha)]_+$, where $[\cdot]_+$ denotes the projection of the argument onto $\mathbb{R}_+$; hence

$$\min_{x_i\geq 0} \left(d_i^2 x_i^2 - (\gamma\hat{r}_i + [F\lambda]_i)x_i\right) = -\frac{1}{4d_i^2}[\gamma\hat{r}_i + [F\lambda]_i]_+^2,$$

which is attained at the unique optimal point

$$x_i = \frac{1}{2d_i^2}[\gamma\hat{r}_i + [F\lambda]_i]_+. \tag{14.37}$$

Thus, we have that

$$\begin{aligned} g(\lambda) &= -\frac{1}{4}\sum_{i=1}^{k} \lambda_i^2 - \frac{1}{4}\sum_{i=1}^{n} \frac{[\gamma\hat{r}_i + [F\lambda]_i]_+^2}{d_i^2} \\ &= -\frac{1}{4}\|\lambda\|_2^2 - \frac{1}{4}\|D^{-1}[\gamma\hat{r} + F\lambda]_+\|_2^2. \end{aligned}$$

2. If λ^* is an optimal solution to the dual problem $\max_\lambda\ g(\lambda)$, then an optimal solution to the primal problem is recovered applying (14.37), as

$$x^* = \frac{1}{2}D^{-2}[\gamma\hat{r} + F\lambda^*]_+.$$

3. Let us compute the subdifferential of $h \doteq -g$:

$$\begin{aligned} \partial h(\lambda) &= \frac{1}{2}\lambda + \frac{1}{4}\sum_{i=1}^{n} \frac{\partial[\gamma\hat{r}_i + [F\lambda]_i]_+^2}{d_i^2} \\ &= \frac{1}{2}\lambda + \frac{1}{4}\sum_{i=1}^{n} \frac{2[\gamma\hat{r}_i + [F\lambda]_i]_+ \partial[\gamma\hat{r}_i + [F\lambda]_i,]_+}{d_i^2} \end{aligned}$$

and

$$\partial[\gamma\hat{r}_i + [F\lambda]_i,]_+ = \begin{cases} f_i & \text{if } \gamma\hat{r}_i + f_i^\top\lambda > 0 \\ 0 & \text{if } \gamma\hat{r}_i + f_i^\top\lambda < 0 \\ \alpha f_i,\ \alpha\in[0,1] & \text{if } \gamma\hat{r}_i + f_i^\top\lambda = 0 \end{cases}$$

A subgradient of $h(\lambda)$ is thus given by

$$\frac{1}{2}\lambda + \frac{1}{2}\sum_{i=1}^{n} \frac{1}{d_i^2}[\gamma \hat{r}_i + f_i^\top \lambda]_+ f_i.$$

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 looses it. His wealth W_k after k bets is a random variable:

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

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 with w and ℓ the (random) number of times where the gambler wins and looses 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[51].

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 $\sum_i p_i = 1$).

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

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.

Solution 14.5 (Kelly's betting strategy)

1. The expected value of W_k is

$$\mathbb{E}\{W_k\} = 2^k W_0 p^k + 0(1-p^k) = (2p)^k W_0,$$

so, if $p > 0.5$, the expected wealth grows at geometric rate. However, such average growth is rather illusory in practice, since the probability of running broke at iteration k also tends to one at geometric rate:

$$\text{Prob}\{W_k = 0\} = 1 - p^k,$$

hence $\lim_{k\to\infty} W_k = 0$ with probability one (whenever $p < 1$).

2. (a) We have

$$\begin{aligned} G &= \lim_{k\to\infty} \frac{1}{k} \log_2 \frac{W_k}{W_0} = \lim_{k\to\infty} \frac{1}{k} \left(w \log_2(1+x) + \ell \log_2(1-x)\right) \\ &= \lim_{k\to\infty} \frac{w}{k} \log_2(1+x) + \lim_{k\to\infty} \frac{\ell}{k} \log_2(1-x) \\ &= p \log_2(1+x) + (1-p)\log_2(1-x). \end{aligned}$$

This function is a positive combination of two concave functions, hence it is concave.

(b) Imposing that the derivative of G w.r.t. x is zero, we obtain that $x = 2p - 1$. If this quantity is positive, it is also the optimal solution we sought, otherwise the optimal solution is zero, that is

$$x^* = \begin{cases} 2p-1 & \text{if } p > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

and the corresponding optimal growth rate is

$$G^* = 1 + p\log_2 p + (1-p)\log_2(1-p),$$

if $p > 0.5$, and $G^* = 0$ otherwise.

3. (a) We have that

$$G = \mathbb{E}\{\log(1+rx)\} = \sum_{i=1}^{m} p_i \log(1+r_i x)$$

(b) We need to solve the univariate optimization problem

$$\max_x \sum_{i=1}^{m} p_i \log(1+r_i x) \quad \text{s.t.: } x \in [0,\, 1].$$

The objective function is concave, with derivative given by

$$g(x) = \sum_{i=1}^{m} \frac{p_i r_i}{1+r_i x}.$$

The optimal fraction can thus be computed via bisection over the interval [0, 1], see Exercise 12.3.

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 sub-gradient of $-\phi$ at w. If ϕ is differentiable at w, what is its gradient at w?

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 $u \in [w, w+\epsilon]$.

Solution 14.6

1. The problem is equivalent to

$$\begin{array}{ll} \displaystyle\max_{u\,:\,u_{n+1}=0} & \displaystyle\sum_{i=1}^{n+1} y_i u_i - \sigma_i^2 u_i^2 - c t_i \\ \text{subject to} & t_i \geq u_i - u_{i-1}, i = 1, \ldots, n+1 \\ & t_i \geq -u_i + u_{i-1}, i = 1, \ldots, n+1, \end{array}$$

The Lagrangian is

$$\begin{aligned}\mathcal{L}(u,t,\lambda,\mu) &= \sum_{i=1}^{n+1}\left\{y_iu_i - \sigma_i^2u_i^2 - ct_i + \lambda_i(t_i - u_i + u_{i-1}) + \mu_i(t_i + u_i - u_{i-1})\right\}\\ &= (\lambda_1 - \mu_1)w + \sum_{i=1}^{n}\left\{(y_i - \lambda_i + \lambda_{i+1} + \mu_i - \mu_{i+1})u_i - \sigma_i^2u_i^2\right\}\\ &+ \sum_{i=1}^{n+1}(\lambda_i + \mu_i - c_i)t_i.\end{aligned}$$

The dual function is

$$\begin{aligned}\max_{u,t} L(u,t,\lambda,\mu) &= (\lambda_1 - \mu_1)w + \sum_{i=1}^{n}\max_{u_i}\left\{(y_i - \lambda_i + \lambda_{i+1} + \mu_i - \mu_{i+1})u_i - \sigma_i^2u_i^2\right\}\\ &+ \sum_{i=1}^{n+1}\max_{t_i}(\lambda_i + \mu_i - c_i)t_i\\ &= \begin{cases}(\lambda_1 - \mu_1)w + \sum_{i=1}^{n}\frac{(y_i - \lambda_i + \lambda_{i+1} + \mu_i - \mu_{i+1})^2}{4\sigma_i^2} & \text{if } \lambda_i + \mu_i - c = 0,\\ +\infty & \text{otherwise.}\end{cases}\end{aligned}$$

Eliminating the variable μ, we obtain the dual problem as:

$$\min_{0\le\lambda\le c} \quad (2\lambda_1 - c)w + \sum_{i=1}^{n}\frac{(y_i - 2\lambda_i + 2\lambda_{i+1})^2}{4\sigma_i^2}.$$

Let $v = 2\lambda - c\mathbf{1}$, the dual problem reduces to:

$$\phi(w) = \min_v \; v_1w + \sum_{i=1}^{n}\frac{(y_i - v_i + v_{i+1})^2}{4\sigma_i^2} \;\; : \;\; \|v\|_\infty \le c.$$

2. The function ϕ is concave. The formula for sub-gradient of maxima of functions[52] shows that for any w, a sub-gradient of $-\phi$ at w is $v_1^*(w)$, where $v^*(w)$ is an optimal dual variable. If ϕ is differentiable at w, then the sub-gradient at w is unique, and equal to the gradient. Hence, $\phi'(w) = v_1^*(w)$ for points w such that ϕ is differentiable at w.

3. Assuming that ϕ is differentiable in $[w, w+\epsilon]$, let $u \in [w, w+\epsilon]$, and let $v_1^*(u)$ be a corresponding gradient of ϕ at u. We then have

$$\begin{aligned}\phi(w+\epsilon) &= \phi(w) + \int_w^{w+\epsilon}\phi'(u)\,du\\ &= \phi(w) + \epsilon\int_0^1\phi'(w+t\epsilon)\,dt.\end{aligned}$$

With the bound $|\phi'(u)| = |v_1^*(u)| \le c$:

$$|\phi(w+\epsilon) - \phi(w)| = \epsilon\left|\int_0^1\phi'(w+t\epsilon)\,dt\right| \le \epsilon c.$$

[52] See Section 8.2.3.1.

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 be $D = \$7000$. 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.

1. Formulate the decision-making problem as an optimization problem. Make sure to define the variables and constraints precisely.

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 $\$8000 = 10\% \times \$80,000$

To describe the tax, use the following constraint:

$$T_i = 0.1 \min(G_i, \alpha) + 0.28 \max(G_i - \alpha, 0) \qquad (14.38)$$

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 (14.38) be replaced by the following set of constraints? Is it the case for our problem? Can you replace (14.38) by (14.39) in your problem? Explain.

$$\begin{aligned} T_i &= 0.1 d_{1,i} + 0.28 d_{2,i} \\ d_{2,i} &\geq G_i - \alpha \\ d_{2,i} &\geq 0 \\ d_{1,i} &\geq G_i - d_{2,i} \\ d_{1,i} &\geq d_{2,i} - \alpha \end{aligned} \qquad (14.39)$$

4. Is the new problem formulation, with (14.39), 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?

Solution 14.7

1. The variables are, for $i = 1, \ldots, 6$: x_i: amount of pay received at the beginning of month; $y_{i,j}$: amount of credit card $j = 1, 2$ paid off at the beginning of month $i = 1, \ldots, 6$; t_i: amount of transfer from credit card 2 to 1 at the beginning of month.

   We also introduce the following redundant variables, for $i = 1, \ldots, 6$, $j, k = 1, 2$: w_i: amount remaining at the end of month after paying credit card debt/transfer; $z_{i,j}$: amount owed to credit card j at the end of of month i; T_k: total tax paid in year k; G_k total gross income in years k; W: total wealth at the end of two fiscal years.

The optimization problem is

$$
\begin{array}{ll}
\max & W \\
s.\,t. & \\
& \text{Month 1 balance equations :} \\
& w_1 = x_1 - y_{1,1} - r_3 t_1 \\
& z_{1,1} = (1 + r_1/12)(D - y_{1,1} - t_1) \\
& z_{1,2} = (1 + r_2/12) t_1 \\
& \\
& \text{Month } 2 \leq k \leq 6 \text{ balance equations :} \\
& w_k = w_{k-1} + x_k - y_{k,1} - y_{k,2} - r_3 t_k \\
& z_{k,1} = (1 + r_1/12)(z_{k-1,1} - y_{k,1} - t_k) \\
& z_{k,2} = (1 + r_2/12)(z_{k-1,2} - y_{k,2} + t_k) \\
& \\
& \text{Pay off all credit card debt in 6 months :} \\
& z_{6,1} = 0 \\
& z_{6,2} = 0 \\
& \\
& \text{Tax Calculation} \\
& G_1 = B + x_1 + x_2 + x_3 \\
& G_2 = B + x_4 + x_5 + x_6 \\
& T_i = 0.1 \min(G_i, \alpha) + 0.28 \max(G_i - \alpha, 0) \quad i = 1, 2 \\
& \\
& \text{Total wealth at end of 2yrs} \\
& W = 2B + w_6 - T_1 - T_2 \\
& \\
& \text{Variable constraints} \\
& \sum_{i=1}^{6} x_i = C \\
& x_i \leq 4/3 \times C/6 \qquad\qquad i = 1, \ldots, 6 \\
& W, t, x, y, z, w, T, G \geq 0
\end{array}
$$

2. No, because the tax equality constraints $T_i = \ldots$ is not affine in variables G_i.

3. Consider solving the problem with constraints (14.39) instead of (14.38). As we maximize $W = 2B + w_6 - T_1 - T_2$, the dummy variable $d_{2,i}$ is forced to be satisfy $d_{2,i} = \max(G_i - \alpha, 0)$, $i = 1, 2$.

   Now if $d_{2,i} = 0 \geq G_i - \alpha$ (income is in the low tax bracket), then (14.39) forces $d_{1,i} \geq G_i$ & $d_{1,i} \geq -\alpha$, and maximization forces $d_{1,i} = G_i$. Now if $d_{2,i} = G_i - \alpha \geq 0$ (income is in the high tax bracket), then (14.39) forces $d_{1,i} \geq \alpha$ & $d_{1,i} \geq G_i - 2\alpha$, and maximization forces $d_{1,i} = \max(\alpha, G_i - 2\alpha)$. For the equivalence of this new formulation to the original problem, we require that: $\alpha \geq G_i - 2\alpha \implies 3\alpha \geq G_i$ at optimality. In the given problem, the gross annual income in any year cannot exceed $B + C = \$100,000$, whereas $\alpha = \$80,000$. So for the given problem parameters, we can safely replace the tax constraints (14.38) by (14.39).

4. The new problem is convex, in fact, a linear program, as constraints are affine inequalities.

5. For $\alpha = 80,000$, the results are (rounding the solution to 2 decimals):

$$\begin{array}{llllllll}
W^* & = 144,198.52 & & & & & \\
x^* & = [6666.67 & 4506.23 & 4397.92 & 4732.92 & 4760.70 & 4936.25] \\
y^*_{:,1} & = [6666.00 & 0 & 0 & 0 & 0 & 0] \\
y^*_{:,2} & = [0 & 334.81 & 0 & 0 & 0 & 0] \\
t_: & = [334.05 & 0 & 0 & 0 & 0 & 0]
\end{array}$$

Note the optimal decision may not be unique.

6. As seen in Fig. (14.15), the optimal schedule is unaffected as α increases beyond a threshold: when α is large enough, only the low tax bracket plays an effect in the optimal schedule.

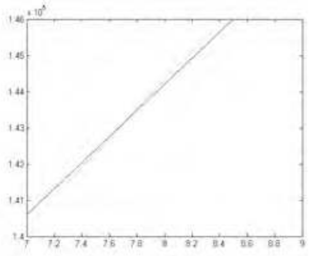

Figure 14.15: Optimal wealth as a function of income α.

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

$$\max_x \hat{r}^\top x - \lambda x^\top C x - c \cdot T(x - x^0) \ : \ x \geq 0, x \in \mathcal{X}, \qquad (14.40)$$

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 $x_1 + \ldots + x_n = 1$.

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 that 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 (14.40) 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 $M = .01, .05, .1, 1, 5$, with $c = 5.10^{-4}$. Comment on the qualitative differences between the optimal portfolio for two different values of $M = .01, 1$.

Solution 14.8

1. We can eliminate the variable v, and get

   $$\begin{aligned} p^* &\doteq \min_{v,w} \; v + w + \frac{w^2}{2M} \; : \; |z| \leq v + w, \;\; v \leq M, \;\; w \geq 0 \\ &= |z| + \frac{1}{2M} \min_{w \geq \gamma} w^2, \end{aligned}$$

   where $\gamma = \max(0, |z| - M)$. If $|z| \leq M$, then $\gamma = 0$, and we obtain

   $$p^* = |z| + \frac{1}{2M} \min_{w \geq 0} w^2 = |z|.$$

If $|z| \geq M$, then

$$\begin{aligned} p^* &= |z| + \frac{1}{2M} \min_{w \geq |z|-M} w^2 \\ &= |z| + \frac{(|z|-M)^2}{2M} \\ &= \frac{z^2+M^2}{2M}. \end{aligned}$$

The representation proves convexity of B_M, since the constraints are convex in z. Indeed, the epigraph condition $B_M(z) \leq t$, for $t \in \mathbb{R}$, is equivalent to the conditions

$$v + w + \frac{w^2}{2M} \leq t, \quad |z| \leq v + w, \quad v \leq M, \quad w \geq 0$$

and those are *jointly* convex in (z, t, v, w).

2. Denoting $z = x - x^0$, we have

$$T(x) = \sum_{i=1}^{n} B_M(z_i) = \min_{v,w} \sum_{i=1}^{n} v_i + w_i + \frac{w_i^2}{2M} \ : \ \begin{array}{l} |z_i| \leq v_i + w_i, \ \ v_i \leq M, \\ w_i \geq 0. \end{array}$$

The above can indeed be written in vector form, as

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

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. The optimization problem in convex format is given by

$$\begin{array}{ll} \text{maximize} & \hat{r}^\top x - \lambda x^\top C x - c \cdot \left(\mathbf{1}^\top (v+w) + \dfrac{1}{2M} w^\top w \right) \\ \text{subject to} & \mathbf{1}^\top x = 1, \ \ |x - x^0| \leq v + w, \\ & x \geq 0, \ \ v \leq M\mathbf{1}, \ \ w \geq 0. \end{array}$$

This optimization problem is a QP.

4. The efficient frontier is shown in Fig. 14.16.

5. The optimal portfolio for $M = 1$ strictly dominates the one for $M = 0.1$, i.e., it gives higher average return for all values of the standard deviation. Moreover, while the average return for the optimal portfolio for $M = 1$ increases with standard deviation, the average return for the optimal portfolio for $M = 0.1$ actually decreases with standard deviation.

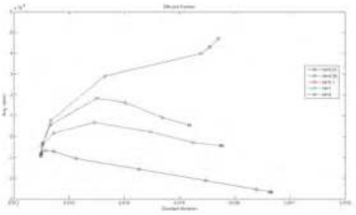

Figure 14.16: Efficient frontier with a reverse Hüber function cost model.

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[53].

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

[53] 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 $V = [v_1, \ldots, v_T]$ 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 semi-definite 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 $D_w \doteq \mathrm{diag}\,(w)\, D$ is diagonal, positive definite, and $F_w \doteq \mathrm{diag}\,(w)\, F$.

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 $C = [c_1, \ldots, c_T] \in \mathbb{R}^{n,T}$ 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.

Solution 14.9

1. Let us first summarize our problem data:

   - $p \in \mathbb{R}^n$: initial price of each asset.
   - $s \in \mathbb{R}^n$: side of each asset (1 to indicate long, -1 to indicate short)
   - $w \in \mathbb{R}^n$: price weight vector, computed as $w = p \circ s$, so that $|w| = p$.
   - Covariance matrix $\Sigma = D^2 + FF^\top$, with
     - $F \in \mathbb{R}^{n,k}$, factor loading matrix, and $F_w = \mathrm{diag}\,(w)\,F \in \mathbb{R}^{n \times k}$, its weighted counterpart;
     - $D \in \mathbb{R}^{n,n}$, positive-definite diagonal matrix in the factor covariance model, and $D_w \doteq \mathrm{diag}\,(w)\,D$, its weighted counterpart.
   - $C = [c_1, \ldots, c_T] \in \mathbb{R}^{n \times T}$, a matrix that accounts for return prediction.
   - $V \in \mathbb{R}_+^{n,T}$, a non-negative matrix that contains market impact parameters.
   - x_0, x_{target}, initial and final list of shares. Since a dollar-neutral strategy is sought, we assume that $w^\top x_0 = w^\top x_{\text{target}}$ are both zero.
   - X^u, X^l and Y^u, Y^l: $n \times T$ matrices that contain upper and lower bounds on the schedule and trading matrices.

2. We formulate the portfolio execution as follows:

   Decision variables. We define two sets of decision variables given by:
   $$x_t, y_t \in \mathbb{R}^n, \quad t = 1, \ldots, T.$$
   Here, x_t encapsulates to all the portfolio execution that takes place at period t, and y_t encapsulates the change in the portfolio between time $t-1$ and t.

   Constraints. The upper and lower bounds on x_t are given by:
   $$x_t^l \leq x_t \leq x_t^u, \quad t = 1, \ldots, T.$$
   The y_t variables are defined by the following constraints:
   $$y_t = x_t - x_{t-1}, \quad t = 1, \ldots, T.$$

Upper and lower bounds on the trading are enforced as follows:

$$0 \leq y_t \leq y_t^u, \;\; t = 1, \ldots, T.$$

The *dollar neutral strategy* is enforced by:

$$w^\top x_t = 0, \;\; t = 1, \ldots, T.$$

Objective function. The objective function is given by

$$\sum_{t=1}^{T} y_t^\top \operatorname{diag}(v_t)\, y_t + x_t^\top (D_w^2 + F_w F_w^\top) x_t - c_t^\top x_t.$$

The above problem has a convex quadratic function that is minimized over a convex set defined by linear constraints. Hence, it has the required QP format.

3. We can take advantage of the factor model to speed up computation as follows. For $t = 1, \cdots, T$, we can define new variables $z_t = F_w^\top x_t$. Here $z_t \in \mathbb{R}^k$, where $k \ll n$. With this in place, the objective function becomes

$$\sum_{t=1}^{\top} (y_t^\top \operatorname{diag}(v_t)\, y_t + x_t^\top D_w^2 x_t + z_t^\top z_t - c_t^\top x_t.$$

We observe that the above function of $(x_t, y_t, z_t)_{t=1}^t$ is decomposable; that is, its Hessian is diagonal. As explained in Exercise 12.9, this leads to great speed-ups. Precisely, the computational time will grows linearly with the number of assets n, as opposed to a cubic growth.

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 $x(t) = \mathrm{e}^{At}x_0$ 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.

Solution 15.1 (Stability and eigenvalues) The continuous-time LTI system (15.20) is, by definition, asymptotically stable if and only if its free response (i.e., the state evolution obtained for input $u(t) = 0$ for all $t \geq 0$) goes to zero as $t \to \infty$, for any initial condition x_0. Since we assume that A is diagonalizable, we can write $A = U\Lambda U^{-1}$ (where Λ is diagonal and contains the eigenvalues λ_i of A, and U contains by columns the eigenvectors u_i), and apply the result of Example 3.8, writing the matrix exponential as $\mathrm{e}^{At} = U\mathrm{e}^{\Lambda t}U^{-1}$. We thus have that

$$\begin{aligned}
\|\mathrm{e}^{At}\|_2 &= \|U\mathrm{e}^{\Lambda t}U^{-1}\|_2 \leq \|U\|_2 \cdot \|U^{-1}\|_2 \cdot \|\mathrm{e}^{\Lambda t}\|_2 \\
&= \kappa\|\mathrm{e}^{\Lambda t}\|_2 = \kappa \max_{i=1,\ldots,n} \mathrm{e}^{\sigma_i t} \\
&= \kappa \mathrm{e}^{\sigma_{\max} t},
\end{aligned}$$

where $\kappa > 0$ is a constant, σ_i is the real part of λ_i, and $\sigma_{\max}$ is the maximum real part of the eigenvalues. If $\sigma_i < 0$ for all $i = 1, \ldots, n$, then clearly $\sigma_{\max} < 0$, and $\|\mathrm{e}^{At}\|_2 \to 0$ for $t \to \infty$[54]. Now, since the free evolution of the state is given by

$$x(t) = \mathrm{e}^{At}x_0,$$

we have that

$$\begin{aligned}
\|x(t)\|_2 &= \|\mathrm{e}^{At}x_0\|_2 \leq \|\mathrm{e}^{At}\|_2 \cdot \|x_0\|_2 \\
&\leq \kappa\|x_0\|_2 \mathrm{e}^{\sigma_{\max} t},
\end{aligned}$$

thus $\|x(t)\|_2 \to 0$ for $t \to \infty$, which proves that if all the eigenvalues of A have negative real parts, then the system is asymptotically stable. We can also prove the converse, that is if the system is asymptotically stable, then all eigenvalues must have negative real parts. We actually prove the following equivalent statement: if there exist

[54] Although we proved this fact only for diagonalizable A, this assumption can be removed, and the result actually holds for all A.

$\sigma_i \geq 0$, then the system is not asympotically stable. To prove this fact, it suffices to choose the initial condition $x_0 = Ue_i$ to see that, for this initial condition,

$$\|x(t)\|_2 = \|U\mathrm{e}^{\Lambda t}U^{-1}x_0\|_2 = \|U\mathrm{e}^{\Lambda t}e_i\|_2 = \|u_i\mathrm{e}^{\lambda_i t}\|_2 = \|u_i\|_2\mathrm{e}^{\sigma_i t},$$

thus $\|x(t)\|_2$ doesn't go to zero as $t \to \infty$ (since $\sigma_i \geq 0$).

The corresponding result for the discrete-time LTI system (15.28) can be proved analogously, by considering the free evolution $x(k) = A^k x_0 = U\Lambda^k U^{-1}x_0$.

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

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

Solution 15.2 (Signal norms)

1. For a complex vector signal $w(t)$, we have that

   $$\|w(t)\|_2^2 = w^\star(t)w(t),$$

   where * denotes the transpose conjugate. Thus, for the harmonic signal $w(t) = v\mathrm{e}^{\jmath\omega t}$ we have $\|w(t)\|_2^2 = \|v\|_2^2$, and the energy is

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

   hence the harmonic signal has infinite energy.

   For the causal exponential signal $w(t) = v\mathrm{e}^{at}$, $a < 0$, we have instead

   $$E(w) = \|v\|_2^2 \int_0^{\infty} \mathrm{e}^{2at}\mathrm{d}t = \frac{\|v\|_2^2}{2|a|}.$$

2. The harmonic signal $w(t) = v e^{\jmath \omega t}$ has period $T = 2\pi/\omega$; its power is given by

$$P(w) \doteq \frac{1}{T}\int_{t_0}^{t_0+T} \|w(t)\|_2^2 \mathrm{d}t = \|v\|_2^2 \frac{1}{2\pi/\omega}\int_0^{2\pi/\omega} \mathrm{d}t = \|v\|_2^2.$$

The power of the sinusoidal signal $w(t) = v\sin(\omega t)$ is instead

$$P(w) \doteq \|v\|_2^2 \frac{1}{2\pi/\omega}\int_0^{2\pi/\omega} \sin^2(\omega t)\mathrm{d}t = \|v\|_2^2 \frac{\pi}{\omega}.$$

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 exist 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.

Solution 15.3 (Energy upper bound on the system's state evolution)

1. Using an argument analogous to the one introduced in the solution of Exericse 15.1, one can prove that $\|y(t)\|_2 \leq c\|x_0\|_2 e^{\sigma_{\max} t}$, where $\sigma_{\max}$ is the maximum real part of the eigenvalues λ_i of A. If the system is stable, then $\mathrm{Re}\,(\lambda_i) < 0$ for all i, hence $\|y(t)\|_2$ is upper bounded by a decreasing exponential, and thus it is integrable over 0 to ∞, hence $J(x_0) < \infty$, for any given x_0.

2. Define $V(x) \doteq x^\top P x$. Evaluating V along the trajectory of the system we have $V(x(t)) = x(t)^\top P x(t)$, and

$$\begin{aligned}\dot{V}(x(t)) &= \frac{\mathrm{d}V(x(t))}{\mathrm{d}t} = \nabla_x^\top V \cdot \frac{\mathrm{d}x(t)}{\mathrm{d}t} = 2x(t)^\top P \dot{x}(t) \\ &= 2x(t)^\top PAx(t) = x(t)^\top PAx(t) + x(t)^\top A^\top P x(t) \\ &= x(t)^\top (PA + A^\top P)x(t).\end{aligned}$$

Assume now there exist $P \succeq 0$ such that

$$PA + A^\top P + Q \preceq 0. \tag{15.41}$$

multiplying this inequality on the left by $x(t)^\top$ and on the right by $x(t)$, we have

$$x(t)^\top (PA + A^\top P)x(t) + x(t)^\top Qx(t) \leq 0.$$

Recognizing that the first term in the above inequality is equal to $\dot{V}(x(t))$, we rewrite it as

$$x(t)^\top Qx(t) \leq -\dot{V}(x(t)).$$

Integrating from 0 to ∞, we have

$$\begin{aligned} J(x_0) &= \int_0^\infty x(t)^\top Qx(t)\mathrm{d}t \leq -\int_0^\infty \dot{V}(x(t))\mathrm{d}t \\ &= V(x_0) - V(x(\infty)) = V(x_0) = x_0^\top Px_0, \end{aligned}$$

where we have used the hypothesis of stability, which implies that $x(\infty) = 0$.

3. From the previous point, any $P \succeq 0$ that satisfies (15.41) provides an upper bound on the state energy: $J(x_0) \leq x_0^\top Px_0$. We can thus find the minimal upper bound by minimizing $x_0^\top Px_0$ over $P \succeq 0$ that satisfy (15.41), which amounts to solving the following SDP

$$\begin{aligned} \min_P \quad & x_0^\top Px_0 \\ \text{s.t.:} \quad & P \succeq 0 \\ & PA + A^\top P + Q \preceq 0. \end{aligned}$$

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 with 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 exist $P \succeq 0$ such that

$$\begin{bmatrix} A^\top P + PA + C^\top C & PB \\ B^\top P & -\gamma^2 I \end{bmatrix} \preceq 0 \qquad (15.42)$$

then it holds that

$$\|H\|_\infty \leq \gamma.$$

Then, 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 $V(x) = x^\top P x$, and observe that the derivative in time of V, along the trajectories of system (15.26) is

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

Then, show that the LMI condition (15.42) 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 $\|H\|_\infty \leq \gamma$.

Solution 15.4 (System gain)

1. We recall that, for a stable system, the Fourier transform $\hat{Y}(\omega)$ of the outpuy $y(t)$ is given by[55] $\hat{Y}(\omega) = H(\jmath\omega)\hat{U}(\omega)$, where $U(\omega)$ is the Fourier transform of the input $u(t)$ (which exists, since u is assumed to be of finite energy). Using Parseval's identity, we

[55] See Remark 15.1.

rewrite the energy gain as

$$
\begin{aligned}
\text{energy gain} &= \sup_{\|\hat{U}\|_2^2<\infty, \hat{U}\neq 0} \frac{\int_{-\infty}^{\infty} \|H(\jmath\omega)\hat{U}(\omega)\|_2^2 \mathrm{d}\omega}{\int_{-\infty}^{\infty} \|U(\omega)\|_2^2 \mathrm{d}\omega} \\
&\leq \sup_{\|\hat{U}\|_2^2<\infty, \hat{U}\neq 0} \frac{\int_{-\infty}^{\infty} \|H(\jmath\omega)\|_2^2 \|\hat{U}(\omega)\|_2^2 \mathrm{d}\omega}{\int_{-\infty}^{\infty} \|U(\omega)\|_2^2 \mathrm{d}\omega} \\
&\leq \sup_{\|\hat{U}\|_2^2<\infty, \hat{U}\neq 0} \frac{\sup_{\omega} \|H(\jmath\omega)\|_2^2 \int_{-\infty}^{\infty} \|\hat{U}(\omega)\|_2^2 \mathrm{d}\omega}{\int_{-\infty}^{\infty} \|U(\omega)\|_2^2 \mathrm{d}\omega} \\
&= \sup_{\omega} \|H(\jmath\omega)\|_2^2.
\end{aligned}
$$

Further, since $H(\jmath\omega)$ is the Fourier transform of real signals (the impulse response of the system), its norm is centrally symmetric with respect to ω, hence the supremum above can be restricted to the nonnegative ω, thus proving the claim that energy gain $=$ $\sup_{\omega\geq 0} \|H(\jmath\omega)\|_2^2$.

2. Following the hint, we let $V(x(t)) = x(t)^\top P x(t)$, and evaluate

$$
\begin{aligned}
\dot{V}(x(t)) &\doteq \frac{\mathrm{d}V(x(t))}{\mathrm{d}t} = x(t)^\top P\dot{x}(t) + \dot{x}(t)^\top P x(t) \\
&= x(t)^\top P(Ax(t) + Bu(t)) + (Ax(t) + Bu(t))^\top P x(t) \\
&= x(t)^\top (PA + A^\top P)x(t) + 2x(t)^\top PBu(t).
\end{aligned}
$$

Next, multiplying Eq. (15.42) on the right by (x, u) and on the left by its transpose, and recalling that $y(t) = Cx(t)$, we have

$$
\begin{aligned}
0 &\geq \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}^\top \begin{bmatrix} A^\top P + PA + C^\top C & PB \\ B^\top P & -\gamma^2 I \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \\
&= x(t)^\top (PA + A^\top P)x(t) + 2x(t)^\top PBu(t) + \|y(t)\|_2^2 - \gamma^2 \|u(t)\|_2^2 \\
&= \dot{V}(x(t)) + \|y(t)\|_2^2 - \gamma^2 \|u(t)\|_2^2.
\end{aligned}
$$

Integrating over t from 0 to ∞, we obtain that

$$
\begin{aligned}
\int_0^\infty \|y(t)\|_2^2 \mathrm{d}t - \gamma^2 \int_0^\infty \|u(t)\|_2^2 \mathrm{d}t &\leq -\int_0^\infty \dot{V}(x(t))\mathrm{d}t \\
&= V(x(0)) - V(x(\infty)) \\
&= -V(x(\infty)) \leq 0,
\end{aligned}
$$

where we have used the fact that $x(0) = 0$ and that $V(x) \geq 0$ for all x. We thus have that, for all u with finite energy, the current hypotheses imply that

$$
\|y\|_2^2 \leq \gamma^2 \|u\|_2^2,
$$

hence

$$\frac{\|y\|_2}{\|u\|_2} \le \gamma, \quad \forall u \neq 0, \ \|u\|_2 < \infty,$$

which implies that $\|H\|_\infty \le \gamma$, as desired.

The best upper bound γ^* on $\|H\|_\infty$ can thus be obtained by solving an SDP problem in variables (P, γ^2), where one minimizes γ^2, subject to $P \succeq 0$ and to the LMI (15.42). It could actually be proved (we did not do this) that this optimal upper bound is tight, i.e., γ^* is equal to $\|H\|_\infty$.

Exercise 15.5 (Extended superstable matrices) A matrix $A \in \mathbb{R}^{n,n}$ is said to be continuous-time *extended superstable*[56] (which we denote with $A \in E_c$) if there exist $d \in \mathbb{R}^n$ 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 with $A \in E_d$) if there exist $d \in \mathbb{R}^n$ 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 $|\delta_{ij}| \le \rho r_{ij}$, for given $r_{ij} \ge 0$. Define the radius of extended superstability as the largest value ρ^* of $\rho \ge 0$ such that A is extended superstable for all the admissible uncertainties. Describe a computational approach for determining such ρ^*.

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

Solution 15.5 (Extended superstable matrices)

1. The controlled system is described by the linear differential equation
$$\dot{x} = (A - BK)x.$$
This system is continuous-time extended superstable, if there exist $d \in \mathbb{R}^n$, $d > 0$, such that
$$\sum_{j=1}^{n} |(A - BK)_{ij}| d_j < -(A - BK)_{ii} d_i, \quad i = 1, \ldots, n.$$
Since $d_j \geq 0$, we can absorb it inside the absolute value on the left, obtaining
$$\sum_{j=1}^{n} |a_{ij} d_j - (BK)_{ij} d_j| < -a_{ii} d_i + (BK)_{ii} d_i, \quad i = 1, \ldots, n.$$
Further, since $(BK)_{ij} = \sum_{p=1}^{m} b_{ip} k_{pj}$, we may define new variables $z_{pj} \doteq k_{pj} d_j$, $p = 1, \ldots, m$, $j = 1, \ldots, n$, and write the above condition as
$$\sum_{j=1}^{n} |a_{ij} d_j - \sum_{p=1}^{m} b_{ip} z_{pj}| < -a_{ii} d_i + \sum_{p=1}^{m} b_{ip} z_{pi}, \quad i = 1, \ldots, n,$$
which is a set of linear inequality conditions on the entries of d and $Z \doteq K\text{diag}\,(d)$. Choosing a small $\epsilon > 0$, we can determine a suitable feedback matrix K by first solving the following linear program for variables d and Z
$$\begin{aligned} \min_{d,Z} \quad & \text{trace}(Z) \\ \text{s.t.:} \quad & d \geq \epsilon \mathbf{1} \\ & \textstyle\sum_{j=1}^{n} |a_{ij} d_j - \sum_{p=1}^{m} b_{ip} z_{pj}| \leq -a_{ii} d_i + \sum_{p=1}^{m} b_{ip} z_{pi} - \epsilon, \\ & i = 1, \ldots, n, \end{aligned}$$
and then retrieving K as $K = Z(\text{diag}\,(d))^{-1}$. Notice that the choice of the objective trace(Z) is immaterial in the above problem, since we are actually just required to find a feasible design K.

2. Since the uncertainty on the entries of A is independent and bound in intervals, we have
$$\max_{|\delta_{ij}| \leq \rho r_{ij}} |a_{ij}| = |\hat{a}_{ij}| + \rho r_{ij}.$$
The uncertain matrix is robustly discrete-time extended superstable if if there exist $d \in \mathbb{R}^n$, $d > 0$, such that
$$\max_{|\delta_{ij}| \leq \rho r_{ij}} \sum_{j=1}^{n} |a_{ij}| d_j < d_i, \quad i = 1, \ldots, n,$$

which is equivalent to

$$\sum_{j=1}^{n}(|\hat{a}_{ij}| + \rho r_{ij})d_j < d_i, \quad i = 1, \ldots, n,$$

and this is a set of linear inequality conditions on d. To compute the largest $\rho \geq 0$ such that the matrix is robustly superstable, we may proceed via a bisection approach: fix some small $\epsilon > 0$, let $\rho_l = 0$, and suppose we know in advance that $\rho^* > 0$ and $\rho^* < \rho_r$ (some large value)[57].

Let $\rho = (\rho_r + \rho_l)/2$, and check feasibility (e.g., by solving an LP with zero objective) of the linear inequalities

$$\sum_{j=1}^{n}(|\hat{a}_{ij}| + \rho r_{ij})d_j \leq d_i - \epsilon, \quad i = 1, \ldots, n.$$

If feasible, let $\rho_l \leftarrow \rho$ and repeat, otherwise let $\rho_r \leftarrow \rho$ and repeat. This algorithm will converge at a geometric rate to the optimal value ρ^*.

[57] The first condition can be easily checked by verifying feasibility at $\rho = 0$. The initial ρ_r can be found by progressively increasing ρ until unfeasibility is detected.

16. Engineering Design

Exercise 16.1 (Network congestion control) A network of $n = 6$ peer-to-peer computers is shown in Figure 16.17. Each computer can upload or download data at a certain rate on the connection links shown in the figure.

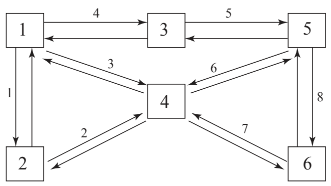

Figure 16.17: A small network.

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.

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.

Solution 16.1 As stated in the exercise, let

$$\begin{aligned} v_{\mathrm{upl}} &= A_+ b^+ - A_- b^- \\ v_{\mathrm{dwl}} &= A_+ b^- - A_- b^+ \\ v_{\mathrm{net}} &= v_{\mathrm{upl}} - v_{\mathrm{dwl}} = A b^+ - A b_- \end{aligned}$$

be, respectively, the total output (upload) rate at the nodes, the total input (download) rate at the nodes, and the net outflow at nodes. The requirements of a maximum download rate of $\bar{v}_{\mathrm{dwl}} = 20$ Mbit/s and maximum upload rate $\bar{v}_{\mathrm{upl}} = 10$ Mbit/s are imposed as linear constraints on the rate vectors b_+, b_-

$$A_+ b^- - A_- b^+ \leq 20 \cdot \mathbf{1}, \quad A_+ b^+ - A_- b^- \leq 10 \cdot \mathbf{1}.$$

The node flows f_i are specified as follows: $f_1 = 9$, $f_5 = -9$ (node 1 transmits to node 5 at rate 9), $f_2 = 8$, $f_6 = -8$ (node 2 transmits to node 6 at rate 8), and $f_i = 0$ for all other nodes. The objective to be minimized is the average congestion level, which is proportional to

$$J = \sum_{i=1}^{8} c_i = \sum_{i=1}^{8} \max(0, b_i^+ + b_i^- - 4),$$

so the problem is written as

$$\begin{aligned} \min_{b^+, b^-} \quad & \textstyle\sum_{i=1}^{8} \max(0, b_i^+ + b_i^- - 4) \\ \text{s.t.:} \quad & A_+ b^- - A_- b^+ \leq 20 \cdot \mathbf{1} \\ & A_+ b^+ - A_- b^- \leq 10 \cdot \mathbf{1} \\ & A b^+ - A b_- = f \\ & b^+ \geq 0,\ b^- \geq 0. \end{aligned}$$

This problems can be recast in the form of an LP, by epigraphic reformulation:

$$\begin{aligned} \min_{b^+, b^-, z} \quad & \textstyle\sum_{i=1}^{8} z_i \\ \text{s.t.:} \quad & z \geq 0,\ b^+ + b^- - 4 \cdot \mathbf{1} \leq z \\ & A_+ b^- - A_- b^+ \leq 20 \cdot \mathbf{1} \\ & A_+ b^+ - A_- b^- \leq 10 \cdot \mathbf{1} \\ & A b^+ - A b_- = f \\ & b^+ \geq 0,\ b^- \geq 0. \end{aligned}$$

The exercise can be solved numerically via the following CVX code. We obtained optimal average congestion level of 1.3750, and rates

$$\begin{aligned} b^+ &= [0\ 7\ 3\ 7\ 7\ 0\ 0\ 2.9233] \\ b^- &= [1\ 0\ 0\ 0\ 0\ 4.7187\ 5.2813\ 0.2047]. \end{aligned}$$

```
%% CVX code for network congestion exercise
A=[1 -1 0 0 0 0;0 1 0 -1 0 0;1 0 0 -1 0 0;...
1 0 -1 0 0 0;0 0 1 0 -1 0;0 0 0 -1 1 0;...
0 0 0 -1 0 1;0 0 0 0 1 -1]’;
[n,p]=size(A);
Ap = max(A,0);
Am = min(A,0);
vmax_dnl=20;
vmax_upl=10;
f=zeros(n,1);
f(1) = 9; f(5)=-f(1);
f(2) = 8; f(6)=-f(2);
%
cvx_begin
variables bp(p) bm(p) z(p)
v_upl = Ap*bp - Am*bm;
v_dnl = Ap*bm - Am*bp;
v_net = v_upl - v_dnl;
minimize ( sum(z) )
subject to
bp >= 0;
bm >= 0;
v_net == f; % flow balance
v_upl <= vmax_upl;
v_dnl <= vmax_dnl;
bp+bm-4 <= z;
0 <= z;
cvx_end
```

Exercise 16.2 (Design of a water reservoir) We need to design a water reservoir for water and energy storage, as depicted in Figure 16.18.

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

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

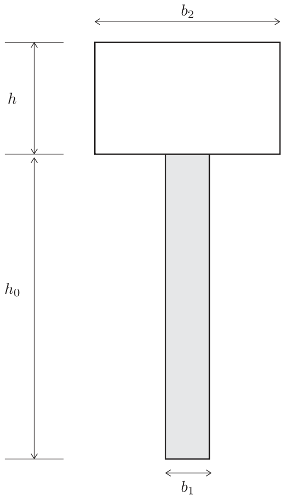

Figure 16.18: A water reservoir on concrete basement.

and, in the positive case, find the optimal design.

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

Solution 16.2 (Design of a water reservoir) All specs in this problem can be formulated in the form of monomial inequalities. The problem can thus be posed as a GP, as shown in the following CVX code. Using the provided numerical data, we obtained an optimal design with

$$b_1 = 0.78, \quad b_2 = 4.7, \quad h_0 = 21.38, \quad h = 2.95.$$

```
%% water reservoir design exercise
ro_w = 10e3; % N/m^3 specific weigths
ro_s = 25e3;
Y=30e9; % elastic modulus in N/m^2
g = 9.8;
Wlimit = 1e5*g; % max weight
cvx_begin gp
variables h0 h b1 b2
m_w = ro_w*b2^2*h; % weight of water
m_s = ro_s*h0*b1^2; % weigth of basement
E = m_w*h0; % energy
maximize (E)
subject to
% structural limit
J = b1^4/12; %basement moment of inertia
Ncr = pi^2*J*Y/(2*h0)^2; % structural limit
Ncr >= 2*m_w;
% form factor of reservoir
b2 >= h;
b2 <= 2*h;
% total heigth
h0+h <=30;
%
h0 <= 35*b1^2;
```

```
m_w+m_s <= Wlimit;
cvx_end
```

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 16.19.

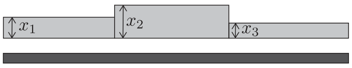

Figure 16.19: A wire is represented as a sequence of rectangular surfaces on a substrate. Lenghts ℓ_i are fixed, and the widths x_i of the segments are the decision variables. This example has three wire segments.

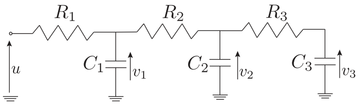

Figure 16.20: RC model of a three-segment wire.

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 16.20.

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

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

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 it 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 to minimize the total area occupied by the wire, while guaranteeing that the dominant time constant does not exceed an assigned level $\eta > 0$.

Solution 16.3 (Wire sizing in circuit design) The request that $\tau \leq \eta$ is written as

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

which holds if and only if

$$\lambda_{\min}(\mathcal{C}(x)^{-1}\mathcal{S}(x)) \geq \frac{1}{\eta}.$$

This condition is in turn equivalent to

$$\lambda_i(\mathcal{C}(x)^{-1}\mathcal{S}(x)) \geq \frac{1}{\eta}, \quad \forall i,$$

where λ_i denotes the i-th eigenvalue of the symmetric matrix $\mathcal{C}(x)^{-1}\mathcal{S}(x)$. This latter condition can be expressed in matrix inequality form as

$$\mathcal{C}(x)^{-1}\mathcal{S}(x) \succeq \frac{1}{\eta}I,$$

which is finally written as an LMI in the variable x:

$$\mathcal{S}(x) \succeq \frac{1}{\eta}\mathcal{C}(x).$$

The wire-sizing problem can thus be cast (and hence efficiently solved) in SDP format as follows:

$$\begin{array}{cc} \min\limits_{x} & \sum_{i=1}^{3} \ell_i x_i \\ \text{s.t.:} & x \geq 0 \\ & \mathcal{S}(x) \succeq \frac{1}{\eta}\mathcal{C}(x). \end{array}$$