Convex Optimization

Stephen Boyd *Lieven Vandenberghe*

August 7, 1999

Course Reader for EECS290N
UC Berkeley
Fall 1999

August 7, 1999

Contents

Chapter 1

Convex sets

1.1 Affine and convex sets

1.1.1 Lines and line segments

Suppose $x_1 \neq x_2$ are two points in $\mathbf{R}^n$. The *line* passing through these points is given by

$$\{\theta x_1 + (1-\theta)x_2 \mid \theta \in \mathbf{R}\}.$$

Values of the parameter θ between 0 and 1 correspond to the (closed) *line segment* between x_1 and x_2, denoted $[x_1, x_2]$:

$$[x_1, x_2] = \{\theta x_1 + (1-\theta)x_2 \mid 0 \leq \theta \leq 1\}.$$

(If $x_1 = x_2$ the line segment reduces to a point: $[x_1, x_2] = \{x_1\}$.) This is illustrated in figure 1.1.

1.1.2 Affine sets

A set $C \subseteq \mathbf{R}^n$ is *affine* if the line through any two distinct points in C lies in C, *i.e.*, if for any $x_1,\ x_2 \in C$ and $\theta \in \mathbf{R}$, we have

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

Affine sets are also called *flats*.

If $x_1, \ldots, x_k \in C$, where C is an affine set, and $\theta_1 + \cdots + \theta_k = 1$, then the point $\theta_1 x_1 + \cdots + \theta_k x_k$ also belongs to C. This is readily shown by induction from the definition of an affine set. We will illustrate the idea for $k = 3$, leaving the general case to the reader. Suppose that $x_1, x_2, x_3 \in C$, and $\theta_1 + \theta_2 + \theta_3 = 1$. 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 $\theta_3 \neq 1$. Then we can write

$$y = \mu_1(\mu_2 x_1 + (1-\mu_2)x_2) + (1-\mu_1)x_3 \tag{1.1}$$

where $\mu_1 = 1 - \theta_3$ and $\mu_2 = \theta_1/(1-\theta_3)$. Since C is affine and $x_1, x_2 \in C$, we conclude that $\mu_2 x_1 + (1-\mu_2)x_2 \in C$. Since this point and x_3 are in C, (1.1) shows that $y \in C$. This is illustrated in figure 1.2.

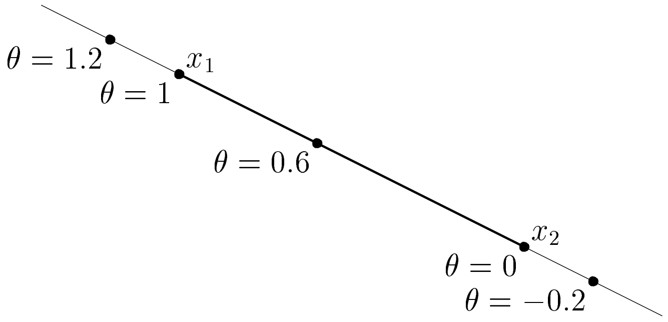

Figure 1.1: Line passing through x_1 and x_2, described parametrically as $\theta x_1 + (1-\theta)x_2$, where θ varies over $\mathbf{R}$. The line segment between x_1 and x_2, which corresponds to θ between 0 and 1, is shown darker.

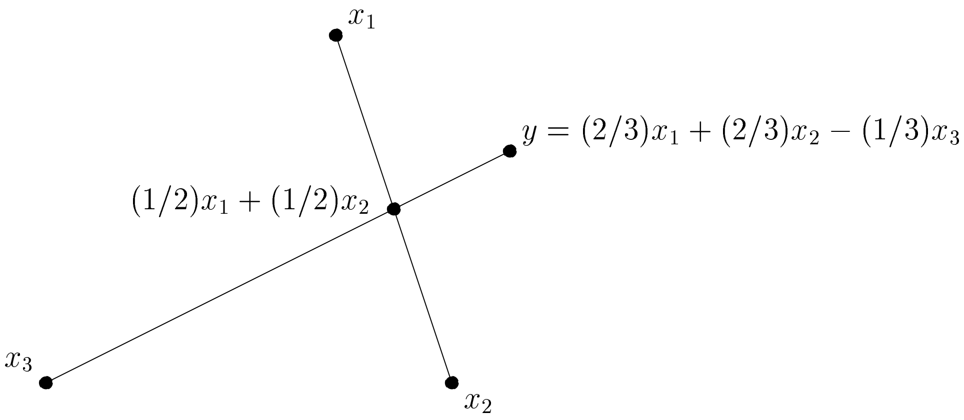

Figure 1.2: Points x_1, x_2, and x_3, and the affine combination $y = (2/3)x_1 + (2/3)x_2 - (1/3)x_3$. y can be expressed as $y = (4/3)((1/2)x_1 + (1/2)x_2) - (1/3)x_3$.

A point of the form $\theta_1 x_1 + \cdots + \theta_k x_k$, where $\theta_1 + \cdots + \theta_k = 1$, is called an *affine combination* of the points $x_1, \ldots, x_k$. The construction above shows that a set is affine if and only if it contains every affine combination of its points.

If C is an affine set and $x_0 \in C$, then the set

$$V = C - x_0 = \{x - x_0 | x \in C\}$$

is a subspace. Indeed, if $v_1, v_2 \in V$, then for all α, β

$$\alpha v_1 + \beta v_2 + x_0 = \alpha(v_1 + x_0) + \beta(v_2 + x_0) + (1 - \alpha - \beta)x_0 \in C,$$

and therefore $\alpha v_1 + \beta v_2 \in V$, *i.e.*, V is a subspace. The affine set C can then be expressed as

$$C = V + x_0 = \{v + x_0 | v \in V\},$$

i.e., as a subspace plus an offset. We define the *dimension* of an affine set C as the dimension of the subspace $C - x_0$, where x_0 is any element of C. (It can be shown that this dimension does not depend on the choice of x_0.)

The set of all affine combinations of points in some set C is called the *affine hull* of C, and denoted **aff** C:

$$\mathbf{aff}\, C = \{\theta_1 x_1 + \cdots + \theta_k x_k | x_i \in C,\ \theta_1 + \cdots + \theta_k = 1\}. \tag{1.2}$$

The affine hull is the smallest affine set that contains C, in the following sense: if S is any affine set with $C \subseteq S$, then $\mathbf{aff}\, C \subseteq S$.

1.1.3 Affine dimension and relative interior

We define the *affine dimension* of a set C as the dimension of its affine hull. Affine dimension is useful in the context of convex analysis and optimization, but is not always consistent with other definitions of dimension. As an example consider the unit circle in $\mathbf{R}^2$, *i.e.*, $\{x \in \mathbf{R}^2 | x_1^2 + x_2^2 = 1\}$. Its affine hull is all of $\mathbf{R}^2$, and hence its affine dimension is 2. By most definitions of dimension (*e.g.*, topological, Hausdorf), the unit circle has dimension 1.

If the affine dimension of a set $C \subseteq \mathbf{R}^n$ is smaller than n, then the set lies in the affine set $\mathbf{aff}\, C \neq \mathbf{R}^n$. We define the *relative interior* of the set C, denoted $\mathbf{relint}\, C$, as its interior relative to $\mathbf{aff}\, C$:

$$\mathbf{relint}\, C = \{x \in C \mid B(x, \epsilon) \cap \mathbf{aff}\, C \subseteq C \text{ for some } \epsilon > 0\},$$

where $B(x, \epsilon)$ is the ball of radius ϵ and center x, *i.e.*, $B(x, \epsilon) = \{y | \|x - y\| \leq \epsilon\}$. We can then define the *relative boundary* of a set C as $\overline{C} - \mathbf{relint}\, C$, where $\overline{C}$ is the closure of C.

Example. Consider for example a square in the (x_1, x_2)-plane in $\mathbf{R}^3$, defined as

$$C = \{x \in \mathbf{R}^3 \mid -1 \leq x_1 \leq 1,\ -1 \leq x_2 \leq 1,\ x_3 = 0\}.$$

The affine hull is the (x_1, x_2)-plane, *i.e.*,

$$\mathbf{aff}\, C = \{x \in \mathbf{R}^3 \mid x_3 = 0\}.$$

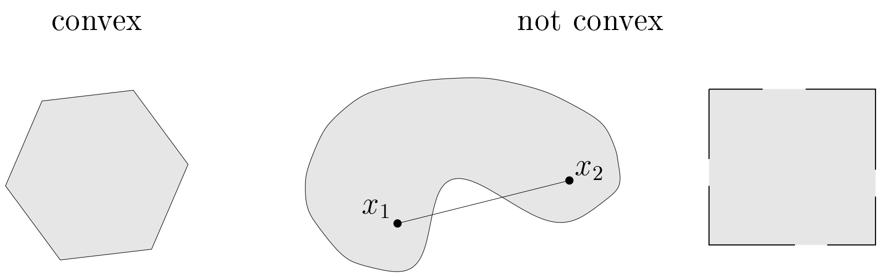

Figure 1.3: Some convex and non-convex sets

The interior of C is empty, but the relative interior is

$$\textbf{relint}\, C = \{x \in \mathbf{R}^3 \mid -1 < x_1 < 1,\ \ -1 < x_2 < 1,\ x_3 = 0\}.$$

Its boundary (in $\mathbf{R}^3$) is itself; its relative boundary is the wire-frame outline,

$$\{x \in \mathbf{R}^3 \mid \max\{|x_1|, |x_2|\} = 1,\ x_3 = 0\}.$$

1.1.4 Convex sets

A set C is *convex* if the line segment between any two points in C lies in C, *i.e.*, if for any $x_1,\ x_2 \in C$ and any θ with $0 \leq \theta \leq 1$, we have

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

Evidently every affine set is convex. Figure 1.3 shows some convex and nonconvex sets.

We call a point of the form $\theta_1 x_1 + \cdots + \theta_k x_k$, where $\theta_1 + \cdots + \theta_k = 1$ and $\theta_i \geq 0$, $i = 1, \ldots, k$, a *convex combination* of the points $x_1, \ldots, x_k$. Using the construction described above for an affine combination of more than two points, it can be shown that a set is convex if and only if it contains every convex combination of its points.

A convex combination of points can be thought of as a mixture or composition of the points, with $100\theta_i$ the percentage of x_i in the mixture. Figure 1.4 shows an example.

The *convex hull* of a set S, denoted $\mathbf{Co}S$, is the set of all convex combinations of points in S:

$$\mathbf{Co}S = \{\theta_1 x_1 + \cdots + \theta_k x_k | x_i \in S,\ \theta_i \geq 0,\ \theta_1 + \cdots + \theta_k = 1\}.$$

As the name suggests, $\mathbf{Co}S$ is convex. It is the smallest convex set that contains S: If C is convex and $C \supseteq S$, then $C \supseteq \mathbf{Co}S$. Figure 1.5 illustrates the definition of convex hull.

The idea of a convex combination can be generalized to include infinite sums, integrals, and, in the most general form, probability distributions. Suppose $\theta_1, \theta_2, \ldots$ satisfy

$$\theta_i \geq 0, \qquad \sum_{i=1}^{\infty} \theta_i = 1,$$

and $x_1, x_2, \ldots \in C$, which is convex. Then

$$\sum_{i=1}^{\infty} \theta_i x_i \in C,$$

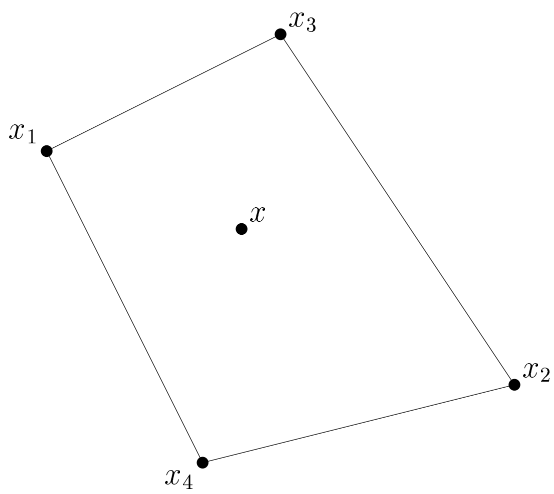

Figure 1.4: Four points x_1, x_2, x_3, x_4, and a convex combination $x = (1/4)x_1 + (1/6)x_2 + (1/3)x_3 + (1/4)x_4$.

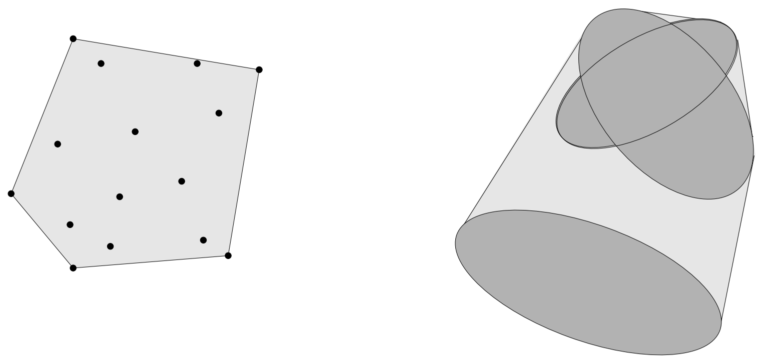

Figure 1.5: The convex hulls of two sets in $\mathbf{R}^2$. The left figure shows the convex hull of a set of fifteen points. The right figure shows the convex hull of the union of three ellipsoids.

if the series converges. More generally, suppose $p : C \to \mathbf{R}$ satisfies $p(x) \geq 0$ for all $x \in C$ and $\int_C p(x)\, dx = 1$, where C is convex. Then

$$\int p(x)x\, dx \in C$$

if the integral exists.

In the most general form, suppose C is convex and x is a random variable with $x \in C$ with probability one. Then $\mathbf{E}\, x \in C$. Indeed, this form includes all the others as special cases. For example, suppose the random variable x only takes on the two values x_1 and x_2, with $\mathbf{Prob}(x = x_1) = \theta$ and $\mathbf{Prob}(x = x_2) = 1-\theta$ (where $0 \leq \theta \leq 1$). Then $\mathbf{E}\, x = \theta x_1 + (1-\theta)x_2$; we are back to a simple convex combination of two points.

1.1.5 Cones

For given x_0 and $v \neq 0$, we call the set $\{x_0 + tv \mid t \geq 0\}$ the *ray* with *direction* v and *base* x_0. A set C is called a *cone* if for every $x \in C$ the ray with base 0 and direction x lies in C, *i.e.*, for any $x \in C$ and $\theta \geq 0$ we have $\theta x \in C$.

A set C is a *convex cone* if for any $x_1,\ x_2 \in C$ and $\theta_1,\ \theta_2 \geq 0$, we have

$$\theta_1 x_1 + \theta_2 x_2 \in C.$$

A point of the form $\theta_1 x_1 + \cdots + \theta_k x_k$ with $\theta_1, \ldots, \theta_k \geq 0$ is called a *conic combination* (or a *nonnegative linear combination*) of $x_1, \ldots, x_k$. If x_i are in a convex cone C, then every conic combination of x_i is in C. Conversely, a set C is a convex cone if and only if it contains all conic combinations of its elements.

The *conic hull* of S is the set of all conic combinations of points in S, *i.e.*,

$$\{\ \theta_1 x_1 + \cdots + \theta_k x_k \mid x_i \in S,\ \theta_i \geq 0\ \},$$

which is also the smallest convex cone that contains S.

1.2 Some important examples

In this section we describe some simple but important examples of convex sets which we will encounter throughout the rest of the book. Further details can be found in the appendix sections on linear algebra and on analytic geometry.

We start with some simple examples.

- The empty set $\emptyset$, any single point (*i.e.*, singleton) $\{x_0\}$, and the whole space $\mathbf{R}^n$ are affine (hence, convex) subsets of $\mathbf{R}^n$.
- Any line is affine.
- A line segment is convex, but not affine (unless it reduces to a point).
- A ray is convex, but not affine.
- Any subspace is affine, hence convex.

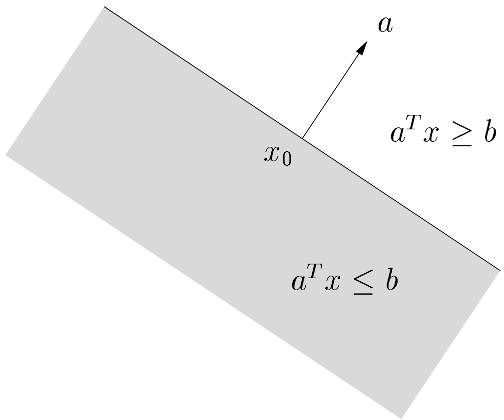

Figure 1.6: Hyperplane in $\mathbf{R}^2$. Two halfspaces.

1.2.1 Hyperplanes and halfspaces

A *hyperplane* is a set of the form

$$\{x \mid a^Tx = b\},$$

where $a \neq 0$. Analytically it is the solution set of a nontrivial linear equation among the components of x. Geometrically it can be interpreted as the set of points with a constant inner product to a given vector a, or as a hyperplane with normal vector a; the constant $b \in \mathbf{R}$ determines the 'offset' of the hyperplane from the origin. The hyperplane can also be expressed as

$$\{x \mid a^T(x - x_0) = 0\},$$

where x_0 is any point in the hyperplane (*i.e.*, satisfies $a^Tx_0 = b$). Hyperplanes are affine, hence also convex. These definitions are illustrated in figure 1.6.

A hyperplane divides $\mathbf{R}^n$ into two *halfspaces*. A (closed) halfspace is a set of the form

$$\{x \mid a^Tx \leq b\}, \tag{1.3}$$

where $a \neq 0$, *i.e.*, the solution set of one (nontrivial) linear inequality. This halfspace can also be expressed as

$$\{x \mid a^T(x - x_0) \leq 0\},$$

where x_0 is any point on the associated hyperplane, *i.e.*, satisfies $a^Tx_0 = b$ (see figure 1.7). The boundary of the halfspace (1.3) is the hyperplane $\{x \mid a^Tx = b\}$. The set $\{x \mid a^Tx < b\}$ is called an *open halfspace*. Halfspaces are convex, but not affine.

1.2.2 Euclidean balls and ellipsoids

A *(Euclidean) ball* (or just ball) in $\mathbf{R}^n$ has the form

$$B(x_c, r) = \{x \mid \|x - x_c\| \leq r\} = \{x \mid (x - x_c)^T(x - x_c) \leq r^2\}, \tag{1.4}$$

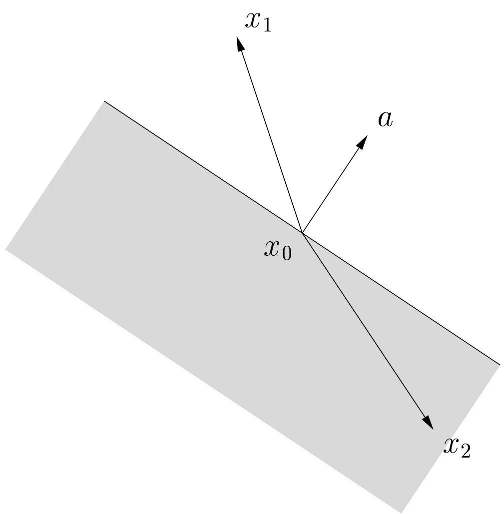

Figure 1.7: The shaded set is the halfspace $\mathcal{H} = \{x \mid a^T(x - x_0) \leq 0\}$. The vector $x_1 - x_0$ makes an acute angle with a, and so x_1 is not in $\mathcal{H}$. The vector $x_2 - x_0$ makes an obtuse angle with a, and so x_2 is in $\mathcal{H}$.

where $r > 0$. The vector x_c is the *center* of the ball and the scalar r is its *radius*; $B(x_c, r)$ consists of all points within a distance r of the center x_c.

A Euclidean ball is a convex set: if $\|x_1 - x_c\| \leq r$ and $\|x_2 - x_c\| \leq r$, then

$$\begin{aligned} \|\theta x_1 + (1-\theta)x_2 - x_c\| &= \|\theta(x_1 - x_c) + (1-\theta)(x_2 - x_c)\| \\ &\leq \theta\|x_1 - x_c\| + (1-\theta)\|x_2 - x_c\| \leq r. \end{aligned}$$

(Here we use the homogeneity property and triangle inequality for $\|\cdot\|$; see §A.1.2.)

A related family of convex sets is the *ellipsoids*, which have the form

$$\mathcal{E} = \{x \mid (x - x_c)^T P^{-1}(x - x_c) \leq 1\} \tag{1.5}$$

where $P = P^T \succ 0$. The vector $x_c \in \mathbf{R}^n$ is the *center* of the ellipsoid. The matrix P determines how far the ellipsoid extends in every direction from x_c; the lengths of the semi-axes of $\mathcal{E}$ are given by $\sqrt{\lambda_i}$, where λ_i are the eigenvalues of P. A ball is an ellipsoid with $P = r^2 I$.

Figure 1.8 shows an ellipsoid in $\mathbf{R}^2$.

1.2.3 Norm balls and cones

Suppose $\|\cdot\|$ is any norm on $\mathbf{R}^n$ (see §A.1.2). From the general properties of norms it is easily shown that a *norm ball* of radius r and center x_c, given by $\{x \mid \|x - x_c\| \leq r\}$, is convex.

The *norm cone* associated with the norm $\|\cdot\|$ is the set

$$C = \{(x, t) \mid \|x\| \leq t\}.$$

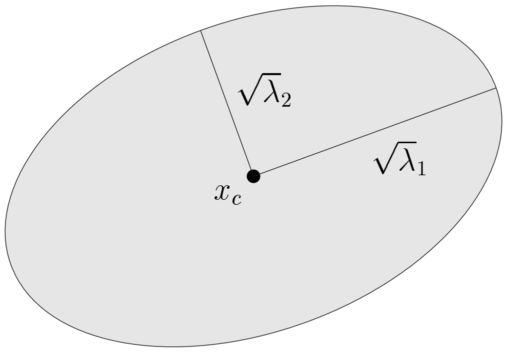

Figure 1.8: Ellipsoid in $\mathbf{R}^2$ with the two semi-axes.

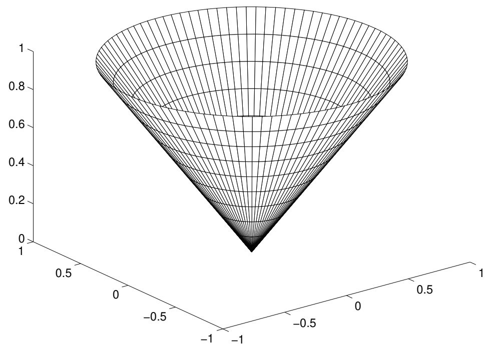

Figure 1.9: Boundary of the second-order cone in $\mathbf{R}^3$.

It is (as the name implies) a convex cone.

Example. *The second-order cone* is the norm cone for the Euclidean norm, *i.e.*,

$$\begin{aligned} C &= \left\{ (x,t) \in \mathbf{R}^{n+1} \mid \|x\| \le t \right\} \\ &= \left\{ (x,t) \;\middle|\; \begin{bmatrix} x \\ t \end{bmatrix}^T \begin{bmatrix} I & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ t \end{bmatrix} \le 0,\ t \ge 0 \right\}. \end{aligned}$$

The second-order cone is also known by several other names. It is called the *quadratic cone*, since it is defined by a quadratic inequality. (It is also called the *Lorentz cone* or *ice-cream cone.*) This is illustrated in figure 1.9.

1.2.4 Polyhedra and polytopes

A *polyhedron* is defined as the solution set of a finite number of linear equalities and inequalities:

$$\mathcal{P} = \{x \mid a_j^T x \le b_j,\ j = 1, \ldots, m,\ c_j^T x = d_j,\ j = 1, \ldots, p\} \tag{1.6}$$

A polyhedron is thus the intersection of a finite number of halfspaces and hyperplanes. Affine sets (*e.g.*, subspaces, hyperplanes, lines), rays, line segments, and halfspaces are all

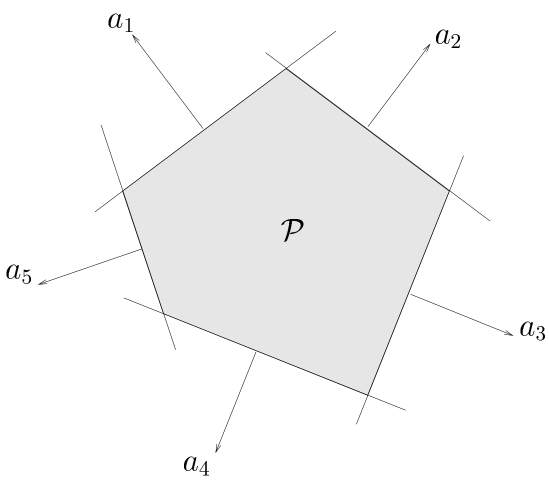

Figure 1.10: The polyhedron $\mathcal{P}$ is the intersection of five halfspaces.

polyhedra. It is easily shown that polyhedra are convex sets. A bounded polyhedron is sometimes called a *polytope*, but some authors use the opposite convention (*i.e.*, polytope for any set of the form (1.6), and polyhedron when it is bounded). Figure 1.10 shows an example of a polyhedron defined as the intersection of five halfspaces.

It will be convenient to use the compact notation

$$\mathcal{P} = \{x \mid Ax \preceq b,\ Cx = d\}, \tag{1.7}$$

for (1.6), where

$$A = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix}, \quad C = \begin{bmatrix} c_1^T \\ \vdots \\ c_p^T \end{bmatrix},$$

and the symbol $\preceq$ denotes *vector inequality* or *componentwise inequality* in $\mathbf{R}^m$: $u \preceq v$ means $u_i \leq v_i$ for $i = 1, \ldots, m$.

Example. The *nonnegative orthant* is the set of points with nonnegative components, *i.e.*,

$$\mathbf{R}_+^n = \{\, x \in \mathbf{R}^n \mid x_i \geq 0, i = 1, \ldots, n \,\} = \{\, x \in \mathbf{R}^n \mid x \succeq 0 \,\}.$$

It is a polyhedron and a cone (hence, called a *polyhedral cone*).

Simplexes

Simplexes are another important family of polyhedra. Suppose the $k+1$ points $v_0, \ldots, v_k \in \mathbf{R}^n$ are affinely independent, which means $v_1 - v_0, \ldots, v_k - v_0$ are linearly independent. The simplex determined by them is given by

$$C = \mathbf{Co}\{v_0, \ldots, v_k\} = \left\{ \theta_0 v_0 + \cdots + \theta_k v_k \;\middle|\; \theta_i \geq 0,\ i = 0, \ldots, k,\ \sum_{i=0}^{k} \theta_i = 1 \right\}, \tag{1.8}$$

i.e., the set of all convex combinations of the $k+1$ points $v_0, \ldots, v_k$. The (affine) dimension of this simplex is k, so it is sometimes referred to as a k-dimensional simplex in $\mathbf{R}^n$.

Examples. A 1-dimensional simplex is a line segment; a 2 dimensional simplex is a triangle (including its interior); and a 3-dimensional simplex is a tetrahedron.

The *unit simplex* is the n-dimensional simplex determined by the zero vector and the unit vectors, *i.e.*, $e_1, \ldots, e_n, 0 \in \mathbf{R}^n$. It can be expressed as the set of vectors that satisfy

$$x \succeq 0, \quad \mathbf{1}^T x \leq 1,$$

where $\mathbf{1}$ denotes the vector all of whose components are one.

The *probability simplex* is the $(n-1)$-dimensional simplex determined by the unit vectors $e_1, \ldots, e_n \in \mathbf{R}^n$. It is the set of vectors that satisfy

$$x \succeq 0, \quad \mathbf{1}^T x = 1.$$

Vectors in the probability simplex correspond to possible probability distributions on a set with n elements.

To describe the simplex (1.8) as a polyhedron, *i.e.*, in the form (1.6), we proceed as follows. First we find F and g such that the affine hull of $v_0, \ldots, v_k$ can be expressed as

$$\left\{ \sum_{i=0}^{k} \theta_i v_i | \sum_{i=0}^{k} \theta_i = 1 \right\} = \{x | Fx = g\}.$$

Clearly every vector in the simplex C must satisfy the equality constraints $Fx = g$.

Now we note that affine independence of the points v_i implies that the matrix $A \in \mathbf{R}^{(n+1)\times(k+1)}$ given by

$$A = \left[\begin{array}{cccc} v_0 & v_1 & \cdots & v_k \\ 1 & 1 & \cdots & 1 \end{array} \right] = \left[\begin{array}{cccc} v_0 & v_1 - v_0 & \cdots & v_k - v_0 \\ 1 & 0 & \cdots & 0 \end{array} \right] \left[\begin{array}{cc} 1 & \mathbf{1}^T \\ 0 & I \end{array} \right]$$

has full column rank (*i.e.*, $k+1$). Therefore A has a left inverse $B \in \mathbf{R}^{(k+1)\times(n+1)}$.

Now $x \in \mathbf{R}^n$ is in the affine hull of $v_0, \ldots, v_k$ if and only if $(x, 1)$ is in the range of A, *i.e.*, $(x, 1) = A\theta$ for some $\theta \in \mathbf{R}^{k+1}$. Since the columns of A are independent, such vectors can be expressed in only one way, *i.e.*, $\theta = B(x, 1)$.

We claim that $x \in C$ if and only if it satisfies

$$x \in C \iff Fx = g, \quad \theta = B(x, 1) \succeq 0,$$

which is a set of linear equalities amd inequalities, and so describes a polyhedron. To see this, first suppose that $x \in C$. Then x is in the affine hull of $v_0, \ldots, v_k$, and so satisfies $Fx = g$. It can be represented $(x, 1) = A\theta$ for only one θ, $\theta = B(x, 1)$. Therefore $\theta \succeq 0$. The converse follows similarly.

Vertex description of polytopes and polyhedra

The convex hull of the set $\{v_1, \ldots, v_k\}$ is

$$\mathbf{Co}\{v_1, \ldots, v_k\} = \left\{ \theta_1 v^1 + \cdots + \theta_k v^k \ \mid\ \theta_1 + \cdots + \theta_k = 1,\ \theta_i \geq 0,\ i = 1, \ldots, k \right\},$$

i.e., the set of all convex combinations of $v^1, \ldots, v^k$. This set is a polyhedron, and bounded (*i.e.*, a polytope), but (except in special cases, *e.g.*, a simplex) it is not simple to express it in the form (1.6), *i.e.*, by a set of linear equalities and inequalities.

A generalization of this description is

$$\left\{ \theta_1 v_1 + \cdots + \theta_k v_k \;\middle|\; \sum_{i=1}^{m} \theta_i = 1,\; \theta_i \geq 0,\; i = 1, \ldots, m \right\}. \tag{1.9}$$

(Here we consider linear combinations of v_i, but only the first m coefficients are required to be nonnegative and sum to one.) In fact, every polyhedron can be represented in this form.

The question of how a polyhedron is represented is subtle, and has very important practical consequences. As a simple example consider the *unit cube* in $\mathbf{R}^n$,

$$C = \{\; x \mid |x_i| \leq 1,\; i = 1, \ldots, n \;\}.$$

C can be described in the form (1.6) with $2n$ linear inequalities, *i.e.*, $\pm e_i^T x \leq 1$. To describe it in the vertex form (1.9) requires at least 2^n points:

$$C = \mathbf{Co}\{\; v_1, \ldots, v_{2^n} \;\},$$

where $v_1, \ldots, v_{2^n}$ are the 2^n vectors all of whose components are 1 or -1. Thus the size of the two descriptions differs greatly, for large n. We will come back to this topic several times.

1.2.5 The positive semidefinite cone

The set of positive semidefinite matrices,

$$\{\; X \in \mathbf{R}^{n \times n} \mid X = X^T \succeq 0 \;\}$$

is a convex cone: if $\theta_1, \theta_2 \geq 0$ and $A = A^T \succeq 0$ and $B = B^T \succeq 0$, then $\theta_1 A + \theta_2 B \succeq 0$. This can be seen directly from the definition of positive semidefiniteness: for any $x \in \mathbf{R}^n$, we have

$$x^T(\theta_1 A + \theta_2 B)x = \theta_1 x^T A x + \theta_2 x^T B x \geq 0$$

if $A \succeq 0$, $B \succeq 0$ and θ_1, $\theta_2 \geq 0$.

1.3 Operations that preserve convexity

In this section we describe a number of operations that preserve convexity of sets, or allow us to construct convex sets from others. These operations, together with the simple examples described in §1.2, form a calculus of convex sets that is useful for determining or establishing convexity of sets.

1.3.1 Intersection

Convexity is preserved under intersection: if S_1 and S_2 are convex, then $S_1 \cap S_2$ is convex. This property extends to the intersection of an infinite number of sets: if S_α is convex for every $\alpha \in \mathcal{A}$, then $\bigcap_{\alpha \in \mathcal{A}} S_\alpha$ is convex.

A similar property holds for subspaces, affine sets, convex cones, *i.e.*,

$$S_\alpha \text{ is } \left(\begin{array}{l} \text{a subspace} \\ \text{affine} \\ \text{convex} \\ \text{convex cone} \end{array} \right) \text{ for } \alpha \in \mathcal{A} \Longrightarrow \bigcap_{\alpha \in \mathcal{A}} S_\alpha \text{ is } \left(\begin{array}{l} \text{a subspace} \\ \text{affine} \\ \text{convex} \\ \text{convex cone} \end{array} \right).$$

As a simple example, a polyhedron is the intersection of halfspaces and hyperplanes (which are convex), and hence is convex.

Example. The positive semidefinite cone

$$\mathcal{P} = \{ X \in \mathbf{R}^{n \times n} \mid X = X^T, \ X \succeq 0 \},$$

can be expressed as

$$\bigcap_{z \in \mathbf{R}^n, \ z \neq 0} \{ X \in \mathbf{R}^{n \times n} \mid X = X^T, \ z^T X z \geq 0 \}.$$

For each $z \neq 0$, $z^T X z$ is a (not identically zero) linear function of X, so the sets

$$\{ X \in \mathbf{R}^{n \times n} \mid X = X^T, \ z^T X z \geq 0 \}$$

are, in fact, halfspaces in $\mathbf{R}^{n \times n}$. Thus the positive semidefinite cone is the intersection of an infinite number of halfspaces, hence is convex.

Example. The set

$$S = \{ (a_0, \ldots, a_n) \in \mathbf{R}^{n+1} \mid |a_0 + a_1 t + a_2 t^2 + \cdots + a_n t^n| \leq 1.3 \text{ for } |t| \leq 1.3 \}$$

is convex. It is the intersection of an infinite number of halfspaces (two for each value of t)

$$-1.3 \leq (1, t, \ldots, t^n)^T (a_0, a_1, \ldots, a_n) \leq 1.3.$$

Example. The set

$$S = \{ a \in \mathbf{R}^m \mid |p(t)| \leq 1 \text{ for } |t| \leq \pi/3 \},$$

where $p(t) = \sum_{k=1}^m a_k \cos kt$, is convex. The definition and the set are illustrated in figure 1.11. The set S can be expressed as intersection of slabs: $S = \bigcap_{|t| \leq \pi/3} S_t$,

$$S_t = \{ a \mid -1 \leq (\cos t, \ldots, \cos mt)^T a \leq 1 \}.$$

In the examples above we establish convexity of a set by expressing it as a (possibly infinite) intersection of halfspaces. In fact, we will see later (in §1.5.1) that *every* closed convex set S is a (usually infinite) intersection of halfspaces: a closed convex set is the intersection of all halfspaces that contain it, *i.e.*,

$$S = \bigcap \{ \mathcal{H} \mid \mathcal{H} \text{ halfspace}, \ S \subseteq \mathcal{H} \}.$$

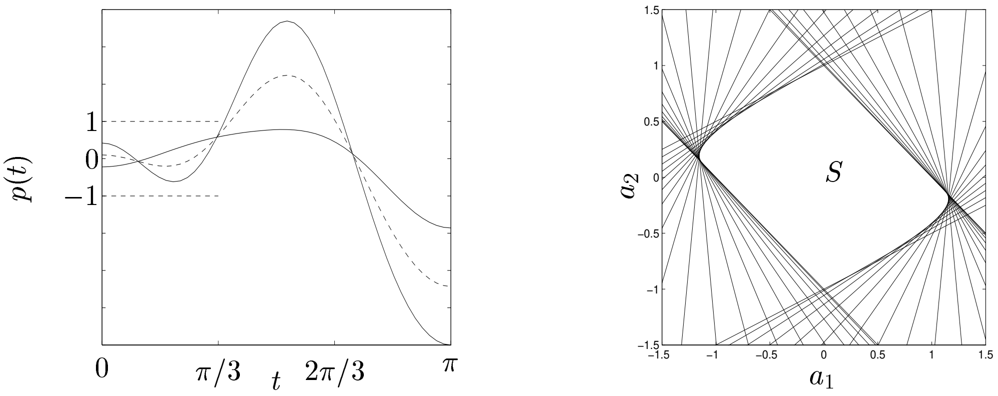

Figure 1.11: The elements of the set $S \subset \mathbf{R}^m$ are the coefficient vectors a of all trigonometric polynomials $p(t) = \sum_{k=1}^m a_k \cos kt$ that have magnitude less than one on the interval $|t| \le \pi/3$. The set is the intersection of an infinite number of slabs, hence convex.

1.3.2 Image and inverse image under affine functions

Suppose $S \subseteq \mathbf{R}^n$ is convex and $f : \mathbf{R}^n \to \mathbf{R}^m$ is an affine function. Then the image of S under f,

$$f(S) = \{f(x) \mid x \in S\},$$

is convex. Similarly, if $g : \mathbf{R}^k \to \mathbf{R}^n$ is an affine function, the *inverse image* of S under f,

$$f^{-1}(S) = \{x | f(x) \in S\}$$

is convex.

Two simple examples are *scaling* and *translation*. If $S \subseteq \mathbf{R}^n$ is convex, $\alpha \in \mathbf{R}$, and $a \in \mathbf{R}^n$, then the sets αS and $S + a$ are convex.

The *projection* of a convex set onto some of its coordinates is convex: if $S \subseteq \mathbf{R}^{m+n}$ is convex, then

$$T = \{x_1 \in \mathbf{R}^m \mid \exists\, x_2 \in \mathbf{R}^n,\ (x_1, x_2) \in S\}$$

is convex.

The *sum* of two sets is defined as

$$S_1 + S_2 = \{x + y \mid x \in S_1,\ y \in S_2\}.$$

If S_1 and S_2 are convex, then $S_1 + S_2$ is convex. To see this, if S_1 and S_2 are convex, then so is the direct or Cartesian product

$$S_1 \times S_2 = \{(x_1, x_2) \mid x_1 \in S_1,\ x_2 \in S_2\}.$$

The image of this set under the linear function $f(x_1, x_2) = x_1 + x_2$ is the sum $S_1 + S_2$.

In a similar way we can interpret the intersection of S_1 and S_2 as the image under the affine mapping $f(x, y) = x$ of the set

$$\{(x, y) \mid x \in S_1, y \in S_2, x = y\}$$

which is itself the inverse image of $\{0\}$ under the mapping $f(x,y) = x - y$, defined on $S_1 \times S_2$.

We can also consider the *(partial) sum* of $S_1 \in \mathbf{R}^{m+n}$ and $S_2 \in \mathbf{R}^{m+n}$, defined as

$$S = \{(x, y_1 + y_2) \mid (x, y_1) \in S_1,\ (x, y_2) \in S_2\},$$

where $x \in \mathbf{R}^n$ and $y_i \in \mathbf{R}^m$. For $m = 0$, the partial sum gives the intersection of S_1 and S_2; for $n = 0$, it is set addition. Partial sums of convex sets are convex; the proof is a straightforward extension of the proofs for sum and intersection.

Example. *Polyhedron.* The polyhedron $\{x | Ax \preceq b, Cx = d\}$ can be expressed as the inverse image of the Cartesian product of the nonnegative orthant and the origin under the affine function $f(x) = (b - Ax, d - Cx)$:

$$\{x | Ax \preceq b, Cx = d\} = \{x | f(x) \in \mathbf{R}_+^m \times \{0\}\}.$$

Example. *Solution set of linear matrix inequality.* The condition

$$A(x) = x_1 A_1 + \cdots + x_n A_n \preceq B, \tag{1.10}$$

where $B = B^T$, $A_i = A_i^T \in \mathbf{R}^{m \times m}$, is called a *linear matrix inequality* in x. (Note the similarity to an ordinary linear inequality,

$$a^T x = x_1 a_1 + \cdots + x_n a_n \leq b,$$

with $b,\ a_i \in \mathbf{R}$.)

The solution set of a linear matrix inequality, $\{x | A(x) \preceq B\}$, is convex. Indeed, it is the inverse image of the positive semidefinite cone under the affine function $f(x) = B - A(x)$.

Example. *Hyperbolic cone.* The set

$$\{x | x^T P x \leq (c^T x)^2,\ c^T x \geq 0\}$$

where $P = P^T \succ 0$ and $c \in \mathbf{R}^n$, is convex, since it is the inverse image of the second-order cone,

$$\{(z,t) | z^T z \leq t^2, t \geq 0\}$$

under the affine function $f(x) = (P^{1/2} x, c^T x)$.

Example. *Ellipsoid.* The ellipsoid

$$C = \{x \mid (x - x_c) P^{-1} (x - x_c) \leq 1\}.$$

where $P = P^T \succ 0$, is the image of the unit ball $\{u \mid \|u\| \leq 1\}$ under the affine mapping $f(u) = P^{1/2} u + x_c$. (It is also the inverse image of the unit ball under the affine mapping $g(x) = P^{-1/2}(x - x_c)$.)

1.3.3 Image and inverse image under linear-fractional and perspective functions

In this section we explore a class of functions, called linear-fractional, that is more general than affine but still preserves convexity.

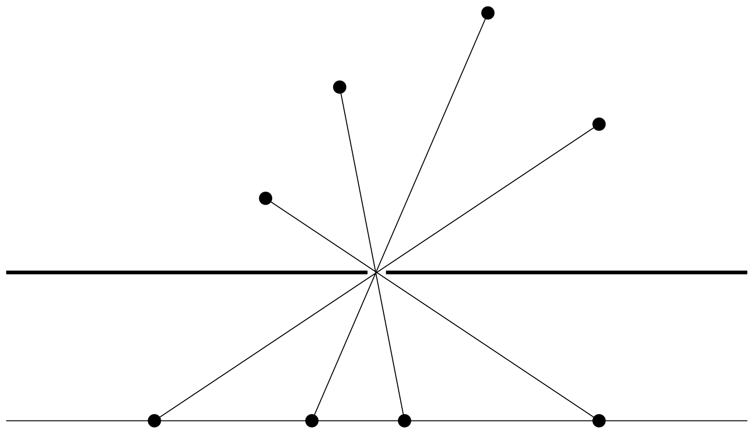

Figure 1.12: Interpretation of perspective function.

The perspective function

We define the *perspective function* $P : \mathbf{R}^{n+1} \to \mathbf{R}^n$, with $\mathbf{dom}\, P = \mathbf{R}^n \times \{t \in \mathbf{R} | t > 0\}$, as $P(z,t) = z/t$. The perspective function scales or normalizes vectors so the last component is one, and then drops the last component.

> **Example.** We can interpret the perspective function as the action of a *pin-hole camera*. A pin-hole camera (in $\mathbf{R}^3$) consists of an opaque horizontal plane $x_3 = 0$, with a single pin-hole at the origin, through which light can pass, and a horizontal image plane $x_3 = -1$. An object at x, above the camera (*i.e.*, with $x_3 > 0$), forms an image at the point $-(x_1/x_3, x_2/x_3, 1)$ on the image plane. Dropping the last component of the image point (since it is always -1), the image of a point at x appears at $y = -(x_1/x_3, -x_2/x_3) = -P(x)$ on the image plane. (See figure 1.12.)

If $C \subseteq \mathbf{dom}\, f$ is convex, then the image

$$P(C) = \{P(x) | x \in C\}$$

is convex. This result is certainly intuitive: a convex object, viewed through a pin-hole camera, yields a convex image.

To establish this fact we will show that line segments are mapped to line segments under the perspective function. (Which, again, makes sense: a line segment viewed through a pin-hole camera yields a line segment.) Suppose that $x,\ y \in \mathbf{R}^{n+1}$ with $x_{n+1} > 0$, $y_{n+1} > 0$. Then for $0 \le \theta \le 1$,

$$P(\theta x + (1-\theta)y) = \mu P(x) + (1-\mu)P(y),$$

where

$$\mu = \frac{\theta x_{n+1}}{\theta x_{n+1} + (1-\theta)y_{n+1}} \in [0,1].$$

This correspondence between θ and μ is monotonic; as varies between 0 and 1 (which sweeps out the line segment $[x,y]$), μ varies between 0 and 1 (which sweeps out the line segment

$[P(x), P(y)])$. This shows that $P([x,y]) = [P(x), P(y)]$. (Several other types of sets are preserved under the perspective mapping, *e.g.*, polyhedra, ellipsoids, and solution sets of linear matrix inequalities; see the exercises.)

Now suppose C is convex with $C \subseteq \mathbf{dom}\, P$ (*i.e.*, $x_{n+1} > 0$ for all $x \in C$), and $x,\ y \in C$. To establish convexity of $P(C)$ we need to show that the line segment $[P(x), P(y)]$ is in $P(C)$. But this line segment is the image of the line segment $[x, y]$ under P, and so lies in $P(C)$.

The inverse image of a convex set under the perspective function is also convex: if $C \subseteq \mathbf{R}^n$, then

$$P^{-1}(C) = \{(x,t) \in \mathbf{R}^{n+1} | x/t \in C,\ t > 0\}$$

is convex. To show this, suppose $(x,t) \in P^{-1}(C)$, $(y,s) \in P^{-1}(C)$, and $0 \leq \theta \leq 1$. We need to show that

$$\theta(x,t) + (1-\theta)(y,s) \in P^{-1}(C),$$

i.e., that

$$\frac{\theta x + (1-\theta)y}{\theta t + (1-\theta)s} \in C$$

$(\theta t + (1-\theta)s > 0$ is obvious). This follows from

$$\frac{\theta x + (1-\theta)y}{\theta t + (1-\theta)s} = \mu(x/t) + (1-\mu)(y/s),$$

where

$$\mu = \frac{\theta t}{\theta t + (1-\theta)s} \in [0,1].$$

Linear-fractional functions

Composing the perspective function with an affine function yields a general *linear-fractional function.* Suppose $g : \mathbf{R}^n \to \mathbf{R}^{m+1}$ is affine, *i.e.*,

$$g(x) = \begin{bmatrix} A \\ c^T \end{bmatrix} x + \begin{bmatrix} b \\ d \end{bmatrix}, \tag{1.11}$$

where $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$, $c \in \mathbf{R}^n$, and $d \in \mathbf{R}$. The function $f : \mathbf{R}^n \to \mathbf{R}^m$ given by $f = Pg$, *i.e.*,

$$f(x) = (Ax+b)/(c^Tx+d), \quad \mathbf{dom}\, f = \{x | c^Tx + d > 0\}, \tag{1.12}$$

is called linear-fractional (or *projective*) function. If $c = 0$, we can take the domain of f as $\mathbf{R}^n$, in which case f is affine. So we can think of affine and linear functions as special cases of linear-fractional functions.

Like the perspective function, linear-fractional functions preserve convexity. If C is convex and lies in the domain of f (*i.e.*, $c^Tx + d > 0$ for $x \in C$), then its image $f(C)$ is convex. This follows immediately from results above: the image of C under the affine mapping (1.11) is convex; and the image of the resulting set under the perspective function P (which yields $f(C)$) is convex. Similarly, if $C \subseteq \mathbf{R}^m$ is convex, then the inverse image $f^{-1}(C)$ is convex.

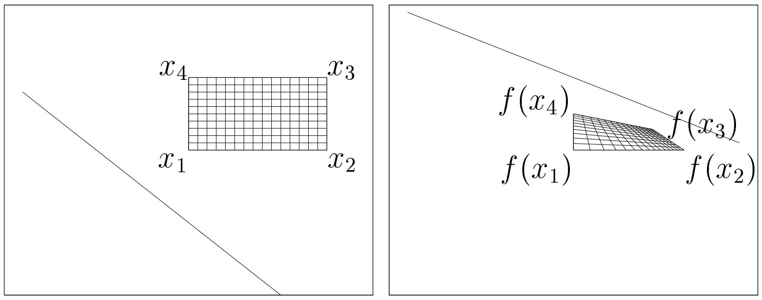

Figure 1.13: A grid in $\mathbf{R}^2$ (left) and its transformation under a linear-fractional function (right). The lines show the boundary of the domains.

Example. *Conditional probabilities.* Suppose u and v are random variables that take on the values $1, \ldots, n$ and $1, \ldots, m$, respectively, and let p_{ij} denote $\mathbf{Prob}(u = i, v = j)$. Then the conditional probability $f_{ij} = \mathbf{Prob}(u = i | v = j)$ is given by

$$f_{ij} = \frac{p_{ij}}{\sum_{k=1}^{m} p_{kj}}.$$

Thus f is obtained by a linear-fractional mapping from p.

It follows that if $\mathcal{P}$ is a convex set of joint probabilities for (u, v), then the associated set of conditional probabilities of u given v is also convex.

Figure 1.13 shows an example of a linear-fractional mapping from $\mathbf{R}^2$ to $\mathbf{R}^2$ and the image of a rectangle.

1.4 Convex cones and generalized inequalities

Convex cones can be used to define *generalized inequalities*, which are partial orderings in $\mathbf{R}^n$ that have many of the properties of the standard ordering on $\mathbf{R}$.

1.4.1 Partial order induced by a convex cone

Let $K \subseteq \mathbf{R}^n$ be a convex cone. Then K defines (or induces) a partial ordering on $\mathbf{R}^n$: we define $x \preceq_K y$ to mean $y - x \in K$. When $K = \mathbf{R}_+$, this is the usual ordering on $\mathbf{R}$, *i.e.*, $x \preceq_K y$ means $x \leq y$. We also write $x \succeq_K y$ for $y \preceq_K x$.

Example. The partial order induced by the nonnegative orthant, *i.e.*, $K = \mathbf{R}^n_+$, corresponds to componentwise inequality between vectors. This partial ordering arises so frequently that we drop the subscript $\mathbf{R}^n_+$; it is understood when the symbol $\preceq_K$ appears between vectors.

Example. The positive semidefinite cone $\{X = X^T \in \mathbf{R}^{n\times n} \mid X \succeq 0\}$ defines a partial order on the space of symmetric matrices: $X \preceq_K Y$ if and only if $Y - X$ is positive semidefinite. Here, too, the partial ordering arises so frequently that we drop the subscript: for symmetric matrices we write simply $X \preceq Y$.

Example. Let K be defined as

$$K = \{c \in \mathbf{R}^n \mid c_1 + c_2 t + \cdots + c_n t^{n-1} \geq 0 \text{ for } t \in [0,1]\}, \tag{1.13}$$

i.e., K is the cone of (coefficients of) polynomials of degree $n-1$ that are nonnegative on the interval $[0,1]$. Two vectors $c, d \in \mathbf{R}^n$ satisfy $c \preceq_K d$ if and only if

$$c_1 + c_1 t + \cdots + c_n t^{n-1} \leq d_1 + d_1 t + \cdots + d_n t^{n-1}$$

for all $t \in [0,1]$.

Example. Let $K \subseteq \mathbf{R}^n$ be a subspace (hence, also a cone). Then $x \preceq_K y$ means $y - x \in K$. In this example the partial ordering is symmetric: $x \preceq_K y$ if and only if $y \preceq_K x$. In fact here $\preceq_K$ is an equivalence relation (despite the asymmetry suggested by the symbol $\preceq_K$).

Remark. The partial ordering $\preceq_K$ satisfies the following properties:

- *preserved under addition*: if $x \preceq_K y$ and $u \preceq_K v$ then $x + u \preceq_K y + v$. This implies *transitivity*: if $x \preceq_K y$ and $y \preceq_K z$ then $x \preceq_K z$.
- *preserved under nonnegative scaling*: if $x \preceq_K y$ and $\alpha \geq 0$ then $\alpha x \preceq_K \alpha y$.
- $x \preceq_K x$.

Conversely, let $\preceq$ denote any partial order that satisfies these properties, and define $K = \{x \mid x \succeq 0\}$. Then K is a convex cone, which induces the partial ordering $\preceq$.

1.4.2 Generalized inequalities

The partial ordering induced by a cone is especially useful when the cone satisfies some special properties:

- K is closed.
- K has nonempty interior.
- K is *pointed*, which means that K contains no line (or equivalently, $x \in K, \ -x \in K \ \Rightarrow \ x = 0$).

In this case we refer to the induced partial ordering $\preceq_K$ as a *generalized inequality* (associated with K).

For a generalized inequality we define the associated *generalized strict inequality* by

$$x \prec_K y \iff y - x \in \mathbf{int}\, K.$$

(We refer to $\preceq_K$ as the generalized *nonstrict inequality* if we need to distinguish it from $\prec_K$.)

The first two cones described above, $\mathbf{R}^n_+$ and the cone of positive semidefinite matrices, induce generalized inequalities. For $K = \mathbf{R}^n_+$, $x \prec_K y$ means $x_i < y_i$ for $i = 1, \ldots, n$, *i.e.*, componentwise strict inequality between vectors. When K is the positive semidefinite cone, $\mathbf{int}\, K$ is the set of positive definite matrices, and the associated strict inequality $X \prec_K Y$ means $Y - X$ is positive definite, *i.e.*, $X \prec Y$. These two generalized inequalities occur so often that we drop the subscript: for x and y vectors, $x \prec y$ means componentwise strict inequality, and for X and Y symmetric matrices, $X \prec Y$ means $Y - X$ is positive definite.

(Strict and nonstrict) generalized inequalities satisfy several further properties in addition to those listed above, for example:

- if $x \preceq_K y$ and $y \preceq_K x$, then $x = y$
- if $x_i \to x$, $y_i \to y$ as $i \to \infty$, and $x_i \preceq_K y_i$, then $x \preceq_K y$
- if $x \prec_K y$, then for u and v small enough we have $x + u \prec_K y + v$
- $x \not\prec_K x$
- if $x \prec_K y$, $u \preceq_K v$, then $x + u \prec_K y + v$
- if $x \prec_K y$ and $t > 0$, then $tx \prec_K ty$
- $\{z | x \preceq_K z \preceq_K y\}$ is bounded

These and other properties are explored in the exercises.

1.4.3 Minimum and minimal elements

The notation of generalized inequality (*i.e.*, $\preceq_K$, $\prec_K$) is meant to suggest the analogy to ordinary inequality on $\mathbf{R}$ (*i.e.*, $\leq$, $<$). While many properties of ordinary inequality do hold for generalized inequalities, some important ones do not. The most obvious difference is that $\leq$ on $\mathbf{R}$ is a *linear ordering*: any two points are *comparable*, meaning either $x \leq y$ or $y \leq x$. This property does not hold for other generalized inequalities. One implication is that concepts such as minimum, maximum, and (lower and upper) bounds are more complicated in the context of generalized inequalities. We briefly discuss this in this section.

We say that $x \in S$ is the *minimum* element of S (with respect to the generalized inequality $\preceq_K$) if for every $y \in S$ we have $x \preceq_K y$. We define the maximum element of a set S, with respect to a generalized inequality, in a similar way. If a set has a minimum (maximum) element, then it is unique.

A related concept is *minimal element*. We say that $x \in S$ is a *minimal* element of S (with respect to the generalized inequality $\preceq_K$) if $y \in S$, $y \preceq_K x$ only if $y = x$. We define maximal element in a similar way. A set can have many different minimal (maximal) elements.

We can describe minimal and maximal elements using simple set notation. A point $x \in S$ is the minimum element of S if and only if

$$S \subseteq x + K.$$

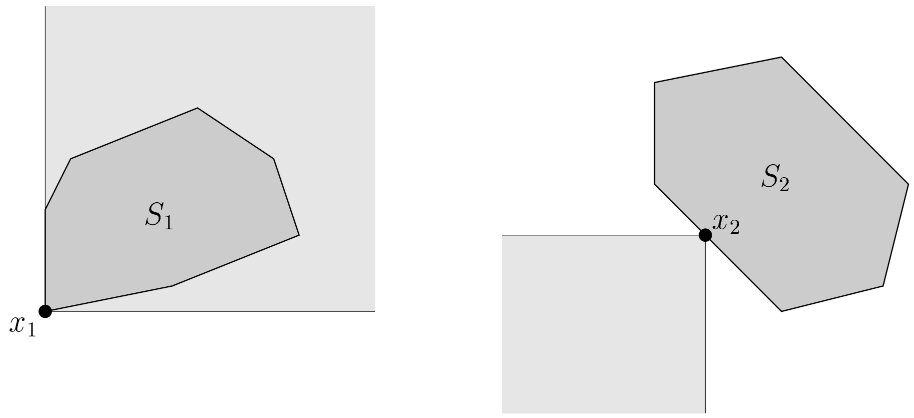

Figure 1.14: The set S_1 in the lefthand figure has a minimum element x_1. The set $x_1 + K$ is shaded lightly; x_1 is the minimum element of S_1 since $S_1 \subseteq x + K$. The righthand figure shows another set S_2 and a minimal point x. The set $x_2 - K$ is shown lightly shaded. x_2 is minimal because $x_2 - K$ and S_2 intersect only at x_2.

Here $x + K$ denotes all the points that are comparable to x and larger than or equal to x (according to $\preceq_K$). A point $x \in S$ is a minimal element if and only if

$$(x - K) \cap S = \{x\}.$$

Here $x - K$ denotes all the points that are comparable to x and less than or equal to x (according to $\preceq_K$); the only point in common with S is x.

For $K = \mathbf{R}_+$, which induces the usual ordering on $\mathbf{R}$, the concepts of minimal and minimum are the same, and agree with the usual definition of the minimum element of a set.

Example. Consider the cone $\mathbf{R}^2_+$, which induces componentwise inequality in $\mathbf{R}^2$. Here we can give some simple geometric descriptions of minimal and minimum elements. The inequality $x \preceq y$ means y is above and to the right of x. To say that x is the minimum element of a set S means that S contains a point such that all other points of S lie above and to the right. To say that x is a minimal element of a set S means that no other point of S lies to the left or below x. This is illustrated in figure 1.14.

Example. *Minimum and minimal elements of a set of symmetric matrices.* We associate with each positive definite $A = A^T \in \mathbf{R}^{n \times n}$ an ellipsoid centered at the origin, given by

$$E_A = \{\ x \mid x^T A^{-1} x \leq 1\ \}.$$

We have $A \preceq B$ if and only if $E_A \subseteq E_B$.

Let $v_1, \ldots, v_k \in \mathbf{R}^n$ be given and define

$$S = \{\ P = P^T \mid v_i^T P^{-1} v_i \leq 1,\ P \succ 0\ \},$$

which corresponds to the set of ellipsoids that contain the points $v_1, \ldots, v_k$. S does not have a minimum element; for any ellipsoid that contains the v_i we can find another

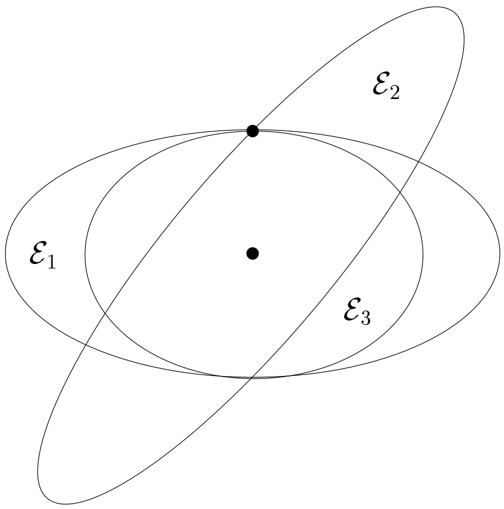

Figure 1.15: A point in $\mathbf{R}^2$ and three ellipsoids that contain it. The ellipsoid $\mathcal{E}_1$ is not minimal, since there exist enclosing ellipsoids that are smaller (*e.g.*, $\mathcal{E}_3$). $\mathcal{E}_3$ is not minimal for the same reason. The ellipsoid $\mathcal{E}_2$ is minimal, since no other enclosing ellipsoid is contained in $\mathcal{E}_2$.

one that is not comparable to it. An ellipsoid is minimal if it contains the points, but no smaller ellipsoid does. Figure 1.15 shows an example in $\mathbf{R}^2$ with $k = 1$.

We say x is a *lower bound* on a set S if $x \preceq_K y$ for all $y \in S$. Similarly, x is an *upper bound* on a set S if $x \succeq_K y$ for all $y \in S$. We say a set S is *bounded below* (or bounded above) if there is an x (not necessarily in S) which is a lower (upper) bound for S. We say that S is bounded if it is bounded below and bounded above. (It can be shown that this is the same as the usual notion of (norm) bounded, *i.e.*, that there is an M such that $\|y\| \leq M$ for all $y \in S$; see exercises.)

A basic result of real analysis is that on $\mathbf{R}$, a set which is closed and bounded below has a minimum element. For generalized inequalities, however, this assertion is false: a set can be closed and bounded below, but not have a minimum element. The correct extension to generalized inequalities is: a set which is closed and bounded below has a *minimal* element.

1.5 Separating and supporting hyperplanes

1.5.1 Separating hyperplanes

We briefly mention an idea that will be important later: the use of hyperplanes or affine functions to separate convex sets that do not intersect. The *separating hyperplane theorem* is the following: Suppose C and D are two convex sets that do not intersect, *i.e.*, $C \cap D = \emptyset$. Then there exist $a \neq 0$ and b such that $a^T x \leq b$ for all $x \in C$ and $a^T x \geq b$ for all $x \in D$. In other words, the affine function $a^T x - b$ is nonpositive on C and nonnegative on D. The

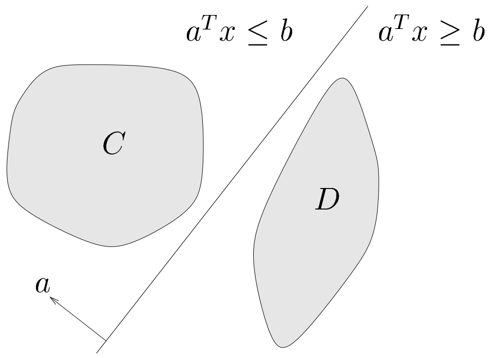

Figure 1.16: A hyperplane separating disjoint convex sets C and D. The affine function $a^Tx - d$ is nonpositive on C and nonnegative on D.

hyperplane $\mathcal{H} = \{x \mid a^Tx = d\}$ is called a *separating hyperplane* for the sets C and D. This is illustrated in figure 1.16.

We will not give a general proof of the separating hyperplane theorem here, but only consider a special case that illustrates the idea. We assume that the *distance* between C and D, defined as

$$\mathbf{dist}(C, D) = \inf\{\|u - v\| \mid u \in C, v \in D\},$$

is positive, and that there exist points $c \in C$ and $d \in D$ that achieve the minimum distance, *i.e.*, $\|c - d\| = \mathbf{dist}(C, D)$. (These conditions are satisfied, for example, when C and D are closed and one set is bounded.)

Define

$$a = c - d, \quad b = \frac{\|c\|^2 - \|d\|^2}{2}.$$

We will show that the affine function

$$f(x) = a^Tx - b = (d - c)^T(x - \frac{d + c}{2})$$

is nonpositive on C and nonnegative on D, *i.e.*, separates C and D. The corresponding hyperplane $f(x) = 0$ is perpendicular to the line segment $[c, d]$ and passes through its midpoint; see figure 1.17.

We first show that f is nonnegative on D. The proof that f is nonpositive on C is similar (or follows by swapping C and D and considering $-f$). Suppose there were a point $u \in D$ for which

$$f(u) = (d - c)^T(u - (d + c)/2) < 0. \tag{1.14}$$

We can express $f(u)$ as

$$f(u) = (d - c)^T(u - d + \frac{1}{2}(d - c)) = (d - c)^T(u - d) + \frac{1}{2}\|d - c\|^2.$$

We see that (1.14) implies $(d - c)^T(u - d) < 0$, *i.e.*, the line segments $[c, d]$ and $[d, u]$ make an acute angle. For small $t > 0$ we therefore have

$$\|d + t(u - d) - c\| < \|d - c\|,$$

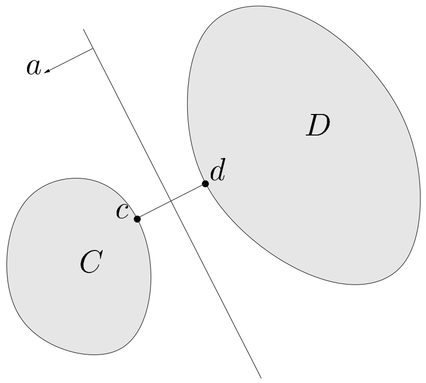

Figure 1.17: Construction of a separating hyperplane between two convex sets.

i.e., there are points on the line segment $[d, u]$ that lie closer to c than d. This contradicts the assumption that $\mathbf{dist}(C, D) = \|c - d\|$, since, by convexity of C, the entire line segment $[d, u]$ lies in C,

Example. *Separation of an affine and a convex set.* Suppose C is convex and D is affine, *i.e.*, $D = \{Fu + g \mid u \in \mathbf{R}^m\}$ where $F \in \mathbf{R}^{n\times m}$. Suppose C and D are disjoint, so by the separating hyperplane theorem there are $a \neq 0$ and b such that $a^T x \leq b$ for all $x \in C$ and $a^T x \geq b$ for all $x \in D$.

Now $a^T x \geq b$ for all $x \in D$ means $a^T F u \geq b - a^T g$ for all $u \in \mathbf{R}^m$. But a linear function is only bounded below on $\mathbf{R}^m$ when it is zero, so we conclude $a^T F = 0$ (and hence, $b \leq a^T g$).

Thus we conclude that there exists $a \neq 0$ such $F^T a = 0$ and $a^T g \leq a^T x$ for all $x \in C$.

Strict separation

The hyperplane we have constructed satisfies the stronger condition $a^T x < b$ for all $x \in C$ and $a^T x > b$ for all $x \in D$, which is called *strict separation* of the sets C and D. Simple examples show that in general, disjoint convex sets need not be strictly separable by a hyperplane. In many special cases, however, strict separation can be established.

Example. *Strict separation of a point and a closed convex set.* Let C be a closed convex set and $x_0 \notin C$. Then there exists a hyperplane that strictly separates x_0 from C.

To see this, note that the two sets C and $B(x_0, \epsilon)$ do not intersect for some $\epsilon > 0$. By the separating hyperplane theorem, there exist $a \neq 0$ and b such that $a^T x \leq b$ for $x \in C$ and $a^T x \geq b$ for $x \in B(x_0, \epsilon)$.

Using $B(x_0, \epsilon) = \{x_0 + u | \|u\| \leq \epsilon\}$, the second condition can be expressed as

$$a^T(x_0 + u) \geq b \quad \text{for all} \quad \|u\| \leq \epsilon.$$

By the Cauchy-Schwarz inequality, the u that minimizes the lefthand side is $u = -\epsilon a/\|a\|$, so we have

$$a^T x_0 - \epsilon\|a\| \geq b.$$

Therefore the affine function

$$f(x) = a^T x - b - \epsilon\|a\|/2$$

is negative on C and positive at x_0.

As an immediate consequence we can establish a fact that we already mentioned above: a closed convex set is the intersection of all halfspaces that contain it. Indeed, let C be closed and convex, and let S be the intersection of all halfspaces containing C. Obviously $x \in C \Rightarrow x \in S$. To show the converse, suppose there exists $x \in S$, $x \notin C$. By the strict separation result there exists a hyperplane that strictly separates x from C, *i.e.*, there is a halfspace containing C but not x. In other words, $x \notin S$.

Converse separating hyperplane theorems

The converse of the separating hyperplane theorem (*i.e.*, existence of a separating hyperplane implies that C and D do not intersect) is not true, unless one imposes additional constraints on C or D. As a simple counterexample, consider $C = D = \{0\} \subseteq \mathbf{R}$. Here the affine function $x = x$ separates C and D, since it is zero on both sets.

By adding conditions on C and D various converse separation theorems can be derived. As a very simple example, suppose C and D are convex sets, with C open, and there exists an affine function f that is nonpositive on C and nonnegative on D. Then C and D are disjoint. (To see this we first note that f must be positive on C; for if f were zero at a point of C then f would take on negative values near the point, which is a contradiction. But then C and D must be disjoint since f is positive on C and nonpositive on D.) Putting this converse together with the separating hyperplane theorem, we have the following result: for any two convex sets C and D, at least one of which is open, they are disjoint if and only if there exists a separating hyperplane.

Example. *Theorem of alternatives for strict linear inequalities.* We derive the necessary and sufficient conditions for solvability of a system of strict linear inequalities

$$Ax \prec b, \tag{1.15}$$

(*i.e.*, $a_i^T x < b_i \; i = 1, \ldots, m$). These inequalities are infeasible if and only if the (convex) sets

$$C = \{b - Ax \mid x \in \mathbf{R}^m\}, \quad D = \{y \mid y \succ 0\},$$

do not intersect. (Here a_i^T are the rows of A.) The set C is open; D is an affine set. Hence by the result above, C and D are disjoint if and only if there exists a separating hyperplane, *i.e.*, a nonzero $\lambda \in \mathbf{R}^m$ and $\mu \in \mathbf{R}$ such that $\lambda^T x \le \mu$ on C and $\lambda^T x \ge \mu$ on D.

Each of these conditions can be simplified. The first means $\lambda^T(b - Ax) \le \mu$ for all x. This implies (as above) that $A^T\lambda = 0$ and $\lambda^T b \le \mu$. The second inequality means $\lambda^T y \ge \mu$ for all $y \succ 0$. This implies $\mu \le 0$ and $\lambda \succeq 0$, $\lambda \ne 0$.

Putting it all together, we find that the set of strict inequalities (1.15) is infeasible if and only if there exists $\lambda \in \mathbf{R}^m$ such that

$$\lambda \ne 0, \quad \lambda \succeq 0, \quad A^T\lambda = 0, \quad \lambda^T b \le 0. \tag{1.16}$$

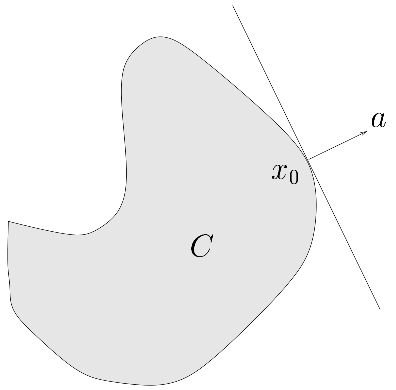

Figure 1.18: The hyperplane $\{x \mid a^T x = a^T x_0\}$ supports C at x_0.

This is also system of linear inequalities and linear equations in the variable $\lambda \in \mathbf{R}^m$. We say that (1.15) and (1.16) form a pair of *alternatives*: for any data A and b, exactly one of them is solvable.

1.5.2 Supporting hyperplanes

Suppose x_0 is a point in the boundary of a set $C \subseteq \mathbf{R}^n$. If $a \neq 0$ satisfies $a^T x \leq a^T x_0$ for all $x \in C$, then the hyperplane $\{x \mid a^T x = a^T x_0\}$ is called a *supporting hyperplane* to C at the point x_0. This is equivalent to saying that the point x_0 and the set C can be separated by a hyperplane; geometrically it means that the halfspace $\{x \mid a^T x \geq a^T x_0\}$ is tangent to C at x_0. This is illustrated in figure 1.18.

A basic result, called the *supporting hyperplane theorem*, is that for any nonempty convex set C, and any $x_0 \in \partial C$, there exists a supporting hyperplane to C at x_0.

The supporting hyperplane theorem is readily proved from the separating hyperplane theorem. We distinguish two cases. If the interior of C is nonempty, the result follows immediately by applying the separating hyperplane theorem to the sets $\{x_0\}$ and $\mathbf{int}\, C$. If the interior of C is empty, then C must lie in an affine set of dimension less than n, and any hyperplane containing that affine set contains C and x_0, and is a (trivial) supporting hyperplane.

1.6 Dual cones and generalized inequalities

1.6.1 Dual cones

Let K be a convex cone. The set

$$K^* = \{y \mid x^T y \geq 0 \text{ for all } x \in K\} \tag{1.17}$$

is called the *dual cone* of K. (As the name suggests, K^* is indeed a cone; see exercises.) Geometrically, $y \in K^*$ if and only if $-y$ is the normal of a hyperplane that supports K at the origin. This is illustrated in figure 1.19.

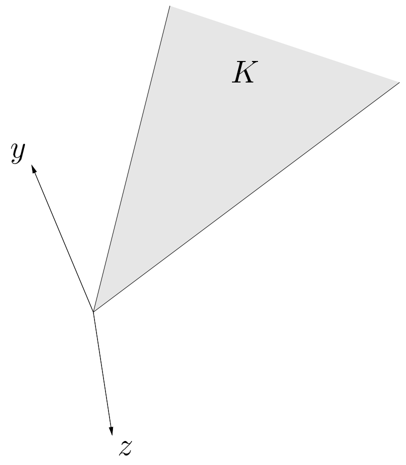

Figure 1.19: A cone K and two points: $y \in K^*$, and $z \notin K^*$.

Example. The dual of a subspace $V \subseteq \mathbf{R}^n$ (which is a cone) is its orthogonal complement $V^\perp = \{ y \mid y^T v = 0 \text{ for all } v \in V \}$.

Example. The nonnegative orthant $\mathbf{R}^n_+$ is its own dual:

$$y^T x \geq 0 \ \ \forall x \succeq 0 \Longleftrightarrow y \succeq 0.$$

We call such a cone *self-dual*.

Example. The positive semidefinite cone is also self-dual, *i.e.*, for $X = X^T$, $Y = Y^T \in \mathbf{R}^{n \times n}$,

$$\mathbf{Tr}\, XY \geq 0 \ \ \forall X \succeq 0 \Longleftrightarrow Y \succeq 0.$$

(Recall that $\mathbf{Tr}\, XY$ is the inner product we use for two symmetric matrices X and Y.) We will establish this fact.

Suppose Y is not positive semidefinite. Then there exists $q \in \mathbf{R}^n$ with $q^T Y q = \mathbf{Tr}\, qq^T Y < 0$. Hence the positive semidefinite matrix $X = qq^T$ satisfies $\mathbf{Tr}\, XY < 0$. It follows that Y is not in the dual of the cone of positive semidefinite matrices.

Now suppose $Y = Y^T \succeq 0$, and let $X = X^T \succeq 0$ be any positive semidefinite matrix. We can express X in terms of its eigenvalue decomposition as $X = \sum_i \lambda_i q_i q_i^T$, where (the eigenvalues) $\lambda_i \geq 0$. Then we have

$$\mathbf{Tr}\, YX = \mathbf{Tr}\, Y \sum_{i=1}^n \lambda_i q_i q_i^T = \sum_{i=1}^n \lambda_i q_i^T Y q_i$$

(using the general fact $\mathbf{Tr}\, AB = \mathbf{Tr}\, BA$). Since Y is positive semidefinite, each term in this sum is nonnegative, so $\mathbf{Tr}\, YX \geq 0$. This proves that Y is in the dual of the cone of positive semidefinite matrices.

Example. *Dual of a norm cone.* Let $\|\cdot\|$ be an arbitrary norm on $\mathbf{R}^n$. The dual of the associated cone $K = \{(x,t) \in \mathbf{R}^{n+1} \mid \|x\| \leq t\}$ is the cone defined by the dual norm, *i.e.*,

$$K^* = \{(u,v) \in \mathbf{R}^{n+1} \mid \|u\|_* \leq v\}.$$

This follows from the definition of the dual norm: $\|u\|_* = \sup\{u^T x \mid \|x\| \le 1\}$.

To prove the result we have to show that

$$x^T u + tv \ge 0 \text{ if } \|x\| \le t \Longleftrightarrow \|u\|_* \le v.$$

Let us start by showing that the condition on (u, v) is sufficient. Suppose $\|u\|_* \le v$, and $\|x\| \le t$ for some $t > 0$. (If $t = 0$, x must be zero, so obviously $u^T x + vt \ge 0$.) Applying the definition of the dual norm, and the fact that $\|-x/t\| \le 1$, we have

$$u^T(-x/t) \le \|u\|_* \le v$$

and therefore $u^T x + vt \ge 0$.

Next we show that $\|u\|_* \le v$ is necessary. Suppose $\|u\|_* > v$. Then by the definition of the dual norm, there exists an x with $\|x\| \le 1$ and $x^T u > v$. Hence there is at least one vector $(-x, 1)$ in the cone K that makes a negative inner product with (u, v), since $\|-x\| \le 1$ and

$$u^T(-x) + v < 0.$$

Example. *Dual of cone of nonnegative polynomials on* $[0, 1]$. Consider the cone of polynomials that are nonnegative on $[0, 1]$, defined in (1.13). Here the dual cone K^* turns out to be the closure of

$$\left\{ y \in \mathbf{R}^n \;\middle|\; y_i = \int_0^1 t^{i-1} w(t)\, dt,\ w(t) \ge 0 \text{ for } t \in [0, 1] \right\},$$

which is the cone of moments of nonnegative functions on $[0, 1]$ (see exercises).

Dual cones satisfy several properties such as:

- $K_1 \subseteq K_2$ implies $K_2^* \subseteq K_1^*$.
- If K is pointed then K^* has nonempty interior, and vice-versa.
- K^* is always closed.
- K^{**} is the closure of K. (Hence if K is closed, $K^{**} = K$.)

(See exercises.)

1.6.2 Dual generalized inequalities

Now suppose that the convex cone K induces a generalized inequality $\preceq_K$. Then the dual cone K^* also satisfies the conditions required to induce a generalized inequality, *i.e.*, K^* is pointed, closed, and has nonempty interior. We refer to the generalized inequality $\preceq_{K^*}$ as the *dual* of the generalized inequality $\preceq_K$.

Some important properties of the dual generalized inequality are listed below.

- $x \preceq_K y$ if and only if $\lambda^T x \le \lambda^T y$ for all $\lambda \succeq_{K^*} 0$.

- $x \prec_K y$ if and only if $\lambda^T x < \lambda^T y$ for all $\lambda \succeq_{K^*} 0$, $\lambda \neq 0$.

The dual generalized inequality associated with $\preceq_{K^*}$ is $\preceq_K$ (since K is closed, hence $K^{**} = K$), so these properties hold if the generalized inequality and its dual are swapped. As a specific example, we have $\lambda \preceq_{K^*} \mu$ if and only if $\lambda^T x \leq \mu^T x$ for all $x \succeq_K 0$.

Example. *Theorem of alternatives for linear strict generalized inequalities.* Suppose K defines a generalized linear inequality in $\mathbf{R}^m$. Consider the strict generalized inequality

$$Ax \prec_K b, \tag{1.18}$$

where $x \in \mathbf{R}^n$.

We will derive a theorem of alternatives for this inequality. It is infeasible if and only if the affine set $\{b - Ax | x \in \mathbf{R}^n\}$ does not intersect the open convex set $\mathbf{int}\, K$. This occurs if and only if there is a separating hyperplane, *i.e.*, a nonzero $\lambda \in \mathbf{R}^m$ and $\mu \in \mathbf{R}$ such that $\lambda^T(b - Ax) \leq \mu$ for all x, and $\lambda^T y \geq \mu$ for all $y \in \mathbf{int}\, K$. From the first condition we get $A^T\lambda = 0$ and $\lambda^T b \leq \mu$. The second implies that $\lambda^T y \geq \mu$ for all $y \in K$, which occurs only if $\lambda \in K^*$ and $\mu \leq 0$.

Putting it all together we get the condition for infeasibility of (1.18): there exists λ such that

$$\lambda \neq 0, \quad \lambda \succeq_{K^*} 0, \quad A^T\lambda = 0, \quad \lambda^T b \leq 0 \tag{1.19}$$

The inequality systems (1.18) and (1.19) are alternatives: for any data A, b, exactly one of them is feasible.

(Note the similarity to the alternatives (1.15), (1.16) for the special case $K = \mathbf{R}^m_+$.)

1.6.3 Characterization of minimum and minimal elements via dual inequalities

We can use dual generalized inequalities to characterize minimum and minimal elements of a (possibly nonconvex) set $S \subseteq \mathbf{R}^m$ (with respect to a generalized inequality induced by a cone K). We first consider a characterization of the *minimum* element: x is the minimum element of S if and only if for all $\lambda \succ_{K^*} 0$, x is the unique minimizer of $\lambda^T z$ over $z \in S$.

Proof. Suppose x is the minimum element of S, *i.e.*, $x \preceq_K z$ for all $z \in S$. Then we have

$$\begin{aligned} \lambda \succ_{K^*} 0 \quad &\Longrightarrow \quad \lambda^T(z - x) > 0 \text{ for all } z \succeq_K x,\ z \neq x \\ &\Longrightarrow \quad \lambda^T(z - x) > 0 \text{ for all } z \in S,\ z \neq x. \end{aligned}$$

Conversely, suppose that for all nonzero $\lambda \succ_{K^*} 0$, x is the unique minimizer of $\lambda^T z$ over $z \in S$, but that x is not the minimum element, *i.e.*, there exists a $z \in S$, $z \not\succeq_{K^*} x$. Then we have

$$\begin{aligned} z - x \not\succeq_{K^*} 0 \quad &\Longrightarrow \quad \exists \lambda \succeq_{K^*} 0,\ \lambda^T(z - x) < 0 \\ &\Longrightarrow \quad \exists \lambda \succ_{K^*} 0,\ \lambda^T(z - x) < 0, \end{aligned}$$

which contradicts the assumption that x is the unique minimizer of $\lambda^T z$ over S.

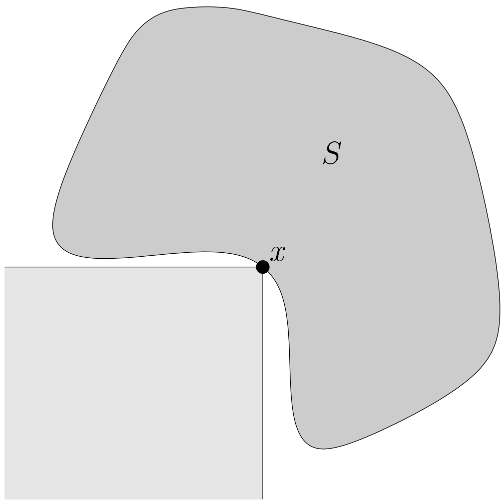

Figure 1.20: The point x is a minimal element of S. However there exists no $\lambda \succ 0$ for which x minimizes $\lambda^T x$ over S.

We now turn to a similar characterization of *minimal elements*. Here there is a gap between the necessary and sufficient conditions. If $\lambda \succ_{K^*} 0$ and x minimizes $\lambda^T z$ over $z \in S$, then x is minimal.

Proof. Suppose x is not minimal, *i.e.*, there exists a $z \in S$, $z \neq x$, and $z \preceq_K x$. Then

$$\lambda^T(x - z) > 0,$$

which contradicts our assumption that x is the minimizer of $\lambda^T z$ over S.

The converse is in general false; there can be minimal points of S that are not minimizers of $\lambda^T x$ for any λ (see figure 1.20).

The figure 1.20 suggests that convexity plays an important role in the converse, which is correct. Provided the set S is convex, we can say that for any minimal element x there exists a nonzero $\lambda \succeq_{K^*} 0$ such that x minimizes $\lambda^T z$ over $z \in S$.

Proof. x is minimal if and only if

$$((x - K) \setminus \{x\}) \cap S = \emptyset.$$

Applying the weak separation theorem to the convex sets $x - K$ and S, we conclude that

$$\exists \lambda \neq 0 : \lambda^T(x - K) \leq \mu, \lambda^T S \geq \mu.$$

Since $x \in S$ and $x \in x - K$, this implies that $\lambda^T x = \mu$, and that μ is the minimum of $\lambda^T z$ over S. Therefore, $\lambda^T K \geq 0$, *i.e.*, $\lambda \succeq_{K^*} 0$.

Figure 1.21 illustrates these results. The weaker converse with positive λ cannot be strengthened, as the example below shows.

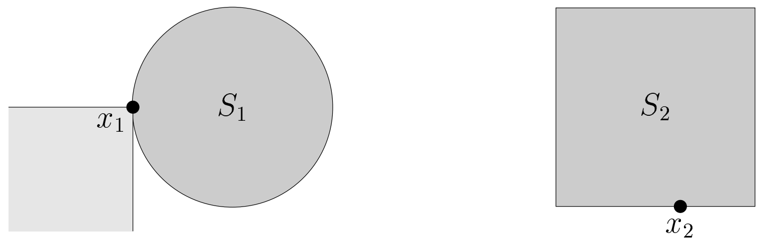

Figure 1.21: Two examples that illustrate the gap between necessary and sufficient conditions that characterize minimal elements. The example on the left shows that for $x \in S_1$ to be minimal, it is not necessary that there exista a $\lambda \succ 0$ for which x minimizes $\lambda^T z$ over S: the point x_1 is a minimal element of S_1, and there exists no $\lambda \succ 0$ for which x_1 minimizes $\lambda^T z$ over S_1. The example on the right shows that for x to be minimal, it is not sufficient that there exists a $\lambda \succeq 0$ for which x minimizes $\lambda^T z$ over S: the point x_2 is not a minimal element of S_2 although it minimizes $\lambda^T z$ over S_2 for $\lambda = (0, -1)$.

Chapter 2

Convex functions

2.1 Basic properties and examples

2.1.1 Definition

A function $f : \mathbf{R}^n \to \mathbf{R}$ is *convex* if $\mathbf{dom}\, f$ is a convex set and if for all x, $y \in \mathbf{dom}\, f$, and θ with $0 \le \theta \le 1$, we have

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y). \tag{2.1}$$

Geometrically, this inequality means that the line segment between $(x, f(x))$ and $(y, f(y))$ (*i.e.*, the *chord* from x to y) lies above the graph of f (figure 2.1). A function f is *strictly convex* if strict inequality holds in (2.1) whenever $x \ne y$ and $0 < \theta < 1$. We say f is *concave* if $-f$ is convex, and *strictly concave* if $-f$ is strictly convex.

For an affine function we always have equality in (2.1), so all affine (and therefore also linear) functions are both convex and concave. Conversely, any function that is convex and concave is affine.

A function is convex if and only if it is convex when restricted to any line that intersects its domain. In other words f is convex if and only if for all $x \in \mathbf{dom}\, f$ and all v, the function $h(t) = f(x+tv)$ is convex (on its domain, $\{t \mid x + tv \in \mathbf{dom}\, f\}$). This property is very useful, since it allows us to check whether a function is convex by restricting it to a line.

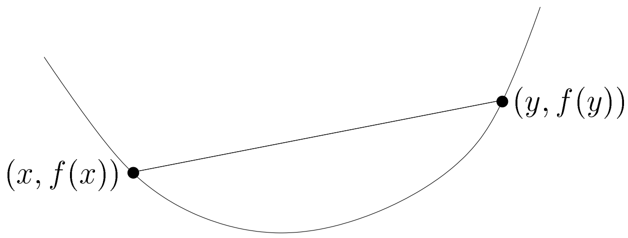

Figure 2.1: Graph of a convex function. The chord (*i.e.*, line segment) between any two points on the graph lies above the graph.

The *analysis* of convex functions is a well developed field, which we will not pursue in any depth. One simple result, for example, is that a convex function is continuous on the relative interior of its domain; it can have discontinuities only on its relative boundary.

2.1.2 Extended-valued extensions

It is often convenient to extend a convex function to all of $\mathbf{R}^n$ by defining its value to be ∞ outside its domain. If f is convex we define its extended-valued extension $\tilde{f} : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ by

$$\tilde{f}(x) = \begin{cases} f(x) & x \in \mathbf{dom}\, f, \\ +\infty & x \notin \mathbf{dom}\, f. \end{cases}$$

The extension $\tilde{f}$ is defined on all $\mathbf{R}^n$, and takes values in $\mathbf{R} \cup \{\infty\}$. We can recover the domain of the original function f from the extension $\tilde{f}$ as $\mathbf{dom}\, f = \{x | \tilde{f}(x) < \infty\}$.

The extension can simplify notation, since we do not have to explicitly describe the domain, or add the qualifier 'for all $x \in \mathbf{dom}\, f$' every time we refer to $f(x)$. Consider, for example, the basic defining inequality (2.1). In terms of the extension $\tilde{f}$, we can express it as: for $0 < \theta < 1$,

$$\tilde{f}(\theta x + (1-\theta)y) \le \theta \tilde{f}(x) + (1-\theta)\tilde{f}(y)$$

for *any* x and y. Of course here we must interpret the inequality using extended arithmetic and ordering. For x and y both in $\mathbf{dom}\, f$, this inequality coincides with (2.1); if either is outside $\mathbf{dom}\, f$, then the righthand side is ∞, and the inequality therefore holds. As another example of this notational device, suppose f_1 and f_2 are two convex functions on $\mathbf{R}^n$. The pointwise sum $f = f_1 + f_2$ is the function with domain $\mathbf{dom}\, f = \mathbf{dom}\, f_1 \cap \mathbf{dom}\, f_2$, with $f(x) = f_1(x) + f_2(x)$ for any $x \in \mathbf{dom}\, f$. Using extended valued extensions we can simply say that for any x, $\tilde{f}(x) = \tilde{f}_1(x) + \tilde{f}_2(x)$. In this equation the domain of f has been automatically defined as $\mathbf{dom}\, f = \mathbf{dom}\, f_1 \cap \mathbf{dom}\, f_2$, since $\tilde{f}(x) = \infty$ whenever $x \notin \mathbf{dom}\, f_1$ or $x \notin \mathbf{dom}\, f_2$. In this example we are relying on extended arithmetic to automatically define the domain.

In this book we will use the same symbol to denote a convex function and its extension, whenever there is no harm from the ambiguity. This is the same as assuming that all convex functions are implicitly extended, *i.e.*, are defined as ∞ outside their domains.

> **Example.** *Indicator function of a convex set.* Let $C \subseteq \mathbf{R}^n$ be a convex set, and consider the (convex) function f with domain C and $f(x) = 0$ for all $x \in C$. In other words, the function is identically zero on the set C. Its extended valued extension is given by
>
> $$\tilde{f}(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$$
>
> The convex function $\tilde{f}$ is called the *indicator function* of the set C.
>
> We can play several notational tricks with the indicator function $\tilde{f}$. For example the problem of minimizing a function g (defined on all of $\mathbf{R}^n$, say) on the set C is the same as minimizing the function $g + \tilde{f}$ over all of $\mathbf{R}^n$. Indeed, the function $g + \tilde{f}$ is (by our convention) g restricted to the set C.

In a similar way we can extend a concave function by defining it to be $-\infty$ outside its domain.

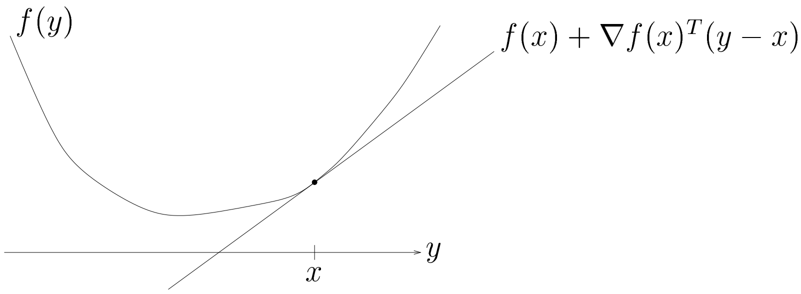

Figure 2.2: If f is convex and differentiable, then $f(x) + \nabla f(x)^T(y - x) \leq f(y)$ for all $x,\ y \in \mathbf{dom}\, f$.

2.1.3 First order conditions

Suppose f is differentiable (*i.e.*, its gradient ∇f exists at each point in $\mathbf{dom}\, f$, which is open). Then f is convex if and only if

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) \tag{2.2}$$

holds for all $x,\ y \in \mathbf{dom}\, f$. This inequality is illustrated in figure 2.2.

The affine function of y given by $f(x) + \nabla f(x)^T(y - x)$ is, of course, the first order Taylor approximation of f near x. The inequality (2.2) states that for a convex function, the first order Taylor approximation is in fact *global underestimator* of the function. Conversely, if the first order Taylor approximation of a function is always a global underestimator of the function, then the function is convex.

The inequality (2.2) shows that from *local information* about a convex function (*i.e.*, its derivative at a point) we can derive *global information* (*i.e.*, a global underestimator of it). This is perhaps the most important property of convex functions, and explains some of the remarkable properties of convex functions and convex optimization problems.

Proof. To prove (2.2), we first consider the case $n = 1$, *i.e.*, show that a differentiable function $f : \mathbf{R} \to \mathbf{R}$ is convex if and only if

$$f(y) \geq f(x) + f'(x)(y - x) \tag{2.3}$$

for all x and y.

Assume first that f is convex. Then for all $0 < t \leq 1$ we have

$$f(x + t(y - x)) \leq (1 - t)f(x) + tf(y).$$

If we divide both sides by t, we obtain

$$f(y) \geq f(x) + \frac{f(x + t(y - x)) - f(x)}{t},$$

and the limit as $t \to 0$ yields (2.3).

To show sufficiency, assume the function satisfies (2.3) for all x and y. Choose any $x \neq y$, and $0 \leq \theta \leq 1$, and let $z = \theta x + (1-\theta)y$. Applying (2.3) twice yields

$$f(x) \geq f(z) + f'(z)(x-z), \qquad f(y) \geq f(z) + f'(z)(y-z).$$

Multiplying the first inequality by θ, the second by $1-\theta$, and adding them yields

$$\theta f(x) + (1-\theta) f(y) \geq f(z),$$

which proves that f is convex.

Now we can prove the general case, with $f : \mathbf{R}^n \to \mathbf{R}$. Let $x,\ y \in \mathbf{R}^n$ and consider f restricted to the line passing through them, *i.e.*, the function defined by $g(t) = f(ty + (1-t)x)$, so $g'(t) = \nabla f(ty + (1-t)x)^T(y-x)$.

First assume f is convex, which implies g is convex, so by the argument above we have $g(1) \geq g(0) + g'(0)$, which means

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

Now assume that this inequality holds for any x and y, so if $ty + (1-t)x \in \mathbf{dom}\, f$ and $\tilde{t}y + (1-\tilde{t})x \in \mathbf{dom}\, f$, we have

$$f(ty + (1-t)x) \geq f(\tilde{t}y - (1-\tilde{t})x) + \nabla f(\tilde{t}y - (1-\tilde{t})x)^T(y-x)(t-\tilde{t}),$$

i.e., $g(t) \geq g(\tilde{t}) + g'(\tilde{t})(t-\tilde{t})$. We have seen that this implies that g is convex.

For concave functions we have the corresponding characterization: f is concave if and only if

$$f(y) \leq f(x) + \nabla f(x)^T(y-x)$$

for all $x,\ y \in \mathbf{dom}\, f$.

Strict convexity can also be characterized by a first-order condition: f is strictly convex if and only if for $x,\ y \in \mathbf{dom}\, f$, $x \neq y$, we have

$$f(y) > f(x) + \nabla f(x)^T(y-x). \tag{2.4}$$

2.1.4 Second order conditions

We now assume that f is twice differentiable, *i.e.*, its Hessian or second derivative $\nabla^2 f$ exists at each point in $\mathbf{dom}\, f$, which is open. Then f is convex if and only if its Hessian is positive semidefinite:

$$\nabla^2 f(x) \succeq 0$$

for all $x \in \mathbf{dom}\, f$. For a function on $\mathbf{R}$, this reduces to the simple condition $f''(x) \geq 0$. (The general second order condition is readily proved by reducing it to the case of $f : \mathbf{R} \to \mathbf{R}$). Similarly, f is concave if and only $\nabla^2 f(x) \preceq 0$ for all $x \in \mathbf{dom}\, f$.

Strict convexity can be partially characterized by second order conditions. If $\nabla^2 f(x) \succ 0$ for all $x \in \mathbf{dom}\, f$, then f is strictly convex. The converse, however, is not true: the function $f : \mathbf{R} \to \mathbf{R}$ given by $f(x) = x^4$ is strictly convex but has zero second derivative at $x = 0$.

Example. *Quadratic functions.* Consider the quadratic function $f : \mathbf{R}^n \to \mathbf{R}$, with $\mathbf{dom}\, f = \mathbf{R}^n$, given by

$$f(x) = x^T Px + 2q^T x + r,$$

with $P = P^T \in \mathbf{R}^{n\times n}$, $q \in \mathbf{R}^n$, and $r \in \mathbf{R}$. Since $\nabla^2 f(x) = 2P$ for all x, f is convex if and only if $P \succeq 0$ (and concave if and only if $P \preceq 0$).

For quadratic functions, strict convexity is easily characterized: f is strictly convex if and only if $P \succ 0$ (and strictly concave if and only if $P \prec 0$).

2.1.5 Examples

We have already mentioned that all linear and affine functions are convex (and concave), and have described the convex and concave quadratic functions. In this section we give a few more examples of convex and concave functions. We start with some functions on $\mathbf{R}$, with variable x.

- *Exponential.* e^{ax} is convex on $\mathbf{R}$.
- *Powers.* x^a is convex on $\{x|x > 0\}$ when $a \geq 1$ or $a \leq 0$, and concave for $0 \leq a \leq 1$.
- *Powers of absolute value.* $|x|^p$, for $p \geq 1$, is convex on $\mathbf{R}$.
- *Logarithm.* $\log x$ is concave on $\{x|x > 0\}$.
- *Negative entropy.* $x \log x$ (either on $\{x|x > 0\}$, or defined as 0 for $x = 0$) is convex.
- $\log \mathbf{erfc}(x) = \log \left(\frac{2}{\sqrt{\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du \right)$ is concave on $\mathbf{R}$.
- $\log \Gamma(x) = \log \left(\int_0^\infty u^{x-1} e^{-u}\, du \right)$ is convex for $x \geq 1$.

Convexity or concavity of these examples is easily shown by verifying the basic inequality (2.1), or, when the function is differentiable, by checking that the second derivative is nonnegative or nonpositive. For example, with $f(x) = \mathbf{erfc}(x)$, we have

$$f'(x) = e^{-x^2/2} A^{-1}, \qquad f''(x) = -xe^{-x^2/2} A^{-1} - e^{-x^2} A^{-2},$$

where

$$A = \int_{-\infty}^{x} e^{-a^2/2}\, da.$$

Both terms in the expression for f'' are nonpositive, so f is concave. Indeed, $f''(x) < 0$ for all x, so f is strictly concave.

We now give a few interesting examples of functions on $\mathbf{R}^n$.

- *Norms.* Every norm on $\mathbf{R}^n$ is convex.
- *Max function.* $f(x) = \max_i x_i$ is convex on $\mathbf{dom}\, f = \mathbf{R}^n$.

- *Log-sum-exp.* The function $f(x) = \log(e^{x_1} + \cdots + e^{x_n})$ (with $\mathbf{dom}\, f = \mathbf{R}^n$) is convex. This function can be interpreted as a smooth approximation of the max function, since
$$\max_i x_i \le f(x) \le \log n + \max_i x_i$$
for all x. (This inequality is sharp when all components of x are equal.)

- *Geometric mean.* The geometric mean $f(x) = \left(\prod_{i=1}^n x_i\right)^{1/n}$ is concave on $\mathbf{dom}\, f = \{x | x \succ 0\}$.

- *Log determinant.* The function $f(x) = \log\det X^{-1}$ is convex on $\mathbf{dom}\, f = \{X \in \mathbf{R}^{n\times n} | X = X^T \succ 0\}$.

Convexity of the first two functions can be shown by directly verifying the inequality (2.1). If $f : \mathbf{R}^n \to \mathbf{R}$ is a norm, and $0 \le \theta \le 1$, then
$$f(\theta x + (1-\theta)y) \le f(\theta x) + f((1-\theta)y) = \theta f(x) + (1-\theta) f(y),$$
since by definition a norm is homogeneous and satisfies the triangle inequality.

The function $f(x) = \max_i x_i$ satisfies, for $0 \le \theta \le 1$,
$$\begin{aligned} f(\theta x + (1-\theta)y) &= \max_i(\theta x_i + (1-\theta)y_i) \\ &\le \theta \max_i x_i + (1-\theta)\max_i y_i \\ &= \theta f(x) + (1-\theta) f(y). \end{aligned}$$

The log-sum-exp function is smooth and it is quite easy to show that $\nabla^2 f(x) \succeq 0$ for all x. The Hessian at x is
$$\nabla^2 f(x) = \frac{1}{(\mathbf{1}^T z)^2}\left((\mathbf{1}^T z)\,\mathbf{diag}(z) - zz^T\right)$$
where $z = (e^{x_1}, \ldots, e^{x_n})$. To verify that $\nabla^2 f(x) \succeq 0$ we must show that $v^T \nabla^2 f(x) v \ge 0$ for all v, *i.e.*,
$$v^T \nabla^2 f(x) v = \frac{1}{(\mathbf{1}^T z)^2}\left(\left(\sum_i z_i\right)\left(\sum_i v_i^2 z_i\right) - \left(\sum_i v_i z_i^2\right)\right) \ge 0.$$
But this follows from the Cauchy-Schwarz inequality $(a^T a)(b^T b) \ge (a^T b)^2$ applied to the vectors with components $a_i = v_i\sqrt{z_i}$, $b_i = \sqrt{z_i}$.

In a similar way we can show that the geometric mean $f(x) = \left(\prod_{i=1}^n x_i\right)^{1/n}$ is concave on $\mathbf{dom}\, f = \{x | x \succ 0\}$. Its Hessian $\nabla^2 f(x)$ is given by
$$\frac{\partial^2 f(x)}{\partial x_k^2} = -(n-1)\frac{\left(\prod_i x_i\right)^{1/n}}{n^2 x_k^2}, \quad \frac{\partial^2 f(x)}{\partial x_k \partial x_l} = \frac{\left(\prod_i x_i\right)^{1/n}}{n x_k x_l} \quad (k \ne l),$$
and can be expressed as
$$\nabla^2 f(x) = -\frac{\prod_i x_i^{1/n}}{n^2}\left(n\,\mathbf{diag}(1/x_1^2, \ldots, 1/x_n^2) - qq^T\right)$$

where $q_i = 1/x_i$. We must show that $\nabla^2 f(x) \preceq 0$, *i.e.*, that

$$v^T \nabla^2 f(x) v = -\frac{\prod_i x_i^{1/n}}{n^2}\left(n\sum_i v_i^2/x_i^2 - \left(\sum_i v_i/x_i\right)^2\right) \leq 0$$

for all v. Again this follows from the Cauchy-Schwarz inequality $(a^Ta)(b^Tb) \geq (a^Tb)^2$, applied to the vectors $a = \mathbf{1}$ and $b_i = v_i/x_i$.

For the function $f(X) = \log\det X^{-1}$, we can verify convexity by considering an arbitrary line, given by $X = Z + tV$, where $Z = Z^T \succ 0$ and $V = V^T$. With $g(t) = f(Z + tV)$ we have

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

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $Z^{-1/2}VZ^{-1/2}$. Therefore we have

$$g'(t) = -\sum_{i=1}^n \frac{\lambda_i}{1 + t\lambda_i}, \quad g''(t) = \sum_{i=1}^n \frac{\lambda_i^2}{1 + t\lambda_i^2},$$

so

$$g''(0) = \sum_{i=1}^n \lambda_i^2 \geq 0,$$

which proves f is convex.

2.1.6 Sublevel sets

The *α-sublevel set* of a function $f : \mathbf{R}^n \to \mathbf{R}$ is defined as

$$C_\alpha = \{x \in \mathbf{dom}\, f \mid f(x) \leq \alpha\}.$$

Sublevel sets of a convex function are convex, for any value of α. The proof is immediate from the definition of convexity: if $x,\ y \in C_\alpha$, *i.e.*, $f(x) \leq \alpha$, and $f(y) \leq \alpha$, then $f(\theta x + (1-\theta)y) \leq \alpha$ for $0 \leq \theta \leq 1$, and hence $\theta x + (1-\theta)y \in C_\alpha$.

The converse is not true; a function can have all its sublevel sets convex, but not be a convex function. For example, $f(x) = -e^x$ is not convex on $\mathbf{R}$ (indeed, it is strictly concave) but all its sublevel sets are convex.

If f is concave, then its *α-superlevel set*, given by $\{x \mid f(x) \geq \alpha\}$, is convex.

Example. The geometric and arithmetic means of $x \in \mathbf{R}_+^n$ are, respectively,

$$G(x) = \left(\prod_{i+1}^n x_i\right)^{1/n}, \quad A(x) = \frac{1}{n}\sum_{i+1}^n x_i.$$

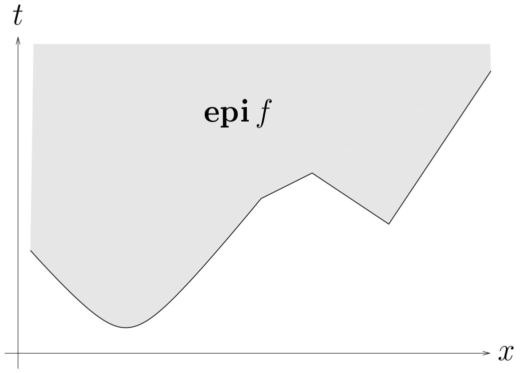

Figure 2.3: The epigraph of a function.

The geometric-arithmetic mean inequality states that $G(x) \leq A(x)$.

Suppose $0 \leq \alpha \leq 1$, and consider the set

$$\{\ x \in \mathbf{R}^n_+ \ \mid G(x) \geq \alpha A(x)\ \},$$

i.e., the set of vectors with geometric mean no more than a factor α smaller than arithmetic mean. This set is convex, since it is the 0-superlevel set of the function $G(x) - \alpha A(x)$, which is concave. In fact, the set is positively homogeneous, so it is a convex cone.

2.1.7 Epigraph

The graph of a function $f : \mathbf{R}^n \to \mathbf{R}$ is defined as

$$\{(x, f(x)) \mid x \in \mathbf{dom}\, f\},$$

which is a subset of $\mathbf{R}^{n+1}$. The *epigraph* of a function $f : \mathbf{R}^n \to \mathbf{R}$ is defined as

$$\mathbf{epi}\, f = \{(x, t) \mid x \in \mathbf{dom}\, f,\ f(x) \leq t\ \}.$$

('Epi' means 'above' so epigraph means 'above the graph'.) The definition is illustrated in figure 2.3.

The link between convex sets and convex functions is via the epigraph: A function is convex if and only if its epigraph is a convex set. Similarly, a function is strictly convex if and only if its epigraph is a strictly convex set.

A function is concave if and only if its *hypograph*, defined as

$$\mathbf{hypo}\, f = \{(x, t) \mid t \leq f(x)\},$$

is a convex set.

Many results for convex functions can be proved (or interpreted) geometrically using epigraphs, and applying results for convex sets. For example, we can interpret the inequality (2.2) geometrically as follow:

$$\begin{aligned} (y, t) \in \mathbf{epi}\, f \quad &\Longrightarrow \quad t \geq f(x) + \nabla f(x)^T (y - x) \\ &\Longleftrightarrow \quad \begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}^T \begin{bmatrix} y - x \\ t - f(x) \end{bmatrix} \leq 0. \end{aligned}$$

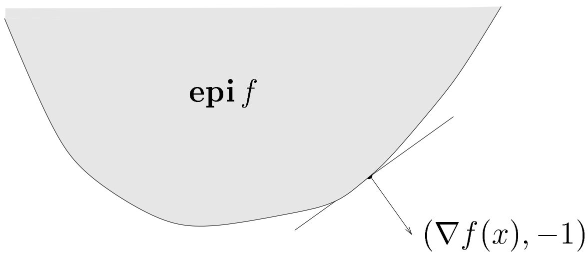

Figure 2.4: The gradient $\nabla f(x)$ defines a supporting hyperplane to the epigraph of f at x.

This means that the hyperplane

$$\{(y,t) \in \mathbf{R}^{n+1} \mid (y-x)^T \nabla f(x) - t = f(x)\}$$

supports $\mathbf{epi}\, f$ at the boundary point $(x, f(x))$ (see figure 2.4).

2.1.8 Jensen's inequality and extensions

Inequality (2.1) is sometimes called *Jensen's inequality.* It is easily extended to convex combinations of more than two points: If f is convex, $x_1, \ldots, x_k \in \mathbf{dom}\, f$ and $\theta_1, \ldots, \theta_k \geq 0$ with $\theta_1 + \cdots + \theta_k = 1$, then

$$f(\theta_1 x_1 + \cdots + \theta_k x_k) \leq \theta_1 f(x_1) + \cdots + \theta_k f(x_k).$$

As in the case of convex sets, the inequality extends to infinite sums, integrals, and expected values. For example, if $p(x) \geq 0$ on $S \subseteq \mathbf{dom}\, f$, $\int_S p(x)\, dx = 1$, then

$$f\left(\int_S p(x) x \; dx\right) \leq \int_S f(x) p(x) \; dx$$

provided the integrals exist. In the most general case we can take any probability measure with support in $\mathbf{dom}\, f$. If x is a random variable such that $x \in \mathbf{dom}\, f$ with probability one, and f is convex, then we have

$$f(\mathbf{E}\, x) \leq \mathbf{E}\, f(x), \tag{2.5}$$

provided the expectations exist. We can recover the basic inequality (2.1) from this general form, by taking the random variable x to have support $\{x_1, x_2\}$, with $\mathbf{Prob}(x = x_1) = \theta$, $\mathbf{Prob}(x = x_2) = 1 - \theta$. Thus the inequality (2.5) characterizes convexity: If f is not convex, there is a random variable x, with $x \in \mathbf{dom}\, f$ with probability one, such that $f(\mathbf{E}\, x) > \mathbf{E}\, f(x)$.

All of these inequalities are now called *Jensen's inequality*, even though the inequality studied by Jensen was the very simple one

$$f((x+y)/2) \leq \frac{f(x) + f(y)}{2}.$$

Remark. We can interpret (2.5) as follows. Suppose $x \in \mathbf{dom}\, f \subseteq \mathbf{R}^n$ and z is any zero-mean random vector in $\mathbf{R}^n$. Then we have

$$\mathbf{E}\, f(x+z) \geq f(x).$$

Thus, randomization or dithering cannot decrease the value of a convex function on average.

2.1.9 Inequalities

Many famous inequalities can be derived by applying Jensen's inequality to some appropriate convex function. (Indeed, convexity and Jensen's inequality can be made the foundation of a theory of inequalities.) As a simple example, consider the arithmetic-geometric mean inequality:

$$\sqrt{ab} \leq \frac{1}{2}(a+b) \tag{2.6}$$

for $a, b \geq 0$. The function $f(x) = -\log x$ is convex; Jensen's inequality with $\theta = 1/2$ yields

$$f((a+b)/2) = -\log(a+b)/2 \leq \frac{-\log a - \log b}{2} = \frac{f(a)+f(b)}{2}.$$

Taking the exponential of both sides yields (2.6).

As a less trivial example we prove Hölder's inequality: for $p > 1$, $1/p + 1/q = 1$,

$$\sum_{i=1}^n x_i y_i \leq \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \left(\sum_{i=1}^n |y_i|^q\right)^{1/q}.$$

By convexity of $-\log x$, and Jensen's inequality with general θ, we obtain the more general arithmetic-geometric mean inequality

$$a^\theta b^{1-\theta} \leq \theta a + (1-\theta) b,$$

valid for $a, b \geq 0$ and $0 \leq \theta \leq 1$. Applying this with

$$a = \frac{|x_i|^p}{\sum_{j=1}^n |x_j|^p}, \qquad b = \frac{|y_i|^q}{\sum_{j=1}^n |y_j|^q}, \qquad \theta = 1/p,$$

yields

$$\left(\frac{|x_i|^p}{\sum_{j=1}^n |x_j|^p}\right)^{1/p} \left(\frac{|y_i|^q}{\sum_{j=1}^n |y_j|^q}\right)^{1/q} \leq \frac{|x_i|^p}{p\sum_{j=1}^n |x_j|^p} + \frac{|y_i|^q}{q\sum_{j=1}^n |y_j|^q}.$$

Summing over i then yields the Hölder inequality.

2.2 Operations that preserve convexity

In this section we describe some operations that preserve convexity or concavity of functions, or allow us to construct new convex and concave functions. We start with some simple operations such as addition, scaling, and pointwise supremum, and then describe some more sophisticated operations (some of which include the simple operations as special cases).

2.2.1 Nonnegative weighted sums

Evidently if f is a convex function and $\alpha \geq 0$, then the function αf is convex. If f_1 and f_2 are both convex functions, then so is their sum $f_1 + f_2$. Combining positive scaling and addition, we see that the set of convex functions on a set is itself a cone: a nonnegative weighted sum of convex functions,

$$f = w_1 f_1 + \cdots + w_m f_m,$$

is convex. Similarly, a nonnegative weighted sum of concave functions is concave. A nonnegative, nonzero weighted sum of strictly convex (concave) functions is strictly convex (concave).

These properties extend to infinite sums and integrals. For example if $f(x, y)$ is convex in x for each $y \in \mathcal{A}$, and $w(y) \geq 0$ for each $y \in \mathcal{A}$, then the function g defined as

$$g(x) = \int_{\mathcal{A}} w(y) f(x, y)\, dy$$

is convex in x.

The fact that convexity is preserved under nonnegative scaling and addition is easily verified directly, or can be seen in terms of the associated epigraphs. For example, if $w \geq 0$ and f convex, we have

$$\mathbf{epi}(wf) = \begin{bmatrix} I & 0 \\ 0 & w \end{bmatrix} \mathbf{epi}\, f,$$

which is convex because the image of a convex set under a linear mapping is convex.

2.2.2 Composition with an affine mapping

Suppose $f : \mathbf{R}^n \rightarrow \mathbf{R}$, $A \in \mathbf{R}^{n \times m}$, and $b \in \mathbf{R}^n$. Define $g : \mathbf{R}^m \rightarrow \mathbf{R}$ by

$$g(x) = f(Ax + b)$$

with $\mathbf{dom}\, g = \{x | Ax + b \in \mathbf{dom}\, f\}$. Then if f is convex, so is g; if f is concave, so is g.

It is easy to directly establish the defining inequality (2.1). We can also establish the result using epigraphs. Suppose f is convex, so $\mathbf{epi}\, f$ is convex. The epigraph of g is the partial sum of an affine set and $\mathbf{epi}\, f$, hence is convex:

$$\mathbf{epi}\, g = \{\ (x, t) \mid (x, y) \in \mathcal{A},\ (y, t) \in \mathbf{epi}\, f \text{ for some } y\ \},$$

where $\mathcal{A} = \{(x, y) | y = Ax + b\}$. Hence, g is convex.

2.2.3 Pointwise maximum and supremum

If f_1 and f_2 are convex functions then their *pointwise maximum* f, defined by

$$f(x) = \max\{f_1(x), f_2(x)\},$$

(with $\mathbf{dom}\, f = \mathbf{dom}\, f_1 \cap \mathbf{dom}\, f_2$) is also convex. This property is easily verified: if $0 \le \theta \le 1$ and $x,\ y, then \in \mathbf{dom}\, f$, then

$$\begin{array}{rcl} f(\theta x + (1-\theta)y) & = & \max\{f_1(\theta x + (1-\theta)y), f_2(\theta x + (1-\theta)y)\} \\ & \le & \max\{\theta f_1(x) + (1-\theta)f_1(y), \theta f_2(x) + (1-\theta)f_2(y)\} \\ & \le & \theta \max\{f_1(x), f_2(x)\} + (1-\theta)\max\{f_1(y), f_2(y)\} \\ & = & \theta f(x) + (1-\theta)f(y), \end{array}$$

which establishes convexity of f. It is easily shown that if $f_1, \ldots, f_m$ are convex, then their pointwise maximum

$$f(x) = \max\{f_1(x), \ldots, f_m(x)\}$$

is also convex.

Example. *Piecewise linear functions.* The function

$$f(x) = \max_{i=1,\ldots,L} \{a_i^T x + b_i\}$$

defines a piecewise linear (or really, affine) function (with L or fewer regions). It is convex since it is the pointwise maximum of affine functions.

The converse can also be shown: any piecewise linear convex function with L or fewer regions can be expressed in this form. (See exercises.)

Example. *Sum of r largest components.* For $x \in \mathbf{R}^n$ we denote by $x_{[i]}$ the ith largest component of x, *i.e.*, $x_{[1]},\ x_{[2]}, \ldots, x_{[n]}$ are the components of x sorted in decreasing order. Then the function

$$f(x) = \sum_{i=1}^{r} x_{[i]},$$

i.e., the sum of the r largest elements of x, is a convex function. This can be seen by writing it 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 $\begin{pmatrix} n \\ r \end{pmatrix}$ linear functions.

As an extension it can be shown that the function $\sum_{i=1}^{r} w_i x_{[i]}$ is convex, provided $w_1 \ge w_2 \ge \cdots \ge w_n \ge 0$. (See exercises.)

The pointwise maximum property extends to the pointwise supremum over an infinite set of convex functions. If for each $y \in \mathcal{A}$, $f(x,y)$ is convex in x, then the function g, defined as

$$g(x) = \sup_{y \in \mathcal{A}} f(x,y) \qquad (2.7)$$

is convex in x. Here the domain of g is (by our extended valued function convention)

$$\mathbf{dom}\, g = \left\{ x \;\middle|\; (x,y) \in \mathbf{dom}\, f \text{ for all } y \in \mathcal{A},\ \sup_{y \in \mathcal{A}} f(x,y) < \infty \right\}.$$

Similarly, the pointwise infimum of a set of concave functions is a concave function.

In terms of epigraphs, the pointwise supremum of functions corresponds to the intersection of epigraphs: with f, g, and $\mathcal{A}$ as defined in (2.7), we have

$$\mathbf{epi}\, g = \bigcap_{y \in \mathcal{A}} \mathbf{epi}\, f(\cdot, y).$$

Thus, the result follows from the fact that the intersection of a family of convex sets is convex.

Example. *Support function of a set.* Let $S \subseteq \mathbf{R}^n$, with $S \neq \emptyset$. The *support function* associated with the set S is defined as

$$f(x) = \sup\{\ x^T y \mid y \in S\ \}$$

(and, naturally, $\mathbf{dom}\, f = \{\ x \mid \sup_{y \in S} x^T y < \infty\ \}$).

For each $y \in S$, $x^T y$ is a linear function of x, so f (which is the pointwise supremum of a family of linear functions) is convex.

Example. *Distance to farthest point of a set.* Let $S \subseteq \mathbf{R}^n$. The distance to the farthest point of S,

$$f(x) = \sup_{y \in S} \|x - y\|,$$

is convex. To see this, note that for any y, the function $\|x - y\|$ is convex in x. Since f is the pointwise supremum of a family of convex functions (indexed by $y \in S$), it is a convex function of x.

Example. *Least-squares cost as function of weights.* Let $a_1, \ldots, a_n \in \mathbf{R}^m$. In a weighted least-squares problem we minimize the objective function $\sum_{i=1}^n w_i (a_i^T x - b_i)^2$ over $x \in \mathbf{R}^m$. We refer to w_i as *weights*, and allow negative w_i (which opens the possibility that the objective function is unbounded below).

We define the (optimal) *weighted least-squares cost* as

$$g(w) = \inf_x \sum_{i=1}^n w_i (a_i^T x - b_i)^2$$

(which is $-\infty$ if $\sum_i w_i (a_i^T x - b_i)^2$ is unbounded below).

Then g is a concave function of $w \in \mathbf{R}^n$, since it is the infimum of a family of linear functions of w (indexed by $x \in \mathbf{R}^m$).

In fact we can derive an explicit expression for g:

$$g(w) = \begin{cases} \sum_{i=1}^n w_i b_i^2 \left(1 - w_i a_i^T F^\dagger a_i\right), & F = \sum_{i=1}^n w_i a_i a_i^T \succeq 0, \\ -\infty & \text{otherwise,} \end{cases}$$

where $X^\dagger$ denotes the pseudo-inverse of a matrix X (see the appendix on linear algebra). Convexity of g from this expression is hardly obvious.

Example. *Maximum eigenvalue of a symmetric matrix.* Let $f(X) = \lambda_{\max}(X)$, with $\mathbf{dom}\, f = \{X \in \mathbf{R}^{m\times m} | X = X^T\}$. Then f is convex.

We can express f as

$$f(X) = \sup\left\{ y^T X y \;\middle|\; \|y\| = 1 \right\},$$

i.e., as the pointwise supremum of (an infinite number of) linear functions of X (*i.e.*, $y^T X y$) indexed by y. Convexity of f follows.

Example. *Norm of a matrix.* Consider $f(X) = \|X\|$ with $\mathbf{dom}\, f = \mathbf{R}^{p\times q}$, and the norm denotes the spectral norm or maximum singular value. Convexity of f follows from

$$f(X) = \sup\{u^T X v | \|u\| \leq 1,\ \|v\| \leq 1\},$$

which shows it is the pointwise supremum of a family of affine functions of X.

As a generalization suppose $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on $\mathbf{R}^p$ and $\mathbf{R}^q$, respectively. Then the induced norm

$$\|X\|_{a,b} = \sup_{v\neq 0} \frac{\|Xv\|_a}{\|v\|_b}$$

is a convex function of $X \in \mathbf{R}^{p\times q}$, since it can be expressed as

$$f(X) = \sup\{u^T X v | \|u\|_{a*} \leq 1,\ \|v\| \leq 1\},$$

where $\|\cdot\|_{a*}$ is the dual norm of $\|\cdot\|_a$.

These examples illustrate a good method for establishing convexity of a function, *i.e.*, by expressing it as the pointwise supremum of a family of affine functions. Except for a technical condition, a sort of converse holds, *i.e.*, (almost) every convex function can be expressed as the pointwise supremum of a family of affine functions. As an example of a simple result, if $f : \mathbf{R}^n \to \mathbf{R}$ is convex, with $\mathbf{dom}\, f = \mathbf{R}^n$, then

$$f(x) = \sup\{\ g(x) \mid g \text{ affine},\ g(z) \leq f(z) \text{ for all } z\ \}. \tag{2.8}$$

In other words, f is the pointwise supremum of the set of all affine global underestimators of it. We can give the general idea of the proof here, and leave the details (and case where $\mathbf{dom}\, f \neq \mathbf{R}^n$) to the exercises.

The epigraph of f is, of course, a convex set. Hence we can find a supporting hyperplane to it at $(x, f(x))$. Since $\mathbf{dom}\, f$ is all of $\mathbf{R}^n$, this hyperplane cannot be vertical. Hence it is the graph of an affine function g. It can then be shown that g is a global underestimator of f, and that $g(x) = f(x)$. The assertion (2.8) follows.

2.2.4 Composition

In this section we examine conditions on $h : \mathbf{R}^k \to \mathbf{R}$ and $g : \mathbf{R}^n \to \mathbf{R}^k$ that guarantee convexity or concavity of their composition $f : \mathbf{R}^n \to \mathbf{R}$, defined by $f(x) = h(g(x))$ (and, of course, $\mathbf{dom}\, f = \{\ x \in \mathbf{dom}\, g \mid g(x) \in \mathbf{dom}\, h\ \}$).

We first consider the case $k = 1$, so $h : \mathbf{R} \to \mathbf{R}$ and $g : \mathbf{R}^n \to \mathbf{R}$. We can restrict ourselves to the case $n = 1$ (since convexity is determined by the behavior of a function on arbitrary lines that intersect its domain). We start by assuming that h and g are twice differentiable, so

$$f'' = (g')^2 h'' + g'' h', \tag{2.9}$$

and convexity (concavity) of f simply means $f'' \geq 0$ ($f'' \leq 0$). From (2.9) we can see that

- f is convex if g is convex and h is convex and nondecreasing
- f is convex if g is concave and h is convex and nonincreasing
- f is concave if g is concave and h is concave and nondecreasing
- f is concave if g is convex and h is concave and nonincreasing

In fact it is easy to directly prove these composition rules in the general case $n > 1$, without assuming differentiability of f and g. For example, suppose g is convex and h is convex and nondecreasing. With $x,\ y \in \mathbf{R}^n$ and $0 \leq \theta \leq 1$ we have

$$\begin{aligned} f(\theta x + (1-\theta)y) &= h(g(\theta x + (1-\theta)y)) \\ &\leq h(\theta g(x) + (1-\theta)g(y)) \\ &\leq \theta h(g(x)) + (1-\theta)h(g(y)). \end{aligned}$$

The first inequality follows from convexity of g and monotonicity of h; the second inequality follows from convexity of h.

Examples.

- If g is convex then $f(x) = \exp g(x)$ is convex.
- If g is concave and positive, then $\log g(x)$ is concave.
- If g is concave and positive, then $1/g(x)$ is convex.
- If g is convex and nonnegative and $p \geq 1$, then $g(x)^p$ is convex.
- If f is convex then $-\log(-f(x))$ is convex on $\{x \mid f(x) < 0\}$.

These composition results can also be derived using the epigraphs of the functions. For example, suppose that $h : \mathbf{R} \to \mathbf{R}$ is convex and nondecreasing, and $g : \mathbf{R}^n \to \mathbf{R}$ is convex, so $f(x) = h(g(x))$ is convex. Here $\mathbf{epi}\, f$ is the partial sum of $\mathbf{epi}\, h$ and $\mathbf{epi}\, g$, *i.e.*,

$$\mathbf{epi}\, f = \{(x,t) | \exists u\ (x,u) \in \mathbf{epi}\, g,\ (u,t) \in \mathbf{epi}\, h\},$$

and hence is convex (see §1.3.2).

We now turn to the more complicated case when $k \geq 1$. Suppose

$$f(x) = h(g_1(x), \ldots, g_k(x))$$

with $h : \mathbf{R}^k \to \mathbf{R}$, $g_i : \mathbf{R}^n \to \mathbf{R}$. Again without loss of generality we can assume $n = 1$.

As in the case $k = 1$, we start by assuming the functions are twice differentiable, in order to discover the composition rules. We have

$$f'' = \nabla h^T \begin{bmatrix} g_1'' \\ \vdots \\ g_k'' \end{bmatrix} + \begin{bmatrix} g_1' \\ \vdots \\ g_k' \end{bmatrix}^T \nabla^2 h \begin{bmatrix} g_1' \\ \vdots \\ g_k' \end{bmatrix}. \tag{2.10}$$

Again the issue is to determine conditions under which $f'' \geq 0$ (or $f'' \leq 0$ for concavity). From (2.10) we can derive many rules, for example:

- If h is convex and nondecreasing in each argument (*i.e.*, $h_i' \geq 0$), and g_i are convex, then f is convex.
- If h is convex and nonincreasing in each argument, and g_i are concave, then f is convex.
- If h is concave and nondecreasing in each argument, and g_i are convex, then f is concave.
- If g_i are convex and nondecreasing, and h is increasing in each argument and satisfies $(\nabla^2 h)_{ij} \geq 0$, then f is convex.

As in the scalar case, it is possible to prove these composition results directly, without assuming differentiability or $n = 1$.

Examples.

- Let $h(z) = z_{[1]} + \cdots + z_{[r]}$, *i.e.*, h is the sum of the r largest components of z. Then h is convex and nondecreasing in each argument. Suppose $g_1, \ldots, g_k$ are convex functions on $\mathbf{R}^n$. Then the composition function $f = h(g)$, *i.e.*, the pointwise sum of the r largest g_i's, is convex.
- $h(z) = \log \sum_{i=1,\ldots,k} \exp z_i$ is convex and nondecreasing in each argument, so $\log \sum_{i=1,\ldots,m} \exp g_i$ is convex whenever g_i are.
- For $0 < p \leq 1$, $h(z) = (\sum z_i^p)^{1/p}$ is concave and nondecreasing in each component on $\mathbf{R}_+^n$. So if g_i are convex and positive, $f(x) = (\sum g_i(x)^p)^{1/p}$ is concave.
- For $p \geq 1$, $h(z) = (\sum z_i^p)^{1/p}$ is convex and nondecreasing on $\mathbf{R}_+^n$. Therefore $h(x) = (\sum g_i(x)^2)^{1/2}$ is convex if g_i are convex and positive.

2.2.5 Minimization

We have seen that the maximum or supremum of an arbitrary family of convex functions is convex. It turns out that some special forms of minimization also yield convex functions. If f is convex in (x, y), and C is a convex set, then the function

$$g(x) = \inf_{(x,y) \in C} f(x, y) \tag{2.11}$$

is convex in x. Here, of course, we take

$$\mathbf{dom}\, f = \{x \mid \exists y,\ (x, y) \in C,\ \inf_{x,y \in C} f(x, y) > -\infty\}$$

(which can be shown to be convex).

We prove this by verifying Jensen's inequality for $x_1,\ x_2 \in \mathbf{dom}\, g$. Let $\epsilon > 0$. Then there are $y_1,\ y_2 \in C$ such that $f(x_i, y_i) \leq g(x_i) + \epsilon$ for $i = 1,\ 2$. Now let $\theta \in [0,1]$. Then we have

$$\begin{aligned} g(\theta x_1 + (1-\theta)x_2)) &= \inf_{y \in C} f(\theta x_1 + (1-\theta)x_2, y) \\ &\leq f(\theta x_1 + (1-\theta)x_2, \theta y_1 + (1-\theta)y_2) \\ &\leq \theta f(x_1, y_1) + (1-\theta) f(x_2, y_2) \\ &= \theta g(x_1) + (1-\theta) g(x_2) + \epsilon. \end{aligned}$$

Since this holds for any $\epsilon > 0$, we have

$$g(\theta x_1 + (1-\theta)x_2) \leq \theta g(x_1) + (1-\theta) g(x_2).$$

The result can also be seen in terms of epigraphs. With f, g, and C defined as in (2.11), we have

$$\mathbf{epi}\, g = \{(x,t) | \exists y,\ (x,y,t) \in \mathbf{epi}\, f,\ y \in C\}.$$

Thus $\mathbf{epi}\, g$ is convex, since it is the projection of a convex set on a subset of its components.

Example. *Schur complement.* Suppose the quadratic function $f(x,y) = x^T Ax + 2x^T By + y^T Cy$ is convex, *i.e.*,

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \succeq 0.$$

We can express $g(x) = \inf_y f(x,y)$ as

$$g(x) = x^T (A - BC^\dagger B^T) x,$$

where $C^\dagger$ is the pseudo-inverse of C. By the minimization rule, g is convex, so we conclude that $A - BC^\dagger B^T \succeq 0$.

(If C is invertible, *i.e.*, $C \succ 0$, then the matrix $A - BC^{-1}B^T$ is called the *Schur complement* of C in the matrix $\begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$.)

Example. *Distance to a set.* The distance of a point x to a set $S \subseteq \mathbf{R}^n$ is defined as

$$\mathrm{dist}(x, S) = \inf_{y \in S} \|x - y\|.$$

The function $\|x - y\|$ is convex in (x,y), so if the set S is convex, the distance function $\mathrm{dist}(x, S)$ is a convex function of x.

Example. Suppose h is convex. Then the function g defined as

$$g(x) = \inf\{h(y) \mid Ay = x\}$$

is convex. To see this we define f by

$$f(x,y) = \begin{cases} h(y) & \text{if } Ay = x \\ +\infty & \text{otherwise,} \end{cases}$$

which is convex in (x,y). Then g is the minimum of f over y, and hence is convex. (It is not hard to show directly that g is convex.)

2.2.6 Perspective of a function

If $f : \mathbf{R}^n \to \mathbf{R}$, then the *perspective* of f is the function $g : \mathbf{R}^{n+1} \to \mathbf{R}$ defined by

$$g(x,t) = tf(x/t)$$

for $t > 0$. Its domain is, naturally,

$$\mathbf{dom}\, g = \{(x,t) | x/t \in \mathbf{dom}\, f, t > 0\}$$

(which is easily shown to be convex). If f is convex, then the perspective function g is convex.

This can be proved several ways, for example, direct verification of the defining inequality (2.1) (see exercises). We give a short proof here using epigraphs and the perspective mapping on $\mathbf{R}^{n+1}$ described in §1.3.3 (which will also explain the name 'perspective'). For $t > 0$ we have

$$\begin{aligned} (x,t,s) \in \mathbf{epi}\, g &\iff tf(x/t) \leq s \\ &\iff f(x/t) \leq s/t \\ &\iff (x/t, s/t) \in \mathbf{epi}\, f. \end{aligned}$$

Therefore $\mathbf{epi}\, g$ is the inverse image of $\mathbf{epi}\, f$ under the perspective mapping that takes (u,v,w) to $(u,v)/w$. It follows (see §1.3.3) that $\mathbf{epi}\, g$ is convex, so the function g is convex.

Example. *The quadratic-over-linear function.* Consider the convex function $f(x) = x^T x$ on $\mathbf{R}^n$. The perspective of f is given by

$$g(x,t) = t(x/t)^T(x/t) = \frac{x^T x}{t},$$

(for $t > 0$), which is convex in (x,t). As a consequence, we see that

$$f(x,t) = -\log(t^2 - x^T x)$$

is convex on $\mathbf{dom}\, f = \{(x,t) \mid \|x\| < t\}$. The function can be written as a sum of two convex functions

$$f(x,t) = -\log t - \log(t - x^T x/t).$$

Convexity of the second term follows from convexity of g and the composition rules mentioned above.

Example. $f(x) = -\log x$ on $\{x | x > 0\}$. Then

$$g(x,t) = -t\log(x/t) = t\log t - t\log x$$

is convex (in (x,t)). For $x = 1$, this reduces to the negative entropy function.

From convexity of g we can establish convexity of several interesting related functions. It follows that the *Kullback-Leibler* distance between $x > 0$ and $t > 0$, given by

$$D_{\mathrm{kl}}(x,t) = x\log(x/t) - x + t,$$

is convex. ($D_{\rm kl}$ also satisfies $D_{\rm kl} \geq 0$ and $D_{\rm kl} = 0$ if and only if $x = t$; see exercises.)

It also follows that the function

$$\sum_{i=1}^{n} (t_i \log t_i - t_i \log x)$$

is convex in $(x, t_1, \ldots, t_n)$ (for $x,\ t_i > 0$). Taking $x = \sum_{i=1}^n t_i$, we conclude that the *normalized negative entropy* function,

$$\sum_{i=1}^{n} t_i \log t_i - \left(\sum_{i=1}^{n} t_i\right) \log \sum_{i=1}^{n} t_i = \sum_{i=1}^{n} \frac{t_i}{\sum_{j=1}^n t_j} \log \frac{t_i}{\sum_{j=1}^n t_j}$$

is convex. (Note that $t_i / \sum_{j=1}^n t_j$ normalizes t to be a probability distribution; the normalized entropy is the entropy of this distribution.)

Example. Suppose $f : \mathbf{R}^m \to \mathbf{R}$ is convex, and $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$, $c \in \mathbf{R}^n$, and $d \in \mathbf{R}$. We define

$$g(x) = (c^T x + d) f \left((Ax + b)/(c^T x + d) \right)$$

with

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

Then g is convex.

2.3 The conjugate function

In this section we introduce an operation that will play an important role in later chapters.

2.3.1 Definition and examples

Let $f : \mathbf{R}^n \to \mathbf{R}$. The function $f^\star : \mathbf{R}^n \to \mathbf{R}$, defined as

$$f^\star(y) = \sup_{x \in \mathbf{dom}\, f} \left(y^T x - f(x) \right), \tag{2.12}$$

is called the *conjugate* of the function f. The domain of the conjugate function consists of $y \in \mathbf{R}^n$ for which the supremum is finite, *i.e.*, for which the difference $x^T y - f(x)$ is bounded above on $\mathbf{dom}\, f$.

We see immediately that $f^\star$ is a convex function, since it is the pointwise supremum of a family of convex (indeed, affine) functions of y. This is true whether or not f is convex. (Note that when f is convex, the subscript $x \in \mathbf{dom}\, f$ is not necessary, since by convention, $y^T x - f(x) = -\infty$ for $x \notin \mathbf{dom}\, f$.)

We start with some simple examples, and then describe some rules for conjugating functions. This allows us to derive an analytical expression for the conjugate of many common convex functions.

Examples. We derive the conjugate of some convex functions on $\mathbf{R}$.

- *Affine function.* $f(x) = ax + b$. As a function of x, $yx - ax - b$ is bounded if and only if $y = a$, in which case it is constant. Therefore the domain of the conjugate function $f^\star$ is the singleton $\{a\}$, and $f^\star(a) = -b$.
- *Logarithm.* $f(x) = -\log x$, with $\mathbf{dom}\, f = \{x \mid x > 0\}$. The function $xy + \log x$ is unbounded above if $y \geq 0$ and reaches its maximum at $x = -1/y$ otherwise. Therefore, $\mathbf{dom}\, f^\star = \{y \mid y < 0\}$ and $f^\star(y) = -\log(-y) - 1$ for $y < 0$.
- *Exponential.* $f(x) = e^x$. $xy - e^x$ is unbounded if $y \leq 0$. For $y > 0$, $xy - e^x$ reaches its maximum at $x = \log y$, so we have $f^\star(y) = y \log y - y$. For $y = 0$, $f^\star(y) = \sup_x -e^x = 0$. In summary, $\mathbf{dom}\, f^\star = \mathbf{R}_+$ and $f^\star(y) = y\log y - y$ (with the interpretation $0 \log 0 = 0$).
- *Negative entropy.* $f(x) = x \log x$, with $\mathbf{dom}\, f = \mathbf{R}_+$ (and $f(0) = 0$). The function $xy - x\log x$ is bounded above on $\mathbf{R}_+$ for all y, hence $\mathbf{dom}\, f^\star = \mathbf{R}$. It attains its maximum at $x = e^{y-1}$, and substituting we find $f^\star(y) = e^{y-1}$.
- *Inverse.* $f(x) = 1/x$ on $\{x | x > 0\}$. For $y > 0$, $yx - 1/x$ is unbounded above. For $y = 0$ this function has supremum 0; for $y < 0$ the supremum is attained at $x = (-y)^{-1/2}$. Therefore we have $f^\star(y) = -2(-y)^{1/2}$, with $\mathbf{dom}\, f^\star = \mathbf{R}_-$.

Example. *Strictly convex quadratic function.* Consider $f(x) = \frac{1}{2}x^T Qx$, with $Q = Q^T \succ 0$. The function $x^T y - \frac{1}{2}x^T Qx$ is bounded above as a function of x for all y. It attains its maximum at $x = Q^{-1}y$, so

$$f^\star(y) = \frac{1}{2} y^T Q^{-1} y.$$

Example. *Log determinant.* $f(X) = \log\det X^{-1}$ on $\{X = X^T \in \mathbf{R}^{n\times n} \mid X \succ 0\}$. The conjugate function is defined as

$$f^\star(Y) = \sup_{X \succ 0} \mathbf{Tr}\, XY + \log\det X.$$

The argument of the supremum is unbounded above unless $Y \prec 0$; when $Y \prec 0$ we can find the maximizing X by setting the gradient with respect to X equal to zero:

$$\nabla_X (\mathbf{Tr}\, XY + \log\det X) = Y + X^{-1} = 0$$

(see §A.3.2), which yields $X = -Y^{-1}$.

Therefore we have

$$f^\star(Y) = \log\det(-Y)^{-1} - n$$

on $\mathbf{dom}\, f^\star = \{Y = Y^T \mid Y \prec 0\}$.

Example. *Indicator function.* Suppose f is the indicator function of a (not necessarily convex) set $S \subseteq \mathbf{R}^n$, *i.e.*, $f(x) = 0$ on $\mathbf{dom}\, f = S$. Its conjugate is

$$f^\star(y) = \sup_{x\in S} y^T x,$$

which is the support function of the set S.

2.3.2 Basic properties

Fenchel's inequality

From the definition of conjugate function, we immediately obtain the inequality

$$f(x) + f^\star(y) \geq x^T y$$

for all x, y. This is called *Fenchel's inequality* (or *Young's inequality* if f is differentiable).

For example with $f(x) = (1/2)x^T Qx$, where $Q = Q^T \succ 0$, we obtain the inequality

$$x^T y \leq (1/2)x^T Qx + (1/2)y^T Q^{-1} y.$$

Conjugate of the conjugate

The examples above, and the name 'conjugate', suggest that the conjugate of the conjugate of a convex function is the original function. This is the case provided a technical condition holds: if f is convex, and $\mathbf{epi}\, f$ is a closed set, then $f^{\star\star} = f$. For example, if $\mathbf{dom}\, f = \mathbf{R}^n$, then we have $f^{\star\star} = f$, *i.e.*, the conjugate of the conjugate of f is f again (see exercises).

This means that if f is convex and $\mathbf{epi}\, f$ is closed, then for each x there exists a y such that Fenchel's inequality is tight.

2.3.3 Elementary calculus of conjugation

The task of finding an analytical expression for the conjugate of a function can be simplified by a few simple rules which we describe in this section. Further rules are explored in the exercises.

Differentiable functions

The conjugate of a differentiable function f is also called the *Legendre transform* of f. (To distinguish the general definition from the differentiable case, the term *Fenchel conjugate* is also sometimes used instead of conjugate.)

Suppose f is convex and differentiable, with $\mathbf{dom}\, f = \mathbf{R}^n$. Any maximizer $x^\star$ of $x^T y - f(x)$ satisfies $y = \nabla f(x^\star)$, and conversely, if $x^\star$ satisfies $y = \nabla f(x^\star)$, then $x^\star$ maximizes $x^T y - f(x)$.

Therefore, if $y = \nabla f(x^\star)$, we have

$$f^\star(y) = x^{\star T} \nabla f(x^\star) - f(x^\star).$$

This allows us to determine $f^\star(y)$ for any y for which we can solve the gradient equation $y = \nabla f(z)$ for z.

We can express this another way. Let $z \in \mathbf{R}^n$ be arbitrary and define $y = \nabla f(z)$. Then we have

$$f^\star(y) = z^T \nabla f(z) - f(z).$$

Nonnegative scaling

For $a > 0$, the conjugate of $af(x)$ is $af^\star(y/a)$.

Composition with affine transformation

Suppose $A \in \mathbf{R}^{n\times n}$ is nonsingular and $b \in \mathbf{R}^n$. Then the conjugate of $g(x) = f(Ax+b)$ is

$$g^\star(y) = f^\star(A^{-T}y) - b^TA^{-T}y,$$

and $\mathbf{dom}\, g^\star = A^T\, \mathbf{dom}\, f^\star$. For example,

- the conjugate of $f(ax)$ ($a \neq 0$) is $f^\star(y/a)$,
- the conjugate of $f(x+b)$ is $f^\star(y) - b^Ty$.

As an extension, suppose $A \in \mathbf{R}^{m\times n}$ with $\mathbf{Rank}\, A = m$, and g is defined as $g(x) = f(Ax+b)$. Then

$$g^\star(y) = \begin{cases} f^\star(z) - b^Tz & y = A^Tz, z \in \mathbf{dom}\, f^\star \\ +\infty & \text{otherwise.} \end{cases}$$

The proof is as follows. First suppose $y = A^Tz$ for some z. By our assumption on the rank of A, z is unique. We can write the conjugate as

$$\begin{aligned} g^\star(y) &= \sup_{Ax+b\in\mathbf{dom}\, f} y^Tx - f(Ax+b) \\ &= \sup_{Ax+b\in\mathbf{dom}\, f} z^T(Ax+b) - f(Ax+b) - b^Tz \\ &= \begin{cases} f^\star(z) - b^Tz & z \in \mathbf{dom}\, f^\star \\ +\infty & \text{otherwise.} \end{cases} \end{aligned}$$

Next, assume that y does not lie in the range of A^T. This means there exists a v with $Av = 0$ and $y^Tv \neq 0$. Let x_0 be any point that satisfies $Ax_0 + b \in \mathbf{dom}\, f$. We have

$$\begin{aligned} g^\star(y) &\geq \sup_t y^T(x_0+tv) - f(A(x_0+tv)+b) \\ &= \sup_t y^Tx_0 + tv^Ty - f(Ax_0+b) \\ &= +\infty, \end{aligned}$$

wich shows that $y \notin \mathbf{dom}\, g^\star$.

Sums of independent functions

If $f(u,v) = f_1(u) + f_2(v)$, where f_1 and f_2 are convex functions with conjugates $f_1^\star$ and $f_2^\star$, respectively, then

$$f^\star(w,z) = f_1^\star(w) + f_2^\star(z).$$

In other words, the conjugate of the sum of *independent* convex functions is the sum of the conjugates. ('Independent' means they are functions of different variables.)

2.4 Quasiconvex functions

2.4.1 Definition and examples

A function f is called *quasiconvex* (or *unimodal*) if its domain and all its sublevel sets

$$S_\alpha = \{x \in \mathbf{dom}\, f \mid f(x) \le \alpha\}$$

are convex. A function is *quasiconcave* if $-f$ is quasiconvex, *i.e.*, every superlevel set $\{x \mid f(x) \ge \alpha\}$ is convex. A function that is both quasiconvex and quasiconcave is called *quasilinear*.

Convex functions have convex sublevel sets, and hence are quasiconvex. But simple examples show that the converse is not true.

Examples. Some examples on $\mathbf{R}$:

- *Logarithm.* $\log x$ on $\{x \mid x > 0\}$ is quasiconvex (and quasiconcave, hence quasilinear).
- *Ceiling function.* $\mathrm{ceil}(x) = \min\{z \in \mathbf{Z} | z \ge x\}$ is quasiconvex (and quasiconcave).

These examples show that quasiconvex functions can be concave, or discontinuous (indeed, integer valued). We now give some examples on $\mathbf{R}^n$.

Example. *Length of a vector.* We define the *length* of $x \in \mathbf{R}^n$ as the largest index of a nonzero component, *i.e.*,

$$f(x) = \max\{\ k \le n \mid x_i = 0 \text{ for } i = k+1, \ldots, n\ \}.$$

This function is quasiconvex on $\mathbf{R}^n$; its sublevel sets are subspaces.

Example. Consider $f : \mathbf{R}^2 \to \mathbf{R}$, with $\mathbf{dom}\, f = \mathbf{R}^2_+$ and $f(x_1, x_2) = x_1 x_2$. The function is neither convex or concave since its Hessian

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

is indefinite; it has one positive and one negative eigenvalue. The function f is quasiconcave, however, since the superlevel sets

$$\left\{ x \in \mathbf{R}^2_+ \;\middle|\; x_1 x_2 \ge \alpha \right\}$$

are convex sets for all α.

Example. *Linear-fractional function.* The function

$$f(x) = \frac{a^T x + b}{c^T x + d},$$

with $\mathbf{dom}\, f = \{x | c^T x + d > 0\}$, is quasiconvex (and quasiconcave).

Its α-sublevel set is

$$\begin{aligned} S_\alpha &= \{x \mid c^T x + d > 0,\ (a^T x + b)/(c^T x + d) \le \alpha\} \\ &= \{x \mid c^T x + d > 0,\ (a - \alpha c)^T x + b - \alpha d \le 0\}, \end{aligned}$$

which is convex.

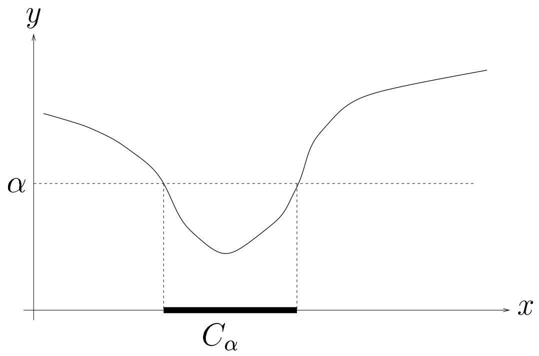

Figure 2.5: Graph and sublevel set of a quasiconvex function.

Example. *Distance ratio function.* Suppose $a, b \in \mathbf{R}^n$, and define

$$f(x) = \frac{\|x-a\|}{\|x-b\|},$$

i.e., the ratio of the Euclidean distance to a to the distance to b. Then f is quasiconvex on the halfspace $\{x \mid \|x-a\| \leq \|x-b\|\}$. To see this, we consider the α-sublevel set of f, with $\alpha \leq 1$ since $f(x) \leq 1$ on the halfspace $\{x \mid \|x-a\| \leq \|x-b\|\}$. This sublevel set is the set of points satisfying

$$\|x-a\| \leq \alpha\|x-b\|.$$

Squaring both sides, and rearranging terms, we see that this equivalent to

$$(1-\alpha)x^T x - 2(a-\alpha b)^T x + a^T a - \alpha b^T b \leq 0.$$

This describes a convex set (in fact a ball) if $\alpha \leq 1$.

2.4.2 Basic properties

The examples above show that quasiconvexity is a considerable generalization of convexity. Still, many of the properties of convex functions hold, or have analogs, for quasiconvex functions. For example, a function f is quasiconvex if and only if its restriction to any line intersecting its domain is quasiconvex.

There is a variation on Jensen's inequality that characterizes quasiconvexity. A function f is quasiconvex if and only if for any $x_,\ x_2 \in \mathbf{dom}\, f$ and $0 \leq \theta \leq 1$,

$$f(\theta x_1 + (1-\theta)x_2) \leq \max\{f(x_1), f(x_2)\}, \tag{2.13}$$

i.e., the graph of f on the segment $[x_1, x_2]$ lies below the horizontal line segment $\max(f(x_1), f(x_2))$. The inequality (2.13) is sometimes called Jensen's inequality for quasiconvex functions. Figure (2.5) illustrates this.

Example. *Rank of positive semidefinite matrix.* The function $\mathbf{Rank}(X)$ is quasiconcave on $\{\ X \in \mathbf{R}^{n\times n} \mid X = X^T \succeq 0\ \}$. This follows from the modified Jensen inequality (2.13),

$$\mathbf{Rank}(X+Y) \geq \max\{\mathbf{Rank}(X), \mathbf{Rank}(Y)\}$$

which holds for symmetric, positive semidefinite X, Y. We can also see the result by examining the superlevel sets

$$C_k = \{ X \in \mathbf{R}^{n \times n} \mid X = X^T \succeq 0, \ \mathbf{Rank}(X) \geq k \}$$

which are convex cones with the point 0 removed, generated by all matrices of the form FF^T with $F \in \mathbf{R}^{n \times k}$, $\mathbf{Rank}(F) = k$.

Remark. *Nested convex set characterization.* Any function is characterized by its sublevel sets, in the following way. The sublevel sets S_α of any function are nested, *i.e.*, $S_{\alpha_1} \subseteq S_{\alpha_2}$ whenever $\alpha_1 \leq \alpha_2$.

Conversely, if T_α is any nested family of sets, we can associate the function defined as

$$f(x) = \inf\{\alpha \mid x \in T_\alpha\},$$

and $\mathbf{dom}\, f = \bigcup_\alpha T_\alpha$. (Here we either allow f to take on the value $-\infty$, or assume $\bigcap_\alpha T_\alpha = \emptyset$.) The sublevel sets of f are T_α.

For a quasiconvex function, the sublevel sets form a family of nested *convex* sets. This correspondence is one-to-one; to any nested family of convex sets we can associate the function whose sublevel sets are the family.

2.4.3 Differentiable quasiconvex functions

Suppose f is differentiable. Then f is quasiconvex if and only if for all x, $y \in \mathbf{dom}\, f$

$$f(y) \leq f(x) \Longrightarrow \nabla f(x)^T (y - x) \leq 0. \tag{2.14}$$

This is the analog of inequality (2.2), for quasiconvex functions.

Now suppose f is twice differentiable. If f is quasiconvex, then for all $x \in \mathbf{dom}\, f$,

$$y^T \nabla f(x) = 0 \Longrightarrow y^T \nabla^2 f(x) y \geq 0. \tag{2.15}$$

As a (partial) converse, if f satisfies

$$y^T \nabla f(x) = 0 \Longrightarrow y^T \nabla^2 f(x) y > 0 \tag{2.16}$$

for all $x \in \mathbf{dom}\, f$, then f is quasiconvex.

Proof. *Necessary condition.* Suppose the condition is not necessary, *i.e.*, there are $x \in \mathbf{dom}\, f$ and y with $y^T \nabla f(x) = 0$ and $y^T \nabla^2 f(x) y < 0$. Consider the function $h(t) = f(x + ty)$. From

$$h'(t) = y^T \nabla f(x + ty), \quad h''(t) = y^T \nabla^2 f(x + ty) y$$

we see that $h'(0) = 0$ and $h''(0) < 0$. This implies h is increasing in t for t small and negative, and decreasing in t for t small and positive. Therefore h cannot be quasiconvex.

Sufficient condition. Let $h(t) = f(x + ty)$ where $x \in \mathbf{dom}\, f$. Let $I = \{t \mid x + ty \in \mathbf{dom}\, f\}$. Three cases are possible under the condition: (a) $h'(t) > 0$ for all $t \in I$; (b) $h'(t) < 0$ for all $t \in I$; (c) $h'(t_0) = 0$ for some $t_0 \in I$, and $h''(t_0) > 0$. In the first case, h is strictly increasing on I. In the second case, h is strictly decreasing on I. In the third case, h is decreasing for $t < t_0$ and increasing for $t > t_0$. In all three cases, h is quasiconvex on I.

Since x and y were arbitrary, we have shown that f is quasiconvex when restricted to any line. Hence it is quasiconvex.

Example. *Quadratic function on a set.* When is a quadratic function $f(x) = x^T Ax + 2b^T x + c$ quasiconvex on a convex set C? For simplicity we assume C is open.

We first consider a quadratic function of one variable $h(t) = at^2 + 2bt + c$ on an open interval I. h is quasiconvex on I if it is nonincreasing ($at + b \leq 0$ for all $t \in I$), or nondecreasing ($at + b \geq 0$ for all $t \in I$), or if there is a point $t_0 \in I$, such that $at + b \leq 0$ for $t \leq t_0$ and $at + b \geq 0$ for $t \geq t_0$ (which implies $a \geq 0$). We can summarize all three cases by saying: if $at + b = 0$ for some $t \in I$, then $a \geq 0$.

Applying this condition to the restriction of f to an arbitrary line $\hat{x} + ty$ in the open set C, we can say that f is quasiconvex on C if for all $\hat{x} \in C$ and for all y, either $y^T Ay \geq 0$, or $y^T(Ax + b)$ does not change sign on the the interval $\{x \mid x = \hat{x} + ty \in C\}$. In other words, if for all $x \in C$:

$$y^T(Ax + b) = 0 \Longrightarrow y^T Ay \geq 0. \tag{2.17}$$

We can express this condition in two equivalent ways (see exercises).

- For all $x \in C$, there exists a $\lambda \geq 0$ such that $A + \lambda(Ax + b)(Ax + b)^T \succeq 0$.
- For all $x \in C$ the matrix

$$\begin{bmatrix} A & Ax + b \\ (Ax + b)^T & 0 \end{bmatrix}$$

  has at most one negative eigenvalue.

As an example, we verify that these conditions hold for $f(x) = -x_1 x_2$ on $C = \{x \in \mathbf{R}^2 \mid x \succ 0\}$. The first condition, (2.17), is obviously true: for all $x_1, x_2 > 0$,

$$y_1 x_2 + y_2 x_1 = 0 \Longrightarrow 2y_1 y_2 \leq 0.$$

To verify the second condition we note that $\lambda = 1/(x_1 x_2)$ gives

$$\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} + \frac{1}{x_1 x_2} \begin{bmatrix} -x_2 \\ -x_1 \end{bmatrix} \begin{bmatrix} -x_2 & -x_1 \end{bmatrix} = \begin{bmatrix} x_2/x_1 & 0 \\ 0 & x_1/x_2 \end{bmatrix} \succeq 0.$$

The third condition states that the matrix

$$B = \begin{bmatrix} 0 & -1 & -x_2 \\ -1 & 0 & -x_1 \\ -x_2 & -x_1 & 0 \end{bmatrix}$$

can have a most one negative eigenvalue. To verify this, we can use the property that a congruence transformation preserves the signature of a matrix, and note that

$$R^TBR = \mathbf{diag}(x_2/x_1, x_1/x_2, -2x_1x_2)$$

where

$$R = \begin{bmatrix} x_2/(2x_1) & -1/2 & x_2 \\ -1/2 & -x_1/(2x_2) & x_1 \\ -x_2 & -x_1 & -2x_1x_2 \end{bmatrix}.$$

2.4.4 Operations that preserve quasiconvexity

Nonnegative weighted maximum

A nonnegative weighted maximum of quasiconvex functions, *i.e.*,

$$f = \max\{w_1f_1, \ldots, w_mf_m\},$$

with $w_i \geq 0$ and f_i quasiconvex, is quasiconvex. The property extends to the general pointwise supremum

$$f(x) = \sup_{y \in C} w(y)g(x, y)$$

where $w(y) \geq 0$ and $g(x, y)$ is quasiconvex in x for each y. This fact can be easily verified: $f(x) \leq \alpha$ if and only if

$$w(y)g(x, y) \leq \alpha, \text{ for all } y \in C,$$

i.e., the α-sublevel set of f is the intersection of the α-sublevel sets of the functions $w(y)g(x, y)$ in the variable x.

Example. *Generalized eigenvalue.* The *largest generalized eigenvalue* of a pair of symmetric matrices (X, Y), with $Y \succ 0$, is defined as

$$\lambda_{\max}(X, Y) = \sup_{u \neq 0} \frac{u^TXu}{u^TYu} = \max\{\lambda | \det(\lambda Y - X) = 0\}.$$

This function is quasiconvex (on its domain, $\{(X, Y) | X = X^T,\ Y = Y^T \succ 0\}$).

To see this we note that

$$\lambda_{\max}(X, Y) = \sup_{u \neq 0} \frac{u^TXu}{u^TYu}.$$

For each $u \neq 0$, the function u^TXu/u^TYu is linear fractional, hence quasiconvex, so $\lambda_{\max}$ is the supremum of a family of quasiconvex functions.

Composition

If $h : \mathbf{R}^n \to \mathbf{R}$ is quasiconvex and $g : \mathbf{R} \to \mathbf{R}$ is nondecreasing, then $f(x) = g(h(x))$ is quasiconvex.

The composition of a quasiconvex function with an affine or linear-fractional transformation yields a quasiconvex function. If g is quasiconvex, then $f(x) = g(Ax+b)$ is quasiconvex, and $\tilde{f}(x) = g((Ax + b)/(c^Tx + d))$ is quasiconvex on the set

$$\left\{x \;\middle|\; c^Tx + d > 0,\ (Ax + b)/(c^Tx + d) \in \mathbf{dom}\, g\right\}.$$

Minimization

If $g(x, y)$ is quasiconvex jointly in x and y and C is a convex set, then the function

$$f(x) = \inf_{y \in C} g(x, y)$$

is quasiconvex.

Proof. From the definition of f, $f(x) \leq \alpha$ if and only if for any $\epsilon > 0$ there exists a $y \in C$ with $g(x, y) \leq \alpha + \epsilon$. Let x_1 and x_2 be two points in the α-sublevel set of f. Then for any $\epsilon > 0$, there exists y_1, $y_2 \in C$ with

$$g(x_1, y_1) \leq \alpha + \epsilon, \quad g(x_2, y_2) \leq \alpha + \epsilon,$$

and since g is quasiconvex in x and y, we also have

$$g(\theta x_1 + (1 - \theta)x_2, \theta y_1 + (1 - \theta)y_2) \leq \alpha + \epsilon,$$

for $0 \leq \theta \leq 1$. Hence $f(\theta x_1 + (1 - \theta)x_2) \leq \alpha$.

2.5 Log-concave and log-convex functions

2.5.1 Definition

A function $f : \mathbf{R}^n \to \mathbf{R}$ is *logarithmically concave* or *log-concave* if $f(x) > 0$ for all $x \in \mathbf{dom}\, f$ and $\log f$ is concave. It is said to be *logarithmically convex* or *log-convex* if $\log f$ is convex. Thus f is log-convex if and only if $1/f$ is log-concave. It is convenient to allow f to take on the value zero, in which case we take $\log f(x) = -\infty$. In this case we say f is log-concave if the extended-valued function $\log f$ is concave.

From the composition rules we know that e^h is convex if h is convex, so a log-convex function is convex. Similarly, a concave function is log-concave. It is also clear that a log-convex function is quasiconvex and a log-concave function is quasiconcave.

Examples. Some simple examples are:

- *Powers.* $f(x) = x^a$, on $\{x | x > 0\}$ is log-convex for $a \leq 0$ and log-concave for $a \geq 0$.
- *Exponentials.* $f(x) = e^{ax}$ is log-convex and log-concave.
- $\mathbf{erfc}(x)$ and $\Gamma(x)$ (see §2.1.5).
- *Determinant.* $\det X$ is log concave on $\{X \in \mathbf{R}^{n \times n} | X = X^T \succ 0\}$.
- *Elementary symmetric functions.* The kth degree elementary symmetric function of $x \in \mathbf{R}^n$, *i.e.*,

  $$S_k(x) = \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} x_{i_1} x_{i_2} \cdots x_{i_k},$$

  is log concave on $\{x \mid x \succ 0\}$.

Examples. Many common probability density functions are log-concave. Two examples are the multivariate normal distribution,

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-\frac{1}{2}(x-\bar{x})^T \Sigma^{-1}(x-\bar{x})},$$

and the exponential distribution ($\lambda_i > 0$) on $\mathbf{R}_+^n$,

$$f(x) = \left(\prod_{i=1}^{n} \lambda_i \right) e^{-(\lambda_1 x_1 + \cdots + \lambda_n x_n)}.$$

Another example is the uniform distribution over a convex set C,

$$f(x) = \begin{cases} 1/\alpha & x \in C \\ 0 & x \notin C \end{cases}$$

where α is the Lebesgue measure of C. In this case $\log f$ takes on the value $-\infty$ outside C, and $-\log \alpha$ on C, hence is concave.

As a more exotic example consider the Wishart distribution, defined as follows. Let $x_1, \ldots, x_p \in \mathbf{R}^n$ be independent Gaussian random vectors with zero mean and covariance $\Sigma \in \mathbf{R}^{n \times n}$, with $p > n$. The random matrix $X = \sum_{i=1}^{p} x_i x_i^T$ has the Wishart density

$$f(X) = \begin{cases} a\,(\det X)^{(p-n-1)/2}\, e^{-\frac{1}{2}\mathbf{Tr}\,\Sigma^{-1} X} & X = X^T \succ 0 \\ 0 & \text{otherwise,} \end{cases}$$

which is log-concave.

2.5.2 Properties

Twice differentiable log-convex/concave functions

Suppose f is twice differentiable. Then f is log-convex if and only if for all $x \in \mathbf{dom}\, f$,

$$f(x)\nabla^2 f(x) \succeq \nabla f(x) \nabla f(x)^T,$$

which can be expressed as the matrix inequality

$$\left[\begin{array}{cc} \nabla^2 f(x) & \nabla f(x) \\ \nabla f(x)^T & f(x) \end{array} \right] \succeq 0.$$

The function f is log-concave if and only if for all $x \in \mathbf{dom}\, f$,

$$f(x)\nabla^2 f(x) \preceq \nabla f(x) \nabla f(x)^T.$$

Multiplication, addition, and integration

Log-convexity and log-concavity are evidently closed under multiplication and positive scaling.

Let f and g be log-convex functions, *i.e.*, $F = \log f$ and $G = \log g$ are convex. From the composition rules for convex functions, it follows that

$$\log(\exp F + \exp G) = \log(f + g)$$

is convex. Therefore the sum of two log-convex functions is log-convex.

More generally, if $f(x, y)$ is log-convex in x for each $y \in C$ then

$$\int_C f(x, y)\, dy$$

is log-convex.

Example. *Laplace transform of nonnegative function.* Suppose $p : \mathbf{R}^n \to \mathbf{R}$ satisfies $p(x) \geq 0$ for all x. The Laplace transform of p,

$$P(z) = \int p(x) e^{-z^T x}\, dx,$$

is log-convex on $\mathbf{R}^n$. (Here $\mathbf{dom}\, P$ is, naturally, $\{z | P(z) < \infty\}$.)

Indeed, if p is a density, *i.e.*, satisfies $\int p(x)\, dx = 1$, then $\log P(z)$ is the *moment generating function* for p. It satisfies

$$\nabla \log P(0) = \mathbf{E}\, v, \quad \nabla^2 \log P(0) = \mathbf{E}(v - \mathbf{E}\, v)(v - \mathbf{E}\, v)^T,$$

where v is a random vector with density p. The first and second derivatives of $\log P$ give the mean and covariance of v. (Convexity of $\log P$ follows since the convariance of v is positive semidefinite.)

The sum of log-concave functions is not, in general, log-concave.

Integration of log-concave functions

In some special cases log-concavity is preserved by integration. If $f : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ is log-concave, then

$$g(x) = \int f(x, y)\, dy \tag{2.18}$$

is a log-concave function of x (on $\mathbf{R}^n$). (The integration here is over $\mathbf{R}^m$.) A proof of this result is not simple; see the notes and references.

This result has many important consequences, some of which we describe in the rest of this section. It implies, for example, that marginal distributions of log-concave probability densities are log-concave.

It also implies that log-concavity is closed under convolution, *i.e.*, if f and g are log-concave on $\mathbf{R}^n$, then the convolution

$$(f * g)(x) = \int f(x - y) g(y)\, dy$$

(To see this, note that $g(y)$ and $f(x - y)$ are log-concave in (x, y), hence the product $f(x - y)g(y)$ is; then the result (2.18) applies.)

Suppose $C \subseteq \mathbf{R}^n$ is a convex set and w is a random vector in $\mathbf{R}^n$ with log-concave probability density p. Then the function

$$f(x) = \mathbf{Prob}(x + w \in C)$$

is log-concave in x. To see this, express f as

$$f(x) = \int g(x+z)p(z)\, dz$$

where g is defined as

$$g(u) = \begin{cases} 1 & u \in C \\ 0 & u \notin C \end{cases}$$

(which is log-concave) and apply (2.18).

2.6 Convexity with respect to generalized inequalities

We now consider generalizations of the notions of monotonicity and convexity, using generalized inequalities instead of the usual ordering on $\mathbf{R}$.

2.6.1 Monotonicity with respect to a generalized inequality

Suppose $K \subseteq \mathbf{R}^n$ and $L \subseteq \mathbf{R}^m$ are convex cones that induce generalized inequalities $\preceq_K$ and $\preceq_L$. A function $f : \mathbf{R}^n \to \mathbf{R}^m$ is called K, L*-nondecreasing* if

$$x \preceq_K y \Longrightarrow f(x) \preceq_L f(y),$$

and K, L*-increasing* if

$$x \preceq_K y,\ x \neq y \Longrightarrow f(x) \prec_L f(y).$$

We define K, L*-nonincreasing* and K, L*-decreasing* functions in a similar way. We also say a function is K, L*-monotone* if it is K, L-nonincreasing or K, L-nondecreasing, and *strictly* K, L*-monotone* if it is K, L-decreasing or K, L-increasing.

We will restrict the rest of this section to the most important special case $m = 1$ (with the standard ordering $L = \mathbf{R}_+$). We then call a K, L-monotone function simply K-monotone, or monotone with respect to K (and similarly for nondecresaing, increasing, etc.).

Example. A function $f : \mathbf{R}^n \to \mathbf{R}$ is nondecreasing with respect to $\mathbf{R}^n_+$ if and only if

$$x_1 \leq y_1, \ldots, x_n \leq y_n \;\Rightarrow\; f(x) \leq f(y)$$

for all x, y. This is the same as saying that f, when restricted to any component x_i (*i.e.*, x_i is considered the variable while x_j for $j \neq i$ are fixed), is nondecreasing.

Example. *Monotone functions of symmetric matrices.* Some examples of monotone functions of the variable $X = X^T \in \mathbf{R}^{n \times n}$:

- $\mathbf{Tr}\, WX$, where $W = W^T$, is K-nondecreasing if $W \succeq 0$, and K-increasing if $W \succ 0$ (it is K-nonincreasing if $W \preceq 0$, and K-decreasing if $W \prec 0$).
- $\mathbf{Tr}\, X^{-1}$ is K-decreasing on the set of positive definite matrices.
- $\det X$ is K-increasing on the set of positive definite matrices.

Gradient conditions for monotonicity

Recall that a differentiable function $f : \mathbf{R} \to \mathbf{R}$ is nonincreasing if and only if $f' \geq 0$ for all $x \in \mathbf{dom}\, f$, and increasing if $f'(x) > 0$ for all $x \in \mathbf{dom}\, f$ (but the converse is not true). These conditions are readily extended to the case of monotonicity with respect to a generalized inequality. A differentiable function f is K-nondecreasing if and only if

$$\nabla f(x) \succeq_{K^*} 0 \tag{2.19}$$

for all $x \in \mathbf{dom}\, f$. Note the difference with the simple scalar case: the gradient must be nonnegative in the *dual* inequality. For the strict case, we have the following: If

$$\nabla f(x) \succ_{K^*} 0 \tag{2.20}$$

for all $x \in \mathbf{dom}\, f$ then f is K-increasing. As in the scalar case, the converse is not true.

Proof. First, assume that f satisfies (2.19) for all x, but is not K-nondecreasing, *i.e.*, there exist x, y with $x \preceq_K y$ and $f(y) < f(x)$. By differentiability of f there exists a $t \in [0, 1]$ with

$$\frac{d}{dt} f(x + t(y - x)) = \nabla f(x + t(y - x))^T (y - x) < 0.$$

Since $y - x \in K$ this means

$$-\nabla f(x + t(y - x)) \notin K^*,$$

which contradicts our assumption that (2.19) is satisfied everywhere. In a similar way it can be shown that (2.20) implies f is K-increasing.

It is also quite straightforward to see that it is necessary that (2.19) holds everywhere. Assume (2.19) does not hold for $x = z$. By the definition of dual cone this means there exists a $v \in K$ with

$$\nabla g(z)^T v < 0.$$

Now consider $h(t) = g(z + tv)$ as a function of t. We have $h'(0) = \nabla g(z)^T v < 0$, and therefore there exists $t > 0$ with $h(t) = g(z + tv) < h(0) = g(z)$, which means g is not K-nondecreasing.

2.6.2 Convexity with respect to a generalized inequality

Definition and examples

Suppose $K \subseteq \mathbf{R}^m$ is a convex cone that induces a generalized inequality $\preceq_K$. We say $f : \mathbf{R}^n \to \mathbf{R}^m$ is *K-convex* if for all x, y, and $0 \leq \theta \leq 1$,

$$f(\theta x + (1 - \theta)y) \;\preceq_K\; \theta f(x) + (1 - \theta)f(y).$$

The function is *strictly K-convex* if

$$f(\theta x + (1 - \theta)y) \;\prec_K\; \theta f(x) + (1 - \theta)f(y).$$

for all $x \neq y$ and $0 < \theta < 1$. These definitions reduce to ordinary convexity and strict convexity when $m = 1$ (and $K = \mathbf{R}_+$).

Example. *Convexity with respect to componentwise inequality.* A function $f : \mathbf{R}^n \to \mathbf{R}^m$ is convex with respect to componentwise inequality (*i.e.*, the generalized inequality induced by $\mathbf{R}^m_+$) if and only if for all $x,\ y$ and $0 \leq \theta \leq 1$,

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

i.e., each component f_i is a convex function. f is strictly convex with respect to componentwise inequality if and only if each component f_i is strictly convex.

Example. *Convexity with respect to matrix inequality.* Suppose f is a symmetric matrix valued functions, *i.e.*, $f : \mathbf{R}^n \to \mathbf{R}^{m\times m}$, with $f(x) = f(x)^T$ for all x. f is convex with respect to matrix inequality if

$$f(\theta x + (1-\theta)y) \preceq \theta f(x) + (1-\theta)f(y)$$

for any x and y, and for $\theta \in [0,1]$. This is sometimes called *matrix-convexity.* An equivalent definition is that the scalar function $z^T f(x) z$ is convex for all vectors z. (This is often a good way to prove matrix-convexity). A matrix function is strictly matrix-convex if

$$f(\theta x + (1-\theta)y) \prec \theta f(x) + (1-\theta)f(y)$$

when $x \neq y$ and $0 < \theta t < 1$ or, equivalently, if $z^T f z$ is strictly convex for every $z \neq 0$.

Some examples:

- $f(X) = XX^T$ where $X \in \mathbf{R}^{n\times m}$ is convex, since for fixed z the function $z^T XX^T z = \|X^T z\|^2$ is a convex quadratic function of the components of X. For the same reason, $f(X) = X^2$ is convex on $\{X | X = X^T \in \mathbf{R}^{n\times n}\}$.
- The functions X^p is matrix convex on the set of positive definite matrices for $1 \leq p \leq 2$ or $-1 \leq p \leq 0$, and matrix concave for $0 \leq p \leq 1$.
- The function $f(X) = e^X$ is not convex on the space of symmetric matrices.

From the definition it is clear that K-convexity can be verified along a line in the same way as convexity in the usual sense: a function is K-convex if and only if its restriction to any line in its domain is K-convex.

Dual characterization of K-convexity

A function f is K-convex if and only if for every $w \succeq_{K^*} 0$, the (real-valued) function $w^T f$ is convex (in the ordinary sense); f is strictly K-convex if and only if for every nonzero $w \succeq_{K^*} 0$ the function $w^T f$ is strictly convex. (These follow directly from the definitions and properties of dual inequality.)

Differentiable K-convex functions

A differentiable function f is K-convex if and only if for all $x,\ y \in \mathbf{dom}\, f$,

$$f(y) \succeq_K f(x) + \frac{\partial f}{\partial x}(x)(y-x).$$

(Here $\frac{\partial f}{\partial x}(x) \in \mathbf{R}^{m\times n}$ is the first derivative or Jacobian matrix of f at x; see §A.3.2.)

The function f is strictly K-convex if for all $x,\ y \in \mathbf{dom}\, f$ with $x \neq y$,

$$f(y) \succ_K f(x) + \frac{\partial f}{\partial x}(x)(y-x).$$

(As in the scalar case, a function can be strictly K-convex without satisfying this inequality everywhere.)

A twice differentiable function f is K-convex if for all $x \in \mathbf{dom}\, f$ and all $y \in \mathbf{R}^n$,

$$\sum_{i,j=1,\ldots,n} \frac{\partial^2 f}{\partial x_i \partial x_j} y_i y_j \succeq_K 0,$$

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.3.2.)

Sublevel sets and epigraph

For $\alpha \in \mathbf{R}^m$, the α-sublevel set of $f : \mathbf{R}^n \to \mathbf{R}^m$, with respect to the generalized inequality induced by $K \subseteq \mathbf{R}^m$, is

$$C_\alpha = \{\ x \in \mathbf{R}^n \mid f(x) \preceq_K \alpha\ \}.$$

Sublevel sets of K-convex functions are convex sets.

The epigraph of f, with respect to the generalized inequality, is the set

$$\mathbf{epi}_K f = \{\ (x,t) \in \mathbf{R}^{n\times m} \mid f(x) \preceq_K t\ \}.$$

A function is K-convex if and only if $\mathbf{epi}_K f$ is a convex set.

Operations that preserve K-convexity

Some of the operations that preserve convexity have counterparts for K-convexity. For example, nonnegative weighted sums of K-convex functions are K-convex.

The pointwise maximum of two (or more) K-convex functions is K-convex, but the situation is far trickier here than in the scalar case. Recall that $a,\ b \in \mathbf{R}^m$ need not, in general, have a maximum with respect to K; in other words, there need not exist a $c \in \mathbf{R}^m$ such that $a \preceq_K c$, $b \preceq_K c$, and

$$a \preceq_K d, \quad b \preceq_K d \implies c \preceq_K d.$$

Therefore the pointwise maximum of f_1 and f_2, given by

$$f(x) = \max_K \{f_1(x), f_2(x)\},$$

is defined only if for each x the points $f_1(x)$ and $f_2(x)$ have a maximum.

Proof. Let's show that when the pointwise maximum of two K-convex functions f_1, f_2 does exist, it is K-convex. Assume $x_1,\ x_2 \in \mathbf{R}^n$, $0 \le \theta \le 1$, and let F denote the pointwise maximum of f_1 and f_2. Since f_1 and f_2 are K-convex,

$$f_1(\theta x_1 + (1-\theta)x_2) \preceq_K \theta f_1(x) + (1-\theta) f_1(x_2)$$
$$f_2(\theta x_1 + (1-\theta)x_2) \preceq_K \theta f_2(x_1) + (1-\theta) f_2(x_2).$$

Since $F(x) \succeq_K f_1(x)$ and $F(x) \succeq_K f_2(x)$, we have

$$f_1(\theta x_1 + (1-\theta)x_2) \preceq_K \theta F(x_1) + (1-\theta) F(x_2)$$
$$f_2(\theta x_1 + (1-\theta)x_2) \preceq_K \theta F(x_1) + (1-\theta) F(x_2),$$

i.e., $\theta F(x_1) + (1-\theta)F(x_2)$ is an upper bound on $f_1(\theta x_1 + (1-\theta)x_2)$ and $f_2(\theta x_1 + (1-\theta)x_2)$. By definition of F, we have

$$F(\theta x_1 + (1-\theta)x_2) \preceq_K \theta F(x_1) + (1-\theta) F(x_2).$$

There is also a generalized *composition* theorem. If h is L-convex and (K, L)-nondecreasing, and g is K-convex, Then $h(g(x))$ is L-convex.

Proof. The function g is K-convex. Therefore, $g(\theta x_1 + (1-\theta)x_2) \preceq_K \theta g(x_1) + (1-\theta)g(x_2)$. The function h is (K, L)-increasing, and L-convex. Therefore

$$\begin{aligned} h(g(\theta x_1 + (1-\theta)x_2)) &\preceq_L h(\theta g(x_1) + (1-\theta) g(x_2)) \\ &\preceq_L \theta h(g(x_1)) + (1-\theta) h(g(x_2)). \end{aligned}$$

Suppose $g(x, y)$ is K-convex, jointly in x and y, and C is a convex set. Suppose for each x, the set

$$\{g(x, y) | y \in C\}$$

has a minimum element (with respect to K), which we denote as $f(y)$. Then f is K-convex.

Chapter 3

Convex optimization problems

3.1 Optimization problems

3.1.1 Basic terminology

We use the notation

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array} \tag{3.1}$$

to describe the problem of finding an x that minimizes $f_0(x)$ among all x that satisfy the conditions $f_i(x) \leq 0$, $i = 1, \ldots, m$, and $h_i(x) = 0$, $i = 1, \ldots, p$. We call $x \in \mathbf{R}^n$ the *optimization variable* and the function $f_0 : \mathbf{R}^n \to \mathbf{R}$ the *objective function* or *cost function*. The inequalities $f_i(x) \leq 0$ are called *inequality constraints*, and the corresponding functions $f_i : \mathbf{R}^n \to \mathbf{R}$ are called the *inequality constraint functions*. The equations $h_i(x) = 0$ are called the *equality constraints*, and the functions $h_i : \mathbf{R}^n \to \mathbf{R}$ are the *equality constraint functions*. If there are no constraints we say the problem (3.1) is *unconstrained*.

The set of points for which the objective and all constraint functions are defined, *i.e.*,

$$D = \bigcap_{i=0}^{m} \mathbf{dom}\, f_i \ \cap \ \bigcap_{i=1}^{p} \mathbf{dom}\, h_i$$

is called the *domain* of the optimization problem (3.1). A point $x \in D$ is *feasible* if it satisfies the constraints $f_i(x) \leq 0$ and $h_i(x) = 0$. The problem (3.1) is said to be feasible if there exists at least one feasible point, and *infeasible* otherwise. The set of all feasible points is called the *feasible set* or the *constraint set*.

The *optimal value* (or minimal value) $f^\star$ of the problem (3.1) is defined as

$$f^\star = \inf \{ \, f_0(x) \mid f_i(x) \leq 0, \ i = 1, \ldots, m, \ h_i(x) = 0, \ i = 1, \ldots, p \, \}.$$

We allow $f^\star$ to take on the extended values $\pm\infty$. If the problem is infeasible, we have $f^\star = +\infty$ (using the standard convention that the infimum of the empty set is $+\infty$). If there are feasible points x_k with $f_0(x_k) \to -\infty$ as $k \to \infty$, then $f^\star = -\infty$, and we say the problem (3.1) is *unbounded below*.

We say $x^\star$ is an *optimal point*, or solves the problem (3.1), if $x^\star$ is feasible and $f(x^\star) = f^\star$. The set of all optimal points is the *optimal set*, denoted

$$X_{\rm opt} = \{ \; x \mid f_i(x) \le 0, \; i = 1, \ldots, m, \; h_i(x) = 0, \; i = 1, \ldots, p, \; f_0(x) = f^\star \; \}.$$

If there exists an optimal point for the problem (3.1), we say the optimal value is *attained* or *achieved*, and the problem is *solvable*. If $X_{\rm opt}$ is empty, we say the optimal value is not attained or not achieved. A feasible point x with $f_0(x) \le f^\star + \epsilon$ (where $\epsilon > 0$) is called *ϵ-suboptimal*, and the set of all ϵ-suboptimal points is called the *ϵ-suboptimal set* for the problem (3.1).

We say a feasible point x is *locally optimal* if there is an $R > 0$ such that

$$f(x) = \inf\{ \; f_0(z) \mid f_i(z) \le 0, \; i = 1, \ldots, m, \; h_i(z) = 0, \; i = 1, \ldots, p, \; \|z - x\| \le R \; \},$$

or, in other words, x solves the optimization problem

$$\begin{array}{ll} \mbox{minimize} & f_0(z) \\ \mbox{subject to} & f_i(z) \le 0, \quad i = 1, \ldots, m \\ & h_i(z) = 0, \quad i = 1, \ldots, p \\ & \|z - x\| \le R \end{array}$$

with variable z. Roughly speaking, this means x minimizes f_0 over nearby points in the feasible set. The term 'globally optimal' is sometimes used for 'optimal' to distinguish between 'locally optimal' and 'optimal'. Throughout this book, however, optimal will mean globally optimal.

If x is feasible and $f_i(x) = 0$, we say the constraint $f_i(x) \le 0$ is *active* at x. If $f_i(x) < 0$, we say the constraint $f_i(x) \le 0$ is *inactive*. (The equality constraints are active at all feasible points.) We say that a constraint is *redundant* if it is implied by the other constraints (so that deleting it does not change the constraint set).

If the objective function is identically zero, the optimal value is either zero (if the feasible set is nonempty) or $+\infty$ (if the feasible set is empty). The problem (3.1), therefore, is really to determine whether the problem is feasible or not. We call this the *feasibility problem*, and will sometimes write it as

$$\begin{array}{ll} \mbox{find} & x \\ \mbox{subject to} & f_i(x) \le 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p. \end{array} \tag{3.2}$$

The feasibility problem is thus to determine whether the inequalities are consistent, and if so, find a point that satisfies them.

Example. We illustrate these definitions with a few simple examples where $f_0 : \mathbf{R} \to \mathbf{R}$, with $\mathbf{dom}\, f_0 = \{ \; x \in \mathbf{R} \mid x > 0 \; \}$, and $m = p = 0$.

- $f_0(x) = 1/x$: $f^\star = 0$, but optimum is not achieved.
- $f_0(x) = -\log x$: $f^\star = -\infty$.
- $f_0(x) = x \log x$: $f^\star = -1/e$, achieved at the optimal point $x^\star = 1/e$.

3.1.2 Expressing problems in standard form

We refer to (3.1) as an optimization problem in *standard form*. In the standard form problem we choose the convention that the righthand side of the inequality and equality constraints are zero. This can always be arranged by subtracting any nonzero righthand side: we represent the equality constraint $g_i(x) = \tilde{g}_i(x)$, for example, as $h_i(x) = 0$, where $h_i(x) = g_i(x) - \tilde{g}_i(x)$. In a similar way we express inequalities of the form $f_i(x) \geq 0$ as $-f_i(x) \leq 0$.

Example. *Box constraints.* Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & l_i \leq x_i \leq u_i, \;\; i = 1, \ldots, m, \end{array}$$

where $x \in \mathbf{R}^n$ is the variable. The constraints are called *variable bounds* (since they give lower and upper bounds for each x_i) or box constraints (since the feasible set is a box).

We can express this problem in standard form as

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & l_i - x_i \leq 0, \;\; i = 1, \ldots, m, \\ & x_i - u_i \leq 0, \;\; i = 1, \ldots, m. \end{array}$$

There are $2m$ inequality constraint functions:

$$f_i(x) = l_i - x_i, \quad i = 1, \ldots, m,$$

and

$$f_i(x) = x_{n-i} - u_{n-i}, \quad i = n+1, \ldots, m.$$

Maximization problems

We concentrate on the minimization problem by convention. We can solve the *maximization* problem

$$\begin{array}{ll} \text{maximize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \;\; i = 1, \ldots, m \\ & h_i(x) = 0, \;\; i = 1, \ldots, p \end{array} \qquad (3.3)$$

by minimizing the function $-f_0$ subject to the constraints. By this correspondence we can define all the terms above for the maximization problem (3.3). For example the optimal value of (3.3) is defined as

$$f^\star = \sup \{ \, f_0(x) \mid f_i(x) \leq 0, \; i = 1, \ldots, m, \; h_i(x) = 0, \; i = 1, \ldots, p \, \},$$

and a feasible point x is ϵ-suboptimal if $f_0(x) \geq f^\star - \epsilon$. When the maximization problem is considered, the objective is sometimes called the *utility* or *satisfaction level* instead of the cost.

3.1.3 Equivalent problems

In this book we will use the notion of equivalence of optimization problems in an informal way. We call two problems *equivalent* if from a solution of one, a solution of the other is readily found, and vice versa. (It is possible, but quite complicated, to give a formal definition of equivalence.)

As a simple example, consider the problem

$$\begin{array}{ll} \text{minimize} & \tilde{f}(x) = \alpha_0 f_0(x) \\ \text{subject to} & \tilde{f}_i(x) = \alpha_i f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & \tilde{h}_i(x) = \beta_i h_i(x) = 0 \quad i = 1, \ldots, p, \end{array} \tag{3.4}$$

where $\alpha_i > 0$ and $\beta_i \neq 0$. This problem is obtained from the standard form problem (3.1) by scaling the objective and inequality constraint functions by positive constants, and the equality constraint functions by nonzero constants. The feasible set of the problem (3.4) and the original problem (3.1) are identical. A point x is optimal for the original problem (3.1) if and only if it is optimal for the scaled problem (3.4), so we say the two problems are equivalent. The two problems (3.1) and (3.4) are not, however, the same (unless α_i and β_i are all equal to one), since the objective function and the inequality and constraint functions differ.

We now describe some general transformations that yield equivalent problems.

Change of variables

Suppose $\phi : \mathbf{R}^n \to \mathbf{R}^n$ is one-to-one on the problem domain D, with image $\tilde{D} = f(D)$. We define functions $\tilde{f}_i$ and $\tilde{h}_i$ as

$$\tilde{f}_i(z) = f_i(\phi(z)), \ i = 1, \ldots, m, \quad \tilde{h}_i(z) = h_i(\phi(z)), \ i = 1, \ldots, p.$$

Now consider the problem

$$\begin{array}{ll} \text{minimize} & \tilde{f}_0(z) \\ \text{subject to} & \tilde{f}_i(z) \leq 0, \quad i = 1, \ldots, m \\ & \tilde{h}_i(z) = 0 \quad i = 1, \ldots, p, \end{array} \tag{3.5}$$

with variable z. We say that the standard form problem (3.1) and the problem (3.5) are related by the change or substitution of variable $x = \phi(z)$.

The two problems are clearly equivalent: if x solves the problem (3.1), then $z = \phi^{-1}(x)$ solves the problem (3.5); if z solves the problem (3.5), then $x = \phi(z)$ solves the problem (3.1).

Transformation of objective and constraint functions

Suppose that $\psi_0 : \mathbf{R} \to \mathbf{R}$ is monotonic increasing, $\psi_1, \ldots, \psi_m : \mathbf{R} \to \mathbf{R}$ satisfy $\psi_i(u) \leq 0$ if and only if $u \leq 0$, and $\psi_{m+1}, \ldots, \psi_{m+p} : \mathbf{R} \to \mathbf{R}$ satisfy $\psi_i(u) = 0$ if and only if $u = 0$. We define functions $\tilde{f}_i$ and $\tilde{h}_i$ as the compositions

$$\tilde{f}_i(x) = \psi_i(f_i(x)), \quad i = 1, \ldots, m, \quad \tilde{h}_i(x) = \psi_{m+i}(g_i(x)), \quad i = 1, \ldots, p.$$

Evidently the associated problem

$$\begin{array}{ll} \text{minimize} & \tilde{f}_0(x) \\ \text{subject to} & \tilde{f}_i(x) \leq 0, \quad i = 1, \ldots, m \\ & \tilde{h}_i(x) = 0 \quad i = 1, \ldots, p. \end{array} \tag{3.6}$$

and the standard form problem (3.1) are equivalent; indeed, the feasible sets are identical, and the optimal points are identical.

The example above, in which the objective and constraint functions are scaled by appropriate constants, is a special case.

Example. *Least-norm and least-norm-squared problems.* As a simple example consider the unconstrained Euclidean norm minimization problem

$$\text{minimize} \quad \|Ax - b\|, \tag{3.7}$$

with variable $x \in \mathbf{R}^n$. Since the norm is always nonnegative, we can just as well solve the problem

$$\text{minimize} \quad \|Ax - b\|^2 = (Ax - b)^T(Ax - b), \tag{3.8}$$

in which we minimize the square of the Euclidean norm. The problems (3.7) and (3.8) are clearly equivalent; the optimal points are the same. The two problems are not the same, however. For example, the objective in (3.7) is not differentiable at any x with $Ax - b = 0$, whereas the objective in (3.8) is differentiable (in fact, quadratic) for all x.

Slack variables

Consider the problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & s_i \geq 0, \quad i = 1, \ldots, m \\ & f_i(x) + s_i = 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0 \quad i = 1, \ldots, p, \end{array} \tag{3.9}$$

where the variables are $x \in \mathbf{R}^n$ and $s \in \mathbf{R}^m$. This problem has $n+m$ variables, m inequality constraints (the nonnegativity constraints on s_i), and $n+p$ equality constraints. The new variable s_i is called the *slack variable* associated with the original inequality constraint $f_i(x) \leq 0$. Introducing slack variables replaces each inequality constraint with an equality constraint, and a nonnegativity constraint.

The problem (3.9) is equivalent to the original standard form problem (3.1). Indeed, if (x, s) is feasible for the problem (3.9), then x is feasible for the original problem, since $s_i = -f_i(x) \geq 0$. Conversely, if x is feasible for the original problem, then (x, s) is feasible for the problem (3.9), where we take $s_i = -f_i(x)$. Similarly, x is optimal for the original problem (3.1) if and only (x, s) is optimal for the problem (3.9), where $s_i = -f_i(x)$.

Eliminating equality constraints

If we can explicitly parametrize all solutions of the equality constraints

$$h_i(x) = 0, \quad i = 1, \ldots, p, \tag{3.10}$$

using some parameter $z \in \mathbf{R}^k$, then we can *eliminate* the equality constraints from the problem, as follows. Suppose the function $\phi : \mathbf{R}^k \to \mathbf{R}^n$ is such that x satisfies (3.10) if and only there is some $z \in \mathbf{R}^k$ such that $x = \phi(z)$. The optimization problem

$$\begin{array}{ll} \text{minimize} & \tilde{f}_0(z) = f(\phi(z)) \\ \text{subject to} & \tilde{f}_i(z) = f_i(\phi(z)) \leq 0, \;\; i = 1, \ldots, m \end{array}$$

is then equivalent to the original problem (3.1). This transformed problem has variable $z \in \mathbf{R}^k$, m inequality constraints, and no equality constraints. If z is optimal for the transformed problem, then $x = \phi(z)$ is optimal for the original problem. Conversely, if x is optimal for the original problem, then (since x is feasible) there is at least one z such that $x = \phi(z)$. Any such z is optimal for the transformed problem.

The process of eliminating variables can be described more explicitly, and easily carried out numerically, when the equality constraints are all linear, *i.e.*, have the form $Ax = b$. If $Ax = b$ is inconsistent, *i.e.*, $b \notin \mathcal{R}(A)$ (where $\mathcal{R}(A)$ stands for the range of A), then the original problem is infeasible. Assuming this is not the case, let x_0 denote any solution of the equality constraints, *i.e.*, any point that satisfies $Ax_0 = b$. Let $F \in \mathbf{R}^{n \times k}$ be any matrix with $\mathcal{R}(F) = \mathcal{N}(A)$ ($\mathcal{N}(A)$ stands for the nullspace of A), so the general solution of the linear equations $Ax = b$ is given by $Fz + x_0$, where $z \in \mathbf{R}^k$. (We can choose F to be full rank, in which case we have $k = n - \mathbf{Rank}(A)$.)

Substituting $x = Fz + x_0$ into the original problem yields

$$\begin{array}{ll} \text{minimize} & f_0(Fz + x_0) \\ \text{subject to} & f_i(Fz + x_0) \leq 0, \;\; i = 1, \ldots, m, \end{array}$$

which is equivalent to the original problem, has no equality constraints, and $\mathbf{Rank}\, A$ fewer variables.

Introducing equality constraints

We can also *introduce* equality constraints and new variables into a problem. Instead of describing the general case, which is complicated and not very illuminating, we give a typical example that will be useful later. Consider the problem

$$\begin{array}{ll} \text{minimize} & f_0(A_0 x + b_0) \\ \text{subject to} & f_i(A_i x + b_i) \leq 0, \;\; i = 1, \ldots, m, \\ & h_i(x) = 0, \;\; i = 1, \ldots, p. \end{array}$$

In this problem the objective and constraint functions are given as compositions of the functions f_i with affine transformations defined by $A_i x + b_i$. We introduce new variables

$y_i = A_i x + b_i$, for $i = 0, \ldots, m$, and form the equivalent problem

$$\begin{array}{ll} \text{minimize} & f_0(y_0) \\ \text{subject to} & f_i(y_i) \leq 0, \quad i = 1, \ldots, m, \\ & y_i = A_i x + b_i, \quad i = 0, \ldots, m. \\ & h_i(x) = 0, \quad i = 1, \ldots, p, \end{array}$$

which has $m+1$ more variables, and $m+1$ more (linear) equality constraints. The inequality constraints in this problem are 'independent', *i.e.*, involve different optimization variables.

Optimizing over some variables

Recall that we always have

$$\inf_{x,y} f(x,y) = \inf_x \tilde{f}(x)$$

where $\tilde{f}(x) = \inf_y f(x,y)$. In other words, we can minimize a function by first minimizing over some of the variables, and then minimizing over the remaining ones. This simple and general principle can be used to transform problems into equivalent forms. The general case is cumbersome to describe and not illuminating, so we describe instead an example.

Suppose the variable $x \in \mathbf{R}^n$ is partitioned as $x = (x_1, x_2)$, with $x_1 \in \mathbf{R}^{n_1}$, $x_2 \in \mathbf{R}^{n_2}$, and $n_1 + n_2 = n$. We consider the problem

$$\begin{array}{ll} \text{minimize} & f_0(x_1, x_2) \\ \text{subject to} & f_i(x_1) \leq 0, \quad i = 1, \ldots, m_1, \\ & \tilde{f}_i(x_2) \leq 0, \quad i = 1, \ldots, m_2 \end{array} \tag{3.11}$$

in which the constraints are 'independent', *i.e.*, each constraint function depends on x_1 or x_2. We first minimize over x_2. Define the function $\tilde{f}_0$ of x_1 by

$$\tilde{f}_0(x_1) = \inf\{\ f_0(x_1, x_2) \mid \tilde{f}_i(x_2) \leq 0, \quad i = 1, \ldots, m_2\ \}.$$

The problem (3.11) is equivalent to

$$\begin{array}{ll} \text{minimize} & \tilde{f}_0(x_1) \\ \text{subject to} & f_i(x_1) \leq 0, \quad i = 1, \ldots, m_1. \end{array} \tag{3.12}$$

Example. *Minimizing a quadratic with constraints on some variables.* Consider a problem with strictly convex quadratic objective, with some of the variables unconstrained:

$$\begin{array}{ll} \text{minimize} & \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right]^T \left[\begin{array}{cc} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] \\ \text{subject to} & f_i(x_1) \leq 0, \quad i = 1, \ldots, m \end{array}$$

where $x = (x_1, x_2) \in \mathbf{R}^n$. Here we can analytically minimize over x_2:

$$\inf_{x_2} \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right]^T \left[\begin{array}{cc} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] = x_1^T \left(P_{11} - P_{12} P_{22}^{-1} P_{12}^T\right) x_1.$$

Therefore the original problem is equivalent to

$$\begin{array}{ll} \text{minimize} & x_1^T \left(P_{11} - P_{12} P_{22}^{-1} P_{12}^T\right) x_1 \\ \text{subject to} & f_i(x_1) \leq 0, \quad i = 1, \ldots, m. \end{array}$$

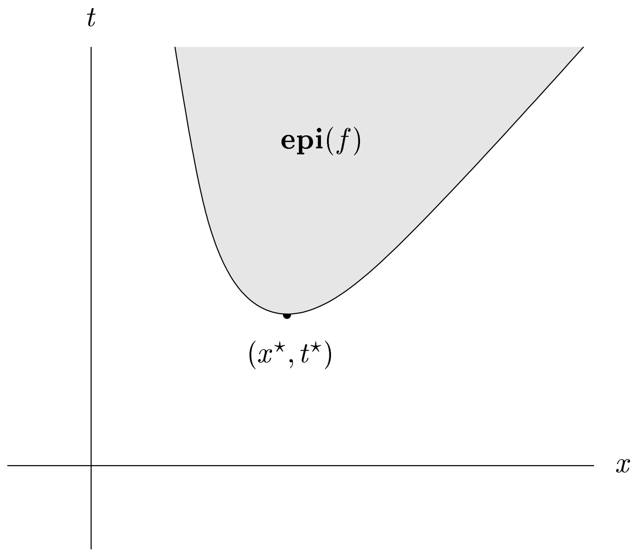

Figure 3.1: Geometrical interpretation of epigraph form problem, for a problem with no constraints. The problem is to find the point in the epigraph that minimizes t, *i.e.*, the 'lowest' point in the epigraph. The optimal point is $(x^\star, t^\star)$.

Epigraph problem form

The *epigraph form* of the standard problem (3.1) is the problem

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & f_0(x) - t \leq 0 \\ & f_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & h_i(x) = 0, \quad i = 1, \ldots, p, \end{array} \tag{3.13}$$

with variables $x \in \mathbf{R}^n$ and $t \in \mathbf{R}$. We can easily see that it is equivalent to the original problem: (x, t) is optimal for (3.13) if and only if x is optimal for (3.15) and $t = f_0(x)$. Note that the objective function of the epigraph form problem is *linear*.

The epigraph form problem (3.13) can be interpreted geometrically as an optimization problem in the 'graph space' (x, t): we minimize t over the epigraph of f_0, subject to the constraints on x.

Implicit and explicit constraints

By a simple trick already mentioned in the previous chapter, we can include any of the constraints *implicitly* in the objective function, by redefining its domain. As an extreme example, the standard form problem can be expressed as the 'unconstrained' problem

$$\text{minimize} \quad \tilde{f}_0(x) \tag{3.14}$$

where we define

$$\tilde{f}_0(x) = f_0(x), \quad \mathbf{dom}\, \tilde{f}_0 = \{\, x \in \mathbf{dom}\, f_0 \mid f_i(x) \leq 0,\ i = 1, \ldots, m,\ h_i(x) = 0,\ i = 1, \ldots, p \,\}.$$

(Equivalently, we can define $\tilde{f}_0(x)$ to have value ∞ for x not feasible.) The problems (3.1) and (3.14) are clearly equivalent: they have the same feasible sets, optimal points, and optimal values.

Of course this transformation is nothing more than a notational trick; making the constraints implicit has not made the problem any easier to analyze or solve, even though the problem (3.14) is, at least nominally, unconstrained. In some ways the transformation makes the problem more difficult. Suppose, for example, that the objective f_0 in the original problem is differentiable, so in particular its domain is open. The restricted objective function $\tilde{f}_0$ is probably not differentiable, since its domain is likely not to be open.

Conversely, we will encounter problems with implicit constraints, which we can then make explicit.

3.1.4 Parameter and oracle problem descriptions

For a problem in the standard form (3.1), there is still the question of how, exactly, the objective and constraint functions are specified. In the simplest case, these functions have some analytic form, or are given by a formula. Suppose, for example, the objective is quadratic, so it has the form $f_0(x) = x^T Px + q^T x + r$. To specify the objective function we give the coefficients (also called *problem parameters* or *problem data*) $P = P^T \in \mathbf{R}^{n\times n}$, $q \in \mathbf{R}^n$, and $r \in \mathbf{R}$.

In other cases the objective and constraint functions are described by *oracle* models (which are also called *black box* or *subroutine* models). In an oracle model, we do not know f explicitly, but can evaluate $f(x)$ (and usually also some derivatives) at any $x \in \mathbf{dom}\, f$. This is referred to as *querying the oracle*, and is usually associated with some cost, *e.g.*, time. We are also given some prior information about the function, such as convexity and a bound on its values. An algorithm is then judged, in part, on the number of times the oracle must be queried before an ϵ-suboptimal solution is found. As a concrete example of an oracle model, suppose f and its gradient ∇f are evaluated in a subroutine. We can call the subroutine at any $x \in \mathbf{dom}\, f$, but do not have access to its source code. Calling the subroutine with argument x yields (when the subroutine returns) $f(x)$ and $\nabla f(x)$. Note that in the oracle model, we never really know the function; we only know the function value (and some derivatives) at the points at which we have queried the oracle.

An oracle model is sometimes appropriate even when the functions have an analytic form, *e.g.*, affine or quadratic. As a common example suppose f_i, $i = 1, \ldots, m$, are affine and *sparse*, *i.e.*, each f_i depends only on a few of the x_j. Thus f_i, which have the form $f_i(x) = a_i^T x + b_i$ for some $a_i \in \mathbf{R}^n$ and $b_i \in \mathbf{R}$, can in principle be described by explicitly giving the coefficients of the vectors a_i and scalars b_i (most of which are zero). If m and n are very large, explicitly forming the vectors a_i may be impossible. On the other hand $f_i(x)$ are readily evaluated, given any x, by sparse matrix multiplication.

3.2 Convex optimization

3.2.1 Convex optimization problems in standard form

A *convex optimization problem* is one of the form

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & a_i^T x = b_i, \quad i = 1, \ldots, p, \end{array} \tag{3.15}$$

where $f_0, \ldots, f_m$ are convex functions. Comparing (3.15) with the general standard form problem (3.1), the convex problem has three additional requirements:

- the objective is convex
- the inequality constraint functions are convex
- the equality constraint functions $g_i(x) = a_i^T x - b_i$ are affine

We immediately note an important property: the feasible set of a convex optimization problem is convex, since it is the intersection of the domain of the problem

$$D = \cap_{i=0}^m \,\mathbf{dom}\, f_i,$$

which is a convex set, with m (convex) sublevel sets $\{x \mid f_i(x) \leq 0\}$ and p hyperplanes $\{x \mid a_i^T x = b_i\}$.

It is important to note a subtlety in this definition. Consider the example with $x \in \mathbf{R}^2$

$$\begin{array}{ll} \text{minimize} & f_0(x) = x_1^2 + x_2^2 \\ \text{subject to} & f_1(x) = x_1/(1 + x_2^2) \leq 0, \\ & h_1(x) = (x_1 + x_2)^2 = 0, \end{array}$$

which is in the standard form (3.1). This problem is *not* a convex optimization problem in standard form since the equality constraint function h_1 is not affine, and the inequality constraint function f_1 is not convex. Nevertheless the constraint set, which is $\{x | x_1 \leq 0,\ x_1 + x_2 = 0\}$, is convex. So although in this problem we are minimizing a convex function f_0 over a convex set, it is not a convex problem by our definition.

Of course, the problem is readily reformulated as

$$\begin{array}{ll} \text{minimize} & f_0(x) = x_1^2 + x_2^2 \\ \text{subject to} & \tilde{f}_1(x) = x_1 \leq 0, \\ & \tilde{g}_1(x) = x_1 + x_2 = 0, \end{array}$$

which is in standard convex optimization form, since $\tilde{f}_1$ is convex and $\tilde{g}_1$ is affine.

If f_0 is quasiconvex instead of convex, we say the problem (3.15) is a (standard form) *quasiconvex optimization problem.*

With a slight abuse of notation, we will also refer to

$$\begin{array}{ll} \text{maximize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & a_i^T x = b_i, \quad i = 1, \ldots, p, \end{array} \tag{3.16}$$

as a convex optimization problem if the objective function f_0 is concave (and the inequality constraint functions $f_1, \ldots, f_m$ are convex), since it is readily solved by minimizing the convex objective function $-f_0$. All of the results, conclusions, or algorithms that we describe for the minimization problem are easily transposed to the maximization case. In a similar way the maximization problem (3.16) is called *quasiconvex* if f_0 is quasiconcave.

Since the sublevel sets of a convex or quasiconvex function are convex, we conclude that for a convex or quasiconvex optimization problem the ϵ-suboptimal sets are convex. In particular, the optimal set is convex. If the objective is strictly convex, then the optimal set contains at most one point.

3.2.2 Local and global optima

A fundamental property of convex optimization problems is that any locally optimal point is also (globally) optimal. To see this, suppose that x is locally optimal for a convex optimization problem, *i.e.*, x is feasible and

$$f_0(x) = \inf\{ f_0(z) \mid z \text{ feasible}, \ \|z - x\| \leq R \}, \tag{3.17}$$

for some $R > 0$. Now suppose that x is *not* globally optimal, *i.e.*, there is a feasible y such that $f_0(y) < f_0(x)$. Evidently $\|y - x\| > R$, since otherwise $f_0(x) \leq f_0(y)$. Consider the point z given by

$$z = (1 - \theta)x + \theta y, \quad \theta = \frac{R}{2\|y - x\|}.$$

Then we have $\|z - x\| = R/2 < R$, and by convexity of the feasible set, z is feasible. By convexity of f_0 we have

$$f_0(z) \leq (1 - \theta) f_0(x) + \theta f_0(y) < f_0(x),$$

which contradicts (3.17). Hence there exists no feasible y with $f_0(y) < f_0(x)$, *i.e.*, x is globally optimal.

It is not true that locally optimal points of quasiconvex optimization problems are globally optimal. But a locally optimal point of a quasiconvex optimization problem with strictly quasiconvex objective is globally optimal.

3.2.3 An optimality criterion for differentiable f_0

Suppose that f_0 is differentiable (and convex), so that for all $x, y \in \mathbf{dom}\, f_0$,

$$f_0(y) \geq f_0(x) + \nabla f_0(x)^T (y - x) \tag{3.18}$$

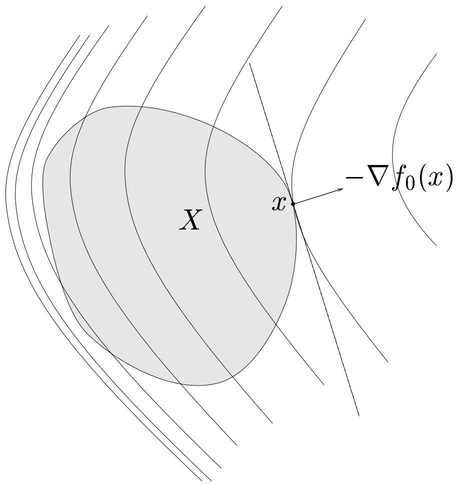

Figure 3.2: Geometrical interpretation of the optimality condition (3.19). The feasible set X is shown shaded. Some level curves of f_0 are shown. The point x is optimal: $-\nabla f_0(x)$ defines a supporting hyperplane (shown as dashed line) to X at x.

(see §2.1.3). Let X denote the feasible set, *i.e.*,

$$X = \{ \ x \mid f_i(x) \leq 0, \ i = 1, \dots, m, \ h(x) = 0, \ i = 1, \dots, p \ \}.$$

Then x is optimal if and only if $x \in X$ and

$$\nabla f_0(x)^T (y - x) \geq 0 \text{ for all } y \in X \tag{3.19}$$

This optimality criterion can be understood geometrically: it means that $-\nabla f_0(x)$ defines a supporting hyperplane to the feasible set at x (see figure 3.2).

Proof. First suppose $x \in X$ and satisfies (3.19). Then if $y \in X$ we have, by (3.18), $f_0(y) \geq f_0(x)$. This shows x is an optimal point for (3.1).

Conversely, suppose x is optimal, but the condition (3.19) does not hold, *i.e.*, for some $y \in X$ we have

$$\nabla f_0(x)^T (y - x) < 0.$$

Consider the point $z = \lambda y + (1 - \lambda)x$, where $0 \leq \lambda \leq 1$. By convexity of the feasible set, z is feasible. For small λ, we have

$$f_0(z) = f_0(x) + \lambda \nabla f_0(x)^T (y - x) + o(\lambda),$$

so for small, positive λ we have $f_0(z) < f_0(x)$, which contradicts optimality of x.

We will pursue the topic of optimality conditions in more depth in chapter 4, but here we examine a few simple examples.

Unconstrained problems

One obvious example is an unconstrained problem (*i.e.*, $m = p = 0$) where the condition is the well known necessary and sufficient condition

$$\nabla f_0(x) = 0 \tag{3.20}$$

for x to be optimal.

There are several possible situations. If there are no solutions of (3.20), then there are no optimal points; the optimal value of the problem is not attained. Here we can distinguish between two cases: the problem is unbounded below, or the optimal value is finite, but not attained. In either case, we can conclude that some sublevel set is not closed.

On the other hand we can have multiple solutions of the equation (3.20), in which case each such solution is a minimizer of f_0.

Example. *Unconstrained quadratic optimization.* Consider the problem of minimizing the quadratic function

$$f_0(x) = x^T Px + 2q^T x + r,$$

where $P = P^T \succeq 0$ (which makes f_0 convex). The necessary and sufficient condition for x to be optimal, *i.e.*, a minimizer of f_0, is

$$\nabla f_0(x) = 2Px + 2q = 0.$$

Several cases can occur, depending on whether this (linear) equation has no solutions, one solution, or many solutions.

- If $q \notin \mathcal{R}(P)$, then there is no solution. In this case f_0 is unbounded below.
- If $P \succ 0$ (which is the condition for f_0 to be strictly convex), then there is a unique minimizer, $x^{\rm opt} = -P^{-1}q$.
- If P is singular, but $q \in \mathcal{R}(P)$, then the set of optimal points is the (affine) set $x = -P^\dagger q + \mathcal{N}(P)$, where $P^\dagger$ denotes the pseudo-inverse of P.

Example. Consider the (unconstrained) problem of minimizing the function $f_0 : \mathbf{R}^n \to \mathbf{R}$, defined as

$$f_0(x) = \begin{cases} -\sum_{i=1}^m \log(b_i - a_i^T x) & \text{if } Ax \prec b,\ \ i = 1, \ldots, m \\ +\infty & \text{otherwise,} \end{cases}$$

where $A = [a_1 \ \ldots \ a_m]^T$ (and $\mathbf{dom}\, f_0 = \{x | Ax \prec b\}$).

f_0 is differentiable, so the necessary and sufficient conditions for x to be optimal are

$$Ax \prec b, \quad \nabla f_0(x) = -\sum_{i=1}^m \frac{1}{b_i - a_i^T x} a_i = 0. \tag{3.21}$$

If $Ax \prec b$ is infeasible, then the domain of f_0 is empty. Assuming $Ax \prec b$ is feasible, there are still several possible cases:

- There are no solutions of (3.21), and hence no optimal points for the problem. (This occurs if and only if f_0 is unbounded below).
- There are many solutions of (3.21). In this case it can be shown that the solutions form an affine set. This occurs if the feasible set is unbounded.
- There is a unique solution of (3.21), *i.e.*, a unique minimizer of f_0. This occurs if and only if the open polyhedron $\{x | Ax \prec b\}$ is nonempty and bounded.

Problems with equality constraints only

Consider the case where there are equality constraints but no inequality constraints, *i.e.*,

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & Ax = b. \end{array}$$

Here the feasible set is affine. We assume that it is nonempty; otherwise the problem is infeasible.

The optimality condition for a feasible x becomes

$$\nabla f_0(x)^T (y - x) \geq 0$$

for all y satisfying $Ay = b$. Since x is feasible, every feasible y has the form $y = x + v$ for some $v \in \mathcal{N}(A)$. The optimality condition can therefore be expressed as:

$$\nabla f_0(x)^T v \geq 0 \text{ for all } v \in \mathcal{N}(A).$$

If a linear function is nonnegative on a subspace, then it must be zero, so it follows that $\nabla f_0(x)^T v = 0$ for all $v \in \mathcal{N}(A)$. In other words,

$$\nabla f_0(x) \perp \mathcal{N}(A).$$

This optimality condition can also be expressed as $\nabla f_0 \in \mathcal{R}(A^T)$, *i.e.*, there exists a $\nu \in \mathbf{R}^p$ such that

$$\nabla f_0(x) + A^T \nu = 0.$$

This is the classical Lagrange multiplier optimality condition, which we will study in greater detail in chapter 4.

Minimization over the nonnegative orthant

As another example we consider the problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & x \succeq 0, \end{array}$$

where the only inequality constraints are nonnegativity constraints on the variables.

The optimality condition (3.19) is then

$$x \succeq 0, \quad \nabla f_0(x)^T (y - x) \geq 0 \text{ for all } y \succeq 0.$$

The term $\nabla f_0(x)^T y$, which is a linear function of y, is unbounded below on $y \succeq 0$, unless we have $\nabla f_0(x) \succeq 0$. The condition then reduces to $-\nabla f_0(x)^T x \geq 0$. But $x \succeq 0$ and $\nabla f_0(x) \succeq 0$, so we must have $\nabla f_0(x)^T x = 0$, which in turn implies each of its components is zero.

The optimality condition can therefore be expressed as

$$x \succeq 0, \quad \nabla f_0(x) \succeq 0, \quad x_i \left(\nabla f_0(x)\right)_i = 0.$$

3.2.4 Equivalent convex problems

It is useful to see which of the transformations described in §3.1.3 preserve convexity.

Eliminating equality constraints

For a convex problem the equality constraints must be linear, *i.e.*, of the form $Ax = b$. In this case they can be eliminated by finding a particular solution x_0 of $Ax = b$, and a matrix F whose range is the nullspace of A, which results in the problem

$$\begin{array}{ll} \text{minimize} & f_0(Fz + x_0) \\ \text{subject to} & f_i(Fz + x_0) \leq 0, \quad i = 1, \ldots, m, \end{array}$$

with variable z. Since the composition of a convex function with an affine function is convex, eliminating equality constraints preserves convexity of a problem. Moreover, the process of eliminating equality constraints (and reconstructing the solution of the original problem from the solution of the transformed problem) involves standard linear algebra operations.

At least in principle, this means we can restrict our attention to convex optimization problems which have no equality constraints. In some cases, however, it can be better to leave the equality constraints, since eliminating them can make the problem harder to understand and analyse, or ruin the efficiency of an algorithm that solves it. This is true, for example, when the variable x has very large dimension, and eliminating the equality constraints would destroy sparsity or some other useful structure of the problem.

Slack variables

By introducing slack variables we have the new constraints $f_i(x) + s_i = 0$. Since equality constraint functions must be affine in a convex problem, we must have f_i affine. In other words: introducing slack variables for *linear inequalities* preserves convexity of a problem.

Epigraph problem form

The epigraphic form of the convex optimization problem (3.15) is

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & f_0(x) - t \leq 0 \\ & f_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & a_i^T x = b_i, \quad i = 1, \ldots, p. \end{array}$$

The objective is linear, hence convex; and the new constraint function $f_0(x) - t$, is also convex in (x, t), so the epigraph form problem is convex as well.

It is sometimes said that a linear objective is universal for convex optimization, since any convex optimization problem is readily transformed to one with linear objective. The epigraphic form of a convex problem has several practical uses. By assuming the objective of a convex optimization problem is linear, we can simplify theoretical analysis. It can also simplify algorithm development, since an algorithm that solves convex optimization problems with linear objective can, using the transformation above, solve any convex optimization problem (provided it can handle the constraint $f_0(x) - t \le 0$).

Minimizing over some variables

Minimizing a convex function over some variables preserves convexity. Therefore, if f_0 in (3.11) is jointly convex in x_1 and x_2, and f_i, $i = 1, \ldots, m_1$ and $\tilde{f}_2$, $i = 1, \ldots, m_2$, are convex, then the equivalent problem (3.12) is convex.

3.3 Linear optimization problems

When the functions f_i and h_i in (3.15) are affine, the problem is called a *linear program* (LP). We will use the notation

$$\begin{array}{ll} \text{minimize} & c^T x + d \\ \text{subject to} & Gx \preceq h \\ & Ax = b, \end{array} \tag{3.22}$$

where $G \in \mathbf{R}^{m\times n}$ and $A \in \mathbf{R}^{p\times n}$. (Obviously we can omit the constant d in the objective function, *i.e.*, by omitting d we obtain an equivalent problem.) LPs are, of course, convex optimization problems. We can also maximize an affine objective $c^T x + d$, by minimizing $-c^T x - d$ (which is still convex).

The geometrical interpretation is illustrated in figure 3.3. The feasible set of (3.22) is a polyhedron $\mathcal{P}$; the problem is to minimize the affine function $c^T x + d$ (or, equivalently, the linear function $c^T x$) over $\mathcal{P}$.

Two special cases of the LP (3.22) are so widely encountered that they have been given separate names. In a *standard form LP* the only inequalities are componentwise nonnegativity constraints $x \succeq 0$:

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b \\ & x \succeq 0. \end{array} \tag{3.23}$$

If the LP has no equality constraints, it is called an *inequality form LP*, usually written as

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b. \end{array} \tag{3.24}$$

Remark. *Converting LPs to standard form.* It is sometimes useful to be able to transform a general LP (3.22) to one in standard form (3.23) (for example in order to

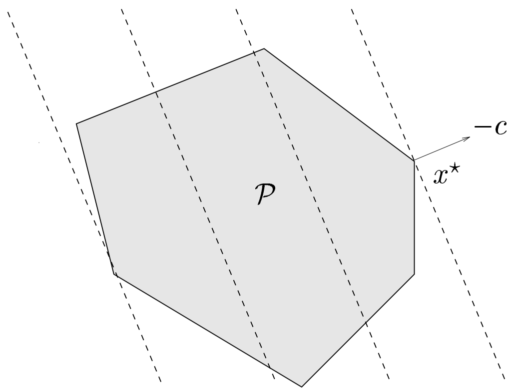

Figure 3.3: Geometrical interpretation of an LP. The feasible set $\mathcal{P}$, which is a polyhedron, is shaded. The objective $c^T x$ is linear, so its level curves are hyperplanes orthogonal to c (shown as dashed lines). The point $x^\star$ is optimal; it is the point in $\mathcal{P}$ as far as possible in the direction $-c$.

use an algorithm for standard form LPs). The first step is to introduce slack variables s_i for the inequalities, which results in

$$\begin{array}{ll} \text{minimize} & c^T x + d \\ \text{subjct to} & Gx + s = h \\ & Ax = b \\ & s \succeq 0. \end{array}$$

The second step is to express the variable x as the difference of two nonnegative variables x^+ and x^-, *i.e.*, $x = x^+ - x^-$, $x^+, x^- \succeq 0$. This yields the problem

$$\begin{array}{ll} \text{minimize} & c^T x^+ - c^T x^- + d \\ \text{subjct to} & Gx^+ - Gx^- + s = h \\ & Ax^+ - Ax^- = b \\ & x^+,\ x^-,\ s \succeq 0, \end{array}$$

which is an LP in standard form, with variables x^+, x^-, and s.

3.3.1 Examples

LPs arise in a vast number of fields and applications; here we give a handful of examples.

Diet problem

A healthy diet contains m different nutrients in quantities at least equal to $b_1, \ldots, b_m$. We can compose such a diet by choosing nonnegative quantities $x_1, \ldots, x_n$ or n different foods. One unit quantity of food j contains an amount a_{ij} of nutrient i, and has a cost of c_j. We want to determine the cheapest diet that satisfies the nutritional requirements. This problem can be formulated as the LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \succeq b \\ & x \succeq 0. \end{array}$$

Chebychev center

We consider the problem of finding the largest Euclidean ball inscribed in a polyhedron described by linear inequalities,

$$\mathcal{P} = \{x \in \mathbf{R}^n \mid a_i^T x \leq b_i, \;\; i = 1, \ldots, m\}.$$

(The center of the optimal ball is called the *Chebychev center* of the polyhedron; it is the point deepest inside the polyhedron, *i.e.*, farthest from the boundary.) We represent the ball as

$$\mathcal{B} = \{x_c + u \mid \|u\| \leq r\}.$$

The variables in the problem are the center $x_c \in \mathbf{R}^n$ and the radius r; we wish to maximize r subject to the constraint $\mathcal{B} \subseteq \mathcal{P}$.

We start by expressing the simpler constraint that $\mathcal{B}$ lies in one halfspace $a_i^T x \leq b_i$, *i.e.*,

$$\|u\| \leq r \Longrightarrow a_i^T (x_c + u) \leq b_i. \tag{3.25}$$

Since

$$\sup\{a_i^T u \mid \|u\| \leq r\} = r\|a_i\|$$

we can write (3.25) as

$$a_i^T x_c + r\|a_i\| \leq b_i, \tag{3.26}$$

which is a linear inequality in x_c and r. In other words, the constraint that the ball lies in the halfspace determined by the inequality $a_i^T x \leq b_i$ can be written as a linear inequality.

Therefore $\mathcal{B} \subseteq \mathcal{P}$ if and only if (3.26) holds for all $i = 1, \ldots, m$. Hence the Chebychev center can be determined by solving the LP

$$\begin{array}{ll} \text{maximize} & r \\ \text{subject to} & a_i^T x_c + r\|a_i\| \leq b_i, \;\; i = 1, \ldots, m, \end{array}$$

in the variables r and x_c.

Dynamic activity planning

We consider the problem of choosing, or planning, the activity levels of n activities, or sectors of an economy, over N time periods. We let $x_j(t) \geq 0$, $t = 1, \ldots, N$ denote the activity level of sector j, in period t. The activities both consume and produce products or goods in proportion to their activity levels. The amount of good i produced per unit of activity j, is given by a_{ij}. Similarly, the amount of good i consumed per unit of activity j is b_{ij}. The total amount of goods produced is in period t is given by $Ax(t) \in \mathbf{R}^m$, and the amount of goods consumed is $Bx(t) \in \mathbf{R}^m$, (Although we refer to these products as 'goods', they can also include unwanted products such as pollutants.)

The goods consumed in a period cannot exceed those produced in the previous period, *i.e.*, we must have $Bx(t+1) \preceq Ax(t)$ for $t = 1, \ldots, N$. A vector $g_0 \in \mathbf{R}^m$ of initial goods is given, which constrains the first period acitivity levels: $Bx(1) \preceq g_0$.

The (vectors of) excess goods not consumed by the activities are given by

$$\begin{aligned} s(0) &= g_0 - Bx(1), \\ s(t) &= Ax(t) - Bx(t+1), \quad t = 1, \ldots, N-1 \\ s(N) &= Ax(N). \end{aligned}$$

The objective is to maximize a discounted total value of excess goods:

$$c^T s(0) + \gamma c^T s(1) + \cdots + \gamma^N c^T s(N),$$

where $c \in \mathbf{R}^m$ gives the values of the goods, and $\gamma > 0$ is a discount factor. (The value c_i is negative if the ith product is unwanted, *i.e.*, a pollutant; $|c_i|$ is then the cost of disposal per unit.)

Putting it all together we arrive at the LP

$$\begin{array}{ll} \text{minimize} & c^T s(0) + \gamma c^T s(1) + \cdots + \gamma^T c^T s(N) \\ \text{subject to} & x(t) \succeq 0, \ t = 1, \ldots, N \\ & s(t) \succeq 0, \ t = 0, \ldots, N \\ & s(0) = g_0 - Bx(1) \\ & s(t) = Ax(t) - Bx(t+1), \quad t = 1, \ldots, N-1 \\ & s(N) = Ax(N) \end{array}$$

with variables $x(1), \ldots, x(N), s(0), \ldots, s(N)$. This problem is in LP standard form; the variables $s(t)$ are the slack variables associated with the constraints $Bx(t+1) \preceq Ax(t)$.

Chebychev inequalities

We consider a probability distribution for a random variable x on a set $\{u_1, \ldots, u_n\} \subseteq \mathbf{R}$ with n elements. We describe the distribution of x by a vector $p \in \mathbf{R}^n$, with

$$p_i = \mathbf{Prob}(x = u_i),$$

so p satisfies $p \succeq 0$ and $\mathbf{1}^T p = 1$. Conversely, if p satisfies $p \succeq 0$ and $\mathbf{1}^T p = 1$, then it defines a possible distribution for x. We assume that u_i are known and fixed, but the distribution, p, is not known.

If f is any function of x, then

$$\mathbf{E} f = \sum_{i=1}^n p_i f(u_i)$$

is a linear function of p. If S is any subset of $\mathbf{R}$, then

$$\mathbf{Prob}(x \in S) = \sum_{u_i \in S} p_i$$

is a linear function of p.

Although we do not know p, we are given prior knowledge of the following form. We know upper and lower bounds on expected values of some functions of x, and probabilities of some subsets of $\mathbf{R}$. This prior knowledge can be expressed as linear inequality constraints on p,

$$l_i \le a_i^T p \le u_i, \quad i = 1, \ldots, m.$$

The problem is to give lower and upper bounds on $\mathbf{E}\, f_0(x) = a_0^T p$, where f_0 is some function of x.

To find a lower bound we solve the LP

$$\begin{array}{ll} \text{minimize} & a_0^T p \\ \text{subject to} & p \succeq 0, \quad \mathbf{1}^T p = 1, \\ & l_i \le a_i^T p \le u_i, \quad i = 1, \ldots, m. \end{array}$$

The optimal value of this LP gives the lowest possible value of $\mathbf{E}\, f_0(X)$ for any distribution that is consistent with the prior information. Moreover, the bound is sharp: the optimal solution gives a distribution that is consistent with the prior information and achieves the lower bound.

Similarly, we can find the best upper bound by maximizing $a_0^T p$ subject to the same constraints.

Piecewise linear minimization

Consider the (unconstrained) problem of minimizing the piecewise linear, convex function

$$f(x) = \max_{i=1,\ldots,m} (a_i^T x + b_i).$$

This problem can be transformed to an equivalent LP by first forming the epigraph problem,

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \max_{i=1,\ldots,m} (a_i^T x + b_i) \le t, \end{array}$$

and then expressing the inequality as a set of m separate inequalities:

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & a_i^T x + b_i \le t, \quad i = 1, \ldots, m. \end{array}$$

This is an LP (in inequality form).

3.3.2 Linear fractional programming

The problem of minimizing a ratio of affine functions over a polyhedron is called a *linear fractional program*:

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & Gx \preceq h \\ & Ax = b \end{array} \tag{3.27}$$

where

$$f_0(x) = \begin{cases} \dfrac{c^Tx+d}{e^Tx+f} & \text{if } e^Tx+f>0 \\ +\infty & \text{otherwise} \end{cases}$$

The objective function is quasiconvex (in fact, quasilinear) on the halfspace where $e^Tx+f>0$, so linear fractional programs are quasiconvex optimization problems.

In fact, if the feasible set

$$\{x \mid Gx \preceq h,\ Ax=b,\ e^Tx+f>0\}$$

is bounded, the linear fractional program (3.27) can be transformed to an equivalent linear program

$$\begin{array}{ll} \text{minimize} & c^Ty+dz \\ \text{subject to} & Gy-hz \preceq 0 \\ & Ay-bz=0 \\ & e^Ty+fz=1 \\ & z \geq 0 \end{array} \tag{3.28}$$

with variables y, z.

To show the equivalence, we first note that if (y,z) is feasible in (3.28), then $z \neq 0$. If there were a feasible $(y,0)$, we would have $Gy \preceq 0$, $Ay=0$ and $e^Ty=1$. This would contradict the assumption that the feasible set of (3.27) is bounded, since for any x feasible in (3.27) and any $t \geq 0$, $x+ty$ would also be feasible.

With that in mind, it is easy to show that both problems are equivalent. If x is feasible in (3.27) then

$$y = \frac{x}{e^Tx+f}, \quad z = \frac{1}{e^Tx+f}$$

is feasible in (3.28), with the same objective value $c^Ty+dz=f_0(x)$. Conversely, if (y,z) is feasible in (3.28), then $x=y/z$ is feasible in (3.27), with the same objective value.

Generalized linear fractional programming

A generalization of the linear fractional program (3.27) is the *generalized linear fractional program*

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & Cx+d \preceq t(Ex+f) \\ & Ex+f \succeq 0 \\ & Ax=b. \end{array}$$

Example. *Von Neumann growth problem.* We consider again an economy with n sectors, and activity levels $x_i(t) \geq 0$ at period $t=1,\ldots$. There are m goods which are consumed, and also produced, by the activity: An activity level x consumes goods $Bx \in \mathbf{R}^m$, and produces goods Ax. The goods consumed in a period cannot exceed the goods produced in the last period, *i.e.*, $Bx(t+1) \preceq Ax(t)$. The *growth rate* in sector i is given by $x_i(t+1)/x_i(t)$.

Von Neumann's growth problem is to find an activity level vector $x(t)$ that maximizes the minimum growth rate across all sectors. Denoting $x(t)$ as x and $x(t+1)$ as y, the problem can be expressed as

$$\begin{array}{ll} \text{minimize} & \min_{i=1}^n y_i/x_i \\ \text{subject to} & x \succeq 0, \;\; y \succeq 0 \\ & By \preceq Ax. \end{array}$$

This problem is quasiconvex, since the objective is quasiconcave. It can be expressed as the generalized linear fractional problem

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & x \preceq ty \\ & x \succeq 0, \;\; y \succeq 0 \\ & By \preceq Ax. \end{array}$$

(The optimal value gives the inverse of the optimal growth rate.)

3.4 Quadratic optimization problems

The convex optimization problem (3.15) is called a *quadratic program* (QP) if the objective function is (convex) quadratic, and the constraint functions are affine. A quadratic program can be expressed in the form

$$\begin{array}{ll} \text{minimize} & x^T P x + 2q^T x + r \\ \text{subject to} & Gx \preceq h \\ & Ax = b \end{array} \tag{3.29}$$

where $P = P^T \succeq 0$. In a quadratic program, we minimize a convex quadratic function over a polyhderon, as illustrated in figure 3.4.

If the objective in (3.15) as well as the inequality constraint functions are (convex) quadratic, as in

$$\begin{array}{ll} \text{minimize} & x^T P_0 x + 2q_0^T x + r_0 \\ \text{subject to} & x^T P_i x + 2q_i^T x + r_i \leq 0, \quad i = 1, \ldots, m \\ & Ax = b \end{array} \tag{3.30}$$

(with $P_i = P_i^T \succeq 0$ for $i = 0, 1 \ldots, m$) the problem is called a *quadratically constrained quadratic program* (QCQP). In a QCQP, we minimize a convex quadratic function over a feasible region that is the intersection of (possibly degenerate) ellipsoids.

3.4.1 Examples

Least-squares and regression

The problem of minimizing the convex quadratic function

$$\|Ax - b\|^2 = x^T A^T A x - 2b^T A x + b^T b$$

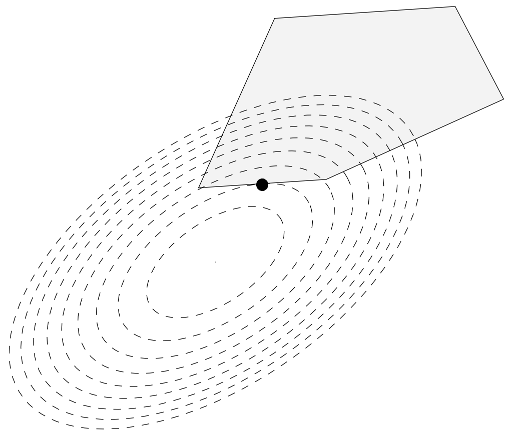

Figure 3.4: Geometric illustration of QP. The feasible set, which is a polyhedron, is shown shaded. The contour lines of the objective function, which is convex quadratic, are shown as dashed curves. The point shown in black is optimal.

is an (unconstrained) QP. It arises in many fields and has many names, *e.g.*, regression analysis or least-squares approximation. This problem, of course, is simple enough to have an analytical solution $x = A^\dagger b$, so calling it a QP seems a bit pedantic.

When linear inequality constraints are added, the problem is called *constrained regression* or *constrained least-squares*, and there is no longer a simple analytical solution. As a simple example we can consider regression with bounds on the variables, *i.e.*,

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|^2 \\ \text{subject to} & l_i \leq x_i \leq u_i, \quad i = 1, \ldots, n, \end{array}$$

which is a QP.

Distance between polyhedra

The distance between two polyhedra $\mathcal{P}_1 = \{x | Ax \preceq b\}$ and $\mathcal{P}_2 = \{x | Fx \preceq g\}$ is defined as

$$\mathrm{dist}(\mathcal{P}_1, \mathcal{P}_2) = \inf \{ \; \|x_1 - x_2\| \mid x_1 \in \mathcal{P}_1, \; x_2 \in \mathcal{P}_2 \; \}.$$

If the polyhedra intersect, the distance is zero.

To find the distance between $\mathcal{P}_1$ and $\mathcal{P}_2$, we can solve the QP

$$\begin{array}{ll} \text{minimize} & \|x_1 - x_2\|^2 \\ \text{subject to} & Ax_1 \preceq b \\ & Fx_2 \preceq g. \end{array}$$

This problem is infeasible if and only if one of the polyhedra is empty. The optimal value is zero if and only if the polyhedra intersect, in which case the optimal x_1 and x_2 are equal (and is a point in the intersection $\mathcal{P}_1 \cap \mathcal{P}_2$). Otherwise the optimal x_1 and x_2 are the points in $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively, that are closest to each other. (This allows us to construct a strictly separating hyperplane between $\mathcal{P}_1$ and $\mathcal{P}_2$; see §1.5.1).

Bounding variance

We consider again the situation described above in the section on Chebychev inequalities. The variance of a random variable $f(x)$ is given by

$$\mathbf{E}\, f^2 - (\mathbf{E}\, f)^2 = \sum_{i=1}^n f_i^2 p_i - \left(\sum_{i=1}^n f_i p_i\right)^2,$$

(where $f_i = f(u_i)$), is a concave quadratic function of p.

It follows that we can determine the best lower bound on the variance of $f(x)$, subject to the given prior information, by solving the QP

$$\begin{array}{ll} \text{minimize} & -\sum_{i=1}^n f_i^2 p_i + \left(\sum_{i=1}^n f_i p_i\right)^2 \\ \text{subject to} & p \succeq 0, \quad \mathbf{1}^T p = 1, \\ & l_i \leq a_i^T p \leq u_i, \quad i = 1, \ldots, m. \end{array}$$

LP with random cost

Consider the LP (3.22). We suppose that the cost function (vector) c is *random*, with mean value c_0 and covariance $\mathbf{E}(c - c_0)(c - c_0)^T = \Sigma$. (We assume for simplicity that the other problem parameters are deterministic.)

For a given $x \in \mathbf{R}^n$, the cost $c^T x$ is a (scalar) random variable with mean $\mathbf{E}\, c^T x = c_0^T x$ and variance

$$\mathbf{var}(c^T x) = \mathbf{E}\left(c^T x - \mathbf{E}\, c^T x\right)^2 = x^T \Sigma x.$$

In general there is a tradeoff between small expected cost and small cost variance. One way to take variance into account is to minimize a linear combination of the expected value and the variance of the cost, *i.e.*,

$$\mathbf{E}\, c^T x + \gamma\, \mathbf{var}(c^T x),$$

which is called the 'risk-sensitive cost'. The parameter $\gamma \geq 0$ is called the *risk-aversion parameter*, since it sets the relative values of cost variance and expected value. (For $\gamma > 0$, we are willing to tradeoff an increase in expected cost for a sufficiently large decrease in cost variance).

To minimize the risk-sensitive cost we solve the QP

$$\begin{array}{ll} \text{minimize} & c_0^T x + \gamma x^T \Sigma x \\ \text{subject to} & Gx \preceq h \\ & Ax \preceq b. \end{array}$$

3.4.2 Second-order cone programming

A problem that is closely-related to quadratic programming (and sometimes also called a quadratic program) is the *second-order cone program* (SOCP):

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \|A_i x + b_i\| \leq c_i^T x + d_i \quad i = 1, \ldots, N \\ & Fx = g, \end{array} \tag{3.31}$$

where $x \in \mathbf{R}^n$ is the optimization variable, and the norm appearing in the constraints is the Euclidean norm. We call a constraint of the form

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

where $A \in \mathbf{R}^{k \times m}$, a *second-order cone constraint*, since it is the same as requiring the affine function $(Ax + b, c^T x + d)$ to lie in the second-order cone in $\mathbf{R}^{k+1}$.

If $c_i = 0$, the SOCP (3.31) is equivalent to a QCQP, but the SOCP is more general. Similarly, if A_i and b_i are zero, the SOCP (3.31) reduces to a (general) LP.

Robust linear programming

We consider a linear program,

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & a_i^T x \leq b_i, \quad i = 1, \ldots, m, \end{array}$$

in which there is some uncertainty or variation in the parameters c, a_i, b_i. To simplify the exposition we will assume that c and b_i are fixed, and that a_i are known to lie in given ellipsoids:

$$a_i \in \mathcal{E}_i = \{\overline{a}_i + P_i u \mid \|u\| \leq 1\},$$

where $P_i = P_i^T \succeq 0$. (If P_i is singular we obtain 'flat' ellipsoids, of dimension $\mathbf{Rank}(P_i)$; $P_i = 0$ means that a_i is known perfectly.)

In a worst-case framework, we require that the constraints be satisfied for all possible values of the parameters a_i, which leads us to the *robust linear program*

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & a_i^T x \leq b_i, \text{ for all } a_i \in \mathcal{E}_i, \quad i = 1, \ldots, m. \end{array} \tag{3.32}$$

The robust linear constraint $a_i^T x \leq b_i$ for all $a_i \in \mathcal{E}_i$ can be expressed as

$$\sup\{\, a_i^T x \mid a_i \in \mathcal{E}_i \,\} = \overline{a}_i^T x + \|P_i x\| \leq b_i,$$

which is evidently a second-order cone constraint. Hence the robust LP (3.32) can be expressed as the SOCP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & a_i^T x + \|P_i x\| \leq b_i, \quad i = 1, \ldots, m. \end{array}$$

Note that the additional norm terms act as 'regularization terms', discouraging large x in directions with considerable uncertainty in the parameters a_i. Note that conversely, we can interpret a general SOCP with $b_i = 0$ as a robust LP.

Linear programming with random constraints

The robust LP described above can also be considered in a statistical framework. Here we suppose that the parameters a_i are independent Gaussian random vectors, with mean $\overline{a}_i$ and covariance Σ_i. We require that each constraint $a_i^T x \leq b_i$ should hold with a probability (confidence) exceeding η, where $\eta \geq 0.5$, *i.e.*,

$$\mathrm{Prob}(a_i^T x \leq b_i) \geq \eta. \tag{3.33}$$

We will show that this probability constraint can be expressed as an second-order cone constraint.

Letting $u = a_i^T x$, with σ denoting its variance, this constraint can be written as

$$\mathrm{Prob}\left(\frac{u - \overline{u}}{\sqrt{\sigma}} \leq \frac{b_i - \overline{u}}{\sqrt{\sigma}}\right) \geq \eta.$$

Since $(u - \overline{u})/\sqrt{\sigma}$ is a zero mean unit variance Gaussian variable, the probability above is simply $\Phi((b_i - \overline{u})/\sqrt{\sigma})$, where

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$$

is the CDF of a zero mean unit variance Gaussian random variable. Thus the probability constraint (3.33) can be expressed as

$$\frac{b_i - \overline{u}}{\sqrt{\sigma}} \geq \Phi^{-1}(\eta),$$

or, equivalently,

$$\overline{u} + \Phi^{-1}(\eta)\sqrt{\sigma} \leq b_i.$$

From $\overline{u} = \overline{a}_i^T x$ and $\sigma = x^T \Sigma_i x$ we obtain

$$\overline{a}_i^T x + \Phi^{-1}(\eta) \|\Sigma_i^{1/2} x\| \leq b_i.$$

Now, provided $\eta \geq 1/2$ (*i.e.*, $\Phi^{-1}(\eta) \geq 0$), this constraint is a second-order cone constraint.

In summary, the problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \mathrm{Prob}\,(a_i^T x \leq b_i) \geq \eta, \quad i = 1, \ldots, m \end{array}$$

can be expressed as the SOCP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \overline{a}_i^T x + \Phi^{-1}(\eta) \|\Sigma_i^{1/2} x\| \leq b_i, \quad i = 1, \ldots, m. \end{array}$$

Example. *Portfolio optimization with loss risk constraints.* We consider a classical portfolio problem with n assets or stocks held over one period. We let x_i denote the amount of asset i held at the beginning of (and throughout) the period, and p_i will denote the price change of asset i over the period, so the return is $r = p^T x$. The optimization variable is the portfolio vector $x \in \mathbf{R}^n$. The simplest assumptions are $x_i \geq 0$ (*i.e.*, no short positions) and $x_1 + \cdots + x_n = 1$ (*i.e.*, unit total budget).

We take a simple stochastic model for price changes: $p \in \mathbf{R}^n$ is Gaussian, with known mean $\overline{p}$ and covariance Σ. Therefore with portfolio $x \in \mathbf{R}^n$, the return r is a (scalar) Gaussian random variable with mean $\overline{r} = \overline{p}^T x$ and variance $\sigma_r = x^T \Sigma x$. The choice of portfolio x involves the (classical, Markowitz) tradeoff between return mean and variance.

Using SOCP, we can directly handle constraints that limit the risk of various levels of loss. Consider a loss risk constraint of the form

$$\mathbf{Prob}(r \leq \alpha) \leq \beta, \tag{3.34}$$

where α is a given unwanted return level (*e.g.*, an excessive loss) and β is a given maximum probability. As in the stochastic interpretation of the robust LP of §3.4.2, we can express this constraint using the CDF Φ of a unit Gaussian random variable. The inequality (3.34) is equivalent to

$$\overline{p}^T x + \Phi^{-1}(\beta) \left\| \Sigma^{1/2} x \right\| \geq \alpha. \tag{3.35}$$

Provided $\beta \leq 1/2$ (*i.e.*, $\Phi^{-1}(\beta) \leq 0$), this loss risk constraint is a second-order cone constraint. (If $\beta > 1/2$, the loss risk constraint becomes concave in x.)

The problem of maximizing the expected return subject to a bound on the loss risk (with $\beta \leq 1/2$), can therefore be cast as a simple SOCP with one second-order cone constraint:

$$\begin{array}{ll} \text{maximize} & \overline{p}^T x \\ \text{subject to} & \overline{p}^T x + \Phi^{-1}(\beta) \left\| \Sigma^{1/2} x \right\| \geq \alpha \\ & x \geq 0, \;\; \sum_{i=1}^n x_i = 1. \end{array}$$

(The solution of this problem coincides with the risk-averse solution described above, for some value of the risk-aversion parameter γ.)

There are many extensions of this simple problem. For example, we can impose several loss risk constraints, *i.e.*,

$$\mathrm{Prob}(r \leq \alpha_i) \leq \beta_i, \quad i = 1, \ldots, k,$$

(where $\beta_i \leq 1/2$), which expresses the risks (β_i) we are willing to accept for various levels of loss (α_i).

As another (standard) extension, we can allow short positions, *i.e.*, $x_i < 0$. To do this we introduce variables x_{long} and x_{short}, with

$$x_{\mathrm{long}} \succeq 0, \quad x_{\mathrm{short}} \succeq 0, \quad x = x_{\mathrm{long}} - x_{\mathrm{short}}, \quad \mathbf{1}^T x_{\mathrm{short}} \leq \eta \mathbf{1}^T x_{\mathrm{long}}.$$

(The last constraint limits the total short position to some fraction η of the total long position.)

3.5 Geometric programming

3.5.1 Monomials and posynomials

A function $f : \mathbf{R}^n \to \mathbf{R}$ with $\mathbf{dom}\, f = \{x | x \succ 0\}$, defined as

$$f(x) = c x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}, \tag{3.36}$$

where $c \geq 0$ and $a_i \in \mathbf{R}$, is called a *monomial function*. Note that exponents can be any real numbers, including fractional or negative. (Therefore the term 'monomial' conflicts with the standard definition from algebra, in which the coefficients must be nonnegative integers, but this will not cause any confusion.) A sum of monomials, *i.e.*, a function of the form

$$f(x) = \sum_{k=1}^{K} c_k x_1^{a_{1k}} x_2^{a_{2k}} \cdots x_n^{a_{nk}}, \tag{3.37}$$

where $c_j \geq 0$, is called a *posynomial function* (with K terms).

Note that posynomials are closed under addition, multiplication, and nonnegative scaling. Monomials are closed under multiplication and division.

3.5.2 Geometric programming

An optimization problem of the form

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 1, \quad i = 1, \ldots, m, \\ & h_i(x) = 1, \quad i = 1, \ldots, p \end{array} \tag{3.38}$$

where f_i are posynomial functions and h_i are monomials, is called a *geometric program* (GP). The domain of this problem is $D = \{x \mid x \succ 0\}$; the constraint $x \succ 0$ is implicit.

Several extensions are readily handled. If f is a posynomial and g is a monomial, then the constraint $f(x) \leq g(x)$ can be handled by expressing it as $f(x)/g(x) \leq 1$ (since f/g is posynomial). For example, we can handle constraints of the form $f(x) \leq a$, where f is posynomial and $a > 0$. In a similar way if g_1 and g_2 are both monomial functions, then we can handle the equality constraint $g_1(x) = g_2(x)$ by expressing it as $g_1(x)/g_2(x) = 1$ (since g_1/g_2 is monomial). We can maximize a monomial, by minimizing its inverse (which is also a monomial).

3.5.3 Geometric program in convex form

Geometric programs are not (in general) convex optimization problems, but they can be transformed to convex problems by a change of variables and a transformation of the objective and constraint functions.

We will use the variables defined as $y_i = \log x_i$, so $x_i = e^{y_i}$. If f is the monomial function of x given in (3.36), then

$$\tilde{f}(y) = f(x)$$

$$
\begin{aligned}
&= f(e^{y_1}, \ldots, e^{y_n}) \\
&= c\,(e^{y_1})^{a_1} \cdots (e^{y_n})^{a_n} \\
&= e^{a^T y + b},
\end{aligned}
$$

where $b = \log c$. The change of variables $y_i = \log x_i$ turns a monomial function into the exponential of an affine function.

Similarly, if f is the posynomial given by (3.37), then the associated function of y is given by

$$
\tilde{f}(y) = f(x) = \sum_{k=1}^{K} e^{a_k^T y + b_k},
$$

where $a_k = (a_{1k}, \ldots, a_{nk})$ and $b_k = \log c_k$. After the change of variables, a posynomial becomes a sum of exponentials of affine functions.

The geometric program (3.38) can be expressed in terms of the new variable y as

$$
\begin{array}{ll}
\text{minimize} & \sum_{k=1}^{K_0} e^{a_{0k}^T y + b_{0k}} \\
\text{subject to} & \sum_{k=1}^{K_i} e^{a_{ik}^T y + b_{ik}} \le 1, \quad i = 1, \ldots, m \\
& e^{g_i^T y + h_i} = 1, \quad i = 1, \ldots, p,
\end{array}
$$

where $a_{ik} \in \mathbf{R}^n$, $i = 0, \ldots, m$, contain the exponents in the posynomials of the original geometric program.

Now we transform the objective and constraint functions, by taking the logarithm. This results in the problem

$$
\begin{array}{ll}
\text{minimize} & \tilde{f}_0(y) = \log \sum_{k=1}^{K_0} e^{a_{0k}^T y + b_{0k}} \\
\text{subject to} & \tilde{f}_i(y) = \log \sum_{k=1}^{K_i} e^{a_{ik}^T y + b_{ik}} \le 0, \quad i = 1, \ldots, m \\
& \tilde{h}_i(y) = g_i^T y + h_i = 0, \quad i = 1, \ldots, p.
\end{array}
\tag{3.39}
$$

Since the functions $\tilde{f}_i$ are convex, and $\tilde{h}_i$ are affine, this problem is a convex optimization problem. We refer to it as a *geometric program in convex form.* To distinguish it from the original geometric program, we refer to (3.38) as a *geometric program in posynomial form.*

Note that the transformation between the posynomial form geometric program (3.38) and the convex form geometric program (3.39) does not involve any computation; the problem data for the two problems is the same. It simply changes the form of the objective and constraint functions.

> **Remark.** If the posynomial objective and constraint functions all have only one term, *i.e.*, are monomials, then the convex form geometric program (3.39) reduces to a (general) linear program. We can therefore consider geometric programming to be a generalization, or extension, of linear programming.

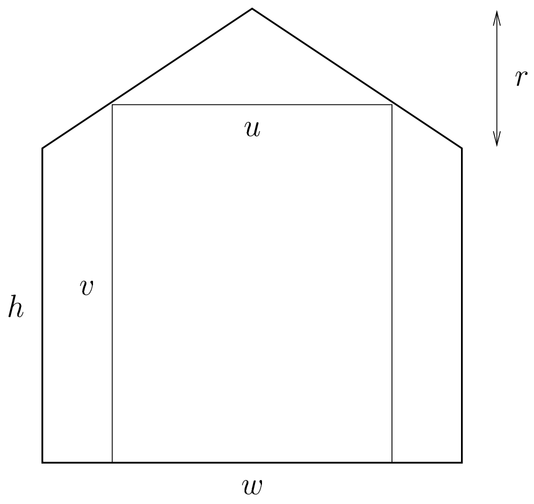

Figure 3.5: House to be optimized. The lengths w, h, and r are design variables to be optimized; u and v are the dimensions of an enclosed rectangle.

3.5.4 Examples

Frobenius norm diagonal scaling

Consider a matrix $M \in \mathbf{R}^{n\times n}$, and the associated linear function that maps u into Mu. If we scale our coordinates, *i.e.*, change variables to $\tilde{u} = Du$, where D is diagonal, with $D_{ii} > 0$. In the new coordinates the linear function is given by $\tilde{u} = DMD^{-1}\tilde{u}$.

Now suppose we want to choose the scaling in such a way that the resulting matrix, DMD^{-1}, is small. If we use the Frobenius norm to measure the size of the matrix, the resulting problem is a geometric program. The problem is

$$\begin{array}{ll} \text{minimize} & \|DMD^{-1}\|_F^2 = \mathbf{Tr}\,(DMD^{-1})^T\,(DMD^{-1}) \\ \text{subject to} & D \text{ diagonal}, \quad D \succ 0. \end{array}$$

This is an unconstrained geometric program in the variables $d_1, \ldots, d_n$, where $d_i = D_{ii}$:

$$\text{minimize} \quad \sum_{i,j=1}^{n} M_{ij}^2 d_i^2 / d_j^2.$$

In this geometric program, the only exponents that occur are 0, 2, and -2.

A simple geometrical problem

We consider the design of the house-like shape shown in figure 3.5. The variables are the width w, wall height h, and roof height r. The house is required to enclose a rectangle of width u and height v, which leads to the inequality $h + ru/w \leq v$.

The cost of the house is

$$\alpha w + 2\beta h + 2\gamma\sqrt{r^2 + (w/2)^2},$$

where α is the cost per unit length of the floor, β is the cost per unit length of the walls, and γ is the cost per unit length of the roof. The cost must not exceed a given budget B.

We are given lower and upper bounds on the variables:

$$w_{\min} \le w \le w_{\max}, \quad h_{\min} \le h \le h_{\max}, \quad r_{\min} \le r \le r_{\max},$$

as well as bounds on the roof pitch, defined as $2r/w$:

$$p_{\min} \le 2r/w \le p_{\max}.$$

(We assume all the lower limits are nonnegative.)

As objective, we are to maximize the area of the enclosed rectangle, *i.e.*, uv. The problem is then

$$\begin{array}{ll} \text{maximize} & uv \\ \text{subject to} & w_{\min} \le w \le w_{\max},\ h_{\min} \le h \le h_{\max},\ r_{\min} \le r \le r_{\max} \\ & p_{\min} \le 2r/w \le p_{\max} \\ & \alpha w + 2\beta h + 2\gamma\sqrt{r^2 + (w/2)^2} \le B \\ & h + ru/w \le v. \end{array}$$

To formulate this problem as a geometric program, we introduce a new variable t, which is an upper bound on the length of the roof, *i.e.*, satisfies $t \ge \sqrt{r^2 + (w/2)^2}$. The problem becomes

$$\begin{array}{ll} \text{maximize} & uv \\ \text{subject to} & w_{\min}/w \le 1,\ h_{\min}/h \le 1,\ r_{\min}/r \le 1,\ p_{\min} w/(2r) \le 1 \\ & w/w_{\max} \le 1,\ h/h_{\max} \le 1,\ r/r_{\max} \le 1,\ 2r/(w p_{\max}) \le 1 \\ & \alpha w/B + 2\beta h/B + 2\gamma t/B \le 1 \\ & h/v + ru/(wv) \le 1 \\ & r^2/t^2 + (w/2)^2/t^2 \le 1. \end{array}$$

which is a geometric program (in posynomial form).

3.6 Generalized inequality constraints

It is very useful to generalize the standard form convex optimization problem (3.15) by allowing the inequality constraint functions to be vector valued, and using generalized inequalities in the constraints:

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \preceq_{K_i} 0, \quad i = 1, \ldots, L, \\ & Ax = b, \end{array} \tag{3.40}$$

where $f_0 : \mathbf{R}^n \to \mathbf{R}$, $K_i \subseteq \mathbf{R}^{m_i}$ are cones that induce generalized inequalities, and $f_i : \mathbf{R}^n \to \mathbf{R}^{m_i}$ are K_i-convex. We refer to this problem as a (standard form) *convex optimization problem with generalized inequality constraints*. Problem (3.15) is a special case with $L = m$ and $K_i = \mathbf{R}_+$, $i = 1, \ldots, L$.

Many of the results for ordinary convex optimization problems hold for problems with generalized inequalities. Some examples are:

- The feasible set, any sublevel set, and the optimal set are convex.
- Any point that is locally optimal for the problem (3.40) is globally optimal.
- The optimality condition for differentiable f_0, given in §3.2.3, holds without any change.

We will also see later that convex optimization problems with generalized inequality constraints can usually be solved as easily as ordinary convex optimization problems.

3.6.1 Conic form problems

Among the simplest convex optimization problems with generalized inequalities are the *conic form problems*, which have a linear objective and one inequality constraint function, which is affine (hence K-convex):

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Fx + g \preceq_K 0 \\ & Ax = b. \end{array} \tag{3.41}$$

Although this problem looks simple, it includes linear programming as a special case (*i.e.*, $K = \mathbf{R}^m_+$).

Continuing the analogy to linear programming, we refer to the conic form problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & x \succeq_K 0 \\ & Ax = b \end{array} \tag{3.42}$$

as a *conic problem in standard form.* Similarly, the problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Fx + g \preceq_K 0 \end{array} \tag{3.43}$$

is called a *conic problem in inequality form.*

Example. *Semidefinite programming.* When K is the cone of positive semidefinite matrices in $\mathbf{R}^{k\times k}$, the associated conic problem is called a *semidefinite program* (SDP), and has the following form:

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & x_1F_1 + \cdots + x_nF_n + G \preceq 0 \\ & Ax = b \end{array}$$

where $F_i = F_i^T \in \mathbf{R}^{k\times k}$, $i = 0, 1, \ldots, n$.

The inequality

$$x_1F_1 + \cdots + x_nF_n + G \preceq 0$$

is called a *linear matrix inequality* (LMI). (The term 'affine matrix inequality' is more accurate, but 'linear matrix inequality' suggests the analogy to the 'linear inequality'

$$x_1f_1 + \cdots + x_nf_n + g \leq 0$$

(where $f_i \in \mathbf{R}$, $g \in \mathbf{R}$.)

3.6.2 Examples

Moment problems

Let t be a random variable in $\mathbf{R}$. The expected values $\mathbf{E}\, t^k$ (assuming they exist) are called the (power) *moments* of the distribution of t. The following classical result gives a characterization of a moment sequence: There exists a probability distribution on $\mathbf{R}$ such that $x_k = \mathbf{E}\, t^k$, $k = 0, \ldots, 2n$, if and only if $x_0 = 1$ and

$$H(x_0, \ldots, x_{2n}) = \begin{bmatrix} x_0 & x_1 & x_2 & \cdots & x_{n-1} & x_n \\ x_1 & x_2 & x_3 & \cdots & x_n & x_{n+1} \\ x_2 & x_3 & x_4 & \cdots & x_{n+1} & x_{n+2} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ x_{n-1} & x_n & x_{n+1} & \cdots & x_{2n-2} & x_{2n-1} \\ x_n & x_{n+1} & x_{n+2} & \cdots & x_{2n-1} & x_{2n} \end{bmatrix} \succeq 0. \tag{3.44}$$

It is easy to see that the condition is necessary: Let $x_i = \mathbf{E}\, t^i$, $i = 0, \ldots, 2n$ be the moments of some distribution, and let $y = [y_0\ y_1\ \cdots\ y_n]^T \in \mathbf{R}^{n+1}$. Then we have

$$y^T H(x_0, \ldots, x_{2n}) y = \sum_{i,j=0}^{n} y_i y_j\, \mathbf{E}\, t^{i+j} = \mathbf{E}\left(y_0 + y_1 t^1 + \cdots + y_n t^n\right)^2 \geq 0.$$

Sufficiency is less obvious; see the Notes and References.

Thus, the condition that $x_0, \ldots, x_{2n}$ be the moments of some distribution on $\mathbf{R}$ can be expressed as the linear matrix inequality (3.44) in the variable x. Using this fact, we can cast some interesting moment problems as SDPs.

Suppose t is a random variable on $\mathbf{R}$. We do not know its distribution, but we do know some bounds on the moments, *i.e.*,

$$\underline{\mu}_k \leq \mathbf{E}\, t^k \leq \overline{\mu}_k$$

(which includes, as a special case, knowing exact values of some of the moments). Let $p(t) = c_0 + c_1 t + \cdots + c_{2n} t^{2n}$ be a given polynomial in t. The expected value of $p(t)$ is linear in the moments $\mathbf{E}\, t^i$:

$$\mathbf{E}\, p(t) = \sum_{i=0}^{2n} c_i\, \mathbf{E}\, t^i = \sum_{i=0}^{2n} c_i x_i.$$

We can compute upper and lower bounds for $\mathbf{E}\, p(t)$,

$$\begin{array}{ll} \text{minimize (maximize)} & \mathbf{E}\, p(t) \\ \text{subject to} & \underline{\mu}_k \leq \mathbf{E}\, t^k \leq \overline{\mu}_k, \quad k = 1, \ldots, 2n, \end{array}$$

over all probability distributions that satisfy the given moment bounds, by solving the SDP

$$\begin{array}{ll} \text{minimize (maximize)} & c_1 x_1 + \cdots + c_{2n} x_{2n} \\ \text{subject to} & \underline{\mu}_k \leq x_k \leq \overline{\mu}_k, \quad k = 1, \ldots, 2n \\ & H(1, x_1, \ldots, x_{2n}) \succeq 0 \end{array}$$

with variables $x_1, \ldots, x_{2n}$. This gives bounds on $\mathbf{E}\, p(t)$, over all probability distributions that satisfy the known moment constraints. The bounds are sharp in the sense that there are distributions, whose moments satisfy the given moment bounds, for which $\mathbf{E}\, p(t)$ takes on the upper and lower bounds found by these SDPs.

3.7 Vector optimization

3.7.1 Pareto optimization

In §3.6 we extended the standard form problem (3.1) to include vector-valued constraint functions f_i. In this section we investigate the meaning of a vector-valued *objective function*.

We denote a general *vector optimization problem* as

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{with respect to} & K \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \end{array} \tag{3.45}$$

where $K \subseteq \mathbf{R}^p$ is a convex cone that induces a generalized inequality, $f_0 : \mathbf{R}^n \to \mathbf{R}^p$, $f_i : \mathbf{R}^n \to \mathbf{R}$. (For simplicity we assume there are no equality or generalized inequality constraints, which are readily incorporated.)

Suppose x and y are two feasible points. Their associated objective values, $f_0(x)$ and $f_0(y)$, are to be compared using the generalized inequality $\preceq_K$. We interpret $f_0(x) \preceq_K f_0(y)$ as meaning that x is 'better than or equal' in value to y (as judged by the objective f_0, with respect to K). The confusing aspect of vector optimization is that the two objective values $f_0(x)$ and $f_0(y)$ need not be comparable; we can have neither $f_0(x) \preceq_K f_0(y)$ nor $f_0(y) \preceq_K f_0(y)$, *i.e.*, neither is better than the other. This, of course, cannot happen in a scalar objective optimization problem.

Consider the set of objective values of feasible points,

$$\mathcal{A} = \{\ f_0(x) \mid f_i(x) \leq 0,\ i = 1, \ldots, m\ \}. \tag{3.46}$$

If this set has a minimum element (see §1.4.3), *i.e.*, there is a feasible x such that $f_0(x) \preceq_K f_0(y)$ for all feasible y, then we say x is *optimal* for the problem (3.45), and refer to $f_0(x)$ as the *optimal value* of the problem. In this case $f_0(x)$, the objective at x, can be compared to the objective at every other feasible point, and is better than or equal to it. This situation sometimes occurs.

In most cases, however, the set $\mathcal{A}$ does not have a minimum element. In such cases minimal elements play an important role. We say that a feasible x is *Pareto optimal* if $f_0(x)$ is a minimal element of the set $\mathcal{A}$. In this case we say that $f_0(x)$ is a *Pareto optimal value* for the vector optimization problem (3.45).

Thus, x is Pareto optimal if it is feasible and, for any feasible y, $f_0(y) \preceq_K f_0(x)$ implies $f_0(y) = f_0(x)$.

Example. *Multicriterion optimization.* The most common application of vector optimization uses $K = \mathbf{R}^p_+$, in which case the problem (3.45) is called a *multicriterion* or *multiobjective* optimization problem; the components of $f_0 = (\phi_1, \dots, \phi_m)$ are interpreted as competing objectives.

In this case an optimal point x satisfies

$$\phi(x) \leq \phi(y), \quad i = 1, \dots, m$$

for every feasible y. In other words, x is simultaneously optimal for each of the scalar problems with objectives ϕ_i.

A Pareto optimal x satisfies the following: if y is feasible and $\phi_i(y) \leq \phi_i(x)$ for $i = 1, \dots, m$, then $\phi_i(x) = \phi_i(y)$.

We say the problem (3.45) is a *convex vector optimization problem* if the constraint functions f_i are convex and f_0 is K-convex.

Example. Suppose $y = Ax + v$ where $v \sim \mathcal{N}(0, I)$ is measurement noise, y is a vector of measurements, and x is a quantity to be determined. A *linear estimator* of x has the form $\widehat{x} = Fy$. The estimator is called *unbiased* if $\mathbf{E}\,\widehat{x} = x$, *i.e.*, if $FA = I$. The error covariance of an unbiased estimator is

$$\mathbf{E}(\widehat{x} - x)(\widehat{x} - x)^T = \mathbf{E}\, Fvv^T F^T = FF^T.$$

We can compare error covariances by matrix inequality, *i.e.*, with respect to the positive semidefinite cone, and consider the vector optimization problem

$$\begin{array}{ll} \text{minimize} & FF^T \\ \text{subject to} & FA = I. \end{array} \tag{3.47}$$

This partial order has the following interpretation. Suppose $\widehat{x}_1 = F_1 y$, $\widehat{x}_2 = F_2 y$ are two unbiased estimators. Then $F_1 F_1^T \preceq F_2 F_2^T$ if and only if for all c

$$\mathbf{E}(c^T \widehat{x}_1 - c^T x)^2 \leq \mathbf{E}(c^T \widehat{x}_2 - c^T x)^2,$$

i.e., the variance of the linear function $c^T \widehat{x}_1$ is always smaller than the variance of $c^T \widehat{x}_2$.

Note that the objective FF^T is convex with respect to the positive semidefinite cone. An easy way to see this is to observe that $v^T FF^T v = \|Fv\|^2$ is a convex function of F for any fixed v.

It can be shown that problem (3.47) has a minimum solution, the least-squares or minimum-variance estimator $F = A^\dagger$: for any F with $FA = I$, we have $FF^T \succeq A^\dagger A^{\dagger T}$. (If we add any constraints, however, the problem generally will not have a minimum element.)

3.7.2 Scalarization

A standard technique can be used to find Pareto optimal (or optimal) points for the problem (3.45). Choose any $\lambda \succ_{K^*} 0$, *i.e.*, any vector that is positive in the dual generalized inequality. Now consider the *scalar* optimization problem

$$\begin{array}{ll} \text{minimize} & \lambda^T f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \; i = 1, \dots, m \end{array} \tag{3.48}$$

and let x be an optimal point. Then x is Pareto optimal for the vector optimization problem (3.45). (This follows from the dual inequality characterization of minimal points; see §1.6.3.) Thus, we can find Pareto optimal points by solving an ordinary scalar optimization problem.

Note that if the vector optimization problem (3.45) is convex, then the scalarized problem (3.48) is also convex, since $\lambda^T f_0$ is a (scalar-valued) convex function. Moreover the set $\mathcal{A}$ is convex, so (again by the results in §1.6.3) we have a partial converse: any Pareto optimal point x solves the scalarized problem (3.48) for some $\lambda \succeq_{K^*} 0$. Roughly speaking, for convex problems, the method of scalarization yields all Pareto optimal points, as the weight λ varies.

Example. *Minimal upper bound of a set matrices.* We consider a vector optimization problem with respect to the the positive semidefinite cone:

$$\begin{array}{ll} \text{minimize} & X \\ \text{with respect to} & \text{pos. semidefinite cone} \\ \text{subject to} & X \succeq A_i, \quad i = 1, \ldots, L, \end{array} \tag{3.49}$$

where $A_i = A_i^T \in \mathbf{R}^{p \times p}$, $i = 1, \ldots, L$ are given matrices. We can interpret a minimal solution of (3.49) as a *minimal upper bound* on the L matrices A_i. In contrast to the previous example, there is no reason why such a minimal solution should be unique.

To compute a minimal lower bound, one can solve the semidefinite programming problem

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}\, WX \\ \text{subject to} & X \succeq A_i, \quad i = 1, \ldots, L \end{array}$$

for some $W \succ 0$. Different W will give different solutions in general.

Chapter 4

Duality

4.1 The Lagrange dual function

4.1.1 The Lagrangian

We consider an optimization problem in the standard form (3.1):

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p. \end{array} \tag{4.1}$$

We assume its domain $D = \bigcap_{i=0}^{m} \mathbf{dom}\, f_i \,\cap\, \bigcap_{i=1}^{p} \mathbf{dom}\, h_i$ is nonempty, and denote the optimal value of (4.1) by $p^\star$. We do not assume the problem (4.1) is convex.

The basic idea in Lagrangian duality is to take the constraints in (4.1) into account by augmenting the objective function with a weighted sum of the constraint functions. We define the *Lagrangian* $L : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$ associated with the problem (4.1) as

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x), \tag{4.2}$$

with $\mathbf{dom}\, L = D \times \mathbf{R}^{m+p}$. We refer to λ_i as the *Lagrange multiplier* associated with the ith inequality constraint $f_i(x) \leq 0$; similarly we refer to ν_i as the Lagrange multiplier associated with the ith equality constraint $h_i(x) = 0$. The vectors λ and ν are called the *dual variables* or *Lagrange multiplier vectors* associated with the problem (4.1).

4.1.2 The Lagrange dual function

We define the *Lagrange dual function* (or just dual function) $g : \mathbf{R}^{m+p} \to \mathbf{R}$ as the minimum value of the Lagrangian over x, *i.e.*, for $\lambda \in \mathbf{R}^m$, $\nu \in \mathbf{R}^p$,

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) = \inf_{x \in D} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right). \tag{4.3}$$

When the Lagrangian is unbounded below in x, the dual function takes on the value $-\infty$.

Since the dual function is the pointwise infimum of a family of affine functions of (λ, ν), it is concave (even when the problem (4.1) is not convex).

4.1.3 Lower bounds on optimal value

The dual function yields lower bounds on the optimal value $p^\star$ of the problem (4.1): for any $\lambda \succeq 0$ and any ν we have

$$g(\lambda, \nu) \leq p^\star. \tag{4.4}$$

This important property is easily verified. If $\tilde{x}$ is any feasible point for the problem (4.1), *i.e.*, $f_i(\tilde{x}) \leq 0$ and $h_i(\tilde{x}) = 0$, and $\lambda \succeq 0$, then we have

$$L(\tilde{x}, \lambda, \nu) = f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{x}) \leq f_0(\tilde{x}),$$

since $\lambda_i \geq 0$, $f_i(\tilde{x}) \leq 0$ and $h_i(\tilde{x}) = 0$. Hence

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) \leq L(\tilde{x}, \lambda, \nu) \leq f_0(\tilde{x}).$$

Since $g(\lambda, \nu) \leq f_0(\tilde{x})$ holds for every feasible point $\tilde{x}$, the inequality (4.4) follows.

The inequality (4.4) holds, but is vacuous, when $g(\lambda, \nu) = -\infty$. The dual function gives a nontrivial lower bound on $p^\star$ only when $(\lambda, \nu) \in \mathbf{dom}\, g$, *i.e.*, $g(\lambda, \nu) > -\infty$.

4.1.4 Examples

In this section we give some examples for which we can derive an analytical expression for the Lagrange dual function.

Equality constrained least-squares

We consider the problem

$$\begin{array}{ll} \text{minimize} & x^T x \\ \text{subject to} & Ax = b \end{array} \tag{4.5}$$

where $A \in \mathbf{R}^{p \times n}$. This problem has no inequality constraints and p (linear) equality constraints.

The Lagrangian is $L(x, \nu) = x^T x + \nu^T (Ax - b)$, with domain $\mathbf{R}^n \times \mathbf{R}^p$. The dual function is given by $g(\nu) = \inf_x L(x, \nu)$. Since $L(x, \nu)$ is a convex quadratic (hence smooth) function of x, we can find the minimizing x from the optimality condition

$$\nabla_x L(x, \nu) = 2x + A^T \nu = 0,$$

which yields $\tilde{x} = -(1/2) A^T \nu$. Therefore the dual function is

$$g(\nu) = L(-(1/2)A^T \nu, \nu) = -\frac{1}{4} \nu^T A A^T \nu - b^T \nu.$$

g is a strictly concave quadratic function, with domain $\mathbf{R}^p$.

Standard form LP

Consider an LP in standard form,

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \\ & x \succeq 0, \end{array} \tag{4.6}$$

which in standard form has inequality constraint functions $f_i(x) = -x_i$, $i = 1, \ldots, n$. To form the Lagrangian we introduce multipliers λ_i for the n inequality constraints and multipliers ν_i for the equality constraints, and obtain

$$L(x, \lambda, \nu) = c^T x - \sum_{i=1}^n \lambda_i x_i + \nu^T (Ax - b) = -b^T \nu + (c + A^T \nu - \lambda)^T x.$$

The dual function is

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu) = -b^T \nu + \inf_x (c + A^T \nu - \lambda)^T x,$$

which is easily determined analytically, since a linear function is bounded below only when it is identically zero. Thus, $g(\lambda, \nu) = -\infty$ except when $c + A^T \nu - \lambda = 0$, in which case it is $-b^T \nu$:

$$g(\lambda, \nu) = \begin{cases} -b^T \nu & \text{if } A^T \nu - \lambda + c = 0 \\ -\infty & \text{otherwise.} \end{cases} \tag{4.7}$$

Note that the dual function g is finite only on a proper affine subset of (λ, ν). We will see that this is a common occurence.

A nonconvex quadratic problem

We consider the (nonconvex) problem

$$\begin{array}{ll} \text{minimize} & x^T W x \\ \text{subject to} & x_i^2 = 1, \; i = 1, \ldots, n \end{array} \tag{4.8}$$

where $W = W^T \in \mathbf{R}^{n \times n}$. The constraints restrict the values of x_i to 1 or -1, so the problem is equivalent to finding the vector with components ± 1 that minimizes $x^T W x$.

The Lagrangian is

$$\begin{aligned} L(x, \nu) &= x^T W x + \sum_{i=1}^n \nu_i (x_i^2 - 1) \\ &= x^T (W + \mathbf{diag}(\nu)) x - \mathbf{1}^T \nu. \end{aligned}$$

We obtain the Lagrange dual function by minimizing over x:

$$\begin{aligned} g(\nu) &= \inf_x x^T (W + \mathbf{diag}(\nu)) x - \mathbf{1}^T \nu \\ &= \begin{cases} -\mathbf{1}^T \nu & W + \mathbf{diag}(\nu) \succeq 0 \\ -\infty & \text{otherwise.} \end{cases} \end{aligned}$$

4.1.5 The Lagrange dual function and conjugate functions

Recall from §2.3 that the conjugate $f^\star$ of a function $f : \mathbf{R}^n \to \mathbf{R}$ is given by

$$f^\star(y) = \sup_x \left(y^T x - f(x) \right).$$

The conjugate function and Lagrange dual are closely related. To see one simple connection between them, consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x = 0 \end{array}$$

(which is not very interesting, and solvable by inspection). This problem has Lagrangian $L(x, \nu) = f_0(x) + \nu^T x$, and dual function

$$g(\nu) = \inf_x \left(f_0(x) + \nu^T x \right) = -\sup_x \left((-\nu)^T x - f_0(x) \right) = -f^\star(-\nu).$$

More generally (and more usefully), consider an optimization problem with linear inequality and equality constraints,

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & Ax \preceq b, \\ & Cx = d. \end{array} \tag{4.9}$$

Using the conjugate of f_0 we can write the dual function as

$$\begin{array}{rcl} g(\lambda) & = & \inf_x \left(f_0(x) + \lambda^T (Ax - b) + \nu^T (Cx - d) \right) \\ & = & -b^T \lambda - d^T \nu + \inf_x \left(f_0(x) + (A^T \lambda + C^T \nu)^T x \right) \\ & = & -b^T \lambda - d^T \nu - f_0^\star(-A^T \lambda - C^T \nu). \end{array}$$

The domain of g follows from the domain of $f_0^\star$:

$$\mathbf{dom}\, g = \{ (\lambda, \nu) \mid -A^T \lambda - C^T \nu \in \mathbf{dom}\, f_0^\star \}.$$

Let us illustrate this with a few examples.

Linearly constrained norm minimization

Consider the problem

$$\begin{array}{ll} \text{minimize} & \|x\| \\ \text{subject to} & Ax = b, \end{array} \tag{4.10}$$

where $\|\cdot\|$ is any norm. Recall that the conjugate of $f_0 = \|\cdot\|$ is given by

$$f_0^\star(y) = \begin{cases} 0 & \|y\|_* \le 1 \\ \infty & \text{otherwise,} \end{cases}$$

i.e., the indicator function of the dual norm ball.

Using the result above, the dual function for the problem (4.10) is given by

$$g(\nu) = -b^T \nu - f_0^\star(-A^T \nu) = \begin{cases} -b^T \nu & \|A^T \nu\|_* \le 1 \\ -\infty & \text{otherwise.} \end{cases}$$

Entropy maximization

Consider the entropy maximization problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & Ax \preceq b \\ & \mathbf{1}^T x = 1 \end{array} \tag{4.11}$$

where $f_0(x) = \sum_{i=1}^n x_i \log x_i$ on $\mathbf{dom}\, f_0 = \{x \mid x \succ 0\}$. We derived the conjugate of this function in §2.3, and found that

$$f_0^\star(y) = \sum_{i=1}^n e^{y_i - 1},$$

with $\mathbf{dom}\, f_0^\star = \mathbf{R}^n$. Therefore the dual function of (4.11) is given by

$$g(\lambda, \nu) = -b^T\lambda - \nu - \sum_{i=1}^n e^{-a_i^T\lambda - \nu - 1} = -b^T\lambda - \nu - e^{-\nu-1} \sum_{i=1}^n e^{-a_i^T\lambda}$$

where $A = [a_1\ a_2\ \cdots\ a_n]$.

Minimum volume covering ellipsoid

Consider the problem with matrix variable $X = X^T \in \mathbf{R}^{n\times n}$,

$$\begin{array}{ll} \text{minimize} & f_0(X) \\ \text{subject to} & a_i^T X a_i \leq 1, \quad i = 1, \ldots, m \end{array} \tag{4.12}$$

where $\mathbf{dom}\, f_0 = \{X = X^T \in \mathbf{R}^{n\times n} \mid X \succ 0\}$ and, for $X \succ 0$,

$$f_0(X) = \log\det X^{-1}.$$

The problem (4.12) has a simple geometrical interpretation: With each $X = X^T \succ 0$ we associate the ellipsoid, centered at the origin,

$$\mathcal{E}_X = \{\ z \mid z^T X z \leq 1\ \}.$$

The volume of this ellipsoid is proportional to $\det X^{-1}$, so the objective of (4.12) is, except for a constant, the logarithm of the volume of $\mathcal{E}_X$. The constraints of the problem (4.12) are that $a_i \in \mathcal{E}_X$. Thus the problem (4.12) is to determine the minimum volume ellipsoid, centered at the origin, that includes the points $a_1, \ldots, a_m$.

The inequality constraints in problem (4.12) are affine; they can be expressed as $\mathbf{Tr}(a_i a_i^T)X \leq 1$. In §2.3 we found the conjugate of f_0 is

$$f_0^\star(Y) = \log\det(-Y)^{-1} - n,$$

with $\mathbf{dom}\, f_0^\star = \{Y = Y^T \in \mathbf{R}^{n\times n} \mid Y \prec 0\}$. Applying our result above, the dual function for the problem (4.12) is given by

$$g(\lambda) = \begin{cases} -\log\det\left(\sum_{i=1}^m \lambda_i a_i a_i^T\right) - \sum_{i=1}^m \lambda_i + n & \text{if } \sum_{i=1}^m \lambda_i a_i a_i^T \succ 0 \\ -\infty & \text{otherwise.} \end{cases} \tag{4.13}$$

Thus, for any $\lambda \succeq 0$ with $\sum_{i=1}^m \lambda_i a_i a_i^T \succ 0$, the number

$$-\log\det\left(\sum_{i=1}^m \lambda_i a_i a_i^T\right) - \sum_{i=1}^m \lambda_i + n$$

is a lower bound on the optimal value of the problem (4.12).

4.2 The Lagrange dual problem

For each pair λ, ν with $\lambda \succeq 0$, the Lagrange dual function gives us a lower bound on the optimal value $p^\star$ of the optimization problem (4.1). Thus we have a lower bound that depends on some parameters, *i.e.*, the Lagrange multipliers λ, ν. A natural question is: what is the *best* lower bound that can be obtained from the Lagrange dual function?

This leads to the optimization problem

$$\begin{array}{ll} \text{maximize} & g(\lambda,\nu) \\ \text{subject to} & \lambda \succeq 0. \end{array} \tag{4.14}$$

This problem is called the *Lagrange dual problem* associated with the problem (4.1). In this context the original problem (4.1) is sometimes called the *primal problem.* We say (λ,ν) are *dual feasible* if $\lambda \succeq 0$ and $g(\lambda,\nu) > -\infty$. We refer to (λ,ν) as *dual optimal* or *optimal Lagrange multipliers* if they are optimal for the problem (4.14).

The Lagrange dual problem (4.14) is a convex optimization problem, since the objective to be maximized is concave and the constraint is convex. This is the case whether or not the primal problem (4.1) is convex.

4.2.1 Making dual constraints explicit

The examples above show that it is not uncommon for the domain of the dual function,

$$\mathbf{dom}\, g = \{\ (\lambda,\nu) \mid g(\lambda,\nu) > -\infty\ \},$$

to have dimension smaller than $m+p$. In many cases we can identify the affine hull of $\mathbf{dom}\, g$, and decribe it as a set of linear equality constraints. Roughly speaking, this means we can identify the equality constraints that are 'hidden' or 'implicit' in the objective g of the dual problem (4.1). In this case we can form an equivalent problem, in which these equality constraints are given explicitly as constraints. The following examples demonstrate this idea.

Lagrange dual of standard form LP

In §4.1.4 we found that the Lagrange dual function for the standard form LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \\ & x \succeq 0 \end{array} \tag{4.15}$$

is given by

$$g(\lambda, \nu) = \begin{cases} -b^T\nu & \text{if } A^T\nu - \lambda + c = 0 \\ -\infty & \text{otherwise.} \end{cases}$$

Strictly speaking, the Lagrange dual problem of the standard form LP is to maximize this dual function g subject to $\lambda \succeq 0$, *i.e.*,

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) = \begin{cases} -b^T\nu & \text{if } A^T\nu - \lambda + c = 0 \\ -\infty & \text{otherwise.} \end{cases} \\ \text{subject to} & \lambda \succeq 0. \end{array} \tag{4.16}$$

Here g is finite only when $A^T\nu - \lambda + c = 0$. We can form an equivalent problem by making these equality constraints explicit:

$$\begin{array}{ll} \text{maximize} & -b^T\nu \\ \text{subject to} & A^T\nu - \lambda + c = 0 \\ & \lambda \succeq 0. \end{array} \tag{4.17}$$

This problem, in turn, can be expressed as

$$\begin{array}{ll} \text{maximize} & -b^T\nu \\ \text{subject to} & A^T\nu + c \succeq 0. \end{array} \tag{4.18}$$

Note the subtle distinctions between these three problems. The Lagrange dual of the standard form LP (4.15) is the problem (4.16), which is equivalent (but not the same as) the problems (4.17) and (4.18). With some abuse of terminology, we refer to the problem (4.17) or the problem (4.18) as the Lagrange dual of the standard form LP (4.15).

Lagrange dual of inequality form LP

In a similar way we can find the Lagrange dual problem of a linear program in inequality form

$$\begin{array}{ll} \text{minimize} & c^Tx \\ \text{subject to} & Ax \preceq b. \end{array} \tag{4.19}$$

The Lagrangian is

$$L(x, \lambda) = c^Tx + \lambda^T(Ax - b) = -b^T\lambda + (A^T\lambda + c)^Tx,$$

so the dual function is

$$g(\lambda) = \inf_x L(x, \lambda) = -b^T\lambda + \inf_x (A^T\lambda + c)^Tx.$$

The infimum of a linear function is $-\infty$, except in the special case when it is identically zero, so the dual function is

$$g(\lambda) = \begin{cases} -b^T\lambda & \text{if } A^T\lambda + c = 0 \\ -\infty & \text{otherwise.} \end{cases} \tag{4.20}$$

The dual variable λ is dual feasible if $\lambda \succeq 0$ and $A^T\lambda + c = 0$.

The Lagrange dual of the LP (4.19) is to maximize g over all $\lambda \succeq 0$. Again we can reformulate this by explicitly including the dual feasibility conditions as constraints, as in

$$\begin{array}{ll} \text{maximize} & -b^T\lambda \\ \text{subject to} & \lambda \succeq 0 \\ & A^T\lambda + c = 0, \end{array} \tag{4.21}$$

which is an LP in standard form.

Note the interesting symmetry between the standard and inequality form LPs and their duals: The dual of a standard form LP is an LP with only inequality constraints, and vice-versa. One can also verify that the Lagrange dual of (4.21) is (equivalent to) the primal problem (4.19).

4.2.2 Weak duality

The optimal value of the Lagrange dual problem, which we denote $d^\star$, is, by definition, the best lower bound on $p^\star$ that can be obtained from the Lagrange dual function. In particular, we have the simple but important inequality

$$d^\star \leq p^\star \tag{4.22}$$

(which holds even if the original problem is not convex). This property is called *weak duality.*

The weak duality inequality (4.22) holds when $d^\star$ and $p^\star$ are infinite. For example, if the primal problem is unbounded below, so that $p^\star = -\infty$, we must have $d^\star = -\infty$, *i.e.*, the Lagrange dual problem is infeasible. Conversely, if the dual problem is unbounded above, so that $d^\star = +\infty$, we must have $p^\star = +\infty$, *i.e.*, the primal problem is infeasible.

We refer to the difference $p^\star - d^\star$ as the *optimal duality gap* of the original problem, since it gives the 'gap' between the optimal value of the primal problem and the best (*i.e.*, largest) lower bound on it that can be obtained from the Lagrange dual function. The optimal duality gap is always nonnegative.

4.2.3 Strong duality and constraint qualifications

If the equality

$$d^\star = p^\star, \tag{4.23}$$

holds, *i.e.*, the optimal duality gap is zero, then we say that *strong duality* holds. This means that the bounds that can be obtained from the Lagrange dual function are sharp.

Strong duality does not, in general, hold. But if the primal problem (4.1) is convex, *i.e.*, of the form

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & Ax = b, \end{array} \tag{4.24}$$

with $f_0, \ldots, f_m$ convex, we usually (but not always) have strong duality. There are many results that establish conditions, beyond convexity, on the problem under which strong duality holds. These conditions are called *constraint qualifications.*

Slater's constraint qualification

One simple constraint qualification is *Slater's condition*: there exists an $x \in \mathbf{relint}\, D$ such that

$$f_i(x) < 0, \ i = 1, \ldots, m, \quad Ax = b. \tag{4.25}$$

Such a point is sometimes called *strictly feasible*.

A refinement of Slater's condition is possible when some of the inequality constraint functions f_i are affine. If the first m_1 constraint functions f_i are affine, then strong duality holds provided the following weaker condition holds: there exists an $x \in \mathbf{relint}\, D$ with

$$f_i(x) \leq 0, \ \ i = 1, \ldots, m_1, \quad f_i(x) < 0, \ \ i = m_1 + 1, \ldots, m, \quad Ax = b. \tag{4.26}$$

In other words, the affine inequalities do not have to hold with strict inequality. Note that the condition (4.26) reduces to feasibility when the constraints are all linear equalities and inequalities, and $\mathbf{dom}\, f_0$ is open.

Slater's condition (and the refinement (4.26)) not only implies strong duality, *i.e.*, $p^\star = d^\star$. It also implies that the dual optimal value is attained, *i.e.*, there exists a dual feasible $(\lambda^\star, \nu^\star)$ with $g(\lambda^\star) = d^\star = p^\star$.

We will prove that strong duality obtains, when Slater's condition holds, in §4.2.5.

4.2.4 Examples

Equality constrained least-squares

Recall the equality constrained least-squares problem (4.5). The associated dual problem is

$$\text{maximize} \quad -\frac{1}{4}\nu^T AA^T \nu - b^T \nu,$$

which is an unconstrained concave quadratic maximization problem.

Slater's condition is simply that the primal problem is feasible, so $p^\star = d^\star$ provided $b \in \text{range}(A)$, *i.e.*, $p^\star < \infty$. In fact for this problem we always have strong duality, even when $p^\star = \infty$. This is the case when $b \notin \text{range}(A)$, so there is a v with $A^T v = 0$, $b^T v \neq 0$. It follows that the dual function is unbounded above along the line $\{ tv \mid t \in \mathbf{R} \}$, so $d^\star = \infty$ as well.

For this problem we can also solve the primal and dual problems analytically, and directly verify strong duality.

Lagrange dual of LP

By the weaker form of Slater's condition, we find that strong duality holds for any LP (in standard or inequality form) provided the primal problem is feasible. Applying this result to the duals, we conclude that strong duality holds for LPs if the dual is feasible. This leaves only one possible situation in which strong duality for LPs can fail: both the primal and dual problems are infeasible. In fact, this pathological case can occur.

Lagrange dual of convex QCQP

We consider the convex QCQP

$$\begin{array}{ll} \text{minimize} & x^T P_0 x + 2q_0^T x + r_0 \\ \text{subject to} & x^T P_i x + 2q_i^T x + r_i \leq 0, \quad i = 1, \ldots, m, \end{array} \tag{4.27}$$

with $P_0 \succ 0$, and $P_i \succeq 0$, $i = 1, \ldots, m$. The Lagrangian is

$$L(x, \lambda) = x^T P(\lambda) x + 2q(\lambda)^T x + r(\lambda)$$

where

$$P(\lambda) = P_0 + \sum_{i=1}^m \lambda_i P_i, \quad q(\lambda) = q_0 + \sum_{i=1}^m \lambda_i q_i, \quad r(\lambda) = r_0 + \sum_{i=1}^m \lambda_i r_i.$$

It is possible to derive an expression for $g(\lambda)$ for general λ, but it is quite complicated. If $\lambda \succeq 0$, however, we have $P(\lambda) \succ 0$ and

$$g(\lambda) = \inf_x L(x, \lambda) = -q(\lambda)^T P(\lambda)^{-1} q(\lambda) + r(\lambda).$$

We can therefore express the dual problem as

$$\begin{array}{ll} \text{maximize} & -q(\lambda)^T P(\lambda)^{-1} q(\lambda) + r(\lambda) \\ \text{subject to} & \lambda \succeq 0. \end{array} \tag{4.28}$$

The Slater condition says that strong duality between (4.28) and (4.27) holds if the quadratic inequality constraints are strictly feasible, *i.e.*, there exists an x with

$$x^T P_i x + 2q_i^T x + r_i < 0, \quad i = 1, \ldots, m.$$

Entropy maximization

Our next example is the entropy maximization problem (4.11). The dual problem is

$$\begin{array}{ll} \text{maximize} & -b^T \lambda - \nu - e^{-\nu - 1} \sum_{i=1}^m e^{-a_i^T \lambda} \\ \text{subject to} & \lambda \succeq 0, \end{array} \tag{4.29}$$

with variables $\lambda \in \mathbf{R}^m$, $\nu \in \mathbf{R}$. The (weaker) Slater condition for (4.11) tells us that the optimal duality gap is zero if there exists an $x \succ 0$ with $Ax \preceq b$ and $\mathbf{1}^T x = 1$.

We can simplify the dual problem (4.29) by maximizing over the dual variable ν analytically. For fixed λ, the objective function is maximized when the derivative with respect to ν is zero, *i.e.*,

$$\nu = \log \sum_{i=1}^m e^{-a_i^T \lambda} - 1.$$

Substituting this optimal value of ν into the dual problem gives

$$\begin{array}{ll} \text{maximize} & -b^T \lambda - \log \sum_{i=1}^m e^{a_i^T \lambda} \\ \text{subject to} & \lambda \succeq 0, \end{array}$$

which is a geometric program (in convex form) with nonnegativity constraints.

Minimum volume covering ellipsoid

We consider the problem (4.12). The Lagrange dual function is given by (4.13), so the dual problem can be expressed as

$$\begin{array}{ll} \text{maximize} & -\log\det\left(\sum_{i=1}^m \lambda_i a_i a_i^T\right) - \sum_{i=1}^m \lambda_i + n \\ \text{subject to} & \lambda \succeq 0 \end{array} \tag{4.30}$$

where we take $\log\det X = -\infty$ if $X \not\succ 0$.

The (weaker) Slater condition for the problem (4.12) is that there exists an $X = X^T \succ 0$ with $a_i^T X a_i \leq 1$, for $i = 1, \ldots, m$. This is always satisfied, so strong duality always obtains between (4.12) and the dual problem (4.30).

4.2.5 Proof of strong duality under constraint qualification

In this section we prove that Slater's constraint qualification guarantees strong duality (and that the dual optimum is attained). We consider the primal problem (4.24), with f_i convex. In order to simplify the proof, we make two additional assumptions: first that D has nonempty interior (hence, $\mathbf{relint}\, D = \mathbf{int}\, D$) and second, that A is full rank, *i.e.*, has rank m. We assume Slater's condition holds: there exists $\tilde{x} \in \mathbf{int}\, D$ with $f_i(\tilde{x}) < 0$, $i = 1, \ldots, m$, and $A\tilde{x} = b$.

We assume $p^\star$ is finite. (Since there is a feasible point, we can only have $p^\star = -\infty$ or $p^\star$ finite; if $p^\star = -\infty$, then $d^\star = -\infty$ by weak duality.)

Consider the two sets

$$\mathcal{A} = \left\{ (u, v, t) \in \mathbf{R}^{m+p+1} \;\middle|\; \exists x \in D \; f_i(x) \leq u_i,\; i = 1, \ldots, m,\; Ax - b = v,\; f_0(x) \leq t \right\}, \tag{4.31}$$

and

$$\mathcal{B} = \left\{ (0, 0, s) \in \mathbf{R}^{m+p+1} \;\middle|\; s < p^\star \right\}, \tag{4.32}$$

which are readily shown to be convex. These sets do not intersect, *i.e.*, $\mathcal{A} \cap \mathcal{B} = \emptyset$. To see this suppose $(u, v, t) \in \mathcal{A} \cap \mathcal{B}$. Since $(u, v, t) \in \mathcal{B}$ we have $u = 0$, $v = 0$, and $t < p^\star$. Since $(u, v, t) \in \mathcal{A}$, there exists an x with $f_i(x) \leq 0$, $i = 1, \ldots, m$, $Ax - b = 0$, and $f_0(x) \leq t < p^\star$, which is impossible since x is feasible.

By the separation theorem there exists $(\tilde{\lambda}, \tilde{\nu}, \mu) \neq 0$ and α such that

$$(u, v, t) \in \mathcal{A} \;\Rightarrow\; \tilde{\lambda}^T u + \tilde{\nu}^T v + \mu t \geq \alpha, \tag{4.33}$$

and

$$(u, v, t) \in \mathcal{B} \;\Rightarrow\; \tilde{\lambda}^T u + \tilde{\nu}^T v + \mu t \leq \alpha. \tag{4.34}$$

From (4.33) we conclude that $\tilde{\lambda} \succeq 0$ and $\mu \geq 0$. (If that is not true, then $\tilde{\lambda}^T u + \mu t$ would be unbounded above, so (4.33) could not hold.)

The condition (4.34) simply means that $\mu t \leq \alpha$ for all $t < p^\star$, and hence, $\mu p^\star \leq \alpha$. Together with (4.33) we conclude that for any $x \in D$,

$$\sum_{i=1}^m \tilde{\lambda}_i f_i(x) + \tilde{\nu}^T(Ax - b) + \mu f_0(x) \geq \alpha \geq \mu p^\star. \tag{4.35}$$

First assume that $\mu > 0$. In that case we can divide (4.35) by μ to obtain

$$L(x, \tilde{\lambda}/\mu, \tilde{\nu}/\mu) \geq p^\star$$

for all $x \in D$, from which it follows, by minimizing over x, that $g(\lambda, \nu) \geq p^\star$, where we define

$$\lambda = \tilde{\lambda}/\mu, \quad \nu = \tilde{\nu}/\mu.$$

By weak duality we have $g(\lambda, \nu) \leq p^\star$, so in fact $g(\lambda, \nu) = p^\star$. This shows that strong duality holds, and that the dual optimum is attained, at least in the case when $\mu > 0$.

Now consider the case $\mu = 0$. From (4.35), we conclude that for all $x \in D$,

$$\sum_{i=1}^m \tilde{\lambda}_i f_i(x) + \tilde{\nu}^T(Ax - b) \geq 0.$$

Applying this to the point $\tilde{x}$ that satisfies the Slater condition, we have

$$\sum_{i=1}^m \tilde{\lambda}_i f_i(\tilde{x}) \geq 0.$$

Since $f_i(\tilde{x}) < 0$ and $\tilde{\lambda}_i \geq 0$, we conclude that $\tilde{\lambda} = 0$. From $(\tilde{\lambda}, \tilde{\nu}, \mu) \neq 0$ and $\tilde{\lambda} = 0$, $\mu = 0$, we conclude that $\tilde{\nu} \neq 0$. Then (4.35) implies that for all $x \in D$, $\tilde{\nu}^T(Ax - b) \geq 0$. But $\tilde{x}$ satisfies $\tilde{\nu}^T(A\tilde{x} - b) = 0$, and since $\tilde{x} \in \mathbf{int}\, D$, there are points in D with $\tilde{\nu}^T(Ax - b) < 0$ unless $A^T\nu = 0$. This, of course, contradicts our assumption that A is full rank.

4.3 Interpretations

In this section we give several interpretations of Lagrange duality. The material of this section will not be used in the sequel.

4.3.1 Max-min characterization of strong and weak duality

We can express the primal and the dual optimization problems in a form that is more symmetric. To simplify the discussion we assume there are no equality constraints; the results are easily extended to cover them.

First note that

$$\sup_{\lambda \succeq 0} L(x, \lambda) = \sup_{\lambda \succeq 0} \left(f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) \right) = \begin{cases} f_0(x) & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ +\infty & \text{otherwise.} \end{cases}$$

Indeed, suppose x is not feasible, and $f_i(x) > 0$ for some i. Then $\sup_{\lambda \succeq 0} L(x, \lambda) = +\infty$, as can be seen by choosing $\lambda_j = 0$, $j \neq i$, and $\lambda_i \to \infty$. On the other hand, if $f_i(x) \leq 0$, $i = 1, \ldots, m$, then the optimal choice of λ is $\lambda = 0$ and $\sup_{\lambda \succeq 0} L(x, \lambda) = f_0(x)$. This means that we can express the optimal value of the primal problem as

$$p^\star = \inf_x \sup_{\lambda \succeq 0} L(x, \lambda).$$

On the other hand, by the definition of the dual function, we have

$$d^\star = \sup_{\lambda \succeq 0} \inf_x L(x, \lambda).$$

Thus, weak duality can be expressed as the inequality

$$d^\star = \sup_{\lambda \succeq 0} \inf_x L(x, \lambda) \leq p^\star = \inf_x \sup_{\lambda \succeq 0} L(x, \lambda). \tag{4.36}$$

In fact, this inequality does not depend on any properties of L:

$$\sup_{w \in W} \inf_{z \in Z} f(w, z) \leq \inf_{z \in Z} \sup_{w \in W} f(w, z) \tag{4.37}$$

holds for *any* function f, and any sets W, Z. This general inequality is called the *weak max-min inequality*.

Strong duality can be expressed as

$$\sup_{\lambda \succeq 0} \inf_x L(x, \lambda) = \inf_x \sup_{\lambda \succeq 0} L(x, \lambda). \tag{4.38}$$

In other words, strong duality means that the order of the minimization over x and the maximization over $\lambda \succeq 0$ can be switched without affecting the value.

When equality holds in the general case, *i.e.*,

$$\sup_{w \in W} \inf_{z \in Z} f(w, z) = \inf_{z \in Z} \sup_{w \in W} f(w, z) \tag{4.39}$$

we say that f (and W and Z) satisfy the *saddle-point* property. Of course the saddle-point property holds only in special cases, for example, the Lagrangian of a problem for which strong duality obtains.

4.3.2 Game interpretation

We can interpret (4.37) and (4.39) in terms of a zero-sum game, as follows. If the first player chooses $w \in W$ and the second player selects $z \in Z$, then player 1 pays an amount $f(w, z)$ to player 2. Player 1 therefore wants to minimize L, while player 2 wants to maximize L.

First suppose that player 1 makes his choice first, and then player 2, after learning player 1's choice, makes her selection. Player 2 wants to maximize the payoff $L(w, z)$, and so will choose $z \in Z$ to maximize $f(w, z)$. The resulting payoff will be $\sup_{z \in Z} f(w, z)$, which depends on w. Player 1 knows (or assumes) that player 2 will follow this strategy, and so will choose $w \in W$ to make this worst-case payoff to player 2 as small as possible. Thus player 1 chooses

$$\operatorname*{argmin}_{w \in W} \sup_{z \in Z} f(w, z),$$

which results in the payoff

$$\inf_{w \in W} \sup_{z \in Z} f(w, z)$$

from player 1 to player 2.

Now suppose the order of play is reversed: Player 2 must choose $z \in Z$ first, and then player 1 chooses $w \in W$ (with knowledge of z). Following a similar argument, if the players follow the optimal strategy, player 2 should choose $z \in Z$ to maximize $\inf_{w \in W} f(w, z)$, which results in payoff of

$$\sup_{z \in Z} \inf_{w \in W} f(w, z)$$

from player 1 to player 2.

The weak max-min inequality (4.37) states the (intuitively obvious) fact that it is better for a player to go second, or more precisely, for a player to know his or her opponent's choice before choosing. In other words, the payoff to player 2 will be larger if player 1 must choose first.

When the saddle-point property (4.39) holds, there is no advantage to playing second, *i.e.*, there is no advantage to knowing your opponent's choice.

Now consider the special case where the payoff function is the Lagrangian, and $Z = \mathbf{R}^m_+$. Here player 1 chooses the primal variable x, while player 2 chooses the dual variable $\lambda \succeq 0$. By the argument above, the optimal choice for player 1, if he must choose first, is any $x^\star$ which is optimal for the primal problem, which results in payoff to player 2 of $p^\star$. Conversely, if player 2 must choose first, her optimal choice is $\lambda^\star$, which is dual optimal, which results in a payoff of $d^\star$.

The optimal duality gap for the problem is exactly equal to the advantage afforded the player who goes second, *i.e.*, the player who has the advantage of knowing his or her opponent's choice before choosing. If strong duality holds, then there is no advantage to the players of knowing their opponent's choice.

4.3.3 Price or tax interpretation

Lagrange duality has an interesting economic interpretation. Suppose the variable x denotes how an enterprise operates and $f_0(x)$ denotes the cost of operating at x, *i.e.*, $-f_0(x)$ is the profit (say, in dollars) made at the operating condition x. Each constraint $f_i(x) \leq 0$ represents some limit, such as a limit on resources (*e.g.*, warehouse space, personnel) or regulatory limits (*e.g.*, environmental). The operating condition that maximizes profit while respecting the limits can be found by solving the basic problem (4.1). The resulting optimal profit is $-p^\star$.

Now imagine a second scenario in which the limits can be violated, by paying an additional cost which is linear in the amount of violation, measured by f_i. Thus the payment made by the enterprise for the ith limit or constraint is $\lambda_i f_i(x)$. Payments are also made *to* the firm for constraints that are not tight; if $f_i(x) < 0$, then $\lambda_i f_i(x)$ represents a payment to the firm. The coefficient λ_i has the interpretation of price for violating $f_i(x) \leq 0$; its units are dollars per unit violation (as measured by f_i). For the same price the enterprise can sell any 'unused' portion of the ith constraint. We assume $\lambda_i \geq 0$, *i.e.*, the firm must pay for violations (and receives income if a constraint is not tight).

As an example, the original problem may have a limit on warehouse space (say, in square feet). In this new arrangement, we open the possibility that the firm can rent extra warehouse space at a cost of λ_1 dollars per square foot, and also rent out unused space, at the same rate.

Thus the total cost to the firm, for operating condition x, and constraint prices λ_i, is $L(x, \lambda) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x)$. The firm will obviously operate so as to minimize its total cost $L(x, \lambda)$, which yields a cost $g(\lambda)$. The dual function therefore represents the optimal cost to the firm, as a function of the constraint price vector λ. The optimal dual value, $d^\star$, is the optimal cost to the enterprise under the least favorable set of prices.

Using this interpretation we can paraphrase weak duality as follows: the optimal cost to the firm in the second scenario (in which constraint violations can be bought and sold) is less than or equal to the cost in the original situation (which has constraints that cannot be violated), even with the most unfavorable prices. This is obvious: If $x^\star$ is optimal in the first scenario, then the operating cost of $x^\star$ in the second scenario will be lower than $f_0(\tilde{x})$, since some income can be derived from the constraints that are not tight.

The optimal duality gap is then the minimum possible advantage to the enterprise of being allowed to pay for constraint violations (and receive payments for nontight constraints).

Now suppose strong duality holds, and the dual optimum is attained. We can interpret a dual optimal $\lambda^\star$ as a set of prices for which there is no advantage to the firm in being allowed to pay for constraint violations (or receive payments for nontight constraints).

4.4 Optimality conditions

Again we do not assume the problem (4.1) is convex.

4.4.1 Certificate of suboptimality and stopping criteria

If we can find a dual feasible (λ, ν), we establish a lower bound on the optimal value of the primal problem: $p^\star \geq g(\lambda, \nu)$. Thus a dual feasible point (λ, ν) provides a *proof* or *certificate* that $p^\star \geq g(\lambda, \nu)$. Strong duality means there exist arbitrarily good certificates.

This allows us to bound how suboptimal a given feasible point is, without knowing the exact value of $p^\star$. Indeed if x is primal feasible and (λ, ν) is dual feasible, then

$$f_0(x) - p^\star \leq f_0(x) - g(\lambda, \nu).$$

In particular, this establishes that x is ϵ-suboptimal, with $\epsilon = f_0(x) - g(\lambda, \nu)$. (It also establishes that (λ, ν) is ϵ-suboptimal for the dual problem.)

We refer to the gap between primal and dual objectives,

$$f_0(x) - g(\lambda, \nu),$$

as the *duality gap* associated with the primal feasible point x and dual feasible point (λ, ν). A primal dual feasible pair x, (λ, ν) localizes the optimal value of the primal (and dual) problems in an interval:

$$p^\star \in [g(\lambda, \nu), f_0(x)], \quad d^\star \in [g(\lambda, \nu), f_0(x)],$$

the width of which is the duality gap.

If the duality gap of the primal dual feasible pair x, (λ, ν) is zero, *i.e.*, $f_0(x) = g(\lambda, \nu)$, then x is primal optimal and (λ, ν) is dual optimal. We can think of (λ, ν) as a certificate that proves x is optimal (and, similarly, we can think of x as a certificate that proves (λ, ν) is dual optimal).

These observations can be used in optimization algorithms to provide non-heuristic stopping criteria. Suppose an algorithm produces a sequence of primal feasible $x^{(k)}$ and dual feasible $\lambda^{(k)}$, $\nu^{(k)}$, and $\epsilon > 0$ is a given required absolute accuracy. Then the stopping criterion (*i.e.*, the condition for terminating the algorithm)

$$f_0(x^{(k)}) - g(\lambda^{(k)}, \nu^{(k)}) \leq \epsilon$$

guarantees that when the algorithm terminates the absolute error $x^{(k)}$ is ϵ-suboptimal. Indeed, $\lambda^{(k)}$, $\nu^{(k)}$ is a certificate that proves it. (Of course strong duality must hold if this method is to work for arbitrarily small tolerances ϵ.)

A similar condition can be used to guarantee a given relative accuracy. For simplicity we assume that $p^* > 0$. If $\eta > 0$ is the required relative accuracy, the stopping criterion

$$g(\lambda^{(k)}, \nu^{(k)}) > 0, \quad \frac{f_0(x^{(k)}) - g(\lambda^{(k)}, \nu^{(k)})}{g(\lambda^{(k)}, \nu^{(k)})} \leq \eta$$

guarantees that the relative error

$$\frac{f_0(x^{(k)}) - p^\star}{p^\star}$$

is less than η when the algorithm terminates.

4.4.2 Complementary slackness

Suppose that the primal and dual optimal values are attained and equal (so, in particular, strong duality holds). Let $x^\star$ be a primal optimal and $(\lambda^\star, \nu^\star)$ be a dual optimal point. This means that

$$\begin{aligned} f_0(x^\star) &= g(\lambda^\star, \nu^\star) \\ &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^\star f_i(x) + \sum_{i=1}^p \nu_i^\star h_i(x) \right) \\ &\leq f_0(x^\star) + \sum_{i=1}^m \lambda_i^\star f_i(x^\star) + \sum_{i=1}^p \nu_i^\star h_i(x^\star). \end{aligned} \tag{4.40}$$

(The first line states that the optimal duality gap is zero; the second line is the definition of the dual function; and the third follows since the infimum of the Lagrangian over x is less than or equal to its value at $x = x^\star$.) Therefore, and since $h_i(x^\star) = 0$, we have

$$\sum_{i=1}^m \lambda_i^\star f_i(x^\star) \geq 0. \tag{4.41}$$

Since $(\lambda^\star, \nu^\star)$ is dual feasible, $\lambda_i^\star \geq 0$; since $x^\star$ is primal feasible, $f_i(x^\star) \leq 0$; therefore we have $\lambda_i^\star f_i(x^\star) \leq 0$. Combined with (4.41) we conclude that

$$\lambda_i^\star f_i(x^\star) = 0, \quad i = 1, \ldots, m. \tag{4.42}$$

This condition is known as *complementary slackness*; it holds for any primal optimal $x^\star$ and any dual optimal $(\lambda^\star, \nu^\star)$ (when strong duality holds). We can express the complementary slackness condition as

$$\lambda_i^\star > 0 \implies f_i(x^\star) = 0,$$

or, equivalently,

$$f_i(x^\star) < 0 \implies \lambda_i^\star = 0.$$

Roughly speaking, this means the ith optimal Lagrange multiplier is zero unless the ith constraint is active at the optimum.

4.4.3 KKT optimality conditions

We now assume that functions $f_0, \ldots, f_m, h_1, \ldots, h_p$ are differentiable (and, by our convention, have open domains). As above, let $x^\star$ and $(\lambda^\star, \nu^\star)$ be any primal and dual optimal points with zero duality gap.

Combining complementary slackness (*i.e.*, (4.42)) with the inequality (4.40) yields

$$f_0(x^\star) = g(\lambda^\star, \nu^\star) = f_0(x^\star) + \sum_{i=1}^m \lambda_i^\star f_i(x^\star) + \sum_{i=1}^p \nu_i^\star h_i(x^\star)$$

which shows that $x^\star$ minimizes $L(x, \lambda^\star, \nu^\star)$ over x. ($L(x, \lambda^\star, \nu^\star)$ can have other minimizers; $x^\star$ is simply *a* minimizer.) Since $L(x, \lambda^\star, \nu^\star)$ is differentiable it follows that its gradient must vanish at $x^\star$, *i.e.*,

$$\nabla f_0(x^\star) + \sum_{i=1}^m \lambda_i^\star \nabla f_i(x^\star) + \sum_{i=1}^p \nu_i^\star \nabla h_i(x^\star) = 0.$$

Thus we have

$$\begin{array}{rcl} h_i(x^\star) = 0, \quad f_i(x^\star) & \leq & 0 \\ \lambda_i^\star & \geq & 0 \\ \nabla f_0(x^\star) + \sum_{i=1}^m \lambda_i^\star \nabla f_i(x^\star) + \sum_{i=1}^p \nu_i^\star \nabla h_i(x^\star) & = & 0 \\ \lambda_i^\star f_i(x^\star) & = & 0, \end{array} \tag{4.43}$$

which are called the Karush-Kuhn-Tucker (KKT) conditions. We have shown that any primal optimal $x^\star$ and any dual optimal $(\lambda^\star, \nu^\star)$, for a problem in which strong duality obtains, must satisfy the KKT conditions (4.43).

The converse holds, if the (primal) problem is convex. In other words, if $\tilde{x}$, $\tilde{\lambda}$, $\tilde{\nu}$ are any points that satisfy the KKT conditions

$$\begin{array}{rcl} h_i(\tilde{x}) = 0, \quad f_i(\tilde{x}) & \leq & 0 \\ \tilde{\lambda}_i & \geq & 0 \\ \nabla f_0(\tilde{x}) + \sum_{i=1}^m \tilde{\lambda}_i \nabla f_i(\tilde{x}) + \sum_{i=1}^p \tilde{\nu}_i \nabla h_i(\tilde{x}) & = & 0 \\ \tilde{\lambda}_i f_i(\tilde{x}) & = & 0, \end{array}$$

and f_i are convex and h_i are affine, then $\tilde{x}$ and $(\tilde{\lambda}, \tilde{\nu})$ are primal and dual optimal, with zero duality gap.

To see this, note that the first condition states that $\tilde{x}$ is primal feasible. Since $\tilde{\lambda}_i \geq 0$, $L(x, \tilde{\lambda}, \tilde{\nu})$ is convex in x, and the third KKT condition states that its gradient with respect to x vanishes at $x = \tilde{x}$, it follows that $\tilde{x}$ minimizes $L(x, \tilde{\lambda}, \tilde{\nu})$ over x. From this it follows that

$$\begin{aligned} g(\tilde{\lambda}, \tilde{\nu}) &= L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) \\ &= f_0(\tilde{x}) + \sum_{i=1}^m \tilde{\lambda}_i f_i(\tilde{x}) + \sum_{i=1}^p \tilde{\nu}_i h_i(\tilde{x}) \\ &= f_0(\tilde{x}) \end{aligned}$$

where in the last line we use the fact that $h_i(\tilde{x}) = 0$ and $\tilde{\lambda}_i f_i(\tilde{x}) = 0$. This shows that $\tilde{x}$ and $(\tilde{\lambda}, \tilde{\nu})$ have zero duality gap; hence they are primal and dual optimal.

In summary:

- For *any* optimization problem, primal and dual optimal points with zero duality gap (if any exist) must satisfy the KKT conditions.
- For any *convex* optimization problem, any points that satisfy the KKT conditions are primal and dual optimal, and have zero duality gap.

Suppose, for example, that a convex optimization problem satisfies Slater's condition, which implies that the optimal duality gap is zero, and the dual optimum is attained. Then the KKT conditions are necessary and sufficient conditions for optimality, *i.e.*, x is optimal if and only if there are (λ, ν) that, together with x, satisfy the KKT conditions.

The KKT conditions play an important role in optimization. In some very special cases it is possible to solve the KKT conditions analytically. More generally, many algorithms for convex optimization are conceived as, or can be interpreted as, methods for solving the KKT equations.

Example. Consider the convex optimization problem

$$\begin{array}{ll} \text{minimize} & -\sum_{i=1}^n \log(\alpha_i + x_i) \\ \text{subject to} & x \succeq 0 \\ & \mathbf{1}^T x = 1, \end{array}$$

where $\alpha_i > 0$, $i = 1, \ldots, n$, and $D = \{x \mid x + \alpha \succ 0\}$. Introducing Lagrange multipliers λ for the inequality constraints $x \succeq 0$, and a multiplier ν for the equality constraint $\mathbf{1}^T x = 1$, we obtain the KKT conditions

$$\begin{aligned} \mathbf{1}^T x = 1, \quad x &\succeq 0 \\ \lambda &\succeq 0 \\ -\frac{1}{\alpha_i + x_i} - \lambda_i + \nu &= 0, \quad i = 1, \ldots, n \\ \lambda_i x_i &= 0, \quad i = 1, \ldots, n. \end{aligned}$$

We have to solve these equations for x, λ, and ν. We start by noting that λ acts as a slack variable in the third equation, so it can be eliminated, and the last three conditions reduce to

$$\nu \geq \frac{1}{\alpha_i + x_i}, \quad x_i \left(\nu - \frac{1}{\alpha_i + x_i} \right) = 0, \quad i = 1, \ldots, n.$$

Combining these conditions with $x \succeq 0$, we can determine x as a function of ν:

$$\begin{array}{ll} x_i = \dfrac{1}{\nu} - \alpha_i & \text{if } \nu < 1/\alpha_i \\ x_i = 0 & \text{if } \nu \geq 1/\alpha_i, \end{array}$$

i.e., $x_i = \max\left\{0, \frac{1}{\nu} - \alpha_i\right\}$. We now have one remaining variable, ν, which can be determined from

$$\sum_{i=1}^n \max\{0, 1/\nu - \alpha_i\} = 1.$$

The left hand side is a piecewise-linear increasing function of $1/\nu$, with breakpoints at α_i, so the equation has a unique solution which is readily determined.

4.4.4 Solving the primal problem via the dual

We mentioned at the beginning of §4.4.3 that if strong duality holds and a dual optimal solution $(\lambda^\star, \nu^\star)$ exists, then any primal optimal point is also a minimizer of $L(x, \lambda^\star, \nu^\star)$. This fact sometimes allows us to compute the primal optimal solution from the dual optimal solution.

More precisely, suppose we have strong duality and an optimal $(\lambda^\star, \nu^\star)$ is known. Suppose that the minimizer of $L(x, \lambda^\star, \nu^\star)$, *i.e.*, the solution of

$$\text{minimize } f_0(x) + \sum_{i=1}^m \lambda_i^\star f_i(x) + \sum_{i=1}^p \nu_i^\star h_i(x) \tag{4.44}$$

is unique. (For a convex problem this occurs, for example, if $L(x, \lambda^\star, \nu^\star)$ is a strictly convex function of x.) Then if the solution of (4.44) is primal feasible, it must be primal optimal; if it is not primal feasible, then no primal optimal point can exist, *i.e.*, we can conclude that the primal optimum is not attained.

This observation is interesting when the dual problem is easier to solve than the primal problem, for example, because it can be solved analytically.

> **Example.** *Entropy maximization.* As an example, we consider the entropy maximization problem (4.11). Here one might prefer to solve the dual problem (4.29) because it is unconstrained.
>
> We assume Slater's condition holds, *i.e.*, there exists an $x \succ 0$ with $Ax \preceq b$ and $\mathbf{1}^T x = 1$, so strong duality holds and an optimal solution $(\lambda^\star, \nu^\star)$ exists.
>
> Suppose we have solved the dual problem. The Lagrangian at $(\lambda^\star, \nu^\star)$ is
>
> $$L(x, \lambda^\star, \nu^\star) = \sum_{i=1}^n x_i \log x_i + \lambda^\star(Ax - b) + \nu^\star(\mathbf{1}^T x - 1)$$
>
> which is strictly convex on D and bounded below, so it has a unique solution $x^\star$, given by
>
> $$x_i^\star = \exp(a_i^T \lambda^\star + \nu^\star - 1), \quad i = 1, \ldots, n.$$
>
> If $x^\star$ is primal feasible, it must be the optimal solution of the primal problem (4.11). If $x^\star$ is not primal feasible, then we can conclude that the primal optimum is not attained.

4.5 Sensitivity analysis

4.5.1 The perturbed problem

The optimal dual variables for a convex optimization problem give very useful information about the sensitivity of the optimal value with respect to perturbations of the constraints.

We consider the following perturbed version of the original convex problem (4.24):

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq u_i, \quad i = 1, \ldots, m, \\ & Ax - b = v, \end{array} \tag{4.45}$$

where $f_0, \ldots, f_m$ are convex. This problem coincides with the original problem (4.24) when $u = 0$, $v = 0$. When u_i is positive it means that we have relaxed the ith inequality constraint; when u_i is negative, it means that we have tightened the constraint. A nonzero value for v_i means that we translate the hyperplane $a_i^T x = b_i$ along the direction of its normal vector a_i. Thus the perturbed problem (4.45) results from the original problem (4.24) by tightening or relaxing each inequality constraint by u_i, and translating the affine set defined by the equality constraints.

We define $p^\star(u, v)$ as the optimal value of the perturbed problem (4.45):

$$p^\star(u, v) = \inf_x \{\ f_0(x) \ \mid f_i(x) \leq u_i, \ i = 1, \ldots, m, \ Ax - b = v, \ \}.$$

This is a convex function of u and v. We can have $p^\star(u, v) = \infty$, which corresponds to perturbations of the constraints that are infeasible. Note that $p^\star(0, 0) = p^\star$, the optimal value of the unperturbed problem (4.24). (We hope this slight abuse of notation will cause no confusion.)

4.5.2 A global inequality

Now suppose we have strong duality and that the dual optimum is attained (as would be the case if Slater's condition is satisfied). Let $(\lambda^\star, \nu^\star)$ be optimal for the dual (4.14) of the unperturbed problem. Then for all u and v we have

$$p^\star(u, v) \geq p^\star - \lambda^{\star T} u - \nu^{\star T} v. \tag{4.46}$$

Proof. We first observe that for any x

$$p^\star = g(\lambda^\star, \nu^\star) \leq f_0(x) + \sum_{i=1}^m \lambda_i^\star f_i(x) + \nu^{\star T}(Ax - b).$$

This follows from strong duality and the definition of $g(\lambda^\star, \nu^\star)$. If in addition x is feasible for the perturbed problem, *i.e.*, $f_i(x) \leq u_i$ for $i = 1, \ldots, m$, and $Ax - b = v$, we have

$$p^\star \leq f_0(x) + \sum_{i=1}^m \lambda_i^\star f_i(x) + \nu^{\star T}(Ax - b) \leq f_0(x) + \lambda^{\star T} u + \nu^{\star T} v$$

since the components of $\lambda^\star$ are nonnegative. The inequality (4.46) follows by minimizing both sides over x, subject to the perturbed constraints.

Various sensitivity interpretations of the optimal Lagrange variables follow directly from the inequality (4.46). Some of the conclusions are:

- If $\lambda_i^\star$ is large and we tighten the ith constraint (*i.e.*, choose $u_i < 0$), then the optimal value $p^\star(u,v)$ is guaranteed to increase greatly.
- If $\nu_i^\star$ is large and positive and we take $v_i < 0$, or if $\nu_i^\star$ is large and negative and we take $v_i > 0$, then the optimal value $p^\star(u,v)$ is guaranteed to increase greatly.
- If $\lambda_i^\star$ is small, and we loosen the ith constraint ($u_i > 0$), then the optimal value $p^\star$ will not decrease too much.
- If $\nu_i^\star$ is small and positive, and $v_i > 0$, or if $\nu_i^\star$ is small and negative and $v_i < 0$, then the optimal value $p^\star$ will not decrease too much.

The inequality (4.46), and the conclusions listed above, give a *lower bound* on the perturbed optimal value, but no upper bound. Thus the results are *not* symmetric. For example, suppose that $\lambda_i^\star$ is large, and we loosen the ith constraint a bit (*i.e.*, take $u_i > 0$). In this case the inequality (4.46) isn't useful; it does not, for example, imply that the optimal value will decrease considerably.

4.5.3 Local sensitivity analysis

Suppose now that $p^\star(u,v)$ is differentiable at $u = 0$, $v = 0$. Then the optimal dual variables $\lambda^\star$, $\nu^\star$ are related to the gradient at $p^\star(u,v)$ at $u = 0$, $v = 0$:

$$\lambda_i^\star = -\frac{\partial p^\star(0,0)}{\partial u_i}, \quad \nu_i^\star = -\frac{\partial p^\star(0,0)}{\partial v_i}. \tag{4.47}$$

Proof. We can derive this property from (4.46). If $p^\star(u,v)$ is differentiable, then, by definition of the gradient,

$$p^\star(u,v) = p^\star + \nabla_u p^\star(0,0)^T u + \nabla_v p^\star(0,0)^T v + o(u) + o(v).$$

Combining this with (4.46), we obtain

$$\begin{aligned} p^\star(u,v) &= p^\star + (\nabla_u p^\star(0))^T u + (\nabla_v p^\star(0))^T v + o(u) + o(v) \\ &\geq p^\star - \lambda^{\star T} u - \nu^{\star T} v \\ p^\star(-u,-v) &= p^\star - (\nabla_u p^\star(0,0))^T u - (\nabla_v p^\star(0,0))^T v + o(u) + o(v) \\ &\geq p^\star + \lambda^{\star T} u + \nu^{\star T} v \end{aligned}$$

and therefore

$$\begin{aligned} -\nabla_u p^\star(0,0)^T u - \nabla_v p^\star(0,0)^T v &\leq \lambda^{\star T} u + \nu^{\star T} v + o(u) + o(v) \\ &\leq -\nabla_u p^\star(0,0)^T u - \nabla_v p^\star(0,0)^T v, \end{aligned}$$

for all u and v. This is only possible if $\lambda^\star = -\nabla_u p^\star(0,0)$ and $\nu^\star = -\nabla_v p^\star(0,0)$.

Thus, when $p^\star(u,v)$ is differentiable at $u=0$, $v=0$, the optimal Lagrange multipliers are exactly the local sensitivities of the optimal value with respect to constraint perturbations. In contrast to the nondifferentiable case, this interpretation is symmetric: tightening the ith inequality constraint a small amount (*i.e.*, taking u_i small and negative) yields an increase in $p^\star$ of approximately $-\lambda_i^\star u_i$; loosening the ith constraint a small amount (*i.e.*, taking u_i small and positive) yields a decrease in $p^\star$ of approximately $\lambda_i^\star u_i$.

The result (4.47) allows us to assign a numerical value to how 'active' a constraint is at the optimum $x^\star$. If $f_i(x^\star) < 0$, then the constraint is inactive, and it follows that the constraint can be tightened or loosened a small amount without affecting the optimal value. By complementary slackness, the associated optimal Lagrange multiplier must be zero. But now suppose that $f_i(x^\star) = 0$, *i.e.*, the ith constraint is active at the optimum. The ith optimal Lagrange multiplier tells us how active the constraint is: If $\lambda_i^\star$ is small, it means that the constraint can be loosened or tightened a bit without much affect on the optimal value; if $\lambda_i^\star$ is large, it means that if the constraint is loosened or tightened a bit, the effect on the optimal value will be great.

4.6 More examples

In this section we show by example that simple equivalent reformulations of a problem can lead to very different dual problems. We will consider the following types of reformulations:

- Introducing new variables and associated equality constraints.
- Replacing the objective with an increasing function of the original objective.
- Making explicit constraints implicit, *i.e.*, incorporating them into the domain of the objective.

4.6.1 Introducing new variables and equality constraints

Consider an unconstrained problem of the form

$$\text{minimize } f_0(Ax+b). \tag{4.48}$$

Its Lagrange dual function is the constant $p^\star$. So while we do have strong duality, *i.e.*, $p^\star = d^\star$, the Lagrangian dual is neither useful nor interesting.

Now let us reformulate the problem (4.48) as

$$\begin{array}{ll} \text{minimize} & f_0(y) \\ \text{subject to} & Ax+b=y. \end{array} \tag{4.49}$$

Here we have introduced new variables y, as well as new equality constraints $Ax+b=y$. The problems (4.48) and (4.49) are clearly equivalent.

Now let us find the Lagrange dual of the reformulated problem. The Lagrangian is $L(x,\nu) = f_0(y) + \nu^T(Ax+b-y)$. To find the dual function we minimize L over x and y. Minimizing over x we find that $g(\nu) = -\infty$ unless $A^T\nu = 0$, in which case we are left with

$$g(\nu) = b^T\nu + \inf_y (f_0(y) - \nu^T y) = b^T\nu - f_0^\star(\nu),$$

where $f_0^\star$ is the conjugate of f_0.

The dual problem of (4.49) can therefore be expressed as

$$\begin{array}{ll} \text{maximize} & b^T\nu - f_0^\star(\nu) \\ \text{subject to} & A^T\nu = 0. \end{array} \tag{4.50}$$

Thus, the dual of the reformulated problem (4.49) is considerably more useful than the dual of the original problem (4.48).

Example. *Unconstrained geometric programming.* Consider the unconstrained geometric program

$$\text{minimize } \log \sum_{i=1}^m \exp(a_i^T x + b_i).$$

We first reformulate it by introducing new variables and equality constraints:

$$\begin{array}{ll} \text{minimize} & \log \sum_{i=1}^m \exp y_i \\ \text{subject to} & Ax + b = y. \end{array}$$

In chapter 2 we found the conjugate of the objective to be

$$f^\star(\nu) = \begin{cases} \sum_i \nu_i \log \nu_i & \text{if } \nu \succeq 0,\ \mathbf{1}^T\nu = 1 \\ +\infty & \text{otherwise.} \end{cases}$$

The dual of the reformulated problem can therefore be expressed as

$$\begin{array}{ll} \text{maximize} & b^T\nu - \sum_{i=1}^m \nu_i \log \nu_i \\ \text{subject to} & \mathbf{1}^T\nu = 1 \\ & A^T\nu = 0 \\ & \nu \succeq 0, \end{array} \tag{4.51}$$

which is a maximum entropy problem.

Example. *Minimum norm problems.* We consider the unconstrained minimum norm problem

$$\text{minimize} \quad \|Ax - b\|, \tag{4.52}$$

where $\|\cdot\|$ is some (not necessarily Euclidean) norm. Here too the Lagrange dual function is constant, equal to the optimal value of (4.52), and therefore not useful.

Once again we reformulate the problem as

$$\begin{array}{ll} \text{minimize} & \|y\| \\ \text{subject to} & y = Ax - b. \end{array}$$

The Lagrange dual problem is, following (4.50),

$$\begin{array}{ll} \text{maximize} & b^T\nu \\ \text{subject to} & \|\nu\|_* \le 1 \\ & A^T\nu = 0, \end{array} \tag{4.53}$$

where we use the fact the conjugate of a norm is the indicator function of the dual norm unit ball; see §2.3).

The idea of introducing new equality constraints can be applied to the constraint functions as well. Consider, for example, the problem

$$\begin{array}{ll} \text{minimize} & f_0(A_0x+b_0) \\ \text{subject to} & f_i(A_ix+b_i) \leq 0, \quad i=1,\ldots,m, \end{array} \tag{4.54}$$

where $A_i \in \mathbf{R}^{k_i\times n}$ and $f_i : \mathbf{R}^{k_i} \to \mathbf{R}$ are convex. (For simplicity we do not include equality constraints here.)

We introduce a new variable $y_i \in \mathbf{R}^{k_i}$, for $i = 0,\ldots,m$, and reformulate the problem as

$$\begin{array}{ll} \text{minimize} & f_0(y_0) \\ \text{subject to} & f_i(y_i) \leq 0, \quad i=1,\ldots,m, \\ & A_ix+b_i = y_i, \quad i=0,\ldots,m. \end{array} \tag{4.55}$$

The Lagrangian for this problem is

$$L(x,\lambda,\nu_0,\ldots,\nu_m) = f_0(y_0) + \sum_{i=1}^m \lambda_i f_i(y_i) + \sum_{i=0}^m \nu_i^T(A_ix+b_i-y_i).$$

To find the dual function we minimize over x and y_i. The minimum over x is $-\infty$ unless

$$\sum_{i=0}^m A_i^T\nu_i = 0,$$

in which case we have

$$\begin{aligned} g(\lambda,\nu_0,\ldots,\nu_m) &= \sum_{i=0}^m \nu_i^Tb_i + \inf_{y_0,\ldots,y_m}\left(f_0(y_0) + \sum_{i=1}^m \lambda_i f_i(y_i) - \sum_{i=0}^m \nu_i^Ty_i\right) \\ &= \sum_{i=0}^m \nu_i^Tb_i + \inf_{y_0}\left(f_0(y_0) - \nu_0^Ty_0\right) + \sum_{i=1}^m \inf_{y_i}\left(\lambda_i f_i(y_i) - \nu_i^Ty_i\right) \\ &= -f_0^\star(\nu_0) + \sum_{i=0}^m \left(\nu_i^Tb_i - \lambda_i f_i^\star(\nu_i/\lambda_i)\right). \end{aligned}$$

(The last expression involves the perspective of the conjugate function, and is therefore concave in the dual variables. Values $\lambda_i = 0$ are allowed; the correct interpretation is that $\lambda_i f_i^\star(\nu_i/\lambda_i) = 0$ if $\lambda_i = 0$ and $\nu_i = 0$, and $\lambda_i f_i^\star(\nu_i/\lambda_i) = -\infty$ if $\lambda_i = 0$ and $\nu_i \neq 0$.)

Therefore we can express the dual of the problem (4.55) as

$$\begin{array}{ll} \text{maximize} & \nu_0^Tb_0 - f_0^\star(\nu_0) + \sum_{i=1}^m \left(\nu_i^Tb_i - \lambda_i f_i^\star(\nu_i/\lambda_i)\right) \\ \text{subject to} & \lambda \succeq 0, \\ & \sum_{i=0}^m A_i^T\nu_i = 0. \end{array} \tag{4.56}$$

Example. *Constrained geometric program.* The constrained geometric program

$$\begin{array}{ll} \text{minimize} & \log \sum_{j=1}^{k_0} e^{a_{0j}^T x + b_{0j}} \\ \text{subject to} & \log \sum_{j=1}^{k_i} e^{a_{ij}^T x + b_{ij}} \leq 0, \quad i = 1, \dots, m \end{array}$$

is of the form (4.54) with $f_i : \mathbf{R}^{k_i} \to \mathbf{R}$ given by $f_i(y) = \log \sum_{j=1}^{k_i} e^{y_j}$. The conjugate of this function is

$$f_i^\star(\nu) = \begin{cases} \sum_{j=1}^{k_i} \nu_i \log \nu_i & \text{if } \nu_i \succeq 0, \ \ \mathbf{1}^T \nu_i = 1 \\ +\infty & \text{otherwise.} \end{cases}$$

Using (4.56) we can immediately write down the dual problem as

$$\begin{array}{ll} \text{maximize} & b_0^T \nu_0 - \sum_{j=1}^{k_0} \nu_{0j} \log \nu_{0j} + \sum_{i=1}^{p} \left(b_i^T \nu_i - \sum_{j=1}^{k_i} \nu_{ij} \log(\nu_{ij}/\lambda_i) \right) \\ \text{subject to} & \nu_0 \succeq 0, \ \ \mathbf{1}^T \nu_0 = 1 \\ & \nu_i \succeq 0, \ \ \mathbf{1}^T \nu_i = \lambda_i, \ \ i = 1, \dots, m \\ & \lambda_i \geq 0, \ \ i = 1, \dots, m \\ & \sum_{i=0}^{m} A_i^T \nu_i = 0, \end{array}$$

which further simplifies to

$$\begin{array}{ll} \text{maximize} & b_0^T \nu_0 - \sum_{j=1}^{k_0} \nu_{0j} \log \nu_{0j} + \sum_{i=1}^{p} \left(b_i^T \nu_i - \sum_{j=1}^{k_i} \nu_{ij} \log \dfrac{\nu_{ij}}{\mathbf{1}^T \nu_i} \right) \\ \text{subject to} & \nu_i \succeq 0, \ \ i = 0, \dots, m \\ & \mathbf{1}^T \nu_0 = 1, \\ & \sum_{i=0}^{m} A_i^T \nu_i = 0. \end{array}$$

4.6.2 Transforming the objective

If we replace the objective f_0 by an increasing function of f_0, the resulting problem is clearly equivalent. The dual of this equivalent problem, however, can be very different from the dual of the original problem.

Example. We consider again the minimum norm problem

$$\text{minimize} \quad \|Ax - b\|,$$

where $\|\cdot\|$ is some norm. We reformulate this problem as

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|y\|^2 \\ \text{subject to} & Ax - b = y. \end{array}$$

Here we have introduced new variables, and replaced the objective by half its square. Evidently it is equivalent to the original problem.

The dual of the reformulated problem is

$$\begin{array}{ll} \text{maximize} & -\frac{1}{2}\|\nu\|_*^2 + b^T\nu \\ \text{subject to} & A^T\nu = 0, \end{array} \tag{4.57}$$

where we used the fact that the conjugate of $(1/2)\|\cdot\|^2$ is $(1/2)\|\cdot\|_*^2$.

Note that this dual problem is not the same as the dual problem (4.53) derived earlier.

4.6.3 Implicit constraints

The next simple reformulation we study is to include some of the constraints in the objective function, by modifying the objective function to be infinite when the constraint is violated.

Example. *Linear program with box constraints.* We consider the linear program

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b \\ & l \preceq x \preceq u \end{array} \tag{4.58}$$

where $A \in \mathbf{R}^{p\times n}$ and $l \prec u$. The constraints $l \preceq x \preceq u$ are sometimes called *box constraints* or *variable bounds.*

We can, of course, derive the dual of this linear program. The dual will have a Lagrange multiplier ν associated with the equality constraint, λ_1 associated with the inequality constraint $x \preceq u$, and λ_2 associated with the inequality constraint $l \preceq x$. The dual is

$$\begin{array}{ll} \text{maximize} & -b^T\nu - \lambda_1^T u + \lambda_2^T l \\ \text{subject to} & A^T\nu + \lambda_1 - \lambda_2 + c = 0 \\ & \lambda_1 \succeq 0, \ \lambda_2 \succeq 0. \end{array} \tag{4.59}$$

Instead, let us first reformulate the problem (4.58) as

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & Ax = b, \end{array}$$

where we define

$$f_0(x) = \begin{cases} c^T x & \text{if } l \preceq x \preceq u, \\ \infty & \text{otherwise.} \end{cases}$$

This problem is clearly equivalent to (4.58); we have merely made the explicit box constraints implicit.

The dual function for the problem (4.58) is

$$\begin{aligned} g(\nu) &= \inf_{l\preceq x\preceq u} \left(c^T x + \nu^T(Ax - b)\right) \\ &= -b^T\nu - u^T(A^T\nu + c)^- + l^T(A^T\nu + c)^+ \end{aligned}$$

where $y_i^+ = \max\{y_i, 0\}$, $y_i^- = \max\{-y_i, 0\}$. So here we are able to derive an analytical formula for g, which is a concave piecewise linear function.

The dual problem is the unconstrained problem

$$\text{maximize} \ -b^T\nu - u^T(A^T\nu + c)^- + l^T(A^T\nu + c)^+, \tag{4.60}$$

which has a quite different form from the dual of the original problem.

(The problems (4.59) and (4.60) are closely related, in fact, equivalent.

4.7 Systems of inequalities

4.7.1 Weak alternatives via the dual function

In this section we apply Lagrange duality theory to the problem of determining feasibility of systems of inequalities and equalities,

$$f_i(x) \leq 0, \ i = 1, \ldots, m, \quad h_i(x) = 0, \ i = 1, \ldots, p. \tag{4.61}$$

We assume the domain of the inequality system (4.61),

$$D = \cap_{i=0}^m \mathbf{dom}\, f_i \bigcap \cap_{i=1}^p \mathbf{dom}\, h_i,$$

is nonempty.

We can think of (4.61) as our standard problem (4.1), with objective $f_0 = 0$, *i.e.*,

$$\begin{array}{ll} \text{minimize} & 0 \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p. \end{array} \tag{4.62}$$

This problem has optimal value

$$p^\star = \begin{cases} 0 & \text{if (4.61) is feasible} \\ \infty & \text{if (4.61) is infeasible.} \end{cases} \tag{4.63}$$

The dual function

We associate with the inequality system (4.61) the dual function

$$g(\lambda, \nu) = \inf_{x \in D} \left(\sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x) \right),$$

which is the same as the dual function for the optimization problem (4.62). Since $f_0 = 0$, the dual function is nonnegative homogeneous in (λ, ν): for $\alpha \geq 0$, $g(\alpha\lambda, \alpha\nu) = \alpha g(\lambda, \nu)$. The dual problem associated with (4.62) is to maximize $g(\lambda, \nu)$ subject to $\lambda \succeq 0$. Since g is homogenous here, the optimal value of this dual problem is given by

$$d^\star = \begin{cases} \infty & \text{if } \lambda \succeq 0, \ g(\lambda, \nu) > 0 \text{ is feasible} \\ 0 & \text{if } \lambda \succeq 0, \ g(\lambda, \nu) > 0 \text{ is infeasible.} \end{cases} \tag{4.64}$$

Weak duality tells us that $d^\star \le p^\star$. Combining this fact with (4.63) and (4.64) yields the following: if the inequality system

$$\lambda \succeq 0, \quad g(\lambda,\nu) > 0 \tag{4.65}$$

is feasible (which means $d^\star = \infty$), then the inequality system (4.61) is infeasible (since we then have $p^\star = \infty$). Indeed, we can interpret any solution (λ,ν) of the inequalities (4.65) as a *proof* or *certificate* of infeasibility of the system (4.61).

This relation is symmetric: if the original inequality system (4.61) is feasible, then the inequality system (4.65) must be infeasible. We can interpret an x which satisfies (4.61) as a certificate establishing infeasibility of the inequality system (4.65).

Two systems of inequalities (and equalities) are called *weak alternatives* if at most one of the two is feasible. Thus, the systems (4.61) and (4.65) are weak alternatives. This is true whether or not the inequalities (4.61) are convex (*i.e.*, f_i convex, h_i affine); moreover, the alternative inequality system (4.65) is always convex (*i.e.*, g is concave and the constraints $\lambda_i \ge 0$ are convex).

Strict inequalities

We can also study feasibility of the *strict* inequality system

$$f_i(x) < 0,\ i = 1,\dots,m, \quad h_i(x) = 0,\ i = 1,\dots,p. \tag{4.66}$$

With g as defined as for the nonstrict inequality system, we have the alternative inequality system

$$\lambda \succeq 0, \quad \lambda \ne 0, \quad g(\lambda,\nu) \ge 0. \tag{4.67}$$

We can show directly that (4.66) and (4.67) are weak alternatives. Suppose there exists an $\tilde{x}$ with $f_i(\tilde{x}) < 0$, $h_i(\tilde{x}) = 0$. Then for any $\lambda \succeq 0$, $\lambda \ne 0$, and ν,

$$\lambda_1 f_1(\tilde{x}) + \cdots + \lambda_m f_m(\tilde{x}) + \nu_1 h_1(\tilde{x}) + \cdots + \nu_p h_p(\tilde{x}) < 0.$$

It follows that

$$\begin{aligned} g(\lambda,\nu) &= \inf_{x\in D}\left(\sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x)\right) \\ &\le \sum_{i=1}^m \lambda_i f_i(\tilde{x}) + \sum_{i=1}^p \nu_i h_i(\tilde{x}) < 0. \end{aligned}$$

Therefore, feasibility of (4.66) implies that there does not exist (λ,ν) satisfying (4.67).

In other words, we can prove infeasibility of (4.66) by producing a solution of the system (4.67); we can prove infeasibility of (4.67) by producing a solution of the system (4.66).

4.7.2 Strong alternatives

When the original inequality system is convex, *i.e.*, f_i are convex and h_i are affine, and some type of constraint qualification holds, then the pairs of weak alternatives described above are *strong alternatives*, which means that *exactly one* of the two alternatives holds. In other words, each of the inequality systems is feasible if and only if the other is infeasible.

In this section we assume that f_i are convex and h_i are affine, so the inequality (4.61) can be expressed as

$$f_i(x) \leq 0,\ i = 1, \ldots, m, \quad Ax = b. \tag{4.68}$$

Strict inequalities

We first study the strict inequality system

$$f_i(x) < 0,\ i = 1, \ldots, m, \quad Ax = b, \tag{4.69}$$

and its alternative

$$\lambda \succeq 0, \quad \lambda \neq 0, \quad g(\lambda, \nu) \geq 0. \tag{4.70}$$

We need one technical condition: There exists an $x \in \mathbf{relint}\, D$ with $Ax = b$. In other words we not only assume that the linear equality constraints are consistent, but also that they have a solution in $\mathbf{relint}\, D$. (Very often $D = \mathbf{R}^n$, so the condition is satisfied if the equality constraints are consistent.) Under this condition, exactly one of the inequality systems (4.69) and (4.70) is feasible. In other words, the inequality systems (4.69) and (4.70) are strong alternatives.

We will establish this result by considering the related optimization problem

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & f_i(x) - t \leq 0,\ i = 1, \ldots, m \\ & Ax = b \end{array} \tag{4.71}$$

with variables x, t, and domain $D \times \mathbf{R}$. The optimal value $p^\star$ of this problem is negative if and only if there exists a solution to the strict inequality system (4.69).

The Lagrange dual function for the problem (4.71) is

$$\inf_{x \in D, t} \left(t + \sum_{i=1}^m \lambda_i (f_i(x) - t) + \nu^T (Ax - b) \right) = \left\{ \begin{array}{ll} g(\lambda, \nu) & \text{if } \mathbf{1}^T \lambda = 1 \\ -\infty & \text{otherwise.} \end{array} \right.$$

Therefore we can express the dual problem of (4.71) as

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T \lambda = 1. \end{array}$$

Now we observe that Slater's condition holds for the problem (4.71). By the hypothesis there exists an $\tilde{x} \in \mathbf{relint}\, D$ with $A\tilde{x} = b$. Choosing any $\tilde{t} > \max_i f_i(\tilde{x})$ yields a point $(\tilde{x}, \tilde{t})$ which is strictly feasible for (4.71). Therefore we have $d^\star = p^\star$, and the dual optimum $d^\star$ is attained. In other words, there exist $(\lambda^\star, \nu^\star)$ such that

$$g(\lambda^\star, \nu^\star) = p^\star, \quad \lambda^\star \succeq 0, \quad \mathbf{1}^T \lambda^\star = 1. \tag{4.72}$$

Now suppose that the strict inequality system (4.69) is infeasible, which means that $p^\star \geq 0$. Then $(\lambda^\star, \nu^\star)$ from (4.72) satisfy the alternate inequality system (4.70). Thus, the inequality systems (4.69) and (4.70) are strong alternatives; one is feasible if and only if the other is not.

Nonstrict inequalities

We now consider the nonstrict inequality system

$$f_i(x) \leq 0, \ i = 1, \ldots, m, \quad Ax = b, \tag{4.73}$$

and its alternative

$$\lambda \succeq 0, \quad A^T\nu = 0, \quad g(\lambda, \nu) > 0. \tag{4.74}$$

We will show these are strong alternatives, provided the following conditions hold: There exists an $x \in \mathbf{relint}\, D$ with $Ax = b$, and the optimal value $p^\star$ of (4.71) is attained. This holds, for example, if $D = \mathbf{R}^n$ and $\max_i f_i(x) \to \infty$ as $x \to \infty$.

With these assumptions we have, as in the strict case, that $p^\star = d^\star$, and that both the primal and dual optimal values are attained. Now suppose that the nonstrict inequality system (4.69) is infeasible, which means that $p^\star > 0$. (Here we use the assumption that the primal optimal value is attained.) Then $(\lambda^\star, \nu^\star)$ from (4.72) satisfy the alternate inequality system (4.74). Thus, the inequality systems (4.73) and (4.74) are strong alternatives; one is feasible if and only if the other is not.

Finally we mention that it is possible to work out alternatives for mixtures of strict and nonstrict inequalities, and work out more sophisticated (and complicated) qualifications that ensure the resulting alternatives are strong; see the notes and references.

4.7.3 Examples

Linear inequalities

Consider the system of linear inequalities $Ax \preceq b$. The dual function is

$$g(\lambda) = \inf_x \lambda^T(Ax - b) = \begin{cases} -b^T\lambda & \text{if } A^T\lambda = 0 \\ -\infty & \text{otherwise.} \end{cases}$$

The alternative inequality system is therefore

$$\lambda \succeq 0, \quad A^T\lambda = 0, \quad b^T\lambda < 0.$$

These are, in fact, strong alternatives. This follows since the optimum in the related problem (4.71) is achieved, unless it is unbounded below.

We now consider the system of strict linear inequalities $Ax \prec b$, which has the strong alternative system

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

In fact we have encountered (and proved) this result before, in §1.5.1; see (1.15) and (1.16).

Intersection of ellipsoids

We consider m ellipsoids, described as

$$\mathcal{E}_i = \{ \ x \mid f_i(x) \leq 0 \ \},$$

with $f_i(x) = x^T A_i x + 2b_i^T x + c_i$, $i = 1, \ldots, m,$, where $A_i = A_i^T \succ 0$. We ask when the intersection of these ellipsoids has nonempty interior. This is equivalent to feasibility of the set of strict quadratic inequalities

$$f_i(x) = x^T A_i x + 2b_i^T x + c_i < 0, \;\; i = 1, \ldots, m. \tag{4.75}$$

The dual function g is

$$\begin{aligned} g(\lambda) &= \inf_x \left(x^T A(\lambda) x + 2b(\lambda)^T x + c(\lambda) \right) \\ &= \begin{cases} -b(\lambda^T) A(\lambda)^\dagger b(\lambda) + c(\lambda) & \text{if } A(\lambda) \succeq 0 \text{ and } b(\lambda) \in \operatorname{range} A(\lambda) \\ -\infty & \text{otherwise} \end{cases} \end{aligned}$$

where

$$A(\lambda) = \sum_{i=1}^m \lambda_i A_i, \quad b(\lambda) = \sum_{i=1}^m \lambda_i b_i, \quad c(\lambda) = \sum_{i=1}^m \lambda_i c_i.$$

The strong alternative of the system (4.75) is

$$\lambda \succeq 0, \quad \lambda \neq 0, \quad A(\lambda) \succeq 0, \quad b(\lambda) \in \operatorname{range} A(\lambda), \quad -b(\lambda^T) A(\lambda)^\dagger b(\lambda) + c(\lambda) \geq 0.$$

We can simplify this system of inequalities, since $\lambda \succeq 0$, $\lambda \neq 0$ implies $A(\lambda) \succ 0$:

$$\lambda \succeq 0, \quad \lambda \neq 0, \quad -b(\lambda^T) A(\lambda)^{-1} b(\lambda) + c(\lambda) \geq 0. \tag{4.76}$$

This can also be expressed in terms of a linear matrix inequality as

$$\lambda \succeq 0, \quad \lambda \neq 0, \quad \begin{bmatrix} A(\lambda) & b(\lambda) \\ b(\lambda)^T & c(\lambda) \end{bmatrix} \succeq 0$$

We can give a simple geometrical interpretation of this pair of strong alternatives. For any nonzero $\lambda \succeq 0$, the (possibly empty) ellipsoid

$$\mathcal{E}_\lambda = \{ \; x \mid x^T A(\lambda) x + 2b(\lambda)^T x + c(\lambda) \leq 0 \; \}$$

contains $\mathcal{E}_1 \cap \cdots \cap \mathcal{E}_m$ (since $f_i \leq 0$ implies $\sum_{i=1}^m f_i \leq 0$). Now, $\mathcal{E}_\lambda$ has empty interior if and only if

$$\inf_x \left(x^T A(\lambda) x + 2b(\lambda)^T x + c(\lambda) \right) = -b(\lambda^T) A(\lambda)^{-1} b(\lambda) + c(\lambda) \geq 0.$$

Therefore the alternative system (4.76) means that $\mathcal{E}_\lambda$ has empty interior.

Weak duality is obvious: if (4.76) holds, then $\mathcal{E}_\lambda$ contains the intersection $\mathcal{E}_1 \cap \cdots \cap \mathcal{E}_m$, and has empty interior, so naturally the intersection has empty interior. The fact that these are strong alternatives states the (nonobvious) fact that if the intersection $\mathcal{E}_1 \cap \cdots \cap \mathcal{E}_m$ has empty interior, then we can construct an ellipsoid $\mathcal{E}_\lambda$ that contains the intersection and has empty interior.

4.8 Generalized inequalities

In this section we examine how Lagrange duality extends to a problem with generalized inequality constraints

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \preceq_{K_i} 0, \quad i = 1, \ldots, L \\ & h_i(x) = 0, \quad i = 1, \ldots, p, \end{array} \tag{4.77}$$

where $K_i \subseteq \mathbf{R}^{m_i}$ are cones that induce generalized inequalities. For now, we do not assume convexity of the problem (4.77).

We assume the domain of (4.77)

$$D = \cap_{i=0}^{m} \mathbf{dom}\, f_i \bigcap \cap_{i=1}^{p} \mathbf{dom}\, h_i$$

is nonempty.

4.8.1 The Lagrange dual

With each generalized inequality $f_i(x) \preceq_{K_i} 0$ in (4.77) we associate a Lagrange multiplier *vector* $\lambda_i \in \mathbf{R}^{m_i}$ and define the associated Lagrangian as

$$L(x, \lambda, \nu) = f_0(x) + \lambda_1^T f_1(x) + \cdots + \lambda_L^T f_L(x) + \nu_1 h_1(x) + \cdots + \nu_p h_p(x)$$

where $\lambda = (\lambda_1, \ldots, \lambda_L)$ and $\nu = (\nu_1, \ldots, \nu_p)$. The dual function is defined exactly as in the standard problem, *i.e.*,

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) = \inf_{x \in D} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^T f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right).$$

Since the Lagrangian is affine in the dual variables (λ, ν), and the dual function is a pointwise infimum of the Lagrangian, the dual function is concave.

As in the standard problem, the dual function gives lower bounds on $p^\star$, the optimal value of the primal problem (4.77). In the standard problem formulation, we require $\lambda_i \geq 0$. Here the nonnegativity requirement on the dual variables is replaced by the condition

$$\lambda_i \succeq_{K_i^\star} 0, \quad i = 1, \ldots, L,$$

where $K_i^\star$ denotes the dual cone of K_i. In other words, the Lagrange multipliers associated with inequalities must be dual nonneqative.

Weak duality follows immediately from the definition of dual cone. If $\lambda_i \succeq_{K_i^\star} 0$ and $f_i(\tilde{x}) \preceq_{K_i} 0$, then $\lambda_i^T f_i(\tilde{x}) \leq 0$. Therefore for any primal feasible point $\tilde{x}$ and any $\lambda_i \succeq_{K_i\star} 0$, we have

$$f_0(\tilde{x}) + \sum_{i=1}^{L} \lambda_i^T f_i(\tilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{x}) \leq f_0(\tilde{x}).$$

Taking the infimum over $\tilde{x}$ yields weak duality, *i.e.*, $g(\lambda) \leq p^\star$.

The Lagrange dual optimization problem is

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda_i \succeq_{K_i^\star} 0, \quad i = 1, \ldots, L. \end{array} \tag{4.78}$$

We always have *weak duality*, *i.e.*, $d^\star \leq p^\star$, where $d^\star$ denotes the optimal value of the dual problem (4.78), whether or not the primal problem (4.77) is convex.

As might be expected, *strong* duality ($d^\star = p^\star$) holds when the primal problem is convex and satisfies appropriate constraint qualifications. For example, a generalized version of Slater's condition for the problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \preceq_{K_i} 0, \quad i = 1, \ldots, L \\ & Ax = b, \end{array}$$

where f_0 is convex and f_i is K_i-convex, is that there exists an $x \in \mathbf{relint}\, D$ with $Ax = b$ and $f_i(x) \prec_{K_i} 0$, $i = 1, \ldots, L$. This condition implies strong duality (and also, that the dual optimum is attained).

Example. *Lagrange dual of semidefinite program.* As an example we derive the dual of the semidefinite program in inequality form,

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & F_0 + x_1 F_1 + \cdots + x_n F_n \preceq 0, \end{array} \tag{4.79}$$

where $F_i = F_i^T \in \mathbf{R}^{p \times p}$, $i = 1, \ldots, n$. (Here f_1 is affine, and K_1 is the nonnegative definite cone.)

We associate with the constraint a dual variable or multiplier $Z = Z^T \in \mathbf{R}^{p \times p}$, and the Lagrangian

$$\begin{array}{rcl} L(x, Z) & = & c^T x + \mathbf{Tr}\, Z\,(F_0 + x_1 F_1 + \cdots + x_n F_n) \\ & = & \mathbf{Tr}\, Z F_0 + x_1(c_1 + \mathbf{Tr}\, Z F_1) + \cdots + x_n(c_n + \mathbf{Tr}\, Z F_n), \end{array}$$

which is affine in x. The dual function is given by

$$g(Z) = \inf_x L(x, Z) = \left\{ \begin{array}{ll} \mathbf{Tr}\, F_0 Z & \text{if } \mathbf{Tr}\, F_i Z + c_i = 0, \quad i = 1, \ldots, n \\ -\infty & \text{otherwise.} \end{array} \right.$$

The dual problem can therefore be expressed as

$$\begin{array}{ll} \text{maximize} & \mathbf{Tr}\, F_0 Z \\ \text{subject to} & \mathbf{Tr}\, F_i Z + c_i = 0, \; i = 1, \ldots, n \\ & Z = Z^T \succeq 0. \end{array} \tag{4.80}$$

(We use the fact that the nonnegative definite cone is self dual, *i.e.*, $K_1^* = K_1$; see §1.6.)

Strong duality obtains if the semidefinite program (4.79) is strictly feasible, *i.e.*, if there exists an x with

$$F_0 + x_1 F_1 + \cdots + x_n F_n \prec 0.$$

Example. *Lagrange dual of conic program in equality form.*

We consider the conic form problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b \\ & x \succeq_K 0 \end{array}$$

with $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, and K is a cone in $\mathbf{R}^m$ that induces a generalized inequality.

We associate with the equality constraint a multiplier $\nu \in \mathbf{R}^m$ and with the nonnegativity constraint a multiplier $\lambda \in \mathbf{R}^n$. The Lagrangian is

$$L(x, \lambda, \nu) = c^T x + \lambda^T x + \nu^T (Ax - b),$$

so the dual function is

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu) = \begin{cases} -b^T \nu & \text{if } A^T \nu + \lambda + c = 0 \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem can be expressed as

$$\begin{array}{ll} \text{maximize} & -b^T \nu \\ \text{subject to} & A^T \nu + \lambda + c = 0 \\ & \lambda \succeq_{K^*} 0, \end{array}$$

which can be expressed as

$$\begin{array}{ll} \text{maximize} & -b^T \nu \\ \text{subject to} & A^T \nu + c \preceq_{K^*} 0, \end{array}$$

which is a conic problem in inequality form, involving the dual generalized inequality.

Strong duality obtains if the Slater condition holds, *i.e.*, there is an $x \succ_K 0$ with $Ax = b$.

4.8.2 Theorems of alternatives

We can also derive theorems of alternatives for systems of generalized inequalities and equalities

$$\begin{array}{l} f_i(x) \preceq_{K_i} 0, \quad i = 1, \ldots, L, \\ h_i(x) = 0, \quad i = 1, \ldots, p, \end{array} \tag{4.81}$$

where $K_i \subseteq \mathbf{R}^{m_i}$ are convex cones that induce generalized inequalities. We will also consider systems with strict inequalities

$$\begin{array}{l} f_i(x) \prec_{K_i} 0, \quad i = 1, \ldots, L, \\ h_i(x) = 0, \quad i = 1, \ldots, p. \end{array} \tag{4.82}$$

We assume that

$$D = \cap_{i=0}^m \mathbf{dom}\, f_i \bigcap \cap_{i=1}^p \mathbf{dom}\, h_i$$

is nonempty.

Weak alternatives

We associate with the sytems (4.81) and (4.82) the dual function

$$g(\lambda,\nu) = \inf_{x\in D}\left(\sum_{i=1}^{L}\lambda_i^T f_i(x) + \sum_{i=1}^{p}\nu_i h_i(x)\right)$$

where $\lambda = (\lambda_1,\ldots,\lambda_L)$ with $\lambda_i \in \mathbf{R}^{m_i}$ and $\nu \in \mathbf{R}^p$. In analogy with (4.65), we claim that

$$\lambda_i \succeq_{K_i^\star} 0,\ \ i=1,\ldots,m, \quad g(\lambda,\nu) > 0, \tag{4.83}$$

is a weak alternative to the system (4.81). To verify this, we have to show that at least one of the two systems is infeasible. Suppose there exists an x satisfying (4.81) and (λ,ν) satisfying (4.83). Then we have a contradiction:

$$0 < g(\lambda,\nu) \le \lambda_1^T f_1(x) + \cdots + \lambda_L^T f_L(x) + \nu_1 h_1(x) + \cdots + \nu_p h_p(x) \le 0.$$

Therefore at least one of the two systems (4.81) and (4.83) must be infeasible, *i.e.*, the two systems are weak alternatives.

In a similar way, we can prove that (4.83) and the system

$$\lambda_i \succeq_{K_i^\star},\ \ i=1,\ldots,m, \quad \lambda \ne 0, \quad g(\lambda,\nu) \ge 0. \tag{4.84}$$

form a pair of weak alternatives.

Strong alternatives

We now assume that functions f_i are convex with respect to the cones K_i and the functions h_i are affine. We first consider a system with strict inequalities

$$f_i(x) \prec_{K_i} 0,\ i=1,\ldots,L, \quad Ax=b \tag{4.85}$$

and its alternative

$$\lambda_i \succeq_{K_i^\star} 0, \quad \lambda \ne 0, \quad g(\lambda,\nu) \ge 0. \tag{4.86}$$

We have already seen that (4.85) and (4.86) are weak alternatives. They are also strong alternatives provided the following constraint qualification holds: there exists an $\tilde{x} \in \textbf{relint}\, D$ with $A\tilde{x} = b$. To prove this, we select a set of vectors $y_i \succ_{K_i} 0$, and consider the problem

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & f_i(x) \preceq_{K_i} t y_i \\ & Ax = b \end{array} \tag{4.87}$$

with variables x and $t \in \mathbf{R}$. Slater's condition holds since $(\tilde{x},\tilde{t})$ satisfies the strict inequalities $f_i(\tilde{x}) \prec_{K_i} \tilde{t} y_i$ provided $\tilde{t}$ is large enough.

The dual of (4.87) is

$$\begin{array}{ll} \text{maximize} & g(\lambda,\nu) \\ \text{subject to} & \lambda_i \succeq_{K_i^\star} 0,\ \ i=1,\ldots,L \\ & \sum_i y_i^T \lambda_i = 1 \end{array} \tag{4.88}$$

with variables $\lambda = (\lambda_1, \ldots, \lambda_L)$ and ν.

Now suppose the system (4.85) is infeasible. Then the optimal value of (4.87) is nonnegative. Since Slater's condition is satisfied, we have strong duality and the dual optimum is attained. Therefore there exist $(\tilde{\lambda}, \tilde{\nu})$ that satisfy the constraints of (4.88) and $g(\tilde{\lambda}, \tilde{\nu}) \geq 0$, *i.e.*, the system (4.86) has a solution.

As we noted in the case of scalar inequalities, existence of an $x \in \mathbf{relint}\, D$ with $Ax = b$ is not sufficient for the system of strict inequalities

$$f_i(x) \preceq_{K_i} 0, \; i = 1, \ldots, L, \quad Ax = b \tag{4.89}$$

and its alternative

$$\lambda_i \succeq_{K_i^\star} 0, \quad g(\lambda, \nu) > 0 \tag{4.90}$$

to be strong alternatives. An additional condition is required, *e.g.*, that the optimal value of (4.87) is attained.

Example. *Feasibility of a linear matrix inequality.* The following systems are strong alternatives:

$$F(x) = F_0 + x_1 F_1 + \cdots + x_m F_m \succ 0, \tag{4.91}$$

where $F_i = F_i^T \in \mathbf{R}^{n \times n}$, and

$$Z = Z^T \succeq 0, \; Z \neq 0, \; \mathbf{Tr}\, F_0 Z \leq 0, \; \mathbf{Tr}\, F_i Z = 0, \; i = 1, \ldots, m. \tag{4.92}$$

This follows from the general result, if we take for K the positive semidefinite cone, and

$$g(Z) = \inf_x \left(-\mathbf{Tr}\, Z F(x)\right) = \begin{cases} -\mathbf{Tr}\, F_0 Z & \text{if } \mathbf{Tr}\, F_i Z = 0, \quad i = 1, \ldots, m \\ -\infty & \text{otherwise.} \end{cases}$$

The nonstrict inequality case is slightly more involved, and we need an extra assumption on the matrices F_i to have strong alternatives. One such condition is

$$\sum_{i=1}^m v_i F_i \succeq 0 \Longrightarrow \sum_{i=1}^m v_i F_i = 0. \tag{4.93}$$

If (4.93) holds, the following systems are strong alternatives:

$$F(x) = F_0 + x_1 F_1 + \cdots + x_m F_m \succeq 0, \tag{4.94}$$

where $F_i = F_i^T \in \mathbf{R}^{n \times n}$, and

$$Z = Z^T \succeq 0, \; \mathbf{Tr}\, F_0 Z < 0. \; \mathbf{Tr}\, F_i Z = 0, \; i = 1, \ldots, m. \tag{4.95}$$

(see the notes and references).

Chapter 5

Smooth unconstrained minimization methods

5.1 Unconstrained minimization and extensions

In this chapter we discuss methods for solving unconstrained optimization problems

$$\text{minimize } f(x) \tag{5.1}$$

where $f : \mathbf{R}^n \to \mathbf{R}$ is convex and twice continuously differentiable (which, we remind the reader, implies that $\mathbf{dom}\, f$ is open). We denote the optimal value of this problem, *i.e.*, $\inf_x f(x)$, as $f^\star$.

Since f is differentiable, a necessary and sufficient condition for a point $x^\star$ to be optimal is

$$\nabla f(x^\star) = 0 \tag{5.2}$$

(see §3.2.3). Thus, solving the problem (5.1) is the same as computing a solution of the equation (5.2). In a few special cases, we can find an analytical solution to the problem (5.1), but usually the problem has to be solved by an iterative algorithm. By this we mean an algorithm that computes a sequence of points $x^{(0)}$, $x^{(1)}, \ldots$ with $f(x^{(k)}) \to f^\star$ as $k \to \infty$. Such a sequence of points is called a *minimizing sequence* for the problem (5.1). The algorithm is terminated when $f(x^{(k)}) - f^\star \le \epsilon$, where $\epsilon > 0$ is some specified tolerance, or when it is determined that the problem is unbounded below.

Initial point and sublevel set

The methods described in this chapter require a suitable starting point $x^{(0)}$. The starting point must lie in $\mathbf{dom}\, f$, and in addition the sublevel set

$$S = \{\, x \in \mathbf{dom}\, f \mid f(x) \le f(x^{(0)}) \,\} \tag{5.3}$$

must be closed. The importance of this second condition will become clear later; it is needed to rule out the possibility that the algorithm converges to a point in the boundary of $\mathbf{dom}\, f$. If $\mathbf{dom}\, f = \mathbf{R}^n$, the condition is satisfied by any $x^{(0)}$.

5.1.1 Examples

Least-squares and unconstrained quadratic minimization

Two important unconstrained minimization problems are least-squares problems

$$\text{minimize } \|Ax - b\|^2, \tag{5.4}$$

and the more general convex quadratic minimization problems

$$\text{minimize } x^T Px + 2q^T x + c,$$

where $P = P^T \succeq 0$.

These problems can be solved by solving the optimality conditions, which are $A^T Ax^\star = A^T b$ and $Px^\star = -q$, respectively. These are just linear equations, which can be solved by a variety of numerical algorithms.

Unconstrained geometric programming

As a second example, we consider an unconstrained geometric program in convex form,

$$\text{minimize } \log \left(\sum_{i=1}^m \exp \left(a_i^T x + b_i \right) \right).$$

The optimality condition is

$$\nabla f(x) = \frac{1}{\sum_{j=1}^m \exp(a_j^T x + b_j)} \sum_{i=1}^m \exp(a_i^T x + b_i) a_i = 0,$$

which in general has no analytical solution, so here we must resort to an iterative algorithm. For this problem, $\mathbf{dom}\, f = \mathbf{R}^n$, so any point can be chosen as the initial point $x^{(0)}$.

Analytic center of linear inequalities

Consider the optimization problem

$$\text{minimize } \quad -\sum_{i=1}^m \log(b_i - a_i^T x) \tag{5.5}$$

where we take $\log x = -\infty$ for $x \leq 0$. The domain of the objective function is the open set

$$\mathbf{dom}\, f = \{ \ x \mid a_i^T x < b_i, \ i = 1, \ldots, m \ \}.$$

The objective function in this problem is called the *logarithmic barrier* for the inequalities $a_i^T x \leq b_i$. The solution of (5.5), if it exists, is called the *analytic center* of the inequalities.

The initial point $x^{(0)}$ must satisfy the strict inequalities $a_i^T x^{(0)} < b_i$, $i = 1, \ldots, m$. The sublevel set S for any such point is closed, since $f(x) \to \infty$ as $x \to \partial\, \mathbf{dom}\, f$.

Analytic center of a linear matrix inequality

A related problem is

$$\text{minimize} \quad \log\det F(x)^{-1} \tag{5.6}$$

where $F(x) = F_0 + x_1F_1 + \cdots + x_mF_m$ and $F_i = F_i^T$, and we take $\log\det F(x)^{-1} = \infty$ for $F(x) \not\succ 0$. The objective function is called the *logarithmic barrier* for the linear matrix inequality $F(x) \succeq 0$, and the solution (if it exists) is called the *analytic center* of the linear matrix inequality.

Here the initial point $x^{(0)}$ must satisfy the strict linear matrix inequality $F(x^{(0)}) \succ 0$. As in the previous example, the sublevel set of any such point will be closed, since $f(x) \to \infty$ as $x \to \partial\, \mathbf{dom}\, f$.

5.1.2 Strong convexity and implications

In much of this chapter (with the exception of §5.6) we assume that the objective function is *strongly convex* on S, which means that there exists an $m > 0$ such that

$$\nabla^2 f(x) \succeq mI \tag{5.7}$$

for all $x \in S$. Strong convexity has several interesting consequences. For $x,\ y \in S$ we have

$$f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T\nabla^2 f(z)(y-x)$$

for some z on the line segment $[x, y]$. By the strong convexity assumption (5.7), the last term on the righthand side is at least $(m/2)\|y-x\|^2$, so we have the inequality

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|^2 \tag{5.8}$$

for all x and y in S. When $m = 0$, we recover the basic inequality characterizing convexity; for $m > 0$ we obtain a better lower bound on $f(y)$ than follows from convexity alone.

We will first show that the inequality (5.8) can be used to bound $f(x) - f^\star$, which is the suboptimality of the point x, in terms of $\|\nabla f(x)\|$. The righthand side of (5.8) is a convex quadratic function of y (for fixed x). Setting the gradient with respect to y equal to zero, we find that $\tilde{y} = x - (1/m)\nabla f(x)$ minimizes the righthand side. Therefore we have

$$\begin{aligned} f(y) &\geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|^2 \\ &\geq f(x) + \nabla f(x)^T(\tilde{y}-x) + \frac{m}{2}\|\tilde{y}-x\|^2 \\ &= f(x) - \frac{1}{2m}\|\nabla f(x)\|^2. \end{aligned}$$

Since this holds for any $y \in S$, we have

$$f^\star \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|^2. \tag{5.9}$$

This inequality shows that if the gradient is small at a point, then the point is nearly optimal. The inequality (5.9) can also be interpreted as a condition for *suboptimality* which generalizes the optimality condition (5.2):

$$\|\nabla f(x)\| \leq (2m\epsilon)^{1/2} \Longrightarrow f(x) - f^\star \leq \epsilon. \tag{5.10}$$

We can also derive a bound on $\|x - x^\star\|$, the distance between x and any optimal point $x^\star$, in terms of $\|\nabla f(x)\|$:

$$\|x - x^\star\| \leq \frac{2}{m}\|\nabla f(x)\|. \tag{5.11}$$

To see this, we apply (5.8) with $y = x^\star$ to obtain

$$\begin{aligned} f^\star = f(x^\star) &\geq f(x) + \nabla f(x)^T(x^\star - x) + \frac{m}{2}\|x^\star - x\|^2 \\ &\geq f(x) - \|\nabla f(x)\|\|x^\star - x\| + \frac{m}{2}\|x^\star - x\|^2, \end{aligned}$$

where we use the Cauchy-Schwarz inequality in the second inequality. Since $f^\star \leq f(x)$, we must have

$$-\|\nabla f(x)\| \, \|x^\star - x\| + \frac{m}{2}\|x^\star - x\|^2 \leq 0,$$

from which (5.11) follows. One consequence of (5.11) is that the optimal point $x^\star$ is unique.

Upper bound on $\nabla^2 f(x)$

Strong convexity of f on S implies that the sublevel sets contained in S are bounded, so in particular, S is bounded. Therefore the maximum eigenvalue of $\nabla^2 f(x)$, which is a continuous function of x on S, is bounded above on S, *i.e.*, there exists a constant M such that

$$\nabla^2 f(x) \preceq MI \tag{5.12}$$

for all $x \in S$. This upper bound on the Hessian implies for any $x,\ y \in S$,

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{M}{2}\|y - x\|^2, \tag{5.13}$$

which is analogous to (5.8). Minimizing each side over y yields

$$f^\star \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|^2,$$

the counterpart of (5.9). Similarly, we can give a lower bound on the distance to the optimal point, in terms of $\|\nabla f(x)\|$:

$$\|x^\star - x\| \geq \frac{2}{M}\|\nabla f(x)\|^2.$$

The strong convexity constants

It must be kept in mind that the constants m and M are known only in rare cases, so the inequality (5.10) is not a practical stopping criterion. It can be considered a *conceptual* stopping criterion; it shows that if the gradient of f is small enough, then the difference between $f(x)$ and $f^\star$ is small. If we terminate an algorithm when $\|\nabla f(x^{(k)})\| \leq \eta$, where η is chosen small enough to be (very likely) smaller than $(m\epsilon)^{1/2}$, then we have $f(x^{(k)}) - f^\star \leq \epsilon$ (very likely).

In the following sections we give convergence proofs for algorithms, which include bounds on the number of iterations required before $f(x^{(k)}) - f^\star \leq \epsilon$, where ϵ is some positive tolerance. Many of these bounds involve the (usually unknown) constants m and M, so the same comments apply. These results are at least conceptually useful; they establish that the algorithm converges, even if the bound on the number of iterations required to reach a given accuracy depends on constants that are unknown.

We will encounter one important exception to this situation. In §5.6 we will study a special class of convex functions, called *self-concordant*, for which we can provide a complete and practically useful convergence analysis that does not depend on any unknown constants.

5.2 Descent methods

The algorithms described in this chapter produce a minimizing sequence $x^{(k)}$, $k = 1, \ldots,$ where

$$x^{(k+1)} = x^{(k)} + t^{(k)} v^{(k)}$$

and $t^{(k)} > 0$ (except when $x^{(k)}$ is optimal). Here k denotes the iteration number. When we focus on one iteration of an algorithm, we sometimes drop the superscripts and use the lighter notation $x^+ = x + tv$, or $x := x + tv$ in place of $x^{(k+1)} = x^{(k)} + t^{(k)} v^{(k)}$.

The vector $v^{(k)}$ is called the *search direction* (even though it need not have unit norm) and the scalar $t^{(k)} \geq 0$ is called the *step size* or *step length* at iteration k (even though it is not equal to $\|x^{(k+1)} - x^{(k)}\|$ unless $\|v^{(k)}\| = 1$). Perhaps the terms 'search step' and 'scale factor' are more accurate terms, but search direction and step length are widely used.

All the methods we study are *descent methods* which means that

$$f(x^{(k+1)}) < f(x^{(k)})$$

(except when $x^{(k)}$ is optimal). From convexity we know that $\nabla f(x^{(k)})^T (y - x^{(k)}) \geq 0$ implies $f(y) \geq f(x^{(k)})$, so the search direction in a descent method must satisfy

$$\nabla f(x^{(k)})^T v^{(k)} < 0,$$

i.e., it must make an acute angle with the negative gradient. We call such a direction a *descent direction* (for f, at $x^{(k)}$).

The outline of a general descent method is as follows. It alternates between two steps: the computation of the descent direction v, and the selection of a step size t.

General descent method.

given a starting point $x \in \mathbf{dom}\, f$.

repeat

1. *Determine a descent direction* v.
2. *Line search.* Choose a step size $t > 0$.
3. *Update.* $x := x + tv$.

until stopping criterion is satisfied.

The second step is called the *line search* since selection of the step size t determines where along the line $\{ x + tv \mid t \in \mathbf{R} \}$ the next iterate will be. (A more accurate term might be *ray search.*)

A practical descent method has the same general structure, but might be organized d-ifferently. For example, the stopping criterion is often checked while, or immediately after, the descent direction v is being computed. The stopping criterion is usually of the form $\|\nabla f(x)\| \leq \eta$, where η is small and positive, as suggested by the suboptimality condition (5.9).

In the following sections, we present several methods for selecting descent directions and step sizes. We first briefly describe two common types of line searches, and how they can be carried out efficiently in some cases.

5.2.1 Line search types

Exact line search

One line search method often used in practice is *exact line search*, in which t is chosen to minimize f along the ray $\{x + tv | t \geq 0\}$:

$$t = \mathrm{argmin}_s\, f(x + sv). \tag{5.14}$$

An exact line search is used when the cost of the minimization problem with one variable, required in (5.14), is low compared to the cost of computing the search direction itself. In some special cases the minimizer along the ray can be found analytically; in others it can be computed efficiently.

Backtracking line search

Most line searches used in practice are *inexact*: the step length is chosen to approximately minimize f along the ray $\{x + tv | t \geq 0\}$, or even to just reduce f 'enough'.

One line search method that is very simple and quite effective is called *backtracking* line search. It depends on two constants α, β with $0 < \alpha < 0.5$, $0 < \beta < 1$.

Backtracking line search.

$t := 1$.

while $(f(x + tv) > f(x) + \alpha t \nabla f(x)^T v)$

  $t := \beta t$.

end

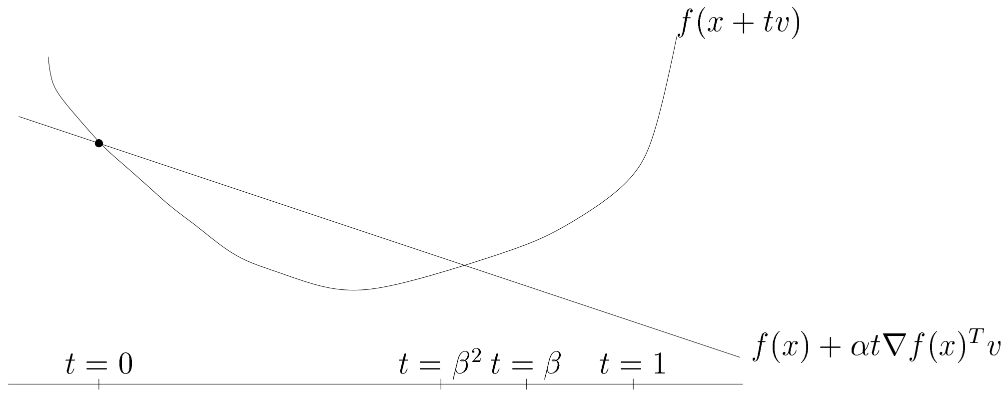

Figure 5.1: Backtracking line search. In this example we have $\beta = 0.8$ and $\alpha = 0.3$. Two backtracking steps are required before the inequality is satisfied; the line search terminates with $t = \beta^2$.

The line search is called backtracking because it starts with unit step size and then reduces it by the factor β until the condition $f(x+tv) \leq f(x) + \alpha t \nabla f(x)^T v$ holds.

Since v is a descent direction, we have $\nabla f(x)^T v < 0$, so for small enough t

$$f(x+tv) \approx f(x) + t\nabla f(x)^T v < f(x) + \alpha t \nabla f(x)^T v.$$

Thus the backtracking eventually terminates.

The constant α can be interpreted as the fraction of the decrease in f predicted by linear extrapolation that we will settle for. The reason for imposing the upper bound $\alpha < 0.5$ will become clear later.

When $\mathbf{dom}\, f$ is not all of $\mathbf{R}^n$, the condition $f(x+tv) \leq f(x) + \alpha t \nabla f(x)^T v$ in the backtracking line search must be interpreted carefully. By our convention that f is infinite outside its domain, the inequality implies that $x+tv \in \mathbf{dom}\, f$. In a practical implementation, we first multiply t by β until $x+tv \in \mathbf{dom}\, f$; then we start to check whether the inequality $f(x+tv) \leq f(x) + \alpha t \nabla f(x)^T v$ holds.

The parameter α is typically chosen between 0.1 and 0.3, meaning that we accept a decrease in f between 10% and 30% of the prediction based on the linear extrapolation. The parameter β is often chosen to be between 0.1 (which corresponds to a very crude search) and 0.5 (which corresponds to less crude search).

Pre-computation for line searches

In the simplest implementation of a line search, $f(x+tv)$ is evaluated for each value of t in the same way that $f(z)$ is evaluated for any $z \in \mathbf{dom}\, f$. But in some cases we can exploit the fact that f (and its derivatives, in an exact line search) are to be evaluated at many points along the ray $\{x+tv \mid t \geq 0\}$ to reduce the computational effort.

This usually requires some pre-computation, which is often on the same order as computing f at any point, after which f (and its derivatives) can be computed more efficiently along the ray.

Suppose that $x \in \mathbf{dom}\, f$ and $v \in \mathbf{R}^n$, and define g as f restricted to the line or ray determined by x and v, *i.e.*, $g(t) = f(x + tv)$. In a backtracking line search we must evaluate g for several, and possibly many, values of t; in an exact line search method we must evaluate g and one or more derivatives at a number of values of t. In the simple method described above, we evaluate $g(t)$ by first forming $z = x + tv$, and then evaluating $f(z)$. To evaluate $g'(t)$, we form $z = x + tv$, then evaluate $\nabla f(z)$, and then compute $g'(t) = \nabla f(z)^T v$.

To demonstrate how a line search can be carried out more efficiently we consider the problem (5.6) of computing the analytic center of a linear matrix inequality, *i.e.*, minimizing $\log\det F(x)^{-1}$, where $x \in \mathbf{R}^m$ and $F \in \mathbf{R}^{p\times p}$. Along the line through x with direction v we have

$$g(t) = \log\det(F(x + tv))^{-1} = -\log\det(A + tB)$$

where $A = F(x)$ and $B = v_1F_1 + \cdots + v_nF_n$. Therefore we can express g as

$$g(t) = -\log\det\left(A^{1/2}(I + tA^{-1/2}BA^{-1/2})A^{1/2}\right) = -\log\det A - \sum_{i=1}^{p} \log(1 + t\lambda_i)$$

where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $A^{-1/2}BA^{-1/2}$. Once these eigenvalues are computed, we can evaluate $g(t)$, for any t, with $O(p)$ simple artihmetic computations. Indeed we can evaluate $g'(t)$ (and similarly, any higher derivative) in $O(p)$ computations, using the formula

$$g'(t) = -\sum_{i=1}^{p} \frac{\lambda_i}{1 + t\lambda_i}.$$

Let us compare the two methods for carrying out a line search that requires evaluation of $f(x + tv)$ for k values of t. In the simple method, for each value of t we form $F(x + tv)$, and then evaluate $f(x + tv)$ as $-\log\det F(x)$, for example, by finding the Cholesky factorization of F. The cost is $O(np^2)$ to form $F(x + tv)$, plus $O(p^3)$ to factorize $F(x + tv)$. The total cost is $O(kp^2(p + n))$.

Using the method outlined above, we first form A, which costs $O(np^2)$, and factorize it, which costs $O(p^3)$. We also form B (which costs $O(np^2)$), and also $A^{-1/2}BQ^{-1/2}$, which costs $O(p^3)$. The eigenvalues of this matrix are then computed, at a cost of $O(p^3)$. Having finished the pre-computation, we can now evaluate $g(t)$ for each value of t at a cost of $O(p)$. The total cost is then $O(p^2(p + n) + kp)$. Assuming k is small compared to $p(p + n)$, this means the entire line search can be carried out at the same effort as simply evaluating f. Depending on the values of k, p, and n, the savings over the simple method can be as large as $O(k)$.

5.3 Gradient and steepest descent methods

5.3.1 Gradient method

A natural choice for the search direction is the negative gradient $-\nabla f(x)$. The resulting algorithm is called the *gradient algorithm.*

Gradient descent method.

given a starting point x.

repeat

1. *Compute search direction* $v = -\nabla f(x)$.
2. *Line search.* Choose step size t via exact or backtracking line search.
3. *Update.* $x := x + tv$.

until stopping criterion is satisfied.

The stopping criterion is usually of the form $\|\nabla f(x)\| \leq \eta$, where η is small and positive. In a real implementation, this condition would be checked after step 1.

5.3.2 Convergence analysis of gradient method

In this section we present a simple analysis of the convergence of the gradient method, using the notation $x^+ = x + tv$ for $x^{(k+1)} = x^{(k)} + t^{(k)}v^{(k)}$.

We assume f is strongly convex on S, so there are positive constants m and M such that $mI \preceq \nabla^2 f(x) \preceq MI$. Now define the function $g : \mathbf{R} \to \mathbf{R}$ by $g(t) = f(x - t\nabla f(x))$, *i.e.*, f as a function of the step length t in the negative gradient direction. In the following discussion we will only consider t for which $x - t\nabla f(x) \in S$.

From the inequality (5.13), with $y = x - t\nabla f(x)$, we obtain a quadratic upper bound on g:

$$g(t) \leq f(x) - t\|\nabla f(x)\|^2 + \frac{Mt^2}{2}\|\nabla f(x)\|^2. \tag{5.15}$$

Now we minimize over t both sides of this inequality. On the lefthand side we get $g(t_{\mathrm{exact}})$, where t_{exact} is the step length that minimizes g. The righthand side is a simple quadratic, which is minimized by $t = 1/M$, and has minimum value $f(x) - (1/2M)\|\nabla f(x)\|^2$. Therefore we have

$$f(x^+) = g(t_{\mathrm{exact}}) \leq f(x) - \frac{1}{2M}\|\nabla(f(x))\|^2.$$

Subtracting $f^\star$ from both sides, we get

$$f(x^+) - f^\star \leq f(x) - f^\star - \frac{1}{2M}\|\nabla f(x)\|^2.$$

We combine this with $\|\nabla f(x)\|^2 \geq 2m(f(x) - f^\star)$ (which follows from (5.9)) to conclude

$$f(x^+) - f^\star \leq (1 - m/M)(f(x) - f^\star).$$

Applying this inequality recursively, we find that

$$f(x^{(k)}) - f^\star \leq c^k \left(f(x^{(0)}) - f^\star\right)$$

where $c = 1 - m/M < 1$.

This implies that $f(x^{(k)})$ converges to $f^\star$. In particular, we must have $f(x^{(k)}) - f^\star \leq \epsilon$ after at most

$$\frac{\log((f(x^{(0)}) - f^\star)/\epsilon)}{\log(1/c)}$$

iterations of the gradient method with exact line search.

Analysis for backtracking line search

Now we consider the case where a backtracking line search is used. We will show that the backtracking exit condition,

$$g(t) \leq f(x) - \alpha t \|\nabla f(x)\|^2,$$

is satisfied whenever $0 \leq t \leq 1/M$. First note that

$$0 \leq t \leq 1/M \implies -t + \frac{Mt^2}{2} \leq -t/2$$

(which follows from convexity of $-t + Mt^2/2$). Using this result and the bound (5.15), we have, for $0 \leq t \leq 1/M$,

$$\begin{aligned} g(t) &\leq f(x) - t\|\nabla f(x)\|^2 + \frac{Mt^2}{2}\|\nabla(f(x))\|^2 \\ &\leq f(x) - t/2\|\nabla f(x)\|^2 \\ &\leq f(x) - \alpha t\|\nabla f(x)\|^2, \end{aligned}$$

since $\alpha \leq 1/2$.

Therefore the backtracking line search terminates either with $t = 1$ or with a value $t \geq \beta/M$. This provides a lower bound on the decrease in the objective function. In the first case we have

$$f(x^+) \leq f(x) - \alpha\|\nabla f(x)\|^2,$$

and in the second case we have

$$f(x^+) \leq f(x) - \beta\alpha/M\|\nabla f(x)\|^2.$$

Putting these together, we always have

$$f(x^+) \leq f(x) - \min\{\alpha, \beta\alpha/M\}\|\nabla f(x)\|^2.$$

Now we can proceed exactly as in the case of exact line search. We subtract $f^\star$ from both sides to get

$$f(x^+) - f^\star \leq f(x) - f^\star - \min\{\alpha, \beta\alpha/M\}\|\nabla f(x)\|^2,$$

and combine this with $\|\nabla f(x)\|^2 \geq 2m(f(x) - f^\star)$ to obtain

$$f(x^+) - f^\star \leq (1 - \min\{2m\alpha, 2\beta\alpha m/M\})(f(x) - f^\star).$$

From this we conclude

$$f(x^{(k)}) - f^\star \leq c^k \left(f(x^{(0)}) - f^\star\right)$$

where

$$c = 1 - \min\{2m\alpha, 2\beta\alpha m/M\} < 1.$$

In particular, $f(x^{(k)})$ converges to $f^\star$.

5.3.3 The steepest descent method

The first order Taylor approximation at x, of $f(x+z)$, is

$$f(x+z) \approx f(x) + \nabla f(x)^T z.$$

The second term on the righthand side, $\nabla f(x)^T z$, is the *directional derivative* of f at x in the direction z. It gives the approximate change in f for a small step z. The step z is a descent direction if the directional derivative is negative.

We now address the question of how to choose z to make the directional derivative as small as possible. Since the directional derivative $\nabla f(x)^T z$ is linear in z, it can be made as small as we like by taking z large (provided z is a descent direction, *i.e.*, $\nabla f(x)^T z < 0$). To make the question sensible we have to limit the size of z, or normalize by the length of z.

Let $\|\cdot\|$ be any norm on $\mathbf{R}^n$. We define a *normalized steepest descent direction* (with respect to the norm $\|\cdot\|$) as

$$v_{\mathrm{nsd}} = \mathrm{argmin}\{\nabla f(x)^T z \mid \|z\| = 1\}. \tag{5.16}$$

We say 'a' steepest descent direction because there can be multiple minimizers. The normalized steepest descent direction v_{nsd} is the step of unit norm that gives the largest decrease in the linear approximation of f.

The normalized steepest descent direction can be interpreted geometrically as follows. We can just as well define v_{nsd} as

$$v_{\mathrm{nsd}} = \mathrm{argmin}\{\ \nabla f(x)^T z \mid \|z\| \leq 1\ \},$$

i.e., as the direction in the unit ball of $\|\cdot\|$ that extends farthest in the direction $-\nabla f(x)$.

It is also convenient to consider a steepest descent step v_{sd} that is *unnormalized*, by scaling the normalized steepest descent direction in a particular way:

$$v_{\mathrm{sd}} = v_{\mathrm{nsd}} \|\nabla f(x)\|_*, \tag{5.17}$$

where $\|\cdot\|_*$ denotes the dual norm. The steepest descent step v_{sd} can also be described as

$$v_{\mathrm{sd}} = \mathrm{argmin}\left(\nabla f(x)^T z + (1/2)\|z\|^2\right).$$

The first term in the expression on the right is the directional derivative, and the second term penalizes long steps. Note that for the steepest descent step, we have

$$\nabla f(x)^T v_{\mathrm{sd}} = -\|\nabla f(x)\|_*^2.$$

The *steepest descent method* uses the steepest descent direction as search direction.

STEEPEST DESCENT METHOD.

given a starting point $x \in \mathbf{dom}\, f$.

repeat

1. *Compute steepest descent direction* v_{sd}.
2. *Line search.* Determine t via backtracking or exact line search.
3. *Update.* $x := x + tv$.

until stopping criterion is satisfied.

When exact line search is used, scale factors in the descent direction have no effect, so the normalized or unnormalized direction can be used.

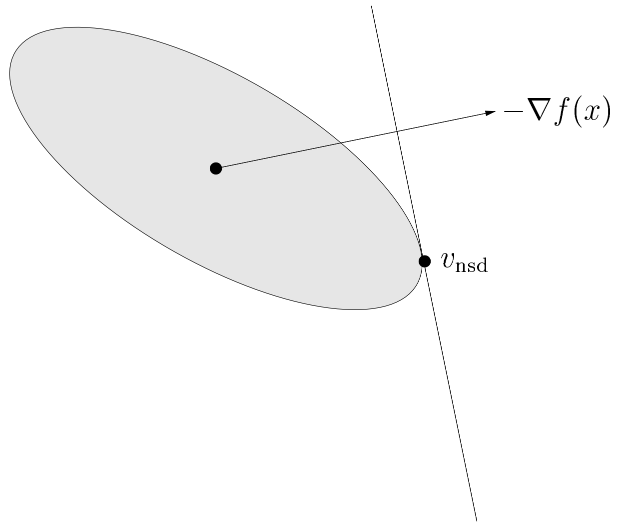

Figure 5.2: Normalized steepest descent direction for a quadratic norm. The ellipsoid shown is the unit ball of the norm. The normalized steepest descent direction $v_{\rm nsd}$ at x extends as far as possible in the direction $-\nabla f(x)$ while staying in the unit ball.

Steepest descent for Euclidean norm

If we take the norm $\|\cdot\|$ to be the Euclidean norm we find that the steepest descent direction is simply the negative gradient, *i.e.*, $v_{\rm sd} = -\nabla f(x)$. The steepest descent method for the Euclidean norm coincides with the gradient descent method.

Steepest descent for quadratic norm

We consider the quadratic norm

$$\|z\|_P = \left(z^T P z\right)^{1/2} = \left\|P^{1/2} z\right\|,$$

where $P = P^T \succ 0$. The normalized steepest descent direction is given by

$$v_{\rm nsd} = -\left(\nabla f(x)^T P^{-1} \nabla f(x)\right)^{-1/2} P^{-1/2} \nabla f(x).$$

The dual norm is given by $\|z\|_* = \|P^{-1/2} z\|$, so the steepest descent step with respect to $\|\cdot\|_P$ is given by

$$v_{\rm sd} = -P^{-1/2} \nabla f(x).$$

The normalized steepest descent direction for a quadratic norm is illustrated in figure 5.2.

Steepest descent for ℓ_1-norm

As another example, we consider the steepest descent method for the ℓ_1-norm. A normalized steepest descent direction,

$$v_{\rm nsd} = \mathop{\rm argmin}\{\nabla f(x)^T z \mid \|z\|_\infty \le 1\},$$

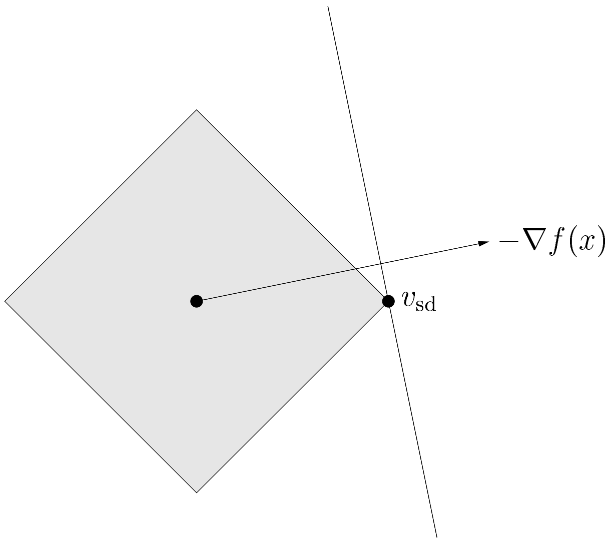

Figure 5.3: Normalized steepest descent direction for the ℓ_1-norm. The diamond is the unit ball of the ℓ_1-norm. The normalized steepest descent direction with respect to the ℓ_1-norm extends as far as possible in the direction of the negative gradient while staying in the unit ball. It can always be chosen in the direction of a unit vector; in this example we have $v_{\rm nsd} = e_1$.

is easily characterized: let i be any index for which $\|\nabla f(x)\|_\infty = |(\nabla f(x))_i|$. Then a normalized steepest descent direction $v_{\rm nsd}$ for the ℓ_1 norm is given by

$$v_{\rm nsd} = -\mathrm{sign}\left(\frac{\partial f}{\partial x_i}\right) e_i,$$

where e_i is the ith unit vector. The unnormalized steepest descent step is then

$$v_{\rm sd} = v_{\rm nsd}\|\nabla f(x)\|_\infty = -\frac{\partial f}{\partial x_i} e_i.$$

Thus, the steepest descent step in ℓ_1 norm can always be chosen to be a unit vector (or a negative unit vector). It is the coordinate axis direction long which the approximate decrease in f is greatest. This is illustrated in figure 5.3.

The steepest descent algorithm in the ℓ_1-norm has a very natural interpretation: at each iteration we select a component of $\nabla f(x)$ with maximum absolute value; and then decrease or increase the corresponding component of x, according to the sign of $(\nabla f(x))_i$. The algorithm is sometimes called a *coordinate-descent* algorithm, since only one component of the variable x is updated at each iteration. This can greatly simplify, or even trivialize, the line search.

Example. *Frobenius norm scaling.* In chapter 3 we encountered the following unconstrained geometric program

$$\text{minimize} \sum_{i,j=1}^{n} M_{ij}^2 d_i^2/d_j^2.$$

where $M \in \mathbf{R}^{n\times n}$ is given, and the variable is $d \in \mathbf{R}^n$. Using the change of variables $x_i = 2\log d_i$ we can express this geometric program in convex form

$$\text{minimize } f(x) = \log \sum_{i,j=1}^{n} M_{ij}^2 e^{x_i - x_j}.$$

We show that it is easy to minimize f one component at a time. Keeping all components except the kth fixed, we can write $f(x) = \alpha_k + \beta_k e^{-x_k} + \gamma_k e^{x_k}$, where

$$\alpha_k = M_{kk}^2 + \sum_{i,j\neq k}^{n} M_{ij}^2 e^{x_i - x_j}, \quad \beta_k = \sum_{i\neq k} M_{ik}^2 e^{x_i}, \quad \gamma_k = \sum_{j\neq k} M_{kj}^2 e^{-x_j}.$$

The minimum of $f(x_1, x_2, \ldots, x_n)$ as a function of x_k is obtained for $x_k = \log(\beta_k/\gamma_k)/2$. So for this problem an exact line search can be carried out using a simple analytical formula.

The ℓ_1-steepest descent algorithm with exact line search consists of repeating the following steps.

1. Compute the gradient $(\nabla f(x))_i = -\beta_i e^{-x_i} + \gamma_i e^{x_i}$, $i = 1, \ldots, n$.
2. Select the largest (in absolute value) component of $\nabla f(x)$: $|\nabla f(x)|_k = \|\nabla f(x)\|_\infty$.
3. Let $t = \operatorname{argmin}_t f(x + te_k)$, and update $x := x + te_k$.

5.3.4 Convergence analysis of the steepest descent method

In this section we extend the simple convergence analysis for the gradient method to the general steepest descent method.

Again we assume f is strongly convex on the initial sublevel set S. The upper bound $\nabla^2 f(x) \preceq MI$ implies an upper bound on the function $f(x+tv)$ as a function of t:

$$\begin{aligned} f(x+tv) &\leq f(x) + t\nabla f(x)^T v + \frac{M\|v\|^2}{2}t^2 \\ &= f(x) - t\|\nabla f(x)\|_*^2 + \frac{M}{2}t^2\|\nabla f(x)\|_*^2. \end{aligned} \tag{5.18}$$

The step size $\hat{t} = 1/M$ (which minimizes the quadratic upper bound (5.18)) satisfies the exit condition for the backtracking line search:

$$f(x+\hat{t}v) \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|_*^2 \leq f(x) - \frac{\alpha}{M}\|\nabla f(x)\|_*^2, \tag{5.19}$$

The line search therefore returns a step size $t \geq \min\{1, \beta/M\}$, and we have

$$f(x^+) = f(x+tv) \leq f(x) - \alpha\min\{1, \beta/M\}\|\nabla f(x)\|_*^2.$$

Since $\|\nabla f(x)\|_*$ is a norm it can be bounded below in terms of the Euclidean norm of $\nabla f(x)$, *i.e.*, there exists a γ $(0 < \gamma \leq 1)$ such that

$$\|\nabla f(x)\|_* \geq \gamma\|\nabla f(x)\|$$

for all x. Substituting this in (5.19), we obtain

$$f(x^+) \le f(x) - \alpha\gamma^2 \min\{1, \beta/M\} \|\nabla f(x)\|^2$$

and

$$f(x^{(k)}) - f^\star \le c^k (f(x^{(0)}) - f^\star)$$

where $c = 1 - m\alpha\gamma^2 \min\{1, \beta/M\}$. Except for the factor γ^2, this is identical to the estimate for the gradient method.

5.4 Newton's method

5.4.1 The Newton step

The vector

$$v = -\nabla^2 f(x)^{-1} \nabla f(x).$$

is called the *Newton step* (for f, at x). The Newton step can be interpreted and motivated in several ways.

The first interpretation is that $x + v$ minimizes the second-order approximation $\hat{f}$ of f at x, which can be written as

$$\hat{f}(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x). \tag{5.20}$$

If f is quadratic, then $x + v$ is the exact minimizer of f. The idea of the method is that any twice differentiable convex function f is nearly quadratic near its optimum, and therefore $x + v$ will be a much better approximation of $x^\star$ than x, provided x is already close to $x^\star$. We will make this argument more precise below.

The Newton step can also be interpreted as the steepest-descent direction at x, for the quadratic norm defined by the Hessian $\nabla^2 f(x)$, *i.e.*,

$$v = \operatorname{argmin} \left\{ v^T \nabla f(x) \;\middle|\; v^T \nabla^2 f(x) v \le 1 \right\}.$$

Positive definiteness of $\nabla^2 f(x)$ implies that

$$\nabla f(x)^T v = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) < 0$$

unless $\nabla f(x) = 0$, so the Newton step is always a descent direction.

A third interpretation of the Newton step is that $x + v$ solves the linear equation

$$\nabla f(x) + \nabla^2 f(x)(y - x) = 0,$$

which is the linearization around x of the nonlinear equation $\nabla f(x^\star) = 0$ (*i.e.*, the optimality condition).

Affine invariance of the Newton step

An important feature of the Newton step is that it is independent of affine changes of coordinates. Suppose $A \in \mathbf{R}^{n\times n}$ is nonsingular, and define $\tilde{f}(y) = f(Ay+b)$. Then we have

$$\nabla \tilde{f}(y) = A^T \nabla f(x), \quad \nabla^2 \tilde{f}(y) = A^T \nabla^2 f(x) A,$$

where $x = Ay + b$. The Newton step for $\tilde{f}$ at y is therefore

$$\tilde{v} = -\left(A^T \nabla^2 f(x) A\right)^{-1}\left(A^T \nabla f(x)\right) = -A^{-1}\nabla^2 f(x)^{-1}\nabla f(x) = A^{-1}v,$$

if v is the Newton step for f at x. Hence the Newton steps of f and $\tilde{f}$ are related by the same linear transformation, and

$$x + v = A(y + \tilde{v}) + b$$

The Newton decrement

The quantity

$$\lambda(x) = \left(\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)\right)^{1/2} = \left\|\nabla^2 f(x)^{-1/2}\nabla f(x)\right\|$$

is called the *Newton decrement* at x. We will see that the Newton decrement plays an important role it the analysis of Newton's method, and is also useful as a stopping criterion.

The Newton decrement can be interpreted several ways. The defining formula shows that it is the length of the gradient, after a linear change of coordinates that makes the Hessian the identity matrix. As another interpretation, we can relate the Newton decrement to the the quantity $f(x) - \inf_y \hat{f}(y)$, where $\hat{f}$ is the second-order approximation of f at x:

$$f(x) - \inf_y \hat{f}(y) = f(x) - \hat{f}(x+v) = -\frac{1}{2}\lambda(x)^2.$$

Thus, $\lambda(x)^2/2 = f(x) - f(x+v)$ when f is exactly quadratic.

5.4.2 Newton's method

Newton's method, as outlined below, is sometimes called the *damped* Newton method or *guarded* Newton method, to distinguish it from the *pure* Newton method, which uses a fixed step size $t = 1$.

Newton's method.

given a starting point $x \in \mathbf{dom}\, f$.

repeat

1. *Compute the Newton step.* $v = -\nabla^2 f(x)^{-1}\nabla f(x)$.
2. *Line search:* compute step size t by backtracking line search.
3. *Update.* $x := x + tv$.

until stopping criterion is satisfied.

5.4.3 Convergence analysis

We assume, as before, that f is twice continuously differentiable, and strongly convex with constant m, *i.e.*, $\nabla^2 f(x) \succeq mI$ for $x \in S$. We have seen that this also implies that there exists an $M > 0$ such that $\nabla^2 f(x) \preceq MI$ for all $x \in S$.

In addition, we assume that the Hessian of f is Lipschitz continuous on S with constant L, *i.e.*,

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \leq L\|x - y\|$$

for all x, $y \in S$. The coefficient L is zero for a quadratic function; for general functions it measures how well f can be approximated by a quadratic model.

Idea and outline of convergence proof

We first give the idea and outline of the convergence proof, and the main conclusion, and then the details of the proof.

We will show there are numbers η_1 and η_2 with $0 < \eta_1 \leq m^2/L$ and $\eta_2 > 0$ such that the following hold.

- if $\|\nabla f(x^k)\| \geq \eta_1$, then
$$f(x^{k+1}) - f(x^{(k)}) \leq -\eta_2. \tag{5.21}$$
- if $\|\nabla f(x^{(k)})\| < \eta_1$, then the backtracking line search selects $t = 1$ and
$$\frac{L}{2m^2}\|\nabla f(x^{(l)}\| \leq \left(\frac{L}{2m^2}\|\nabla f(x^{(l-1)})\|\right)^2 \tag{5.22}$$
for $l > k$. This has the following implications:
  - since $\eta_1 \leq m^2/L$, we have $\|\nabla f(x^{(l)})\| < \eta_1$ for all $l \geq k$
  - the algorithm takes full Newton steps $t = 1$ for all $l \geq k$
  - $f(x^{(l)})$ converges quadratically to $f^\star$, *i.e.*, there exists a $\gamma > 0$ such that
  $$f(x^{(l)}) - f^\star \leq \gamma\left(f(x^{(l-1)}) - f^\star\right)^2$$
  for all $l > k$. (This follows directly from (5.22), (5.48) and (5.9).)

We refer to the first stage as the *damped Newton phase*, because the algorithm uses a step size $t \leq 1$. The second stage, where $\|\nabla f(x^k)\| < \eta_1$ is called the *quadratically convergent* or *pure Newton phase*.

Based on these facts we can estimate the total complexity. First, there is an upper bound on the number of damped Newton steps. Since f decreases by η_2 at each iteration, the number of damped Newton steps cannot exceed

$$\frac{f(x^{(0)}) - f^\star}{\eta_2},$$

since if it did, f would be less than $f^\star$.

We can bound the number of iterations in the quadratically convergent phase by observing that if $\|\nabla f(x^{(k)})\| < \eta_1$ and $l \geq k$, then

$$\frac{L}{2m^2}\|\nabla f(x^{(l)}\| \leq \left(\frac{L}{2m^2}\|\nabla f(x^k)\|\right)^{2^{l-k}} \leq \left(\frac{1}{2}\right)^{2^{l-k}}$$

and hence

$$f(x^{(l)}) - f^\star \leq \frac{1}{m}\|\nabla f(x^{(l)})\|^2 \leq \frac{4m^3}{L^2}\left(\frac{1}{2}\right)^{2^{l-k}}.$$

We therefore have $f(x) - f^\star \leq \epsilon$ after no more than

$$\log_2 \log_2(\epsilon_0/\epsilon)$$

iterations in the quadratically convergent phase, where $\epsilon_0 = 4m^3/L^2$.

Overall, then, the number of iterations until $f(x) - f^\star \leq \epsilon$, is bounded above by

$$\frac{f(x^{(0)}) - f^\star}{\eta_2} + \log_2 \log_2(\epsilon_0/\epsilon).$$

The term $\log_2 \log_2(\epsilon_0/\epsilon)$, which bounds the number of iterations in the quadratically convergent phase, grows *extremely slowly* with required accuracy ϵ, and can be considered a constant for practical purposes. Not quite accurately, then, we can say that the number of iterations required to find the minimum of f is bounded above by

$$\frac{f(x^{(0)}) - f^\star}{\eta_2} + c$$

where c is a constant (say, five or six). (The precise statement is that this is the number of iterations to compute a very good approximation of the solution.)

One important practical consequence is that the stopping criterion for Newton's method is straightforward. We can stop whenever the gradient, or some other measure such as the Newton decrement, is smaller than some very small tolerance.

It remains to prove the two facts above.

Damped Newton phase

Assume $\|\nabla f(x)\| \geq \eta_1$. We first derive a lower bound on the step size selected by the line search. Strong convexity implies $\nabla^2 f(x) \preceq MI$ on S, and therefore

$$\begin{aligned} f(x+tv) &\leq f(x) + t\nabla f(x)^T v + \frac{M\|v\|^2}{2}t^2 \\ &\leq f(x) - t\lambda(x)^2 + \frac{M}{2m}t^2\lambda(x)^2. \end{aligned}$$

The step size $\hat{t} = m/M$ satisfies the exit condition of the line search, since

$$f(x+\hat{t}v) \leq f(x) - \frac{m}{2M}\lambda(x)^2 \leq f(x) - \alpha\hat{t}\lambda(x)^2.$$

Therefore the line search returns a step size $t \geq \beta m/M$, resulting in a decrease of the objective function

$$f(x^+) - f(x) \leq -\alpha t \lambda(x)^2 \leq -\alpha\beta\eta_1^2 \frac{m}{M^2},$$

i.e., (5.21) is satisfied with

$$\eta_2 = \alpha\beta\eta_1^2 \frac{m}{M^2}$$

Quadratically convergent phase

We first show that the backtracking line search selects unit steps, provided

$$\eta_1 \leq 3(1-2\alpha)\frac{m^2}{L}.$$

The Lipschitz condition implies that for all $t \geq 0$,

$$-tL\|v\|^3 \leq v^T \left(\nabla^2 f(x+tv) - \nabla^2 f(x)\right) v \leq tL\|v\|^3.$$

This allows us to determine an upper bound on $g(t) = f(x+tv)$. We have $-tL\|v\|^3 \leq g''(t) - g''(0) \leq tL\|v\|^3$, and by integrating twice, we obtain the upper bound

$$f(x+tv) \leq f(x) - t\nabla f(x)^T v + \frac{t^2}{2} v^T \nabla^2 f(x) v + \frac{t^3}{6} L\|v\|^3. \tag{5.23}$$

The first three terms are the second-order approximation of $f(x+tv)$ around $t=0$; the fourth term gives an upper bound on the deviation from the quadratic function. The bound (5.23) holds for any direction v. Specialized to the Newton direction, it simplifies to

$$\begin{aligned} f(x+tv) &\leq f(x) - t\left(1 - \frac{t}{2}\right) \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) + \frac{t^3}{6} L\|\nabla^2 f(x)^{-1}\nabla f(x)\|^3 \\ &\leq f(x) - t\left(1-\frac{t}{2}\right)\lambda(x)^2 + t^3 \frac{L}{6m^{3/2}}\lambda(x)^3 \end{aligned} \tag{5.24}$$

where $\lambda(x)$ is the Newton decrement of f at x. (We used strong convexity in the second step.)

Now suppose

$$\|\nabla f(x)\| \leq \eta_1 \leq 3(1-2\alpha)\frac{m^2}{L}.$$

Then by strong convexity, $\lambda(x) \leq 3(1-2\alpha)\frac{m^{3/2}}{L}$, and

$$f(x+v) \leq f(x) - \frac{1}{2}\lambda(x)^2 + \frac{L}{6m^{3/2}}\lambda(x)^3 \leq f(x) - \alpha\lambda(x)^2 = f(x) + \alpha \nabla f(x)^T v,$$

i.e., the unit step $t=1$ is accepted.

Let us now examine the rate of convergence. Applying the Lipschitz condition, we have

$$\begin{aligned} \|\nabla f(x^+)\| &= \|\nabla f(x+v) - \nabla f(x) - \nabla^2 f(x) v\| = \left\| \int_0^1 \left(\nabla^2 f(x+tv) - \nabla^2 f(x)\right) v dt \right\| \\ &\leq \frac{L}{2}\|v\|^2 = \frac{L}{2}\|\nabla^2 f(x)^{-1}\nabla f(x)\|^2 \leq \frac{L}{2m^2}\|\nabla f(x)\|^2, \end{aligned}$$

i.e., the inequality (5.22).

In conclusion, the algorithm selects unit steps and satisfies (5.22) if

$$\eta_1 = \min\{1, 3(1-2\alpha)\}\frac{m^2}{L}.$$

5.4.4 The pure Newton method

Newton's method is often used with a fixed step size $t = 1$. We call this method the *pure Newton method* or Newton's method with full step size.

PURE NEWTON METHOD.

given a starting point x close to $x^\star$.

repeat

1. *Compute Newton step.* $v = -\nabla^2 f(x)^{-1}\nabla f(x)$.
2. *Update.* $x := x + v$.

until stopping criterion is satisfied.

Our first interpretation of the Newton direction suggests that if x is close to $x^\star$, then this iteration converges very rapidly to the optimum. This is confirmed by our analysis of the quadratically convergent region of the guarded Newton method.

However, simple examples show the pure Newton method can diverge if the initial point is not close to $x^\star$. If $\mathbf{dom}\, f \neq \mathbf{R}^n$, it can even produce infeasible iterates.

Examples.

- $f(x) = \log(e^x + e^{-x})$ 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	$x^{(k)}$	$f(x^{(k)}) - f^\star$
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	$x^{(k)}$	$f(x^{(k)}) - f^\star$
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 $\mathbf{dom}\, f$.

5.5 Variable metric methods

The disadvantages of Newton's method are the relatively high cost of computing the search direction and the need to evaluate the second derivatives of the function. The first disadvantage is often mitigated in practice by the structure in the Newton equations (for example sparsity). Exploiting this structure can reduce the computational cost considerably. Nevertheless, situations arise where Newton's method is too expensive, or the second derivatives are not readily available. Variable metric algorithms are a class of methods that can be used in those situations. They combine some of the advantages of the gradient method and Newton algorithms: they require only first derivatives but converge much faster than the gradient descent method; the convergence is slower than Newton's method but the iterations are cheaper. Some variants also require less storage than Newton's method.

The search directions used in these methods have the form $v = -H^{-1}\nabla f(x)$ where $H = H^T$ is a positive definite matrix that is changed at each iteration. We can interpret v as the steepest descent direction in a norm defined by H, or as an approximate Newton direction, where the Hessian is replaced with an approximation H. Note that the condition $H \succ 0$ guarantees that the direction v is a descent direction: $v^T\nabla f(x) = -\nabla f(x)^T H^{-1}\nabla f(x) < 0$.

We will discuss two classes of variable metric methods: quasi-Newton methods and the conjugate gradients method.

5.5.1 The Broyden-Fletcher-Goldfarb-Shanno method

In a quasi-Newton method, we use $f(x^+)$ and $\nabla f(x^+)$ to update H. Several variants for this update exist. To be practical, the update should satisfy several properties. First, H should remain positive definite, since otherwise the direction $-H^{-1}\nabla f(x)$ might not be a descent direction. Secondly, the update should be cheap to compute (in any case cheaper than evaluating the Hessian of f). Moreover, it is desirable that the system of equations $H^+v = -\nabla f(x^+)$ is easier to solve than the Newton system $\nabla^2 f(x^+)v = -\nabla f(x^+)$. Finally, we imposes the following so-called *secant condition* on H^+:
$$\nabla f(x^+) - \nabla f(x) = H^+(x^+ - x).$$
To motivate this, consider the quadratic function
$$r(\tilde{x}) = f(x^+) + \nabla f(x^+)^T(\tilde{x} - x^+) + \frac{1}{2}(\tilde{x} - x^+)^T H^+(\tilde{x} - x^+).$$
We would like to choose the update H^+ in such a way that r is a good approximation to f at x^+. An obvious condition is that H^+ is symmetric and positive definite. A second condition is that the gradient of r at x coincides with $\nabla f(x)$, which is exactly the secant condition.

We will discuss only one method, the so-called *Broyden-Fletcher-Goldfarb-Shanno method* (or BFGS-method), which is generally believed to be the most efficient quasi-Newton method. In the BFGS method we update H as follows:

$$H^{+} = H + \frac{yy^T}{y^T s} - \frac{Hss^T H}{s^T H s}, \tag{5.25}$$

where $y = \nabla f(x^+) - \nabla f(x)$ and $s = x^+ - x$. It is readily verified that H^+ is positive definite if the function f is convex and that it satisfies the secant equation. Moreover H^+ is a simple rank-two update of H, and this can be exploited to solve the linear system $H^+ v = -\nabla f(x^+)$ more efficiently.

Broyden-Fletcher-Goldfarb-Shanno (BFGS) method.

given a starting point $x \in \mathbf{dom}\, f$ and a positive definite matrix H.

repeat

1. *Compute quasi-Newton direction* $v = -H^{-1}\nabla f(x)$.
2. *Line search.* Determine t by backtracking line search.
3. *Update H.* $H := H + \frac{yy^T}{y^T s} - \frac{Hss^T H}{s^T H s}$ where $y = \nabla f(x + tv) - \nabla f(x)$ and $s = tv$.
4. *Update x.* $x := x + tv$.

until stopping criterion is satisfied.

For the initial H we can use any positive definite approximation of $\nabla^2 f(x^{(0)})$, or, if none is available, simply take $H = I$.

A straightforward implementation of the algorithm requires, beside the evaluation of the function and its derivatives, $O(n^3)$ operations to solve the linear system $Hv = -\nabla f(x)$. It is possible to save an order of magnitude by updating the *inverse* of H instead of H. Let $B = H^{-1}$ and $B^+ = (H^+)^{-1}$. Using the Sherman-Morrison formula we can show that

$$B^{+} = B + \left(1 + \frac{y^T B y}{s^T y}\right)\frac{ss^T}{s^T y} - \frac{Bys^T + sy^T B}{y^T s}. \tag{5.26}$$

Updating B instead of H reduces the cost of computing the search direction to $O(n^2)$, since the search direction follows from a matrix-vector product $v = B\nabla f(x)$.

Both update rules (5.25) and (5.26) suffer from the disadvantage that roundoff errors may cause the matrices B^+ or H^+ to become negative definite (even though positive definiteness is guaranteed in exact arithmetic). For numerical reasons it is therefore recommended to propagate the Cholesky factor of H instead of H of B. Let $H = RR^T$ be the Cholesky decomposition of H. To see how we can update the Cholesky factor R to the Cholesky factor R^+ of H^+, note that (5.25) can be written as

$$\begin{aligned} H^{+} &= R\left(I + \frac{\bar{y}\bar{y}}{\bar{s}^T\bar{y}} - \frac{\bar{s}\bar{s}^T}{\bar{s}^T\bar{s}}\right)R^T \\ &= R\left(I + \frac{(\bar{y} - \sqrt{\alpha}\bar{s})\bar{s}^T}{\sqrt{\alpha}\bar{s}^T\bar{s}}\right)\left(I + \frac{(\bar{y} - \sqrt{\alpha}\bar{s})\bar{s}^T}{\sqrt{\alpha}\bar{s}^T\bar{s}}\right)^T R^T \end{aligned}$$

where $\bar{y} = R^{-1}y$, $\bar{s} = R^T s$, and $\alpha = \bar{y}^T\bar{s}/\bar{s}^T\bar{s}$. (Note incidentally that this formula proves that $H^+ \succ 0$ whenever $y^T s > 0$). To compute R^+ from R, we reduce the matrix

$$\bar{R} = R\left(I + \frac{(\bar{y} - \sqrt{\alpha}\bar{s})\bar{s}^T}{\sqrt{\alpha}\bar{s}^T\bar{s}}\right)$$

to triangular form. This can be done in $O(n^2)$ operations, since $\bar{R}$ is a rank-one update of R.

5.5.2 Convergence analysis of BFGS method

The assumptions are identical to the assumptions in §5.4.3, *i.e.*, strong convexity and Lipschitz continuity of the Hessian.

Global convergence

We first show that the BFGS method with backtracking converges. The proof relies on the following nontrivial fact (proved in the notes and references). There exist positive constants $q_1 \le \min\{1, M\}$ and $q_2 \ge 1$ (q_1 and q_2 depend only on m and M) such that at least half of the iterations satisfy

$$\frac{-v^T\nabla f(x)}{\|v\|\,\|\nabla f(x)\|} \ge me^{1-M}, \tag{5.27}$$

and

$$q_1 \le \frac{-v^T\nabla f(x)}{\|v\|^2} \le q_2. \tag{5.28}$$

Let us estimate the decrease in objective function achieved in an iteration where (5.27) and (5.28) hold. We first derive a lower bound on the step size, by noting that the step size

$$\hat{t} = -\frac{v^T\nabla f(x)}{M\|v\|^2}$$

satisfies the exit condition for the backtracking line search:

$$f(x + \hat{t}v) \le f(x) + \nabla f(x)^T v\hat{t} + \frac{M\|v\|^2}{2}\hat{t}^2 = f(x) - \frac{(\nabla f(x)^T v)^2}{2M\|v\|^2} \le f(x) + \alpha\hat{t}\nabla f(x)^T v,$$

since $\alpha < 0.5$. The line search therefore selects a step size

$$t \ge \min\{1, \frac{-\beta v^T\nabla f(x)}{M\|v\|^2}\} \ge \frac{\beta q_1}{M},$$

which results in a decrease of the objective function by

$$f(x + tv) - f(x) \le \alpha t\nabla f(x)^T v \le \frac{\alpha\beta q_1}{M}\nabla f(x)^T v \le -\frac{\alpha\beta q_1(me^{1-M})^2}{Mq_2}\|\nabla f(x)\|^2.$$

Hence

$$f(x^+) - f^\star \le r(f(x) - f^\star) \tag{5.29}$$

with

$$r = 1 - \frac{\alpha\beta q_1 m e^{2(1-M)}}{M q_2}.$$

Since (5.29) holds for at at least $\lfloor k/2 \rfloor$ of the iterates $0, \ldots, k-1$, we have

$$f(x^{(k)}) - f^\star \le r^{\lfloor k/2 \rfloor}(f(x^{(0)}) - f^\star).$$

Superlinear local convergence

There is a local convergence result parallel to the quadratic conergence result of Newton's method. It states that the asympotic convergence is superlinear, *i.e.*,

$$\frac{\|f(x^{(k+1)}) - f^\star\|}{\|f(x^{(k)}) - f^\star\|} \to 0.$$

(See notes and references.)

5.5.3 Conjugate gradients

The conjugate gradients method is a useful alternative to the BFGS method for problems that are so large that storing and updating the factors of the approximate Hessian H is too expensive. The method can be interpreted as a variant of the BFGS method with exact line search, where the matrix H is reset to I at each iteration. We will see that this leads to a very simple update rule for x.

Consider the inverse update rule (5.26) with $B = I$:

$$B^+ = I + \frac{ss^T}{s^T y}\left(1 + \frac{y^T y}{s^T y}\right) - \frac{ys^T + sy^T}{y^T s},$$

where $s = x^+ - x$, $y = \nabla f(x^+) - \nabla f(x)$. Suppose we determined x^+ by an exact line search, *i.e.*, $x^+ = x + tv$ where t minimizes $g(t) = f(x + tv)$. Then

$$g'(t) = \nabla f(x^+)^T (x^+ - x) = \nabla f(x^+)^T s = 0.$$

This simplifies the expression of the search direction v^+ at x^+, since $v^+ = -B^+ \nabla f(x^+)$ reduces to

$$v^+ = -\nabla f(x^+) + \frac{y^T \nabla f(x^+)}{y^T s} s. \tag{5.30}$$

Note that the search direction is a simple linear combination of the negative gradient and the previous search direction s, and can be computed at a cost of $O(n)$ operations. This update rule (5.30) is due to Hesteness and Stiefel.

CONJUGATE GRADIENTS METHOD.

given a starting point $x \in \mathbf{dom}\, f$. Take $\alpha = 0$.
repeat

1. *Compute search direction.* $v := -\nabla f(x) + \alpha s$.
2. *Line search.* Determine t by exact line search.
3. *Update s and α.* $s := tv$, $\alpha := y^T \nabla f(x+s)/y^T s$ $(y = \nabla f(x+s) - \nabla f(x))$.
4. *Update x.* $x := x + tv$.

until stopping criterion is satisfied.

5.6 Self-concordance

The convergence analysis in the previous sections is based on the assumption of strong convexity and Lipschitz continuity of the Hessian. We have derived complexity estimates that involved three constants m, M, and L. The main shortcoming of the analysis is that these constants are usually not known in practice.

In this section we replace these assumptions by a simpler and more elegant assumption, called *self-concordance*, and analyze the convergence of Newton's method applied to self-concordant functions. Self-concordant functions are important for several reasons.

- They include logarithmic barrier functions that will play an important role in the next chapter, and that do not satisfy the strong convexity assumption or the Lipschitz condition (or at least, the constants m, M, and L depend in a complicated way on the starting point).

- The analysis of Newton's method for self-concordant functions does not depend on any unknown constants.

- Self-concordance is an affine-invariant property, *i.e.*, if we apply a linear transformation of variables to a self-concordant function, we obtain a self-concordant function. Therefore the complexity estimate that we will obtain for Newton's method applied to a self-concordant function does not depend on the coordinate system.

  This is in contrast with the analysis we presented above, which is not affine-invariant. In other words, if we apply a linear transformation of variables, Newton's method will produce identical iterates, and converge in the same number of iterations. However, the constants m, M and L will be different, and therefore also the bound on the number of iterations.

5.6.1 Definition and examples

A convex function $f : \mathbf{R} \to \mathbf{R}$ is *self-concordant* if

$$|f'''(x)| \leq 2 f''(x)^{3/2} \tag{5.31}$$

for all $x \in \mathbf{dom}\, f$.

Examples.

- Linear and quadratic functions are evidently self-concordant.
- $f(x) = -\log x$ is self-concordant:
$$f''(x) = -1/x^2, \quad f'''(x) = 2/x^3.$$
- $f(x) = -\log(-\alpha x^2 + \beta x + \gamma)$ where $\alpha \geq 0$ is self-concordant on $\{ x \mid -\alpha x^2 + \beta x + \gamma > 0 \}$. (For a proof, see the notes and references.)
- $f(x) = x \log x - \log x$ is self-concordant on $\{ x \mid x > 0 \}$.
$$f''(x) = \frac{x+1}{x^2}, \quad f'''(x) = -\frac{x+2}{x^3},$$
and the ratio $(x+2)/(x+1)^{3/2}$ is maximum for $x = 0$, where it is equal to 2.

Remark. The coefficient 2 in the definition (5.31) is chosen for convenience, in order to simplify the formulas later on. Any other positive constant could be used instead: if g satisfies
$$|g'''(x)| \leq a g''(x)^{3/2}$$
for some $a > 0$, then $f(x) = a^2 g(x)/4$ is self-concordant.

A function $f : \mathbf{R}^n \to \mathbf{R}$ is self-concordant if it is self-concordant along every line in its domain, *i.e.*, if the function $g(t) \stackrel{\Delta}{=} f(x + tv)$ is a self-concordant function of t for all $x \in \mathbf{dom}\, f$ and for all v.

Examples.

- The function $f(x) = -\log(-x^T Ax - b^T x - c)$ where $A = A^T \in \mathbf{R}^{n\times n}$, and $A \succeq 0$, is self-concordant. If we restrict f to a line, we obtain a function of the form
$$f(x + tv) = -\log(-(v^T Av)t^2 - (b^T v + 2v^T Ax)t - x^T Ax - b^T x - c),$$
and as we have seen above, this is a self-concordant function of t.

5.6.2 Self-concordant calculus

Addition and linear transformation of variables

If f is self-concordant, then af with $a \geq 1$ is self-concordant.

If f_1, f_2 are self-concordant, then $f(x) = f_1(x) + f_2(x)$ is self-concordant. To show this, it is sufficient to consider functions $f_i : \mathbf{R} \to \mathbf{R}$. We have
$$\begin{aligned} |f_1'''(x) + f_2'''(x)| &\leq |f_1'''(x)| + |f_2'''(x)| \\ &\leq 2(f_1''(x)^{3/2} + f_2''(x)^{3/2}) \\ &\leq 2(f_1''(x) + f_2''(x))^{3/2}. \end{aligned}$$
If f_1, f_2 are self-concordant, then $f(x_1, x_2) = f_1(x_1) + f_2(x_2)$ is self-concordant.

If $f : \mathbf{R}^m \to \mathbf{R}$ is self-concordant, and $A \in \mathbf{R}^{n\times m}$, $b \in \mathbf{R}^n$, then $f(Ax + b)$ is self-concordant.

Examples.

- The function $f(x) = -\sum_{i=1}^m \log(b_i - a_i^T x)$, defined on $\mathbf{dom}\, f = \{\ x \mid a_i^T x < b_i,\ i = 1, \ldots, m\ \}$ is self-concordant. Each term $-\log(b_i - a_i^T x)$ is the composition of $-\log y$ with the affine transformation $y = b_i - a_i^T x$, and hence self-concordant. Therefore the sum f is also self-concordant.
- $f(X) = -\log\det X$ is self-concordant on $\mathbf{dom}\, f = \{\ X = X^T \in \mathbf{R}^{n\times n} \mid X \succ 0\ \}$. The function $g(t) = f(X + tV)$, where $X \succ 0$ and V are given, can be written as

$$\begin{aligned} f(X+tV) &= -\log\det(X^{1/2}(I + tX^{-1/2}VX^{-1/2})X^{1/2}) \\ &= -\log\det X - \log\det(I + tX^{-1/2}VX^{-1/2}) \\ &= -\log\det X - \sum_{i=1}^n (1 + t\lambda_i) \end{aligned}$$

where λ_i are the eigenvalues of $X^{-1/2}VX^{-1/2}$. In other words $f(X + tV)$ is a sum of self-concordant functions of t, and therefore self-concordant itself.

Composition with logarithm

Let $g : \mathbf{R} \to \mathbf{R}$ be a convex function with $\mathbf{dom}\, g = \{\ x \mid x > 0\ \}$, and

$$|g'''(x)| \leq 3\frac{g''(x)}{x} \tag{5.32}$$

for all x. Then

$$f(x) = -\log(-g(x)) - \log x$$

is self-concordant on $\{\ x \mid x > 0,\ g(x) < 0\ \}$. (For a proof, see the notes and references.)

Examples. The following functions g satisfy the condition (5.32), for any $a, b \in \mathbf{R}$.

- $g(x) = ax + b - x^p$ for $0 < p \leq 1$.
- $g(x) = ax + b - \log x$.
- $g(x) = ax + b + x\log x$.
- $g(x) = ax + b + x^p$ for $-1 \leq p \leq 0$.
- $g(x) = (ax + b)^2/x$.

Examples

By combining the calculus rules we can prove self-concordance for a wide variety of functions.

- $f(x, y) = -\log(y^2 - x^T x)$ is self-concordant on $\{\ (x, y) \mid \|x\| < y\ \}$.
- The function $f(x, y) = -2\log y - \log(y^{2/p} - x^2)$, with $p \geq 1$, is self-concordant on $\{\ (x, y) \in \mathbf{R}^2 \mid |x|^p \leq y\ \}$.
- $f(x, y) = -\log y - \log(\log y - x)$ on $\{\ (x, y) \mid y \geq e^x\ \}$.

5.6.3 Properties of self-concordant functions

In §5.1.2 we used strong convexity to derive bounds on the suboptimality of a point x and on its distance to the optimum. The bounds were expressed in terms of the norm of the gradient at x. For self-concordant functions, we can obtain scale-independent bounds in terms of the Newton decrement

$$\lambda(x) = \left(\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)\right)^{1/2}.$$

For future reference we note that the Newton decrement can also be expressed as

$$\lambda(x) = \sup_{v \neq 0} \frac{-v^T \nabla f(x)}{\sqrt{v^T \nabla^2 f(x) v}} \tag{5.33}$$

(see exercises).

Upper and lower bounds on second derivatives

Suppose $g : \mathbf{R} \to \mathbf{R}$ is a strictly convex self-concordant function. We can write (5.31) as

$$\left| \frac{d}{dt} \left(\frac{1}{\sqrt{g''(t)}} \right) \right| \leq 1 \tag{5.34}$$

for all $t \in \mathbf{dom}\, g$. Assuming $0 \in \mathbf{dom}\, g$ we can integrate (5.34) to find upper and lower bounds on $g''(t)$:

$$\frac{g''(0)}{\left(1 + t\sqrt{g''(0)}\right)^2} \leq g''(t) \leq \frac{g''(0)}{\left(1 - t\sqrt{g''(0)}\right)^2} \tag{5.35}$$

for $t \geq 0$. (The upper bound is valid for $t < 1/\sqrt{g''(0)}$.)

Bound on suboptimality and distance to the optimum

Let $f : \mathbf{R}^n \to \mathbf{R}$ be a strictly convex self-concordant function, and let v be a descent direction (*i.e.*, any direction satisfying $v^T \nabla f(x) < 0$; not necessarily the Newton direction). Define $g : \mathbf{R} \to \mathbf{R}$ as $g(t) = f(x + tv)$. By definition, the function g is self-concordant.

Integrating the lower bound in (5.35) twice yields lower bounds on $g'(t)$ and $g(t)$:

$$g'(t) \geq g'(0) + g''(0) \frac{t}{1 + \sqrt{g''(0)}t}, \tag{5.36}$$

and

$$g(t) \geq g(0) + g'(0)t + \sqrt{g''(0)}t - \log(1 + \sqrt{g''(0)}t) \tag{5.37}$$

for $t \geq 0$. The right-hand side reaches its minimum for

$$t = \frac{-g'(0)}{g''(0) + \sqrt{g''(0)}g'(0)}.$$

The minimum value of the right hand side provides a lower bound on g:

$$\begin{aligned}\min_{t\geq 0} g(t) &\geq g(0) - g'(0)/\sqrt{g''(0)} + \log(1 + g'(0)/\sqrt{g''(0)})\\ &\geq g(0) + \lambda(x) + \log(1-\lambda(x)),\end{aligned}$$

where we made use of (5.33), and the fact that $x + \log(1-x)$ is a monotonically decreasing function. The lower bound holds for any descent direction v. Therefore

$$f(x^\star) \geq f(x) + \lambda(x) + \log(1-\lambda(x))$$

if $\lambda(x) < 1$. In particular, if $\lambda(x) \leq 0.81$, then

$$f(x) - f(x^\star) \leq \lambda(x). \tag{5.38}$$

We can also bound $x - x^\star$ in terms of $\lambda(x)$. From (5.36) we see that if $t^\star$ is the minimizer of $g(t)$, then

$$g'(0) + g''(0)\frac{t^\star}{1+\sqrt{g''(0)}t^\star} \leq 0$$

i.e.,

$$\sqrt{g''(0)}t^\star \leq \frac{-g'(0)/\sqrt{g''(0)}}{1 + g'(0)/\sqrt{g''(0)}} \leq \frac{\lambda(x)}{1-\lambda(x)}$$

where again we used (5.33). 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

$$\left((x-x^\star)^T\nabla^2 f(x)(x-x^\star)\right)^{1/2} \leq \frac{\lambda(x)}{1-\lambda(x)} \tag{5.39}$$

if $\lambda(x) < 1$.

5.6.4 Analysis of Newton's method for self-concordant functions

We now use the definition and properties of self-concordant functions to analyze the guarded Newton method with backtracking line search. We will study the following variation. Note that the validity of the stopping criterion follows directly from (5.38).

Newton's method for self-concordant functions.

given a starting point $x \in \mathbf{dom}\, f$, and a tolerance $\epsilon \leq 0.25$.
repeat

1. *Compute Newton direction.* $v = -\nabla^2 f(x)^{-1}\nabla f(x)$.
2. *Line search.* Determine t from backtracking line search.
3. *Update.* $x := x + tv$.

until $\lambda(x) \leq \epsilon$.

The analysis and proofs are very similar to the analysis of §5.4.2. In particular, the analysis distinguishes two phases: we will show that there exist numbers η_1 $(0 < \eta_1 \leq 1/4)$ and $\eta_2 > 0$ such that the following holds:

- if $\lambda(x^{(k)} > \eta_1$, then
$$f(x^{(k+1)}) - f(x^{(k)}) \leq -\eta_2. \tag{5.40}$$
- if $\lambda(x^{(k)}) \leq \eta_1$, then the backtracking line search selects $t = 1$ and
$$\frac{\lambda(x^{(l)})}{2} \leq \left(\frac{\lambda(x^{(k)})}{2}\right)^{2^{l-k}} \tag{5.41}$$
for $l \geq k$. As a consequence, if $l - k \geq \log_2 \log_2(2/\epsilon)$, we have
$$\lambda(x^{(l)}) \leq 2\left(\frac{\lambda(x^{(k)})}{2}\right)^{2^{l-k}} \leq 2\left(\frac{\eta_1}{2}\right)^{2^{l-k}} \leq 2\left(\frac{1}{8}\right)^{2^{l-k}} \leq 2\left(\frac{1}{2}\right)^{2^{l-k}} \leq \epsilon,$$
and the algorithm terminates.

The numbers η_1 and η_2 depend only on the parameters α and β used in the line search.

Combining these two facts, we can bound the number of iterations required to obtain an accuracy $f(x) - f^\star \leq \epsilon$, starting at a point $x^{(0)}$, by
$$\frac{f(x^{(0)}) - f^\star}{\eta_2} + \log_2 \log_2(2/\epsilon).$$
In other words we obtain a similar formula as in the previous analysis based on strong convexity. The difference here is that the constants are the same for all self-concordant functions.

Damped Newton phase

Let $g(t) = f(x+tv)$ where v is a descent direction. If we integrate the upper bound in (5.35) twice, we obtain an upper bound for $g(t)$:
$$g(t) \leq g(0) + tg'(0) - t\sqrt{g''(0)} - \log\left(1 - t\sqrt{g''(0)}\right).$$
If v is the Newton direction at x, we have $g'(0) = -g''(0) = -\lambda(x)^2$, and the upper bound simplifies to
$$g(t) \leq g(0) - t\lambda(x)^2 - t\lambda(x) - \log(1 - t\lambda(x)) \tag{5.42}$$
for $0 \leq t < 1/\lambda(x)$.

We can use this bound to show the backtracking line search always results in a step size $t \geq \beta/(1 + \lambda(x))$. To prove this we simply note that the point $\hat{t} = 1/(1 + \lambda(x))$ satisfies the exit condition of the line search:
$$\begin{aligned} g(\hat{t}) &\leq g(0) - \hat{t}\lambda(x)^2 - \hat{t}\lambda(x) - \log(1 - \hat{t}\lambda(x)) \\ &= g(0) - \lambda(x) + \log(1 + \lambda(x)) \\ &\leq g(0) - \alpha\frac{\lambda(x)^2}{1 + \lambda(x)} \\ &= g(0) - \alpha\lambda(x)^2\hat{t}. \end{aligned}$$

(The second inequality follows from the fact that $-x + \log(1+x) + \frac{x^2}{2(1+x)} \leq 0$ for $x \geq 0$.)

Since $t \geq \beta/(1+\lambda(x))$, we have

$$g(t) - g(0) \leq \alpha\beta \frac{\lambda(x)^2}{1+\lambda(x)},$$

so (5.40) holds with

$$\eta_2 = \alpha\beta \frac{\eta_1^2}{1+\eta_1}.$$

Quadratically convergent phase

We will show that we can take

$$\eta_1 = (1-2\alpha)/4,$$

i.e., if $\lambda(x^{(k)}) \leq (1-2\alpha)/4$, then the backtracking line search accepts the unit step and (5.41) holds.

We first note that the upper bound (5.42) implies that a unit step $t = 1$ yields a point in $\mathbf{dom}\, f$ if $\lambda(x) < 1$. Moreover, if $\lambda(x) \leq (1-2\alpha)/2$, we have, using (5.42),

$$\begin{aligned} g(1) &\leq g(0) - \lambda(x)^2 - \lambda(x) - \log(1-\lambda(x)) \\ &\leq g(0) - \frac{1}{2}\lambda(x)^2 + \lambda(x)^3 \\ &\leq g(0) - \alpha\lambda(x)^2, \end{aligned}$$

so the unit step satisfies the condition of sufficient decrease. (The second line follows from the fact that $-x - \log(1-x) \leq \frac{1}{2}x^2 + x^3$ for $0 \leq x \leq 1/2$.)

The inequality (5.41) follows from the following fact, proved in the notes and references: If $\lambda(x) < 1$, and $x^+ = x - \nabla^2 f(x)^{-1} \nabla f(x)$, then

$$\lambda(x^+) \leq \frac{1+\lambda(x)}{1-\lambda(x)} \lambda(x)^2. \tag{5.43}$$

In particular, if $\lambda(x) \leq 1/3$,

$$\lambda(x^+) \leq 2\lambda(x)^2,$$

which proves that (5.41) holds when $\lambda(x^{(k)}) \leq \eta_1$.

5.7 Problems with equality constraints

Simple extensions of the methods in this chapter can also handle problems with equality contraints, *i.e.*,

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b, \end{array} \tag{5.44}$$

where $f : \mathbf{R}^n \to \mathbf{R}$ is twice continuously differentiable and convex, and $A \in \mathbf{R}^{p \times n}$. (For simplicity we will assume that A has full rank, *i.e.*, $\mathbf{Rank}\, A = p$.)

As described in §3.2.4, the problem (5.44) can be solved by eliminating the equality constraints, yielding a problem of the form (5.1)

$$\text{minimize } f(Fz + \hat{x}). \tag{5.45}$$

Here $\hat{x}$ is any particular solution of $Ax = b$, and $F \in \mathbf{R}^{n\times(n-p)}$ is chosen so that range(F) is the nullspace of A. The basic methods for unconstrained minimization can then be applied to (5.45). Methods that explicitly handle the equality constraints in (5.44), however, can be easier to understand, and can have practical advantages as well. Assume, for example, that A is sparse, n is very large, and $p \ll n$. Then F will be very large and (in general) dense, so calculating and storing F may be very expensive, and it may be better to handle the equality constraints directly.

Again the methods will require a suitable starting point $x^{(0)}$. It must satisfy $x^{(0)} \in \mathbf{dom}\, f$, $Ax^{(0)} = b$, and the sublevel set

$$S = \{\, x \in \mathbf{dom}\, f \mid f(x) \leq f(x^{(0)}),\ Ax = b \,\}$$

must be closed. If $\mathbf{dom}\, f = \mathbf{R}^n$, any point satisfying $Ax^{(0)} = b$ can be used.

Optimality condition

Recall from §3.2.3 that a point $x^\star$ is optimal for (5.44) if and only if $Ax^\star = b$ and there is a ν with

$$\nabla f(x^\star) + A^T\nu = 0. \tag{5.46}$$

We can also derive this from the optimality conditions for (5.45), which are

$$F^T\nabla f(Fz^\star + \hat{x}) = 0,$$

by noting that $F^Ty = 0$ if and only if $y = A^T\nu$ for some ν.

Example. Consider the equality constrained least-squares problem

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|^2 \\ \text{subject to} & Cx = d. \end{array}$$

The optimality conditions are

$$A^T(Ax^\star - b) + C^T\nu = 0, \quad Cx^\star = d,$$

which can also be expressed as

$$\begin{bmatrix} A^TA & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} x^\star \\ \nu \end{bmatrix} = \begin{bmatrix} A^Tb \\ d \end{bmatrix}.$$

These optimality conditions are linear equations and can be directly solved.

Strong convexity and implications

We can easily extend the bounds on suboptimality (5.11) and (5.48) to take into account equality constraints. We assume f is strongly convex on S, *i.e.*, $\nabla^2 f(x) \succeq mI$ for all x in S.

Since $x^\star$ satisfies the optimality conditions (5.46), we have for all v that satisfy $Av = 0$, and for all y,

$$\begin{aligned} 0 = v^T \nabla f(x^*) &= v^T \nabla f(x) + \int_0^1 v^T \nabla^2 f(x + t(x^\star - x))(x^\star - x)\, dt \\ &= v^T \nabla f(x) + v^T H(x^\star - x) \\ &= v^T(\nabla f(x) + A^T y) + v^T H(x^\star - x) \end{aligned}$$

The matrix H is defined as

$$H = \int_0^1 \nabla^2 f(x + t(x^\star - x))\, dt$$

and satisfies $H \succeq mI$. Therefore

$$\|v\| \|\nabla f(x) + A^T y\| \geq |v^T H(x^\star - x)| \geq m\|v\| \|x^\star - x\|,$$

for all y, *i.e.*,

$$\|x^\star - x\| \leq \frac{1}{m} \inf_y \|\nabla f(x) + A^T y\|. \tag{5.47}$$

We can also bound $f(x) - f^\star$ where $Ax = b$. By convexity, we have for all y,

$$\begin{aligned} f^\star &\geq f(x) + \nabla f(x)^T (x^\star - x) \\ &= f(x) + (\nabla f(x) + A^T y)^T (x^\star - x) \\ &\geq f(x) - \|\nabla f(x) + A^T y\| \|x^\star - x\| \end{aligned}$$

from which we obtain

$$f(x) - f^\star \leq \frac{1}{m^2} \inf_y \|\nabla f(x) + A^T y\|^2. \tag{5.48}$$

This inequality provides an approximate stopping criterion:

$$Ax = b, \;\; \inf_y \|\nabla f(x) + A^T y\| \leq \sqrt{m\epsilon} \Longrightarrow f(x) - f^\star \leq \epsilon.$$

Descent methods

We now review the modifications needed to explicitly handle the equality constraints in the methods we described above. For problems with equality constraints the search directions v have to satisfy the additional condition $Av = 0$, so that the iterates remains feasible, *i.e.*, $Ax^{(k)} = b$ for $k = 0, 1, \ldots$. Such a search direction is called a *feasible* search direction.

- *Gradient method.* As search direction in the gradient method we use the projection of the negative gradient on the set of feasible directions, *i.e.*,

  $$v = \operatorname{argmin}\{\ \|\nabla f(x) + y\| \mid Ay = 0\ \}.$$

  This search direction can be computed in several ways, *e.g.*, by solving the equations

  $$\begin{bmatrix} I & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}$$

  for v and w.

- *Steepest descent method.* The normalized steepest descent direction for a norm $\|\cdot\|$ is defined as

  $$v_{\rm sd} = \operatorname{argmin}\left\{ \nabla f(x)^T v \;\middle|\; \|v\| = 1,\ Av = 0 \right\}.$$

  Again this can be computed several ways. For example, it can be shown that

  $$v_{\rm sd} = \frac{-\nabla f(x) + A^T w}{\|-\nabla f(x) + A^T w\|},$$

  where

  $$w = \operatorname{argmin} \|\nabla f(x) + A^T y\|_*.$$

- *Newton's method.* To handle equality constraints in the Newton method we define the Newton direction as the solution of

  $$\begin{array}{ll} \text{minimize} & v^T \nabla f(x) + \frac{1}{2} v^T \nabla^2 f(x) v \\ \text{subject to} & Av = 0, \end{array}$$

  i.e., the feasible direction that minimizes the second-order approximation of f around x. It can be shown that v can be computed by solving the equations

  $$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}$$

  in v and w.

Chapter 6

Sequential unconstrained minimization methods

6.1 Introduction and assumptions

In this chapter we discuss methods for the standard optimization problem (3.15)

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & Ax = b \end{array} \tag{6.1}$$

where $f_i : \mathbf{R}^n \to \mathbf{R}$ and $A \in \mathbf{R}^{p \times n}$. We assume the functions f_i are convex and have continuous second derivatives (which implies that the domain D of the problem is open).

The methods we discuss in this chapter are called *sequential unconstrained minimization methods*, or *interior-point methods*. They are based on using smooth minimization techniques, often Newton's method, to solve or approximately solve a sequence of smooth unconstrained (or equality-constrained) problems. Provided the sequence of smooth problems is chosen appropriately, the resulting algorithms are very efficient both in theory and practice. The term 'interior-point' refers to the fact the iterates $x^{(k)}$ are all strictly feasible, *i.e.*, they satisfy $Ax^{(k)} = b$, $f_i(x^{(k)}) < 0$, $i = 1, \ldots, m$, and therefore they are in the relative interior of the feasible set. (Of course, they converge to optimal points which are usually not strictly feasible.) In this chapter we study one family of interior-point methods, called *sequential unconstrained minimization*, *barrier methods,* or *path-following methods*. Other interior-point methods are related, and are based on the same ideas and concepts.

6.1.1 Initial point and sublevel set

We assume that a strictly feasible point $x^{(0)}$ is given, *i.e.*, $x^{(0)} \in D$, $Ax^{(0)} = b$ and $f_i(x^{(0)}) < 0$, $i = 1, \ldots, m$. We discuss techniques for computing such a strictly feasible point, if one is not already known, in §6.4.

Existence of a strictly feasible point implies that Slater's constraint qualification holds, and, as mentioned in chapter 4, it implies that the KKT conditions are necessary and sufficient for optimality: a point x is optimal if and only if it is feasible and there exists a $\lambda \succeq 0$

and ν such that

$$\nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + A^T\nu = 0, \quad \lambda_i f_i(x) = 0, \ \ i = 1, \ldots, m. \tag{6.2}$$

We will see that the barrier method has a natural interpretation as an algorithm for solving the KKT conditions.

We also make two assumptions that may require some reformulation of the problem. First, we assume that the sublevel set

$$\left\{x \mid f_0(x) \le f_0(x^{(0)}), \quad f_i(x) \le 0, \ i = 1, \ldots, m, \ Ax = b\right\}$$

is bounded. If this assumption is not satisfied, we can add an extra constraint such as $x^T x \le R^2$, where R is sufficiently large (and known to be larger than $\|x^\star\|$).

The second assumption is that every strictly feasible point is in the domain of the objective, *i.e.*,

$$\{x \mid f_i(x) < 0, \quad i = 1, \ldots, m, \quad Ax = b\} \subseteq \mathbf{dom}\, f_0. \tag{6.3}$$

This assumption may require adding extra constraints to make the implicit constraint $x \in \mathbf{dom}\, f_0$ explicit.

Example. The entropy maximization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} x_i \log x_i \\ \text{subject to} & Ax = b \end{array} \tag{6.4}$$

has domain $D = \mathbf{dom}\, f_0 = \{x \mid x \succ 0\}$. The condition (6.3) reduces to

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

which is only satisfied when $x^{(0)}$ is the unique solution of $Ax = b$. In general, we have to make the implicit constraint $x \in \mathbf{dom}\, f_0$ explicit by considering the equivalent problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} x_i \log x_i \\ \text{subject to} & x \succeq 0 \\ & Ax = b. \end{array} \tag{6.5}$$

Example. If the objective function or some of the constraint functions are not differentiable, the methods in this chapter do not apply, and we have to reformulate the problem. As an example, consider the ℓ^p-approximation problem

$$\text{minimize } \|Ax + b\|_p = \left(\sum_{i=1}^{m} |a_i^T x + b_i|^p\right)^{1/p}$$

where $p \ge 1$. We can reformulate this problem as

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{m} y_i^p \\ \text{subject to} & -y \preceq Ax + b \preceq y \end{array}$$

with variables x, y and $\mathbf{dom}\, f_0 = \{\ y \mid y \succ 0\ \}$. The objective and constraint functions in this second problem are smooth and convex, so our methods apply.

6.1.2 Idea of barrier method

Let us first rewrite the standard problem (6.1) as

$$\begin{array}{ll} \text{minimize} & f_0(x) + I_{\text{feas}}(x) \\ \text{subject to} & Ax = b \end{array} \tag{6.6}$$

where $I_{\text{feas}} : \mathbf{R}^n \to \mathbf{R}$ is defined as

$$I_{\text{feas}}(x) = \begin{cases} 0 & f_i(x) \le 0, \quad i = 1, \ldots, m \\ +\infty & \text{otherwise.} \end{cases}$$

(I_{feas} is called the *indicator function* of the set $\{\ x \mid f_i(x) \le 0, \quad i = 1, \ldots, m\}$.) Of course we have only reformulated the problem; we have not made it any easier to solve. Although the problem (6.6) has no inequality constraints, its objective function is not differentiable so none of the methods of chapter 5 can be applied.

The basic idea of barrier methods is to *approximate* the indicator function I_{feas} with a function $\widehat{I} : \mathbf{R}^n \to \mathbf{R}$ that has the following properties:

- $\widehat{I}$ is convex and smooth
- $\mathbf{dom}\,\widehat{I} = \{x \mid f_i(x) < 0, \quad i = 1, \ldots, m\}$
- If $x^{(k)} \in \mathbf{dom}\,\widehat{I}$, $x^{(k)} \to \partial\,\mathbf{dom}\,\widehat{I}$, then $\widehat{I}(x^{(k)}) \to \infty$, *i.e.*, $\widehat{I}(x)$ grows without bound as x approaches the relative boundary of the feasible set.

We refer to such a function as a *barrier function* for the problem (6.1).

Substituting the barrier function $\widehat{I}$ for the indicator fucntion I_{feas}, we obtain the problem

$$\begin{array}{ll} \text{minimize} & f_0(x) + \widehat{I}(x) \\ \text{subject to} & Ax = b \end{array} \tag{6.7}$$

which is smooth, convex, and has only equality constraints. It also satisfies the assumption at the beginning of §5.7: we know a point $x^{(0)} \in \mathbf{dom}(f_0 + \widehat{I})$ that satisfies $Ax^{(0)} = b$ and such that the sublevel set

$$S = \{x \mid f_0(x) + \widehat{I}(x) \le f_0(x^{(0)}) + \widehat{I}(x^{(0)}), \quad Ax = b\}$$

is closed. (This follows from the fact that $f_0(x) + \widehat{I}(x) \to \infty$ as $x \to \partial\,\mathbf{dom}\,\widehat{I}$.) Therefore we can solve the modified problem (6.7) using the methods of chapter 5.

When a descent method is applied to the problem (6.7), starting at a strictly feasible point $x^{(0)}$, successive iterates are all strictly feasible, since $\widehat{I}$ increases without bound as x approaches the boundary of $\mathbf{dom}\,\widehat{I}$. Thus $\widehat{I}$ prevents the iterates from leaving the (strictly) feasible set. This property gives the name 'barrier' to $\widehat{I}$.

Of course, the problem (6.7) is only an approximation of the original problem (6.6), so one question that arises immediately is how well a solution $\widehat{x}$ of (6.7) approximates a solution of the original problem (6.6). Intuition suggests that the quality of the approximation depends on how well the barrier function $\hat{I}$ approximates the indicator function I_{feas}. If $\widehat{I}$ is near zero over most of the strictly feasible set and rises rapidly only near the boundary, then, it seems, $\widehat{x}$ should be a good approximate solution of the original problem.

By scaling the barrier function we can make it approximate $I_{\rm feas}$ as well as we like, except on the boundary of $\mathbf{dom}\, I_{\rm feas}$. For any $t > 0$, the function $(1/t)\widehat{I}$ is also a barrier for the problem. The larger the value of t, the better (at least intuitively) the barrier function $(1/t)\widehat{I}$ approximates the indicator function $I_{\rm feas}$. Certainly, the functions $(1/t)\widehat{I}$ converge pointwise to the indicator function of the strictly feasible set as $t \to \infty$. This suggests that by solving the smooth unconstrained problem

$$\begin{array}{ll} \text{minimize} & f_0(x) + (1/t)\widehat{I}(x) \\ \text{subject to} & Ax = b \end{array} \tag{6.8}$$

for t large enough will yield a good approximate solution of the original problem. We will soon see that this is correct.

On the other hand, by making the parameter t large, the function $f_0 + (1/t)\widehat{I}$ becomes 'harder' to minimize by smooth unconstrained methods. When t is large, its Hessian varies rapidly near the boundary, which suggests that Newton's method will require a large number of iterations to converge. This intuition is confirmed in practice: using Newton's method to minimize $f_0 + (1/t)\widehat{I}$ for some large value of t generally requires a large number of iterations.

We can develop a more sophisticated approach by noting that (6.8) defines a family of smooth unconstrained minimization problems, indexed by the parameter t. Letting $\widehat{x}(t)$ denote the solution of the smooth unconstrained problem (6.8), we expect (and will soon confirm) that the curve $\hat{x}(t)$ converges to an optimal point of the original problem as t increases to ∞. We can therefore employ a *homotopy* or *curve following* method: we solve for $\widehat{x}(t)$ for a sequence of increasing values of t, using the last iterate as the starting point. Our hope is that when t is large we can solve the 'hard' problem of minimizing $f_0 + (1/t)\widehat{I}$ because we have a good starting point, *i.e.*, the minimizer of this function for a slightly smaller value of t.

It turns out that, provided the barrier satisfies certain properties, and the parameter t is updated appropriately, this path following approach works very well, both in theory and practice.

Example. This simple idea is illustrated in figure 6.1, for the problem

$$\begin{array}{ll} \text{minimize} & x^2 + 1 \\ \text{subject to} & 2 \le x \le 4, \end{array}$$

which has optimal point $x^\star = 2$. We use the barrier function

$$\widehat{I}(x) = \begin{cases} -\log(x-2) - \log(4-x) & \text{if } 2 < x < 4 \\ +\infty & \text{otherwise.} \end{cases}$$

Figure 6.1 shows the function $f_0 + (1/t)\widehat{I}$ for a sequence of positive values of t. For fixed t, $f_0 + (1/t)\widehat{I}$ is a smooth function with the open interval $(2, 4)$ as its domain. For $t = 1$, the minimum is attained at $x = 3$; as t increases, the minimizer moves towards the optimal point $x^\star = 2$. No matter how large the parameter t is, the minimizer of $f_0 + (1/t)\widehat{I}$ is always in the open interval $(2, 4)$ since $(1/t)\widehat{I}$ converges to ∞ as x approaches the boundary points 2 and 4.

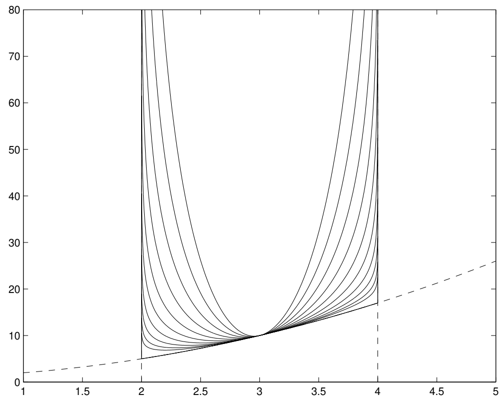

Figure 6.1: The function $f_0 + (1/t)\widehat{I}$ for $f_0(x) = x^2 + 1$, with barrier function $\widehat{I}(x) = -\log(x-2) - \log(4-x)$, and different values of t.

6.2 Logarithmic barrier function and central path

6.2.1 Logarithmic barrier

The most widely used barrier function is the *logarithmic barrier*, defined as the function $\phi : \mathbf{R}^n \to \mathbf{R}$, with

$$\phi(x) = \begin{cases} -\sum_{i=1}^m \log(-f_i(x)) & \text{if } f_i(x) < 0, \quad i = 1, \ldots, m \\ +\infty & \text{otherwise.} \end{cases}$$

It is clear that ϕ is a barrier: it is smooth and convex on its domain

$$\mathbf{dom}\, \phi = \{x \mid f_i(x) < 0, \quad i = 1, \ldots, m\},$$

and it tends to infinity as x approaches the boundary of $\mathbf{dom}\, \phi$.

For future reference, the gradient and Hessian of ϕ are given by

$$\nabla \phi(x) = \sum_{i=1}^m \frac{1}{-f_i(x)} \nabla f_i(x), \qquad \nabla^2 \phi(x) = \sum_{i=1}^m \frac{1}{f_i(x)^2} \nabla f_i(x) \nabla f_i(x)^T + \sum_{i=1}^m \frac{1}{-f_i(x)} \nabla^2 f_i(x). \tag{6.9}$$

There are many other ways to form a barrier for the problem. For example the *inverse barrier* is given by

$$\phi_{\text{inv}}(x) = \begin{cases} \sum_{i=1}^m -1/f_i(x) & f_i(x) < 0, \quad i = 1, \ldots, m \\ \infty & \text{otherwise} \end{cases}$$

Later in this chapter we will see that the logarithmic barrier has some important advantages over others, at least for some common forms of constraint functions f_i (including linear and quadratic functions).

6.2.2 Central path

We now consider in more detail the minimization problem

$$\begin{array}{ll} \text{minimize} & f_0(x) + (1/t)\phi(x) \\ \text{subject to} & Ax = b, \end{array}$$

where $t > 0$, that arises in the barrier method. It will simplify notation later on if we multiply the objective by t, and consider the problem

$$\begin{array}{ll} \text{minimize} & tf_0(x) + \phi(x) \\ \text{subject to} & Ax = b \end{array} \tag{6.10}$$

which obviously has the same minimizers. It can be shown that the optimal value of (6.10) is finite and attained, under our assumption that the sublevel sets of (6.1) are bounded.

We assume that for $t > 0$, the Hessian of the function $tf_0 + \phi$, *i.e.*,

$$t\nabla^2 f_0 + \sum_{i=1}^m \frac{1}{f_i(x)^2}\nabla f_i(x)\nabla f_i(x)^T + \sum_{i=1}^m \frac{1}{-f_i(x)}\nabla^2 f_i(x),$$

is positive definite everywhere on the strictly feasible set, and as a consequence, the solution of (6.10) is unique for all $t > 0$. This assumption is not crucial; most of the theory can be developed without it. It can also be enforced by adding a constraint that bounds x, such as $x^T x \le R^2$, where R is large. This constraint adds a term $-\log(R^2 - x^T x)$ to the logarithmic barrier, which contributes the terms

$$\frac{2}{R^2 - x^T x} I + \frac{4}{(R^2 - x^T x)^2} x x^T \succeq (2/R^2) I$$

to the Hessian of $tf_0 + \phi$. Thus it guarantees strong convexity of the augmented function $tf_0 + \phi$, since we have

$$\nabla^2(tf_0 + \phi) \succeq (2/R^2) I.$$

For $t > 0$ we define $x^\star(t)$ as the minimizer of the augmented function, *i.e.*,

$$x^\star(t) = \mathop{\rm argmin}_{Ax=b} \left(tf_0(x) + \phi(x)\right). \tag{6.11}$$

The *central path* associated with problem (6.1) is defined as the set of points $x^\star(t)$, $t > 0$, which we call the *central points*. We will soon see that as t increases (which puts more weight on the objective term compared to the barrier term), $f_0(x^\star(t))$ approaches $p^\star$.

Points on the central path are characterized by the following necessary and sufficient conditions: $x^\star(t)$ is strictly feasible, and there exists a $\hat{\nu} \in \mathbf{R}^p$ such that

$$\begin{aligned} t\nabla f_0(x^\star(t)) + \nabla\phi(x^\star(t)) + A^T\hat{\nu} &= t\nabla f_0(x^\star(t)) + \sum_{i=1}^m \frac{1}{-f_i(x)}\nabla f_i(x) + A^T\hat{\nu} \\ &= 0. \end{aligned} \tag{6.12}$$

Example. *Linear programming.* We consider an LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \end{array} \tag{6.13}$$

where $A = [a_1 \cdots a_m]^T \in \mathbf{R}^{m \times n}$. The logarithmic barrier function is defined as

$$\phi(x) = -\sum_{i=1}^{m} \log(b_i - a_i^T x).$$

Its gradient and hessian are

$$\nabla \phi(x) = \sum_{i=1}^{m} \frac{1}{b_i - a_i^T x} a_i, \quad \nabla^2 \phi(x) = \sum_{i=1}^{m} \frac{1}{(b_i - a_i^T x)^2} a_i a_i^T$$

or, more compactly,

$$\nabla \phi(x) = A^T d, \quad \nabla^2 \phi(x) = A^T \, \mathbf{diag}(d)^2 A,$$

where d is defined as

$$d = (1/(b_1 - a_1^T x), 1/(b_2 - a_2^T x), \ldots, 1/(b_m - a_m^T x)).$$

If x is strictly feasible, then $d \succ 0$, and the Hessian of ϕ is nonsingular if and only if A has rank n.

The optimality condition (6.12) reduces to

$$tc + \sum_{i=1}^{m} \frac{1}{b_i - a_i^T x} a_i = tc + A^T d = 0. \tag{6.14}$$

The rank condition on A is not at all restrictive. Assume A does not have full rank, *i.e.*, there exists a nonzero v, such that $Av = 0$. There are two possibilities. If $c^T v \neq 0$ and the LP is feasible, then it is unbounded below: if x is feasible, then $A(x + tv) \preceq b$ for all t, while $c^T(x + tv) \to -\infty$ as t goes to plus infinity (if $c^T v < 0$) or minus infinity (if $c^T v > 0$).

If on the other hand $c^T v = 0$, then we can eliminate one of the variables. Suppose $v_j \neq 0$. Then we can assume without loss of generality that $x_j = 0$: if x is feasible, then $\tilde{x} = x - (x_j/v_j)v$ satisfies $Ax = A\tilde{x}$, $c^T x = c^T \tilde{x}$, and $\tilde{x}_j = 0$, *i.e.*, $\tilde{x}$ is feasible with the same objective value as x, and its jth component is zero.

In other words, if A does not have full rank, then the problem is either unbounded below, or we can delete linearly dependent columns of A and the corresponding coefficients of c until A is full rank.

Example. *Entropy maximization.* The logarithmic barrier function for (6.5) is

$$\phi(x) = -\sum_{i=1}^{n} \log x_i,$$

and its gradient and Hessian are

$$\nabla\phi(x) = -(1/x_1, \ldots, 1/x_n), \quad \nabla^2\phi(x) = \mathbf{diag}(1/x_1, \ldots, 1/x_n)^2.$$

The gradient and Hessian of the objective function are

$$\nabla f_0(x) = (1 + \log x_1, \ldots, 1 + \log x_n), \quad \nabla^2 f_0(x) = \mathbf{diag}(1/x_1, \ldots, 1/x_n).$$

It is clear that $t\nabla^2 f_0(x) + \nabla^2\phi(x) \succ 0$ for all feasible x, since

$$t\nabla^2 f_0(x) + \nabla^2\phi(x) \succeq \mathbf{diag}(1/x_1, \ldots, 1/x_n)^2 \succ 0.$$

Points $x^\star(t)$ on the central path are characterized by the conditions $x^\star(t) \succ 0$, $Ax^\star(t) = b$ and

$$t(1 + \log x_i^\star(t)) - 1/x_i^\star(t) + a_i^T\hat{\nu} = 0, \quad i = 1, \ldots, n, \tag{6.15}$$

for some $\hat{\nu} \in \mathbf{R}^p$, where a_i is the ith column of A.

Dual points from central path

From (6.12) we can derive an important property of the central path. Every central point yields dual feasible points, and hence a lower bound on the optimal value $p^\star$. More specifically, define

$$\lambda_i^\star(t) = -\frac{1}{tf_i(x^\star(t))}, \quad i = 1, \ldots, m, \quad \nu^\star(t) = \hat{\nu}/t. \tag{6.16}$$

We claim that $\lambda^\star(t)$ and $\nu^\star(t)$ are dual feasible.

First, it is clear that $\lambda^\star(t) \succ 0$ because $f_i(x^\star(t)) < 0$, $i = 1, \ldots, m$. By writing out the optimality conditions (6.12) as

$$\nabla f_0(x^\star(t)) + \sum_{i=1}^m \lambda_i^\star(t)\nabla f_i(x^\star(t)) + A^T\nu^\star(t) = 0,$$

we also see that $x^\star(t)$ minimizes the Lagrangian

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \nu^T(Ax - b),$$

for $\lambda = \lambda^\star(t)$ and $\nu = \nu^\star(t)$. Therefore $g(\lambda^\star(t), \nu^\star(t))$ is finite, and

$$\begin{aligned} g(\lambda^\star(t), \nu^\star(t)) &= f_0(x^\star(t)) + \sum_{i=1}^m \lambda_i^\star(t) f_i(x^\star(t)) + \nu^\star(t)^T(Ax^\star(t) - b) \\ &= f_0(x^\star(t)) - m/t, \end{aligned}$$

so the duality gap associated with $x^\star(t)$ and the dual feasible point $\lambda^\star(t)$, $\nu^\star(t)$ is simply m/t. As an important consequence we have

$$f_0(x^\star(t)) - f^\star \leq m/t,$$

which confirms the intuitive idea that $x^\star(t)$ converges to an optimal point as $t \to \infty$.

Example. *Linear programming.* The dual of the LP (6.13) is

$$\begin{array}{ll} \text{maximize} & -b^T z \\ \text{subject to} & A^T z + c = 0 \\ & z \succeq 0. \end{array}$$

From the optimality conditions (6.14), it is clear that

$$z_i^\star(t) = \frac{1}{t(b_1 - a_1^T x^\star(t))}, \quad i = 1, \ldots, m,$$

is dual feasible and that

$$-b^T z^\star(t) = c^T x^\star(t) + (Ax^\star(t) - b)^T z^\star(t) = c^T x^\star(t) - m/t.$$

Example. *Entropy maximization.* The dual of the entropy maximization problem (6.5) is

$$\begin{array}{ll} \text{maximize} & -b^T \nu + \sum_{i=1}^n e^{-1-a_i^T \nu + \lambda_i} \\ \text{subject to} & \lambda \succeq 0, \end{array}$$

where a_i is the ith column of A. The variable ν is the dual variable associated with the equality constraint $Ax = b$; λ is the dual variable associated with the inequality constraint $x \succeq 0$.

Points on the central path are characterized by (6.15). If we define $\lambda_i^\star(t) = 1/(tx_i^\star(t))$ and $\nu^\star(t) = \hat{\nu}/t$, we can write those conditions as

$$\log x_i^\star(t) = -1 - a_i^T \nu^\star(t) + \frac{1}{tx_i^\star(t)},$$

and therefore

$$\begin{array}{rcl} \sum_{i=1}^n x_i^\star(t) \log x_i^\star(t) & = & -\sum_{i=1}^n x_i^\star(t) - b^T \nu^\star(t) - n/t \\ & = & \sum_{i=1}^n e^{-1-a_i^T \nu^\star(t) + \lambda_i^\star(t)} - b^T \nu^\star(t) - n/t, \end{array}$$

i.e., the duality gap associated with between $x^\star(t)$ and $\lambda^\star(t)$, $\nu^\star(t)$ is equal to n/t.

Interpretation via KKT conditions

We can also interpret the central path conditions (6.12) as a continuous deformation of the KKT optimality conditions for problem (6.1). We have seen in chapter 4 that a point x solves the optimization problem (6.1) if and only if there exists $\lambda \in \mathbf{R}^m$, $\nu \in \mathbf{R}^p$ such that

$$\begin{array}{l} Ax = b, \quad f_i(x) \le 0, \quad \lambda \succeq 0 \\ \nabla f_0(x) + \sum_{i=1}^m \lambda_i \nabla f_i(x) + A^T \nu = 0 \\ -\lambda_i f_i(x) = 0, \quad i = 1, \ldots, m. \end{array} \qquad (6.17)$$

The centrality condition (6.12) can be written in a very similar form: a point x is equal to $x^\star(t)$ if and only if there exists λ, ν such that

$$\begin{array}{l} Ax = b, \quad f_i(x) \le 0, \quad \lambda \succeq 0 \\ \nabla f_0(x) + \sum_{i=1}^m \lambda_i \nabla f_i(x) + A^T \nu = 0 \\ -\lambda_i f_i(x) = 1/t, \quad i = 1, \dots, m. \end{array} \tag{6.18}$$

The only difference between (6.17) and (6.18) is that the complementarity condition $-\lambda_i f_i(x) = 0$ is replaced by the condition $-\lambda_i f_i(x) = 1/t$. In particular, for large t, $x^\star(t)$ and the associated dual point $\lambda^\star(t)$, $\nu^\star(t)$ 'almost' satisfy the KKT optimality conditions for (6.1).

Force field interpretation

We can give a simple mechanical interpretation of the central path in terms of potential forces acting on a particle in the strictly feasible set C. For simplicity we assume that there are no equality constraints.

We associate with each constraint the force

$$F_i(x) = \nabla \left(-\log(-f_i(x)) \right) = \frac{1}{-f_i(x)} \nabla f_i(x)$$

acting on the particle when it is at position x. The potential associated with the total force field generated by the constraints is ϕ. As the particle moves towards the boumndary of the feasible set, it is strongly repelled by the forces generated by the constraints.

Now we imagine another force acting on the particle, given by

$$F_0(x) = -t \nabla f_0(x),$$

when the particle is at position x. This objective force field acts to pull the particle in the negative gradient direction, *i.e.*, toward smaller f_0. The parameter t scales the objective force, relative to the constraint forces.

The central point $x^*(t)$ is the point where the constraint forces exactly balance the objective force felt by the particle. As the parameter t increases, the particle is more strongly pulled towards the optimal point, but it always trapped in C by the barrier potential, which becomes infinite as the particle approaches the boundary.

6.3 Sequential unconstrained minimization method

6.3.1 Unconstrained minimization method

We have seen that the point $x^\star(t)$ is m/t-suboptimal, and that a certificate of this accuracy is provided by the dual feasible points $\lambda^\star(t)$, $\nu^\star(t)$. This suggests a very straightforward method for solving the original problem (6.1) with a guaranteed specified accuracy ϵ: we simply take $t = m/\epsilon$ and solve the unconstrained problem

$$\begin{array}{ll} \text{minimize} & (m/\epsilon) f_0(x) + \phi(x) \\ \text{subject to} & Ax = b \end{array}$$

using, for example, Newton's method. This method could be called the *unconstrained minimization method*, since it allows us to solve the constrained optimization problem (6.1) to a guaranteed accruacy by solving an unconstrained problem.

Although this method can work well for very small problems, good starting points, and moderate accuracy (*i.e.*, ϵ not too small), it does not work well in other cases. As a result it is rarely, if ever, used.

6.3.2 Sequential unconstrained minimization method

A simple extension of the unconstrained minimization method does work well. It is based on solving the constrained problem (6.1) by solving a sequence of unconstrained minimization problems, using the last point found as the starting point for the next unconstrained minimization problem. In other words, we compute $x^\star(t)$ for a sequence of increasing values of t, until $t \geq m/\epsilon$, which guarantees that we have an ϵ-suboptimal solution of the original problem. This method is called the *sequential unconstrained minimization method*, since it solves the constrained problem (6.1) by solving a sequence of smooth unconstrained problems. When the method was first proposed by Fiacco and McCormick in the 1960s, it was called the *sequential unconstrained minimization technique* (SUMT).

A simple version of the method is as follows.

SUMT

given strictly feasible x, $t := t^{(0)} > 0$, $\mu > 1$, $\epsilon > 0$.
repeat {
1. Compute $x^\star(t)$ by minimizing $tf_0 + \phi$, starting at x.
2. $x := x^\star(t)$.
3. If $m/t < \epsilon$, return(x).
4. $t := \mu t$.

}

At each iteration (except the first one) we compute the central point $x^\star(t)$ starting from the previously computed central point, and then increase t by a factor $\mu > 1$. The algorithm can also return $\lambda = \lambda^\star(t)$, and $\nu = \nu^\star(t)$, a dual ϵ-suboptimal point, or certificate for x.

We will refer to each execution of step 1 as a *centering step* (since a central point is being computed) or an *outer iteration*. We refer to the iterations or steps executed during the centering step (*e.g.*, Newton steps) as *inner iterations*.

We should make some comments on the accuracy to which we solve the centering problems. Computing $x^\star(t)$ exactly is not necessary since the central path has no significance beyond the fact that it leads to a solution of the original problem as $t \to \infty$; inexact centering will still yield a sequence of points $x^{(k)}$ that converges to an optimal point. Inexact centering, however, means that the points $\lambda^\star(t)$, $\nu^\star(t)$, computed from (6.16), are not exactly dual feasible. This can be corrected by adding a small correction term to the formula (6.16), which yields a dual feasible point provided the computed x is near the central path, *i.e.*, $x^\star(t)$.

On the other hand if Newton's method is used for the centering steps, then the cost of computing an *extremely accurate* minimizer of $tf_0 + \phi$, as compared to the cost of computing a *good* minimizer of $tf_0 + \phi$, is only marginally more, *i.e.*, a few Newton steps at most. For this reason it is not unreasonable to assume exact centering, at least when Newton's method is used for the minimizations.

We will refer to the first centering step (the computation of $x^\star(t^{(0)})$) as the *initial centering step*. Each subsequent centering step computes the central point $x^\star(\mu^k t^{(0)})$ for $k = 1, 2, \ldots$, so the duality gap after $k+1$ centering steps is $m/(\mu^k t^{(0)})$. Therefore the desired accuracy ϵ is achieved after *exactly*

$$\left\lceil \frac{\log(\epsilon t^{(0)}/m)}{\log \mu} \right\rceil$$

centering steps, plus the initial centering step.

Tradeoff in choice of μ

The choice of the parameter μ involves a tradeoff. If μ is small (*i.e.*, near 1) then at each outer iteration t increases by a small factor. As a result the initial point for the Newton process, *i.e.*, the previous iterate x, is a very good starting point, and the number of Newton steps needed to compute the next iterate is small. Thus for small μ we expect a small number of Newton steps per outer iteration, but of course a large number of outer iterations since each outer iteration reduces the gap by only a small amount. In this case the iterates (including even the iterates of the inner iterations) closely follow the central path. In this situation the sequential unconstrained minimization method is called a 'path-following' algorithm since the iterates closely follow the central path toward the solution.

On the other hand if μ is large we have the opposite situation. After each outer iteration t increases a large amount, so the current iterate is probably not a very good approximation of the next iterate. Thus we expect many more inner iterations. This 'aggressive' updating of t results in fewer outer iterations, since the duality gap is reduced by the large factor μ at each outer iteration, but more inner iterations. With μ large, the iterates are widely separated on the central path; the inner iterates veer way off the central path.

This tradeoff in the choice of μ is confirmed both in practice and, as we will see, in theory.

Choice of $t^{(0)}$

Another important issue is the choice of initial value of t. Here the tradeoff is simple: if t is chosen too large, the first inner iteration will require too many iterations. If t is chosen too small, the algorithm will require extra outer iterations.

Since $m/t^{(0)}$ is the duality gap that will result from the first centering step, one reasonable choice is to choose $t^{(0)}$ so that $m/t^{(0)}$ is approximately of the same order as $f_0(x^{(0)}) - f^\star$. For example, if a dual feasible point λ, ν is known, with duality gap $\eta = f_0(x^{(0)}) - g(\lambda, \nu)$, then we can take $t^{(0)} = m/\eta$. Thus, in the first outer iteration we simply compute a pair with the same duality gap as the initial primal and dual feasible points.

Another possibility is suggested by the central path condition (6.12). We can interpret

$$\inf_\nu \left\| t\nabla f_0(x^{(0)}) + \nabla\phi(x^{(0)}) + A^T\nu \right\|, \tag{6.19}$$

as a measure for the deviation of $x^{(0)}$ to the point $x^\star(t)$, and choose for $t^{(0)}$ the value that minimizes (6.19).

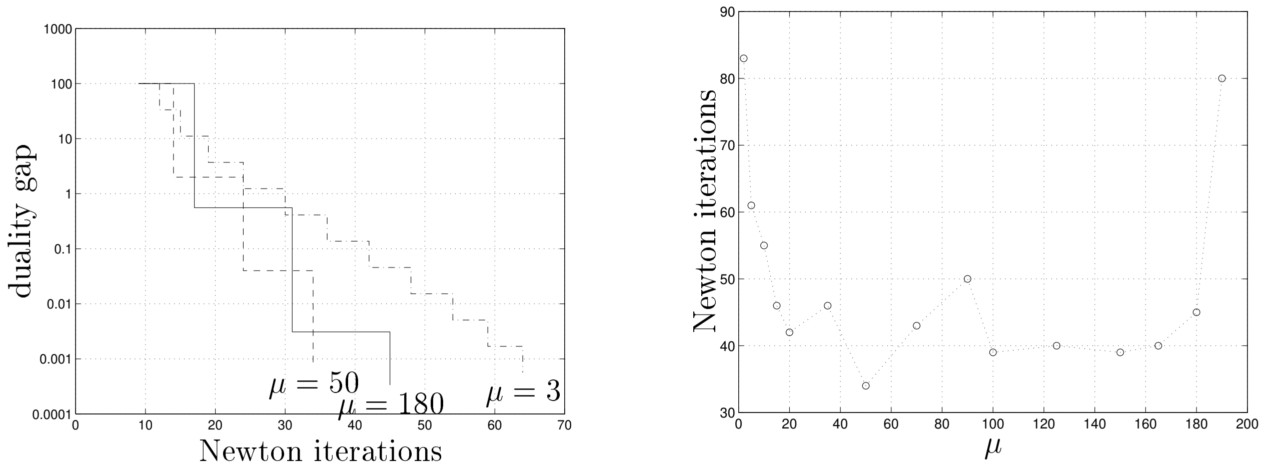

Figure 6.2: Example illustrating the tradeoff in the choice of μ.

6.3.3 Examples

Linear programming

We consider an LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \end{array}$$

with $A \in \mathbf{R}^{100\times 50}$.

The lefthand plot in figure 6.2 shows duality gap versus inner iterations, for several values of μ. The righthand figure shows the total number of Newton steps required to reduce the duality gap by a factor of 10^5, as a function of the parameter μ.

Geometric programming

The logarithmic barrier for the geometric program

$$\begin{array}{ll} \text{minimize} & \log \sum_{j=1}^{k_0} \exp(a_{0j}^T x + b_{0j}) \\ \text{subject to} & \log \sum_{j=1}^{k_i} \exp(a_{ij}^T x + b_{ij}) \leq 0, \quad i = 1, \ldots, m \end{array} \tag{6.20}$$

is

$$\phi(x) = -\sum_{i=1}^{m} \log \left(-\log \sum_{j=1}^{k_i} \exp(a_{ij}^T x + b_{ij}) \right). \tag{6.21}$$

Second-order cone programming

Consider the SOCP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \|A_i x + b_i\| \leq c_i^T x + d_i, \quad i = 1, \ldots, N. \end{array}$$

The constraint functions are not differentiable, so we cannot directly apply the barrier method. To do so we will reformulate the problem as

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \dfrac{\|A_i x + b_i\|^2}{c_i^T x + d_i} \leq c_i^T x + d_i, \quad i = 1, \ldots, N. \\ & c_i^T x + d_i \geq 0, \quad i = 1, \ldots, N, \end{array}$$

in which the constraint functions are smooth. The associated logarithmic barrier function is

$$\begin{aligned} \phi(x) &= -\sum_{i=1}^N \log(c_i^T x + d_i) - \sum_{i=1}^N \log\left(c_i^T x + d_i - \frac{\|A_i x + b_i\|^2}{c_i^T x + d_i}\right) \\ &= \sum_{i=1}^N \log((c_i^T x + d_i)^2 - \|A_i x + b_i\|^2). \end{aligned} \tag{6.22}$$

6.4 Feasibility and phase-I methods

SUMT requires a strictly feasible starting point $x^{(0)}$. In some cases such a starting point is readily available. In most cases, however, the method has to be preceded by a preliminary stage (referred to as *phase I*) to compute a starting point.

In this section we first describe and analyze a simple method for solving a set of convex inequalities $f_i(x) < 0$. We then describe a phase-I method for problem (6.1). The phase-I method computes a point that is not only strictly feasible, but lies on the central path, which allows us to skip the initial centering phase in SUMT.

6.4.1 Feasibility problems

We consider a set of inequalities and equalities

$$\begin{array}{l} f_i(x) \leq 0, \quad i = 1, \ldots, m \\ Ax = b \end{array} \tag{6.23}$$

where $f_i : \mathbf{R}^n \to \mathbf{R}$ are convex functions with continous second derivatives. We assume that the feasible set

$$\{\ x \mid f_i(x) \leq 0, i = 1, \ldots, m, \quad Ax = b\ \}. \tag{6.24}$$

is bounded, and that we are given a point $x^{(0)} \in D$, with $Ax^{(0)} = b$.

We are interested in finding a strictly feasible solution or determining that none exists. To achieve that goal we can form the following optimization problem:

$$\begin{array}{ll} \text{minimize} & s \\ \text{subject to} & f_i(x) \leq s, \quad i = 1, \ldots, m \\ & Ax = b \end{array} \tag{6.25}$$

in the variables x, s. This problem is always strictly feasible, since we can choose $x = x^{(0)}$, $s > \max_{i=1,\ldots,m} f_i(x^{(0)})$. Problem (6.25) also satisfies all the assumptions mentioned in §6.1. In particular, boundedness of (6.24) implies that the initial sublevel set

$$\{\ (x, s) \mid f_i(x) \leq s,\ Ax = b,\ s \leq s^{(0)}\}$$

of (6.25) is bounded. We can therefore apply SUMT to solve (6.25)

We can distinguish three cases depending on the sign of the optimal value $\bar{p}^\star$ of (6.25).

- If $\bar{p}^\star < 0$, then (6.23) has a strictly feasible solution. Moreover if (x, s) is feasible in (6.25) with $s < 0$, then x satisfies $f_i(x) < 0$. This means we do not need to solve the optimization problem (6.25) with high accuracy: we can terminate when $s < 0$.
- If $\bar{p}^\star > 0$, then (6.23) is infeasible.
- If $\bar{p}^\star = 0$ and the minimum is attained at $x^\star$ and $s^\star = 0$, then the set of inequalities is feasible, but not strictly feasible. If $\bar{p}^\star = 0$ and the minimum is not attained, then the inequalities are infeasible.

  In practice it is impossible to determine exactly that $\bar{p}^\star = 0$. Instead, an optimization algorithm applied to (6.25) will terminate with the conclusion that $|\bar{p}^\star| < \epsilon$ for some small, positive ϵ. This allows us to conclude that the inequalities $f_i(x) \leq -\epsilon$ are infeasible, while the inequalities $f_i(x) \leq \epsilon$ are feasible.

Obviously there are many variations on the idea we just described. For example, we can form the auxiliary problem

$$\begin{array}{ll} \text{minimize} & s \\ \text{subject to} & f_i(x) \leq 0, \quad i \in I \\ & f_i(x) \leq s, \quad i \notin I \\ & Ax = b. \end{array}$$

where $f_i(x^{(0)}) < 0$ for $i \in I \subseteq \{1, \ldots, m\}$.

6.4.2 Phase I

Suppose we want to apply SUMT to (6.1), but we do not know a suitable starting point, so we first have to find a strictly feasible solution of (6.23). Then the following variation of the method of the previous paragraph is useful.

We assume a point $x^{(0)} \in D$, with $Ax^{(0)} = b$ is given. We form an auxiliary problem as in the previous paragraph, but add an inequality $f_0(x) \leq M$

$$\begin{array}{ll} \text{minimize} & s \\ \text{subject to} & f_i(x) \leq s, \quad i = 1, \ldots, m \\ & f_0(x) \leq M \\ & Ax = b, \end{array} \tag{6.26}$$

where $M > \max\{f_0(x^{(0)}), p^\star\}$. The bounded sublevel set condition of §6.1 implies that the initial sublevel set

$$\{\ (x, s) \mid f_0(x) \leq M,\ f_i(x) \leq s, i = 1, \ldots, m,\ Ax = b\ \}$$

of (6.26) is bounded, so the SUMT method applies.

The central path of (6.26) has the interesting property that it intersects the central path of (6.1). This follows from the optimality conditions that characterize the central path of (6.26):

$$\sum_{i=1}^{m} \frac{1}{s - f_i(x)} = t$$

and

$$\frac{1}{A - f_0(x)} \nabla f_0(x) + \sum_{i=1}^{m} \frac{1}{s - f_i(x)} \nabla f_i(x) = 0.$$

If (x, s) is on the central path and $s = 0$, then x also minimizes

$$t f_0 - \sum_i \log(-f_i)$$

for $t = 1/(A - f_0(x))$.

6.5 Complexity analysis of SUMT

6.5.1 Self-concordance assumption

We now analyze the complexity of the SUMT algorithm and show that under certain conditions on the functions f_i it leads to a method with polynomial complexity.

We have seen that the number of centering steps needed to achieve a required duality gap ϵ, is exactly equal to

$$\left\lceil \frac{\log(m/t^{(0)}\epsilon)}{\log \mu} \right\rceil \tag{6.27}$$

plus one initial centering step. In order to estimate the total complexity, we need an upper bound on the amount of work per centering step. We will assume centering is done using Newton's method, and use the convergence theory of Newton's method for self-concordant functions described in chapter 5. There we have seen that the number of Newton iterations required to minimize a self-concordant function f is bounded by

$$\frac{f(x) - f^\star}{\eta_2} + c,$$

where η_2 and c are constants, x is the starting point for Newton's method, and $f^\star$ is the optimal value. Here we are interested in applying this result to $f = t f_0 + \phi$ so we will need the assumption that $t f_0 + \phi$ is self-concordant for all $t \geq t^{(0)}$.

This assumption restricts the complexity analysis to a very particular class of functions. It is important to emphasize, however, that SUMT works well in general. In particular, the bound (6.27) holds regardless of the self-concordance. The self-concordance assumption is needed only to obtain a rigorous bound on the total number of Newton iterations.

It is also interesting to point out that the complexity analysis actually covers a wide variety of convex problems, as illustrated by the following examples.

Example. *Linear and quadratic problems.* If the functions f_i are linear or quadratic, then

$$tf_0 - \sum_{i=1}^{m} \log(-f_i)$$

is self-concordant for all values of $t \geq 0$ (see §5.6). The complexity analysis given below therefore applies to LP and QCQP.

Example. *Second-order cone programming.* In chapter 5 we have mentioned that the function $f(w,y) = -\log(y^2 - w^T w)$ is self-concordant on $\{ (w,y) \mid \|w\| < y \}$.

Each term in the logarithmic barrier (6.23) for the SOCP is a composition of f with an affine mapping $(w,y) = (A_i x + b_i, c_i^T x + d_i)$, and is self-concordant. Therefore $tc^T x + \phi(x)$ is self-concordant for all $t \geq 0$, and the complexity analysis of the SUMT method will apply.

Example. *Entropy maximization.* It is easily verified that the function $ty\log y - \log y$ is self-concordant on $\{ y \mid y > 0 \}$, for all $t \geq 0$. Using the calculus rules of self-concordant functions, we can conclude

$$t\sum_{i=1}^{n} x_i \log x_i - \sum_{i=1}^{n} \log x_i$$

is self-concordant on $\{ x \mid x \succ 0 \}$. The complexity analysis therefore also applies to the entropy maximization problem (6.5).

Example. *Geometric programming.* It is not clear whether or not the logarithmic barrier (6.21) is self-concordant, so although the SUMT method works for (6.20), the bounds from the complexity analysis may not hold.

However, suppose we reformulate the problem (6.20) as follows:

$$\begin{array}{ll} \text{minimize} & \sum_{j=1}^{k_0} y_{0j} \\ \text{subject to} & \sum_{j=1}^{k_i} y_{ij} \leq 1, \quad i = 1,\ldots,m \\ & a_{ji}^T x + b_{ji} - \log y_{ji} \leq 0, \quad i = 1,\ldots,m, \quad j = 0,\ldots,k_i \\ & y_{ji} \geq 0, \quad i = 1,\ldots,m, \quad j = 0,\ldots,k_i, \end{array} \tag{6.28}$$

where the variables are x and y_{ji}. The logarithmic barrier for this problem is

$$\sum_{i=0}^{m}\sum_{j=1}^{k_i} \Big(-\log y_{ji} - \log(\log y_{ji} - a_{ji}^T x - b_{ji})\Big) - \sum_{i=1}^{m} \log(1 - \sum_{j=1}^{k_i} y_{ij}).$$

It follows from the properties of §5.6 that this function is self-concordant, so the results from the complexity analysis hold for SUMT applied to (6.28).

6.5.2 The number of Newton iterations per centering step

The self-concordance assumption implies that

$$\frac{\mu t f_0(x^\star(t)) + \phi(x^\star(t)) - \mu t f_0(x^\star(\mu t)) - \phi(x^\star(\mu t))}{\eta_2} + c \tag{6.29}$$

is an upper bound on the number of Newton steps required to compute $x^\star(\mu t)$ starting at $x^\star(t)$. Unfortunately we do not know $x^\star(\mu t)$, and hence the upper bound (6.29), until we actually compute $x^\star(\mu t)$, *i.e.*, carry out the Newton algorithm (whereupon we know the *exact* number of Newton steps required to compute $x^\star(\mu t)$, which defeats the purpose!).

However we can bound (6.29) as follows. Using the notation $\lambda = \lambda^\star(t)$, $\nu = \nu^\star(t)$, we can write

$$\begin{aligned}
&\mu t f_0(x^\star(t)) + \phi(x^\star(t)) - \mu t f_0(x^\star(\mu t)) - \phi(x^\star(\mu t)) \\
&\quad = \mu t f_0(x^\star(t)) - \mu t f_0(x^\star(\mu t))) + \sum_{i=1}^m \log(-\mu t \lambda_i f_i(x^\star(\mu t))) - m \log \mu \\
&\quad \leq \mu t f_0(x^\star(t)) - \mu t f_0(x^\star(\mu t)) + \mu t \sum_{i=1}^m \lambda_i f_i(x^\star(\mu t)) - m - m \log \mu \\
&\quad = \mu t f_0(x^\star(t)) - \mu t \left(f_0(x^\star(\mu t)) + \sum_{i=1}^m \lambda_i f_i(x^\star(\mu t)) + \nu^T(Ax^\star(\mu t) - b) \right) - m - m \log \mu \\
&\quad \leq \mu t f_0(x^\star(t)) - \mu t g(\lambda, \nu) - m - m \log \mu \\
&\quad = m(\mu - 1 - \log \mu).
\end{aligned}$$

In the first inequality we used the fact that $\log a \leq a - 1$ for $a > 0$. The second inequality follows from the definition of the dual function:

$$\begin{aligned}
g(\lambda, \nu) &= \inf_x (f_0(x) + \sum_i \lambda_i f_i(x) + \nu^T(Ax - b)) \\
&\leq f_0(x^\star(\mu t)) + \sum_i \lambda_i f_i(x^\star(\mu t)) + \nu^T(Ax^\star(\mu t) - b).
\end{aligned}$$

The last line follows from $g(\lambda, \nu) = f_0(x^\star(t)) - m/t$.

The conclusion is that

$$\frac{m(\mu - 1 - \log \mu)}{\eta_2} + c. \tag{6.30}$$

is an explicit and easily computed upper bound on (6.29).

The function $\mu - 1 - \log \mu$ is shown in figure 6.3.

6.5.3 The total number of Newton iterations

The SUMT method has two distinct parts: the initial centering step (which in general does not start on the central path), and the subsequent centering steps (which all start on the central path). The number of Newton iterations in the initial centering step depends on the initial x and on the initial value of t. We only analyze the computational complexity of SUMT, when started on the central path.

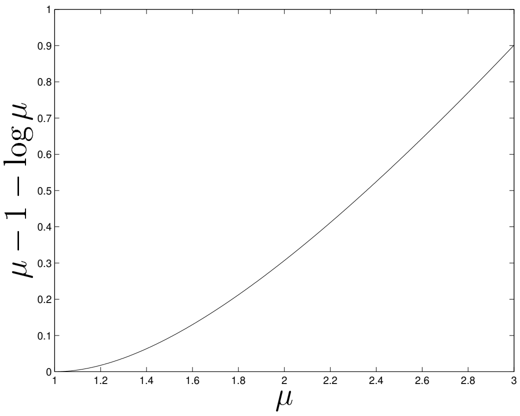

Figure 6.3: $\mu - 1 - \log\mu$ versus μ.

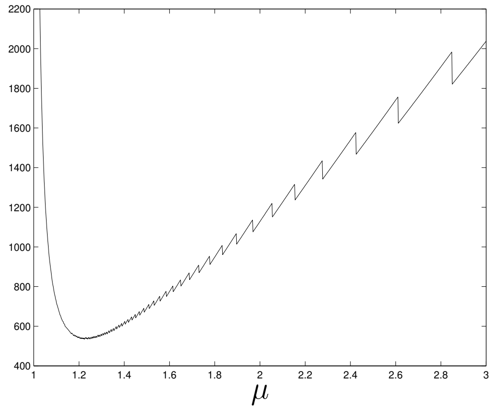

Figure 6.4: The upper bound on the total number of Newton iterations, given by equation (6.31), for $c = 5$, $\eta_2 = 20$ and $m/(t^{(0)}\epsilon) = 10^5$.

Each of the centering steps computes $x^\star(\mu t)$ using Newton's method starting at $x^\star(t)$. To obtain an upper bound on the total number of Newton steps we combine (6.30) and (6.27), and obtain

$$\left\lceil \frac{\log(m/t^{(0)}\epsilon)}{\log\mu} \right\rceil \left(\frac{m(\mu - 1 - \log\mu)}{\eta_2} + c \right). \tag{6.31}$$

This amazing formula shows that the seqeuential unconstrained minimization method not only works, in the sense of computing an ϵ-suboptimal solution, but does so with a total number of Newton steps that (ignoring the initial centering) depends only on the number of inequalities m, the algorithm parameter μ, the required duality gap reduction $t^{(0)}\epsilon$, and η_2, which depends on the functions f_i and is in many cases known. The parameter c depends in a very weak way on the functions f_i, but can be reasonably taken to be a constant such as 5 or 6.

Example. Figure 6.4 shows the bound (6.31) versus μ for

$$c = 5, \quad \eta_2 = 20, \quad m/(t^{(0)}\epsilon) = 10^5.$$

This figure confirms the tradeoff in the choice of μ. The first term in (6.31), which gives the required number of centering steps, grows very rapidly as $\mu \to 1$. For large values of μ, the bound also increases, due to the second term, which bounds the required number of Newton iterations per centering steps.

We can also note some important difference from the experimental curve in figure 6.2. The bound in figure 6.4 has a clear minimum, near $\mu = 1$. In the experimental results we observed that the total number of iterations is almost constant for a wide range values of μ. In practice, we can use a larger value of μ than the one suggested by the bound.

We also note that the value of the bound is quite conservative. In practice the number of Newton iterations is much smaller than predicted by the bound.

Strategies for choosing μ as a function of m

Based on the bound (6.31) we can estimate the worst-case complexity of the SUMT method.

First, suppose we use the same value of μ for all m. Then the total number of Newton iterations needed to reduce the duality gap from the initial value $\epsilon^{(0)} = m/t^{(0)}$ to a value ϵ or less, is

$$O(m \log(\epsilon^{(0)}/\epsilon)),$$

i.e., the worst-case complexity grows linearly with m.

It turns out we can do better, at least from a worst-case complexity standpoint, by decreasing m for large values of m. Suppose we choose $\mu = 1 + 1/\sqrt{m}$. Then we can bound the second term in (6.31) as

$$\mu - 1 - \log \mu = 1/\sqrt{m} - \log(1 + 1/\sqrt{m}) \leq 1/\sqrt{m} - 1/\sqrt{m} + 1/2m = 1/2m$$

(using $-\log(1+a) \leq -a + a^2/2$ for $a \geq 0$), and, using concavity of the logarithm, we also have

$$\log \mu = \log(1 + 1/\sqrt{m}) \geq (\log 2)/\sqrt{m}.$$

Putting everything together we can say that the upper bound (6.31) is

$$O(\sqrt{m} \log(\epsilon^{(0)}/\epsilon). \tag{6.32}$$

Multiplied with the cost of one iteration, which is dominated by the cost of forming and solving the Newton equations (*i.e.*, a polynomial function of n, m, and p), this expression gives us an upper bound on the total number of operations in SUMT.

Note that the upper bound in (6.32) depends only on the dimension m, and not on the variables n or the number of equality constraints p. Moreover it grows very slowly with m, which confirms our empirical observation in §6.3.3 that the number of iterations is nearly constant over a wide range of m.

6.6 Extension to generalized inequalities

In the remainder of this chapter we extend SUMT to convex problems with generalized inequalities

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \preceq_{K_i} 0, \ i = 1, \ldots, L. \end{array} \tag{6.33}$$

The function $f_0 : \mathbf{R}^n \to \mathbf{R}$ is convex, $K_i \subseteq \mathbf{R}^{m_i}$ are cones that induce generalized inequalities, and the functions $f_i : \mathbf{R}^n \to \mathbf{R}^{m_i}$, $i = 1, \ldots, L$ are K_i-convex. We make the same assumptions as in §6.1.

- The functions f_i have continuous second derivatives.
- We know a strictly feasible point $x^{(0)}$ *i.e.*, $x^{(0)} \in D$ and $f_i(x) \prec_{K_i} 0$, $i = 1, \ldots, L$.
- The sublevel set
$$\{ \ x \mid f_0(x) \leq f_0(x^{(0)}), \ f_i(x) \preceq_{K_i} 0, \ \ i = 1, \ldots, L \ \}$$
is bounded.
- $\{ \ x \mid f_i(x) \ \prec_{K_i} 0, \ i = 1, \ldots, L \ \} \subseteq \mathbf{dom}\, f_0$.

The development of the method is completely parallel to the previous sections. We will define the counterparts of the logarithmic barrier function, the central path, associated dual feasible points, etc. Conversely, we can recover what we have done so far by specializing the results of this section to $K_i = \mathbf{R}_+$ and $L = m$.

6.6.1 Logarithmic barrier

We first define the analog of $\log x$. Let K be a cone that induces a generalized inequality. A barrier function ψ of K is called a *logarithmic* barrier if there exists a $\theta > 0$ such that for all $y \succ_K 0$, and all $s > 0$,

$$\psi(sy) = \psi(y) - \theta \log s.$$

In other words, ψ behaves like a logarithm along every ray in the cone K. We call θ the *degree* of ψ (motivated by the fact that $\exp(-\psi)$ is a homogeneous function with degree θ).

We will need two properties of logarithmic barrier functions.

- $\nabla\psi(y) \prec_{K^*} 0$ for all $y \succ_K 0$, *i.e.*, ψ is K-decreasing (see §2.6.1).
- $y^T \nabla\psi(y) = -\theta$ for all $y \succ_K 0$. This follows immediately from differentiating $\psi(sy) = \psi(y) - \theta \log s$ with respect to s.

Returning to problem (6.33), assume ψ_i, $i = 1, \ldots, L$, are logarithmic barriers for the cones K_i. Then we define the *logarithmic barrier function* for the constraints $f_i(x) \preceq_{K_i} 0$ as

$$\phi(x) = \begin{cases} \displaystyle\sum_{i=1}^{L} \psi_i(-f_i(x)) & \text{if } f_i(x) \prec_{K_i} 0, \ \ i = 1, \ldots, L \\ +\infty & \text{otherwise.} \end{cases}$$

Convexity of ϕ follows from the fact that the functions ψ_i are K_i-increasing and the function f_i are K_i-convex (see the composition rule of §2.6.2).

Example. For the cone of positive-semidefinite matrices

$$K = \{X = X^T \in \mathbf{R}^{p\times p} \mid X \succ 0\}$$

we have the logarithmic barrier function

$$\psi(X) = \begin{cases} \log\det X^{-1} & \text{if } X \succ 0 \\ +\infty & \text{otherwise.} \end{cases}$$

The degree is equal to p, since $\log\det(-sX)^{-1} = \log\det X^{-1} - p\log s$.

The gradient of ψ at a matrix $X = X^T \succ 0$ is equal to

$$\nabla\psi(X) = -X^{-1}.$$

We can therefore easily verify the two general properties: $\nabla\psi(X) = -X^{-1} \prec 0$, and the inner product of X and $\nabla\psi(X)$ is equal to $-\mathbf{Tr}\, XX^{-1} = -p$.

Example. *The nonnegative orthant.* $K = \mathbf{R}^n_+$. $\psi(x) = -\sum_{i=1}^n \log x_i$ is logarithmically homogeneous with degree n. For $x \succ 0$, $(\nabla\psi(x))_i = -1/x_i$. Hence $\nabla\psi(x) \prec 0$, and $x^T\nabla\psi(x) = -n$.

Example. *The second-order cone.* $K = \{x \in \mathbf{R}^{n+1} \mid (\sum_{i=1}^n x_i^2)^{1/2} \leq x_{n+1}\}$. $\psi(x) = -\log(x_{n+1}^2 - \sum_{i=1}^n x_i^2)$ is logarithmically homogeneous with degree 2. Suppose $x_{n+1} > (x_1^2 + \cdots + x_n)^{1/2}$. Then

$$\begin{aligned} (\nabla\psi(x))_i &= \frac{2x_i}{(x_{n+1}^2 - \sum_{i=1}^n x_i^2)^{1/2}}, \quad i = 1,\ldots,n \\ (\nabla\psi(x))_{n+1} &= \frac{-2x_{n+1}}{(x_{n+1}^2 - \sum_{i=1}^n x_i^2)^{1/2}}, \end{aligned}$$

and one can verify that $-\nabla\psi(x) \in \mathbf{int}\, K^* = \mathbf{int}\, K$ and $x^T\nabla\psi(x) = -2$.

6.6.2 The central path

Definition

The next step is to define the central path for problem (6.33). As before, we define the central points $x^\star(t)$ as the minimizers of $tf_0 + \phi$, *i.e.*, as

$$x^\star(t) = \operatorname*{argmin}_{f_i(x)\prec_{K_i}0,\ i=1,\ldots,m} \left(tf_0(x) + \sum_{i=1}^{L} \psi(-f_i(x)) \right).$$

The assumption of bounded sublevel sets of (6.33) guarantees that the minimum of $tf_0 + \phi$ is attained. We also impose the slightly stronger condition that the Hessian of $tf_0 + \phi$ is nonsingular everywhere. (If it is not, we can add a constraint, *e.g.*, $x^T x \leq R^2$, which adds a strictly convex term $-\log(R^2 - x^T x)$ to the barrier.)

Central points are characterized by the optimality conditions:

$$t\nabla f_0(x) + \nabla\phi(x) = t\nabla f_0(x) - \sum_{i=1}^{L} Df_i(x)^T \nabla\psi_i(-f_i(x)) = 0, \tag{6.34}$$

where $Df_i(x)$ is the Jacobian matrix of the function f_i at x.

Dual points on central path

For $i = 1, \ldots, L$, define

$$\lambda_i^\star(t) = -\frac{1}{t}\nabla\psi_i(-f_i(x^\star(t))). \tag{6.35}$$

We will show that $\lambda_i^\star(t)$ is strictly dual feasible. First, by the general monotonicity property of logarithmic barrier functions, we have $\lambda_i^\star(t) \succ_{K_i^*} 0$. Secondly, it follows from (6.34) that the Lagrangian

$$L(x, \lambda^\star(t)) = f_0(x) + \sum_{i=1}^{L} \lambda_i^\star(t)^T f_i(x)$$

is bounded below as a function of x, and that the minimizer is $x^\star(t)$. The dual function g evaluated at $\lambda^\star(t)$ is equal to

$$g(\lambda^\star(t)) = f_0(x^\star(t)) - \sum_{i=1}^{L} \lambda^\star(t)_i^T f_i(x^\star(t)) = f_0(x^\star(t)) - m/t$$

where $m = \sum_{i=1}^{L} \theta_i$, and θ_i is the degree of ψ_i.

Example. For the SDP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & F(x) = F_0 + x_1 F_1 + \cdots + x_n F_n \preceq 0, \end{array}$$

with $x \in \mathbf{R}^n$ and $F_i = F_i^T \in \mathbf{R}^{p\times p}$, we define the barrier function

$$\phi(x) = \begin{cases} \log\det(-F(x)^{-1}) & F(x) \prec 0 \\ +\infty & \text{otherwise.} \end{cases}$$

For strictly feasible x, the gradient of ϕ is equal to

$$(\nabla\phi(x))_i = \mathbf{Tr}\, F(x)^{-1} F_i, \quad i = 1, \ldots, m,$$

which gives us the optimality conditions that characterize central points:

$$tc_i - \mathbf{Tr}\, F(x^\star(t))^{-1} F_i = 0, \quad i = 1, \ldots, m.$$

Hence the matrix

$$Z^\star(t) = \frac{1}{t}\left(-F(x^\star(t))\right)^{-1}.$$

is strictly dual feasible, and the duality gap associated with $x^\star(t)$ and $Z^\star(t)$ is p/t.

6.6.3 SUMT for generalized inequalities

The conclusion so far is that the following key properties of the central path generalize in a straightforward way:

- computing a point on the central path is an unconstrained minimization problem
- the central path leads to an optimal point as $t \to \infty$
- with each central point $x^\star(t)$ we can associate a dual feasible point $\lambda^\star(t)$ and a duality gap m/t.

This means we can apply the SUMT method to problem (6.33). The outline of the method is identical to the outline in §6.3. The number of centering steps required to compute a central point with duality gap ϵ starting at $x^\star(t^{(0)})$ is equal to

$$\left\lceil \frac{\log(m/t^{(0)}\epsilon)}{\log \mu} \right\rceil.$$

6.6.4 Complexity analysis

In order to obtain a bound on the total complexity, we need an upper bound on the number of Newton steps required to compute $x^\star(\mu t)$ starting from $x^\star(t)$. We will therefore assume that $tf_0 + \phi$ is self-concordant for all $t > t^{(0)}$, which implies that the upper bound (6.29) is also valid here. It turns out that it is also possible to extend the upper bound (6.30).

The dual logarithmic barrier

Let ψ be a logarithmic barrier for a cone K, with degree θ. It can be shown that the function ψ^*, defined as

$$\psi^*(y) = \begin{cases} \sup_x \left(-x^T y - \psi(x)\right) & \text{if } y \succ_{K^*} 0 \\ +\infty & \text{otherwise} \end{cases}$$

is a logarithmic barrier for K^*, with the same degree θ. From this definition we immediately obtain *Young's inequality*:

$$x \succ_K 0, y \succ_{K^*} 0 \Longrightarrow \psi(x) + \psi^*(y) \geq -x^T y, \tag{6.36}$$

with equality if $x = -\nabla\psi^*(y)$ (or equivalently, $y = -\nabla\psi(x)$). In fact, for logarithmic barriers a stronger property holds, which we may call the *extended Young's inequality*:

$$\psi(x) + \psi^*(y) \geq -\theta \log \frac{y^T x}{\theta} - \theta. \tag{6.37}$$

To prove this, we first apply Young's inequality to x and sy ($s > 0$), which yields

$$\psi(x) \geq -sx^T y - \psi^*(sy) = -sx^T y - \psi^*(y) + \theta \log s.$$

Maximizing the right hand side over s, we find that the optimal value s is $s = \theta/x^T y$, from which we obtain (6.37). One can also verify that the inequality holds with equality if $x = -\nabla\psi^*(y)$, or equivalently, $y = -\nabla\psi(x)$.

Examples.

- $K = \mathbf{R}^m_+$. For given y, one can analytically optimize $x^T y + \sum_{i=1}^m \log x_i$ over x. The maximizer is $x_i = -1/y_i$ and substitution yields

$$\psi^*(y) = -\sum_{i=1}^m \log(-y_i) + m.$$

- For the positive semidefinite cone in $\mathbf{R}^{p\times p}$ we have

$$\psi^*(Y) = \log\det Y^{-1} - p.$$

The number of Newton iterations per centering step

The properties of the dual barrier allow us to extend the bound (6.30) on the number of Newton steps required in a centering operation.

For $\lambda_i \succ_{K_i^*} 0$, and $f_i(x) \prec_{K_i} 0$ we have

$$\begin{aligned}
& tf_0(x) + \phi(x) - tg(\lambda) + \sum_{i=1}^L \psi_i^*(\lambda_i) - m\log t \\
= \quad & t(f_0(x) - g(\lambda)) + \sum_{i=1}^L (\psi_i(-f_i(x)) - \theta_i \log t + \psi_i^*(\lambda_i)) \\
= \quad & t(f_0(x) - g(\lambda)) + \sum_{i=1}^L (\psi_i(-tf_i(x)) + \psi_i^*(\lambda_i)) \\
\geq \quad & t(f_0(x) - g(\lambda)) - \sum_{i=1}^L \theta_i \left(\log \frac{-t\lambda_i^T f_i(x)}{\theta_i} + 1\right) \\
\geq \quad & t\left(f_0(x) - g(\lambda) + \sum_{i=1}^L \lambda_i^T f_i(x)\right) \geq 0.
\end{aligned}$$

(The first inequality follows from the extended Young inequality. The second inequality from $-\log a \geq -a + 1$, and the last inequality follows from the definition of g.) Therefore,

$$\mu t f_0(x^\star(\mu t)) + \phi(x^\star(\mu t)) \geq \mu t g(\lambda) - \sum_{i=1}^L \psi_i^*(-\lambda_i) + m\log(\mu t), \tag{6.38}$$

and if we evaluate this bound for $\lambda_i = \lambda_i^\star(t)$, defined in (6.35), we obtain

$$\mu t f_0(x^\star(\mu t)) + \phi(x^\star(\mu t)) \geq \mu t f_0(x^\star(t)) + \phi(x^\star(t)) - m(\mu - 1 - \log\mu).$$

In conclusion, the upper bounds (6.30) and (6.31) on the number of Newton iterations also hold for the SUMT method with generalized inequalities (if $tf_0 + \phi$ is self-concordant).

Appendix A

Mathematical background

A.1 Inner products and norms

A.1.1 Inner products

All vector spaces in this course are finite-dimensional. Moreover we work almost exclusively with the following three (real) vector spaces:

- $\mathbf{R}^n$ is the space of real vectors of dimension n
- $\mathbf{R}^{m\times n}$ is the space of real matrices of dimension $m \times n$
- S_n is the space of symmetric real matrices of dimension $n \times n$. We will not use the symbol S_n very often. Instead we usually write '$A \in S_n$' as '$A = A^T \in \mathbf{R}^{n\times n}$'. ($A^T$ denotes the transpose of the matrix A).

The notation $\langle x, y\rangle$ denotes a general *inner product* of vectors x and y. Recall that an inner product on a real vector space $\mathcal{V}$ is a function $\langle\cdot,\cdot\rangle$ from $\mathcal{V}\times\mathcal{V}$ to $\mathbf{R}$ with the following properties. If $x, y, z \in \mathcal{V}$, then

- $\langle x, y\rangle = \langle y, x\rangle$
- $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$
- $\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$ for all $\alpha \in \mathbf{R}$
- $\langle x, x\rangle \geq 0$, and $\langle x, x\rangle = 0$ if and only if $x = 0$

We say x and y are *orthogonal* if $\langle x, y\rangle = 0$.

We use the following inner products in the spaces $\mathbf{R}^n$, $\mathbf{R}^{m\times n}$ and S_n.

- The inner product of two vectors $x, y \in \mathbf{R}^n$, is

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i = x^T y.$$

- The inner product of two matrices $X, Y \in \mathbf{R}^{m\times n}$, is

$$\langle X, Y\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}Y_{ij} = \mathbf{Tr}\, X^TY = \mathbf{Tr}\, Y^TX, \tag{A.1}$$

  where $\mathbf{Tr}$ stands for *trace*, *i.e.*, $\mathbf{Tr}\, X^TY$ is the sum of the diagonal elements of the matrix X^TY. The notation $\mathbf{Tr}\, X^TY$ is convenient but somewhat misleading. It suggests evaluating the inner product by forming the matrix product X^TY and then taking its trace, which requires a total of n^2m multiplications. The second expression in (A.1) is less compact but clearly shows that only mn multiplications are needed.

- The inner product of two matrices $X = X^T$, $Y = Y^T \in \mathbf{R}^{n\times n}$ is

$$\langle X, Y\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}Y_{ij} = \mathbf{Tr}\, XY = \mathbf{Tr}\, YX.$$

A.1.2 Norms

Let $\mathcal{V}$ be a vector space. A function f from $\mathcal{V}$ to $\mathbf{R}$ is a *norm* on $\mathcal{V}$ if

- f is nonnegative, *i.e.*, $f(x) \geq 0$ for all x
- f is definite, *i.e.*, $f(x) = 0$ only if $x = 0$
- f is homogeneous, *i.e.*, $f(tx) = |t|f(x)$, for all t
- f satisfies the triangle inequality $f(x + y) \leq f(x) + f(y)$

We often use the notation $f(x) = \|x\|$ (which emphasizes that a norm is a generalization of the absolute value on $\mathbf{R}$). When we must specify a particular norm, or the norm meant is not clear from context, we use the notation $\|x\|_{\mathrm{symb}}$, where the subscript is a mnemonic to indicate which norm is meant.

A norm is a measure for the *length* of x; the four properties in the definition ensure that f has the geometric properties that we expect from a measure of length. We can measure the *distance* $d(x, y)$ between two elements x and y of $\mathcal{V}$ as the length of their difference, *i.e.*,

$$d(x, y) = \|x - y\|.$$

Thus norms can be used to measure closeness of vectors in $\mathcal{V}$, which allows us to define open and closed sets, limits, continuity of functions, etc.

Simple examples of norms are $|\cdot|$ on $\mathbf{R}$, and the *Euclidean* or 2-norm on $\mathbf{R}^n$:

$$\|x\|_2 = \sqrt{x^Tx} = \sqrt{x_1^2 + \cdots + x_n^2}.$$

The Euclidean norm is so common that we often write it without subscript, as $\|x\|$.

The set of all elements of $\mathcal{V}$ with norm less than or equal to one,

$$\mathcal{B} = \{x \in \mathcal{V} \mid f(x) \leq 1\}$$

is called the *unit ball* of the norm f. The four properties of a norm imply the following properties of the unit ball:

- $\mathcal{B}$ is symmetric about the origin, *i.e.*, $x \in \mathcal{B}$ if and only if $-x \in \mathcal{B}$
- $\mathcal{B}$ is convex
- $\mathcal{B}$ is closed, bounded, and has nonempty interior

Conversely, if $C \subseteq \mathcal{V}$ is any set satisfying these three conditions, then it is the unit ball of a norm, which is given by

$$f(x) = \left(\sup\{t \geq 0 \mid f(tx) \in C\}\right)^{-1}.$$

There is a one-to-one correspondence between norms and sets satisfying the conditions above.

A.1.3 Examples

We start with some common norms on $\mathbf{R}^n$. We have already encountered the Euclidean norm. Two other frequently used norms on $\mathbf{R}^n$ are the *sum-absolute-value*, or 1-*norm*, given by

$$\|x\|_1 = |x_1| + \cdots + |x_n|,$$

and the *Chebychev*, *peak*, *max-*, or ∞-*norm*, given by

$$\|x\|_\infty = \max_i |x_i|.$$

In fact these three norms are related; they are part of a family of norms parametrized by $p \geq 1$. The *p-norm* is defined by

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$$

for $p \geq 1$. This yields the 1-norm when $p = 1$ and the Euclidean norm when $p = 2$. It is easy to show that for any $x \in \mathbf{R}^n$,

$$\lim_{p\to\infty} \|x\|_p = \max_i |x_i|,$$

so the ∞-norm also fits in this family (at least, as a limit).

The unit balls for several p-norms are shown in figure A.1. The Euclidean ball is the usual unit disk; the ∞-ball is a square, and the 1-ball is a diamond.

There are many ways to construct norms. One simple and general method is based on linear transformations. If $A \in \mathbf{R}^{k\times n}$ and has rank n (so $k \geq n$), and f is a norm on $\mathbf{R}^k$, then $g(x) = f(Ax)$ is a norm on $\mathbf{R}^n$. This yields, for example, the piecewise-linear norm

$$\max_{i=1,\ldots,k} |a_i^T x|$$

(and conversely, every piecewise-linear norm can be represented this way). When $A \in \mathbf{R}^{n\times n}$ (and nonsingular), the norm $g(x) = f(Ax)$ is called the A-weighted or A-scaled f norm. An important special case is when A is diagonal; A_{ii}^{-1} can be interpreted as the scale or unit used to measure x_i.

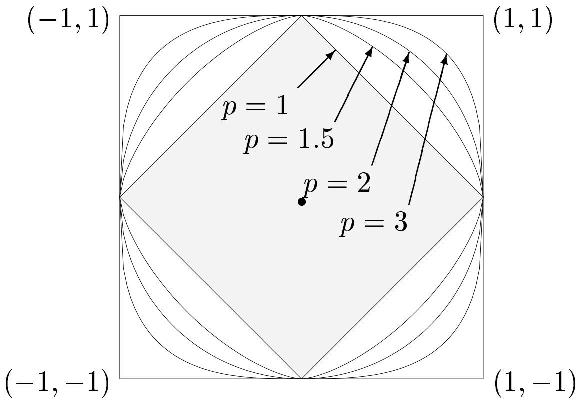

Figure A.1: Norm balls for $\| \cdot \|_p$ with $p = 1$, 1.5, 2, 3, and for the infinity-norm.

As an example of a norm on a vector space other than $\mathbf{R}^n$, we have the *Frobenius norm* of a matrix $X \in \mathcal{V} = \mathbf{R}^{m\times n}$, defined as

$$\|X\|_F = \left(\mathbf{Tr}\, X^T X\right)^{1/2} = \left(\sum_{i,j} X_{ij}^2\right)^{1/2},$$

which can be interpreted as the Euclidean norm of the matrix, when considered as an element of $\mathbf{R}^{mn}$.

Equivalence of norms

We say two norms f and g on a vector space $\mathcal{V}$ are *equivalent* if there exist positive α, β such that for all $x \in \mathcal{V}$,

$$\alpha f(x) \leq g(x) \leq \beta f(x).$$

This property can be verified to be an equivalence relation, *i.e.*, it is symmetric, transitive, and reflexive. When two norms are equivalent, they induce the same topology, *i.e.*, the same set of open subsets of $\mathcal{V}$, the same set of convergent sequences, etc.

A basic fact is that if $\mathcal{V}$ is finite-dimensional, then *any* two norms are equivalent. One implication is that when we can speak of an open subset of, for example, $\mathbf{R}^n$, without worrying about what the underlying norm is.

As an example, on $\mathbf{R}^n$ we have

$$\|x\| \leq \|x\|_1 \leq \sqrt{n}\|x\|$$

(which shows the 1-norm and the Euclidean norm are equivalent),

$$\|x\|_\infty \leq \|x\| \leq \sqrt{n}\|x\|_\infty$$

(equivalence between Euclidean norm and infinity-norm), and

$$\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty$$

(equivalence between 1-norm and infinity-norm). (Moreover, the constants in these inequalities are the tightest possible.)

A.1.4 Operator norms

Let $\mathcal{V}$ and $\mathcal{W}$ be two finite-dimensional vector spaces, with norms $\|\cdot\|_v$ and $\|\cdot\|_w$, respectively. The set $\mathcal{L}(\mathcal{V},\mathcal{W})$ of all linear functions (or operators) from $\mathcal{V}$ into $\mathcal{W}$ forms a vector space, with pointwise addition and scalar multiplication. We define the *operator norm* of $X \in \mathcal{L}(\mathcal{V},\mathcal{W})$ by

$$\|X\| = \sup\left\{\|X(u)\|_w \mid \|u\|_v \le 1\right\}.$$

We say this norm is *induced* by the norms $\|\cdot\|_v$ and $\|\cdot\|_w$.

Let's consider the case in which $\mathcal{V} = \mathbf{R}^n$ and $\mathcal{W} = \mathbf{R}^m$, so we can identify $\mathcal{L}(\mathcal{V},\mathcal{W})$ with $\mathbf{R}^{m\times n}$. Difference choices of norms on $\mathbf{R}^m$ and $\mathbf{R}^n$ yield different operator norms on $\mathbf{R}^{m\times n}$.

We start by using the Euclidean norm on both $\mathbf{R}^n$ and $\mathbf{R}^m$. In this case the induced norm is given by

$$\|A\| = \sigma_{\max}(A) = (\lambda_{\max}(A^TA))^{1/2},$$

where $\sigma_{\max}$ is the maximum singular value and $\lambda_{\max}$ is the maximum eigenvalue. This norm is called the *spectral norm*, the *singular value norm*, or the *2-norm* of A. It is so common that we take it as the default or standard norm on $\mathbf{R}^{m\times n}$. Note that if we identify n-vectors ($\mathbf{R}^n$) as $n\times 1$ matrices ($\mathbf{R}^{n\times 1}$), the spectral norm agrees with the Euclidean norm, so there is no confusion.

The norm induced by the infinity norm on $\mathbf{R}^m$ and $\mathbf{R}^n$ is the max-row-sum norm, defined as

$$\|X\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^n |X_{ij}|$$

(which explains the subscript in $\|A\|_\infty$). In a similar way, the norm induced by the 1-norm is the *max-column-sum* norm, *i.e.*,

$$\|X\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^m |X_{ij}|.$$

Not all norms on matrices are induced by vector norms. The Frobenius norm, for example, is not induced by any vector norm.

A.1.5 Dual norm

Let $\mathcal{V}$ be a vector space with inner product denoted as $\langle\cdot,\cdot\rangle$, and let $\|\cdot\|$ be a norm on $\mathcal{V}$. The *dual norm* $\|\cdot\|_\star$ corresponding to $\|\cdot\|$ is defined as

$$\|z\|_\star = \sup\{\langle z, x\rangle \mid \|x\| \le 1\}. \tag{A.2}$$

It is straightforward to verify that the definition satisfies the four properties of a norm. As an example, we verify the triangle inequality. We have

$$\begin{aligned}
\|x+y\|_\star &= \sup\{\langle z, x+y\rangle \mid \|z\| \le 1\} \\
&\le \sup\{\langle z, x\rangle \mid \|z\| \le 1\} + \sup\{\langle z, y\rangle \mid \|z\| \le 1\} \\
&= \|x\|_\star + \|y\|_\star.
\end{aligned}$$

The dual norm can be interpreted as the operator norm of the linear operator

$$H : \mathcal{V} \to \mathbf{R} : x \mapsto \langle z, x \rangle,$$

induced by the norm f on $\mathcal{V}$ and the absolute value norm on $\mathbf{R}$.

From the definition we also note the important inequality

$$\langle z, x \rangle \leq \|x\| \|z\|_\star, \tag{A.3}$$

which can be verified as follows. If $\|x\| = 0$, the inequality is clearly satisfied (with equality). If $\|x\| \neq 0$, we have from the definition of dual norm

$$\langle z, x/\|x\| \rangle \leq \|z\|_\star,$$

and hence (A.3). It can be shown that the inequality is tight.

We now illustrate the definition of dual norm with some examples on $\mathbf{R}^n$.

The dual of the Euclidean norm is the Euclidean norm itself, since

$$\|y\| = \sup\{x^T y \mid \|x\| \leq 1\|.$$

This follows from the Cauchy-Schwarz inequality: we have $x^T y \leq \|x\| \|y\|$, and therefore

$$\|y\| \geq x^T y$$

for all $\|x\| \leq 1$. Moreover equality holds for $y = x/\|x\|$.

In a similar way, one can use the Hölder inequality *i.e.*, if $p, q \geq 1$ with $1/p + 1/q = 1$, then

$$-\|x\|_p \|y\|_q \leq x^T y \leq \|x\|_p \|y\|_q,$$

to show that if $p \geq 1$, then

$$\|y\|_q = \sup\{x^T y \mid \|x\|_p \leq 1\|,$$

where $1/q + 1/p = 1$. In other words the dual of the $\|\cdot\|_p$ is the norm $\|\cdot\|_q$ with $1/p + 1/q = 1$. This includes the limiting case $p = 1$, $q = \infty$, *i.e.*, the dual of the 1-norm is the infinity-norm.

A.2 Analysis

A.2.1 Open and closed sets

Suppose $C \subseteq \mathcal{V}$, where $\mathcal{V}$ is a finite-dimensional vector space.

An element $x \in C$ is called an *interior* point of C if there exists an $\epsilon > 0$ for which

$$\{y \mid \|y - x\| \leq \epsilon\} \subseteq C,$$

i.e., there exists a small ball centered at x that lies entirely in C. As usual $\|\cdot\|$ denotes the Euclidean norm, although any other norm could be used instead. The set of all points interior to C is called the *interior* of C and is denoted $\mathbf{int}\, C$.

A set C is *open* if $\mathbf{int}\, C = C$, *i.e.*, every point in C is an interior point. Some examples: $(0,1)$, $\emptyset$, $\mathcal{V}$, $\mathbf{R}^3$, $\{x \in \mathbf{R}^3 \mid \mathbf{1} \prec x \prec \mathbf{1}\}$, A set C is *closed* if its complement $\mathcal{V} \setminus C = \{x \in \mathcal{V} \mid x \notin C\}$ is open. Some examples: $[0,1]$, $\emptyset$, $\mathcal{V}$, $\{x \in \mathbf{R}^3 \mid \mathbf{1} \preceq x \preceq \mathbf{1}\}$.

The *closure* of a set C is defined as

$$\overline{C} = \mathcal{V} \setminus \mathbf{int}(\mathcal{V} \setminus C),$$

i.e., the complement of the interior of the complement of C. A point x is in the closure of C if for every $\epsilon > 0$, there is a $y \in C$ with $\|x - y\| \leq \epsilon$.

We can also describe closed sets and the closure in terms of convergent sequences (see below) and limit points. C is closed if and only it contains the limit point of every convergent sequence in it. In other words, if $x_1, x_2, \ldots$ converges to x, and $x_i \in C$, then $x \in C$. The closure of C is the set of all limit points of convergent sequences in C.

The *boundary* of the set C is defined as

$$\partial C = \overline{C} \setminus \mathbf{int}\, C.$$

A *boundary point* x (*i.e.*, a point $x \in \partial C$) satisfies the following property: for all $\epsilon > 0$, there exists $y \in C$ and $z \notin C$ with

$$\|y - x\| \leq \epsilon, \qquad \|z - x\| \leq \epsilon,$$

i.e., there exist arbitrarily close points in C, and also arbitrarily close points not in C.

We can characterize closed and open sets in terms of the boundary operation: C is *closed* if it contains its boundary, *i.e.*, $\partial C \subseteq C$. It is *open* if it contains no boundary points, *i.e.*, all elements of C are interior.

A.2.2 Limits and continuity

Suppose A is a set. A *sequence* in A is a function $h : \mathbf{Z}_+ \to A$. We often write $h(k)$ as h_k, and sometimes denote the whole sequence (*i.e.*, the function h) as $\{h_k\}$, or as the infinite ordered list $h_1, h_2, h_3, \ldots$.

Now suppose A is a finite-dimensional vector space, so we have a norm or notion of distance. We say that a sequence h *converges* to $z \in A$ if for every $\epsilon > 0$ there exists an

$N \in \mathbf{Z}_+$ such that $k \geq N$ implies $\|h_k - z\| \leq \epsilon$. It can be shown there is at most one limit of any sequence, so we refer to z as the limit of the sequence, and write it as $z = \lim_{k\to\infty} h_k$.

Suppose $f : \mathcal{V} \to \mathbf{R}$ is a function on a finite-dimensional vector space $\mathcal{V}$, and let $x \in \overline{\mathbf{dom}\, f}$. If for any convergent sequence $\{x_k\}$, with $x_k \in \mathbf{dom}\, f$ and x as limit, the sequence $\{f(x_k)\}$ converges to the same value z, then we say z is the *limit of f at x*, and write this as

$$\lim_{y\in\mathbf{dom}\, f,\; y\to x} f(y) = z.$$

Equivalently, the limit of f at x is equal to z if, for all $\epsilon > 0$ there exists a δ such that

$$y \in \mathbf{dom}\, f, \quad \|y - x\| \leq \delta \Longrightarrow |z - f(y)| \leq \epsilon.$$

We say f is *continuous* at a point $x \in \mathbf{dom}\, f$ if

$$f(x) = \lim_{y\in\mathbf{dom}\, f, y\to x} f(y).$$

A.2.3 Supremum, infimum, maximum, and minimum

Our usage of the notation $\sup C$, $\max C$, $\inf C$, and $\min C$ is different from the standard in optimization literature.

We use $\sup C$, where $C \subseteq \mathbf{R}$, to denote the smallest upper bound on C, *i.e.*, $\sup C$ is the smallest number α that satisfies $\alpha \geq \beta$ for all $\beta \in C$. We use this notation regardless of whether the bound is achieved or not. In other words, $\sup C$ may or may not be an element of C. The notation $\max C$ will be used only when C is a finite set. In both cases we make the convention that $\sup \emptyset = -\infty$ and $\max \emptyset = -\infty$.

Our convention for $\inf C$ and $\min C$ is similar. The greatest lower bound to C is written as $\inf C$, regardless of whether the bound is attained or not; $\min C$ denotes the smallest element of a finite set C. The infimum or minimum of an empty set is equal to $+\infty$ by convention.

A.3 Functions and derivatives

These notes set out the mechanics of derivatives and differentiation. Mostly it sets out our notation, which is (almost) standard. More details are found in books or courses on advanced calculus or analysis.

A.3.1 Function notation

Our notation for functions is mostly standard, with one exception. When we write

$$f : A \to B$$

we mean that f is a function on the set $\mathbf{dom}\, f \subseteq A$ into the set B; in particular we can have $\mathbf{dom}\, f$ a proper subset of the set A. Thus the notation $f : \mathbf{R}^m \to \mathbf{R}^n$ means that f maps (some) m-vectors into n-vectors; it does not mean that $f(x)$ is defined for every $x \in \mathbf{R}^m$.

This convention is similar to function declarations in computer languages. Specifying the data types of the input and output arguments of a function gives the *syntax* of that function; it does not guarantee that any input arguments with the specified data type are valid.

As an example consider the function $f : S_m \to \mathbf{R}$, where $S_m = \{X \in \mathbf{R}^{m\times m} \mid X = X^T\}$, given by

$$f(X) = \log\det X, \tag{A.4}$$

with $\mathbf{dom}\, f = \{X \in S_m \mid X \succ 0\}$. The notation $f : S_m \to \mathbf{R}$ specifies the *syntax* of f: it takes as argument a symmetric $m \times m$ matrix, and returns a real number. The notation $\mathbf{dom}\, f = \{X \in S_m \mid X \succ 0\}$ specifies which symmetric $m \times m$ matrices are valid input arguments for f (*i.e.*, only positive definite ones). The formula (A.4) specifies what $f(X)$ is, for $X \in \mathbf{dom}\, f$.

When $f : A \to \mathbf{R}$ we say the function f is real-valued, and sometimes call it a *functional* (on A). Conversely, when $f : \mathbf{R} \to B$ we can think of f as parametrizing a curve or path in B, and (for $t \in \mathbf{int}\,\mathbf{dom}\, f$) refer to $Df(t)$ as the tangent to the curve at the point $f(t)$.

A.3.2 Derivative of a function

Suppose $f : \mathcal{V} \to \mathcal{W}$, where $\mathcal{V}$ and $\mathcal{W}$ are (finite-dimensional) vector spaces (*e.g.*, $\mathbf{R}^k$, $\mathbf{R}^{p\times q}$, $S_m = \{X \in \mathbf{R}^{m\times m} \mid X = X^T\}$), and $x \in \mathbf{int}\,\mathbf{dom}\, f$. A linear function $D : \mathcal{V} \to \mathcal{W}$ is the *derivative* of f at x if

$$\lim_{z\in\mathbf{dom}\, f,\ z\to x} \frac{\|f(z) - f(x) - D(z-x)\|}{\|z-x\|} = 0 \tag{A.5}$$

(which does not depend on which norm used). In this case we say f is *differentiable* at x. It can be shown that there is at most one linear function D satisfying (A.5), so (when f is differentiable at x) we refer to it as *the* derivative, and denote it as $Df(x)$. (The argument specifies the point x.)

The affine function of z given by

$$f(x) + Df(x)(z-x) \tag{A.6}$$

is called the *first-order approximation* of f at (or near) x. Evidently this function agrees with f at $z = x$; the definition of derivative (A.5) means that when z is *close* to x, this affine function is *very close* to f. The limit condition (A.5) is often written compactly as

$$|f(z) - f(x) - Df(x)(z - x)| = o(\|z - x\|).$$

In most cases we can describe $Df(x)$ in a concrete way. In the most common case we have $\mathcal{V} = \mathbf{R}^m$ and $\mathcal{W} = \mathbf{R}^n$, *i.e.*, $f : \mathbf{R}^m \to \mathbf{R}^n$. In this case any linear function from $\mathcal{V}$ into $\mathcal{W}$ can be expressed as matrix multiplication by an $n \times m$ matrix, and we use the same symbol, $Df(x)$, to denote the matrix. Thus, we can interpret (A.5) and (A.6) with $Df(x)$ an $n \times m$ matrix.

We can express the entries of the matrix $Df(x)$ in terms of the partial derivatives of f:

$$Df(x)_{ij} = \frac{\partial f_i}{\partial x_j},$$

evaluated at x.

When f is real-valued (*i.e.*, $f : \mathbf{R}^m \to \mathbf{R}$) the derivative matrix Df is in $\mathbf{R}^{1\times m}$, *i.e.*, it is a *row* vector. Its transpose is called the *gradient* of the function:

$$\nabla f(x) = Df(x)^T,$$

which is a (column) vector, *i.e.*, in $\mathbf{R}^m$. We have

$$\nabla f(x)_i = \frac{\partial f}{\partial x_i},$$

evaluated at x. The first order approximation of a real-valued function f at a point x can be expressed as (the affine function of z)

$$f(x) + \nabla f(x)^T(z - x).$$

As a simple example consider the (real-valued) quadratic function

$$f(x) = x^T Px + 2q^T r + s$$

on $\mathbf{R}^n$ (where $P = P^T \in \mathbf{R}^{n\times n}$, $q \in \mathbf{R}^n$, and $s \in \mathbf{R}$ are given). Its derivative at x is given by the row vector $Df(x) = 2x^T P + 2q^T$, and its gradient is given by

$$\nabla f(x) = 2Px + 2q.$$

Let us consider another example, in which $\mathcal{V}$ is not $\mathbf{R}^m$. Consider the function f defined above in (A.4), *i.e.*, $f(X) = \log\det X$. If X is a symmetric positive definite matrix, the derivative $Df(X)$ is a real-valued linear function on S_n which satisfies

$$\lim_{Z\in\mathbf{dom}\, f,\, Z\to X} \frac{|\log\det Z - \log\det X - Df(X)(Z - X)|}{\|Z - X\|} = 0,$$

where the norm in the denominator is any matrix norm. The linear function $Df(X)$ is, in fact, given by

$$Df(X)(S) = \mathbf{Tr}\, X^{-1}S$$

for $S \in S_m$. In other words, we have

$$\lim_{Z\in\mathbf{dom}\, f,\ Z\to X} \frac{|\log\det Z - \log\det X - \mathbf{Tr}(Z-X)X^{-1}|}{\|Z-X\|} = 0.$$

We can give a more concrete representation of $Df(X)$ by noting that every linear functional L on S_m can be represented uniquely in the form $L(S) = \mathbf{Tr}\, FS$ for some $F \in S_m$. Using this representation we have the compact formula

$$Df(X) = X^{-1}.$$

A.3.3 Chain rule

Suppose $f : \mathcal{U} \to \mathcal{V}$ is differentiable at $x \in \mathbf{int\,dom}\, f$ and $g : \mathcal{V} \to \mathcal{W}$ is differentiable at $f(x) \in \mathbf{int\,dom}\, g$, where $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$ are vector spaces. Define the composition $h : \mathcal{U} \to \mathcal{W}$ by $h(z) = g(f(z))$. Then h is differentiable at x, with derivative

$$Dh(x) = Dg(f(x))Df(x), \tag{A.7}$$

where juxtaposition of the linear functions $Dg(f(x))$ and $Df(x)$ means composition. In the simple case when $\mathcal{U} = \mathbf{R}^m$, $\mathcal{V} = \mathbf{R}^n$, and $\mathcal{W} = \mathbf{R}^p$, the derivative $Dg(f(x))$ is a $p \times n$ matrix, the derivative $Df(x)$ is an $n \times m$ matrix, and juxtaposition on the righthand side of (A.7) denotes matrix multiplication.

We examine several special cases. First suppose that f is affine, *i.e.*, $f(x) = Ax + b$. Then $h(x) = g(Ax+b)$, and $Dh(x) = Dg(Ax+b)A$. If h is real-valued, we have

$$\nabla h(x) = A^T \nabla g(Ax+b).$$

As another special case we consider the restriction of a real-valued function f to a line given by $\{x + tv \mid t \in \mathbf{R}\}$, *i.e.*, the function $\tilde{f} : \mathbf{R} \to \mathbf{R}$ given by $\tilde{f}(t) = f(x+tv)$. Here we have

$$D\tilde{f}(t) = \tilde{f}'(t) = \nabla f(x+tv)^T v.$$

Now suppose that the function g in the composition is affine, *i.e.*, $h(x) = Af(x) + b$. Then its derivative at x is given by $Dh(x) = ADf(x)$. As an example suppose $f : \mathbf{R}^m \to \mathbf{R}^n$ is differentiable at $x \in \mathbf{int\,dom}\, f$, and we define $h : \mathbf{R}^m \to \mathbf{R}$ by $h(x) = c^T f(x)$, where $c \in \mathbf{R}^n$. Then we have $Dh(x) = c^T Df(x)$, so $\nabla h(x) = Df(x)^T c$.

A.3.4 Second derivative of a function

Suppose $f : \mathcal{V} \to \mathcal{W}$, where $\mathcal{V}$ and $\mathcal{W}$ are vector spaces, and $x \in \mathbf{dom}\, f$. We define $D^2 f(x)$, the second derivative of f at x, in terms of a quadratic (second-order) approximation of f at x.

A *quadratic function* $g : \mathcal{V} \to \mathcal{W}$ is one of the form

$$g(x) = g_0 + g_1(x) + g_2(x, x)$$

where $g_0 \in \mathcal{W}$, $g_1 : \mathcal{V} \to \mathcal{W}$ is a linear function, and $g_2 : \mathcal{V} \times \mathcal{V} \to \mathcal{W}$ is a symmetric bilinear function from $\mathcal{V}$ into $\mathcal{W}$. (This means $g_2(x, y) = g_2(y, x)$ for all $x, y \in \mathcal{V}$, and for each x, $g_2(x, y)$ is a linear function of y.) We say that g is a second-order approximation of f at x if

$$\lim_{z \in \mathbf{dom}\, f,\; z \to x} \frac{\|f(z) - g(z)\|}{\|z - x\|^2} = 0. \tag{A.8}$$

It can be shown that there is at most one quadratic function g that satisfies (A.8), so (when it exists) we refer to it as *the* second-order approximation of f at x. It is easy to show that if g is the second-order approximation of f at x, then we must have $g_0 = f(x)$ and $g_1 = Df(x)$. We *define* the second derivative of f at x by $D^2 f(x) = g_2$. Thus, the second derivative is a symmetric, bilinear function from $\mathcal{V}$ into $\mathcal{W}$. It can be concretely described in several ways, *e.g.*, using tensor notation.

For the special case in which $\mathcal{V} = \mathbf{R}^m$ and $\mathcal{W} = \mathbf{R}$ (which is the case most often encountered in this course) we can give a simple concrete representation of $Df^2(x)$, using the fact that any symmetric bilinear function $g_2 : \mathbf{R}^m \to \mathbf{R}$ can be expressed uniquely in the form

$$g_2(z) = z^T P z$$

where $P \in S_m$. When g_2 is the second derivative of f at x we refer to the associated symmetric matrix P as the *Hessian* of f at x, and denote it by $\nabla^2 f(x)$. It can be expressed in terms of partial derivatives as

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j},$$

evaluated at the point x.

Not surprisingly, the second derivative can be interpreted as the derivative of the first derivative. For example consider a real-valued differentiable function f on $\mathcal{V} = \mathbf{R}^m$. Consider its associated gradient mapping, *i.e.*, the function g that maps $x \in \mathbf{R}^m$ into $\nabla f(x) \in \mathbf{R}^m$. Then we have $Dg(x) = \nabla^2 f(x)$.

As a more complicated example, consider the function $f : S_m \to \mathbf{R}$ defined above, *i.e.*, $f(X) = \log\det X$. Here the second derivative, which is a symmetric bilinear function on $S_m \times S_m$, is given by

$$D^2 f(X)(Y, Z) = -\,\mathbf{Tr}\, X^{-1} Y X^{-1} Z$$

for $Y, Z \in S_m$.

A general chain rule for the second derivative is in principle simple, but cumbersome in most cases, so we will state it only for some special cases that we will need.

Suppose $f : \mathbf{R}^n \to \mathbf{R}$, $A \in \mathbf{R}^{n \times n}$, and $b \in \mathbf{R}^n$. Define $g : \mathbf{R}^m \to \mathbf{R}$ by $g(x) = f(Ax + b)$. Then we have

$$\nabla^2 g(x) = A^T \nabla^2 f(Ax + b) A.$$

As an example, consider the restriction of a real-valued function f to a line, *i.e.*, the function $\tilde{f}(t) = f(x + tv)$, where x and v are fixed. Then we have

$$\nabla^2 \tilde{f}(t) = f''(t) = v^T \nabla^2 f(x + tv) v.$$