Convex Optimization

Solutions Manual

Stephen Boyd **Lieven Vandenberghe**

January 4, 2006

Chapter 2

Convex sets

Exercises

Definition of convexity

2.1 Let $C \subseteq \mathbf{R}^n$ be a convex set, with $x_1, \ldots, x_k \in C$, and let $\theta_1, \ldots, \theta_k \in \mathbf{R}$ satisfy $\theta_i \geq 0$, $\theta_1 + \cdots + \theta_k = 1$. Show that $\theta_1 x_1 + \cdots + \theta_k x_k \in C$. (The definition of convexity is that this holds for $k = 2$; you must show it for arbitrary k.) *Hint.* Use induction on k.

Solution. This is readily shown by induction from the definition of convex set. We illustrate the idea for $k = 3$, leaving the general case to the reader. Suppose that $x_1, x_2, x_3 \in C$, and $\theta_1 + \theta_2 + \theta_3 = 1$ with $\theta_1, \theta_2, \theta_3 \geq 0$. We will show that $y = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \in C$. At least one of the θ_i is not equal to one; without loss of generality we can assume that $\theta_1 \neq 1$. Then we can write

$$y = \theta_1 x_1 + (1 - \theta_1)(\mu_2 x_2 + \mu_3 x_3)$$

where $\mu_2 = \theta_2/(1 - \theta_1)$ and $\mu_2 = \theta_3/(1 - \theta_1)$. Note that $\mu_2, \mu_3 \geq 0$ and

$$\mu_1 + \mu_2 = \frac{\theta_2 + \theta_3}{1 - \theta_1} = \frac{1 - \theta_1}{1 - \theta_1} = 1.$$

Since C is convex and $x_2, x_3 \in C$, we conclude that $\mu_2 x_2 + \mu_3 x_3 \in C$. Since this point and x_1 are in C, $y \in C$.

2.2 Show that a set is convex if and only if its intersection with any line is convex. Show that a set is affine if and only if its intersection with any line is affine.

Solution. We prove the first part. The intersection of two convex sets is convex. Therefore if S is a convex set, the intersection of S with a line is convex.

Conversely, suppose the intersection of S with any line is convex. Take any two distinct points x_1 and $x_2 \in S$. The intersection of S with the line through x_1 and x_2 is convex. Therefore convex combinations of x_1 and x_2 belong to the intersection, hence also to S.

2.3 *Midpoint convexity.* A set C is *midpoint convex* if whenever two points a, b are in C, the average or midpoint $(a + b)/2$ is in C. Obviously a convex set is midpoint convex. It can be proved that under mild conditions midpoint convexity implies convexity. As a simple case, prove that if C is closed and midpoint convex, then C is convex.

Solution. We have to show that $\theta x + (1 - \theta)y \in C$ for all $\theta \in [0, 1]$ and $x, y \in C$. Let $\theta^{(k)}$ be the binary number of length k, *i.e.*, a number of the form

$$\theta^{(k)} = c_1 2^{-1} + c_2 2^{-2} + \cdots + c_k 2^{-k}$$

with $c_i \in \{0, 1\}$, closest to θ. By midpoint convexity (applied k times, recursively), $\theta^{(k)} x + (1 - \theta^{(k)})y \in C$. Because C is closed,

$$\lim_{k \to \infty} (\theta^{(k)} x + (1 - \theta^{(k)})y) = \theta x + (1 - \theta)y \in C.$$

2.4 Show that the convex hull of a set S is the intersection of all convex sets that contain S. (The same method can be used to show that the conic, or affine, or linear hull of a set S is the intersection of all conic sets, or affine sets, or subspaces that contain S.)

Solution. Let H be the convex hull of S and let $\mathcal{D}$ be the intersection of all convex sets that contain S, *i.e.*,

$$\mathcal{D} = \bigcap \{D \mid D \text{ convex}, \ D \supseteq S\}.$$

We will show that $H = \mathcal{D}$ by showing that $H \subseteq \mathcal{D}$ and $\mathcal{D} \subseteq H$.

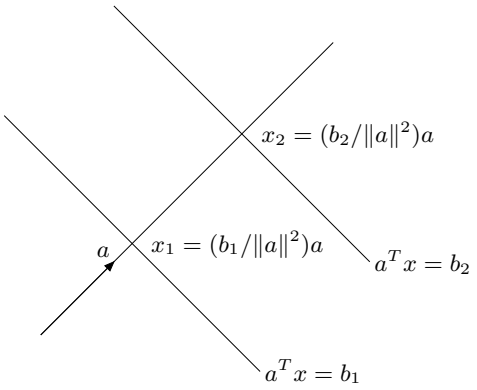
First we show that $H \subseteq \mathcal{D}$. Suppose $x \in H$, *i.e.*, x is a convex combination of some points $x_1, \ldots, x_n \in S$. Now let D be any convex set such that $D \supseteq S$. Evidently, we have $x_1, \ldots, x_n \in D$. Since D is convex, and x is a convex combination of $x_1, \ldots, x_n$, it follows that $x \in D$. We have shown that for any convex set D that contains S, we have $x \in D$. This means that x is in the intersection of all convex sets that contain S, *i.e.*, $x \in \mathcal{D}$.

Now let us show that $\mathcal{D} \subseteq H$. Since H is convex (by definition) and contains S, we must have $H = D$ for some D in the construction of $\mathcal{D}$, proving the claim.

Examples

2.5 What is the distance between two parallel hyperplanes $\{x \in \mathbf{R}^n \mid a^T x = b_1\}$ and $\{x \in \mathbf{R}^n \mid a^T x = b_2\}$?

Solution. The distance between the two hyperplanes is $|b_1 - b_2|/\|a\|_2$. To see this, consider the construction in the figure below.



The distance between the two hyperplanes is also the distance between the two points x_1 and x_2 where the hyperplane intersects the line through the origin and parallel to the normal vector a. These points are given by

$$x_1 = (b_1/\|a\|_2^2)a, \qquad x_2 = (b_2/\|a\|_2^2)a,$$

and the distance is

$$\|x_1 - x_2\|_2 = |b_1 - b_2|/\|a\|_2.$$

2.6 *When does one halfspace contain another?* Give conditions under which

$$\{x \mid a^T x \leq b\} \subseteq \{x \mid \tilde{a}^T x \leq \tilde{b}\}$$

(where $a \neq 0$, $\tilde{a} \neq 0$). Also find the conditions under which the two halfspaces are equal.

Solution. Let $\mathcal{H} = \{x \mid a^T x \leq b\}$ and $\tilde{\mathcal{H}} = \{x \mid \tilde{a}^T x \leq \tilde{b}\}$. The conditions are:

- $\mathcal{H} \subseteq \tilde{\mathcal{H}}$ if and only if there exists a $\lambda > 0$ such that $\tilde{a} = \lambda a$ and $\tilde{b} \geq \lambda b$.
- $\mathcal{H} = \tilde{\mathcal{H}}$ if and only if there exists a $\lambda > 0$ such that $\tilde{a} = \lambda a$ and $\tilde{b} = \lambda b$.

Let us prove the first condition. The condition is clearly sufficient: if $\tilde{a} = \lambda a$ and $\tilde{b} \geq \lambda b$ for some $\lambda > 0$, then

$$a^T x \leq b \implies \lambda a^T x \leq \lambda b \implies \tilde{a}^T x \leq \tilde{b},$$

i.e., $\mathcal{H} \subseteq \tilde{\mathcal{H}}$.

To prove necessity, we distinguish three cases. First suppose a and $\tilde{a}$ are not parallel. This means we can find a v with $\tilde{a}^T v = 0$ and $a^T v \neq 0$. Let $\hat{x}$ be any point in the intersection of $\mathcal{H}$ and $\tilde{\mathcal{H}}$, *i.e.*, $a^T \hat{x} \leq b$ and $\tilde{a}^T x \leq \tilde{b}$. We have $a^T(\hat{x} + tv) = a^T \hat{x} \leq b$ for all $t \in \mathbf{R}$. However $\tilde{a}^T(\hat{x} + tv) = \tilde{a}^T \hat{x} + t\tilde{a}^T v$, and since $\tilde{a}^T v \neq 0$, we will have $\tilde{a}^T(\hat{x} + tv) > \tilde{b}$ for sufficiently large $t > 0$ or sufficiently small $t < 0$. In other words, if a and $\tilde{a}$ are not parallel, we can find a point $\hat{x} + tv \in \mathcal{H}$ that is not in $\tilde{\mathcal{H}}$, *i.e.*, $\mathcal{H} \not\subseteq \tilde{\mathcal{H}}$.

Next suppose a and $\tilde{a}$ are parallel, but point in opposite directions, *i.e.*, $\tilde{a} = \lambda a$ for some $\lambda < 0$. Let $\hat{x}$ be any point in $\mathcal{H}$. Then $\hat{x} - ta \in \mathcal{H}$ for all $t \geq 0$. However for t large enough we will have $\tilde{a}^T(\hat{x} - ta) = \tilde{a}^T \hat{x} + t\lambda\|a\|_2^2 > \tilde{b}$, so $\hat{x} - ta \notin \tilde{\mathcal{H}}$. Again, this shows $\mathcal{H} \not\subseteq \tilde{\mathcal{H}}$.

Finally, we assume $\tilde{a} = \lambda a$ for some $\lambda > 0$ but $\tilde{b} < \lambda b$. Consider any point $\hat{x}$ that satisfies $a^T\hat{x} = b$. Then $\tilde{a}^T\hat{x} = \lambda a^T\hat{x} = \lambda b > \tilde{b}$, so $\hat{x} \notin \tilde{\mathcal{H}}$.

The proof for the second part of the problem is similar.

2.7 *Voronoi description of halfspace.* Let a and b be distinct points in $\mathbf{R}^n$. Show that the set of all points that are closer (in Euclidean norm) to a than b, *i.e.*, $\{x \mid \|x-a\|_2 \leq \|x-b\|_2\}$, is a halfspace. Describe it explicitly as an inequality of the form $c^Tx \leq d$. Draw a picture.

Solution. Since a norm is always nonnegative, we have $\|x-a\|_2 \leq \|x-b\|_2$ if and only if $\|x-a\|_2^2 \leq \|x-b\|_2^2$, so

$$\begin{array}{lll} \|x-a\|_2^2 \leq \|x-b\|_2^2 & \iff & (x-a)^T(x-a) \leq (x-b)^T(x-b) \\ & \iff & x^Tx - 2a^Tx + a^Ta \leq x^Tx - 2b^Tx + b^Tb \\ & \iff & 2(b-a)^Tx \leq b^Tb - a^Ta. \end{array}$$

Therefore, the set is indeed a halfspace. We can take $c = 2(b-a)$ and $d = b^Tb - a^Ta$. This makes good geometric sense: the points that are equidistant to a and b are given by a hyperplane whose normal is in the direction $b-a$.

2.8 Which of the following sets S are polyhedra? If possible, express S in the form $S = \{x \mid Ax \preceq b,\ Fx = g\}$.

(a) $S = \{y_1a_1 + y_2a_2 \mid -1 \leq y_1 \leq 1,\ -1 \leq y_2 \leq 1\}$, where $a_1, a_2 \in \mathbf{R}^n$.

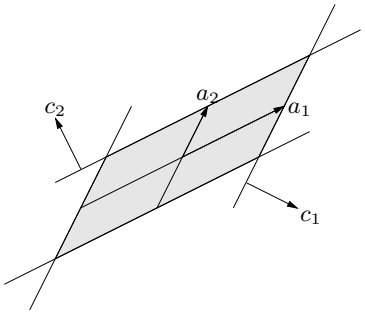
(b) $S = \{x \in \mathbf{R}^n \mid x \succeq 0,\ \mathbf{1}^Tx = 1,\ \sum_{i=1}^n x_ia_i = b_1,\ \sum_{i=1}^n x_ia_i^2 = b_2\}$, where $a_1, \ldots, a_n \in \mathbf{R}$ and $b_1, b_2 \in \mathbf{R}$.

(c) $S = \{x \in \mathbf{R}^n \mid x \succeq 0,\ x^Ty \leq 1 \text{ for all } y \text{ with } \|y\|_2 = 1\}$.

(d) $S = \{x \in \mathbf{R}^n \mid x \succeq 0,\ x^Ty \leq 1 \text{ for all } y \text{ with } \sum_{i=1}^n |y_i| = 1\}$.

Solution.

(a) S is a polyhedron. It is the parallelogram with corners $a_1 + a_2$, $a_1 - a_2$, $-a_1 + a_2$, $-a_1 - a_2$, as shown below for an example in $\mathbf{R}^2$.



For simplicity we assume that a_1 and a_2 are independent. We can express S as the intersection of three sets:

- S_1: the plane defined by a_1 and a_2
- $S_2 = \{z + y_1a_1 + y_2a_2 \mid a_1^Tz = a_2^Tz = 0, -1 \leq y_1 \leq 1\}$. This is a slab parallel to a_2 and orthogonal to S_1
- $S_3 = \{z + y_1a_1 + y_2a_2 \mid a_1^Tz = a_2^Tz = 0, -1 \leq y_2 \leq 1\}$. This is a slab parallel to a_1 and orthogonal to S_1

Each of these sets can be described with linear inequalities.

- S_1 can be described as

$$v_k^Tx = 0, \quad k = 1, \ldots, n-2$$

where v_k are $n-2$ independent vectors that are orthogonal to a_1 and a_2 (which form a basis for the nullspace of the matrix $[a_1\ a_2]^T$).

- Let c_1 be a vector in the plane defined by a_1 and a_2, and orthogonal to a_2. For example, we can take
$$c_1 = a_1 - \frac{a_1^T a_2}{\|a_2\|_2^2} a_2.$$
Then $x \in S_2$ if and only if
$$-|c_1^T a_1| \leq c_1^T x \leq |c_1^T a_1|.$$
- Similarly, let c_2 be a vector in the plane defined by a_1 and a_2, and orthogonal to a_1, *e.g.*,
$$c_2 = a_2 - \frac{a_2^T a_1}{\|a_1\|_2^2} a_1.$$
Then $x \in S_3$ if and only if
$$-|c_2^T a_2| \leq c_2^T x \leq |c_2^T a_2|.$$

Putting it all together, we can describe S as the solution set of $2n$ linear inequalities

$$\begin{array}{rcl} v_k^T x & \leq & 0, \quad k = 1, \ldots, n-2 \\ -v_k^T x & \leq & 0, \quad k = 1, \ldots, n-2 \\ c_1^T x & \leq & |c_1^T a_1| \\ -c_1^T x & \leq & |c_1^T a_1| \\ c_2^T x & \leq & |c_2^T a_2| \\ -c_2^T x & \leq & |c_2^T a_2|. \end{array}$$

(b) S is a polyhedron, defined by linear inequalities $x_k \geq 0$ and three equality constraints.

(c) S is not a polyhedron. It is the intersection of the unit ball $\{x \mid \|x\|_2 \leq 1\}$ and the nonnegative orthant $\mathbf{R}_+^n$. This follows from the following fact, which follows from the Cauchy-Schwarz inequality:

$$x^T y \leq 1 \text{ for all } y \text{ with } \|y\|_2 = 1 \iff \|x\|_2 \leq 1.$$

Although in this example we define S as an intersection of halfspaces, it is not a polyhedron, because the definition requires infinitely many halfspaces.

(d) S is a polyhedron. S is the intersection of the set $\{x \mid |x_k| \leq 1, \quad k = 1, \ldots, n\}$ and the nonnegative orthant $\mathbf{R}_+^n$. This follows from the following fact:

$$x^T y \leq 1 \text{ for all } y \text{ with } \sum_{i=1}^n |y_i| = 1 \iff |x_i| \leq 1, \quad i = 1, \ldots, n.$$

We can prove this as follows. First suppose that $|x_i| \leq 1$ for all i. Then

$$x^T y = \sum_i x_i y_i \leq \sum_i |x_i||y_i| \leq \sum_i |y_i| = 1$$

if $\sum_i |y_i| = 1$.

Conversely, suppose that x is a nonzero vector that satisfies $x^T y \leq 1$ for all y with $\sum_i |y_i| = 1$. In particular we can make the following choice for y: let k be an index for which $|x_k| = \max_i |x_i|$, and take $y_k = 1$ if $x_k > 0$, $y_k = -1$ if $x_k < 0$, and $y_i = 0$ for $i \neq k$. With this choice of y we have

$$x^T y = \sum_i x_i y_i = y_k x_k = |x_k| = \max_i |x_i|.$$

Therefore we must have $\max_i |x_i| \leq 1$.

All this implies that we can describe S by a finite number of linear inequalities: it is the intersection of the nonnegative orthant with the set $\{x \mid -\mathbf{1} \preceq x \preceq \mathbf{1}\}$, *i.e.*, the solution of $2n$ linear inequalities

$$\begin{array}{rcll} -x_i & \leq & 0, & i = 1, \ldots, n \\ x_i & \leq & 1, & i = 1, \ldots, n. \end{array}$$

Note that as in part (c) the set S was given as an intersection of an infinite number of halfspaces. The difference is that here most of the linear inequalities are redundant, and only a finite number are needed to characterize S.

None of these sets are affine sets or subspaces, except in some trivial cases. For example, the set defined in part (a) is a subspace (hence an affine set), if $a_1 = a_2 = 0$; the set defined in part (b) is an affine set if $n = 1$ and $S = \{1\}$; etc.

2.9 *Voronoi sets and polyhedral decomposition.* Let $x_0, \ldots, x_K \in \mathbf{R}^n$. Consider the set of points that are closer (in Euclidean norm) to x_0 than the other x_i, *i.e.*,

$$V = \{x \in \mathbf{R}^n \mid \|x - x_0\|_2 \leq \|x - x_i\|_2,\ i = 1, \ldots, K\}.$$

V is called the *Voronoi region* around x_0 with respect to $x_1, \ldots, x_K$.

(a) Show that V is a polyhedron. Express V in the form $V = \{x \mid Ax \preceq b\}$.

(b) Conversely, given a polyhedron P with nonempty interior, show how to find $x_0, \ldots, x_K$ so that the polyhedron is the Voronoi region of x_0 with respect to $x_1, \ldots, x_K$.

(c) We can also consider the sets

$$V_k = \{x \in \mathbf{R}^n \mid \|x - x_k\|_2 \leq \|x - x_i\|_2,\ i \neq k\}.$$

The set V_k consists of points in $\mathbf{R}^n$ for which the closest point in the set $\{x_0, \ldots, x_K\}$ is x_k.

The sets $V_0, \ldots, V_K$ give a polyhedral decomposition of $\mathbf{R}^n$. More precisely, the sets V_k are polyhedra, $\bigcup_{k=0}^K V_k = \mathbf{R}^n$, and $\mathbf{int}\, V_i \cap \mathbf{int}\, V_j = \emptyset$ for $i \neq j$, *i.e.*, V_i and V_j intersect at most along a boundary.

Suppose that $P_1, \ldots, P_m$ are polyhedra such that $\bigcup_{i=1}^m P_i = \mathbf{R}^n$, and $\mathbf{int}\, P_i \cap \mathbf{int}\, P_j = \emptyset$ for $i \neq j$. Can this polyhedral decomposition of $\mathbf{R}^n$ be described as the Voronoi regions generated by an appropriate set of points?

Solution.

(a) x is closer to x_0 than to x_i if and only if

$$\begin{array}{rcl} \|x - x_0\|_2 \leq \|x - x_i\|_2 & \iff & (x - x_0)^T(x - x_0) \leq (x - x_i)^T(x - x_i) \\ & \iff & x^T x - 2x_0^T x + x_0^T x_0 \leq x^T x - 2x_i^T x + x_i^T x_i \\ & \iff & 2(x_i - x_0)^T x \leq x_i^T x_i - x_0^T x_0, \end{array}$$

which defines a halfspace. We can express V as $V = \{x \mid Ax \preceq b\}$ with

$$A = 2\begin{bmatrix} x_1 - x_0 \\ x_2 - x_0 \\ \vdots \\ x_K - x_0 \end{bmatrix}, \qquad b = \begin{bmatrix} x_1^T x_1 - x_0^T x_0 \\ x_2^T x_2 - x_0^T x_0 \\ \vdots \\ x_K^T x_K - x_0^T x_0 \end{bmatrix}.$$

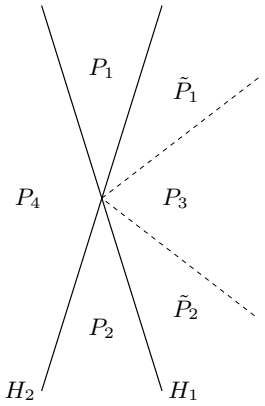
(b) Conversely, suppose $V = \{x \mid Ax \preceq b\}$ with $A \in \mathbf{R}^{K\times n}$ and $b \in \mathbf{R}^K$. We can pick any $x_0 \in \{x \mid Ax \prec b\}$, and then construct K points x_i by taking the mirror image of x_0 with respect to the hyperplanes $\{x \mid a_i^T x = b_i\}$. In other words, we choose x_i of the form $x_i = x_0 + \lambda a_i$, where λ is chosen in such a way that the distance of x_i to the hyperplane defined by $a_i^T x = b_i$ is equal to the distance of x_0 to the hyperplane:

$$b_i - a_i^T x_0 = a_i^T x_i - b_i.$$

Solving for λ, we obtain $\lambda = 2(b_i - a_i^T x_0)/\|a_i\|_2^2$, and

$$x_i = x_0 + \frac{2(b_i - a_i^T x_0)}{\|a_i\|^2} a_i.$$

(c) A polyhedral decomposition of $\mathbf{R}^n$ can not always be described as Voronoi regions generated by a set of points $\{x_1, \ldots, x_m\}$. The figure shows a counterexample in $\mathbf{R}^2$.



$\mathbf{R}^2$ is decomposed into 4 polyhedra $P_1, \ldots, P_4$ by 2 hyperplanes H_1, H_2. Suppose we arbitrarily pick $x_1 \in P_1$ and $x_2 \in P_2$. $x_3 \in P_3$ must be the mirror image of x_1 and x_2 with respect to H_2 and H_1, respectively. However, the mirror image of x_1 with respect to H_2 lies in $\tilde{P}_1$, and the mirror image of x_2 with respect to H_1 lies in $\tilde{P}_2$, so it is impossible to find such an x_3.

2.10 *Solution set of a quadratic inequality.* Let $C \subseteq \mathbf{R}^n$ be the solution set of a quadratic inequality,

$$C = \{x \in \mathbf{R}^n \mid x^T Ax + b^T x + c \leq 0\},$$

with $A \in \mathbf{S}^n$, $b \in \mathbf{R}^n$, and $c \in \mathbf{R}$.

(a) Show that C is convex if $A \succeq 0$.

(b) Show that the intersection of C and the hyperplane defined by $g^T x + h = 0$ (where $g \neq 0$) is convex if $A + \lambda g g^T \succeq 0$ for some $\lambda \in \mathbf{R}$.

Are the converses of these statements true?

Solution. A set is convex if and only if its intersection with an arbitrary line $\{\hat{x} + tv \mid t \in \mathbf{R}\}$ is convex.

(a) We have

$$(\hat{x} + tv)^T A(\hat{x} + tv) + b^T(\hat{x} + tv) + c = \alpha t^2 + \beta t + \gamma$$

where

$$\alpha = v^T Av, \qquad \beta = b^T v + 2\hat{x}^T Av, \qquad \gamma = c + b^T \hat{x} + \hat{x}^T A\hat{x}.$$

The intersection of C with the line defined by $\hat{x}$ and v is the set

$$\{\hat{x} + tv \mid \alpha t^2 + \beta t + \gamma \leq 0\},$$

which is convex if $\alpha \geq 0$. This is true for any v, if $v^T A v \geq 0$ for all v, *i.e.*, $A \succeq 0$. The converse does not hold; for example, take $A = -1$, $b = 0$, $c = -1$. Then $A \not\succeq 0$, but $C = \mathbf{R}$ is convex.

(b) Let $H = \{x \mid g^T x + h = 0\}$. We define α, β, and γ as in the solution of part (a), and, in addition,

$$\delta = g^T v, \qquad \epsilon = g^T \hat{x} + h.$$

Without loss of generality we can assume that $\hat{x} \in H$, *i.e.*, $\epsilon = 0$. The intersection of $C \cap H$ with the line defined by $\hat{x}$ and v is

$$\{\hat{x} + tv \mid \alpha t^2 + \beta t + \gamma \leq 0,\ \delta t = 0\}.$$

If $\delta = g^T v \neq 0$, the intersection is the singleton $\{\hat{x}\}$, if $\gamma \leq 0$, or it is empty. In either case it is a convex set. If $\delta = g^T v = 0$, the set reduces to

$$\{\hat{x} + tv \mid \alpha t^2 + \beta t + \gamma \leq 0\},$$

which is convex if $\alpha \geq 0$. Therefore $C \cap H$ is convex if

$$g^T v = 0 \Longrightarrow v^T A v \geq 0. \tag{2.10.A}$$

This is true if there exists λ such that $A + \lambda g g^T \succeq 0$; then (2.10.A) holds, because then

$$v^T A v = v^T (A + \lambda g g^T) v \geq 0$$

for all v satisfying $g^T v = 0$.

Again, the converse is not true.

2.11 *Hyperbolic sets.* Show that the *hyperbolic set* $\{x \in \mathbf{R}^2_+ \mid x_1 x_2 \geq 1\}$ is convex. As a generalization, show that $\{x \in \mathbf{R}^n_+ \mid \prod_{i=1}^n x_i \geq 1\}$ is convex. *Hint.* If $a, b \geq 0$ and $0 \leq \theta \leq 1$, then $a^\theta b^{1-\theta} \leq \theta a + (1-\theta) b$; see §3.1.9.

Solution.

(a) We prove the first part without using the hint. Consider a convex combination z of two points (x_1, x_2) and (y_1, y_2) in the set. If $x \succeq y$, then $z = \theta x + (1-\theta) y \succeq y$ and obviously $z_1 z_2 \geq y_1 y_2 \geq 1$. Similar proof if $y \succeq x$.

Suppose $y \not\succeq 0$ and $x \not\succeq y$, *i.e.*, $(y_1 - x_1)(y_2 - x_2) < 0$. Then

$$\begin{aligned}
&(\theta x_1 + (1-\theta) y_1)(\theta x_2 + (1-\theta) y_2) \\
&= \theta^2 x_1 x_2 + (1-\theta)^2 y_1 y_2 + \theta(1-\theta) x_1 y_2 + \theta(1-\theta) x_2 y_1 \\
&= \theta x_1 x_2 + (1-\theta) y_1 y_2 - \theta(1-\theta)(y_1 - x_1)(y_2 - x_2) \\
&\geq 1.
\end{aligned}$$

(b) Assume that $\prod_i x_i \geq 1$ and $\prod_i y_i \geq 1$. Using the inequality in the hint, we have

$$\prod_i (\theta x_i + (1-\theta) y_i) \geq \prod x_i^\theta y_i^{1-\theta} = (\prod_i x_i)^\theta (\prod_i y_i)^{1-\theta} \geq 1.$$

2.12 Which of the following sets are convex?

(a) A *slab*, *i.e.*, a set of the form $\{x \in \mathbf{R}^n \mid \alpha \leq a^T x \leq \beta\}$.

(b) A *rectangle*, *i.e.*, a set of the form $\{x \in \mathbf{R}^n \mid \alpha_i \leq x_i \leq \beta_i,\ i = 1, \ldots, n\}$. A rectangle is sometimes called a *hyperrectangle* when $n > 2$.

(c) A *wedge*, *i.e.*, $\{x \in \mathbf{R}^n \mid a_1^T x \leq b_1,\ a_2^T x \leq b_2\}$.

(d) The set of points closer to a given point than a given set, *i.e.*,

$$\{x \mid \|x - x_0\|_2 \leq \|x - y\|_2 \text{ for all } y \in S\}$$

where $S \subseteq \mathbf{R}^n$.

(e) The set of points closer to one set than another, *i.e.*,

$$\{x \mid \mathbf{dist}(x, S) \leq \mathbf{dist}(x, T)\},$$

where $S, T \subseteq \mathbf{R}^n$, and

$$\mathbf{dist}(x, S) = \inf\{\|x - z\|_2 \mid z \in S\}.$$

(f) [HUL93, volume 1, page 93] The set $\{x \mid x + S_2 \subseteq S_1\}$, where $S_1, S_2 \subseteq \mathbf{R}^n$ with S_1 convex.

(g) The set of points whose distance to a does not exceed a fixed fraction θ of the distance to b, *i.e.*, the set $\{x \mid \|x - a\|_2 \leq \theta\|x - b\|_2\}$. You can assume $a \neq b$ and $0 \leq \theta \leq 1$.

Solution.

(a) A slab is an intersection of two halfspaces, hence it is a convex set (and a polyhedron).

(b) As in part (a), a rectangle is a convex set and a polyhedron because it is a finite intersection of halfspaces.

(c) A wedge is an intersection of two halfspaces, so it is convex set. It is also a polyhedron. It is a cone if $b_1 = 0$ and $b_2 = 0$.

(d) This set is convex because it can be expressed as

$$\bigcap_{y \in S} \{x \mid \|x - x_0\|_2 \leq \|x - y\|_2\},$$

i.e., an intersection of halfspaces. (For fixed y, the set

$$\{x \mid \|x - x_0\|_2 \leq \|x - y\|_2\}$$

is a halfspace; see exercise 2.9).

(e) In general this set is not convex, as the following example in $\mathbf{R}$ shows. With $S = \{-1, 1\}$ and $T = \{0\}$, we have

$$\{x \mid \mathbf{dist}(x, S) \leq \mathbf{dist}(x, T)\} = \{x \in \mathbf{R} \mid x \leq -1/2 \text{ or } x \geq 1/2\}$$

which clearly is not convex.

(f) This set is convex. $x + S_2 \subseteq S_1$ if $x + y \in S_1$ for all $y \in S_2$. Therefore

$$\{x \mid x + S_2 \subseteq S_1\} = \bigcap_{y \in S_2} \{x \mid x + y \in S_1\} = \bigcap_{y \in S_2} (S_1 - y),$$

the intersection of convex sets $S_1 - y$.

(g) The set is convex, in fact a ball.

$$\begin{aligned} &\{x \mid \|x - a\|_2 \leq \theta\|x - b\|_2\} \\ &\quad = \quad \{x \mid \|x - a\|_2^2 \leq \theta^2\|x - b\|_2^2\} \\ &\quad = \quad \{x \mid (1 - \theta^2)x^T x - 2(a - \theta^2 b)^T x + (a^T a - \theta^2 b^T b) \leq 0\} \end{aligned}$$

If $\theta = 1$, this is a halfspace. If $\theta < 1$, it is a ball

$$\{x \mid (x - x_0)^T(x - x_0) \leq R^2\},$$

with center x_0 and radius R given by

$$x_0 = \frac{a - \theta^2 b}{1 - \theta^2}, \qquad R = \left(\frac{\theta^2\|b\|_2^2 - \|a\|_2^2}{1 - \theta^2} - \|x_0\|_2^2\right)^{1/2}.$$

2.13 *Conic hull of outer products.* Consider the set of rank-k *outer products*, defined as $\{XX^T \mid X \in \mathbf{R}^{n\times k},\ \mathbf{rank}\, X = k\}$. Describe its conic hull in simple terms.

Solution. We have $XX^T \succeq 0$ and $\mathbf{rank}(XX^T) = k$. A positive combination of such matrices can have rank up to n, but never less than k. Indeed, Let A and B be positive semidefinite matrices of rank k, with $\mathbf{rank}(A + B) = r < k$. Let $V \in \mathbf{R}^{n\times(n-r)}$ be a matrix with $\mathcal{R}(V) = \mathcal{N}(A + B)$, *i.e.*,

$$V^T(A + B)V = V^TAV + V^TBV = 0.$$

Since $A, B \succeq 0$, this means

$$V^TAV = V^TBV = 0,$$

which implies that $\mathbf{rank}\, A \leq r$ and $\mathbf{rank}\, B \leq r$. We conclude that $\mathbf{rank}(A + B) \geq k$ for any A, B such that $\mathbf{rank}(A, B) = k$ and $A, B \succeq 0$.

It follows that the conic hull of the set of rank-k outer products is the set of positive semidefinite matrices of rank greater than or equal to k, along with the zero matrix.

2.14 *Expanded and restricted sets.* Let $S \subseteq \mathbf{R}^n$, and let $\|\cdot\|$ be a norm on $\mathbf{R}^n$.

(a) For $a \geq 0$ we define S_a as $\{x \mid \mathbf{dist}(x, S) \leq a\}$, where $\mathbf{dist}(x, S) = \inf_{y\in S}\|x - y\|$. We refer to S_a as S *expanded* or *extended* by a. Show that if S is convex, then S_a is convex.

(b) For $a \geq 0$ we define $S_{-a} = \{x \mid B(x, a) \subseteq S\}$, where $B(x, a)$ is the ball (in the norm $\|\cdot\|$), centered at x, with radius a. We refer to S_{-a} as S *shrunk* or *restricted* by a, since S_{-a} consists of all points that are at least a distance a from $\mathbf{R}^n\backslash S$. Show that if S is convex, then S_{-a} is convex.

Solution.

(a) Consider two points $x_1, x_2 \in S_a$. For $0 \leq \theta \leq 1$,

$$\begin{aligned}
\mathbf{dist}(\theta x_1 + (1 - \theta)x_2, X) &= \inf_{y\in S}\|\theta x_1 + (1 - \theta)x_2 - y\| \\
&= \inf_{y_1,y_2\in S}\|\theta x_1 + (1 - \theta)x_2 - \theta y_1 - (1 - \theta)y_2\| \\
&= \inf_{y_1,y_2\in S}\|\theta(x_1 - y_1) + (1 - \theta)(x_2 - y_2)\| \\
&\leq \inf_{y_1,y_2\in S}(\theta\|x_1 - y_1\| + (1 - \theta)\|x_2 - y_2\|) \\
&= \theta \inf_{y_1\in S}\|x_1 - y_1\| + (1 - \theta)\inf_{y_2\in s}\|x_2 - y_2\|) \\
&\leq a,
\end{aligned}$$

so $\theta x_1 + (1 - \theta)x_2 \in S_a$, proving convexity.

(b) Consider two points $x_1, x_2 \in S_{-a}$, so for all u with $\|u\| \leq a$,

$$x_1 + u \in S, \qquad x_2 + u \in S.$$

For $0 \leq \theta \leq 1$ and $\|u\| \leq a$,

$$\theta x_1 + (1 - \theta)x_2 + u = \theta(x_1 + u) + (1 - \theta)(x_2 + u) \in S,$$

because S is convex. We conclude that $\theta x_1 + (1 - \theta)x_2 \in S_{-a}$.

2.15 *Some sets of probability distributions.* Let x be a real-valued random variable with $\mathbf{prob}(x = a_i) = p_i$, $i = 1, \ldots, n$, where $a_1 < a_2 < \cdots < a_n$. Of course $p \in \mathbf{R}^n$ lies in the standard probability simplex $P = \{p \mid \mathbf{1}^T p = 1,\ p \succeq 0\}$. Which of the following conditions are convex in p? (That is, for which of the following conditions is the set of $p \in P$ that satisfy the condition convex?)

(a) $\alpha \leq \mathbf{E}\, f(x) \leq \beta$, where $\mathbf{E}\, f(x)$ is the expected value of $f(x)$, *i.e.*, $\mathbf{E}\, f(x) = \sum_{i=1}^n p_i f(a_i)$. (The function $f : \mathbf{R} \to \mathbf{R}$ is given.)

(b) $\mathbf{prob}(x > \alpha) \leq \beta$.

(c) $\mathbf{E}\,|x^3| \leq \alpha\, \mathbf{E}\,|x|$.

(d) $\mathbf{E}\, x^2 \leq \alpha$.

(e) $\mathbf{E}\, x^2 \geq \alpha$.

(f) $\mathbf{var}(x) \leq \alpha$, where $\mathbf{var}(x) = \mathbf{E}(x - \mathbf{E}\, x)^2$ is the variance of x.

(g) $\mathbf{var}(x) \geq \alpha$.

(h) $\mathbf{quartile}(x) \geq \alpha$, where $\mathbf{quartile}(x) = \inf\{\beta \mid \mathbf{prob}(x \leq \beta) \geq 0.25\}$.

(i) $\mathbf{quartile}(x) \leq \alpha$.

Solution. We first note that the constraints $p_i \geq 0$, $i = 1, \ldots, n$, define halfspaces, and $\sum_{i=1}^n p_i = 1$ defines a hyperplane, so P is a polyhedron.
The first five constraints are, in fact, linear inequalities in the probabilities p_i.

(a) $\mathbf{E}\, f(x) = \sum_{i=1}^n p_i f(a_i)$, so the constraint is equivalent to two linear inequalities

$$\alpha \leq \sum_{i=1}^n p_i f(a_i) \leq \beta.$$

(b) $\mathbf{prob}(x \geq \alpha) = \sum_{i:\, a_i \geq \alpha} p_i$, so the constraint is equivalent to a linear inequality

$$\sum_{i:\, a_i \geq \alpha} p_i \leq \beta.$$

(c) The constraint is equivalent to a linear inequality

$$\sum_{i=1}^n p_i(|a_i^3| - \alpha |a_i|) \leq 0.$$

(d) The constraint is equivalent to a linear inequality

$$\sum_{i=1}^n p_i a_i^2 \leq \alpha.$$

(e) The constraint is equivalent to a linear inequality

$$\sum_{i=1}^n p_i a_i^2 \geq \alpha.$$

The first five constraints therefore define convex sets.

(f) The constraint

$$\mathbf{var}(x) = \mathbf{E}\, x^2 - (\mathbf{E}\, x)^2 = \sum_{i=1}^n p_i a_i^2 - (\sum_{i=1}^n p_i a_i)^2 \leq \alpha$$

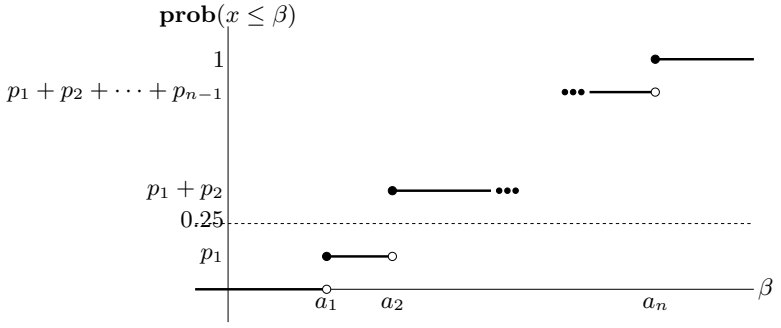
is not convex in general. As a counterexample, we can take $n = 2$, $a_1 = 0$, $a_2 = 1$, and $\alpha = 1/5$. $p = (1, 0)$ and $p = (0, 1)$ are two points that satisfy $\mathbf{var}(x) \leq \alpha$, but the convex combination $p = (1/2, 1/2)$ does not.

(g) This constraint is equivalent to

$$\sum_{i=1}^n a_i^2 p_i + (\sum_{i=1}^n a_i p_i)^2 = b^T p + p^T A p \leq \alpha$$

where $b_i = a_i^2$ and $A = aa^T$. This defines a convex set, since the matrix aa^T is positive semidefinite.

Let us denote $\mathbf{quartile}(x) = f(p)$ to emphasize it is a function of p. The figure illustrates the definition. It shows the cumulative distribution for a distribution p with $f(p) = a_2$.



(h) The constraint $f(p) \geq \alpha$ is equivalent to

$$\mathbf{prob}(x \leq \beta) < 0.25 \text{ for all } \beta < \alpha.$$

If $\alpha \leq a_1$, this is always true. Otherwise, define $k = \max\{i \mid a_i < \alpha\}$. This is a fixed integer, independent of p. The constraint $f(p) \geq \alpha$ holds if and only if

$$\mathbf{prob}(x \leq a_k) = \sum_{i=1}^k p_i < 0.25.$$

This is a strict linear inequality in p, which defines an open halfspace.

(i) The constraint $f(p) \leq \alpha$ is equivalent to

$$\mathbf{prob}(x \leq \beta) \geq 0.25 \text{ for all } \beta \geq \alpha.$$

This can be expressed as a linear inequality

$$\sum_{i=k+1}^n p_i \geq 0.25.$$

(If $\alpha \leq a_1$, we define $k = 0$.)

Operations that preserve convexity

2.16 Show that if S_1 and S_2 are convex sets in $\mathbf{R}^{m\times n}$, then so is their partial sum

$$S = \{(x, y_1 + y_2) \mid x \in \mathbf{R}^m,\ y_1,\ y_2 \in \mathbf{R}^n, (x, y_1) \in S_1,\ (x, y_2) \in S_2\}.$$

Solution. We consider two points $(\bar{x}, \bar{y}_1 + \bar{y}_2)$, $(\tilde{x}, \tilde{y}_1 + \tilde{y}_2) \in S$, *i.e.*, with

$$(\bar{x}, \bar{y}_1) \in S_1, \qquad (\bar{x}, \bar{y}_2) \in S_2, \qquad (\tilde{x}, \tilde{y}_1) \in S_1, \qquad (\tilde{x}, \tilde{y}_2) \in S_2.$$

For $0 \le \theta \le 1$,

$$\theta(\bar{x}, \bar{y}_1 + \bar{y}_2) + (1-\theta)(\tilde{x}, \tilde{y}_1 + \tilde{y}_2) = (\theta\bar{x} + (1-\theta)\tilde{x}, (\theta\bar{y}_1 + (1-\theta)\tilde{y}_1) + (\theta\bar{y}_2 + (1-\theta)\tilde{y}_2))$$

is in S because, by convexity of S_1 and S_2,

$$(\theta\bar{x} + (1-\theta)\tilde{x}, \theta\bar{y}_1 + (1-\theta)\tilde{y}_1) \in S_1, \qquad (\theta\bar{x} + (1-\theta)\tilde{x}, \theta\bar{y}_2 + (1-\theta)\tilde{y}_2) \in S_2.$$

2.17 *Image of polyhedral sets under perspective function.* In this problem we study the image of hyperplanes, halfspaces, and polyhedra under the perspective function $P(x, t) = x/t$, with $\mathbf{dom}\, P = \mathbf{R}^n \times \mathbf{R}_{++}$. For each of the following sets C, give a simple description of

$$P(C) = \{v/t \mid (v, t) \in C,\ t > 0\}.$$

(a) The polyhedron $C = \mathbf{conv}\{(v_1, t_1), \ldots, (v_K, t_K)\}$ where $v_i \in \mathbf{R}^n$ and $t_i > 0$.
Solution. The polyhedron

$$P(C) = \mathbf{conv}\{v_1/t_1, \ldots, v_K/t_K\}.$$

We first show that $P(C) \subseteq \mathbf{conv}\{v_1/t_1, \ldots, v_K/t_K\}$. Let $x = (v, t) \in C$, with

$$v = \sum_{i=1}^K \theta_i v_i, \qquad t = \sum_{i=1}^K \theta_i t_i,$$

and $\theta \succeq 0$, $\mathbf{1}^T\theta = 1$. The image $P(x)$ can be expressed as

$$P(x) = v/t = \frac{\sum_{i=1}^K \theta_i v_i}{\sum_{i=1}^K \theta_i t_i} = \sum_{i=1}^K \mu_i v_i/t_i$$

where

$$\mu_i = \frac{\theta_i t_i}{\sum_{k=1}^K \theta_k t_k}, \quad i = 1, \ldots, K.$$

It is clear that $\mu \succeq 0$, $\mathbf{1}^T\mu = 1$, so we can conclude that $P(x) \in \mathbf{conv}\{v_1/t_1, \ldots, v_K/t_K\}$ for all $x \in C$.
Next, we show that $P(C) \supseteq \mathbf{conv}\{v_1/t_1, \ldots, v_K/t_K\}$. Consider a point

$$z = \sum_{i=1}^K \mu_i v_i/t_i$$

with $\mu \succeq 0$, $\mathbf{1}^T\mu = 1$. Define

$$\theta_i = \frac{\mu_i}{t_i \sum_{j=1}^K \mu_j/t_j}, \quad i = 1, \ldots, K.$$

It is clear that $\theta \succeq 0$ and $\mathbf{1}^T\theta = 1$. Moreover, $z = P(v, t)$ where

$$t = \sum_i \theta_i t_i = \frac{\sum_i \mu_i}{\sum_j \mu_j/t_j} = \frac{1}{\sum_j \mu_j/t_j}, \qquad v = \sum_i \theta_i v_i,$$

i.e., $(v, t) \in C$.

(b) The hyperplane $C = \{(v,t) \mid f^T v + gt = h\}$ (with f and g not both zero).
Solution.

$$\begin{aligned} P(C) &= \{z \mid f^T z + g = h/t \text{ for some } t > 0\} \\ &= \begin{cases} \{z \mid f^T z + g = 0\} & h = 0 \\ \{z \mid f^T z + g > 0\} & h > 0 \\ \{z \mid f^T z + g < 0\} & h < 0. \end{cases} \end{aligned}$$

(c) The halfspace $C = \{(v,t) \mid f^T v + gt \leq h\}$ (with f and g not both zero).
Solution.

$$\begin{aligned} P(C) &= \{z \mid f^T z + g \leq h/t \text{ for some } t > 0\} \\ &= \begin{cases} \{z \mid f^T z + g \leq 0\} & h = 0 \\ \mathbf{R}^n & h > 0 \\ \{z \mid f^T z + g < 0\} & h < 0. \end{cases} \end{aligned}$$

(d) The polyhedron $C = \{(v,t) \mid Fv + gt \preceq h\}$.
Solution.

$$P(C) = \{z \mid Fz + g \preceq (1/t)h \text{ for some } t > 0\}.$$

More explicitly, $z \in P(C)$ if and only if it satisfies the following conditions:

- $f_i^T z + g_i \leq 0$ if $h_i = 0$
- $f_i^T z + g_i < 0$ if $h_i < 0$
- $(f_i^T z + g_i)/h_i \leq (f_k^T z + g_k)/h_k$ if $h_i > 0$ and $h_k < 0$.

2.18 *Invertible linear-fractional functions.* Let $f : \mathbf{R}^n \to \mathbf{R}^n$ be the linear-fractional function

$$f(x) = (Ax + b)/(c^T x + d), \qquad \mathbf{dom}\, f = \{x \mid c^T x + d > 0\}.$$

Suppose the matrix

$$Q = \begin{bmatrix} A & b \\ c^T & d \end{bmatrix}$$

is nonsingular. Show that f is invertible and that f^{-1} is a linear-fractional mapping. Give an explicit expression for f^{-1} and its domain in terms of A, b, c, and d. *Hint.* It may be easier to express f^{-1} in terms of Q.

Solution. This follows from remark 2.2 on page 41. The inverse of f is given by

$$f^{-1}(x) = \mathcal{P}^{-1}(Q^{-1}\mathcal{P}(x)),$$

so f^{-1} is the projective transformation associated with Q^{-1}.

2.19 *Linear-fractional functions and convex sets.* Let $f : \mathbf{R}^m \to \mathbf{R}^n$ be the linear-fractional function

$$f(x) = (Ax + b)/(c^T x + d), \qquad \mathbf{dom}\, f = \{x \mid c^T x + d > 0\}.$$

In this problem we study the inverse image of a convex set C under f, *i.e.*,

$$f^{-1}(C) = \{x \in \mathbf{dom}\, f \mid f(x) \in C\}.$$

For each of the following sets $C \subseteq \mathbf{R}^n$, give a simple description of $f^{-1}(C)$.

(a) The halfspace $C = \{y \mid g^T y \leq h\}$ (with $g \neq 0$).
Solution.

$$\begin{aligned} f^{-1}(C) &= \{x \in \mathbf{dom}\, f \mid g^T f(x) \leq h\} \\ &= \{x \mid g^T(Ax+b)/(c^T x+d) \leq h,\ c^T x + d > 0\} \\ &= \{x \mid (A^T g - hc)^T x \leq hd - g^T b,\ c^T x + d > 0\}, \end{aligned}$$

which is another halfspace, intersected with $\mathbf{dom}\, f$.

(b) The polyhedron $C = \{y \mid Gy \preceq h\}$.
Solution. The polyhedron

$$\begin{aligned} f^{-1}(C) &= \{x \in \mathbf{dom}\, f \mid Gf(x) \preceq h\} \\ &= \{x \mid G(Ax+b)/(c^T x+d) \preceq h,\ c^T x + d > 0\} \\ &= \{x \mid (GA - hc^T)x \leq hd - Gb,\ c^T x + d > 0\}, \end{aligned}$$

a polyhedron intersected with $\mathbf{dom}\, f$.

(c) The ellipsoid $\{y \mid y^T P^{-1} y \leq 1\}$ (where $P \in \mathbf{S}^n_{++}$).
Solution.

$$\begin{aligned} f^{-1}(C) &= \{x \in \mathbf{dom}\, f \mid f(x)^T P^{-1} f(x) \leq 1\} \\ &= \{x \in \mathbf{dom}\, f \mid (Ax+b)^T P^{-1}(Ax+b) \leq (c^T x + d)^2\}, \\ &= \{x \mid x^T Q x + 2q^T x \leq r,\ c^T x + d > 0\}. \end{aligned}$$

where $Q = A^T P^{-1} A - cc^T$, $q = b^T P^{-1} A + dc$, $r = d^2 - b^T P^{-1} b$. If $A^T P^{-1} A \succ cc^T$ this is an ellipsoid intersected with $\mathbf{dom}\, f$.

(d) The solution set of a linear matrix inequality, $C = \{y \mid y_1 A_1 + \cdots + y_n A_n \preceq B\}$, where $A_1, \ldots, A_n, B \in \mathbf{S}^p$.
Solution. We denote by a_i^T the ith row of A.

$$\begin{aligned} f^{-1}(C) &= \{x \in \mathbf{dom}\, f \mid f_1(x)A_1 + f_2(x)A_2 + \cdots + f_n(x)A_n \preceq B\} \\ &= \{x \in \mathbf{dom}\, f \mid (a_1^T x + b_1)A_1 + \cdots + (a_n^T x + b_n)A_n \preceq (c^T x + d)B\} \\ &= \{x \in \mathbf{dom}\, f \mid G_1 x_1 + \cdots + G_m x_m \preceq H,\ c^T x + d > 0\} \end{aligned}$$

where

$$G_i = a_{1i}A_1 + a_{2i}A_2 + \cdots + a_{ni}A_n - c_i B, \qquad H = dB - b_1 A_1 - b_2 A_2 - \cdots - b_n A_n.$$

$f^{-1}(C)$ is the intersection of $\mathbf{dom}\, f$ with the solution set of an LMI.

Separation theorems and supporting hyperplanes

2.20 *Strictly positive solution of linear equations.* Suppose $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$, with $b \in \mathcal{R}(A)$. Show that there exists an x satisfying

$$x \succ 0, \qquad Ax = b$$

if and only if there exists no λ with

$$A^T \lambda \succeq 0, \qquad A^T \lambda \neq 0, \qquad b^T \lambda \leq 0.$$

Hint. First prove the following fact from linear algebra: $c^T x = d$ for all x satisfying $Ax = b$ if and only if there is a vector λ such that $c = A^T \lambda$, $d = b^T \lambda$.

Solution. We first prove the result in the hint. Suppose that there exists a λ such that $c = A^T\lambda$, $d = b^T\lambda$. It is clear that if $Ax = b$ then

$$c^T x = \lambda^T Ax = \lambda^T b = d.$$

Conversely, suppose $Ax = b$ implies $c^T x = d$, and that $\mathbf{rank}\, A = r$. Let $F \in \mathbf{R}^{n\times(n-r)}$ be a matrix with $\mathcal{R}(F) = \mathcal{N}(A)$, and let x_0 be a solution of $Ax = b$. Then $Ax = b$ if and only if $x = Fy + x_0$ for some y, and $c^T x = d$ for all $x = Fy + x_0$ implies

$$c^T Fy + c^T x_0 = d$$

for all y. This is only possible if $F^T c = 0$, *i.e.*, $c \in \mathcal{N}(F^T) = \mathcal{R}(A^T)$, *i.e.*, there exists a λ such that $c = A^T\lambda$. The condition $c^T Fy + c^T x_0 = d$ then reduces to $c^T x_0 = d$, *i.e.*, $\lambda^T Ax_0 = \lambda^T b = d$. In conclusion, if $c^T x = d$ for all x with $Ax = b$, then there there exists a λ such that $c = A^T\lambda$ and $d = b^T\lambda$.

To prove the main result, we use a standard separating hyperplane argument, applied to the sets $C = \mathbf{R}^n_{++}$ and $D = \{x \mid Ax = b\}$. If they are disjoint, there exists $c \neq 0$ and d such that $c^T x \geq d$ for all $x \in C$ and $c^T x \leq d$ for all $x \in D$. The first condition means that $c \succeq 0$ and $d \leq 0$. Since $c^T x \leq d$ on D, which is an affine set, we must have $c^T x$ constant on D. (If $c^T x$ weren't constant on D, it would take on all values.) We can relabel d to be this constant value, so we have $c^T x = d$ on D. Now using the hint, there is some λ such that $c = A^T\lambda$, $d = b^T\lambda$.

2.21 *The set of separating hyperplanes.* Suppose that C and D are disjoint subsets of $\mathbf{R}^n$. Consider the set of $(a, b) \in \mathbf{R}^{n+1}$ for which $a^T x \leq b$ for all $x \in C$, and $a^T x \geq b$ for all $x \in D$. Show that this set is a convex cone (which is the singleton $\{0\}$ if there is no hyperplane that separates C and D).

Solution. The conditions $a^T x \leq b$ for all $x \in C$ and $a^T x \geq b$ for all $x \in D$, form a set of homogeneous linear inequalities in (a, b). Therefore K is the intersection of halfspaces that pass through the origin. Hence it is a convex cone.

Note that this does not require convexity of C or D.

2.22 Finish the proof of the separating hyperplane theorem in §2.5.1: Show that a separating hyperplane exists for two disjoint convex sets C and D. You can use the result proved in §2.5.1, *i.e.*, that a separating hyperplane exists when there exist points in the two sets whose distance is equal to the distance between the two sets.

Hint. If C and D are disjoint convex sets, then the set $\{x - y \mid x \in C,\ y \in D\}$ is convex and does not contain the origin.

Solution. Following the hint, we first confirm that

$$S = \{x - y \mid x \in C,\ y \in D\},$$

is convex, since it is the sum of two convex sets.

Since C and D are disjoint, $0 \notin S$. We distinguish two cases. First suppose $0 \notin \mathbf{cl}\, S$. The partial separating hyperplane in §2.5.1 applies to the sets $\{0\}$ and $\mathbf{cl}\, S$, so there exists an $a \neq 0$ such that

$$a^T(x - y) > 0$$

for all $x - y \in \mathbf{cl}\, S$. In particular this also holds for all $x - y \in S$, *i.e.*, $a^T x > a^T y$ for all $x \in C$ and $y \in D$.

Next, assume $0 \in \mathbf{cl}\, S$. Since $0 \notin S$, it must be in the boundary of S. If S has empty interior, it is contained in a hyperplane $\{z \mid a^T z = b\}$, which must include the origin, hence $b = 0$. In other words, $a^T x = a^T y$ for all $x \in C$ and all $y \in D$, so we have a trivial separating hyperplane.

If S has nonempty interior, we consider the set

$$S_{-\epsilon} = \{z \mid B(z, \epsilon) \subseteq S\},$$

where $B(z,\epsilon)$ is the Euclidean ball with center z and radius $\epsilon > 0$. $S_{-\epsilon}$ is the set S, shrunk by ϵ (see exercise 2.14). $\mathbf{cl}\, S_{-\epsilon}$ is closed and convex, and does not contain 0, so by the partial separating hyperplane result, it is strictly separated from $\{0\}$ by at least one hyperplane with normal vector $a(\epsilon)$:

$$a(\epsilon)^T z > 0 \text{ for all } z \in S_{-\epsilon}.$$

Without loss of generality we assume $\|a(\epsilon)\|_2 = 1$. Now let ϵ_k, $k = 1, 2, \ldots$ be a sequence of positive values of ϵ_k with $\lim_{k\to\infty} \epsilon_k = 0$. Since $\|a(\epsilon_k)\|_2 = 1$ for all k, the sequence $a(\epsilon_k)$ contains a convergent subsequence, and we will denote its limit by $\bar{a}$. We have

$$a(\epsilon_k)^T z > 0 \text{ for all } z \in S_{-\epsilon_k}$$

for all k, and therefore $\bar{a}^T z > 0$ for all $z \in \mathbf{int}\, S$, and $\bar{a}^T z \geq 0$ for all $z \in S$, *i.e.*,

$$\bar{a}^T x \geq \bar{a}^T y$$

for all $x \in C$, $y \in D$.

2.23 Give an example of two closed convex sets that are disjoint but cannot be strictly separated.

Solution. Take $C = \{x \in \mathbf{R}^2 \mid x_2 \leq 0\}$ and $D = \{x \in \mathbf{R}^2_+ \mid x_1 x_2 \geq 1\}$.

2.24 *Supporting hyperplanes.*

(a) Express the closed convex set $\{x \in \mathbf{R}^2_+ \mid x_1 x_2 \geq 1\}$ as an intersection of halfspaces.

Solution. The set is the intersection of all supporting halfspaces at points in its boundary, which is given by $\{x \in \mathbf{R}^2_+ \mid x_1 x_2 = 1\}$. The supporting hyperplane at $x = (t, 1/t)$ is given by

$$x_1/t^2 + x_2 = 2/t,$$

so we can express the set as

$$\bigcap_{t>0} \{x \in \mathbf{R}^2 \mid x_1/t^2 + x_2 \geq 2/t\}.$$

(b) Let $C = \{x \in \mathbf{R}^n \mid \|x\|_\infty \leq 1\}$, the ℓ_∞-norm unit ball in $\mathbf{R}^n$, and let $\hat{x}$ be a point in the boundary of C. Identify the supporting hyperplanes of C at $\hat{x}$ explicitly.

Solution. $s^T x \geq s^T \hat{x}$ for all $x \in C$ if and only if

$$\begin{array}{ll} s_i < 0 & \hat{x}_i = 1 \\ s_i > 0 & \hat{x}_i = -1 \\ s_i = 0 & -1 < \hat{x}_i < 1. \end{array}$$

2.25 *Inner and outer polyhedral approximations.* Let $C \subseteq \mathbf{R}^n$ be a closed convex set, and suppose that $x_1, \ldots, x_K$ are on the boundary of C. Suppose that for each i, $a_i^T(x - x_i) = 0$ defines a supporting hyperplane for C at x_i, *i.e.*, $C \subseteq \{x \mid a_i^T(x - x_i) \leq 0\}$. Consider the two polyhedra

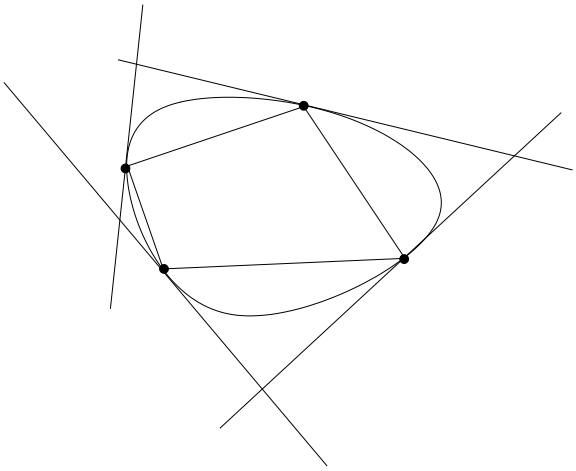
$$P_{\text{inner}} = \mathbf{conv}\{x_1, \ldots, x_K\}, \qquad P_{\text{outer}} = \{x \mid a_i^T(x - x_i) \leq 0,\ i = 1, \ldots, K\}.$$

Show that $P_{\text{inner}} \subseteq C \subseteq P_{\text{outer}}$. Draw a picture illustrating this.

Solution. The points x_i are in C because C is closed. Any point in $P_{\text{inner}} = \mathbf{conv}\{x_1, \ldots, x_K\}$ is also in C because C is convex. Therefore $P_{\text{inner}} \subseteq C$.

If $x \in C$ then $a_i^T(x - x_i) \leq 0$ for $i = 1, \ldots, K$, *i.e.*, $x \in P_{\text{outer}}$. Therefore $C \subseteq P_{\text{outer}}$.

The figure shows an example with $K = 4$.



2.26 *Support function.* The support function of a set $C \subseteq \mathbf{R}^n$ is defined as

$$S_C(y) = \sup\{y^T x \mid x \in C\}.$$

(We allow $S_C(y)$ to take on the value $+\infty$.) Suppose that C and D are closed convex sets in $\mathbf{R}^n$. Show that $C = D$ if and only if their support functions are equal.

Solution. Obviously if $C = D$ the support functions are equal. We show that if the support functions are equal, then $C = D$, by showing that $D \subseteq C$ and $C \subseteq D$.

We first show that $D \subseteq C$. Suppose there exists a point $x_0 \in D$, $x \notin C$. Since C is closed, x_0 can be strictly separated from C, *i.e.*, there exists an $a \neq 0$ with $a^T x_0 > b$ and $a^T x < b$ for all $x \in C$. This means that

$$\sup_{x \in C} a^T x \leq b < a^T x_0 \leq \sup_{x \in D} a^T x,$$

which implies that $S_C(a) \neq S_D(a)$. By repeating the argument with the roles of C and D reversed, we can show that $C \subseteq D$.

2.27 *Converse supporting hyperplane theorem.* Suppose the set C is closed, has nonempty interior, and has a supporting hyperplane at every point in its boundary. Show that C is convex.

Solution. Let H be the set of all halfspaces that contain C. H is a closed convex set, and contains C by definition.

The support function S_C of a set C is defined as $S_C(y) = \sup_{x \in C} y^T x$. The set H and its interior can be defined in terms of the support function as

$$H = \bigcap_{y \neq 0} \{x \mid y^T x \leq S_C(y)\}, \qquad \mathbf{int}\, H = \bigcap_{y \neq 0} \{x \mid y^T x < S_C(y)\},$$

and the boundary of H is the set of all points in H with $y^T x = S_C(y)$ for at least one $y \neq 0$.

By definition $\mathbf{int}\, C \subseteq \mathbf{int}\, H$. We also have $\mathbf{bd}\, C \subseteq \mathbf{bd}\, H$: if $\bar{x} \in \mathbf{bd}\, C$, then there exists a supporting hyperplane at $\bar{x}$, *i.e.*, a vector $a \neq 0$ such that $a^T \bar{x} = S_C(a)$, *i.e.*, $\bar{x} \in \mathbf{bd}\, H$.

We now show that these properties imply that C is convex. Consider an arbitrary line intersecting $\mathbf{int}\, C$. The intersection is a union of disjoint open intervals I_k, with endpoints in $\mathbf{bd}\, C$ (hence also in $\mathbf{bd}\, H$), and interior points in $\mathbf{int}\, C$ (hence also in $\mathbf{int}\, H$). Now $\mathbf{int}\, H$ is a convex set, so the interior points of two different intervals I_1 and I_2 can not be separated by boundary points (since boundary points are in $\mathbf{bd}\, H$, not in $\mathbf{int}\, H$). Therefore there can be at most one interval, *i.e.*, $\mathbf{int}\, C$ is convex.

Convex cones and generalized inequalities

2.28 *Positive semidefinite cone for $n = 1,\ 2,\ 3$.* Give an explicit description of the positive semidefinite cone $\mathbf{S}^n_+$, in terms of the matrix coefficients and ordinary inequalities, for $n = 1,\ 2,\ 3$. To describe a general element of $\mathbf{S}^n$, for $n = 1,\ 2,\ 3$, use the notation

$$x_1, \qquad \begin{bmatrix} x_1 & x_2 \\ x_2 & x_3 \end{bmatrix}, \qquad \begin{bmatrix} x_1 & x_2 & x_3 \\ x_2 & x_4 & x_5 \\ x_3 & x_5 & x_6 \end{bmatrix}.$$

Solution. For $n = 1$ the condition is $x_1 \geq 0$. For $n = 2$ the condition is

$$x_1 \geq 0, \qquad x_3 \geq 0, \qquad x_1x_3 - x_2^2 \geq 0.$$

For $n = 3$ the condition is

$$x_1 \geq 0, \qquad x_2 \geq 0, \qquad x_3 \geq 0, \qquad x_1x_4 - x_2^2 \geq 0, \qquad x_4x_6 - x_5^2 \geq 0, \qquad x_1x_6 - x_3^2 \geq 0$$

and

$$x_1x_4x_6 + 2x_2x_3x_5 - x_1x_5^2 - x_6x_2^2 - x_4x_3^2 \geq 0,$$

i.e., all principal minors must be nonnegative.
We give the proof for $n = 3$, assuming the result is true for $n = 2$. The matrix

$$X = \begin{bmatrix} x_1 & x_2 & x_3 \\ x_2 & x_4 & x_5 \\ x_3 & x_5 & x_6 \end{bmatrix}$$

is positive semidefinite if and only if

$$z^T X z = x_1z_1^2 + 2x_2z_1z_2 + 2x_3z_1z_3 + x_4z_2^2 + 2x_5z_2z_3 + x_6z_3^2 \geq 0$$

for all z.
If $x_1 = 0$, we must have $x_2 = x_3 = 0$, so $X \succeq 0$ if and only if

$$\begin{bmatrix} x_4 & x_5 \\ x_5 & x_6 \end{bmatrix} \succeq 0.$$

Applying the result for the 2×2-case, we conclude that if $x_1 = 0$, $X \succeq 0$ if and only if

$$x_2 = x_3 = 0, \qquad x_4 \geq 0, \qquad x_6 \geq 0, \qquad x_4x_6 - x_5^2 \geq 0.$$

Now assume $x_1 \neq 0$. We have

$$z^T X z = x_1(z_1 + (x_2/x_1)z_2 + (x_3/x_1)z_3)^2 + (x_4 - x_2^2/x_1)z_2^2 + (x_6 - x_3^2/x_1)z_3^2 + 2(x_5 - x_2x_3/x_1)z_2z_3,$$

so it is clear that we must have $x_1 > 0$ and

$$\begin{bmatrix} x_4 - x_2^2/x_1 & x_5 - x_2x_3/x_1 \\ x_5 - x_2x_3/x_1 & x_6 - x_3^2/x_1 \end{bmatrix} \succeq 0.$$

By the result for 2×2-case studied above, this is equivalent to

$$x_1x_4 - x_2^2 \geq 0, \qquad x_1x_6 - x_3^2 \geq 0, \qquad (x_4 - x_2^2/x_1)(x_6 - x_3^2/x_1) - (x_5 - x_2x_3/x_1)^2 \geq 0.$$

The third inequality simplifies to

$$(x_1x_4x_6 - 2x_2x_3x_5 - x_1x_5^2 - x_6x_2^2 - x_4x_3^2)/x_1 \geq 0.$$

Therefore, if $x_1 > 0$, then $X \succeq 0$ if and only if

$$x_1x_4 - x_2^2 \geq 0, \qquad x_1x_6 - x_3^2 \geq 0, \qquad (x_1x_4x_6 - 2x_2x_3x_5 - x_1x_5^2 - x_6x_2^2 - x_4x_3^2)/x_1 \geq 0.$$

We can combine the conditions for $x_1 = 0$ and $x_1 > 0$ by saying that all 7 principal minors must be nonnegative.

Exercises

2.29 *Cones in* $\mathbf{R}^2$. Suppose $K \subseteq \mathbf{R}^2$ is a closed convex cone.

(a) Give a simple description of K in terms of the polar coordinates of its elements ($x = r(\cos\phi, \sin\phi)$ with $r \geq 0$).

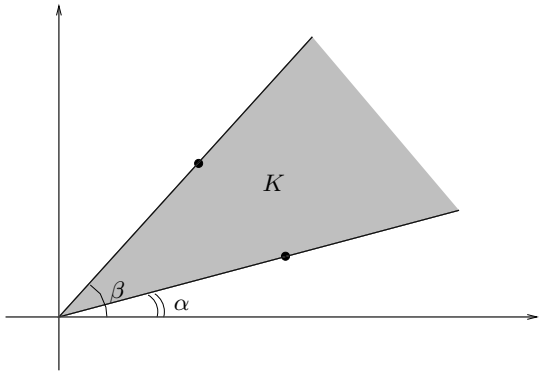
(b) Give a simple description of K^*, and draw a plot illustrating the relation between K and K^*.

(c) When is K pointed?

(d) When is K proper (hence, defines a generalized inequality)? Draw a plot illustrating what $x \preceq_K y$ means when K is proper.

Solution.

(a) In $\mathbf{R}^2$ a cone K is a "pie slice" (see figure).



In terms of polar coordinates, a pointed closed convex cone K can be expressed

$$K = \{(r\cos\phi, r\sin\phi) \mid r \geq 0, \alpha \leq \phi \leq \beta\}$$

where $0 \leq \beta - \alpha < 180^\circ$. When $\beta - \alpha = 180^\circ$, this gives a non-pointed cone (a halfspace). Other possible non-pointed cones are the entire plane

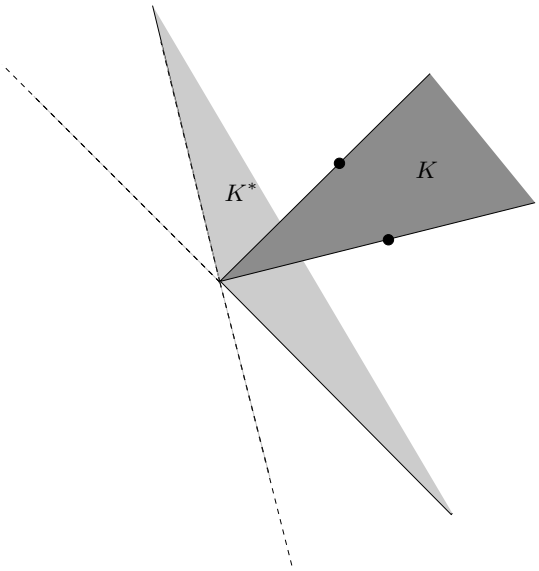
$$K = \{(r\cos\phi, r\sin\phi) \mid r \geq 0, 0 \leq \phi \leq 2\pi\} = \mathbf{R}^2,$$

and lines through the origin

$$K = \{(r\cos\alpha, r\sin\alpha) \mid r \in \mathbf{R}\}.$$

(b) By definition, K^* is the intersection of all halfspaces $x^T y \geq 0$ where $x \in K$. However, as can be seen from the figure, if K is pointed, the two halfspaces defined by the extreme rays are sufficient to define K^*, *i.e.*,

$$K^* = \{y \mid y_1\cos\alpha + y_2\sin\alpha \geq 0, y_1\cos\beta + y_2\sin\beta \geq 0\}.$$



If K is a halfspace, $K = \{x \mid v^T x \geq 0\}$, the dual cone is the ray

$$K^* = \{tv \mid t \geq 0\}.$$

If $K = \mathbf{R}^2$, the dual cone is $K^* = \{0\}$. If K is a line $\{tv \mid t \in \mathbf{R}\}$ through the origin, the dual cone is the line perpendicular to v

$$K^* = \{y \mid v^T y = 0\}.$$

(c) See part (a).

(d) K must be closed convex and pointed, and have nonempty interior. From part (a), this means K can be expressed as

$$K = \{(r\cos\phi, r\sin\phi) \mid r \geq 0, \alpha \leq \phi \leq \beta\}$$

where $0 < \beta - \alpha < 180^\circ$.

$x \preceq_K y$ means $y \in x + K$.

2.30 *Properties of generalized inequalities.* Prove the properties of (nonstrict and strict) generalized inequalities listed in §2.4.1.

Solution.

Properties of generalized inequalities.

(a) $\preceq_K$ *is preserved under addition.* If $y - x \in K$ and $v - u \in K$, where K is a convex cone, then the conic combination $(y - x) + (v - u) \in K$, *i.e.*, $x + u \preceq_K y + v$.

(b) $\preceq_K$ *is transitive.* If $y - x \in K$ and $z - y \in K$ then the conic combination $(y - x) + (z - y) = z - x \in K$, *i.e.*, $x \preceq_K z$.

(c) $\preceq_K$ *is preserved under nonnegative scaling.* Follows from the fact that K is a cone.

(d) $\preceq_K$ *is reflexive.* Any cone contains the origin.

(e) $\preceq_K$ *is antisymmetric.* If $y - x \in K$ and $x - y \in K$, then $y - x = 0$ because K is pointed.

(f) $\preceq_K$ *is preserved under limits.* If $y_i - x_i \in K$ and K is closed, then $\lim_{i\to\infty}(y_i - x_i) \in K$.

Properties of strict inequality.

(a) *If $x \prec_K y$ then $x \preceq_K y$.* Every set contains its interior.

(b) *If $x \prec_K y$ and $u \preceq_K v$ then $x + u \prec_K y + v$.* If $y - x \in \mathbf{int}\, K$, then $(y - x) + z \in K$ for all sufficiently small nonzero z. Since K is a convex cone and $v - u \in K$, $(y - x) + z + (u - v) \in K$ for all sufficiently small u, *i.e.*, $x + u \prec_K y + v$.

(c) *If $x \prec_K y$ and $\alpha > 0$ then $\alpha x \prec_K \alpha y$.* If $y - x + z \in K$ for sufficiently small nonzero z, then $\alpha(y - x + z) \in K$ for all $\alpha > 0$, *i.e.*, $\alpha(y - x) + \tilde{z} \in K$ for all sufficiently small nonzero $\tilde{z}$.

(d) $x \not\prec_K x$. $0 \notin \mathbf{int}\, K$ because K is a pointed cone.

(e) *If $x \prec_K y$, then for u and v small enough, $x + u \prec_K y + v$.* If $y - x \in \mathbf{int}\, K$, then $(y - x) + (v - u) \in \mathbf{int}\, K$ for sufficiently small u and v.

2.31 *Properties of dual cones.* Let K^* be the dual cone of a convex cone K, as defined in (2.19). Prove the following.

(a) K^* is indeed a convex cone.

Solution. K^* is the intersection of a set of homogeneous halfspaces (meaning, halfspaces that include the origin as a boundary point). Hence it is a closed convex cone.

(b) $K_1 \subseteq K_2$ implies $K_2^* \subseteq K_1^*$.

Solution. $y \in K_2^*$ means $x^T y \geq 0$ for all $x \in K_2$, which is includes K_1, therefore $x^T y \geq 0$ for all $x \in K_1$.

(c) K^* is closed.

Solution. See part (a).

(d) The interior of K^* is given by $\mathbf{int}\, K^* = \{y \mid y^T x > 0 \text{ for all } x \in K\}$.

Solution. If $y^T x > 0$ for all $x \in K$ then $(y + u)^T x > 0$ for all $x \in K$ and all sufficiently small u; hence $y \in \mathbf{int}\, K$.

Conversely if $y \in K^*$ and $y^T x = 0$ for some $x \in K$, then $y \notin \mathbf{int}\, K^*$ because $(y - tx)^T x < 0$ for all $t > 0$.

(e) If K has nonempty interior then K^* is pointed.

Solution. Suppose K^* is not pointed, *i.e.*, there exists a nonzero $y \in K^*$ such that $-y \in K^*$. This means $y^T x \geq 0$ and $-y^T x \geq 0$ for all $x \in K$, *i.e.*, $y^T x = 0$ for all $x \in K$, hence K has empty interior.

(f) K^{**} is the closure of K. (Hence if K is closed, $K^{**} = K$.)

Solution. By definition of K^*, $y \neq 0$ is the normal vector of a (homogeneous) halfspace containing K if and only if $y \in K^*$. The intersection of all homogeneous halfspaces containing a convex cone K is the closure of K. Therefore the closure of K is

$$\mathbf{cl}\, K = \bigcap_{y \in K^*} \{x \mid y^T x \geq 0\} = \{x \mid y^T x \geq 0 \text{ for all } y \in K^*\} = K^{**}.$$

(g) If the closure of K is pointed then K^* has nonempty interior.

Solution. If K^* has empty interior, there exists an $a \neq 0$ such that $a^T y = 0$ for all $y \in K^*$. This means a and $-a$ are both in K^{**}, which contradicts the fact that K^{**} is pointed.

As an example that shows that it is not sufficient that K is pointed, consider $K = \{0\} \cup \{(x_1, x_2) \mid x_1 > 0\}$. This is a pointed cone, but its dual has empty interior.

2.32 Find the dual cone of $\{Ax \mid x \succeq 0\}$, where $A \in \mathbf{R}^{m \times n}$.

Solution. $K^* = \{y \mid A^T y \succeq 0\}$.

2.33 *The monotone nonnegative cone.* We define the *monotone nonnegative cone* as

$$K_{\mathrm{m}+} = \{x \in \mathbf{R}^n \mid x_1 \geq x_2 \geq \cdots \geq x_n \geq 0\}.$$

i.e., all nonnegative vectors with components sorted in nonincreasing order.

(a) Show that $K_{\mathrm{m}+}$ is a proper cone.

(b) Find the dual cone $K_{\mathrm{m}+}^*$. *Hint.* Use the identity

$$\begin{aligned}\sum_{i=1}^n x_i y_i &= (x_1 - x_2)y_1 + (x_2 - x_3)(y_1 + y_2) + (x_3 - x_4)(y_1 + y_2 + y_3) + \cdots \\ &\quad + (x_{n-1} - x_n)(y_1 + \cdots + y_{n-1}) + x_n(y_1 + \cdots + y_n).\end{aligned}$$

Solution.

(a) The set $K_{\mathrm{m}+}$ is defined by n homogeneous linear inequalities, hence it is a closed (polyhedral) cone.

The interior of $K_{\mathrm{m}+}$ is nonempty, because there are points that satisfy the inequalities with strict inequality, for example, $x = (n, n-1, n-2, \ldots, 1)$.

To show that $K_{\mathrm{m}+}$ is pointed, we note that if $x \in K_{\mathrm{m}+}$, then $-x \in K_{\mathrm{m}+}$ only if $x = 0$. This implies that the cone does not contain an entire line.

(b) Using the hint, we see that $y^T x \geq 0$ for all $x \in K_{\mathrm{m}+}$ if and only if

$$y_1 \geq 0, \qquad y_1 + y_2 \geq 0, \quad \ldots, y_1 + y_2 + \cdots + y_n \geq 0.$$

Therefore

$$K_{\mathrm{m}+}^* = \{y \mid \sum_{i=1}^k y_i \geq 0,\ k = 1, \ldots, n\}.$$

2.34 *The lexicographic cone and ordering.* The *lexicographic cone* is defined as

$$K_{\mathrm{lex}} = \{0\} \cup \{x \in \mathbf{R}^n \mid x_1 = \cdots = x_k = 0,\ x_{k+1} > 0,\ \text{for some } k,\ 0 \leq k < n\},$$

i.e., all vectors whose first nonzero coefficient (if any) is positive.

(a) Verify that K_{lex} is a cone, but *not* a proper cone.

(b) We define the *lexicographic ordering* on $\mathbf{R}^n$ as follows: $x \leq_{\mathrm{lex}} y$ if and only if $y - x \in K_{\mathrm{lex}}$. (Since K_{lex} is not a proper cone, the lexicographic ordering is not a generalized inequality.) Show that the lexicographic ordering is a *linear ordering*: for any $x,\ y \in \mathbf{R}^n$, either $x \leq_{\mathrm{lex}} y$ or $y \leq_{\mathrm{lex}} x$. Therefore any set of vectors can be sorted with respect to the lexicographic cone, which yields the familiar sorting used in dictionaries.

(c) Find K_{lex}^*.

Solution.

(a) K_{lex} is not closed. For example, $(\epsilon, -1, 0, \ldots, 0) \in K_{\mathrm{lex}}$ for all $\epsilon > 0$, but not for $\epsilon = 0$.

(b) If $x \neq y$ then $x \leq_{\mathrm{lex}} y$ and $y \leq_{\mathrm{lex}} x$. If not, let $k = \min\{i \in \{1, \ldots, n\} \mid x_i \neq y_i\}$, be the index of the first component in which x and y differ. If $x_k < y_k$, we have $x \leq_{\mathrm{lex}} y$. If $x_k > y_k$, we have $x \geq_{\mathrm{lex}} y$.

(c) $K_{\mathrm{lex}}^* = \mathbf{R}_+ e_1 = \{(t, 0, \ldots, 0) \mid t \geq 0\}$. To prove this, first note that if $y = (t, 0, \ldots, 0)$ with $t \geq 0$, then obviously $y^T x = t x_1 \geq 0$ for all $x \in K_{\mathrm{lex}}$.

Conversely, suppose $y^T x \geq 0$ for all $x \in K_{\mathrm{lex}}$. In particular $y^T e_1 \geq 0$, so $y_1 \geq 0$. Furthermore, by considering $x = (\epsilon, -1, 0, \ldots, 0)$, we have $\epsilon y_1 - y_2 \geq 0$ for all $\epsilon > 0$, which is only possible if $y_2 = 0$. Similarly, one can prove that $y_3 = \cdots = y_n = 0$.

Exercises

2.35 *Copositive matrices.* A matrix $X \in \mathbf{S}^n$ is called *copositive* if $z^T X z \geq 0$ for all $z \succeq 0$. Verify that the set of copositive matrices is a proper cone. Find its dual cone.

Solution. We denote by K the set of copositive matrices in $\mathbf{S}^n$. K is a closed convex cone because it is the intersection of (infinitely many) halfspaces defined by homogeneous inequalities

$$z^T X z = \sum_{i,j} z_i z_j X_{ij} \geq 0.$$

K has nonempty interior, because it includes the cone of positive semidefinite matrices, which has nonempty interior. K is pointed because $X \in K$, $-X \in K$ means $z^T X z = 0$ for all $z \succeq 0$, hence $X = 0$.

By definition, the dual cone of a cone K is the set of normal vectors of all homogeneous halfspaces containing K (plus the origin). Therefore,

$$K^* = \mathbf{conv}\{z z^T \mid z \succeq 0\}.$$

2.36 *Euclidean distance matrices.* Let $x_1, \ldots, x_n \in \mathbf{R}^k$. The matrix $D \in \mathbf{S}^n$ defined by $D_{ij} = \|x_i - x_j\|_2^2$ is called a *Euclidean distance matrix*. It satisfies some obvious properties such as $D_{ij} = D_{ji}$, $D_{ii} = 0$, $D_{ij} \geq 0$, and (from the triangle inequality) $D_{ik}^{1/2} \leq D_{ij}^{1/2} + D_{jk}^{1/2}$. We now pose the question: When is a matrix $D \in \mathbf{S}^n$ a Euclidean distance matrix (for some points in $\mathbf{R}^k$, for some k)? A famous result answers this question: $D \in \mathbf{S}^n$ is a Euclidean distance matrix if and only if $D_{ii} = 0$ and $x^T D x \leq 0$ for all x with $\mathbf{1}^T x = 0$. (See §8.3.3.)

Show that the set of Euclidean distance matrices is a convex cone. Find the dual cone.

Solution. The set of Euclidean distance matrices in $\mathbf{S}^n$ is a closed convex cone because it is the intersection of (infinitely many) halfspaces defined by the following homogeneous inequalities:

$$e_i^T D e_i \leq 0, \qquad e_i^T D e_i \geq 0, \qquad x^T D x = \sum_{j,k} x_j x_k D_{jk} \leq 0,$$

for all $i = 1, \ldots, n$, and all x with $\mathbf{1}^T x = 1$.

It follows that dual cone is given by

$$K^* = \mathbf{conv}(\{-x x^T \mid \mathbf{1}^T x = 1\} \bigcup \{e_1 e_1^T, -e_1 e_1^T, \ldots, e_n e_n^T, -e_n e_n^T\}).$$

This can be made more explicit as follows. Define $V \in \mathbf{R}^{n \times (n-1)}$ as

$$V_{ij} = \begin{cases} 1 - 1/n & i = j \\ -1/n & i \neq j. \end{cases}$$

The columns of V form a basis for the set of vectors orthogonal to $\mathbf{1}$, *i.e.*, a vector x satisfies $\mathbf{1}^T x = 0$ if and only if $x = Vy$ for some y. The dual cone is

$$K^* = \{V W V^T + \mathbf{diag}(u) \mid W \preceq 0, u \in \mathbf{R}^n\}.$$

2.37 *Nonnegative polynomials and Hankel LMIs.* Let K_{pol} be the set of (coefficients of) nonnegative polynomials of degree $2k$ on $\mathbf{R}$:

$$K_{\text{pol}} = \{x \in \mathbf{R}^{2k+1} \mid x_1 + x_2 t + x_3 t^2 + \cdots + x_{2k+1} t^{2k} \geq 0 \text{ for all } t \in \mathbf{R}\}.$$

(a) Show that K_{pol} is a proper cone.

(b) A basic result states that a polynomial of degree $2k$ is nonnegative on $\mathbf{R}$ if and only if it can be expressed as the sum of squares of two polynomials of degree k or less. In other words, $x \in K_{\rm pol}$ if and only if the polynomial

$$p(t) = x_1 + x_2 t + x_3 t^2 + \cdots + x_{2k+1} t^{2k}$$

can be expressed as

$$p(t) = r(t)^2 + s(t)^2,$$

where r and s are polynomials of degree k.
Use this result to show that

$$K_{\rm pol} = \left\{ x \in \mathbf{R}^{2k+1} \;\middle|\; x_i = \sum_{m+n=i+1} Y_{mn} \text{ for some } Y \in \mathbf{S}_+^{k+1} \right\}.$$

In other words, $p(t) = x_1 + x_2 t + x_3 t^2 + \cdots + x_{2k+1} t^{2k}$ is nonnegative if and only if there exists a matrix $Y \in \mathbf{S}_+^{k+1}$ such that

$$\begin{aligned} x_1 &= Y_{11} \\ x_2 &= Y_{12} + Y_{21} \\ x_3 &= Y_{13} + Y_{22} + Y_{31} \\ &\vdots \\ x_{2k+1} &= Y_{k+1,k+1}. \end{aligned}$$

(c) Show that $K_{\rm pol}^* = K_{\rm han}$ where

$$K_{\rm han} = \{ z \in \mathbf{R}^{2k+1} \mid H(z) \succeq 0 \}$$

and

$$H(z) = \begin{bmatrix} z_1 & z_2 & z_3 & \cdots & z_k & z_{k+1} \\ z_2 & z_3 & z_4 & \cdots & z_{k+1} & z_{k+2} \\ z_3 & z_4 & z_5 & \cdots & z_{k+2} & z_{k+4} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ z_k & z_{k+1} & z_{k+2} & \cdots & z_{2k-1} & z_{2k} \\ z_{k+1} & z_{k+2} & z_{k+3} & \cdots & z_{2k} & z_{2k+1} \end{bmatrix}.$$

(This is the *Hankel matrix* with coefficients $z_1, \ldots, z_{2k+1}$.)

(d) Let $K_{\rm mom}$ be the conic hull of the set of all vectors of the form $(1, t, t^2, \ldots, t^{2k})$, where $t \in \mathbf{R}$. Show that $y \in K_{\rm mom}$ if and only if $y_1 \geq 0$ and

$$y = y_1 (1, \mathbf{E}\, u, \mathbf{E}\, u^2, \ldots, \mathbf{E}\, u^{2k})$$

for some random variable u. In other words, the elements of $K_{\rm mom}$ are nonnegative multiples of the moment vectors of all possible distributions on $\mathbf{R}$. Show that $K_{\rm pol} = K_{\rm mom}^*$.

(e) Combining the results of (c) and (d), conclude that $K_{\rm han} = \mathbf{cl}\, K_{\rm mom}$.
As an example illustrating the relation between $K_{\rm mom}$ and $K_{\rm han}$, take $k = 2$ and $z = (1, 0, 0, 0, 1)$. Show that $z \in K_{\rm han}$, $z \notin K_{\rm mom}$. Find an explicit sequence of points in $K_{\rm mom}$ which converge to z.

Solution.

(a) It is a closed convex cone, because it is the intersection of (infinitely many) closed halfspaces, and also obviously a cone.
It has nonempty interior because $(1, 0, 1, 0, \ldots, 0, 1) \in \mathbf{int}\, K_{\rm pol}$ (*i.e.*, the polynomial $1 + t^2 + t^4 + \cdots + t^{2k}$). It is pointed because $p(t) \geq 0$ and $-p(t) \geq 0$ imply $p(t) = 0$.

(b) First assume that $x_i = \sum_{m+n=i+1} Y_{mn}$ for some $Y \succeq 0$. It easily verified that, for all $t \in \mathbf{R}$,

$$\begin{aligned} p(t) = x_1 + x_2 t + \cdots + x_{2k+1} t^{2k} &= \sum_{i=1}^{2k+1} \sum_{m+n=i+1} Y_{mn} t^{i-1} \\ &= \sum_{m,n=1}^{k+1} Y_{mn} t^{m+n-2} \\ &= \sum_{m,n=1}^{k+1} Y_{mn} t^{m-1} t^{n-1} \\ &= v^T Y v \end{aligned}$$

where $v = (1, t, t^2, \ldots, t^k)$. Therefore $p(t) \geq 0$.
Conversely, assume $x \in K_{\rm pol}$. By the theorem, we can express the corresponding polynomial $p(t)$ as $p(t) = r(t)^2 + s(t)^2$, where

$$r(t) = a_1 + a_2 t + \cdots + a_{k+1} t^k, \qquad s(t) = b_1 + b_2 t + \cdots + b_{k+1} t^k,$$

The coefficient of t^{i-1} in $r(t)^2 + s(t)^2$ is $\sum_{m+n=i+1} (a_m a_n + b_m b_n)$. Therefore,

$$x_i = \sum_{m+n=i+1} (a_m a_n + b_m b_n) = \sum_{m+n=i+1} Y_{mn}$$

for $Y = aa^T + bb^T$.

(c) $z \in K_{\rm pol}^*$ if and only if $x^T z \geq 0$ for all $x \in K_{\rm pol}$. Using the previous result, this is equivalent to the condition that for all $Y \succeq 0$,

$$\sum_{i=1}^{2k+1} z_i \sum_{m+n=i+1} Y_{mn} = \sum_{m,n=1}^{k+1} Y_{mn} z_{m+n-1} = \mathbf{tr}(YH(z)) \geq 0,$$

i.e., $H(z) \succeq 0$.

(d) The conic hull of the vectors of the form $(1, t, \ldots, t^{2k})$ is the set of nonnegative multiples of all convex combinations of vectors of the form $(1, t, \ldots, t^{2k})$, *i.e.*, nonnegative multiples of vectors of the form

$$\mathbf{E}(1, t, t^2, \ldots, t^{2k}).$$

$x^T z \geq 0$ for all $z \in K_{\rm mom}$ if and only if

$$\mathbf{E}(x_1 + x_2 t + x_3 t^2 + \cdots + x_{2k+1} t^{2k}) \geq 0$$

for all distributions on $\mathbf{R}$. This is true if and only if

$$x_1 + x_2 t + x_3 t^2 + \cdots + x_{2k+1} t^{2k} \geq 0$$

for all t.

(e) This follows from the last result in §2.6.1, and the fact that we have shown that $K_{\rm han} = K_{\rm pol}^* = K_{\rm mom}^{**}$.
For the example, note that $\mathbf{E}\, t^2 = 0$ means that the distribution concentrates probability one at $t = 0$. But then we cannot have $\mathbf{E}\, t^4 = 1$. The associated Hankel matrix is $H = \mathbf{diag}(1, 0, 1)$, which is clearly positive semidefinite.
Let's put probability p_k at $t = 0$, and $(1 - p_k)/2$ at each of the points $t = \pm k$. Then we have, for all k, $\mathbf{E}\, t = \mathbf{E}\, t^3 = 0$. We also have $\mathbf{E}\, t^2 = (1 - p_k)k^2$ and $\mathbf{E}\, t^4 = (1 - p_k)k^4$. Let's now choose $p_k = 1 - 1/k^4$, so we have $\mathbf{E}\, t^4 = 1$, and $\mathbf{E}\, t^2 = 1/k^2$. Thus, the moments of this sequence of measures converge to 1, 0, 0, 1.

2.38 [Roc70, pages 15, 61] *Convex cones constructed from sets.*

(a) The *barrier cone* of a set C is defined as the set of all vectors y such that $y^T x$ is bounded above over $x \in C$. In other words, a nonzero vector y is in the barrier cone if and only if it is the normal vector of a halfspace $\{x \mid y^T x \leq \alpha\}$ that contains C. Verify that the barrier cone is a convex cone (with no assumptions on C).

Solution. Take two points x_1, x_2 in the barrier cone. We have

$$\sup_{y \in C} x_1^T y < \infty, \qquad \sup_{y \in C} x_2^T y < \infty,$$

so for all $\theta_1, \theta_2 \geq 0$,

$$\sup_{y \in C} (\theta_1 x_1 + \theta_2 x_2)^T y \leq \sup_{y \in C} (\theta_1 x_1^T y) + \sup_{y \in C} (\theta_2 x_2^T y) < \infty.$$

Therefore $\theta x_1 + \theta_2 x_2$ is also in the barrier cone.

(b) The *recession cone* (also called *asymptotic cone*) of a set C is defined as the set of all vectors y such that for each $x \in C$, $x - ty \in C$ for all $t \geq 0$. Show that the recession cone of a convex set is a convex cone. Show that if C is nonempty, closed, and convex, then the recession cone of C is the dual of the barrier cone.

Solution. It is clear that the recession cone is a cone. We show that it is convex if C is convex.

Let y_1, y_2 be in the recession cone, and suppose $0 \leq \theta \leq 1$. Then if $x \in C$

$$x - t(\theta y_1 + (1-\theta) y_2) = \theta(x - t y_1) + (1-\theta)(x - t y_2) \in C,$$

for all $t \geq 0$, because C is convex and $x - t y_1 \in C$, $x - t y_2 \in C$ for all $t \geq 0$. Therefore $\theta y_1 + (1-\theta) y_2$ is in the recession cone.

Before establishing the second claim, we note that if C is closed and convex, then its recession cone R_C can be defined by choosing any arbitrary point $\hat{x} \in C$, and letting

$$R_C = \{y \mid \hat{x} - ty \in C \ \forall t \geq 0\}.$$

This follows from the following observation. For $x \in C$, define

$$R_C(x) = \{y \mid x - ty \in C \ \forall t \geq 0\}.$$

We want to show that $R_C(x_1) = R_C(x_2)$ for any $x_1, x_2 \in C$. We first show $R_C(x_1) \subseteq R_C(x_2)$. If $y \in R_C(x_1)$, then $x_1 - (t/\theta) y \in C$ for all $t \geq 0$, $0 < \theta < 1$, so by convexity of C,

$$\theta(x_1 - (t/\theta) y) + (1-\theta) x_2 \in C.$$

Since C is closed,

$$x_2 - ty = \lim_{\theta \searrow 0} (\theta(x_1 - (t/\theta) y) + (1-\theta) x_2) \in C.$$

This holds for any $t \geq 0$, *i.e.*, $y \in R_C(x_2)$. The reverse inclusion $R_C(x_2) \subseteq R_C(x_1)$ follows similarly.

We now show that the recession cone is the dual of the barrier cone. Let $S_C(y) = \sup_{x \in C} y^T x$. By definition of the barrier cone, $S_C(y)$ is finite if and only if y is in the barrier cone, and every halfspace that contains C can be expressed as

$$y^T x \leq S_C(y)$$

for some nonzero y in the barrier cone. A closed convex set C is the intersection of all halfspaces that contain it. Therefore

$$C = \{x \mid y^T x \leq S_C(y) \text{ for all } y \in B_C\},$$

Let $\hat{x} \in C$. A vector v is in the recession cone if and only if $\hat{x} - tv \in C$ for all $t \geq 0$, *i.e.*,

$$y^T(\hat{x} - tv) \leq S_C(y) \text{ for all } y \in B_C.$$

This is true if and only if $y^T v \geq 0$ for all $y \in B_C$, *i.e.*, if and only if v is in the dual cone of B_C.

(c) The *normal cone* of a set C at a boundary point x_0 is the set of all vectors y such that $y^T(x - x_0) \leq 0$ for all $x \in C$ (*i.e.*, the set of vectors that define a supporting hyperplane to C at x_0). Show that the normal cone is a convex cone (with no assumptions on C). Give a simple description of the normal cone of a polyhedron $\{x \mid Ax \preceq b\}$ at a point in its boundary.

Solution. The normal cone is defined by a set of homogeneous linear inequalities in y, so it is a closed convex cone.

Let x_0 be a boundary point of $\{x \mid Ax \preceq b\}$. Suppose A and b are partitioned as

$$A = \begin{bmatrix} A_1^T \\ A_2^T \end{bmatrix}, \qquad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

in such a way that

$$A_1 x_0 = b_1, \qquad A_2 x_0 \prec b_2.$$

Then the normal at x_0 is

$$\{A_1^T \lambda \mid \lambda \succeq 0\},$$

i.e., it is the conic hull of the normal vectors of the constraints that are active at x_0.

2.39 *Separation of cones.* Let K and $\tilde{K}$ be two convex cones whose interiors are nonempty and disjoint. Show that there is a nonzero y such that $y \in K^*$, $-y \in \tilde{K}^*$.

Solution. Let $y \neq 0$ be the normal vector of a separating hyperplane separating the interiors: $y^T x \geq \alpha$ for $x \in \mathbf{int}\, K_1$ and $y^T x \leq \alpha$ for $x \in \mathbf{int}\, K_2$. We must have $\alpha = 0$ because K_1 and K_2 are cones, so if $x \in \mathbf{int}\, K_1$, then $tx \in \mathbf{int}\, K_1$ for all $t > 0$.

This means that

$$y \in (\mathbf{int}\, K_1)^* = K_1^*, \qquad -y \in (\mathbf{int}\, K_2)^* = K_2^*.$$

Chapter 3

Convex functions

Exercises

Definition of convexity

3.1 Suppose $f : \mathbf{R} \to \mathbf{R}$ is convex, and $a,\ b \in \mathbf{dom}\, f$ with $a < b$.

(a) Show that

$$f(x) \leq \frac{b-x}{b-a} f(a) + \frac{x-a}{b-a} f(b)$$

for all $x \in [a, b]$.

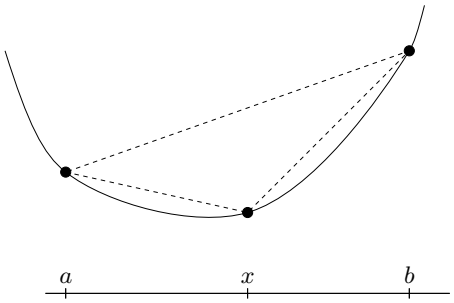
Solution. This is Jensen's inequality with $\lambda = (b-x)/(b-a)$.

(b) Show that

$$\frac{f(x) - f(a)}{x-a} \leq \frac{f(b) - f(a)}{b-a} \leq \frac{f(b) - f(x)}{b-x}$$

for all $x \in (a, b)$. Draw a sketch that illustrates this inequality.

Solution. We obtain the first inequality by subtracting $f(a)$ from both sides of the inequality in (a). The second inequality follows from subtracting $f(b)$. Geometrically, the inequalities mean that the slope of the line segment between $(a, f(a))$ and $(b, f(b))$ is larger than the slope of the segment between $(a, f(a))$ and $(x, f(x))$, and smaller than the slope of the segment between $(x, f(x))$ and $(b, f(b))$.



(c) Suppose f is differentiable. Use the result in (b) to show that

$$f'(a) \leq \frac{f(b) - f(a)}{b-a} \leq f'(b).$$

Note that these inequalities also follow from (3.2):

$$f(b) \geq f(a) + f'(a)(b-a), \qquad f(a) \geq f(b) + f'(b)(a-b).$$

Solution. This follows from (b) by taking the limit for $x \to a$ on both sides of the first inequality, and by taking the limit for $x \to b$ on both sides of the second inequality.

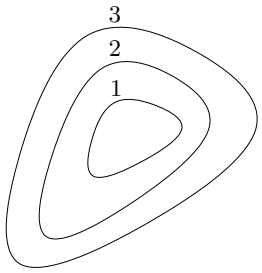
(d) Suppose f is twice differentiable. Use the result in (c) to show that $f''(a) \geq 0$ and $f''(b) \geq 0$.

Solution. From part (c),

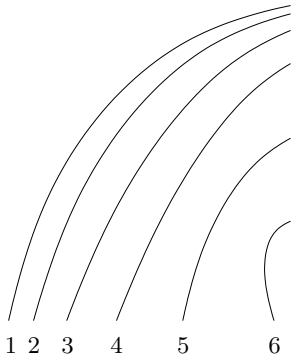
$$\frac{f'(b) - f'(a)}{b-a} \geq 0,$$

and taking the limit for $b \to a$ shows that $f''(a) \geq 0$.

3.2 *Level sets of convex, concave, quasiconvex, and quasiconcave functions.* Some level sets of a function f are shown below. The curve labeled 1 shows $\{x \mid f(x) = 1\}$, etc.

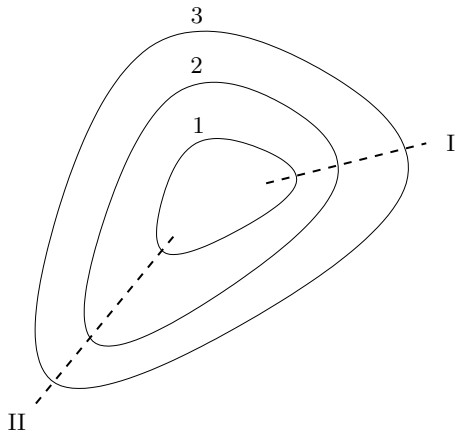


Could f be convex (concave, quasiconvex, quasiconcave)? Explain your answer. Repeat for the level curves shown below.

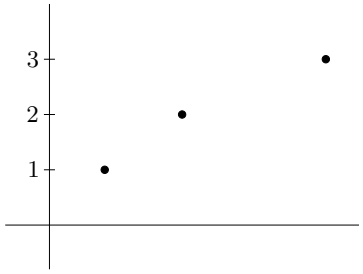


Solution. The first function could be quasiconvex because the sublevel sets appear to be convex. It is definitely not concave or quasiconcave because the superlevel sets are not convex.

It is also not convex, for the following reason. We plot the function values along the dashed line labeled I.



Along this line the function passes through the points marked as black dots in the figure below. Clearly along this line segment, the function is not convex.



If we repeat the same analysis for the second function, we see that it could be concave (and therefore it could be quasiconcave). It cannot be convex or quasiconvex, because the sublevel sets are not convex.

3.3 *Inverse of an increasing convex function.* Suppose $f : \mathbf{R} \to \mathbf{R}$ is increasing and convex on its domain (a, b). Let g denote its inverse, *i.e.*, the function with domain $(f(a), f(b))$ and $g(f(x)) = x$ for $a < x < b$. What can you say about convexity or concavity of g?

Solution. g is concave. Its hypograph is

$$\begin{aligned}
\mathbf{hypo}\, g &= \{(y,t) \mid t \le g(y)\} \\
&= \{(y,t) \mid f(t) \le f(g(y))\} \quad \text{(because } f \text{ is increasing)} \\
&= \{(y,t) \mid f(t) \le y)\} \\
&= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \mathbf{epi}\, f.
\end{aligned}$$

For differentiable g, f, we can also prove the result as follows. Differentiate $g(f(x)) = x$ once to get

$$g'(f(x)) = 1/f'(x).$$

so g is increasing. Differentiate again to get

$$g''(f(x)) = -\frac{f''(x)}{f'(x)^3},$$

so g is concave.

3.4 [RV73, page 15] Show that a continuous function $f : \mathbf{R}^n \to \mathbf{R}$ is convex if and only if for every line segment, its average value on the segment is less than or equal to the average of its values at the endpoints of the segment: For every $x,\ y \in \mathbf{R}^n$,

$$\int_0^1 f(x + \lambda(y - x))\, d\lambda \le \frac{f(x) + f(y)}{2}.$$

Solution. First suppose that f is convex. Jensen's inequality can be written as

$$f(x + \lambda(y - x)) \le f(x) + \lambda(f(y) - f(x))$$

for $0 \le \lambda \le 1$. Integrating both sides from 0 to 1 we get

$$\int_0^1 f(x + \lambda(y - x))\, d\lambda \le \int_0^1 \left(f(x) + \lambda(f(y) - f(x))\right)\, d\lambda = \frac{f(x) + f(y)}{2}.$$

Now we show the converse. Suppose f is not convex. Then there are x and y and $\theta_0 \in (0, 1)$ such that

$$f(\theta_0 x + (1 - \theta_0)y) > \theta_0 f(x) + (1 - \theta_0) f(y).$$

Consider the function of θ given by

$$F(\theta) = f(\theta x + (1-\theta)y) - \theta f(x) - (1-\theta)f(y),$$

which is continuous since f is. Note that F is zero for $\theta = 0$ and $\theta = 1$, and positive at θ_0. Let α be the largest zero crossing of F below θ_0 and let β be the smallest zero crossing of F above θ_0. Define $u = \alpha x + (1-\alpha)y$ and $v = \beta x + (1-\beta)y$. On the interval (α, β), we have

$$F(\theta) = f(\theta x + (1-\theta)y) > \theta f(x) + (1-\theta)f(y),$$

so for $\theta \in (0,1)$,

$$f(\theta u + (1-\theta)v) > \theta f(u) + (1-\theta)f(v).$$

Integrating this expression from $\theta = 0$ to $\theta = 1$ yields

$$\int_0^1 f(u + \theta(u-v))\, d\theta > \int_0^1 (f(u) + \theta(f(u) - f(v)))\, d\theta = \frac{f(u)+f(v)}{2}.$$

In other words, the average of f over the interval $[u, v]$ exceeds the average of its values at the endpoints. This proves the converse.

3.5 [RV73, page 22] *Running average of a convex function.* Suppose $f : \mathbf{R} \to \mathbf{R}$ is convex, with $\mathbf{R}_+ \subseteq \mathbf{dom}\, f$. Show that its *running average* F, defined as

$$F(x) = \frac{1}{x}\int_0^x f(t)\, dt, \qquad \mathbf{dom}\, F = \mathbf{R}_{++},$$

is convex. You can assume f is differentiable.

Solution. F is differentiable with

$$\begin{aligned} F'(x) &= -(1/x^2)\int_0^x f(t)\, dt + f(x)/x \\ F''(x) &= (2/x^3)\int_0^x f(t)\, dt - 2f(x)/x^2 + f'(x)/x \\ &= (2/x^3)\int_0^x (f(t) - f(x) - f'(x)(t-x))\, dt. \end{aligned}$$

Convexity now follows from the fact that

$$f(t) \geq f(x) + f'(x)(t-x)$$

for all $x, t \in \mathbf{dom}\, f$, which implies $F''(x) \geq 0$.

3.6 *Functions and epigraphs.* When is the epigraph of a function a halfspace? When is the epigraph of a function a convex cone? When is the epigraph of a function a polyhedron?

Solution. If the function is affine, positively homogeneous ($f(\alpha x) = \alpha f(x)$ for $\alpha \geq 0$), and piecewise-affine, respectively.

3.7 Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is convex with $\mathbf{dom}\, f = \mathbf{R}^n$, and bounded above on $\mathbf{R}^n$. Show that f is constant.

Solution. Suppose f is not constant, *i.e.*, there exist x, y with $f(x) < f(y)$. The function

$$g(t) = f(x + t(y-x))$$

is convex, with $g(0) < g(1)$. By Jensen's inequality

$$g(1) \leq \frac{t-1}{t}g(0) + \frac{1}{t}g(t)$$

for all $t > 1$, and therefore

$$g(t) \geq tg(1) - (t-1)g(0) = g(0) + t(g(1) - g(0)),$$

so g grows unboundedly as $t \to \infty$. This contradicts our assumption that f is bounded.

3.8 *Second-order condition for convexity.* Prove that a twice differentiable function f is convex if and only if its domain is convex and $\nabla^2 f(x) \succeq 0$ for all $x \in \mathbf{dom}\, f$. *Hint.* First consider the case $f : \mathbf{R} \to \mathbf{R}$. You can use the first-order condition for convexity (which was proved on page 70).

Solution. We first assume $n = 1$. Suppose $f : \mathbf{R} \to \mathbf{R}$ is convex. Let $x, y \in \mathbf{dom}\, f$ with $y > x$. By the first-order condition,

$$f'(x)(y - x) \leq f(y) - f(x) \leq f'(y)(y - x).$$

Subtracting the righthand side from the lefthand side and dividing by $(y - x)^2$ gives

$$\frac{f'(y) - f'(x)}{y - x} \geq 0.$$

Taking the limit for $y \to x$ yields $f''(x) \geq 0$.

Conversely, suppose $f''(z) \geq 0$ for all $z \in \mathbf{dom}\, f$. Consider two arbitrary points $x, y \in \mathbf{dom}\, f$ with $x < y$. We have

$$\begin{aligned} 0 &\leq \int_x^y f''(z)(y - z)\, dz \\ &= (f'(z)(y - z))\big|_{z=x}^{z=y} + \int_x^y f'(z)\, dz \\ &= -f'(x)(y - x) + f(y) - f(x), \end{aligned}$$

i.e., $f(y) \geq f(x) + f'(x)(y - x)$. This shows that f is convex.

To generalize to $n > 1$, we note that a function is convex if and only if it is convex on all lines, *i.e.*, the function $g(t) = f(x_0 + tv)$ is convex in t for all $x_0 \in \mathbf{dom}\, f$ and all v. Therefore f is convex if and only if

$$g''(t) = v^T \nabla^2 f(x_0 + tv) v \geq 0$$

for all $x_0 \in \mathbf{dom}\, f$, $v \in \mathbf{R}^n$, and t satisfying $x_0 + tv \in \mathbf{dom}\, f$. In other words it is necessary and sufficient that $\nabla^2 f(x) \succeq 0$ for all $x \in \mathbf{dom}\, f$.

3.9 *Second-order conditions for convexity on an affine set.* Let $F \in \mathbf{R}^{n \times m}$, $\hat{x} \in \mathbf{R}^n$. The *restriction* of $f : \mathbf{R}^n \to \mathbf{R}$ to the affine set $\{Fz + \hat{x} \mid z \in \mathbf{R}^m\}$ is defined as the function $\tilde{f} : \mathbf{R}^m \to \mathbf{R}$ with

$$\tilde{f}(z) = f(Fz + \hat{x}), \qquad \mathbf{dom}\, \tilde{f} = \{z \mid Fz + \hat{x} \in \mathbf{dom}\, f\}.$$

Suppose f is twice differentiable with a convex domain.

(a) Show that $\tilde{f}$ is convex if and only if for all $z \in \mathbf{dom}\, \tilde{f}$

$$F^T \nabla^2 f(Fz + \hat{x}) F \succeq 0.$$

(b) Suppose $A \in \mathbf{R}^{p \times n}$ is a matrix whose nullspace is equal to the range of F, *i.e.*, $AF = 0$ and $\mathbf{rank}\, A = n - \mathbf{rank}\, F$. Show that $\tilde{f}$ is convex if and only if for all $z \in \mathbf{dom}\, \tilde{f}$ there exists a $\lambda \in \mathbf{R}$ such that

$$\nabla^2 f(Fz + \hat{x}) + \lambda A^T A \succeq 0.$$

Hint. Use the following result: If $B \in \mathbf{S}^n$ and $A \in \mathbf{R}^{p \times n}$, then $x^T B x \geq 0$ for all $x \in \mathcal{N}(A)$ if and only if there exists a λ such that $B + \lambda A^T A \succeq 0$.

Solution.

(a) The Hessian of $\tilde{f}$ must be positive semidefinite everywhere:

$$\nabla^2 \tilde{f}(z) = F^T \nabla^2 f(Fz + \hat{x}) F \succeq 0.$$

(b) The condition in (a) means that $v^T \nabla^2 f(Fz + \hat{x}) v \geq 0$ for all v with $Av = 0$, *i.e.*,

$$v^T A^T A v = 0 \Longrightarrow v^T \nabla^2 f(Fz + \hat{x}) v \geq 0.$$

The result immediately follows from the hint.

3.10 *An extension of Jensen's inequality.* One interpretation of Jensen's inequality is that randomization or dithering hurts, *i.e.*, raises the average value of a convex function: For f convex and v a zero mean random variable, we have $\mathbf{E}\, f(x_0 + v) \geq f(x_0)$. This leads to the following conjecture. If f_0 is convex, then the larger the variance of v, the larger $\mathbf{E}\, f(x_0 + v)$.

(a) Give a counterexample that shows that this conjecture is false. Find zero mean random variables v and w, with $\mathbf{var}(v) > \mathbf{var}(w)$, a convex function f, and a point x_0, such that $\mathbf{E}\, f(x_0 + v) < \mathbf{E}\, f(x_0 + w)$.

(b) The conjecture is true when v and w are scaled versions of each other. Show that $\mathbf{E}\, f(x_0 + tv)$ is monotone increasing in $t \geq 0$, when f is convex and v is zero mean.

Solution.

(a) Define $f : \mathbf{R} \to \mathbf{R}$ as

$$f(x) = \begin{cases} 0, & x \leq 0 \\ x, & x > 0, \end{cases}$$

$x_0 = 0$, and scalar random variables

$$w = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases} \qquad v = \begin{cases} 4 & \text{with probability } 1/10 \\ -4/9 & \text{with probability } 9/10. \end{cases}$$

w and v are zero-mean and

$$\mathbf{var}(v) = 16/9 > 1 = \mathbf{var}(w).$$

However,

$$\mathbf{E}\, f(v) = 2/5 < 1/2 = \mathbf{E}\, f(w).$$

(b) $f(x_0 + tv)$ is convex in t for fixed v, hence if v is a random variable, $g(t) = \mathbf{E}\, f(x_0 + tv)$ is a convex function of t. From Jensen's inequality,

$$g(t) = \mathbf{E}\, f(x_0 + tv) \geq f(x_0) = g(0).$$

Now consider two points a, b, with $0 < a < b$. If $g(b) < g(a)$, then

$$\frac{b-a}{b} g(0) + \frac{a}{b} g(b) < \frac{b-a}{b} g(a) + \frac{a}{b} g(a) = g(a)$$

which contradicts Jensen's inequality. Therefore we must have $g(b) \geq g(a)$.

3.11 *Monotone mappings.* A function $\psi : \mathbf{R}^n \to \mathbf{R}^n$ is called *monotone* if for all $x,\ y \in \mathbf{dom}\, \psi$,

$$(\psi(x) - \psi(y))^T (x - y) \geq 0.$$

(Note that 'monotone' as defined here is not the same as the definition given in §3.6.1. Both definitions are widely used.) Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is a differentiable convex function. Show that its gradient ∇f is monotone. Is the converse true, *i.e.*, is every monotone mapping the gradient of a convex function?

Solution. Convexity of f implies

$$f(x) \geq f(y) + \nabla f(y)^T(x-y), \qquad f(y) \geq f(x) + \nabla f(x)^T(y-x)$$

for arbitrary $x, y \in \mathbf{dom}\, f$. Combining the two inequalities gives

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq 0,$$

which shows that ∇f is monotone.

The converse not true in general. As a counterexample, consider

$$\psi(x) = \begin{bmatrix} x_1 \\ x_1/2 + x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1/2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

ψ is monotone because

$$(x-y)^T \begin{bmatrix} 1 & 0 \\ 1/2 & 1 \end{bmatrix} (x-y) = (x-y)^T \begin{bmatrix} 1 & 1/4 \\ 1/4 & 1 \end{bmatrix} (x-y) \geq 0$$

for all x, y.

However, there does not exist a function $f : \mathbf{R}^2 \to \mathbf{R}$ such that $\psi(x) = \nabla f(x)$, because such a function would have to satisfy

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial \psi_1}{\partial x_2} = 0, \qquad \frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial \psi_2}{\partial x_1} = 1/2.$$

3.12 Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is convex, $g : \mathbf{R}^n \to \mathbf{R}$ is concave, $\mathbf{dom}\, f = \mathbf{dom}\, g = \mathbf{R}^n$, and for all x, $g(x) \leq f(x)$. Show that there exists an affine function h such that for all x, $g(x) \leq h(x) \leq f(x)$. In other words, if a concave function g is an underestimator of a convex function f, then we can fit an affine function between f and g.

Solution. We first note that $\mathbf{int}\, \mathbf{epi}\, f$ is nonempty (since $\mathbf{dom}\, f = \mathbf{R}^n$), and does not intersect $\mathbf{hypo}\, g$ (since $f(x) < t$ for $(x,t) \in \mathbf{int}\, \mathbf{epi}\, f$ and $t \geq g(x)$ for $(x,t) \in \mathbf{hypo}\, g$). The two sets can therefore be separated by a hyperplane, *i.e.*, there exist $a \in \mathbf{R}^n, b \in \mathbf{R}$, not both zero, and $c \in \mathbf{R}$ such that

$$a^T x + bt \geq c \geq a^T y + bv$$

if $t > f(x)$ and $v \leq g(y)$. We must have $b \neq 0$, since otherwise the condition would reduce to $a^T x \geq a^T y$ for all x and y, which is only possible if $a = 0$. Choosing $x = y$, and using the fact that $f(x) \geq g(x)$, we also see that $b > 0$.

Now we apply the separating hyperplane conditions to a point $(x,t) \in \mathbf{int}\, \mathbf{epi}\, f$, and $(y,v) = (x, g(x)) \in \mathbf{hypo}\, g$, and obtain

$$a^T x + bt \geq c \geq a^T x + bg(x),$$

and dividing by b,

$$t \geq (c - a^T x)/b \geq g(x),$$

for all $t > f(x)$. Therefore the affine function $h(x) = (c - a^T x)/b$ lies between f and g.

3.13 *Kullback-Leibler divergence and the information inequality.* Let D_{kl} be the Kullback-Leibler divergence, as defined in (3.17). Prove the *information inequality*: $D_{\mathrm{kl}}(u,v) \geq 0$ for all $u,\ v \in \mathbf{R}^n_{++}$. Also show that $D_{\mathrm{kl}}(u,v) = 0$ if and only if $u = v$.

Hint. The Kullback-Leibler divergence can be expressed as

$$D_{\mathrm{kl}}(u,v) = f(u) - f(v) - \nabla f(v)^T(u-v),$$

where $f(v) = \sum_{i=1}^n v_i \log v_i$ is the negative entropy of v.

Solution. The negative entropy is strictly convex and differentiable on $\mathbf{R}^n_{++}$, hence

$$f(u) > f(v) + \nabla f(v)^T (u - v)$$

for all $u, v \in \mathbf{R}^n_{++}$ with $u \neq v$. Evaluating both sides of the inequality, we obtain

$$\begin{aligned} \sum_{i=1}^n u_i \log u_i &> \sum_{i=1}^n v_i \log v_i + \sum_{i=1}^n (\log v_i + 1)(u_i - v_i) \\ &= \sum_{i=1}^n u_i \log v_i + \mathbf{1}^T (u - v). \end{aligned}$$

Re-arranging this inequality gives the desired result.

3.14 *Convex-concave functions and saddle-points.* We say the function $f : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ is *convex-concave* if $f(x, z)$ is a concave function of z, for each fixed x, and a convex function of x, for each fixed z. We also require its domain to have the product form $\mathbf{dom}\, f = A \times B$, where $A \subseteq \mathbf{R}^n$ and $B \subseteq \mathbf{R}^m$ are convex.

(a) Give a second-order condition for a twice differentiable function $f : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ to be convex-concave, in terms of its Hessian $\nabla^2 f(x, z)$.

(b) Suppose that $f : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ is convex-concave and differentiable, with $\nabla f(\tilde{x}, \tilde{z}) = 0$. Show that the *saddle-point property* holds: for all x, z, we have

$$f(\tilde{x}, z) \leq f(\tilde{x}, \tilde{z}) \leq f(x, \tilde{z}).$$

Show that this implies that f satisfies the *strong max-min property*:

$$\sup_z \inf_x f(x, z) = \inf_x \sup_z f(x, z)$$

(and their common value is $f(\tilde{x}, \tilde{z})$).

(c) Now suppose that $f : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ is differentiable, but not necessarily convex-concave, and the saddle-point property holds at $\tilde{x}, \tilde{z}$:

$$f(\tilde{x}, z) \leq f(\tilde{x}, \tilde{z}) \leq f(x, \tilde{z})$$

for all x, z. Show that $\nabla f(\tilde{x}, \tilde{z}) = 0$.

Solution.

(a) The condition follows directly from the second-order conditions for convexity and concavity: it is

$$\nabla^2_{xx} f(x, z) \succeq 0, \qquad \nabla^2_{zz} f(x, z) \preceq 0,$$

for all x, z. In terms of $\nabla^2 f$, this means that its $1, 1$ block is positive semidefinite, and its $2, 2$ block is negative semidefinite.

(b) Let us fix $\tilde{z}$. Since $\nabla_x f(\tilde{x}, \tilde{z}) = 0$ and $f(x, \tilde{z})$ is convex in x, we conclude that $\tilde{x}$ minimizes $f(x, \tilde{z})$ over x, *i.e.*, for all z, we have

$$f(\tilde{x}, \tilde{z}) \leq f(x, \tilde{z}).$$

This is one of the inequalities in the saddle-point condition. We can argue in the same way about $\tilde{z}$. Fix $\tilde{x}$, and note that $\nabla_z f(\tilde{x}, \tilde{z}) = 0$, together with concavity of this function in z, means that $\tilde{z}$ maximizes the function, *i.e.*, for any x we have

$$f(\tilde{x}, \tilde{z}) \geq f(\tilde{x}, z).$$

(c) To establish this we argue the same way. If the saddle-point condition holds, then $\tilde{x}$ minimizes $f(x, \tilde{z})$ over all x. Therefore we have $\nabla f_x(\tilde{x}, \tilde{z}) = 0$. Similarly, since $\tilde{z}$ maximizes $f(\tilde{x}, z)$ over all z, we have $\nabla f_z(\tilde{x}, \tilde{z}) = 0$.

Exercises

Examples

3.15 *A family of concave utility functions.* For $0 < \alpha \leq 1$ let

$$u_\alpha(x) = \frac{x^\alpha - 1}{\alpha},$$

with $\mathbf{dom}\, u_\alpha = \mathbf{R}_+$. We also define $u_0(x) = \log x$ (with $\mathbf{dom}\, u_0 = \mathbf{R}_{++}$).

(a) Show that for $x > 0$, $u_0(x) = \lim_{\alpha \to 0} u_\alpha(x)$.

(b) Show that u_α are concave, monotone increasing, and all satisfy $u_\alpha(1) = 0$.

These functions are often used in economics to model the benefit or utility of some quantity of goods or money. Concavity of u_α means that the marginal utility (*i.e.*, the increase in utility obtained for a fixed increase in the goods) decreases as the amount of goods increases. In other words, concavity models the effect of *satiation*.

Solution.

(a) In this limit, both the numerator and denominator go to zero, so we use l'Hopital's rule:

$$\lim_{\alpha \to 0} u_\alpha(x) = \lim_{\alpha \to 0} \frac{(d/d\alpha)(x^\alpha - 1)}{(d/d\alpha)\alpha} = \lim_{\alpha \to 0} \frac{x^\alpha \log x}{1} = \log x.$$

(b) By inspection we have

$$u_\alpha(1) = \frac{1^\alpha - 1}{\alpha} = 0.$$

The derivative is given by

$$u'_\alpha(x) = x^{\alpha - 1},$$

which is positive for all x (since $0 < \alpha < 1$), so these functions are increasing. To show concavity, we examine the second derivative:

$$u''_\alpha(x) = (\alpha - 1)x^{\alpha - 2}.$$

Since this is negative for all x, we conclude that u_α is strictly concave.

3.16 For each of the following functions determine whether it is convex, concave, quasiconvex, or quasiconcave.

(a) $f(x) = e^x - 1$ on $\mathbf{R}$.
Solution. Strictly convex, and therefore quasiconvex. Also quasiconcave but not concave.

(b) $f(x_1, x_2) = x_1 x_2$ on $\mathbf{R}^2_{++}$.
Solution. The Hessian of f is

$$\nabla^2 f(x) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

which is neither positive semidefinite nor negative semidefinite. Therefore, f is neither convex nor concave. It is quasiconcave, since its superlevel sets

$$\{(x_1, x_2) \in \mathbf{R}^2_{++} \mid x_1 x_2 \geq \alpha\}$$

are convex. It is not quasiconvex.

(c) $f(x_1, x_2) = 1/(x_1 x_2)$ on $\mathbf{R}^2_{++}$.
Solution. The Hessian of f is

$$\nabla^2 f(x) = \frac{1}{x_1 x_2} \begin{bmatrix} 2/(x_1^2) & 1/(x_1 x_2) \\ 1/(x_1 x_2) & 2/x_2^2 \end{bmatrix} \succeq 0$$

Therefore, f is convex and quasiconvex. It is not quasiconcave or concave.

(d) $f(x_1, x_2) = x_1/x_2$ on $\mathbf{R}^2_{++}$.
Solution. The Hessian of f is

$$\nabla^2 f(x) = \begin{bmatrix} 0 & -1/x_2^2 \\ -1/x_2^2 & 2x_1/x_2^3 \end{bmatrix}$$

which is not positive or negative semidefinite. Therefore, f is not convex or concave. It is quasiconvex and quasiconcave (*i.e.*, quasilinear), since the sublevel and superlevel sets are halfspaces.

(e) $f(x_1, x_2) = x_1^2/x_2$ on $\mathbf{R} \times \mathbf{R}_{++}$.
Solution. f is convex, as mentioned on page 72. (See also figure 3.3). This is easily verified by working out the Hessian:

$$\nabla^2 f(x) = \begin{bmatrix} 2/x_2 & -2x_1/x_2^2 \\ -2x_1/x_2^2 & 2x_1^2/x_2^3 \end{bmatrix} = (2/x_2) \begin{bmatrix} 1 \\ -2x_1/x_2 \end{bmatrix} \begin{bmatrix} 1 & -2x_1/x_2 \end{bmatrix} \succeq 0.$$

Therefore, f is convex and quasiconvex. It is not concave or quasiconcave (see the figure).

(f) $f(x_1, x_2) = x_1^\alpha \, x_2^{1-\alpha}$, where $0 \leq \alpha \leq 1$, on $\mathbf{R}^2_{++}$.
Solution. Concave and quasiconcave. The Hessian is

$$\begin{aligned} \nabla^2 f(x) &= \begin{bmatrix} \alpha(\alpha-1)x_1^{\alpha-2}x_2^{1-\alpha} & \alpha(1-\alpha)x_1^{\alpha-1}x_2^{-\alpha} \\ \alpha(1-\alpha)x_1^{\alpha-1}x_2^{-\alpha} & (1-\alpha)(-\alpha)x_1^\alpha x_2^{-\alpha-1} \end{bmatrix} \\ &= \alpha(1-\alpha)x_1^\alpha x_2^{1-\alpha} \begin{bmatrix} -1/x_1^2 & 1/x_1x_2 \\ 1/x_1x_2 & -1/x_2^2 \end{bmatrix} \\ &= -\alpha(1-\alpha)x_1^\alpha x_2^{1-\alpha} \begin{bmatrix} 1/x_1 \\ -1/x_2 \end{bmatrix} \begin{bmatrix} 1/x_1 \\ -1/x_2 \end{bmatrix}^T \\ &\preceq 0. \end{aligned}$$

f is not convex or quasiconvex.

3.17 Suppose $p < 1$, $p \neq 0$. Show that the function

$$f(x) = \left(\sum_{i=1}^n x_i^p \right)^{1/p}$$

with $\mathbf{dom}\, f = \mathbf{R}^n_{++}$ is concave. This includes as special cases $f(x) = (\sum_{i=1}^n x_i^{1/2})^2$ and the *harmonic mean* $f(x) = (\sum_{i=1}^n 1/x_i)^{-1}$. *Hint.* Adapt the proofs for the log-sum-exp function and the geometric mean in §3.1.5.
Solution. The first derivatives of f are given by

$$\frac{\partial f(x)}{\partial x_i} = (\sum_{i=1}^n x_i^p)^{(1-p)/p} x_i^{p-1} = \left(\frac{f(x)}{x_i} \right)^{1-p}.$$

The second derivatives are

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{1-p}{x_i} \left(\frac{f(x)}{x_i} \right)^{-p} \left(\frac{f(x)}{x_j} \right)^{1-p} = \frac{1-p}{f(x)} \left(\frac{f(x)^2}{x_i x_j} \right)^{1-p}$$

for $i \neq j$, and

$$\frac{\partial^2 f(x)}{\partial x_i^2} = \frac{1-p}{f(x)} \left(\frac{f(x)^2}{x_i^2} \right)^{1-p} - \frac{1-p}{x_i} \left(\frac{f(x)}{x_i} \right)^{1-p}.$$

We need to show that

$$y^T\nabla^2 f(x)y = \frac{1-p}{f(x)}\left(\left(\sum_{i=1}^n \frac{y_i f(x)^{1-p}}{x_i^{1-p}}\right)^2 - \sum_{i=1}^n \frac{y_i^2 f(x)^{2-p}}{x_i^{2-p}}\right) \leq 0$$

This follows by applying the Cauchy-Schwarz inequality $a^T b \leq \|a\|_2\|b\|_2$ with

$$a_i = \left(\frac{f(x)}{x_i}\right)^{-p/2}, \qquad b_i = y_i\left(\frac{f(x)}{x_i}\right)^{1-p/2},$$

and noting that $\sum_i a_i^2 = 1$.

3.18 Adapt the proof of concavity of the log-determinant function in §3.1.5 to show the following.

(a) $f(X) = \mathbf{tr}\left(X^{-1}\right)$ is convex on $\mathbf{dom}\, f = \mathbf{S}^n_{++}$.

(b) $f(X) = (\det X)^{1/n}$ is concave on $\mathbf{dom}\, f = \mathbf{S}^n_{++}$.

Solution.

(a) Define $g(t) = f(Z + tV)$, where $Z \succ 0$ and $V \in \mathbf{S}^n$.

$$\begin{aligned} g(t) &= \mathbf{tr}((Z+tV)^{-1}) \\ &= \mathbf{tr}\left(Z^{-1}(I + tZ^{-1/2}VZ^{-1/2})^{-1}\right) \\ &= \mathbf{tr}\left(Z^{-1}Q(I+t\Lambda)^{-1}Q^T\right) \\ &= \mathbf{tr}\left(Q^TZ^{-1}Q(I+t\Lambda)^{-1}\right) \\ &= \sum_{i=1}^n (Q^TZ^{-1}Q)_{ii}(1+t\lambda_i)^{-1}, \end{aligned}$$

where we used the eigenvalue decomposition $Z^{-1/2}VZ^{-1/2} = Q\Lambda Q^T$. In the last equality we express g as a positive weighted sum of convex functions $1/(1+t\lambda_i)$, hence it is convex.

(b) Define $g(t) = f(Z + tV)$, where $Z \succ 0$ and $V \in \mathbf{S}^n$.

$$\begin{aligned} g(t) &= (\det(Z+tV))^{1/n} \\ &= \left(\det Z^{1/2}\det(I + tZ^{-1/2}VZ^{-1/2})\det Z^{1/2}\right)^{1/n} \\ &= (\det Z)^{1/n}\left(\prod_{i=1}^n (1+t\lambda_i)\right)^{1/n} \end{aligned}$$

where λ_i, $i = 1, \ldots, n$, are the eigenvalues of $Z^{-1/2}VZ^{-1/2}$. From the last equality we see that g is a concave function of t on $\{t \mid Z + tV \succ 0\}$, since $\det Z > 0$ and the geometric mean $(\prod_{i=1}^n x_i)^{1/n}$ is concave on $\mathbf{R}^n_{++}$.

3.19 *Nonnegative weighted sums and integrals.*

(a) Show that $f(x) = \sum_{i=1}^r \alpha_i x_{[i]}$ is a convex function of x, where $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_r \geq 0$, and $x_{[i]}$ denotes the ith largest component of x. (You can use the fact that $f(x) = \sum_{i=1}^k x_{[i]}$ is convex on $\mathbf{R}^n$.)
Solution. We can express f as

$$\begin{aligned} f(x) = {} & \alpha_r(x_{[1]} + x_{[2]} + \cdots + x_{[r]}) + (\alpha_{r-1} - \alpha_r)(x_{[1]} + x_{[2]} + \cdots + x_{[r-1]}) \\ & + (\alpha_{r-2} - \alpha_{r-1})(x_{[1]} + x_{[2]} + \cdots + x_{[r-2]}) + \cdots + (\alpha_1 - \alpha_2)x_{[1]}, \end{aligned}$$

which is a nonnegative sum of the convex functions

$$x_{[1]}, \qquad x_{[1]} + x_{[2]}, \qquad x_{[1]} + x_{[2]} + x_{[3]}, \qquad \ldots, \qquad x_{[1]} + x_{[2]} + \cdots + x_{[r]}.$$

(b) Let $T(x, \omega)$ denote the trigonometric polynomial

$$T(x, \omega) = x_1 + x_2 \cos\omega + x_3 \cos 2\omega + \cdots + x_n \cos(n-1)\omega.$$

Show that the function

$$f(x) = -\int_0^{2\pi} \log T(x, \omega)\, d\omega$$

is convex on $\{x \in \mathbf{R}^n \mid T(x, \omega) > 0,\ 0 \le \omega \le 2\pi\}$.

Solution. The function

$$g(x, \omega) = -\log(x_1 + x_2 \cos\omega + x_3 \cos 2\omega + \cdots + + x_n \cos(n-1)\omega)$$

is convex in x for fixed ω. Therefore

$$f(x) = \int_0^{2\pi} g(x, \omega) d\omega$$

is convex in x.

3.20 *Composition with an affine function.* Show that the following functions $f : \mathbf{R}^n \to \mathbf{R}$ are convex.

(a) $f(x) = \|Ax - b\|$, where $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, and $\|\cdot\|$ is a norm on $\mathbf{R}^m$.

Solution. f is the composition of a norm, which is convex, and an affine function.

(b) $f(x) = -\left(\det(A_0 + x_1 A_1 + \cdots + x_n A_n)\right)^{1/m}$, on $\{x \mid A_0 + x_1 A_1 + \cdots + x_n A_n \succ 0\}$, where $A_i \in \mathbf{S}^m$.

Solution. f is the composition of the convex function $h(X) = -(\det X)^{1/m}$ and an affine transformation. To see that h is convex on $\mathbf{S}^m_{++}$, we restrict h to a line and prove that $g(t) = -\det(Z + tV)^{1/m}$ is convex:

$$\begin{aligned} g(t) &= -(\det(Z + tV))^{1/m} \\ &= -(\det Z)^{1/m} (\det(I + tZ^{-1/2} V Z^{-1/2}))^{1/m} \\ &= -(\det Z)^{1/m} (\prod_{i=1}^m (1 + t\lambda_i))^{1/m} \end{aligned}$$

where $\lambda_1, \ldots, \lambda_m$ denote the eigenvalues of $Z^{-1/2} V Z^{-1/2}$. We have expressed g as the product of a negative constant and the geometric mean of $1 + t\lambda_i$, $i = 1, \ldots, m$. Therefore g is convex. (See also exercise 3.18.)

(c) $f(X) = \mathbf{tr}\,(A_0 + x_1 A_1 + \cdots + x_n A_n)^{-1}$, on $\{x \mid A_0 + x_1 A_1 + \cdots + x_n A_n \succ 0\}$, where $A_i \in \mathbf{S}^m$. (Use the fact that $\mathbf{tr}(X^{-1})$ is convex on $\mathbf{S}^m_{++}$; see exercise 3.18.)

Solution. f is the composition of $\mathbf{tr}\, X^{-1}$ and an affine transformation

$$x \mapsto A_0 + x_1 A_1 + \cdots + x_n A_n.$$

3.21 *Pointwise maximum and supremum.* Show that the following functions $f : \mathbf{R}^n \to \mathbf{R}$ are convex.

(a) $f(x) = \max_{i=1,\ldots,k} \|A^{(i)}x - b^{(i)}\|$, where $A^{(i)} \in \mathbf{R}^{m\times n}$, $b^{(i)} \in \mathbf{R}^m$ and $\|\cdot\|$ is a norm on $\mathbf{R}^m$.

Solution. f is the pointwise maximum of k functions $\|A^{(i)}x - b^{(i)}\|$. Each of those functions is convex because it is the composition of an affine transformation and a norm.

(b) $f(x) = \sum_{i=1}^r |x|_{[i]}$ on $\mathbf{R}^n$, where $|x|$ denotes the vector with $|x|_i = |x_i|$ (*i.e.*, $|x|$ is the absolute value of x, componentwise), and $|x|_{[i]}$ is the ith largest component of $|x|$. In other words, $|x|_{[1]}$, $|x|_{[2]}$, $\ldots$, $|x|_{[n]}$ are the absolute values of the components of x, sorted in nonincreasing order.

Solution. Write f as

$$f(x) = \sum_{i=1}^{r} |x|_{[i]} = \max_{1\le i_1<i_2<\cdots<i_r\le n} |x_{i_1}| + \cdots + |x_{i_r}|$$

which is the pointwise maximum of $n!/(r!(n-r)!)$ convex functions.

3.22 *Composition rules.* Show that the following functions are convex.

(a) $f(x) = -\log(-\log(\sum_{i=1}^m e^{a_i^T x+b_i}))$ on $\mathbf{dom}\, f = \{x \mid \sum_{i=1}^m e^{a_i^T x+b_i} < 1\}$. You can use the fact that $\log(\sum_{i=1}^n e^{y_i})$ is convex.

Solution. $g(x) = \log(\sum_{i=1}^m e^{a_i^T x+b_i})$ is convex (composition of the log-sum-exp function and an affine mapping), so $-g$ is concave. The function $h(y) = -\log y$ is convex and decreasing. Therefore $f(x) = h(-g(x))$ is convex.

(b) $f(x,u,v) = -\sqrt{uv - x^Tx}$ on $\mathbf{dom}\, f = \{(x,u,v) \mid uv > x^Tx,\ u,\ v > 0\}$. Use the fact that x^Tx/u is convex in (x,u) for $u > 0$, and that $-\sqrt{x_1x_2}$ is convex on $\mathbf{R}^2_{++}$.

Solution. We can express f as $f(x,u,v) = -\sqrt{u(v - x^Tx/u)}$. The function $h(x_1,x_2) = -\sqrt{x_1x_2}$ is convex on $\mathbf{R}^2_{++}$, and decreasing in each argument. The functions $g_1(u,v,x) = u$ and $g_2(u,v,x) = v - x^Tx/u$ are concave. Therefore $f(u,v,x) = h(g(u,v,x))$ is convex.

(c) $f(x,u,v) = -\log(uv - x^Tx)$ on $\mathbf{dom}\, f = \{(x,u,v) \mid uv > x^Tx,\ u,\ v > 0\}$.

Solution. We can express f as

$$f(x,u,v) = -\log u - \log(v - x^Tx/u).$$

The first term is convex. The function $v - x^Tx/u$ is concave because v is linear and x^Tx/u is convex on $\{(x,u) \mid u > 0\}$. Therefore the second term in f is convex: it is the composition of a convex decreasing function $-\log t$ and a concave function.

(d) $f(x,t) = -(t^p - \|x\|_p^p)^{1/p}$ where $p > 1$ and $\mathbf{dom}\, f = \{(x,t) \mid t \ge \|x\|_p\}$. You can use the fact that $\|x\|_p^p/u^{p-1}$ is convex in (x,u) for $u > 0$ (see exercise 3.23), and that $-x^{1/p}y^{1-1/p}$ is convex on $\mathbf{R}^2_+$ (see exercise 3.16).

Solution. We can express f as

$$f(x,t) = -\left(t^{p-1}\left(t - \frac{\|x\|_p^p}{t^{p-1}}\right)\right)^{1/p} = -t^{1-1/p}\left(t - \frac{\|x\|_p^p}{t^{p-1}}\right)^{1/p}.$$

This is the composition of $h(y_1,y_2) = -y_1^{1/p}y_2^{1-1/p}$ (convex and decreasing in each argument) and two concave functions

$$g_1(x,t) = t^{1-1/p}, \qquad g_2(x,t) = t - \frac{\|x\|_p^p}{t^{p-1}}.$$

(e) $f(x,t) = -\log(t^p - \|x\|_p^p)$ where $p > 1$ and $\mathbf{dom}\, f = \{(x,t) \mid t > \|x\|_p\}$. You can use the fact that $\|x\|_p^p/u^{p-1}$ is convex in (x,u) for $u > 0$ (see exercise 3.23).

Solution. Express f as

$$\begin{aligned} f(x,t) &= -\log t^{p-1} - \log(t - \|x\|_p^p/t^{p-1}) \\ &= -(p-1)\log t - \log(t - \|x\|_p^p/t^{p-1}). \end{aligned}$$

The first term is convex. The second term is the composition of a decreasing convex function and a concave function, and is also convex.

3.23 *Perspective of a function.*

(a) Show that for $p > 1$,

$$f(x,t) = \frac{|x_1|^p + \cdots + |x_n|^p}{t^{p-1}} = \frac{\|x\|_p^p}{t^{p-1}}$$

is convex on $\{(x,t) \mid t > 0\}$.

Solution. This is the perspective function of $\|x\|_p^p = |x_1|^p + \cdots + |x_n|^p$.

(b) Show that

$$f(x) = \frac{\|Ax+b\|_2^2}{c^T x + d}$$

is convex on $\{x \mid c^T x + d > 0\}$, where $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$, $c \in \mathbf{R}^n$ and $d \in \mathbf{R}$.

Solution. This function is the composition of the function $g(y,t) = y^T y/t$ with an affine transformation $(y,t) = (Ax+b, c^T x + d)$. Therefore convexity of f follows from the fact that g is convex on $\{(y,t) \mid t > 0\}$.

For convexity of g one can note that it is the perspective of $x^T x$, or directly verify that the Hessian

$$\nabla^2 g(y,t) = \begin{bmatrix} I/t & -y/t^2 \\ -y^T/t & y^T y/t^3 \end{bmatrix}$$

is positive semidefinite, since

$$\begin{bmatrix} v \\ w \end{bmatrix}^T \begin{bmatrix} I/t & -y/t^2 \\ -y^T/t & y^T y/t^3 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \|tv - yw\|_2^2/t^3 \geq 0$$

for all v and w.

3.24 *Some functions on the probability simplex.* Let x be a real-valued random variable which takes values in $\{a_1, \ldots, a_n\}$ where $a_1 < a_2 < \cdots < a_n$, with $\mathbf{prob}(x = a_i) = p_i$, $i = 1, \ldots, n$. For each of the following functions of p (on the probability simplex $\{p \in \mathbf{R}_+^n \mid \mathbf{1}^T p = 1\}$), determine if the function is convex, concave, quasiconvex, or quasiconcave.

(a) $\mathbf{E}\, x$.

Solution. $\mathbf{E}\, x = p_1 a_1 + \cdots + p_n a_n$ is linear, hence convex, concave, quasiconvex, and quasiconcave

(b) $\mathbf{prob}(x \geq \alpha)$.

Solution. Let $j = \min\{i \mid a_i \geq \alpha\}$. Then $\mathbf{prob}(x \geq \alpha) = \sum_{i=j}^n p_i$, This is a linear function of p, hence convex, concave, quasiconvex, and quasiconcave.

(c) $\mathbf{prob}(\alpha \leq x \leq \beta)$.

Solution. Let $j = \min\{i \mid a_i \geq \alpha\}$ and $k = \max\{i \mid a_i \leq \beta\}$. Then $\mathbf{prob}(\alpha \leq x \leq \beta) = \sum_{i=j}^k p_i$. This is a linear function of p, hence convex, concave, quasiconvex, and quasiconcave.

(d) $\sum_{i=1}^n p_i \log p_i$, the negative entropy of the distribution.

Solution. $p \log p$ is a convex function on $\mathbf{R}_+$ (assuming $0 \log 0 = 0$), so $\sum_i p_i \log p_i$ is convex (and hence quasiconvex).

The function is not concave or quasiconcave. Consider, for example, $n = 2$, $p^1 = (1, 0)$ and $p^2 = (0, 1)$. Both p^1 and p^2 have function value zero, but the convex combination $(0.5, 0.5)$ has function value $\log(1/2) < 0$. This shows that the superlevel sets are not convex.

(e) $\mathbf{var}\, x = \mathbf{E}(x - \mathbf{E}\, x)^2$.

Solution. We have

$$\mathbf{var}\, x = \mathbf{E}\, x^2 - (\mathbf{E}\, x)^2 = \sum_{i=1}^n p_i a_i^2 - (\sum_{i=1}^n p_i a_i)^2,$$

so $\mathbf{var}\, x$ is a concave quadratic function of p.

The function is not convex or quasiconvex. Consider the example with $n = 2$, $a_1 = 0$, $a_2 = 1$. Both $(p_1, p_2) = (1/4, 3/4)$ and $(p_1, p_2) = (3/4, 1/4)$ lie in the probability simplex and have $\mathbf{var}\, x = 3/16$, but the convex combination $(p_1, p_2) = (1/2, 1/2)$ has a variance $\mathbf{var}\, x = 1/4 > 3/16$. This shows that the sublevel sets are not convex.

(f) $\mathbf{quartile}(x) = \inf\{\beta \mid \mathbf{prob}(x \leq \beta) \geq 0.25\}$.

Solution. The sublevel and the superlevel sets of $\mathbf{quartile}(x)$ are convex (see problem 2.15), so it is quasiconvex and quasiconcave.

$\mathbf{quartile}(x)$ is not continuous (it takes values in a discrete set $\{a_1, \ldots, a_n\}$, so it is not convex or concave. (A convex or a concave function is always continuous on the relative interior of its domain.)

(g) The cardinality of the smallest set $\mathcal{A} \subseteq \{a_1, \ldots, a_n\}$ with probability $\geq 90\%$. (By cardinality we mean the number of elements in $\mathcal{A}$.)

Solution. f is integer-valued, so it can not be convex or concave. (A convex or a concave function is always continuous on the relative interior of its domain.)

f is quasiconcave because its superlevel sets are convex. We have $f(p) \geq \alpha$ if and only if

$$\sum_{i=1}^k p_{[i]} < 0.9,$$

where $k = \max\{i = 1, \ldots, n \mid i < \alpha\}$ is the largest integer less than α, and $p_{[i]}$ is the ith largest component of p. We know that $\sum_{i=1}^k p_{[i]}$ is a convex function of p, so the inequality $\sum_{i=1}^k p_{[i]} < 0.9$ defines a convex set.

In general, $f(p)$ is not quasiconvex. For example, we can take $n = 2$, $a_1 = 0$ and $a_2 = 1$, and $p^1 = (0.1, 0.9)$ and $p^2 = (0.9, 0.1)$. Then $f(p^1) = f(p^2) = 1$, but $f((p^1 + p^2)/2) = f(0.5, 0.5) = 2$.

(h) The minimum width interval that contains 90% of the probability, *i.e.*,

$$\inf\{\beta - \alpha \mid \mathbf{prob}(\alpha \leq x \leq \beta) \geq 0.9\}.$$

Solution. The minimum width interval that contains 90% of the probability must be of the form $[a_i, a_j]$ with $1 \leq i \leq j \leq n$, because

$$\mathbf{prob}(\alpha \leq x \leq \beta) = \sum_{k=i}^j p_k = \mathbf{prob}(a_i \leq x \leq a_k)$$

where $i = \min\{k \mid a_k \geq \alpha\}$, and $j = \max\{k \mid a_k \leq \beta\}$.

We show that the function is quasiconcave. We have $f(p) \geq \gamma$ if and only if all intervals of width less than γ have a probability less than 90%,

$$\sum_{k=i}^{j} p_k < 0.9$$

for all i, j that satisfy $a_j - a_i < \gamma$. This defines a convex set.

The function is not convex, concave nor quasiconvex in general. Consider the example with $n = 3$, $a_1 = 0$, $a_2 = 0.5$ and $a_3 = 1$. On the line $p_1 + p_3 = 0.95$, we have

$$f(p) = \begin{cases} 0 & p_1 + p_3 = 0.95, \quad p_1 \in [0.05, 0.1] \cup [0.9, 0.95] \\ 0.5 & p_1 + p_3 = 0.95, \quad p_1 \in (0.1, 0.15] \cup [0.85, 0.9) \\ 1 & p_1 + p_3 = 0.95, \quad p_1 \in (0.15, 0.85) \end{cases}$$

It is clear that f is not convex, concave nor quasiconvex on the line.

3.25 *Maximum probability distance between distributions.* Let $p,\ q \in \mathbf{R}^n$ represent two probability distributions on $\{1, \ldots, n\}$ (so $p,\ q \succeq 0,\ \mathbf{1}^T p = \mathbf{1}^T q = 1$). We define the *maximum probability distance* $d_{\mathrm{mp}}(p, q)$ between p and q as the maximum difference in probability assigned by p and q, over all events:

$$d_{\mathrm{mp}}(p, q) = \max\{|\,\mathbf{prob}(p, C) - \mathbf{prob}(q, C)| \mid C \subseteq \{1, \ldots, n\}\}.$$

Here $\mathbf{prob}(p, C)$ is the probability of C, under the distribution p, *i.e.*, $\mathbf{prob}(p, C) = \sum_{i \in C} p_i$.

Find a simple expression for d_{mp}, involving $\|p - q\|_1 = \sum_{i=1}^n |p_i - q_i|$, and show that d_{mp} is a convex function on $\mathbf{R}^n \times \mathbf{R}^n$. (Its domain is $\{(p, q) \mid p,\ q \succeq 0,\ \mathbf{1}^T p = \mathbf{1}^T q = 1\}$, but it has a natural extension to all of $\mathbf{R}^n \times \mathbf{R}^n$.)

Solution. Noting that

$$\mathbf{prob}(p, C) - \mathbf{prob}(q, C) = -(\mathbf{prob}(p, \tilde{C}) - \mathbf{prob}(q, \tilde{C})),$$

where $\tilde{C} = \{1, \ldots, n\} \setminus C$, we can just as well express d_{mp} as

$$d_{\mathrm{mp}}(p, q) = \max\{\mathbf{prob}(p, C) - \mathbf{prob}(q, C) \mid C \subseteq \{1, \ldots, n\}\}.$$

This shows that d_{mp} is convex, since it is the maximum of 2^n linear functions of (p, q).

Let's now identify the (or a) subset C that maximizes

$$\mathbf{prob}(p, C) - \mathbf{prob}(q, C) = \sum_{i \in C} (p_i - q_i).$$

The solution is

$$C^\star = \{i \in \{1, \ldots, n\} \mid p_i > q_i\}.$$

Let's show this. The indices for which $p_i = q_i$ clearly don't matter, so we will ignore them, and assume without loss of generality that for each index, $p_i > q_i$ or $p_i < q_i$. Now consider any other subset C. If there is an element k in $C^\star$ but not C, then by adding k to C we increase $\mathbf{prob}(p, C) - \mathbf{prob}(q, C)$ by $p_k - q_k > 0$, so C could not have been optimal. Conversely, suppose that $k \in C \setminus C^\star$, so $p_k - q_k < 0$. If we remove k from C, we'd increase $\mathbf{prob}(p, C) - \mathbf{prob}(q, C)$ by $q_k - p_k > 0$, so C could not have been optimal. Thus, we have $d_{\mathrm{mp}}(p, q) = \sum_{p_i > q_i} (p_i - q_i)$. Now let's express this in terms of $\|p - q\|_1$. Using

$$\sum_{p_i > q_i} (p_i - q_i) + \sum_{p_i \leq q_i} (p_i - q_i) = \mathbf{1}^T p - \mathbf{1}^T q = 0,$$

we have

$$\sum_{p_i>q_i}(p_i-q_i) = -\left(\sum_{p_i\leq q_i}(p_i-q_i)\right),$$

so

$$\begin{aligned} d_{\mathrm{mp}}(p,q) &= (1/2)\sum_{p_i>q_i}(p_i-q_i) - (1/2)\sum_{p_i\leq q_i}(p_i-q_i) \\ &= (1/2)\sum_{i=1}^{n}|p_i-q_i| \\ &= (1/2)\|p-q\|_1. \end{aligned}$$

This makes it very clear that d_{mp} is convex.

The best way to interpret this result is as an interpretation of the ℓ_1-norm for probability distributions. It states that the ℓ_1-distance between two probability distributions is twice the maximum difference in probability, over all events, of the distributions.

3.26 *More functions of eigenvalues.* Let $\lambda_1(X) \geq \lambda_2(X) \geq \cdots \geq \lambda_n(X)$ denote the eigenvalues of a matrix $X \in \mathbf{S}^n$. We have already seen several functions of the eigenvalues that are convex or concave functions of X.

- The maximum eigenvalue $\lambda_1(X)$ is convex (example 3.10). The minimum eigenvalue $\lambda_n(X)$ is concave.
- The sum of the eigenvalues (or trace), $\mathbf{tr}\, X = \lambda_1(X) + \cdots + \lambda_n(X)$, is linear.
- The sum of the inverses of the eigenvalues (or trace of the inverse), $\mathbf{tr}(X^{-1}) = \sum_{i=1}^n 1/\lambda_i(X)$, is convex on $\mathbf{S}^n_{++}$ (exercise 3.18).
- The geometric mean of the eigenvalues, $(\det X)^{1/n} = (\prod_{i=1}^n \lambda_i(X))^{1/n}$, and the logarithm of the product of the eigenvalues, $\log\det X = \sum_{i=1}^n \log\lambda_i(X)$, are concave on $X \in \mathbf{S}^n_{++}$ (exercise 3.18 and page 74).

In this problem we explore some more functions of eigenvalues, by exploiting variational characterizations.

(a) *Sum of k largest eigenvalues.* Show that $\sum_{i=1}^k \lambda_i(X)$ is convex on $\mathbf{S}^n$. *Hint.* [HJ85, page 191] Use the variational characterization

$$\sum_{i=1}^{k}\lambda_i(X) = \sup\{\mathbf{tr}(V^TXV) \mid V \in \mathbf{R}^{n\times k},\ V^TV = I\}.$$

Solution. The variational characterization shows that f is the pointwise supremum of a family of linear functions $\mathbf{tr}(V^TXV)$.

(b) *Geometric mean of k smallest eigenvalues.* Show that $(\prod_{i=n-k+1}^n \lambda_i(X))^{1/k}$ is concave on $\mathbf{S}^n_{++}$. *Hint.* [MO79, page 513] For $X \succ 0$, we have

$$\left(\prod_{i=n-k+1}^{n}\lambda_i(X)\right)^{1/k} = \frac{1}{k}\inf\{\mathbf{tr}(V^TXV) \mid V \in \mathbf{R}^{n\times k},\ \det V^TV = 1\}.$$

Solution. f is the pointwise infimum of a family of linear functions $\mathbf{tr}(V^TXV)$.

(c) *Log of product of k smallest eigenvalues.* Show that $\sum_{i=n-k+1}^n \log\lambda_i(X)$ is concave on $\mathbf{S}^n_{++}$. *Hint.* [MO79, page 513] For $X \succ 0$,

$$\prod_{i=n-k+1}^{n}\lambda_i(X) = \inf\left\{\prod_{i=1}^{k}(V^TXV)_{ii} \;\middle|\; V \in \mathbf{R}^{n\times k},\ V^TV = I\right\}.$$

Solution. f is the pointwise infimum of a family of concave functions

$$\log \prod_i (V^T X V)_{ii} = \sum_i \log(V^T X V)_{ii}.$$

3.27 *Diagonal elements of Cholesky factor.* Each $X \in \mathbf{S}^n_{++}$ has a unique *Cholesky factorization* $X = LL^T$, where L is lower triangular, with $L_{ii} > 0$. Show that L_{ii} is a concave function of X (with domain $\mathbf{S}^n_{++}$).

Hint. L_{ii} can be expressed as $L_{ii} = (w - z^T Y^{-1} z)^{1/2}$, where

$$\begin{bmatrix} Y & z \\ z^T & w \end{bmatrix}$$

is the leading $i \times i$ submatrix of X.

Solution. The function $f(z, Y) = z^T Y^{-1} z$ with $\mathbf{dom}\, f = \{(z, Y) \mid Y \succ 0\}$ is convex jointly in z and Y. To see this note that

$$(z, Y, t) \in \mathbf{epi}\, f \qquad \Longleftrightarrow \qquad Y \succ 0, \quad \begin{bmatrix} Y & z \\ z^T & t \end{bmatrix} \succeq 0,$$

so $\mathbf{epi}\, f$ is a convex set. Therefore, $w - z^T Y^{-1} z$ is a concave function of X. Since the squareroot is an increasing concave function, it follows from the composition rules that $l_{kk} = (w - z^T Y^{-1} z)^{1/2}$ is a concave function of X.

Operations that preserve convexity

3.28 *Expressing a convex function as the pointwise supremum of a family of affine functions.* In this problem we extend the result proved on page 83 to the case where $\mathbf{dom}\, f \neq \mathbf{R}^n$. Let $f : \mathbf{R}^n \to \mathbf{R}$ be a convex function. Define $\tilde{f} : \mathbf{R}^n \to \mathbf{R}$ as the pointwise supremum of all affine functions that are global underestimators of f:

$$\tilde{f}(x) = \sup\{g(x) \mid g \text{ affine}, \; g(z) \leq f(z) \text{ for all } z\}.$$

(a) Show that $f(x) = \tilde{f}(x)$ for $x \in \mathbf{int}\, \mathbf{dom}\, f$.

(b) Show that $f = \tilde{f}$ if f is closed (*i.e.*, $\mathbf{epi}\, f$ is a closed set; see §A.3.3).

Solution.

(a) The point $(x, f(x))$ is in the boundary of $\mathbf{epi}\, f$. (If it were in $\mathbf{int}\, \mathbf{epi}\, f$, then for small, positive ϵ we would have $(x, f(x) - \epsilon) \in \mathbf{epi}\, f$, which is impossible.) From the results of §2.5.2, we know there is a supporting hyperplane to $\mathbf{epi}\, f$ at $(x, f(x))$, *i.e.*, $a \in \mathbf{R}^n$, $b \in \mathbf{R}$ such that

$$a^T z + bt \geq a^T x + bf(x) \;\text{ for all }\; (z, t) \in \mathbf{epi}\, f.$$

Since t can be arbitrarily large if $(z, t) \in \mathbf{epi}\, f$, we conclude that $b \geq 0$.

Suppose $b = 0$. Then

$$a^T z \geq a^T x \text{ for all } z \in \mathbf{dom}\, f$$

which contradicts $x \in \mathbf{int}\, \mathbf{dom}\, f$. Therefore $b > 0$. Dividing the above inequality by b yields

$$t \geq f(x) + (a/b)^T (x - z) \;\text{ for all }\; (z, t) \in \mathbf{epi}\, f.$$

Therefore the affine function

$$g(z) = f(x) + (a/b)^T (x - z)$$

is an affine global underestimator of f, and hence by definition of $\tilde{f}$,

$$f(x) \geq \tilde{f}(x) \geq g(x).$$

However $g(x) = f(x)$, so we must have $f(x) = \tilde{f}(x)$.

(b) A closed convex set is the intersection of all halfspaces that contain it (see chapter 2, example 2.20). We will apply this result to $\mathbf{epi}\, f$. Define

$$H = \{(a,b,c) \in \mathbf{R}^{n+2} \mid (a,b) \neq 0, \inf_{(x,t)\in \mathbf{epi}\, f}(a^T x + bt) \geq c\}.$$

Loosely speaking, H is the set of all halfspaces that contain $\mathbf{epi}\, f$. By the result in chapter 2,

$$\mathbf{epi}\, f = \bigcap_{(a,b,c)\in H} \{(x,t) \mid a^T x + bt \geq c\}. \tag{3.28.A}$$

It is clear that all elements of H satisfy $b \geq 0$. If in fact $b > 0$, then the affine function

$$h(x) = -(a/b)^T x + c/b,$$

minorizes f, since

$$t \geq f(x) \geq -(a/b)^T x + c/t = h(x)$$

for all $(x,t) \in \mathbf{epi}\, f$. Conversely, if $h(x) = -a^T x + c$ minorizes f, then $(a,1,c) \in H$. We need to prove that

$$\mathbf{epi}\, f = \bigcap_{(a,b,c)\in H,\ b>0} \{(x,t) \mid a^T x + bt \geq c\}.$$

(In words, $\mathbf{epi}\, f$ is the intersection of all 'non-vertical' halfspaces that contain $\mathbf{epi}\, f$.) Note that H may contain elements with $b = 0$, so this does not immediately follow from (3.28.A).

We will show that

$$\bigcap_{(a,b,c)\in H,\ b>0} \{(x,t) \mid a^T x + bt \geq c\} = \bigcap_{(a,b,c)\in H} \{(x,t) \mid a^T x + bt \geq c\}. \tag{3.28.B}$$

It is obvious that the set on the left includes the set on the right. To show that they are identical, assume $(\bar{x},\bar{t})$ lies in the set on the left, *i.e.*,

$$a^T \bar{x} + b\bar{t} \geq c$$

for all halfspaces $a^T x + bt \geq c$ that are nonvertical (*i.e.*, $b > 0$) and contain $\mathbf{epi}\, f$. Assume that $(\bar{x},\bar{t})$ is not in the set on the right, *i.e.*, there exist $(\tilde{a},\tilde{b},\tilde{c}) \in H$ (necessarily with $\tilde{b} = 0$), such that

$$\tilde{a}^T \bar{x} < \tilde{c}.$$

H contains at least one element (a_0,b_0,c_0) with $b_0 > 0$. (Otherwise $\mathbf{epi}\, f$ would be an intersection of vertical halfspaces.) Consider the halfspace defined by $(\tilde{a},0,\tilde{c}) + \epsilon(a_0,b_0,c_0)$ for small positive ϵ. This halfspace is nonvertical and it contains $\mathbf{epi}\, f$:

$$(\tilde{a} + \epsilon a_0)^T x + \epsilon b_0 t \geq \tilde{a}^T x + \epsilon(a_0^T x + b_0 t) \geq \tilde{c} + \epsilon c_0,$$

for all $(x,t) \in \mathbf{epi}\, f$, because the halfspaces $\tilde{a}^T x \geq \tilde{c}$ and $a_0^T x + b_0 t \geq c_0$ both contain $\mathbf{epi}\, f$. However,

$$(\tilde{a} + \epsilon a_0)^T \bar{x} + \epsilon b_0 \bar{t} = \tilde{a}^T \bar{x} + \epsilon(a_0^T \bar{x} + b_0 \bar{t}) < \tilde{c} + \epsilon c_0$$

for small ϵ, so the halfspace does not contain $(\bar{x},\bar{t})$. This contradicts our assumption that $(\bar{x},\bar{t})$ is in the intersection of all nonvertical halfspaces containing $\mathbf{epi}\, f$. We conclude that the equality (3.28.B) holds.

3.29 *Representation of piecewise-linear convex functions.* A function $f : \mathbf{R}^n \to \mathbf{R}$, with $\mathbf{dom}\, f = \mathbf{R}^n$, is called *piecewise-linear* if there exists a partition of $\mathbf{R}^n$ as

$$\mathbf{R}^n = X_1 \cup X_2 \cup \cdots \cup X_L,$$

where $\mathbf{int}\, X_i \neq \emptyset$ and $\mathbf{int}\, X_i \cap \mathbf{int}\, X_j = \emptyset$ for $i \neq j$, and a family of affine functions $a_1^T x + b_1, \ldots, a_L^T x + b_L$ such that $f(x) = a_i^T x + b_i$ for $x \in X_i$.

Show that this means that $f(x) = \max\{a_1^T x + b_1, \ldots, a_L^T x + b_L\}$.

Solution. By Jensen's inequality, we have for all x, $y \in \mathbf{dom}\, f$, and $t \in [0, 1]$,

$$f(y + t(x - y)) \leq f(y) + t(f(x) - f(y)),$$

and hence

$$f(x) \geq f(y) + \frac{f(y + t(x - y)) - f(y)}{t}.$$

Now suppose $x \in X_i$. Choose any $y \in \mathbf{int}\, X_j$, for some j, and take t sufficiently small so that $y + t(x - y) \in X_j$. The above inequality reduces to

$$a_i^T x + b_i \geq a_j^T y + b_j + \frac{(a_j^T (y + t(x - y)) + b_j - a_j^T y - b_j)}{t} = a_j^T x + b_j.$$

This is true for any j, so $a_i^T x + b_i \geq \max_{j=1,\ldots,L}(a_j^T x + b_j)$. We conclude that

$$a_i^T x + b_i = \max_{j=1,\ldots,L}(a_j^T x + b_j).$$

3.30 *Convex hull or envelope of a function.* The *convex hull* or *convex envelope* of a function $f : \mathbf{R}^n \to \mathbf{R}$ is defined as

$$g(x) = \inf\{t \mid (x, t) \in \mathbf{conv}\, \mathbf{epi}\, f\}.$$

Geometrically, the epigraph of g is the convex hull of the epigraph of f.

Show that g is the largest convex underestimator of f. In other words, show that if h is convex and satisfies $h(x) \leq f(x)$ for all x, then $h(x) \leq g(x)$ for all x.

Solution. It is clear that g is convex, since by construction its epigraph is a convex set.

Let h be a convex lower bound on f. Since h is convex, $\mathbf{epi}\, h$ is a convex set. Since h is a lower bound on f, $\mathbf{epi}\, f \subseteq \mathbf{epi}\, h$. By definition the convex hull of a set is the intersection of all the convex sets that contain the set. It follows that $\mathbf{conv}\, \mathbf{epi}\, f = \mathbf{epi}\, g \subseteq \mathbf{epi}\, h$, *i.e.*, $g(x) \geq h(x)$ for all x.

3.31 [Roc70, page 35] *Largest homogeneous underestimator.* Let f be a convex function. Define the function g as

$$g(x) = \inf_{\alpha > 0} \frac{f(\alpha x)}{\alpha}.$$

(a) Show that g is homogeneous ($g(tx) = tg(x)$ for all $t \geq 0$).

(b) Show that g is the largest homogeneous underestimator of f: If h is homogeneous and $h(x) \leq f(x)$ for all x, then we have $h(x) \leq g(x)$ for all x.

(c) Show that g is convex.

Solution.

(a) If $t > 0$,

$$g(tx) = \inf_{\alpha > 0} \frac{f(\alpha t x)}{\alpha} = t \inf_{\alpha > 0} \frac{f(\alpha t x)}{t\alpha} = tg(x).$$

For $t = 0$, we have $g(tx) = g(0) = 0$.

(b) If h is a homogeneous underestimator, then

$$h(x) = \frac{h(\alpha x)}{\alpha} \leq \frac{f(\alpha x)}{\alpha}$$

for all $\alpha > 0$. Taking the infimum over α gives $h(x) \leq g(x)$.

(c) We can express g as

$$g(x) = \inf_{t>0} tf(x/t) = \inf_{t>0} h(x,t)$$

where h is the perspective function of f. We know h is convex, jointly in x and t, so g is convex.

3.32 *Products and ratios of convex functions.* In general the product or ratio of two convex functions is not convex. However, there are some results that apply to functions on $\mathbf{R}$. Prove the following.

(a) If f and g are convex, both nondecreasing (or nonincreasing), and positive functions on an interval, then fg is convex.

(b) If f, g are concave, positive, with one nondecreasing and the other nonincreasing, then fg is concave.

(c) If f is convex, nondecreasing, and positive, and g is concave, nonincreasing, and positive, then f/g is convex.

Solution.

(a) We prove the result by verifying Jensen's inequality. f and g are positive and convex, hence for $0 \leq \theta \leq 1$,

$$\begin{aligned} f(\theta x + (1-\theta)y)\, g(\theta x + (1-\theta)y) &\leq (\theta f(x) + (1-\theta)f(y))\,(\theta g(x) + (1-\theta)g(y)) \\ &= \theta f(x)g(x) + (1-\theta)f(y)g(y) \\ &\quad + \theta(1-\theta)(f(y) - f(x))(g(x) - g(y)). \end{aligned}$$

The third term is less than or equal to zero if f and g are both increasing or both decreasing. Therefore

$$f(\theta x + (1-\theta)y)\, g(\theta x + (1-\theta)y) \leq \theta f(x)g(x) + (1-\theta)f(y)g(y).$$

(b) Reverse the inequalities in the solution of part (a).

(c) It suffices to note that $1/g$ is convex, positive and increasing, so the result follows from part (a).

3.33 *Direct proof of perspective theorem.* Give a direct proof that the perspective function g, as defined in §3.2.6, of a convex function f is convex: Show that $\mathbf{dom}\, g$ is a convex set, and that for (x,t), $(y,s) \in \mathbf{dom}\, g$, and $0 \leq \theta \leq 1$, we have

$$g(\theta x + (1-\theta)y, \theta t + (1-\theta)s) \leq \theta g(x,t) + (1-\theta)g(y,s).$$

Solution. The domain $\mathbf{dom}\, g = \{(x,t) \mid x/t \in \mathbf{dom}\, f, t > 0\}$ is the inverse image of $\mathbf{dom}\, f$ under the perspective function $P : \mathbf{R}^{n+1} \to \mathbf{R}^n$, $P(x,t) = x/t$ for $t > 0$, so it is convex (see §2.3.3).

Jensen's inequality can be proved directly as follows. Suppose $s, t > 0$, $x/t \in \mathbf{dom}\, f$, $y/s \in \mathbf{dom}\, f$, and $0 \leq \theta \leq 1$. Then

$$\begin{aligned} & g(\theta x + (1-\theta)y, \theta t + (1-\theta)s) \\ &= (\theta t + (1-\theta)s) f((\theta x + (1-\theta)y)/(\theta t + (1-\theta)s)) \\ &= (\theta t + (1-\theta)s) f((\theta t(x/t) + (1-\theta)s(y/s))/(\theta t + (1-\theta)s)) \\ &\leq \theta t f(x/t) + (1-\theta)s f(y/s). \end{aligned}$$

3.34 *The Minkowski function.* The *Minkowski function* of a convex set C is defined as

$$M_C(x) = \inf\{t > 0 \mid t^{-1}x \in C\}.$$

(a) Draw a picture giving a geometric interpretation of how to find $M_C(x)$.

(b) Show that M_C is homogeneous, *i.e.*, $M_C(\alpha x) = \alpha M_C(x)$ for $\alpha \geq 0$.

(c) What is $\mathbf{dom}\, M_C$?

(d) Show that M_C is a convex function.

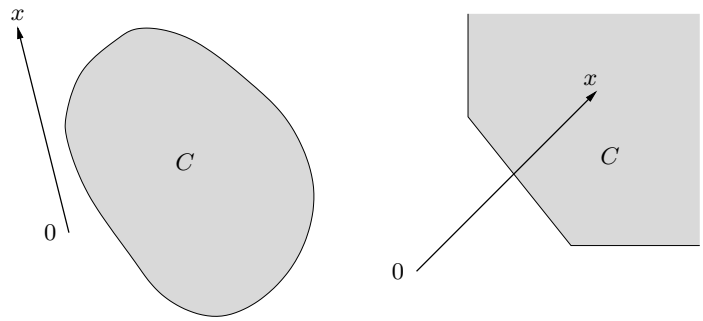
(e) Suppose C is also closed, symmetric (if $x \in C$ then $-x \in C$), and has nonempty interior. Show that M_C is a norm. What is the corresponding unit ball?

Solution.

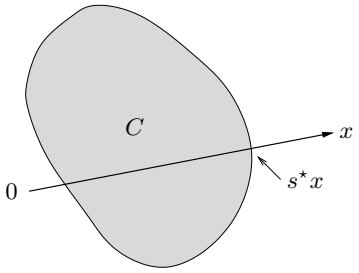
(a) Consider the ray, excluding 0, generated by x, *i.e.*, sx for $s > 0$. The intersection of this ray and C is either empty (meaning, the ray doesn't intersect C), a finite interval, or another ray (meaning, the ray enters C and stays in C).

In the first case, the set $\{t > 0 \mid t^{-1}x \in C\}$ is empty, so the infimum is ∞. This means $M_C(x) = \infty$. This case is illustrated in the figure below, on the left.

In the third case, the set $\{s > 0 \mid sx \in C\}$ has the form $[a, \infty)$ or (a, ∞), so the set $\{t > 0 \mid t^{-1}x \in C\}$ has the form $(0, 1/a]$ or $(0, 1/a)$. In this case we have $M_C(x) = 0$. That is illustrated in the figure below to the right.



In the second case, the set $\{s > 0 \mid sx \in C\}$ is a bounded , interval with endpoints $a \leq b$, so we have $M_C(x) = 1/b$. That is shown below. In this example, the optimal scale factor is around $s^\star \approx 3/4$, so $M_C(x) \approx 4/3$.



In any case, if $x = 0 \in C$ then $M_C(0) = 0$.

(b) If $\alpha > 0$, then

$$\begin{aligned} M_C(\alpha x) &= \inf\{t > 0 \mid t^{-1}\alpha x \in C\} \\ &= \alpha \inf\{t/\alpha > 0 \mid t^{-1}\alpha x \in C\} \\ &= \alpha M_C(x). \end{aligned}$$

If $\alpha = 0$, then

$$M_C(\alpha x) = M_C(0) = \begin{cases} 0 & 0 \in C \\ \infty & 0 \notin C. \end{cases}$$

(c) $\mathbf{dom}\, M_C = \{x \mid x/t \in C \text{ for some } t > 0\}$. This is also known as the conic hull of C, except that $0 \in \mathbf{dom}\, M_C$ only if $0 \in C$.

(d) We have already seen that $\mathbf{dom}\, M_C$ is a convex set. Suppose $x, y \in \mathbf{dom}\, M_C$, and let $\theta \in [0,1]$. Consider any $t_x, t_y > 0$ for which $x/t_x \in C$, $y/t_y \in C$. (There exists at least one such pair, because $x, y \in \mathbf{dom}\, M_C$.) It follows from convexity of C that

$$\frac{\theta x + (1-\theta)y}{\theta t_x + (1-\theta)t_y)} = \frac{\theta t_x (x/t_x) + (1-\theta)t_y(y/t_y)}{\theta t_x + (1-\theta)t_y} \in C$$

and therefore

$$M_C(\theta x + (1-\theta)y) \leq \theta t_x + (1-\theta)t_y.$$

This is true for any $t_x, t_y > 0$ that satisfy $x/t_x \in C$, $y/t_y \in C$. Therefore

$$\begin{array}{rcl} M_C(\theta x + (1-\theta)y) & \leq & \theta \inf\{t_x > 0 \mid x/t_x \in C\} + (1-\theta)\inf\{t_y > 0 \mid y/t_y \in C\} \\ & = & \theta M_C(x) + (1-\theta)M_C(y). \end{array}$$

Here is an alternative snappy, modern style proof:

- The indicator function of C, *i.e.*, I_C, is convex.
- The perspective function, $tI_C(x/t)$ is convex in (x,t). But this is the same as $I_C(x/t)$, so $I_C(x/t)$ is convex in (x,t).
- The function $t + I_C(x/t)$ is convex in (x,t).
- Now let's minimize over t, to obtain $\inf_t(t + I_C(x/t)) = M_C(x)$, which is convex by the minimization rule.

(e) It is the norm with unit ball C.

(a) Since by assumption, $0 \in \mathbf{int}\, C$, $M_C(x) > 0$ for $x \neq 0$. By definition $M_C(0) = 0$.

(b) Homogeneity: for $\lambda > 0$,

$$\begin{array}{rcl} M_C(\lambda x) & = & \inf\{t > 0 \mid (t\lambda)^{-1}x \in C\} \\ & = & \lambda \inf\{u > 0 \mid u^{-1}x \in C\} \\ & = & \lambda M_C(x). \end{array}$$

By symmetry of C, we also have $M_C(-x) = -M_C(x)$.

(c) Triangle inequality. By convexity (part d), and homogeneity,

$$M_C(x+y) = 2M_C((1/2)x + (1/2)y) \leq M_C(x) + M_C(y).$$

3.35 *Support function calculus.* Recall that the support function of a set $C \subseteq \mathbf{R}^n$ is defined as $S_C(y) = \sup\{y^T x \mid x \in C\}$. On page 81 we showed that S_C is a convex function.

(a) Show that $S_B = S_{\mathbf{conv}\, B}$.

(b) Show that $S_{A+B} = S_A + S_B$.

(c) Show that $S_{A \cup B} = \max\{S_A, S_B\}$.

(d) Let B be closed and convex. Show that $A \subseteq B$ if and only if $S_A(y) \leq S_B(y)$ for all y.

Solution.

(a) Let $A = \mathbf{conv}\, B$. Since $B \subseteq A$, we obviously have $S_B(y) \leq S_A(y)$. Suppose we have strict inequality for some y, *i.e.*,

$$y^T u < y^T v$$

for all $u \in B$ and some $v \in A$. This leads to a contradiction, because by definition v is the convex combination of a set of points $u_i \in B$, *i.e.*, $v = \sum_i \theta_i u_i$, with $\theta_i \geq 0$, $\sum_i \theta_i = 1$. Since

$$y^T u_i < y^T v$$

for all i, this would imply

$$y^T v = \sum_i \theta_i y^T u_i < \sum_i \theta_i y^T v = y^T v.$$

We conclude that we must have equality $S_B(y) = S_A(y)$.

(b) Follows from

$$\begin{aligned} S_{A+B}(y) &= \sup\{y^T(u+v) \mid u \in A, v \in B\} \\ &= \sup\{y^T u \mid u \in A\} + \sup\{y^T v \mid u \in B\} \\ &= S_A(y) + S_B(y). \end{aligned}$$

(c) Follows from

$$\begin{aligned} S_{A\cup B}(y) &= \sup\{y^T u \mid u \in A \cup B\} \\ &= \max\{\sup\{y^T u \mid u \in A\}, \sup\{y^T v \mid u \in B\} \\ &= \max\{S_A(y), S_B(y)\}. \end{aligned}$$

(d) Obviously, if $A \subseteq B$, then $S_A(y) \leq S_B(y)$ for all y. We need to show that if $A \not\subseteq B$, then $S_A(y) > S_B(y)$ for some y.
Suppose $A \not\subseteq B$. Consider a point $\bar{x} \in A$, $\bar{x} \notin B$. Since B is closed and convex, $\bar{x}$ can be strictly separated from B by a hyperplane, *i.e.*, there is a $y \neq 0$ such that

$$y^T \bar{x} > y^T x$$

for all $x \in B$. It follows that $S_B(y) < y^T \bar{x} \leq S_A(y)$.

Conjugate functions

3.36 Derive the conjugates of the following functions.

(a) *Max function.* $f(x) = \max_{i=1,\ldots,n} x_i$ on $\mathbf{R}^n$.
Solution. We will show that

$$f^*(y) = \begin{cases} 0 & \text{if } y \succeq 0,\ \ \mathbf{1}^T y = 1 \\ \infty & \text{otherwise.} \end{cases}$$

We first verify the domain of f^*. First suppose y has a negative component, say $y_k < 0$. If we choose a vector x with $x_k = -t$, $x_i = 0$ for $i \neq k$, and let t go to infinity, we see that

$$x^T y - \max_i x_i = -t y_k \to \infty,$$

so y is not in $\mathbf{dom}\, f^*$. Next, assume $y \succeq 0$ but $\mathbf{1}^T y > 1$. We choose $x = t\mathbf{1}$ and let t go to infinity, to show that

$$x^T y - \max_i x_i = t\mathbf{1}^T y - t$$

is unbounded above. Similarly, when $y \succeq 0$ and $\mathbf{1}^T y < 1$, we choose $x = -t\mathbf{1}$ and let t go to infinity.

The remaining case for y is $y \succeq 0$ and $\mathbf{1}^T y = 1$. In this case we have

$$x^T y \leq \max_i x_i$$

for all x, and therefore $x^T y - \max_i x_i \leq 0$ for all x, with equality for $x = 0$. Therefore $f^*(y) = 0$.

(b) *Sum of largest elements.* $f(x) = \sum_{i=1}^r x_{[i]}$ on $\mathbf{R}^n$.

Solution. The conjugate is

$$f^*(y) = \begin{cases} 0 & 0 \preceq y \preceq \mathbf{1}, \quad \mathbf{1}^T y = r \\ \infty & \text{otherwise,} \end{cases}$$

We first verify the domain of f^*. Suppose y has a negative component, say $y_k < 0$. If we choose a vector x with $x_k = -t$, $x_i = 0$ for $i \neq k$, and let t go to infinity, we see that

$$x^T y - f(x) = -ty_k \to \infty,$$

so y is not in $\mathbf{dom}\, f^*$.

Next, suppose y has a component greater than 1, say $y_k > 1$. If we choose a vector x with $x_k = t$, $x_i = 0$ for $i \neq k$, and let t go to infinity, we see that

$$x^T y - f(x) = ty_k - t \to \infty,$$

so y is not in $\mathbf{dom}\, f^*$.

Finally, assume that $\mathbf{1}^T x \neq r$. We choose $x = t\mathbf{1}$ and find that

$$x^T y - f(x) = t\mathbf{1}^T y - tr$$

is unbounded above, as $t \to \infty$ or $t \to -\infty$.

If y satisfies all the conditions we have

$$x^T y \leq f(x)$$

for all x, with equality for $x = 0$. Therefore $f^*(y) = 0$.

(c) *Piecewise-linear function on* $\mathbf{R}$. $f(x) = \max_{i=1,\ldots,m}(a_i x + b_i)$ on $\mathbf{R}$. You can assume that the a_i are sorted in increasing order, *i.e.*, $a_1 \leq \cdots \leq a_m$, and that none of the functions $a_i x + b_i$ is redundant, *i.e.*, for each k there is at least one x with $f(x) = a_k x + b_k$.

Solution. Under the assumption, the graph of f is a piecewise-linear, with breakpoints $(b_i - b_{i+1})/(a_{i+1} - a_i)$, $i = 1, \ldots, m-1$. We can write f^* as

$$f^*(y) = \sup_x \left(xy - \max_{i=1,\ldots,m} (a_i x + b_i) \right)$$

We see that $\mathbf{dom}\, f^* = [a_1, a_m]$, since for y outside that range, the expression inside the supremum is unbounded above. For $a_i \leq y \leq a_{i+1}$, the supremum in the definition of f^* is reached at the breakpoint between the segments i and $i+1$, *i.e.*, at the point $(b_{i+1} - b_i)/(a_{i+1} - a_i)$, so we obtain

$$f^*(y) = -b_i - (b_{i+1} - b_i)\frac{y - a_i}{a_{i+1} - a_i}$$

where i is defined by $a_i \leq y \leq a_{i+1}$. Hence the graph of f^* is also a piecewise-linear curve connecting the points $(a_i, -b_i)$ for $i = 1, \ldots, m$. Geometrically, the epigraph of f^* is the epigraphical hull of the points $(a_i, -b_i)$.

(d) *Power function.* $f(x) = x^p$ on $\mathbf{R}_{++}$, where $p > 1$. Repeat for $p < 0$.

Solution. We'll use standard notation: we define q by the equation $1/p + 1/q = 1$, *i.e.*, $q = p/(p-1)$.

We start with the case $p > 1$. Then x^p is strictly convex on $\mathbf{R}_+$. For $y < 0$ the function $yx - x^p$ achieves its maximum for $x > 0$ at $x = 0$, so $f^*(y) = 0$. For $y > 0$ the function achieves its maximum at $x = (y/p)^{1/(p-1)}$, where it has value

$$y(y/p)^{1/(p-1)} - (y/p)^{p/(p-1)} = (p-1)(y/p)^q.$$

Therefore we have

$$f^*(y) = \begin{cases} 0 & y \leq 0 \\ (p-1)(y/p)^q & y > 0. \end{cases}$$

For $p < 0$ similar arguments show that $\mathbf{dom}\, f^* = -\mathbf{R}_{++}$ and $f^*(y) = \frac{-p}{q}(-y/p)^q$.

(e) *Geometric mean.* $f(x) = -(\prod x_i)^{1/n}$ on $\mathbf{R}^n_{++}$.

Solution. The conjugate function is

$$f^*(y) = \begin{cases} 0 & \text{if } y \preceq 0, \ \left(\prod_i (-y_i)\right)^{1/n} \geq 1/n \\ \infty & \text{otherwise.} \end{cases}$$

We first verify the domain of f^*. Assume y has a positive component, say $y_k > 0$. Then we can choose $x_k = t$ and $x_i = 1$, $i \neq k$, to show that

$$x^T y - f(x) = ty_k + \sum_{i \neq k} y_i - t^{1/n}$$

is unbounded above as a function of $t > 0$. Hence the condition $y \preceq 0$ is indeed required.

Next assume that $y \preceq 0$, but $(\prod_i (-y_i))^{1/n} < 1/n$. We choose $x_i = -t/y_i$, and obtain

$$x^T y - f(x) = -tn - t\left(\prod_i (-\frac{1}{y_i})\right)^{1/n} \to \infty$$

as $t \to \infty$. This demonstrates that the second condition for the domain of f^* is also needed.

Now assume that $y \preceq 0$ and $\left(\prod_i (-y_i)\right)^{1/n} \geq 1/n$, and $x \succeq 0$. The arithmetic-geometric mean inequality states that

$$\frac{x^T y}{n} \geq \left(\prod_i (-y_i x_i)\right)^{1/n} \geq \frac{1}{n}\left(\prod_i x_i\right)^{1/n},$$

i.e., $x^T y \geq f(x)$ with equality for $x_i = -1/y_i$. Hence, $f^*(y) = 0$.

(f) *Negative generalized logarithm for second-order cone.* $f(x,t) = -\log(t^2 - x^T x)$ on $\{(x,t) \in \mathbf{R}^n \times \mathbf{R} \mid \|x\|_2 < t\}$.

Solution.

$$f^*(y,u) = -2 + \log 4 - \log(u^2 - y^T y), \qquad \mathbf{dom}\, f^* = \{(y,u) \mid \|y\|_2 < -u\}.$$

We first verify the domain. Suppose $\|y\|_2 \geq -u$. Choose $x = sy$, $t = s(\|x\|_2 + 1) > s\|y\|_2 \geq -su$, with $s \geq 0$. Then

$$y^T x + tu > sy^T y - su^2 = s(u^2 - y^T y) \geq 0,$$

so $y^x + tu$ goes to infinity, at a linear rate, while the function $-\log(t^2 - x^T x)$ goes to $-\infty$ as $-\log s$. Therefore

$$y^T x + tu + \log(t^2 - x^T x)$$

is unbounded above.
Next, assume that $\|y\|_2 < u$. Setting the derivative of

$$y^T x + ut + \log(t^2 - x^T x)$$

with respect to x and t equal to zero, and solving for t and x we see that the maximizer is

$$x = \frac{2y}{u^2 - y^T y}, \qquad t = -\frac{2u}{u^2 - y^T y}.$$

This gives

$$\begin{aligned} f^*(y,u) &= ut + y^T x + \log(t^2 - x^T x) \\ &= -2 + \log 4 - \log(y^2 - u^t u). \end{aligned}$$

3.37 Show that the conjugate of $f(X) = \mathbf{tr}(X^{-1})$ with $\mathbf{dom}\, f = \mathbf{S}^n_{++}$ is given by

$$f^*(Y) = -2\,\mathbf{tr}(-Y)^{1/2}, \qquad \mathbf{dom}\, f^* = -\mathbf{S}^n_+.$$

Hint. The gradient of f is $\nabla f(X) = -X^{-2}$.
Solution. We first verify the domain of f^*. Suppose Y has eigenvalue decomposition

$$Y = Q\Lambda Q^T = \sum_{i=1}^n \lambda_i q_i q_i^T$$

with $\lambda_1 > 0$. Let $X = Q\,\mathbf{diag}(t, 1, \ldots, 1)Q^T = tq_1 q_1^T + \sum_{i=2}^n q_i q_i^T$. We have

$$\mathbf{tr}\, XY - \mathbf{tr}\, X^{-1} = t\lambda_1 + \sum_{i=2}^n \lambda_i - 1/t - (n-1),$$

which grows unboundedly as $t \to \infty$. Therefore $Y \notin \mathbf{dom}\, f^*$.
Next, assume $Y \preceq 0$. If $Y \prec 0$, we can find the maximum of

$$\mathbf{tr}\, XY - \mathbf{tr}\, X^{-1}$$

by setting the gradient equal to zero. We obtain $Y = -X^{-2}$, *i.e.*, $X = (-Y)^{-1/2}$, and

$$f^*(Y) = -2\,\mathbf{tr}(-Y)^{1/2}.$$

Finally we verify that this expression remains valid when $Y \preceq 0$, but Y is singular. This follows from the fact that conjugate functions are always closed, *i.e.*, have closed epigraphs.

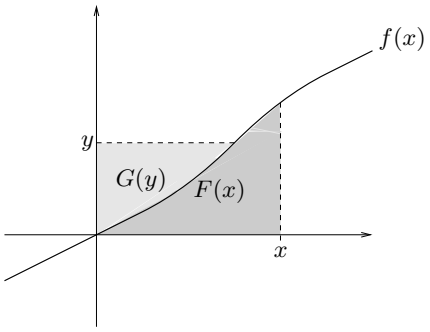
3.38 *Young's inequality.* Let $f : \mathbf{R} \to \mathbf{R}$ be an increasing function, with $f(0) = 0$, and let g be its inverse. Define F and G as

$$F(x) = \int_0^x f(a)\,da, \qquad G(y) = \int_0^y g(a)\,da.$$

Show that F and G are conjugates. Give a simple graphical interpretation of Young's inequality,

$$xy \le F(x) + G(y).$$

Solution. The inequality $xy \le F(x) + G(y)$ has a simple geometric meaning, illustrated below.



$F(x)$ is the shaded area under the graph of f, from 0 to x. $G(y)$ is the area above the graph of f, from 0 to y. For fixed x and y, $F(x) + G(y)$ is the total area below the graph, up to x, and above the graph, up to y. This is at least equal to xy, the area of the rectangle defined by x and y, hence

$$F(x) + G(y) \geq xy$$

for all x, y.

It is also clear that $F(x) + G(y) = xy$ if and only if $y = f(x)$. In other words

$$G(y) = \sup_x(xy - F(x)), \qquad F(x) = \sup_y(xy - G(y)),$$

i.e., the functions are conjugates.

3.39 *Properties of conjugate functions.*

(a) *Conjugate of convex plus affine function.* Define $g(x) = f(x) + c^T x + d$, where f is convex. Express g^* in terms of f^* (and c, d).

Solution.

$$\begin{aligned} g^*(y) &= \sup(y^T x - f(x) - c^T x - d) \\ &= \sup((y-c)^T x - f(x)) - d \\ &= f^*(y-c) - d. \end{aligned}$$

(b) *Conjugate of perspective.* Express the conjugate of the perspective of a convex function f in terms of f^*.

Solution.

$$\begin{aligned} g^*(y,s) &= \sup_{x/t \in \mathbf{dom}\, f, t>0} (y^T x + st - tf(x/t)) \\ &= \sup_{t>0} \sup_{x/t \in \mathbf{dom}\, f} (t(y^T(x/t) + s - f(x/t))) \\ &= \sup_{t>0} t(s + \sup_{x/t \in \mathbf{dom}\, f} (y^T(x/t) - f(x/t))) \\ &= \sup_{t>0} t(s + f^*(y)) \\ &= \begin{cases} 0 & s + f^*(y) \leq 0 \\ \infty & \text{otherwise.} \end{cases} \end{aligned}$$

(c) *Conjugate and minimization.* Let $f(x,z)$ be convex in (x,z) and define $g(x) = \inf_z f(x,z)$. Express the conjugate g^* in terms of f^*.

As an application, express the conjugate of $g(x) = \inf_z\{h(z) \mid Az+b = x\}$, where h is convex, in terms of h^*, A, and b.

Solution.

$$\begin{aligned} g^*(y) &= \sup_x(x^Ty - \inf_z f(x,z)) \\ &= \sup_{x,z}(x^Ty - f(x,z)) \\ &= f^*(y,0). \end{aligned}$$

To answer the second part of the problem, we apply the previous result to

$$f(x,z) = \begin{cases} h(z) & Az+b = x \\ \infty & \text{otherwise.} \end{cases}$$

We have

$$\begin{aligned} f^*(y,v) &= \inf(y^Tx - v^Tz - f(x,z)) \\ &= \inf_{Az+b=x}(y^Tx - v^Tz - h(z)) \\ &= \inf_z(y^T(Az+b) - v^Tz - h(z)) \\ &= b^Ty + \inf_z(y^TAz - v^Tz - h(z)) \\ &= b^Ty + h^*(A^Ty - v). \end{aligned}$$

Therefore

$$g^*(y) = f^*(y,0) = b^Ty + h^*(A^Ty).$$

(d) *Conjugate of conjugate.* Show that the conjugate of the conjugate of a closed convex function is itself: $f = f^{**}$ if f is closed and convex. (A function is closed if its epigraph is closed; see §A.3.3.) *Hint.* Show that f^{**} is the pointwise supremum of all affine global underestimators of f. Then apply the result of exercise 3.28.

Solution. By definition of f^*,

$$f^*(y) = \sup_x(y^Tx - f(x)).$$

If $y \in \mathbf{dom}\, f^*$, then the affine function $h(x) = y^Tx - f^*(y)$, minorizes f. Conversely, if $h(x) = a^Tx + b$ minorizes f, then $a \in \mathbf{dom}\, f^*$ and $f^*(a) \leq -b$. The set of all affine functions that minorize f is therefore exactly equal to the set of all functions $h(x) = y^Tx + c$ where

$$y \in \mathbf{dom}\, f^*, \qquad c \leq -f^*(y).$$

Therefore, by the result of exercise 3.28,

$$f(x) = \sup_{y \in \mathbf{dom}\, f^*}(y^Tx - f^*(y)) = f^{**}(y).$$

3.40 *Gradient and Hessian of conjugate function.* Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is convex and twice continuously differentiable. Suppose $\bar{y}$ and $\bar{x}$ are related by $\bar{y} = \nabla f(\bar{x})$, and that $\nabla^2 f(\bar{x}) \succ 0$.

(a) Show that $\nabla f^*(\bar{y}) = \bar{x}$.

(b) Show that $\nabla^2 f^*(\bar{y}) = \nabla^2 f(\bar{x})^{-1}$.

Solution. We use the implicit function theorem: Suppose $F : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ satisfies

- $F(\bar{u}, \bar{v}) = 0$
- F is continuously differentiable and $D_v F(u, v)$ is nonsingular in a neighborhood of $(\bar{u}, \bar{v})$.

Then there exists a continuously differentiable function $\phi : \mathbf{R}^n \to \mathbf{R}^m$, that satisfies $\bar{v} = \phi(\bar{u})$ and

$$F(u, \phi(u)) = 0$$

in a neighborhood of $\bar{u}$.

Applying this to $u = y$, $v = x$, and $F(u, v) = \nabla f(x) - y$, we see that there exists a continuously differentiable function g such that

$$\bar{x} = g(\bar{y}),$$

and

$$\nabla f(g(y)) = y$$

in a neighborhood around $\bar{y}$. Differentiating both sides with respect to y gives

$$\nabla^2 f(g(y)) Dg(y) = I,$$

i.e., $Dg(y) = \nabla^2 f(g(y))^{-1}$, in a neighborhood of $\bar{y}$.

Now suppose y is near $\bar{y}$. The maximum in the definition of $f^*(y)$,

$$f^*(y) = \sup_x (\tilde{y}^T x - f(x)),$$

is attained at $x = g(y)$, and the maximizer is unique, by the fact that $\nabla^2 f(\bar{x}) \succ 0$. We therefore have

$$f^*(y) = y^T g(y) - f(g(y)).$$

Differentiating with respect to y gives

$$\begin{aligned} \nabla f^*(y) &= g(y) + Dg(y)^T y - Dg(y)^T \nabla f(g(y)) \\ &= g(y) + Dg(y)^T y - Dg(y)^T y \\ &= g(y) \end{aligned}$$

and

$$\nabla^2 f^*(y) = Dg(y) = \nabla^2 f(g(y))^{-1}.$$

In particular,

$$\nabla f^*(\bar{y}) = \bar{x}, \qquad \nabla^2 f^*(\bar{y}) = \nabla^2 f(\bar{x})^{-1}.$$

3.41 *Domain of conjugate function.* Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is a twice differentiable convex function and $x \in \mathbf{dom}\, f$. Show that for small enough u we have

$$y = \nabla f(x) + \nabla^2 f(x) u \in \mathbf{dom}\, f^*,$$

i.e., $y^T x - f(x)$ is bounded above. It follows that $\dim(\mathbf{dom}\, f^*) \geq \mathbf{rank}\, \nabla^2 f(x)$.

Hint. Consider $\nabla f(x + tv)$, where t is small, and v is any vector in $\mathbf{R}^n$.

Solution. Clearly $\nabla f(x) \in \mathbf{dom}\, f^*$, since $\nabla f(x)$ maximizes $\nabla f(x)^T z - f(z)$ over z. Let $v \in \mathbf{R}^n$. For t small enough, we have $x + tv \in \mathbf{dom}\, f$, and therefore $w(t) = \nabla f(x + tv) \in \mathbf{dom}\, f^*$, since $x + tv$ maximizes $w(t)^T z - f(z)$ over z. Thus, $w(t) = \nabla f(x + tv)$ defines a curve (or just a point), passing through $\nabla f(x)$, that lies in $\mathbf{dom}\, f^*$. The tangent to the curve at $\nabla f(x)$ is given by

$$w'(0) = \frac{d}{dt} \nabla f(x + tv) \Big|_{t=0} = \nabla^2 f(x) v.$$

Now in general, the tangent to a curve that lies in a convex set must lie in the linear part of the affine hull of the set, since it is a limit of (scaled) differences of points in the set. (Differences of two points in a convex set lie in the linear part of its affine hull.) It follows that for s small enough, we have $\nabla f(x) + s\nabla^2 f(x)v \in \mathbf{dom}\, f^*$.
Examples:

- $f = a^T x + b$ linear: $\mathbf{dom}\, f^* = \{a\}$.
- functions with $\mathbf{dom}\, f^* = \mathbf{R}^n$
- $f = \log \sum \exp(x)$: $\mathbf{dom}\, f^* = \{y \succeq 0 \mid \mathbf{1}^T y = 1\}$ and

$$\nabla^2 f(x) = -(1/\mathbf{1}^T z)^2 z z^T + (1/\mathbf{1}^T z)\,\mathbf{diag}(z))$$

  where $\mathbf{1}^T z = 1$.
- $f = x^T P x + q^T x + r$: $\mathbf{dom}\, f^* = q + \mathcal{R}(P)$

Quasiconvex functions

3.42 *Approximation width.* Let $f_0, \ldots, f_n : \mathbf{R} \to \mathbf{R}$ be given continuous functions. We consider the problem of approximating f_0 as a linear combination of $f_1, \ldots, f_n$. For $x \in \mathbf{R}^n$, we say that $f = x_1 f_1 + \cdots + x_n f_n$ approximates f_0 with tolerance $\epsilon > 0$ over the interval $[0, T]$ if $|f(t) - f_0(t)| \leq \epsilon$ for $0 \leq t \leq T$. Now we choose a fixed tolerance $\epsilon > 0$ and define the *approximation width* as the largest T such that f approximates f_0 over the interval $[0, T]$:

$$W(x) = \sup\{T \mid |x_1 f_1(t) + \cdots + x_n f_n(t) - f_0(t)| \leq \epsilon \text{ for } 0 \leq t \leq T\}.$$

Show that W is quasiconcave.
Solution. To show that W is quasiconcave we show that the sets $\{x \mid W(x) \geq \alpha\}$ are convex for all α. We have $W(x) \geq \alpha$ if and only if

$$-\epsilon \leq x_1 f_1(t) + \cdots + x_n f_n(t) - f_0(t) \leq \epsilon$$

for all $t \in [0, \alpha)$. Therefore the set $\{x \mid W(x) \geq \alpha\}$ is an intersection of infinitely many halfspaces (two for each t), hence a convex set.

3.43 *First-order condition for quasiconvexity.* Prove the first-order condition for quasiconvexity given in §3.4.3: A differentiable function $f : \mathbf{R}^n \to \mathbf{R}$, with $\mathbf{dom}\, f$ convex, is quasiconvex if and only if for all $x, y \in \mathbf{dom}\, f$,

$$f(y) \leq f(x) \Longrightarrow \nabla f(x)^T (y - x) \leq 0.$$

Hint. It suffices to prove the result for a function on $\mathbf{R}$; the general result follows by restriction to an arbitrary line.
Solution. First suppose f is a differentiable function on $\mathbf{R}$ and satisfies

$$f(y) \leq f(x) \Longrightarrow f'(x)(y - x) \leq 0. \tag{3.43.A}$$

Suppose $f(x_1) \geq f(x_2)$ where $x_1 \neq x_2$. We assume $x_2 > x_1$ (the other case can be handled similarly), and show that $f(z) \leq f(x_1)$ for $z \in [x_1, x_2]$. Suppose this is false, *i.e.*, there exists a $z \in [x_1, x_2]$ with $f(z) > f(x_1)$. Since f is differentiable, we can choose a z that also satisfies $f'(z) < 0$. By (3.43.A), however, $f(x_1) < f(z)$ implies $f'(z)(x_1 - z) \leq 0$, which contradicts $f'(z) < 0$.
To prove sufficiency, assume f is quasiconvex. Suppose $f(x) \geq f(y)$. By the definition of quasiconvexity $f(x + t(y - x)) \leq f(x)$ for $0 < t \leq 1$. Dividing both sides by t, and taking the limit for $t \to 0$, we obtain

$$\lim_{t \to 0} \frac{f(x + t(y - x)) - f(x)}{t} = f'(x)(y - x) \leq 0,$$

which proves (3.43.A).

3.44 *Second-order conditions for quasiconvexity.* In this problem we derive alternate representations of the second-order conditions for quasiconvexity given in §3.4.3. Prove the following.

(a) A point $x \in \mathbf{dom}\, f$ satisfies (3.21) if and only if there exists a σ such that

$$\nabla^2 f(x) + \sigma \nabla f(x) \nabla f(x)^T \succeq 0. \tag{3.26}$$

It satisfies (3.22) for all $y \neq 0$ if and only if there exists a σ such

$$\nabla^2 f(x) + \sigma \nabla f(x) \nabla f(x)^T \succ 0. \tag{3.27}$$

Hint. We can assume without loss of generality that $\nabla^2 f(x)$ is diagonal.

(b) A point $x \in \mathbf{dom}\, f$ satisfies (3.21) if and only if either $\nabla f(x) = 0$ and $\nabla^2 f(x) \succeq 0$, or $\nabla f(x) \neq 0$ and the matrix

$$H(x) = \begin{bmatrix} \nabla^2 f(x) & \nabla f(x) \\ \nabla f(x)^T & 0 \end{bmatrix}$$

has exactly one negative eigenvalue. It satisfies (3.22) for all $y \neq 0$ if and only if $H(x)$ has exactly one nonpositive eigenvalue.

Hint. You can use the result of part (a). The following result, which follows from the eigenvalue interlacing theorem in linear algebra, may also be useful: If $B \in \mathbf{S}^n$ and $a \in \mathbf{R}^n$, then

$$\lambda_n \left(\begin{bmatrix} B & a \\ a^T & 0 \end{bmatrix} \right) \geq \lambda_n(B).$$

Solution.

(a) We prove the equivalence of (3.21) and (3.26). If $\nabla f(x) = 0$, both conditions reduce to $\nabla^2 f(x) \succeq 0$, and they are obviously equivalent. We prove the result for $\nabla f(x) \neq 0$.

To simplify the proof, we adopt the following notation. Let $a \in \mathbf{R}^n$, $a \neq 0$, and $B \in \mathbf{S}^n$. We show that

$$a^T x = 0 \Longrightarrow x^T B x \geq 0 \tag{3.44.A}$$

if and only if there exists a σ such that $B + \sigma a a^T \succeq 0$.

It is obvious that the condition is sufficient: if $B + \sigma a a^T \succeq 0$, then

$$a^T x = 0 \Longrightarrow x^T B x = x^T (B + \sigma a a^T) x \geq 0.$$

Conversely, suppose (3.44.A) holds for all y. Without loss of generality we can assume that B is diagonal, $B = \mathbf{diag}(b)$, with the elements of b sorted in decreasing order ($b_1 \geq b_2 \geq \cdots \geq b_n$). We know that

$$a^T x = 0 \Longrightarrow \sum_{i=1}^n b_i x_i^2 \geq 0.$$

If $b_n \geq 0$, there is nothing to prove: $\mathbf{diag}(b) + \sigma a a^T \succeq 0$ for all $\sigma \geq 0$.

Suppose $b_n < 0$. Then we must have $a_n \neq 0$. (Otherwise, $x = e_n$ would satisfy $a^T x = 0$ and $x^T \mathbf{diag}(b) x = b_n < 0$, a contradiction.) Moreover, we must have $b_{n-1} \geq 0$. Otherwise, the vector x with

$$x_1 = \cdots = x_{n-2} = 0, \qquad x_{n-1} = 1, \qquad x_n = -a_{n-1}/a_n,$$

would satisfy $a^T x = 0$ and $x^T \mathbf{diag}(b)x = b_{n-1} + b_n(a_{n-1}/a_n)^2 < 0$, which is a contradiction. In summary,

$$a_n \neq 0, \qquad b_n < 0, \qquad b_1 \geq \cdots \geq b_{n-1} \geq 0. \tag{3.44.B}$$

We can derive conditions on σ guaranteeing that

$$C = \mathbf{diag}(b) + \sigma a a^T \succeq 0.$$

Define $\bar{a} = (a_1, \ldots, a_{n-1})$, $\bar{b} = (b_1, \ldots, b_{n-1})$. We have $C_{nn} = b_n + \sigma a_n^2 > 0$ if $\sigma > -b_n/a_n^2$. The Schur complement of C_{nn} is

$$\mathbf{diag}(\bar{b}) + \sigma \bar{a}\bar{a}^T - \frac{a_n^2}{b_n + \sigma a_n^2}\bar{a}\bar{a}^T = \mathbf{diag}(\bar{b}) + \frac{a_n^2\sigma^2 + b_n\sigma - a_n^2}{b_n + \sigma a_n^2}\bar{a}\bar{a}^T$$

and is positive semidefinite if if $a_n^2\sigma^2 + b_n\sigma - a_n^2 \geq 0$, *i.e.*,

$$\sigma \geq \frac{-b_n}{2a_n^2} + \sqrt{\frac{b_n^2}{4a_n^4} + 1}.$$

Next, we prove the equivalence of (3.22) and (3.27). We need to show that

$$a^T x = 0 \Longrightarrow x^T B x > 0 \tag{3.44.C}$$

if and only if there exists a σ such that $B + \sigma a a^T \succ 0$.

Again, it is obvious that the condition is sufficient: if $B + \sigma a a^T \succ 0$, then

$$a^T x = 0 \Longrightarrow x^T B x = x^T (B + \sigma a a^T) x > 0.$$

for all nonzero x.

Conversely, suppose (3.44.C) holds for all $x \neq 0$. We use the same notation as above and assume B is diagonal. If $b_n > 0$ there is nothing to prove. If $b_n \leq 0$, we must have $a_n \neq 0$ and $b_{n-1} > 0$. Indeed, if $b_{n-1} \leq 0$, choosing

$$x_1 = \cdots = x_{n-2} = 0, \qquad x_{n-1} = 1, \qquad x_n = -a_{n-1}/a_n$$

would provide a vector with $a^T x = 0$ and $x^T B x \leq 0$. Therefore,

$$a_n \neq 0, \qquad b_n \leq 0, \qquad b_1 \geq \cdots \geq b_{n-1} > 0. \tag{3.44.D}$$

We can now proceed as in the proof above and construct a σ satisfying $B + \sigma a a^T \succ 0$.

(b) We first consider (3.21). If $\nabla f(x) = 0$, both conditions reduce to $\nabla^2 f(x) \succeq 0$, so they are obviously equivalent. We prove the result for $\nabla f(x) \neq 0$. We use the same notation as in part (a), and consider the matrix

$$C = \begin{bmatrix} B & a \\ a^T & 0 \end{bmatrix} \in \mathbf{S}^{n+1}$$

with $a \neq 0$. We need to show that C has exactly one negative eigenvalue if and only if (3.44.A) holds, or equivalently, if and only if there exists a σ such that $B + \sigma a a^T \succeq 0$.

We first note that C has at least one negative eigenvalue: the vector $v = (a, t)$ with $t < a^T B a/(2\|a\|_2^2)$ satisfies

$$v^T C v = a^T B a + 2t a^T a < 0.$$

Assume that C has exactly one negative eigenvalue. Suppose (3.44.A) does not hold, *i.e.*, there exists an x satisfying $a^T x = 0$ and $x^T B x < 0$. The vector $u = (x, 0)$ satisfies

$$u^T C u = u^T B u < 0.$$

We also note that u is orthogonal to the vector v defined above. So we have two orthogonal vectors u and v with $u^T C u < 0$ and $v^T C v < 0$, which contradicts our assumption that C has only one negative eigenvalue.

Conversely, suppose (3.44.A) holds, or, equivalently, $B + \sigma a a^T \succeq 0$ for some σ. Define

$$C(\sigma) = \begin{bmatrix} I & \sqrt{\sigma} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} B & a \\ a^T & 0 \end{bmatrix} \begin{bmatrix} I & 0 \\ \sqrt{\sigma} & 1 \end{bmatrix} = \begin{bmatrix} B + \sigma a a^T & a \\ a^T & 0 \end{bmatrix}.$$

Since $B+\sigma a a^T \succeq 0$, it follows from the hint that $\lambda_n(C(\sigma)) \geq 0$, *i.e.*, $C(\sigma)$ has exactly one negative eigenvalue. Since the inertia of a symmetric matrix is preserved under a congruence, C has exactly one negative eigenvalue.

The equivalence of (3.21) and (3.26) follows similarly. Note that if $\nabla f(x) = 0$, both conditions reduce to $\nabla^2 f(x) \succ 0$. If $\nabla f(x) \neq 0$, $H(x)$ has at least one negative eigenvalue, and we need to show that the other eigenvalues are positive.

3.45 Use the first and second-order conditions for quasiconvexity given in §3.4.3 to verify quasiconvexity of the function $f(x) = -x_1 x_2$, with $\mathbf{dom}\, f = \mathbf{R}^2_{++}$.

Solution. The first and second derivatives of f are

$$\nabla f(x) = \begin{bmatrix} -x_2 \\ -x_1 \end{bmatrix}, \qquad \nabla^2 f(x) = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}.$$

We start with the first-order condition

$$f(x) \leq f(y) \Longrightarrow \nabla f(x)^T (y - x) \leq 0,$$

which in this case reduces to

$$-y_1 y_2 \leq -x_1 x_2 \Longrightarrow -x_2(y_1 - x_1) - x_1(y_2 - x_2) \leq 0$$

for $x, y \succ 0$. Simplifying each side we get

$$y_1 y_2 \geq x_1 x_2 \Longrightarrow 2x_1 x_2 \leq x_1 y_2 + x_2 y_1,$$

and dividing by $x_1 x_2$ (which is positive) we get the equivalent statement

$$(y_1/x_1)(y_2/x_2) \geq 1 \Longrightarrow 1 \leq \left((y_2/x_2) + (y_1/x_1)\right)/2,$$

which is true (it is the arithmetic-geometric mean inequality).

The second-order condition is

$$y^T \nabla f(x) = 0,\ y \neq 0 \Longrightarrow y^T \nabla^2 f(x) y > 0,$$

which reduces to

$$-y_1 x_2 - y_2 x_1 = 0,\ y \neq 0 \Longrightarrow -2y_1 y_2 > 0$$

for $x \succ 0$, *i.e.*,

$$y_2 = -y_1 x_2 / x_1 \Longrightarrow -2y_1 y_2 > 0,$$

which is correct if $x \succ 0$.

3.46 *Quasilinear functions with domain* $\mathbf{R}^n$. A function on $\mathbf{R}$ that is quasilinear (*i.e.*, quasiconvex and quasiconcave) is monotone, *i.e.*, either nondecreasing or nonincreasing. In this problem we consider a generalization of this result to functions on $\mathbf{R}^n$.

Suppose the function $f : \mathbf{R}^n \to \mathbf{R}$ is quasilinear and continuous with $\mathbf{dom}\, f = \mathbf{R}^n$. Show that it can be expressed as $f(x) = g(a^T x)$, where $g : \mathbf{R} \to \mathbf{R}$ is monotone and $a \in \mathbf{R}^n$. In other words, a quasilinear function with domain $\mathbf{R}^n$ must be a monotone function of a linear function. (The converse is also true.)

Solution. The sublevel set $\{x \mid f(x) \leq \alpha\}$ are closed and convex (note that f is continuous), and their complements $\{x \mid f(x) > \alpha\}$ are also convex. Therefore the sublevel sets are closed halfspaces, and can be expressed as

$$\{x \mid f(x) \leq \alpha\} = \{x \mid a(\alpha)^T x \leq b(\alpha)\}$$

with $\|a(\alpha)\|_2 = 1$.

The sublevel sets are nested, *i.e.*, they have the same normal vector $a(\alpha) = a$ for all α, and $b(\alpha_1) \geq b(\alpha_2)$ if $\alpha_1 > \alpha_2$. In other words,

$$\{x \mid f(x) \leq \alpha\} = \{x \mid a^T x \leq b(\alpha)\}$$

where b is nondecreasing. If b is in fact increasing, we can define $g = b^{-1}$ and say that

$$\{x \mid f(x) \leq \alpha\} = \{x \mid g(a^T x) \leq \alpha\}$$

and by continuity of f, $f(x) = g(a^T x)$. If b is merely nondecreasing, we define

$$g(t) = \sup\{\alpha \mid b(\alpha) \leq t\}.$$

Log-concave and log-convex functions

3.47 Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is differentiable, $\mathbf{dom}\, f$ is convex, and $f(x) > 0$ for all $x \in \mathbf{dom}\, f$. Show that f is log-concave if and only if for all $x, y \in \mathbf{dom}\, f$,

$$\frac{f(y)}{f(x)} \leq \exp\left(\frac{\nabla f(x)^T (y - x)}{f(x)}\right).$$

Solution. This is the basic inequality

$$h(y) \geq h(x) + \nabla h(x)^T (y - x)$$

applied to the convex function $h(x) = -\log f(x)$, combined with $\nabla h(x) = (1/f(x))\nabla f(x)$.

3.48 Show that if $f : \mathbf{R}^n \to \mathbf{R}$ is log-concave and $a \geq 0$, then the function $g = f - a$ is log-concave, where $\mathbf{dom}\, g = \{x \in \mathbf{dom}\, f \mid f(x) > a\}$.

Solution. We have for $x, y \in \mathbf{dom}\, f$ with $f(x) > a$, $f(y) > a$, and $0 \leq \theta \leq 1$,

$$\begin{aligned} f(\theta x + (1-\theta)y) - a &\geq f(x)^\theta f(y)^{1-\theta)} - a \\ &\geq (f(x) - a)^\theta (f(y) - a)^{1-\theta}. \end{aligned}$$

The last inequality follows from Hölder's inequality

$$u_1 v_1 + u_2 v_2 \leq (u_1^{1/\theta} + u_2^{1/\theta})^\theta (v_1^{1/(1-\theta)} + v_2^{1/(1-\theta)})^{1-\theta},$$

applied to

$$u_1 = (f(x) - a)^\theta, \qquad v_1 = (f(y) - a)^{1-\theta}, \qquad u_2 = a^\theta, \qquad v_2 = a^{1-\theta},$$

which yields

$$f(x)^\theta f(y)^{1-\theta} \geq (f(x) - a)^\theta (f(y) - a)^{1-\theta} + a.$$

3.49 Show that the following functions are log-concave.

(a) *Logistic function:* $f(x) = e^x/(1+e^x)$ with $\mathbf{dom}\, f = \mathbf{R}$.
Solution. We have

$$\log(e^x/(1+e^x)) = x - \log(1+e^x).$$

The first term is linear, hence concave. Since the function $\log(1+e^x)$ is convex (it is the log-sum-exp function, evaluated at $x_1 = 0$, $x_2 = x$), the second term above is concave. Thus, $e^x/(1+e^x)$ is log-concave.

(b) *Harmonic mean:*

$$f(x) = \frac{1}{1/x_1 + \cdots + 1/x_n}, \qquad \mathbf{dom}\, f = \mathbf{R}^n_{++}.$$

Solution. The first and second derivatives of

$$h(x) = \log f(x) = -\log(1/x_1 + \cdots + 1/x_n)$$

are

$$\begin{aligned}
\frac{\partial h(x)}{\partial x_i} &= \frac{1/x_i^2}{1/x_1 + \cdots + 1/x_n} \\
\frac{\partial^2 h(x)}{\partial x_i^2} &= \frac{-2/x_i^3}{1/x_1 + \cdots + 1/x_n} + \frac{1/x_i^4}{(1/x_1 + \cdots + 1/x_n)^2} \\
\frac{\partial^2 h(x)}{\partial x_i \partial x_j} &= \frac{1/(x_i^2 x_j^2)}{(1/x_1 + \cdots + 1/x_n)^2} \qquad (i \neq j).
\end{aligned}$$

We show that $y^T \nabla^2 h(x) y \prec 0$ for all $y \neq 0$, *i.e.*,

$$(\sum_{i=1}^n y_i/x_i^2)^2 < 2(\sum_{i=1}^n 1/x_i)(\sum_{i=1}^n y_i^2/x_i^3)$$

This follows from the Cauchy-Schwarz inequality $(a^T b)^2 \leq \|a\|_2^2 \|b\|_2^2$, applied to

$$a_i = \frac{1}{\sqrt{x}_i}, \qquad b_i = \frac{y_i}{x_i \sqrt{x}_i}.$$

(c) *Product over sum:*

$$f(x) = \frac{\prod_{i=1}^n x_i}{\sum_{i=1}^n x_i}, \qquad \mathbf{dom}\, f = \mathbf{R}^n_{++}.$$

Solution. We must show that

$$f(x) = \sum_{i=1}^n \log x_i - \log \sum_{i=1}^n x_i$$

is concave on $x \succ 0$. Let's consider a line described by $x + tv$, where and $x,\ v \in \mathbf{R}^n$ and $x \succ 0$: define

$$\tilde{f}(t) = \sum_i \log(x_i + tv_i) - \log \sum_i (x_i + tv_i).$$

The first derivative is

$$\tilde{f}'(t) = \sum_i \frac{v_i}{x_i + tv_i} - \frac{\mathbf{1}^T v}{\mathbf{1}^T x + t\mathbf{1}^T v},$$

and the second derivative is

$$\tilde{f}''(t) = -\sum_i \frac{v_i^2}{(x_i + tv_i)^2} + \frac{(\mathbf{1}^T v)^2}{(\mathbf{1}^T x + t\mathbf{1}^T v)^2}.$$

Therefore to establish concavity of f, we need to show that

$$\tilde{f}''(0) = -\sum_i \frac{v_i^2}{x_i^2} + \frac{(\mathbf{1}^T v)^2}{(\mathbf{1}^T x)^2} \leq 0$$

holds for all v, and all $x \succ 0$.
The inequality holds if $\mathbf{1}^T v = 0$. If $\mathbf{1}^T v \neq 0$, we note that the inequality is homogeneous of degree two in v, so we can assume without loss of generality that $\mathbf{1}^T v = \mathbf{1}^T x$. This reduces the problem to verifying that

$$\sum_i \frac{v_i^2}{x_i^2} \geq 1$$

holds whenever $x \succ 0$ and $\mathbf{1}^T v = \mathbf{1}^T x$.
To establish this, let's fix x, and minimize the convex, quadratic form over $\mathbf{1}^T v = \mathbf{1}^T x$. The optimality conditions give

$$\frac{v_i}{x_i^2} = \lambda,$$

so we have $v_i = \lambda x_i^2$. From $\mathbf{1}^T v = \mathbf{1}^T x$ we can obtain λ, which gives

$$v_i^\star = \frac{\sum_k x_k}{\sum_k x_k^2} x_i^2.$$

Therefore the minimum value of $\sum_i v_i^2/x_i^2$ over $\mathbf{1}^T v = \mathbf{1}^T x$ is

$$\sum_i (\frac{v_i^\star}{x_i})^2 = \left(\frac{\sum_k x_k}{\sum_k x_k^2}\right)^2 \sum_i x_i^2 = \left(\frac{\mathbf{1}^T x}{\|x\|_2}\right)^2 \geq 1,$$

because $\|x\|_2 \leq \|x\|_1$. This proves the inequality.

(d) *Determinant over trace:*

$$f(X) = \frac{\det X}{\mathbf{tr}\, X}, \qquad \mathbf{dom}\, f = \mathbf{S}^n_{++}.$$

Solution. We prove that

$$h(X) = \log f(X) = \log\det X - \log \mathbf{tr}\, X$$

is concave. Consider the restriction on a line $X = Z + tV$ with $Z \succ 0$, and use the eigenvalue decomposition $Z^{-1/2} V Z^{-1/2} = Q\Lambda Q^T = \sum_{i=1}^n \lambda_i q_i q_i^T$:

$$\begin{aligned}
h(Z+tV) &= \log\det(Z+tV) - \log \mathbf{tr}(Z+tV) \\
&= \log\det Z - \log\det(I + tZ^{-1/2}VZ^{-1/2}) - \log \mathbf{tr}\, Z(I + tZ^{-1/2}VZ^{1/2}) \\
&= \log\det Z - \sum_{i=1}^n \log(1+t\lambda_i) - \log \sum_{i=1}^n (q_i^T Z q_i)(1+t\lambda_i)) \\
&= \log\det Z + \sum_{i=1}^n \log(q_i^T Z q_i) - \sum_{i=1}^n \log((q_i^T Z q_i)(1+t\lambda_i)) \\
&\quad - \log \sum_{i=1}^n ((q_i^T Z q_i)(1+t\lambda_i)),
\end{aligned}$$

which is a constant, plus the function

$$\sum_{i=1}^{n} \log y_i - \log \sum_{i=1}^{n} y_i$$

(which is concave; see (c)), evaluated at $y_i = (q_i^T Z q_i)(1 + t\lambda_i)$.

3.50 *Coefficients of a polynomial as a function of the roots.* Show that the coefficients of a polynomial with real negative roots are log-concave functions of the roots. In other words, the functions $a_i : \mathbf{R}^n \to \mathbf{R}$, defined by the identity

$$s^n + a_1(\lambda)s^{n-1} + \cdots + a_{n-1}(\lambda)s + a_n(\lambda) = (s - \lambda_1)(s - \lambda_2)\cdots(s - \lambda_n),$$

are log-concave on $-\mathbf{R}^n_{++}$.

Hint. The function

$$S_k(x) = \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} x_{i_1} x_{i_2} \cdots x_{i_k},$$

with $\mathbf{dom}\, S_k \in \mathbf{R}^n_+$ and $1 \le k \le n$, is called the kth elementary symmetric function on $\mathbf{R}^n$. It can be shown that $S_k^{1/k}$ is concave (see [ML57]).

Solution. The coefficients are given by $a_k(\lambda) = S_k(-\lambda)$. The result follows from the hint, because the logarithm of a nonnegative concave function is log-concave.

3.51 [BL00, page 41] Let p be a polynomial on $\mathbf{R}$, with all its roots real. Show that it is log-concave on any interval on which it is positive.

Solution. We assume the polynomial has the form

$$p(x) = \alpha(x - s_1)(x - s_2)\ldots(x - s_n),$$

with $s_1 \le s_2 \le \cdots \le s_n$, and $\alpha > 0$. (The case $\alpha < 0$ can be handled similarly).

Suppose p is positive on the interval (s_k, s_{k+1}), which means $n - k$ (the number of roots to the right of the interval) must be even. We can write $\log p$ as

$$\begin{aligned}\log p(x) = \log \alpha &+ \sum_{i=1}^{n} \log(x - s_k) \\ &+ \log((x - s_{k+1})(x - s_{k+2})) \\ &+ \log((x - s_{k+3})(x - s_{k+4})) \\ &+ \cdots + \log((x - s_{n-1})(x - s_n)).\end{aligned}$$

The first terms are obviously concave. We need to show that

$$f(x) = \log((x - a)(x - b)) = \log(x^2 - (a + b)x + ab)$$

is concave if $x < a \le b$. We have

$$f'(x) = \frac{2x - (a + b)}{x^2 - (a + b) + ab}, \qquad f''(x) = \frac{2(x - a)(x - b) - (2x - (a + b))^2}{(x^2 - (a + b)x + ab)^2}.$$

It is easily shown that the second derivative is less than or equal to zero:

$$\begin{aligned}&2(x - a)(x - b) - ((x - a) + (x - b))^2 \\ &\quad\le\quad 2(x - a)(x - b) - (x - a)^2 - (x - b)^2 - 2(x - a)(x - b) \\ &\quad=\quad -(x - a)^2 - (x - b)^2 \\ &\quad\le\quad 0.\end{aligned}$$

3.52 [MO79, §3.E.2] *Log-convexity of moment functions.* Suppose $f : \mathbf{R} \to \mathbf{R}$ is nonnegative with $\mathbf{R}_+ \subseteq \mathbf{dom}\, f$. For $x \geq 0$ define

$$\phi(x) = \int_0^\infty u^x f(u)\, du.$$

Show that ϕ is a log-convex function. (If x is a positive integer, and f is a probability density function, then $\phi(x)$ is the xth moment of the distribution.)

Use this to show that the Gamma function,

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u}\, du,$$

is log-convex for $x \geq 1$.

Solution. $g(x, u) = u^x f(u)$ is log-convex (as well as log-concave) in x for all $u > 0$. It follows directly from the property on page 106 that

$$\phi(x) = \int_0^\infty g(x, u)\, du = \int_0^\infty u^x f(u)\, du$$

is log-convex.

3.53 Suppose x and y are independent random vectors in $\mathbf{R}^n$, with log-concave probability density functions f and g, respectively. Show that the probability density function of the sum $z = x + y$ is log-concave.

Solution. The probability density function of $x + y$ is $f * g$.

3.54 *Log-concavity of Gaussian cumulative distribution function.* The cumulative distribution function of a Gaussian random variable,

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2}\, dt,$$

is log-concave. This follows from the general result that the convolution of two log-concave functions is log-concave. In this problem we guide you through a simple self-contained proof that f is log-concave. Recall that f is log-concave if and only if $f''(x)f(x) \leq f'(x)^2$ for all x.

(a) Verify that $f''(x)f(x) \leq f'(x)^2$ for $x \geq 0$. That leaves us the hard part, which is to show the inequality for $x < 0$.

(b) Verify that for any t and x we have $t^2/2 \geq -x^2/2 + xt$.

(c) Using part (b) show that $e^{-t^2/2} \leq e^{x^2/2 - xt}$. Conclude that

$$\int_{-\infty}^x e^{-t^2/2}\, dt \leq e^{x^2/2} \int_{-\infty}^x e^{-xt}\, dt.$$

(d) Use part (c) to verify that $f''(x)f(x) \leq f'(x)^2$ for $x \leq 0$.

Solution. The derivatives of f are

$$f'(x) = e^{-x^2/2}/\sqrt{2\pi}, \qquad f''(x) = -xe^{-x^2/2}/\sqrt{2\pi}.$$

(a) $f''(x) \leq 0$ for $x \geq 0$.

(b) Since $t^2/2$ is convex we have

$$t^2/2 \geq x^2/2 + x(t - x) = xt - x^2/2.$$

This is the general inequality

$$g(t) \geq g(x) + g'(x)(t - x),$$

which holds for any differentiable convex function, applied to $g(t) = t^2/2$.

(c) Take exponentials and integrate.

(d) This basic inequality reduces to

$$-xe^{-x^2/2}\int_{-\infty}^{x} e^{-t^2/2}\,dt \le e^{-x^2}$$

i.e.,

$$\int_{-\infty}^{x} e^{-t^2/2}\,dt \le \frac{e^{-x^2/2}}{-x}.$$

This follows from part (c) because

$$\int_{-\infty}^{x} e^{-xt}\,dt = \frac{e^{-x^2}}{-x}.$$

3.55 *Log-concavity of the cumulative distribution function of a log-concave probability density.* In this problem we extend the result of exercise 3.54. Let $g(t) = \exp(-h(t))$ be a differentiable log-concave probability density function, and let

$$f(x) = \int_{-\infty}^{x} g(t)\,dt = \int_{-\infty}^{x} e^{-h(t)}\,dt$$

be its cumulative distribution. We will show that f is log-concave, *i.e.*, it satisfies $f''(x)f(x) \le (f'(x))^2$ for all x.

(a) Express the derivatives of f in terms of the function h. Verify that $f''(x)f(x) \le (f'(x))^2$ if $h'(x) \ge 0$.

(b) Assume that $h'(x) < 0$. Use the inequality

$$h(t) \ge h(x) + h'(x)(t-x)$$

(which follows from convexity of h), to show that

$$\int_{-\infty}^{x} e^{-h(t)}\,dt \le \frac{e^{-h(x)}}{-h'(x)}.$$

Use this inequality to verify that $f''(x)f(x) \le (f'(x))^2$ if $h'(x) \ge 0$.

Solution.

(a) $f(x) = \int_{-\infty}^{x} e^{-h(t)}\,dt$, $f'(x) = e^{-h(x)}$, $f''(x) = -h'(x)e^{-h(x)}$. Log-concavity means

$$-h'(x)e^{-h(x)}\int_{-\infty}^{x} e^{-h(t)}\,dt \le e^{-2h(x)},$$

which is obviously true if $-h'(x) \le 0$.

(b) Take exponentials and integrate both sides of $-h(t) \le -h(x) - h'(x)(t-x)$:

$$\begin{aligned}
\int_{-\infty}^{x} e^{-h(t)}\,dt &\le e^{xh'(x)-h(x)}\int_{-\infty}^{x} e^{-th'(x)}\,dt \\
&= e^{xh'(x)-h(x)}e^{-xh'(x)}/(-h'(x)) \\
&= \frac{e^{-h(x)}}{-h'(x)} \\
(-h'(x))\int_{-\infty}^{x} e^{-h(t)}\,dt &\le e^{-h(x)}.
\end{aligned}$$

3.56 *More log-concave densities.* Show that the following densities are log-concave.

(a) [MO79, page 493] The *gamma density*, defined by

$$f(x) = \frac{\alpha^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\alpha x},$$

with $\mathbf{dom}\, f = \mathbf{R}_+$. The parameters λ and α satisfy $\lambda \geq 1$, $\alpha > 0$.
Solution.

$$\log f(x) = \log((\alpha^\lambda/\Gamma(\lambda)) + (\lambda - 1)\log x - \alpha x.$$

(b) [MO79, page 306] The *Dirichlet density*

$$f(x) = \frac{\Gamma(\mathbf{1}^T\lambda)}{\Gamma(\lambda_1)\cdots\Gamma(\lambda_{n+1})} x_1^{\lambda_1-1}\cdots x_n^{\lambda_n-1}\left(1 - \sum_{i=1}^n x_i\right)^{\lambda_{n+1}-1}$$

with $\mathbf{dom}\, f = \{x \in \mathbf{R}^n_{++} \mid \mathbf{1}^T x < 1\}$. The parameter λ satisfies $\lambda \succeq \mathbf{1}$.
Solution.

$$\begin{aligned}\log f(x) &= \log(\Gamma(\lambda)/(\Gamma(\lambda_1)\cdots\Gamma(\lambda_{n+1}))) + \sum_{i=1}^n (\lambda_i - 1)\log x_i + (\lambda_{n+1} - 1)\log(1 - \mathbf{1}^T x).\end{aligned}$$

Convexity with respect to a generalized inequality

3.57 Show that the function $f(X) = X^{-1}$ is matrix convex on $\mathbf{S}^n_{++}$.
Solution. We must show that for arbitrary $v \in \mathbf{R}^n$, the function

$$g(X) = v^T X^{-1} v.$$

is convex in X on $\mathbf{S}^n_{++}$. This follows from example 3.4.

3.58 *Schur complement.* Suppose $X \in \mathbf{S}^n$ partitioned as

$$X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix},$$

where $A \in \mathbf{S}^k$. The *Schur complement* of X (with respect to A) is $S = C - B^T A^{-1} B$ (see §A.5.5). Show that the Schur complement, viewed as function from $\mathbf{S}^n$ into $\mathbf{S}^{n-k}$, is matrix concave on $\mathbf{S}^n_{++}$.
Solution. Let $v \in \mathbf{R}^{n-k}$. We must show that the function

$$v^T(C - B^T A^{-1} B)v$$

is concave in X on $\mathbf{S}^n_{++}$. This follows from example 3.4.

3.59 *Second-order conditions for K-convexity.* Let $K \subseteq \mathbf{R}^m$ be a proper convex cone, with associated generalized inequality $\preceq_K$. Show that a twice differentiable function $f : \mathbf{R}^n \to \mathbf{R}^m$, with convex domain, is K-convex if and only if for all $x \in \mathbf{dom}\, f$ and all $y \in \mathbf{R}^n$,

$$\sum_{i,j=1}^n \frac{\partial^2 f(x)}{\partial x_i \partial x_j} y_i y_j \succeq_K 0,$$

i.e., the second derivative is a K-nonnegative bilinear form. (Here $\partial^2 f/\partial x_i \partial x_j \in \mathbf{R}^m$, with components $\partial^2 f_k/\partial x_i \partial x_j$, for $k = 1, \ldots, m$; see §A.4.1.)

Solution. f is K-convex if and only if $v^T f$ is convex for all $v \preceq_{K^*} 0$. The Hessian of $v^T f(x)$ is

$$\nabla^2(v^T f(x)) = \sum_{k=1}^{n} v_i \nabla^2 f_k(x).$$

This is positive semidefinite if and only if for all y

$$y^T \nabla^2(v^T f(x))y = \sum_{i,j=1}^{n}\sum_{k=1}^{n} v_k \nabla^2 f_k(x) y_i y_j = \sum_{k=1}^{n} v_k (\sum_{i,j=1}^{n} \nabla^2 f_k(x) y_i y_j) \geq 0,$$

which is equivalent to

$$\sum_{i,j=1}^{n} \nabla^2 f_k(x) y_i y_j \succeq_K 0$$

by definition of dual cone.

3.60 *Sublevel sets and epigraph of K-convex functions.* Let $K \subseteq \mathbf{R}^m$ be a proper convex cone with associated generalized inequality $\preceq_K$, and let $f : \mathbf{R}^n \to \mathbf{R}^m$. For $\alpha \in \mathbf{R}^m$, the α-sublevel set of f (with respect to $\preceq_K$) is defined as

$$C_\alpha = \{x \in \mathbf{R}^n \mid f(x) \preceq_K \alpha\}.$$

The epigraph of f, with respect to $\preceq_K$, is defined as the set

$$\mathbf{epi}_K f = \{(x,t) \in \mathbf{R}^{n+m} \mid f(x) \preceq_K t\}.$$

Show the following:

(a) If f is K-convex, then its sublevel sets C_α are convex for all α.

(b) f is K-convex if and only if $\mathbf{epi}_K f$ is a convex set.

Solution.

(a) For any $x, y \in C_\alpha$, and $0 \leq \theta \leq 1$,

$$f(\theta x + (1-\theta)y) \preceq_K \theta f(x) + (1-\theta) f(y) \preceq_K \alpha.$$

(b) For any $(x,u),(y,v) \in \mathbf{epi}\, f$, and $0 \leq \theta \leq 1$,

$$f(\theta x + (1-\theta)y) \preceq_K \theta f(x) + (1-\theta) f(y) \preceq_K \theta u + (1-\theta) v.$$

Chapter 4

Convex optimization problems

Exercises

Basic terminology and optimality conditions

4.1 Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & f_0(x_1, x_2) \\ \text{subject to} & 2x_1 + x_2 \geq 1 \\ & x_1 + 3x_2 \geq 1 \\ & x_1 \geq 0, \quad x_2 \geq 0. \end{array}$$

Make a sketch of the feasible set. For each of the following objective functions, give the optimal set and the optimal value.

(a) $f_0(x_1, x_2) = x_1 + x_2$.

(b) $f_0(x_1, x_2) = -x_1 - x_2$.

(c) $f_0(x_1, x_2) = x_1$.

(d) $f_0(x_1, x_2) = \max\{x_1, x_2\}$.

(e) $f_0(x_1, x_2) = x_1^2 + 9x_2^2$.

Solution. The feasible set is the convex hull of $(0, \infty)$, $(0, 1)$, $(2/5, 1/5)$, $(1, 0)$, $(\infty, 0)$.

(a) $x^\star = (2/5, 1/5)$.

(b) Unbounded below.

(c) $X_{\rm opt} = \{(0, x_2) \mid x_2 \geq 1\}$.

(d) $x^\star = (1/3, 1/3)$.

(e) $x^\star = (1/2, 1/6)$. This is optimal because it satisfies $2x_1 + x_2 = 7/6 > 1$, $x_1 + 3x_2 = 1$, and

$$\nabla f_0(x^\star) = (1, 3)$$

is perpendicular to the line $x_1 + 3x_2 = 1$.

4.2 Consider the optimization problem

$$\text{minimize} \quad f_0(x) = -\sum_{i=1}^m \log(b_i - a_i^T x)$$

with domain $\mathbf{dom}\, f_0 = \{x \mid Ax \prec b\}$, where $A \in \mathbf{R}^{m \times n}$ (with rows a_i^T). We assume that $\mathbf{dom}\, f_0$ is nonempty.
Prove the following facts (which include the results quoted without proof on page 141).

(a) $\mathbf{dom}\, f_0$ is unbounded if and only if there exists a $v \neq 0$ with $Av \preceq 0$.

(b) f_0 is unbounded below if and only if there exists a v with $Av \preceq 0$, $Av \neq 0$. *Hint.* There exists a v such that $Av \preceq 0$, $Av \neq 0$ if and only if there exists no $z \succ 0$ such that $A^T z = 0$. This follows from the theorem of alternatives in example 2.21, page 50.

(c) If f_0 is bounded below then its minimum is attained, *i.e.*, there exists an x that satisfies the optimality condition (4.23).

(d) The optimal set is affine: $X_{\rm opt} = \{x^\star + v \mid Av = 0\}$, where $x^\star$ is any optimal point.

Solution. We assume $x_0 \in \mathbf{dom}\, f$.

(a) If such a v exists, then $\mathbf{dom}\, f_0$ is clearly unbounded, since $x_0 + tv \in \mathbf{dom}\, f_0$ for all $t \geq 0$.
Conversely, suppose x^k is a sequence of points in $\mathbf{dom}\, f_0$ with $\|x^k\|_2 \to \infty$. Define $v^k = x^k / \|x^k\|_2$. The sequence has a convergent subsequence because $\|v^k\|_2 = 1$ for all k. Let v be its limit. We have $\|v\|_2 = 1$ and, since $a_i^T v^k < b_i / \|x^k\|_2$ for all k, $a_i^T v \leq 0$. Therefore $Av \preceq 0$ and $v \neq 0$.

(b) If there exists such a v, then f_0 is clearly unbounded below. Let j be an index with $a_j^T v < 0$. For $t \geq 0$,

$$\begin{aligned} f_0(x_0 + tv) &= -\sum_{i=1}^m \log(b_i - a_i^T x_0 - t a_i^T v) \\ &\leq -\sum_{i \neq j} \log(b_i - a_i^T x_0) - \log(b_j - a_j^T x_0 - t a_j^T v), \end{aligned}$$

and the righthand side decreases without bound as t increases.
Conversely, suppose f is unbounded below. Let x^k be a sequence with $b - Ax^k \succ 0$, and $f_0(x^k) \to -\infty$. By convexity,

$$f_0(x^k) \geq f_0(x_0) + \sum_{i=1}^m \frac{1}{b_i - a_i^T x_0} a_i^T (x^k - x_0) = f_0(x_0) + m - \sum_{i=1}^m \frac{b_i - a_i^T x^k}{b_i - a_i^T x_0}$$

so if $f_0(x^k) \to -\infty$, we must have $\max_i (b_i - a_i^T x^k) \to \infty$.
Suppose there exists a z with $z \succ 0$, $A^T z = 0$. Then

$$z^T b = z^T (b - Ax^k) \geq z_i \max_i (b_i - a_i^T x^k) \to \infty.$$

We have reached a contradiction, and conclude that there is no such z. Using the theorem of alternatives, there must be a v with $Av \preceq 0$, $Av \neq 0$.

(c) We can assume that $\mathbf{rank}\, A = n$.
If $\mathbf{dom}\, f_0$ is bounded, then the result follows from the fact that the sublevel sets of f_0 are closed.
If $\mathbf{dom}\, f_0$ is unbounded, let v be a direction in which it is unbounded, *i.e.*, $v \neq 0$, $Av \preceq 0$. Since $\mathbf{rank}\, A = 0$, we must have $Av \neq 0$, but this implies f_0 is unbounded. We conclude that if $\mathbf{rank}\, A = n$, then f_0 is bounded below if and only if its domain is bounded, and therefore its minimum is attained.

(d) Again, we can limit ourselves to the case in which $\mathbf{rank}\, A = n$. We have to show that f_0 has at most one optimal point. The Hessian of f_0 at x is

$$\nabla^2 f(x) = A^T \mathbf{diag}(d) A, \qquad d_i = \frac{1}{(b_i - a_i^T x)^2}, \quad i = 1, \ldots, m,$$

which is positive definite if $\mathbf{rank}\, A = n$, *i.e.*, f_0 is strictly convex. Therefore the optimal point, if it exists, is unique.

4.3 Prove that $x^\star = (1, 1/2, -1)$ is optimal for the optimization problem

$$\begin{array}{ll} \text{minimize} & (1/2) x^T P x + q^T x + r \\ \text{subject to} & -1 \leq x_i \leq 1, \quad i = 1, 2, 3, \end{array}$$

where

$$P = \begin{bmatrix} 13 & 12 & -2 \\ 12 & 17 & 6 \\ -2 & 6 & 12 \end{bmatrix}, \qquad q = \begin{bmatrix} -22.0 \\ -14.5 \\ 13.0 \end{bmatrix}, \qquad r = 1.$$

Solution. We verify that $x^\star$ satisfies the optimality condition (4.21). The gradient of the objective function at $x^\star$ is

$$\nabla f_0(x^\star) = (-1, 0, 2).$$

Therefore the optimality condition is that

$$\nabla f_0(x^\star)^T (y - x) = -1(y_1 - 1) + 2(y_2 + 1) \geq 0$$

for all y satisfying $-1 \leq y_i \leq 1$, which is clearly true.

4.4 [P. Parrilo] *Symmetries and convex optimization.* Suppose $\mathcal{G} = \{Q_1, \ldots, Q_k\} \subseteq \mathbf{R}^{n\times n}$ is a group, *i.e.*, closed under products and inverse. We say that the function $f : \mathbf{R}^n \to \mathbf{R}$ is $\mathcal{G}$-*invariant*, or *symmetric with respect to* $\mathcal{G}$, if $f(Q_i x) = f(x)$ holds for all x and $i = 1, \ldots, k$. We define $\overline{x} = (1/k)\sum_{i=1}^k Q_i x$, which is the average of x over its $\mathcal{G}$-orbit. We define the *fixed subspace* of $\mathcal{G}$ as

$$\mathcal{F} = \{x \mid Q_i x = x,\ i = 1, \ldots, k\}.$$

(a) Show that for any $x \in \mathbf{R}^n$, we have $\overline{x} \in \mathcal{F}$.

(b) Show that if $f : \mathbf{R}^n \to \mathbf{R}$ is convex and $\mathcal{G}$-invariant, then $f(\overline{x}) \leq f(x)$.

(c) We say the optimization problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \end{array}$$

is $\mathcal{G}$-invariant if the objective f_0 is $\mathcal{G}$-invariant, and the feasible set is $\mathcal{G}$-invariant, which means

$$f_1(x) \leq 0, \ldots, f_m(x) \leq 0 \implies f_1(Q_i x) \leq 0, \ldots, f_m(Q_i x) \leq 0,$$

for $i = 1, \ldots, k$. Show that if the problem is convex and $\mathcal{G}$-invariant, and there exists an optimal point, then there exists an optimal point in $\mathcal{F}$. In other words, we can adjoin the equality constraints $x \in \mathcal{F}$ to the problem, without loss of generality.

(d) As an example, suppose f is convex and symmetric, *i.e.*, $f(Px) = f(x)$ for every permutation P. Show that if f has a minimizer, then it has a minimizer of the form $\alpha\mathbf{1}$. (This means to minimize f over $x \in \mathbf{R}^n$, we can just as well minimize $f(t\mathbf{1})$ over $t \in \mathbf{R}$.)

Solution.

(a) $Q_j\overline{x} = (1/k)\sum_{i=1}^k Q_j Q_i x \in \mathcal{F}$, because for each $Q_l \in \mathcal{G}$ there exists a $Q_i \in \mathcal{G}$ s.t. $Q_j Q_i = Q_l$.

(b) Using convexity and invariance of f,

$$f(\overline{x}) \leq (1/k)\sum_{i=1}^k f(Q_i x) = (1/k)\sum_{i=1}^k f(x) = f(x).$$

(c) Suppose $x^\star$ is an optimal solution. Then $\overline{x^\star}$ is feasible, with

$$\begin{aligned} f_0(\overline{x^\star}) &= f_0((1/k)\sum_{i=1}^k Q_i x) \\ &\leq (1/k)\sum_{i=1}^k f_0(Q_i x) \\ &= f_0(x^\star). \end{aligned}$$

Therefore $\overline{x^\star}$ is also optimal.

(d) Suppose $x^\star$ is a minimizer of f. Let $\overline{x} = (1/n!)\sum_P Px^\star$, where the sum is over all permutations. Since $\overline{x}$ is invariant under any permutation, we conclude that $\overline{x} = \alpha\mathbf{1}$ for some $\alpha \in \mathbf{R}$. By Jensen's inequality we have

$$f(\overline{x}) \leq (1/n!)\sum_P f(Px^\star) = f(x^\star),$$

which shows that $\overline{x}$ is also a minimizer.

4.5 *Equivalent convex problems.* Show that the following three convex problems are equivalent. Carefully explain how the solution of each problem is obtained from the solution of the other problems. The problem data are the matrix $A \in \mathbf{R}^{m\times n}$ (with rows a_i^T), the vector $b \in \mathbf{R}^m$, and the constant $M > 0$.

(a) The *robust least-squares problem*

$$\text{minimize} \quad \sum_{i=1}^m \phi(a_i^T x - b_i),$$

with variable $x \in \mathbf{R}^n$, where $\phi : \mathbf{R} \to \mathbf{R}$ is defined as

$$\phi(u) = \begin{cases} u^2 & |u| \le M \\ M(2|u| - M) & |u| > M. \end{cases}$$

(This function is known as the *Huber penalty function*; see §6.1.2.)

(b) The *least-squares problem with variable weights*

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m (a_i^T x - b_i)^2/(w_i + 1) + M^2 \mathbf{1}^T w \\ \text{subject to} & w \succeq 0, \end{array}$$

with variables $x \in \mathbf{R}^n$ and $w \in \mathbf{R}^m$, and domain $\mathcal{D} = \{(x, w) \in \mathbf{R}^n \times \mathbf{R}^m \mid w \succ -\mathbf{1}\}$.

Hint. Optimize over w assuming x is fixed, to establish a relation with the problem in part (a).

(This problem can be interpreted as a weighted least-squares problem in which we are allowed to adjust the weight of the ith residual. The weight is one if $w_i = 0$, and decreases if we increase w_i. The second term in the objective penalizes large values of w, *i.e.*, large adjustments of the weights.)

(c) The *quadratic program*

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m (u_i^2 + 2Mv_i) \\ \text{subject to} & -u - v \preceq Ax - b \preceq u + v \\ & 0 \preceq u \preceq M\mathbf{1} \\ & v \succeq 0. \end{array}$$

Solution.

(a) *Problems (a) and (b).* For fixed u, the solution of the minimization problem

$$\begin{array}{ll} \text{minimize} & u^2/(w + 1) + M^2 w \\ \text{subject to} & w \succeq 0 \end{array}$$

is given by

$$w = \begin{cases} |u|/M - 1 & |u| \ge M \\ 0 & \text{otherwise.} \end{cases}$$

($w = 0|u|/M-1$ is the unconstrained minimizer of the objective function. If $|u|/M - 1 \ge 0$ it is the optimum. Otherwise $w = 0$ is the optimum.) The optimal value is

$$\inf_{w\succeq 0} \left(u^2/(w + 1) + M^2 w\right) = \begin{cases} M(2|u| - M) & |u| \ge M \\ u^2 & \text{otherwise.} \end{cases}$$

It follows that the optimal value of x in both problems is the same. The optimal w in the second problem is given by

$$w_i = \begin{cases} |a_i^T x - b_i|/M - 1 & |a_i^T x - b_i| \ge M \\ 0 & \text{otherwise.} \end{cases}$$

(b) *Problems (a) and (c).* Suppose we fix x in problem (c).

First we note that at the optimum we must have $u_i + v_i = |a_i^T x - b_i|$. Otherwise, *i.e.*, if u_i, v_i satisfy $u_i + v_i > |a_i^T x + b_i$ with $0 \leq u_i \leq M$ and $v_i \geq 0$, then, since u_i and v_i are not both zero, we can decrease u_i and/or v_i without violating the constraints. This also decreases the objective.

At the optimum we therefore have

$$v_i = |a_i^T x - b_i| - u_i.$$

Eliminating v yields the equivalent problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m (u_i^2 - 2Mu_i + 2M|a_i^T x - b_i|) \\ \text{subject to} & 0 \leq u_i \leq \min\{M, |a_i^T x - b_i|\} \end{array}$$

If $|a_i^T x - b_i| \leq M$, the optimal choice for u_i is $u_i = |a_i^T x - b_i|$. In this case the ith term in the objective function reduces to $|a_i^T x - b_i|$. If $|a_i^T x - b_i| > M$, we choose $u_i = M$, and the ith term in the objective function reduces to $2M|a_i^T x - b_i| - M^2$. We conclude that, for fixed x, the optimal value of the problem in (c) is given by

$$\sum_{i=1}^m \phi(a_i^T x - b_i).$$

4.6 *Handling convex equality constraints.* A convex optimization problem can have only *linear* equality constraint functions. In some special cases, however, it is possible to handle convex equality constraint functions, *i.e.*, constraints of the form $g(x) = 0$, where g is convex. We explore this idea in this problem.

Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h(x) = 0, \end{array} \tag{4.65}$$

where f_i and h are convex functions with domain $\mathbf{R}^n$. Unless h is affine, this is *not* a convex optimization problem. Consider the related problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & h(x) \leq 0, \end{array} \tag{4.66}$$

where the convex equality constraint has been relaxed to a convex inequality. This problem is, of course, convex.

Now suppose we can guarantee that at any optimal solution $x^\star$ of the convex problem (4.66), we have $h(x^\star) = 0$, *i.e.*, the inequality $h(x) \leq 0$ is always active at the solution. Then we can solve the (nonconvex) problem (4.65) by solving the convex problem (4.66).

Show that this is the case if there is an index r such that

- f_0 is monotonically increasing in x_r
- $f_1, \ldots, f_m$ are nonincreasing in x_r
- h is monotonically decreasing in x_r.

We will see specific examples in exercises 4.31 and 4.58.

Solution. Suppose $x^\star$ is optimal for the relaxed problem, and $h(x^\star) < 0$. By the last property, we can decrease x_r while staying in the boundary of g. By decreasing x_r we decrease the objective, preserve the inequalities $f_i(x) \leq 0$, and increase the function h.

4.7 *Convex-concave fractional problems.* Consider a problem of the form

$$\begin{array}{ll} \text{minimize} & f_0(x)/(c^T x + d) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & Ax = b \end{array}$$

where $f_0, f_1, \ldots, f_m$ are convex, and the domain of the objective function is defined as $\{x \in \mathbf{dom}\, f_0 \mid c^T x + d > 0\}$.

(a) Show that this is a quasiconvex optimization problem.

Solution. The domain of the objective is convex, because f_0 is convex. The sublevel sets are convex because $f_0(x)/(c^T x + d) \leq \alpha$ if and only if $c^T x + d > 0$ and $f_0(x) \leq \alpha(c^T x + d)$.

(b) Show that the problem is equivalent to

$$\begin{array}{ll} \text{minimize} & g_0(y, t) \\ \text{subject to} & g_i(y, t) \leq 0, \quad i = 1, \ldots, m \\ & Ay = bt \\ & c^T y + dt = 1, \end{array}$$

where g_i is the perspective of f_i (see §3.2.6). The variables are $y \in \mathbf{R}^n$ and $t \in \mathbf{R}$. Show that this problem is convex.

Solution. Suppose x is feasible in the original problem. Define $t = 1/(c^T x + d)$ (a positive number), $y = x/(c^T x + d)$. Then $t > 0$ and it is easily verified that t, y are feasible in the transformed problem, with the objective value $g_0(y, t) = f_0(x)/(c^T x + d)$.

Conversely, suppose y, t are feasible for the transformed problem. We must have $t > 0$, by definition of the domain of the perspective function. Define $x = y/t$. We have $x \in \mathbf{dom}\, f_i$ for $i = 0, \ldots, m$ (again, by definition of perspective). x is feasible in the original problem, because

$$f_i(x) = g_i(y, t)/t \leq 0, \quad i = 1, \ldots, m \qquad Ax = A(y/t) = b.$$

From the last equality, $c^T x + d = (c^T y + dt)/t = 1/t$, and hence,

$$t = 1/(c^T x + d), \qquad f_0(x)/(c^T x + d) = t f_0(x) = g_0(y, t).$$

Therefore x is feasible in the original problem, with the objective value $g_0(y, t)$.

In conclusion, from any feasible point of one problem we can derive a feasible point of the other problem, with the same objective value.

(c) Following a similar argument, derive a convex formulation for the *convex-concave* fractional problem

$$\begin{array}{ll} \text{minimize} & f_0(x)/h(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & Ax = b \end{array}$$

where $f_0, f_1, \ldots, f_m$ are convex, h is concave, the domain of the objective function is defined as $\{x \in \mathbf{dom}\, f_0 \cap \mathbf{dom}\, h \mid h(x) > 0\}$ and $f_0(x) \geq 0$ everywhere.

As an example, apply your technique to the (unconstrained) problem with

$$f_0(x) = (\mathbf{tr}\, F(x))/m, \qquad h(x) = (\det(F(x))^{1/m},$$

with $\mathbf{dom}(f_0/h) = \{x \mid F(x) \succ 0\}$, where $F(x) = F_0 + x_1 F_1 + \cdots + x_n F_n$ for given $F_i \in \mathbf{S}^m$. In this problem, we minimize the ratio of the arithmetic mean over the geometric mean of the eigenvalues of an affine matrix function $F(x)$.

Solution.

(a) We first verify that the problem is quasiconvex. The domain of the objective function is convex, and its sublevel sets are convex because for $\alpha \geq 0$, $f_0(x)/h(x) \leq \alpha$ if and only if $f_0(x) - \alpha h(x) \leq 0$, which is a convex inequality. For $\alpha < 0$, the sublevel sets are empty.

(b) The convex formulation is

$$\begin{array}{ll} \text{minimize} & g_0(y,t) \\ \text{subject to} & g_i(y,t) \leq 0, \quad i = 1, \ldots, m \\ & Ay = bt \\ & \tilde{h}(y,t) \leq -1 \end{array}$$

where g_i is the perspective of f_i and $\tilde{h}$ is the perspective of $-h$.
To verify the equivalence, assume first that x is feasible in the original problem. Define $t = 1/h(x)$ and $y = x/h(x)$. Then $t > 0$ and

$$g_i(y,t) = tf_i(y/t) = tf_i(x) \leq 0, \quad i = 1, \ldots, m, \qquad Ay = Ax/h(x) = bt.$$

Moreover, $\tilde{h}(y,t) = th(y/t) = h(x)/h(x) = 1$ and

$$g_0(y,t) = tf_0(y/t) = f_0(x)/h(x).$$

We see that for every feasible point in the original problem we can find a feasible point in the transformed problem, with the same objective value.
Conversely, assume y, t are feasible in the transformed problem. By definition of perspective, $t > 0$. Define $x = y/t$. We have

$$f_i(x) = f_i(y/t) = g_i(y,t)/t \leq 0, \quad i = 1, \ldots, m, \qquad Ax = A(y/t) = b.$$

From the last inequality, we have

$$\tilde{h}(y,t) = -th(y/t) = -th(x) \leq -1.$$

This implies that $h(x) > 0$ and $th(x) \geq 1$. And finally, the objective is

$$f_0(x)/h(x) = g_0(y,t)/(th(x)) \leq g_0(y,t).$$

We conclude that with every feasible point in the transformed problem there is a corresponding feasible point in the original problem with the same or lower objective value.
Putting the two parts together, we can conclude that the two problems have the same optimal value, and that optimal solutions for one problem are optimal for the other (if both are solvable).

(c)

$$\begin{array}{ll} \text{minimize} & (1/m)\,\mathbf{tr}(tF_0 + y_1F_1 + \cdots + y_nF_n) \\ \text{subject to} & \det(tF_0 + y_1F_1 + \cdots + y_nF_n)^{1/m} \geq 1 \end{array}$$

with domain

$$\{(y,t) \mid t > 0,\, tF_0 + y_1F_1 + \cdots + y_nF_n \succ 0\}.$$

Linear optimization problems

4.8 *Some simple LPs.* Give an explicit solution of each of the following LPs.

(a) *Minimizing a linear function over an affine set.*

$$\begin{array}{ll} \text{minimize} & c^Tx \\ \text{subject to} & Ax = b. \end{array}$$

Solution. We distinguish three possibilities.

- The problem is infeasible ($b \notin \mathcal{R}(A)$). The optimal value is ∞.
- The problem is feasible, and c is orthogonal to the nullspace of A. We can decompose c as
$$c = A^T\lambda + \hat{c}, \qquad A\hat{c} = 0.$$
($\hat{c}$ is the component in the nullspace of A; $A^T\lambda$ is orthogonal to the nullspace.) If $\hat{c} = 0$, then on the feasible set the objective function reduces to a constant:
$$c^T x = \lambda^T Ax + \hat{c}^T x = \lambda^T b.$$
The optimal value is $\lambda^T b$. All feasible solutions are optimal.
- The problem is feasible, and c is not in the range of A^T ($\hat{c} \neq 0$). The problem is unbounded ($p^\star = -\infty$). To verify this, note that $x = x_0 - t\hat{c}$ is feasible for all t; as t goes to infinity, the objective value decreases unboundedly.

In summary,
$$p^\star = \begin{cases} +\infty & b \notin \mathcal{R}(A) \\ \lambda^T b & c = A^T\lambda \text{ for some } \lambda \\ -\infty & \text{otherwise.} \end{cases}$$

(b) *Minimizing a linear function over a halfspace.*
$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & a^T x \leq b, \end{array}$$
where $a \neq 0$.

Solution. This problem is always feasible. The vector c can be decomposed into a component parallel to a and a component orthogonal to a:
$$c = a\lambda + \hat{c},$$
with $a^T\hat{c} = 0$.

- If $\lambda > 0$, the problem is unbounded below. Choose $x = -ta$, and let t go to infinity:
$$c^T x = -tc^T a = -t\lambda a^T a \to -\infty$$
and
$$a^T x - b = -ta^T a - b \leq 0$$
for large t, so x is feasible for large t. Intuitively, by going very far in the direction $-a$, we find feasible points with arbitrarily negative objective values.
- If $\hat{c} \neq 0$, the problem is unbounded below. Choose $x = ba - t\hat{c}$ and let t go to infinity.
- If $c = a\lambda$ for some $\lambda \leq 0$, the optimal value is $c^T ab = \lambda b$.

In summary, the optimal value is
$$p^\star = \begin{cases} \lambda b & c = a\lambda \text{ for some } \lambda \leq 0 \\ -\infty & \text{otherwise.} \end{cases}$$

(c) *Minimizing a linear function over a rectangle.*
$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & l \preceq x \preceq u, \end{array}$$
where l and u satisfy $l \preceq u$.

Solution. The objective and the constraints are separable: The objective is a sum of terms $c_i x_i$, each dependent on one variable only; each constraint depends on only one

variable. We can therefore solve the problem by minimizing over each component of x independently. The optimal $x_i^\star$ minimizes $c_i x_i$ subject to the constraint $l_i \leq x_i \leq u_i$. If $c_i > 0$, then $x_i^\star = l_i$; if $c_i < 0$, then $x_i^\star = u_i$; if $c_i = 0$, then any x_i in the interval $[l_i, u_i]$ is optimal. Therefore, the optimal value of the problem is

$$p^\star = l^T c^+ + u^T c^-,$$

where $c_i^+ = \max\{c_i, 0\}$ and $c_i^- = \max\{-c_i, 0\}$.

(d) *Minimizing a linear function over the probability simplex.*

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \mathbf{1}^T x = 1, \quad x \succeq 0. \end{array}$$

What happens if the equality constraint is replaced by an inequality $\mathbf{1}^T x \leq 1$?
We can interpret this LP as a simple portfolio optimization problem. The vector x represents the allocation of our total budget over different assets, with x_i the fraction invested in asset i. The return of each investment is fixed and given by $-c_i$, so our total return (which we want to maximize) is $-c^T x$. If we replace the budget constraint $\mathbf{1}^T x = 1$ with an inequality $\mathbf{1}^T x \leq 1$, we have the option of not investing a portion of the total budget.

Solution. Suppose the components of c are sorted in increasing order with

$$c_1 = c_2 = \cdots = c_k < c_{k+1} \leq \cdots \leq c_n.$$

We have

$$c^T x \geq c_1(\mathbf{1}^T x) = c_{\min}$$

for all feasible x, with equality if and only if

$$x_1 + \cdots + x_k = 1, \qquad x_1 \geq 0, \ldots, x_k \geq 0, \qquad x_{k+1} = \cdots = x_n = 0.$$

We conclude that the optimal value is $p^\star = c_1 = c_{\min}$. In the investment interpretation this choice is quite obvious. If the returns are fixed and known, we invest our total budget in the investment with the highest return.
If we replace the equality with an inequality, the optimal value is equal to

$$p^\star = \min\{0, c_{\min}\}.$$

(If $c_{\min} \leq 0$, we make the same choice for x as above. Otherwise, we choose $x = 0$.)

(e) *Minimizing a linear function over a unit box with a total budget constraint.*

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \mathbf{1}^T x = \alpha, \quad 0 \preceq x \preceq \mathbf{1}, \end{array}$$

where α is an integer between 0 and n. What happens if α is not an integer (but satisfies $0 \leq \alpha \leq n$)? What if we change the equality to an inequality $\mathbf{1}^T x \leq \alpha$?

Solution. We first consider the case of integer α. Suppose

$$c_1 \leq \cdots \leq c_{i-1} < c_i = \cdots = c_\alpha = \cdots = c_k < c_{k+1} \leq \cdots \leq c_n.$$

The optimal value is

$$c_1 + c_2 + \cdots + c_\alpha$$

i.e., the sum of the smallest α elements of c. x is optimal if and only if

$$x_1 = \cdots = x_{i-1} = 1, \qquad x_i + \cdots + x_k = \alpha - i + 1, \qquad x_{k+1} = \cdots = x_n = 0.$$

If α is not an integer, the optimal value is

$$p^\star = c_1 + c_2 + \cdots + c_{\lfloor\alpha\rfloor} + c_{1+\lfloor\alpha\rfloor}(\alpha - \lfloor\alpha\rfloor).$$

In the case of an inequality constraint $\mathbf{1}^T x \leq \alpha$, with α an integer between 0 and n, the optimal value is the sum of the α smallest nonpositive coefficients of c.

(f) *Minimizing a linear function over a unit box with a weighted budget constraint.*

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & d^T x = \alpha, \quad 0 \preceq x \preceq \mathbf{1}, \end{array}$$

with $d \succ 0$, and $0 \leq \alpha \leq \mathbf{1}^T d$.

Solution. We make a change of variables $y_i = d_i x_i$, and consider the problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n (c_i/d_i) y_i \\ \text{subject to} & \mathbf{1}^T x = \alpha, \quad 0 \preceq y \preceq d. \end{array}$$

Suppose the ratios c_i/d_i have been sorted in increasing order:

$$\frac{c_1}{d_1} \leq \frac{c_2}{d_2} \leq \cdots \leq \frac{c_n}{d_n}.$$

To minimize the objective, we choose

$$y_1 = d_1, \quad y_2 = d_2, \quad \ldots, \quad y_k = d_k,$$

$$y_{k+1} = \alpha - (d_1 + \cdots + d_k), \qquad y_{k+2} = \cdots = y_n = 0,$$

where $k = \max\{i \in \{1, \ldots, n\} \mid d_1 + \cdots + d_i \leq \alpha\}$ (and $k = 0$ if $d_1 > \alpha$). In terms of the original variables,

$$x_1 = \cdots = x_k = 1, \qquad x_{k+1} = (\alpha - (d_1 + \cdots + d_k))/d_{k+1}, \qquad x_{k+2} = \cdots = x_n = 0.$$

4.9 *Square LP.* Consider the LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \end{array}$$

with A square and nonsingular. Show that the optimal value is given by

$$p^\star = \begin{cases} c^T A^{-1} b & A^{-T} c \preceq 0 \\ -\infty & \text{otherwise.} \end{cases}$$

Solution. Make a change of variables $y = Ax$. The problem is equivalent to

$$\begin{array}{ll} \text{minimize} & c^T A^{-1} y \\ \text{subject to} & y \preceq b. \end{array}$$

If $A^{-T} c \preceq 0$, the optimal solution is $y = b$, with $p^\star = c^T A^{-1} b$. Otherwise, the LP is unbounded below.

4.10 *Converting general LP to standard form.* Work out the details on page 147 of §4.3. Explain in detail the relation between the feasible sets, the optimal solutions, and the optimal values of the standard form LP and the original LP.

Solution. Suppose x is feasible in (4.27). Define

$$x_i^+ = \min\{0, x_i\}, \qquad x_i^- = \min\{0, -x_i\}, \qquad s = h - Gx.$$

It is easily verified that x^+, x^-, s are feasible in the standard form LP, with objective value

$$c^T x^+ - c^T x^- + d = c^T x - d.$$

Hence, for each feasible point in (4.27) we can find a feasible point in the standard form LP with the same objective value. In particular, this implies that the optimal value of the standard form LP is less than or equal to the optimal value of (4.27).

Conversely, suppose x^+, x^-, s are feasible in the standard form LP. Define $x = x^+ - x^-$. It is clear that x is feasible for (4.27), with objective value $c^T x + d = c^T x^+ - c^T x^- + d$. Hence, for each feasible point in the standard form LP we can find a feasible point in (4.27) with the same objective value. This implies that the optimal value of the standard form LP is greater than or equal to the optimal value of (4.27).

We conclude that the optimal values are equal.

4.11 *Problems involving ℓ_1- and ℓ_∞-norms.* Formulate the following problems as LPs. Explain in detail the relation between the optimal solution of each problem and the solution of its equivalent LP.

(a) Minimize $\|Ax - b\|_\infty$ (ℓ_∞-norm approximation).

(b) Minimize $\|Ax - b\|_1$ (ℓ_1-norm approximation).

(c) Minimize $\|Ax - b\|_1$ subject to $\|x\|_\infty \leq 1$.

(d) Minimize $\|x\|_1$ subject to $\|Ax - b\|_\infty \leq 1$.

(e) Minimize $\|Ax - b\|_1 + \|x\|_\infty$.

In each problem, $A \in \mathbf{R}^{m\times n}$ and $b \in \mathbf{R}^m$ are given. (See §6.1 for more problems involving approximation and constrained approximation.)

Solution.

(a) Equivalent to the LP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & Ax - b \preceq t\mathbf{1} \\ & Ax - b \geq -t\mathbf{1}. \end{array}$$

in the variables x, t. To see the equivalence, assume x is fixed in this problem, and we optimize only over t. The constraints say that

$$-t \leq a_k^T x - b_k \leq t$$

for each k, *i.e.*, $t \geq |a_k^T x - b_k|$, *i.e.*,

$$t \geq \max_k |a_k^T x - b_k| = \|Ax - b\|_\infty.$$

Clearly, if x is fixed, the optimal value of the LP is $p^\star(x) = \|Ax - b\|_\infty$. Therefore optimizing over t and x simultaneously is equivalent to the original problem.

(b) Equivalent to the LP

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T s \\ \text{subject to} & Ax - b \preceq s \\ & Ax - b \geq -s. \end{array}$$

Assume x is fixed in this problem, and we optimize only over s. The constraints say that

$$-s_k \leq a_k^T x - b_k \leq s_k$$

for each k, *i.e.*, $s_k \geq |a_k^T x - b_k|$. The objective function of the LP is separable, so we achieve the optimum over s by choosing

$$s_k = |a_k^T x - b_k|,$$

and obtain the optimal value $p^\star(x) = \|Ax - b\|_1$. Therefore optimizing over t and s simultaneously is equivalent to the original problem.

(c) Equivalent to the LP

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T y \\ \text{subject to} & -y \preceq Ax - b \preceq y \\ & -\mathbf{1} \leq x \leq \mathbf{1}, \end{array}$$

with variables $x \in \mathbf{R}^n$ and $y \in \mathbf{R}^m$.

(d) Equivalent to the LP

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T y \\ \text{subject to} & -y \leq x \leq y \\ & -\mathbf{1} \leq Ax - b \leq \mathbf{1} \end{array}$$

with variables x and y.

Another good solution is to write x as the difference of two nonnegative vectors $x = x^+ - x^-$, and to express the problem as

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T x^+ + \mathbf{1}^T x^- \\ \text{subject to} & -\mathbf{1} \preceq Ax^+ - Ax^- - b \preceq \mathbf{1} \\ & x^+ \succeq 0, \quad x^- \succeq 0, \end{array}$$

with variables $x^+ \in \mathbf{R}^n$ and $x^- \in \mathbf{R}^n$.

(e) Equivalent to

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T y + t \\ \text{subject to} & -y \preceq Ax - b \preceq y \\ & -t\mathbf{1} \preceq x \preceq t\mathbf{1}, \end{array}$$

with variables x, y, and t.

4.12 *Network flow problem.* Consider a network of n nodes, with directed links connecting each pair of nodes. The variables in the problem are the flows on each link: x_{ij} will denote the flow from node i to node j. The cost of the flow along the link from node i to node j is given by $c_{ij}x_{ij}$, where c_{ij} are given constants. The total cost across the network is

$$C = \sum_{i,j=1}^{n} c_{ij}x_{ij}.$$

Each link flow x_{ij} is also subject to a given lower bound l_{ij} (usually assumed to be nonnegative) and an upper bound u_{ij}.

The external supply at node i is given by b_i, where $b_i > 0$ means an external flow enters the network at node i, and $b_i < 0$ means that at node i, an amount $|b_i|$ flows out of the network. We assume that $\mathbf{1}^T b = 0$, *i.e.*, the total external supply equals total external demand. At each node we have conservation of flow: the total flow into node i along links and the external supply, minus the total flow out along the links, equals zero.

The problem is to minimize the total cost of flow through the network, subject to the constraints described above. Formulate this problem as an LP.

Solution. This can be formulated as the LP

$$\begin{array}{ll} \text{minimize} & C = \sum_{i,j=1}^{n} c_{ij}x_{ij} \\ \text{subject to} & b_i + \sum_{j=1}^{n} x_{ij} - \sum_{j=1}^{n} x_{ji} = 0, \quad i = 1, \ldots, n \\ & l_{ij} \leq x_{ij} \leq u_{ij}. \end{array}$$

4.13 *Robust LP with interval coefficients.* Consider the problem, with variable $x \in \mathbf{R}^n$,

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \text{ for all } A \in \mathcal{A}, \end{array}$$

where $\mathcal{A} \subseteq \mathbf{R}^{m \times n}$ is the set

$$\mathcal{A} = \{A \in \mathbf{R}^{m \times n} \mid \bar{A}_{ij} - V_{ij} \leq A_{ij} \leq \bar{A}_{ij} + V_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n\}.$$

(The matrices $\bar{A}$ and V are given.) This problem can be interpreted as an LP where each coefficient of A is only known to lie in an interval, and we require that x must satisfy the constraints for all possible values of the coefficients.

Express this problem as an LP. The LP you construct should be efficient, *i.e.*, it should not have dimensions that grow exponentially with n or m.

Solution. The problem is equivalent to

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \bar{A}x + V|x| \preceq b \end{array}$$

where $|x| = (|x_1|, |x_2|, \ldots, |x_n|)$. This in turn is equivalent to the LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \bar{A}x + Vy \preceq b \\ & -y \preceq x \preceq y \end{array}$$

with variables $x \in \mathbf{R}^n$, $y \in \mathbf{R}^n$.

4.14 *Approximating a matrix in infinity norm.* The ℓ_∞-norm induced norm of a matrix $A \in \mathbf{R}^{m\times n}$, denoted $\|A\|_\infty$, is given by

$$\|A\|_\infty = \sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}|.$$

This norm is sometimes called the max-row-sum norm, for obvious reasons (see §A.1.5).

Consider the problem of approximating a matrix, in the max-row-sum norm, by a linear combination of other matrices. That is, we are given $k+1$ matrices $A_0, \ldots, A_k \in \mathbf{R}^{m\times n}$, and need to find $x \in \mathbf{R}^k$ that minimizes

$$\|A_0 + x_1 A_1 + \cdots + x_k A_k\|_\infty.$$

Express this problem as a linear program. Explain the significance of any extra variables in your LP. Carefully explain how your LP formulation solves this problem, *e.g.*, what is the relation between the feasible set for your LP and this problem?

Solution. The problem can be formulated as an LP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & -S \preceq_K A_0 + x_1 A_1 + \cdots + x_k a_k \preceq_K S \\ & S\mathbf{1} \preceq t\mathbf{1}, \end{array}$$

with variables $S \in \mathbf{R}^{m\times n}$, $t \in \mathbf{R}$ and $x \in \mathbf{R}^k$. The inequality $\preceq_K$ denotes componentwise inequality between matrices, *i.e.*, with respect to the cone

$$K = \{X \in \mathbf{R}^{m\times n} \mid X_{ij} \geq 0,\ i = 1, \ldots, m,\ j = 1 \ldots, n\}.$$

To see the equivalence, suppose x and S are feasible in the LP. The last constraint means that

$$t \geq \sum_{j=1}^{n} s_{ij}, \quad i = 1, \ldots, m,$$

so the optimal choice of t is

$$t = \max_i \sum_{j=1}^{n} S_{ij}.$$

This shows that the LP is equivalent to

$$\begin{array}{ll} \text{minimize} & \max_i(\sum_{j=1}^n S_{ij}) \\ \text{subject to} & -S \preceq_K A_0 + x_1A_1 + \cdots + x_ka_k \preceq_K S. \end{array}$$

Suppose x is given in this problem, and we optimize over S. The constraints in the LP state that

$$-S_{ij} \leq A(x)_{ij} \leq S_{ij},$$

(where $A(x) = A_0 + x_1A_1 + \cdots + x_kA_k$), and since the objective is monotone increasing in S_{ij}, the optimal choice for S_{ij} is

$$S_{ij} = |A(x)_{ij}|.$$

The problem is now reduced to the original problem

$$\text{minimize} \quad \max_{i=1,\ldots,m} \sum_{j=1}^n |A(x)_{ij}|.$$

4.15 *Relaxation of Boolean LP.* In a *Boolean linear program*, the variable x is constrained to have components equal to zero or one:

$$\begin{array}{ll} \text{minimize} & c^Tx \\ \text{subject to} & Ax \preceq b \\ & x_i \in \{0,1\}, \quad i = 1,\ldots,n. \end{array} \tag{4.67}$$

In general, such problems are very difficult to solve, even though the feasible set is finite (containing at most 2^n points).

In a general method called *relaxation*, the constraint that x_i be zero or one is replaced with the linear inequalities $0 \leq x_i \leq 1$:

$$\begin{array}{ll} \text{minimize} & c^Tx \\ \text{subject to} & Ax \preceq b \\ & 0 \leq x_i \leq 1, \quad i = 1,\ldots,n. \end{array} \tag{4.68}$$

We refer to this problem as the *LP relaxation* of the Boolean LP (4.67). The LP relaxation is far easier to solve than the original Boolean LP.

(a) Show that the optimal value of the LP relaxation (4.68) is a lower bound on the optimal value of the Boolean LP (4.67). What can you say about the Boolean LP if the LP relaxation is infeasible?

(b) It sometimes happens that the LP relaxation has a solution with $x_i \in \{0,1\}$. What can you say in this case?

Solution.

(a) The feasible set of the relaxation includes the feasible set of the Boolean LP. It follows that the Boolean LP is infeasible if the relaxation is infeasible, and that the optimal value of the relaxation is less than or equal to the optimal value of the Boolean LP.

(b) The optimal solution of the relaxation is also optimal for the Boolean LP.

4.16 *Minimum fuel optimal control.* We consider a linear dynamical system with state $x(t) \in \mathbf{R}^n$, $t = 0,\ldots,N$, and actuator or input signal $u(t) \in \mathbf{R}$, for $t = 0,\ldots,N-1$. The dynamics of the system is given by the linear recurrence

$$x(t+1) = Ax(t) + bu(t), \quad t = 0,\ldots,N-1,$$

where $A \in \mathbf{R}^{n\times n}$ and $b \in \mathbf{R}^n$ are given. We assume that the initial state is zero, *i.e.*, $x(0) = 0$.

The *minimum fuel optimal control problem* is to choose the inputs $u(0), \ldots, u(N-1)$ so as to minimize the total fuel consumed, which is given by

$$F = \sum_{t=0}^{N-1} f(u(t)),$$

subject to the constraint that $x(N) = x_{\rm des}$, where N is the (given) time horizon, and $x_{\rm des} \in \mathbf{R}^n$ is the (given) desired final or target state. The function $f : \mathbf{R} \to \mathbf{R}$ is the *fuel use map* for the actuator, and gives the amount of fuel used as a function of the actuator signal amplitude. In this problem we use

$$f(a) = \begin{cases} |a| & |a| \leq 1 \\ 2|a| - 1 & |a| > 1. \end{cases}$$

This means that fuel use is proportional to the absolute value of the actuator signal, for actuator signals between -1 and 1; for larger actuator signals the marginal fuel efficiency is half.

Formulate the minimum fuel optimal control problem as an LP.

Solution.

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T t \\ \text{subject to} & Hu = x_{\rm des} \\ & -y \preceq u \preceq y \\ & t \succeq y \\ & t \succeq 2y - \mathbf{1} \end{array}$$

where

$$H = \begin{bmatrix} A^{N-1}b & A^{N-2}b & \cdots & Ab & b \end{bmatrix}.$$

4.17 *Optimal activity levels.* We consider the selection of n nonnegative activity levels, denoted $x_1, \ldots, x_n$. These activities consume m resources, which are limited. Activity j consumes $A_{ij}x_j$ of resource i, where A_{ij} are given. The total resource consumption is additive, so the total of resource i consumed is $c_i = \sum_{j=1}^n A_{ij}x_j$. (Ordinarily we have $A_{ij} \geq 0$, *i.e.*, activity j consumes resource i. But we allow the possibility that $A_{ij} < 0$, which means that activity j actually *generates* resource i as a by-product.) Each resource consumption is limited: we must have $c_i \leq c_i^{\max}$, where $c_i^{\max}$ are given. Each activity generates revenue, which is a piecewise-linear concave function of the activity level:

$$r_j(x_j) = \begin{cases} p_j x_j & 0 \leq x_j \leq q_j \\ p_j q_j + p_j^{\rm disc}(x_j - q_j) & x_j \geq q_j. \end{cases}$$

Here $p_j > 0$ is the basic price, $q_j > 0$ is the quantity discount level, and $p_j^{\rm disc}$ is the quantity discount price, for (the product of) activity j. (We have $0 < p_j^{\rm disc} < p_j$.) The total revenue is the sum of the revenues associated with each activity, *i.e.*, $\sum_{j=1}^n r_j(x_j)$. The goal is to choose activity levels that maximize the total revenue while respecting the resource limits. Show how to formulate this problem as an LP.

Solution. The basic problem can be expressed as

$$\begin{array}{ll} \text{maximize} & \sum_{j=1}^n r_j(x_j) \\ \text{subject to} & x \succeq 0 \\ & Ax \preceq c^{\max}. \end{array}$$

This is a convex optimization problem since the objective is concave and the constraints are a set of linear inequalities. To transform it to an equivalent LP, we first express the revenue functions as

$$r_j(x_j) = \min\{p_j x_j, p_j q_j + p_j^{\rm disc}(x_j - q_j)\},$$

which holds since r_j is concave. It follows that $r_j(x_j) \geq u_j$ if and only if

$$p_j x_j \geq u_j, \qquad p_j q_j + p_j^{\text{disc}}(x_j - q_j) \geq u_j.$$

We can form an LP as

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^T u \\ \text{subject to} & x \succeq 0 \\ & Ax \preceq c^{\max} \\ & p_j x_j \geq u_j, \quad p_j q_j + p_j^{\text{disc}}(x_j - q_j) \geq u_j, \quad j = 1, \ldots, n, \end{array}$$

with variables x and u.

To show that this LP is equivalent to the original problem, let us fix x. The last set of constraints in the LP ensure that $u_i \leq r_i(x)$, so we conclude that for every feasible x, u in the LP, the LP objective is less than or equal to the total revenue. On the other hand, we can always take $u_i = r_i(x)$, in which case the two objectives are equal.

4.18 *Separating hyperplanes and spheres.* Suppose you are given two sets of points in $\mathbf{R}^n$, $\{v^1, v^2, \ldots, v^K\}$ and $\{w^1, w^2, \ldots, w^L\}$. Formulate the following two problems as LP feasibility problems.

(a) Determine a hyperplane that separates the two sets, *i.e.*, find $a \in \mathbf{R}^n$ and $b \in \mathbf{R}$ with $a \neq 0$ such that

$$a^T v^i \leq b, \quad i = 1, \ldots, K, \qquad a^T w^i \geq b, \quad i = 1, \ldots, L.$$

Note that we require $a \neq 0$, so you have to make sure that your formulation excludes the trivial solution $a = 0$, $b = 0$. You can assume that

$$\mathbf{rank} \begin{bmatrix} v^1 & v^2 & \cdots & v^K & w^1 & w^2 & \cdots & w^L \\ 1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 \end{bmatrix} = n + 1$$

(*i.e.*, the affine hull of the $K + L$ points has dimension n).

(b) Determine a sphere separating the two sets of points, *i.e.*, find $x_c \in \mathbf{R}^n$ and $R \geq 0$ such that

$$\|v^i - x_c\|_2 \leq R, \quad i = 1, \ldots, K, \qquad \|w^i - x_c\|_2 \geq R, \quad i = 1, \ldots, L.$$

(Here x_c is the center of the sphere; R is its radius.)

(See chapter 8 for more on separating hyperplanes, separating spheres, and related topics.)

Solution.

(a) The conditions

$$a^T v^i \leq b, \quad i = 1, \ldots, K, \qquad a^T w^i \geq b, \quad i = 1, \ldots, L$$

form a set of $K + L$ linear inequalities in the variables a, b, which we can write in matrix form as

$$Bx \succeq 0$$

where

$$B = \begin{bmatrix} -(v^1)^T & 1 \\ \vdots & \vdots \\ -(v^K)^T & 1 \\ -(w^1)^T & -1 \\ \vdots & \vdots \\ -(w^L)^T & -1 \end{bmatrix} \in \mathbf{R}^{(K+L)\times(n+1)}, \qquad x = \begin{bmatrix} a \\ b \end{bmatrix}.$$

We are interested in nonzero solutions of $Bx \succeq 0$.

The rank assumption implies that $\mathbf{rank}\, B = n+1$. Therefore, its nullspace contains only the zero vector, *i.e.*, $x \neq 0$ implies $Bx \neq 0$. We can force x to be nonzero by adding a constraint $\mathbf{1}^T Bx = 1$. (On the right hand side we could choose any other positive constraint instead of 1.) This forces at least one component of Bx to be positive. In other words we can find nonzero solution to $Bx \succeq 0$ by solving the LP feasibility problem

$$Bx \succeq 0, \qquad \mathbf{1}^T Bx = 1.$$

(b) We begin by writing the inequalities as

$$\begin{array}{l} \|v^i\|_2^2 - 2(v^i)^T x_c + \|x_c\|_2^2 \leq R^2, \quad i = 1, \ldots, K, \\ \|w^i\|_2^2 - 2(w^i)^T x_c + \|x_c\|_2^2 \geq R^2, \quad i = 1, \ldots, L. \end{array}$$

These inequalities are not linear in x_c and R. However, if we use as variables x_c and $\gamma = R^2 - \|x_c\|_2^2$, then they reduce to

$$\|v^i\|_2^2 - 2(v^i)^T x_c \leq \gamma, \quad i = 1, \ldots, K, \qquad \|w^i\|_2^2 - 2(w^i)^T x_c \geq \gamma, \quad i = 1, \ldots, L,$$

which is a set of linear inequalities in $x_c \in \mathbf{R}^n$ and $\gamma \in \mathbf{R}$. We can solve this feasibility problem for x_c and γ, and compute R as

$$R = \sqrt{\gamma + \|x_c\|_2^2}.$$

We can be certain that $\gamma + \|x_c\|^2 \geq 0$: If x_c and γ are feasible, then

$$\gamma + \|x_c\|_2^2 \geq \|v^i\|_2^2 - 2(v^i)^T x_c + \|x_c\|_2^2 = \|v^i - x_c\|_2^2 \geq 0.$$

4.19 Consider the problem

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|_1/(c^T x + d) \\ \text{subject to} & \|x\|_\infty \leq 1, \end{array}$$

where $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, $c \in \mathbf{R}^n$, and $d \in \mathbf{R}$. We assume that $d > \|c\|_1$, which implies that $c^T x + d > 0$ for all feasible x.

(a) Show that this is a quasiconvex optimization problem.

(b) Show that it is equivalent to the convex optimization problem

$$\begin{array}{ll} \text{minimize} & \|Ay - bt\|_1 \\ \text{subject to} & \|y\|_\infty \leq t \\ & c^T y + dt = 1, \end{array}$$

with variables $y \in \mathbf{R}^n$, $t \in \mathbf{R}$.

Solution.

(a) $f_0(x) \leq \alpha$ if and only if

$$\|Ax - b\|_1 - \alpha(c^T x + d) \leq 0,$$

which is a convex constraint.

(b) Suppose $\|x\|_\infty \leq 1$. We have $c^T x + d > 0$, because $d > \|c\|_1$. Define

$$y = x/(c^T x + d), \quad t = 1/(c^T x + d).$$

Then y and t are feasible in the convex problem with objective value

$$\|Ay - bt\|_1 = \|Ax - b\|_1/(c^T x + d).$$

Conversely, suppose y, t are feasible for the convex problem. We must have $t > 0$, since $t = 0$ would imply $y = 0$, which contradicts $c^T y + dt = 1$. Define

$$x = y/t.$$

Then $\|x\|_\infty \leq 1$, and $c^T x + d = 1/t$, and hence

$$\|Ax - b\|_1/(c^T x + d) = \|Ay - bt\|_1.$$

4.20 *Power assignment in a wireless communication system.* We consider n transmitters with powers $p_1, \ldots, p_n \geq 0$, transmitting to n receivers. These powers are the optimization variables in the problem. We let $G \in \mathbf{R}^{n\times n}$ denote the matrix of *path gains* from the transmitters to the receivers; $G_{ij} \geq 0$ is the path gain from transmitter j to receiver i. The *signal power* at receiver i is then $S_i = G_{ii}p_i$, and the *interference power* at receiver i is $I_i = \sum_{k\neq i} G_{ik}p_k$. The *signal to interference plus noise ratio*, denoted SINR, at receiver i, is given by $S_i/(I_i + \sigma_i)$, where $\sigma_i > 0$ is the (self-) noise power in receiver i. The objective in the problem is to maximize the minimum SINR ratio, over all receivers, *i.e.*, to maximize

$$\min_{i=1,\ldots,n} \frac{S_i}{I_i + \sigma_i}.$$

There are a number of constraints on the powers that must be satisfied, in addition to the obvious one $p_i \geq 0$. The first is a maximum allowable power for each transmitter, *i.e.*, $p_i \leq P_i^{\max}$, where $P_i^{\max} > 0$ is given. In addition, the transmitters are partitioned into groups, with each group sharing the same power supply, so there is a total power constraint for each group of transmitter powers. More precisely, we have subsets $K_1, \ldots, K_m$ of $\{1, \ldots, n\}$ with $K_1 \cup \cdots \cup K_m = \{1, \ldots, n\}$, and $K_j \cap K_l = 0$ if $j \neq l$. For each group K_l, the total associated transmitter power cannot exceed $P_l^{\text{gp}} > 0$:

$$\sum_{k\in K_l} p_k \leq P_l^{\text{gp}}, \quad l = 1, \ldots, m.$$

Finally, we have a limit $P_k^{\text{rc}} > 0$ on the total received power at each receiver:

$$\sum_{k=1}^{n} G_{ik}p_k \leq P_i^{\text{rc}}, \quad i = 1, \ldots, n.$$

(This constraint reflects the fact that the receivers will saturate if the total received power is too large.)

Formulate the SINR maximization problem as a generalized linear-fractional program.

Solution.

$$\begin{array}{ll} \text{minimize} & \max_i(\sum_{k\neq i} G_{ik}p_k + \sigma_i)/(G_{ii}p_i) \\ & 0 \leq p_i \leq P_i^{\max} \\ & \sum_{k\in K_l} p_k \leq P_l^{\text{gp}} \\ & \sum_{k=1}^{n} G_{ik}p_k \leq P_i^{\text{rc}} \end{array}$$

Quadratic optimization problems

4.21 *Some simple QCQPs.* Give an explicit solution of each of the following QCQPs.

(a) *Minimizing a linear function over an ellipsoid centered at the origin.*

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

where $A \in \mathbf{S}^n_{++}$ and $c \neq 0$. What is the solution if the problem is not convex ($A \notin \mathbf{S}^n_+$)?

Solution. If $A \succ 0$, the solution is

$$x^\star = -\frac{1}{\sqrt{c^T A^{-1} c}} A^{-1} c, \qquad p^\star = -\|A^{-1/2} c\|_2 = -\sqrt{c^T A^{-1} c}.$$

This can be shown as follows. We make a change of variables $y = A^{1/2} x$, and write $\tilde{c} = A^{-1/2} c$. With this new variable the optimization problem becomes

$$\begin{array}{ll} \text{minimize} & \tilde{c}^T y \\ \text{subject to} & y^T y \leq 1, \end{array}$$

i.e., we minimize a linear function over the unit ball. The answer is $y^\star = -\tilde{c}/\|\tilde{c}\|_2$. In the general case, we can make a change of variables based on the eigenvalue decomposition

$$A = Q\,\mathbf{diag}(\lambda) Q^T = \sum_{i=1}^n \lambda_i q_i q_i^T.$$

We define $y = Qx$, $b = Qc$, and express the problem as

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n b_i y_i \\ \text{subject to} & \sum_{i=1}^n \lambda_i y_i^2 \leq 1. \end{array}$$

If $\lambda_i > 0$ for all i, the problem reduces to the case we already discussed. Otherwise, we can distinguish several cases.

- $\lambda_n < 0$. The problem is unbounded below. By letting $y_n \to \pm\infty$, we can make any point feasible.
- $\lambda_n = 0$. If for some i, $b_i \neq 0$ and $\lambda_i = 0$, the problem is unbounded below.
- $\lambda_n = 0$, and $b_i = 0$ for all i with $\lambda_i = 0$. In this case we can reduce the problem to a smaller one with all $\lambda_i > 0$.

(b) *Minimizing a linear function over an ellipsoid.*

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & (x - x_c)^T A (x - x_c) \leq 1, \end{array}$$

where $A \in \mathbf{S}^n_{++}$ and $c \neq 0$.

Solution. We make a change of variables

$$y = A^{1/2}(x - x_c), \qquad x = A^{-1/2} y + x_c,$$

and consider the problem

$$\begin{array}{ll} \text{minimize} & c^T A^{-1/2} y + c^T x_c \\ \text{subject to} & y^T y \leq 1. \end{array}$$

The solution is

$$y^\star = -(1/\|A^{-1/2} c\|_2) A^{-1/2} c, \qquad x^\star = x_c - (1/\|A^{-1/2} c\|_2) A^{-1} c.$$

(c) *Minimizing a quadratic form over an ellipsoid centered at the origin.*

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

where $A \in \mathbf{S}^n_{++}$ and $B \in \mathbf{S}^n_+$. Also consider the nonconvex extension with $B \notin \mathbf{S}^n_+$. (See §B.1.)

Solution. If $B \succeq 0$, then the optimal value is obviously zero (since $x^T B x \geq 0$ for all x, with equality if $x = 0$).

In the general case, we use the following fact from linear algebra. The smallest eigenvalue of $B \in \mathbf{S}^n$, can be characterized as

$$\lambda_{\min}(B) = \inf_{x^T x = 1} x^T B x.$$

To solve the optimization problem

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

with $A \succ 0$, we make a change of variables $y = A^{1/2} x$. This is possible since $A \succ 0$, so $A^{1/2}$ is defined and nonsingular. In the new variables the problem becomes

$$\begin{array}{ll}\text{minimize} & y^T A^{-1/2} B A^{-1/2} y \\ \text{subject to} & y^T y \leq 1.\end{array}$$

If the constraint $y^T y \leq 1$ is active at the optimum ($y^T y = 1$), then the optimal value is

$$\lambda_{\min}(A^{-1/2} B A^{-1/2}),$$

by the result mentioned above. If $y^T y < 1$ at the optimum, then it must be at a point where the gradient of the objective function vanishes, *i.e.*, $By = 0$. In that case the optimal value is zero.

To summarize, the optimal value is

$$p^\star = \begin{cases} \lambda_{\min}(A^{-1/2} B A^{-1/2}) & \lambda_{\min}(A^{-1/2} B A^{-1/2}) \leq 0 \\ 0 & \text{otherwise.} \end{cases}$$

In the first case any (normalized) eigenvector of $A^{-1/2} B A^{-1/2}$ corresponding to the smallest eigenvalue is an optimal y. In the second case $y = 0$ is optimal.

4.22 Consider the QCQP

$$\begin{array}{ll}\text{minimize} & (1/2) x^T P x + q^T x + r \\ \text{subject to} & x^T x \leq 1,\end{array}$$

with $P \in \mathbf{S}^n_{++}$. Show that $x^\star = -(P + \lambda I)^{-1} q$ where $\lambda = \max\{0, \bar{\lambda}\}$ and $\bar{\lambda}$ is the largest solution of the nonlinear equation

$$q^T (P + \lambda I)^{-2} q = 1.$$

Solution. x is optimal if and only if

$$x^T x < 1, \qquad P x + q = 0$$

or

$$x^T x = 1, \qquad P x + q = -\lambda x$$

for some $\lambda \geq 0$. (Geometrically, either x is in the interior of the ball and the gradient vanishes, or x is on the boundary, and the negative gradient is parallel to the outward pointing normal.)

The algorithm goes as follows. First solve $Px = -q$. If the solution has norm less than or equal to one ($\|P^{-1}q\|_2 \leq 1$), it is optimal. Otherwise, from the optimality conditions, x must satisfy $\|x\|_2 = 1$ and $(P + \lambda)x = -q$ for some $\lambda \geq 0$. Define

$$f(\lambda) = \|(P + \lambda)^{-1}q\|_2^2 = \sum_{i=1}^{n} \frac{q_i^2}{(\lambda + \lambda_i)^2},$$

where $\lambda_i > 0$ are the eigenvalues of P. (Note that $P + \lambda I \succ 0$ for all $\lambda \geq 0$ because $P \succ 0$.) We have $f(0) = \|P^{-1}q\|_2^2 > 1$. Also f monotonically decreases to zero as $\lambda \to \infty$. Therefore the nonlinear equation $f(\lambda) = 1$ has exactly one nonnegative solution $\bar{\lambda}$. Solve for $\bar{\lambda}$. The optimal solution is $x^\star = -(P + \bar{\lambda}I)^{-1}q$.

4.23 *ℓ_4-norm approximation via QCQP.* Formulate the ℓ_4-norm approximation problem

$$\text{minimize} \quad \|Ax - b\|_4 = (\textstyle\sum_{i=1}^m (a_i^T x - b_i)^4)^{1/4}$$

as a QCQP. The matrix $A \in \mathbf{R}^{m\times n}$ (with rows a_i^T) and the vector $b \in \mathbf{R}^m$ are given.

Solution.

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m z_i^2 \\ \text{subject to} & a_i^T x - b_i = y_i, \;\; i = 1, \ldots, m \\ & y_i^2 \leq z_i, \;\; i = 1, \ldots, m \end{array}$$

4.24 *Complex ℓ_1-, ℓ_2- and ℓ_∞-norm approximation.* Consider the problem

$$\text{minimize} \quad \|Ax - b\|_p,$$

where $A \in \mathbf{C}^{m\times n}$, $b \in \mathbf{C}^m$, and the variable is $x \in \mathbf{C}^n$. The complex ℓ_p-norm is defined by

$$\|y\|_p = \left(\sum_{i=1}^m |y_i|^p \right)^{1/p}$$

for $p \geq 1$, and $\|y\|_\infty = \max_{i=1,\ldots,m} |y_i|$. For $p = 1$, 2, and ∞, express the complex ℓ_p-norm approximation problem as a QCQP or SOCP with real variables and data.

Solution.

(a) Minimizing $\|Ax - b\|_2$ is equivalent to minimizing its square. So, let us expand $\|Ax - b\|_2^2$ around the real and complex parts of $Ax - b$:

$$\begin{aligned} \|Ax - b\|_2^2 &= \|\Re(Ax - b)\|_2^2 + \|\Im(Ax - b)\|_2^2 \\ &= \|\Re A \Re x - \Im A \Im x - \Re b\|_2^2 + \|\Re A \Im x + \Im A \Re x - \Im b\|_2^2. \end{aligned}$$

If we define $z^T = [\Re x^T \; \Im x^T]$ as requested, then this becomes

$$\begin{aligned} \|Ax - b\|_2^2 &= \|[\Re A \;\; -\Im A]z - \Re b\|^2 + \|[\Im A \;\; \Re A]z - \Im b\|^2 \\ &= \left\| \begin{bmatrix} \Re A & -\Im A \\ \Im A & \Re A \end{bmatrix} z - \begin{bmatrix} \Re b \\ \Im b \end{bmatrix} \right\|_2^2. \end{aligned}$$

The values of F and g can be extracted from the above expression.

(b) First, let's write out the optimization problem term-by-term:

$$\text{minimize} \quad \|Ax-b\|_\infty$$

is equivalent to

$$\begin{array}{ll}\text{minimize} & t \\ \text{subject to} & |a_i^T x - b| < t \\ & i = 1, \ldots, m,\end{array}$$

where $a_1^T, \ldots, a_m^T$ are the rows of A. We have introduced a new optimization variable t.

Each term $|a_i^T x - b|$ must now be written in terms of real variables (we'll use the same z as before):

$$\begin{aligned} |a_i^T x - b|^2 &= (\Re a_i^T \Re x - \Im a_i^T \Im x - \Re b)^2 + (\Re a_i^T \Im x + \Im a_i^T \Re x - \Im b)^2 \\ &= \left\| \begin{bmatrix} \Re a_i^T & -\Im a_i^T \\ \Im a_i^T & \Re a_i^T \end{bmatrix} z - \begin{bmatrix} \Re b \\ \Im b \end{bmatrix} \right\|_2^2. \end{aligned}$$

So now we have reduced the problem to the real minimization,

$$\begin{array}{ll}\text{minimize} & t \\ \text{subject to} & \left\| \begin{bmatrix} \Re a_i^T & -\Im a_i^T \\ \Im a_i^T & \Re a_i^T \end{bmatrix} z - \begin{bmatrix} \Re b \\ \Im b \end{bmatrix} \right\|_2 < t \\ & i = 1, \ldots, m.\end{array}$$

This is a minimization over a second-order cone. It can be converted into a QCQP by squaring both sides of the constraint and defining $\lambda = t^2$:

$$\begin{array}{ll}\text{minimize} & \lambda \\ \text{subject to} & \left\| \begin{bmatrix} \Re a_i^T & -\Im a_i^T \\ \Im a_i^T & \Re a_i^T \end{bmatrix} z - \begin{bmatrix} \Re b \\ \Im b \end{bmatrix} \right\|_2^2 < \lambda \\ & i = 1, \ldots, m.\end{array}$$

(c) The ℓ_1-norm minimization problem is to minimize $\|Ax-b\|_1$, *i.e.*,

$$\text{minimize} \quad \textstyle\sum_{i=1}^m |a_i^T x - b|$$

Let us introduce new variables $t_1, \ldots, t_m$, and rewrite the minimization as follows:

$$\begin{array}{ll}\text{minimize} & \sum_{i=1}^m t_i \\ \text{subject to} & |a_i^T x - b| < t_i, \\ & i = 1, \ldots, m.\end{array}$$

The conversion to second-order constraints is similar to part (b):

$$\begin{array}{ll}\text{minimize} & \sum_{i=1}^m t_i \\ \text{subject to} & \left\| \begin{bmatrix} \Re a_i^T & -\Im a_i^T \\ \Im a_i^T & \Re a_i^T \end{bmatrix} z - \begin{bmatrix} \Re b \\ \Im b \end{bmatrix} \right\|_2 < t_i, \quad i = 1, \ldots, m.\end{array}$$

4.25 *Linear separation of two sets of ellipsoids.* Suppose we are given $K+L$ ellipsoids

$$\mathcal{E}_i = \{P_i u + q_i \mid \|u\|_2 \leq 1\}, \quad i = 1, \ldots, K+L,$$

where $P_i \in \mathbf{S}^n$. We are interested in finding a hyperplane that strictly separates $\mathcal{E}_1, \ldots, \mathcal{E}_K$ from $\mathcal{E}_{K+1}, \ldots, \mathcal{E}_{K+L}$, *i.e.*, we want to compute $a \in \mathbf{R}^n$, $b \in \mathbf{R}$ such that

$$a^T x + b > 0 \text{ for } x \in \mathcal{E}_1 \cup \cdots \cup \mathcal{E}_K, \qquad a^T x + b < 0 \text{ for } x \in \mathcal{E}_{K+1} \cup \cdots \cup \mathcal{E}_{K+L},$$

or prove that no such hyperplane exists. Express this problem as an SOCP feasibility problem.

Solution. We first note that the problem is homogeneous in a and b, so we can replace the strict inequalities $a^T x + b > 0$ and $a^T x + b < 0$ with $a^T x + b \geq 1$ and $a^T x + b \leq -1$, respectively.

The variables a and b must satisfy

$$\inf_{\|u\|_2 \leq 1} (a^T P_i u + a^T q_i) \geq 1, \quad 1, \ldots, L$$

and

$$\sup_{\|u\|_2 \leq 1} (a^T P_i u + a^T q_i) \leq -1, \quad i = K+1, \ldots, K+L.$$

The lefthand sides can be expressed as

$$\inf_{\|u\|_2 \leq 1} (a^T P_i u + a^T q_i) = -\|P_i^T a\|_2 + a^T q_i + b, \qquad \sup_{\|u\|_2 \leq 1} (a^T P_i u + a^T q_i) = \|P_i^T a\|_2 + a^T q_i + b.$$

We therefore obtain a set of second-order cone constraints in a, b:

$$\begin{array}{l} -\|P_i^T a\|_2 + a^T q_i + b \geq 1, \quad i = 1, \ldots, L \\ \|P_i^T a\|_2 + a^T q_i + b \leq -1, \quad i = K+1, \ldots, K+L. \end{array}$$

4.26 *Hyperbolic constraints as SOC constraints.* Verify that $x \in \mathbf{R}^n$, $y, z \in \mathbf{R}$ satisfy

$$x^T x \leq yz, \qquad y \geq 0, \qquad z \geq 0$$

if and only if

$$\left\| \begin{bmatrix} 2x \\ y - z \end{bmatrix} \right\|_2 \leq y + z, \qquad y \geq 0, \qquad z \geq 0.$$

Use this observation to cast the following problems as SOCPs.

(a) *Maximizing harmonic mean.*

$$\text{maximize} \quad \left(\sum_{i=1}^m 1/(a_i^T x - b_i) \right)^{-1},$$

with domain $\{x \mid Ax \succ b\}$, where a_i^T is the ith row of A.

(b) *Maximizing geometric mean.*

$$\text{maximize} \quad \left(\prod_{i=1}^m (a_i^T x - b_i) \right)^{1/m},$$

with domain $\{x \mid Ax \succeq b\}$, where a_i^T is the ith row of A.

Solution.

(a) The problem is equivalent to

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T t \\ \text{subject to} & t_i(a_i^T x + b_i) \geq 1, \quad i = 1, \ldots, m \\ & t \succeq 0. \end{array}$$

Writing the hyperbolic constraints as SOC constraints yields an SOCP

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T t \\ \text{subject to} & \left\| \begin{bmatrix} 2 \\ a_i^T x + b_i - t_i \end{bmatrix} \right\|_2 \leq a_i^T x + b_i + t_i, \quad i = 1, \ldots, m \\ & t_i \geq 0, \quad a_i^T x + b_i \geq 0, \quad i = 1, \ldots, m. \end{array}$$

(b) We can assume without loss of generality that $m = 2^K$ for some positive integer K. (If not, define $a_i = 0$ and $b_i = -1$ for $i = m+1, \ldots, 2^K$, where 2^K is the smallest power of two greater than m.)
Let us first take $m = 4$ ($K = 2$) as an example. The problem is equivalent to

$$\begin{array}{ll} \text{maximize} & y_1 y_2 y_3 y_4 \\ \text{subject to} & y = Ax - b \\ & y \succeq 0, \end{array}$$

which we can write as

$$\begin{array}{ll} \text{maximize} & t_1 t_2 \\ \text{subject to} & y = Ax - b \\ & y_1 y_2 \geq t_1^2 \\ & y_3 y_4 \geq t_2^2 \\ & y \succeq 0, \quad t_1 \geq 0, \quad t_2 \geq 0, \end{array}$$

and also as

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & y = Ax - b \\ & y_1 y_2 \geq t_1^2 \\ & y_3 y_4 \geq t_2^2 \\ & t_1 t_2 \geq t^2 \\ & y \succeq 0, \quad t_1, t_2, t \geq 0. \end{array}$$

Expressing the three hyperbolic constraints

$$y_1 y_2 \geq t_1^2, \qquad y_3 y_4 \geq t_2^2, \qquad t_1 t_2 \geq t^2$$

as SOC constraints yields an SOCP:

$$\begin{array}{ll} \text{minimize} & -t \\ \text{subject to} & \left\| \begin{bmatrix} 2t_1 \\ y_1 - y_2 \end{bmatrix} \right\|_2 \leq y_1 + y_2, \quad y_1 \geq 0, \quad y_2 \geq 0 \\ & \left\| \begin{bmatrix} 2t_2 \\ y_3 - y_4 \end{bmatrix} \right\|_2 \leq y_3 + y_4, \quad y_3 \geq 0, \quad y_4 \geq 0 \\ & \left\| \begin{bmatrix} 2t \\ t_1 - t_2 \end{bmatrix} \right\|_2 \leq t_1 + t_2, \quad t_1 \geq 0, \quad t_2 \geq 0 \\ & y = Ax - b. \end{array}$$

We can express the problem as

$$\begin{array}{ll} \text{maximize} & y_{00} \\ \text{subject to} & y_{K-1,j-1} = a_j^T x - b_j, \quad j = 1, \ldots, m \\ & y_{ik}^2 \leq y_{i+1,2^k} y_{i+1,2^k+1}, \quad i = 0, \ldots, K-2, \quad k = 0, \ldots 2^i - 1 \\ & Ax \succeq b, \end{array}$$

where we have introduced auxiliary variables y_{ij} for $i = 0, \ldots, K-1$, $j = 0, \ldots, 2^i - 1$. Expressing the hyperbolic constraints as SOC constraints yields an SOCP.
The equivalence can be proved by recursively expanding the objective function:

$$\begin{array}{rcl} y_{00} & \leq & y_{10} y_{11} \\ & \leq & (y_{20} y_{21})\,(y_{22} y_{23}) \end{array}$$

$$\begin{aligned} &\leq (y_{30}y_{31})(y_{32}y_{33})(y_{34}y_{35})(y_{36}y_{37}) \\ &\quad \cdots \\ &\leq y_{K-1,0}\, y_{K-1,1} \cdots y_{K-1,2^K-1} \\ &= (a_1^T x - b_1) \cdots (a_m^T x - b_m). \end{aligned}$$

4.27 *Matrix fractional minimization via SOCP.* Express the following problem as an SOCP:

$$\begin{array}{ll} \text{minimize} & (Ax+b)^T(I + B\,\mathbf{diag}(x)B^T)^{-1}(Ax+b) \\ \text{subject to} & x \succeq 0, \end{array}$$

with $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$, $B \in \mathbf{R}^{m\times n}$. The variable is $x \in \mathbf{R}^n$.
Hint. First show that the problem is equivalent to

$$\begin{array}{ll} \text{minimize} & v^T v + w^T\,\mathbf{diag}(x)^{-1} w \\ \text{subject to} & v + Bw = Ax + b \\ & x \succeq 0, \end{array}$$

with variables $v \in \mathbf{R}^m$, $w, x \in \mathbf{R}^n$. (If $x_i = 0$ we interpret w_i^2/x_i as zero if $w_i = 0$ and as ∞ otherwise.) Then use the results of exercise 4.26.
Solution. To show the equivalence with the problem in the hint, we assume $x \succeq 0$ is fixed, and optimize over v and w. This is a quadratic problem with equality constraints. The optimality conditions are

$$v = \nu, \qquad w = \mathbf{diag}(x)B^T\nu$$

for some ν. Substituting in the equality constraint, we see that ν must satisfy

$$(I + B\,\mathbf{diag}(x)B^T)\nu = Ax + b,$$

and, since the matrix on the left is invertible for $x \succeq 0$,

$$v = \nu = (I+B\,\mathbf{diag}(x)B^T)^{-1}(Ax+b), \qquad w = \mathbf{diag}(x)B^T(I+B\,\mathbf{diag}(x)B^T)^{-1}(Ax+b).$$

Substituting in the objective of the problem in the hint, we obtain

$$v^T v + w^T\,\mathbf{diag}(x)^{-1} w = (Ax+b)^T(I + B\,\mathbf{diag}(x)B^T)^{-1}(Ax+b).$$

This shows that the problem is equivalent to the problem in the hint.
As in exercise 4.26, we now introduce hyperbolic constraints and formulate the problem in the hint as

$$\begin{array}{ll} \text{minimize} & t + \mathbf{1}^T s \\ \text{subject to} & v^T v \leq t \\ & w_i^2 \leq s_i x_i, \quad i = 1, \ldots, n \\ & x \succeq 0 \end{array}$$

with variables $t \in \mathbf{R}$, $s, x, w \in \mathbf{R}^n$, $v \in \mathbf{R}^m$. Converting the hyperbolic constraints into SOC constraints results in an SOCP.

4.28 *Robust quadratic programming.* In §4.4.2 we discussed robust linear programming as an application of second-order cone programming. In this problem we consider a similar robust variation of the (convex) *quadratic* program

$$\begin{array}{ll} \text{minimize} & (1/2)x^T P x + q^T x + r \\ \text{subject to} & Ax \preceq b. \end{array}$$

For simplicity we assume that only the matrix P is subject to errors, and the other parameters (q, r, A, b) are exactly known. The robust quadratic program is defined as

$$\begin{array}{ll} \text{minimize} & \sup_{P\in\mathcal{E}}((1/2)x^T P x + q^T x + r) \\ \text{subject to} & Ax \preceq b \end{array}$$

where $\mathcal{E}$ is the set of possible matrices P.
For each of the following sets $\mathcal{E}$, express the robust QP as a convex problem. Be as specific as you can. If the problem can be expressed in a standard form (*e.g.*, QP, QCQP, SOCP, SDP), say so.

(a) A finite set of matrices: $\mathcal{E} = \{P_1, \ldots, P_K\}$, where $P_i \in \mathbf{S}^n_+$, $i = 1, \ldots, K$.

(b) A set specified by a nominal value $P_0 \in \mathbf{S}^n_+$ plus a bound on the eigenvalues of the deviation $P - P_0$:

$$\mathcal{E} = \{P \in \mathbf{S}^n \mid -\gamma I \preceq P - P_0 \preceq \gamma I\}$$

where $\gamma \in \mathbf{R}$ and $P_0 \in \mathbf{S}^n_+$,

(c) An ellipsoid of matrices:

$$\mathcal{E} = \left\{ P_0 + \sum_{i=1}^K P_i u_i \;\middle|\; \|u\|_2 \leq 1 \right\}.$$

You can assume $P_i \in \mathbf{S}^n_+$, $i = 0, \ldots, K$.

Solution.

(a) The objective function is a maximum of convex function, hence convex.
We can write the problem as

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & (1/2)x^T P_i x + q^T x + r \leq t, \;\; i = 1, \ldots, K \\ & Ax \preceq b, \end{array}$$

which is a QCQP in the variable x and t.

(b) For given x, the supremum of $x^T \Delta P x$ over $-\gamma I \preceq \Delta P \preceq \gamma I$ is given by

$$\sup_{-\gamma I \preceq \Delta P \preceq \gamma I} x^T \Delta P x = \gamma x^T x.$$

Therefore we can express the robust QP as

$$\begin{array}{ll} \text{minimize} & (1/2)x^T (P_0 + \gamma I)x + q^T x + r \\ \text{subject to} & Ax \preceq b \end{array}$$

which is a QP.

(c) For given x, the quadratic objective function is

$$\begin{aligned} & (1/2)\left(x^T P_0 x + \sup_{\|u\|_2 \leq 1} \sum_{i=1}^K u_i (x^T P_i x) \right) + q^T x + r \\ = \;\; & (1/2) x^T P_0 x + (1/2)\left(\sum_{i=1}^K (x^T P_i x)^2 \right)^{1/2} + q^T x + r. \end{aligned}$$

This is a convex function of x: each of the functions $x^T P_i x$ is convex since $P_i \succeq 0$. The second term is a composition $h(g_1(x), \ldots, g_K(x))$ of $h(y) = \|y\|_2$ with $g_i(x) = x^T P_i x$. The functions g_i are convex and nonnegative. The function h is convex and, for $y \in \mathbf{R}^K_+$, nondecreasing in each of its arguments. Therefore the composition is convex.
The resulting problem can be expressed as

$$\begin{array}{ll} \text{minimize} & (1/2)x^T P_0 x + \|y\|_2 + q^T x + r \\ \text{subject to} & (1/2)x^T P_i x \leq y_i, \quad i = 1, \ldots, K \\ & Ax \preceq b \end{array}$$

which can be further reduced to an SOCP

$$\begin{array}{ll} \text{minimize} & u + t \\ \text{subject to} & \left\| \begin{bmatrix} P_0^{1/2} x \\ 2u - 1/4 \end{bmatrix} \right\|_2 \leq 2u + 1/4 \\ & \left\| \begin{bmatrix} P_i^{1/2} x \\ 2y_i - 1/4 \end{bmatrix} \right\|_2 \leq 2y_i + 1/4, \quad i = 1, \ldots, K \\ & \|y\|_2 \leq t \\ & Ax \preceq b. \end{array}$$

The variables are x, u, t, and $y \in \mathbf{R}^K$.

Note that if we square both sides of the first inequality, we obtain

$$x^T P_0 x + (2u - 1/4)^2 \leq (2u + 1/4)^2,$$

i.e., $x^T P_0 x \leq 2u$. Similarly, the other constraints are equivalent to $(1/2)x^T P_i x \leq y_i$.

4.29 *Maximizing probability of satisfying a linear inequality.* Let c be a random variable in $\mathbf{R}^n$, normally distributed with mean $\bar{c}$ and covariance matrix R. Consider the problem

$$\begin{array}{ll} \text{maximize} & \mathbf{prob}(c^T x \geq \alpha) \\ \text{subject to} & Fx \preceq g, \quad Ax = b. \end{array}$$

Find the conditions under which this is equivalent to a convex or quasiconvex optimization problem. When these conditions hold, formulate the problem as a QP, QCQP, or SOCP (if the problem is convex), or explain how you can solve it by solving a sequence of QP, QCQP, or SOCP feasibility problems (if the problem is quasiconvex).

Solution. The problem can be expressed as a convex or quasiconvex problem if $\alpha < \bar{c}^T x$ for all feasible x.

Before working out the details, we first consider the special case with $\bar{c} = 0$. In this case $c^T x$ is a random variable, normally distributed with $\mathbf{E}(c^T x) = 0$ and $\mathbf{E}(c^T x)^2 = x^T R x$. If $\alpha < 0$, maximizing $\mathbf{prob}(c^T x \geq \alpha)$ means minimizing the variance, *i.e.*, minimizing $x^T R x$, subject to the constraints on x, which is a convex problem (in fact a QP). On the other hand, if $\alpha > 0$, we maximize $\mathbf{prob}(c^T x \geq \alpha)$ by *maximizing* the variance $x^T R x$, which is very difficult.

We now turn to the general case with $\bar{c} \neq 0$. Define $u = c^T x$, a scalar random variable, normally distributed with $\mathbf{E}\, u = \bar{c}^T x$ and $\mathbf{E}(u - \mathbf{E}\, u)^2 = x^T R x$. The random variable

$$\frac{u - \bar{c}^T x}{\sqrt{x^T R x}}$$

has a normal distribution with mean zero, and unit variance, so

$$\mathbf{prob}(u \geq \alpha) = \mathbf{prob}\left(\frac{u - \bar{c}^T x}{\sqrt{x^T R x}} \geq \frac{\alpha - \bar{c}^T x}{\sqrt{x^T R x}} \right) = 1 - \Phi\left(\frac{\alpha - \bar{c}^T x}{\sqrt{x^T R x}} \right),$$

where $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$, a monotonically increasing function.

To maximize $1 - \Phi$, we can minimize $(\alpha - \bar{c}^T x)/\sqrt{x^T R x}$, *i.e.*, solve the problem

$$\begin{array}{ll} \text{maximize} & (\bar{c}^T x - \alpha)/\sqrt{x^T R x} \\ \text{subject to} & Fx \preceq g \\ & Ax = b. \end{array}$$

Equivalently, if $\bar{c}^T x > \alpha$ for all feasible x, we can also minimize the reciprocal of the objective function:

$$\begin{array}{ll} \text{minimize} & \sqrt{x^T R x}/(\bar{c}^T x - \alpha) \\ \text{subject to} & Fx \preceq g \\ & Ax = b. \end{array}$$

If $\bar{c}^T x > \alpha$ for all feasible x, this is a quasiconvex optimization problem, which we can solve by bisection. Each bisection step requires the solution of of an SOCP feasibility problem

$$\sqrt{x^T R x} \leq t(\bar{c}^T x - \alpha), \qquad Fx \preceq g, \qquad Ax = b.$$

The problem can also be expressed as a convex problem, by making a change of variables

$$y = \frac{x}{\bar{c}^T x - \alpha}, \qquad t = \frac{1}{\bar{c}^T x - \alpha}.$$

This yields the problem

$$\begin{array}{ll} \text{minimize} & \sqrt{y^T R y} \\ \text{subject to} & Fy \preceq gt \\ & Ay = bt \\ & c_0^T y - \alpha t = 1 \\ & t \geq 0. \end{array}$$

If we square the objective this is a quadratic program.

Geometric programming

4.30 A heated fluid at temperature T (degrees above ambient temperature) flows in a pipe with fixed length and circular cross section with radius r. A layer of insulation, with thickness $w \ll r$, surrounds the pipe to reduce heat loss through the pipe walls. The design variables in this problem are T, r, and w.

The heat loss is (approximately) proportional to Tr/w, so over a fixed lifetime, the energy cost due to heat loss is given by $\alpha_1 Tr/w$. The cost of the pipe, which has a fixed wall thickness, is approximately proportional to the total material, *i.e.*, it is given by $\alpha_2 r$. The cost of the insulation is also approximately proportional to the total insulation material, *i.e.*, $\alpha_3 rw$ (using $w \ll r$). The total cost is the sum of these three costs.

The heat flow down the pipe is entirely due to the flow of the fluid, which has a fixed velocity, *i.e.*, it is given by $\alpha_4 Tr^2$. The constants α_i are all positive, as are the variables T, r, and w.

Now the problem: maximize the total heat flow down the pipe, subject to an upper limit $C_{\max}$ on total cost, and the constraints

$$T_{\min} \leq T \leq T_{\max}, \qquad r_{\min} \leq r \leq r_{\max}, \qquad w_{\min} \leq w \leq w_{\max}, \quad w \leq 0.1r.$$

Express this problem as a geometric program.

Solution. The problem is

$$\begin{array}{ll} \text{maximize} & \alpha_4 T r^2 \\ \text{subject to} & \alpha_1 T w^{-1} + \alpha_2 r + \alpha_3 rw \leq C_{\max} \\ & T_{\min} \leq T \leq T_{\max} \\ & r_{\min} \leq r \leq r_{\max} \\ & w_{\min} \leq w \leq w_{\max} \\ & w \leq 0.1r. \end{array}$$

This is equivalent to the GP

$$\begin{array}{ll}\text{minimize} & (1/\alpha_4)T^{-1}r^{-2} \\ \text{subject to} & (\alpha_1/C_{\max})Tw^{-1} + (\alpha_2/C_{\max})r + (\alpha_3/C_{\max})rw \leq 1 \\ & (1/T_{\max})T \leq 1, \quad T_{\min}T^{-1} \leq 1 \\ & (1/r_{\max})r \leq 1, \quad r_{\min}r^{-1} \leq 1 \\ & (1/w_{\max})w \leq 1, \quad w_{\min}w^{-1} \leq 1 \\ & 10wr^{-1} \leq 1.\end{array}$$

4.31 *Recursive formulation of optimal beam design problem.* Show that the GP (4.46) is equivalent to the GP

$$\begin{array}{ll}\text{minimize} & \sum_{i=1}^N w_i h_i \\ \text{subject to} & w_i/w_{\max} \leq 1, \quad w_{\min}/w_i \leq 1, \quad i = 1, \ldots, N \\ & h_i/h_{\max} \leq 1, \quad h_{\min}/h_i \leq 1, \quad i = 1, \ldots, N \\ & h_i/(w_i S_{\max}) \leq 1 \quad i = 1, \ldots, N \\ & 6iF/(\sigma_{\max} w_i h_i^2) \leq 1, \quad i = 1, \ldots, N \\ & (2i-1)d_i/v_i + v_{i+1}/v_i \leq 1, \quad i = 1, \ldots, N \\ & (i-1/3)d_i/y_i + v_{i+1}/y_i + y_{i+1}/y_i \leq 1, \quad i = 1, \ldots, N \\ & y_1/y_{\max} \leq 1 \\ & E w_i h_i^3 d_i/(6F) = 1, \quad i = 1, \ldots, N.\end{array}$$

The variables are w_i, h_i, v_i, d_i, y_i for $i = 1, \ldots, N$.

Solution. The problem is then

$$\begin{array}{ll}\text{minimize} & \sum_{i=1}^N w_i h_i \\ \text{subject to} & w_{\min} \leq w_i \leq w_{\max}, \quad i = 1, \ldots, N \\ & h_{\min} \leq h_i \leq h_{\max}, \quad i = 1, \ldots, N \\ & S_{\min} \leq h_i/w_i \leq S_{\max} \quad i = 1, \ldots, N \\ & 6iF/(w_i h_i^2) \leq \sigma_{\max}, \quad i = 1, \ldots, N \\ & v_i = (2i-1)d_i + v_{i+1}, \quad i = 1, \ldots, N \\ & y_i = (i-1/3)d_i + v_{i+1} + y_{i+1}, \quad i = 1, \ldots, N \\ & y_1 \leq y_{\max} \\ & d_i = 6F/(E w_i h_i^3), \quad i = 1, \ldots, N,\end{array} \tag{4.31.A}$$

where to simplify notation we use variables $d_i = 6F/(E w_i h_i^3)$, and define $y_{N+1} = d_{N+1} = 0$. The variables in the problem are w_i, h_i, v_i, y_i, d_i, for $i = 1, \ldots, N$.

This problem is not a GP, since the equalities that define v_i and y_i are not monomial inequalities. (The objective and other constraints, however, are fine.) Two approaches can be used to transform the problem (4.31.A) into an equivalent GP. One simple approach is to eliminate $v_1, \ldots, v_N$ and $y_2, \ldots, y_N$, using the recursion (4.45). This recursion shows that y_i and v_i are all posynomials in the variables w_i, h_i, and in particular, the constraint $y_1 \leq y_{\max}$ is a posynomial inequality.

We now describe another method, that would be better in practice if the number of segments is more than a small number, since it preserves the problem structure. To express this as a GP, we replace the equalities that define v_i and y_i by the inequalities

$$v_i \geq (2i-1)d_i + v_{i+1}, \qquad y_i \geq (i-1/3)d_i + v_{i+1} + y_{i+1}, \quad i = 1, \ldots, N. \tag{4.31.B}$$

This can be done without loss of generality. To see this, suppose we substitute the inequalities (4.31.B) in (4.31.A), and suppose h, w, v, y, d are feasible. The variables v_1 and y_1 appear in the following four inequalities

$$v_1 \geq d_1, \qquad y_1 \geq (2/3)d_1, \qquad v_2 \geq 3d_2 + v_1, \qquad y_2 \geq (5/3)d_2 + v_1 + y_1.$$

It is clear that setting $v_1 = d_1$ and $y_1 = (2/3)d_1$, without changing any of the other variables, yields a feasible point with the same objective value. Next, consider the four inequalities that involve v_2 and y_2:

$$v_2 \geq 3d_2 + v_1, \qquad y_2 \geq (5/3)d_2 + v_1 + y_1, \qquad v_3 \geq 5d_3 + v_2, \qquad y_3 \geq (7/3)d_3 + v_2 + y_2.$$

Again, it is clear that we can decrease v_2 and y_2 until the first two inequalities are tight, without changing the objective value. Continuing the argument, we conclude that the two problems are equivalent.

It is now straightforward to express the problem as the GP

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^N w_i h_i \\ \text{subject to} & w_i/w_{\max} \leq 1, \quad w_{\min}/w_i \leq 1, \quad i = 1, \ldots, N \\ & h_i/h_{\max} \leq 1, \quad h_{\min}/h_i \leq 1, \quad i = 1, \ldots, N \\ & h_i/(w_i S_{\max}) \leq 1 \quad i = 1, \ldots, N \\ & 6iF/(\sigma_{\max} w_i h_i^2) \leq 1, \quad i = 1, \ldots, N \\ & (2i-1)d_i/v_i + v_{i+1}/v_i \leq 1, \quad i = 1, \ldots, N \\ & (i-1/3)d_i/y_i + v_{i+1}/y_i + y_{i+1}/y_i \leq 1, \quad i = 1, \ldots, N \\ & y_1/y_{\max} \leq 1 \\ & E w_i h_i^3/(6Fd_i) = 1, \quad i = 1, \ldots, N. \end{array}$$

4.32 *Approximating a function as a monomial.* Suppose the function $f : \mathbf{R}^n \to \mathbf{R}$ is differentiable at a point $x_0 \succ 0$, with $f(x_0) > 0$. How would you find a monomial function $\hat{f} : \mathbf{R}^n \to \mathbf{R}$ such that $f(x_0) = \hat{f}(x_0)$ and for x near x_0, $\hat{f}(x)$ is very near $f(x)$?

Solution. We'll give two ways to solve this problem. They both end up with the same solution.

Let the monomial approximant have the form

$$\hat{f}(x) = c x_1^{a_1} \cdots x_n^{a_n},$$

where $c > 0$.

Method 1. First-order matching. To make $\hat{f}(x)$ very near $f(x)$ in the vicinity of x_0, we will make the function values agree, and also set the gradient of both functions equal at the point x_0:

$$\hat{f}(x_0) = f(x_0), \qquad \frac{\partial \hat{f}}{\partial x_i}\bigg|_{x_0} = \frac{\partial f}{\partial x_i}\bigg|_{x_0}.$$

We have

$$\frac{\partial \hat{f}}{\partial x_i} = c a_i x_1^{a_1} \cdots x_i^{a_i - 1} \cdots x_n^{a_n} = a_i x_i^{-1} \hat{f}(x),$$

which gives us an explicit expression for the exponents a_i:

$$a_i = \frac{x_i}{\hat{f}(x)} \frac{\partial f}{\partial x_i}\bigg|_{x_0}.$$

All that is left is to find the coefficient c of the monomial approximant. To do this we use the condition $\hat{f}(x_0) = f(x_0)$:

$$c = \frac{f(x)}{x_1^{a_1} \cdots x_n^{a_n}}\bigg|_{x_0}.$$

Method 2. Log transformation. As is done to transform a GP to convex form, we take the log of the function f and the variables, to get

$$g(y) = \log f(y), \qquad y_i = \log x_i,$$

and similarly for the approximating monomial:

$$\hat{g}(y) = \log \hat{f}(y) = \tilde{c} + a^T y,$$

where $\tilde{c} = \log c$. Note that the transformation takes the monomial into an affine function. After this transformation, the problem is this: find an affine function that fits $g(y)$ very well near the point $y_0 = \log x_0$. That's easy — the answer is to form the first-order Taylor approximation of g at y_0:

$$g(y_0) + \nabla g(y_0)^T (y - y_0) = \tilde{c} + a^T y.$$

This implies

$$\tilde{c} = g(y_0) - \nabla g(y_0)^T y_0, \qquad a = \nabla g(y_0).$$

If we work out what this means in terms of f, we end up with the same formulas for c and a_i as in method 1 above.

4.33 Express the following problems as convex optimization problems.

(a) Minimize $\max\{p(x), q(x)\}$, where p and q are posynomials.

(b) Minimize $\exp(p(x)) + \exp(q(x))$, where p and q are posynomials.

(c) Minimize $p(x)/(r(x) - q(x))$, subject to $r(x) > q(x)$, where p, q are posynomials, and r is a monomial.

Solution.

(a) This is equivalent to the GP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & p(x)/t \leq 1, \quad q(x)/t \leq 1. \end{array}$$

Now make the logarithmic change of variables $x_i = e^{y_i}$.

(b) Equivalent to

$$\begin{array}{ll} \text{minimize} & \exp(t_1) + \exp(t_2) \\ \text{subject to} & p(x) \leq t_1, \quad q(x) \leq t_2. \end{array}$$

Now make the logarithmic change of variables $x_i = e^{y_i}$ (but not to t_1, t_2).

(c) Equivalent to

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & p(x) \leq t(r(x) - q(x)), \end{array}$$

and

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & (p(x)/t + q(x))/r(x) \leq 1, \end{array}$$

which is a GP.

4.34 *Log-convexity of Perron-Frobenius eigenvalue.* Let $A \in \mathbf{R}^{n \times n}$ be an elementwise positive matrix, *i.e.*, $A_{ij} > 0$. (The results of this problem hold for irreducible nonnegative matrices as well.) Let $\lambda_{\rm pf}(A)$ denotes its Perron-Frobenius eigenvalue, *i.e.*, its eigenvalue of largest magnitude. (See the definition and the example on page 165.) Show that $\log \lambda_{\rm pf}(A)$ is a convex function of $\log A_{ij}$. This means, for example, that we have the inequality

$$\lambda_{\rm pf}(C) \leq \left(\lambda_{\rm pf}(A)\lambda_{\rm pf}(B)\right)^{1/2},$$

where $C_{ij} = (A_{ij}B_{ij})^{1/2}$, and A and B are elementwise positive matrices.

Hint. Use the characterization of the Perron-Frobenius eigenvalue given in (4.47), or, alternatively, use the characterization

$$\log \lambda_{\rm pf}(A) = \lim_{k \to \infty} (1/k) \log(\mathbf{1}^T A^k \mathbf{1}).$$

Solution. Define $\alpha_{ij} = \log A_{ij}$. From the characterization in the text

$$\begin{aligned}\log \lambda_{\rm pf}(A) &= \inf_{v \succ 0} \max_{i=1,\ldots,n} \log(\sum_{j=1}^n e^{\alpha_{ij}} v_j/v_i) \\ &= \inf_{y} \max_{i=1,\ldots,n} \left(\log(\sum_{j=1}^n e^{\alpha_{ij}+y_j}) - y_i \right)\end{aligned}$$

where we made a change of variables $v_i = e^{y_i}$. The functions

$$\log \left(\sum_{j=1}^n e^{\alpha_{ij}+y_j} \right) - y_i$$

are convex, jointly in α and y, so

$$\max_i \log \left(\sum_{j=1}^n e^{\alpha_{ij}+y_j} \right) - y_i$$

is jointly convex in α and y. Minimizing over y therefore gives a convex function of α. From the characterization in the hint

$$\log \lambda_{\rm pf}(A) = \lim_{k \to \infty} (1/k) \log(\sum_{i,j} (A^k)_{ij}).$$

A^k expanded as a sum of exponentials of linear functions of α_{ij}. So $\log \lambda_{\rm pf}(A)$ is the pointwise limit of a set of convex functions.

4.35 *Signomial and geometric programs.* A *signomial* is a linear combination of monomials of some positive variables $x_1, \ldots, x_n$. Signomials are more general than posynomials, which are signomials with all positive coefficients. A *signomial program* is an optimization problem of the form

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p, \end{array}$$

where $f_0, \ldots, f_m$ and $h_1, \ldots, h_p$ are signomials. In general, signomial programs are very difficult to solve.

Some signomial programs can be transformed to GPs, and therefore solved efficiently. Show how to do this for a signomial program of the following form:

- The objective signomial f_0 is a posynomial, *i.e.*, its terms have only positive coefficients.
- Each inequality constraint signomial $f_1, \ldots, f_m$ has exactly one term with a negative coefficient: $f_i = p_i - q_i$ where p_i is posynomial, and q_i is monomial.
- Each equality constraint signomial $h_1, \ldots, h_p$ has exactly one term with a positive coefficient and one term with a negative coefficient: $h_i = r_i - s_i$ where r_i and s_i are monomials.

Solution. For the inequality constraints, move the single negative term to the righthand side, then divide by it, to get a posynomial inequality: $f_i(x) \leq 0$ is equivalent to $p_i/q_i \leq 1$. For the equality constraints, move the negative term to the righthand side, then divide by it, to get a monomial equality: $h_i(x) = 0$ is equivalent to $r_i/s_i = 1$.

4.36 Explain how to reformulate a general GP as an equivalent GP in which every posynomial (in the objective and constraints) has at most two monomial terms. *Hint.* Express each sum (of monomials) as a sum of sums, each with two terms.

Solution. Consider a posynomial inequality with $t > 2$ terms,

$$f(x) = \sum_{i=1}^{t} g_i(x) \leq 1,$$

where g_i are monomials. We introduce new variables $s_1, \ldots, s_{t-2}$, and express the posynomial inequality as the set of posynomial inequalities

$$\begin{array}{rcl} g_1(x) + g_2(x) & \leq & s_1 \\ g_3(x) + s_1 & \leq & s_2 \\ & \vdots & \\ g_t(x) + s_{t-2} & \leq & 1. \end{array}$$

By dividing by the righthand side, these become posynomial inequalities with two terms each. They are clearly equivalent to the original posynomial inequality. Clearly s_i is an upper bound on $\sum_{j=1}^{i+1} g_j(x)$, so the last inequality, $g_t(x) + s_{t-2} \leq 1$, implies the original posynomial inequality. Conversely, we can always take $s_i = \sum_{j=1}^{i+1} g_j(x)$, so if the original posynomial is satisfied, there are $s_1, \ldots, s_{t-2}$ that satisfy the two-term posynomial inequalities above.

4.37 *Generalized posynomials and geometric programming.* Let $x_1, \ldots, x_n$ be positive variables, and suppose the functions $f_i : \mathbf{R}^n \to \mathbf{R}$, $i = 1, \ldots, k$, are posynomials of $x_1, \ldots, x_n$. If $\phi : \mathbf{R}^k \to \mathbf{R}$ is a polynomial with nonnegative coefficients, then the composition

$$h(x) = \phi(f_1(x), \ldots, f_k(x)) \tag{4.69}$$

is a posynomial, since posynomials are closed under products, sums, and multiplication by nonnegative scalars. For example, suppose f_1 and f_2 are posynomials, and consider the polynomial $\phi(z_1, z_2) = 3z_1^2 z_2 + 2z_1 + 3z_2^3$ (which has nonnegative coefficients). Then $h = 3f_1^2 f_2 + 2f_1 + f_2^3$ is a posynomial.

In this problem we consider a generalization of this idea, in which ϕ is allowed to be a posynomial, *i.e.*, can have fractional exponents. Specifically, assume that $\phi : \mathbf{R}^k \to \mathbf{R}$ is a posynomial, with all its exponents nonnegative. In this case we will call the function h defined in (4.69) a *generalized posynomial*. As an example, suppose f_1 and f_2 are posynomials, and consider the posynomial (with nonnegative exponents) $\phi(z_1, z_2) = 2z_1^{0.3} z_2^{1.2} + z_1 z_2^{0.5} + 2$. Then the function

$$h(x) = 2f_1(x)^{0.3} f_2(x)^{1.2} + f_1(x) f_2(x)^{0.5} + 2$$

is a generalized posynomial. Note that it is *not* a posynomial, however (unless f_1 and f_2 are monomials or constants).

A *generalized geometric program* (GGP) is an optimization problem of the form

$$\begin{array}{ll} \text{minimize} & h_0(x) \\ \text{subject to} & h_i(x) \leq 1, \quad i = 1, \ldots, m \\ & g_i(x) = 1, \quad i = 1, \ldots, p, \end{array} \tag{4.70}$$

where $g_1, \ldots, g_p$ are monomials, and $h_0, \ldots, h_m$ are generalized posynomials.

Show how to express this generalized geometric program as an equivalent geometric program. Explain any new variables you introduce, and explain how your GP is equivalent to the GGP (4.70).

Solution.
We first start by transforming to the epigraph form, by introducing a variable t and introducing a new inequality constraint $h_0(x) \leq t$, which can be written as $h_0(x)/t \leq 1$, which is a valid generalized posynomial inequality constraint. Now we'll show how to deal with the generalized posynomial inequality constraint

$$\phi(f_1(x), \ldots, f_k(x)) \leq 1, \tag{4.37.A}$$

where ϕ is a posynomial with nonnegative exponents, and $f_1, \ldots, f_k$ are posynomials.
We'll use the standard trick: introduce new variables $t_1, \ldots, t_k$, and replace the single generalized posynomial inequality constraint (4.37.A) with

$$\phi(t_1, \ldots, t_k) \leq 1, \qquad f_1(x) \leq t_1, \ldots, f_k(x) \leq t_k, \tag{4.37.B}$$

which is easily transformed to a set of $k+1$ ordinary posynomial inequalities (by dividing the last inequalities by t_i). We claim that this set of posynomial inequalities is equivalent to the original generalized posynomial inequality. To see this, suppose that $x, t_1, \ldots, x_k$ satisfy (4.37.B). Now we use the fact that the function ϕ is *monotone nondecreasing in each argument* (since its exponents are all nonnegative), which implies that

$$\phi(f_1(x), \ldots, f_k(x)) \leq 1.$$

Conversely, suppose that (4.37.A) holds. Then, defining $t_i = f_i(x)$, $i = 1, \ldots, k$, we find that

$$\phi(t_1, \ldots, t_k) \leq 1, \qquad f_1(x) = t_1, \ldots, f_k(x) = t_k$$

holds, which implies (4.37.B).
If we carry out this procedure for each generalized posynomial inequality, we obtain a GP. Since the inequalities are each equivalent, using the argument above, the two problems are equivalent.

Semidefinite programming and conic form problems

4.38 *LMIs and SDPs with one variable.* The *generalized eigenvalues* of a matrix pair (A, B), where $A, B \in \mathbf{S}^n$, are defined as the roots of the polynomial $\det(\lambda B - A)$ (see §A.5.3). Suppose B is nonsingular, and that A and B can be simultaneously diagonalized by a congruence, *i.e.*, there exists a nonsingular $R \in \mathbf{R}^{n\times n}$ such that

$$R^T A R = \mathbf{diag}(a), \qquad R^T B R = \mathbf{diag}(b),$$

where $a, b \in \mathbf{R}^n$. (A sufficient condition for this to hold is that there exists t_1, t_2 such that $t_1 A + t_2 B \succ 0$.)

(a) Show that the generalized eigenvalues of (A, B) are real, and given by $\lambda_i = a_i/b_i$, $i = 1, \ldots, n$.

(b) Express the solution of the SDP

$$\begin{array}{ll} \text{minimize} & ct \\ \text{subject to} & tB \preceq A, \end{array}$$

with variable $t \in \mathbf{R}$, in terms of a and b.

Solution.

(a) If B is nonsingular, $R^T B R$ must be nonsingular, *i.e.*, $b_i \neq 0$ for all i. We have

$$\det(\lambda B - A) = (\det R)^2 \prod(\lambda b_i - a_i) = 0$$

so λ is a generalized eigenvalue if and only if $\lambda = a_i/b_i$ for some i.

(b) We have $tB \preceq A$ if and only if $tb \preceq a$, *i.e.*,

$$\begin{cases} t \le a_i/b_i & b_i > 0 \\ t \ge a_i/b_i & b_i < 0. \end{cases}$$

The feasible set is an interval defined by,

$$\max_{b_i<0} a_i/b_i \le t \le \min_{b_i>0} a_i/b_i.$$

If the interval is nonempty and bounded, the optimal solution is one of the endpoints, depending on the sign of c.

4.39 *SDPs and congruence transformations.* Consider the SDP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & x_1F_1 + x_2F_2 + \cdots + x_nF_n + G \preceq 0, \end{array}$$

with $F_i, G \in \mathbf{S}^k$, $c \in \mathbf{R}^n$.

(a) Suppose $R \in \mathbf{R}^{k\times k}$ is nonsingular. Show that the SDP is equivalent to the SDP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & x_1\tilde{F}_1 + x_2\tilde{F}_2 + \cdots + x_n\tilde{F}_n + \tilde{G} \preceq 0, \end{array}$$

where $\tilde{F}_i = R^T F_i R$, $\tilde{G} = R^T G R$.

(b) Suppose there exists a nonsingular R such that $\tilde{F}_i$ and $\tilde{G}$ are diagonal. Show that the SDP is equivalent to an LP.

(c) Suppose there exists a nonsingular R such that $\tilde{F}_i$ and $\tilde{G}$ have the form

$$\tilde{F}_i = \begin{bmatrix} \alpha_i I & a_i \\ a_i^T & \alpha_i \end{bmatrix}, \quad i = 1, \ldots, n, \qquad \tilde{G} = \begin{bmatrix} \beta I & b \\ b & \beta \end{bmatrix},$$

where $\alpha_i, \beta \in \mathbf{R}$, $a_i, b \in \mathbf{R}^{k-1}$. Show that the SDP is equivalent to an SOCP with a single second-order cone constraint.

Solution.

(a) Let $A \in \mathbf{S}^n$ and $R \in \mathbf{R}^{n\times n}$ with R nonsingular. $A \succeq 0$ if and only if $x^T A x \ge 0$ for all x. Hence, with $x = Ry$, $y^T R^T A R y \ge 0$ for all y, *i.e.*, $y^T R^T A R \succeq 0$.

(b) A diagonal matrix is positive semidefinite if and only if its diagonal elements are nonnegative.

(c) The LMI is equivalent to

$$\tilde{F}(x) = \begin{bmatrix} (\alpha^t x + \beta)I & Ax + b \\ (Ax + b)^T & (\alpha^T x + \beta)I \end{bmatrix} \succeq 0.$$

where A has columns a_i, *i.e.*, $\|Ax + b\|_2 \le \alpha^T x + \beta$.

4.40 *LPs, QPs, QCQPs, and SOCPs as SDPs.* Express the following problems as SDPs.

(a) The LP (4.27).

Solution.

$$\begin{array}{ll} \text{minimize} & c^T x + d \\ \text{subject to} & \mathbf{diag}(Gx - h) \preceq 0 \\ & Ax = b. \end{array}$$

(b) The QP (4.34), the QCQP (4.35) and the SOCP (4.36). *Hint.* Suppose $A \in \mathbf{S}^r_{++}$, $C \in \mathbf{S}^s$, and $B \in \mathbf{R}^{r\times s}$. Then

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \succeq 0 \iff C - B^T A^{-1} B \succeq 0.$$

For a more complete statement, which applies also to singular A, and a proof, see §A.5.5.

Solution.

(a) QP. Express $P = WW^T$ with $W \in \mathbf{R}^{n\times r}$.

$$\begin{array}{ll} \text{minimize} & t + 2q^T x + r \\ \text{subject to} & \begin{bmatrix} I & W^T x \\ x^T W & tI \end{bmatrix} \succeq 0 \\ & \mathbf{diag}(Gx - h) \preceq 0 \\ & Ax = b, \end{array}$$

with variables x, $t \in \mathbf{R}$.

(b) QCQP. Express $P_i = W_i W_i^T$ with $W_i \in \mathbf{R}^{n\times r_i}$.

$$\begin{array}{ll} \text{minimize} & t_0 + 2q_0^T x + r_0 \\ \text{subject to} & t_i + 2q_i^T x + r_i \le 0, \quad i = 1, \ldots, m \\ & \begin{bmatrix} I & W_i^T x \\ x^T W_i & t_i I \end{bmatrix} \succeq 0, \quad i = 0, 1, \ldots, m \\ & Ax = b, \end{array}$$

with variables x, $t_i \in \mathbf{R}$.

(c) SOCP.

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \begin{bmatrix} (c_i^T x + d_i) I & A_i x + b_i \\ (Ax_i + b_i)^T & (c_i^T x + d_i) I \end{bmatrix} \succeq 0, \quad i = 1, \ldots, N \\ & Fx = g. \end{array}$$

By the result in the hint, the constraint is equivalent with $\|A_i x + b_i\|_2 < c_i^T x + d_i$ when $c_i^T x + d_i > 0$. We have to check the case $c_i^T x + d_i = 0$ separately. In this case, the LMI constraint means $A_i x + b_i = 0$, so we can conclude that the LMI constraint and the SOC constraint are equivalent.

(c) The matrix fractional optimization problem

$$\text{minimize} \quad (Ax + b)^T F(x)^{-1} (Ax + b)$$

where $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$,

$$F(x) = F_0 + x_1 F_1 + \cdots + x_n F_n,$$

with $F_i \in \mathbf{S}^m$, and we take the domain of the objective to be $\{x \mid F(x) \succ 0\}$. You can assume the problem is feasible (there exists at least one x with $F(x) \succ 0$).

Solution.

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \begin{bmatrix} F(x) & Ax + b \\ (Ax + b)^T & t \end{bmatrix} \succeq 0 \end{array}$$

with variables x, $t \in \mathbf{R}$. The LMI constraint is equivalent to

$$(Ax + b)^T F(x)^{-1} (Ax + b) \le t$$

if $F(x) \succ 0$.
More generally, let

$$f_0(x) = (Ax+b)^T F(x)^{-1}(Ax+b), \qquad \mathbf{dom}\, f_0(x) = \{x \mid F(x) \succ 0\}.$$

We have

$$\mathbf{epi}\, f_0 = \left\{ (x,t) \;\middle|\; F(x) \succ 0,\ \begin{bmatrix} F(x) & Ax+b \\ (Ax+b)^T & t \end{bmatrix} \succeq 0 \right\}.$$

Then $\mathbf{cl}(\mathbf{epi}\, f_0) = \mathbf{epi}\, g$ where g is defined by

$$\begin{aligned} \mathbf{epi}\, g &= \left\{ (x,t) \;\middle|\; \begin{bmatrix} F(x) & Ax+b \\ (Ax+b)^T & t \end{bmatrix} \succeq 0 \right\} \\ g(x) &= \inf \left\{ t \;\middle|\; \begin{bmatrix} F(x) & Ax+b \\ (Ax+b)^T & t \end{bmatrix} \succeq 0 \right\}. \end{aligned}$$

We conclude that both problems have the same optimal values. An optimal solution for the matrix fractional problem is optimal for the SDP. An optimal solution for the SDP, with $F(x) \succ 0$, is optimal for the matrix fractional problem. If $F(x)$ is singular at the optimal solution of the SDP, then the optimum for the matrix fractional problem is not attained.

4.41 *LMI tests for copositive matrices and P_0-matrices.* A matrix $A \in \mathbf{S}^n$ is said to be *copositive* if $x^T A x \geq 0$ for all $x \succeq 0$ (see exercise 2.35). A matrix $A \in \mathbf{R}^{n\times n}$ is said to be a *P_0-matrix* if $\max_{i=1,\ldots,n} x_i (Ax)_i \geq 0$ for all x. Checking whether a matrix is copositive or a P_0-matrix is very difficult in general. However, there exist useful sufficient conditions that can be verified using semidefinite programming.

(a) Show that A is copositive if it can be decomposed as a sum of a positive semidefinite and an elementwise nonnegative matrix:

$$A = B + C, \qquad B \succeq 0, \qquad C_{ij} \geq 0, \quad i,j = 1,\ldots,n. \tag{4.71}$$

Express the problem of finding B and C that satisfy (4.71) as an SDP feasibility problem.

(b) Show that A is a P_0-matrix if there exists a positive diagonal matrix D such that

$$DA + A^T D \succeq 0. \tag{4.72}$$

Express the problem of finding a D that satisfies (4.72) as an SDP feasibility problem.

Solution.

(a) Suppose A satisfies (4.71). Let $x \succeq 0$. Then

$$x^T A x = x^T B x + x^T C x \geq 0,$$

because $B \succeq 0$ and $C_{ij} \geq 0$ for all i,j.

(b) Suppose A satisfies (4.72). Then

$$x^T (DA + A^T D) x = 2 \sum_{k=1}^n d_k x_k (Ax_k) \geq 0$$

for all x. Since $d_k > 0$, we must have $x_k(Ax_k) \geq 0$ for at least one k.

4.42 *Complex LMIs and SDPs.* A complex LMI has the form

$$x_1F_1 + \cdots + x_nF_n + G \preceq 0$$

where $F_1, \ldots, F_n,\ G$ are complex $n \times n$ Hermitian matrices, *i.e.*, $F_i^H = F_i$, $G^H = G$, and $x \in \mathbf{R}^n$ is a real variable. A complex SDP is the problem of minimizing a (real) linear function of x subject to a complex LMI constraint.

Complex LMIs and SDPs can be transformed to real LMIs and SDPs, using the fact that

$$X \succeq 0 \iff \begin{bmatrix} \Re X & -\Im X \\ \Im X & \Re X \end{bmatrix} \succeq 0,$$

where $\Re X \in \mathbf{R}^{n\times n}$ is the real part of the complex Hermitian matrix X, and $\Im X \in \mathbf{R}^{n\times n}$ is the imaginary part of X.

Verify this result, and show how to pose a complex SDP as a real SDP.

Solution. For a Hermitian matrix $\Re X = (\Re X)^T$ and $\Im X = -\Im X^T$. Now let $z = u + iv$, where u, v are real vectors, and $i = \sqrt{-1}$. We have

$$\begin{aligned} z^H X z &= (u - iv)^T(\Re X + i\Im X)(u + iv) \\ &= u^T\Re X u + v^T\Re X v - u^T\Im X v + v^T\Im X u \\ &= \begin{bmatrix} u^T & v^T \end{bmatrix} \begin{bmatrix} \Re X & -\Im X \\ \Im X & \Re X \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}. \end{aligned}$$

Therefore $z^H X z \geq 0$ for all z if and only if the $2n \times 2n$ real (symmetric) matrix above is positive semidefinite.

Thus, we can convert a complex LMI into a real LMI with twice the size. The conversion is linear, a complex LMI becomes a real LMI, of twice the size.

To pose

4.43 *Eigenvalue optimization via SDP.* Suppose $A : \mathbf{R}^n \to \mathbf{S}^m$ is affine, *i.e.*,

$$A(x) = A_0 + x_1A_1 + \cdots + x_nA_n$$

where $A_i \in \mathbf{S}^m$. Let $\lambda_1(x) \geq \lambda_2(x) \geq \cdots \geq \lambda_m(x)$ denote the eigenvalues of $A(x)$. Show how to pose the following problems as SDPs.

(a) Minimize the maximum eigenvalue $\lambda_1(x)$.

(b) Minimize the spread of the eigenvalues, $\lambda_1(x) - \lambda_m(x)$.

(c) Minimize the condition number of $A(x)$, subject to $A(x) \succ 0$. The condition number is defined as $\kappa(A(x)) = \lambda_1(x)/\lambda_m(x)$, with domain $\{x \mid A(x) \succ 0\}$. You may assume that $A(x) \succ 0$ for at least one x.

Hint. You need to minimize λ/γ, subject to

$$0 \prec \gamma I \preceq A(x) \preceq \lambda I.$$

Change variables to $y = x/\gamma$, $t = \lambda/\gamma$, $s = 1/\gamma$.

(d) Minimize the sum of the absolute values of the eigenvalues, $|\lambda_1(x)| + \cdots + |\lambda_m(x)|$.

Hint. Express $A(x)$ as $A(x) = A_+ - A_-$, where $A_+ \succeq 0$, $A_- \succeq 0$.

Solution.

(a) We use the property that $\lambda_1(x) \leq t$ if and only if $A(x) \preceq tI$. We minimize the maximum eigenvalue by solving the SDP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & A(x) \preceq tI. \end{array}$$

The variables are $x \in \mathbf{R}^n$ and $t \in \mathbf{R}$.

(b) $\lambda_1(x) \leq t_1$ if and only if $A(x) \preceq t_1 I$ and $\lambda_m(A(x)) \geq t_2$ if and only if $A(x) \succeq t_2 I$, so we can minimize $\lambda_1 - \lambda_m$ by solving

$$\begin{array}{ll} \text{minimize} & t_1 - t_2 \\ \text{subject to} & t_2 I \preceq A(x) \preceq t_1 I. \end{array}$$

This is an SDP with variables $t_1 \in \mathbf{R}$, $t_2 \in \mathbf{R}$, and $x \in \mathbf{R}^n$.

(c) We first note that the problem is equivalent to

$$\begin{array}{ll} \text{minimize} & \lambda/\gamma \\ \text{subject to} & \gamma I \preceq A(x) \preceq \lambda I \end{array} \tag{4.43.A}$$

if we take as domain of the objective $\{(\lambda, \gamma) \mid \gamma > 0\}$. This problem is quasiconvex, and can be solved by bisection: The optimal value is less than or equal to α if and only if the inequalities

$$\lambda \leq \gamma\alpha, \qquad \gamma I \preceq A(x) \preceq \lambda I, \qquad \gamma > 0$$

(with variables γ, λ, x) are feasible.

Following the hint we can also pose the problem as the SDP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & I \preceq sA_0 + y_1 A_1 + \cdots + y_n A_n \preceq tI \\ & s \geq 0. \end{array} \tag{4.43.B}$$

We now verify more carefully that the two problems are equivalent. Let $p^\star$ be the optimal value of (4.43.A), and $p^\star_{\text{sdp}}$ is the optimal value of the SDP (4.43.B).

We first show that $p^\star \geq p^\star_{\text{sdp}}$. Let λ/γ be the objective value of (4.43.A), evaluated at a feasible point (γ, λ, x). Define $s = 1/\gamma$, $y = x/\gamma$, $t = \lambda/\gamma$. This yields a feasible point in (4.43.B), with objective value $t = \lambda/\gamma$. This proves that $p^\star \geq p^\star_{\text{sdp}}$.

Next, we show that $p^*_{\text{sdp}} \geq p^*$. Suppose that s, y, t are feasible in (4.43.B). If $s > 0$, then $\gamma = 1/s$, $x = y/s$, $\lambda = t/s$ are feasible in (4.43.A) with objective value t. If $s = 0$, we have

$$I \preceq y_1 A_1 + \cdots + y_n A_n \preceq tI.$$

Choose $x = \tau y$, with τ sufficiently large so that $A(\tau y) \succeq A_0 + \tau I \succ 0$. We have

$$\lambda_1(\tau y) \leq \lambda_1(0) + \tau t, \qquad \lambda_m(\tau y) \geq \lambda_m(0) + \tau$$

so for τ sufficiently large,

$$\kappa(x_0 + \tau y) \leq \frac{\lambda_1(0) + t\tau}{\lambda_m(0) + \tau}.$$

Letting τ go to infinity, we can construct feasible points in (4.43.A), with objective value arbitrarily close to t. We conclude that $t \geq p^\star$ if (s, y, t) are feasible in (4.43.B). Minimizing over t yields $p^\star_{\text{sdp}} \geq p^\star$.

(d) This problem can be expressed as the SDP

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}\, A^+ + \mathbf{tr}\, A^- \\ \text{subject to} & A(x) = A^+ - A^- \\ & A^+ \succeq 0, \quad A^- \succeq 0, \end{array} \tag{4.43.C}$$

with variables x, A^+, A^-. We can show the equivalence as follows. First assume x is fixed in (4.43.C), and that A^+ and A^- are the only variables. We will show that

the optimal A^+ and A^- are easily constructed from the eigenvalue decomposition of $A(x)$, and that at the optimum we have

$$\mathbf{tr}\, A^+ + \mathbf{tr}\, A^- = \sum_{i=1}^{n} |\lambda_i(A(x))|.$$

Let $A(x) = Q\Lambda Q^T$ be the eigenvalue decomposition of $A(x)$. Defining $\tilde{A}^+ = Q^T A^+ Q$, $\tilde{A}^- = Q^T A^- Q$, we can write problem (4.43.C) as

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}\, \tilde{A}^+ + \mathbf{tr}\, \tilde{A}^- \\ \text{subject to} & \Lambda = \tilde{A}^+ - \tilde{A}^- \\ & \tilde{A}^+ \succeq 0, \quad \tilde{A}^- \succeq 0, \end{array} \tag{4.43.D}$$

with variables $\tilde{A}^+$ and $\tilde{A}^-$. Here we have used the fact that

$$\mathbf{tr}\, A^+ = \mathbf{tr}\, QQ^T A^+ = \mathbf{tr}\, Q^T A^+ Q = \mathbf{tr}\, \tilde{A}^+.$$

When solving problem (4.43.D), we can assume without loss of generality that the matrices $\tilde{A}^+$ and $\tilde{A}^-$ are diagonal. (If they are not diagonal, we can set the off-diagonal elements equal to zero, without changing the objective value and without changing feasibility.) The optimal values for the diagonal elements are:

$$\tilde{A}^+_{ii} = \max\{\lambda_i, 0\}, \quad \tilde{A}^-_{ii} = \max\{-\lambda_i, 0\},$$

and the optimal value $\sum_i |\lambda_i|$. Going back to the problem (4.43.C), we have shown that if we fix x, and optimize over A^+ and A^-, the optimal value of the problem is

$$\sum_{i=1}^{m} |\lambda_i(A(x))|.$$

Since the constraints are linear in x, we can allow x to be a variable. Minimizing over x, A^+, and A^- jointly is equivalent to minimizing $\sum_{i=1}^{m} |\lambda_i(A(x))|$.

4.44 *Optimization over polynomials.* Pose the following problem as an SDP. Find the polynomial $p : \mathbf{R} \to \mathbf{R}$,

$$p(t) = x_1 + x_2 t + \cdots + x_{2k+1} t^{2k},$$

that satisfies given bounds $l_i \leq p(t_i) \leq u_i$, at m specified points t_i, and, of all the polynomials that satisfy these bounds, has the greatest minimum value:

$$\begin{array}{ll} \text{maximize} & \inf_t p(t) \\ \text{subject to} & l_i \leq p(t_i) \leq u_i, \quad i = 1, \ldots, m. \end{array}$$

The variables are $x \in \mathbf{R}^{2k+1}$.

Hint. Use the LMI characterization of nonnegative polynomials derived in exercise 2.37, part (b).

Solution. First reformulate the problem as

$$\begin{array}{ll} \text{maximize} & \gamma \\ \text{subject to} & p(t) - \gamma \geq 0, \quad t \in \mathbf{R} \\ & l_i \leq p(t_i) \leq u_i, \quad i = 1, \ldots, m \end{array}$$

(variables x, γ). Now use the LMI characterization to get an SDP:

$$\begin{array}{ll} \text{maximize} & \gamma \\ \text{subject to} & x_1 - \gamma = Y_{11} \\ & x_i = \sum_{m+n=i+1} Y_{mn}, \quad i = 2, \ldots, 2k+1 \\ & l_i \leq \sum_i p(t_i) \leq u_i, \quad i = 1, \ldots, m \\ & Y \succeq 0. \end{array}$$

The variables are $x \in \mathbf{R}^{2k+1}$, $\gamma \in \mathbf{R}$, $Y \in \mathbf{S}^{k+1}$.

4.45 [Nes00, Par00] *Sum-of-squares representation via LMIs.* Consider a polynomial $p : \mathbf{R}^n \to \mathbf{R}$ of degree $2k$. The polynomial is said to be positive semidefinite (PSD) if $p(x) \geq 0$ for all $x \in \mathbf{R}^n$. Except for special cases (*e.g.*, $n = 1$ or $k = 1$), it is extremely difficult to determine whether or not a given polynomial is PSD, let alone solve an optimization problem, with the coefficients of p as variables, with the constraint that p be PSD.

A famous sufficient condition for a polynomial to be PSD is that it have the form

$$p(x) = \sum_{i=1}^{r} q_i(x)^2,$$

for some polynomials q_i, with degree no more than k. A polynomial p that has this sum-of-squares form is called SOS.

The condition that a polynomial p be SOS (viewed as a constraint on its coefficients) turns out to be equivalent to an LMI, and therefore a variety of optimization problems, with SOS constraints, can be posed as SDPs. You will explore these ideas in this problem.

(a) Let $f_1, \ldots, f_s$ be all monomials of degree k or less. (Here we mean monomial in the standard sense, *i.e.*, $x_1^{m_1} \cdots x_n^{m_n}$, where $m_i \in \mathbf{Z}_+$, and not in the sense used in geometric programming.) Show that if p can be expressed as a positive semidefinite quadratic form $p = f^T V f$, with $V \in \mathbf{S}_+^s$, then p is SOS. Conversely, show that if p is SOS, then it can be expressed as a positive semidefinite quadratic form in the monomials, *i.e.*, $p = f^T V f$, for some $V \in \mathbf{S}_+^s$.

(b) Show that the condition $p = f^T V f$ is a set of linear equality constraints relating the coefficients of p and the matrix V. Combined with part (a) above, this shows that the condition that p be SOS is equivalent to a set of linear equalities relating V and the coefficients of p, and the matrix inequality $V \succeq 0$.

(c) Work out the LMI conditions for SOS explicitly for the case where p is polynomial of degree four in two variables.

Solution.

(a) Factor V as $V = WW^T$, where $W \in \mathbf{R}^{s \times r}$ and let w_i denote the ith column of W. We have

$$p = f^T \sum_{i=1}^{r} w_i w_i^T f = \sum_{i=1}^{r} (w_i^T f)^2,$$

i.e., p is SOS.

Conversely, if p is SOS, it can be expressed as $p = \sum_{i=1}^{r} (w_i^T f)^2$, so $p = f^T V F$ for $V = \sum_{i=1}^{r} w_i w_i^T$.

(b) Expanding the quadratic form gives

$$p = \sum_{i,j=1}^{s} V_{ij} f_i f_j,$$

and equating coefficients on both sides proves the result.

(c) Solution for degree 2: The monomials of degree 2 or less are

$$f_1 = 1, \qquad f_2 = x_1, \qquad f_3 = x_2, \qquad f_5 = x_1^2, \qquad f_6 = x_1 x_2, \qquad f_7 = x_2^2$$

and the general expression for p

$$\begin{aligned} p(x) \;=\; & c_1 + c_2 x_1 + c_3 x_2 + c_4 x_1^2 + c_5 x_1 x_2 + c_6 x_2^2 + c_7 x_1^3 + c_8 x_1^2 x_2 \\ & + c_9 x_1 x_2^2 + c_{10} x_2^3 + c_{11} x_1^4 + c_{12} x_1^3 x_2 + c_{13} x_1^2 x_2^2 + c_{14} x_1 x_2^3 + c_{15} x_2^4 \end{aligned}$$

The equality constraints are

$$c_1 = V_{11}, \quad c_2 = 2V_{12}, \quad c_3 = 2V_{13}, \quad c_4 = V_{22} + 2V_{15}, \quad c_5 = 2V_{23} + 2V_{16},$$
$$c_6 = V_{33} + 2V_{17}, \quad c_7 = 2V_{25}, \quad c_8 = 2V_{26} + 2V_{25}, \quad c_9 = 2V_{27} + 2V_{36}, \quad c_{10} = 2V_{37},$$
$$c_{11} = V_{55}, \quad c_{12} = 2V_{56}, \quad c_{13} = 2V_{57}, \quad c_{14} = 2V_{67}, \quad c_{15} = V_{77}.$$

These, together with $V \in \mathbf{S}^7_+$, are the (necessary and sufficient) LMI conditions for p to be SOS.

4.46 *Multidimensional moments.* The moments of a random variable t on $\mathbf{R}^2$ are defined as $\mu_{ij} = \mathbf{E}\, t_1^i t_2^j$, where i, j are nonnegative integers. In this problem we derive necessary conditions for a set of numbers μ_{ij}, $0 \leq i, j \leq 2k$, $i + j \leq 2k$, to be the moments of a distribution on $\mathbf{R}^2$.

Let $p : \mathbf{R}^2 \to \mathbf{R}$ be a polynomial of degree k with coefficients c_{ij},

$$p(t) = \sum_{i=0}^{k} \sum_{j=0}^{k-i} c_{ij} t_1^i t_2^j,$$

and let t be a random variable with moments μ_{ij}. Suppose $c \in \mathbf{R}^{(k+1)(k+2)/2}$ contains the coefficients c_{ij} in some specific order, and $\mu \in \mathbf{R}^{(k+1)(2k+1)}$ contains the moments μ_{ij} in the same order. Show that $\mathbf{E}\, p(t)^2$ can be expressed as a quadratic form in c:

$$\mathbf{E}\, p(t)^2 = c^T H(\mu) c,$$

where $H : \mathbf{R}^{(k+1)(2k+1)} \to \mathbf{S}^{(k+1)(k+2)/2}$ is a linear function of μ. From this, conclude that μ must satisfy the LMI $H(\mu) \succeq 0$.

Remark: For random variables on $\mathbf{R}$, the matrix H can be taken as the Hankel matrix defined in (4.52). In this case, $H(\mu) \succeq 0$ is a necessary and sufficient condition for μ to be the moments of a distribution, or the limit of a sequence of moments. On $\mathbf{R}^2$, however, the LMI is only a necessary condition.

Solution.

$$y = (c_{00}, c_{10}, c_{01}, c_{20}, c_{11}, c_{02}, c_{30}, c_{21}, c_{12}, c_{03}, \ldots, c_{k0}, c_{k-1,1}, \ldots, c_{0k})$$

$$\begin{aligned}
\mathbf{E}\, p(t)^2 &= \mathbf{E} \left(\sum_{i=0}^{k} \sum_{j=0}^{k-i} c_{ij} t_1^i t_2^j \right)^2 \\
&= \mathbf{E} \sum_{i=0}^{k} \sum_{j=0}^{k-i} \sum_{m=0}^{k} \sum_{n=0}^{k-m} c_{ij} c_{mn} (t_1^{i+m} t_2^{j+n}) \\
&= \sum_{i=0}^{k} \sum_{j=0}^{k-i} \sum_{m=0}^{k} \sum_{n=0}^{k-m} c_{ij} c_{mn} \mu_{i+m,j+n},
\end{aligned}$$

i.e.,

$$H_{ij,mn} = \mu_{i+m,j+n}.$$

For example, with $k = 2$,

$$\begin{aligned}
&\mathbf{E}(c_{00} + c_{10} t_1 + c_{01} t_2 + c_{20} t_1^2 + c_{11} t_1 t_2 + c_{02} t_2^2)^2 \\
&= \begin{bmatrix} c_{00} & c_{10} & c_{01} & c_{20} & c_{11} & c_{02} \end{bmatrix}
\begin{bmatrix}
\mu_{00} & \mu_{10} & \mu_{01} & \mu_{20} & \mu_{11} & \mu_{02} \\
\mu_{10} & \mu_{20} & \mu_{11} & \mu_{30} & \mu_{21} & \mu_{12} \\
\mu_{01} & \mu_{11} & \mu_{02} & \mu_{21} & \mu_{12} & \mu_{03} \\
\mu_{20} & \mu_{30} & \mu_{21} & \mu_{40} & \mu_{31} & \mu_{22} \\
\mu_{11} & \mu_{21} & \mu_{12} & \mu_{31} & \mu_{22} & \mu_{13} \\
\mu_{02} & \mu_{12} & \mu_{03} & \mu_{22} & \mu_{13} & \mu_{04}
\end{bmatrix}
\begin{bmatrix} c_{00} \\ c_{10} \\ c_{01} \\ c_{20} \\ c_{11} \\ c_{02} \end{bmatrix}.
\end{aligned}$$

4.47 *Maximum determinant positive semidefinite matrix completion.* We consider a matrix $A \in \mathbf{S}^n$, with some entries specified, and the others not specified. The *positive semidefinite matrix completion problem* is to determine values of the unspecified entries of the matrix so that $A \succeq 0$ (or to determine that such a completion does not exist).

(a) Explain why we can assume without loss of generality that the diagonal entries of A are specified.

(b) Show how to formulate the positive semidefinite completion problem as an SDP feasibility problem.

(c) Assume that A has at least one completion that is positive definite, and the diagonal entries of A are specified (*i.e.*, fixed). The positive definite completion with largest determinant is called the *maximum determinant completion*. Show that the maximum determinant completion is unique. Show that if $A^\star$ is the maximum determinant completion, then $(A^\star)^{-1}$ has zeros in all the entries of the original matrix that were not specified. *Hint.* The gradient of the function $f(X) = \log\det X$ is $\nabla f(X) = X^{-1}$ (see §A.4.1).

(d) Suppose A is specified on its tridiagonal part, *i.e.*, we are given $A_{11}, \ldots, A_{nn}$ and $A_{12}, \ldots, A_{n-1,n}$. Show that if there exists a positive definite completion of A, then there is a positive definite completion whose inverse is tridiagonal.

Solution.

(a) If a diagonal entry, say A_{ii}, were not specified, then we would take it to be infinitely large, *i.e.*, we would take $A_{ii} \to \infty$. Then, the condition that $A \succeq 0$ reduces to $\tilde{A} \succeq 0$, where $\tilde{A}$ is the matrix A with ith row and column removed. Repeating this procedure for each unspecified diagonal entry of A, we see that we can just as well consider the submatrix of A corresponding to rows and columns with specified diagonal entries.

(b) The problem is evidently an LMI, since A is clearly an affine function of its unspecified entries, and we require $A \succeq 0$.

(c) We can just as well minimize $f(A) = -\log\det A$, which is a strictly convex function of A (provided $A \succ 0$. Since the objective is strictly convex, there is at most one optimum point. The objective grows unboundedly as A approaches the boundary of the positive definite set, and the set of feasible entries for the matrix is bounded (since the diagonal entries are fixed, and for a matrix to be positive definite, no entry can exceed the maximum diagonal entry). Therefore, there is exactly one minimizer of $-\log\det A$, and it occurs away from the boundary. The optimality condition is simple: it is that the gradient vanishes. Now suppose the i, j entry of A is unspecified (*i.e.*, a variable). Then we have, at the optimal $A^\star$,

$$\frac{\partial f}{\partial A_{ij}} = 2\,\mathbf{tr}(A^\star)^{-1}E_{ij} = 0.$$

But this is nothing more than twice the i, j entry of $(A^\star)^{-1}$. Thus, all entries of $(A^\star)^{-1}$ corresponding to unspecified entries in A must vanish.

(d) The maximum determinant positive definite completion will be tridiagonal, by part (c).

4.48 *Generalized eigenvalue minimization.* Recall (from example 3.37, or §A.5.3) that the largest generalized eigenvalue of a pair of matrices $(A, B) \in \mathbf{S}^k \times \mathbf{S}^k_{++}$ is given by

$$\lambda_{\max}(A, B) = \sup_{u \neq 0} \frac{u^T A u}{u^T B u} = \max\{\lambda \mid \det(\lambda B - A) = 0\}.$$

As we have seen, this function is quasiconvex (if we take $\mathbf{S}^k \times \mathbf{S}^k_{++}$ as its domain).
We consider the problem

$$\text{minimize} \quad \lambda_{\max}(A(x), B(x)) \tag{4.73}$$

where $A, B : \mathbf{R}^n \to \mathbf{S}^k$ are affine functions, defined as

$$A(x) = A_0 + x_1A_1 + \cdots + x_nA_n, \qquad B(x) = B_0 + x_1B_1 + \cdots + x_nB_n.$$

with $A_i, B_i \in \mathbf{S}^k$.

(a) Give a family of convex functions $\phi_t : \mathbf{S}^k \times \mathbf{S}^k \to \mathbf{R}$, that satisfy

$$\lambda_{\max}(A, B) \leq t \iff \phi_t(A, B) \leq 0$$

for all $(A, B) \in \mathbf{S}^k \times \mathbf{S}^k_{++}$. Show that this allows us to solve (4.73) by solving a sequence of convex feasibility problems.

(b) Give a family of matrix-convex functions $\Phi_t : \mathbf{S}^k \times \mathbf{S}^k \to \mathbf{S}^k$ that satisfy

$$\lambda_{\max}(A, B) \leq t \iff \Phi_t(A, B) \preceq 0$$

for all $(A, B) \in \mathbf{S}^k \times \mathbf{S}^k_{++}$. Show that this allows us to solve (4.73) by solving a sequence of convex feasibility problems with LMI constraints.

(c) Suppose $B(x) = (a^Tx+b)I$, with $a \neq 0$. Show that (4.73) is equivalent to the convex problem

$$\begin{array}{ll} \text{minimize} & \lambda_{\max}(sA_0 + y_1A_1 + \cdots + y_nA_n) \\ \text{subject to} & a^Ty + bs = 1 \\ & s \geq 0, \end{array}$$

with variables $y \in \mathbf{R}^n$, $s \in \mathbf{R}$.

Solution.

(a) Take $\phi_t(A, B) = \lambda_{\max}(A - tB)$. $f_0(A, B) \leq t$ if and only if

$$\begin{array}{rcl} B^{-1/2}AB^{-1/2} \preceq tI & \iff & tB - A \succeq 0 \\ & \iff & \lambda_{\max}(A - tB) \leq 0. \end{array}$$

(b) Take $\Phi_t(A, B) = A - tB$.

(c) We will refer to the generalized eigenvalue minimization problem as the GEVP, and to the eigenvalue optimization problem as the EVP.

The GEVP is feasible because $a \neq 0$, so there exist x with $a^Tx + b > 0$.

Suppose x is feasible for the GEVP. Then

$$y = (1/(a^Tx + b))x, \qquad s = 1/(a^Tx + b)$$

is feasible for the EVP ($a^Ty + bs = 1$ and $s \geq 0$). The objective value of (y, s) in the EVP is equal to the objective value of x in the GEVP:

$$\lambda_{\max}\left(\frac{1}{a^Tx + b)}(A_0 + x_1A_1 + \cdots + x_nA_n)\right) = \lambda_{\max}(A(x), (a^Tx + b)I).$$

Conversely, suppose y, s are feasible for the EVP. If $s \neq 0$, then $x = y/s$ satisfies $a^Tx + b = 1/s > 0$, so x is feasible for the GEVP. Moreover,

$$\lambda_{\max}(A(x), (a^Tx + b)I) = \lambda_{\max}(\frac{1}{(a^Tx + b)}A(x)) = \lambda_{\max}(sA_0 + y_1A_1 + \cdots + y_nA_n),$$

i.e., the objective values are the same.

If y, s are feasible for the EVP with $s = 0$, then for all $\hat{x}$ with $a^T\hat{x} + b > 0$, $a^T(\hat{x}+ty)+b = a^T\hat{x}+b+t > 0$, so $x = \hat{x} = ty$ is feasible in the GEVP for all $t \geq 0$. The objective value of x is

$$\begin{aligned}\lambda_{\max}(A(\hat{x}+ty),(a^T(x_0+ty)+b)I) &= \sup_{u\neq 0}\frac{u^T(A(\hat{x})+t(y_1A_1+\cdots+y_nA_n))u}{(a^T\hat{x}+b+t)u^Tu}\\ &\to \sup_{u\neq 0}\frac{tu^T(y_1A_1+\cdots+y_nA_n))u}{tu^Tu}\\ &= \lambda_{\max}(y_1A_1+\cdots+y_nA_n)\end{aligned}$$

so there are feasible points in the GEVP with objective values arbitrarily close to the objective value of y, s in the EVP.

We conclude that the optimal values of the EVP and the GEVP are equal.

4.49 *Generalized fractional programming.* Let $K \in \mathbf{R}^m$ be a proper cone. Show that the function $f_0 : \mathbf{R}^n \to \mathbf{R}^m$, defined by

$$f_0(x) = \inf\{t \mid Cx + d \preceq_K t(Fx + g)\}, \qquad \mathbf{dom}\, f_0 = \{x \mid Fx + g \succ_K 0\},$$

with $C, F \in \mathbf{R}^{m\times n}$, $d, g \in \mathbf{R}^m$, is quasiconvex.

A quasiconvex optimization problem with objective function of this form is called a *generalized fractional program*. Express the generalized linear-fractional program of page 152 and the generalized eigenvalue minimization problem (4.73) as generalized fractional programs.

Solution.

(a) $f_0(x) \leq \alpha$ if and only if $Cx + d \preceq_K \alpha(Fx + g)$ and $Fx + g \succ_K 0$.
To see this, we first note that if $Cx + d \preceq_K \alpha(Fx + g)$, and $Fx + g \succ_K 0$, then obviously $f_0(x) \leq \alpha$.
Conversely, if $f_0(x) \leq \alpha$ and $Fx + g \succ_K 0$, then $Cx + d \preceq_K \hat{t}(Fx + g)$ for at least one $\hat{t} \leq \alpha$, and therefore (since $Fx + g \succ_K 0$),

$$Cx + d \preceq_K t(Fx + g)$$

for all $t \geq \hat{t}$. In particular, $Cx + d \preceq_K \alpha(Fx + g)$.

(b) Choose $K = \mathbf{R}^r_+$.

$$Cx + d \preceq t(Fx + g),\ Fx + g \succ 0 \iff t \geq \max_i \frac{c_i^Tx + d_i}{f_i^Tx + g_i}.$$

(c) Choose $K \in \mathbf{S}^k_+$.

$$A(x) \preceq tB(x),\ B(x) \succ 0 \iff \lambda_{\max}(A(x), B(x)) \leq t.$$

Vector and multicriterion optimization

4.50 *Bi-criterion optimization.* Figure 4.11 shows the optimal trade-off curve and the set of achievable values for the bi-criterion optimization problem

$$\text{minimize (w.r.t. } \mathbf{R}^2_+) \quad (\|Ax - b\|^2, \|x\|_2^2),$$

for some $A \in \mathbf{R}^{100\times 10}$, $b \in \mathbf{R}^{100}$. Answer the following questions using information from the plot. We denote by x_{ls} the solution of the least-squares problem

$$\text{minimize} \quad \|Ax - b\|_2^2.$$

(a) What is $\|x_{\rm ls}\|_2$?

(b) What is $\|Ax_{\rm ls} - b\|_2$?

(c) What is $\|b\|_2$?

(d) Give the optimal value of the problem

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|_2^2 \\ \text{subject to} & \|x\|_2^2 = 1. \end{array}$$

(e) Give the optimal value of the problem

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|_2^2 \\ \text{subject to} & \|x\|_2^2 \leq 1. \end{array}$$

(f) Give the optimal value of the problem

$$\text{minimize } \|Ax - b\|_2^2 + \|x\|_2^2.$$

(g) What is the rank of A?

Solution.

(a) $\|x_{\rm ls}\|_2 = 3$.

(b) $\|Ax_{\rm ls} - b\|_2^2 = 2$.

(c) $\|b\|_2 = \sqrt{10}$.

(d) About 5.

(e) About 5.

(f) About $3 + 4$.

(g) $\mathbf{rank}\, A = 10$, since the LS solution is unique.

4.51 *Monotone transformation of objective in vector optimization.* Consider the vector optimization problem (4.56). Suppose we form a new vector optimization problem by replacing the objective f_0 with $\phi \circ f_0$, where $\phi : \mathbf{R}^q \to \mathbf{R}^q$ satisfies

$$u \preceq_K v,\ u \neq v \Longrightarrow \phi(u) \preceq_K \phi(v),\ \phi(u) \neq \phi(v).$$

Show that a point x is Pareto optimal (or optimal) for one problem if and only if it is Pareto optimal (optimal) for the other, so the two problems are equivalent. In particular, composing each objective in a multicriterion problem with an increasing function does not affect the Pareto optimal points.

Solution. Follows from

$$f_0(x) \preceq_K f_0(y) \iff \phi(f_0(x)) \preceq_K \phi(f_0(y))$$

with equality only if $f_0(x) = f_0(y)$.

4.52 *Pareto optimal points and the boundary of the set of achievable values.* Consider a vector optimization problem with cone K. Let $\mathcal{P}$ denote the set of Pareto optimal values, and let $\mathcal{O}$ denote the set of achievable objective values. Show that $\mathcal{P} \subseteq \mathcal{O} \cap \mathbf{bd}\, \mathcal{O}$, *i.e.*, every Pareto optimal value is an achievable objective value that lies in the boundary of the set of achievable objective values.

Solution. $\mathcal{P} \subseteq \mathcal{O}$, because that is part of the definition of Pareto optimal points. Suppose $f_0(x) \in \mathcal{P}$, $f_0(x) \in \mathbf{int}\, O$. Then $f_0(x) + z \in \mathcal{O}$ for all sufficiently small z, including small values of $z \prec_K 0$. This means that $f_0(x)$ is not a Pareto optimal value.

4.53 Suppose the vector optimization problem (4.56) is convex. Show that the set

$$\mathcal{A} = \mathcal{O} + K = \{t \in \mathbf{R}^q \mid f_0(x) \preceq_K t \text{ for some feasible } x\},$$

is convex. Also show that the minimal elements of $\mathcal{A}$ are the same as the minimal points of $\mathcal{O}$.

Solution. If $f_0(x_1) \preceq_K t_1$ and $f_0(x_2) \preceq_K t_2$ for feasible x_1, x_2, then for $0 \leq \theta \leq 1$, $\theta x_1 + (1-\theta)x_2$ is feasible, and

$$\begin{array}{lll} f_0(\theta x_1 + (1-\theta)x_2) & \preceq_K & \theta f_0(x_1) + (1-\theta) f_0(y_1) \\ & \preceq_K & \theta t_1 + (1-\theta) t_2, \end{array}$$

i.e., $\theta t_1 + (1-\theta)t_2 \in \mathcal{A}$.

Suppose u is minimal for $\mathcal{A}$, *i.e.*,

$$v \in \mathcal{A},\ v \preceq_K u \Longrightarrow v = u.$$

We can express u as $u = \hat{u} + z$, where $\hat{u} \in \mathcal{O}$ and $z \succeq_K 0$. We must have $z = 0$, otherwise the point $v = \hat{u} + z/2 \in \mathcal{A}$, $v \preceq_K u$ and $v \neq u$. In other words, $u \in \mathcal{O}$. Furthermore, u is minimal in $\mathcal{O}$, because

$$v \in \mathcal{O},\ v \preceq_K u \Longrightarrow v \in \mathcal{A},\ v \preceq_K u \Longrightarrow v = u.$$

Conversely, suppose u is minimal for $\mathcal{O}$, *i.e.*,

$$v \in \mathcal{O},\ v \preceq_K u \Longrightarrow v = u.$$

Then for all $v = \hat{v} + z \in \mathcal{A}$, with $\hat{v} \in \mathcal{O}$, $z \succeq_K 0$,

$$\begin{array}{lll} \hat{v} + z \preceq_K u,\ \hat{v} \in \mathcal{O},\ z \succeq_K 0 & \Longrightarrow & \hat{v} \preceq_K u,\ \hat{v} \in \mathcal{O} \\ & \Longrightarrow & \hat{v} = u,\ z = 0. \end{array}$$

4.54 *Scalarization and optimal points.* Suppose a (not necessarily convex) vector optimization problem has an optimal point $x^\star$. Show that $x^\star$ is a solution of the associated scalarized problem for any choice of $\lambda \succ_{K^*} 0$. Also show the converse: If a point x is a solution of the scalarized problem for any choice of $\lambda \succ_{K^*} 0$, then it is an optimal point for the (not necessarily convex) vector optimization problem.

Solution. Follows from the dual characterization of minimum elements in §2.6.3: $f_0(x^\star)$ is the minimum element of the achievable set $\mathcal{O}$, if and only if for all $\lambda \succ_{K^*} 0$, $\lambda^T f_0(x^\star)$ is the unique minimizer of $\lambda^T z$ over $\mathcal{O}$.

4.55 *Generalization of weighted-sum scalarization.* In §4.7.4 we showed how to obtain Pareto optimal solutions of a vector optimization problem by replacing the vector objective $f_0 : \mathbf{R}^n \to \mathbf{R}^q$ with the scalar objective $\lambda^T f_0$, where $\lambda \succ_{K^*} 0$. Let $\psi : \mathbf{R}^q \to \mathbf{R}$ be a K-increasing function, *i.e.*, satisfying

$$u \preceq_K v,\ u \neq v \Longrightarrow \psi(u) < \psi(v).$$

Show that any solution of the problem

$$\begin{array}{lll} \text{minimize} & \psi(f_0(x)) & \\ \text{subject to} & f_i(x) \leq 0, & i = 1, \ldots, m \\ & h_i(x) = 0, & i = 1, \ldots, p \end{array}$$

is Pareto optimal for the vector optimization problem

$$\begin{array}{lll} \text{minimize (w.r.t. } K) & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0, & i = 1, \ldots, m \\ & h_i(x) = 0, & i = 1, \ldots, p. \end{array}$$

Note that $\psi(u) = \lambda^T u$, where $\lambda \succ_{K^*} 0$, is a special case.

As a related example, show that in a multicriterion optimization problem (*i.e.*, a vector optimization problem with $f_0 = F : \mathbf{R}^n \to \mathbf{R}^q$, and $K = \mathbf{R}^q_+$), a *unique* solution of the scalar optimization problem

$$\begin{array}{ll} \text{minimize} & \max_{i=1,\ldots,q} F_i(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p, \end{array}$$

is Pareto optimal.

Solution. Suppose $x^\star$ is a solution of the scalar problem. Now, suppose

$$u \in \mathcal{O}, \qquad u \preceq_K f_0(x^\star), \qquad u \neq f_0(x^\star).$$

Because ψ is increasing, $\psi(u) < \psi(f_0(x^\star))$. However, this contradicts the fact that $x^\star$ is minimizes $\psi \circ f_0$.

Miscellaneous problems

4.56 [P. Parrilo] We consider the problem of minimizing the convex function $f_0 : \mathbf{R}^n \to \mathbf{R}$ over the convex hull of the union of some convex sets, $\mathbf{conv}\left(\bigcup_{i=1}^q C_i\right)$. These sets are described via convex inequalities,

$$C_i = \{x \mid f_{ij}(x) \leq 0, \ j = 1, \ldots, k_i\},$$

where $f_{ij} : \mathbf{R}^n \to \mathbf{R}$ are convex. Our goal is to formulate this problem as a convex optimization problem.

The obvious approach is to introduce variables $x_1, \ldots, x_q \in \mathbf{R}^n$, with $x_i \in C_i$, $\theta \in \mathbf{R}^q$ with $\theta \succeq 0$, $\mathbf{1}^T\theta = 1$, and a variable $x \in \mathbf{R}^n$, with $x = \theta_1 x_1 + \cdots + \theta_q x_q$. This equality constraint is not affine in the variables, so this approach does not yield a convex problem.

A more sophisticated formulation is given by

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & s_i f_{ij}(z_i/s_i) \leq 0, \quad i = 1, \ldots, q, \quad j = 1, \ldots, k_i \\ & \mathbf{1}^T s = 1, \quad s \succeq 0 \\ & x = z_1 + \cdots + z_q, \end{array}$$

with variables $z_1, \ldots, z_q \in \mathbf{R}^n$, $x \in \mathbf{R}^n$, and $s_1, \ldots, s_q \in \mathbf{R}$. (When $s_i = 0$, we take $s_i f_{ij}(z_i/s_i)$ to be 0 if $z_i = 0$ and ∞ if $z_i \neq 0$.) Explain why this problem is convex, and equivalent to the original problem.

Solution. Since f_{ij} are convex functions, so are the perspectives $s_i f_{ij}(z_i/s_i)$. Thus the problem is convex.

Now we show it is equivalent to the original problem. First, suppose that x is feasible for the original problem, and can be expressed as $x = \theta_1 x_1 + \cdots + \theta_q x_q$, where $x_i \in C_i$, and $\theta \succeq 0$, $\mathbf{1}^T\theta = 1$. Define $z_i = \theta_i x_i$, and $s_i = \theta_i$. We claim that $z_1, \ldots, z_q$, $s_1, \ldots, s_q$, x are feasible for the reformulated problem. Clearly we have $x = z_1 + \cdots + z_q$, and $s \succeq 0$, $\mathbf{1}^T s = 1$. For $s_i > 0$, we have $z_i/s_i = x_i \in C_i$, so

$$f_{ij}(z_i/s_i) \leq 0, \quad j = 1, \ldots, k_i.$$

Multiplying by s_i yields the inequalities in the reformulated problem. For $s_i = 0$, the inequalities hold since we take $s_i f_{ij}(z_i/s_i) = 0$.

Conversely, let $z_1, \ldots, z_q$, $s_1, \ldots, s_q$, x be feasible for the reformulated problem. When $s_i = 0$, we must also have $z_i = 0$, so we can ignore these, and assume without loss of generality that all $s_i > 0$. Define $x_i = z_i/s_i$. Dividing the inequalities

$$f_{ij}(z_i/s_i) \leq 0, \quad j = 1, \ldots, k_i$$

by s_i yields

$$f_{ij}(x_i) \leq 0, \quad j = 1, \ldots, k_i,$$

which shows $x_i \in C_i$. From

$$x = z_1 + \cdots + z_q = s_1 x_1 + \cdots + s_q x_q$$

we see that x is a convex combination of $x_1, \ldots, x_q$, and therefore is feasible for the original problem.

It follows that the two problems are equivalent.

4.57 *Capacity of a communication channel.* We consider a communication channel, with input $X(t) \in \{1, \ldots, n\}$, and output $Y(t) \in \{1, \ldots, m\}$, for $t = 1, 2, \ldots$ (in seconds, say). The relation between the input and the output is given statistically:

$$p_{ij} = \mathbf{prob}(Y(t) = i | X(t) = j), \quad i = 1, \ldots, m, \quad j = 1, \ldots, n.$$

The matrix $P \in \mathbf{R}^{m \times n}$ is called the *channel transition matrix*, and the channel is called a *discrete memoryless channel*.

A famous result of Shannon states that information can be sent over the communication channel, with arbitrarily small probability of error, at any rate less than a number C, called the *channel capacity*, in bits per second. Shannon also showed that the capacity of a discrete memoryless channel can be found by solving an optimization problem. Assume that X has a probability distribution denoted $x \in \mathbf{R}^n$, *i.e.*,

$$x_j = \mathbf{prob}(X = j), \quad j = 1, \ldots, n.$$

The *mutual information* between X and Y is given by

$$I(X; Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} x_j p_{ij} \log_2 \frac{p_{ij}}{\sum_{k=1}^{n} x_k p_{ik}}.$$

Then the channel capacity C is given by

$$C = \sup_x I(X; Y),$$

where the supremum is over all possible probability distributions for the input X, *i.e.*, over $x \succeq 0$, $\mathbf{1}^T x = 1$.

Show how the channel capacity can be computed using convex optimization.

Hint. Introduce the variable $y = Px$, which gives the probability distribution of the output Y, and show that the mutual information can be expressed as

$$I(X; Y) = c^T x - \sum_{i=1}^{m} y_i \log_2 y_i,$$

where $c_j = \sum_{i=1}^{m} p_{ij} \log_2 p_{ij}$, $j = 1, \ldots, n$.

Solution. The capacity is the optimal value of the problem

$$\begin{array}{ll} \text{maximize} & f_0(x) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_j p_{ij} \log \frac{p_{ij}}{\sum_{k=1}^{m} x_k p_{ik}} \\ \text{subject to} & x \succeq 0, \quad \mathbf{1}^T x = 1, \end{array}$$

with variable x. It is possible to argue directly that the objective f_0 (which is the mutual information between X and Y) is concave in x. This can be done several ways, starting from the example 3.19.

Another (related) approach is to follow the hint given, and introduce $y = Px$ as another variable. We can express the mutual information in terms of x and y as

$$\begin{aligned} I(X;Y) &= \sum_{i,j} x_j p_{ij} \log \frac{p_{ij}}{\sum_k x_k p_{ik}} \\ &= \sum_j x_j \sum_i p_{ij} \log p_{ij} - \sum_i y_i \log y_i \\ &= -c^T x - \sum_i y_i \log y_i, \end{aligned}$$

where $c_j = -\sum_i p_{ij} \log p_{ij}$. Therefore the channel capacity problem can be expressed as

$$\begin{array}{ll} \text{maximize} & I(X;Y) = -c^T x - \sum_i y_i \log y_i \\ \text{subject to} & x \succeq 0, \quad \mathbf{1}^T x = 1 \\ & y = Px, \end{array}$$

with variables x and y. The objective is a constant plus the entropy of y, hence concave, so this is a convex optimization problem.

4.58 *Optimal consumption.* In this problem we consider the optimal way to consume (or spend) an initial amount of money (or other asset) k_0 over time. The variables are $c_1, \ldots, c_T$, where $c_t \geq 0$ denotes the *consumption* in period t. The utility derived from a consumption level c is given by $u(c)$, where $u : \mathbf{R} \to \mathbf{R}$ is an increasing concave function. The present value of the utility derived from the consumption is given by

$$U = \sum_{t=1}^{T} \beta^t u(c_t),$$

where $0 < \beta < 1$ is a *discount factor*.

Let k_t denote the amount of money available for investment in period t. We assume that it earns an investment return given by $f(k_t)$, where $f : \mathbf{R} \to \mathbf{R}$ is an increasing, concave *investment return function*, which satisfies $f(0) = 0$. For example if the funds earn simple interest at rate R percent per period, we have $f(a) = (R/100)a$. The amount to be consumed, *i.e.*, c_t, is withdrawn at the end of the period, so we have the recursion

$$k_{t+1} = k_t + f(k_t) - c_t, \quad t = 0, \ldots, T.$$

The initial sum $k_0 > 0$ is given. We require $k_t \geq 0$, $t = 1, \ldots, T+1$ (but more sophisticated models, which allow $k_t < 0$, can be considered).

Show how to formulate the problem of maximizing U as a convex optimization problem. Explain how the problem you formulate is equivalent to this one, and exactly how the two are related.

Hint. Show that we can replace the recursion for k_t given above with the inequalities

$$k_{t+1} \leq k_t + f(k_t) - c_t, \quad t = 0, \ldots, T.$$

(Interpretation: the inequalities give you the option of throwing money away in each period.) For a more general version of this trick, see exercise 4.6.

Solution. We start with the problem

$$\begin{array}{ll} \text{maximize} & U = \sum_{t=1}^{T} \beta^t u(c_t) \\ \text{subject to} & k_{t+1} = k_t + f(k_t) - c_t, \ t = 0, \ldots, T \\ & k_t \geq 0, \ t = 1, \ldots, T+1, \end{array}$$

with variables $c_1, \ldots, c_T$ and $k_1, \ldots, k_{T+1}$. The objective is concave, since it is a positive weighted sum of concave functions. But the budget recursion constraints are *not* convex,

since they are *equality* constraints involving the (possibly) nonlinear function f. The hint explains what to do: we look instead at the modified problem

$$\begin{array}{ll} \text{maximize} & U = \sum_{t=1}^{T} \beta^t u(c_t) \\ \text{subject to} & k_{t+1} \leq k_t + f(k_t) - c_t, \ t = 0, \ldots, T \\ & k_t \geq 0, \ t = 1, \ldots, T+1. \end{array}$$

This problem *is* convex, since the budget inequalities can be written as

$$k_{t+1} - k_t - f(k_t) + c_t \leq 0,$$

where the lefthand side is a convex function of the variables c and k.

We will now show that when we solve the modified problem with the inequality constraints, for any optimal solution we actually get *equality* for each of the budget constraints. This means that the solution of the modified problem is actually optimal for the original problem as well. To see this, we note that by changing the equality constraints into inequalities, we are *relaxing* the constraints (*i.e.*, making them looser), and therefore, if anything, we improve the objective compared to the original problem.

Let $c^\star$ and $k^\star$ be optimal for the modified problem. Suppose that at some period s, we have

$$k^\star_{s+1} < k^\star_s + f(k^\star_s) - c^\star_s.$$

This looks pretty suspicious, since it means that in period t, we are actually throwing away money (*i.e.*, we are not investing or consuming all of our available funds). Now consider a new consumption stream $\tilde{c}$ defined as

$$\tilde{c}_t = \begin{cases} c^\star_t & t \neq s \\ c^\star_t + \epsilon & t = s \end{cases}$$

where ϵ is a small positive number such that

$$k^\star_{s+1} \leq k^\star_s + f(k^\star_s) - c^\star_s$$

holds. In words, $\tilde{c}$ is the same consumption stream as $c^\star$, except in the period when we throw away some money (in $c^\star$) we just consume a little more. Clearly we have $U(\tilde{c}) \geq U(c^\star)$, since the two streams consume the same amount for every period except one, in which we consume more with $\tilde{c}$. (Here we use the fact that U is increasing.)

Let $\tilde{k}$ be the asset stream that results from the consumption stream $\tilde{c}$. Then all the constraints of the original problem are satisfied for $\tilde{c}$ and $\tilde{k}$, and yet $c^\star$ has a lower objective value than $\tilde{c}$. That contradicts optimality of $c^\star$. We conclude that for $c^\star$, we have

$$k^\star_{t+1} = k^\star_t + f(k^\star_t) - c^\star_t.$$

4.59 *Robust optimization.* In some optimization problems there is uncertainty or variation in the objective and constraint functions, due to parameters or factors that are either beyond our control or unknown. We can model this situation by making the objective and constraint functions $f_0, \ldots, f_m$ functions of the optimization variable $x \in \mathbf{R}^n$ *and* a parameter vector $u \in \mathbf{R}^k$ that is unknown, or varies. In the *stochastic optimization* approach, the parameter vector u is modeled as a random variable with a known distribution, and we work with the expected values $\mathbf{E}_u\, f_i(x, u)$. In the *worst-case analysis* approach, we are given a set U that u is known to lie in, and we work with the maximum or worst-case values $\sup_{u \in U} f_i(x, u)$. To simplify the discussion, we assume there are no equality constraints.

(a) *Stochastic optimization.* We consider the problem

$$\begin{array}{ll} \text{minimize} & \mathbf{E}\, f_0(x, u) \\ \text{subject to} & \mathbf{E}\, f_i(x, u) \leq 0, \quad i = 1, \ldots, m, \end{array}$$

where the expectation is with respect to u. Show that if f_i are convex in x for each u, then this stochastic optimization problem is convex.

(b) *Worst-case optimization.* We consider the problem

$$\begin{array}{ll} \text{minimize} & \sup_{u\in U} f_0(x,u) \\ \text{subject to} & \sup_{u\in U} f_i(x,u) \le 0, \quad i=1,\ldots,m. \end{array}$$

Show that if f_i are convex in x for each u, then this worst-case optimization problem is convex.

(c) *Finite set of possible parameter values.* The observations made in parts (a) and (b) are most useful when we have analytical or easily evaluated expressions for the expected values $\mathbf{E}\, f_i(x,u)$ or the worst-case values $\sup_{u\in U} f_i(x,u)$.

Suppose we are given the set of possible values of the parameter is finite, *i.e.*, we have $u \in \{u_1,\ldots,u_N\}$. For the stochastic case, we are also given the probabilities of each value: $\mathbf{prob}(u=u_i)=p_i$, where $p\in\mathbf{R}^N$, $p\succeq 0$, $\mathbf{1}^T p=1$. In the worst-case formulation, we simply take $U\in\{u_1,\ldots,u_N\}$.

Show how to set up the worst-case and stochastic optimization problems explicitly (*i.e.*, give explicit expressions for $\sup_{u\in U} f_i$ and $\mathbf{E}_u\, f_i$).

Solution.

(a) Follows from the fact that the inequality

$$f_i(\theta x+(1-\theta)y,u) \le \theta f(x,u)+(1-\theta)f(y,u)$$

is preserved when we take expectations on both sides.

(b) If $f_i(x,u)$ is convex in x for fixed u, then $\sup_u f_i(x,u)$ is convex in x.

(c) Stochastic formulation:

$$\begin{array}{ll} \text{minimize} & \sum_i p_k f_0(x,u_k) \\ \text{subject to} & \sum_k p_k f_i(x,u_k) \le 0, \quad i=1,\ldots,m. \end{array}$$

Worst-case formulation:

$$\begin{array}{ll} \text{minimize} & \max_k f_0(x,u_k) \\ \text{subject to} & \max_k f_i(x,u_k) \le 0, \quad i=1,\ldots,m. \end{array}$$

4.60 *Log-optimal investment strategy.* We consider a portfolio problem with n assets held over N periods. At the beginning of each period, we re-invest our total wealth, redistributing it over the n assets using a fixed, constant, allocation strategy $x\in\mathbf{R}^n$, where $x\succeq 0$, $\mathbf{1}^T x=1$. In other words, if $W(t-1)$ is our wealth at the beginning of period t, then during period t we invest $x_i W(t-1)$ in asset i. We denote by $\lambda(t)$ the total return during period t, *i.e.*, $\lambda(t)=W(t)/W(t-1)$. At the end of the N periods our wealth has been multiplied by the factor $\prod_{t=1}^N \lambda(t)$. We call

$$\frac{1}{N}\sum_{t=1}^N \log\lambda(t)$$

the *growth rate* of the investment over the N periods. We are interested in determining an allocation strategy x that maximizes growth of our total wealth for large N.

We use a discrete stochastic model to account for the uncertainty in the returns. We assume that during each period there are m possible scenarios, with probabilities π_j, $j=1,\ldots,m$. In scenario j, the return for asset i over one period is given by p_{ij}.

Therefore, the return $\lambda(t)$ of our portfolio during period t is a random variable, with m possible values $p_1^T x, \ldots, p_m^T x$, and distribution

$$\pi_j = \mathbf{prob}(\lambda(t) = p_j^T x), \quad j = 1, \ldots, m.$$

We assume the same scenarios for each period, with (identical) independent distributions. Using the law of large numbers, we have

$$\lim_{N\to\infty} \frac{1}{N} \log\left(\frac{W(N)}{W(0)}\right) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \log \lambda(t) = \mathbf{E} \log \lambda(t) = \sum_{j=1}^{m} \pi_j \log(p_j^T x).$$

In other words, with investment strategy x, the long term growth rate is given by

$$R_{\mathrm{lt}} = \sum_{j=1}^{m} \pi_j \log(p_j^T x).$$

The investment strategy x that maximizes this quantity is called the *log-optimal investment strategy*, and can be found by solving the optimization problem

$$\begin{array}{ll} \text{maximize} & \sum_{j=1}^{m} \pi_j \log(p_j^T x) \\ \text{subject to} & x \succeq 0, \quad \mathbf{1}^T x = 1, \end{array}$$

with variable $x \in \mathbf{R}^n$.

Show that this is a convex optimization problem.

Solution. Actually, there's not much to do in this problem. The constraints, $x \succeq 0$, $\mathbf{1}^T x = 1$, are clearly convex, so we just need to show that the objective is concave (since it is to be maximized). We can do that in just a few steps: First, note that log is concave, so $\log(p_j^T x)$ is concave in x (on the domain, which is the open halfspace $\{x \mid p_j^T x \succ 0\}$). Since $\pi_j \geq 0$, we conclude that the sum of concave functions

$$\sum_{j=1}^{m} \pi_j \log(p_j^T x)$$

is concave.

4.61 *Optimization with logistic model.* A random variable $X \in \{0, 1\}$ satisfies

$$\mathbf{prob}(X = 1) = p = \frac{\exp(a^T x + b)}{1 + \exp(a^T x + b)},$$

where $x \in \mathbf{R}^n$ is a vector of variables that affect the probability, and a and b are known parameters. We can think of $X = 1$ as the event that a consumer buys a product, and x as a vector of variables that affect the probability, *e.g.*, advertising effort, retail price, discounted price, packaging expense, and other factors. The variable x, which we are to optimize over, is subject to a set of linear constraints, $Fx \preceq g$.

Formulate the following problems as convex optimization problems.

(a) *Maximizing buying probability.* The goal is to choose x to maximize p.

(b) *Maximizing expected profit.* Let $c^T x + d$ be the profit derived from selling the product, which we assume is positive for all feasible x. The goal is to maximize the expected profit, which is $p(c^T x + d)$.

Solution.

(a) The function $e^u/(1+e^u)$ is monotonically increasing in u, so we can maximize $\exp(a^T x+b)/(1+\exp(a^T x+b))$ by maximizing $a^T x+b$, which leads to the LP

$$\begin{array}{ll}\text{maximize} & a^T x+b\\ \text{subject to} & Fx \preceq g.\end{array}$$

(b) Here we have to maximize $p(c^T x+d)$, or equivalently, its logarithm:

$$\begin{array}{ll}\text{maximize} & a^T x+b-\log\left(1+\exp(a^T x+b)\right)+\log(c^T x+d)\\ \text{subject to} & Fx \preceq g.\end{array}$$

This is a convex problem, since the objective is a concave function of x. (Recall that $f(x)=\log\sum_{i=1}^m \exp(a_i^T x+b_i)$ is convex.)

4.62 *Optimal power and bandwidth allocation in a Gaussian broadcast channel.* We consider a communication system in which a central node transmits messages to n receivers. ('Gaussian' refers to the type of noise that corrupts the transmissions.) Each receiver channel is characterized by its (transmit) power level $P_i \geq 0$ and its bandwidth $W_i \geq 0$. The power and bandwidth of a receiver channel determine its *bit rate* R_i (the rate at which information can be sent) via

$$R_i = \alpha_i W_i \log(1+\beta_i P_i/W_i),$$

where α_i and β_i are known positive constants. For $W_i=0$, we take $R_i=0$ (which is what you get if you take the limit as $W_i \to 0$).

The powers must satisfy a total power constraint, which has the form

$$P_1+\cdots+P_n = P_{\text{tot}},$$

where $P_{\text{tot}}>0$ is a given total power available to allocate among the channels. Similarly, the bandwidths must satisfy

$$W_1+\cdots+W_n = W_{\text{tot}},$$

where $W_{\text{tot}}>0$ is the (given) total available bandwidth. The optimization variables in this problem are the powers and bandwidths, *i.e.*, $P_1,\ldots,P_n$, $W_1,\ldots,W_n$.

The objective is to maximize the total utility,

$$\sum_{i=1}^n u_i(R_i),$$

where $u_i : \mathbf{R} \to \mathbf{R}$ is the utility function associated with the ith receiver. (You can think of $u_i(R_i)$ as the revenue obtained for providing a bit rate R_i to receiver i, so the objective is to maximize the total revenue.) You can assume that the utility functions u_i are nondecreasing and concave.

Pose this problem as a convex optimization problem.

Solution. If we substitute the expression for R_i in the objective, we obtain

$$\begin{array}{ll}\text{maximize} & \sum_{i=1}^n u\left(\alpha_i W_i \log(1+\beta_i P_i/W_i)\right)\\ \text{subject to} & \mathbf{1}^T P = P_{\text{tot}}, \quad \mathbf{1}^T W = W_{\text{tot}}\\ & P \succeq 0, \quad W \succeq 0\end{array}$$

with variables $P,\ W \in \mathbf{R}^n$. We show that R_i is a concave function of (P_i, W_i). It will follow that $u(R_i)$ is concave since it is a nondecreasing concave function of a concave function. The total utility U is then concave since it is the sum of concave functions.

To show that R_i is concave in (P_i, W_i) we can derive the Hessian, which is

$$\nabla^2 R_i = \frac{-\alpha_i \beta_i^2}{W_i(1+\beta_i P_i/W_i)^2} \begin{bmatrix} 1 \\ -P_i \end{bmatrix} \begin{bmatrix} 1 \\ -P_i \end{bmatrix}^T.$$

Since α_i, β_i, W_i, and P_i are positive, $\nabla^2 R_i$ is negative semidefinite.

An alternative proof follows fromt the fact that $t\log(1+x/t)$ is concave in (x,t) for $t>0$, since it is the perspective of $\log(1+x)$, and $\log(1+x)$ is concave.

Another approach is to relax the bit-rate equality constraint, and write the problem as

$$\begin{array}{ll} \text{maximize} & U = \sum_{i=1}^n u(R_i) \\ \text{subject to} & R_i \le \alpha_i W_i \log(1+\beta_i P_i/W_i) \\ & \mathbf{1}^T P = P_{\text{tot}}, \quad \mathbf{1}^T W = W_{\text{tot}}, \end{array}$$

with variables P_i, W_i, and R_i. The bit-rate inequality is convex, since the lefthand side is a convex function of the variables (actually, linear), and the righthand side is a concave function of the variables. Since the objective is concave, this is a convex optimization problem. We need to show now is that when we solve this convex optimization problem, we end up with equality in the bit-rate inequality constraints. But this is easy: for each variable R_i, the objective is monotonically increasing in R_i, so we want each R_i are large as possible. Examining the constraints, we see that this occurs when

$$R_i = \alpha_i W_i \log(1+\beta_i P_i/W_i).$$

4.63 *Optimally balancing manufacturing cost and yield.* The vector $x \in \mathbf{R}^n$ denotes the nominal parameters in a manufacturing process. The yield of the process, *i.e.*, the fraction of manufactured goods that is acceptable, is given by $Y(x)$. We assume that Y is log-concave (which is often the case; see example 3.43). The cost per unit to manufacture the product is given by $c^T x$, where $c \in \mathbf{R}^n$. The cost per acceptable unit is $c^T x/Y(x)$. We want to minimize $c^T x/Y(x)$, subject to some convex constraints on x such as a linear inequalities $Ax \preceq b$. (You can assume that over the feasible set we have $c^T x > 0$ and $Y(x) > 0$.)

This problem is *not* a convex or quasiconvex optimization problem, but it can be solved using convex optimization and a one-dimensional search. The basic ideas are given below; you must supply all details and justification.

(a) Show that the function $f : \mathbf{R} \to \mathbf{R}$ given by

$$f(a) = \sup\{Y(x) \mid Ax \preceq b,\ c^T x = a\},$$

which gives the maximum yield versus cost, is log-concave. This means that by solving a convex optimization problem (in x) we can evaluate the function f.

(b) Suppose that we evaluate the function f for enough values of a to give a good approximation over the range of interest. Explain how to use these data to (approximately) solve the problem of minimizing cost per good product.

Solution. We first verify that the objective is not convex or quasiconvex. For $c^T x/Y(x)$ to be quasiconvex, we need the constraint

$$c^T x/Y(x) \le t \iff \log(c^T x) - \log Y(x) \le \log t$$

to be convex. By assumption, $-\log Y(x)$ is convex, but in general we can't assume that the sum with $\log(c^T x)$ is convex.

(a) The function $f(a)$ is log-concave because

$$\log f(a) = \sup_x F(x, a)$$

where

$$F(x,a) = \begin{cases} \log Y(x) & Ax \preceq b, \quad c^T x = a \\ -\infty & \text{otherwise.} \end{cases}$$

F has domain $\{(x, a) \mid Ax \preceq b,\ c^T x = a\}$, which is a convex set. On its domain it is equal to $\log Y(x)$, a concave function. Therefore F is concave, and maximizing over a gives another concave function.

(b) We would like to solve the problem

$$\begin{array}{ll} \text{maximize} & \log(Y(x)/c^T x) \\ \text{subject to} & Ax \preceq b. \end{array}$$

or, equivalently,

$$\begin{array}{ll} \text{maximize} & \log Y(x) - \log a \\ \text{subject to} & Ax \preceq b \\ & c^T x = a, \end{array}$$

with variables x and a. By first optimizing over x and then over a, we can write the problem as

$$\text{maximize} \quad \log f(a) - \log a,$$

with variable a. The objective function is the sum of a concave and a convex function. By evaluating $\log f(a) - \log a$ for a large set of values of a, we can approximately solve the problem.

Another useful observation is as follows. If we evaluate the objective function at some $a = \hat{a}$. This yields not only the value, but also a concave lower bound

$$\begin{aligned} \log f(a) - \log a &\geq \log f(a) - \log \hat{a} - (a - \hat{a})/\hat{a} \\ &= \log f(a) - a/\hat{a} - \log \hat{a} + 1. \end{aligned}$$

By repeatedly maximizing the lower bound and linearizing, we can find a local maximum of $f(a)/a$.

4.64 *Optimization with recourse.* In an optimization problem with recourse, also called *two-stage optimization*, the cost function and constraints depend not only on our choice of variables, but also on a discrete random variable $s \in \{1, \ldots, S\}$, which is interpreted as specifying which of S *scenarios* occurred. The scenario random variable s has known probability distribution π, with $\pi_i = \mathbf{prob}(s = i)$, $i = 1, \ldots, S$.

In two-stage optimization, we are to choose the values of two variables, $x \in \mathbf{R}^n$ and $z \in \mathbf{R}^q$. The variable x must be chosen *before* the particular scenario s is known; the variable z, however, is chosen *after* the value of the scenario random variable is known. In other words, z is a function of the scenario random variable s. To describe our choice z, we list the values we would choose under the different scenarios, *i.e.*, we list the vectors

$$z_1, \ldots, z_S \in \mathbf{R}^q.$$

Here z_3 is our choice of z when $s = 3$ occurs, and so on. The set of values

$$x \in \mathbf{R}^n, \qquad z_1, \ldots, z_S \in \mathbf{R}^q$$

is called the *policy*, since it tells us what choice to make for x (independent of which scenario occurs), and also, what choice to make for z in each possible scenario.

The variable z is called the *recourse variable* (or *second-stage variable*), since it allows us to take some action or make a choice after we know which scenario occurred. In

contrast, our choice of x (which is called the *first-stage variable*) must be made without any knowledge of the scenario.

For simplicity we will consider the case with no constraints. The cost function is given by

$$f : \mathbf{R}^n \times \mathbf{R}^q \times \{1, \ldots, S\} \to \mathbf{R},$$

where $f(x, z, i)$ gives the cost when the first-stage choice x is made, second-stage choice z is made, and scenario i occurs. We will take as the overall objective, to be minimized over all policies, the expected cost

$$\mathbf{E}\, f(x, z_s, s) = \sum_{i=1}^{S} \pi_i f(x, z_i, i).$$

Suppose that f is a convex function of (x, z), for each scenario $i = 1, \ldots, S$. Explain how to find an optimal policy, *i.e.*, one that minimizes the expected cost over all possible policies, using convex optimization.

Solution. The variables in the problem are

$$x, \quad z_1, \ldots, z_q,$$

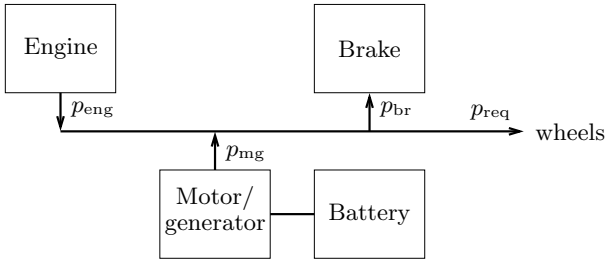
i.e., the policy. The (total) dimension of the variables is $n + Sq$. Our problem is nothing more than

$$\text{minimize} \quad F(x) = \textstyle\sum_{i=1}^{S} \pi_i f(x, z_i, i),$$

which is convex since for each i, $f(x, z, i)$ is convex in (x, z_i), and $\pi_i \geq 0$.

4.65 *Optimal operation of a hybrid vehicle.* A hybrid vehicle has an internal combustion engine, a motor/generator connected to a storage battery, and a conventional (friction) brake. In this exercise we consider a (highly simplified) model of a *parallel hybrid vehicle*, in which both the motor/generator and the engine are directly connected to the drive wheels. The engine can provide power to the wheels, and the brake can take power from the wheels, turning it into heat. The motor/generator can act as a motor, when it uses energy stored in the battery to deliver power to the wheels, or as a generator, when it takes power from the wheels or engine, and uses the power to charge the battery. When the generator takes power from the wheels and charges the battery, it is called *regenerative braking*; unlike ordinary friction braking, the energy taken from the wheels is *stored*, and can be used later. The vehicle is judged by driving it over a known, fixed test track to evaluate its fuel efficiency.

A diagram illustrating the power flow in the hybrid vehicle is shown below. The arrows indicate the direction in which the power flow is considered positive. The engine power p_{eng}, for example, is positive when it is delivering power; the brake power p_{br} is positive when it is taking power from the wheels. The power p_{req} is the required power at the wheels. It is positive when the wheels require power (*e.g.*, when the vehicle accelerates, climbs a hill, or cruises on level terrain). The required wheel power is negative when the vehicle must decelerate rapidly, or descend a hill.



All of these powers are functions of time, which we discretize in one second intervals, with $t = 1, 2, \ldots, T$. The required wheel power $p_{\rm req}(1), \ldots, p_{\rm req}(T)$ is given. (The speed of the vehicle on the track is specified, so together with known road slope information, and known aerodynamic and other losses, the power required at the wheels can be calculated.)

Power is conserved, which means we have

$$p_{\rm req}(t) = p_{\rm eng}(t) + p_{\rm mg}(t) - p_{\rm br}(t), \quad t = 1, \ldots, T.$$

The brake can only dissipate power, so we have $p_{\rm br}(t) \geq 0$ for each t. The engine can only provide power, and only up to a given limit $P_{\rm eng}^{\rm max}$, *i.e.*, we have

$$0 \leq p_{\rm eng}(t) \leq P_{\rm eng}^{\rm max}, \quad t = 1, \ldots, T.$$

The motor/generator power is also limited: $p_{\rm mg}$ must satisfy

$$P_{\rm mg}^{\rm min} \leq p_{\rm mg}(t) \leq P_{\rm mg}^{\rm max}, \quad t = 1, \ldots, T.$$

Here $P_{\rm mg}^{\rm max} > 0$ is the maximum motor power, and $-P_{\rm mg}^{\rm min} > 0$ is the maximum generator power.

The battery charge or energy at time t is denoted $E(t)$, $t = 1, \ldots, T + 1$. The battery energy satisfies

$$E(t+1) = E(t) - p_{\rm mg}(t) - \eta|p_{\rm mg}(t)|, \quad t = 1, \ldots, T + 1,$$

where $\eta > 0$ is a known parameter. (The term $-p_{\rm mg}(t)$ represents the energy removed or added the battery by the motor/generator, ignoring any losses. The term $-\eta|p_{\rm mg}(t)|$ represents energy lost through inefficiencies in the battery or motor/generator.)

The battery charge must be between 0 (empty) and its limit $E_{\rm batt}^{\rm max}$ (full), at all times. (If $E(t) = 0$, the battery is fully discharged, and no more energy can be extracted from it; when $E(t) = E_{\rm batt}^{\rm max}$, the battery is full and cannot be charged.) To make the comparison with non-hybrid vehicles fair, we fix the initial battery charge to equal the final battery charge, so the net energy change is zero over the track: $E(1) = E(T + 1)$. We do not specify the value of the initial (and final) energy.

The objective in the problem is the total fuel consumed by the engine, which is

$$F_{\rm total} = \sum_{t=1}^{T} F(p_{\rm eng}(t)),$$

where $F : \mathbf{R} \to \mathbf{R}$ is the *fuel use characteristic* of the engine. We assume that F is positive, increasing, and convex.

Formulate this problem as a convex optimization problem, with variables $p_{\rm eng}(t)$, $p_{\rm mg}(t)$, and $p_{\rm br}(t)$ for $t = 1, \ldots, T$, and $E(t)$ for $t = 1, \ldots, T + 1$. Explain why your formulation is equivalent to the problem described above.

Solution. We first collect the given objective and constraints to form the problem

$$\begin{array}{ll} \text{minimize} & \sum_{t=1}^{T} F(p_{\rm eng}(t)) \\ \text{subject to} & p_{\rm req}(t) = p_{\rm eng}(t) + p_{\rm mg}(t) - p_{\rm br}(t) \\ & E(t+1) = E(t) - p_{\rm mg}(t)) - \eta|p_{\rm mg}(t))| \\ & 0 \leq E(t) \leq E_{\rm batt}^{\rm max} \\ & E(1) = E(T+1) \\ & 0 \leq p_{\rm eng}(t) \leq P_{\rm eng}^{\rm max} \\ & P_{\rm mg}^{\rm min} \leq p_{\rm mg}(t) \leq P_{\rm mg}^{\rm max} \\ & 0 \leq p_{\rm br}(t), \end{array}$$

where each constraint is imposed for the appropriate range of t. The fuel use function F is convex, so the objective function is convex. With the exception of the battery charge

equations, each constraint is a linear equality or linear inequality. So in this form the problem is *not* convex.

We need to show how to deal with the nonconvex constraints

$$E(t+1) = E(t) - p_{\mathrm{mg}}(t)) - \eta|p_{\mathrm{mg}}(t))|.$$

One approach is to replace this constraint with the relaxation,

$$E(t+1) \leq E(t) - p_{\mathrm{mg}}(t)) - \eta|p_{\mathrm{mg}}(t))|,$$

which is convex, in fact, two linear inequalities. Intuitively, this relaxation means that we open the possibility of throwing energy from the battery away at each step. This sounds like a bad idea, when fuel efficiency is the goal, and indeed, it is easy to see that if we solve the problem with the relaxed battery charge constraints, the optimal $E^\star$ satisfies

$$E^\star(t+1) = E^\star(t) - p_{\mathrm{mg}}(t)) - \eta|p_{\mathrm{mg}}(t))|,$$

and therefore solves the original problem. To argue formally that this is the case, suppose that the solution of the relaxed problem does throw away some energy at some step t. We then construct a new trajectory, where we do not throw away the extra energy, and instead, use the energy to power the wheels, and reduce the engine power. This reduces the fuel consumption since the fuel consumption characteristic is increasing, which shows that the original could not have been optimal.

Chapter 5

Duality

Exercises

Basic definitions

5.1 *A simple example.* Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & x^2 + 1 \\ \text{subject to} & (x-2)(x-4) \leq 0, \end{array}$$

with variable $x \in \mathbf{R}$.

(a) *Analysis of primal problem.* Give the feasible set, the optimal value, and the optimal solution.

(b) *Lagrangian and dual function.* Plot the objective x^2+1 versus x. On the same plot, show the feasible set, optimal point and value, and plot the Lagrangian $L(x,\lambda)$ versus x for a few positive values of λ. Verify the lower bound property ($p^\star \geq \inf_x L(x,\lambda)$ for $\lambda \geq 0$). Derive and sketch the Lagrange dual function g.

(c) *Lagrange dual problem.* State the dual problem, and verify that it is a concave maximization problem. Find the dual optimal value and dual optimal solution $\lambda^\star$. Does strong duality hold?

(d) *Sensitivity analysis.* Let $p^\star(u)$ denote the optimal value of the problem

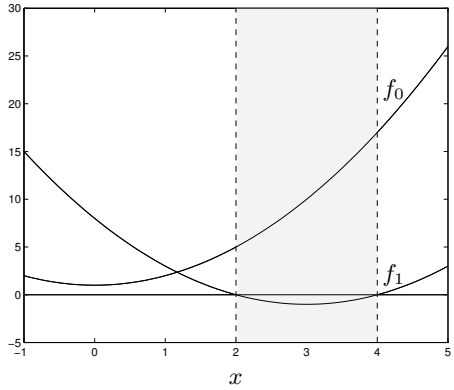
$$\begin{array}{ll} \text{minimize} & x^2 + 1 \\ \text{subject to} & (x-2)(x-4) \leq u, \end{array}$$

as a function of the parameter u. Plot $p^\star(u)$. Verify that $dp^\star(0)/du = -\lambda^\star$.

Solution.

(a) The feasible set is the interval $[2,4]$. The (unique) optimal point is $x^\star = 2$, and the optimal value is $p^\star = 5$.

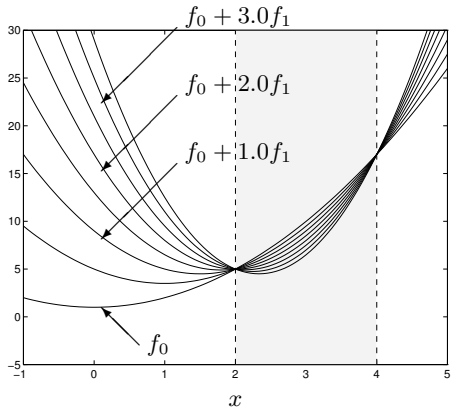
The plot shows f_0 and f_1.



(b) The Lagrangian is

$$L(x,\lambda) = (1+\lambda)x^2 - 6\lambda x + (1+8\lambda).$$

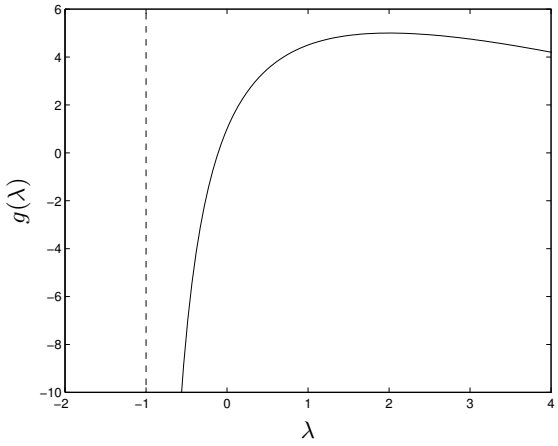
The plot shows the Lagrangian $L(x,\lambda) = f_0 + \lambda f_1$ as a function of x for different values of $\lambda \geq 0$. Note that the minimum value of $L(x,\lambda)$ over x (*i.e.*, $g(\lambda)$) is always less than $p^\star$. It increases as λ varies from 0 toward 2, reaches its maximum at $\lambda = 2$, and then decreases again as λ increases above 2. We have equality $p^\star = g(\lambda)$ for $\lambda = 2$.



For $\lambda > -1$, the Lagrangian reaches its minimum at $\tilde{x} = 3\lambda/(1+\lambda)$. For $\lambda \leq -1$ it is unbounded below. Thus

$$g(\lambda) = \begin{cases} -9\lambda^2/(1+\lambda) + 1 + 8\lambda & \lambda > -1 \\ -\infty & \lambda \leq -1 \end{cases}$$

which is plotted below.



We can verify that the dual function is concave, that its value is equal to $p^\star = 5$ for $\lambda = 2$, and less than $p^\star$ for other values of λ.

(c) The Lagrange dual problem is

$$\begin{array}{ll} \text{maximize} & -9\lambda^2/(1+\lambda) + 1 + 8\lambda \\ \text{subject to} & \lambda \geq 0. \end{array}$$

The dual optimum occurs at $\lambda = 2$, with $d^\star = 5$. So for this example we can directly observe that strong duality holds (as it must — Slater's constraint qualification is satisfied).

(d) The perturbed problem is infeasible for $u < -1$, since $\inf_x(x^2 - 6x + 8) = -1$. For $u \geq -1$, the feasible set is the interval

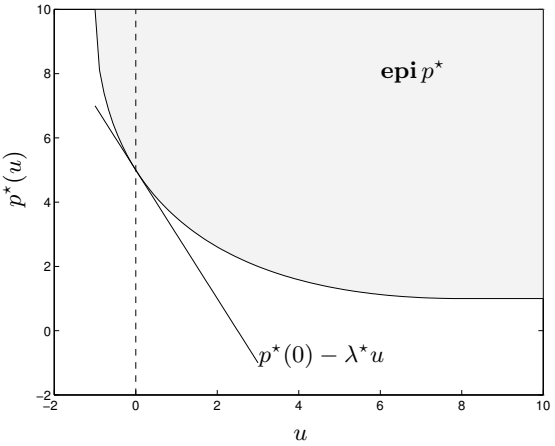
$$[3 - \sqrt{1+u}, 3 + \sqrt{1+u}],$$

given by the two roots of $x^2 - 6x + 8 = u$. For $-1 \leq u \leq 8$ the optimum is $x^\star(u) = 3 - \sqrt{1+u}$. For $u \geq 8$, the optimum is the unconstrained minimum of f_0,

i.e., $x^\star(u) = 0$. In summary,

$$p^\star(u) = \begin{cases} \infty & u < -1 \\ 11 + u - 6\sqrt{1+u} & -1 \le u \le 8 \\ 1 & u \ge 8. \end{cases}$$

The figure shows the optimal value function $p^\star(u)$ and its epigraph.



Finally, we note that $p^\star(u)$ is a differentiable function of u, and that

$$\frac{dp^\star(0)}{du} = -2 = -\lambda^\star.$$

5.2 *Weak duality for unbounded and infeasible problems.* The weak duality inequality, $d^\star \le p^\star$, clearly holds when $d^\star = -\infty$ or $p^\star = \infty$. Show that it holds in the other two cases as well: If $p^\star = -\infty$, then we must have $d^\star = -\infty$, and also, if $d^\star = \infty$, then we must have $p^\star = \infty$.

Solution.

(a) $p^\star = -\infty$. The primal problem is unbounded, *i.e.*, there exist feasible x with arbitrarily small values of $f_0(x)$. This means that

$$L(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x)$$

is unbounded below for all $\lambda \succeq 0$, *i.e.*, $g(\lambda) = -\infty$ for $\lambda \succeq 0$. Therefore the dual problem is infeasible ($d^\star = -\infty$).

(b) $d^\star = \infty$. The dual problem is unbounded above. This is only possible if the primal problem is infeasible. If it were feasible, with $f_i(\tilde{x}) \le 0$ for $i = 1, \ldots, m$, then for all $\lambda \succeq 0$,

$$g(\lambda) = \inf(f_0(x) + \sum_i \lambda_i f_i(x)) \le f_0(\tilde{x}) + \sum_i \lambda_i f_i(\tilde{x}),$$

so the dual problem is bounded above.

5.3 *Problems with one inequality constraint.* Express the dual problem of

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & f(x) \le 0, \end{array}$$

with $c \neq 0$, in terms of the conjugate f^*. Explain why the problem you give is convex. We do not assume f is convex.

Solution. For $\lambda = 0$, $g(\lambda) = \inf c^T x = -\infty$. For $\lambda > 0$,

$$\begin{aligned} g(\lambda) &= \inf(c^T x + \lambda f(x)) \\ &= \lambda \inf((c/\lambda)^T x + \lambda f(x)) \\ &= -\lambda f_1^*(-c/\lambda), \end{aligned}$$

i.e., for $\lambda \geq 0$, $-g$ is the perspective of f_1^*, evaluated at $-c/\lambda$. The dual problem is

$$\begin{array}{ll} \text{minimize} & -\lambda f_1^*(-c/\lambda) \\ \text{subject to} & \lambda \geq 0. \end{array}$$

Examples and applications

5.4 *Interpretation of LP dual via relaxed problems.* Consider the inequality form LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b, \end{array}$$

with $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$. In this exercise we develop a simple geometric interpretation of the dual LP (5.22).

Let $w \in \mathbf{R}^m_+$. If x is feasible for the LP, *i.e.*, satisfies $Ax \preceq b$, then it also satisfies the inequality

$$w^T Ax \leq w^T b.$$

Geometrically, for any $w \succeq 0$, the halfspace $H_w = \{x \mid w^T Ax \leq w^T b\}$ contains the feasible set for the LP. Therefore if we minimize the objective $c^T x$ over the halfspace H_w we get a lower bound on $p^\star$.

(a) Derive an expression for the minimum value of $c^T x$ over the halfspace H_w (which will depend on the choice of $w \succeq 0$).

(b) Formulate the problem of finding the best such bound, by maximizing the lower bound over $w \succeq 0$.

(c) Relate the results of (a) and (b) to the Lagrange dual of the LP, given by (5.22).

Solution.

(a) The optimal value is

$$\inf_{x \in H_w} c^T x = \begin{cases} \lambda w^T b & c = \lambda A^T w \text{ for some } \lambda \leq 0 \\ -\infty & \text{otherwise.} \end{cases}$$

(See exercise 4.8.)

(b) We maximize the lower bound by solving

$$\begin{array}{ll} \text{maximize} & \lambda w^T b \\ \text{subject to} & c = \lambda A^T w \\ & \lambda \leq 0, \quad w \succeq 0 \end{array}$$

with variables λ and w. Note that, as posed, this is not a convex problem.

(c) Defining $z = -\lambda w$, we obtain the equivalent problem

$$\begin{array}{ll} \text{maximize} & -b^T z \\ \text{subject to} & A^T z + c = 0 \\ & z \succeq 0. \end{array}$$

This is the dual of the original LP.

5.5 *Dual of general LP.* Find the dual function of the LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Gx \preceq h \\ & Ax = b. \end{array}$$

Give the dual problem, and make the implicit equality constraints explicit.
Solution.

(a) The Lagrangian is

$$\begin{array}{rcl} L(x, \lambda, \nu) & = & c^T x + \lambda^T (Gx - h) + \nu^T (Ax - b) \\ & = & (c^T + \lambda^T G + \nu^T A)x - h\lambda^T - \nu^T b, \end{array}$$

which is an affine function of x. It follows that the dual function is given by

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu) = \begin{cases} -\lambda^T h - \nu^T b & c + G^T \lambda + A^T \nu = 0 \\ -\infty & \text{otherwise.} \end{cases}$$

(b) The dual problem is

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0. \end{array}$$

After making the implicit constraints explicit, we obtain

$$\begin{array}{ll} \text{maximize} & -\lambda^T h - \nu^T b \\ \text{subject to} & c + G^T \lambda + A^T \nu = 0 \\ & \lambda \succeq 0. \end{array}$$

5.6 *Lower bounds in Chebyshev approximation from least-squares.* Consider the Chebyshev or ℓ_∞-norm approximation problem

$$\text{minimize} \quad \|Ax - b\|_\infty, \tag{5.103}$$

where $A \in \mathbf{R}^{m \times n}$ and $\mathbf{rank}\, A = n$. Let x_{ch} denote an optimal solution (there may be multiple optimal solutions; x_{ch} denotes one of them).
The Chebyshev problem has no closed-form solution, but the corresponding least-squares problem does. Define

$$x_{\text{ls}} = \operatorname{argmin} \|Ax - b\|_2 = (A^T A)^{-1} A^T b.$$

We address the following question. Suppose that for a particular A and b we have computed the least-squares solution x_{ls} (but not x_{ch}). How suboptimal is x_{ls} for the Chebyshev problem? In other words, how much larger is $\|Ax_{\text{ls}} - b\|_\infty$ than $\|Ax_{\text{ch}} - b\|_\infty$?

(a) Prove the lower bound

$$\|Ax_{\text{ls}} - b\|_\infty \leq \sqrt{m}\, \|Ax_{\text{ch}} - b\|_\infty,$$

using the fact that for all $z \in \mathbf{R}^m$,

$$\frac{1}{\sqrt{m}} \|z\|_2 \leq \|z\|_\infty \leq \|z\|_2.$$

(b) In example 5.6 (page 254) we derived a dual for the general norm approximation problem. Applying the results to the ℓ_∞-norm (and its dual norm, the ℓ_1-norm), we can state the following dual for the Chebyshev approximation problem:

$$\begin{array}{ll} \text{maximize} & b^T\nu \\ \text{subject to} & \|\nu\|_1 \leq 1 \\ & A^T\nu = 0. \end{array} \tag{5.104}$$

Any feasible ν corresponds to a lower bound $b^T\nu$ on $\|Ax_{\text{ch}} - b\|_\infty$.
Denote the least-squares residual as $r_{\text{ls}} = b - Ax_{\text{ls}}$. Assuming $r_{\text{ls}} \neq 0$, show that

$$\hat{\nu} = -r_{\text{ls}}/\|r_{\text{ls}}\|_1, \qquad \tilde{\nu} = r_{\text{ls}}/\|r_{\text{ls}}\|_1,$$

are both feasible in (5.104). By duality $b^T\hat{\nu}$ and $b^T\tilde{\nu}$ are lower bounds on $\|Ax_{\text{ch}} - b\|_\infty$. Which is the better bound? How do these bounds compare with the bound derived in part (a)?

Solution.

(a) Simple manipulation yields

$$\|Ax_{\text{cheb}} - b\|_\infty \geq \frac{1}{\sqrt{m}}\|Ax_{\text{cheb}} - b\|_2 \geq \frac{1}{\sqrt{m}}\|Ax_{\text{ls}} - b\|_2 \geq \frac{1}{\sqrt{m}}\|Ax_{\text{ls}} - b\|_\infty.$$

(b) From the expression $x_{\text{ls}} = (A^TA)^{-1}A^Tb$ we note that

$$A^Tr_{\text{ls}} = A^T(b - A(A^TA)^{-1}b) = A^Tb - A^Tb = 0.$$

Therefore $A^T\hat{\nu} = 0$ and $A^T\tilde{\nu} = 0$. Obviously we also have $\|\hat{\nu}\|_1 = 1$ and $\|\tilde{\nu}\|_1 = 1$, so $\hat{\nu}$ and $\tilde{\nu}$ are dual feasible.
We can write the dual objective value at $\hat{\nu}$ as

$$b^T\hat{\nu} = \frac{-b^Tr_{\text{ls}}}{\|r_{\text{ls}}\|_1} = \frac{(Ax_{\text{ls}} - b)^Tr_{\text{ls}}}{\|r_{\text{ls}}\|_1} = -\frac{\|r_{\text{ls}}\|_2^2}{\|r_{\text{ls}}\|_1}$$

and, similarly,

$$b^T\tilde{\nu} = \frac{\|r_{\text{ls}}\|_2^2}{\|r_{\text{ls}}\|_1}.$$

Therefore $\tilde{\nu}$ gives a better bound than $\hat{\nu}$.
Finally, to show that the resulting lower bound is better than the bound in part (a), we have to verify that

$$\frac{\|r_{\text{ls}}\|_2^2}{\|r_{\text{ls}}\|_1} \geq \frac{1}{\sqrt{m}}\|r_{\text{ls}}\|_\infty.$$

This follows from the inequalities

$$\|x\|_1 \leq \sqrt{m}\|x\|_2, \qquad \|x\|_\infty \leq \|x\|_2$$

which hold for general $x \in \mathbf{R}^m$.

5.7 *Piecewise-linear minimization.* We consider the convex piecewise-linear minimization problem

$$\text{minimize} \quad \max_{i=1,\ldots,m}(a_i^Tx + b_i) \tag{5.105}$$

with variable $x \in \mathbf{R}^n$.

(a) Derive a dual problem, based on the Lagrange dual of the equivalent problem

$$\begin{array}{ll} \text{minimize} & \max_{i=1,\ldots,m} y_i \\ \text{subject to} & a_i^T x + b_i = y_i, \quad i = 1, \ldots, m, \end{array}$$

with variables $x \in \mathbf{R}^n$, $y \in \mathbf{R}^m$.

(b) Formulate the piecewise-linear minimization problem (5.105) as an LP, and form the dual of the LP. Relate the LP dual to the dual obtained in part (a).

(c) Suppose we approximate the objective function in (5.105) by the smooth function

$$f_0(x) = \log\left(\sum_{i=1}^m \exp(a_i^T x + b_i)\right),$$

and solve the unconstrained geometric program

$$\text{minimize} \quad \log\left(\textstyle\sum_{i=1}^m \exp(a_i^T x + b_i)\right). \tag{5.106}$$

A dual of this problem is given by (5.62). Let $p^\star_{\text{pwl}}$ and $p^\star_{\text{gp}}$ be the optimal values of (5.105) and (5.106), respectively. Show that

$$0 \leq p^\star_{\text{gp}} - p^\star_{\text{pwl}} \leq \log m.$$

(d) Derive similar bounds for the difference between $p^\star_{\text{pwl}}$ and the optimal value of

$$\text{minimize} \quad (1/\gamma)\log\left(\textstyle\sum_{i=1}^m \exp(\gamma(a_i^T x + b_i))\right),$$

where $\gamma > 0$ is a parameter. What happens as we increase γ?

Solution.

(a) The dual function is

$$g(\lambda) = \inf_{x,y}\left(\max_{i=1,\ldots,m} y_i + \sum_{i=1}^m \lambda_i(a_i^T x + b_i - y_i)\right).$$

The infimum over x is finite only if $\sum_i \lambda_i a_i = 0$. To minimize over y we note that

$$\inf_y(\max_i y_i - \lambda^T y) = \begin{cases} 0 & \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1 \\ -\infty & \text{otherwise.} \end{cases}$$

To prove this, we first note that if $\lambda \succeq 0$, $\mathbf{1}^T\lambda = 1$, then

$$\lambda^T y = \sum_j \lambda_j y_j \leq \sum_j \lambda_j \max_i y_i = \max_i y_i,$$

with equality if $y = 0$, so in that case

$$\inf_y(\max_i y_i - \lambda^T y) = 0.$$

If $\lambda \not\succeq 0$, say $\lambda_j < 0$, then choosing $y_i = 0$, $i \neq j$, and $y_j = -t$, with $t \geq 0$, and letting t go to infinity, gives

$$\max_i y_i - \lambda^T y = 0 + t\lambda_k \to -\infty.$$

Finally, if $\mathbf{1}^T\lambda \neq 1$, choosing $y = t\mathbf{1}$, gives

$$\max_i y_i - \lambda^T y = t(1 - \mathbf{1}^T\lambda) \to -\infty,$$

if $t \to \infty$ and $1 < \mathbf{1}^T\lambda$, or if $t \to -\infty$ and $1 > \mathbf{1}^T\lambda$.
Summing up, we have

$$g(\lambda) = \begin{cases} b^T\lambda & \sum_i \lambda_i a_i = 0, \quad \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1 \\ -\infty & \text{otherwise.} \end{cases}$$

The resulting dual problem is

$$\begin{array}{ll} \text{maximize} & b^T\lambda \\ \text{subject to} & A^T\lambda = 0 \\ & \mathbf{1}^T\lambda = 1 \\ & \lambda \succeq 0. \end{array}$$

(b) The problem is equivalent to the LP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & Ax + b \preceq t\mathbf{1}. \end{array}$$

The dual problem is

$$\begin{array}{ll} \text{maximize} & b^T z \\ \text{subject to} & A^T z = 0, \quad \mathbf{1}^T z = 1, \quad z \succeq 0, \end{array}$$

which is identical to the dual derived in (a).

(c) Suppose $z^\star$ is dual optimal for the dual GP (5.62),

$$\begin{array}{ll} \text{maximize} & b^T z - \sum_{i=1}^m z_i \log z_i \\ \text{subject to} & \mathbf{1}^T z = 1 \\ & A^T z = 0. \end{array}$$

Then $z^\star$ is also feasible for the dual of the piecewise-linear formulation, with objective value

$$b^T z = p^\star_{\text{gp}} + \sum_{i=1}^m z_i^\star \log z_i^\star.$$

This provides a lower bound on $p^\star_{\text{pwl}}$:

$$p^\star_{\text{pwl}} \geq p^\star_{\text{gp}} + \sum_{i=1}^m z_i^\star \log z_i^\star \geq p^\star_{\text{gp}} - \log m.$$

The bound follows from

$$\inf_{\mathbf{1}^T z = 1} \sum_{i=1}^m z_i \log z_i = -\log m.$$

On the other hand, we also have

$$\max_i (a_i^T x + b_i) \leq \log \sum_i \exp(a_i^T x + b_i)$$

for all x, and therefore $p^\star_{\text{pwl}} \leq p^\star_{\text{gp}}$.
In conclusion,

$$p^\star_{\text{gp}} - \log m \leq p^\star_{\text{pwl}} \leq p^\star_{\text{gp}}.$$

(d) We first reformulate the problem as

$$\begin{array}{ll} \text{minimize} & (1/\gamma)\log\sum_{i=1}^m \exp(\gamma y_i) \\ \text{subject to} & Ax+b=y. \end{array}$$

The Lagrangian is

$$L(x,y,z) = \frac{1}{\gamma}\log\sum_{i=1}^m \exp(\gamma y_i) + z^T(Ax+b-y).$$

L is bounded below as a function of x only if $A^Tz=0$. To find the optimum over y, we set the gradient equal to zero:

$$\frac{e^{\gamma y_i}}{\sum_{i=1}^m e^{\gamma y_i}} = z_i.$$

This is solvable for y_i if $\mathbf{1}^Tz=1$ and $z\succeq 0$. The Lagrange dual function is

$$g(z) = b^Tz - \frac{1}{\gamma}\sum_{i=1}^m z_i\log z_i,$$

and the dual problem is

$$\begin{array}{ll} \text{maximize} & b^Tz - (1/\gamma)\sum_{i=1}^m z_i\log z_i \\ \text{subject to} & A^Tz=0 \\ & \mathbf{1}^Tz=1. \end{array}$$

Let $p^\star_{\text{gp}}(\gamma)$ be the optimal value of the GP. Following the same argument as above, we can conclude that

$$p^\star_{\text{gp}}(\gamma) - \frac{1}{\gamma}\log m \le p^\star_{\text{pwl}} \le p^\star_{\text{gp}}(\gamma).$$

In other words, $p^\star_{\text{gp}}(\gamma)$ approaches $p^\star_{\text{pwl}}$ as γ increases.

5.8 Relate the two dual problems derived in example 5.9 on page 257.

Solution. Suppose for example that ν is feasible in (5.71). Then choosing $\lambda_1 = (A^T\nu + c)^-$ and $\lambda_2 = (A^T\nu+c)^+$, yields a feasible solution in (5.69), with the same objective value. Conversely, suppose ν, λ_1 and λ_2 are feasible in (5.69). The equality constraint implies that

$$\lambda_1 = (A^T\nu+c)^- + v, \quad \lambda_2 = (A^T\nu+c)^+ + v,$$

for some $v\succeq 0$. Therefore we can write (5.69) as

$$\begin{array}{ll} \text{maximize} & -b^T\nu - u^T(A^T\nu+c)^- + l^T(A^T\nu+c)^+ - (u-l)^Tv \\ \text{subject to} & v\succeq 0, \end{array}$$

and it is clear that at the optimum $v=0$. Therefore the optimum ν in (5.69) is also optimal in (5.71).

5.9 *Suboptimality of a simple covering ellipsoid.* Recall the problem of determining the minimum volume ellipsoid, centered at the origin, that contains the points $a_1,\ldots,a_m\in\mathbf{R}^n$ (problem (5.14), page 222):

$$\begin{array}{ll} \text{minimize} & f_0(X) = \log\det(X^{-1}) \\ \text{subject to} & a_i^TXa_i \le 1, \quad i=1,\ldots,m, \end{array}$$

with $\mathbf{dom}\, f_0 = \mathbf{S}^n_{++}$. We assume that the vectors $a_1,\ldots,a_m$ span $\mathbf{R}^n$ (which implies that the problem is bounded below).

(a) Show that the matrix

$$X_{\rm sim} = \left(\sum_{k=1}^m a_k a_k^T\right)^{-1},$$

is feasible. *Hint.* Show that

$$\left[\begin{array}{cc} \sum_{k=1}^m a_k a_k^T & a_i \\ a_i^T & 1 \end{array}\right] \succeq 0,$$

and use Schur complements (§A.5.5) to prove that $a_i^T X a_i \leq 1$ for $i = 1, \ldots, m$.
Solution.

$$\left[\begin{array}{cc} \sum_{k=1}^m a_k a_k^T & a_k \\ a_i^T & 1 \end{array}\right] = \left[\begin{array}{cc} \sum_{k\neq i} a_k a_k^T & 0 \\ 0 & 0 \end{array}\right] + \left[\begin{array}{c} a_i \\ 1 \end{array}\right]\left[\begin{array}{c} a_i \\ 1 \end{array}\right]^T$$

is the sum of two positive semidefinite matrices, hence positive semidefinite. The Schur complement of the $1,1$ block of this matrix is therefore also positive semidefinite:

$$1 - a_i^T \left(\sum_{k=1}^m a_k a_k^T\right)^{-1} a_i \geq 0,$$

which is the desired conclusion.

(b) Now we establish a bound on how suboptimal the feasible point $X_{\rm sim}$ is, via the dual problem,

$$\begin{array}{ll} \text{maximize} & \log\det\left(\sum_{i=1}^m \lambda_i a_i a_i^T\right) - \mathbf{1}^T\lambda + n \\ \text{subject to} & \lambda \succeq 0, \end{array}$$

with the implicit constraint $\sum_{i=1}^m \lambda_i a_i a_i^T \succ 0$. (This dual is derived on page 222.) To derive a bound, we restrict our attention to dual variables of the form $\lambda = t\mathbf{1}$, where $t > 0$. Find (analytically) the optimal value of t, and evaluate the dual objective at this λ. Use this to prove that the volume of the ellipsoid $\{u \mid u^T X_{\rm sim} u \leq 1\}$ is no more than a factor $(m/n)^{n/2}$ more than the volume of the minimum volume ellipsoid.
Solution. The dual function evaluated at $\lambda = t\mathbf{1}$ is

$$g(\lambda) = \log\det\left(\sum_{i=1}^m a_i a_i^T\right) + n\log t - mt + n.$$

Now we'll maximize over $t > 0$ to get the best lower bound. Setting the derivative with respect to t equal to zero yields the optimal value $t = n/m$. Using this λ we get the dual objective value

$$g(\lambda) = \log\det\left(\sum_{i=1}^m a_i a_i^T\right) + n\log(n/m).$$

The primal objective value for $X = X_{\rm sim}$ is given by

$$-\log\det\left(\sum_{i=1}^m a_i a_i^T\right)^{-1},$$

so the duality gap associated with $X_{\rm sim}$ and λ is $n\log(m/n)$. (Recall that $m \geq n$, by our assumption that $a_1, \ldots, a_m$ span $\mathbf{R}^n$.) It follows that, in terms of the objective function, $X_{\rm sim}$ is no more than $n\log(m/n)$ suboptimal. The volume V of the ellipsoid $\mathcal{E}$ associated with the matrix X is given by $V = \exp(-O/2)$, where O is the associated objective function, $O = -\log\det X$. The bound follows.

5.10 *Optimal experiment design.* The following problems arise in experiment design (see §7.5).

(a) *D-optimal design.*

$$\begin{array}{ll} \text{minimize} & \log\det\left(\sum_{i=1}^p x_i v_i v_i^T\right)^{-1} \\ \text{subject to} & x \succeq 0, \quad \mathbf{1}^T x = 1. \end{array}$$

(b) *A-optimal design.*

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}\left(\sum_{i=1}^p x_i v_i v_i^T\right)^{-1} \\ \text{subject to} & x \succeq 0, \quad \mathbf{1}^T x = 1. \end{array}$$

The domain of both problems is $\{x \mid \sum_{i=1}^p x_i v_i v_i^T \succ 0\}$. The variable is $x \in \mathbf{R}^p$; the vectors $v_1, \ldots, v_p \in \mathbf{R}^n$ are given.

Derive dual problems by first introducing a new variable $X \in \mathbf{S}^n$ and an equality constraint $X = \sum_{i=1}^p x_i v_i v_i^T$, and then applying Lagrange duality. Simplify the dual problems as much as you can.

Solution.

(a) *D-optimal design.*

$$\begin{array}{ll} \text{minimize} & \log\det(X^{-1}) \\ \text{subject to} & X = \sum_{i=1}^p x_i v_i v_i^T \\ & x \succeq 0, \quad \mathbf{1}^T x = 1. \end{array}$$

The Lagrangian is

$$\begin{aligned} L(x, Z, z, \nu) &= \log\det(X^{-1}) + \mathbf{tr}(ZX) - \sum_{i=1}^p x_i v_i^T Z v_i - z^T x + \nu(\mathbf{1}^T x - 1) \\ &= \log\det(X^{-1}) + \mathbf{tr}(ZX) + \sum_{i=1}^p x_i(-v_i^T Z v_i - z_i + \nu) - \nu. \end{aligned}$$

The minimum over x_i is bounded below only if $\nu - v_i^T Z v_i = z_i$. Setting the gradient with respect to X equal to zero gives $X^{-1} = Z$. We obtain the dual function

$$g(Z, z) = \begin{cases} \log\det Z + n - \nu & \nu - v_i^T Z v_i = z_i, \quad i = 1, \ldots, p \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem is

$$\begin{array}{ll} \text{maximize} & \log\det Z + n - \nu \\ \text{subject to} & v_i^T Z v_i \leq \nu, \quad i = 1, \ldots, p, \end{array}$$

with domain $\mathbf{S}^n_{++} \times \mathbf{R}$. We can eliminate ν by first making a change of variables $W = (1/\nu)Z$, which gives

$$\begin{array}{ll} \text{maximize} & \log\det W + n + n\log\nu - \nu \\ \text{subject to} & v_i^T \hat{W} v_i \leq 1, \quad i = 1, \ldots, p. \end{array}$$

Finally, we note that we can easily optimize $n\log\nu - \nu$ over ν. The optimum is $\nu = n$, and substituting gives

$$\begin{array}{ll} \text{maximize} & \log\det W + n\log n \\ \text{subject to} & v_i^T W v_i \leq 1, \quad i = 1, \ldots, p. \end{array}$$

(b) *A-optimal design.*

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(X^{-1}) \\ \text{subject to} & X = \left(\sum_{i=1}^p x_i v_i v_i^T\right)^{-1} \\ & x \succeq 0, \quad \mathbf{1}^T x = 1. \end{array}$$

The Lagrangian is

$$\begin{aligned} L(X, Z, z, \nu) &= \mathbf{tr}(X^{-1}) + \mathbf{tr}(ZX) - \sum_{i=1}^p x_i v_i^T Z v_i - z^T x + \nu(\mathbf{1}^T x - 1) \\ &= \mathbf{tr}(X^{-1}) + \mathbf{tr}(ZX) + \sum_{i=1}^p x_i(-v_i^T Z v_i - z_i + \nu) - \nu. \end{aligned}$$

The minimum over x is unbounded below unless $v_i^T Z v_i + z_i = \nu$. The minimum over X can be found by setting the gradient equal to zero: $X^{-2} = Z$, or $X = Z^{-1/2}$ if $Z \succ 0$, which gives

$$\inf_{X \succ 0} (\mathbf{tr}(X^{-1}) + \mathbf{tr}(ZX)) = \begin{cases} 2\,\mathbf{tr}(Z^{1/2}) & Z \succeq 0 \\ -\infty & \text{otherwise.} \end{cases}$$

The dual function is

$$g(Z, z, \nu) = \begin{cases} -\nu + 2\,\mathbf{tr}(Z^{1/2}) & Z \succeq 0, \quad v_i^T Z v_i + z_i = \nu \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem is

$$\begin{array}{ll} \text{maximize} & -\nu + 2\,\mathbf{tr}(Z^{1/2}) \\ \text{subject to} & v_i^T Z v_i \leq nu, \quad i = 1, \ldots, p \\ & Z \succeq 0. \end{array}$$

As a first simplification, we define $W = (1/\nu)Z$, and write the problem as

$$\begin{array}{ll} \text{maximize} & -\nu + 2\sqrt{\nu}\,\mathbf{tr}(W^{1/2}) \\ \text{subject to} & v_i^T W v_i \leq 1, \quad i = 1, \ldots, p \\ & W \succeq 0. \end{array}$$

By optimizing over $\nu > 0$, we obtain

$$\begin{array}{ll} \text{maximize} & (\mathbf{tr}(W^{1/2}))^2 \\ \text{subject to} & v_i^T W v_i \leq 1, \quad i = 1, \ldots, p \\ & W \succeq 0. \end{array}$$

5.11 Derive a dual problem for

$$\text{minimize} \quad \textstyle\sum_{i=1}^N \|A_i x + b_i\|_2 + (1/2)\|x - x_0\|_2^2.$$

The problem data are $A_i \in \mathbf{R}^{m_i \times n}$, $b_i \in \mathbf{R}^{m_i}$, and $x_0 \in \mathbf{R}^n$. First introduce new variables $y_i \in \mathbf{R}^{m_i}$ and equality constraints $y_i = A_i x + b_i$.

Solution. The Lagrangian is

$$L(x, z_1, \ldots, z_N) = \sum_{i=1}^N \|y_i\|_2 + \frac{1}{2}\|x - x_0\|_2^2 - \sum_{i=1}^N z_i^T (y_i - A_i x - b_i).$$

We first minimize over y_i. We have

$$\inf_{y_i}(\|y_i\|_2 + z_i^T y_i) = \begin{cases} 0 & \|z_i\|_2 \leq 1 \\ -\infty & \text{otherwise.} \end{cases}$$

(If $\|z_i\|_2 > 1$, choose $y_i = -tz_i$ and let $t \to \infty$, to show that the function is unbounded below. If $\|z_i\|_2 \leq 1$, it follows from the Cauchy-Schwarz inequality that $\|y_i\|_2 + z_i^T y_i \geq 0$, so the minimum is reached when $y_i = 0$.)
We can minimize over x by setting the gradient with respect to x equal to zero. This yields

$$x = x_0 + \sum_{i=1}^N A_i^T z.$$

Substituting in the Lagrangian gives the dual function

$$g(z_1, \ldots, z_N) = \begin{cases} \sum_{i=1}^N (A_i x_0 + b_i)^T z_i - \frac{1}{2}\| \sum_{i=1}^N A_i^T z_i\|_2^2 & \|z_i\|_2 \leq 1, \quad i = 1, \ldots, N \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem is

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^N (A_i x_0 + b_i)^T z_i - \frac{1}{2}\| \sum_{i=1}^N A_i^T z_i\|^2 \\ \text{subject to} & \|z_i\|_2 \leq 1, \quad i = 1, \ldots, N. \end{array}$$

5.12 *Analytic centering.* Derive a dual problem for

$$\text{minimize} \quad -\sum_{i=1}^m \log(b_i - a_i^T x)$$

with domain $\{x \mid a_i^T x < b_i,\ i = 1, \ldots, m\}$. First introduce new variables y_i and equality constraints $y_i = b_i - a_i^T x$.
(The solution of this problem is called the *analytic center* of the linear inequalities $a_i^T x \leq b_i$, $i = 1, \ldots, m$. Analytic centers have geometric applications (see §8.5.3), and play an important role in barrier methods (see chapter 11).)
Solution. We derive the dual of the problem

$$\begin{array}{ll} \text{minimize} & -\sum_{i=1}^m \log y_i \\ \text{subject to} & y = b - Ax, \end{array}$$

where $A \in \mathbf{R}^{m \times n}$ has a_i^T as its ith row. The Lagrangian is

$$L(x, y, \nu) = -\sum_{i=1}^m \log y_i + \nu^T (y - b + Ax)$$

and the dual function is

$$g(\nu) = \inf_{x,y} \left(-\sum_{i=1}^m \log y_i + \nu^T (y - b + Ax) \right).$$

The term $\nu^T Ax$ is unbounded below as a function of x unless $A^T \nu = 0$. The terms in y are unbounded below if $\nu \not\succ 0$, and achieve their minimum for $y_i = 1/\nu_i$ otherwise. We therefore find the dual function

$$g(\nu) = \begin{cases} \sum_{i=1}^m \log \nu_i + m - b^T \nu & A^T \nu = 0, \quad \nu \succ 0 \\ -\infty & \text{otherwise} \end{cases}$$

and the dual problem

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^m \log \nu_i - b^T \nu + m \\ \text{subject to} & A^T \nu = 0. \end{array}$$

5.13 *Lagrangian relaxation of Boolean LP.* A *Boolean linear program* is an optimization problem of the form

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \\ & x_i \in \{0,1\}, \quad i = 1, \ldots, n, \end{array}$$

and is, in general, very difficult to solve. In exercise 4.15 we studied the LP relaxation of this problem,

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \\ & 0 \leq x_i \leq 1, \quad i = 1, \ldots, n, \end{array} \tag{5.107}$$

which is far easier to solve, and gives a lower bound on the optimal value of the Boolean LP. In this problem we derive another lower bound for the Boolean LP, and work out the relation between the two lower bounds.

(a) *Lagrangian relaxation.* The Boolean LP can be reformulated as the problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \\ & x_i(1 - x_i) = 0, \quad i = 1, \ldots, n, \end{array}$$

which has quadratic equality constraints. Find the Lagrange dual of this problem. The optimal value of the dual problem (which is convex) gives a lower bound on the optimal value of the Boolean LP. This method of finding a lower bound on the optimal value is called *Lagrangian relaxation.*

(b) Show that the lower bound obtained via Lagrangian relaxation, and via the LP relaxation (5.107), are the same. *Hint.* Derive the dual of the LP relaxation (5.107).

Solution.

(a) The Lagrangian is

$$\begin{array}{rcl} L(x, \mu, \nu) & = & c^T x + \mu^T (Ax - b) - \nu^T x + x^T \mathbf{diag}(\nu) x \\ & = & x^T \mathbf{diag}(\nu) x + (c + A^T \mu - \nu)^T x - b^T \mu. \end{array}$$

Minimizing over x gives the dual function

$$g(\mu, \nu) = \begin{cases} -b^T \mu - (1/4) \sum_{i=1}^n (c_i + a_i^T \mu - \nu_i)^2 / \nu_i & \nu \succeq 0 \\ -\infty & \text{otherwise} \end{cases}$$

where a_i is the ith column of A, and we adopt the convention that $a^2/0 = \infty$ if $a \neq 0$, and $a^2/0 = 0$ if $a = 0$.
The resulting dual problem is

$$\begin{array}{ll} \text{maximize} & -b^T \mu - (1/4) \sum_{i=1}^n (c_i + a_i^T \mu - \nu_i)^2 / \nu_i \\ \text{subject to} & \nu \succeq 0. \end{array}$$

In order to simplify this dual, we optimize analytically over ν, by noting that

$$\begin{array}{rcl} \sup\limits_{\nu_i \geq 0} \left(-\dfrac{(c_i + a_i^T \mu - \nu_i)^2}{\nu_i} \right) & = & \begin{cases} (c_i + a_i^T \mu) & c_i + a_i^T \mu \leq 0 \\ 0 & c_i + a_i^T \mu \geq 0 \end{cases} \\ & = & \min\{0, (c_i + a_i^T \mu)\}. \end{array}$$

This allows us to eliminate ν from the dual problem, and simplify it as

$$\begin{array}{ll} \text{maximize} & -b^T \mu + \sum_{i=1}^n \min\{0, c_i + a_i^T \mu\} \\ \text{subject to} & \mu \succeq 0. \end{array}$$

(b) We follow the hint. The Lagrangian and dual function of the LP relaxation re

$$\begin{aligned} L(x,u,v,w) &= c^Tx + u^T(Ax-b) - v^Tx + w^T(x-\mathbf{1}) \\ &= (c + A^Tu - v + w)^Tx - b^Tu - \mathbf{1}^Tw \\ g(u,v,w) &= \begin{cases} -b^Tu - \mathbf{1}^Tw & A^Tu - v + w + c = 0 \\ -\infty & \text{otherwise.} \end{cases} \end{aligned}$$

The dual problem is

$$\begin{array}{ll} \text{maximize} & -b^Tu - \mathbf{1}^Tw \\ \text{subject to} & A^Tu - v + w + c = 0 \\ & u \succeq 0, v \succeq 0, w \succeq 0, \end{array}$$

which is equivalent to the Lagrange relaxation problem derived above. We conclude that the two relaxations give the same value.

5.14 *A penalty method for equality constraints.* We consider the problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & Ax = b, \end{array} \tag{5.108}$$

where $f_0 : \mathbf{R}^n \to \mathbf{R}$ is convex and differentiable, and $A \in \mathbf{R}^{m\times n}$ with $\mathbf{rank}\, A = m$.
In a *quadratic penalty method*, we form an auxiliary function

$$\phi(x) = f(x) + \alpha\|Ax - b\|_2^2,$$

where $\alpha > 0$ is a parameter. This auxiliary function consists of the objective plus the *penalty term* $\alpha\|Ax-b\|_2^2$. The idea is that a minimizer of the auxiliary function, $\tilde{x}$, should be an approximate solution of the original problem. Intuition suggests that the larger the penalty weight α, the better the approximation $\tilde{x}$ to a solution of the original problem.
Suppose $\tilde{x}$ is a minimizer of ϕ. Show how to find, from $\tilde{x}$, a dual feasible point for (5.108). Find the corresponding lower bound on the optimal value of (5.108).
Solution. If $\tilde{x}$ minimizes ϕ, then

$$\nabla f_0(\tilde{x}) + 2\alpha A^T(A\tilde{x} - b) = 0.$$

Therefore $\tilde{x}$ is also a minimizer of

$$f_0(x) + \nu^T(Ax - b)$$

where $\nu = 2\alpha(A\tilde{x} - b)$. Therefore ν is dual feasible with

$$\begin{aligned} g(\nu) &= \inf_x (f_0(x) + \nu^T(Ax - b)) \\ &= f_0(\tilde{x}) + 2\alpha\|A\tilde{x} - b\|_2^2. \end{aligned}$$

Therefore,

$$f_0(x) \geq f_0(\tilde{x}) + 2\alpha\|A\tilde{x} - b\|_2^2$$

for all x that satisfy $Ax = b$.

5.15 Consider the problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m, \end{array} \tag{5.109}$$

where the functions $f_i : \mathbf{R}^n \to \mathbf{R}$ are differentiable and convex. Let $h_1, \ldots, h_m : \mathbf{R} \to \mathbf{R}$ be increasing differentiable convex functions. Show that

$$\phi(x) = f_0(x) + \sum_{i=1}^{m} h_i(f_i(x))$$

is convex. Suppose $\tilde{x}$ minimizes ϕ. Show how to find from $\tilde{x}$ a feasible point for the dual of (5.109). Find the corresponding lower bound on the optimal value of (5.109).

Solution. $\tilde{x}$ satisfies

$$0 = \nabla f_0(\tilde{x}) + \sum_{i=1}^{m} (h_i'(f_i(\tilde{x})))\nabla f_i(\tilde{x})) = \nabla f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i \nabla f_i(\tilde{x}))$$

where $\lambda_i = h_i'(f_i(\tilde{x}))$. λ is dual feasible: $\lambda_i \geq 0$, since h_i is increasing, and

$$\begin{aligned} g(\lambda) &= f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) \\ &= f_0(\tilde{x}) + \sum_{i=1}^{m} h_i'(f_i(\tilde{x})) f_i(\tilde{x}). \end{aligned}$$

5.16 *An exact penalty method for inequality constraints.* Consider the problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m, \end{array} \tag{5.110}$$

where the functions $f_i : \mathbf{R}^n \to \mathbf{R}$ are differentiable and convex. In an exact penalty method, we solve the auxiliary problem

$$\text{minimize} \quad \phi(x) = f_0(x) + \alpha \max_{i=1,\ldots,m} \max\{0, f_i(x)\}, \tag{5.111}$$

where $\alpha > 0$ is a parameter. The second term in ϕ penalizes deviations of x from feasibility. The method is called an *exact* penalty method if for sufficiently large α, solutions of the auxiliary problem (5.111) also solve the original problem (5.110).

(a) Show that ϕ is convex.

(b) The auxiliary problem can be expressed as

$$\begin{array}{ll} \text{minimize} & f_0(x) + \alpha y \\ \text{subject to} & f_i(x) \leq y, \quad i = 1, \ldots, m \\ & 0 \leq y \end{array}$$

where the variables are x and $y \in \mathbf{R}$. Find the Lagrange dual of this problem, and express it in terms of the Lagrange dual function g of (5.110).

(c) Use the result in (b) to prove the following property. Suppose $\lambda^\star$ is an optimal solution of the Lagrange dual of (5.110), and that strong duality holds. If $\alpha > \mathbf{1}^T\lambda^\star$, then any solution of the auxiliary problem (5.111) is also an optimal solution of (5.110).

Solution.

(a) The first term is convex. The second term is convex since it can be expressed as

$$\max\{f_1(x), \ldots, f_m(x), 0\},$$

i.e., the pointwise maximum of a number of convex functions.

(b) The Lagrangian is

$$L(x,y,\lambda,\mu) = f_0(x) + \alpha y + \sum_{i=1}^m \lambda_i(f_i(x) - y) - \mu y.$$

The dual function is

$$\begin{aligned}\inf_{x,y} L(x,y,\lambda,\mu) &= \inf_{x,y} f_0(x) + \alpha y + \sum_{i=1}^m \lambda_i(f_i(x) - y) - \mu y \\ &= \inf_x (f_0(x) + \sum_{i=1}^m \lambda_i f_i(x)) + \inf_y (\alpha - \sum_{i=1}^m \lambda_i - \mu) y \\ &= \begin{cases} g(\lambda) & \mathbf{1}^T\lambda + \mu = \alpha \\ -\infty & \text{otherwise,} \end{cases}\end{aligned}$$

and the dual problem is

$$\begin{array}{ll} \text{maximize} & g(\lambda) \\ \text{subject to} & \mathbf{1}^T\lambda + \mu = \alpha \\ & \lambda \succeq 0, \quad \mu \geq 0, \end{array}$$

or, equivalently,

$$\begin{array}{ll} \text{maximize} & g(\lambda) \\ \text{subject to} & \mathbf{1}^T\lambda \leq \alpha \\ & \lambda \succeq 0. \end{array}$$

(c) If $\mathbf{1}^T\lambda^\star < \alpha$, then $\lambda^\star$ is also optimal for the dual problem derived in part (b). By complementary slackness $y = 0$ in any optimal solution of the primal problem, so the optimal x satisfies $f_i(x) \leq 0$, $i = 1, \ldots, m$, *i.e.*, it is feasible in the original problem, and therefore also optimal.

5.17 *Robust linear programming with polyhedral uncertainty.* Consider the robust LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \sup_{a \in \mathcal{P}_i} a^T x \leq b_i, \quad i = 1, \ldots, m, \end{array}$$

with variable $x \in \mathbf{R}^n$, where $\mathcal{P}_i = \{a \mid C_i a \preceq d_i\}$. The problem data are $c \in \mathbf{R}^n$, $C_i \in \mathbf{R}^{m_i \times n}$, $d_i \in \mathbf{R}^{m_i}$, and $b \in \mathbf{R}^m$. We assume the polyhedra $\mathcal{P}_i$ are nonempty.
Show that this problem is equivalent to the LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & d_i^T z_i \leq b_i, \quad i = 1, \ldots, m \\ & C_i^T z_i = x, \quad i = 1, \ldots, m \\ & z_i \succeq 0, \quad i = 1, \ldots, m \end{array}$$

with variables $x \in \mathbf{R}^n$ and $z_i \in \mathbf{R}^{m_i}$, $i = 1, \ldots, m$. *Hint.* Find the dual of the problem of maximizing $a_i^T x$ over $a_i \in \mathcal{P}_i$ (with variable a_i).
Solution. The problem can be expressed as

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & f_i(x) \leq b_i, \quad i = 1, \ldots, m \end{array}$$

if we define $f_i(x)$ as the optimal value of the LP

$$\begin{array}{ll} \text{maximize} & x^T a \\ \text{subject to} & C_i a \preceq d, \end{array}$$

where a is the variable, and x is treated as a problem parameter. It is readily shown that the Lagrange dual of this LP is given by

$$\begin{array}{ll} \text{minimize} & d_i^T z \\ \text{subject to} & C_i^T z = x \\ & z \succeq 0. \end{array}$$

The optimal value of this LP is also equal to $f_i(x)$, so we have $f_i(x) \leq b_i$ if and only if there exists a z_i with

$$d_i^T z \leq b_i, \qquad C_i^T z_i = x, \qquad z_i \succeq 0.$$

5.18 *Separating hyperplane between two polyhedra.* Formulate the following problem as an LP or an LP feasibility problem. Find a separating hyperplane that strictly separates two polyhedra

$$\mathcal{P}_1 = \{x \mid Ax \preceq b\}, \qquad \mathcal{P}_2 = \{x \mid Cx \preceq d\},$$

i.e., find a vector $a \in \mathbf{R}^n$ and a scalar γ such that

$$a^T x > \gamma \text{ for } x \in \mathcal{P}_1, \qquad a^T x < \gamma \text{ for } x \in \mathcal{P}_2.$$

You can assume that $\mathcal{P}_1$ and $\mathcal{P}_2$ do not intersect.

Hint. The vector a and scalar γ must satisfy

$$\inf_{x \in \mathcal{P}_1} a^T x > \gamma > \sup_{x \in \mathcal{P}_2} a^T x.$$

Use LP duality to simplify the infimum and supremum in these conditions.

Solution. Define $p_1^\star(a)$ and $p_2^\star(a)$ as

$$p_1^\star(a) = \inf\{a^T x \mid Ax \preceq b\}, \qquad p_2^\star(a) = \sup\{a^T x \mid Cx \preceq d\}.$$

A hyperplane $a^T x = \gamma$ strictly separates the two polyhedra if

$$p_2^\star(a) < \gamma < p_1^\star(a).$$

For example, we can find a by solving

$$\begin{array}{ll} \text{maximize} & p_1^\star(a) - p_2^\star(a) \\ \text{subject to} & \|a\|_1 \leq 1 \end{array}$$

and selecting $\gamma = (p_1^\star(a) + p_2^\star(a))/2$. (The bound $\|a\|_1$ is added because the objective is homogeneous in a, so it unbounded unless we add a constraint on a.)

Using LP duality we have

$$\begin{array}{rcl} p_1^\star(a) & = & \sup\{-b^T z_1 \mid A^T z_1 + a = 0,\ z_1 \succeq 0\} \\ p_2^\star(a) & = & \inf\{-a^T x \mid Cx \preceq d\} \\ & = & \sup\{-d^T z_2 \mid C^T z_2 - a = 0,\ z_2 \succeq 0\}, \end{array}$$

so we can reformulate the problem as

$$\begin{array}{ll} \text{maximize} & -b^T z_1 - d^T z_2 \\ \text{subject to} & A^T z_1 + a = 0 \\ & C^T z_2 - a = 0 \\ & z_1 \succeq 0, \quad z_2 \succeq 0 \\ & \|a\|_1 \leq 1. \end{array}$$

The variables are a, z_1 and z_2.

Another solution is based on theorems of alternative. The hyperplane separates the two polyhedra if the following two sets of linear inequalities are infeasible:

- $Ax \preceq b$, $a^T x \leq \gamma$
- $Cx \preceq d$, $a^T x \geq \gamma$.

Using a theorem of alternatives this is equivalent to requiring that the following two sets of inequalities are both feasible:

- $z_1 \succeq 0$, $w_1 \geq 0$, $A^T z_1 + a w_1 = 0$, $b^T z_1 - \gamma w_1 < 0$
- $z_2 \succeq 0$, $w_2 \geq 0$, $C^T z_2 - a w_2 = 0$, $d^T z_2 + \gamma w_2 < 0$

w_1 and w_2 must be nonzero. If $w_1 = 0$, then $A^T z_1 = 0$, $b^T z_1 < 0$. which means $\mathcal{P}_1$ is empty, and similarly, $w_2 = 0$ means $\mathcal{P}_2$ is empty. We can therefore simplify the two conditions as

- $z_1 \succeq 0$, $A^T z_1 + a = 0$, $b^T z_1 < \gamma$
- $z_2 \succeq 0$, $C^T z_2 - a = 0$, $d^T z_2 < -\gamma$,

which is basically the same as the conditions derived above.

5.19 *The sum of the largest elements of a vector.* Define $f : \mathbf{R}^n \to \mathbf{R}$ as

$$f(x) = \sum_{i=1}^{r} x_{[i]},$$

where r is an integer between 1 and n, and $x_{[1]} \geq x_{[2]} \geq \cdots \geq x_{[r]}$ are the components of x sorted in decreasing order. In other words, $f(x)$ is the sum of the r largest elements of x. In this problem we study the constraint

$$f(x) \leq \alpha.$$

As we have seen in chapter 3, page 80, this is a convex constraint, and equivalent to a set of $n!/(r!(n-r)!)$ linear inequalities

$$x_{i_1} + \cdots + x_{i_r} \leq \alpha, \quad 1 \leq i_1 < i_2 < \cdots < i_r \leq n.$$

The purpose of this problem is to derive a more compact representation.

(a) Given a vector $x \in \mathbf{R}^n$, show that $f(x)$ is equal to the optimal value of the LP

$$\begin{array}{ll} \text{maximize} & x^T y \\ \text{subject to} & 0 \preceq y \preceq \mathbf{1} \\ & \mathbf{1}^T y = r \end{array}$$

with $y \in \mathbf{R}^n$ as variable.

(b) Derive the dual of the LP in part (a). Show that it can be written as

$$\begin{array}{ll} \text{minimize} & rt + \mathbf{1}^T u \\ \text{subject to} & t\mathbf{1} + u \succeq x \\ & u \succeq 0, \end{array}$$

where the variables are $t \in \mathbf{R}$, $u \in \mathbf{R}^n$. By duality this LP has the same optimal value as the LP in (a), *i.e.*, $f(x)$. We therefore have the following result: x satisfies $f(x) \leq \alpha$ if and only if there exist $t \in \mathbf{R}$, $u \in \mathbf{R}^n$ such that

$$rt + \mathbf{1}^T u \leq \alpha, \qquad t\mathbf{1} + u \succeq x, \qquad u \succeq 0.$$

These conditions form a set of $2n+1$ linear inequalities in the $2n+1$ variables x, u, t.

(c) As an application, we consider an extension of the classical Markowitz portfolio optimization problem

$$\begin{array}{ll} \text{minimize} & x^T\Sigma x \\ \text{subject to} & \overline{p}^T x \geq r_{\min} \\ & \mathbf{1}^T x = 1, \quad x \succeq 0 \end{array}$$

discussed in chapter 4, page 155. The variable is the portfolio $x \in \mathbf{R}^n$; $\overline{p}$ and Σ are the mean and covariance matrix of the price change vector p.

Suppose we add a *diversification constraint*, requiring that no more than 80% of the total budget can be invested in any 10% of the assets. This constraint can be expressed as

$$\sum_{i=1}^{\lfloor 0.1n \rfloor} x_{[i]} \leq 0.8.$$

Formulate the portfolio optimization problem with diversification constraint as a QP.

Solution.

(a) See also chapter 4, exercise 4.8.

For simplicity we assume that the elements of x are sorted in decreasing order:

$$x_1 \geq x_2 \geq \cdots \geq x_n.$$

It is easy to see that the optimal value is

$$x_1 + x_2 + \cdots + x_r,$$

obtained by choosing $y_1 = y_2 = \cdots = y_r = 1$ and $y_{r+1} = \cdots = y_n = 0$.

(b) We first change the objective from maximization to minimization:

$$\begin{array}{ll} \text{minimize} & -x^T y \\ \text{subject to} & 0 \preceq y \preceq \mathbf{1} \\ & \mathbf{1}^T y = r. \end{array}$$

We introduce a Lagrange multiplier λ for the lower bound, u for the upper bound, and t for the equality constraint. The Lagrangian is

$$\begin{array}{rcl} L(y, \lambda, u, t) & = & -x^T y - \lambda^T y + u^T(y - \mathbf{1}) + t(\mathbf{1}^T y - r) \\ & = & -\mathbf{1}^T u - rt + (-x - \lambda + u + t\mathbf{1})^T y. \end{array}$$

Minimizing over y yields the dual function

$$g(\lambda, u, t) = \begin{cases} -\mathbf{1}^T u - rt & -x - \lambda + u + t\mathbf{1} = 0 \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem is to maximize g subject to $\lambda \succeq 0$ and $u \succeq 0$:

$$\begin{array}{ll} \text{maximize} & -\mathbf{1}^T u - rt \\ \text{subject to} & -\lambda + u + t\mathbf{1} = x \\ & \lambda \succeq 0, \quad u \succeq 0, \end{array}$$

or after changing the objective to minimization (*i.e.*, undoing the sign change we started with),

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T u + rt \\ \text{subject to} & u + t\mathbf{1} \succeq x \\ & u \succeq 0. \end{array}$$

We eliminated λ by noting that it acts as a slack variable in the first constraint.

(c)

$$\begin{array}{ll} \text{minimize} & x^T\Sigma x \\ \text{subject to} & \overline{p}^T x \geq r_{\min} \\ & \mathbf{1}^T x = 1, \quad x \succeq 0 \\ & \lfloor n/20 \rfloor t + \mathbf{1}^T u \leq 0.9 \\ & \lambda \mathbf{1} + u \succeq 0 \\ & u \succeq 0, \end{array}$$

with variables x, u, t, v.

5.20 *Dual of channel capacity problem.* Derive a dual for the problem

$$\begin{array}{ll} \text{minimize} & -c^T x + \sum_{i=1}^m y_i \log y_i \\ \text{subject to} & Px = y \\ & x \succeq 0, \quad \mathbf{1}^T x = 1, \end{array}$$

where $P \in \mathbf{R}^{m \times n}$ has nonnegative elements, and its columns add up to one (*i.e.*, $P^T\mathbf{1} = \mathbf{1}$). The variables are $x \in \mathbf{R}^n$, $y \in \mathbf{R}^m$. (For $c_j = \sum_{i=1}^m p_{ij} \log p_{ij}$, the optimal value is, up to a factor $\log 2$, the negative of the capacity of a discrete memoryless channel with channel transition probability matrix P; see exercise 4.57.)
Simplify the dual problem as much as possible.
Solution. The Lagrangian is

$$\begin{aligned} L(x, y, \lambda, \nu, z) &= -c^T x + \sum_{i=1}^m y_i \log y_i - \lambda^T x + \nu(\mathbf{1}^T x - 1) + z^T(Px - y) \\ &= (-c - \lambda + \nu\mathbf{1} + P^T z)^T x + \sum_{i=1}^m y_i \log y_i - z^T y - \nu. \end{aligned}$$

The minimum over x is bounded below if and only if

$$-c - \lambda + \nu\mathbf{1} + P^T z = 0.$$

To minimize over y, we set the derivative with respect to y_i equal to zero, which gives $\log y_i + 1 - z_i = 0$, and conclude that

$$\inf_{y_i \geq 0} (y_i \log y_i - z_i y_i) = -e^{z_i - 1}.$$

The dual function is

$$g(\lambda, \nu, z) = \begin{cases} -\sum_{i=1}^m e^{z_i - 1} - \nu & -c - \lambda + \nu\mathbf{1} + P^T z = 0 \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem is

$$\begin{array}{ll} \text{maximize} & -\sum_{i=1}^m \exp(z_i - 1) - \nu \\ \text{subject to} & P^T z - c + \nu\mathbf{1} \succeq 0. \end{array}$$

This can be simplified by introducing a variable $w = z + \nu\mathbf{1}$ (and using the fact that $\mathbf{1} = P^T\mathbf{1}$), which gives

$$\begin{array}{ll} \text{maximize} & -\sum_{i=1}^m \exp(w_i - \nu - 1) - \nu \\ \text{subject to} & P^T w \succeq c. \end{array}$$

Finally we can easily maximize the objective function over ν by setting the derivative equal to zero (the optimal value is $\nu = -\log(\sum_i e^{1-w_i})$, which leads to

$$\begin{array}{ll} \text{maximize} & -\log(\sum_{i=1}^m \exp w_i) - 1 \\ \text{subject to} & P^T w \succeq c. \end{array}$$

This is a geometric program, in convex form, with linear inequality constraints (*i.e.*, monomial inequality constraints in the associated geometric program).

Strong duality and Slater's condition

5.21 *A convex problem in which strong duality fails.* Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & e^{-x} \\ \text{subject to} & x^2/y \leq 0 \end{array}$$

with variables x and y, and domain $\mathcal{D} = \{(x, y) \mid y > 0\}$.

(a) Verify that this is a convex optimization problem. Find the optimal value.

(b) Give the Lagrange dual problem, and find the optimal solution $\lambda^\star$ and optimal value $d^\star$ of the dual problem. What is the optimal duality gap?

(c) Does Slater's condition hold for this problem?

(d) What is the optimal value $p^\star(u)$ of the perturbed problem

$$\begin{array}{ll} \text{minimize} & e^{-x} \\ \text{subject to} & x^2/y \leq u \end{array}$$

as a function of u? Verify that the global sensitivity inequality

$$p^\star(u) \geq p^\star(0) - \lambda^\star u$$

does not hold.

Solution.

(a) $p^\star = 1$.

(b) The Lagrangian is $L(x, y, \lambda) = e^{-x} + \lambda x^2/y$. The dual function is

$$g(\lambda) = \inf_{x,y>0}(e^{-x} + \lambda x^2/y) = \begin{cases} 0 & \lambda \geq 0 \\ -\infty & \lambda < 0, \end{cases}$$

so we can write the dual problem as

$$\begin{array}{ll} \text{maximize} & 0 \\ \text{subject to} & \lambda \geq 0, \end{array}$$

with optimal value $d^\star = 0$. The optimal duality gap is $p^\star - d^\star = 1$.

(c) Slater's condition is not satisfied.

(d) $p^\star(u) = 1$ if $u = 0$, $p^\star(u) = 0$ if $u > 0$ and $p^\star(u) = \infty$ if $u < 0$.

5.22 *Geometric interpretation of duality.* For each of the following optimization problems, draw a sketch of the sets

$$\begin{aligned} \mathcal{G} &= \{(u, t) \mid \exists x \in \mathcal{D},\ f_0(x) = t,\ f_1(x) = u\}, \\ \mathcal{A} &= \{(u, t) \mid \exists x \in \mathcal{D},\ f_0(x) \leq t,\ f_1(x) \leq u\}, \end{aligned}$$

give the dual problem, and solve the primal and dual problems. Is the problem convex? Is Slater's condition satisfied? Does strong duality hold?

The domain of the problem is $\mathbf{R}$ unless otherwise stated.

(a) Minimize x subject to $x^2 \leq 1$.

(b) Minimize x subject to $x^2 \leq 0$.

(c) Minimize x subject to $|x| \leq 0$.

(d) Minimize x subject to $f_1(x) \leq 0$ where

$$f_1(x) = \begin{cases} -x+2 & x \geq 1 \\ x & -1 \leq x \leq 1 \\ -x-2 & x \leq -1. \end{cases}$$

(e) Minimize x^3 subject to $-x+1 \leq 0$.

(f) Minimize x^3 subject to $-x+1 \leq 0$ with domain $\mathcal{D} = \mathbf{R}_+$.

Solution. For the first four problems $\mathcal{G}$ is the curve

$$\mathcal{G} = \{(u,t) \mid u \in \mathcal{D},\ u = f_1(t)\}.$$

For problem (e), $\mathcal{G}$ is the curve

$$\mathcal{G} = \{(u,t) \mid t = (1-u)^3\}.$$

For problem (f), $\mathcal{G}$ is the curve

$$\mathcal{G} = \{(u,t) \mid u \leq 1,\ t = (1-u)^3\}.$$

$\mathcal{A}$ is the set of points above and to the right of $\mathcal{G}$.

(a) $x^\star = -1$. $\lambda^\star = 1$. $p^\star = -1$. $d^\star = -1$. Convex. Strong duality. Slater's condition holds.
This is the generic convex case.

(b) $x^\star = 0$. $p^\star = 0$. $d^\star = 0$. Dual optimum is not achieved. Convex. Strong duality. Slater's condition does not hold.
We have strong duality although Slater's condition does not hold. However the dual optimum is not attained.

(c) $x^\star = 0$. $p^\star = 0$. $\lambda^\star = 1$. $d^\star = 0$. Convex. Strong duality. Slater's condition not satisfied.
We have strong duality and the dual is attained, although Slater's condition does not hold.

(d) $x^\star = -2$. $p^\star = -2$. $\lambda^\star = 1$. $d^\star = -2$. Not convex. Strong duality.
We have strong duality, although this is a very nonconvex problem.

(e) $x^\star = 1$. $p^\star = 1$. $d^\star = -\infty$. Not convex. No strong duality.
The problem has a convex feasibility set, and the objective is convex on the feasible set. However the problem is *not* convex, according to the definition used in this book. Lagrange duality gives a trivial bound $-\infty$.

(f) $x^\star = 1$. $p^\star = 1$. $\lambda^\star = 1$. $d^\star = 1$. Convex. Strong duality. Slater's condition is satisfied.
Adding the domain condition seems redundant at first. However the new problem is convex (according to our definition). Now strong duality holds and the dual optimum is attained.

5.23 *Strong duality in linear programming.* We prove that strong duality holds for the LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \end{array}$$

and its dual

$$\begin{array}{ll} \text{maximize} & -b^T z \\ \text{subject to} & A^T z + c = 0, \quad z \succeq 0, \end{array}$$

provided at least one of the problems is feasible. In other words, the only possible exception to strong duality occurs when $p^\star = \infty$ and $d^\star = -\infty$.

(a) Suppose $p^\star$ is finite and $x^\star$ is an optimal solution. (If finite, the optimal value of an LP is attained.) Let $I \subseteq \{1, 2, \ldots, m\}$ be the set of active constraints at $x^\star$:

$$a_i^T x^\star = b_i, \quad i \in I, \qquad a_i^T x^\star < b_i, \quad i \notin I.$$

Show that there exists a $z \in \mathbf{R}^m$ that satisfies

$$z_i \geq 0, \quad i \in I, \qquad z_i = 0, \quad i \notin I, \qquad \sum_{i \in I} z_i a_i + c = 0.$$

Show that z is dual optimal with objective value $c^T x^\star$.

Hint. Assume there exists no such z, *i.e.*, $-c \notin \{\sum_{i \in I} z_i a_i \mid z_i \geq 0\}$. Reduce this to a contradiction by applying the strict separating hyperplane theorem of example 2.20, page 49. Alternatively, you can use Farkas' lemma (see §5.8.3).

(b) Suppose $p^\star = \infty$ and the dual problem is feasible. Show that $d^\star = \infty$. *Hint.* Show that there exists a nonzero $v \in \mathbf{R}^m$ such that $A^T v = 0$, $v \succeq 0$, $b^T v < 0$. If the dual is feasible, it is unbounded in the direction v.

(c) Consider the example

$$\begin{array}{ll} \text{minimize} & x \\ \text{subject to} & \left[\begin{array}{c} 0 \\ 1 \end{array}\right] x \preceq \left[\begin{array}{c} -1 \\ 1 \end{array}\right]. \end{array}$$

Formulate the dual LP, and solve the primal and dual problems. Show that $p^\star = \infty$ and $d^\star = -\infty$.

Solution.

(a) Without loss of generality we can assume that $I = \{1, 2, \ldots, k\}$. Let $\bar{A} \in \mathbf{R}^{k \times n}$ be the matrix formed by the first k rows of A. We assume there is no $\bar{z} \succeq 0$ such that $c + \bar{A}^T \bar{z} = 0$, *i.e.*,

$$-c \notin S = \{\bar{A}^T \bar{z} \mid \bar{z} \succeq 0\}.$$

By the strict separating hyperplane theorem, applied to $-c$ and S, there exists a u such that

$$-u^T c > u^T \bar{A}^T \bar{z}$$

for all $\bar{z} \succeq 0$. This means $c^T u < 0$ (evaluate the righthand side at $\bar{z} = 0$), and $\bar{A} u \preceq 0$.

Now consider $x = x^\star + tu$. We have

$$a_i^T x = a_i^T x^\star + t a_i^T u = b_i + t a_i^T u \leq b_i, \quad i \in I,$$

for all $t \geq 0$, and

$$a_i^T x = a_i^T x^\star + t a_i^T u < b_i + t a_i^T u < b_i, \quad i \notin I,$$

for sufficiently small positive t. Finally

$$c^T x = c^T x^\star + t c^T u < c^T x^\star$$

for all positive t. This is a contradiction, because we have constructed primal feasible points with a lower objective value than $x^\star$.

We conclude that there exists a $\bar{z} \succeq 0$ with $\bar{A}^T \bar{z} + c = 0$. Choosing $z = (\bar{z}, 0)$ yields a dual feasible point. Its objective value is

$$-b^T z = -(x^\star)^T \bar{A}^T z = c^T x^\star.$$

(b) The primal problem is infeasible, *i.e.*,

$$-b \notin S = \{Ax + s \mid s \succeq 0\}.$$

The righthand side is a closed convex set, so we can apply the strict separating hyperplane theorem and conclude there exists a $v \in \mathbf{R}^m$ such that $-v^T b > v^T(Ax + s)$ for all x and all $s \succeq 0$. This is equivalent to

$$b^T v < 0, \qquad A^T v = 0, \qquad v \succeq 0.$$

This only leaves two possibilities. Either the dual problem is infeasible, or it is feasible and unbounded above. (If z_0 is dual feasible, then $z = z_0 + tv$ is dual feasible for all $t \geq 0$, with $-b^T z = -b^T z_0 + tb^T v$).

(c) The dual LP is

$$\begin{array}{ll} \text{maximize} & z_1 - z_2 \\ \text{subject to} & z_2 + 1 = 0 \\ & z_1, z_2 \geq 0, \end{array}$$

which is also infeasible ($d^\star = -\infty$).

5.24 *Weak max-min inequality.* Show that the weak max-min inequality

$$\sup_{z \in Z} \inf_{w \in W} f(w, z) \leq \inf_{w \in W} \sup_{z \in Z} f(w, z)$$

always holds, with no assumptions on $f : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$, $W \subseteq \mathbf{R}^n$, or $Z \subseteq \mathbf{R}^m$.

Solution. If W and Z are empty, the inequality reduces to $-\infty \leq \infty$.

If W is nonempty, with $\tilde{w} \in W$, we have

$$\inf_{w \in W} f(w, z) \leq f(\tilde{w}, z)$$

for all $z \in Z$. Taking the supremum over $z \in Z$ on both sides we get

$$\sup_{z \in Z} \inf_{w \in W} f(w, z) \leq \sup_{z \in Z} f(\tilde{w}, z).$$

Taking the inf over $\tilde{w} \in W$ we get the max-min inequality.

The proof for nonempty Z is similar.

5.25 [BL00, page 95] *Convex-concave functions and the saddle-point property.* We derive conditions under which the saddle-point property

$$\sup_{z \in Z} \inf_{w \in W} f(w, z) = \inf_{w \in W} \sup_{z \in Z} f(w, z) \tag{5.112}$$

holds, where $f : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$, $W \times Z \subseteq \mathbf{dom}\, f$, and W and Z are nonempty. We will assume that the function

$$g_z(w) = \begin{cases} f(w, z) & w \in W \\ \infty & \text{otherwise} \end{cases}$$

is closed and convex for all $z \in Z$, and the function

$$h_w(z) = \begin{cases} -f(w, z) & z \in Z \\ \infty & \text{otherwise} \end{cases}$$

is closed and convex for all $w \in W$.

(a) The righthand side of (5.112) can be expressed as $p(0)$, where

$$p(u) = \inf_{w \in W} \sup_{z \in Z} (f(w,z) + u^T z).$$

Show that p is a convex function.

(b) Show that the conjugate of p is given by

$$p^*(v) = \begin{cases} -\inf_{w \in W} f(w,v) & v \in Z \\ \infty & \text{otherwise.} \end{cases}$$

(c) Show that the conjugate of p^* is given by

$$p^{**}(u) = \sup_{z \in Z} \inf_{w \in W} (f(w,z) + u^T z).$$

Combining this with (a), we can express the max-min equality (5.112) as $p^{**}(0) = p(0)$.

(d) From exercises 3.28 and 3.39 (d), we know that $p^{**}(0) = p(0)$ if $0 \in \mathbf{int}\,\mathbf{dom}\, p$. Conclude that this is the case if W and Z are bounded.

(e) As another consequence of exercises 3.28 and 3.39, we have $p^{**}(0) = p(0)$ if $0 \in \mathbf{dom}\, p$ and p is closed. Show that p is closed if the sublevel sets of g_z are bounded.

Solution.

(a) For fixed z, $F_z(u,w) = g_z(w) - u^T z$ is a (closed) convex function of (w,u). Therefore

$$F(w,u) = \sup_{z \in Z} (g_z(w) + u^T z)$$

is a convex function of (w,u). (It is also closed because it epigraph is the intersection of closed sets, the epigraphs of the functions F_z.)
Minimizing F over w yields a convex function

$$\begin{aligned} \inf_w F(w,u) &= \inf_w \sup_{z \in Z} (g_z(w) + u^T z) \\ &= \inf_{w \in W} \sup_{z \in Z} (f(w,z) + u^T z) \\ &= p(u). \end{aligned}$$

(b) The conjugate is

$$\begin{aligned} p^*(v) &= \sup_u (v^T u - p(u)) \\ &= \sup_u (v^T u - \inf_{w \in W} \sup_{z \in Z} (f(w,z) + u^T z)) \\ &= \sup_u \sup_{w \in W} (v^T u - \sup_{z \in Z} (f(w,z) + u^T z)) \\ &= \sup_u \sup_{w \in W} (-\sup_{z \in Z} (f(w,z) + (z-v)^T u)) \\ &= \sup_u \sup_{w \in W} \inf_{z \in Z} (-f(w,z) + (v-z)^T u) \\ &= \sup_{w \in W} \sup_u \inf_{z \in Z} (-f(w,z) + (v-z)^T u). \end{aligned}$$

By assumption, for all w, the set

$$C_w = \mathbf{epi}\, h_w = \{(z,t) \mid z \in Z,\ t \geq -f(z,w)\}$$

is closed and convex. We show that this implies that

$$\sup_u \inf_{z\in Z}(-f(w,z)+(z-v)^T u) = \begin{cases} -f(w,v) & v \in Z \\ \infty & \text{otherwise.} \end{cases}$$

First assume $v \in Z$. It is clear that

$$\inf_{z\in Z}(-f(w,z)+z^T u) \leq -f(w,v)+v^T u \tag{5.25.A}$$

for all u. Since h_w is closed and convex, there exists a nonvertical supporting hyperplane to its epigraph C_w at the point $(z, f(z,w))$, *i.e.*, there exists a $\tilde{u}$ such that

$$\inf_{z\in Z}(\tilde{u}^T z - f(z,w)) = \inf_{(z,t)\in C_w}(\tilde{u}^T z - t) = \tilde{u}^T v - f(v,w). \tag{5.25.B}$$

Combining (5.25.A) and (5.25.B) we conclude that

$$\inf_{z\in Z}(-f(w,z)+(z-v)^T u) \leq -f(w,v)$$

for all u, with equality for $u = \tilde{u}$. Therefore

$$\sup_u \inf_{z\in Z}(-f(w,z)+z^T u - v^T u) = -f(w,v).$$

Next assume $v \neq Z$. For all w, and all t, $(v,t) \neq C_w$, hence it can be strictly separated from C_w by a nonvertical hyperplane: for all t and $w \in W$ there exists a u such that

$$t + u^T v < \inf_{z\in Z}(-f(w,z)+u^T z),$$

i.e.,

$$t < \inf_{z\in Z}(-f(w,z)+u^T(z-v)).$$

This holds for all t, so

$$\sup_u \inf_{z\in Z}(-f(w,z)+u^T(z-v)) = \infty.$$

(c) The conjugate of p^* is

$$\begin{aligned} p^{**}(u) &= \sup_{v\in Z}(u^T v + \inf_{w\in W} f(w,v)) \\ &= \sup_{v\in Z}\inf_{w\in W}(f(w,v)+u^T v). \end{aligned}$$

(d) We noted in part (a) that $F(w,u) = \sup_{z\in Z}(f(w,z)+z^T u)$ is a closed convex function. If Z is bounded, then the maximum in the definition is attained for all $(w,u) \in W \times \mathbf{R}^m$, so $W \times \mathbf{R}^m \subseteq \mathbf{dom}\, F_z$.

If W is bounded, the minimum in $p(u) = \inf_{w\in W} F(w,u)$ is also attained for all u, so $\mathbf{dom}\, p = \mathbf{R}^m$.

(e) $\mathbf{epi}\, p$ is the projection of $\mathbf{epi}\, F \subseteq \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}$ (a closed set) on $\mathbf{R}^m \times \mathbf{R}$.

Now in general, the projection of a closed convex set $C \in \mathbf{R}^p \times \mathbf{R}^q$ on $\mathbf{R}^p$ is closed if C does not contain any half-lines of the form $\{(\bar{x}, \bar{y}+sv) \in \mathbf{R}^p \times \mathbf{R}^q \mid s \geq 0\}$ with $v \neq 0$ (*i.e.*, no directions of recession of the form $(0,v)$).

Applying this result to the epigraph of F and its projection $\mathbf{epi}\, p$, we conclude that $\mathbf{epi}\, p$ is closed if $\mathbf{epi}\, F$ does not contain any half-lines $\{(\bar{w}, \bar{u}, \bar{t}) + s(v,0,0) \mid s \geq 0\}$. This is the case if the sublevel sets of g_z are bounded.

Optimality conditions

5.26 Consider the QCQP

$$\begin{array}{ll} \text{minimize} & x_1^2 + x_2^2 \\ \text{subject to} & (x_1-1)^2 + (x_2-1)^2 \leq 1 \\ & (x_1-1)^2 + (x_2+1)^2 \leq 1 \end{array}$$

with variable $x \in \mathbf{R}^2$.

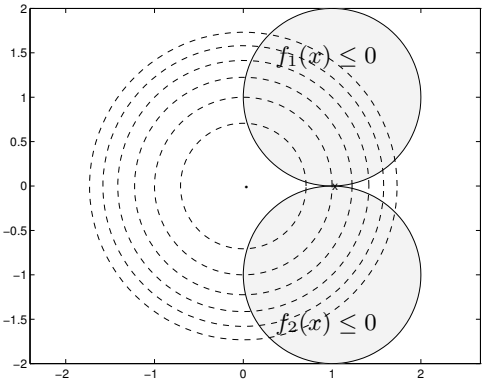
(a) Sketch the feasible set and level sets of the objective. Find the optimal point $x^\star$ and optimal value $p^\star$.

(b) Give the KKT conditions. Do there exist Lagrange multipliers $\lambda_1^\star$ and $\lambda_2^\star$ that prove that $x^\star$ is optimal?

(c) Derive and solve the Lagrange dual problem. Does strong duality hold?

Solution.

(a) The figure shows the feasible set (the intersection of the two shaded disks) and some contour lines of the objective function. There is only one feasible point, $(1, 0)$, so it is optimal for the primal problem, and we have $p^\star = 1$.



(b) The KKT conditions are

$$\begin{gathered} (x_1-1)^2 + (x_2-1)^2 \leq 1, \quad (x_1-1)^2 + (x_2+1)^2 \leq 1, \\ \lambda_1 \geq 0, \qquad \lambda_2 \geq 0 \\ 2x_1 + 2\lambda_1(x_1-1) + 2\lambda_2(x_1-1) = 0 \\ 2x_2 + 2\lambda_1(x_2-1) + 2\lambda_2(x_2+1) = 0 \\ \lambda_1((x_1-1)^2 + (x_2-1)^2 - 1) = \lambda_2((x_1-1)^2 + (x_2+1)^2 - 1) = 0. \end{gathered}$$

At $x = (1, 0)$, these conditions reduce to

$$\lambda_1 \geq 0, \qquad \lambda_2 \geq 0, \qquad 2 = 0, \qquad -2\lambda_1 + 2\lambda_2 = 0,$$

which (clearly, in view of the third equation) have no solution.

(c) The Lagrange dual function is given by

$$g(\lambda_1, \lambda_2) = \inf_{x_1, x_2} L(x_1, x_2, \lambda_1, \lambda_2)$$

where

$$\begin{aligned} & L(x_1, x_2, \lambda_1, \lambda_2) \\ &= x_1^2 + x_2^2 + \lambda_1((x_1-1)^2 + (x_2-1)^2 - 1) + \lambda_2((x_1-1)^2 + (x_2+1)^2 - 1) \\ &= (1+\lambda_1+\lambda_2)x_1^2 + (1+\lambda_1+\lambda_2)x_2^2 - 2(\lambda_1+\lambda_2)x_1 - 2(\lambda_1-\lambda_2)x_2 + \lambda_1 + \lambda_2. \end{aligned}$$

L reaches its minimum for

$$x_1 = \frac{\lambda_1 + \lambda_2}{1 + \lambda_1 + \lambda_2}, \qquad x_2 = \frac{\lambda_1 - \lambda_2}{1 + \lambda_1 + \lambda_2},$$

and we find

$$g(\lambda_1, \lambda_2) = \begin{cases} -\frac{(\lambda_1+\lambda_2)^2+(\lambda_1-\lambda_2)^2}{1+\lambda_1+\lambda_2} + \lambda_1 + \lambda_2 & 1 + \lambda_1 + \lambda_2 \geq 0 \\ -\infty & \text{otherwise,} \end{cases}$$

where we interpret $a/0 = 0$ if $a = 0$ and as $-\infty$ if $a < 0$. The Lagrange dual problem is given by

$$\begin{array}{ll} \text{maximize} & (\lambda_1 + \lambda_2 - (\lambda_1 - \lambda_2)^2)/(1 + \lambda_1 + \lambda_2) \\ \text{subject to} & \lambda_1, \lambda_2 \geq 0. \end{array}$$

Since g is symmetric, the optimum (if it exists) occurs with $\lambda_1 = \lambda_2$. The dual function then simplifies to

$$g(\lambda_1, \lambda_1) = \frac{2\lambda_1}{2\lambda_1 + 1}.$$

We see that $g(\lambda_1, \lambda_2)$ tends to 1 as $\lambda_1 \to \infty$. We have $d^\star = p^\star = 1$, but the dual optimum is not attained.

Recall that the KKT conditions only hold if (1) strong duality holds, (2) the primal optimum is attained, and (3) the dual optimum is attained. In this example, the KKT conditions fail because the dual optimum is not attained.

5.27 *Equality constrained least-squares.* Consider the equality constrained least-squares problem

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|_2^2 \\ \text{subject to} & Gx = h \end{array}$$

where $A \in \mathbf{R}^{m\times n}$ with $\mathbf{rank}\, A = n$, and $G \in \mathbf{R}^{p\times n}$ with $\mathbf{rank}\, G = p$.

Give the KKT conditions, and derive expressions for the primal solution $x^\star$ and the dual solution $\nu^\star$.

Solution.

(a) The Lagrangian is

$$\begin{aligned} L(x, \nu) &= \|Ax - b\|_2^2 + \nu^T(Gx - h) \\ &= x^T A^T Ax + (G^T\nu - 2A^T b)^T x - \nu^T h, \end{aligned}$$

with minimizer $x = -(1/2)(A^T A)^{-1}(G^T\nu - 2A^T b)$. The dual function is

$$g(\nu) = -(1/4)(G^T\nu - 2A^T b)^T (A^T A)^{-1}(G^T\nu - 2A^T b) - \nu^T h$$

(b) The optimality conditions are

$$2A^T(Ax^\star - b) + G^T\nu^\star = 0, \qquad Gx^\star = h.$$

(c) From the first equation,

$$x^\star = (A^T A)^{-1}(A^T b - (1/2)G^T\nu^\star).$$

Plugging this expression for $x^\star$ into the second equation gives

$$G(A^T A)^{-1}A^T b - (1/2)G(A^T A)^{-1}G^T\nu^\star = h$$

i.e.,

$$\nu^\star = -2(G(A^T A)^{-1}G^T)^{-1}(h - G(A^T A)^{-1}A^T b).$$

Substituting in the first expression gives an analytical expression for $x^\star$.

5.28 Prove (without using any linear programming code) that the optimal solution of the LP

$$\begin{array}{ll} \text{minimize} & 47x_1 + 93x_2 + 17x_3 - 93x_4 \\ \text{subject to} & \begin{bmatrix} -1 & -6 & 1 & 3 \\ -1 & -2 & 7 & 1 \\ 0 & 3 & -10 & -1 \\ -6 & -11 & -2 & 12 \\ 1 & 6 & -1 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \preceq \begin{bmatrix} -3 \\ 5 \\ -8 \\ -7 \\ 4 \end{bmatrix} \end{array}$$

is unique, and given by $x^\star = (1, 1, 1, 1)$.

Solution.

Clearly, $x^\star = (1, 1, 1, 1)$ is feasible (it satisfies the first four constraints with equality). The point $z^\star = (3, 2, 2, 7, 0)$ is a certificate of optimality of $x = (1, 1, 1, 1)$:

- $z^\star$ is dual feasible: $z^\star \succeq 0$ and $A^T z^\star + c = 0$.
- $z^\star$ satisfies the complementary slackness condition:

$$z_i^\star (a_i^T x - b_i) = 0, \quad i = 1, \ldots, m,$$

since the first four components of $Ax - b$ and the last component of $z^\star$ are zero.

5.29 The problem

$$\begin{array}{ll} \text{minimize} & -3x_1^2 + x_2^2 + 2x_3^2 + 2(x_1 + x_2 + x_3) \\ \text{subject to} & x_1^2 + x_2^2 + x_3^2 = 1, \end{array}$$

is a special case of (5.32), so strong duality holds even though the problem is not convex. Derive the KKT conditions. Find all solutions x, ν that satisfy the KKT conditions. Which pair corresponds to the optimum?

Solution.

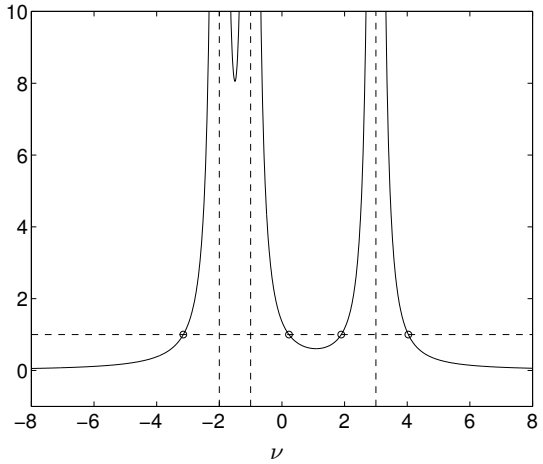
(a) The KKT conditions are

$$x_1^2 + x_2^2 + x_3^2 = 1, \qquad (-3+\nu)x_1 + 1 = 0, \qquad (1+\nu)x_2 + 1 = 0, \qquad (2+\nu)x_3 + 1 = 0.$$

(b) A first observation is that the KKT conditions imply $\nu \neq 2$, $\nu \neq -1$, $\nu \neq 3$. We can therefore eliminate x and reduce the KKT conditions to a nonlinear equation in ν:

$$\frac{1}{(-3+\nu)^2} + \frac{1}{(1+\nu)^2} + \frac{1}{(2+\nu)^2} = 1$$

The lefthand side is plotted in the figure.



There are four solutions:

$$\nu = -3.15, \qquad \nu = 0.22, \qquad \nu = 1.89, \qquad \nu = 4.04,$$

corresponding to

$$x = (0.16, 0.47, -0.87), \qquad x = (0.36, -0.82, 0.45),$$

$$x = (0.90, -0.35, 0.26), \qquad x = (-0.97, -0.20, 0.17).$$

(c) $\nu^\star$ is the largest of the four values: $\nu^\star = 4.0352$. This can be seen several ways. The simplest way is to compare the objective values of the four solutions x, which are

$$f_0(x) = 1.17, \qquad f_0(x) = 0.67, \qquad f_0(x) = -0.56, \qquad f_0(x) = -4.70.$$

We can also evaluate the dual objective at the four candidate values for ν. Finally we can note that we must have

$$\nabla^2 f_0(x^\star) + \nu^\star \nabla^2 f_1^\star(x^\star) \succeq 0,$$

because $x^\star$ is a minimizer of $L(x, \nu^\star)$. In other words

$$\begin{bmatrix} -3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} + \nu^\star \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \succeq 0,$$

and therefore $\nu^\star \geq 3$.

5.30 Derive the KKT conditions for the problem

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}\, X - \log\det X \\ \text{subject to} & Xs = y, \end{array}$$

with variable $X \in \mathbf{S}^n$ and domain $\mathbf{S}^n_{++}$. $y \in \mathbf{R}^n$ and $s \in \mathbf{R}^n$ are given, with $s^T y = 1$. Verify that the optimal solution is given by

$$X^\star = I + yy^T - \frac{1}{s^T s} ss^T.$$

Solution. We introduce a Lagrange multiplier $z \in \mathbf{R}^n$ for the equality constraint. The KKT optimality conditions are:

$$X \succ 0, \qquad Xs = y, \qquad X^{-1} = I + \frac{1}{2}(zs^T + sz^T). \tag{5.30.A}$$

We first determine z from the condition $Xs = y$. Multiplying the gradient equation on the right with y gives

$$s = X^{-1}y = y + \frac{1}{2}(z + (z^T y)s). \tag{5.30.B}$$

By taking the inner product with y on both sides and simplifying, we get $z^T y = 1 - y^T y$. Substituting in (5.30.B) we get

$$z = -2y + (1 + y^T y)s,$$

and substituting this expression for z in (5.30.A) gives

$$\begin{aligned} X^{-1} &= I + \frac{1}{2}(-2ys^T - 2sy^T + 2(1 + y^T y)ss^T) \\ &= I + (1 + y^T y)ss^T - ys^T - sy^T. \end{aligned}$$

Finally we verify that this is the inverse of the matrix $X^\star$ given above:

$$\begin{aligned}
&\left(I + (1+y^Ty)ss^T - ys^T - sy^T\right) X^\star \\
&= (I + yy^T - (1/s^Ts)ss^T) + (1+y^Ty)(ss^T + sy^T - ss^T) \\
&\quad - (ys^T + yy^T - ys^T) - (sy^T + (y^Ty)sy^T - (1/s^Ts)ss^T) \\
&= I.
\end{aligned}$$

To complete the solution, we prove that $X^\star \succ 0$. An easy way to see this is to note that

$$X^\star = I + yy^T - \frac{ss^T}{s^Ts} = \left(I + \frac{ys^T}{\|s\|_2} - \frac{ss^T}{s^Ts}\right)\left(I + \frac{ys^T}{\|s\|_2} - \frac{ss^T}{s^Ts}\right)^T.$$

5.31 *Supporting hyperplane interpretation of KKT conditions.* Consider a convex problem with no equality constraints,

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m. \end{array}$$

Assume that $x^\star \in \mathbf{R}^n$ and $\lambda^\star \in \mathbf{R}^m$ satisfy the KKT conditions

$$\begin{aligned}
f_i(x^\star) &\leq 0, \quad i = 1, \ldots, m \\
\lambda_i^\star &\geq 0, \quad i = 1, \ldots, m \\
\lambda_i^\star f_i(x^\star) &= 0, \quad i = 1, \ldots, m \\
\nabla f_0(x^\star) + \textstyle\sum_{i=1}^m \lambda_i^\star \nabla f_i(x^\star) &= 0.
\end{aligned}$$

Show that

$$\nabla f_0(x^\star)^T(x - x^\star) \geq 0$$

for all feasible x. In other words the KKT conditions imply the simple optimality criterion of §4.2.3.

Solution. Suppose x is feasible. Since f_i are convex and $f_i(x) \leq 0$ we have

$$0 \geq f_i(x) \geq f_i(x^\star) + \nabla f_i(x^\star)^T(x - x^\star), \quad i = 1, \ldots, m.$$

Using $\lambda_i^\star \geq 0$, we conclude that

$$\begin{aligned}
0 &\geq \sum_{i=1}^m \lambda_i^\star \left(f_i(x^\star) + \nabla f_i(x^\star)^T(x - x^\star)\right) \\
&= \sum_{i=1}^m \lambda_i^\star f_i(x^\star) + \sum_{i=1}^m \lambda_i^\star \nabla f_i(x^\star)^T(x - x^\star) \\
&= -\nabla f_0(x^\star)^T(x - x^\star).
\end{aligned}$$

In the last line, we use the complementary slackness condition $\lambda_i^\star f_i(x^\star) = 0$, and the last KKT condition. This shows that $\nabla f_0(x^\star)^T(x - x^\star) \geq 0$, *i.e.*, $\nabla f_0(x^\star)$ defines a supporting hyperplane to the feasible set at $x^\star$.

Perturbation and sensitivity analysis

5.32 *Optimal value of perturbed problem.* Let $f_0, f_1, \ldots, f_m : \mathbf{R}^n \to \mathbf{R}$ be convex. Show that the function

$$p^\star(u, v) = \inf\{f_0(x) \mid \exists x \in \mathcal{D},\ f_i(x) \leq u_i,\ i = 1, \ldots, m,\ Ax - b = v\}$$

is convex. This function is the optimal cost of the perturbed problem, as a function of the perturbations u and v (see §5.6.1).

Solution. Define the function

$$G(x,u,v) = \begin{cases} f_0(x) & f_i(x) \leq u_i,\ i=1,\ldots,m, \quad Ax-b=v \\ \infty & \text{otherwise.} \end{cases}$$

G is convex on its domain

$$\mathbf{dom}\, G = \{(x,u,v) \mid x \in \mathcal{D},\ f_i(x) \leq u_i,\ i=1,\ldots,m,\ Ax-b=v\},$$

which is easily shown to be convex. Therefore G is convex, jointly in x,u,v. Therefore

$$p^\star(u,v) = \inf_x G(x,u,v)$$

is convex.

5.33 *Parametrized ℓ_1-norm approximation.* Consider the ℓ_1-norm minimization problem

$$\text{minimize} \quad \|Ax+b+\epsilon d\|_1$$

with variable $x \in \mathbf{R}^3$, and

$$A = \begin{bmatrix} -2 & 7 & 1 \\ -5 & -1 & 3 \\ -7 & 3 & -5 \\ -1 & 4 & -4 \\ 1 & 5 & 5 \\ 2 & -5 & -1 \end{bmatrix}, \qquad b = \begin{bmatrix} -4 \\ 3 \\ 9 \\ 0 \\ -11 \\ 5 \end{bmatrix}, \qquad d = \begin{bmatrix} -10 \\ -13 \\ -27 \\ -10 \\ -7 \\ 14 \end{bmatrix}.$$

We denote by $p^\star(\epsilon)$ the optimal value as a function of ϵ.

(a) Suppose $\epsilon = 0$. Prove that $x^\star = \mathbf{1}$ is optimal. Are there any other optimal points?

(b) Show that $p^\star(\epsilon)$ is affine on an interval that includes $\epsilon = 0$.

Solution. The dual problem of

$$\text{minimize} \quad \|Ax+b\|_1$$

is given by

$$\begin{array}{ll} \text{maximize} & b^T z \\ \text{subject to} & A^T z = 0 \\ & \|z\|_\infty \leq 1. \end{array}$$

If x and z are both feasible, then

$$\|Ax+b\|_1 \geq z^T(Ax+b) = b^T z$$

(this follows from the inequality $u^T v \leq \|u\|_\infty \|v\|_1$). We have equality ($\|Ax+b\|_1 = b^T z$) only if $z_i(Ax+b)_i = |(Ax+b)_i|$ for all i. In other words, the optimality conditions are: x and z are optimal if and only if $A^T z = 0$, $\|z\|_\infty \leq 1$ and the following 'complementarity conditions' hold:

$$\begin{array}{lcl} -1 < z_i < 1 & \Longrightarrow & (Ax+b)_i = 0 \\ (Ax+b)_i > 0 & \Longrightarrow & z_i = 1 \\ (Ax+b)_i < 0 & \Longrightarrow & z_i = -1. \end{array}$$

(a) $b + Ax = (2, 0, 0, -1, 0, 1)$, so the optimality conditions tell us that the dual optimal solution must satisfy $z_1 = 1$, $z_4 = -1$, and $z_5 = 1$. It remains to find the other 3 components z_2, z_3, z_6. We can do this by solving

$$A^T z = \begin{bmatrix} -5 & -7 & 1 \\ -1 & 3 & 5 \\ 3 & -5 & 5 \end{bmatrix} \begin{bmatrix} z_2 \\ z_3 \\ z_5 \end{bmatrix} + \begin{bmatrix} -2 & -1 & 2 \\ 7 & 4 & -5 \\ 1 & -4 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = 0,$$

in the three variables z_2, z_3, z_6. The solution is $z^\star = (1, -0.5, 0.5, -1, 0, 1)$. By construction $z^\star$ satisfies $A^T z^\star = 0$, and the complementarity conditions. It also satisfies $\|z^\star\|_\infty \leq 1$, hence it is optimal.

(b) All primal optimal points x must satisfy the complementarity conditions with the dual optimal $z^\star$ we have constructed. This implies that

$$(Ax + b)_2 = (Ax + b)_3 = (Ax + b)_5 = 0.$$

This forms a set of three linearly independent equations in three variables. Therefore the solution is unique.

(c) $z^\star$ remains dual feasible for nonzero ϵ. It will be optimal as long as at the optimal $x^\star(\epsilon)$,

$$(b + \epsilon d + Ax^\star(\epsilon))_k = 0, \quad k = 2, 3, 5.$$

Solving this three equations for $x^\star(\epsilon)$ yields

$$x^\star(\epsilon) = (1, 1, 1) + \epsilon(-3, 2, 0).$$

To find the limits on ϵ, we note that $z^\star$ and $x^\star(\epsilon)$ are optimal as long as

$$\begin{aligned} (A(x^\star(\epsilon) + b + \epsilon d)_1 &= 2 + 10\epsilon \geq 0 \\ (A(x^\star(\epsilon) + b + \epsilon d)_4 &= -1 + \epsilon \leq 0 \\ (A(x^\star(\epsilon) + b + \epsilon d)_6 &= 1 - 2\epsilon \geq 0 \end{aligned}$$

i.e., $-1/5 \leq \epsilon \leq 1/2$.
The optimal value is

$$p^\star(\epsilon) = (b + \epsilon d)^T z^\star = 4 + 7\epsilon.$$

5.34 Consider the pair of primal and dual LPs

$$\begin{array}{ll} \text{minimize} & (c + \epsilon d)^T x \\ \text{subject to} & Ax \preceq b + \epsilon f \end{array}$$

and

$$\begin{array}{ll} \text{maximize} & -(b + \epsilon f)^T z \\ \text{subject to} & A^T z + c + \epsilon d = 0 \\ & z \succeq 0 \end{array}$$

where

$$A = \begin{bmatrix} -4 & 12 & -2 & 1 \\ -17 & 12 & 7 & 11 \\ 1 & 0 & -6 & 1 \\ 3 & 3 & 22 & -1 \\ -11 & 2 & -1 & -8 \end{bmatrix}, \qquad b = \begin{bmatrix} 8 \\ 13 \\ -4 \\ 27 \\ -18 \end{bmatrix}, \qquad f = \begin{bmatrix} 6 \\ 15 \\ -13 \\ 48 \\ 8 \end{bmatrix},$$

$c = (49, -34, -50, -5)$, $d = (3, 8, 21, 25)$, and ϵ is a parameter.

(a) Prove that $x^\star = (1, 1, 1, 1)$ is optimal when $\epsilon = 0$, by constructing a dual optimal point $z^\star$ that has the same objective value as $x^\star$. Are there any other primal or dual optimal solutions?

(b) Give an explicit expression for the optimal value $p^\star(\epsilon)$ as a function of ϵ on an interval that contains $\epsilon = 0$. Specify the interval on which your expression is valid. Also give explicit expressions for the primal solution $x^\star(\epsilon)$ and the dual solution $z^\star(\epsilon)$ as a function of ϵ, on the same interval.

Hint. First calculate $x^\star(\epsilon)$ and $z^\star(\epsilon)$, assuming that the primal and dual constraints that are active at the optimum for $\epsilon = 0$, remain active at the optimum for values of ϵ around 0. Then verify that this assumption is correct.

Solution.

(a) All constraints except the first are active at $x = (1, 1, 1, 1)$, so complementary slackness implies that $z_1 = 0$ at the dual optimum.

For this problem, the complementary slackness condition uniquely determines z: We must have

$$\bar{A}^T \bar{z} + c = 0,$$

where

$$\bar{A} = \begin{bmatrix} -17 & 12 & 7 & 11 \\ 1 & 0 & -6 & 1 \\ 3 & 3 & 22 & -1 \\ -11 & 2 & -1 & -8 \end{bmatrix}, \quad \bar{z} = \begin{bmatrix} z_2 \\ z_3 \\ z_4 \\ z_5 \end{bmatrix}$$

$\bar{A}$ is nonsingular, so $\bar{A}^T \bar{z} + c = 0$ has a unique solution: $\bar{z} = (2, 1, 2, 2)$. All components are nonnegative, so we conclude that $z = (0, 2, 1, 2, 2)$ is dual feasible.

(b) We expect that for small ϵ the same primal and dual constraints remain active. Let us first construct $x^\star(\epsilon)$ and $z^\star(\epsilon)$ under that assumption, and then verify using complementary slackness that they are optimal for the perturbed problem.

To keep the last four constraints of $x^\star(\epsilon)$ active, we must have

$$x^\star(\epsilon) = (1, 1, 1, 1) + \epsilon \Delta x$$

where $\bar{A}\Delta x = (f_2, f_3, f_4, f_5)$. We find $\Delta x = (0, 1, 2, -1)$. $x^\star(\epsilon)$ is primal feasible as long as

$$A((1, 1, 1, 1) + \epsilon(0, 1, 2, -1) \leq b + \epsilon f.$$

By construction, this holds with equality for constraints 2–5. For the first inequality we obtain

$$7 + 7\epsilon \leq 8 + 6\epsilon.$$

i.e., $\epsilon \leq 1$.

If we keep the first component of $z^\star(\epsilon)$ zero, the other components follow from $A^T z^\star(\epsilon) + c + \epsilon d = 0$. We must have

$$z^\star(\epsilon) = (0, 2, 1, 2, 2) + \epsilon \Delta z$$

where $A^T \Delta z + f = 0$ and $\Delta z_1 = 0$. We find $\Delta z = (0, -1, 2, 0, 2)$. By construction, $z^\star(\epsilon)$ satisfies the equality constraints $A^T z^\star(\epsilon) + c + \epsilon f = 0$, so it is dual feasible if its components are nonnegative:

$$z^\star(\epsilon) = (0, 2 - \epsilon, 1 + 2\epsilon, 2, 2 + 2\epsilon) \geq 0,$$

i.e., $-1/2 \leq \epsilon \leq 2$.

In conclusion, we constructed $x^\star(\epsilon)$ and $z^\star(\epsilon)$ that are primal and dual feasible for the perturbed problem, and complementary. Therefore they must be optimal for the perturbed problems in the interval $-1/2 \leq \epsilon \leq 1$..

(c) The optimal value is quadratic

$$p^\star(\epsilon) = (c + \epsilon d)^T x^\star(\epsilon) = -(b + \epsilon f)^T z^\star(\epsilon) = -40 - 72\epsilon + 25\epsilon^2.$$

5.35 *Sensitivity analysis for GPs.* Consider a GP

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 1, \quad i = 1, \ldots, m \\ & h_i(x) = 1, \quad i = 1, \ldots, p, \end{array}$$

where $f_0, \ldots, f_m$ are posynomials, $h_1, \ldots, h_p$ are monomials, and the domain of the problem is $\mathbf{R}^n_{++}$. We define the perturbed GP as

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq e^{u_i}, \quad i = 1, \ldots, m \\ & h_i(x) = e^{v_i}, \quad i = 1, \ldots, p, \end{array}$$

and we denote the optimal value of the perturbed GP as $p^\star(u, v)$. We can think of u_i and v_i as relative, or fractional, perturbations of the constraints. For example, $u_1 = -0.01$ corresponds to tightening the first inequality constraint by (approximately) 1%.

Let $\lambda^\star$ and $\nu^\star$ be optimal dual variables for the convex form GP

$$\begin{array}{ll} \text{minimize} & \log f_0(y) \\ \text{subject to} & \log f_i(y) \leq 0, \quad i = 1, \ldots, m \\ & \log h_i(y) = 0, \quad i = 1, \ldots, p, \end{array}$$

with variables $y_i = \log x_i$. Assuming that $p^\star(u, v)$ is differentiable at $u = 0$, $v = 0$, relate $\lambda^\star$ and $\nu^\star$ to the derivatives of $p^\star(u, v)$ at $u = 0$, $v = 0$. Justify the statement "Relaxing the ith constraint by α percent will give an improvement in the objective of around $\alpha\lambda_i^\star$ percent, for α small."

Solution. $-\lambda^\star$, $-\nu^\star$ are 'shadow prices' for the perturbed problem

$$\begin{array}{ll} \text{minimize} & \log f_0(y) \\ \text{subject to} & \log f_i(y) \leq u_i, \quad i = 1, \ldots, m \\ & \log h_i(y) = v_i, \quad i = 1, \ldots, p, \end{array}$$

i.e., if the optimal value $\log p^\star(u, v)$ is differentiable at the origin, they are the derivatives of the optimal value,

$$-\lambda_i^\star = \frac{\partial \log p^\star(0,0)}{\partial u_i} = \frac{\partial p^\star(0,0)/\partial u_i}{p^\star(0,0)} \qquad -\nu_i^\star = \frac{\partial \log p^*(0,0)}{\partial v_i} = \frac{\partial p^\star(0,0)/\partial v_i}{p^\star(0,0)}.$$

Theorems of alternatives

5.36 *Alternatives for linear equalities.* Consider the linear equations $Ax = b$, where $A \in \mathbf{R}^{m\times n}$. From linear algebra we know that this equation has a solution if and only $b \in \mathcal{R}(A)$, which occurs if and only if $b \perp \mathcal{N}(A^T)$. In other words, $Ax = b$ has a solution if and only if there exists no $y \in \mathbf{R}^m$ such that $A^T y = 0$ and $b^T y \neq 0$.

Derive this result from the theorems of alternatives in §5.8.2.

Solution. We first note that we can't directly apply the results on strong alternatives for systems of the form

$$f_i(x) \leq 0, \quad i = 1, \ldots, m, \qquad Ax = b$$

or

$$f_i(x) < 0, \quad i = 1, \ldots, m, \qquad Ax = b,$$

because the theorems all *assume* that $Ax = b$ is feasible.

We can apply the theorem for strict inequalities to

$$t < -1, \qquad Ax + bt = b. \tag{5.36.A}$$

This is feasible if and only if $Ax = b$ is feasible: Indeed, if $A\tilde{x} = b$ is feasible, then $A(3\tilde{x}) - 2b = b$. so $x = 3\tilde{x}$, $t = -2$ satisfies (5.36.A). Conversely, if $\tilde{x}$, $\tilde{t}$ satisfies (5.36.A) then $1 - \tilde{t} > 2$ and

$$A(\tilde{x}/(1 - \tilde{t})) = b,$$

so $Ax = b$ is feasible.

Moreover $Ax + bt = b$ is always feasible (choose $x = 0$, $t = 1$, so we can apply the theorem of alternatives for strict inequalities to (5.36.A). The dual function is

$$g(\lambda, \nu) = \inf_{x,t}(\lambda(t + 1) + \nu^T(Ax + bt - b)) = \begin{cases} \lambda - b^T\nu & A^T\nu = 0, \quad \lambda + b^T\nu = 0 \\ -\infty & \text{otherwise.} \end{cases}$$

The alternative reduces to

$$A^T\nu = 0, \qquad b^T\nu < 0.$$

5.37 [BT97] *Existence of equilibrium distribution in finite state Markov chain.* Let $P \in \mathbf{R}^{n\times n}$ be a matrix that satisfies

$$p_{ij} \geq 0, \quad i, j = 1, \ldots, n, \qquad P^T\mathbf{1} = \mathbf{1},$$

i.e., the coefficients are nonnegative and the columns sum to one. Use Farkas' lemma to prove there exists a $y \in \mathbf{R}^n$ such that

$$Py = y, \qquad y \succeq 0, \qquad \mathbf{1}^T y = 1.$$

(We can interpret y as an equilibrium distribution of the Markov chain with n states and transition probability matrix P.)

Solution. Suppose there exists no such y, *i.e.*,

$$\begin{bmatrix} P - I \\ \mathbf{1}^T \end{bmatrix} y = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad y \succeq 0,$$

is infeasible. From Farkas' lemma there exist $z \in \mathbf{R}^n$ and $w \in \mathbf{R}$ such that

$$(P - I)^T z + w\mathbf{1} \succeq 0, \qquad w < 0,$$

i.e.,

$$P^T z \succ z.$$

Since the elements of P are nonnegative with unit column sums we must have

$$(P^T z)_i \leq \max_j z_j$$

which contradicts $P^T z \succ 1$.

5.38 [BT97] *Option pricing.* We apply the results of example 5.10, page 263, to a simple problem with three assets: a riskless asset with fixed return $r > 1$ over the investment period of interest (for example, a bond), a stock, and an option on the stock. The option gives us the right to purchase the stock at the end of the period, for a predetermined price K.

We consider two scenarios. In the first scenario, the price of the stock goes up from S at the beginning of the period, to Su at the end of the period, where $u > r$. In this scenario, we exercise the option only if $Su > K$, in which case we make a profit of $Su - K$.

Otherwise, we do not exercise the option, and make zero profit. The value of the option at the end of the period, in the first scenario, is therefore $\max\{0, Su - K\}$.

In the second scenario, the price of the stock goes down from S to Sd, where $d < 1$. The value at the end of the period is $\max\{0, Sd - K\}$.

In the notation of example 5.10,

$$V = \begin{bmatrix} r & uS & \max\{0, Su - K\} \\ r & dS & \max\{0, Sd - K\} \end{bmatrix}, \qquad p_1 = 1, \qquad p_2 = S, \qquad p_3 = C,$$

where C is the price of the option.

Show that for given r, S, K, u, d, the option price C is uniquely determined by the no-arbitrage condition. In other words, the market for the option is complete.

Solution. The condition $V^T y = p$ reduces to

$$y_1 + y_2 = 1/r, \qquad uy_1 + dy_2 = 1, \qquad y_1 \max\{0, Su - K\} + y_2 \max\{0, Sd - K\} = C.$$

The first two equations determine y_1 and y_2 uniquely:

$$y_1 = \frac{r - d}{r(u - d)}, \qquad y_2 = \frac{u - r}{r(u - d)},$$

and these values are positive because $u > r > d$. Hence

$$C = \frac{(r - d)\max\{0, Su - K\} + (u - r)\max\{0, Sd - K\}}{r(u - d)}.$$

Generalized inequalities

5.39 *SDP relaxations of two-way partitioning problem.* We consider the two-way partitioning problem (5.7), described on page 219,

$$\begin{array}{ll} \text{minimize} & x^T W x \\ \text{subject to} & x_i^2 = 1, \quad i = 1, \ldots, n, \end{array} \tag{5.113}$$

with variable $x \in \mathbf{R}^n$. The Lagrange dual of this (nonconvex) problem is given by the SDP

$$\begin{array}{ll} \text{maximize} & -\mathbf{1}^T \nu \\ \text{subject to} & W + \mathbf{diag}(\nu) \succeq 0 \end{array} \tag{5.114}$$

with variable $\nu \in \mathbf{R}^n$. The optimal value of this SDP gives a lower bound on the optimal value of the partitioning problem (5.113). In this exercise we derive another SDP that gives a lower bound on the optimal value of the two-way partitioning problem, and explore the connection between the two SDPs.

(a) *Two-way partitioning problem in matrix form.* Show that the two-way partitioning problem can be cast as

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(WX) \\ \text{subject to} & X \succeq 0, \quad \mathbf{rank}\, X = 1 \\ & X_{ii} = 1, \quad i = 1, \ldots, n, \end{array}$$

with variable $X \in \mathbf{S}^n$. *Hint.* Show that if X is feasible, then it has the form $X = xx^T$, where $x \in \mathbf{R}^n$ satisfies $x_i \in \{-1, 1\}$ (and vice versa).

(b) *SDP relaxation of two-way partitioning problem.* Using the formulation in part (a), we can form the relaxation

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(WX) \\ \text{subject to} & X \succeq 0 \\ & X_{ii} = 1, \quad i = 1, \ldots, n, \end{array} \tag{5.115}$$

with variable $X \in \mathbf{S}^n$. This problem is an SDP, and therefore can be solved efficiently. Explain why its optimal value gives a lower bound on the optimal value of the two-way partitioning problem (5.113). What can you say if an optimal point $X^\star$ for this SDP has rank one?

(c) We now have two SDPs that give a lower bound on the optimal value of the two-way partitioning problem (5.113): the SDP relaxation (5.115) found in part (b), and the Lagrange dual of the two-way partitioning problem, given in (5.114). What is the relation between the two SDPs? What can you say about the lower bounds found by them? *Hint:* Relate the two SDPs via duality.

Solution.

(a) Follows from $\mathbf{tr}(Wxx^T) = x^TWx$ and $(xx^T)_{ii} = x_i^2$.

(b) It gives a lower bound because we minimize the same objective over a larger set. If X is rank one, it is optimal.

(c) We write the problem as a minimization problem

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T\nu \\ \text{subject to} & W + \mathbf{diag}(\nu) \succeq 0. \end{array}$$

Introducing a Lagrange multiplier $X \in \mathbf{S}^n$ for the matrix inequality, we obtain the Lagrangian

$$\begin{aligned} L(\nu, X) &= \mathbf{1}^T\nu - \mathbf{tr}(X(W + \mathbf{diag}(\nu))) \\ &= \mathbf{1}^T\nu - \mathbf{tr}(XW) - \sum_{i=1}^n \nu_i X_{ii} \\ &= -\mathbf{tr}(XW) + \sum_{i=1}^n \nu_i(1 - X_{ii}). \end{aligned}$$

This is bounded below as a function of ν only if $X_{ii} = 1$ for all i, so we obtain the dual problem

$$\begin{array}{ll} \text{maximize} & -\mathbf{tr}(WX) \\ \text{subject to} & X \succeq 0 \\ & X_{ii} = 1, \quad i = 1, \ldots, n. \end{array}$$

Changing the sign again, and switching from maximization to minimization, yields the problem in part (a).

5.40 *E-optimal experiment design.* A variation on the two optimal experiment design problems of exercise 5.10 is the *E-optimal design* problem

$$\begin{array}{ll} \text{minimize} & \lambda_{\max}\left(\sum_{i=1}^p x_i v_i v_i^T\right)^{-1} \\ \text{subject to} & x \succeq 0, \quad \mathbf{1}^T x = 1. \end{array}$$

(See also §7.5.) Derive a dual for this problem, by first reformulating it as

$$\begin{array}{ll} \text{minimize} & 1/t \\ \text{subject to} & \sum_{i=1}^p x_i v_i v_i^T \succeq tI \\ & x \succeq 0, \quad \mathbf{1}^T x = 1, \end{array}$$

with variables $t \in \mathbf{R}$, $x \in \mathbf{R}^p$ and domain $\mathbf{R}_{++} \times \mathbf{R}^p$, and applying Lagrange duality. Simplify the dual problem as much as you can.

Solution.

$$\begin{array}{ll} \text{minimize} & 1/t \\ \text{subject to} & \sum_{i=1}^p x_i v_i v_i^T \succeq tI \\ & x \succeq 0, \quad \mathbf{1}^T x = 1. \end{array}$$

The Lagrangian is

$$\begin{aligned} L(t,x,Z,z,\nu) &= 1/t - \mathbf{tr}\left(Z(\sum_{i=1}^{p} x_i v_i v_i^T - tI)\right) - z^T x + \nu(\mathbf{1}^T x - 1) \\ &= 1/t + t\,\mathbf{tr}\,Z + \sum_{i=1}^{p} x_i(-v_i^T Z v_i - z_i + \nu) - \nu. \end{aligned}$$

The minimum over x_i is bounded below only if $-v_i^T Z v_i - z_i + \nu = 0$. To minimize over t we note that

$$\inf_{t>0}(1/t + t\,\mathbf{tr}\,Z) = \begin{cases} 2\sqrt{\mathbf{tr}\,Z} & Z \succeq 0 \\ -\infty & \text{otherwise.} \end{cases}$$

The dual function is

$$g(Z,z,\nu) = \begin{cases} 2\sqrt{\mathbf{tr}\,Z} - \nu & v_i^T Z v_i + z_i = \nu, \quad Z \succeq 0 \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem is

$$\begin{array}{ll} \text{maximize} & 2\sqrt{\mathbf{tr}\,Z} - \nu \\ \text{subject to} & v_i^T Z v_i \leq \nu, \quad i = 1,\ldots,p \\ & Z \succeq 0. \end{array}$$

We can define $W = (1/\nu)Z$,

$$\begin{array}{ll} \text{maximize} & 2\sqrt{\nu}\sqrt{\mathbf{tr}\,W} - \nu \\ \text{subject to} & v_i^T W v_i \geq 1, \quad i = 1,\ldots,p \\ & W \succeq 0. \end{array}$$

Finally, optimizing over ν, gives $\nu = \mathbf{tr}\,W$, so the problem simplifies further to

$$\begin{array}{ll} \text{maximize} & \mathbf{tr}\,W \\ \text{subject to} & v_i^T W v_i \leq 1, \quad i = 1,\ldots,p, \\ & W \succeq 0. \end{array}$$

5.41 *Dual of fastest mixing Markov chain problem.* On page 174, we encountered the SDP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & -tI \preceq P - (1/n)\mathbf{1}\mathbf{1}^T \preceq tI \\ & P\mathbf{1} = \mathbf{1} \\ & P_{ij} \geq 0, \quad i,j = 1,\ldots,n \\ & P_{ij} = 0 \text{ for } (i,j) \notin \mathcal{E}, \end{array}$$

with variables $t \in \mathbf{R}$, $P \in \mathbf{S}^n$.
Show that the dual of this problem can be expressed as

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^T z - (1/n)\mathbf{1}^T Y \mathbf{1} \\ \text{subject to} & \|Y\|_{2*} \leq 1 \\ & (z_i + z_j) \leq Y_{ij} \text{ for } (i,j) \in \mathcal{E} \end{array}$$

with variables $z \in \mathbf{R}^n$ and $Y \in \mathbf{S}^n$. The norm $\|\cdot\|_{2*}$ is the dual of the spectral norm on $\mathbf{S}^n$: $\|Y\|_{2*} = \sum_{i=1}^n |\lambda_i(Y)|$, the sum of the absolute values of the eigenvalues of Y. (See §A.1.6, page 639.)

Solution. We represent the Lagrange multiplier for the last constraint as $\Lambda \in \mathbf{S}^n$, with $\lambda_{ij} = 0$ for $(i,j) \in \mathcal{E}$.
The Lagrangian is

$$\begin{aligned} & L(t,P,U,V,z,W,\Lambda) \\ &= t + \mathbf{tr}(U(-tI - P + (1/n)\mathbf{1}\mathbf{1}^T)) + \mathbf{tr}(V(P - (1/n)\mathbf{1}\mathbf{1}^T - tI)) \\ &\quad + z^T(\mathbf{1} - P\mathbf{1}) - \mathbf{tr}(WP) + \mathbf{tr}(\Lambda P) \\ &= (1 - \mathbf{tr}\, U - \mathbf{tr}\, V)t + \mathbf{tr}(P(-U + V - W + \Lambda - (1/2)(\mathbf{1}z^T - z\mathbf{1}^T)) \\ &\quad + \mathbf{1}^T z + (1/n)(\mathbf{1}^T U \mathbf{1} - \mathbf{1}^T V \mathbf{1}). \end{aligned}$$

Minimizing over t and P gives the conditions

$$\mathbf{tr}\, U + \mathbf{tr}\, V = 1, \qquad (1/2)(\mathbf{1}z^T + z\mathbf{1}^T)) = V - U - W + \Lambda.$$

The dual problem is

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^T z - (1/n)\mathbf{1}^T(V-U)\mathbf{1} \\ \text{subject to} & U \succeq 0, \quad V \succeq 0, \quad \mathbf{tr}(U+V) = 1 \\ & (z_i + z_j) \leq V_{ij} - U_{ij} \text{ for } (i,j) \in \mathcal{E}. \end{array}$$

This problem is equivalent to

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^T z - (1/n)\mathbf{1}^T Y \mathbf{1} \\ \text{subject to} & \|Y\|_* \leq 1 \\ & (z_i + z_j) \leq Y_{ij} \text{ for } (i,j) \in \mathcal{E} \end{array}$$

with variables $z \in \mathbf{R}^n$, $Y \in \mathbf{S}^n$.

5.42 *Lagrange dual of conic form problem in inequality form.* Find the Lagrange dual problem of the conic form problem in inequality form

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq_K b \end{array}$$

where $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$, and K is a proper cone in $\mathbf{R}^m$. Make any implicit equality constraints explicit.
Solution. We associate with the inequality a multiplier $\lambda \in \mathbf{R}^m$, and form the Lagrangian

$$L(x,\lambda) = c^T x + \lambda^T(Ax - b).$$

The dual function is

$$\begin{aligned} g(\lambda) &= \inf_x \left(c^T x + \lambda^T(Ax-b)\right) \\ &= \begin{cases} -b^T\lambda & A^T\lambda + c = 0 \\ -\infty & \text{otherwise.} \end{cases} \end{aligned}$$

The dual problem is to maximize $g(\lambda)$ over all $\lambda \preceq_{K^*} 0$ or, equivalently,

$$\begin{array}{ll} \text{maximize} & -b^T\lambda \\ \text{subject to} & A^T\lambda + c = 0 \\ & \lambda \succeq_{K^*} 0. \end{array}$$

5.43 *Dual of SOCP.* Show that the dual of the SOCP

$$\begin{array}{ll} \text{minimize} & f^T x \\ \text{subject to} & \|A_i x + b_i\|_2 \leq c_i^T x + d_i, \quad i = 1, \ldots, m, \end{array}$$

with variables $x \in \mathbf{R}^n$, can be expressed as

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^m (b_i^T u_i + d_i v_i) \\ \text{subject to} & \sum_{i=1}^m (A_i^T u_i + c_i v_i) + f = 0 \\ & \|u_i\|_2 \leq v_i, \quad i = 1, \ldots, m, \end{array}$$

with variables $u_i \in \mathbf{R}^{n_i}$, $v_i \in \mathbf{R}$, $i = 1, \ldots, m$. The problem data are $f \in \mathbf{R}^n$, $A_i \in \mathbf{R}^{n_i \times n}$, $b_i \in \mathbf{R}^{n_i}$, $c_i \in \mathbf{R}$ and $d_i \in \mathbf{R}$, $i = 1, \ldots, m$.
Derive the dual in the following two ways.

(a) Introduce new variables $y_i \in \mathbf{R}^{n_i}$ and $t_i \in \mathbf{R}$ and equalities $y_i = A_i x + b_i$, $t_i = c_i^T x + d_i$, and derive the Lagrange dual.

(b) Start from the conic formulation of the SOCP and use the conic dual. Use the fact that the second-order cone is self-dual.

Solution.

(a) We introduce the new variables, and write the problem as

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \|y_i\|_2 \leq t_i, \quad i = 1, \ldots, m \\ & y_i = A_i x + b_i, \quad i = 1, \ldots, m \\ & t_i = c_i^T x + d_i, \quad i = 1, \ldots, m \end{array}$$

The Lagrangian is

$$\begin{aligned} & L(x, y, t, \lambda, \nu, \mu) \\ &= c^T x + \sum_{i=1}^m \lambda_i (\|y_i\|_2 - t_i) + \sum_{i=1}^m \nu_i^T (y_i - A_i x - b_i) + \sum_{i=1}^m \mu_i (t_i - c_i^T x - d_i) \\ &= (c - \sum_{i=1}^m A_i^T \nu_i - \sum_{i=1}^m \mu_i c_i)^T x + \sum_{i=1}^m (\lambda_i \|y_i\|_2 + \nu_i^T y_i) + \sum_{i=1}^m (-\lambda_i + \mu_i) t_i \\ &\quad - \sum_{i=1}^n (b_i^T \nu_i + d_i \mu_i). \end{aligned}$$

The minimum over x is bounded below if and only if

$$\sum_{i=1}^m (A_i^T \nu_i + \mu_i c_i) = c.$$

To minimize over y_i, we note that

$$\inf_{y_i} (\lambda_i \|y_i\|_2 + \nu_i^T y_i) = \begin{cases} 0 & \|\nu_i\|_2 \leq \lambda_i \\ -\infty & \text{otherwise.} \end{cases}$$

The minimum over t_i is bounded below if and only if $\lambda_i = \mu_i$. The Lagrangian is

$$g(\lambda, \nu, \mu) = \begin{cases} -\sum_{i=1}^n (b_i^T \nu_i + d_i \mu_i) & \sum_{i=1}^m (A_i^T \nu_i + \mu_i c_i) = c, \\ & \|\nu_i\|_2 \leq \lambda_i, \quad \mu = \lambda \\ -\infty & \text{otherwise} \end{cases}$$

which leads to the dual problem

$$\begin{array}{ll} \text{maximize} & -\sum_{i=1}^n (b_i^T \nu_i + d_i \lambda_i) \\ \text{subject to} & \sum_{i=1}^m (A_i^T \nu_i + \lambda_i c_i) = c \\ & \|\nu_i\|_2 \leq \lambda_i, \quad i = 1, \ldots, m. \end{array}$$

(b) We express the SOCP as a conic form problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & -(A_i x + b_i, c_i^T x + d_i) \preceq_{K_i} 0, \quad i = 1, \ldots, m. \end{array}$$

The conic dual is

$$\begin{array}{ll} \text{maximize} & -\sum_{i=1}^n (b_i^T u_i + d_i v_i) \\ \text{subject to} & \sum_{i=1}^m (A_i^T u_i + v_i c_i) = c \\ & (u_i, v_i) \succeq_{K_i^*} 0, \quad i = 1, \ldots, m. \end{array}$$

5.44 *Strong alternatives for nonstrict LMIs.* In example 5.14, page 270, we mentioned that the system

$$Z \succeq 0, \qquad \mathbf{tr}(GZ) > 0, \qquad \mathbf{tr}(F_i Z) = 0, \quad i = 1, \ldots, n, \tag{5.116}$$

is a strong alternative for the nonstrict LMI

$$F(x) = x_1 F_1 + \cdots + x_n F_n + G \preceq 0, \tag{5.117}$$

if the matrices F_i satisfy

$$\sum_{i=1}^n v_i F_i \succeq 0 \implies \sum_{i=1}^n v_i F_i = 0. \tag{5.118}$$

In this exercise we prove this result, and give an example to illustrate that the systems are not always strong alternatives.

(a) Suppose (5.118) holds, and that the optimal value of the auxiliary SDP

$$\begin{array}{ll} \text{minimize} & s \\ \text{subject to} & F(x) \preceq sI \end{array}$$

is positive. Show that the optimal value is attained. If follows from the discussion in §5.9.4 that the systems (5.117) and (5.116) are strong alternatives.

Hint. The proof simplifies if you assume, without loss of generality, that the matrices $F_1, \ldots, F_n$ are independent, so (5.118) may be replaced by $\sum_{i=1}^n v_i F_i \succeq 0 \Rightarrow v = 0$.

(b) Take $n = 1$, and

$$G = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad F_1 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Show that (5.117) and (5.116) are both infeasible.

Solution.

(a) Suppose that the optimal value is finite but not attained, *i.e.*, there exists a sequence $(x^{(k)}, s^{(k)})$, $k = 0, 1, 2, \ldots$, with

$$x_1^{(k)} F_1 + \cdots + x_n^{(k)} F_n + G \preceq s^{(k)} I \tag{5.44.A}$$

for all k, and $s^{(k)} \to s^\star > 0$. We show that the norms $\|x^{(k)}\|_2$ are bounded.

Suppose they are not. Dividing (5.44.A) by $\|x^{(k)}\|_2$, we have

$$(1/\|x^{(k)}\|_2)G + v_1^{(k)}F_1 + \cdots + v_n^{(k)}F_n \preceq w^{(k)}I,$$

where $v^{(k)} = x^{(k)}/\|x^{(k)}\|_2$, $w^{(k)} = s^{(k)}/\|x^{(k)}\|_2$. The sequence $(v^{(k)}, w^{(k)})$ is bounded, so it has a convergent subsequence. Let $\bar{v}$, $\bar{w}$ be its limit. We have

$$\bar{v}_1F_1 + \cdots + \bar{v}_nF_n \preceq 0,$$

since $\bar{w}$ must be zero. By assumption, this implies that $v = 0$, which contradicts our assumption that the sequence $x^{(k)}$ is unbounded.

Since it is bounded, the sequence $x^{(k)}$ must have a convergent subsequence. Taking limits in (5.44.A), we get

$$\bar{x}_1F_1 + \cdots + \bar{x}_nF_n + G \preceq s^\star I,$$

i.e., the optimum is attained.

(b) The LMI is

$$\begin{bmatrix} x_1 & 1 \\ 1 & 0 \end{bmatrix} \preceq 0,$$

which is infeasible. The alternative system is

$$\begin{bmatrix} z_{11} & z_{12} \\ z_{12} & z_{22} \end{bmatrix} \succeq 0, \qquad z_{22} = 0, \qquad z_{12} > 0,$$

which is also impossible.

Chapter 6

Approximation and fitting

Exercises

Norm approximation and least-norm problems

6.1 *Quadratic bounds for log barrier penalty.* Let $\phi : \mathbf{R} \to \mathbf{R}$ be the log barrier penalty function with limit $a > 0$:

$$\phi(u) = \begin{cases} -a^2 \log(1 - (u/a)^2) & |u| < a \\ \infty & \text{otherwise.} \end{cases}$$

Show that if $u \in \mathbf{R}^m$ satisfies $\|u\|_\infty < a$, then

$$\|u\|_2^2 \leq \sum_{i=1}^m \phi(u_i) \leq \frac{\phi(\|u\|_\infty)}{\|u\|_\infty^2} \|u\|_2^2.$$

This means that $\sum_{i=1}^m \phi(u_i)$ is well approximated by $\|u\|_2^2$ if $\|u\|_\infty$ is small compared to a. For example, if $\|u\|_\infty/a = 0.25$, then

$$\|u\|_2^2 \leq \sum_{i=1}^m \phi(u_i) \leq 1.033 \cdot \|u\|_2^2.$$

Solution. The left inequality follows from $\log(1 + x) \leq x$ for all $x > -1$.
The right inequality follows from convexity of $-\log(1 - x)$:

$$-\log(1 - u_i^2/a^2) \leq -\frac{u_i^2}{\|u\|_\infty^2} \log(1 - \|u\|_\infty^2/a^2)$$

and therefore

$$-a^2 \sum_{i=1}^m \log(1 - u_i^2/a^2) \leq -a^2 \frac{\|u\|_2^2}{\|u\|_\infty^2} \log(1 - \|u\|_\infty^2/a^2).$$

6.2 *ℓ_1-, ℓ_2-, and ℓ_∞-norm approximation by a constant vector.* What is the solution of the norm approximation problem with one scalar variable $x \in \mathbf{R}$,

$$\text{minimize} \quad \|x\mathbf{1} - b\|,$$

for the ℓ_1-, ℓ_2-, and ℓ_∞-norms?
Solution.

(a) ℓ_2-norm: the average $\mathbf{1}^T b/m$.

(b) ℓ_1-norm: the (or a) median of the coefficients of b.

(c) ℓ_∞-norm: the midrange point $(\max b_i - \min b_i)/2$.

6.3 Formulate the following approximation problems as LPs, QPs, SOCPs, or SDPs. The problem data are $A \in \mathbf{R}^{m \times n}$ and $b \in \mathbf{R}^m$. The rows of A are denoted a_i^T.

(a) *Deadzone-linear penalty approximation*: minimize $\sum_{i=1}^m \phi(a_i^T x - b_i)$, where

$$\phi(u) = \begin{cases} 0 & |u| \leq a \\ |u| - a & |u| > a, \end{cases}$$

where $a > 0$.

(b) *Log-barrier penalty approximation*: minimize $\sum_{i=1}^m \phi(a_i^T x - b_i)$, where

$$\phi(u) = \begin{cases} -a^2 \log(1-(u/a)^2) & |u| < a \\ \infty & |u| \geq a, \end{cases}$$

with $a > 0$.

(c) *Huber penalty approximation*: minimize $\sum_{i=1}^m \phi(a_i^T x - b_i)$, where

$$\phi(u) = \begin{cases} u^2 & |u| \leq M \\ M(2|u| - M) & |u| > M, \end{cases}$$

with $M > 0$.

(d) *Log-Chebyshev approximation*: minimize $\max_{i=1,\ldots,m} |\log(a_i^T x) - \log b_i|$. We assume $b \succ 0$. An equivalent convex form is

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & 1/t \leq a_i^T x/b_i \leq t, \quad i = 1, \ldots, m, \end{array}$$

with variables $x \in \mathbf{R}^n$ and $t \in \mathbf{R}$, and domain $\mathbf{R}^n \times \mathbf{R}_{++}$.

(e) Minimizing the sum of the largest k residuals:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^k |r|_{[i]} \\ \text{subject to} & r = Ax - b, \end{array}$$

where $|r|_{[1]} \geq |r|_{[2]} \geq \cdots \geq |r|_{[m]}$ are the numbers $|r_1|$, $|r_2|$, $\ldots$, $|r_m|$ sorted in decreasing order. (For $k = 1$, this reduces to ℓ_∞-norm approximation; for $k = m$, it reduces to ℓ_1-norm approximation.) *Hint.* See exercise 5.19.

Solution.

(a) *Deadzone-linear.*

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T y \\ \text{subject to} & -y - a\mathbf{1} \preceq Ax - b \preceq y + a\mathbf{1} \\ & y \succeq 0. \end{array}$$

An LP with variables $y \in \mathbf{R}^m$, $x \in \mathbf{R}^n$.

(b) *Log-barrier penalty.* We can express the problem as

$$\begin{array}{ll} \text{maximize} & \prod_{i=1}^m t_i^2 \\ \text{subject to} & (1 - y_i/a)(1 + y_i/a) \geq t_i^2, \quad i = 1, \ldots, m \\ & -1 \leq y_i/a \leq 1, \quad i = 1, \ldots, m \\ & y = Ax - b, \end{array}$$

with variables $t \in \mathbf{R}^m$, $y \in \mathbf{R}^m$, $x \in \mathbf{R}^n$.

We can now proceed as in exercise 4.26 (maximizing geometric mean), and reduce the problem to an SOCP or an SDP.

(c) *Huber penalty.* See exercise 4.5 (c), and also exercise 6.6.

(d) *Log-Chebyshev approximation.*

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & 1/t \leq a_i^T x/b_i \leq t, \quad i = 1, \ldots, m \end{array}$$

over $x \in \mathbf{R}^n$ and $t \in \mathbf{R}$. The left inequalities are hyperbolic constraints

$$t a_i^T x \geq b_i, \qquad t \geq 0, \qquad a_i^T x \geq 0$$

that can be formulated as LMI constraints

$$\begin{bmatrix} t & \sqrt{b_i} \\ \sqrt{b_i} & a_i^T x \end{bmatrix} \succeq 0,$$

or SOC constraints

$$\left\| \begin{bmatrix} 2\sqrt{b_i} \\ t - a_i^T x \end{bmatrix} \right\|_2 \leq t + a_i^T x.$$

(e) *Sum of largest residuals.*

$$\begin{array}{ll} \text{minimize} & kt + \mathbf{1}^T z \\ \text{subject to} & -t\mathbf{1} - z \preceq Ax - b \preceq t\mathbf{1} + z \\ & z \succeq 0, \end{array}$$

with variables $x \in \mathbf{R}^n$, $t \in \mathbf{R}$, $z \in \mathbf{R}^m$.

6.4 *A differentiable approximation of ℓ_1-norm approximation.* The function $\phi(u) = (u^2+\epsilon)^{1/2}$, with parameter $\epsilon > 0$, is sometimes used as a differentiable approximation of the absolute value function $|u|$. To approximately solve the ℓ_1-norm approximation problem

$$\text{minimize} \quad \|Ax - b\|_1, \tag{6.26}$$

where $A \in \mathbf{R}^{m\times n}$, we solve instead the problem

$$\text{minimize} \quad \sum_{i=1}^m \phi(a_i^T x - b_i), \tag{6.27}$$

where a_i^T is the ith row of A. We assume $\mathbf{rank}\, A = n$.
Let $p^\star$ denote the optimal value of the ℓ_1-norm approximation problem (6.26). Let $\hat{x}$ denote the optimal solution of the approximate problem (6.27), and let $\hat{r}$ denote the associated residual, $\hat{r} = A\hat{x} - b$.

(a) Show that $p^\star \geq \sum_{i=1}^m \hat{r}_i^2/(\hat{r}_i^2 + \epsilon)^{1/2}$.

(b) Show that

$$\|A\hat{x} - b\|_1 \leq p^\star + \sum_{i=1}^m |\hat{r}_i| \left(1 - \frac{|\hat{r}_i|}{(\hat{r}_i^2 + \epsilon)^{1/2}}\right).$$

(By evaluating the righthand side after computing $\hat{x}$, we obtain a bound on how suboptimal $\hat{x}$ is for the ℓ_1-norm approximation problem.)

Solution. One approach is based on duality. The point $\hat{x}$ minimizes the differentiable convex function $\sum_{i=1}^m \phi(a_i^T x - b_i)$, so its gradient vanishes:

$$\sum_{i=1}^m \phi'(\hat{r}_i) a_i = \sum_{i=1}^m \hat{r}_i (\hat{r}_i^2 + \epsilon)^{-1/2} a_i = 0.$$

Now, the dual of the ℓ_1-norm approximation problem is

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^m b_i \lambda_i \\ \text{subject to} & |\lambda_i| \leq 1, \quad i = 1, \ldots, m \\ & \sum_{i=1}^m \lambda_i a_i = 0. \end{array}$$

Thus, we see that the vector

$$\lambda_i = -\frac{\hat{r}_i}{(\hat{r}_i^2 + \epsilon)^{-1/2}}, \quad i = 1, \ldots, m,$$

is dual feasible. It follows that its dual function value,

$$\sum_{i=1}^{m} -b_i\lambda_i = \frac{-b_i\hat{r}_i}{(\hat{r}_i^2+\epsilon)^{-1/2}},$$

provides a lower bound on $p^\star$. Now we use the fact that $\sum_{i=1}^m \lambda_i a_i = 0$ to obtain

$$\begin{aligned} p^\star &\geq \sum_{i=1}^{m} -b_i\lambda_i \\ &= \sum_{i=1}^{m} (a_i^T\hat{x} - b_i)\lambda_i \\ &= \sum_{i=1}^{m} \hat{r}_i\lambda_i \\ &= \frac{\hat{r}_i^2}{(\hat{r}_i^2+\epsilon)^{-1/2}}. \end{aligned}$$

Now we establish part (b). We start with the result above,

$$p^\star \geq \sum_{i=1}^{m} \hat{r}_i^2/(\hat{r}_i^2+\epsilon)^{1/2},$$

and subtract $\|A\hat{x}-b\|_1 = \sum_{i=1}^m |\hat{r}_i|$ from both sides to get

$$p^\star - \|A\hat{x}-b\|_1 \geq \sum_{i=1}^{m} \left(\hat{r}_i^2/(\hat{r}_i^2+\epsilon)^{1/2} - |\hat{r}_i|\right).$$

Re-arranging gives the desired result,

$$\|A\hat{x}-b\|_1 \leq p^\star + \sum_{i=1}^{m} |r_i| \left(1 - \frac{|r_i|}{(r_i^2+\epsilon)^{1/2}}\right).$$

6.5 *Minimum length approximation.* Consider the problem

$$\begin{array}{ll} \text{minimize} & \text{length}(x) \\ \text{subject to} & \|Ax-b\| \leq \epsilon, \end{array}$$

where $\text{length}(x) = \min\{k \mid x_i = 0 \text{ for } i > k\}$. The problem variable is $x \in \mathbf{R}^n$; the problem parameters are $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$, and $\epsilon > 0$. In a regression context, we are asked to find the minimum number of columns of A, taken in order, that can approximate the vector b within ϵ.

Show that this is a quasiconvex optimization problem.

Solution. $\text{length}(x) \leq \alpha$ if and only if $x_k = 0$ for $k > \alpha$. Thus, the sublevel sets of length are convex, so length is quasiconvex.

6.6 *Duals of some penalty function approximation problems.* Derive a Lagrange dual for the problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m \phi(r_i) \\ \text{subject to} & r = Ax - b, \end{array}$$

for the following penalty functions $\phi : \mathbf{R} \to \mathbf{R}$. The variables are $x \in \mathbf{R}^n$, $r \in \mathbf{R}^m$.

(a) *Deadzone-linear penalty* (with deadzone width $a = 1$),

$$\phi(u) = \begin{cases} 0 & |u| \leq 1 \\ |u| - 1 & |u| > 1. \end{cases}$$

(b) *Huber penalty* (with $M = 1$),

$$\phi(u) = \begin{cases} u^2 & |u| \leq 1 \\ 2|u| - 1 & |u| > 1. \end{cases}$$

(c) *Log-barrier* (with limit $a = 1$),

$$\phi(u) = -\log(1 - x^2), \qquad \mathbf{dom}\,\phi = (-1, 1).$$

(d) *Relative deviation from one,*

$$\phi(u) = \max\{u, 1/u\} = \begin{cases} u & u \geq 1 \\ 1/u & u \leq 1, \end{cases}$$

with $\mathbf{dom}\,\phi = \mathbf{R}_{++}$.

Solution. We first derive a dual for general penalty function approximation. The Lagrangian is

$$L(x, r, \lambda) = \sum_{i=1}^{m} \phi(r_i) + \nu^T(Ax - b - r).$$

The minimum over x is bounded if and only if $A^T\nu = 0$, so we have

$$g(\nu) = \begin{cases} -b^T\nu + \sum_{i=1}^m \inf_{r_i}(\phi(r_i) - \nu_i r_i) & A^T\nu = 0 \\ -\infty & \text{otherwise.} \end{cases}$$

Using

$$\inf_{r_i}(\phi(r_i) - \nu_i r_i) = -\sup_{r_i}(\nu_i r_i - \phi(r_i)) = -\phi^*(\nu_i),$$

we can express the general dual as

$$\begin{array}{ll} \text{maximize} & -b^T\nu - \sum_{i=1}^m \phi^*(\nu_i) \\ \text{subject to} & A^T\nu = 0. \end{array}$$

Now we'll work out the conjugates of the given penalty functions.

(a) *Deadzone-linear penalty.* The conjugate of the deadzone-linear function is

$$\phi^*(z) = \begin{cases} |z| & |z| \leq 1 \\ \infty & |z| > 1, \end{cases}$$

so the dual of the dead-zone linear penalty function approximation problem is

$$\begin{array}{ll} \text{maximize} & -b^T\nu - \|\nu\|_1 \\ \text{subject to} & A^T\nu = 0, \quad \|\nu\|_\infty \leq 1. \end{array}$$

(b) *Huber penalty.*

$$\phi^*(z) = \begin{cases} z^2/4 & |z| \leq 2 \\ \infty & \text{otherwise,} \end{cases}$$

so we get the dual problem

$$\begin{array}{ll} \text{maximize} & -(1/4)\|\nu\|_2^2 - b^T\nu \\ \text{subject to} & A^T\nu = 0 \\ & \|\nu\|_\infty \leq 2. \end{array}$$

(c) *Log-barrier.* The conjugate of ϕ is

$$\begin{aligned}\phi^*(z) &= \sup_{|x|<1}\left(xz + \log(1-x^2)\right) \\ &= -1 + \sqrt{1+z^2} + \log(-1+\sqrt{1+z^2}) - 2\log|z| + \log 2.\end{aligned}$$

(d) *Relative deviation from one.* Here we have

$$\phi^*(z) = \sup_{x>0}(xz - \max\{x, 1/x\}) = \begin{cases} -2\sqrt{-z} & z \le -1 \\ z-1 & -1 \le z \le 1 \\ -\infty & z > 1. \end{cases}$$

Plugging this in the dual problem gives

$$\begin{array}{ll}\text{maximize} & -b^T\nu + \sum_{i=1}^m s(\nu_i) \\ \text{subject to} & A^T\nu = 0, \quad \nu \preceq \mathbf{1},\end{array}$$

where

$$s(\nu_i) = \begin{cases} 2\sqrt{-\nu_i} & \nu_i \le -1 \\ 1-\nu_i & \nu_i \ge -1. \end{cases}$$

Regularization and robust approximation

6.7 *Bi-criterion optimization with Euclidean norms.* We consider the bi-criterion optimization problem

$$\text{minimize (w.r.t. } \mathbf{R}^2_+) \quad (\|Ax-b\|_2^2, \|x\|_2^2),$$

where $A \in \mathbf{R}^{m\times n}$ has rank r, and $b \in \mathbf{R}^m$. Show how to find the solution of each of the following problems from the singular value decomposition of A,

$$A = U\,\mathbf{diag}(\sigma)V^T = \sum_{i=1}^r \sigma_i u_i v_i^T$$

(see §A.5.4).

(a) *Tikhonov regularization:* minimize $\|Ax-b\|_2^2 + \delta\|x\|_2^2$.

(b) Minimize $\|Ax-b\|_2^2$ subject to $\|x\|_2^2 = \gamma$.

(c) Maximize $\|Ax-b\|_2^2$ subject to $\|x\|_2^2 = \gamma$.

Here δ and γ are positive parameters.

Your results provide efficient methods for computing the optimal trade-off curve and the set of achievable values of the bi-criterion problem.

Solution. Define

$$\tilde{x} = (V^Tx, V_2^Tx), \qquad \tilde{b} = (U^Tb, U_2^Tb).$$

where $V_2 \in \mathbf{R}^{n\times(n-r)}$ satisfies $V_2^TV_2 = I$, $V_2^TV = 0$, and $U_2 \in \mathbf{R}^{m\times(m-r)}$ satisfies $U_2^TU_2 = I$, $U_2^TU = 0$. We have

$$\|Ax-b\|_2^2 = \sum_{i=1}^r(\sigma_i\tilde{x}_i - \tilde{b}_i)^2 + \sum_{i=r+1}^m \tilde{b}_i^2, \qquad \|x\|_2^2 = \sum_{i=1}^n \tilde{x}_i^2.$$

We will use $\tilde{x}$ as variable.

(a) *Tikhonov regularization.* Setting the gradient (with respect to $\tilde{x}$) to zero gives

$$(\sigma_i^2 + \delta)\tilde{x}_i = \sigma_i \tilde{b}_i, \quad i = 1, \ldots, r, \qquad \tilde{x}_i = 0, \quad i = r+1, \ldots, n.$$

The solution is

$$\tilde{x}_i = \frac{\tilde{b}_i \sigma_i}{\delta + \sigma_i^2}, \quad i = 1, \ldots, r, \qquad \tilde{x}_i = 0, \quad i = r+1, \ldots, n.$$

In terms of the original variables,

$$x = \sum_{i=1}^{r} \frac{\sigma_i}{\delta + \sigma_i^2} (u_i^T b) v_i.$$

If $\delta = 0$, this is the least-squares solution

$$x = A^\dagger b = V \Sigma^{-1} U^T b = \sum_{i=1}^{r} (1/\sigma_i)(u_i^T b) v_i.$$

If $\delta > 0$, each component $(u_i^T b)v_i$ receives a weight $\sigma_i/(\delta + \sigma_i^2)$. The function $\sigma/(\delta + \sigma^2)$ is zero if $\sigma = 0$, goes through a maximum of $1/(1+\delta)$ at $\sigma = \delta$, and decreases to zero as $1/\sigma$ for $\sigma \to \infty$.

In other words, if σ_i is large ($\sigma_i \gg \delta$), we keep the ith term in the LS solution. For small σ_i ($\sigma_i \approx \delta$ or less), we dampen its weight, replacing $1/\sigma_i$ by $\sigma_i/(\delta + \sigma_i^2)$.

(b) After the change of variables, this problem is

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{r} (\sigma_i \tilde{x}_i - \tilde{b}_i)^2 + \sum_{i=r+1}^{m} \tilde{b}_i^2 \\ \text{subject to} & \sum_{i=1}^{n} \tilde{x}_i^2 = \gamma. \end{array}$$

Although the problem is not convex, it is clear that a necessary and sufficient condition for a feasible $\tilde{x}$ to be optimal is that either the gradient of the objective vanishes at $\tilde{x}$, or the gradient is normal to the sphere through $\tilde{x}$, and pointing toward the interior of the sphere. In other words, the optimality conditions are that $\|\tilde{x}\|_2^2 = \gamma$ and there exists a $\nu \geq 0$, such that

$$(\sigma_i^2 + \nu)\tilde{x}_i = \sigma_i \tilde{b}_i, \quad i = 1, \ldots, r, \qquad \nu \tilde{x}_i = 0, \quad i = r+1, \ldots, n.$$

We distinguish two cases.

- If $\sum_{i=1}^{r} (\tilde{b}_i/\sigma_i)^2 \leq \gamma$, then $\nu = 0$ and

  $$\tilde{x}_i = \tilde{b}_i \sigma_i, \quad i = 1, \ldots, r,$$

  (*i.e.*, the unconstrained minimum) is optimal. For the other variables we can choose any $\tilde{x}_i$, $i = r+1, \ldots, n$ that gives $\|\tilde{x}\|_2^2 = \gamma$.
- If $\sum_{i=1}^{r} (\tilde{b}_i/\sigma_i)^2 > \gamma$, we must take $\nu > 0$, and

  $$\tilde{x}_i = \frac{\tilde{b}_i \sigma_i}{\sigma_i^2 + \nu}, \quad i = 1, \ldots, r, \qquad \tilde{x}_i = 0, \quad i = r+1, \ldots, n.$$

  We determine $\nu > 0$ by solving the nonlinear equation

  $$\sum_{i=1}^{n} \tilde{x}_i^2 = \sum_{i=1}^{r} \left(\frac{\tilde{b}_i \sigma_i}{\sigma_i^2 + \nu} \right)^2 = \gamma.$$

  The left hand side is monotonically decreasing with ν, and by assumption it is greater than γ at $\nu = 0$, so the equation has a unique positive solution.

(c) After the change of variables to $\tilde{x}$, this problem reduces to

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{r}(\sigma_i\tilde{x}_i - \tilde{b}_i)^2 + \sum_{i=r+1}^{m}\tilde{b}_i^2 \\ \text{subject to} & \sum_{i=1}^{n}\tilde{x}_i^2 = \gamma. \end{array}$$

Without loss of generality we can replace the equality with an inequality, since a convex function reaches its maximum over a compact convex on the boundary. As shown in §B.1, strong duality holds for quadratic optimization problems with one inequality constraint.

In this case, however, it is also easy to derive this result directly, without appealing to the general result in §B.1. We will first derive and solve the dual, and then show strong duality by establishing a feasible $\tilde{x}$ with the same primal objective value as the dual optimum.

The Lagrangian of the problem above (after switching the sign of the objective) is

$$\begin{aligned} L(\tilde{x},\nu) &= -\sum_{i=1}^{r}(\sigma_i\tilde{x}_i - \tilde{b}_i)^2 - \sum_{i=r+1}^{n}\tilde{b}_i^2 + \nu(\sum_{i=1}^{n}\tilde{x}_i^2 - \gamma) \\ &= \sum_{i=1}^{r}(\nu - \sigma_i^2)\tilde{x}_i^2 + 2\sum_{i=1}^{r}\sigma_i\tilde{b}_i\tilde{x}_i - \sum_{i=1}^{n}\tilde{b}_i^2 - \nu\gamma. \end{aligned}$$

L is bounded below as a function of $\tilde{x}$ only if $\nu > \sigma_1^2$, or if $\nu = \sigma_1^2$ and $\tilde{b}_1 = 0$. The infimum is

$$\inf_{\tilde{x}} L(\tilde{x},\nu) = -\sum_{i=1}^{r}\frac{(\sigma_i\tilde{b}_i)^2}{\nu - \sigma_i^2} - \sum_{i=1}^{n}\tilde{b}_i^2 - \nu\gamma,$$

with domain $[\sigma_1^2,\infty)$, and where for $\nu = \sigma_1^2$ we interpret $\tilde{b}_1^2/(\nu - \sigma_1^2)$ as ∞ if $\tilde{b}_1 \neq 0$, and as 0 if $\tilde{b}_1 = 0$. The dual problem is therefore (after switching back to maximization)

$$\begin{array}{ll} \text{minimize} & g(\nu) = \sum_{i=1}^{r}(\tilde{b}_i\sigma_i)^2/(\nu - \sigma_i^2) + \nu\gamma + \sum_{i=1}^{n}\tilde{b}_i^2 \\ \text{subject to} & \nu \geq \sigma_1^2. \end{array}$$

The derivative of g is

$$g'(\nu) = -\sum_{i=1}^{r}\frac{(\tilde{b}_i\sigma_i)^2}{(\nu - \sigma_i^2)^2} + \gamma.$$

We can distinguish three cases. We assume that the first singular value is repeated k times where $k \leq r$.

- $g(\sigma_1^2) = \infty$. This is the case if at least one of the coefficients $\tilde{b}_1, \ldots, \tilde{b}_k$ is nonzero.
  In this case g first decreases as we increase $\nu > \sigma_1^2$ and then increases as ν goes to infinity. There is therefore a unique $\nu > \sigma_1^2$ where the derivative is zero:

  $$\sum_{i=1}^{r}\frac{(\tilde{b}_i\sigma_i)^2}{(\nu - \sigma_i^2)^2} = \gamma.$$

  From ν we compute the optimal primal $\tilde{x}$ as

  $$\tilde{x}_i = \frac{-\sigma_i\tilde{b}_i}{\nu - \sigma_i^2}, \quad i = 1,\ldots,r, \qquad \tilde{x}_i = 0, \quad i = r+1,\ldots,n.$$

This point satisfies $\|\tilde{x}\|^2 = \gamma$ and its objective value is

$$\begin{aligned}\sum_{i=1}^{r}\sigma_i^2\tilde{x}_i^2 - 2\sum_{i=1}^{r}\sigma_i\tilde{b}_i\tilde{x}_i + \sum_{i=1}^{n}\tilde{b}_i^2 &= \sum_{i=1}^{r}(\sigma_i^2-\nu)\tilde{x}_i^2 - 2\sum_{i=1}^{r}\sigma_i\tilde{b}_i\tilde{x}_i + \sum_{i=1}^{n}\tilde{b}_i^2 + \nu\gamma \\ &= \sum_{i=1}^{r}\frac{\sigma_i^2\tilde{b}_i^2}{\nu-\sigma_i^2} + \sum_{i=1}^{n}\tilde{b}_i^2 + \nu\gamma \\ &= g(\nu).\end{aligned}$$

By weak duality, this means $\tilde{x}$ is optimal.

- $g(\sigma_1^2)$ is finite and $g'(\sigma_1^2) < 0$. This is the case when $\tilde{b}_1 = \cdots = \tilde{b}_k = 0$ and

$$g'(\sigma_1^2) = -\sum_{i=k+1}^{r}\frac{(\tilde{b}_i\sigma_i)^2}{(\sigma_1^2-\sigma_i^2)^2} + \gamma < 0.$$

  As we increase $\nu > \sigma_1^2$, the dual objective first decreases, and then increases as ν goes to infinity. The solution is the same as in the previous case: we compute ν by solving $g'(\nu) = 0$, and then calculate $\tilde{x}$ as above.
- $g(\sigma_1^2)$ is finite and $g'(\sigma_1^2) \geq 0$. This is the case when $\tilde{b}_1 = \cdots = \tilde{b}_k = 0$ and

$$g'(\sigma_1^2) = -\sum_{i=k+1}^{r}\frac{(\tilde{b}_i\sigma_i)^2}{(\sigma_1^2-\sigma_i^2)^2} + \gamma \geq 0.$$

  In this case $\nu = \sigma_1^2$ is optimal. A primal optimal solution is

$$\tilde{x}_i = \begin{cases} \sqrt{g'(\nu)} & i = 1 \\ 0 & i = 1,\ldots,k \\ -\tilde{b}_i\sigma_i/(\sigma_1^2-\sigma_i^2) & i = k+1,\ldots,r \\ 0 & i = r+1,\ldots,n. \end{cases}$$

  (The first k coefficients are arbitrary as long as their squares add up to $g'(\nu)$.) To verify that $\tilde{x}$ is optimal, we note that it is feasible, *i.e.*,

$$\|\tilde{x}\|_2^2 = g'(\nu) + \sum_{i=k+1}^{r}\frac{\tilde{b}_i^2\sigma_i^2}{(\sigma_1^2-\sigma_i^2)^2} = \gamma,$$

  and that its objective value equals $g(\sigma_1^2)$:

$$\begin{aligned}\sum_{i=1}^{r}(\sigma_i^2\tilde{x}_i^2 - 2\sigma_i\tilde{b}_i\tilde{x}_i) &= \sigma_1^2 g'(\sigma_1^2) + \sum_{i=k+1}^{r}(\sigma_i^2\tilde{x}_i^2 - 2\sigma_i\tilde{b}_i\tilde{x}_i) \\ &= \sigma_1^2\left(g'(\sigma_1^2) + \sum_{i=k+1}^{r}\tilde{x}_i^2\right) + \sum_{i=k+1}^{r}\left((\sigma_i^2-\sigma_1^2)\tilde{x}_i^2 - 2\sigma_i\tilde{b}_i\tilde{x}_i\right) \\ &= \sigma_1^2\gamma + \sum_{i=k+1}^{r}\left((\sigma_i^2-\sigma_1^2)\tilde{x}_i^2 - 2\sigma_i\tilde{b}_i\tilde{x}_i\right) \\ &= \sigma_1^2\gamma + \sum_{i=k+1}^{r}\frac{(\tilde{b}_i\sigma_i)^2}{\sigma_1^2-\sigma_i^2} \\ &= g(\sigma_1^2) - \sum_{i=1}^{n}\tilde{b}_i^2.\end{aligned}$$

6.8 Formulate the following robust approximation problems as LPs, QPs, SOCPs, or SDPs. For each subproblem, consider the ℓ_1-, ℓ_2-, and the ℓ_∞-norms.

(a) *Stochastic robust approximation with a finite set of parameter values, i.e.*, the sum-of-norms problem

$$\text{minimize} \quad \sum_{i=1}^k p_i \|A_i x - b\|$$

where $p \succeq 0$ and $\mathbf{1}^T p = 1$. (See §6.4.1.)

Solution.

- ℓ_1-norm:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^k p_i \mathbf{1}^T y_i \\ \text{subject to} & -y_i \preceq A_i x - b \preceq y_i, \quad i = 1, \ldots, k. \end{array}$$

  An LP with variables $x \in \mathbf{R}^n$, $y_i \in \mathbf{R}^m$, $i = 1, \ldots, k$.

- ℓ_2-norm:

$$\begin{array}{ll} \text{minimize} & p^T y \\ \text{subject to} & \|A_i x - b\|_2 \leq y_i, \quad i = 1, \ldots, k. \end{array}$$

  An SOCP with variables $x \in \mathbf{R}^n$, $y \in \mathbf{R}^k$.

- ℓ_∞-norm:

$$\begin{array}{ll} \text{minimize} & p^T y \\ \text{subject to} & -y_i \mathbf{1} \preceq A_i x - b \leq y_i \mathbf{1}, \quad i = 1, \ldots, k. \end{array}$$

  An LP with variables $x \in \mathbf{R}^n$, $y \in \mathbf{R}^k$.

(b) *Worst-case robust approximation with coefficient bounds:*

$$\text{minimize} \quad \sup_{A \in \mathcal{A}} \|Ax - b\|$$

where

$$\mathcal{A} = \{A \in \mathbf{R}^{m \times n} \mid l_{ij} \leq a_{ij} \leq u_{ij}, \ i = 1, \ldots, m, \ j = 1, \ldots, n\}.$$

Here the uncertainty set is described by giving upper and lower bounds for the components of A. We assume $l_{ij} < u_{ij}$.

Solution. We first note that

$$\begin{aligned} \sup_{l_{ij} \leq a_{ij} \leq u_{ij}} |a_i^T x - b_i| &= \sup_{l_{ij} \leq a_{ij} \leq u_{ij}} \max\{a_i^T x - b_i, -a_i^T x + b_i\} \\ &= \max\{\sup_{l_{ij} \leq a_{ij} \leq u_{ij}} (a_i^T x - b_i), \sup_{l_{ij} \leq a_{ij} \leq u_{ij}} (-a_i^T x + b_i)\}. \end{aligned}$$

Now,

$$\sup_{l_{ij} \leq a_{ij} \leq u_{ij}} (\sum_{j=1}^n a_{ij} x_j - b_i) = \bar{a}_i^T x - b_i + \sum_{j=1}^n v_{ij} |x_j|$$

where $\bar{a}_{ij} = (l_{ij} + u_{ij})/2$, and $v_{ij} = (u_{ij} - l_{ij})/2$, and

$$\sup_{l_{ij} \leq a_{ij} \leq u_{ij}} (-\sum_{j=1}^n a_{ij} x_j + b_i) = -\bar{a}_i^T x + b_i + \sum_{j=1}^n v_{ij} |x_j|.$$

Therefore

$$\sup_{l_{ij} \leq a_{ij} \leq u_{ij}} |a_i^T x - b_i| = |\bar{a}_i^T x - b_i| + \sum_{j=1}^n v_{ij} |x_j|.$$

- ℓ_1-norm:
$$\text{minimize} \quad \sum_{i=1}^m \left(|\bar{a}_i^T x - b_i| + \sum_{j=1}^n v_{ij}|x_j| \right).$$
This can be expressed as an LP
$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T(y + Vw) \\ & -y \preceq \bar{A}x - b \preceq y \\ & -w \preceq x \preceq w. \end{array}$$
The variables are $x \in \mathbf{R}^n$, $y \in \mathbf{R}^m$, $w \in \mathbf{R}^n$.
- ℓ_2-norm:
$$\text{minimize} \quad \sum_{i=1}^m \left(|\bar{a}_i^T x - b_i| + \sum_{j=1}^n v_{ij}|x_j| \right)^2.$$
This can be expressed as an SOCP
$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & -y \preceq \bar{A}x - b \preceq y \\ & -w \preceq x \preceq w \\ & \|y + Vw\|_2 \leq t. \end{array}$$
The variables are $x \in \mathbf{R}^n$, $y \in \mathbf{R}^m$, $w \in \mathbf{R}^n$, $t \in \mathbf{R}$.
- ℓ_∞-norm:
$$\text{minimize} \quad \max_{i=1,\ldots,m} \left(|\bar{a}_i^T x - b_i| + \sum_{j=1}^n v_{ij}|x_j| \right).$$
This can be expressed as an LP
$$\begin{array}{ll} \text{minimize} & t \\ & -y \preceq \bar{A}x - b \preceq y \\ & -w \preceq x \preceq w \\ & -t\mathbf{1} \preceq y + Vw \leq t\mathbf{1}. \end{array}$$
The variables are $x \in \mathbf{R}^n$, $y \in \mathbf{R}^m$, $w \in \mathbf{R}^n$, $t \in \mathbf{R}$.

(c) *Worst-case robust approximation with polyhedral uncertainty:*
$$\text{minimize} \quad \sup_{A \in \mathcal{A}} \|Ax - b\|$$
where
$$\mathcal{A} = \{[a_1 \ \cdots \ a_m]^T \mid C_i a_i \preceq d_i, \ i = 1, \ldots, m\}.$$
The uncertainty is described by giving a polyhedron $\mathcal{P}_i = \{a_i \mid C_i a_i \preceq d_i\}$ of possible values for each row. The parameters $C_i \in \mathbf{R}^{p_i \times n}$, $d_i \in \mathbf{R}^{p_i}$, $i = 1, \ldots, m$, are given. We assume that the polyhedra $\mathcal{P}_i$ are nonempty and bounded.

Solution. $\mathcal{P}_i = \{a \mid C_i a \preceq d_i\}$.
$$\begin{aligned} \sup_{a_i \in \mathcal{P}_i} |a_i^T x - b_i| &= \sup_{a_i \in \mathcal{P}_i} \max\{a_i^T x - b_i, -a_i^T x + b_i\} \\ &= \max\{\sup_{a_i \in \mathcal{P}_i} (a_i^T x) - b_i, \ \sup_{a_i \in \mathcal{P}_i} (-a_i^T x) + b_i\}. \end{aligned}$$
By LP duality,
$$\begin{aligned} \sup_{a_i \in \mathcal{P}_i} a_i^T x &= \inf\{d_i^T v \mid C_i^T v = x, \ v \succeq 0\} \\ \sup_{a_i \in \mathcal{P}_i} (-a_i^T x) &= \inf\{d_i^T w \mid C_i^T w = -x, \ w \succeq 0\}. \end{aligned}$$

Therefore, $t_i \geq \sup_{a_i \in \mathcal{P}_i} |a_i^T x - b_i|$ if and only if there exist v, w, such that

$$v, w \succeq 0, \qquad x = C_i^T v = -C_i^T w, \qquad d_i^T v \leq t_i, \qquad d_i^T w \leq t_i.$$

This allows us to pose the robust approximation problem as

$$\begin{array}{ll} \text{minimize} & \|t\| \\ \text{subject to} & x = C_i^T v_i, \quad x = -C_i^T w_i, \quad i = 1, \ldots, m \\ & d_i^T v_i \leq t_i, \qquad d_i^T w_i \leq t_i, \quad i = 1, \ldots, m \\ & v_i, w_i \succeq 0, \quad i = 1, \ldots, m. \end{array}$$

- ℓ_1-norm:

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T t \\ \text{subject to} & x = C_i^T v_i, \quad x = -C_i^T w_i, \quad i = 1, \ldots, m \\ & d_i^T v_i \leq t_i, \qquad d_i^T w_i \leq t_i, \quad i = 1, \ldots, m \\ & v_i, w_i \succeq 0, \quad i = 1, \ldots, m. \end{array}$$

- ℓ_2-norm:

$$\begin{array}{ll} \text{minimize} & u \\ \text{subject to} & x = C_i^T v_i, \quad x = -C_i^T w_i, \quad i = 1, \ldots, m \\ & d_i^T v_i \leq t_i, \qquad d_i^T w_i \leq t_i, \quad i = 1, \ldots, m \\ & v_i, w_i \succeq 0, \quad i = 1, \ldots, m \\ & \|t\|_2 \leq u. \end{array}$$

- ℓ_∞-norm:

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & x = C_i^T v_i, \quad x = -C_i^T w_i, \quad i = 1, \ldots, m \\ & d_i^T v_i \leq t, \qquad d_i^T w_i \leq t, \quad i = 1, \ldots, m \\ & v_i, w_i \succeq 0, \quad i = 1, \ldots, m. \end{array}$$

Function fitting and interpolation

6.9 *Minimax rational function fitting.* Show that the following problem is quasiconvex:

$$\text{minimize} \quad \max_{i=1,\ldots,k} \left| \frac{p(t_i)}{q(t_i)} - y_i \right|$$

where

$$p(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_m t^m, \qquad q(t) = 1 + b_1 t + \cdots + b_n t^n,$$

and the domain of the objective function is defined as

$$D = \{(a, b) \in \mathbf{R}^{m+1} \times \mathbf{R}^n \mid q(t) > 0, \ \alpha \leq t \leq \beta\}.$$

In this problem we fit a rational function $p(t)/q(t)$ to given data, while constraining the denominator polynomial to be positive on the interval $[\alpha, \beta]$. The optimization variables are the numerator and denominator coefficients a_i, b_i. The interpolation points $t_i \in [\alpha, \beta]$, and desired function values y_i, $i = 1, \ldots, k$, are given.

Solution. Let's show the objective is quasiconvex. Its domain is convex. Since $q(t_i) > 0$ for $i = 1, \ldots, k$, we have

$$\max_{i=1,\ldots,k} |p(t_i)/q(t_i) - y_i| \leq \gamma$$

if and only if

$$-\gamma q(t_i) \leq p(t_i) - y_i q(t_i) \leq \gamma q(t_i), \quad i = 1, \ldots, k,$$

which is a pair of linear inequalities.

6.10 *Fitting data with a concave nonnegative nondecreasing quadratic function.* We are given the data

$$x_1, \ldots, x_N \in \mathbf{R}^n, \qquad y_1, \ldots, y_N \in \mathbf{R},$$

and wish to fit a quadratic function of the form

$$f(x) = (1/2)x^T Px + q^T x + r,$$

where $P \in \mathbf{S}^n$, $q \in \mathbf{R}^n$, and $r \in \mathbf{R}$ are the parameters in the model (and, therefore, the variables in the fitting problem).

Our model will be used only on the box $\mathcal{B} = \{x \in \mathbf{R}^n \mid l \preceq x \preceq u\}$. You can assume that $l \prec u$, and that the given data points x_i are in this box.

We will use the simple sum of squared errors objective,

$$\sum_{i=1}^{N} (f(x_i) - y_i)^2,$$

as the criterion for the fit. We also impose several constraints on the function f. First, it must be concave. Second, it must be nonnegative on $\mathcal{B}$, *i.e.*, $f(z) \geq 0$ for all $z \in \mathcal{B}$. Third, f must be nondecreasing on $\mathcal{B}$, *i.e.*, whenever $z,\ \tilde{z} \in \mathcal{B}$ satisfy $z \preceq \tilde{z}$, we have $f(z) \leq f(\tilde{z})$.

Show how to formulate this fitting problem as a convex problem. Simplify your formulation as much as you can.

Solution. The objective function is a convex quadratic function of the function parameters, which are the variables in the fitting problem, so we need only consider the constraints. The function f is concave if and only if $P \preceq 0$, which is a convex constraint, in fact, a linear matrix inequality. The nonnegativity constraint states that $f(z) \geq 0$ for each $z \in \mathcal{B}$. For each such z, the constraint is a linear inequality in the variables $P,\ q,\ r$, so the constraint is the intersection of an infinite number of linear inequalities (one for each $z \in \mathcal{B}$) and therefore convex. But we can derive a much simpler representation for this constraint. Since we will impose the condition that f is nondecreasing, it follows that the lowest value of f must be attained at the point l. Thus, f is nonnegative on $\mathcal{B}$ if and only if $f(l) \geq 0$, which is a single linear inequality.

Now let's look at the monotonicity constraint. We claim this is equivalent to $\nabla f(z) \succeq 0$ for $z \in \mathcal{B}$. Let's show that first. Suppose f is monotone on $\mathcal{B}$ and let $z \in \mathbf{int}\,\mathcal{B}$. Then for small positive $t \in \mathbf{R}$, we have $f(z + te_i) \geq f(z)$. Subtracting, and taking the limit as $t \to 0$ gives the conclusion $\nabla f(z)_i \geq 0$. To show the converse, suppose that $\nabla f(z) \succeq 0$ on $\mathcal{B}$, and let $z,\ \tilde{z} \in \mathcal{B}$, with $z \preceq \tilde{z}$. Define $g(t) = f(z + t(\tilde{z} - z))$. Then we have

$$\begin{aligned} f(\tilde{z}) - f(z) &= g(1) - g(0) \\ &= \int_0^1 g'(t)\, dt \\ &= \int_0^1 (\tilde{z} - z)^T \nabla f(z + t(\tilde{z} - z))\, dt \\ &\geq 0, \end{aligned}$$

since $\tilde{z} - z \succeq 0$ and $\nabla f \succeq 0$ on $\mathcal{B}$. (Note that this result doesn't depend on f being quadratic.)

For our function, monotonicity is equivalent to $\nabla f(z) = Pz + q \succeq 0$ for $z \in \mathcal{B}$. This too is convex, since for each z, it is a set of linear inequalities in the parameters of the function. We replace this abstract constraint with 2^n constraints, by insisting that $\nabla f(z) = Pz + q \succeq 0$ must hold at the 2^n vertices of $\mathcal{B}$ (obtained by setting each component equal to l_i or u_i). But there is a *far better* description of the monotonicity constraint.

Let us express P as $P = P_+ - P_-$, where P_+ and P_- are the elementwise positive and negative parts of P, respectively:

$$(P_+)_{ij} = \max\{P_{ij}, 0\}, \qquad (P_-)_{ij} = \max\{-P_{ij}, 0\}.$$

Then

$$Pz + q \succeq 0 \quad \text{for all } l \preceq z \preceq u$$

holds if and only if

$$P_+ l - P_- u + q \succeq 0.$$

Note that in contrast to our set of 2^n linear inequalities, this representation involves $n(n+1)$ new variables, and n linear inequality constraints.

(Another method to get a compact representation of the monotonicity constraint is based on deriving the alternative inequality to the condition that $Pz + q \succeq 0$ for $l \preceq z \preceq u$; this results in an equivalent formulation.)

Finally, we can express the problem as

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^N \left((1/2)x_i^T P x_i + q^T x_i + r - y_i\right)^2 \\ \text{subject to} & P \preceq 0 \\ & (1/2)l^T P l + q^T l + r \geq 0 \\ & P = P_+ - P_-, \quad (P_+)_{ij} \geq 0, \quad (P_-)_{ij} \geq 0 \\ & P_+ l - P_- u + q \succeq 0, \end{array}$$

with variables $P,\ P_+, P_- \in \mathbf{S}^n$, $q \in \mathbf{R}$, and $r \in \mathbf{R}$. The objective is convex quadratic, there is one linear matrix inequality (LMI) constraint, and some linear equality and inequality constraints. This problem can be expressed as an SDP.

We should note one common pitfall. We argue that f is concave, so its gradient must be monotone nonincreasing. Therefore, the argument goes, its 'lowest' value in $\mathcal{B}$ is achieved at the upper corner u. Therefore, for $Pu+q \succeq 0$ is enough to ensure that the monotonicity condition holds. One variation on this argument holds that it is enough to impose the two inequalities $Pl + q \succeq 0$ and $Pu + q \succeq 0$.

This sounds very reasonable, and in fact is true for dimensions $n = 1$ and $n = 2$. But sadly, it is false in general. Here is a counterexample:

$$P = \begin{bmatrix} -1 & 1 & -1 \\ 1 & -10 & 0 \\ -1 & 0 & -10 \end{bmatrix}, \quad l = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \quad u = \begin{bmatrix} 1.1 \\ 1 \\ 1 \end{bmatrix}, \quad q = \begin{bmatrix} 2.1 \\ 20 \\ 20 \end{bmatrix}.$$

It is easily checked that $P \preceq 0$, $Pl + q \succeq 0$, and $Pu + q \succeq 0$. However, consider the point

$$z = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix},$$

which satisfies $l \preceq z \preceq u$. For this point we have

$$Pz + q = \begin{bmatrix} -0.9 \\ 31 \\ 9 \end{bmatrix} \not\succeq 0.$$

6.11 *Least-squares direction interpolation.* Suppose $F_1, \ldots, F_n : \mathbf{R}^k \to \mathbf{R}^p$, and we form the linear combination $F : \mathbf{R}^k \to \mathbf{R}^p$,

$$F(u) = x_1 F_1(u) + \cdots + x_n F_n(u),$$

where x is the variable in the interpolation problem.

In this problem we require that $\angle(F(v_j), q_j) = 0$, $j = 1, \ldots, m$, where q_j are given vectors in $\mathbf{R}^p$, which we assume satisfy $\|q_j\|_2 = 1$. In other words, we require the direction of F to take on specified values at the points v_j. To ensure that $F(v_j)$ is not zero (which makes the angle undefined), we impose the minimum length constraints $\|F(v_j)\|_2 \geq \epsilon$, $j = 1, \ldots, m$, where $\epsilon > 0$ is given.

Show how to find x that minimizes $\|x\|^2$, and satisfies the direction (and minimum length) conditions above, using convex optimization.

Solution. Introduce variables y_i, and constraints

$$F(v_j) = y_j q_j, \qquad y_j \geq \epsilon,$$

and minimize $\|x\|^2$. This is a QP.

6.12 *Interpolation with monotone functions.* A function $f : \mathbf{R}^k \to \mathbf{R}$ is monotone nondecreasing (with respect to $\mathbf{R}^k_+$) if $f(u) \geq f(v)$ whenever $u \succeq v$.

(a) Show that there exists a monotone nondecreasing function $f : \mathbf{R}^k \to \mathbf{R}$, that satisfies $f(u_i) = y_i$ for $i = 1, \ldots, m$, if and only if

$$y_i \geq y_j \text{ whenever } u_i \succeq u_j, \quad i,\ j = 1, \ldots, m.$$

(b) Show that there exists a convex monotone nondecreasing function $f : \mathbf{R}^k \to \mathbf{R}$, with $\mathbf{dom}\, f = \mathbf{R}^k$, that satisfies $f(u_i) = y_i$ for $i = 1, \ldots, m$, if and only if there exist $g_i \in \mathbf{R}^k$, $i = 1, \ldots, m$, such that

$$g_i \succeq 0, \quad i = 1, \ldots, m, \qquad y_j \geq y_i + g_i^T(u_j - u_i), \quad i,\ j = 1, \ldots, m.$$

Solution.

(a) The condition is obviously necessary. It is also sufficient. Define

$$f(x) = \max_{u_i \preceq x} y_i.$$

This function is monotone, because $v \preceq w$ *always* implies

$$f(v) = \max_{u_i \preceq v} y_i \leq \max_{u_i \preceq w} y_i = f(w).$$

f satisfies the interpolation conditions if

$$f(u_i) = \max_{u_j \preceq u_i} y_j = y_i,$$

which is true if $u_i \succeq u_j$ implies $y_i \geq y_j$.

If we want $\mathbf{dom}\, f = \mathbf{R}^k$, we can define f as

$$f(x) = \begin{cases} \min_i y_i & x \not\succeq u_i, \quad i = 1, \ldots, m \\ \max_{u_i \preceq x} y_i & \text{otherwise.} \end{cases}$$

(b) We first show it is necessary. Suppose f is convex, monotone nondecreasing, with $\mathbf{dom}\, f = \mathbf{R}^k$, and satisfies the interpolation conditions. Let g_i be a normal vector to a supporting hyperplane at u_i to f, *i.e.*,

$$f(x) \geq y_i + g_i^T(x - u_i),$$

for all x. In particular, at $x = u_j$, this inequality reduces to

$$y_j \geq y_i + g_i^T(x - u_i),$$

It also follows that $g_i \succeq 0$: If $g_{ik} < 0$, then choosing $x = u_i - e_k$ gives

$$f(x) \geq y_i + g_i^T(x - u_i) = y_i - g_{ij} > y_i,$$

so f is not monotone.

To show that the conditions are sufficient, consider

$$f(x) = \max_{i=1,\ldots,m} \left(y_i + g_i^T(x - u_i)\right).$$

f is convex, satisfies the interpolation conditions, and is monotone: if $v \preceq w$, then

$$y_i + g_i^T(v - u_i) \leq y_i + g_i^T(w - u_i)$$

for all i, and hence $f(v) \leq f(w)$.

6.13 *Interpolation with quasiconvex functions.* Show that there exists a quasiconvex function $f : \mathbf{R}^k \to \mathbf{R}$, that satisfies $f(u_i) = y_i$ for $i = 1, \ldots, m$, if and only if there exist $g_i \in \mathbf{R}^k$, $i = 1, \ldots, m$, such that

$$g_i^T(u_j - u_i) \leq -1 \text{ whenever } y_j < y_i, \quad i,\ j = 1, \ldots, m.$$

Solution. We first show that the condition is necessary. For each $i = 1, \ldots, m$, define $J_i = \{j = 1, \ldots, m \mid y_j < y_i\}$. Suppose the condition does not hold, *i.e.*, for some i, the set of inequalities

$$g_i^T(u_j - u_i) \leq -1, \quad j \in J_i$$

is infeasible. By a theorem of alternatives, there exists $\lambda \succeq 0$ such that

$$\sum_{j \in J_i} \lambda_j (u_j - u_i) = 0, \qquad \sum_{j \in J_i} \lambda_j = 1.$$

This means u_i is a convex combination of u_j, $j \in J_i$. On the other hand, $y_i > y_j$ for $j \in J_i$, so if $f(u_i) = y_i$ and $f(u_j) = y_j$, then f cannot be quasiconvex.

Next we prove the condition is sufficient. Suppose the condition holds. Define $f : \mathbf{R}^k \to \mathbf{R}$ as

$$f(x) = \max\left\{y_{\min}, \max\{y_j \mid g_j^T(x - u_j) \geq 0\}\right\}$$

where $y_{\min} = \min_i y_i$.

We first verify that f satisfies the interpolation conditions $f(u_i) = y_i$. It is immediate from the definition of f that $f(u_i) \geq y_i$. Also, $f(u_i) > y_i$ only if $g_j^T(u_i - u_j) \geq 0$ for some j with $y_j > y_i$. This contradicts the definition of g_j. Therefore $f(u_i) = y_i$.

Finally, we check that f is quasiconvex. The sublevel sets of f are convex because $f(x) \leq \alpha$ if and only if

$$g_j^T(x - u_j) \geq 0 \quad \Longrightarrow \quad y_j \leq \alpha$$

or equivalently, $g_j^T(x - u_j) < 0$ for all j with $y_j > \alpha$.

6.14 [Nes00] *Interpolation with positive-real functions.* Suppose $z_1, \ldots, z_n \in \mathbf{C}$ are n distinct points with $|z_i| > 1$. We define K_{np} as the set of vectors $y \in \mathbf{C}^n$ for which there exists a function $f : \mathbf{C} \to \mathbf{C}$ that satisfies the following conditions.

- f is *positive-real*, which means it is analytic outside the unit circle (*i.e.*, for $|z| > 1$), and its real part is nonnegative outside the unit circle ($\Re f(z) \geq 0$ for $|z| > 1$).
- f satisfies the *interpolation conditions*

$$f(z_1) = y_1, \qquad f(z_2) = y_2, \qquad \ldots, \qquad f(z_n) = y_n.$$

If we denote the set of positive-real functions as $\mathcal{F}$, then we can express K_{np} as

$$K_{\text{np}} = \{y \in \mathbf{C}^n \mid \exists f \in \mathcal{F},\ y_k = f(z_k),\ k = 1, \ldots, n\}.$$

(a) It can be shown that f is positive-real if and only if there exists a nondecreasing function ρ such that for all z with $|z| > 1$,

$$f(z) = i\Im f(\infty) + \int_0^{2\pi} \frac{e^{i\theta} + z^{-1}}{e^{i\theta} - z^{-1}}\, d\rho(\theta),$$

where $i = \sqrt{-1}$ (see [KN77, page 389]). Use this representation to show that $K_{\rm np}$ is a closed convex cone.

Solution. It follows that every element in $K_{\rm np}$ can be expressed as $i\alpha\mathbf{1} + v$ where $\alpha \in \mathbf{R}$ and v is in the conic hull of the vectors

$$v(\theta) = \left(\frac{e^{i\theta} + z_1^{-1}}{e^{i\theta} - z_1^{-1}}, \frac{e^{i\theta} + z_2^{-1}}{e^{i\theta} - z_2^{-1}}, \ldots, \frac{e^{i\theta} + z_n^{-1}}{e^{i\theta} - z_n^{-1}} \right), \quad 0 \leq \theta \leq 2\pi.$$

Therefore $K_{\rm np}$ is the sum of a convex cone and a line, so it is also a convex cone. Closedness is less obvious. The set

$$C = \{v(\theta) \mid 0 \leq \theta \leq 2\pi\}$$

is compact, because v is continuous on $[0, 2\pi]$. The convex hull of a compact set is compact, and the conic hull of a compact set is closed. Therefore $K_{\rm np}$ is the sum of two closed sets (the conic hull of C and the line $i\alpha\mathbf{R}$), hence it is closed.

(b) We will use the inner product $\Re(x^H y)$ between vectors $x, y \in \mathbf{C}^n$, where x^H denotes the complex conjugate transpose of x. Show that the dual cone of $K_{\rm np}$ is given by

$$K_{\rm np}^* = \left\{ x \in \mathbf{C}^n \;\middle|\; \Im(\mathbf{1}^T x) = 0,\ \Re\left(\sum_{l=1}^n x_l \frac{e^{-i\theta} + \bar{z}_l^{-1}}{e^{-i\theta} - \bar{z}_l^{-1}} \right) \geq 0 \ \forall \theta \in [0, 2\pi] \right\}.$$

Solution. $x \in K_{\rm np}^*$ if

$$\Re((i\alpha\mathbf{1} + v)^H x) = \alpha\Im(\mathbf{1}^T x) + \Re(v^H x) \geq 0$$

for all $\alpha \in \mathbf{R}$ and all v in the conic hull of the vectors $v(\theta)$. This condition is equivalent to $\Im(\mathbf{1}^T x) = 0$ and $\Re(v(\theta)^H x) \geq 0$ for all $\theta \in [0, 2\pi]$.

(c) Show that

$$K_{\rm np}^* = \left\{ x \in \mathbf{C}^n \;\middle|\; \exists Q \in \mathbf{H}_+^n,\ x_l = \sum_{k=1}^n \frac{Q_{kl}}{1 - z_k^{-1}\bar{z}_l^{-1}},\ l = 1, \ldots, n \right\}$$

where $\mathbf{H}_+^n$ denotes the set of positive semidefinite Hermitian matrices of size $n \times n$. Use the following result (known as *Riesz-Fejér theorem*; see [KN77, page 60]). A function of the form

$$\sum_{k=0}^n (y_k e^{-ik\theta} + \bar{y}_k e^{ik\theta})$$

is nonnegative for all θ if and only if there exist $a_0, \ldots, a_n \in \mathbf{C}$ such that

$$\sum_{k=0}^n (y_k e^{-ik\theta} + \bar{y}_k e^{ik\theta}) = \left| \sum_{k=0}^n a_k e^{ik\theta} \right|^2.$$

Solution. We first show that any x of the form

$$x_l = \sum_{k=1}^n \frac{Q_{kl}}{1 - z_k^{-1}\bar{z}_l^{-1}}, \quad l = 1, \ldots, n, \tag{6.14.A}$$

where $Q \in \mathbf{H}_+^n$, belongs to $K_{\rm np}^*$. Suppose x satisfies (6.14.A) for some $Q \in \mathbf{H}_+^n$. We have

$$\begin{aligned} 2i\Im(\mathbf{1}^T x) &= \mathbf{1}^T x - \mathbf{1}^T \bar{x} \\ &= \sum_{k=1}^n \sum_{l=1}^n \frac{Q_{kl}}{1 - z_k^{-1}\bar{z}_l^{-1}} - \sum_{k=1}^n \sum_{l=1}^n \frac{Q_{lk}}{1 - \bar{z}_k^{-1} z_l^{-1}} \\ &= \sum_{k=1}^n \sum_{l=1}^n \frac{Q_{kl}}{1 - z_k^{-1}\bar{z}_l^{-1}} - \sum_{l=1}^n \sum_{k=1}^n \frac{Q_{kl}}{1 - \bar{z}_l^{-1} z_k^{-1}} \\ &= 0. \end{aligned}$$

Also,

$$\begin{aligned} &\Re\left(\sum_{l=1}^n x_l \frac{e^{-i\theta} + \bar{z}_l^{-1}}{e^{-i\theta} - \bar{z}_l^{-1}}\right) \\ &= \Re\left(\sum_{k=1}^n \sum_{l=1}^n \frac{Q_{kl}}{1 - z_k^{-1}\bar{z}_l^{-1}} \frac{e^{-i\theta} + \bar{z}_l^{-1}}{e^{-i\theta} - \bar{z}_l^{-1}}\right) \\ &= \frac{1}{2}\sum_{k=1}^n \sum_{l=1}^n \left(\frac{Q_{kl}}{1 - z_k^{-1}\bar{z}_l^{-1}} \frac{e^{-i\theta} + \bar{z}_l^{-1}}{e^{-i\theta} - \bar{z}_l^{-1}} + \frac{Q_{lk}}{1 - \bar{z}_k^{-1} z_l^{-1}} \frac{e^{i\theta} + z_l^{-1}}{e^{i\theta} - z_l^{-1}}\right) \\ &= \frac{1}{2}\sum_{k=1}^n \sum_{l=1}^n \frac{Q_{kl}}{1 - z_k^{-1}\bar{z}_l^{-1}} \left(\frac{e^{-i\theta} + \bar{z}_l^{-1}}{e^{-i\theta} - \bar{z}_l^{-1}} + \frac{e^{i\theta} + z_k^{-1}}{e^{i\theta} - z_k^{-1}}\right) \\ &= \frac{1}{2}\sum_{k=1}^n \sum_{l=1}^n \frac{Q_{kl}}{1 - z_k^{-1}\bar{z}_l^{-1}} \frac{2(1 - z_k^{-1}\bar{z}_l^{-1})}{(e^{i\theta} - z_k^{-1})(e^{-i\theta} - \bar{z}_l^{-1})} \\ &= \sum_{k=1}^n \sum_{l=1}^n \frac{Q_{kl}}{(e^{i\theta} - z_k^{-1})(e^{-i\theta} - \bar{z}_l^{-1})} \\ &\geq 0. \end{aligned}$$

Therefore $x \in K_{\rm np}^*$.

Conversely, suppose $x \in K_{\rm np}^*$, *i.e.*, $\Im(\mathbf{1}^T x) = 0$ and the function

$$R(\theta) = \Re\left(\sum_{l=1}^n x_l \frac{e^{-i\theta} + \bar{z}_l^{-1}}{e^{-i\theta} - \bar{z}_l^{-1}}\right)$$

is nonnegative. We can write R as

$$\begin{aligned} R(\theta) &= \Re\left(\sum_{l=1}^n x_l \frac{e^{-i\theta} + \bar{z}_l^{-1}}{e^{-i\theta} - \bar{z}_l^{-1}}\right) \\ &= \frac{1}{2}\sum_{l=1}^n \left(x_l \frac{e^{-i\theta} + \bar{z}_l^{-1}}{e^{-i\theta} - \bar{z}_l^{-1}} + \bar{x}_l \frac{e^{i\theta} + z_l^{-1}}{e^{i\theta} - z_l^{-1}}\right) \\ &= \frac{\sum_{k=0}^{n-1}\left(y_k e^{-ik\theta} + \bar{y}_k e^{ik\theta}\right)}{\prod_{l=1}^n |e^{i\theta} - z_l^{-1}|^2} \end{aligned}$$

for some y. The last line follows by bringing all terms in the previous line on the same denominator. The absence of a term $k = n$ in the numerator on the last line

requires some explanation. The coefficient of the term $e^{in\theta}/\prod_{l=1}^n |e^{i\theta} - z_l^{-1}|^2$ is

$$\begin{aligned}
\bar{y}_n &= \frac{1}{2}\sum_{l=1}^n \left(x_l \bar{z}_l^{-1} \prod_{k\neq l}(-\bar{z}_k^{-1}) - \bar{x}_l \bar{z}_l^{-1} \prod_{k\neq l}(-\bar{z}_k^{-1}) \right) \\
&= \frac{(-1)^{n-1}}{2} \left(\prod_{k=1}^n \bar{z}_k^{-1} \right) \sum_{l=1}^n (x_l - \bar{x}_l) \\
&= 0
\end{aligned}$$

because $\Im(\mathbf{1}^T x) = 0$.

Applying the Riesz-Fejér theorem to the numerator in the last expression for R we get

$$R(\theta) = \left| \frac{\sum_{k=0}^{n-1} a_k e^{ik\theta}}{\prod_{l=1}^n (e^{i\theta} - z_l^{-1})} \right|^2$$

for some set of coefficients a_k, and hence

$$R(\theta) = \left| \sum_{l=1}^n \frac{b_l}{e^{i\theta} - z_l^{-1}} \right|^2.$$

for some $b \in \mathbf{C}^n$. Therefore

$$\begin{aligned}
R(\theta) &= \sum_{l=1}^n \sum_{k=1}^n \frac{b_k \bar{b}_l}{(e^{i\theta} - z_k^{-1})(e^{-i\theta} - \bar{z}_l^{-1})} \\
&= \frac{1}{2}\sum_{l=1}^n \sum_{k=1}^n \frac{b_k \bar{b}_l}{1 - z_k^{-1}\bar{z}_l^{-1}} \left(\frac{e^{-i\theta} + \bar{z}_l^{-1}}{e^{-i\theta} - \bar{z}_l^{-1}} + \frac{e^{i\theta} + z_k^{-1}}{e^{i\theta} - z_k^{-1}} \right) \\
&= \Re \left(\sum_{l=1}^n \sum_{k=1}^n \frac{b_k \bar{b}_l}{1 - z_k^{-1}\bar{z}_l^{-1}} \frac{e^{-i\theta} + \bar{z}_l^{-1}}{e^{-i\theta} - \bar{z}_l^{-1}} \right).
\end{aligned}$$

Since the functions $(e^{-i\theta} + \bar{z}_l^{-1})/(e^{-i\theta} - \bar{z}_l^{-1})$ are linearly independent, we conclude that

$$x_l = \sum_{k=1}^n \frac{b_k \bar{b}_l}{1 - z_k^{-1}\bar{z}_l^{-1}},$$

i.e., we can choose $Q = bb^H$.

(d) Show that $K_{\rm np} = \{y \in \mathbf{C}^n \mid P(y) \succeq 0\}$ where $P(y) \in \mathbf{H}^n$ is defined as

$$P(y)_{kl} = \frac{y_k + \overline{y}_l}{1 - z_k^{-1}\bar{z}_l^{-1}}, \quad l, k = 1, \ldots, n.$$

The matrix $P(y)$ is called the *Nevanlinna-Pick matrix* associated with the points z_k, y_k.

Hint. As we noted in part (a), $K_{\rm np}$ is a closed convex cone, so $K_{\rm np} = K_{\rm np}^{**}$.

Solution. From the result in (c), $x \in K_{\rm np}^{**}$ if and only if for all $Q \in \mathbf{H}_+^n$,

$$\begin{aligned}
\Re(x^H y) &= \frac{1}{2}(x^H y + y^H x) \\
&= \frac{1}{2}\sum_{l=1}^n \sum_{k=1}^n \left(y_l \frac{Q_{lk}}{1 - \bar{z}_k^{-1} z_l^{-1}} + \bar{y}_l \frac{Q_{kl}}{1 - z_k^{-1}\bar{z}_l^{-1}} \right)
\end{aligned}$$

$$\begin{aligned} &= \frac{1}{2}\sum_{l=1}^{n}\sum_{k=1}^{n} Q_{lk}\left(\frac{y_l+\bar{y}_k}{1-\bar{z}_k^{-1}z_l^{-1}}\right) \\ &= \mathbf{tr}(QP(y)) \\ &\geq 0. \end{aligned}$$

In other words, if and only if $P(y) \succeq 0$.

(e) As an application, pose the following problem as a convex optimization problem:

$$\begin{array}{ll} \text{minimize} & \sum_{k=1}^{n} |f(z_k) - w_k|^2 \\ \text{subject to} & f \in \mathcal{F}. \end{array}$$

The problem data are n points z_k with $|z_k| > 1$ and n complex numbers $w_1, \ldots, w_n$. We optimize over all positive-real functions f.

Solution. We can express this problem as

$$\begin{array}{ll} \text{minimize} & \sum_{k=1}^{n} |y_k - w_k|^2 \\ \text{subject to} & P(y) \succeq 0, \end{array}$$

where $P(y)$ is the Nevanlinna-Pick matrix, and the variable is the (complex) vector y. Since P is linear in y, the constraint is a (complex) LMI, which can be expressed as a real LMI in the real and imaginary parts of y, following exercise 4.42. The objective is (convex) quadratic.

Chapter 7

Statistical estimation

Exercises

Estimation

7.1 *Linear measurements with exponentially distributed noise.* Show how to solve the ML estimation problem (7.2) when the noise is exponentially distributed, with density

$$p(z) = \begin{cases} (1/a)e^{-z/a} & z \geq 0 \\ 0 & z < 0, \end{cases}$$

where $a > 0$.

Solution. Solve the LP

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T(y - Ax) \\ \text{subject to} & Ax \preceq y. \end{array}$$

7.2 *ML estimation and ℓ_∞-norm approximation.* We consider the linear measurement model $y = Ax + v$ of page 352, with a uniform noise distribution of the form

$$p(z) = \begin{cases} 1/(2\alpha) & |z| \leq \alpha \\ 0 & |z| > \alpha. \end{cases}$$

As mentioned in example 7.1, page 352, any x that satisfies $\|Ax - y\|_\infty \leq \alpha$ is a ML estimate.

Now assume that the parameter α is not known, and we wish to estimate α, along with the parameters x. Show that the ML estimates of x and α are found by solving the ℓ_∞-norm approximation problem

$$\text{minimize} \quad \|Ax - y\|_\infty,$$

where a_i^T are the rows of A.

Solution. The log-likelihood function is

$$l(x, \alpha) = \begin{cases} m \log(1/2\alpha) & \|Ax - y\|_\infty \leq \alpha \\ -\infty & \text{otherwise.} \end{cases}$$

Maximizing over α and y is equivalent to solving the ℓ_∞-norm problem.

7.3 *Probit model.* Suppose $y \in \{0, 1\}$ is random variable given by

$$y = \begin{cases} 1 & a^T u + b + v \leq 0 \\ 0 & a^T u + b + v > 0, \end{cases}$$

where the vector $u \in \mathbf{R}^n$ is a vector of explanatory variables (as in the logistic model described on page 354), and v is a zero mean unit variance Gaussian variable.

Formulate the ML estimation problem of estimating a and b, given data consisting of pairs (u_i, y_i), $i = 1, \ldots, N$, as a convex optimization problem.

Solution. We have

$$\mathbf{prob}(y = 1) = Q(a^T u + b), \qquad \mathbf{prob}(y = 0) = 1 - Q(a^T u + b) = P(-a^T u - b)$$

where

$$Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{t^2/2}\, dt.$$

The log-likelihood function is

$$l(a, b) = \sum_{y_i = 1} \log Q(a^T u_i + b) + \sum_{y_i = 0} \log Q(-a^T u_i - b),$$

which is a concave function of a and b.

7.4 *Estimation of covariance and mean of a multivariate normal distribution.* We consider the problem of estimating the covariance matrix R and the mean a of a Gaussian probability density function

$$p_{R,a}(y) = (2\pi)^{-n/2} \det(R)^{-1/2} \exp(-(y-a)^T R^{-1}(y-a)/2),$$

based on N independent samples $y_1, y_2, \ldots, y_N \in \mathbf{R}^n$.

(a) We first consider the estimation problem when there are no additional constraints on R and a. Let μ and Y be the sample mean and covariance, defined as

$$\mu = \frac{1}{N}\sum_{k=1}^{N} y_k, \qquad Y = \frac{1}{N}\sum_{k=1}^{N}(y_k-\mu)(y_k-\mu)^T.$$

Show that the log-likelihood function

$$l(R,a) = -(Nn/2)\log(2\pi) - (N/2)\log\det R - (1/2)\sum_{k=1}^{N}(y_k-a)^T R^{-1}(y_k-a)$$

can be expressed as

$$l(R,a) = \frac{N}{2}\left(-n\log(2\pi) - \log\det R - \mathbf{tr}(R^{-1}Y) - (a-\mu)^T R^{-1}(a-\mu)\right).$$

Use this expression to show that if $Y \succ 0$, the ML estimates of R and a are unique, and given by

$$a_{\mathrm{ml}} = \mu, \qquad R_{\mathrm{ml}} = Y.$$

(b) The log-likelihood function includes a convex term $(-\log\det R)$, so it is not obviously concave. Show that l is concave, jointly in R and a, in the region defined by

$$R \preceq 2Y.$$

This means we can use convex optimization to compute simultaneous ML estimates of R and a, subject to convex constraints, as long as the constraints include $R \preceq 2Y$, *i.e.*, the estimate R must not exceed twice the unconstrained ML estimate.

Solution.

(a) We show that $\sum_{k=1}^{N}(y_k-a)(y_k-a)^T = N(Y-(a-\mu)(a-\mu)^T)$:

$$\begin{aligned}\sum_{k=1}^{N}(y_k-a)(y_k-a)^T &= \sum_{k=1}^{N} y_k y_k^T - Na\mu^T - N\mu a^T + Naa^T \\ &= \sum_{k=1}^{N}(y_k-\mu)(y_k-\mu)^T + N\mu\mu^T - Na\mu^T - N\mu a^T + Naa^T \\ &= NY + N(a-\mu)(a-\mu^T).\end{aligned}$$

This proves that the two expressions for l are equivalent.

Now let's maximize l. It is not (in general) a concave function of R, so we have to be careful. We do know that at the global optimizer, the gradient vanishes (but not conversely). Setting the gradient with respect to R and μ to zero gives

$$-R^{-1} + R^{-1}(Y + (a-\mu)(a-\mu)^T)R^{-1} = 0, \qquad -2R^{-1}(a-\mu) = 0,$$

which has only one solution,

$$Y + (a-\mu)(a-\mu)^T = R, \qquad a = \mu.$$

It must be the global maximizer of l, since l is not unbounded above.

(b) We show that the function

$$f(R) = -\log\det R - \mathbf{tr}(R^{-1}Y)$$

is concave in R for $0 \prec R \preceq 2Y$. This will establish concavity of the log-likelihood function because the remaining term of l is concave in a and R.
The gradient and Hessian of f are given by

$$\begin{aligned}\nabla f(R) &= -R^{-1} + R^{-1}YR^{-1}\\ \nabla^2 f(R)[V] &= R^{-1}VR^{-1} - R^{-1}VR^{-1}YR^{-1} - R^{-1}YR^{-1}VR^{-1}\end{aligned}$$

where by $\nabla^2 f(R)[V]$ we mean

$$\nabla^2 f(R)[V] = \frac{d}{dt}\nabla f(R+tV)\Big|_{t=0}.$$

We show that

$$\mathbf{tr}(V\nabla^2 f(R)[V]) = \frac{d^2}{dt^2} f(R+tV)\Big|_{t=0} \leq 0$$

for all V. We have

$$\begin{aligned}\mathbf{tr}(V\nabla^2 f(R)[V]) &= \mathbf{tr}(VR^{-1}VR^{-1}) - 2\,\mathbf{tr}(VR^{-1}VR^{-1}YR^{-1})\\ &= \mathbf{tr}\left((R^{-1/2}VR^{-1/2})^2(I - 2R^{-1/2}YR^{-1/2})\right)\\ &\leq 0\end{aligned}$$

for all V if

$$2R^{-1/2}YR^{-1/2} \succeq I,$$

i.e., $R \preceq 2Y$.

7.5 *Markov chain estimation.* Consider a Markov chain with n states, and transition probability matrix $P \in \mathbf{R}^{n\times n}$ defined as

$$P_{ij} = \mathbf{prob}(y(t+1) = i \mid y(t) = j).$$

The transition probabilities must satisfy $P_{ij} \geq 0$ and $\sum_{i=1}^n P_{ij} = 1$, $j = 1, \ldots, n$. We consider the problem of estimating the transition probabilities, given an observed sample sequence $y(1) = k_1$, $y(2) = k_2$, $\ldots$, $y(N) = k_n$.

(a) Show that if there are no other prior constraints on P_{ij}, then the ML estimates are the empirical transition frequencies: $\hat{P}_{ij}$ is the ratio of the number of times the state transitioned from j into i, divided by the number of times it was j, in the observed sample.

(b) Suppose that an equilibrium distribution p of the Markov chain is known, *i.e.*, a vector $q \in \mathbf{R}^n_+$ satisfying $\mathbf{1}^T q = 1$ and $Pq = q$. Show that the problem of computing the ML estimate of P, given the observed sequence and knowledge of q, can be expressed as a convex optimization problem.

Solution.

(a) The probability of the sequence $y(2), \ldots, y(N)$, given that we start in $y(1)$ is

$$P_{k_2,k_1}P_{k_3,k_2}\cdots P_{k_n,k_{n-1}} = \prod_{i,k=1}^n P_{ij}^{c_{ij}}$$

where c_{ij} is the number of times the state transitioned from j to i. The ML estimation problem is therefore

$$\begin{array}{ll}\text{maximize} & \sum_{i,j=1}^n c_{ij}\log P_{ij}\\ \text{subject to} & \mathbf{1}^T P = \mathbf{1}^T.\end{array}$$

The problem is separable, and can be solved column by column. Let $p_j = (P_{1j}, \ldots, P_{nj})$ be column j of P. It is the solution of

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^n c_{ij} \log p_{ij} \\ \text{subject to} & \mathbf{1}^T p_j = 1. \end{array}$$

Using Lagrange multipliers we find that

$$P_{ij} = \frac{c_{ij}}{\sum_{i=1}^n c_{ij}}.$$

(b) The ML estimation problem is

$$\begin{array}{ll} \text{maximize} & \sum_{i,j=1}^n c_{ij} \log P_{ij} \\ \text{subject to} & \mathbf{1}^T P = \mathbf{1}^T \\ & Pq = q. \end{array}$$

7.6 *Estimation of mean and variance.* Consider a random variable $x \in \mathbf{R}$ with density p, which is normalized, *i.e.*, has zero mean and unit variance. Consider a random variable $y = (x+b)/a$ obtained by an affine transformation of x, where $a > 0$. The random variable y has mean b and variance $1/a^2$. As a and b vary over $\mathbf{R}_+$ and $\mathbf{R}$, respectively, we generate a family of densities obtained from p by scaling and shifting, uniquely parametrized by mean and variance.

Show that if p is log-concave, then finding the ML estimate of a and b, given samples $y_1, \ldots, y_n$ of y, is a convex problem.

As an example, work out an analytical solution for the ML estimates of a and b, assuming p is a normalized Laplacian density, $p(x) = e^{-2|x|}$.

Solution. The density of y is given by

$$p_y(u) = ap(au - b).$$

The log-likelihood function is given by

$$\log p_y(u) = \log a + \log p(au - b).$$

If p is log-concave, then this log-likelihood function is a concave function of a and b. This allows us to compute ML estimates of the mean and variance of a random variable with a normalized density that is log-concave.

Suppose that n samples $y_1, \ldots, y_n$ are drawn from the distribution of y, which has a log-concave normalized density. To find the ML estimate of the parameters a and b, we maximize the concave function

$$\sum_{i=1}^n p_y(y_i) = n \log a + \sum_{i=1}^n \log p(ay_i - b).$$

For the Laplace distribution, you get

$$\sum_{i=1}^n p_y(y_i) = n \log a - 2 \sum_{i=1}^n |ay_i - b|,$$

so the ML estimates solve

$$\text{minimize} \quad -n \log a + 2 \textstyle\sum_{i=1}^n |ay_i - b|.$$

We can define $c = b/a$, and solve

$$\text{minimize} \quad -n \log a + 2a \textstyle\sum_{i=1}^n |y_i - c|.$$

The solution c is the median of y_i. a can be found by setting the derivative equal to zero:

$$a = \frac{n}{2\sum_{i=1}^{n} |y_i - c|}.$$

7.7 *ML estimation of Poisson distributions.* Suppose x_i, $i = 1, \ldots, n$, are independent random variables with Poisson distributions

$$\mathbf{prob}(x_i = k) = \frac{e^{-\mu_i}\mu_i^k}{k!},$$

with unknown means μ_i. The variables x_i represent the number of times that one of n possible independent events occurs during a certain period. In emission tomography, for example, they might represent the number of photons emitted by n sources.

We consider an experiment designed to determine the means μ_i. The experiment involves m detectors. If event i occurs, it is detected by detector j with probability p_{ji}. We assume the probabilities p_{ji} are given (with $p_{ji} \geq 0$, $\sum_{j=1}^{m} p_{ji} \leq 1$). The total number of events recorded by detector j is denoted y_j,

$$y_j = \sum_{i=1}^{n} y_{ji}, \quad j = 1, \ldots, m.$$

Formulate the ML estimation problem of estimating the means μ_i, based on observed values of y_j, $j = 1, \ldots, m$, as a convex optimization problem.

Hint. The variables y_{ji} have Poisson distributions with means $p_{ji}\mu_i$, *i.e.*,

$$\mathbf{prob}(y_{ji} = k) = \frac{e^{-p_{ji}\mu_i}(p_{ji}\mu_i)^k}{k!}.$$

The sum of n independent Poisson variables with means $\lambda_1, \ldots, \lambda_n$ has a Poisson distribution with mean $\lambda_1 + \cdots + \lambda_n$.

Solution. It follows from the two hints that y_j has a Poisson distribution with mean

$$\sum_{i=1}^{n} p_{ji}\mu_i = p_j^T \mu.$$

Therefore,

$$\log(\mathbf{prob}(y_j = k)) = -p_j^T \mu + k \log(p_j^T \mu) - \log k!.$$

Suppose the observed values of y_j are k_j, $j = 1, \ldots, n$. Then the ML estimation problem is

$$\begin{array}{ll} \text{maximize} & -\sum_{j=1}^{m} p_j^T \mu + \sum_{j=1}^{m} k_j \log(p_j^T \mu) \\ \text{subject to} & \mu \succeq 0, \end{array}$$

which is convex in μ.

For completeness we also prove the two hints. Suppose x is a Poisson random variable with mean μ (number of times that an event occurs). It is well known that the Poisson distribution is the limit of a binomial distribution

$$\mathbf{prob}(x = k) = \frac{e^{-\mu}\mu^k}{k!} = \lim_{n \to \infty,\ nq \to \mu} \left(\begin{array}{c} n \\ k \end{array} \right) q^k (1-q)^{n-k},$$

i.e., we can think of x is the total number of positives in n Bernoulli trials with $q = \mu/n$.

Now suppose y is the total number of positives that is detected, where the probability of detection is p. In the binomial formula, we simply replace q with pq, and in the limit

$$\begin{aligned}\mathbf{prob}(y=k) &= \lim_{n\to\infty,\ nq\to\mu} \left(\begin{array}{c} n \\ k \end{array}\right) (pq)^k (1-(pq))^{n-k} \\ &= \lim_{n\to\infty,\ nq\to p\mu} \left(\begin{array}{c} n \\ k \end{array}\right) q^k (1-q)^{n-k} \\ &= \frac{e^{-p\mu}(p\mu)^k}{k!}.\end{aligned}$$

Assume x and y are independent Poisson variables with means μ and λ. Then

$$\begin{aligned}\mathbf{prob}(x+y=k) &= \sum_{i=0}^{k} \mathbf{prob}(x=i)\,\mathbf{prob}(y=k-i) \\ &= e^{-\mu-\lambda} \sum_{i=0}^{k} \frac{\mu^i \lambda^{k-i}}{i!(k-i)!} \\ &= \frac{e^{-\mu-\lambda}}{k!} \sum_{i=0}^{k} \frac{k!}{i!(k-i)!} \mu^i \lambda^{k-i} \\ &= \frac{e^{-\mu-\lambda}}{k!} (\lambda+\mu)^k.\end{aligned}$$

7.8 *Estimation using sign measurements.* We consider the measurement setup

$$y_i = \mathbf{sign}(a_i^T x + b_i + v_i), \quad i = 1, \ldots, m,$$

where $x \in \mathbf{R}^n$ is the vector to be estimated, and $y_i \in \{-1, 1\}$ are the measurements. The vectors $a_i \in \mathbf{R}^n$ and scalars $b_i \in \mathbf{R}$ are known, and v_i are IID noises with a log-concave probability density. (You can assume that $a_i^T x + b_i + v_i = 0$ does not occur.) Show that maximum likelihood estimation of x is a convex optimization problem.

Solution. We re-order the observations so that $y_i = 1$ for $i = 1, \ldots, k$ and $y_i = 0$ for $i = k+1, \ldots, m$. The probability of this event is

$$\begin{array}{l}\prod_{i=1}^{k} \mathbf{prob}(a_i^T x + b_i + v_i > 0) \cdot \prod_{i=k+1}^{m} \mathbf{prob}(a_i^T x + b_i + v_i < 0) \\ = \prod_{i=1}^{k} F(-a_i^T x - b_i) \cdot \prod_{i=k+1}^{m} (1 - F(-a_i^T x - b_i)),\end{array}$$

where F is the cumulative distribution of the noise density. The integral of a log-concave function is log-concave, so F is log-concave, and so is $1 - F$. The log-likelihood function is

$$l(x) = \sum_{i=1}^{k} \log F(-a_i^T x - b_i) + \sum_{i=k+1}^{m} \log(1 - F(-a_i^T x - b_i)),$$

which is concave. Therefore, maximizing it is a convex problem.

7.9 *Estimation with unknown sensor nonlinearity.* We consider the measurement setup

$$y_i = f(a_i^T x + b_i + v_i), \quad i = 1, \ldots, m,$$

where $x \in \mathbf{R}^n$ is the vector to be estimated, $y_i \in \mathbf{R}$ are the measurements, $a_i \in \mathbf{R}^n$, $b_i \in \mathbf{R}$ are known, and v_i are IID noises with log-concave probability density. The function $f : \mathbf{R} \to \mathbf{R}$, which represents a measurement nonlinearity, is *not* known. However, it is known that $f'(t) \in [l, u]$ for all t, where $0 < l < u$ are given.

Explain how to use convex optimization to find a maximum likelihood estimate of x, as well as the function f. (This is an infinite-dimensional ML estimation problem, but you can be informal in your approach and explanation.)

Solution. For fixed function f and vector x, we observe $y_1, \ldots, y_m$ if and only if

$$f^{-1}(y_i) - a_i^T x - b_i = v_i, \quad i = 1, \ldots, m.$$

(Note that the assumption $0 < l < u$ implies f is invertible.) It follows that the probability of observing $y_1, \ldots, y_m$ is

$$\prod_{i=1}^m p_v \left(f^{-1}(y_i) - a_i^T x - b_i \right).$$

The log of this expression, regarded as a function of x and the function f, is the log-likelihood function:

$$l(x, f) = \sum_{i=1}^m \log p_v \left(z_i - a_i^T x - b_i \right),$$

where $z_i = f^{-1}(y_i)$. This is a concave function of z and x.

The function f only affects the log-likelihood function through the numbers z_i. The constraints can be expressed in terms of the inverse as

$$(d/dt) f^{-1}(t) \in [1/u, 1/l],$$

so we conclude that

$$(1/u)|y_i - y_j| \leq |z_i - z_j| \leq (1/l)|y_i - y_j|,$$

for all i, j. Conversely, if these inequalities hold, then there is a function f that satisfies the inequality, with $f^{-1}(y_i) = z_i$. (Actually, this is true only in the limit, but we're being informal here.)

Therefore, to find the ML estimate, we maximize the concave function of x and z above, subject to the linear inequalities on z.

7.10 *Nonparametric distributions on* $\mathbf{R}^k$. We consider a random variable $x \in \mathbf{R}^k$ with values in a finite set $\{\alpha_1, \ldots, \alpha_n\}$, and with distribution

$$p_i = \mathbf{prob}(x = \alpha_i), \quad i = 1, \ldots, n.$$

Show that a lower bound on the covariance of X,

$$S \preceq \mathbf{E}(X - \mathbf{E}\, X)(X - \mathbf{E}\, X)^T,$$

is a convex constraint in p.

Solution.

$$\mathbf{E}(X - \mathbf{E}\, X)(X - \mathbf{E}\, X)^T = \sum_{i=1}^n p_i \alpha_i \alpha_i^T - \left(\sum_{i=1}^n p_i \alpha_i \right) \left(\sum_{i=1}^n p_i \alpha_i \right)^T \succeq S$$

if and only if

$$\begin{bmatrix} \sum_{i=1}^n p_i \alpha_i \alpha_i^T - S & \sum_{i=1}^n p_i \alpha_i \\ (\sum_{i=1}^n p_i \alpha_i)^T & 1 \end{bmatrix} \succeq 0.$$

Optimal detector design

7.11 *Randomized detectors.* Show that every randomized detector can be expressed as a convex combination of a set of deterministic detectors: If

$$T = \begin{bmatrix} t_1 & t_2 & \cdots & t_n \end{bmatrix} \in \mathbf{R}^{m\times n}$$

satisfies $t_k \succeq 0$ and $\mathbf{1}^T t_k = 1$, then T can be expressed as

$$T = \theta_1 T_1 + \cdots + \theta_N T_N,$$

where T_i is a zero-one matrix with exactly one element equal to one per column, and $\theta_i \geq 0$, $\sum_{i=1}^N \theta_i = 1$. What is the maximum number of deterministic detectors N we may need?

We can interpret this convex decomposition as follows. The randomized detector can be realized as a bank of N deterministic detectors. When we observe $X = k$, the estimator chooses a random index from the set $\{1, \ldots, N\}$, with probability $\mathbf{prob}(j = i) = \theta_i$, and then uses deterministic detector T_j.

Solution. The detector T can be expressed as a convex combination of deterministic detectors as follows:

$$T = \sum_{i_1=1}^m \sum_{i_2=1}^m \cdots \sum_{i_n=1}^m \theta_{i_1,i_2,\ldots,i_m} \begin{bmatrix} e_{i_1} & e_{i_2} & \cdots & e_{i_n} \end{bmatrix}.$$

where

$$\theta_{i_1,i_2,\ldots,i_m} = t_{i_1,1} t_{i_2,2} \cdots t_{i_n,n}.$$

To see this, note that

$$\begin{aligned}
&\sum_{i_1=1}^m \sum_{i_2=1}^m \cdots \sum_{i_n=1}^m \theta_{i_1,i_2,\ldots,i_m} \begin{bmatrix} e_{i_1} & e_{i_2} & \cdots & e_{i_n} \end{bmatrix} \\
&= \sum_{i_n=1}^m \cdots \sum_{i_2=1}^m (t_{i_n,n} \cdots t_{i_2,2}) \left(\sum_{i_1=1}^m t_{i_1,1} \begin{bmatrix} e_{i_1} & e_{i_2} & \cdots & e_{i_n} \end{bmatrix} \right) \\
&= \sum_{i_n=1}^m \cdots \sum_{i_2=1}^m (t_{i_n,n} \cdots t_{i_2,2}) \begin{bmatrix} t_1 & e_{i_2} & \cdots & e_{i_n} \end{bmatrix} \\
&= \sum_{i_n=1}^m \cdots \sum_{i_3=1}^m (t_{i_n,n} \cdots t_{i_3,3}) \left(\sum_{i_2=1}^m t_{i_2,2} \begin{bmatrix} t_1 & e_{i_2} & \cdots & e_{i_n} \end{bmatrix} \right) \\
&= \sum_{i_n=1}^m \cdots \sum_{i_3=1}^m (t_{i_n,n} \cdots t_{i_3,3}) \sum_{i_2=1}^m \begin{bmatrix} t_1 & t_2 & \cdots & e_{i_n} \end{bmatrix} \\
&\vdots \\
&= \sum_{i_n=1}^m t_{i_n,n} \begin{bmatrix} t_1 & t_2 & \cdots & t_{n-1} & e_{i_n} \end{bmatrix} \\
&= \begin{bmatrix} t_1 & t_2 & \cdots & t_{n-1} & t_n \end{bmatrix}.
\end{aligned}$$

It is also clear that

$$\sum_{i_1,i_2,\ldots,i_m} \theta_{i_1,i_2,\ldots,i_m} = 1.$$

The following general argument (familiar from linear programming) shows that every detector can be expressed as a convex combination of no more than $n(m-1)+1$ deterministic detectors.

Suppose $v_1, \ldots, v_N$ are affinely dependent points in $\mathbf{R}^p$, which means that

$$\mathbf{rank} \begin{bmatrix} v_1 & v_2 & \cdots & v_N \\ 1 & 1 & \cdots & 1 \end{bmatrix} < N,$$

and suppose x is a strict convex combination of the points v_k:

$$x = \theta_1 v_1 + \cdots + \theta_N v_N, \qquad 1 = \theta_1 + \cdots + \theta_N, \qquad \theta \succ 0,$$

Then x is a convex combination of a subset of the points v_i. To see this note that the rank condition implies that there exists a $\lambda \neq 0$ such that

$$\sum_{i=1}^N \lambda_i v_i = 0, \qquad \sum_{i=1}^N \lambda_i = 0.$$

Therefore,

$$x = (\theta_1 + t\lambda_1)v_1 + \cdots + (\theta_N + t\lambda_N)v_N, \qquad 1 = (\theta_1 + t\lambda_1)v_1 + \cdots + (\theta_N + t\lambda_N)v_N,$$

for all t. Since λ has at least one negative component and $\theta \succ 0$, the number

$$t_{\max} = \sup\{t \mid \theta + t\lambda \succeq 0\}$$

is finite and positive. Define $\hat{\theta} = \theta + t_{\max}\lambda$. We have

$$x = \hat{\theta}_1 v_1 + \cdots + \hat{\theta}_N v_N, \qquad 1 = \hat{\theta}_1 + \cdots + \hat{\theta}_N, \qquad \hat{\theta} \succeq 0,$$

and at least one of the coefficients of θ is zero. We have expressed x as strict convex combination of a subset of the vectors v_i. Repeating this argument, we can express x as a strict convex combination of an affinely independent subset of $\{v_1, \ldots, v_N\}$.

Applied to the detector problem, this means that every randomized detector can be expressed as a convex combination of affinely independent deterministic detectors. Since the affine hull of the set of all detectors has dimension $n(m-1)$, it is impossible to find more than $n(m-1)+1$ affinely independent deterministic detectors.

7.12 *Optimal action.* In detector design, we are given a matrix $P \in \mathbf{R}^{n\times m}$ (whose columns are probability distributions), and then design a matrix $T \in \mathbf{R}^{m\times n}$ (whose columns are probability distributions), so that $D = TP$ has large diagonal elements (and small off-diagonal elements). In this problem we study the dual problem: Given P, find a matrix $S \in \mathbf{R}^{m\times n}$ (whose columns are probability distributions), so that $\tilde{D} = PS \in \mathbf{R}^{n\times n}$ has large diagonal elements (and small off-diagonal elements). To make the problem specific, we take the objective to be maximizing the minimum element of $\tilde{D}$ on the diagonal.

We can interpret this problem as follows. There are n *outcomes*, which depend (stochastically) on which of m inputs or *actions* we take: P_{ij} is the probability that outcome i occurs, given action j. Our goal is find a (randomized) strategy that, to the extent possible, causes any specified outcome to occur. The strategy is given by the matrix S: S_{ji} is the probability that we take action j, when we want outcome i to occur. The matrix $\tilde{D}$ gives the action error probability matrix: $\tilde{D}_{ij}$ is the probability that outcome i occurs, when we want outcome j to occur. In particular, $\tilde{D}_{ii}$ is the probability that outcome i occurs, when we want it to occur.

Show that this problem has a simple analytical solution. Show that (unlike the corresponding detector problem) there is always an optimal solution that is deterministic.

Hint. Show that the problem is separable in the columns of S.

Solution. Let $\tilde{p}_k^T$ be kth row of P. The problem is then

$$\begin{array}{ll} \text{maximize} & \min_k \tilde{p}_k^T s_k \\ \text{subject to} & s_k \succeq 0, \quad k = 1, \ldots, m \\ & \mathbf{1}^T s_k = 1, \quad k = 1, \ldots, m. \end{array}$$

This problem is separable (when put in epigraph form): we can just as well choose each s_k to maximize $\tilde{p}_k^T s_k$ subject to $s_k \succeq 0$, $\mathbf{1}^T s_k = 1$. But this is easy: we choose an index l of $\tilde{p}_k$ which has maximum entry, and take $s_k = e_l$.

In other words, the optimal strategy is very simple: when the outcome i is desired, simply choose (deterministically) an input that maximizes the probability of the outcome k.

Chebyshev and Chernoff bounds

7.13 *Chebyshev-type inequalities on a finite set.* Assume X is a random variable taking values in the set $\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$, and let S be a subset of $\{\alpha_1, \ldots, \alpha_m\}$. The distribution of X is unknown, but we are given the expected values of n functions f_i:

$$\mathbf{E}\, f_i(X) = b_i, \quad i = 1, \ldots, n. \tag{7.32}$$

Show that the optimal value of the LP

$$\begin{array}{ll} \text{minimize} & x_0 + \sum_{i=1}^n b_i x_i \\ \text{subject to} & x_0 + \sum_{i=1}^n f_i(\alpha) x_i \geq 1, \quad \alpha \in S \\ & x_0 + \sum_{i=1}^n f_i(\alpha) x_i \geq 0, \quad \alpha \notin S, \end{array}$$

with variables $x_0, \ldots, x_n$, is an upper bound on $\mathbf{prob}(X \in S)$, valid for all distributions that satisfy (7.32). Show that there always exists a distribution that achieves the upper bound.

Solution. The best upper bound on $\mathbf{prob}(x \in S)$ is the optimal value of

$$\begin{array}{ll} \text{maximize} & \sum_{\alpha \in S} p_k \alpha \\ \text{subject to} & \sum_{k=1}^m p_k = 1 \\ & \sum_{k=1}^m p_k f_i(\alpha_k) = b_i, \quad i = 1, \ldots, n \\ & p \succeq 0. \end{array}$$

The dual problem is

$$\begin{array}{ll} \text{minimize} & x_0 + \sum_{i=1}^n x_i b_i \\ \text{subject to} & x_0 + \sum_{i=1}^n x_i f_i(\alpha) \geq 1, \quad \alpha \in S \\ & x_0 + \sum_{i=1}^n x_i f_i(\alpha) \geq 0, \quad \alpha \notin S, \end{array}$$

The dual problem is feasible, so strong duality holds. Furthermore, the dual problem is bounded below, so the optimal value is finite, and hence there is a primal optimal solution.

Chapter 8

Geometric problems

Exercises

Projection on a set

8.1 *Uniqueness of projection.* Show that if $C \subseteq \mathbf{R}^n$ is nonempty, closed and convex, and the norm $\|\cdot\|$ is strictly convex, then for every x_0 there is exactly one $x \in C$ closest to x_0. In other words the projection of x_0 on C is unique.

Solution. There is at least one projection (this is true for any norm): Suppose $\hat{x} \in C$, then the projection is found by minimizing the continuous function $\|x - x_0\|$ over a closed bounded set $C \cap \{x \mid \|x - x_0\| \leq \|\hat{x} - x_0\|\}$, so the minimum is attained.

To show that it is unique if the norm is strictly convex, suppose $u, v \in C$ with $u \neq v$ and $\|u - x_0\| = \|v - x_0\| = D$. Then $(1/2)(u + v) \in C$ and

$$\begin{aligned} \|(1/2)(u+v) - x_0\| &= \|(1/2)(u - x_0) + (1/2)(v - x_0)\| \\ &< (1/2)\|u - x_0\| + (1/2)\|v - x_0\| \\ &= D, \end{aligned}$$

so u and v are not the projection of x_0 on C.

8.2 [Web94, Val64] *Chebyshev characterization of convexity.* A set $C \in \mathbf{R}^n$ is called a *Chebyshev set* if for every $x_0 \in \mathbf{R}^n$, there is a unique point in C closest (in Euclidean norm) to x_0. From the result in exercise 8.1, every nonempty, closed, convex set is a Chebyshev set. In this problem we show the converse, which is known as *Motzkin's theorem*.

Let $C \in \mathbf{R}^n$ be a Chebyshev set.

(a) Show that C is nonempty and closed.

(b) Show that P_C, the Euclidean projection on C, is continuous.

(c) Suppose $x_0 \notin C$. Show that $P_C(x) = P_C(x_0)$ for all $x = \theta x_0 + (1 - \theta)P_C(x_0)$ with $0 \leq \theta \leq 1$.

(d) Suppose $x_0 \notin C$. Show that $P_C(x) = P_C(x_0)$ for all $x = \theta x_0 + (1 - \theta)P_C(x_0)$ with $\theta \geq 1$.

(e) Combining parts (c) and (d), we can conclude that all points on the ray with base $P_C(x_0)$ and direction $x_0 - P_C(x_0)$ have projection $P_C(x_0)$. Show that this implies that C is convex.

Solution.

(a) C is nonempty, because it contains the projection of an arbitrary point $x_0 \in \mathbf{R}^n$.

To show that C is closed, let x_k, $k = 1, 2, \ldots$ be a sequence of points in C with limit $\bar{x}$. We have

$$\|\bar{x} - P_C(\bar{x})\|_2 \leq \|\bar{x} - x_k\|_2$$

for all k (by definition of $P_C(\bar{x})$). Taking the limit of the righthand side for $k \to \infty$ gives $\|\bar{x} - P_C(\bar{x})\|_2 = 0$. Therefore $\bar{x} = P_C(\bar{x}) \in C$.

(b) Let x_k, $k = 1, 2, \ldots$, be a sequence of points converging to $\bar{x}$. We have

$$\|x_k - P_C(x_k)\|_2 \leq \|x_k - P_C(\bar{x})\|_2 \leq \|x_k - \bar{x}\|_2 + \|\bar{x} - P_C(\bar{x})\|_2.$$

Taking limits on both sides, we see that

$$\lim_{k\to\infty} \|x_k - P_C(x_k)\|_2 = \lim_{k\to\infty} \|\bar{x} - P_C(x_k)\|_2 \leq \|\bar{x} - P_C(\bar{x})\|_2.$$

Now $\bar{x}$ has a unique projection, and therefore $P_C(\bar{x})$ is the only element of C in the ball $\{x \mid \|x - \bar{x}\|_2 \leq \mathbf{dist}(\bar{x}, C)\}$. Moreover C is a closed set. Therefore

$$\lim_{k\to\infty} \|\bar{x} - P_C(x_k)\|_2 \leq \|\bar{x} - P_C(\bar{x})\|_2$$

is only possible if $P_C(x_k)$ converges to $P_C(\bar{x})$.

(c) Suppose $x = \theta x_0 + (1-\theta)P_C(x_0)$ with $0 \leq \theta < 1$. We have

$$\begin{aligned}\|x_0 - P_C(x)\|_2 &\leq \|x_0 - x\|_2 + \|x - P_C(x)\|_2 \\ &\leq \|x_0 - x\|_2 + \|x - P_C(x_0)\|_2 \\ &= \|(1-\theta)(x_0 - P_C(x_0))\|_2 + \|\theta(x_0 - P_C(x_0))\|_2 \\ &= \|x_0 - P_C(x_0)\|_2.\end{aligned}$$

(The first inequality is the triangle inequality. The second inequality follows from the definition of $P_C(x)$.) Since C is a Chebyshev set, $P_C(x) = P_C(x_0)$.

(d) We will use the following fact (which follows from Brouwer's fixed point theorem): If $g : \mathbf{R}^n \to \mathbf{R}^n$ is continuous and $g(x) \neq 0$ for $\|x\|_2 = 1$, then there exists an x with $\|x\|_2 = 1$ and $g(x)/\|g(x)\|_2 = x$.
Let $x = \theta x_0 + (1-\theta)P_C(x_0)$ with $\theta > 1$. To simplify the notation we assume that $x_0 = 0$ and

$$\|x - x_0\|_2 = (\theta - 1)\|P_C(x_0)\|_2 = 1.$$

The function $g(x) = -P_C(x)$ is continuous (see part (b)). $g(x) \neq 0$ for $x \neq 0$ because $x_0 = 0 \notin C$. Using the fixed point theorem, we conclude that there exists a y with $\|y\|_2 = 1$ such that

$$y = -\frac{P_C(y)}{\|P_C(y)\|_2}.$$

This means that $x_0 = 0$ lies on the line segment between $P_C(y)$ and y. Hence, from (c), $P_C(x_0) = P_C(y)$, and

$$y = -\frac{P_C(x_0)}{\|P_C(x_0)\|_2} = (1-\theta)P_C(x_0) = x.$$

We conclude that $P_C(x) = P_C(x_0)$.

(e) It is sufficient to show that C is midpoint convex. Suppose it is not, *i.e.*, there exist $x_1, x_2 \in C$ with $x_0 = (1/2)(x_1 + x_2) \notin C$. For simplicity we assume that $\|x_1 - x_2\|_2 = 2$, so $\|x_0 - x_2\|_2 = \|x_0 - x_1\|_2 = 1$.
Define $D = \|x_0 - P_C(x_0)\|_2$. We must have $0 < D < 1$. ($D > 0$ because $x_0 \notin C$ and C is closed; $D < 1$ because otherwise x_0 would have two projections, x_1 and x_2, contradicting the fact that C is a Chebyshev set.)
By the result in (c) and (d), all points $x(\theta) = P_C(x_0) + \theta(x_0 - P_C(x_0))$ are projected on $P_C(x_0)$, *i.e.*,

$$\mathbf{dist}(x(\theta), C) = \|P_C(x_0) + \theta(x_0 - P_C(x_0)) - P_C(x_0)\|_2 = \theta\|x_0 - P_C(x_0)\|_2 = \theta D.$$

Without loss of generality, assume that

$$(x_0 - P_C(x_0))^T(x_1 - x_0) \leq 0.$$

(Otherwise, switch the roles of x_1 and x_2). We have for $\theta \geq 1$,

$$\begin{aligned}\theta^2 D^2 &= \mathbf{dist}(x(\theta), C)^2 \\ &< \|x(\theta) - x_1\|_2^2 \\ &= \|x(\theta) - x_0\|_2^2 + \|x_0 - x_1\|_2^2 + 2(x(\theta) - x_0)^T(x_0 - x_1) \\ &= (\theta-1)^2 D^2 + 1 + 2(x(\theta) - x_0)^T(x_0 - x_1) \\ &= (\theta-1)^2 D^2 + 1 + 2(\theta-1)(x_0 - P_C(x_0))^T(x_0 - x_1) \\ &\leq (\theta-1)^2 D^2 + 1.\end{aligned}$$

(The first inequality follows from the fact that $P_C(x_0) \neq x_1$.) Therefore $0 < (1 - 2\theta)D^2 + 1$, which is false for $\theta \geq (1/2)(1 + 1/D^2)$.

8.3 *Euclidean projection on proper cones.*

(a) *Nonnegative orthant.* Show that Euclidean projection onto the nonnegative orthant is given by the expression on page 399.

Solution. The inner product of two nonnegative vectors is zero if and only the componentwise product is zero. We can therefore solve the equations

$$x_{0,i} = x_{+,i} - x_{-,i}, \qquad x_{+,i} \geq 0, \qquad x_{-,i} \geq 0, \qquad x_{+,i}x_{-,i} = 0,$$

for $i = 1, \ldots, n$. If $x_{0,i} > 0$ the solution is $x_{+,i} = x_{0,i}$, $x_{-,i} = 0$. If $x_{0,i} < 0$ the solution is $x_{+,i} = 0$, $x_{-,i} = -x_{0,i}$. If $x_{0,i} = 0$ the solution is $x_{+,i} = x_{-,i} = 0$.

(b) *Positive semidefinite cone.* Show that Euclidean projection onto the positive semidefinite cone is given by the expression on page 399.

Solution. Define $\tilde{X}_+ = V^T X_+ V$, $\tilde{X}_- = V^T X_- V$. These matrices must satisfy

$$\Lambda = \tilde{X}_+ - \tilde{X}_-, \qquad \tilde{X}_+ \succeq 0, \qquad \tilde{X}_- \succeq 0, \qquad \mathbf{tr}(\tilde{X}_+\tilde{X}_-) = 0.$$

The first condition implies that the off-diagonal elements are equal: $(\tilde{X}_+)_{ij} = (\tilde{X}_-)_{ij}$ if $i \neq j$. The third equation implies

$$\mathbf{tr}(\tilde{X}_+ X_-) = \sum_{i=1}^n (\tilde{X}_+)_{ii}(\tilde{X}_-)_{ii} + \sum_{i=1}^n \sum_{j \neq i} (\tilde{X}_+)_{ij}(\tilde{X}_-)_{ij} = 0$$

which is only possible if

$$(\tilde{X}_+)_{ij} = (\tilde{X}_-)_{ij} = 0, \quad i \neq j$$

and

$$(\tilde{X}_+)_{ii}(\tilde{X}_-)_{ii} = 0, \quad i = 1, \ldots, n.$$

In other words, $\tilde{X}_+$ and $\tilde{X}_-$ are diagonal, with a complementary zero-nonzero pattern on the diagonal, *i.e.*,

$$(\tilde{X}_+)_{ii} = \max\{\lambda_i, 0\}, \qquad (\tilde{X}_0)_{ii} = \max\{-\lambda_i, 0\}.$$

(c) *Second-order cone.* Show that the Euclidean projection of (x_0, t_0) on the second-order cone

$$K = \{(x, t) \in \mathbf{R}^{n+1} \mid \|x\|_2 \leq t\}$$

is given by

$$P_K(x_0, t_0) = \begin{cases} 0 & \|x_0\|_2 \leq -t_0 \\ (x_0, t_0) & \|x_0\|_2 \leq t_0 \\ (1/2)(1 + t_0/\|x_0\|_2)(x_0, \|x_0\|_2) & \|x_0\|_2 \geq |t_0|. \end{cases}$$

Solution. The second-order cone is self-dual, so the conditions are

$$x_0 = u - v, \quad t_0 = \mu - \tau, \qquad \|u\|_2 \leq \mu, \quad \|v\|_2 \leq \tau, \qquad u^T v + \mu\tau = 0.$$

It follows from the Cauchy-Schwarz inequality that the last three conditions are satisfied if one of the following three cases holds.

- $\mu = 0$, $u = 0$, $\|v\|_2 \leq \tau$. The first two conditions give $v = -x_0$, $t_0 = -\tau$. The fourth condition implies $t_0 \leq 0$, and $\| - x_0\|_2 \leq -t_0$.
  In this case (x_0, t_0) is in the negative second-order cone, and its projection is the origin.

- $\tau = 0$, $v = 0$, $\|u\|_2 \leq \mu$. The first two conditions give $u = x_0$, $\mu = t_0$. The third condition implies $\|x_0\|_2 \leq t_0$.
  In this case (x_0, t_0) is in the second-order cone, so it is its own projection.
- $\|u\|_2 = \mu > 0$, $\|v\|_2 = \tau > 0$, $\tau u = -\mu v$. We can express v as $v = -(\tau/\mu)u$. From $x_0 = u - v$,
  $$x_0 = (1 + \tau/\mu)u, \qquad \mu = \|u\|_2,$$
  and therefore $\mu + \tau = \|x_0\|_2$. Also, $t_0 = \mu - \tau$. Solving for μ and τ gives
  $$\mu = (1/2)(t_0 + \|x_0\|_2), \qquad \tau = (1/2)(-t_0 + \|x_0\|_2).$$
  τ is only positive if $t_0 < \|x_0\|_2$. We obtain
  $$u = \frac{t_0 + \|x_0\|_2}{2\|x_0\|_2} x_0, \quad \mu = \frac{\|x_0\|_2 + t_0}{2}, \qquad v = \frac{t_0 - \|x_0\|_2}{2\|x_0\|_2} x_0, \quad \tau = \frac{\|x_0\|_2 - t_0}{2}.$$

8.4 The Euclidean projection of a point on a convex set yields a simple separating hyperplane
$$(P_C(x_0) - x_0)^T \left(x - (1/2)(x_0 + P_C(x_0))\right) = 0.$$
Find a counterexample that shows that this construction does not work for general norms.
Solution. We use the ℓ_1-norm, with
$$C = \{x \in \mathbf{R}^2 \mid x_1 + x_2/2 \leq 1\}, \qquad x_0 = (1, 1).$$
The projection is $P_C(x_0) = (1/2, 1)$, so the hyperplane as above,
$$(P_C(x_0) - x_0)^T (x - (1/2)(x_0 + P_C(x_0))) = 0,$$
simplifies to $x_1 = 3/4$. This does not separate $(1, 1)$ from C.

8.5 [HUL93, volume 1, page 154] *Depth function and signed distance to boundary.* Let $C \subseteq \mathbf{R}^n$ be a nonempty convex set, and let $\mathbf{dist}(x, C)$ be the distance of x to C in some norm. We already know that $\mathbf{dist}(x, C)$ is a convex function of x.

(a) Show that the depth function,
$$\mathbf{depth}(x, C) = \mathbf{dist}(x, \mathbf{R}^n \setminus C),$$
is concave for $x \in C$.
Solution. We will show that the depth function can be expressed as
$$\mathbf{depth}(x, C) = \inf_{\|y\|_* = 1} (S_C(y) - y^T x),$$
where S_C is the support function of C. This proves that the depth function is concave because it is the infimum of a family of affine functions of x.
We first prove the following result. Suppose $a \neq 0$. The distance of a point x_0, in the norm $\|\cdot\|$, to the hyperplane defined by $a^T x = b$, is given by $|a^T x - b|/\|a\|_*$. We can show this by applying Lagrange duality for the problem
$$\begin{array}{ll} \text{minimize} & \|x - x_0\| \\ \text{subject to} & a^T x = b. \end{array}$$
The dual function is
$$\begin{aligned} g(\nu) &= \inf_x \left(\|x - x_0\| + \nu(a^T x - b)\right) \\ &= \inf_x \left(\|x - x_0\| + \nu a^T (x - x_0) + \nu(a^T x_0 - b)\right) \\ &= \begin{cases} \nu(a^T x_0 - b) & \|\nu a\|_* \leq 1 \\ -\infty & \text{otherwise} \end{cases} \end{aligned}$$

so we obtain the dual problem

$$\begin{array}{ll} \text{maximize} & \nu(a^T x_0 - b) \\ \text{subject to} & |\nu| \leq 1/\|a\|_*. \end{array}$$

If $a^T x_0 \geq b$, the solution is $\nu^\star = 1/\|a\|_*$. If $a^T x_0 \leq b$, the solution is $\nu^\star = -1/\|a\|_*$. In both cases the optimal value is $|a^T x_0 - b|/\|a\|_*$.

We now give a geometric interpretation and proof of the expression for the depth function. Let $\mathcal{H}$ be the set of all halfspaces defined by supporting hyperplanes of C, and containing C. We can describe any $H \in \mathcal{H}$ by a linear inequality $x^T y \leq S_C(y)$ where y is a nonzero vector in $\mathbf{dom}\, S_C(y)$.

Let $H \in \mathcal{H}$. The function $\mathbf{dist}(x, \mathbf{R}^n \setminus H)$ is affine for all $x \in C$:

$$\mathbf{dist}(x, \mathbf{R}^n \setminus H) = \frac{S_C(y) - x^T y}{\|y\|_*}.$$

The intersection of all H in $\mathcal{H}$ is equal to $\mathbf{cl}\, C$ and therefore

$$\begin{array}{rcl} \mathbf{depth}(x, C) & = & \inf\limits_{H \in \mathcal{H}} \mathbf{dist}(x, \mathbf{R}^n \setminus H) \\ & = & \inf\limits_{y \neq 0} (S_C(y) - x^T y)/\|y\|_* \\ & = & \inf\limits_{\|y\|_* = 1} (S_C(y) - x^T y). \end{array}$$

(b) The *signed distance* to the boundary of C is defined as

$$s(x) = \begin{cases} \mathbf{dist}(x, C) & x \notin C \\ -\mathbf{depth}(x, C) & x \in C. \end{cases}$$

Thus, $s(x)$ is positive outside C, zero on its boundary, and negative on its interior. Show that s is a convex function.

Solution. We will show that if we extend the expression in part (a) to points $x \notin C$, we obtain the signed distance:

$$s(x) = \sup_{\|y\|_* = 1} (y^T x - S_C(y)).$$

In part (a) we have shown that this is true for $x \in C$.

If $x \in \mathbf{bd}\, C$, then $y^T x \leq S_C(y)$ for all unit norm y, with equality if y is the normalized normal vector to a supporting hyperplane at x, so the expression for s holds.

If $x \notin \mathbf{cl}\, C$, then for all y with $\|y\|_* = 1$, $y^T x - S_C(y)$ is the distance of x to a hyperplane supporting C (as proved in part (a)), and therefore

$$y^T x - S_C(y) \leq \mathbf{dist}(x, C).$$

Equality holds if we take y equal to the optimal solution of

$$\begin{array}{ll} \text{maximize} & y^T x - S_C(y) \\ \text{subject to} & \|y\|_* \leq 1 \end{array}$$

with variable y. As we have seen in §8.1.3 the optimal value of this problem is equal to $\mathbf{dist}(x, C)$.

The geometric interpretation is as follows. As in part (a), we let $\mathcal{H}$ be the set of all halfspaces defined by supporting hyperplanes of C, and containing C. From part (a), we already know that for $H \in \mathcal{H}$

$$-\mathbf{depth}(x, C) = \max_{H \in \mathcal{H}} s(x, H),$$

where $s(x, \mathbf{R}^n \setminus H)$ is the signed distance from x to H. We now have to show that for x outside of C

$$\mathbf{dist}(x, C) = \sup_{H \in \mathcal{H}} s(x, H).$$

By construction, we know that for all $G \in \mathcal{H}$, we must have $\mathbf{dist}(x, C) \geq s(x, G)$. Now, let B be a ball of radius $\mathbf{dist}(x, C)$ centered at x. Because both B and C are convex with B closed, there is a separating hyperplane H such that $H \in \mathcal{H}$ and $s(x, H) = \mathbf{dist}(x, C)$, hence

$$\mathbf{dist}(x, C) \leq \sup_{H \in \mathcal{H}} s(x, H),$$

and the desired result.

Distance between sets

8.6 Let C, D be convex sets.

(a) Show that $\mathbf{dist}(C, x + D)$ is a convex function of x.

(b) Show that $\mathbf{dist}(tC, x + tD)$ is a convex function of (x, t) for $t > 0$.

Solution. To prove the first, we note that

$$\mathbf{dist}(C, x + D) = \inf_{u,v} \left(I_C(u) + I_C(x + v) + \|u - (x + v)\| \right).$$

The righthand side is convex in (u, v, x). Therefore $\mathbf{dist}(C, x + D)$ is convex by the minimization rule. To prove the second, we note that

$$\mathbf{dist}(tC, x + tD) = t\,\mathbf{dist}(C, x/t + D).$$

The righthand side is the perspective of the convex function from part (a).

8.7 *Separation of ellipsoids.* Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two ellipsoids defined as

$$\mathcal{E}_1 = \{x \mid (x - x_1)^T P_1^{-1} (x - x_1) \leq 1\}, \qquad \mathcal{E}_2 = \{x \mid (x - x_2)^T P_2^{-1} (x - x_2) \leq 1\},$$

where $P_1,\ P_2 \in \mathbf{S}^n_{++}$. Show that $\mathcal{E}_1 \cap \mathcal{E}_2 = \emptyset$ if and only if there exists an $a \in \mathbf{R}^n$ with

$$\|P_2^{1/2} a\|_2 + \|P_1^{1/2} a\|_2 < a^T (x_1 - x_2).$$

Solution. The two sets are closed and bounded, so the intersection is nonempty if and only if there is an $a \neq 0$ satisfying

$$\inf_{x \in \mathcal{E}_1} a^T x > \sup_{x \in \mathcal{E}_2} a^T x.$$

The infimum is giving by the optimal value of

$$\begin{array}{ll} \text{minimize} & a^T x \\ \text{subject to} & (x - x_1)^T P_1^{-1} (x - x_1) \leq 1. \end{array}$$

A change of variables $y = P_1^{-1/2}(x - x_1)$ yields

$$\begin{array}{ll} \text{minimize} & a^T x_1 + a^T P^{1/2} y \\ \text{subject to} & y^T y \leq 1, \end{array}$$

which has optimal value $a^T x_1 - \|P^{1/2}a\|_2$.
Similarly,

$$\sup_{x \in \mathcal{E}_2} a^T x = a^T x_2 + \|P^{1/2}a\|_2.$$

The condition therefore reduces to

$$a^T x_1 - \|P^{1/2}a\|_2 > a^T x_2 + \|P^{1/2}a\|_2.$$

We can also derive this result directly from duality, without using the separating hyperplane theorem. The distance between the two sets is the optimal value of the problem

$$\begin{array}{ll} \text{minimize} & \|x - y\|_2 \\ \text{subject to} & \|P_1^{-1/2}(x - x_1)\|_2 \leq 1 \\ & \|P_2^{-1/2}(y - x_2)\|_2 \leq 1, \end{array}$$

with variables x and y. The optimal value is positive if and only if the intersection of the ellipsoids is empty, and zero otherwise.
To derive a dual, we first reformulate the problem as

$$\begin{array}{ll} \text{minimize} & \|u\|_2 \\ \text{subject to} & \|v\|_2 \leq 1, \quad \|w\|_2 \leq 1 \\ & P_1^{1/2} v = x - x_1 \\ & P_2^{1/2} w = y - x_2 \\ & u = x - y, \end{array}$$

with new variables u, v, w. The Lagrangian is

$$\begin{array}{rl} & L(x, y, u, v, w, \lambda_1, \lambda_2, z_1, z_2, z) \\ = & \|u\|_2 + \lambda_1(\|v\|_2 - 1) + \lambda_2(\|w\|_2 - 1) + z_1^T(P_1^{1/2}v - x + x_1) \\ & + z_2^T(P_2^{1/2}w - y + x_2) + z^T(u - x + y) \\ = & -\lambda_1 - \lambda_2 + z_1^T x_1 + z_2^T x_2 - (z + z_1)^T x + (z - z_2)^T y \\ & + \|u\|_2 + z^T u + \lambda_1\|v\|_2 + z_1^T P_1^{1/2} v + \lambda_2\|w\|_2 + z_2^T P_2^{1/2} w. \end{array}$$

The minimum over x is unbounded below unless $z_1 = -z$. The minimum over y is unbounded below unless $z_2 = z$. Eliminating z_1 and z_2 we can therefore write the dual function as

$$\begin{array}{rl} g(\lambda_1, \lambda_2, z) = & -\lambda_1 - \lambda_2 + z^T(x_2 - x_1) + \inf_u(\|u\|_2 + z^T u) \\ & + \inf_v(\lambda_1\|v\|_2 - z^T P_1^{1/2} v) + + \inf_w(\lambda_2\|w\|_2 + z^T P_2^{1/2} w). \end{array}$$

We have

$$\inf_u(\|u\|_2 + z^T u) = \begin{cases} 0 & \|z\|_2 \leq 1 \\ -\infty & \text{otherwise.} \end{cases}$$

This follows from the Cauchy-Schwarz inequality: if $\|z\|_2 \leq 1$, then $z^T u \geq -\|z\|_2\|u\|2 \geq -\|u\|_2$, with equality if $u = 0$. If $\|z\|_2 > 1$, we can take $u = -tz$ with $t \to \infty$ to show that $\|u\|_2 + z^T u = t\|z\|_1(1 - \|z\|_2))$ is unbounded below.
We also have

$$\inf_v(\lambda_1\|v\|_2 - z^T P_1^{1/2} v) = \begin{cases} 0 & \|P_1^{1/2} z\|_2 \leq \lambda_1 \\ -\infty & \text{otherwise.} \end{cases}$$

This can be shown by distinguishing two cases: if $\lambda_1 = 0$ then the infimum is zero if $P_1^{1/2} z = 0$ and $-\infty$ otherwise. If $\lambda_1 < 0$ the minimum is $-\infty$. If $\lambda_1 > 0$, we have

$$\begin{aligned} \inf_v(\lambda_1\|v\|_2 - z^T P_1^{1/2} v) &= \lambda_1 \inf_v(\|v\|_2 - (1/\lambda_1) z^T P_1^{1/2} v) \\ &= \begin{cases} 0 & \|P_1^{1/2} z\|_2 \le \lambda_1 \\ -\infty & \text{otherwise.} \end{cases} \end{aligned}$$

Similarly,

$$\inf_w(\lambda_2\|w\|_2 + z^T P_2^{1/2} w) = \begin{cases} 0 & \|P_2^{1/2} z\|_2 \le \lambda_2 \\ -\infty & \text{otherwise.} \end{cases}$$

Putting this all together, we obtain the dual problem

$$\begin{array}{ll} \text{maximize} & -\lambda_1 - \lambda_2 + z^T(x_2 - x_1) \\ \text{subject to} & \|z\|_2 \le 1, \quad \|P_1^{1/2} z\|_2 \le \lambda_1, \quad \|P_2^{1/2} z\|_2 \le \lambda_2, \end{array}$$

which is equivalent to

$$\begin{array}{ll} \text{maximize} & -\|P_1^{1/2} z\|_2 - \|P_2^{1/2} z\|_2 + z^T(x_2 - x_1) \\ \text{subject to} & \|z\|_2 \le 1. \end{array}$$

The intersection of the ellipsoids is empty if and only if the optimal value is positive, *i.e.*, there exists a z with

$$-\|P_1^{1/2} z\|_2 - \|P_2^{1/2} z\|_2 + z^T(x_2 - x_1) > 0.$$

Setting $a = -z$ gives the desired inequality.

8.8 *Intersection and containment of polyhedra.* Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two polyhedra defined as

$$\mathcal{P}_1 = \{x \mid Ax \preceq b\}, \qquad \mathcal{P}_2 = \{x \mid Fx \preceq g\},$$

with $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$, $F \in \mathbf{R}^{p\times n}$, $g \in \mathbf{R}^p$. Formulate each of the following problems as an LP feasibility problem, or a set of LP feasibility problems.

(a) Find a point in the intersection $\mathcal{P}_1 \cap \mathcal{P}_2$.

(b) Determine whether $\mathcal{P}_1 \subseteq P_2$.

For each problem, derive a set of linear inequalities and equalities that forms a strong alternative, and give a geometric interpretation of the alternative.
Repeat the question for two polyhedra defined as

$$\mathcal{P}_1 = \mathbf{conv}\{v_1, \ldots, v_K\}, \qquad \mathcal{P}_2 = \mathbf{conv}\{w_1, \ldots, w_L\}.$$

Solution
Inequality description.

(a) Solve

$$Ax \preceq b, \qquad Fx \preceq g.$$

The alternative is

$$A^T u + F^T v = 0, \qquad u \succeq 0, \qquad v \succeq 0, \qquad b^T u + g^T v < 0.$$

Interpretation: if the sets do not intersect, then they can be separated by a hyperplane with normal vector $a = A^T u = -F^T v$. If $Ax \preceq b$ and $Fy \preceq g$,

$$a^T x = u^T A x \le u^T b < -v^T g \le -v^T F x \le a^T y.$$

(b) $\mathcal{P}_1 \subseteq \mathcal{P}_2$ if and only if

$$\sup_{Ax \preceq b} f_i^T x \leq g_i, \quad i = 1, \ldots, p.$$

We can solve p LPs, and compare the optimal values with g_i. Using LP duality we can write the same conditions as

$$\inf_{A^T z = f_i,\ z \succeq 0} b^T z \leq g_i, \quad i = 1, \ldots, p,$$

which is equivalent to p (decoupled) LP feasibility problems

$$A^T z_i = f_i, \qquad z_i \succeq 0, \qquad b^T z_i \leq g_i$$

with variables z_i. The alternative for this system is

$$Ax \preceq \lambda b, \qquad f_i^T x > \lambda g_i, \qquad \lambda \geq 0.$$

If $\lambda > 0$, this means that $(1/\lambda)x \in \mathcal{P}_1$, $(1/\lambda)x \notin \mathcal{P}_2$.
If $\lambda = 0$, it means that if $\bar{x} \in \mathcal{P}1$, then $\bar{x} + tx \notin \mathcal{P}_2$ for t sufficiently large.

Vertex description.

(a) $\mathcal{P}_1 \cap \mathcal{P}_2 = \emptyset$? Solve

$$\lambda \succeq 0, \qquad \mathbf{1}^T \lambda = 1, \qquad \mu \succeq 0, \qquad \mathbf{1}^T \mu = 1, \qquad V\lambda = W\mu,$$

where V has columns v_i and W has columns w_i.
From Farkas' lemma the alternative is

$$V^T z + t\mathbf{1} \succeq 0, \qquad t < 0, \qquad -W^T z + u\mathbf{1} \succeq 0, \qquad u < 0,$$

i.e., $V^T z \succ 0$, $W^T z \prec 0$. Therefore z defines a separating hyperplane.

(b) $\mathcal{P}_1 \subseteq P_2$? For $i = 1, \ldots, K$,

$$w_i = V\mu_i, \qquad \mu_i \succeq 0, \qquad \mathbf{1}^T \mu_i = 1.$$

The alternative (from Farkas lemma) is

$$V^T z_i + t_i \mathbf{1} \succeq 0, \qquad w_i^T z_i + t_i < 0,$$

i.e., $w_i^T z_i \mathbf{1} < V^T z_i$. Thus, z_i defines a hyperplane separating w_i from $\mathcal{P}_2$.

Euclidean distance and angle problems

8.9 *Closest Euclidean distance matrix to given data.* We are given data $\hat{d}_{ij}$, for $i, j = 1, \ldots, n$, which are corrupted measurements of the Euclidean distances between vectors in $\mathbf{R}^k$:

$$\hat{d}_{ij} = \|x_i - x_j\|_2 + v_{ij}, \quad i, j = 1, \ldots, n,$$

where v_{ij} is some noise or error. These data satisfy $\hat{d}_{ij} \geq 0$ and $\hat{d}_{ij} = \hat{d}_{ji}$, for all i, j. The dimension k is not specified.
Show how to solve the following problem using convex optimization. Find a dimension k and $x_1, \ldots, x_n \in \mathbf{R}^k$ so that $\sum_{i,j=1}^n (d_{ij} - \hat{d}_{ij})^2$ is minimized, where $d_{ij} = \|x_i - x_j\|_2$, $i, j = 1, \ldots, n$. In other words, given some data that are approximate Euclidean distances, you are to find the closest set of actual Euclidean distances, in the least-squares sense.
Solution. The condition that d_{ij} are actual Euclidean distances can be expressed in terms of the associated Euclidean distance matrix, $D_{ij} = d_{ij}^2$:

$$D_{ii} = 0, \quad i = 1, \ldots, n, \qquad D_{ij} \geq 0, \quad i, j = 1, \ldots, n$$

$$(I - (1/n)\mathbf{1}\mathbf{1}^T)D(I - (1/n)\mathbf{1}\mathbf{1}^T) \preceq 0,$$

which is a set of convex conditions on D.

The objective can be expressed in terms of D as

$$\begin{aligned}\sum_{i,j=1}^{n}(d_{ij} - \hat{d}_{ij})^2 &= \sum_{i,j=1}^{n}(D_{ij}^{1/2} - \hat{d}_{ij})^2 \\ &= \sum_{i,j=1}^{n}\left(D_{ij} - 2D_{ij}^{1/2}\hat{d}_{ij} + \hat{d}_{ij}^2\right),\end{aligned}$$

which is a convex function of D (since $D_{ij}^{1/2}\hat{d}_{ij}$ is concave). Thus we minimize this function, subject to the constraints above. We reconstruct x_i as described in the text, using Cholesky factorization.

8.10 *Minimax angle fitting.* Suppose that $y_1, \ldots, y_m \in \mathbf{R}^k$ are affine functions of a variable $x \in \mathbf{R}^n$:

$$y_i = A_i x + b_i, \quad i = 1, \ldots, m,$$

and $z_1, \ldots, z_m \in \mathbf{R}^k$ are given nonzero vectors. We want to choose the variable x, subject to some convex constraints, (*e.g.*, linear inequalities) to minimize the maximum angle between y_i and z_i,

$$\max\{\angle(y_1, z_1), \ldots, \angle(y_m, z_m)\}.$$

The angle between nonzero vectors is defined as usual:

$$\angle(u, v) = \cos^{-1}\left(\frac{u^T v}{\|u\|_2\|v\|_2}\right),$$

where we take $\cos^{-1}(a) \in [0, \pi]$. We are only interested in the case when the optimal objective value does not exceed $\pi/2$.

Formulate this problem as a convex or quasiconvex optimization problem. When the constraints on x are linear inequalities, what kind of problem (or problems) do you have to solve?

Solution. This is a quasiconvex optimization problem. To see this, we note that

$$\begin{aligned}\angle(u, v) = \cos^{-1}\left(\frac{u^T v}{\|u\|_2\|v\|_2}\right) \leq \theta \quad &\Longleftrightarrow \quad \frac{u^T v}{\|u\|_2\|v\|_2} \geq \cos(\theta) \\ &\Longleftrightarrow \quad \cos(\theta)\|u\|_2\|v\|_2 \leq u^T v,\end{aligned}$$

where in the first line we use the fact that $\cos^{-1}$ is monotone decreasing. Now suppose that v is fixed, and u is a variable. For $\theta \leq \pi/2$, the sublevel set of $\angle(u, v)$ (in u) is a convex set, in fact, a simple second-order cone constraint. Thus, $\angle(u, v)$ is a quasiconvex function of u, for fixed v, as long as $u^T v \geq 0$. It follows that the objective in the angle fitting problem,

$$\max\{\angle(y_1, z_1), \ldots, \angle(y_m, z_m)\},$$

is quasiconvex in x, provided it does not exceed $\pi/2$.

To formulate the angle fitting problem, we first check whether the optimal objective value does not exceed $\pi/2$. To do this we solve the inequality system

$$(A_i x + b_i)^T z_i \geq 0, \quad i = 1, \ldots, m,$$

together with inequalities on x, say, $Fx \preceq g$. This can be done via LP. If this set of inequalities is not feasible, then the optimal objective for the angle fitting problem exceeds $\pi/2$, and we quit. If it is feasible, we solve the SOC inequality system

$$Fx \preceq g, \qquad (A_i x + b_i)^T z_i \geq \cos(\theta)\|A_i x + b_i\|_2\|z_i\|_2, \quad i = 1, \ldots, m,$$

to check if the optimal objective is more or less than θ. We can then bisect on θ to find the smallest value for which this system is feasible. Thus, we need to solve a sequence of SOCPs to solve the minimax angle fitting problem.

8.11 *Smallest Euclidean cone containing given points.* In $\mathbf{R}^n$, we define a *Euclidean cone*, with center direction $c \neq 0$, and angular radius θ, with $0 \leq \theta \leq \pi/2$, as the set

$$\{x \in \mathbf{R}^n \mid \angle(c, x) \leq \theta\}.$$

(A Euclidean cone is a second-order cone, *i.e.*, it can be represented as the image of the second-order cone under a nonsingular linear mapping.)

Let $a_1, \ldots, a_m \in \mathbf{R}$. How would you find the Euclidean cone, of smallest angular radius, that contains $a_1, \ldots, a_m$? (In particular, you should explain how to solve the feasibility problem, *i.e.*, how to determine whether there is a Euclidean cone which contains the points.)

Solution. First of all, we can assume that each a_i is nonzero, since the points that are zero lie in all cones, and can be ignored. The points lie in some Euclidean cone if and only if they lie in some halfspace, which is the 'largest' Euclidean cone, with angular radius $\pi/2$. This can be checked by solving a set of linear inequalities:

$$a_i^T x \geq 0, \quad i = 1, \ldots, m.$$

Now, on to finding the smallest possible Euclidean cone. The points lie in a cone of angular radius θ if and only if there is a (nonzero) vector $x \in \mathbf{R}^n$ such that

$$\frac{a_i^T x}{\|a_i\|_2 \|x\|_2} \geq \cos\theta, \quad i = 1, \ldots, m.$$

Since $\theta \leq \pi/2$, this is the same as

$$a_i^T x \geq \|a_i\|_2 \|x\|_2 \cos\theta, \quad i = 1, \ldots, m,$$

which is a set of second-order cone constraints. Thus, we can find the smallest cone by bisecting θ, and solving a sequence of SOCP feasibility problems.

Extremal volume ellipsoids

8.12 Show that the maximum volume ellipsoid enclosed in a set is unique. Show that the Löwner-John ellipsoid of a set is unique.

Solution. Follows from strict convexity of $f(A) = \log\det A^{-1}$.

8.13 *Löwner-John ellipsoid of a simplex.* In this exercise we show that the Löwner-John ellipsoid of a simplex in $\mathbf{R}^n$ must be shrunk by a factor n to fit inside the simplex. Since the Löwner-John ellipsoid is affinely invariant, it is sufficient to show the result for one particular simplex.

Derive the Löwner-John ellipsoid $\mathcal{E}_{\rm lj}$ for the simplex $C = \mathbf{conv}\{0, e_1, \ldots, e_n\}$. Show that $\mathcal{E}_{\rm lj}$ must be shrunk by a factor $1/n$ to fit inside the simplex.

Solution. By symmetry, the center of the LJ ellipsoid must lie in the direction $\mathbf{1}$, and its intersection with any hyperplane orthogonal to $\mathbf{1}$ should be a ball. This means we can describe the ellipsoid by a quadratic inequality

$$(x - \alpha\mathbf{1})^T (I + \beta\mathbf{1}\mathbf{1}^T)(x - \alpha\mathbf{1}) \leq \gamma,$$

parameterized by three parameters α, β, γ.

The extreme points must be in the boundary of the ellipsoid. For $x = 0$, this gives the condition

$$\gamma = \alpha^2 n(1 + n\beta).$$

For $x = e_i$, we get the condition

$$\alpha = \frac{1+\beta}{2(1+n\beta)}.$$

The volume of the ellipsoid is proportional to

$$\gamma^n \det(I + \beta \mathbf{1}\mathbf{1}^T)^{-1} = \frac{\gamma^n}{1+\beta n},$$

and its logarithm is

$$\begin{aligned} n\log\gamma - \log(1+\beta n) &= n\log(\alpha^2 n(1+n\beta)) - \log(1+\beta n) \\ &= n\log\left(\frac{(1+\beta)^2}{4(1+\beta)}\right) - \log(1+\beta n) \\ &= n\log(n/4) + 2n\log(1+\beta) - (n+1)\log(1+n\beta). \end{aligned}$$

Setting the derivative equal to zero gives $\beta = 1$, and hence

$$\alpha = \frac{1}{n+1}, \qquad \beta = 1, \qquad \gamma = \frac{n}{1+n}.$$

We conclude that $\mathcal{E}_{\rm lj}$ is the solution set of the quadratic inequality

$$(x - \frac{1}{n+1}\mathbf{1})^T(I + \mathbf{1}\mathbf{1}^T)(x - \frac{1}{n+1}\mathbf{1}) \leq \frac{n}{1+n},$$

which simplifies to $x^T x + (1 - \mathbf{1}^T x)^2 \leq 1$. The shrunk ellipsoid is the solution set of the quadratic inequality

$$(x - \frac{1}{n+1}\mathbf{1})^T(I + \mathbf{1}\mathbf{1}^T)(x - \frac{1}{n+1}\mathbf{1}) \leq \frac{1}{n(1+n)},$$

which simplifies to

$$x^T x + (1 - \mathbf{1}^T x)^2 \leq \frac{1}{n}.$$

We verify that the shrunk ellipsoid lies in C by maximizing the linear functions $\mathbf{1}^T x$, $-x_i$, $i = 1, \ldots, n$ subject to the quadratic inequality. The solution of

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^T x \\ \text{subject to} & x^T x + (1 - \mathbf{1}^T x)^2 \leq 1/n \end{array}$$

is the point $(1/n)\mathbf{1}$. The solution of

$$\begin{array}{ll} \text{minimize} & x_i \\ \text{subject to} & x^T x + (1 - \mathbf{1}^T x)^2 \leq 1/n \end{array}$$

is the point $(1/n)(\mathbf{1} - e_i)$.

8.14 *Efficiency of ellipsoidal inner approximation.* Let C be a polyhedron in $\mathbf{R}^n$ described as $C = \{x \mid Ax \preceq b\}$, and suppose that $\{x \mid Ax \prec b\}$ is nonempty.

(a) Show that the maximum volume ellipsoid enclosed in C, expanded by a factor n about its center, is an ellipsoid that contains C.

(b) Show that if C is symmetric about the origin, *i.e.*, of the form $C = \{x \mid -\mathbf{1} \preceq Ax \preceq \mathbf{1}\}$, then expanding the maximum volume inscribed ellipsoid by a factor $\sqrt{n}$ gives an ellipsoid that contains C.

Solution.

(a) The ellipsoid $\mathcal{E} = \{Bu + d \mid \|u\|_2 \leq 1\}$ is the maximum volume inscribed ellipsoid, if B and d solve

$$\begin{array}{ll} \text{minimize} & \log\det B^{-1} \\ \text{subject to} & \|Ba_i\|_2 \leq b_i - a_i^T d, \quad i = 1, \ldots, m, \end{array}$$

or in generalized inequality notation

$$\begin{array}{ll} \text{minimize} & \log\det B^{-1} \\ \text{subject to} & (Ba_i, b_i - a_i^T d) \succeq_K 0, \quad i = 1, \ldots, m, \end{array}$$

where K is the second-order cone. The Lagrangian is

$$L(B, d, u, v) = \log\det B^{-1} - \sum_{i=1}^m u_i^T Ba_i - v^T(b - Ad).$$

Minimizing over B and d gives

$$B^{-1} = -\frac{1}{2}\sum_{i=1}^m (a_i u_i^T + u_i a_i^T), \qquad A^T v = 0.$$

The dual problem is

$$\begin{array}{ll} \text{maximize} & \log\det(-(1/2)\sum_{i=1}^m (a_i u_i^T + u_i a_i^T)) - b^T v + n \\ \text{subject to} & A^T v = 0 \\ & \|u_i\|_2 \leq v_i, \quad i = 1, \ldots, m. \end{array}$$

The optimality conditions are: primal and dual feasibility and

$$B^{-1} = -\frac{1}{2}\sum_{i=1}^m (a_i u_i^T + u_i a_i^T), \qquad u_i^T Ba_i + v_i(b_i - a_i^T d) = 0, \quad i = 1, \ldots, m.$$

To simplify the notation we will assume that $B = I$, $d = 0$, so the optimality conditions reduce to

$$\|a_i\|_2 \leq b_i, \quad i = 1, \ldots, m, \qquad A^T v = 0, \qquad \|u_i\|_2 \leq v_i, \quad i = 1, \ldots, m,$$

and

$$I = -\frac{1}{2}\sum_{i=1}^m (a_i u_i^T + u_i a_i^T), \qquad u_i^T a_i + v_i b_i = 0, \quad i = 1, \ldots, m. \tag{8.14.A}$$

From the Cauchy-Schwarz inequality the last inequality, combined with $\|a_i\|_2 \leq b_i$ and $\|u_i\|_2 \leq v_i$, implies that and $u_i = 0$, $v_i = 0$ if $\|a_i\|_2 < b_i$, and

$$u_i = -(\|u_i\|_2/b_i)a_i, \qquad v_i = \|u_i\|_2$$

if $\|a_i\|_2 = b_i$.
We need to show that $\|x\|_2 \leq n$ if $Ax \preceq b$. The optimality conditions (8.14.A) give

$$n = -\sum_{i=1}^m a_i^T u = b^T v$$

and

$$x^T x = -\sum_{i=1}^m (u_i^T x)(a_i^T x) = \sum_{i=1}^m \frac{\|u_i\|_2}{\|a_i\|_2}(a_i^T x)^2 \leq \sum_{i=1}^m \frac{\|u_i\|_2}{\|a_i\|_2} b_i^2.$$

Since $u_i = 0$, $v_i = 0$ if $\|a_i\|_2 < b_i$, the last sum further simplifies and we obtain

$$x^T x \leq \sum_{i=1}^{m} \|u_i\|_2 b_i = b^T v = n.$$

(b) Let $\mathcal{E} = \{x \mid x^T Q^{-1} x \leq 1\}$ be the maximum volume ellipsoid with center at the origin inscribed in C, where $Q \in \mathbf{S}^n_{++}$. We are asked to show that the ellipsoid

$$\sqrt{n}\mathcal{E} = \{x \mid x^T Q^{-1} x \leq n\}$$

contains C.

We first formulate this problem as a convex optimization problem. $x \in \mathcal{E}$ if $x = Q^{1/2}y$ for some y with $\|y\|_2 \leq 1$, so we have $\mathcal{E} \subseteq C$ if and only if for $i = 1, \ldots, p$,

$$\sup_{\|y\|_2 \leq 1} a_i^T Q^{1/2} y = \|Q^{1/2} a_i\|_2 \leq 1, \qquad \inf_{\|y\|_2 \leq 1} a_i^T Q^{1/2} y = -\|Q^{1/2} a_i\|_2 \geq -1,$$

or in other words $a_i^T Q a_i = \|Q^{1/2} a_i\|_2^2 \leq 1$. We find the maximum volume inscribed ellipsoid by solving

$$\begin{array}{ll} \text{minimize} & \log\det Q^{-1} \\ \text{subject to} & a_i^T Q a_i \leq 1, \quad i = 1, \ldots, p. \end{array} \tag{8.14.B}$$

The variable is the matrix $Q \in \mathbf{S}^n$.

The dual function is

$$g(\lambda) = \inf_{Q \succ 0} L(Q, \lambda) = \inf_{Q \succ 0} \left(\log\det Q^{-1} + \sum_{i=1}^{n} \lambda_i (a_i^T Q a_i - 1) \right).$$

Minimizing over Q gives

$$Q^{-1} = \sum_{i=1}^{p} \lambda_i a_i a_i^T,$$

and hence

$$g(\lambda) = \begin{cases} \log\det \left(\sum_{i=1}^{p} \lambda_i a_i a_i^T \right) - \sum_{i=1}^{p} \lambda_i + n & \sum_{i=1}^{p} (\lambda_i a_i a_i^T) \succ 0 \\ -\infty & \text{otherwise.} \end{cases}$$

The resulting dual problem is

$$\begin{array}{ll} \text{maximize} & \log\det \left(\sum_{i=1}^{p} \lambda_i a_i a_i^T \right) - \sum_{i=1}^{p} \lambda_i + n \\ \text{subject to} & \lambda \succeq 0. \end{array}$$

The KKT conditions are primal and dual feasibility ($Q \succ 0$, $a_i^T Q a_i \leq 1$, $\lambda \succeq 0$), plus

$$Q^{-1} = \sum_{i=1}^{p} \lambda_i a_i a_i^T, \qquad \lambda_i (1 - a_i^T Q a_i) = 0, \quad i = 1, \ldots, p. \tag{8.14.C}$$

The third condition (the complementary slackness condition) implies that $a_i^T Q a_i = 1$ if $\lambda_i > 0$. Note that Slater's condition for (8.14.B) holds ($a_i^T Q a_i < 1$ for $Q = \epsilon I$ and $\epsilon > 0$ small enough), so we have strong duality, and the KKT conditions are necessary and sufficient for optimality.

Now suppose Q and λ are primal and dual optimal. If we multiply (8.14.C) with Q on the left and take the trace, we have

$$n = \mathbf{tr}(QQ^{-1}) = \sum_{i=1}^{p} \lambda_i \, \mathbf{tr}(Qa_i a_i^T) = \sum_{i=1}^{p} \lambda_i a_i^T Q a_i = \sum_{i=1}^{p} \lambda_i.$$

The last inequality follows from the fact that $a_i^T Q a_i = 1$ when $\lambda_i \neq 0$. This proves $\mathbf{1}^T \lambda = n$. Finally, we note that (8.14.C) implies that if $x \in C$,

$$x^T Q^{-1} x = \sum_{i=1}^{p} \lambda_i (a_i^T x)^2 \leq \sum_{i=1}^{p} \lambda_i = n.$$

8.15 *Minimum volume ellipsoid covering union of ellipsoids.* Formulate the following problem as a convex optimization problem. Find the minimum volume ellipsoid $\mathcal{E} = \{x \mid (x - x_0)^T A^{-1} (x - x_0) \leq 1\}$ that contains K given ellipsoids

$$\mathcal{E}_i = \{x \mid x^T A_i x + 2b_i^T x + c_i \leq 0\}, \quad i = 1, \ldots, K.$$

Hint. See appendix B.

Solution. $\mathcal{E}$ contains $\mathcal{E}_i$ if

$$\sup_{x \in \mathcal{E}_i} (x - x_0)^T A^{-1} (x - x_0) \leq 1,$$

i.e.,

$$x^T A_i x + 2b_i^T x + c_i \leq 0 \quad \Longrightarrow \quad x^T A^{-1} x - 2x_0^T A^{-1} x + x_0^T A^{-1} x_0 - 1 \leq 0.$$

From the S-procedure in appendix B, this is true if and only if there exists a $\lambda_i \geq 0$ such that

$$\lambda_i \begin{bmatrix} A_i & b_i \\ b_i^T & c_i \end{bmatrix} \succeq \begin{bmatrix} A^{-1} & -A^{-1} x_0 \\ -(A^{-1} x_0)^T & x_0^T A^{-1} x_0 - 1 \end{bmatrix}.$$

In other words,

$$\begin{bmatrix} \lambda_i A_i & \lambda_i b_i \\ \lambda_i b_i^T & 1 + \lambda_i c_i \end{bmatrix} - \begin{bmatrix} I \\ -x_0^T \end{bmatrix} A^{-1} \begin{bmatrix} I & -x_0 \end{bmatrix} \succeq 0,$$

i.e., the LMI

$$\begin{bmatrix} A & I & -x_0^T \\ I & \lambda_i A_i & \lambda_i b_i \\ -x_0 & \lambda_i b_i^T & 1 + \lambda_i c_i \end{bmatrix} \succeq 0$$

holds. We therefore obtain the SDP formulation

$$\begin{array}{ll} \text{minimize} & \log\det A^{-1} \\ \text{subject to} & \begin{bmatrix} A & I & -x_0^T \\ I & \lambda_i A_i & \lambda_i b_i \\ -x_0 & \lambda_i b_i^T & 1 + \lambda_i c_i \end{bmatrix} \succeq 0, \quad i = 1, \ldots, K \\ & \lambda_i \geq 0, \quad i = 1, \ldots, K. \end{array}$$

The variables are $A \in \mathbf{S}^n$, $x_0 \in \mathbf{R}^n$, and λ_i, $i = 1, \ldots, K$.

8.16 *Maximum volume rectangle inside a polyhedron.* Formulate the following problem as a convex optimization problem. Find the rectangle

$$\mathcal{R} = \{x \in \mathbf{R}^n \mid l \preceq x \preceq u\}$$

of maximum volume, enclosed in a polyhedron $\mathcal{P} = \{x \mid Ax \preceq b\}$. The variables are $l, u \in \mathbf{R}^n$. Your formulation should not involve an exponential number of constraints.

Solution. A straightforward, but very inefficient, way to express the constraint $\mathcal{R} \subseteq \mathcal{P}$ is to use the set of $m2^n$ inequalities $Av^i \preceq b$, where v^i are the (2^n) corners of $\mathcal{R}$. (If the corners of a box lie inside a polyhedron, then the box does.) Fortunately it is possible to express the constraint in a far more efficient way. Define

$$a_{ij}^+ = \max\{a_{ij}, 0\}, \qquad a_{ij}^- = \max\{-a_{ij}, 0\}.$$

Then we have $\mathcal{R} \subseteq \mathcal{P}$ if and only if

$$\sum_{i=1}^{n} (a_{ij}^+ u_j - a_{ij}^- l_j) \leq b_i, \quad i = 1, \ldots, m,$$

The maximum volume rectangle is the solution of

$$\begin{array}{ll} \text{maximize} & \left(\prod_{i=1}^n (u_i - l_i)\right)^{1/n} \\ \text{subject to} & \sum_{i=1}^n (a_{ij}^+ u_j - a_{ij}^- l_j) \leq b_i, \quad i = 1, \ldots, m, \end{array}$$

with implicit constraint $u \succeq l$. Another formulation can be found by taking the log of the objective, which yields

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^n \log(u_i - l_i) \\ \text{subject to} & \sum_{i=1}^n (a_{ij}^+ u_j - a_{ij}^- l_j) \leq b_i, \quad i = 1, \ldots, m. \end{array}$$

Centering

8.17 *Affine invariance of analytic center.* Show that the analytic center of a set of inequalities is affine invariant. Show that it is invariant with respect to positive scaling of the inequalities.

Solution. If $x_{\rm ac}$ is the minimizer of $-\sum_{i=1}^m \log(-f_i(x))$ then $y_{\rm ac} = Tx_{\rm ac} + x_0$ is the minimizer of $-\sum_{i=1}^m \log(-f_i(Tx + x_0))$.

Positive scaling of the inequalities adds a constant to the logarithmic barrier function.

8.18 *Analytic center and redundant inequalities.* Two sets of linear inequalities that describe the same polyhedron can have different analytic centers. Show that by adding redundant inequalities, we can make *any* interior point x_0 of a polyhedron

$$\mathcal{P} = \{x \in \mathbf{R}^n \mid Ax \preceq b\}$$

the analytic center. More specifically, suppose $A \in \mathbf{R}^{m \times n}$ and $Ax_0 \prec b$. Show that there exist $c \in \mathbf{R}^n$, $\gamma \in \mathbf{R}$, and a positive integer q, such that $\mathcal{P}$ is the solution set of the $m + q$ inequalities

$$Ax \preceq b, \qquad c^T x \leq \gamma, \qquad c^T x \leq \gamma, \quad \ldots, \quad c^T x \leq \gamma \tag{8.36}$$

(where the inequality $c^T x \leq \gamma$ is added q times), and x_0 is the analytic center of (8.36).

Solution. The optimality conditions are

$$\sum_{i=1}^{m} \frac{1}{b_i - a_i^T x^\star} a_i + \frac{q}{\gamma - c^T x^\star} c = 0$$

so we have to choose

$$c = -\frac{\gamma - c^T x^\star}{q} A^T d$$

where $d_i = 1/(b_i - a_i^T x^\star)$. We can choose $c = -A^T d$, and for q any integer satisfying

$$q \geq \max\{c^T x | Ax \leq b\} - c^T x^\star,$$

and $\gamma = q + c^T x^\star$.

8.19 Let $x_{\rm ac}$ be the analytic center of a set of linear inequalities

$$a_i^T x \leq b_i, \quad i = 1, \ldots, m,$$

and define H as the Hessian of the logarithmic barrier function at $x_{\rm ac}$:

$$H = \sum_{i=1}^m \frac{1}{(b_i - a_i^T x_{\rm ac})^2} a_i a_i^T.$$

Show that the kth inequality is redundant (*i.e.*, it can be deleted without changing the feasible set) if

$$b_k - a_k^T x_{\rm ac} \geq m(a_k^T H^{-1} a_k)^{1/2}.$$

Solution. We have an enclosing ellipsoid defined by

$$(x - x_{\rm ac})^T H (x - x_{\rm ac}) \leq m(m-1).$$

The maximum of $a_k^T x$ over the enclosing ellipsoid is

$$a_k^T x_{\rm ac} + \sqrt{m(m-1)}\sqrt{a_k^T H^{-1} a_k}$$

so if

$$a_k^T x_{\rm ac} + \sqrt{m(m-1)}\sqrt{a_k^T H^{-1} a_k} \leq b_k,$$

the inequality is redundant.

8.20 *Ellipsoidal approximation from analytic center of linear matrix inequality.* Let C be the solution set of the LMI

$$x_1 A_1 + x_2 A_2 + \cdots + x_n A_n \preceq B,$$

where $A_i, B \in \mathbf{S}^m$, and let $x_{\rm ac}$ be its analytic center. Show that

$$\mathcal{E}_{\rm inner} \subseteq C \subseteq \mathcal{E}_{\rm outer},$$

where

$$\begin{aligned} \mathcal{E}_{\rm inner} &= \{x \mid (x - x_{\rm ac})^T H (x - x_{\rm ac}) \leq 1\}, \\ \mathcal{E}_{\rm outer} &= \{x \mid (x - x_{\rm ac})^T H (x - x_{\rm ac}) \leq m(m-1)\}, \end{aligned}$$

and H is the Hessian of the logarithmic barrier function

$$-\log\det(B - x_1 A_1 - x_2 A_2 - \cdots - x_n A_n)$$

evaluated at $x_{\rm ac}$.

Solution. Define $F(x) = B - \sum_i x_i A_i$. and $F_{\rm ac} = F(x_{\rm ac})$ The Hessian is given by

$$H_{ij} = \mathbf{tr}(F_{\rm ac}^{-1} A_i F_{\rm ac}^{-1} A_j),$$

so we have

$$\begin{aligned}(x-x_{\rm ac})^T H(x-x_{\rm ac}) &= \sum_{i,j}(x_i-x_{{\rm ac},i})(x_j-x_{{\rm ac},j})\,\mathbf{tr}(F_{\rm ac}^{-1}A_iF_{\rm ac}^{-1}A_j)\\ &= \mathbf{tr}\left(F_{\rm ac}^{-1}(F(x)-F_{\rm ac})F_{\rm ac}^{-1}(F(x)-F_{\rm ac})\right)\\ &= \mathbf{tr}\left(F_{\rm ac}^{-1/2}(F(x)-F_{\rm ac})F_{\rm ac}^{-1/2}\right)^2.\end{aligned}$$

We first consider the inner ellipsoid. Suppose $x\in\mathcal{E}_{\rm inner}$, *i.e.*,

$$\mathbf{tr}\left(F_{\rm ac}^{-1/2}(F(x)-F_{\rm ac})F_{\rm ac}^{-1/2}\right)^2 = \left\|F_{\rm ac}^{-1/2}F(x)F_{\rm ac}^{-1/2}-I\right\|_F^2\le 1.$$

This implies that

$$-1\le\lambda_i(F_{\rm ac}^{-1/2}F(x)F_{\rm ac}^{-1/2})-1\le 1,$$

i.e.,

$$0\le\lambda_i(F_{\rm ac}^{-1/2}F(x)F_{\rm ac}^{-1/2})\le 2$$

for $i=1,\ldots,m$. In particular, $F(x)\succeq 0$, *i.e.*, $x\in C$.
To prove that $C\subseteq\mathcal{E}_{\rm outer}$, we first note that the gradient of the logarithmic barrier function vanishes at $x_{\rm ac}$, and therefore,

$$\mathbf{tr}(F_{\rm ac}^{-1}A_i)=0,\quad i=1,\ldots,n,$$

and therefore

$$\mathbf{tr}\left(F_{\rm ac}^{-1}(F(x)-F_{\rm ac})\right)=0,\qquad \mathbf{tr}\left(F_{\rm ac}^{-1}F(x)\right)=m.$$

Now assume $x\in C$. Then

$$\begin{aligned}&(x-x_{\rm ac})^T H(x-{\rm ac})\\ &= \mathbf{tr}\left(F_{\rm ac}^{-1/2}(F(x)-F_{\rm ac})F_{\rm ac}^{-1/2}\right)^2\\ &= \mathbf{tr}\left(F_{\rm ac}^{-1}(F(x)-F_{\rm ac})F_{\rm ac}^{-1}(F(x)-F_{\rm ac})\right)\\ &= \mathbf{tr}\left(F_{\rm ac}^{-1}F(x)F_{\rm ac}^{-1}F(x)\right)-2\,\mathbf{tr}\left(F_{\rm ac}^{-1}F(x)\right)+\mathbf{tr}\left(F_{\rm ac}^{-1}F_{\rm ac}F_{\rm ac}^{-1}F_{\rm ac}\right)\\ &= \mathbf{tr}\left(F_{\rm ac}^{-1}F(x)F_{\rm ac}^{-1}F(x)\right)-2m+m\\ &= \mathbf{tr}\left(F_{\rm ac}^{-1/2}F(x)F_{\rm ac}^{-1/2}\right)^2-m\\ &\le \left(\mathbf{tr}(F_{\rm ac}^{-1/2}F(x)F_{\rm ac}^{-1/2})\right)^2-m\\ &= m^2-m.\end{aligned}$$

The inequality follows by applying the inequality $\sum_i\lambda_i^2\le(\sum_i\lambda_i)^2$ for $\lambda\succeq 0$ to the eigenvalues of $F_{\rm ac}^{-1/2}F(x)F_{\rm ac}^{-1/2}$.

8.21 [BYT99] *Maximum likelihood interpretation of analytic center.* We use the linear measurement model of page 352,

$$y=Ax+v,$$

where $A\in\mathbf{R}^{m\times n}$. We assume the noise components v_i are IID with support $[-1,1]$. The set of parameters x consistent with the measurements $y\in\mathbf{R}^m$ is the polyhedron defined by the linear inequalities

$$-\mathbf{1}+y\preceq Ax\preceq\mathbf{1}+y. \tag{8.37}$$

Suppose the probability density function of v_i has the form

$$p(v)=\begin{cases}\alpha_r(1-v^2)^r & -1\le v\le 1\\ 0 & \text{otherwise,}\end{cases}$$

where $r \geq 1$ and $\alpha_r > 0$. Show that the maximum likelihood estimate of x is the analytic center of (8.37).

Solution.

$$L = m \log \alpha_r + r \sum_{i=1}^{m} \left(\log(1 + y_i - a_i^T x) + \log(1 - y_i + a_i^T x) \right).$$

8.22 *Center of gravity.* The center of gravity of a set $C \subseteq \mathbf{R}^n$ with nonempty interior is defined as

$$x_{\rm cg} = \frac{\int_C u \, du}{\int_C 1 \, du}.$$

The center of gravity is affine invariant, and (clearly) a function of the set C, and not its particular description. Unlike the centers described in the chapter, however, it is very difficult to compute the center of gravity, except in simple cases (*e.g.*, ellipsoids, balls, simplexes).

Show that the center of gravity $x_{\rm cg}$ is the minimizer of the convex function

$$f(x) = \int_C \|u - x\|_2^2 \, du.$$

Solution. Setting the gradient equal to zero gives

$$\int_C 2(u - x) \, du = 0$$

i.e.,

$$\int_C u \, du = \left(\int_C 1 \, du \right) x.$$

Classification

8.23 *Robust linear discrimination.* Consider the robust linear discrimination problem given in (8.23).

(a) Show that the optimal value $t^\star$ is positive if and only if the two sets of points can be linearly separated. When the two sets of points can be linearly separated, show that the inequality $\|a\|_2 \leq 1$ is tight, *i.e.*, we have $\|a^\star\|_2 = 1$, for the optimal $a^\star$.

(b) Using the change of variables $\tilde{a} = a/t$, $\tilde{b} = b/t$, prove that the problem (8.23) is equivalent to the QP

$$\begin{array}{ll} \text{minimize} & \|\tilde{a}\|_2 \\ \text{subject to} & \tilde{a}^T x_i - \tilde{b} \geq 1, \quad i = 1, \ldots, N \\ & \tilde{a}^T y_i - \tilde{b} \leq -1, \quad i = 1, \ldots, M. \end{array}$$

Solution.

(a) If $t^\star > 0$, then

$$a^{\star T} x_i \geq t^\star + b^\star > b^\star > b^\star - t^\star \geq a^{\star T} y_i,$$

so $a^\star$, $b^\star$ define a separating hyperplane.

Conversely if a, b define a separating hyperplane, then there is a positive t satisfying the constraints.

The constraint is tight because the other constraints are homogeneous.

(b) Suppose a, b, t are feasible in problem (8.23), with $t > 0$. Then $\tilde{a}$, $\tilde{b}$ are feasible in the QP, with objective value $\|\tilde{a}\|_2 = \|a\|_2/t \leq 1/t$.

Conversely, if $\tilde{a}$, $\tilde{b}$ are feasible in the QP, then $t = 1/\|\tilde{a}\|_2$, $a = \tilde{a}/\|\tilde{a}\|_2$, $b = \tilde{b}/\|\tilde{a}\|_2$, are feasible in problem (8.23), with objective value $t = 1/\|\tilde{a}\|_2$.

8.24 *Linear discrimination maximally robust to weight errors.* Suppose we are given two sets of points $\{x_1, \ldots, x_N\}$ and and $\{y_1, \ldots, y_M\}$ in $\mathbf{R}^n$ that can be linearly separated. In §8.6.1 we showed how to find the affine function that discriminates the sets, and gives the largest gap in function values. We can also consider robustness with respect to changes in the vector a, which is sometimes called the *weight vector*. For a given a and b for which $f(x) = a^T x - b$ separates the two sets, we define the *weight error margin* as the norm of the smallest $u \in \mathbf{R}^n$ such that the affine function $(a+u)^T x - b$ no longer separates the two sets of points. In other words, the weight error margin is the maximum ρ such that

$$(a+u)^T x_i \geq b, \quad i = 1, \ldots, N, \qquad (a+u)^T y_j \leq b, \quad i = 1, \ldots, M,$$

holds for all u with $\|u\|_2 \leq \rho$.

Show how to find a and b that maximize the weight error margin, subject to the normalization constraint $\|a\|_2 \leq 1$.

Solution. The weight error margin is the maximum ρ such that

$$(a+u)^T x_i \geq b, \quad i = 1, \ldots, N, \qquad (a+u)^T y_j \leq b, \quad i = 1, \ldots, M,$$

for all u with $\|u\|_2 \leq \rho$, *i.e.*,

$$a^T x_i - \rho\|x_i\|_2 \geq b_i, \qquad a^T y_i + \rho\|y_i\|_2 \leq b_i.$$

This shows that the weight error margin is given by

$$\min_{\substack{i=1,\ldots,N \\ j=1,\ldots,M}} \left\{ \frac{a^T x_i - b}{\|x_i\|_2}, \frac{b - a^T y_i}{\|y_i\|_2} \right\}.$$

We can maximize the weight error margin by solving the problem

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & a^T x_i - b \geq t\|x_i\|_2, \quad i = 1, \ldots, N \\ & b - a^T y_i \geq t\|y_i\|_2, \quad j = 1, \ldots, M \\ & \|a\|_2 \leq 1 \end{array}$$

with variables a, b, t.

8.25 *Most spherical separating ellipsoid.* We are given two sets of vectors $x_1, \ldots, x_N \in \mathbf{R}^n$, and $y_1, \ldots, y_M \in \mathbf{R}^n$, and wish to find the ellipsoid with minimum eccentricity (*i.e.*, minimum condition number of the defining matrix) that contains the points $x_1, \ldots, x_N$, but not the points $y_1, \ldots, y_M$. Formulate this as a convex optimization problem.

Solution. This can be solved as the SDP

$$\begin{array}{ll} \text{minimize} & \gamma \\ \text{subject to} & x_i^T P x_i + q^T x_i + r \geq 0, \quad i = 1, \ldots, N \\ & y_i^T P y_i + q^T y_i + r \leq 0, \quad i = 1, \ldots, M \\ & I \preceq P \preceq \gamma I, \end{array}$$

with variables $P \in \mathbf{S}^n$, $q \in \mathbf{R}^n$, and $r, \gamma \in \mathbf{R}$.

Placement and floor planning

8.26 *Quadratic placement.* We consider a placement problem in $\mathbf{R}^2$, defined by an undirected graph $\mathcal{A}$ with N nodes, and with quadratic costs:

$$\text{minimize} \quad \sum_{(i,j)\in\mathcal{A}} \|x_i - x_j\|_2^2.$$

The variables are the positions $x_i \in \mathbf{R}^2$, $i = 1, \ldots, M$. The positions x_i, $i = M+1, \ldots, N$ are given. We define two vectors $u, v \in \mathbf{R}^M$ by

$$u = (x_{11}, x_{21}, \ldots, x_{M1}), \qquad v = (x_{12}, x_{22}, \ldots, x_{M2}),$$

containing the first and second components, respectively, of the free nodes.
Show that u and v can be found by solving two sets of linear equations,

$$Cu = d_1, \qquad Cv = d_2,$$

where $C \in \mathbf{S}^M$. Give a simple expression for the coefficients of C in terms of the graph $\mathcal{A}$.
Solution. The objective function is

$$\sum_{(i,j)\in\mathcal{A}} (u_i - u_j)^2 + \sum_{(i,j)\in\mathcal{A}} (v_j - v_j)^2.$$

Setting the gradients with respect to u and v equal to zero gives equations $Cu = d_1$ and $Cv = d_2$ with

$$C_{ij} = \begin{cases} \text{degree of node } i & i = j \\ -(\text{number of arcs between } i \text{ and } j) & i \neq j, \end{cases}$$

and

$$d_{1i} = \sum_{j>M,\ (i,j)\in\mathcal{A}} x_{j1}, \qquad d_{2i} = \sum_{j>M,\ (i,j)\in\mathcal{A}} x_{j2}.$$

8.27 *Problems with minimum distance constraints.* We consider a problem with variables $x_1, \ldots, x_N \in \mathbf{R}^k$. The objective, $f_0(x_1, \ldots, x_N)$, is convex, and the constraints

$$f_i(x_1, \ldots, x_N) \leq 0, \quad i = 1, \ldots, m,$$

are convex (*i.e.*, the functions $f_i : \mathbf{R}^{Nk} \to \mathbf{R}$ are convex). In addition, we have the *minimum distance constraints*

$$\|x_i - x_j\|_2 \geq D_{\min}, \quad i \neq j, \quad i, j = 1, \ldots, N.$$

In general, this is a hard nonconvex problem.
Following the approach taken in floorplanning, we can form a *convex restriction* of the problem, *i.e.*, a problem which is convex, but has a smaller feasible set. (Solving the restricted problem is therefore easy, and any solution is guaranteed to be feasible for the nonconvex problem.) Let $a_{ij} \in \mathbf{R}^k$, for $i < j$, $i, j = 1, \ldots, N$, satisfy $\|a_{ij}\|_2 = 1$.
Show that the restricted problem

$$\begin{array}{ll} \text{minimize} & f_0(x_1, \ldots, x_N) \\ \text{subject to} & f_i(x_1, \ldots, x_N) \leq 0, \quad i = 1, \ldots, m \\ & a_{ij}^T(x_i - x_j) \geq D_{\min}, \quad i < j, \ i, j = 1, \ldots, N, \end{array}$$

is convex, and that every feasible point satisfies the minimum distance constraint.
Remark. There are many good heuristics for choosing the directions a_{ij}. One simple one starts with an approximate solution $\hat{x}_1, \ldots, \hat{x}_N$ (that need not satisfy the minimum distance constraints). We then set $a_{ij} = (\hat{x}_i - \hat{x}_j)/\|\hat{x}_i - \hat{x}_j\|_2$.
Solution. Follows immediately from the Cauchy-Schwarz inequality:

$$1 \leq a^T(u - v) \leq \|a\|_2\|u - v\|_2 = \|u - v\|_2.$$

Miscellaneous problems

8.28 Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two polyhedra described as

$$\mathcal{P}_1 = \{x \mid Ax \preceq b\}, \qquad \mathcal{P}_2 = \{x \mid -\mathbf{1} \preceq Cx \preceq \mathbf{1}\},$$

where $A \in \mathbf{R}^{m\times n}$, $C \in \mathbf{R}^{p\times n}$, and $b \in \mathbf{R}^m$. The polyhedron $\mathcal{P}_2$ is symmetric about the origin. For $t \geq 0$ and $x_c \in \mathbf{R}^n$, we use the notation $t\mathcal{P}_2 + x_c$ to denote the polyhedron

$$t\mathcal{P}_2 + x_c = \{tx + x_c \mid x \in \mathcal{P}_2\},$$

which is obtained by first scaling $\mathcal{P}_2$ by a factor t about the origin, and then translating its center to x_c.
Show how to solve the following two problems, via an LP, or a set of LPs.

(a) Find the largest polyhedron $t\mathcal{P}_2 + x_c$ enclosed in $\mathcal{P}_1$, *i.e.*,

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & t\mathcal{P}_2 + x_c \subseteq \mathcal{P}_1 \\ & t \geq 0. \end{array}$$

(b) Find the smallest polyhedron $t\mathcal{P}_2 + x_c$ containing $\mathcal{P}_1$, *i.e.*,

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \mathcal{P}_1 \subseteq t\mathcal{P}_2 + x_c \\ & t \geq 0. \end{array}$$

In both problems the variables are $t \in \mathbf{R}$ and $x_c \in \mathbf{R}^n$.
Solution.

(a) We can write the problem as

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & \sup_{x\in t\mathcal{P}_2+x_c} a_i^T x \leq b_i, \quad i = 1,\ldots,m \end{array}$$

or

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & a_i^T x_c + \sup_{-t\mathbf{1}\leq Cv\leq t\mathbf{1}} a_i^T v \leq b_i, \quad i = 1,\ldots,m. \end{array} \tag{8.28.A}$$

If we define

$$p(a_i) = \sup_{-\mathbf{1}\leq Cv\leq \mathbf{1}} a_i^T v, \tag{8.28.B}$$

we can write (8.28.A) as

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & a_i^T x_c + tp(a_i) \leq b_i, \quad i = 1,\ldots,m, \end{array} \tag{8.28.C}$$

which is an LP in x_c and t. Note that $p(a_i)$ can be evaluated by solving the LP in the definition (8.28.B).
In summary we can solve the problem by first determining $p(a_i)$ for $i = 1,\ldots,m$, by solving m LPs, and then solving the LP (8.28.C) for t and x_c.

(b) We first note that $x \in t\mathcal{P}_2 + x_c$ if and only

$$-t\mathbf{1} \leq C(x - x_c) \leq t\mathbf{1}.$$

The problem is therefore equivalent to

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \sup_{x\in\mathcal{P}_1} \; c_i^T x - c_i^T x_c \leq t, \quad i = 1,\ldots,l \\ & \inf_{x\in\mathcal{P}_1} \; c_i^T x - c_i^T x_c \geq -t, \quad i = 1,\ldots,l \end{array}$$

or

$$\begin{array}{ll}\text{minimize} & t \\ \text{subject to} & -t + \sup_{Ax \preceq b} c_i^T x \leq c_i^T x_c \leq t + \inf_{Ax \preceq b} c_i^T x, \quad i = 1, \ldots, l.\end{array}$$

If we define $p(c_i)$ and $q(c_i)$ as

$$p(c_i) = \sup_{Ax \preceq b} c_i^T x, \quad q(c_i) = \inf_{Ax \preceq b} c_i^T x \tag{8.28.D}$$

then the problem simplifies to

$$\begin{array}{ll}\text{minimize} & t \\ \text{subject to} & -t + p(c_i) \leq c_i^T x_c \leq t + q(c_i), \quad i = 1, \ldots, l,\end{array} \tag{8.28.E}$$

which is an LP in x_c and t.

In conclusion, we can solve the problem by first determining $p(c_i)$ and $q(c_i)$, $i = 1, \ldots, p$ from the $2l$ LPs in the definition (8.28.D), and then solving the LP (8.28.E).

8.29 *Outer polyhedral approximations.* Let $\mathcal{P} = \{x \in \mathbf{R}^n \mid Ax \preceq b\}$ be a polyhedron, and $C \subseteq \mathbf{R}^n$ a given set (not necessarily convex). Use the support function S_C to formulate the following problem as an LP:

$$\begin{array}{ll}\text{minimize} & t \\ \text{subject to} & C \subseteq t\mathcal{P} + x \\ & t \geq 0.\end{array}$$

Here $t\mathcal{P} + x = \{tu + x \mid u \in \mathcal{P}\}$, the polyhedron $\mathcal{P}$ scaled by a factor of t about the origin, and translated by x. The variables are $t \in \mathbf{R}$ and $x \in \mathbf{R}^n$.

Solution. We have $C \subseteq t\mathcal{P} + x$ if and only if $(1/t)(C - x) \subseteq \mathcal{P}$, *i.e.*,

$$S_{(1/t)(C-x)}(a_i) \leq b_i, \quad i = 1, \ldots, m.$$

Noting that for $t \geq 0$,

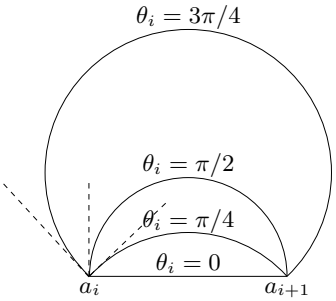
$$S_{(1/t)(C-x)}(a) = \sup_{u \in C} a^T((1/t)(u - x)) = (1/t)(S_C(a) - a^T x),$$

we can express the problem as

$$\begin{array}{ll}\text{minimize} & t \\ \text{subject to} & S_C(a_i) - a_i^T x \leq t b_i, \quad i = 1, \ldots, m \\ & t \geq 0,\end{array}$$

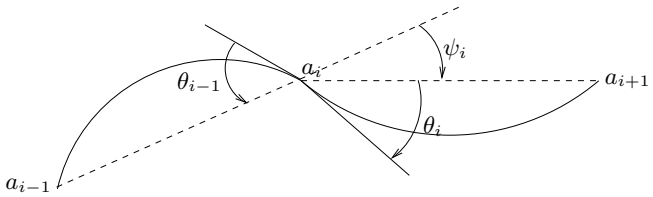
which is an LP in the variables x, t.

8.30 *Interpolation with piecewise-arc curve.* A sequence of points $a_1, \ldots, a_n \in \mathbf{R}^2$ is given. We construct a curve that passes through these points, in order, and is an arc (*i.e.*, part of a circle) or line segment (which we think of as an arc of infinite radius) between consecutive points. Many arcs connect a_i and a_{i+1}; we parameterize these arcs by giving the angle $\theta_i \in (-\pi, \pi)$ between its tangent at a_i and the line segment $[a_i, a_{i+1}]$. Thus, $\theta_i = 0$ means the arc between a_i and a_{i+1} is in fact the line segment $[a_i, a_{i+1}]$; $\theta_i = \pi/2$ means the arc between a_i and a_{i+1} is a half-circle (above the linear segment $[a_1, a_2]$); $\theta_i = -\pi/2$ means the arc between a_i and a_{i+1} is a half-circle (below the linear segment $[a_1, a_2]$). This is illustrated below.



Our curve is completely specified by the angles $\theta_1, \ldots, \theta_n$, which can be chosen in the interval $(-\pi, \pi)$. The choice of θ_i affects several properties of the curve, for example, its *total arc length* L, or the *joint angle discontinuities*, which can be described as follows.

At each point a_i, $i = 2, \ldots, n-1$, two arcs meet, one coming from the previous point and one going to the next point. If the tangents to these arcs exactly oppose each other, so the curve is differentiable at a_i, we say there is no joint angle discontinuity at a_i. In general, we define the joint angle discontinuity at a_i as $|\theta_{i-1}+\theta_i+\psi_i|$, where ψ_i is the angle between the line segment $[a_i, a_{i+1}]$ and the line segment $[a_{i-1}, a_i]$, *i.e.*, $\psi_i = \angle(a_i - a_{i+1}, a_{i-1} - a_i)$. This is shown below. Note that the angles ψ_i are known (since the a_i are known).



We define the *total joint angle discontinuity* as

$$D = \sum_{i=2}^{n} |\theta_{i-1} + \theta_i + \psi_i|.$$

Formulate the problem of minimizing total arc length length L, and total joint angle discontinuity D, as a bi-criterion convex optimization problem. Explain how you would find the extreme points on the optimal trade-off curve.

Solution. The total joint angle discontinuity is

$$D = \sum_{i=2}^{n} |\theta_{i-1} + \theta_i + \psi_i|,$$

which is evidently convex in θ.

The other objective is the total arc length, which turns out to be

$$L = \sum_{i=1}^{n-1} l_i \frac{\theta_i}{\sin \theta_i},$$

where $l_i = \|a_i - a_{i+1}\|_2$. We will show that L is a convex function of θ. Of course we need only show that the function $f(x) = x/\sin x$ is convex over the interval $|x| < \pi$. In fact f is log-convex. With $g = \log(x/\sin x)$, we have

$$g'' = -\frac{1}{x^2} + \frac{1}{\sin^2 x}.$$

Now since $|\sin x| \leq |x|$ for (all) x, we have $1/x^2 \leq 1/\sin^2 x$ for all x, and hence $g'' \geq 0$.

Therefore we find that both objectives D and L are convex. To find the optimal trade-off curve, we minimize various (nonnegative) weighted combinations of D and L, *i.e.*, $D+\lambda L$, for various values of $\lambda \geq 0$.

Now let's consider the extreme points of the trade-off curve. Obviously L is minimized by taking $\theta_i = 0$, *i.e.*, with the curve consisting of the line segments connecting the points. So $\theta = 0$ is one end of the optimal trade-off curve.

We can also say something about the other extreme point, which we claim occurs when the total joint angle discontinuity is zero (which means that the curve is differentiable). This occurs when the recursion

$$\theta_i = -\theta_{i-1} - \psi_i, \quad i = 2, \ldots, n,$$

holds. This shows that once the first angle θ_1 is fixed, the whole curve is fixed. Thus, there is a one-parameter family of piecewise-arc curves that pass through the points, parametrized by θ_1. To find the other extreme point of the optimal trade-off curve, we need to find the curve in this family that has minimum length. This can be found by solving the one-dimensional problem of minimizing L, over θ_1, using the recursion above.

Chapter 9

Unconstrained minimization

Exercises

Unconstrained minimization

9.1 *Minimizing a quadratic function.* Consider the problem of minimizing a quadratic function:

$$\text{minimize} \quad f(x) = (1/2)x^T Px + q^T x + r,$$

where $P \in \mathbf{S}^n$ (but we do not assume $P \succeq 0$).

(a) Show that if $P \not\succeq 0$, *i.e.*, the objective function f is not convex, then the problem is unbounded below.

(b) Now suppose that $P \succeq 0$ (so the objective function is convex), but the optimality condition $Px^\star = -q$ does not have a solution. Show that the problem is unbounded below.

Solution.

(a) If $P \not\succeq 0$, we can find v such that $v^T Pv < 0$. With $x = tv$ we have

$$f(x) = t^2(v^T Pv/2) + t(q^T v) + r,$$

which converges to $-\infty$ as t becomes large.

(b) This means $q \notin \mathcal{R}(P)$. Express q as $q = \tilde{q} + v$, where $\tilde{q}$ is the Euclidean projection of q onto $\mathcal{R}(P)$, and take $v = q - \tilde{q}$. This vector is nonzero and orthogonal to $\mathcal{R}(P)$, *i.e.*, $v^T Pv = 0$. It follows that for $x = tv$, we have

$$f(x) = tq^T v + r = t(\tilde{q} + v)^T v + r = t(v^T v) + r,$$

which is unbounded below.

9.2 *Minimizing a quadratic-over-linear fractional function.* Consider the problem of minimizing the function $f : \mathbf{R}^n \to \mathbf{R}$, defined as

$$f(x) = \frac{\|Ax - b\|_2^2}{c^T x + d}, \qquad \mathbf{dom}\, f = \{x \mid c^T x + d > 0\}.$$

We assume $\mathbf{rank}\, A = n$ and $b \notin \mathcal{R}(A)$.

(a) Show that f is closed.

(b) Show that the minimizer $x^\star$ of f is given by

$$x^\star = x_1 + tx_2$$

where $x_1 = (A^T A)^{-1} A^T b$, $x_2 = (A^T A)^{-1} c$, and $t \in \mathbf{R}$ can be calculated by solving a quadratic equation.

Solution.

(a) Since $b \notin \mathcal{R}(A)$, the numerator is bounded below by a positive number ($\|Ax_{\rm ls} - b\|_2^2$). Therefore $f(x) \to \infty$ as x approaches the boundary of $\mathbf{dom}\, f$.

(b) The optimality conditions are

$$\begin{aligned} \nabla f(x) &= \frac{2}{c^T x - d} A^T (Ax - b) - \frac{\|Ax - b\|_2^2}{(c^T x - d)^2} c \\ &= \frac{2}{c^T x - d}(x - x_1) - \frac{\|Ax - b\|_2^2}{(c^T x - d)^2} x_2 \\ &= 0, \end{aligned}$$

i.e., $x = x_1 + tx_2$ where

$$t = \frac{\|Ax - b\|_2^2}{2(c^T x - d)} = \frac{\|Ax_1 + tAx_2 - b\|_2^2}{2(c^T x_1 + tc^T x_2 - d)}.$$

In other words t must satisfy

$$\begin{aligned} 2t^2 c^T x_2 + 2t(c^T x_1 - d) &= t^2\|Ax_2\|_2^2 + 2t(Ax_1 - b)^T Ax_2 + \|Ax_1 - b\|_2^2 \\ &= t^2 c^T x_2 + \|Ax_1 - b\|_2^2, \end{aligned}$$

which reduces to a quadratic equation

$$t^2 c^T x_2 + 2t(c^T x_1 - d) - \|Ax_1 - b\|_2^2 = 0.$$

We have to pick the root

$$t = \frac{-(c^T x_1 - d) \pm \sqrt{(c^T x_1 - d)^2 + (c^T x_2)\|Ax_1 - b\|_2^2}}{c^T x_2},$$

so that

$$\begin{aligned} c^T(x_1 + tx_2) - d &= c^T x_1 - d - (c^T x_1 - d) + \sqrt{(c^T x_1 - d)^2 + (c^T x_2)\|Ax_1 - b\|_2^2} \\ &= \sqrt{(c^T x_1 - d)^2 + (c^T x_2)\|Ax_1 - b\|_2^2} \\ &> 0. \end{aligned}$$

9.3 *Initial point and sublevel set condition.* Consider the function $f(x) = x_1^2 + x_2^2$ with domain $\mathbf{dom}\, f = \{(x_1, x_2) \mid x_1 > 1\}$.

(a) What is $p^\star$?

(b) Draw the sublevel set $S = \{x \mid f(x) \leq f(x^{(0)})\}$ for $x^{(0)} = (2, 2)$. Is the sublevel set S closed? Is f strongly convex on S?

(c) What happens if we apply the gradient method with backtracking line search, starting at $x^{(0)}$? Does $f(x^{(k)})$ converge to $p^\star$?

Solution.

(a) $p^\star = \lim_{x \to (1,0)} f(x_1.x_2) = 1$.

(b) No, the sublevel set is not closed. The points $(1 + 1/k, 1)$ are in the sublevel set for $k = 1, 2, \ldots$, but the limit, $(1, 1)$, is not.

(c) The algorithm gets stuck at $(1, 1)$.

9.4 Do you agree with the following argument? The ℓ_1-norm of a vector $x \in \mathbf{R}^m$ can be expressed as

$$\|x\|_1 = (1/2) \inf_{y \succ 0} \left(\sum_{i=1}^m x_i^2/y_i + \mathbf{1}^T y \right).$$

Therefore the ℓ_1-norm approximation problem

$$\text{minimize} \quad \|Ax - b\|_1$$

is equivalent to the minimization problem

$$\text{minimize} \quad f(x, y) = \textstyle\sum_{i=1}^m (a_i^T x - b_i)^2/y_i + \mathbf{1}^T y, \tag{9.62}$$

with $\mathbf{dom}\, f = \{(x, y) \in \mathbf{R}^n \times \mathbf{R}^m \mid y \succ 0\}$, where a_i^T is the ith row of A. Since f is twice differentiable and convex, we can solve the ℓ_1-norm approximation problem by applying Newton's method to (9.62).

Solution. The reformulation is valid. The hitch is that the objective function f is not closed.

Exercises

9.5 *Backtracking line search.* Suppose f is strongly convex with $mI \preceq \nabla^2 f(x) \preceq MI$. Let Δx be a descent direction at x. Show that the backtracking stopping condition holds for

$$0 < t \leq -\frac{\nabla f(x)^T \Delta x}{M\|\Delta x\|_2^2}.$$

Use this to give an upper bound on the number of backtracking iterations.

Solution. The upper bound $\nabla^2 f(x) \preceq MI$ implies

$$f(x + t\Delta x) \leq f(x) + t\nabla f(x)^T \Delta x + (M/2)t^2 \Delta x^T \Delta x$$

hence $f(x + t\Delta x) \leq f(x) + \alpha t \nabla f(x)^T \Delta x$ if

$$t(1-\alpha)\nabla f(x)^T \Delta x + (M/2)t^2 \Delta x^T \Delta x \leq 0$$

i.e., the exit condition certainly holds if $0 \leq t \leq t_0$ with

$$t_0 = -2(1-\alpha)\frac{\nabla f(x)^T \Delta x}{M \Delta x^T \Delta x} \geq -\frac{\nabla f(x)^T \Delta x}{M \Delta x^T \Delta x}.$$

Assume $t_0 \leq 1$. Then $\beta^k t \leq t_0$ for $k \geq \log(1/t_0)/\log(1/\beta)$.

Gradient and steepest descent methods

9.6 *Quadratic problem in* $\mathbf{R}^2$. Verify the expressions for the iterates $x^{(k)}$ in the first example of §9.3.2.

Solution. For $k = 0$, we get the starting point $x^{(0)} = (\gamma, 1)$.

The gradient at $x^{(k)}$ is $(x_1^{(k)}, \gamma x_2^{(k)})$, so we get

$$x^{(k)} - t\nabla f(x^{(k)}) = \begin{bmatrix} (1-t)x_1^{(k)} \\ (1-\gamma t)x_2^{(k)} \end{bmatrix} = \left(\frac{\gamma-1}{\gamma+1}\right)^k \begin{bmatrix} (1-t)\gamma \\ (1-\gamma t)(-1)^k \end{bmatrix}$$

and

$$f(x^{(k)} - t\nabla f(x^{(k)})) = (\gamma^2(1-t)^2 + \gamma(1-\gamma t)^2)\left(\frac{\gamma-1}{\gamma+1}\right)^{2k}.$$

This is minimized by $t = 2/(1+\gamma)$, so we have

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - t\nabla f(x^{(k)}) \\ &= \begin{bmatrix} (1-t)x_1^{(k)} \\ (1-\gamma t)\gamma x_2^{(k)} \end{bmatrix} \\ &= \left(\frac{\gamma-1}{\gamma+1}\right)\begin{bmatrix} x_1^{(k)} \\ -x_2^{(k)} \end{bmatrix} \\ &= \left(\frac{\gamma-1}{\gamma+1}\right)^{k+1}\begin{bmatrix} \gamma \\ (-1)^k \end{bmatrix}. \end{aligned}$$

9.7 Let $\Delta x_{\rm sd}$ and $\Delta x_{\rm sd}$ be the normalized and unnormalized steepest descent directions at x, for the norm $\|\cdot\|$. Prove the following identities.

(a) $\nabla f(x)^T \Delta x_{\rm nsd} = -\|\nabla f(x)\|_*$.

(b) $\nabla f(x)^T \Delta x_{\rm sd} = -\|\nabla f(x)\|_*^2$.

(c) $\Delta x_{\rm sd} = \operatorname{argmin}_v(\nabla f(x)^T v + (1/2)\|v\|^2)$.

Solution.

(a) By definition of dual norm.

(b) By (a) and the definition of $\Delta x_{\rm sd}$.

(c) Suppose $v = tw$ with $\|w\| = 1$ and w fixed. We optimize over t and w separately. We have

$$\nabla f(x)^T v + (1/2)\|v\|^2 = t\nabla f(x)^T w + t^2/2.$$

Minimizing over $t \geq 0$ gives the optimum $\hat{t} = -\nabla f(x)^T w$ if $\nabla f(x)^T w \leq 0$, and $\hat{t} = 0$ otherwise. This shows that we should choose w such that $\nabla f(x)^T w \leq 0$. Substituting $\hat{t} = -\nabla f(x)^T w$ gives

$$\hat{t}\nabla f(x)^T w + \hat{t}^2/2 = -(\nabla f(x)^T w)^2/2.$$

We now minimize over w, *i.e.*, solve

$$\begin{array}{ll} \text{minimize} & -(\nabla f(x)^T w)^2/2 \\ \text{subject to} & \|w\| = 1. \end{array}$$

The solution is $w = \Delta x_{\rm nsd}$ by definition. This gives

$$\hat{t} = -\Delta x_{\rm nsd}^T \nabla f(x) = \|\nabla f(x)\|_*,$$

and $v = \hat{t}w = \Delta x_{\rm sd}$.

9.8 *Steepest descent method in ℓ_∞-norm.* Explain how to find a steepest descent direction in the ℓ_∞-norm, and give a simple interpretation.

Solution. The normalized steepest descent direction is given by

$$\Delta x_{\rm nsd} = -\operatorname{sign}(\nabla f(x)),$$

where the sign is taken componentwise. Interpretation: If the partial derivative with respect to x_k is positive we take a step that reduces x_k; if it is positive, we take a step that increases x_k.

The unnormalized steepest descent direction is given by

$$\Delta x_{\rm sd} = -\|\nabla f(x)\|_1 \operatorname{sign}(\nabla f(x)).$$

Newton's method

9.9 *Newton decrement.* Show that the Newton decrement $\lambda(x)$ satisfies

$$\lambda(x) = \sup_{v^T\nabla^2 f(x)v=1} (-v^T\nabla f(x)) = \sup_{v\neq 0} \frac{-v^T\nabla f(x)}{(v^T\nabla^2 f(x)v)^{1/2}}.$$

Solution. The first expression follows from a a change of variables

$$w = \nabla^2 f(x)^{1/2} v, \qquad v = \nabla^2 f(x)^{-1/2} w$$

and from

$$\sup_{\|w\|_2=1} -w^T\nabla^2 f(x)^{-1/2}\nabla f(x) = \|\nabla f(x)^{-1/2}\nabla f(x)\|_2 = \lambda(x).$$

The second expression follows immediately from the first.

9.10 *The pure Newton method.* Newton's method with fixed step size $t = 1$ can diverge if the initial point is not close to $x^\star$. In this problem we consider two examples.

(a) $f(x) = \log(e^x + e^{-x})$ has a unique minimizer $x^\star = 0$. Run Newton's method with fixed step size $t = 1$, starting at $x^{(0)} = 1$ and at $x^{(0)} = 1.1$.

(b) $f(x) = -\log x + x$ has a unique minimizer $x^\star = 1$. Run Newton's method with fixed step size $t = 1$, starting at $x^{(0)} = 3$.

Plot f and f', and show the first few iterates.

Solution.

- $f(x) = \log(e^x + e^{-x})$ is a smooth convex function, with a unique minimum at the origin. The pure Newton method started at $x^{(0)} = 1$ produces the following sequence.

k	$x^{(k)}$	$f(x^{(k)}) - p^\star$
1	$-8.134 \cdot 10^{-01}$	$4.338 \cdot 10^{-1}$
2	$4.094 \cdot 10^{-01}$	$2.997 \cdot 10^{-1}$
3	$-4.730 \cdot 10^{-02}$	$8.156 \cdot 10^{-2}$
4	$7.060 \cdot 10^{-05}$	$1.118 \cdot 10^{-3}$
5	$-2.346 \cdot 10^{-13}$	$2.492 \cdot 10^{-9}$

  Started at $x^{(0)} = 1.1$, the method diverges.

k	$x^{(k)}$	$f(x^{(k)}) - p^\star$
1	$-1.129 \cdot 10^{0}$	$5.120 \cdot 10^{-1}$
2	$1.234 \cdot 10^{0}$	$5.349 \cdot 10^{-1}$
3	$-1.695 \cdot 10^{0}$	$6.223 \cdot 10^{-1}$
4	$5.715 \cdot 10^{0}$	$1.035 \cdot 10^{0}$
5	$-2.302 \cdot 10^{4}$	$2.302 \cdot 10^{4}$

- $f(x) = -\log x + x$ is smooth and convex on $\mathbf{dom}\, f = \{x \mid x > 0\}$, with a unique minimizer at $x = 1$. The pure Newton method started at $x^{(0)} = 3$ gives as first iterate

$$x^{(1)} = 3 - f'(3)/f''(3) = -3$$

  which lies outside $\mathbf{dom}\, f$.

9.11 *Gradient and Newton methods for composition functions.* Suppose $\phi : \mathbf{R} \to \mathbf{R}$ is increasing and convex, and $f : \mathbf{R}^n \to \mathbf{R}$ is convex, so $g(x) = \phi(f(x))$ is convex. (We assume that f and g are twice differentiable.) The problems of minimizing f and minimizing g are clearly equivalent.

Compare the gradient method and Newton's method, applied to f and g. How are the search directions related? How are the methods related if an exact line search is used? *Hint.* Use the matrix inversion lemma (see §C.4.3).

Solution.

(a) *Gradient method.* The gradients are positive multiples

$$\nabla g(x) = \phi'(f(x))\nabla f(x),$$

so with exact line search the iterates are identical for f and g. With backtracking there can be big differences.

(b) *Newton method.* The Hessian of g is

$$\phi''(f(x))\nabla f(x)\nabla f(x)^T + \phi'(f(x))\nabla^2 f(x),$$

so the Newton direction for g is

$$-\left(\phi''(f(x))\nabla f(x)\nabla f(x)^T + \phi'(f(x))\nabla^2 f(x)\right)^{-1} \nabla f(x).$$

From the matrix inversion lemma, we see that this is some positive multiple of the Newton direction for f. Hence with exact line search, the iterates are identical.

Without exact line search, *e.g.*, with Newton step one, there can be big differences. Take *e.g.*, $f(x) = x^2$ and $\phi(x) = x^2$ for $x \geq 0$.

9.12 *Trust region Newton method.* If $\nabla^2 f(x)$ is singular (or very ill-conditioned), the Newton step $\Delta x_{\rm nt} = -\nabla^2 f(x)^{-1}\nabla f(x)$ is not well defined. Instead we can define a search direction $\Delta x_{\rm tr}$ as the solution of

$$\begin{array}{ll} \text{minimize} & (1/2)v^T H v + g^T v \\ \text{subject to} & \|v\|_2 \leq \gamma, \end{array}$$

where $H = \nabla^2 f(x)$, $g = \nabla f(x)$, and γ is a positive constant. The point $x+\Delta x_{\rm tr}$ minimizes the second-order approximation of f at x, subject to the constraint that $\|(x+\Delta x_{\rm tr})-x\|_2 \leq \gamma$. The set $\{v \mid \|v\|_2 \leq \gamma\}$ is called the *trust region*. The parameter γ, the size of the trust region, reflects our confidence in the second-order model.

Show that $\Delta x_{\rm tr}$ minimizes

$$(1/2)v^T H v + g^T v + \hat{\beta}\|v\|_2^2,$$

for some $\hat{\beta}$. This quadratic function can be interpreted as a regularized quadratic model for f around x.

Solution. This follows from duality. If we associate a multiplier β with the constraint, then the optimal v must be a minimizer of the Lagrangian

$$(1/2)v^T H v + g^T v + \beta(\|v\|_2^2 - \gamma).$$

The value of $\hat{\beta}$ can be determined as follows. The optimality conditions are

$$Hv + g + \beta v = 0, \qquad v^T v \leq \gamma, \qquad \beta \geq 0, \qquad \beta(\gamma - v^T v) = 0.$$

- If $H \succ 0$, then $H + \beta I$ is invertible for all $\beta \geq 0$, so from the first equation, $v = -(H + \beta I)^{-1} g$. The norm of v is a decreasing function of β. If $\|H^{-1}g\|_2 \leq \gamma$, then the optimal solution is

  $$v = -H^{-1}g, \qquad \beta = 0.$$

  If $\|H^{-1}g\|_2 > \gamma$, then β is the unique positive solution of the equation $\|(H + \beta I)^{-1}g\|_2 = \gamma$.

- If H is singular, then we have $\beta = 0$ only if $g \in \mathcal{R}(H)$ and $\|H^\dagger g\|_2 \leq \gamma$.
  Otherwise, β is the unique solution positive solution of the equation $\|(H+\beta I)^{-1}g\|_2 = \gamma$.

Self-concordance

9.13 *Self-concordance and the inverse barrier.*

(a) Show that $f(x) = 1/x$ with domain $(0, 8/9)$ is self-concordant.

(b) Show that the function

$$f(x) = \alpha \sum_{i=1}^m \frac{1}{b_i - a_i^T x}$$

with $\mathbf{dom}\, f = \{x \in \mathbf{R}^n \mid a_i^T x < b_i,\ i = 1, \ldots, m\}$, is self-concordant if $\mathbf{dom}\, f$ is bounded and

$$\alpha > (9/8) \max_{i=1,\ldots,m} \sup_{x \in \mathbf{dom}\, f} (b_i - a_i^T x).$$

Solution.

(a) The derivatives are

$$f'(x) = -1/x^2, \qquad f''(x) = 2/x^3, \qquad f'''(x) = -6/x^4,$$

so the self-concordance condition is

$$\frac{6}{x^4} \leq 2\left(\frac{2}{x^3}\right)^{3/2} = \frac{4\sqrt{2}}{x^4\sqrt{x}}.$$

which holds if $\sqrt{x} \leq 4\sqrt{2}/6 = \sqrt{8/9}$.

(b) If we make an affine change of variables $y_i = 8(b_i - a_i^T x)/(9\alpha)$, then $y_i < 8/9$ for all $x \in \mathbf{dom}\, f$. The function f reduces to $\sum_{i=1}^m (1/y_i)$, which is self-concordant by the result in (a).

9.14 *Composition with logarithm.* Let $g : \mathbf{R} \to \mathbf{R}$ be a convex function with $\mathbf{dom}\, g = \mathbf{R}_{++}$, and

$$|g'''(x)| \leq 3\frac{g''(x)}{x}$$

for all x. Prove that $f(x) = -\log(-g(x)) - \log x$ is self-concordant on $\{x \mid x > 0,\ g(x) < 0\}$. *Hint.* Use the inequality

$$\frac{3}{2}rp^2 + q^3 + \frac{3}{2}p^2q + r^3 \leq 1$$

which holds for $p, q, r \in \mathbf{R}_+$ with $p^2 + q^2 + r^2 = 1$.

Solution. The derivatives of f are

$$\begin{aligned}
f'(x) &= -\frac{g'(x)}{g(x)} - \frac{1}{x} \\
f''(x) &= \left(\frac{g'(x)}{g(x)}\right)^2 - \frac{g''(x)}{g(x)} + \frac{1}{x^2} \\
f'''(x) &= -\frac{g'''(x)}{g(x)} - 2\left(\frac{g'(x)}{g(x)}\right)^3 + \frac{3g''(x)g'(x)}{g(x)^2} - \frac{2}{x^3}.
\end{aligned}$$

We have

$$\begin{aligned}
|f'''(x)| &\leq \frac{|g'''(x)|}{-g(x)} + 2\left(\frac{|g'(x)|}{-g(x)}\right)^3 + \frac{3g''(x)|g'(x)|}{g(x)^2} + \frac{2}{x^3} \\
&\leq \frac{3g''(x)}{-xg(x)} + 2\left(\frac{|g'(x)|}{-g(x)}\right)^3 + \frac{3g''(x)|g'(x)|}{g(x)^2} + \frac{2}{x^3}.
\end{aligned}$$

We will show that

$$\frac{3g''(x)}{-xg(x)} + 2\left(\frac{|g'(x)|}{-g(x)}\right)^3 + \frac{3g''(x)|g'(x)|}{g(x)^2} + \frac{2}{x^3} \leq 2\left(\left(\frac{g'(x)}{g(x)}\right)^2 - \frac{g''(x)}{g(x)} + \frac{1}{x^2}\right)^{3/2}.$$

To simplify the formulas we define

$$\begin{aligned}
p &= \frac{(-g''(x)/g(x))^{1/2}}{(-g''(x)/g(x) + g'(x)^2/g(x)^2 + 1/x^2)^{1/2}} \\
q &= \frac{-|g'(x)|/g(x)}{(-g''(x)/g(x) + g'(x)^2/g(x)^2 + 1/x^2)^{1/2}} \\
r &= \frac{1/x}{(-g''(x)/g(x) + g'(x)^2/g(x)^2 + 1/x^2)^{1/2}}.
\end{aligned}$$

Note that $p \geq 0$, $q \geq 0$, $r \geq 0$, and $p^2 + q^2 + r^2 = 1$. With these substitutions, the inequality reduces to the inequality

$$\frac{3}{2}rp^2 + q^3 + \frac{3}{2}p^2q + r^3 \leq 1$$

in the hint.

For completeness we also derive the inequality:

$$\begin{aligned}
\frac{3}{2}rp^2 + q^3 + \frac{3}{2}p^2q + r^3 &= (r+q)(\frac{3}{2}p^2 + q^2 + r^2 - qr) \\
&= (r+q)(\frac{3}{2}(p^2+q^2+r^2) - \frac{1}{2}(r+q)^2) \\
&= \frac{1}{2}(r+q)(3-(r+q)^2) \\
&\leq 1.
\end{aligned}$$

On the last line we use the inequality $(1/2)x(3-x^2) \leq 1$ for $0 \leq x \leq 1$, which is easily verified.

9.15 Prove that the following functions are self-concordant. In your proof, restrict the function to a line, and apply the composition with logarithm rule.

(a) $f(x,y) = -\log(y^2 - x^Tx)$ on $\{(x,y) \mid \|x\|_2 < y\}$.

(b) $f(x,y) = -2\log y - \log(y^{2/p} - x^2)$, with $p \geq 1$, on $\{(x,y) \in \mathbf{R}^2 \mid |x|^p < y\}$.

(c) $f(x,y) = -\log y - \log(\log y - x)$ on $\{(x,y) \mid e^x < y\}$.

Solution.

(a) To prove this, we write f as $f(x,y) = -\log y - \log(y - x^Tx/y)$ and restrict the function to a line $x = \hat{x} + tv$, $y = \hat{y} + tw$,

$$f(\hat{x}+tv, \hat{y}+tw) = -\log\left(\hat{y} + tw - \frac{\hat{x}^T\hat{x}}{\hat{y}+tw} - \frac{2t\hat{x}^Tv}{\hat{y}+tw} - \frac{t^2v^Tv}{\hat{y}+tw}\right) - \log(\hat{y}+tw).$$

If $w = 0$, the argument of the log reduces to a quadratic function of t, which is the case considered in example 9.6.

Otherwise, we can use y instead of t as variable (*i.e.*, make a change of variables $t = (y - \hat{y})/w$). We obtain

$$f(\hat{x}+tv, \hat{y}+tw) = -\log(\alpha + \beta y - \gamma/y) - \log y$$

where

$$\alpha = 2\frac{\hat{y}v^Tv}{w^2} - 2\frac{\hat{x}^Tv}{w}, \qquad \beta = 1 - \frac{v^Tv}{w}, \qquad \gamma = \hat{x}^T\hat{x} - 2\frac{\hat{y}\hat{x}^Tv}{w} + \frac{\hat{y}^2v^Tv}{w^2}.$$

Defining $g(y) = -\alpha - \beta y + \gamma/y$, we have

$$f(\hat{x}+tv, \hat{y}+tw) = -\log(-g(y)) - \log y$$

The function g is convex (since $\gamma > 0$) and satisfies (9.43) because

$$g'''(y) = -6\gamma/y^4, \qquad g''(y) = 2\gamma/y^3.$$

(b) We can write f as a sum of two functions

$$f_1(x,y) = -\log y - \log(y^{1/p} - x), \qquad f_2(x,y) = -\log y - \log(y^{1/p} + x).$$

We restrict the functions to a line $x = \hat{x} + tv$, $y = \hat{y} + tw$. If $w = 0$, both functions reduce to logs of affine functions, so they are self-concordant. If $w \neq 0$, we can use y as variable (*i.e.*, make a change of variables $t = (y - \hat{y})/w$), and reduce the proof to showing that the function

$$-\log y - \log(y^{1/p} + ay + b)$$

is self-concordant. This is true because $g(x) = -ax - b - x^{1/p}$ is convex, with derivatives

$$g'''(x) = -\frac{(1-p)(1-2p)}{p^3}x^{1/p-3}, \qquad g''(x) = \frac{p-1}{p^2}x^{1/p-2},$$

so the inequality (9.43) reduces

$$\frac{(p-1)(2p-1)}{p^3} \leq 3\frac{p-1}{p^2},$$

i.e., $p \geq -1$.

(c) We restrict the function to a line $x = \hat{x} + tv$, $y = \hat{y} + tw$:

$$f(\hat{x} + tv, \hat{y} + tw) = -\log(\hat{y} + tw) - \log(\log(\hat{y} + tw) - \hat{x} - tw).$$

If $w = 0$ the function is obviously self-concordant. If $w \neq 0$, we use y as variable (*i.e.*, use a change of variables $t = (y - \hat{y})/w$), and the function reduces to

$$-\log y - \log(\log y - a - by),$$

so we need to show that $g(y) = a + by - \log y$ satisfies the inequality (9.43). We have

$$g'''(y) = -\frac{2}{y^3}, \qquad g''(y) = \frac{1}{y^2},$$

so (9.43) becomes

$$\frac{2}{y^3} \leq \frac{3}{y^3}.$$

9.16 Let $f : \mathbf{R} \to \mathbf{R}$ be a self-concordant function.

(a) Suppose $f''(x) \neq 0$. Show that the self-concordance condition (9.41) can be expressed as

$$\left|\frac{d}{dx}\left(f''(x)^{-1/2}\right)\right| \leq 1.$$

Find the 'extreme' self-concordant functions of one variable, *i.e.*, the functions f and $\tilde{f}$ that satisfy

$$\frac{d}{dx}\left(f''(x)^{-1/2}\right) = 1, \qquad \frac{d}{dx}\left(\tilde{f}''(x)^{-1/2}\right) = -1,$$

respectively.

(b) Show that either $f''(x) = 0$ for all $x \in \mathbf{dom}\, f$, or $f''(x) > 0$ for all $x \in \mathbf{dom}\, f$.

Solution.

(a) We have

$$\frac{d}{dx}f''(x)^{-1/2} = (-1/2)\frac{f'''(x)}{f''(x)^{3/2}}.$$

Integrating

$$\frac{d}{dx}f''(x)^{-1/2} = 1$$

gives $f(x) = -\log(x + c_0) + c_1 x + c_2$. Integrating

$$\frac{d}{dx} f''(x)^{-1/2} = -1$$

gives

$$f(x) = -\log(-x + c_0) + c_1 x + c_2.$$

(b) Suppose $f''(0) > 0$, $f''(\bar{x}) = 0$ for $\bar{x} > 0$, and $f''(x) > 0$ on the interval between 0 and $\bar{x}$. The inequality

$$-1 \leq \frac{d}{dx} f''(x)^{-1/2} \leq 1$$

holds for x between 0 and $\bar{x}$. Integrating gives

$$f''(\bar{x})^{-1/2} - f''(0)^{-1/2} \leq \bar{x}$$

which contradicts $f''(\bar{x}) = 0$.

9.17 *Upper and lower bounds on the Hessian of a self-concordant function.*

(a) Let $f : \mathbf{R}^2 \to \mathbf{R}$ be a self-concordant function. Show that

$$\begin{aligned} \left| \frac{\partial^3 f(x)}{\partial^3 x_i} \right| &\leq 2 \left(\frac{\partial^2 f(x)}{\partial x_i^2} \right)^{3/2}, \quad i = 1, 2, \\ \left| \frac{\partial^3 f(x)}{\partial x_i^2 \partial x_j} \right| &\leq 2 \frac{\partial^2 f(x)}{\partial x_i^2} \left(\frac{\partial^2 f(x)}{\partial x_j^2} \right)^{1/2}, \quad i \neq j \end{aligned}$$

for all $x \in \mathbf{dom}\, f$.
Hint. If $h : \mathbf{R}^2 \times \mathbf{R}^2 \times \mathbf{R}^2 \to \mathbf{R}$ is a symmetric trilinear form, *i.e.*,

$$\begin{aligned} h(u, v, w) &= a_1 u_1 v_1 w_1 + a_2 (u_1 v_1 w_2 + u_1 v_2 w_1 + u_2 v_1 w_1) \\ &\quad + a_3 (u_1 v_2 w_2 + u_2 v_1 w_1 + u_2 v_2 w_1) + a_4 u_2 v_2 w_2, \end{aligned}$$

then

$$\sup_{u,v,w \neq 0} \frac{h(u, v, w)}{\|u\|_2 \|v\|_2 \|w\|_2} = \sup_{u \neq 0} \frac{h(u, u, u)}{\|u\|_2^3}.$$

Solution. We first note the following generalization of the result in the hint. Suppose $A \in \mathbf{S}^2_{++}$, and h is symmetric and trilinear. Then $h(A^{-1/2}u, A^{-1/2}v, A^{-1/2}w)$ is a symmetric trilinear function, so

$$\sup_{u,v,w \neq 0} \frac{h(A^{-1/2}u, A^{-1/2}v, A^{-1/2}w)}{\|u\|_2 \|v\|_2 \|w\|_2} = \sup_{u \neq 0} \frac{h(A^{-1/2}u, A^{-1/2}u, A^{-1/2}u)}{\|u\|_2^3},$$

i.e.,

$$\sup_{u,v,w \neq 0} \frac{h(u, v, w)}{(u^T A u)^{1/2} (v^T A v)^{1/2} (w^T A w)^{1/2}} = \sup_{u \neq 0} \frac{h(u, u, u)}{(u^T A u)^{3/2}}. \tag{9.17.A}$$

By definition, $f : \mathbf{R}^n \to \mathbf{R}$ is self-concordant if and only if

$$\left| u^T \left(\frac{d}{dt} \nabla^2 f(\hat{x} + tu) \Big|_{t=0} \right) u \right| \leq 2 (u^T \nabla^2 f(\hat{x}) u)^{3/2}.$$

for all u and all $\hat{x} \in \mathbf{dom}\, f$. If $n = 2$ this means that

$$|h(u, u, u)| \leq (u^T A u)^{3/2}$$

for all u, where

$$\begin{aligned} h(u,v,w) &= u^T \left(\frac{d}{dt} \nabla^2 f(\hat{x}+tv)\Big|_{t=0} \right) w \\ &= u_1 v_1 w_1 \frac{\partial^3 f(\hat{x})}{\partial x_1^3} + (u_1 v_1 w_2 + u_1 v_2 w_1 + u_2 v_1 w_1) \frac{\partial^3 f(\hat{x})}{\partial x_1^2 \partial x_2} \\ &\quad + (u_1 v_2 w_2 + u_2 v_1 w_2 + u_2 v_2 w_1) \frac{\partial^3 f(\hat{x})}{\partial x_1 \partial x_2^2} + u_2 v_2 w_2 \frac{\partial^3 f(\hat{x})}{\partial x_2^3} \\ u^T A u &= u_1^2 \frac{\partial^2 f(\hat{x})}{\partial x_1^2} + 2 u_1 u_2 \frac{\partial^2 f(\hat{x})}{\partial x_1 \partial x_2} + u_2^2 \frac{\partial^2 f(\hat{x})}{\partial x_2^2}, \end{aligned}$$

i.e., $A = \nabla^2 f(\hat{x})$. In other words,

$$\sup_{u \neq 0} \frac{h(u,u,u)}{(u^T A u)^{3/2}} \leq 2, \qquad \sup_{u \neq 0} \frac{-h(u,u,u)}{(u^T A u)^{3/2}} \leq 2.$$

Applying (9.17.A) (to h and $-h$), we also have

$$|h(u,v,u)| \leq 2 (u^T A u)(v^T A v)^{1/2} \tag{9.17.B}$$

for all u and v. The inequalities

$$\left| \frac{\partial^3 f(x)}{\partial^3 x_1} \right| \leq 2 \left(\frac{\partial^2 f(x)}{\partial x_1^2} \right)^{3/2}, \qquad \left| \frac{\partial^3 f(x)}{\partial^3 x_2} \right| \leq 2 \left(\frac{\partial^2 f(x)}{\partial x_2^2} \right)^{3/2},$$

follow from (9.17.B) by choosing $u = v = (1,0)$ and $u = v = (0,1)$, respectively. The inequalities

$$\left| \frac{\partial^3 f(x)}{\partial x_1^2 \partial x_2} \right| \leq 2 \frac{\partial^2 f(x)}{\partial x_1^2} \left(\frac{\partial^2 f(x)}{\partial x_2^2} \right)^{1/2}, \qquad \left| \frac{\partial^3 f(x)}{\partial x_1 \partial x_2^2} \right| \leq 2 \left(\frac{\partial^2 f(x)}{\partial x_1^2} \right)^{1/2} \frac{\partial^2 f(x)}{\partial x_1^2},$$

follow by choosing $v = (1,0)$, $w = (0,1)$, and $v = (0,1)$, $w = (1,0)$, respectively.

To complete the proof we relax the assumption that $\nabla^2 f(\hat{x}) \succ 0$. Note that if f is self-concordant then $f(x) + \epsilon x^T x$ is self-concordant for all $\epsilon \geq 0$. Applying the inequalities to $f(x) + \epsilon x^T x$ gives

$$\left| \frac{\partial^3 f(x)}{\partial^3 x_i} \right| \leq 2 \left(\frac{\partial^2 f(x)}{\partial x_i^2} \right)^{3/2} + \epsilon, \qquad \left| \frac{\partial^3 f(x)}{\partial x_i^2 \partial x_j} \right| \leq 2 \frac{\partial^2 f(x)}{\partial x_i^2} \left(\frac{\partial^2 f(x)}{\partial x_j^2} \right)^{1/2} + \epsilon$$

for all $\epsilon > 0$. This is only possible if the inequalities hold for $\epsilon = 0$.

(b) Let $f : \mathbf{R}^n \to \mathbf{R}$ be a self-concordant function. Show that the nullspace of $\nabla^2 f(x)$ is independent of x. Show that if f is strictly convex, then $\nabla^2 f(x)$ is nonsingular for all $x \in \mathbf{dom}\, f$.

Hint. Prove that if $w^T \nabla^2 f(x) w = 0$ for some $x \in \mathbf{dom}\, f$, then $w^T \nabla^2 f(y) w = 0$ for all $y \in \mathbf{dom}\, f$. To show this, apply the result in (a) to the self-concordant function $\tilde{f}(t,s) = f(x + t(y-x) + sw)$.

Solution. Suppose $w^T \nabla^2 f(x) w = 0$. We show that $w^T \nabla^2 f(y) w = 0$ for all $y \in \mathbf{dom}\, f$.

Define $v = y - x$ and let $\tilde{f}$ be the restriction of f to the plane through x and defined by w, v:

$$\tilde{f}(s,t) = f(x + sw + tv).$$

Also define

$$g(t) = w^T \nabla^2 f(x+tv)w = \frac{\partial^2 \tilde{f}(0,t)}{\partial s^2}.$$

$\tilde{f}$ is a self-concordant function of two variables, so from (a),

$$|g'(t)| = \left|\frac{\partial^3 \tilde{f}(0,t)}{\partial t \partial s^2}\right| \leq 2\left(\frac{\partial^2 \tilde{f}(0,t)}{\partial t^2}\right)^{1/2} \frac{\partial^2 \tilde{f}(0,t)}{\partial s^2} = 2\left(\frac{\partial^2 \tilde{f}(0,t)}{\partial s^2}\right)^{1/2} g(t),$$

i.e., if $g(t) \neq 0$, then

$$\frac{d}{dt}\log g(t) \geq -2\left(\frac{\partial^2 \tilde{f}(0,t)}{\partial s^2}\right)^{1/2}.$$

By assumption, $g(0) > 0$ and $g(t) = 0$ for $t = 1$. Assume that $g(\tau) > 0$ for $0 \leq \tau < t$. (If not, replace t with the smallest positive t for which $g(t) = 0$.) Integrating the inequality above, we have

$$\begin{aligned}
\log(g(t)/g(0)) &\geq -2\int_0^t \left(\frac{\partial^2 \tilde{f}(0,\tau)}{\partial s^2}\right)^{1/2} d\tau \\
g(t)/g(0) &\geq \exp\left(-2\int_0^t \left(\frac{\partial^2 \tilde{f}(0,\tau)}{\partial s^2}\right)^{1/2} d\tau\right),
\end{aligned}$$

which contradicts the assumption $g(t) = 0$. We conclude that either $g(t) = 0$ for all t, or $g(t) > 0$ for all t. This is true for arbitrary x and v, so a vector w either satisfies $w^T \nabla^2 f(x) w = 0$ for all x, or $w^T \nabla^2 f(x) w > 0$ for all x.

Finally, suppose f is strictly convex but satisfies $v^T \nabla^2 f(x) v = 0$ for some x and $v \neq 0$. By the previous result, $v^T \nabla^2 f(x+tv) v = 0$ for all t, *i.e.*, f is affine on the line $x + tv$, and not strictly convex.

(c) Let $f : \mathbf{R}^n \to \mathbf{R}$ be a self-concordant function. Suppose $x \in \mathbf{dom}\, f$, $v \in \mathbf{R}^n$. Show that

$$(1-t\alpha)^2 \nabla^2 f(x) \preceq \nabla^2 f(x+tv) \preceq \frac{1}{(1-t\alpha)^2}\nabla^2 f(x)$$

for $x + tv \in \mathbf{dom}\, f$, $0 \leq t < \alpha$, where $\alpha = (v^T \nabla^2 f(x) v)^{1/2}$.

Solution. As in part (b), we can prove that

$$\left|\frac{d}{dt}\log g(t)\right| \leq 2\left(\frac{\partial^2 \tilde{f}(0,t)}{\partial s^2}\right)^{1/2}$$

where $g(t) = w^T \nabla^2 f(x+tv) w$ and $\tilde{f}(s,t) = f(x + sw + tv)$. Applying the upper bound in (9.46) to the self-concordant function $\tilde{f}(0,t) = f(x+tv)$ of one variable, t, we obtain

$$\frac{\partial^2 \tilde{f}(0,t)}{\partial s^2} \leq \frac{\alpha^2}{(1-t\alpha)^2},$$

so

$$\frac{-2\alpha}{(1-t\alpha)} \leq \frac{d}{dt}\log g(t) \leq \frac{2\alpha}{(1-t\alpha)}.$$

Integrating gives

$$2\log(1-t\alpha) \leq \log(g(t)/g(0)) \leq -2\log(1-t\alpha)$$

$$g(0)(1-t\alpha)^2 \leq g(t) \leq \frac{g(0)}{(1-t\alpha)^2}.$$

Finally, observing that $g(0) = \alpha^2$ gives the inequalities

$$(1-t\alpha)^2 w^T \nabla^2 f(x) w \leq w^T \nabla^2 f(x+tv) w \leq \frac{w^T \nabla^2 f(x) w}{(1-t\alpha)^2}.$$

This holds for all w, and hence

$$(1-t\alpha)^2 \nabla^2 f(x) \preceq \nabla^2 f(x+tv) \preceq \frac{1}{(1-t\alpha)^2} \nabla^2 f(x).$$

9.18 *Quadratic convergence.* Let $f : \mathbf{R}^n \to \mathbf{R}$ be a strictly convex self-concordant function. Suppose $\lambda(x) < 1$, and define $x^+ = x - \nabla^2 f(x)^{-1} \nabla f(x)$. Prove that $\lambda(x^+) \leq \lambda(x)^2/(1-\lambda(x))^2$. *Hint.* Use the inequalities in exercise 9.17, part (c).

Solution. Let $v = -\nabla^2 f(x)^{-1} \nabla f(x)$. From exercise 9.17, part (c),

$$(1-t\lambda(x))^2 \nabla^2 f(x) \preceq \nabla^2 f(x+tv) \preceq \frac{1}{(1-t\lambda(x))^2} \nabla^2 f(x).$$

We can assume without loss of generality that $\nabla^2 f(x) = I$ (hence, $v = -\nabla f(x)$), and

$$(1-\lambda(x))^2 I \preceq \nabla^2 f(x^+) \preceq \frac{1}{(1-\lambda(x))^2} I.$$

We can write $\lambda(x^+)$ as

$$\begin{aligned}
\lambda(x^+) &= \|\nabla^2 f(x^+)^{-1} \nabla f(x^+)\|_2 \\
&\leq (1-\lambda(x))^{-1} \|\nabla f(x^+)\|_2 \\
&= (1-\lambda(x))^{-1} \left\| \left(\int_0^1 \nabla^2 f(x+tv) v \, dt + \nabla f(x) \right) \right\|_2 \\
&= (1-\lambda(x))^{-1} \left\| \left(\int_0^1 (\nabla^2 f(x+tv) - I) \, dt \right) v \right\|_2 \\
&\leq (1-\lambda(x))^{-1} \left\| \left(\int_0^1 (\frac{1}{(1-t\lambda(x))^2} - 1) \, dt \right) v \right\|_2 \\
&\leq \|v\|_2 (1-\lambda(x))^{-1} \int_0^1 (\frac{1}{(1-t\lambda(x))^2} - 1) \, dt \\
&= \frac{\lambda(x)^2}{(1-\lambda(x))^2}.
\end{aligned}$$

9.19 *Bound on the distance from the optimum.* Let $f : \mathbf{R}^n \to \mathbf{R}$ be a strictly convex self-concordant function.

(a) Suppose $\lambda(\bar{x}) < 1$ and the sublevel set $\{x \mid f(x) \leq f(\bar{x})\}$ is closed. Show that the minimum of f is attained and

$$\left((\bar{x} - x^\star)^T \nabla^2 f(\bar{x}) (\bar{x} - x^\star) \right)^{1/2} \leq \frac{\lambda(\bar{x})}{1-\lambda(\bar{x})}.$$

(b) Show that if f has a closed sublevel set, and is bounded below, then its minimum is attained.

Solution.

(a) As in the derivation of (9.47) we consider the function $\tilde{f}(t) = f(\hat{x} + tv)$ for an arbitrary descent direction v. Note from (9.44) that

$$1 + \frac{\tilde{f}'(0)}{\tilde{f}''(0)^{1/2}} > 0$$

if $\lambda(\bar{x}) < 1$.

We first argue that $\tilde{f}(t)$ reaches its minimum for some positive (finite) $t^\star$. Let $t_0 = \sup\{t \geq 0 \mid \hat{x} + tv \in \mathbf{dom}\, f\}$. If $t_0 = \infty$ (*i.e.*, $\hat{x} + tv \in \mathbf{dom}\, f$ for all $t \geq 0$), then, from (9.47), $\tilde{f}'(t) > 0$ for

$$t > \bar{t} = \frac{-\tilde{f}'(0)}{\tilde{f}''(0) + \tilde{f}''(0)^{1/2}\tilde{f}'(0)},$$

so $\tilde{f}$ must reach a minimum in the interval $(0, \bar{t})$.

If t_0 is finite, then we must have

$$\lim_{t \to t_0} \tilde{f}(t) > \tilde{f}(0).$$

since the sublevel set $\{t \mid \tilde{f}(t) \leq \tilde{f}(0)\}$ is closed. Therefore $\tilde{f}$ reaches a minimum in the interval $(0, t_0)$.

In both cases,

$$\begin{aligned} t^\star &\leq \frac{-\tilde{f}'(0)}{\tilde{f}''(0) + \tilde{f}''(0)^{1/2}\tilde{f}'(0)} \\ \sqrt{\tilde{f}''(0)}t^\star &\leq \frac{-\tilde{f}'(0)/\sqrt{\tilde{f}''(0)}}{1 + \tilde{f}'(0)/\sqrt{\tilde{f}''(0)}} \\ &\leq \frac{\lambda(x)}{1 - \lambda(x)} \end{aligned}$$

where again we used (9.44). This bound on $t^\star$ holds for any descent vector v. In particular, in the direction $v = x^\star - x$, we have $t^\star = 1$, so we obtain

$$\left((\bar{x} - x^\star)^T \nabla^2 f(\bar{x})(\bar{x} - x^\star)\right)^{1/2} \leq \frac{\lambda(\bar{x})}{1 - \lambda(\bar{x})}.$$

(b) If f is strictly convex, and self-concordant, with a closed sublevel set, then our convergence analysis of Newton's method applies. In other words, after a finite number of iterations, $\lambda(x)$ becomes less than one, and from the previous result this means that the minimum is attained.

9.20 *Conjugate of a self-concordant function.* Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is closed, strictly convex, and self-concordant. We show that its conjugate (or Legendre transform) f^* is self-concordant.

(a) Show that for each $y \in \mathbf{dom}\, f^*$, there is a unique $x \in \mathbf{dom}\, f$ that satisfies $y = \nabla f(x)$. *Hint.* Refer to the result of exercise 9.19.

(b) Suppose $\bar{y} = \nabla f(\bar{x})$. Define

$$g(t) = f(\bar{x} + tv), \qquad h(t) = f^*(\bar{y} + tw)$$

where $v \in \mathbf{R}^n$ and $w = \nabla^2 f(\bar{x})v$. Show that

$$g''(0) = h''(0), \qquad g'''(0) = -h'''(0).$$

Use these identities to show that f^* is self-concordant.

Solution.

(a) $y \in \mathbf{dom}\, f^*$ means that $f(x) - y^T x$ is bounded below as a function of f. From exercise 9.19, part (a), the minimum is attained. The minimizer satisfies $\nabla f(x) = y$, and is unique because $f(x) - y^T x$ is strictly convex.

(b) Let F be the inverse mapping of ∇f, *i.e.*, $x = F(y)$ if and only if $y = \nabla f(x)$. We have $\bar{x} = F(\bar{y})$, and also (from exercise 3.40),

$$\nabla f^*(y) = F(y), \qquad \nabla^2 f^*(y) = \nabla^2 f(F(y))^{-1}$$

for all $y \in \mathbf{dom}\, f^*$.
The first equality follows from $\nabla^2 f^*(\bar{y}) = \nabla^2 f(\bar{x})^{-1}$:

$$g''(0) = v^T \nabla^2 f(\bar{x}) v = w^T \nabla^2 f^*(\bar{y}) w = h''(0).$$

In order to prove the second equality we define

$$G = \frac{d}{dt} \nabla^2 f(\bar{x} + tv)\Big|_{t=0}, \qquad H = \frac{d}{dt} \nabla^2 f^*(\bar{y} + tw)\Big|_{t=0},$$

i.e., we have

$$\nabla^2 f(\bar{x} + tv) \approx \nabla^2 f(\bar{x}) + tG, \qquad \nabla^2 f^*(\bar{y} + tw) \approx \nabla^2 f^*(\bar{y}) + tH$$

for small t, and

$$\begin{aligned} \nabla^2 f^*(\nabla f(\bar{x} + tv)) &\approx \nabla^2 f^*(\nabla f(\bar{x}) + t\nabla^2 f(\bar{x}) v) \\ &= \nabla^2 f^*(\bar{y} + tw) \\ &\approx \nabla^2 f^*(\bar{y}) + tH. \end{aligned}$$

Linearizing both sides of the equation

$$\nabla^2 f^*(\nabla f(\bar{x} + tv)) \nabla^2 f(\bar{x} + tv) = I$$

gives

$$H \nabla^2 f(\bar{x}) + \nabla^2 f^*(\bar{y}) G = 0,$$

i.e., $G = -\nabla^2 f(\bar{x}) H \nabla^2 f(\bar{x})$. Therefore

$$\begin{aligned} g'''(0) &= \frac{d}{dt} v^T \nabla^2 f(\bar{x} + tv) v \Big|_{t=0} \\ &= v^T G v \\ &= -w^T H w \\ &= -\frac{d}{dt} w^T \nabla^2 f^*(\bar{y} + tw) w \Big|_{t=0} \\ &= -h'''(0). \end{aligned}$$

It follows that

$$|h'''(0)| \leq 2h''(0)^{3/2},$$

for any $\bar{y} \in \mathbf{dom}\, f^*$ and all w, so f^* is self-concordant.

9.21 *Optimal line search parameters.* Consider the upper bound (9.56) on the number of Newton iterations required to minimize a strictly convex self-concordant functions. What is the minimum value of the upper bound, if we minimize over α and β?

Solution. Clearly, we should take β near one.
The function

$$\frac{20 - 8\alpha}{\alpha(1 - 2\alpha)^2}$$

reaches its minimum at $\alpha = 0.1748$, with a minimum value of about 252, so the lowest upper bound is

$$252(f(x^{(0)}) - p^\star) + \log_2 \log_2(1/\epsilon).$$

9.22 Suppose that f is strictly convex and satisfies (9.42). Give a bound on the number of Newton steps required to compute $p^\star$ within ϵ, starting at $x^{(0)}$.

Solution.

$$\frac{\tilde{f}(x^{(0)}) - \tilde{p}^\star}{\gamma} + \log_2 \log_2(4\epsilon/k^2)$$

where $\tilde{f} = (k^2/4)f$. In other words

$$(k^2/4)\frac{f(x^{(0)}) - p^\star}{\gamma} + \log_2 \log_2(4\epsilon/k^2).$$

Implementation

9.23 *Pre-computation for line searches.* For each of the following functions, explain how the computational cost of a line search can be reduced by a pre-computation. Give the cost of the pre-computation, and the cost of evaluating $g(t) = f(x + t\Delta x)$ and $g'(t)$ with and without the pre-computation.

(a) $f(x) = -\sum_{i=1}^m \log(b_i - a_i^T x)$.

(b) $f(x) = \log\left(\sum_{i=1}^m \exp(a_i^T x + b_i)\right)$.

(c) $f(x) = (Ax - b)^T(P_0 + x_1P_1 + \cdots + x_nP_n)^{-1}(Ax - b)$, where $P_i \in \mathbf{S}^m$, $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$ and $\mathbf{dom}\, f = \{x \mid P_0 + \sum_{i=1}^n x_iP_i \succ 0\}$.

Solution.

(a) Without pre-computation the cost is order mn.
We can write g as

$$g(t) = -\sum_{i=1}^m \log(b_i - a_i^T x) - \sum_{i=1}^m \log(1 - ta_i^T\Delta x/(b_i - a_i^T x)),$$

so if we pre-compute $w_i = a_i^T\Delta x/(b_i - a_i^T x)$, we can express g as

$$g(t) = g(0) - \sum_{i=1}^m \log(1 - tw_i), \qquad g'(t) = -\sum_{i=1}^m \frac{w_i}{1 - tw_i}.$$

The cost of the pre-computation is $2mn + m$ (if we assume $b - Ax$ is already computed). After the pre-computation the cost of evaluating g and g' is linear in m.

(b) Without pre-computation the cost is order mn. We can write g as

$$\begin{aligned} g(t) &= \log\left(\sum_{i=1}^m \exp(a_i^T x + b_i + ta_i^T\Delta x)\right) \\ &= \log\sum_{i=1}^m e^{t\alpha_i + \beta_i} \end{aligned}$$

where $\alpha_i = a_i^T\Delta x$ and $\beta_i = a_i^T x + b_i$. If we pre-compute α_i and β_i (at a cost that is order mn), we can reduce the cost of computing g and g' to order m.

(c) Without pre-computation the cost is $2mn$ (for computing $Ax-b$), plus $2nm^2$ (for computing $P(x)$), followed by $(1/3)m^3$ (for computing $P(x)^{-1}(Ax-b)$, followed by $2m$ for the inner product. The total cost $2nm^2+(1/3)m^3$.

The following pre-computation steps reduce the complexity:

- Compute the Cholesky factorization $P(x)=LL^T$
- Compute the eigenvalue decomposition $L^{-1}(\sum_{i=1}^n \Delta x_i P_i)L^{-T}=Q\Lambda Q^T$.
- Compute $y=Q^TL^{-1}Ax$, and $v=Q^TL^{-1}A\Delta x$.

The pre-computation involves steps that are order m^3 (Cholesky factorization, eigenvalue decomposition), $2nm^2$ (computing $P(x)$ and $\sum_i \Delta x_i P_i$), and lower order terms.

After the pre-computation we can express g as

$$g(x+t\Delta x)=\sum_{i=1}^m \frac{(y_i+tv_i)^2}{1+t\lambda_i},$$

which can be evaluated and differentiated in order m operations.

9.24 *Exploiting block diagonal structure in the Newton system.* Suppose the Hessian $\nabla^2 f(x)$ of a convex function f is block diagonal. How do we exploit this structure when computing the Newton step? What does it mean about f?

Solution. If the Hessian is block diagonal, then the objective function is *separable*, *i.e.*, a sum of functions of disjoint sets of variables. This means we might as well solve each of the problems separately.

9.25 *Smoothed fit to given data.* Consider the problem

$$\text{minimize} \quad f(x)=\textstyle\sum_{i=1}^n \psi(x_i-y_i)+\lambda\sum_{i=1}^{n-1}(x_{i+1}-x_i)^2$$

where $\lambda>0$ is smoothing parameter, ψ is a convex penalty function, and $x\in\mathbf{R}^n$ is the variable. We can interpret x as a smoothed fit to the vector y.

(a) What is the structure in the Hessian of f?

(b) Extend to the problem of making a smooth fit to two-dimensional data, *i.e.*, minimizing the function

$$\sum_{i,j=1}^n \psi(x_{ij}-y_{ij})+\lambda\left(\sum_{i=1}^{n-1}\sum_{j=1}^{n}(x_{i+1,j}-x_{ij})^2+\sum_{i=1}^{n}\sum_{j=1}^{n-1}(x_{i,j+1}-x_{ij})^2\right),$$

with variable $X\in\mathbf{R}^{n\times n}$, where $Y\in\mathbf{R}^{n\times n}$ and $\lambda>0$ are given.

Solution.

(a) Tridiagonal.

(b) Block-tridiagonal if we store the elements of X columnwise. The blocks have size $n\times n$. The diagonal blocks are tridiagonal. The blocks on the first sub-diagonal are diagonal.

9.26 *Newton equations with linear structure.* Consider the problem of minimizing a function of the form

$$f(x)=\sum_{i=1}^N \psi_i(A_ix+b_i) \tag{9.63}$$

where $A_i\in\mathbf{R}^{m_i\times n}$, $b_i\in\mathbf{R}^{m_i}$, and the functions $\psi_i:\mathbf{R}^{m_i}\to\mathbf{R}$ are twice differentiable and convex. The Hessian H and gradient g of f at x are given by

$$H=\sum_{i=1}^N A_i^TH_iA_i, \qquad g=\sum_{i=1}^N A_i^Tg_i. \tag{9.64}$$

where $H_i = \nabla^2\psi_i(A_ix + b_i)$ and $g_i = \nabla\psi_i(A_ix + b_i)$.

Describe how you would implement Newton's method for minimizing f. Assume that $n \gg m_i$, the matrices A_i are very sparse, but the Hessian H is dense.

Solution.

In many applications, for example, when n is small compared to the dimensions m_i, the simplest and most efficient way to calculate the Newton direction is to evaluate H and g using (9.64), and solve the Newton system with a dense Cholesky factorization.

It is possible, however, that the matrices A_i are very sparse, while H itself is dense. In that case the straightforward method, which involves solving a dense set of linear equations of size n, may not be the most efficient method, since it does not take advantage of sparsity. Specifically, assume that $n \gg m_i$, $\mathbf{rank}\, A_i = m_i$, and $H_i \succ 0$, so the Hessian is a sum of N matrices of rank m_i. We can introduce new variables $y_i = A_i^T v$, and write the Newton system as

$$\sum_{i=1}^N A_i^T y_i = -g, \qquad y_i = H_i A_i^T v, \quad i = 1, \ldots, N.$$

This is an indefinite system of $n + \sum_i m_i$ linear equations in $n + \sum_i m_i$ variables:

$$\begin{bmatrix} -H_1^{-1} & 0 & \cdots & 0 & A_1 \\ 0 & -H_2^{-1} & \cdots & 0 & A_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -H_N^{-1} & A_N \\ A_1^T & A_2^T & \cdots & A_N^T & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -g \end{bmatrix}. \tag{9.26.A}$$

This system is larger than the Newton system, but if $n \gg m_i$, and the matrices A_i are sparse, it may be easier to solve (9.26.A) using a sparse solver than to solve the Newton system directly.

9.27 *Analytic center of linear inequalities with variable bounds.* Give the most efficient method for computing the Newton step of the function

$$f(x) = -\sum_{i=1}^n \log(x_i + 1) - \sum_{i=1}^n \log(1 - x_i) - \sum_{i=1}^m \log(b_i - a_i^T x),$$

with $\mathbf{dom}\, f = \{x \in \mathbf{R}^n \mid -\mathbf{1} \prec x \prec \mathbf{1}, Ax \prec b\}$, where a_i^T is the ith row of A. Assume A is dense, and distinguish two cases: $m \geq n$ and $m \leq n$. (See also exercise 9.30.)

Solution. Note that f has the form (9.60) with $k = n$, $p = m$, $g = b$, $F = -A$, and

$$\psi_0(y) = -\sum_{i=1}^m \log y_i, \qquad \psi_i(x_i) = -\log(1 - x_i^2), \quad i = 1, \ldots, n.$$

The Hessian f at x is given by

$$H = D + A^T \hat{D} A \tag{9.27.A}$$

where $D_{ii} = 1/(1 - x_i)^2 + 1/(x_i + 1)^2$, and $\hat{D}_{ii} = 1/(b_i - a_i^T x)^2$.

The first possibility is to form H as given by (9.27.A), and to solve the Newton system using a dense Cholesky factorization. The cost is mn^2 operations (to form $A^T \hat{D} A$) plus $(1/3)n^3$ for the Cholesky factorization.

A second possibility is to introduce a new variable $y = \hat{D}Av$, and to write the Newton system as

$$D\Delta x_{\mathrm{nt}} + A^T y = -g, \qquad \hat{D}^{-1} y = A\Delta x_{\mathrm{nt}}. \tag{9.27.B}$$

From the first equation, $\Delta x_{\rm nt} = D^{-1}(-g - A^T y)$, and substituting this in the second equation, we obtain

$$(\hat{D}^{-1} + AD^{-1}A^T)y = -AD^{-1}g. \tag{9.27.C}$$

This is a positive definite set of m linear equations in the variable $y \in \mathbf{R}^m$. Given y, we find $\Delta x_{\rm nt}$ by evaluating $\Delta x_{\rm nt} = -D^{-1}(g + A^T y)$. The cost of forming and solving (9.27.C) is $mn^2 + (1/3)m^3$ operations (assuming A is dense). Therefore if $m < n$, this second method is faster than directly solving the Newton system $H\Delta x_{\rm nt} = -g$.

A third possibility is to solve (9.27.B) as an indefinite set of $m + n$ linear equations

$$\begin{bmatrix} D & A^T \\ A & -\hat{D}^{-1} \end{bmatrix} \begin{bmatrix} \Delta x_{\rm nt} \\ y \end{bmatrix} = \begin{bmatrix} -g \\ 0 \end{bmatrix}. \tag{9.27.D}$$

This method is interesting when A is sparse, and the two matrices $D + A^T\hat{D}A$ and $\hat{D}^{-1} + AD^{-1}A^T$ are not. In that case, solving (9.27.D) using a sparse solver may be faster than the two methods above.

9.28 *Analytic center of quadratic inequalities.* Describe an efficient method for computing the Newton step of the function

$$f(x) = -\sum_{i=1}^{m} \log(-x^T A_i x - b_i^T x - c_i),$$

with $\mathbf{dom}\, f = \{x \mid x^T A_i x + b_i^T x + c_i < 0,\ i = 1, \ldots, m\}$. Assume that the matrices $A_i \in \mathbf{S}^n_{++}$ are large and sparse, and $m \ll n$.

Hint. The Hessian and gradient of f at x are given by

$$H = \sum_{i=1}^{m} (2\alpha_i A_i + \alpha_i^2 (2A_i x + b_i)(2A_i x + b_i)^T), \qquad g = \sum_{i=1}^{m} \alpha_i (2A_i x + b_i),$$

where $\alpha_i = 1/(-x^T A_i - b_i^T x - c_i)$.

Solution. We can write H as $H = Q + FF^T$, where

$$Q = 2\sum_{i=1}^{m} \alpha_i A_i, \quad F = \begin{bmatrix} \alpha_1(2A_1x + b_1) & \alpha_2(2A_2x + b_2) & \cdots & \alpha_m(2A_mx + b_m) \end{bmatrix}.$$

In general the Hessian will be dense, even when the matrices A_i are sparse, because of the dense rank-one terms. Finding the Newton direction by building and solving the Newton system $Hv = g$, therefore costs at least $(1/3)n^3$ operations, since we need a dense Cholesky factorization.

An alternative that may be faster when $n \gg m$ is as follows. We introduce a new variable $y \in \mathbf{R}^m$, and write the Newton system as

$$Qv + Fy = -g, \qquad y = F^T v.$$

Substituting $v = -Q^{-1}(g + Fy)$ in the second equation yields

$$(I + F^T Q^{-1} F)y = -F^T Q^{-1} g, \tag{9.28.A}$$

which is a set of m linear equations.

We can therefore also compute the Newton direction as follows. We factor Q using a sparse Cholesky factorization. Then we calculate the matrix $V = Q^{-1}F$ by solving the matrix equation $QV = F$ column by column, using the Cholesky factors of Q. For each colum this involves a sparse forward and backward substitution. We then form the matrix $I + F^T V$ ($m^2 n$ flops), factor it using a dense Cholesky factorization ($(1/3)m^3$ flops), and solve for y. Finally we compute v by solving $Qv = -g - Fy$. The cost of this procedure is $(1/3)m^3 + m^2 n$ operations plus the cost of the sparse Cholesky factorization of Q, and the m sparse forward and backward substitutions. If $n \gg m$ and Q is sparse, the overall cost can be much smaller than solving $Hv = -g$ by a dense method.

9.29 *Exploiting structure in two-stage optimization.* This exercise continues exercise 4.64, which describes optimization with recourse, or two-stage optimization. Using the notation and assumptions in exercise 4.64, we assume in addition that the cost function f is a twice differentiable function of (x, z), for each scenario $i = 1, \ldots, S$.

Explain how to efficiently compute the Newton step for the problem of finding the optimal policy. How does the approximate flop count for your method compare to that of a generic method (which exploits no structure), as a function of S, the number of scenarios?

Solution. The problem to be solved is just

$$\text{minimize} \quad F(x) = \textstyle\sum_{i=1}^{S} \pi_i f(x, z_i, i),$$

which is convex since for each i, $f(x, z, i)$ is convex in (x, z_i), and $\pi_i \geq 0$.

Now let's see how to compute the Newton step efficiently. The Hessian of F has the block-arrow form

$$\nabla^2 F = \begin{bmatrix} \nabla^2_{x,x} F & \nabla^2_{x,z_1} F & \nabla^2_{x,z_2} F & \cdots & \nabla^2_{z_S,x} F \\ \nabla^2_{x,z_1} F^T & \nabla^2_{z_1,z_1} F & 0 & \cdots & 0 \\ \nabla^2_{x,z_2} F^T & 0 & \nabla^2_{z_2,z_2} F & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \nabla^2_{x,z_S} F^T & 0 & 0 & \cdots & \nabla^2_{z_S,z_S} F \end{bmatrix},$$

which we can exploit to compute the Newton step efficiently. First, let's see what happens if we don't exploit this structure. We need to solve the set of $n + Sq$ (symmetric, positive definite) linear equations $\nabla^2 F \Delta_{\text{nt}} = -\nabla F$, so the cost is around $(1/3)(n + Sq)^3$ flops. As a function of the number of scenarios, this grows like S^3.

Now let's exploit the structure to compute Δ_{nt}. We do this by using elimination, eliminating the bottom right block of size $Sq \times Sq$. This block is block diagonal, with S blocks of size $q \times q$, This situation is described on page 677 of the text. The overall complexity is

$$\begin{aligned} & (2/3)Sq^3 + 2nSq^2 + 2n^2Sq + 2n^2Sq + (2/3)n^3 \\ = \quad & \left((2/3)q^3 + 2nq^2 + 2n^2q + 2n^2q\right) S + (2/3)n^3, \end{aligned}$$

which grows *linearly* in S.

Here are the explicit details of how we can exploit structure to solve a block arrow, positive definite symmetric, system of equations:

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1N} \\ A_{12}^T & A_{22} & 0 & \cdots & 0 \\ A_{13}^T & 0 & A_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{1N}^T & 0 & 0 & \cdots & A_{NN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}.$$

We eliminate x_j, for $j = 2, \ldots, N$, to obtain

$$x_j = A_{jj}^{-1}(b_j - A_{1j}^T x_1), \quad j = 2, \ldots, N.$$

The first block equation becomes

$$\left(A_{11} - \sum_{j=2}^{N} A_{1j} A_{jj}^{-1} A_{1j}^T\right) x_1 = b_1 - \sum_{j=2}^{N} A_{1j} A_{jj}^{-1} b_j.$$

We'll solve this equation to find x_1, and then use the equations above to find $x_2, \ldots, x_N$. To do this we first carry out a Cholesky factorization of $A_{22}, \ldots, A_{NN}$, and then compute $A_{22}^{-1}A_{12}^T, \ldots, A_{NN}^{-1}A_{1N}^T$, and $A_{22}^{-1}b_2, \ldots, A_{NN}^{-1}b_N$, by back substitution. We then form the righthand side of the equations above, and the lefthand matrix, which is the Schur complement. We then solve these equations via Cholesky factorization and back substitution.

Numerical experiments

9.30 *Gradient and Newton methods.* Consider the unconstrained problem

$$\text{minimize} \quad f(x) = -\sum_{i=1}^m \log(1 - a_i^T x) - \sum_{i=1}^n \log(1 - x_i^2),$$

with variable $x \in \mathbf{R}^n$, and $\mathbf{dom}\, f = \{x \mid a_i^T x < 1,\ i = 1, \ldots, m,\ |x_i| < 1,\ i = 1, \ldots, n\}$. This is the problem of computing the analytic center of the set of linear inequalities

$$a_i^T x \leq 1, \quad i = 1, \ldots, m, \qquad |x_i| \leq 1, \quad i = 1, \ldots, n.$$

Note that we can choose $x^{(0)} = 0$ as our initial point. You can generate instances of this problem by choosing a_i from some distribution on $\mathbf{R}^n$.

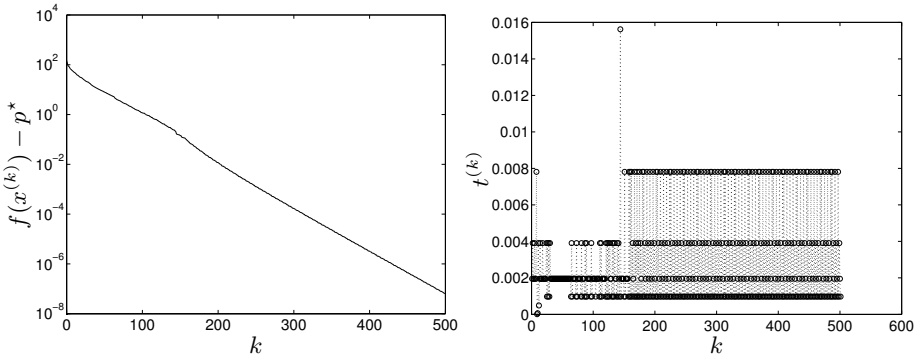
(a) Use the gradient method to solve the problem, using reasonable choices for the backtracking parameters, and a stopping criterion of the form $\|\nabla f(x)\|_2 \leq \eta$. Plot the objective function and step length versus iteration number. (Once you have determined $p^\star$ to high accuracy, you can also plot $f - p^\star$ versus iteration.) Experiment with the backtracking parameters α and β to see their effect on the total number of iterations required. Carry these experiments out for several instances of the problem, of different sizes.

(b) Repeat using Newton's method, with stopping criterion based on the Newton decrement λ^2. Look for quadratic convergence. You do not have to use an efficient method to compute the Newton step, as in exercise 9.27; you can use a general purpose dense solver, although it is better to use one that is based on a Cholesky factorization.

Hint. Use the chain rule to find expressions for $\nabla f(x)$ and $\nabla^2 f(x)$.

Solution.

(a) *Gradient method.* The figures show the function values and step lengths versus iteration number for an example with $m = 200$, $n = 100$. We used $\alpha = 0.01$, $\beta = 0.5$, and exit condition $\|\nabla f(x^{(k)})\|_2 \leq 10^{-3}$.



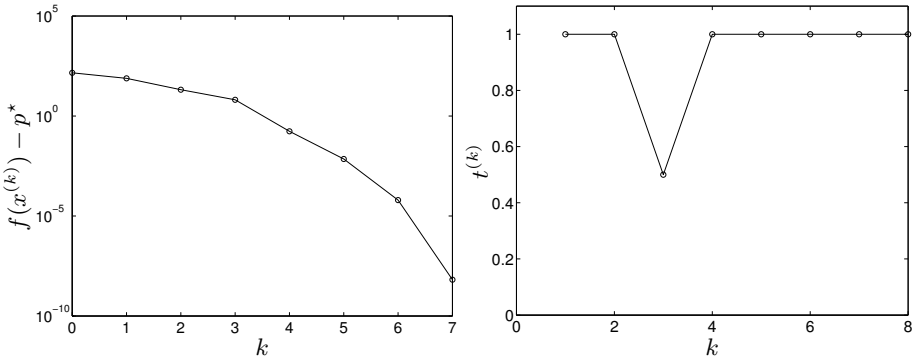
The following is a Matlab implementation.

```
ALPHA = 0.01;
BETA = 0.5;
MAXITERS = 1000;
GRADTOL = 1e-3;

x = zeros(n,1);
for iter = 1:MAXITERS
   val = -sum(log(1-A*x)) - sum(log(1+x)) - sum(log(1-x));
   grad =  A'*(1./(1-A*x)) - 1./(1+x) + 1./(1-x);
   if norm(grad) < GRADTOL, break; end;
```

```
   v = -grad;
   fprime = grad'*v;
   t = 1;  while ((max(A*(x+t*v)) >= 1) | (max(abs(x+t*v)) >= 1)),
      t = BETA*t;
   end;
   while ( -sum(log(1-A*(x+t*v))) - sum(log(1-(x+t*v).^2)) >  ...
           val + ALPHA*t*fprime )
      t = BETA*t;
   end;
   x = x+t*v;
end;
```

(b) *Newton method.* The figures show the function values and step lengths versus iteration number for the same example. We used $\alpha = 0.01$, $\beta = 0.5$, and exit condition $\lambda(x^{(k)})^2 \leq 10^{-8}$.



The following is a Matlab implementation.

```
ALPHA = 0.01;
BETA = 0.5;
MAXITERS = 1000;
NTTOL  = 1e-8;

x = zeros(n,1);
for iter = 1:MAXITERS
   val = -sum(log(1-A*x)) - sum(log(1+x)) - sum(log(1-x));
   d = 1./(1-A*x);
   grad =  A'*d - 1./(1+x) + 1./(1-x);
   hess =  A'*diag(d.^2)*A + diag(1./(1+x).^2 + 1./(1-x).^2);
   v = -hess\grad;
   fprime = grad'*v;
   if abs(fprime) < NTTOL, break; end;
   t = 1;  while ((max(A*(x+t*v)) >= 1) | (max(abs(x+t*v)) >= 1)),
      t = BETA*t;
   end;
   while ( -sum(log(1-A*(x+t*v))) - sum(log(1-(x+t*v).^2)) >  ...
           val + ALPHA*t*fprime )
      t = BETA*t;
   end;
   x = x+t*v;
end;
```

9.31 *Some approximate Newton methods.* The cost of Newton's method is dominated by the cost of evaluating the Hessian $\nabla^2 f(x)$ and the cost of solving the Newton system. For large

problems, it is sometimes useful to replace the Hessian by a positive definite approximation that makes it easier to form and solve for the search step. In this problem we explore some common examples of this idea.

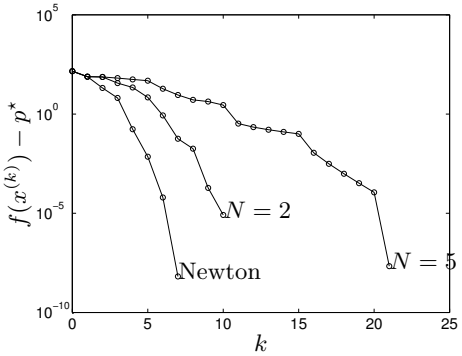
For each of the approximate Newton methods described below, test the method on some instances of the analytic centering problem described in exercise 9.30, and compare the results to those obtained using the Newton method and gradient method.

(a) *Re-using the Hessian.* We evaluate and factor the Hessian only every N iterations, where $N > 1$, and use the search step $\Delta x = -H^{-1}\nabla f(x)$, where H is the last Hessian evaluated. (We need to evaluate and factor the Hessian once every N steps; for the other steps, we compute the search direction using back and forward substitution.)

(b) *Diagonal approximation.* We replace the Hessian by its diagonal, so we only have to evaluate the n second derivatives $\partial^2 f(x)/\partial x_i^2$, and computing the search step is very easy.

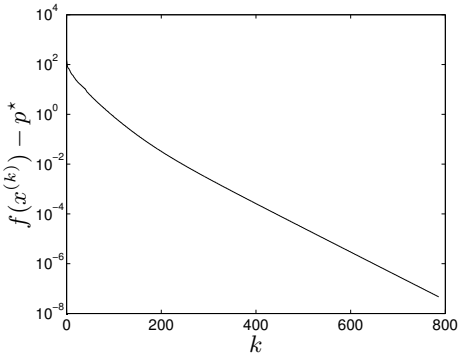
Solution.

(a) The figure shows the function value versus iteration number (for the same example as in the solution of exercise 9.30), for $N = 1$ (*i.e.*, Newton's method), $N = 2$, and $N = 5$.



We see that the speed of convergence deteriorates rapidly as N increases.

(b) The figure shows the function value versus iteration number (for the same example as in the solution of exercise 9.30), for a diagonal approximation of the Hessian. The experiment shows that the algorithm converges very much like the gradient method.



9.32 *Gauss-Newton method for convex nonlinear least-squares problems.* We consider a (nonlinear) least-squares problem, in which we minimize a function of the form

$$f(x) = \frac{1}{2}\sum_{i=1}^{m} f_i(x)^2,$$

where f_i are twice differentiable functions. The gradient and Hessian of f at x are given by

$$\nabla f(x) = \sum_{i=1}^m f_i(x)\nabla f_i(x), \qquad \nabla^2 f(x) = \sum_{i=1}^m \left(\nabla f_i(x)\nabla f_i(x)^T + f_i(x)\nabla^2 f_i(x)\right).$$

We consider the case when f is convex. This occurs, for example, if each f_i is either nonnegative and convex, or nonpositive and concave, or affine.

The *Gauss-Newton method* uses the search direction

$$\Delta x_{\rm gn} = -\left(\sum_{i=1}^m \nabla f_i(x)\nabla f_i(x)^T\right)^{-1}\left(\sum_{i=1}^m f_i(x)\nabla f_i(x)\right).$$

(We assume here that the inverse exists, *i.e.*, the vectors $\nabla f_1(x), \ldots, \nabla f_m(x)$ span $\mathbf{R}^n$.) This search direction can be considered an approximate Newton direction (see exercise 9.31), obtained by dropping the second derivative terms from the Hessian of f.

We can give another simple interpretation of the Gauss-Newton search direction $\Delta x_{\rm gn}$. Using the first-order approximation $f_i(x+v) \approx f_i(x) + \nabla f_i(x)^T v$ we obtain the approximation

$$f(x+v) \approx \frac{1}{2}\sum_{i=1}^m (f_i(x) + \nabla f_i(x)^T v)^2.$$

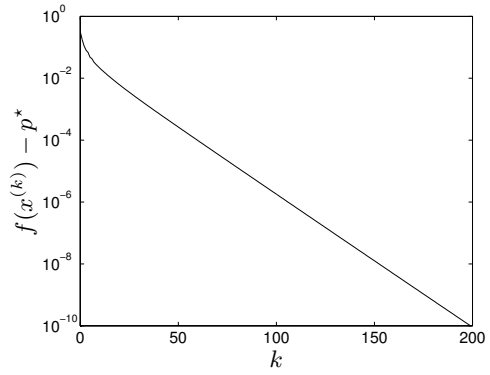
The Gauss-Newton search step $\Delta x_{\rm gn}$ is precisely the value of v that minimizes this approximation of f. (Moreover, we conclude that $\Delta x_{\rm gn}$ can be computed by solving a linear least-squares problem.)

Test the Gauss-Newton method on some problem instances of the form

$$f_i(x) = (1/2)x^T A_i x + b_i^T x + 1,$$

with $A_i \in \mathbf{S}^n_{++}$ and $b_i^T A_i^{-1} b_i \leq 2$ (which ensures that f is convex).

Solution. We generate random $A_i \in \mathbf{S}^n_{++}$, random b_i, and scale A_i and b_i so that $b_i^T A_i^{-1} b_i = 2$. We take $n = 50$, $m = 100$. The figure shows a typical convergence plot.



We note that the Gauss-Newton method converges linearly, and much more slowly than Newton's method (which for this example converged in 2 iterations).

This was to be expected. From the interpretation of the Gauss-Newton method as an approximate Newton method, we expect that it works well if the second term in the expression for the Hessian is small compared to the first term, *i.e.*, if either $\nabla^2 f_i$ is small (f_i is nearly linear), or f_i is small. For this test example neither of these conditions was satisfied.

Chapter 10

Equality constrained minimization

Exercises

Equality constrained minimization

10.1 *Nonsingularity of the KKT matrix.* Consider the KKT matrix

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix},$$

where $P \in \mathbf{S}^n_+$, $A \in \mathbf{R}^{p\times n}$, and $\mathbf{rank}\, A = p < n$.

(a) Show that each of the following statements is equivalent to nonsingularity of the KKT matrix.

- $\mathcal{N}(P) \cap \mathcal{N}(A) = \{0\}$.
- $Ax = 0$, $x \neq 0 \Longrightarrow x^T P x > 0$.
- $F^T P F \succ 0$, where $F \in \mathbf{R}^{n\times(n-p)}$ is a matrix for which $\mathcal{R}(F) = \mathcal{N}(A)$.
- $P + A^T Q A \succ 0$ for some $Q \succeq 0$.

(b) Show that if the KKT matrix is nonsingular, then it has exactly n positive and p negative eigenvalues.

Solution.

(a) The second and third are clearly equivalent. To see this, if $Ax = 0$, $x \neq 0$, then x must have the form $x = Fz$, where $z \neq 0$. Then we have $x^T P x = z^T F^T P F z$.

Similarly, the first and second are equivalent. To see this, if $x \in \mathcal{N}(A) \cap \mathcal{N}(P)$, $x \neq 0$, then $Ax = 0$, $x \neq 0$, but $x^T P x = 0$, contradicting the second statement. Conversely, suppose the second statement fails to hold, *i.e.*, there is an x with $Ax = 0$, $x \neq 0$, but $x^T P x = 0$. Since $P \succeq 0$, we conclude $Px = 0$, *i.e.*, $x \in \mathcal{N}(P)$, which contradicts the first statement.

Finally, the second and fourth statements are equivalent. If the second holds then the last statement holds with $Q = I$. If the last statement holds for some $Q \succeq 0$ then it holds for all $Q \succ 0$, and therefore the second statement holds.

Now let's show that the four statements are equivalent to nonsingularity of the KKT matrix. First suppose that x satisfies $Ax = 0$, $Px = 0$, and $x \neq 0$. Then

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ 0 \end{bmatrix} = 0,$$

which shows that the KKT matrix is singular.

Now suppose the KKT matrix is singular, *i.e.*, there are x, z, not both zero, such that

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} = 0.$$

This means that $Px + A^T z = 0$ and $Ax = 0$, so multiplying the first equation on the left by x^T, we find $x^T P x + x^T A^T z = 0$. Using $Ax = 0$, this reduces to $x^T P x = 0$, so we have $Px = 0$ (using $P \succeq 0$). This contradicts (a), unless $x = 0$. In this case, we must have $z \neq 0$. But then $A^T z = 0$ contradicts $\mathbf{rank}\, A = p$.

(b) From part (a), $P + A^T A \succ 0$. Therefore there exists a nonsingular matrix $R \in \mathbf{R}^{n\times n}$ such that

$$R^T(P + A^T A)R = I.$$

Let $AR = U\Sigma V_1^T$ be the singular value decomposition of AR, with $U \in \mathbf{R}^{p\times p}$, $\Sigma = \mathbf{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbf{R}^{p\times p}$ and $V_1 \in \mathbf{R}^{n\times p}$. Let $V_2 \in \mathbf{R}^{n\times(n-p)}$ be such that

$$V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}$$

is orthogonal, and define

$$S = \begin{bmatrix} \Sigma & 0 \end{bmatrix} \in \mathbf{R}^{p\times n}.$$

We have $AR = USV^T$, so

$$V^TR^T(P + A^TA)RV = V^TR^TPRV + S^TS = I.$$

Therefore $V^TR^TPRV = I - S^TS$ is diagonal. We denote this matrix by Λ:

$$\Lambda = V^TR^TPRV = \mathbf{diag}(1 - \sigma_1^2, \ldots, 1 - \sigma_p^2, 1, \ldots, 1).$$

Applying a congruence transformation to the KKT matrix gives

$$\begin{bmatrix} V^TR^T & 0 \\ 0 & U^T \end{bmatrix} \begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} RV & 0 \\ 0 & U \end{bmatrix} = \begin{bmatrix} \Lambda & S^T \\ S & 0 \end{bmatrix},$$

and the inertia of the KKT matrix is equal to the inertia of the matrix on the right. Applying a permutation to the matrix on the right gives a block diagonal matrix with n diagonal blocks

$$\begin{bmatrix} \lambda_i & \sigma_i \\ \sigma_i & 0 \end{bmatrix}, \quad i = 1, \ldots, p, \qquad \lambda_i = 1, \quad i = p+1, \ldots, n.$$

The eigenvalues of the 2×2-blocks are

$$\frac{\lambda_i \pm \sqrt{\lambda_i^2 + 4\sigma_i^2}}{2},$$

i.e., one eigenvalue is positive and one is negative. We conclude that there are $p + (n - p) = n$ positive eigenvalues and p negative eigenvalues.

10.2 *Projected gradient method.* In this problem we explore an extension of the gradient method to equality constrained minimization problems. Suppose f is convex and differentiable, and $x \in \mathbf{dom}\, f$ satisfies $Ax = b$, where $A \in \mathbf{R}^{p\times n}$ with $\mathbf{rank}\, A = p < n$. The Euclidean projection of the negative gradient $-\nabla f(x)$ on $\mathcal{N}(A)$ is given by

$$\Delta x_{\rm pg} = \operatorname*{argmin}_{Au=0} \|-\nabla f(x) - u\|_2.$$

(a) Let (v, w) be the unique solution of

$$\begin{bmatrix} I & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}.$$

Show that $v = \Delta x_{\rm pg}$ and $w = \operatorname{argmin}_y \|\nabla f(x) + A^Ty\|_2$.

(b) What is the relation between the projected negative gradient $\Delta x_{\rm pg}$ and the negative gradient of the reduced problem (10.5), assuming $F^TF = I$?

(c) The *projected gradient method* for solving an equality constrained minimization problem uses the step $\Delta x_{\rm pg}$, and a backtracking line search on f. Use the results of part (b) to give some conditions under which the projected gradient method converges to the optimal solution, when started from a point $x^{(0)} \in \mathbf{dom}\, f$ with $Ax^{(0)} = b$.

Solution.

(a) These are the optimality conditions for the problem

$$\begin{array}{ll} \text{minimize} & \|-\nabla f(x) - u\|_2^2 \\ \text{subject to} & Au = 0. \end{array}$$

(b) If $F^TF = I$, then $\Delta x_{\rm pg} = -F\nabla\tilde{f}(Fz+\hat{x})$ where $x = Fz+\hat{x}$.

(c) By part (b), running the projected gradient from $x^{(0)}$ is the same as running the gradient method on the reduced problem, assuming $F^TF = I$. This means that the projected gradient method converges if the initial sublevel set $\{x \mid f(x) \leq f(x^{(0)}), Ax = b\}$ is closed and the objective function of the reduced or eliminated problem, $f(Fz+\hat{x})$ is strongly convex.

Newton's method with equality constraints

10.3 *Dual Newton method.* In this problem we explore Newton's method for solving the dual of the equality constrained minimization problem (10.1). We assume that f is twice differentiable, $\nabla^2 f(x) \succ 0$ for all $x \in \mathbf{dom}\, f$, and that for each $\nu \in \mathbf{R}^p$, the Lagrangian $L(x,\nu) = f(x) + \nu^T(Ax-b)$ has a unique minimizer, which we denote $x(\nu)$.

(a) Show that the dual function g is twice differentiable. Find an expression for the Newton step for the dual function g, evaluated at ν, in terms of f, ∇f, and $\nabla^2 f$, evaluated at $x = x(\nu)$. You can use the results of exercise 3.40.

(b) Suppose there exists a K such that

$$\left\| \begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix}^{-1} \right\|_2 \leq K$$

for all $x \in \mathbf{dom}\, f$. Show that g is strongly concave, with $\nabla^2 g(\nu) \preceq -(1/K)I$.

Solution.

(a) By the results of exercise 3.40, g is twice differentiable, with

$$\begin{aligned} \nabla g(\nu) &= A\nabla f^*(-A^T\nu) = Ax(\nu) \\ \nabla^2 g(\nu) &= -A\nabla^2 f^*(-A^T\nu)A^T = -A\nabla^2 f(x(\nu))^{-1}A^T. \end{aligned}$$

Therefore the Newton step for g at ν is given by

$$\Delta\nu_{\rm nt} = (A\nabla^2 f(x(\nu))^{-1}A^T)^{-1}Ax(\nu).$$

(b) Now suppose

$$\left\| \begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix}^{-1} \right\|_2 \leq K$$

for all $x \in x(S) = \{x(\nu) \mid \nu \in S\}$. Using the expression

$$\begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix}^{-1} = \begin{bmatrix} H^{-1} & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} H^{-1}A^T \\ -I \end{bmatrix} (AH^{-1}A^T)^{-1} \begin{bmatrix} AH^{-1} & -I \end{bmatrix}$$

(with $H = \nabla^2 f(x)$), we see that

$$\begin{aligned} \left\| \begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix}^{-1} \right\|_2 &\geq \sup_{\|u\|_2=1} \left\| \begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ u \end{bmatrix} \right\|_2 \\ &= \sup_{\|u\|_2=1} \left\| \begin{bmatrix} H^{-1}A^T \\ -I \end{bmatrix} (AH^{-1}A^T)^{-1}u \right\|_2 \\ &\geq \sup_{\|u\|_2=1} \left\| (AH^{-1}A^T)^{-1}u \right\|_2 \\ &= \|(AH^{-1}A^T)^{-1}\|_2 \end{aligned}$$

for all $x \in x(S)$, which implies that

$$\nabla^2 g(\nu) = -A\nabla^2 f(x(\nu))^{-1} A^T \preceq -(1/K)I$$

for all $\nu \in S$.

10.4 *Strong convexity and Lipschitz constant of the reduced problem.* Suppose f satisfies the assumptions given on page 529. Show that the reduced objective function $\tilde{f}(z) = f(Fz+\hat{x})$ is strongly convex, and that its Hessian is Lipschitz continuous (on the associated sublevel set $\tilde{S}$). Express the strong convexity and Lipschitz constants of $\tilde{f}$ in terms of K, M, L, and the maximum and minimum singular values of F.

Solution. In the text it was shown that $\nabla^2 \tilde{f}(z) \succeq mI$, for $m = \sigma_{\min}(F)^2/(K^2 M)$. Here we establish the other properties of $\tilde{f}$. We have

$$\|\nabla^2 \tilde{f}(z)\|_2 = \|F^T \nabla^2 f(Fz+\hat{x})F\|_2 \leq \|F\|_2^2 M,$$

using $\|\nabla f^2(x)\|_2 \leq M$. Therefore we have $\nabla^2 \tilde{f}(z) \preceq \tilde{M}I$, with $\tilde{M} = \|F\|_2^2 M$.

Now we establish that $\nabla^2 \tilde{f}(z)$ satisfies a Lipschitz condition:

$$\begin{aligned}
\|\nabla^2 \tilde{f}(z) - \nabla^2 \tilde{f}(w)\|_2 &= \|F^T(\nabla^2 f(Fz+\hat{x}) - \nabla^2 f(Fw+\hat{x}))F\|_2 \\
&\leq \|F\|_2^2 \, \|\nabla^2 f(Fz+\hat{x}) - \nabla^2 f(Fw+\hat{x})\|_2 \\
&\leq L\|F\|_2^2 \, \|F(z-w)\|_2 \\
&\leq L\|F\|_2^3 \, \|z-w\|_2.
\end{aligned}$$

Thus, $\nabla^2 \tilde{f}(z)$ satisfies a Lipschitz condition with constant $\tilde{L} = L\|F\|_2^3$.

10.5 *Adding a quadratic term to the objective.* Suppose $Q \succeq 0$. The problem

$$\begin{array}{ll} \text{minimize} & f(x) + (Ax-b)^T Q(Ax-b) \\ \text{subject to} & Ax = b \end{array}$$

is equivalent to the original equality constrained optimization problem (10.1). Is the Newton step for this problem the same as the Newton step for the original problem?

Solution. The Newton step of the new problem satisfies

$$\begin{bmatrix} H + A^T Q A & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ w \end{bmatrix} = \begin{bmatrix} -g - 2A^T Q A x + 2A^T Q b \\ 0 \end{bmatrix}.$$

From the second equation, $A\Delta x = 0$. Therefore,

$$\begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ w \end{bmatrix} = \begin{bmatrix} -g - 2A^T Q A x + 2A^T Q b \\ 0 \end{bmatrix},$$

and

$$\begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \tilde{w} \end{bmatrix} = \begin{bmatrix} -g \\ 0 \end{bmatrix},$$

where $\hat{w} = w + 2QAx - 2Qb$. We conclude that the Newton steps are equal. Note the connection to the last statement in exercise 10.1.

10.6 *The Newton decrement.* Show that (10.13) holds, *i.e.*,

$$f(x) - \inf\{\widehat{f}(x+v) \mid A(x+v) = b\} = \lambda(x)^2/2.$$

Solution. The Newton step is defined by

$$\begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ w \end{bmatrix} = \begin{bmatrix} -g \\ 0 \end{bmatrix}.$$

We first note that this implies that $\Delta x^T H \Delta x = -g^T \Delta x$. Therefore

$$\begin{aligned}\hat{f}(x+\Delta x) &= f(x) + g^T v + (1/2) v^T H v \\ &= f(x) + (1/2) g^T v \\ &= f(x) - (1/2)\lambda(x)^2.\end{aligned}$$

Infeasible start Newton method

10.7 *Assumptions for infeasible start Newton method.* Consider the set of assumptions given on page 536.

(a) Suppose that the function f is closed. Show that this implies that the norm of the residual, $\|r(x,\nu)\|_2$, is closed.

Solution. Recall from §A.3.3 that a continuous function h with an open domain is closed if $h(y)$ tends to infinity as y approaches the boundary of $\mathbf{dom}\, h$. The function $\|r\|_2 : \mathbf{R}^n \times \mathbf{R}^p \to \mathbf{R}$ is clearly continuous (by assumption f is continuously differentiable), and its domain, $\mathbf{dom}\, f \times \mathbf{R}^p$, is open. Now suppose f is closed. Consider a sequence of points $(x^{(k)}, \nu^{(k)}) \in \mathbf{dom}\, \|r\|_2$ converging to a limit $(\bar{x}, \bar{\nu}) \in \mathbf{bd}\, \mathbf{dom}\, \|r\|_2$. Then $\bar{x} \in \mathbf{bd}\, \mathbf{dom}\, f$, and since f is closed, $f(x^{(k)}) \to \infty$, hence $\|\nabla f(x^{(k)})\|_2 \to \infty$, and $\|r(x^{(k)}, \nu^{(k)})\|_2 \to \infty$. We conclude that $\|r\|_2$ is closed.

(b) Show that Dr satisfies a Lipschitz condition if and only if $\nabla^2 f$ does.

Solution. First suppose that $\nabla^2 f$ satisfies the Lipschitz condition

$$\|\nabla^2 f(x) - \nabla^2 f(\tilde{x})\|_2 \leq L\|x - \tilde{x}\|_2$$

for $x,\ \tilde{x} \in S$. From this we get a Lipschitz condition on Dr: If $y = (x, \nu) \in S$, and $\tilde{y} = (\tilde{x}, \tilde{\nu}) \in S$, then

$$\begin{aligned}\|Dr(y) - Dr(\tilde{y})\|_2 &= \left\| \begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} - \begin{bmatrix} \nabla^2 f(\tilde{x}) & A^T \\ A & 0 \end{bmatrix} \right\|_2 \\ &= \left\| \begin{bmatrix} \nabla^2 f(x) - \nabla^2 f(\tilde{x}) & 0 \\ 0 & 0 \end{bmatrix} \right\|_2 \\ &= \|\nabla^2 f(x) - \nabla^2 f(\tilde{x})\|_2 \\ &\leq L\|x - \tilde{x}\|_2 \\ &\leq L\|y - \tilde{y}\|_2.\end{aligned}$$

To show the converse, suppose that Dr satisfies a Lipschitz condition with constant L. Using the equations above this means that

$$\|Dr(y) - Dr(\tilde{y})\|_2 = \|\nabla^2 f(x) - \nabla^2 f(\tilde{x})\|_2 \leq L\|y - \tilde{y}\|_2$$

for all y and $\tilde{y}$. In particular, taking $\nu = \tilde{\nu} = 0$, this reduces to a Lipschitz condition for $\nabla^2 f$, with constant L.

10.8 *Infeasible start Newton method and initially satisfied equality constraints.* Suppose we use the infeasible start Newton method to minimize $f(x)$ subject to $a_i^T x = b_i$, $i = 1, \ldots, p$.

(a) Suppose the initial point $x^{(0)}$ satisfies the linear equality $a_i^T x = b_i$. Show that the linear equality will remain satisfied for future iterates, *i.e.*, if $a_i^T x^{(k)} = b_i$ for all k.

(b) Suppose that one of the equality constraints becomes satisfied at iteration k, *i.e.*, we have $a_i^T x^{(k-1)} \neq b_i$, $a_i^T x^{(k)} = b_i$. Show that at iteration k, *all* the equality constraints are satisfied.

Solution.
Follows easily from

$$r^{(k)} = \left(\prod_{i=0}^{k-1} (1 - t^{(i)}) \right) r^{(0)}.$$

10.9 *Equality constrained entropy maximization.* Consider the equality constrained entropy maximization problem

$$\begin{array}{ll} \text{minimize} & f(x) = \sum_{i=1}^{n} x_i \log x_i \\ \text{subject to} & Ax = b, \end{array} \tag{10.42}$$

with $\mathbf{dom}\, f = \mathbf{R}^n_{++}$ and $A \in \mathbf{R}^{p \times n}$. We assume the problem is feasible and that $\mathbf{rank}\, A = p < n$.

(a) Show that the problem has a unique optimal solution $x^\star$.

(b) Find A, b, and feasible $x^{(0)}$ for which the sublevel set

$$\{x \in \mathbf{R}^n_{++} \mid Ax = b,\ f(x) \leq f(x^{(0)})\}$$

is *not* closed. Thus, the assumptions listed in §10.2.4, page 529, are not satisfied for some feasible initial points.

(c) Show that the problem (10.42) satisfies the assumptions for the infeasible start Newton method listed in §10.3.3, page 536, for any feasible starting point.

(d) Derive the Lagrange dual of (10.42), and explain how to find the optimal solution of (10.42) from the optimal solution of the dual problem. Show that the dual problem satisfies the assumptions listed in §10.2.4, page 529, for *any* starting point.

The results of part (b), (c), and (d) do not mean the standard Newton method will fail, or that the infeasible start Newton method or dual method will work better in practice. It only means our convergence analysis for the standard Newton method does not apply, while our convergence analysis does apply to the infeasible start and dual methods. (See exercise 10.15.)

Solution.

(a) If $p^\star$ is not attained, then either $p^\star$ is attained asymptotically, as x goes to infinity, or in the limit as x goes to $x^\star$, where $x^\star \succeq 0$ with one or more zero components.

The first possibility cannot occur because the entropy goes to infinity as x goes to infinity. The second possibility can also be ruled out, because by assumption the problem is feasible. Suppose $\tilde{x} \succ 0$ and $A\tilde{x} = b$. Define $v = \tilde{x} - x$ and

$$g(t) = \sum_{i=1}^{n} (x_i^\star + tv_i) \log(x_i^\star + tv_i)$$

for $t > 0$. The derivative is

$$g'(t) = \sum_{i=1}^{n} v_i (1 + \log(x_i^\star + tv_i).$$

Now if $x_i^\star = 0$ for some i, then $v_i > 0$, and hence $\lim_{t \to 0} g(t) = -\infty$. This means it is impossible that $\lim_{t \to 0} g(t) = p^\star$.

(b) Consider

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

and starting point $x^{(0)} = (1/20, 9/10, 1/20)$. Eliminating x_2 and x_3 from the two equations

$$2x_1 + x_2 = 1, \qquad x_1 + x_2 + x_3 = 1$$

gives $x_2 = 1 - 2x_1$, $x_3 = x_1$. For $x^{(0)} = (1/20, 9/10, 1/20)$, with $f(x^{(0)}) = -0.3944$ we have $f(x_1, 1 - 2x_1, x_1) \leq f(x^{(0)})$ if and only if $1/20 \leq x_1 < 0.5$, which is not closed.

(c) The dual problem is

$$\text{maximize} \quad -b^T\nu - \textstyle\sum_{i=1}^{p} \exp(-1 - a_i^T\nu)$$

where a_i is the ith column of A. The dual objective function is closed with domain $\mathbf{R}^p$.

(d) We have

$$r(x, \nu) = (\nabla f(x) + A^T\nu, Ax - b)$$

where

$$\nabla f(x)_i = \log x_i + 1, \quad i = 1, \ldots, n.$$

We show that $\|r\|_2$ is a closed function.

Clearly $\|r\|_2$ is continuous on its domain, $\mathbf{R}^n_{++} \times \mathbf{R}^p$.

Suppose $(x^{(k)}, \nu^{(k)})$, $k = 1, 2, \ldots$ is a sequence of points converging to a point $(\bar{x}, \bar{\nu}) \in \mathbf{bd}\,\mathbf{dom}\,\|r\|_2$. We have $\bar{x}_i = 0$ for at least one i, so $\log x_i^{(k)} + 1 + a_i^T\nu^{(k)} \to -\infty$. Hence $\|r(x^{(k)}, \nu^{(k)})\|_2 \to \infty$.

We conclude that r satisfies the sublevel set condition for arbitrary starting points.

10.10 *Bounded inverse derivative condition for strongly convex-concave game.* Consider a convex-concave game with payoff function f (see page 541). Suppose $\nabla^2_{uu} f(u, v) \succeq mI$ and $\nabla^2_{vv} f(u, v) \preceq -mI$, for all $(u, v) \in \mathbf{dom}\, f$. Show that

$$\|Dr(u, v)^{-1}\|_2 = \|\nabla^2 f(u, v)^{-1}\|_2 \leq 1/m.$$

Solution. Let

$$H = \nabla^2 f(u, v) = \begin{bmatrix} D & E \\ E^T & -F \end{bmatrix}$$

where $D \in \mathbf{S}^p$, $F \in \mathbf{S}^q$, $E \in \mathbf{R}^{p\times q}$, and assume $D \succeq mI$, $F \succeq mI$. Let $D^{-1/2}EF^{-1/2} = U_1\Sigma V_1^T$ be the singular value decomposition ($U_1 \in \mathbf{R}^{p\times r}$, $V_1 \in \mathbf{R}^{q\times r}$, $\Sigma \in \mathbf{R}^{r\times r}$, $r = \mathbf{rank}\, E$). Choose $U_2 \in \mathbf{R}^{p\times(p-r)}$ and $V_2 \in \mathbf{R}^{q\times(q-r)}$, so that $U_2^TU_2 = I$, $U_2^TU_1 = 0$ and $V_2^TV_2 = I$, $V_2^TV_1 = 0$. Define

$$U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \in \mathbf{R}^{p\times p}, \qquad V = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \in \mathbf{R}^{p\times p}, \qquad S = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \in \mathbf{R}^{p\times q}.$$

With these definitions we have $D^{-1/2}EF^{-1/2} = USV^T = U_1\Sigma V_1^T$, and

$$H = \begin{bmatrix} D^{1/2}U & 0 \\ 0 & F^{1/2}V \end{bmatrix} \begin{bmatrix} I & S \\ S^T & -I \end{bmatrix} \begin{bmatrix} U^TD^{1/2} & 0 \\ 0 & V^TF^{1/2} \end{bmatrix}.$$

Therefore

$$H^{-1} = \begin{bmatrix} U^TD^{-1/2} & 0 \\ 0 & V^TF^{-1/2} \end{bmatrix} \begin{bmatrix} I & S \\ S^T & -I \end{bmatrix}^{-1} \begin{bmatrix} D^{-1/2}U & 0 \\ 0 & F^{-1/2}V \end{bmatrix}$$

and $\|H^{-1}\|_2 \leq (1/m)\|G^{-1}\|_2$, where

$$G = \begin{bmatrix} I & S \\ S^T & -I \end{bmatrix}.$$

We can permute the rows and columns of G so that it is block diagonal with $\max\{p,q\}-r$ scalar diagonal blocks with value 1, $\max\{p,q\}-r$ scalar diagonal blocks with value -1, and r diagonal blocks of the form

$$\begin{bmatrix} 1 & \sigma_i \\ \sigma_i & -1 \end{bmatrix}.$$

Note that

$$\begin{aligned} \begin{bmatrix} 1 & \sigma_i \\ \sigma_i & -1 \end{bmatrix}^{-1} &= \frac{1}{1+\sigma_i^2}\begin{bmatrix} 1 & \sigma_i \\ \sigma_i & -1 \end{bmatrix} \\ &= \begin{bmatrix} 1/\sqrt{1+\sigma_i^2} & \sigma_i/\sqrt{1+\sigma_i^2} \\ \sigma_i/\sqrt{1+\sigma_i^2} & -1/\sqrt{1+\sigma_i^2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{1+\sigma_i^2}} & 0 \\ 0 & \frac{1}{\sqrt{1+\sigma_i^2}} \end{bmatrix}, \end{aligned}$$

and therefore

$$\left\| \begin{bmatrix} 1 & \sigma_i \\ \sigma_i & -1 \end{bmatrix}^{-1} \right\|_2 = \frac{1}{\sqrt{1+\sigma_i^2}}.$$

If $r \neq \max\{p,q\}$, then $\|G^{-1}\|_2 = 1$. Otherwise

$$\|G^{-1}\|_2 = \max_i (1+\sigma_i^2)^{-1/2} \leq 1.$$

In conclusion,

$$\|H^{-1}\|_2 \leq (1/m)\|G^{-1}\|_2 \leq 1/m.$$

Implementation

10.11 Consider the resource allocation problem described in example 10.1. You can assume the f_i are strongly convex, *i.e.*, $f_i''(z) \geq m > 0$ for all z.

(a) Find the computational effort required to compute a Newton step for the reduced problem. Be sure to exploit the special structure of the Newton equations.

(b) Explain how to solve the problem via the dual. You can assume that the conjugate functions f_i^*, and their derivatives, are readily computable, and that the equation $f_i'(x) = \nu$ is readily solved for x, given ν. What is the computational complexity of finding a Newton step for the dual problem?

(c) What is the computational complexity of computing a Newton step for the resource allocation problem? Be sure to exploit the special structure of the KKT equations.

Solution.

(a) The reduced problem is

$$\text{minimize} \quad \tilde{f}(z) = \textstyle\sum_{i=1}^{n-1} f_i(z_i) + f_n(b - \mathbf{1}^T z).$$

The Newton equation is

$$(D + d\mathbf{1}\mathbf{1}^T)\Delta z = g.$$

where D is diagonal with $D_{ii} = f_i''(z_i)$ and $d = f_n''(b - \mathbf{1}^T z)$.
The cost of computing Δz is order n, if we use the matrix inversion lemma.

(b) The dual problem is

$$\text{maxmize} \quad g(\nu) = -b\nu - \textstyle\sum_{i=1}^n f_i^*(-\nu).$$

From the solution of exercise 10.3,

$$g'(\nu) = \mathbf{1}^T x(\nu), \qquad g''(\nu) = -\mathbf{1}^T \nabla^2 f(x(\nu))^{-1}\mathbf{1},$$

where $\nabla^2 f(x(\nu))$ is diagonal with diagonal elements $f_i''(x_i(\nu))$. The cost of forming $g''(\nu)$ is order n.

(c) The KKT system is

$$\begin{bmatrix} D & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ w \end{bmatrix} = \begin{bmatrix} -g \\ 0 \end{bmatrix},$$

which can be solved in order n operations by eliminating Δx.

10.12 Describe an efficient way to compute the Newton step for the problem

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(X^{-1}) \\ \text{subject to} & \mathbf{tr}(A_i X) = b_i, \quad i = 1, \ldots, p \end{array}$$

with domain $\mathbf{S}_{++}^n$, assuming p and n have the same order of magnitude. Also derive the Lagrange dual problem and give the complexity of finding the Newton step for the dual problem.

Solution.

(a) The gradient of f_0 is $\nabla f_0(X) = -X^{-2}$. The optimality conditions are

$$-X^{-2} + \sum_{i=1}^p w_i A_i = 0, \qquad \mathbf{tr}(A_i X) = b_i, \quad i = 1, \ldots, p.$$

Linearizing around X gives

$$\begin{aligned} -X^{-2} + X^{-1}\Delta X X^{-2} + X^{-2}\Delta X X^{-1} + \sum_{i=1}^p w_i A_i &= 0 \\ \mathbf{tr}(A_i(X + \Delta X)) &= b_i, \quad i = 1, \ldots, p, \end{aligned}$$

i.e.,

$$\begin{aligned} \Delta X X^{-1} + X^{-1}\Delta X + \sum_{i=1}^p w_i (X A_i X) &= I \\ \mathbf{tr}(A_i \Delta X) &= b_i - \mathbf{tr}(A_i X), \quad i = 1, \ldots, p. \end{aligned}$$

We can eliminate ΔX from the first equation by solving $p+1$ Lyapunov equations:

$$\Delta X = Y_0 + \sum_{i=1}^n w_i Y_i$$

where

$$Y_0 X^{-1} + X^{-1} Y_0 = I, \qquad Y_i X^{-1} + X^{-1} Y_i = X A_i X, \quad i = 1, \ldots, p.$$

Substituting in the second equation gives

$$Hw = g,$$

with $H_i = \mathbf{tr}(Y_i Y_j)$, $i, j = 1, \ldots, p$.

The cost is order pn^3 for computing Y_i, $p^2 n^2$ for constructing H and p^3 for solving the equations.

(b) The conjugate of f_0 is given in exercise 3.37:

$$f_0^*(Y) = -2\,\mathbf{tr}(-Y)^{-1/2}, \qquad \mathbf{dom}\, f_0^* = -\mathbf{S}_{++}^n.$$

The dual problem is

$$\text{maximize} \quad g(\nu) = -b^T\nu + 2\,\mathbf{tr}(\textstyle\sum_{i=1}^p \nu_i A_i)^{1/2}$$

with domain $\{\nu \in \mathbf{R}^p \mid \sum_i \nu_i A_i \succeq 0\}$. The optimality conditions are

$$2\,\mathbf{tr}(A_i \nabla g_0(Z)) = b_i, \quad i = 1, \ldots, p, \qquad Z = \sum_{i=1}^p \nu_i A_i, \tag{10.12.A}$$

where $g_0(Z) = \mathbf{tr}\, Z^{1/2}$.
The gradient of g_0 is $\nabla\,\mathbf{tr}(Z^{1/2}) = (1/2)Z^{-1/2}$, as can be seen as follows. Suppose $Z \succ 0$. For small symmetric ΔZ, $(Z + \Delta Z)^{1/2} \approx Z^{1/2} + \Delta Y$ where

$$\begin{aligned} Z + \Delta Z &= (Z^{1/2} + \Delta Y)^2 \\ &\approx Z + Z^{1/2}\Delta Y + \Delta Y Z^{1/2}, \end{aligned}$$

i.e., ΔY is the solution of the Lyapunov equation $\Delta Z = Z^{1/2}\Delta Y + \Delta Y Z^{1/2}$. In particular,

$$\mathbf{tr}\,\Delta Y = \mathbf{tr}(Z^{-1/2}\Delta Z) - \mathbf{tr}(Z^{-1/2}\Delta Y Z^{1/2}) = \mathbf{tr}(Z^{-1/2}\Delta Z) - \mathbf{tr}\,\Delta Y,$$

i.e., $\mathbf{tr}\,\Delta Y = (1/2)\,\mathbf{tr}(Z^{-1/2}\Delta Z)$. Therefore

$$\begin{aligned} \mathbf{tr}(Z + \Delta Z)^{1/2} &\approx \mathbf{tr}\, Z^{1/2} + \mathbf{tr}\,\Delta Y \\ &= \mathbf{tr}\, Z^{1/2} + (1/2)\,\mathbf{tr}(Z^{-1/2}\Delta Z), \end{aligned}$$

i.e., $\nabla_Z\,\mathbf{tr}\, Z^{1/2} = (1/2)Z^{-1/2}$.
We can therefore simplify the optimality conditions (10.12.A) as

$$\mathbf{tr}(A_i Z^{-1/2}) = b_i, \quad i = 1, \ldots, p, \qquad Z = \sum_{i=1}^p \nu_i A_i,$$

Linearizing around Z, ν gives

$$\begin{aligned} \mathbf{tr}(A_i Z^{-1/2}) + \mathbf{tr}(A_i \Delta Y) &= b_i, \quad i = 1, \ldots, p \\ Z^{1/2}\Delta Y + \Delta Y Z^{1/2} &= \Delta Z \\ Z + \Delta Z &= \sum_{i=1}^p \nu_i A_i + \sum_{i=1}^p \Delta\nu_i A_i, \end{aligned}$$

i.e., after a simplification

$$\begin{aligned} \mathbf{tr}(A_i \Delta Y) &= b_i - \mathbf{tr}(A_i Z^{-1/2}), \quad i = 1, \ldots, p \\ Z^{1/2}\Delta Y + \Delta Y Z^{1/2} - \sum_i \Delta\nu_i A_i &= -Z + \sum_{i=1}^p \nu_i A_i. \end{aligned}$$

These equations have the same form as the Newton equations in part (a) (with X replaced with $Z^{-1/2}$).

10.13 *Elimination method for computing Newton step for convex-concave game.* Consider a convex-concave game with payoff function $f : \mathbf{R}^p \times \mathbf{R}^q \to \mathbf{R}$ (see page 541). We assume that f is *strongly convex-concave*, *i.e.*, for all $(u, v) \in \mathbf{dom}\, f$ and some $m > 0$, we have $\nabla^2_{uu} f(u, v) \succeq mI$ and $\nabla^2_{vv} f(u, v) \preceq -mI$.

(a) Show how to compute the Newton step using Cholesky factorizations of $\nabla^2_{uu} f(u, v)$ and $-\nabla^2 f_{vv}(u, v)$. Compare the cost of this method with the cost of using an LDL$^{\mathrm{T}}$ factorization of $\nabla f(u, v)$, assuming $\nabla^2 f(u, v)$ is dense.

(b) Show how you can exploit diagonal or block diagonal structure in $\nabla^2_{uu} f(u, v)$ and/or $\nabla^2_{vv} f(u, v)$. How much do you save, if you assume $\nabla^2_{uv} f(u, v)$ is dense?

Solution.

(a) We use the notation

$$\nabla^2 f(u, v) = \begin{bmatrix} D & E \\ E^T & -F \end{bmatrix},$$

with $D \in \mathbf{S}^p_{++}$, $E \in \mathbf{R}^{p\times q}$, $F \in \mathbf{S}^p_{++}$, and consider the cost of solving a system of the form

$$\begin{bmatrix} D & E \\ E^T & -F \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = - \begin{bmatrix} g \\ h \end{bmatrix}.$$

We have two equations

$$Dv + Ew = -g, \qquad E^T v - Fw = -h.$$

From the first equation we solve for v to obtain

$$v = -D^{-1}(g + Ew).$$

Substituting in the other equation gives $E^T D^{-1}(g + Ew) + Fw = h$, so

$$w = (F + E^T D^{-1} E)^{-1}(h - E^T D^{-1} g).$$

We can implement this method using the Cholesky factorization as follows.

- Factor $D = L_1 L_1^T$ ($(1/3)p^3$ flops).
- Compute $y = D^{-1}g$, and $Y = L_1^{-1}E$ ($p^2(2 + q) \approx p^2 q$ flops).
- Compute $S = F + Y^T Y$ (pq^2 flops) and $d = h - E^T y$ ($2pq$ flops)
- Solve $Sw = d$ via Cholesky factorization ($(1/3)q^3$ flops).

The total number of flops (ignoring lower order terms) is

$$(1/3)p^3 + p^2 q + pq^2 + (1/3)q^3 = (1/3)(p + q)^3.$$

Eliminating w would give the same result.
The cost is the same as using LDL$^{\mathrm{T}}$ factorization of $\nabla f(u, v)$, *i.e.*, $(1/3)(p + q)^3$.
A matrix of the form of $\nabla^2 f(u, v)$ above is called a quasidefinite matrix. It has the special property that it has an LDL$^{\mathrm{T}}$ factorization with diagonal D: with the same notation as above,

$$\begin{bmatrix} D & E \\ E^T & -F \end{bmatrix} = \begin{bmatrix} L_1 & 0 \\ Y^T & L_2 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} \begin{bmatrix} L_1^T & Y \\ 0 & L_2^T \end{bmatrix}.$$

(b) Assume f is the cost of factoring D, and s is the cost of solving a system $Dx = b$ after factoring. Then the cost of the algorithm is

$$f + p^2(s/2) + pq^2 + (1/3)q^3.$$

Numerical experiments

10.14 *Log-optimal investment.* Consider the log-optimal investment problem described in exercise 4.60. Use Newton's method to compute the solution, with the following problem data: there are $n = 3$ assets, and $m = 4$ scenarios, with returns

$$p_1 = \begin{bmatrix} 2 \\ 1.3 \\ 1 \end{bmatrix}, \qquad p_2 = \begin{bmatrix} 2 \\ 0.5 \\ 1 \end{bmatrix}, \qquad p_3 = \begin{bmatrix} 0.5 \\ 1.3 \\ 1 \end{bmatrix}, \qquad p_4 = \begin{bmatrix} 0.5 \\ 0.5 \\ 1 \end{bmatrix}.$$

The probabilities of the four scenarios are given by $\pi = (1/3, 1/6, 1/3, 1/6)$.

Solution. Eliminating x_3 using the equality constraint $x_1 + x_2 + x_3 = 1$ gives the equivalent problem

$$\begin{array}{ll} \text{maximize} & (1/3)\log(1 + x_1 + 0.3x_2) + (1/6)\log(1 + x_1 - 0.5x_2) \\ & + (1/3)\log(1 - 0.5x_1 + 0.3x_2) + (1/6)\log(1 - 0.5x_1 - 0.5x_2), \end{array}$$

with two variables x_1 and x_2. The solution is

$$x_1 = 0.4973, \qquad x_2 = 0.1994, \qquad x_3 = 0.7021.$$

We use Newton's method with backtracking parameters $\alpha = 0.01$, $\beta = 0.5$, stopping criterion $\lambda < 10^{-8}$, and initial point $x = (0, 0, 1)$. The algorithm converges in five steps, with no backtracking necessary.

10.15 *Equality constrained entropy maximization.* Consider the equality constrained entropy maximization problem

$$\begin{array}{ll} \text{minimize} & f(x) = \sum_{i=1}^n x_i \log x_i \\ \text{subject to} & Ax = b, \end{array}$$

with $\mathbf{dom}\, f = \mathbf{R}^n_{++}$ and $A \in \mathbf{R}^{p\times n}$, with $p < n$. (See exercise 10.9 for some relevant analysis.)

Generate a problem instance with $n = 100$ and $p = 30$ by choosing A randomly (checking that it has full rank), choosing $\hat{x}$ as a random positive vector (*e.g.*, with entries uniformly distributed on $[0, 1]$) and then setting $b = A\hat{x}$. (Thus, $\hat{x}$ is feasible.)

Compute the solution of the problem using the following methods.

(a) *Standard Newton method.* You can use initial point $x^{(0)} = \hat{x}$.

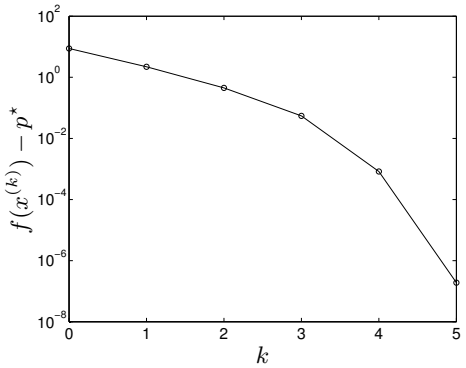
(b) *Infeasible start Newton method.* You can use initial point $x^{(0)} = \hat{x}$ (to compare with the standard Newton method), and also the initial point $x^{(0)} = \mathbf{1}$.

(c) *Dual Newton method*, *i.e.*, the standard Newton method applied to the dual problem.

Verify that the three methods compute the same optimal point (and Lagrange multiplier). Compare the computational effort per step for the three methods, assuming relevant structure is exploited. (Your implementation, however, does not need to exploit structure to compute the Newton step.)

Solution.

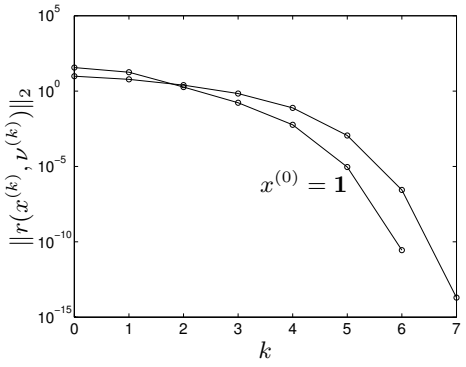
(a) *Standard Newton method.* A typical convergence plot is shown below.



The Matlab code is as follows.

```
MAXITERS = 100;
ALPHA = 0.01;
BETA = 0.5;
NTTOL = 1e-7;
x = x0;
for iter=1:MAXITERS
   val = x'*log(x);
   grad = 1+log(x);
   hess = diag(1./x);
   sol = -[hess A'; A zeros(p,p)] \ [grad; zeros(p,1)];
   v = sol(1:n);
   fprime = grad'*v;
   if (abs(fprime) < NTTOL), break; end;
   t=1;
   while (min(x+t*v) <= 0),  t = BETA*t;  end;
   while ((x+t*v)'*log(x+t*v) >= val + t*ALPHA*fprime), t=BETA*t; end;
   x = x + t*v;
end;
```

(b) *Infeasible start Newton method.* The figure shows the norm of the residual versus $(\nabla(f(x)) + A^T\nu, Ax - b)$ verus iteration number for the same example. The lower curve uses starting point $x^{(0)} = \mathbf{1}$; the other curve uses the same starting point as in part (a).




```
MAXITERS = 100;
ALPHA = 0.01;
BETA = 0.5;
```

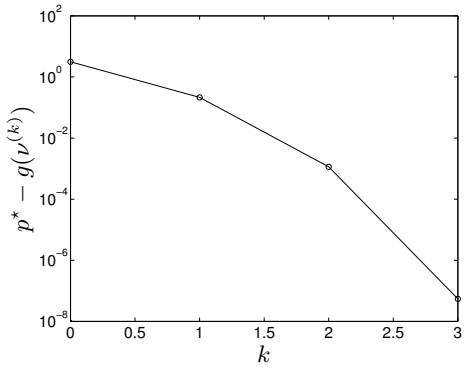
```
RESTOL = 1e-7;

x=x0;  nu=zeros(p,1);
for i=1:MAXITERS
  r = [1+log(x)+A'*nu;  A*x-b];  resdls = [resdls, norm(r)];
  sol = -[diag(1./x) A';  A zeros(p,p)] \ r;
  Dx = sol(1:n);  Dnu = sol(n+[1:p]);
  if (norm(r) < RESTOL), break; end;
  t=1;
  while (min(x+t*Dx) <= 0),  t = BETA*t;  end;
  while norm([1+log(x+t*Dx)+A'*(nu+Dnu);  A*(x+Dx)-b])  > ...
         (1-ALPHA*t)*norm(r),  t=BETA*t;  end;
  x = x + t*Dx;  nu = nu + t*Dnu;
end;
```

(c) *Dual Newton method.* The dual problem is

$$\text{maximize} \quad -b^T\nu - \sum_{i=1}^n e^{-a_i^T\nu-1}$$

where a_i is the ith column of A. The figure shows the dual function value versus iteration number for the same example.




```
MAXITERS = 100;
ALPHA = 0.01;
BETA = 0.5;
NTTOL = 1e-8;
nu = zeros(p,1);
for i=1:MAXITERS
   val = b'*nu + sum(exp(-A'*nu-1));
   grad = b - A*exp(-A'*nu-1);
   hess = A*diag(exp(-A'*nu-1))*A';
   v = -hess\grad;
   fprime = grad'*v;
   if (abs(fprime) < NTTOL), break; end;
   t=1;
   while (b'*(nu+t*v) + sum(exp(-A'*(nu+t*v)-1)) > ...
          val + t*ALPHA*fprime), t = BETA*t;  end;
   nu = nu + t*v;
end;
```

The computational effort is the same for each method. In the standard and infeasible start Newton methods, we solve equations with coefficient matrix

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix},$$

where

$$\nabla^2 f(x) = \mathbf{diag}(x)^{-1}.$$

Block elimination reduces the equation to one with coefficient matrix $A\,\mathbf{diag}(x)A^T$.
In the dual method, we solve an equation with coefficient matrix

$$-\nabla^2 g(\nu) = ADA^T$$

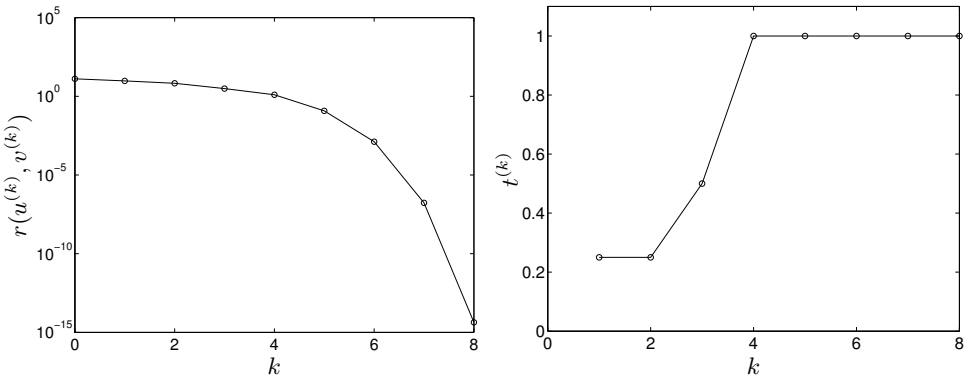
where D is diagonal with $D_{ii} = e^{-a_i^T\nu-1}$.
In all three methods, the main computation in each iteration is therefore the solution of a linear system of the form

$$A^T DAv = -g$$

where D is diagonal with positive diagonal elements.

10.16 *Convex-concave game.* Use the infeasible start Newton method to solve convex-concave games of the form (10.32), with randomly generated data. Plot the norm of the residual and step length versus iteration. Experiment with the line search parameters and initial point (which must satisfy $\|u\|_2 < 1$, $\|v\|_2 < 1$, however).

Solution. See figure 10.5 and the two figures below.



A Matlab implementation, using the notation

$$f(x,y) = x^T Ay + c^T x + d^T y - \log(1 - x^T x) + \log(1 - y^T y),$$

is as follows.

```
BETA = .5;
ALPHA = .01;
MAXITERS = 100;
x = .01*ones(n,1);
y = .01*ones(n,1);

for iters =1:MAXITERS
   r = [ A*y + (2/(1-x'*x))*x + c;  A'*x - (2/(1-y'*y))*y + d];
   if (norm(r) < 1e-8), break;  end;
   Dr = [ ((2/(1-x'*x))*eye(n) + (4/(1-x'*x)^2)*x*x')   A ;
          A'  (-(2/(1-y'*y))*eye(n) - (4/(1-y'*y)^2)*y*y')];
   step = -Dr\r;  dx = step(1:n);  dy = step(n+[1:n]);
   t = 1;
   newx = x+t*dx;  newy = y+t*dy;
   while ((norm(newx) >= 1) | (norm(newy) >= 1)),
      t = BETA*t;  newx = x+t*dx;   newy = y+t*dy;
   end;
```

```
   newr = [ A*newy + (2/(1-newx'*newx))*newx + c;
            A'*newx - (2/(1-newy'*newy))*newy + d ];
   while (norm(newr) > (1-ALPHA*t)*norm(r))
       t = BETA*t;  newx = x+t*dx;   newy = y+t*dy;
       newr = [ A*newy + (2/(1-newx'*newx))*newx + c;
                A'*newx - (2/(1-newy'*newy))*newy + d];
   end;
   x = x+t*dx;  y = y+t*dy;
end;
```

Chapter 11

Interior-point methods

Exercises

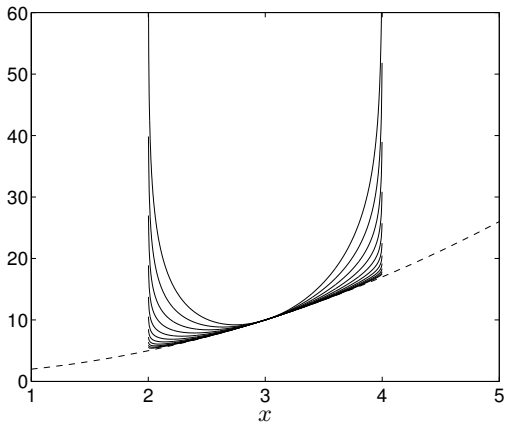
The barrier method

11.1 *Barrier method example.* Consider the simple problem

$$\begin{array}{ll} \text{minimize} & x^2 + 1 \\ \text{subject to} & 2 \leq x \leq 4, \end{array}$$

which has feasible set $[2, 4]$, and optimal point $x^\star = 2$. Plot f_0, and $tf_0 + \phi$, for several values of $t > 0$, versus x. Label $x^\star(t)$.

Solution. The figure shows the function $f_0 + (1/t)\widehat{I}$ for $f_0(x) = x^2 + 1$, with barrier function $\widehat{I}(x) = -\log(x - 2) - \log(4 - x)$, for $t = 10^{-1}$, $10^{-0.8}$, $10^{-0.6}$, $\ldots$, $10^{0.8}$, 10. The inner curve corresponds to $t = 0.1$, and the outer curve corresponds to $t = 10$. The objective function is shown as a dashed curve.



11.2 What happens if the barrier method is applied to the LP

$$\begin{array}{ll} \text{minimize} & x_2 \\ \text{subject to} & x_1 \leq x_2, \quad 0 \leq x_2, \end{array}$$

with variable $x \in \mathbf{R}^2$?

Solution. We need to minimize

$$tf_0(x) + \phi(x) = tx_2 - \log(x_2 - x_1) - \log x_2,$$

but this function is unbounded below (letting $x_1 \to -\infty$), so the first centering step never converges.

11.3 *Boundedness of centering problem.* Suppose the sublevel sets of (11.1),

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & Ax = b, \end{array}$$

are bounded. Show that the sublevel sets of the associated centering problem,

$$\begin{array}{ll} \text{minimize} & tf_0(x) + \phi(x) \\ \text{subject to} & Ax = b, \end{array}$$

are bounded.

Solution. Suppose a sublevel set $\{x \mid tf_0(x)+\phi(x) \leq M\}$ is unbounded. Let $\{x+sv \mid s \geq 0\}$, with with $v \neq 0$ and x strictly feasible, be a ray contained in the sublevel set. We have $A(x+sv) = b$ for all $s \geq 0$ (*i.e.*, $Ax = b$ and $Av = 0$), and $f_i(x+sv) < 0$, $i = 1, \ldots, m$. By assumption, the sublevel sets of (11.1) are bounded, which is only possible if $f_0(x + sv)$ increases with s for sufficiently large s. Without loss of generality, we can choose x such that $\nabla f_0(x)^T v > 0$.
We have

$$\begin{aligned} M &\geq tf_0(x+sv) - \sum_{i=1}^m \log(-f_i(x+sv)) \\ &\geq tf_0(x) + st\nabla f_0(x)^T v - \sum_{i=1}^m \log(-f_i(x) - s\nabla f_i(x)^T v) \end{aligned}$$

for all $s \geq 0$. This is impossible since $\nabla f_0(x)^T v > 0$.

11.4 *Adding a norm bound to ensure strong convexity of the centering problem.* Suppose we add the constraint $x^T x \leq R^2$ to the problem (11.1):

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & Ax = b \\ & x^T x \leq R^2. \end{array}$$

Let $\tilde{\phi}$ denote the logarithmic barrier function for this modified problem. Find $a > 0$ for which $\nabla^2(tf_0(x) + \tilde{\phi}(x)) \succeq aI$ holds, for all feasible x.
Solution. Let ϕ denote the logarithmic barrier of the original problem. The constraint $x^T x \leq R^2$ adds the term $-\log(R^2 - x^T x)$ to the logarithmic barrier, so we have

$$\begin{aligned} \nabla^2(tf_0 + \tilde{\phi}) &= \nabla^2(tf_0 + \phi) + \frac{2}{R^2 - x^T x} I + \frac{4}{(R^2 - x^T x)^2} xx^T \\ &\succeq \nabla^2(tf_0 + \phi) + (2/R^2) I \\ &\succeq (2/R^2) I, \end{aligned}$$

so we can take $m = 2/R^2$.

11.5 *Barrier method for second-order cone programming.* Consider the SOCP (without equality constraints, for simplicity)

$$\begin{array}{ll} \text{minimize} & f^T x \\ \text{subject to} & \|A_i x + b_i\|_2 \leq c_i^T x + d_i, \quad i = 1, \ldots, m. \end{array} \tag{11.63}$$

The constraint functions in this problem are not differentiable (since the Euclidean norm $\|u\|_2$ is not differentiable at $u = 0$) so the (standard) barrier method cannot be applied. In §11.6, we saw that this SOCP can be solved by an extension of the barrier method that handles generalized inequalities. (See example 11.8, page 599, and page 601.) In this exercise, we show how the standard barrier method (with scalar constraint functions) can be used to solve the SOCP.
We first reformulate the SOCP as

$$\begin{array}{ll} \text{minimize} & f^T x \\ \text{subject to} & \|A_i x + b_i\|_2^2 / (c_i^T x + d_i) \leq c_i^T x + d_i, \quad i = 1, \ldots, m \\ & c_i^T x + d_i \geq 0, \quad i = 1, \ldots, m. \end{array} \tag{11.64}$$

The constraint function

$$f_i(x) = \frac{\|A_i x + b_i\|_2^2}{c_i^T x + d_i} - c_i^T x - d_i$$

is the composition of a quadratic-over-linear function with an affine function, and is twice differentiable (and convex), provided we define its domain as $\mathbf{dom}\, f_i = \{x \mid c_i^T x + d_i > 0\}$. Note that the two problems (11.63) and (11.64) are not exactly equivalent. If $c_i^T x^\star + d_i = 0$ for some i, where $x^\star$ is the optimal solution of the SOCP (11.63), then the reformulated problem (11.64) is not solvable; $x^\star$ is not in its domain. Nevertheless we will see that the barrier method, applied to (11.64), produces arbitrarily accurate suboptimal solutions of (11.64), and hence also for (11.63).

(a) Form the log barrier ϕ for the problem (11.64). Compare it to the log barrier that arises when the SOCP (11.63) is solved using the barrier method for generalized inequalities (in §11.6).

(b) Show that if $tf^T x + \phi(x)$ is minimized, the minimizer $x^\star(t)$ is $2m/t$-suboptimal for the problem (11.63). It follows that the standard barrier method, applied to the reformulated problem (11.64), solves the SOCP (11.63), in the sense of producing arbitrarily accurate suboptimal solutions. This is the case even though the optimal point $x^\star$ need not be in the domain of the reformulated problem (11.64).

Solution.

(a) The log barrier ϕ for the problem (11.64) is

$$\begin{aligned} &-\textstyle\sum_{i=1}^m \log\left(c_i^T x + d_i - \frac{\|A_i x + b_i\|_2^2}{c_i^T x + d_i}\right) - \sum_{i=1}^m \log(c_i^T x + d_i) \\ &= -\textstyle\sum_{i=1}^m \log\left((c_i^T x + d_i)^2 - \|A_i x + b_i\|_2^2\right) \end{aligned}$$

The log barrier for the SOCP (11.63), using the generalized logarithm for the second-order cone given in §11.6, is

$$-\sum_{i=1}^m \log\left((c_i^T x + d_i)^2 - \|A_i x + b_i\|_2^2\right),$$

which is exactly the same. The log barriers are the *same*.

(b) The centering problems are the same, and the central paths are the same. The proof is identical to the derivation in example 11.8.

11.6 *General barriers.* The log barrier is based on the approximation $-(1/t)\log(-u)$ of the indicator function $\widehat{I}_-(u)$ (see §11.2.1, page 563). We can also construct barriers from other approximations, which in turn yield generalizations of the central path and barrier method. Let $h : \mathbf{R} \to \mathbf{R}$ be a twice differentiable, closed, increasing convex function, with $\mathbf{dom}\, h = -\mathbf{R}_{++}$. (This implies $h(u) \to \infty$ as $u \to 0$.) One such function is $h(u) = -\log(-u)$; another example is $h(u) = -1/u$ (for $u < 0$).
Now consider the optimization problem (without equality constraints, for simplicity)

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m, \end{array}$$

where f_i are twice differentiable. We define the *h-barrier* for this problem as

$$\phi_h(x) = \sum_{i=1}^m h(f_i(x)),$$

with domain $\{x \mid f_i(x) < 0,\ i = 1, \ldots, m\}$. When $h(u) = -\log(-u)$, this is the usual logarithmic barrier; when $h(u) = -1/u$, ϕ_h is called the *inverse barrier*. We define the *h-central path* as

$$x^\star(t) = \operatorname{argmin}\, tf_0(x) + \phi_h(x),$$

where $t > 0$ is a parameter. (We assume that for each t, the minimizer exists and is unique.)

(a) Explain why $tf_0(x) + \phi_h(x)$ is convex in x, for each $t > 0$.

(b) Show how to construct a dual feasible λ from $x^\star(t)$. Find the associated duality gap.

(c) For what functions h does the duality gap found in part (b) depend only on t and m (and no other problem data)?

Solution.

(a) The composition rules show that $tf_0(x) + \phi_h(x)$ is convex in x, since h is increasing and convex, and f_i are convex.

(b) The minimizer of $tf_0(x)+\phi_h(x)$, $z = x^\star(t)$, satisfies $t\nabla f_0(z)+\nabla\phi(z) = 0$. Expanding this we get

$$t\nabla f_0(z) + \sum_{i=1}^m h'(f_i(z))\nabla f_i(z) = 0.$$

This shows that z minimizes the Lagrangian $f_0(z) + \sum_{i=1}^m \lambda_i f_i(z)$, for

$$\lambda_i = h'(f_i(z))/t, \quad i = 1, \ldots, m.$$

The associated dual function value is

$$g(\lambda) = f_0(z) + \sum_{i=1}^m \lambda_i f_i(z) = f_0(z) + \sum_{i=1}^m h'(f_i(z))f_i(z)/t,$$

so the duality gap is

$$(1/t)\sum_{i=1}^m h'(f_i(z))(-f_i(z)).$$

(c) The only way the expression above does not depend on problem data (except t and m) is for $h'(u)(-u)$ to be constant. This means $h'(u) = a/(-u)$ for some constant a, so $h(u) = -a\log(-u) + b$, for some constant b. Since h must be convex and increasing, we need $a > 0$. Thus, h gives rise to a scaled, offset log barrier. In particular, the central path associated with h is the same as for the standard log barrier.

11.7 *Tangent to central path.* This problem concerns $dx^\star(t)/dt$, which gives the tangent to the central path at the point $x^\star(t)$. For simplicity, we consider a problem without equality constraints; the results readily generalize to problems with equality constraints.

(a) Find an explicit expression for $dx^\star(t)/dt$. *Hint.* Differentiate the centrality equations (11.7) with respect to t.

(b) Show that $f_0(x^\star(t))$ decreases as t increases. Thus, the objective value in the barrier method decreases, as the parameter t is increased. (We already know that the duality gap, which is m/t, decreases as t increases.)

Solution.

(a) Differentiating the centrality equation yields

$$\nabla f_0(x^\star(t)) + \left(t\nabla^2 f_0(x^\star(t)) + \nabla^2\phi(x^\star(t))\right)\frac{dx^\star}{dt} = 0.$$

Thus, the tangent to the central path at $x^\star(t)$ is given by

$$\frac{dx^\star}{dt} = -\left(t\nabla^2 f_0(x^\star(t)) + \nabla^2\phi(x^\star(t))\right)^{-1}\nabla f_0(x^\star(t)). \tag{11.7.A}$$

(b) We will show that $df_0(x^\star(t))/dt < 0$.

$$\begin{aligned}\frac{df_0(x^\star(t))}{dt} &= \nabla f_0(x^\star(t))^T \frac{dx^\star(t)}{dt} \\ &= -\nabla f_0(x^\star(t))^T \left(t\nabla^2 f_0(x^\star(t)) + \nabla^2\phi(x^\star(t))\right)^{-1} \nabla f_0(x^\star(t)) \\ &< 0.\end{aligned}$$

11.8 *Predictor-corrector method for centering problems.* In the standard barrier method, $x^\star(\mu t)$ is computed using Newton's method, starting from the initial point $x^\star(t)$. One alternative that has been proposed is to make an approximation or prediction $\widehat{x}$ of $x^\star(\mu t)$, and then start the Newton method for computing $x^\star(\mu t)$ from $\widehat{x}$. The idea is that this should reduce the number of Newton steps, since $\widehat{x}$ is (presumably) a better initial point than $x^\star(t)$. This method of centering is called a *predictor-corrector method*, since it first makes a *prediction* of what $x^\star(\mu t)$ is, then *corrects* the prediction using Newton's method.

The most widely used predictor is the first-order predictor, based on the tangent to the central path, explored in exercise 11.7. This predictor is given by

$$\widehat{x} = x^\star(t) + \frac{dx^\star(t)}{dt}(\mu t - t).$$

Derive an expression for the first-order predictor $\widehat{x}$. Compare it to the Newton update obtained, *i.e.*, $x^\star(t) + \Delta x_{\rm nt}$, where $\Delta x_{\rm nt}$ is the Newton step for $\mu t f_0(x) + \phi(x)$, at $x^\star(t)$. What can you say when the objective f_0 is linear? (For simplicity, you can consider a problem without equality constraints.)

Solution. The first-order predictor is, using the expression for $dx^\star/dt$ found in exercise 11.7,

$$\begin{aligned}\widehat{x} &= x^\star(t) + \frac{dx^\star(t)}{dt}(\mu t - t) \\ &= x^\star(t) - (\mu - 1)t\left(t\nabla^2 f_0(x^\star(t)) + \nabla^2\phi(x^\star(t))\right)^{-1} \nabla f_0(x^\star(t)).\end{aligned}$$

The Newton step for $\mu t f_0 + \phi$, at the point $x^\star(t)$, is given by

$$\begin{aligned}\Delta x_{\rm nt} &= -\left(\mu t\nabla^2 f_0(x^\star(t)) + \nabla^2\phi(x^\star(t))\right)^{-1} (\mu t\nabla f_0(x^\star(t)) + \nabla\phi(x^\star(t))) \\ &= -(\mu - 1)t\left(\mu t\nabla^2 f_0(x^\star(t)) + \nabla^2\phi(x^\star(t))\right)^{-1} \nabla f_0(x^\star(t)),\end{aligned}$$

where we use $t\nabla f_0(x^\star(t)) + \nabla\phi(x^\star(t)) = 0$. The Newton update is then

$$x^\star(t) + \Delta x_{\rm nt} = x^\star(t) - (\mu - 1)t\left(\mu t\nabla^2 f_0(x^\star(t)) + \nabla^2\phi(x^\star(t))\right)^{-1} \nabla f_0(x^\star(t)).$$

This is similar to, but not quite the same as, the first-order predictor.

Now let's consider the special case when f_0 is linear, say, $f_0(x) = c^T x$. Then the first-order predictor is given by

$$\widehat{x} = x^\star(t) - (\mu - 1)t\nabla^2\phi(x^\star(t))^{-1} c.$$

The Newton update is *exactly the same*. The Newton step for $\mu t f_0 + \phi$ at $x^\star$ is exactly the tangent to the central path. We conclude that when the objective is linear, the fancy sounding predictor-corrector method is exactly the same as the simple method of just starting Newton's method from the current point $x^\star(t)$.

11.9 *Dual feasible points near the central path.* Consider the problem

$$\begin{array}{ll}\text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m,\end{array}$$

with variable $x \in \mathbf{R}^n$. We assume the functions f_i are convex and twice differentiable. (We assume for simplicity there are no equality constraints.) Recall (from §11.2.2, page 565)

that $\lambda_i = -1/(tf_i(x^\star(t)))$, $i = 1, \ldots, m$, is dual feasible, and in fact, $x^\star(t)$ minimizes $L(x, \lambda)$. This allows us to evaluate the dual function for λ, which turns out to be $g(\lambda) = f_0(x^\star(t)) - m/t$. In particular, we conclude that $x^\star(t)$ is m/t-suboptimal.

In this problem we consider what happens when a point x is close to $x^\star(t)$, but not quite centered. (This would occur if the centering steps were terminated early, or not carried out to full accuracy.) In this case, of course, we cannot claim that $\lambda_i = -1/(tf_i(x))$, $i = 1, \ldots, m$, is dual feasible, or that x is m/t-suboptimal. However, it turns out that a slightly more complicated formula does yield a dual feasible point, provided x is close enough to centered.

Let $\Delta x_{\rm nt}$ be the Newton step at x of the centering problem

$$\text{minimize} \quad tf_0(x) - \textstyle\sum_{i=1}^m \log(-f_i(x)).$$

Define

$$\lambda_i = \frac{1}{-tf_i(x)} \left(1 + \frac{\nabla f_i(x)^T \Delta x_{\rm nt}}{-f_i(x)}\right), \quad i = 1, \ldots, m.$$

You will show that for small $\Delta x_{\rm nt}$ (*i.e.*, for x nearly centered), λ is dual feasible (*i.e.*, $\lambda \succeq 0$ and $L(x, \lambda)$ is bounded below).

In this case, the vector x *does not* minimize $L(x, \lambda)$, so there is no general formula for the dual function value $g(\lambda)$ associated with λ. (If we have an analytical expression for the dual objective, however, we can simply evaluate $g(\lambda)$.)

Hint. Use the results in exercise 3.41 to show that when $\Delta x_{\rm nt}$ is small enough, there exist x_0, x_1, $\ldots$, x_m such that

$$\begin{aligned} \nabla f_0(x_0) &= \nabla f_0(x) + \nabla^2 f_0(x)\Delta x_{\rm nt} \\ \nabla f_i(x_i) &= \nabla f_i(x) + (1/\lambda_i)\nabla^2 f_i(x)\Delta x_{\rm nt}, \quad i = 1, \ldots, m. \end{aligned}$$

This implies that

$$\nabla f_0(x_0) + \sum_{i=1}^m \lambda_i \nabla f_i(x_i) = 0.$$

Now use $f_i(z) \geq f_i(x_i) + \nabla f_i(x_i)^T(z - x_i)$, $i = 0, \ldots, m$, to derive a lower bound on $L(z, \lambda)$.

Solution. It is clear that $\lambda \succ 0$ for sufficiently small $\Delta x_{\rm nt}$. We need to show that $f_0 + \sum_i \lambda_i f_i$ is bounded below.

The Newton equations at x are

$$\begin{aligned} \nabla f_0(x) + \sum_{i=1}^m \frac{1}{-tf_i(x)} \nabla f_i(x) + \sum_{i=1}^m \frac{\nabla f_i(x)^T \Delta x_{\rm nt}}{tf_i(x)^2} \nabla f_i(x) \\ + \nabla^2 f_0(x)\Delta x_{\rm nt} + \sum_{i=1}^m \frac{1}{-tf_i(x)} \nabla^2 f_i(x)\Delta x_{\rm nt} = 0 \end{aligned}$$

i.e., using the above definition of λ,

$$\nabla f_0(x) + \sum_{i=1}^m \lambda_i \nabla f_i(x) + \nabla^2 f_0(x)\Delta x_{\rm nt} + \sum_{i=1}^m \frac{1}{-tf_i(x)} \nabla^2 f_i(x)\Delta x_{\rm nt} = 0.$$

Now, from the result in exercise 3.41, if $\Delta x_{\rm nt}$ is small enough, there exist x_0, x_1, $\ldots$, x_m such that

$$\nabla f_0(x_0) = \nabla f_0(x) + \nabla^2 f_0(x)\Delta x_{\rm nt},$$

and

$$\nabla f_i(x_i) = \nabla f_i(x) + (1/\lambda_i)\nabla^2 f_i(x)\Delta x_{\rm nt}, \quad i = 1, \ldots, m.$$

We can therefore write the Newton equation as

$$\nabla f_0(x_0) + \sum_{i=1}^m \lambda_i \nabla f_i(x_i) = 0.$$

Returning to the question of boundedness of $f_0 + \sum_i \lambda_i f_i$, we have

$$\begin{aligned} f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) &\geq f_0(x_0) + \nabla f_0(x_0)^T(x - x_0) + \sum_{i=1}^m \lambda_i (f_i(x_i) + \nabla f_i(x_i)^T(x - x_i)) \\ &= f_0(x_0) + \sum_i \lambda_i f_i(x_i) + \left(\nabla f_0(x_0) + \sum_{i=1}^m \lambda_i \nabla f_i(x_i)\right)^T x \\ &\quad - \nabla f_0(x_0)^T x_0 - \sum_i \lambda_i \nabla f_i(x_i)^T x_i \\ &= f_0(x_0) + \sum_i \lambda_i f_i(x_i) - \nabla f_0(x_0)^T x_0 - \sum_i \lambda_i \nabla f_i(x_i)^T x_i, \end{aligned}$$

which shows that $f_0 + \sum_i \lambda_i f_i$ is bounded below.

11.10 *Another parametrization of the central path.* We consider the problem (11.1), with central path $x^\star(t)$ for $t > 0$, defined as the solution of

$$\begin{array}{ll} \text{minimize} & tf_0(x) - \sum_{i=1}^m \log(-f_i(x)) \\ \text{subject to} & Ax = b. \end{array}$$

In this problem we explore another parametrization of the central path.
For $u > p^\star$, let $z^\star(u)$ denote the solution of

$$\begin{array}{ll} \text{minimize} & -\log(u - f_0(x)) - \sum_{i=1}^m \log(-f_i(x)) \\ \text{subject to} & Ax = b. \end{array}$$

Show that the curve defined by $z^\star(u)$, for $u > p^\star$, is the central path. (In other words, for each $u > p^\star$, there is a $t > 0$ for which $x^\star(t) = z^\star(u)$, and conversely, for each $t > 0$, there is an $u > p^\star$ for which $z^\star(u) = x^\star(t)$).

Solution. $z^\star(u)$ satisfies the optimality conditions

$$\frac{1}{u - f_0(z^\star(u))} \nabla f_0(z^\star(u)) + \sum_{i=1}^m \frac{1}{-f_i(z^\star(u))} \nabla f_i(z^\star(u)) + A^T\nu = 0$$

for some ν. We conclude that $z^\star(u) = x^\star(t)$ for

$$t = \frac{1}{u - f_0(z^\star(u))}.$$

Conversely, for each $t > 0$, $x^\star(t) = z^\star(u)$ with

$$u = \frac{1}{t} + f_0(x^\star(t)) > p^\star.$$

11.11 *Method of analytic centers.* In this problem we consider a variation on the barrier method, based on the parametrization of the central path described in exercise 11.10. For simplicity, we consider a problem with no equality constraints,

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m. \end{array}$$

The method of analytic centers starts with any strictly feasible initial point $x^{(0)}$, and any $u^{(0)} > f_0(x^{(0)})$. We then set

$$u^{(1)} = \theta u^{(0)} + (1-\theta) f_0(x^{(0)}),$$

where $\theta \in (0,1)$ is an algorithm parameter (usually chosen small), and then compute the next iterate as

$$x^{(1)} = z^\star(u^{(1)})$$

(using Newton's method, starting from $x^{(0)}$). Here $z^\star(s)$ denotes the minimizer of

$$-\log(s - f_0(x)) - \sum_{i=1}^m \log(-f_i(x)),$$

which we assume exists and is unique. This process is then repeated.

The point $z^\star(s)$ is the *analytic center* of the inequalities

$$f_0(x) \leq s, \quad f_1(x) \leq 0, \ldots, f_m(x) \leq 0,$$

hence the algorithm name.

Show that the method of centers works, *i.e.*, $x^{(k)}$ converges to an optimal point. Find a stopping criterion that guarantees that x is ϵ-suboptimal, where $\epsilon > 0$.

Hint. The points $x^{(k)}$ are on the central path; see exercise 11.10. Use this to show that

$$u^+ - p^\star \leq \frac{m+\theta}{m+1}(u - p^\star),$$

where u and u^+ are the values of u on consecutive iterations.

Solution. Let $x = z^\star(u)$. From the duality result in exercise 11.10,

$$\begin{array}{rcl} p^\star & \geq & f_0(x) - m(u - f_0(x)) \\ & = & (m+1) f_0(x) - mu, \end{array}$$

and therefore

$$f_0(x) \leq \frac{p^\star + mu}{m+1}.$$

Let $u^+ = \theta u + (1-\theta) f_0(x)$. We have

$$\begin{array}{rcl} u^+ - p^\star & = & \theta u + (1-\theta) f_0(x) - p^\star \\ & \leq & (1-\theta)\dfrac{p^\star + mu}{m+1} + \theta u - p^\star \\ & = & \left(\dfrac{1-\theta}{m+1} - 1\right) p^\star + \left(\dfrac{(1-\theta)m}{m+1} + \theta\right) u \\ & = & \dfrac{m+\theta}{m+1}(u - p^\star). \end{array}$$

11.12 *Barrier method for convex-concave games.* We consider a convex-concave game with inequality constraints,

$$\begin{array}{ll} \text{minimize}_w \text{ maximize}_z & f_0(w,z) \\ \text{subject to} & f_i(w) \leq 0, \quad i=1,\ldots,m \\ & \tilde{f}_i(z) \leq 0, \quad i=1,\ldots,\tilde{m}. \end{array}$$

Here $w \in \mathbf{R}^n$ is the variable associated with minimizing the objective, and $z \in \mathbf{R}^{\tilde{n}}$ is the variable associated with maximizing the objective. The constraint functions f_i and $\tilde{f}_i$ are convex and differentiable, and the objective function f_0 is differentiable and convex-concave, *i.e.*, convex in w, for each z, and concave in z, for each w. We assume for simplicity that $\mathbf{dom}\, f_0 = \mathbf{R}^n \times \mathbf{R}^{\tilde{n}}$.

A *solution* or *saddle-point* for the game is a pair $w^\star$, $z^\star$, for which

$$f_0(w^\star, z) \leq f_0(w^\star, z^\star) \leq f_0(w, z^\star)$$

holds for every feasible w and z. (For background on convex-concave games and functions, see §5.4.3, §10.3.4 and exercises 3.14, 5.24, 5.25, 10.10, and 10.13.) In this exercise we show how to solve this game using an extension of the barrier method, and the infeasible start Newton method (see §10.3).

(a) Let $t > 0$. Explain why the function

$$tf_0(w,z) - \sum_{i=1}^{m} \log(-f_i(w)) + \sum_{i=1}^{\tilde{m}} \log(-\tilde{f}_i(z))$$

is convex-concave in (w,z). We will assume that it has a unique saddle-point, $(w^\star(t), z^\star(t))$, which can be found using the infeasible start Newton method.

(b) As in the barrier method for solving a convex optimization problem, we can derive a simple bound on the suboptimality of $(w^\star(t), z^\star(t))$, which depends only on the problem dimensions, and decreases to zero as t increases. Let W and Z denote the feasible sets for w and z,

$$W = \{w \mid f_i(w) \leq 0,\ i=1,\ldots,m\}, \qquad Z = \{z \mid \tilde{f}_i(z) \leq 0,\ i=1,\ldots,\tilde{m}\}.$$

Show that

$$\begin{aligned} f_0(w^\star(t), z^\star(t)) &\leq \inf_{w \in W} f_0(w, z^\star(t)) + \frac{m}{t}, \\ f_0(w^\star(t), z^\star(t)) &\geq \sup_{z \in Z} f_0(w^\star(t), z) - \frac{\tilde{m}}{t}, \end{aligned}$$

and therefore

$$\sup_{z \in Z} f_0(w^\star(t), z) - \inf_{w \in W} f_0(w, z^\star(t)) \leq \frac{m+\tilde{m}}{t}.$$

Solution.

(a) Follows from the convex-concave property of f_0; convexity of $-\log(-f_i)$, and concavity of $\log(-\tilde{f}_i)$.

(b) Since $(w^\star(t), z^\star(t))$ is a saddle-point of the function

$$tf_0(w,z) - \sum_{i=1}^{m} \log(-f_i(w)) + \sum_{i=1}^{\tilde{m}} \log(-\tilde{f}_i(z)),$$

its gradient with respect to w, and also with respect to z, vanishes there:

$$\begin{aligned} t\nabla_w f_0(w^\star(t), z^\star(t)) + \sum_{i=1}^m \frac{1}{-f_i(w^\star(t))}\nabla f_i(w^\star(t)) &= 0 \\ t\nabla_z f_0(w^\star(t), z^\star(t)) + \sum_{i=1}^{\tilde{m}} \frac{-1}{-\tilde{f}_i(z^\star(t))}\nabla \tilde{f}_i(z^\star(t)) &= 0. \end{aligned}$$

It follows that $w^\star(t)$ minimizes

$$f_0(w, z^\star(t)) + \sum_{i=1}^m \lambda_i f_i(w)$$

over w, where $\lambda_i = 1/(-tf_i(w^\star(t)))$, *i.e.*, for all w, we have

$$f_0(w^\star(t), z^\star(t)) + \sum_{i=1}^m \lambda_i f_i(w^\star(t)) \leq f_0(w, z^\star(t)) + \sum_{i=1}^m \lambda_i f_i(w).$$

The lefthand side is equal to $f_0(w^\star(t), z^\star(t)) - m/t$, and for all $w \in W$, the second term on the righthand side is nonpositive, so we have

$$f_0(w^\star(t), z^\star(t)) \leq \inf_{w\in W} f_0(w, z^\star(t)) + m/t.$$

A similar argument shows that

$$f_0(w^\star(t), z^\star(t)) \geq \sup_{z\in Z} f_0(w^\star(t), z) - m/t.$$

Self-concordance and complexity analysis

11.13 *Self-concordance and negative entropy.*

(a) Show that the negative entropy function $x \log x$ (on $\mathbf{R}_{++}$) is *not* self-concordant.

(b) Show that for any $t > 0$, $tx \log x - \log x$ is self-concordant (on $\mathbf{R}_{++}$).

Solution.

(a) First we consider $f(x) = x \log x$, for which

$$f'(x) = 1 + \log x, \qquad f''(x) = \frac{1}{x}, \qquad f'''(x) = -\frac{1}{x^2}.$$

Thus

$$\frac{|f'''(x)|}{f''(x)^{3/2}} = \frac{1/x^2}{1/x^{3/2}} = \frac{1}{\sqrt{x}}$$

which is unbounded above (as $x \to 0^+$). In particular, the self-concordance inequality $|f'''(x)| \leq 2f''(x)^{3/2}$ fails for $x = 1/5$, so f is *not* self-concordant.

(b) Now we consider $g(x) = tx \log x - \log x$, for which

$$g'(x) = -\frac{1}{x} + t + t \log x, \qquad g''(x) = \frac{1}{x^2} + \frac{t}{x}, \qquad g'''(x) = -\frac{2}{x^3} - \frac{t}{x^2}.$$

Therefore

$$\frac{|g'''(x)|}{g''(x)^{3/2}} = \frac{2/x^3 + t/x^2}{(1/x^2 + t/x)^{3/2}} = \frac{2 + tx}{(1 + tx)^{3/2}}.$$

Define

$$h(a) = \frac{2+a}{(1+a)^{3/2}}$$

so that

$$h(tx) = \frac{|g'''(x)|}{g''(x)^{3/2}}.$$

We have $h(0) = 2$ and we will show that $h'(a) < 0$ for $a > 0$, *i.e.*, h is decreasing for $a > 0$. This will prove that $h(a) \leq h(0) = 2$, and therefore

$$\frac{|g'''(x)|}{g''(x)^{3/2}} \leq 2.$$

We have

$$\begin{aligned} h'(a) &= \frac{(1+a)^{3/2} - (3/2)(1+a)^{1/2}(2+a)}{(1+a)^3} \\ &= \frac{(1+a)^{1/2}((1+a) - (3/2)(2+a))}{(1+a)^3} \\ &= -\frac{(2+a/2)}{(1+a)^{5/2}} \\ &< 0, \end{aligned}$$

for $a > 0$, so we are done.

11.14 *Self-concordance and the centering problem.* Let ϕ be the logarithmic barrier function of problem (11.1). Suppose that the sublevel sets of (11.1) are bounded, and that $tf_0 + \phi$ is closed and self-concordant. Show that $t\nabla^2 f_0(x) + \nabla^2\phi(x) \succ 0$, for all $x \in \mathbf{dom}\,\phi$. *Hint.* See exercises 9.17 and 11.3.

Solution. From exercise 11.3, the sublevel sets of $tf_0 + \phi$ are bounded.

From exercise 9.17, the nullspace of $tf_0 + \phi$ is independent of x. So if the Hessian is not positive definite, $tf_0 + \phi$ is linear along certain lines, which would contradict the fact that the sublevel sets are bounded.

Barrier method for generalized inequalities

11.15 *Generalized logarithm is K-increasing.* Let ψ be a generalized logarithm for the proper cone K. Suppose $y \succ_K 0$.

(a) Show that $\nabla\psi(y) \succeq_{K^*} 0$, *i.e.*, that ψ is K-nondecreasing. *Hint.* If $\nabla\psi(y) \not\succeq_{K^*} 0$, then there is some $w \succ_K 0$ for which $w^T\nabla\psi(y) \leq 0$. Use the inequality $\psi(sw) \leq \psi(y) + \nabla\psi(y)^T(sw - y)$, with $s > 0$.

(b) Now show that $\nabla\psi(y) \succ_{K^*} 0$, *i.e.*, that ψ is K-increasing. *Hint.* Show that $\nabla^2\psi(y) \prec 0$, $\nabla\psi(y) \succeq_{K^*} 0$ imply $\nabla\psi(y) \succ_{K^*} 0$.

Solution.

(a) If $\nabla\psi(y) \not\succeq_{K^*} 0$, there exists a $w \succ_K 0$ such that $w^T\nabla\psi(y) \leq 0$. By concavity of ψ we have

$$\begin{aligned} \psi(sw) &\leq \psi(y) + \nabla\psi(y)^T(sw - y) \\ &= \psi(y) - \theta + sw^T\nabla\psi(y) \\ &\leq \psi(y) - \theta \end{aligned}$$

for all $s > 0$. In particular, $\psi(sw)$ is bounded, for $s \geq 0$. But we have $\psi(sw) = \psi(w) + \theta\log s$, which is unbounded as $s \to \infty$. (We need $w \succ_K 0$ to ensure that $sw \in \mathbf{dom}\,\psi$.)

(b) We now know that $\nabla\psi(y) \succeq_{K^*} 0$. For small v we have

$$\nabla\psi(y+v) \approx \nabla\psi(y) + \nabla^2\psi(y)v,$$

and by part (a) we have $\nabla\psi(y+v) \succeq_{K^*} 0$. Since $\nabla^2\psi(y)$ is nonsingular, we conclude that we must have $\nabla\psi(y) \succ_{K^*} 0$.

11.16 [NN94, page 41] *Properties of a generalized logarithm.* Let ψ be a generalized logarithm for the proper cone K, with degree θ. Prove that the following properties hold at any $y \succ_K 0$.

(a) $\nabla\psi(sy) = \nabla\psi(y)/s$ for all $s > 0$.

(b) $\nabla\psi(y) = -\nabla^2\psi(y)y$.

(c) $y^T\nabla\psi^2(y)y = -\theta$.

(d) $\nabla\psi(y)^T\nabla^2\psi(y)^{-1}\nabla\psi(y) = -\theta$.

Solution.

(a) Differentiate $\psi(sy) = \psi(y) + \theta\log s$ with respect to y to get $s\nabla\psi(sy) = \nabla\psi(y)$.

(b) Differentiating $(y+tv)^T\nabla\psi(y+tv) = \theta$ with respect to t gives

$$\nabla\psi(y+tv)^Tv + (y+tv)^T\nabla^2\psi(y+tv)v = 0.$$

At $t = 0$ we get

$$\nabla\psi(y)^Tv + y^T\nabla^2\psi(y)v = 0.$$

This holds for all v, so $\nabla\psi(y) = -\nabla^2\psi(y)y$.

(c) From part (b),

$$y^T\nabla\psi^2(y)y = -y^T\nabla\psi(y) = -\theta.$$

(d) From part (b),

$$\nabla\psi(y)^T\nabla^2\psi(y)^{-1}\nabla\psi(y) = -\nabla\psi(y)^Ty = -\theta.$$

11.17 *Dual generalized logarithm.* Let ψ be a generalized logarithm for the proper cone K, with degree θ. Show that the dual generalized logarithm $\overline{\psi}$, defined in (11.49), satisfies

$$\overline{\psi}(sv) = \psi(v) + \theta\log s,$$

for $v \succ_{K^*} 0$, $s > 0$.

Solution.

$$\overline{\psi}(sv) = \inf_u \left(sv^Tu - \psi(u)\right) = \inf_{\tilde{u}} \left(v^T\tilde{u} - \psi(\tilde{u}/s)\right)$$

where $\tilde{u} = su$. Using the logarithm property for ψ, we have $\psi(\tilde{u}/s) = \psi(\tilde{u}) - \theta\log s$, so

$$\overline{\psi}(sv) = \inf_{\tilde{u}} \left(v^T\tilde{u} - \psi(\tilde{u})\right) + \theta\log s = \overline{\psi}(u) + \theta\log s.$$

11.18 Is the function

$$\psi(y) = \log\left(y_{n+1} - \frac{\sum_{i=1}^n y_i^2}{y_{n+1}}\right),$$

with $\mathbf{dom}\,\psi = \{y \in \mathbf{R}^{n+1} \mid y_{n+1} > \sum_{i=1}^n y_i^2\}$, a generalized logarithm for the second-order cone in $\mathbf{R}^{n+1}$?

Solution. It is not. It satisfies all the required properties except closedness.

To see this, take any $a > 0$, and suppose y approaches the origin along the path

$$(y_1, \ldots, y_n) = \sqrt{t(t-a)/n}, \qquad y_{n+1} = t$$

where $t > 0$. We have

$$(\sum_{i=1}^{n} y_i^2)^{1/2} = \sqrt{t(t-a)} < y_{n+1}$$

so $y \in \mathbf{int}\, K$. However,

$$\psi(y) = \log(t - t(t-a)/t) = \log a.$$

Therefore we can find sequences of points with any arbitrary limit.

Implementation

11.19 *Yet another method for computing the Newton step.* Show that the Newton step for the barrier method, which is given by the solution of the linear equations (11.14), can be found by solving a *larger* set of linear equations with coefficient matrix

$$\begin{bmatrix} t\nabla^2 f_0(x) + \sum_i \frac{1}{-f_i(x)} \nabla^2 f_i(x) & Df(x)^T & A^T \\ Df(x) & -\mathbf{diag}(f(x))^2 & 0 \\ A & 0 & 0 \end{bmatrix}$$

where $f(x) = (f_1(x), \ldots, f_m(x))$.

For what types of problem structure might solving this larger system be interesting?

Solution.

$$\begin{bmatrix} t\nabla^2 f_0(x) + \sum_i \frac{1}{-f_i(x)} \nabla^2 f_i(x) & Df(x)^T & A^T \\ Df(x) & -\mathbf{diag}(f(x))^2 & 0 \\ A & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\rm nt} \\ y \\ \nu_{\rm nt} \end{bmatrix} = - \begin{bmatrix} g \\ 0 \\ 0 \end{bmatrix}.$$

where $g = t\nabla f_0(x) + \nabla\phi(x)$. From the second equation,

$$y_i = \frac{\nabla f_i(x)^T \Delta x_{\rm nt}}{f_i(x)^2}$$

and substituting in the first equation gives (11.14).

This might be useful if the big matrix is sparse, and the 2×2 block system (obtained by pivoting on the $\mathbf{diag}(f(x))^2$ block) has a dense (1,1) block. For example if the (1,1) block of the big system is block diagonal, $m \ll n$ is small, and $Df(x)$ is dense.

11.20 *Network rate optimization via the dual problem.* In this problem we examine a dual method for solving the network rate optimization problem of §11.8.4. To simplify the presentation we assume that the utility functions U_i are strictly concave, with $\mathbf{dom}\, U_i = \mathbf{R}_{++}$, and that they satisfy $U_i'(x_i) \to \infty$ as $x_i \to 0$ and $U_i'(x_i) \to 0$ as $x_i \to \infty$.

(a) Express the dual problem of (11.62) in terms of the conjugate utility functions $V_i = (-U_i)^*$, defined as

$$V_i(\lambda) = \sup_{x>0}(\lambda x + U_i(x)).$$

Show that $\mathbf{dom}\, V_i = -\mathbf{R}_{++}$, and that for each $\lambda < 0$ there is a unique x with $U_i'(x) = -\lambda$.

(b) Describe a barrier method for the dual problem. Compare the complexity per iteration with the complexity of the method in §11.8.4. Distinguish the same two cases as in §11.8.4 ($A^T A$ is sparse and AA^T is sparse).

Solution.

(a) Suppose $\lambda < 0$. Since U_i is strictly concave and increasing, with $U_i'(x_i) \to \infty$ as $x_i \to 0$ and $U_i'(x_i) \to 0$ as $x_i \to \infty$, there is a unique x with

$$U_i'(x) = -\lambda.$$

After changing problem (11.62) its Lagrangian is

$$\begin{aligned} L(x,\lambda,z) &= \sum_{i=1}^n (-U_i(x)) + \lambda^T(Ax-c) - z^T x \\ &= -\sum_{i=1}^n \left(U_i(x) - (A^T\lambda)_i x_i + z_i x_i\right) - c^T\lambda. \end{aligned}$$

The minimum over x is

$$\begin{aligned} \inf_x L(x,\lambda,z) &= \inf_x \left(-\sum_{i=1}^n (U_i(x) - (A^T\lambda)_i x_i + z_i x_i) - c^T\lambda \right) \\ &= -\sum_{i=1}^n \sup_x (U_i(x) - (A^T\lambda)_i x_i + z_i x_i) - c^T\lambda \\ &= -\sum_{i=1}^n V_i(-(A^T\lambda)_i + z_i) - c^T\lambda, \end{aligned}$$

so the dual problem is (after changing the sign again)

$$\begin{array}{ll} \text{minimize} & c^T\lambda + \sum_{i=1}^n V_i(-(A^T\lambda)_i + z_i) \\ \text{subject to} & \lambda \succeq 0, \quad z \succeq 0. \end{array}$$

The function V_i is increasing on its domain $-\mathbf{R}_{++}$, so $z = 0$ at the optimum and the dual problem simplifies to

$$\begin{array}{ll} \text{minimize} & c^T\lambda + \sum_{i=1}^n V_i(-(A^T\lambda)_i) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

$-\lambda_i$ can be interpreted as the price on link i. $-(A^T\lambda)_i$ is the sum of the prices along the path of flow i.

(b) The Hessian of

$$t\left(c^T\lambda + \sum_{i=1}^n V_i(-(A^T\lambda)_i)\right) - \sum_i \log\lambda_i$$

is

$$H = tA\,\mathbf{diag}(-A^T\lambda)^{-2}A^T + \mathbf{diag}(\lambda)^{-2}.$$

If AA^T is sparse, we solve the Newton equation $H\Delta\lambda = -g$.

If A^TA is sparse, we apply the matrix inversion lemma and compute the Newton step by first solving an equation with coefficient matrix of the form $D_1 + A^TD_2A$, where D_1 and D_2 are diagonal (see §11.8.4).

Numerical experiments

11.21 *Log-Chebyshev approximation with bounds.* We consider an approximation problem: find $x \in \mathbf{R}^n$, that satisfies the variable bounds $l \preceq x \preceq u$, and yields $Ax \approx b$, where $b \in \mathbf{R}^m$. You can assume that $l \prec u$, and $b \succ 0$ (for reasons we explain below). We let a_i^T denote the ith row of the matrix A.

We judge the approximation $Ax \approx b$ by the *maximum fractional deviation*, which is

$$\max_{i=1,\ldots,n} \max\{(a_i^T x)/b_i, b_i/(a_i^T x)\} = \max_{i=1,\ldots,n} \frac{\max\{a_i^T x, b_i\}}{\min\{a_i^T x, b_i\}},$$

when $Ax \succ 0$; we define the maximum fractional deviation as ∞ if $Ax \not\succ 0$.
The problem of minimizing the maximum fractional deviation is called the *fractional Chebyshev approximation problem*, or the *logarithmic Chebyshev approximation problem*, since it is equivalent to minimizing the objective

$$\max_{i=1,\ldots,n} |\log a_i^T x - \log b_i|.$$

(See also exercise 6.3, part (c).)

(a) Formulate the fractional Chebyshev approximation problem (with variable bounds) as a convex optimization problem with twice differentiable objective and constraint functions.

(b) Implement a barrier method that solves the fractional Chebyshev approximation problem. You can assume an initial point $x^{(0)}$, satisfying $l \prec x^{(0)} \prec u$, $Ax^{(0)} \succ 0$, is known.

Solution.

(a) We can formulate the fractional Chebyshev approximation problem with variable bounds as

$$\begin{array}{ll} \text{minimize} & s \\ \text{subject to} & (a_i^T x)/b_i \le s, \quad i = 1, \ldots, m \\ & b_i/(a_i^T x) \le s, \quad i = 1, \ldots, m \\ & a_i^T x \ge 0, \quad i = 1, \ldots, m \\ & l \preceq x \preceq u, \end{array}$$

This is clearly a convex problem, since the inequalities are linear, except for the second group, which involves the inverse.

The sublevel sets are bounded (by the last constraint).

Note that we can, without loss of generality, take $b_i = 1$, and replace a_i with a_i/b_i. We will assume this has been done. To simplify the notation, we will use a_i to denote the scaled version (*i.e.*, a_i/b_i in the original problem data).

(b) In the centering problems we must minimize the function

$$\begin{aligned} ts + \phi(s,x) &= ts - \sum_{i=1}^m \log(s - a_i^T x) - \sum_{i=1}^m \log a_i^T x - \sum_{i=1}^m \log(s - 1/a_i^T x) \\ &\quad - \sum_{i=1}^n \log(u_i - x_i) - \sum_{i=1}^n \log(x_i - l_i) \\ &= \phi_1(s,x) + \phi_2(s,x) + \phi_3(s,x) \end{aligned}$$

with variables x, s, where

$$\begin{aligned} \phi_1(s,x) &= ts - \sum_{i=1}^n \log(u_i - x_i) - \sum_{i=1}^n \log(x_i - l_i) \\ \phi_2(s,x) &= -\sum_{i=1}^m \log(s - a_i^T x) \\ \phi_3(s,x) &= -\sum_{i=1}^m \log(s(a_i^T x) - 1). \end{aligned}$$

The gradient and Hessian of ϕ_1 are

$$\begin{aligned}\nabla\phi_1(s,x) &= \begin{bmatrix} t \\ \mathbf{diag}(u-x)^{-1}\mathbf{1} - \mathbf{diag}(x-l)^{-1}\mathbf{1} \end{bmatrix} \\ \nabla^2\phi_1(s,x) &= \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{diag}(u-x)^{-2} + \mathbf{diag}(x-l)^{-2} \end{bmatrix}.\end{aligned}$$

The gradient and Hessian of ϕ_2 are

$$\begin{aligned}\nabla\phi_2(x) &= \begin{bmatrix} -\mathbf{1}^T \\ A^T \end{bmatrix} \mathbf{diag}(s-Ax)^{-1}\mathbf{1} \\ \nabla^2\phi_2(x) &= \begin{bmatrix} -\mathbf{1}^T \\ A^T \end{bmatrix} \mathbf{diag}(s-Ax)^{-2} \begin{bmatrix} -\mathbf{1} & A \end{bmatrix}.\end{aligned}$$

We can find the gradient and Hessian of ϕ_3 by expressing it as $\phi_3(s,x) = h(s,Ax)$ where

$$h(s,y) = -\sum_{i=1}^m \log(sy_i - 1),$$

and then applying the chain rule. The gradient and Hesian of h are

$$\nabla h(s,y) = -\begin{bmatrix} \sum_{i=1}^m y_i/(sy_i-1) \\ s/(sy_1-1) \\ \vdots \\ s/(sy_m-1) \end{bmatrix} = -\begin{bmatrix} y^T\,\mathbf{diag}(sy-\mathbf{1})^{-1}\mathbf{1} \\ s\,\mathbf{diag}(sy-\mathbf{1})^{-1}\mathbf{1} \end{bmatrix}$$

and

$$\begin{aligned}&\nabla^2 h(s,y) \\ &= \begin{bmatrix} \sum_i y_i^2/(sy_i-1)^2 & 1/(sy_1-1)^2 & 1/(sy_2-1)^2 & \cdots & 1/(sy_m-1)^2 \\ 1/(sy_1-1)^2 & s^2/(sy_1-1)^2 & 0 & \cdots & 0 \\ 1/(sy_2-1)^2 & 0 & s^2/(sy_2-1)^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1/(sy_m-1)^2 & 0 & 0 & \cdots & s^2/(sy_m-1)^2 \end{bmatrix} \\ &= \begin{bmatrix} y^T\,\mathbf{diag}(sy-\mathbf{1})^{-2}y & \mathbf{1}^T\,\mathbf{diag}(sy-\mathbf{1})^{-2} \\ \mathbf{diag}(sy-\mathbf{1})^{-2}\mathbf{1} & s^2\,\mathbf{diag}(sy-\mathbf{1})^{-2} \end{bmatrix}.\end{aligned}$$

We therefore obtain

$$\begin{aligned}\nabla\phi_3(s,x) &= \begin{bmatrix} 1 & 0 \\ 0 & A^T \end{bmatrix} \nabla h(s,Ax) \\ &= -\begin{bmatrix} y^T \\ sA^T \end{bmatrix} \mathbf{diag}(sAx-\mathbf{1})^{-1}\mathbf{1} \\ \nabla^2\phi_3(s,x) &= \begin{bmatrix} 1 & 0 \\ 0 & A^T \end{bmatrix} \nabla^2 h(s,Ax) \begin{bmatrix} 1 & 0 \\ 0 & A \end{bmatrix} \\ &= \begin{bmatrix} x^TA\,\mathbf{diag}(sAx-\mathbf{1})^{-2}Ax & \mathbf{1}^T\,\mathbf{diag}(sAx-\mathbf{1})^{-2}A \\ A^T\,\mathbf{diag}(sAx-\mathbf{1})^{-2}\mathbf{1} & s^2A^T\,\mathbf{diag}(sAx-\mathbf{1})^{-2}A \end{bmatrix}.\end{aligned}$$

A Matlab implementation is given below.

```
MAXITERS = 200;
ALPHA = 0.01;
BETA = 0.5;
NTTOL = 1e-8;   % terminate Newton iterations if lambda^2 < NTTOL
MU = 20;
TOL = 1e-4;     % terminate if duality gap less than TOL

x = x0;  y = A*x;  s = 1.1*max([max(A*x), max(1./y)]);
t = 1;
for iter = 1:MAXITERS
   val = t*s - sum(log(u-x)) - sum(log(x-l)) - sum(log(s-y)) - ...
         sum(log(s*y-1));
   grad = [t-sum(1./(s-y))-sum(y./(s*y-1));
           1./(u-x)-1./(x-l)+A'*(1./(s-y)-s./(s*y-1))];
   hess = [sum((s-y).^(-2)+(y./(s*y-1)).^2)  ...
                  (-(s-y).^(-2) + (s*y-1).^(-2))'*A;
            A'*(-(s-y).^(-2) + (s*y-1).^(-2)) ...
            diag((u-x).^(-2) + (x-l).^(-2)) + ...
                A'*(diag((s-y).^(-2)+(s./(s*y-1)).^2))*A];
   step = -hess\grad;  fprime = grad'*step;
   if (abs(fprime) < NTTOL),
      gap = (3*m+2*n)/t;
      if (gap<TOL); break; end;
      t = MU*t;
   else
      ds = step(1);  dx = step(1+[1:n]);  dy = A*dx;
      tls = 1;
      news = s+tls*ds;  newx = x+tls*dx;  newy = y+tls*dy;
      while (min([news-newy; news-1./newy; newy; newx-l; u-newx]) <= 0),
         tls = BETA*tls;
         news = s+tls*ds;  newx = x+tls*dx;  newy = y+tls*dy;
      end;
      newval = t*news - sum(log(u-newx)) - sum(log(newx-l)) ...
               - sum(log(news-newy)) - sum(log(news*newy-1));
      while (newval >= val + tls*ALPHA*fprime),
         tls = BETA*tls;
         news = s+tls*ds;  newx = x+tls*dx;  newy = y+tls*dy;
         newval = t*news - sum(log(u-newx)) - sum(log(newx-l)) ...
              - sum(log(news-newy)) - sum(log(news*newy-1));
      end;
      x = x+tls*dx;  y = A*x;  s = s+tls*ds;
   end;
end;
```

11.22 *Maximum volume rectangle inside a polyhedron.* Consider the problem described in exercise 8.16, *i.e.*, finding the maximum volume rectangle $\mathcal{R} = \{x \mid l \preceq x \preceq u\}$ that lies in a polyhedron described by a set of linear inequalities, $\mathcal{P} = \{x \mid Ax \preceq b\}$. Implement a barrier method for solving this problem. You can assume that $b \succ 0$, which means that for small $l \prec 0$ and $u \succ 0$, the rectangle $\mathcal{R}$ lies inside $\mathcal{P}$.

Test your implementation on several simple examples. Find the maximum volume rect-

angle that lies in the polyhedron defined by

$$A = \begin{bmatrix} 0 & -1 \\ 2 & -4 \\ 2 & 1 \\ -4 & 4 \\ -4 & 0 \end{bmatrix}, \qquad b = \mathbf{1}.$$

Plot this polyhedron, and the maximum volume rectangle that lies inside it.

Solution. We use the formulation

$$\begin{array}{ll} \text{minimize} & -\sum_{i=1}^n \log(u_i - l_i) \\ \text{subject to} & A^+u - A^-l \preceq b, \end{array}$$

(with implicit constraint $u \succ l$) worked out in exercise 8.16. Here $a_{ij}^+ = \max\{a_{ij}, 0\}$, $a_{ij}^- = \max\{-a_{ij}, 0\}$.

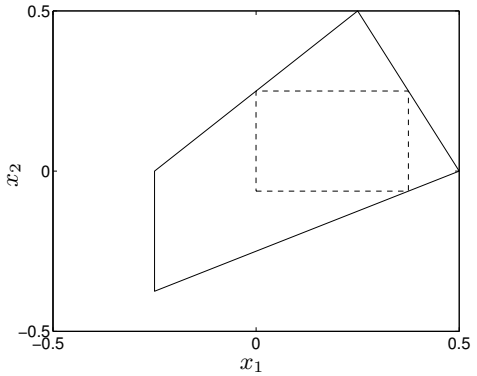
The gradient and Hessian of the function

$$\psi(l,u) = -t\sum_{i=1}^n \log(u_i - l_i) - \sum_{i=1}^n \log((b - A^+u + A^-l)_i)$$

are

$$\begin{aligned} \nabla\psi(l,u) &= t\begin{bmatrix} I \\ -I \end{bmatrix}\mathbf{diag}(u-l)^{-1}\mathbf{1} + \begin{bmatrix} -A^{-T} \\ A^{+T} \end{bmatrix}\mathbf{diag}(b - A^+u + A^-l)^{-1}\mathbf{1} \\ \nabla^2\psi(l,u) &= t\begin{bmatrix} I \\ -I \end{bmatrix}\mathbf{diag}(u-l)^{-2}\begin{bmatrix} I & -I \end{bmatrix} \\ &\quad + \begin{bmatrix} -A^{-T} \\ A^{+T} \end{bmatrix}\mathbf{diag}(b - A^+u + A^-l)^{-2}\begin{bmatrix} -A^- & A^+ \end{bmatrix}. \end{aligned}$$

A plot of the particular polyhedron and the maximum volume box is given below.



An implementation in Matlab is given below.

```
MAXITERS = 200;
ALPHA = 0.01;
BETA = 0.5;
NTTOL = 1e-8;   % terminate Newton iterations if lambda^2 < NTTOL
MU = 20;
TOL = 1e-4;     % terminate if duality gap less than TOL
```

```
Ap = max(A,0);  Am = max(-A,0);
r = max(Ap*ones(n,1) + Am*ones(n,1));
u = (.5/r)*ones(n,1);  l = -(.5/r)*ones(n,1);
t = 1;
for iter = 1:MAXITERS
   y = b+Am*l-Ap*u;
   val = -t*sum(log(u-l)) - sum(log(y));
   grad = t*[1./(u-l); -1./(u-l)] + [-Am'; Ap']*(1./y);
   hess = t*[diag(1./(u-l).^2), -diag(1./(u-l).^2);
             -diag(1./(u-l).^2),  diag(1./(u-l).^2)]  + ...
          [-Am'; Ap']*diag(1./y.^2)*[-Am Ap];
   step = -hess\grad;  fprime = grad'*step;
   if (abs(fprime) < NTTOL),
      gap = (2*m)/t;
      disp(['iter ', int2str(iter), '; gap = ', num2str(gap)]);
      if (gap<TOL); break; end;
      t = MU*t;
   else
      dl = step(1:n);  du = step(n+[1:n]);   dy = Am*dl-Ap*du;
      tls = 1;
      while (min([u-l+tls*(du-dl); y+tls*dy]) <= 0)
         tls = BETA*tls;
      end;
      while (-t*sum(log(u-l+tls*(du-dl))) - sum(log(y+tls*dy)) >= ...
             val + tls*ALPHA*fprime),
         tls = BETA*tls;
      end;
      l = l+tls*dl;  u = u+tls*du;
   end;
end;
```

11.23 *SDP bounds and heuristics for the two-way partitioning problem.* In this exercise we consider the two-way partitioning problem (5.7), described on page 219, and also in exercise 5.39:

$$\begin{array}{ll} \text{minimize} & x^T W x \\ \text{subject to} & x_i^2 = 1, \quad i = 1, \ldots, n, \end{array} \tag{11.65}$$

with variable $x \in \mathbf{R}^n$. We assume, without loss of generality, that $W \in \mathbf{S}^n$ satisfies $W_{ii} = 0$. We denote the optimal value of the partitioning problem as $p^\star$, and $x^\star$ will denote an optimal partition. (Note that $-x^\star$ is also an optimal partition.)

The Lagrange dual of the two-way partitioning problem (11.65) is given by the SDP

$$\begin{array}{ll} \text{maximize} & -\mathbf{1}^T \nu \\ \text{subject to} & W + \mathbf{diag}(\nu) \succeq 0, \end{array} \tag{11.66}$$

with variable $\nu \in \mathbf{R}^n$. The dual of this SDP is

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(WX) \\ \text{subject to} & X \succeq 0 \\ & X_{ii} = 1, \quad i = 1, \ldots, n, \end{array} \tag{11.67}$$

with variable $X \in \mathbf{S}^n$. (This SDP can be interpreted as a relaxation of the two-way partitioning problem (11.65); see exercise 5.39.) The optimal values of these two SDPs are equal, and give a lower bound, which we denote $d^\star$, on the optimal value $p^\star$. Let $\nu^\star$ and $X^\star$ denote optimal points for the two SDPs.

(a) Implement a barrier method that solves the SDP (11.66) and its dual (11.67), given the weight matrix W. Explain how you obtain nearly optimal ν and X, give formulas for any Hessians and gradients that your method requires, and explain how you compute the Newton step. Test your implementation on some small problem instances, comparing the bound you find with the optimal value (which can be found by checking the objective value of all 2^n partitions). Try your implementation on a randomly chosen problem instance large enough that you cannot find the optimal partition by exhaustive search (*e.g.*, $n = 100$).

(b) *A heuristic for partitioning.* In exercise 5.39, you found that if $X^\star$ has rank one, then it must have the form $X^\star = x^\star(x^\star)^T$, where $x^\star$ is optimal for the two-way partitioning problem. This suggests the following simple heuristic for finding a good partition (if not the best): solve the SDPs above, to find $X^\star$ (and the bound $d^\star$). Let v denote an eigenvector of $X^\star$ associated with its largest eigenvalue, and let $\hat{x} = \mathbf{sign}(v)$. The vector $\hat{x}$ is our guess for a good partition.

Try this heuristic on some small problem instances, and the large problem instance you used in part (a). Compare the objective value of your heuristic partition, $\hat{x}^T W \hat{x}$, with the lower bound $d^\star$.

(c) *A randomized method.* Another heuristic technique for finding a good partition, given the solution $X^\star$ of the SDP (11.67), is based on *randomization*. The method is simple: we generate independent samples $x^{(1)}, \ldots, x^{(K)}$ from a normal distribution on $\mathbf{R}^n$, with zero mean and covariance $X^\star$. For each sample we consider the heuristic approximate solution $\hat{x}^{(k)} = \mathbf{sign}(x^{(k)})$. We then take the best among these, *i.e.*, the one with lowest cost. Try out this procedure on some small problem instances, and the large problem instance you considered in part (a).

(d) *A greedy heuristic refinement.* Suppose you are given a partition x, *i.e.*, $x_i \in \{-1, 1\}$, $i = 1, \ldots, n$. How does the objective value change if we move element i from one set to the other, *i.e.*, change x_i to $-x_i$? Now consider the following simple greedy algorithm: given a starting partition x, move the element that gives the largest reduction in the objective. Repeat this procedure until no reduction in objective can be obtained by moving an element from one set to the other.

Try this heuristic on some problem instances, including the large one, starting from various initial partitions, including $x = \mathbf{1}$, the heuristic approximate solution found in part (b), and the randomly generated approximate solutions found in part (c). How much does this greedy refinement improve your approximate solutions from parts (b) and (c)?

Solution.

(a) We implement a barrier method to solve the SDP (11.66). The only constraint in the problem is the LMI $W + \mathbf{diag}(\nu) \succeq 0$, for which we will use the log barrier $-\log\det(W + \mathbf{diag}(\nu))$. To start the barrier method, we need a strictly feasible point, but this is easily found. If $\lambda_{\min}(W)$ is the smallest eigenvalue of the matrix W, then $W + (-\lambda_{\min}(W) + 1)I$ has smallest eigenvalue one, and so is positive definite. Thus, $\nu = (-\lambda_{\min}(W) + 1)\mathbf{1}$ is a strictly feasible starting point.

At each outer iteration, we use Newton's method to minimize

$$f(\nu) = t\mathbf{1}^T\nu - \log\det(W + \mathbf{diag}(\nu)). \tag{11.23.A}$$

We can start with $t = 1$, and at the end of each outer iteration increase t by a factor $\mu = 10$ (say) until the desired accuracy is reached. At the end of each iteration, the duality gap is exactly n/t, with dual feasible point

$$Z = (n/t)(W + \mathbf{diag}(\nu))^{-1}.$$

We will return ν and Z, at the end of the first outer iteration to satisfy $n/t \leq \epsilon$, where ϵ is the required tolerance.

Now we turn to the question of how to compute the gradient and Hessian of f. We know that for $X \in \mathbf{S}^n_{++}$, the gradient of the function $g(X) = \log\det(X)$ at X is given by

$$\nabla g(X) = X^{-1}.$$

We use the chain rule, with

$$X = W + \mathbf{diag}(\nu) = W + \sum_{i=1}^{n} \nu_i E_{ii},$$

where E_{ii} is the matrix with a one in the i,i entry and zeros elsewhere, to obtain

$$\begin{aligned} \nabla f(\nu)_i &= t - \mathbf{tr}((W + \mathbf{diag}(\nu))^{-1} E_{ii}) \\ &= t - \left((W + \mathbf{diag}(\nu))^{-1}\right)_{ii} \end{aligned}$$

for $i = 1, \ldots, n$. Thus we have the simple formula

$$\nabla f(\nu) = t\mathbf{1} - \mathbf{diag}((W + \mathbf{diag}(\nu))^{-1}).$$

The second derivative of $\log\det X$, at $X \in \mathbf{S}^n_{++}$, is given by the bilinear form

$$\nabla^2 g(X)[Y, Z] = -\mathbf{tr}(X^{-1} Y X^{-1} Z).$$

Applying this to our function f yields, with $X = W + \mathbf{diag}(\nu)$,

$$\nabla^2 f(\nu)_{ij} = \mathbf{tr}(X^{-1} E_{ii} X^{-1} E_{jj}) = \left(X^{-1}\right)^2_{ij},$$

for $i, j = 1, \ldots, n$. Thus we have the very simple formula for the Hessian:

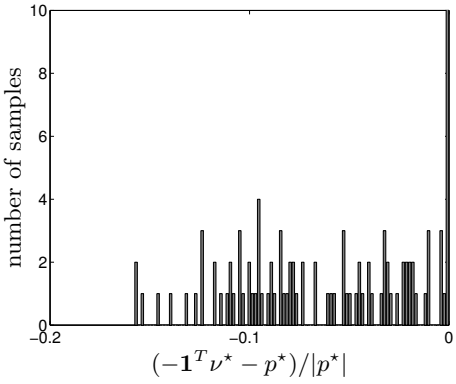
$$\nabla^2 f(\nu) = \left((W + \mathbf{diag}(\nu))^{-1}\right) \circ \left((W + \mathbf{diag}(\nu))^{-1}\right),$$

where for $U,\ V \in \mathbf{S}^n$, the Schur (or Hadamard, or elementwise) product of U and V, denoted $W = U \circ V$, is defined by $W_{ij} = U_{ij} V_{ij}$.

We first test the method on some small problems. We generate random symmetric matrices $W \in \mathbf{S}^{10}$, with off-diagonal elements generated from independent $\mathcal{N}(0, 1)$ distributions, and zero diagonal elements. The figure shows the distribution of the relative error

$$\frac{-\mathbf{1}^T \nu^\star - p^\star}{|p^\star|}$$

for 100 randomly generated matrices.



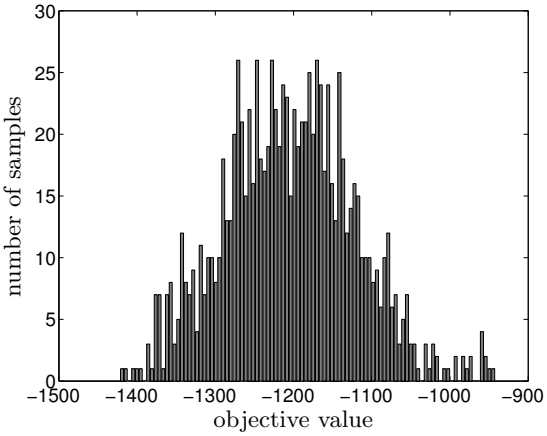
We notice that the lower bound is equal (or very close) to $p^\star$ in 10 cases, and never less than about 15% below $p^\star$.

We also generate a larger problem instance, with $n = 100$. The optimal value of the relaxation is -1687.5. The lower bound from the eigenvalue decomposition of W (see remark 5.1) is $n\lambda_{\min}(W) = -1898.4$.

(b) We first try the heuristic on the family of 100 problems with $n = 10$. The heuristic gave the correct solution in 70 instances. For the larger problem, the heuristic gives the upper bound -1336.5. At this point we can say that the optimal value of the larger problem lies between -1336.5 and -1687.5.

(c) We first try this heuristic, with $K = 10$, on the family of 100 problems with $n = 10$. The heuristic gave the correct solution in 88 instances.

We plot below a histogram of the objective obtained by the randomized heuristic, over 1000 samples.



Many of these samples have an objective value larger than the one found in part (b) above, but some have a lower cost. The minimum value is -1421.7, so $p^\star$ lies between -1421.7 and -1687.5.

(d) The contribution of x_j to the cost is $(\sum_{i=1}^n W_{ij}x_i)x_j$. If this number is positive, then switching the sign of x_j will decrease the objective by $2\sum_{i=1}^n W_{ij}x_i$.

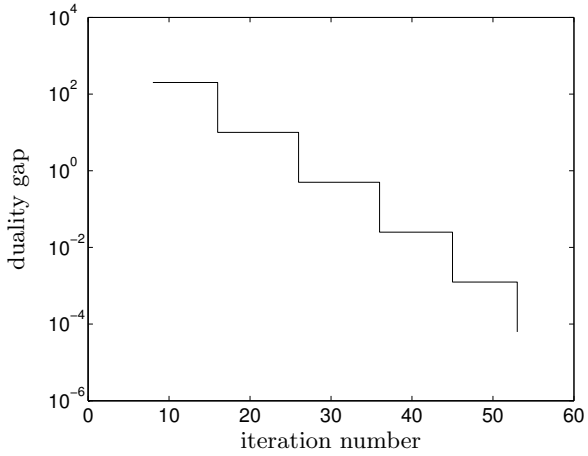
We apply the greedy heuristic to the larger problem instance. For $x = \mathbf{1}$, the cost is reduced from 13.6 to -1344.8. For the solution from part (b), the cost is reduced from -1336.5 to -1440.6. For the solution from part (b), the cost is reduced from -1421.7 to -1440.6.

11.24 *Barrier and primal-dual interior-point methods for quadratic programming.* Implement a barrier method, and a primal-dual method, for solving the QP (without equality constraints, for simplicity)

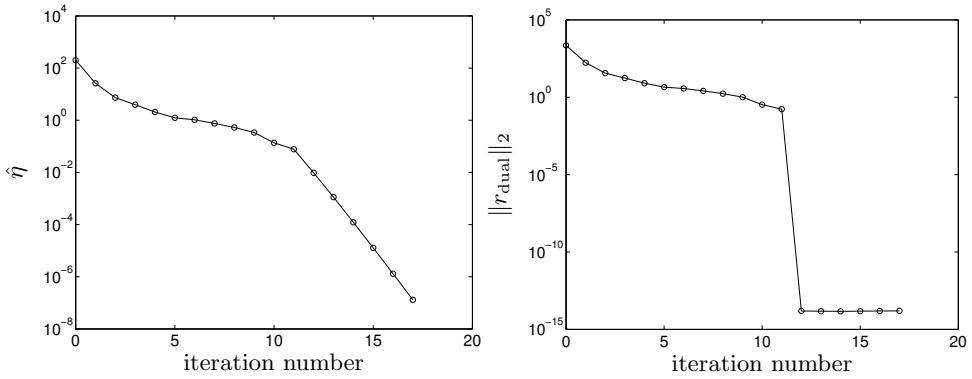
$$\begin{array}{ll} \text{minimize} & (1/2)x^TPx + q^Tx \\ \text{subject to} & Ax \preceq b, \end{array}$$

with $A \in \mathbf{R}^{m\times n}$. You can assume a strictly feasible initial point is given. Test your codes on several examples. For the barrier method, plot the duality gap versus Newton steps. For the primal-dual interior-point method, plot the surrogate duality gap and the norm of the dual residual versus iteration number.

Solution. The first figure shows the progress (duality gap) versus Newton iterations for the barrier method, applied to a randomly generated instance with $n = 100$ variables and $m = 200$ constraints. We use $\mu = 20$, $\alpha = 0.01$, $\beta = 0.5$, and $t^{(0)} = 1$. We choose $b \succ 0$, and use $x^{(0)} = 0$.



The next two figure show the progress (surrogate duality gap η and dual residual norm $\|r_{\rm dual}\|_2$ versus iteration number) of the primal-dual method applied to the same problem instance. We use $\mu = 10$, $\alpha = 0.01$, $\beta = 0.5$, $x^{(0)} = 1$, and $\lambda_i^{(0)} = 1/b_i$.



The Matlab code for the barrier method is as follows.

```
MAXITERS = 200;
ALPHA = 0.01;
BETA = 0.5;
MU = 20;
TOL = 1e-3;
NTTOL = 1e-6;
x = zeros(n,1);  t = 1;
for iter = 1:MAXITERS
   y = b-A*x;
   val = t*(.5*x'*P*x + q'*x) - sum(log(y));
   grad = t*(P*x+q) + A'*(1./y);
   hess = t*P + A'*diag(1./y.^2)*A;
   v = -hess\grad;  fprime = grad'*v;
   s = 1;  dy = -A*v;
   while (min(y+s*dy) <= 0),  s = BETA*s;  end;
   while (t*(.5*(x+s*v)'*P*(x+s*v) + q'*(x+s*v)) - ...
      sum(log(y+s*dy)) >= val + ALPHA*s*fprime),  s=BETA*s;end;
   x = x+s*v;
   if (-fprime < NTTOL),
```

```
      gap = m/t;
      if (gap < TOL),  break;  end;
      t = MU*t;
   end;
end;
```

The Matlab code for the primal-dual method is as follows.

```
MAXITERS = 200;
TOL = 1e-6;
RESTOL = 1e-8;
MU = 10;
ALPHA = 0.01;
BETA = 0.5;
x = zeros(n,1);  s = b-A*x;  z = 1./s;
for iters = 1:MAXITERS
   gap = s'*z;   res = P*x + q + A'*z ;
   if ((gap < TOL) & (norm(res) < RESTOL)), break; end;
   tinv = gap/(m*MU);
   sol = -[ P     A';  A   diag(-s./z) ] \ ...
          [ P*x+q+A'*z;  -s + tinv*(1./z) ];
   dx = sol(1:n);  dz = sol(n+[1:m]);  ds = -A*dx;
   r = [P*x+q+A'*z;  z.*s-tinv];
   step = min(1.0, 0.99/max(-dz./z));
   while (min(s+step*ds) <= 0),  step = BETA*step;  end;
   newz = z+step*dz;  newx = x+step*dx;  news = s+step*ds;
   newr = [P*newx+q+A'*newz;  newz.*news-tinv];
   while (norm(newr) > (1-ALPHA*step)*norm(r))
      step = BETA*step;
      newz = z+step*dz;  newx = x+step*dx;  news = s+step*ds;
      newr = [P*newx+q+A'*newz;  newz.*news-tinv];
   end;
   x = x+step*dx;  z = z +step*dz;  s = b-A*x;
end;
```