Essential Mathematics for Economists

Alexis Akira Toda[1]

[1]Department of Economics, University of California San Diego. Email: atoda@ucsd.edu

Contents

Notations

Symbol	Meaning
$\forall x \ldots$	for all $x \ldots$
$\exists x \ldots$	there exists x such that $\ldots$
$\emptyset$	empty set
$x \in A$ or $A \ni x$	x is an element of the set A
$A \subset B$ or $B \supset A$	A is a subset of B; B contains A
$A \cap B$	intersection of sets A and B
$A \cup B$	union of sets A and B
$A \backslash B$	elements of A but not in B
$\mathbb{R}^N$	set of vectors $x = (x_1, \ldots, x_N)$ with $x_n \in \mathbb{R}$
$\mathbb{R}^N_+$	set of $x = (x_1, \ldots, x_N)$ with $x_n \geq 0$ for all n
$\mathbb{R}^N_{++}$	set of $x = (x_1, \ldots, x_N)$ with $x_n > 0$ for all n
$x \geq y$ or $y \leq x$	$x_n \geq y_n$ for all n; same as $x - y \in \mathbb{R}^N_+$
$x > y$ or $y < x$	$x_n \geq y_n$ for all n and $x_n > y_n$ for some n; same as $x - y \in \mathbb{R}^N_+ \backslash \{0\}$
$x \gg y$ or $y \ll x$	$x_n > y_n$ for all n; same as $x - y \in \mathbb{R}^N_{++}$
$\langle x, y \rangle$	inner product of x and y, $\langle x, y \rangle = x_1 y_1 + \cdots + x_N y_N$
$\|x\|$	norm of x, usually Euclidean norm $\|x\| = \sqrt{x_1^2 + \cdots + x_N^2}$
$\operatorname{cl} A$	closure of A
$\operatorname{int} A$	interior of A
$\operatorname{co} A$	convex hull of A
$[a, b]$	closed interval $\{x \,\vert\, a \leq x \leq b\}$
(a, b)	open interval $\{x \,\vert\, a < x < b\}$
$(a, b]$; $[a, b)$	half open intervals
$f : A \to B$	f is a function defined on A taking values in B
$\operatorname{dom} f$	effective domain of f, $\left\{x \in \mathbb{R}^N \,\middle\vert\, f(x) < \infty\right\}$
$\operatorname{epi} f$	epigraph of f, $\left\{(x, y) \in \mathbb{R}^N \times \mathbb{R} \,\middle\vert\, y \geq f(x)\right\}$
$f \in C(\Omega)$	function f is continuous on Ω
$f \in C^r(\Omega)$	function f is r times continuously differentiable on Ω
$f \leq g$ or $g \geq f$	$f(x) \leq g(x)$ for all x
$\nabla f(x)$	gradient (vector of partial derivatives) of f at x
$\nabla^2 f(x)$	Hessian (matrix of second derivatives) of f at x
$Df(x)$	Jacobian (matrix of partial derivatives) of f at x

Part I

Basics

Chapter 1

Linear Algebra

This chapter covers the most basic topics in linear algebra. This note is too short to cover all the details. Good references are Lax (2007) and Horn and Johnson (2013).

1.1 Linearity

In mathematics, "linear" means that a property is preserved by addition and multiplication by a constant. A *linear space* (more commonly *vector space*) is a set X for which $x+y$ (addition) and αx (multiplication by α) are defined, where $x, y \in X$ and $\alpha \in \mathbb{R}$. In this course we only encounter the *Euclidean space* $\mathbb{R}^N$, which consists of N-tuples of real numbers (called *N-vectors*)

$$x = (x_1, \dots, x_N).$$

Here addition and multiplication by a constant are defined entry-wise:

$$\begin{aligned}(x_1, \dots, x_N) + (y_1, \dots, y_N) &\coloneqq (x_1 + y_1, \dots, x_N + y_N)\\ \alpha(x_1, \dots, x_N) &\coloneqq (\alpha x_1, \dots, \alpha x_N).\end{aligned}$$

(The symbol ":=" means that we define the left-hand side by the right-hand side.)

A *linear function* is a function that preserves linearity. Thus $f : \mathbb{R}^N \to \mathbb{R}$ is linear if

$$\begin{aligned}f(x+y) &= f(x) + f(y),\\ f(\alpha x) &= \alpha f(x).\end{aligned}$$

An obvious example of a linear function $f : \mathbb{R}^N \to \mathbb{R}$ is

$$f(x) = a_1 x_1 + \cdots + a_N x_N = \sum_{n=1}^{N} a_n x_n,$$

where $a_1, \dots, a_N$ are numbers. In fact we can show that all linear functions are of this form.

Proposition 1.1. *If $f : \mathbb{R}^N \to \mathbb{R}$ is linear, then $f(x) = a_1x_1 + \cdots + a_Nx_N$ for some $a_1, \ldots, a_N$.*

Proof. Let $e_n = (0, \ldots, 0, 1, 0, \ldots, 0)$ be the vector whose n-th entry is 1 and all other entries are 0. (These vectors are called *unit* vectors.) By the definition of $\mathbb{R}^N$, we have

$$x = (x_1, \ldots, x_N) = x_1e_1 + \cdots + x_Ne_N.$$

Hence by the linearity of f, we get

$$f(x) = x_1f(e_1) + \cdots + x_Nf(e_N),$$

so $f(x)$ has the desired form by setting $a_n = f(e_n)$. □

A set of vectors $\{x_j\}_{j=1}^J \subset \mathbb{R}^N$ is called *linearly independent* if $\sum_{j=1}^J \alpha_j x_j = 0$ implies $\alpha_1 = \cdots = \alpha_J = 0$. Otherwise (if there is a combination $(\alpha_1, \ldots, \alpha_J) \neq (0, \ldots, 0)$ such that $\sum_{j=1}^J \alpha_j x_j = 0$), the set of vectors $\{x_j\}_{j=1}^J \subset \mathbb{R}^N$ is called *linearly dependent*.

1.2 Inner product and norm

An expression of the form $a_1x_1 + \cdots + a_Nx_N$ appears so often that it deserves a special name and notation. Let $x = (x_1, \ldots, x_N)$ and $y = (y_1, \ldots, y_N)$ be two vectors. Then

$$\langle x, y \rangle \coloneqq x_1y_1 + \cdots + x_Ny_N = \sum_{n=1}^N x_ny_n$$

is called the *inner product* (also *vector product*) of x and y.[1] The *(Euclidean) norm* of x is defined by

$$\|x\| \coloneqq \sqrt{\langle x, x \rangle} = \sqrt{x_1^2 + \cdots + x_N^2}.$$

The Euclidean norm is also called the L^2 norm for a reason that will be clear later.

Fixing x, the inner product $\langle x, y \rangle$ is linear in y, so we have

$$\langle x, y_1 + y_2 \rangle = \langle x, y_1 \rangle + \langle x, y_2 \rangle ,$$
$$\langle x, \alpha y \rangle = \alpha \langle x, y \rangle .$$

The same holds for x as well, fixing y. So the inner product is a *bilinear* function of x and y.

You might remember from high school algebra/geometry that the inner product in a two dimensional space satisfies

$$\langle x, y \rangle = x_1y_1 + x_2y_2 = \|x\| \, \|y\| \cos\theta,$$

where θ is the angle between the vector $x = (x_1, x_2)$ and $y = (y_1, y_2)$. Since

$$\cos\theta \begin{cases} > 0 & \text{if } \theta \text{ is an acute angle,} \\ = 0 & \text{if } \theta \text{ is a right angle,} \\ < 0 & \text{if } \theta \text{ is an obtuse angle,} \end{cases}$$

[1] The term "inner" is weird but this is because there is a notion of "outer product". The inner product of x, y is sometimes denoted by (x, y), $x \cdot y$, and $\langle x \mid y \rangle$, etc.

the vectors x, y are *orthogonal* if $\langle x, y\rangle = 0$ and form an acute (obtuse) angle if $\langle x, y\rangle > 0$ (< 0). Most of us cannot "see" higher dimensional spaces, but geometric intuition is very useful. For any $x, y \in \mathbb{R}^N$, we say that x, y are orthogonal if $\langle x, y\rangle = 0$.

The inner product and norms of vectors x, y satisfy the following Cauchy-Schwarz inequality: $|\langle x, y\rangle| \le \|x\| \, \|y\|$. The proof is in Problem 1.1. The norm $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$ satisfies

1. $\|x\| \ge 0$, with equality if and only if $x = 0$,
2. $\|\alpha x\| = |\alpha| \, \|x\|$ for all $\alpha \in \mathbb{R}$,
3. $\|x + y\| \le \|x\| + \|y\|$.

The last inequality is called the *triangle inequality* because it says that the length of any edge of any triangle is less than or equal to the sum of the length of the remaining two edges. (Draw a picture of a triangle with vertices at points 0, x, and $x + y$.) Proving the triangle inequality is an exercise.

In general, any function from $\mathbb{R}^N$ to $\mathbb{R}$ that satisfies the above three properties is called a norm. Other examples than the Euclidean norm are

$$\begin{aligned}
\|x\|_1 &\coloneqq \sum_{n=1}^{N} |x_n| \,, && (l^1 \text{ norm}) \\
\|x\|_\infty &\coloneqq \max_n |x_n| \,, && (l^\infty \text{ or sup norm}) \\
\|x\|_p &\coloneqq \left(\sum_{n=1}^{N} |x_n|^p \right)^{1/p} . && (l^p \text{ norm for } p \ge 1)
\end{aligned}$$

The proof that $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are norms is straightforward. The l^2 norm is the same as the Euclidean norm. The proof that $\|\cdot\|_p$ is a norm uses Minkowski's inequality, proved in Problem 5.9.

1.3 Matrix

Instead of a linear function $f : \mathbb{R}^N \to \mathbb{R}$, consider a *linear map* $f : \mathbb{R}^N \to \mathbb{R}^M$. This means that for each $x \in \mathbb{R}^N$, f associates a vector $f(x) \in \mathbb{R}^M$, and f is linear (preserves addition and multiplication by a constant): $f(x + y) = f(x) + f(y)$ and $f(\alpha x) = \alpha f(x)$. Let $f_m(x)$ be the m-th element of f, so $f(x) = (f_1(x), \ldots, f_M(x))$. It's easy to see that each $f_m(x)$ is a linear function of x. Hence by Proposition 1.1, we have

$$f_m(x) = a_{m1} x_1 + \cdots + a_{mN} x_N$$

for some numbers $a_{m1}, \ldots, a_{mN}$. Since this is true for any m, a linear map corresponds to numbers (a_{mn}), where $1 \le m \le M$ and $1 \le n \le N$. Conversely, any such array of numbers corresponds to a linear map. We write

$$A = (a_{mn}) = \begin{bmatrix}
a_{11} & \cdots & a_{1n} & \cdots & a_{1N} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
a_{m1} & \cdots & a_{mn} & \cdots & a_{mN} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
a_{M1} & \cdots & a_{Mn} & \cdots & a_{MN}
\end{bmatrix}$$

and call it a *matrix*. For an $M \times N$ matrix A and an N-vector x, we define the M-vector Ax by the vector whose m-th element is

$$a_{m1}x_1 + \cdots + a_{mN}x_N.$$

So $f : \mathbb{R}^N \to \mathbb{R}^M$ defined by $f(x) = Ax$ is a linear map. By defining addition and multiplication by a constant entry-wise, the set of all $M \times N$ matrices can be identified as $\mathbb{R}^{MN}$, the MN-dimensional Euclidean space.

Now consider the linear maps $f : \mathbb{R}^N \to \mathbb{R}^M$ and $g : \mathbb{R}^M \to \mathbb{R}^L$. Since f, g are linear, we can find an $M \times N$ matrix $A = (a_{mn})$ and an $L \times M$ matrix $B = (b_{lm})$ such that $f(x) = Ax$ and $g(y) = By$. We can also consider the composition of these two maps, $h = g \circ f$, where $h(x) \coloneqq g(f(x))$. It is easy to see that h is a linear map from $\mathbb{R}^N$ to $\mathbb{R}^L$, and therefore it can be written as $h(x) = Cx$ with an $L \times N$ matrix $C = (c_{ln})$. Using the definition $h(x) = g(f(x)) = B(Ax)$, it is not hard to see (exercise) that

$$c_{ln} = \sum_{m=1}^{M} b_{lm}a_{mn}.$$

So it makes sense to define the multiplication of matrix $C = BA$ by this rule. You can use all standard rules of algebra such as $B(A_1 + A_2) = BA_1 + BA_2$, $A(BC) = (AB)C$, etc. The proof is immediate by carrying out the algebra or thinking about linear maps. In Matlab, `A+B` and `A*B` return the sum and the product of matrices A, B (if they are well-defined). If A, B have the same size, then `A.*B` returns the entry-wise product (Hadamard product).

1.4 Identity matrix, inverse, determinant

An $M \times N$ matrix is *square* when $M = N$. The identity map $\mathrm{id} : \mathbb{R}^N \to \mathbb{R}^N$ defined by $\mathrm{id}(x) = x$ is clearly linear and has a corresponding matrix I. By simple calculation I is square, its diagonal (off-diagonal) entries are all 1 (0). Clearly $AI = IA = A$ when A is a square matrix (think of maps, or alternatively, do the entry-wise calculation). In Matlab, `eye(N)` returns the N-dimensional identity matrix.

A map $f : \mathbb{R}^N \to \mathbb{R}^N$ is said to be *one-to-one* (or *injective*) if $f(x) \neq f(y)$ whenever $x \neq y$. f is *onto* (or *surjective*) if for each $y \in \mathbb{R}^N$, there exists $x \in \mathbb{R}^N$ such that $y = f(x)$. f is *bijective* if f is both injective and surjective. If f is bijective, for each $y \in \mathbb{R}^N$ there exists a unique $x \in \mathbb{R}^N$ such that $y = f(x)$. Since this x depends only on y, we write $x = f^{-1}(y)$ and say that f^{-1} is the *inverse* of f. Now if $f : \mathbb{R}^N \to \mathbb{R}^N$ is a *bijective* linear map with a corresponding square matrix A (so $f(x) = Ax$), its inverse f^{-1} is also linear and hence has a matrix representation. We write this matrix A^{-1} and call it the *inverse* of A. Clearly $AA^{-1} = A^{-1}A = I$. The inverse of A, if it exists, is unique. To see this, suppose that B, C are both inverse of A. Then $AB = BA = I$ and $AC = CA = I$, so

$$B = BI = B(AC) = (BA)C = IC = C.$$

A matrix that has an inverse is called *regular*, *nonsingular*, *invertible*, etc. In Matlab, `inv(A)` returns the inverse of A.

If $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, then the determinant of A is $\det A = ad - bc$. In general, we can define the determinant of a square matrix inductively as follows. For 1×1 matrix $A = (a)$, we have $\det A = a$. Suppose that the determinant has been defined up to $(N-1) \times (N-1)$ matrices. If $A = (a_{mn})$ is $N \times N$, then

$$\det A = \sum_{n=1}^{N} (-1)^{m+n} a_{mn} M_{mn} = \sum_{m=1}^{N} (-1)^{m+n} a_{mn} M_{mn},$$

where M_{mn} is the determinant of the matrix obtained by removing row m and column n of A. It is well-known that this definition is consistent (i.e., does not depend on m, n). Following are useful properties of the determinant (see textbooks for proofs).

1. A is regular if and only if $\det A \neq 0$. In that case, we have $A^{-1} = \frac{1}{\det A} \tilde{A}$, where $\tilde{A} = (\tilde{a}_{mn})$ satisfies $\tilde{a}_{mn} = (-1)^{m+n} M_{nm}$. For example, if $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, then $A^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.
2. If A, B are square matrices of the same order, then $\det(AB) = (\det A)(\det B)$.
3. If A is partitioned as $A = \begin{bmatrix} A_{11} & A_{12} \\ O & A_{22} \end{bmatrix}$, where A_{11} and A_{22} are square matrices, then $\det A = (\det A_{11})(\det A_{22})$.

In Matlab, `det(A)` returns the determinant of A.

1.5 Transpose, symmetric matrices

When numbers are stacked horizontally like $x = (x_1, \ldots, x_N)$, it is called a *row vector*. When stacked vertically like

$$\begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix},$$

it is a *column vector*. An N-column vector is the same as an $N \times 1$ matrix. An N-row vector is the same as a $1 \times N$ matrix. The notation $f(x) = Ax$ is compatible with the definition of the product of an $M \times N$ matrix A and an $N \times 1$ matrix x. To see this, writing down the entries, we get

$$Ax = \begin{bmatrix} a_{11}x_1 + \cdots + a_{1N}x_N \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mN}x_N \\ \vdots \\ a_{M1}x_1 + \cdots + a_{MN}x_N \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} & \cdots & a_{mN} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{M1} & \cdots & a_{1n} & \cdots & a_{1N} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \\ \vdots \\ x_N \end{bmatrix} = Ax.$$

(The left-most Ax is the linear map; the right-most Ax is the multiplication of the $M \times N$ matrix A and the $N \times 1$ matrix x.)

Unless otherwise specified, vectors are always assumed to be column vectors. However, it is awkward to write down column vectors every time because they take up a lot of space, so we use the notation $(x_1, \dots, x_N)'$ (with a prime) to denote a column vector. The vector $(x_1, \dots, x_N)'$ is called the *transpose* of the row vector $x = (x_1, \dots, x_N)$. Sometimes, to distinguish from derivatives, we also use the symbol $^\top$ instead of $'$ to denote the transpose. Oftentimes we are even more sloppy and don't distinguish between a row and column vector. (After all, what does it mean *mathematically* to stack numbers horizontally or vertically?)

Transpose can be defined for matrices, too. For an $M \times N$ matrix $A = (a_{mn})$, we define its *transpose* by the $N \times M$ matrix $B = (b_{nm})$, where $b_{nm} = a_{mn}$, and we write $B = A'$ (or $B = A^\top$). Thus A' is the matrix obtained by flipping A "diagonally". In Matlab, `A'` returns the transpose of A.

A square matrix P such that $P'P = PP' = I$ is called *orthogonal*, because by definition the column vectors of P are orthogonal and have Euclidean norm 1 (just write down the entries of $P'P$). If P is orthogonal, then clearly $P^{-1} = P'$.

A square matrix A such that $A = A'$ is called *symmetric*, because its entries are symmetric about the diagonal. A is *positive semidefinite* if $\langle x, Ax \rangle \geq 0$ for all x, and *positive definite* if in addition $\langle x, Ax \rangle = 0$ only if $x = 0$. Symmetric matrices have a natural (partial) order (exercise): we write $A \succeq B$ if and only if $A - B$ is positive semidefinite.

There is a simple test for positive definiteness. Let A be square. The determinant of the matrix obtained by keeping the first k-th rows and columns of A is called the k-th *principal minor* of A. For example, if $A = (a_{mn})$ is $N \times N$, then the first principal minor is a_{11}, the second principal minor is $a_{11}a_{22} - a_{12}a_{21}$, and the N-th principal minor is $\det A$, etc.

Proposition 1.2. *Let A be real symmetric. Then A is positive definite if and only if its principal minors are all positive.*

Proof. We prove by mathematical induction on the dimension N of the matrix A. If $N = 1$, the claim is trivial.

Suppose the claim is true up to dimension $N-1$, and let A be N-dimensional. Partition A as $A = \begin{bmatrix} A_1 & b \\ b' & c \end{bmatrix}$, where A_1 is an $(N-1)$-dimensional symmetric matrix, b is an $(N-1)$-dimensional vector, and c is a scalar. Let

$$P = \begin{bmatrix} I & -A_1^{-1}b \\ 0 & 1 \end{bmatrix}.$$

Then by simple algebra we get

$$P'AP = \begin{bmatrix} A_1 & 0 \\ 0 & c - b'A_1^{-1}b \end{bmatrix}.$$

Clearly $\det P = 1$, so P is regular. Since

$$\langle x, Ax \rangle = x'Ax = (P^{-1}x)'(P'AP)(P^{-1}x),$$

A is positive definite if and only if $P'AP$ is. But since $P'AP$ is block diagonal, $P'AP$ is positive definite if and only if A_1 is positive definite and $c - b'A^{-1}b > 0$.

By assumption, A_1 is positive definite if and only if its principal minors are all positive. Furthermore, since $\det P = 1$, we get

$$\det A = \det(P'AP) = (\det A_1)(c - b'A_1^{-1}b).$$

Therefore

$$\begin{aligned} A \succ O &\iff \text{all principal minors of } A_1 \text{ are positive and } c - b'A_1^{-1}b > 0 \\ &\iff \text{all principal minors of } A_1 \text{ are positive and } \det A > 0 \\ &\iff \text{all principal minors of } A \text{ are positive,} \end{aligned}$$

so the claim is true for N as well. □

1.6 Eigenvector, diagonalization

If A is a square matrix and there exist a number α and a nonzero vector v such that $Av = \alpha v$, then we say that v is an *eigenvector* of A associated with *eigenvalue* α. Since

$$Av = \alpha v \iff (\alpha I - A)v = 0,$$

α is an eigenvalue of A if and only if $\det(\alpha I - A) = 0$ (for otherwise $\alpha I - A$ is invertible, which would imply $v = 0$, a contradiction). The polynomial $\Phi_A(x) \coloneqq \det(xI - A)$ is called the *characteristic polynomial* of A. In Matlab, `eig(A)` returns the eigenvalues of A.

Even if A is a real matrix, eigenvalues and eigenvectors need not be real. For complex vectors x, y, the inner product is defined by

$$\langle x, y \rangle = x^* y = \bar{x}' y = \sum_{n=1}^{N} \bar{x}_n y_n,$$

where $\bar{x}$ is the complex conjugate of x and $x^* = \bar{x}'$ is the transpose of the complex conjugate of x (called adjoint). By definition, $\overline{\langle x, y \rangle} = \langle y, x \rangle$. Similarly, for a complex matrix A, its adjoint A^* is defined by the complex conjugate of the transpose. It is easy to see that $\langle x, Ay \rangle = \langle A^* x, y \rangle$, because

$$\langle A^* x, y \rangle = (A^* x)^* y = x^* (A^*)^* y = x^* A y = \langle x, Ay \rangle .$$

Matrices satisfying $A^* = A$ are called *Hermite*. If A is real, then an Hermite matrix is the same as a symmetric matrix. For an Hermite matrix A, the quadratic form $\langle x, Ax \rangle$ is real, for

$$\overline{\langle x, Ax \rangle} = \langle Ax, x \rangle = \langle A^* x, x \rangle = \langle x, Ax \rangle .$$

Proposition 1.3. *The eigenvalues of an Hermite matrix are real.*

Proof. Suppose that $Av = \alpha v$ with $v \neq 0$. Then

$$\langle v, Av \rangle = \langle v, \alpha v \rangle = \alpha \langle v, v \rangle = \alpha \|v\|^2$$

is real, so $\alpha = \langle v, Av \rangle / \|v\|^2$ is also real. □

Since real symmetric matrices are Hermite, the eigenvalues of real symmetric matrices are all real (and so are eigenvectors).

If U is a square matrix such that $U^*U = UU^* = I$, then U is called *unitary*. Real unitary matrices are orthogonal, by definition.

We usually take the standard basis $\{e_1, \dots, e_N\}$ in $\mathbb{R}^N$, but that is not necessary. Suppose we take vectors $\{p_1, \dots, p_N\}$, where the matrix $P = [p_1, \dots, p_N]$ is regular. Let x be any vector and $y = P^{-1}x$. Then

$$x = PP^{-1}x = Py = y_1p_1 + \cdots + y_Np_N,$$

so the entries of y can be interpreted as the coordinates of x when we use the basis P. What does a matrix A look like when we use the basis P? Consider the linear map $x \mapsto Ax$. Using the basis P, this map looks like $P^{-1}x \mapsto P^{-1}Ax = (P^{-1}AP)(P^{-1}x)$, so the linear map $x \mapsto Ax$ has the matrix representation $B = P^{-1}AP$. Oftentimes, it is useful to find a matrix P such that $P^{-1}AP$ is a simple matrix. The simplest matrix of all are diagonal ones. If we can find P such that $P^{-1}AP$ is diagonal, we say that A is diagonalizable. A remarkable property of real symmetric matrices is that they are diagonalizable with some orthogonal matrix.

Theorem 1.4. *Let A be real symmetric. Then there exists a real orthogonal matrix P such that $P^{-1}AP = P'AP = \operatorname{diag}[\alpha_1, \dots, \alpha_N]$, where $\alpha_1, \dots, \alpha_N$ are eigenvalues of A. (*diag *is the symbol for diagonal matrix with specified diagonal entries.)*

Proof. We prove by mathematical induction on the dimension of A. If A is one-dimensional (scalar), then the claim is trivial. (Just take $P = 1$.)

Suppose that the claim is true up to dimension $N-1$. Let α_1 be an eigenvalue of A and p_1 be an associated eigenvector, so $Ap_1 = \alpha_1p_1$. Let W_1 be the set of vectors orthogonal to p_1, so $W_1 = \{x \mid \langle p_1, x\rangle = 0\}$. If $x \in W_1$, then

$$\langle p_1, Ax\rangle = \langle A'p_1, x\rangle = \langle Ap_1, x\rangle = \langle \alpha p_1, x\rangle = \alpha \langle p_1, x\rangle = 0,$$

so $Ax \in W_1$. Pick an orthonormal basis $q_2, \dots, q_N$ in W_1. Letting $P_1 = [p_1, q_2, \dots, q_N]$, then we have

$$AP_1 = P_1 \begin{bmatrix} \alpha_1 & 0 \\ 0 & A_1 \end{bmatrix} \iff P_1'AP_1 = \begin{bmatrix} \alpha_1 & 0 \\ 0 & A_1 \end{bmatrix},$$

where A_1 is some $N-1$-dimensional matrix. To see this, note that

$$AP_1 = [Ap_1, Aq_2, \dots, Aq_N] = [\alpha p_1, Aq_2, \dots, Aq_N].$$

Since each q_n belongs to the space W_1, Aq_n is a linear combination of $q_2, \dots, q_N$, so there exists a matrix A_1 above. Since A is symmetric, $(P_1'AP_1)' = P_1'A(P_1)'' = P_1'AP_1$, so $P_1'AP_1$ is also symmetric. Therefore A_1 is symmetric. Since A_1 is $(N-1)$-dimensional, by assumption we can take an orthogonal matrix P_2 such that $P_2'A_1P_2 = D_1$, where $D_1 = \operatorname{diag}[\alpha_2, \dots, \alpha_N]$ is diagonal. Then $A_1 = P_2D_1P_2'$, so

$$P_1'AP_1 = \begin{bmatrix} \alpha_1 & 0 \\ 0 & P_2D_1P_2' \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & P_2 \end{bmatrix} \begin{bmatrix} \alpha_1 & 0 \\ 0 & D_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & P_2 \end{bmatrix}'.$$

Letting $P = P_1 \begin{bmatrix} 1 & 0 \\ 0 & P_2 \end{bmatrix}$, which is orthogonal, then $P'AP = \begin{bmatrix} \alpha_1 & 0 \\ 0 & D_1 \end{bmatrix} = \operatorname{diag}[\alpha_1, \dots, \alpha_N]$, so $P'AP$ is diagonal. □

Similarly, Hermite matrices can be diagonalized by unitary matrices. Diagonalization is often useful for proving theorems, see for example Problems 1.8, 1.9, and 1.10.

1.7 Jordan canonical form

Two matrices A, B are said to be *similar* if there exists a regular matrix S such that $B = S^{-1}AS$. Sometimes we want to find a simple matrix that is similar to a given matrix. We know from Theorem 1.4 that if A is Hermite, then we can find a unitary matrix U and a diagonal matrix D such that $D = U^*AU = U^{-1}AU$, so we can take $B = D$ and $S = U$. However, in general not all matrices are diagonalizable (Problem 1.12). Jordan's theorem allows us to reduce any matrix to a simple one, called the *Jordan canonical form*. I omit the proof since it is tedious but refer to textbooks (Lax, 2007, Appendix 15).

An $n \times n$ Jordan matrix with diagonal element λ is defined by

$$J_n(\lambda) = \begin{bmatrix} \lambda & 1 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & \lambda \end{bmatrix},$$

so the diagonal entries are λ, the super diagonal entries are 1, and all other entries are 0.

Jordan's theorem. *For any matrix A, there exists a regular matrix S such that*

$$S^{-1}AS = \begin{bmatrix} J_{n_1}(\lambda_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & J_{n_k}(\lambda_k) \end{bmatrix} = D + N,$$

where D is a diagonal matrix and N is a matrix whose super diagonal entries are either 0 or 1 and all other entries are 0, and $DN = ND$.

Jordan's theorem is useful for computing matrix powers analytically. For example, using $DN = ND$ and the binomial theorem, we obtain

$$S^{-1}A^nS = (S^{-1}AS)^n = \sum_{k=0}^{n} \binom{n}{k} D^{n-k}N^k.$$

It is straightforward to show that $N^k = O$ for large enough k.

1.8 Matrix norm, spectral radius

The *matrix norm* is a function that satisfies

1. (positivity) $\|A\| \geq 0$, with equality if and only if $A = O$,
2. (scalar multiplicativity) $\|\alpha A\| = |\alpha| \, \|A\|$,
3. (triangle inequality) $\|A + B\| \leq \|A\| + \|B\|$,

4. (submultiplicativity) $\|AB\| \leq \|A\| \, \|B\|$.

The following observation shows that there is a matrix norm associated with any norm on $\mathbb{R}^N$. Let $\|\cdot\|$ be any norm on $\mathbb{R}^N$. For any $N \times N$ matrix A, define

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

Then it is easy to show (Problem 1.13) that $\|\cdot\|$ is a matrix norm, called the *operator norm*.

Let $\{\alpha_n\}_{n=1}^N$ be the eigenvalues of a matrix A. The quantity

$$\rho(A) = \max_n |\alpha_n| ,$$

the largest modulus of all eigenvalues, is called the *spectral radius* of A. The spectral radius and the matrix norm are related as follows.

Proposition 1.5 (Gelfand spectral radius formula)**.** *Let $\|\cdot\|$ be any matrix norm. Then $\rho(A) \leq \|A^n\|^{1/n}$ and $\rho(A) = \lim_{n\to\infty} \|A^n\|^{1/n}$.*

Proof. Let α be an eigenvalue of A and $x \neq 0$ be a corresponding eigenvector. Then $A^n x = \alpha^n x$ for all n. Let $X = (x, \ldots, x)$ be the matrix obtained by replicating x. Then $A^n X = \alpha^n X$. Taking the norm of both sides, we obtain

$$|\alpha|^n \|X\| = \|A^n X\| \leq \|A^n\| \, \|X\| \implies |\alpha|^n \leq \|A^n\| .$$

Since α is any eigenvalue, it follows that $\rho(A) \leq \|A^n\|^{1/n}$.

Take any $\epsilon > 0$ and define $\tilde{A} = \frac{1}{\rho(A)+\epsilon} A$. Then $\rho(\tilde{A}) = \frac{\rho(A)}{\rho(A)+\epsilon} < 1$, so considering the Jordan canonical form, it follows that $\lim_{n\to\infty} \tilde{A}^n = O$. Therefore $\left\|\tilde{A}^n\right\| < 1$ for large enough n, and hence $\|A^n\| \leq (\rho(A) + \epsilon)^n$. Taking the n-th root, letting $n \to \infty$, and $\epsilon \downarrow 0$, we obtain $\limsup_{n\to\infty} \|A^n\|^{1/n} \leq \rho(A)$. Since $\rho(A) \leq \|A^n\|^{1/n}$, it follows that $\rho(A) = \lim_{n\to\infty} \|A^n\|^{1/n}$. □

Remark. Although the above proof of Proposition 1.5 uses the submultiplicativity of the matrix norm, this condition is actually not necessary. For a proof of the Gelfand formula that does not require submultiplicativity, see Theorem 5.7.10 of Horn and Johnson (2013).

1.9 Nonnegative matrices

Nonnegative matrices, though not usually treated in introductory textbooks of linear algebra, play an important role in economics. This section provides a brief introduction. For a more complete treatment, see Chapter 8 of Horn and Johnson (2013), Berman and Plemmons (1994), or Bapat and Raghavan (1997).

I first discuss a motivating example. Suppose a worker can be either employed or unemployed. If employed, he will be unemployed with probability p next period. If unemployed, he will be employed with probability q next period. Let $x_t = (e_t, u_t)'$ be the probability vector of being employed and unemployed at time t, where $u_t = 1 - e_t$. Then by assumption we have

$$\begin{aligned} e_{t+1} &= (1-p)e_t + qu_t, \\ u_{t+1} &= pe_t + (1-q)u_t. \end{aligned}$$

Putting these equations into vector form, we obtain $x_{t+1} = P'x_t$, where

$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}.$$

Given the initial probability x_0, one might be interested in the probability vector x_t at time t and its behavior as $t \to \infty$. For this example we can easily calculate these as follows. First, note that $x_t = (P')^t x_0$, so it suffices to compute P^t. The characteristic polynomial of P is

$$\begin{aligned}\Phi_P(x) = |xI - P| &= \begin{vmatrix} x-1+p & -p \\ -q & x-1+q \end{vmatrix} \\ &= x^2 + (p+q-2)x + 1 - p - q = (x-1)(x+p+q-1).\end{aligned}$$

Assuming $0 < p, q < 1$, P has two eigenvalues 1 and $1-p-q \in (-1,1)$. Therefore the spectral radius of P is 1. We can easily show that

$$P\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}1\\1\end{bmatrix} \quad \text{and} \quad P\begin{bmatrix}p\\-q\end{bmatrix} = (1-p-q)\begin{bmatrix}p\\-q\end{bmatrix}.$$

Therefore

$$S^{-1}PS = D = \begin{bmatrix} 1 & 0 \\ 0 & 1-p-q \end{bmatrix},$$

where

$$S = \begin{bmatrix} 1 & p \\ 1 & -q \end{bmatrix}.$$

Therefore

$$\begin{aligned}P^t = SD^tS^{-1} &= \frac{1}{p+q}\begin{bmatrix}1 & p\\ 1 & -q\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & (1-p-q)^t\end{bmatrix}\begin{bmatrix}q & p\\ 1 & -1\end{bmatrix} \\ &= \frac{1}{p+q}\begin{bmatrix} q+p(1-p-q)^t & p(1-(1-p-q)^t) \\ q(1-(1-p-q)^t) & p+q(1-p-q)^t \end{bmatrix}.\end{aligned}$$

Since $|1-p-q| < 1$, letting $t \to \infty$, we obtain

$$P^t \to \frac{1}{p+q}\begin{bmatrix}q & p\\ q & p\end{bmatrix}.$$

Thus regardless of $x_0 = (e_0, u_0)$, we obtain

$$x_t = (P')^t x_0 \to \frac{1}{p+q}\begin{bmatrix}q & q\\ p & p\end{bmatrix}\begin{bmatrix}e_0\\ u_0\end{bmatrix} = \frac{1}{p+q}\begin{bmatrix}q\\ p\end{bmatrix},$$

so the worker eventually becomes unemployed with probability $\frac{p}{p+q}$.

The above example can be generalized as follows. We say that a square matrix $A = (a_{mn})$ is *positive* if $a_{mn} > 0$ for all m, n, and we write $A \gg 0$.

Theorem 1.6 (Perron)**.** *Let A be a square positive matrix. Then*

1. *$\rho(A) > 0$, which is an eigenvalue of A (called the Perron root),*
2. *there exist positive vectors x, y (called the right and left Perron vectors) such that $Ax = \rho(A)x$ and $y'A = \rho(A)y'$,*

3. *$\rho(A)$ is geometrically simple (hence x, y are unique up to a multiplicative constant), and*

4. *If x, y are chosen such that $y'x = 1$, then $\lim_{k\to\infty}[\frac{1}{\rho(A)}A]^k = xy'$.*

The proof of Theorem 1.6 is deferred to Chapter 8. When A is only nonnegative, some of the conclusions of Theorem 1.6 hold by taking limits. (See Chapter 2 for the definition of limits.)

Corollary 1.7. *Let A be a square nonnegative matrix. Then $\rho(A)$ is an eigenvalue of A, and there exist nonnegative vectors x, y such that $Ax = \rho(A)x$ and $y'A = \rho(A)y'$.*

Proof. Let $A = (a_{nn'})$ and define $A(\epsilon)$ by $A(\epsilon)_{nn'} = a_{nn'} + \epsilon$, where $\epsilon > 0$. Applying Perron's theorem to $A(\epsilon)$, there exists a positive vector $x(\epsilon)$ with $\|x(\epsilon)\| = 1$ such that $A(\epsilon)x(\epsilon) = \rho(A(\epsilon))x(\epsilon)$. Since roots of polynomials are continuous in the coefficients (Harris and Martin, 1987), we obtain $\rho(A(\epsilon)) \to \rho(A)$ as $\epsilon \downarrow 0$. By taking a subsequence, there exists a nonnegative vector x such that $x(\epsilon) \to x$ as $\epsilon \downarrow 0$. Therefore $Ax = \rho(A)x$. The same is true for the left eigenvector y. □

With an additional assumption called irreducibility, Perron's theorem for nonnegative matrices (Corollary 1.7) can be further strengthened, which is known as the Perron-Frobenius theorem. See Theorem 8.4.4 of Horn and Johnson (2013) for details.

In some applications, we need to deal with square matrices $A = (a_{nn'})$ with nonnegative off-diagonal entries ($a_{nn'} \geq 0$ if $n \neq n'$), although A may have negative diagonal entries. We call such matrices *Metzler*. If A is Metzler, since by definition its off-diagonal entries are nonnegative, the matrix $A + dI$ becomes nonnegative if $d \geq 0$ is large enough. This observation enables us to establish properties of Metzler matrices using the Perron-Frobenius theorem. For Metzler matrices, the role of the spectral radius $\rho(A)$ is replaced by the *spectral abscissa*

$$\zeta(A) = \max\{\operatorname{Re}\alpha \,|\, \alpha \text{ is an eigenvalue of } A\},$$

which is the maximum real part of all eigenvalues.

The following theorem is the analogue of the Perron-Frobenius theorem for Metzler matrices.

Theorem 1.8. *Let A be a Metzler matrix. Then the spectral abscissa $\zeta(A)$ is an eigenvalue of A, and there exist nonnegative vectors x, y such that $Ax = \zeta(A)x$ and $y'A = \zeta(A)y'$. If in addition A has positive off-diagonal entries (more generally, if A is irreducible), then x, y are positive vectors and unique up to a multiplicative constant.*

Proof. Immediate by applying Corollary 1.7 or the Perron-Frobenius theorem to the matrix $A + dI$, where $d \geq 0$ is large enough such that $A + dI \geq 0$. □

Problems

1.1. Let $x = (x_1, \ldots, x_N)$ and $y = (y_1, \ldots, y_N)$ be vectors in $\mathbb{R}^N$. Define $f(t) = \|tx - y\|^2$, where $t \in \mathbb{R}$.

1. Expand f and express it as a quadratic function of t.

2. Prove the Cauchy-Schwarz inequality $|\langle x, y\rangle| \leq \|x\| \, \|y\|$. (Hint: how many solutions does the quadratic equation $f(t) = 0$ have? Make sure to treat the cases $x = 0$ and $x \neq 0$ separately.)

1.2. Prove the triangle inequality $\|x + y\| \leq \|x\| + \|y\|$. (Hint: Cauchy-Schwarz inequality.)

1.3. Let A, B, C be matrices with appropriate dimensions so that the following expressions are well defined. Prove that $A(B + C) = AB + AC$, $A(BC) = (AB)C$, $(AB)^{-1} = B^{-1}A^{-1}$, and $(AB)' = B'A'$.

1.4. 1. Let A be a 2×2 block upper triangular matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}.$$

If A is invertible, explicitly compute A^{-1}.

2. Repeat the above problem if A is 3×3 block upper triangular. What if A is $N \times N$ block upper triangular?

1.5. Let A be an $M \times N$ matrix and write $A = [a_1, \ldots, a_N]$, where a_n is the n-th column vector of A. Show that the set of vectors $\{a_n\}_{n=1}^N$ are linearly independent if and only if the linear map $x \mapsto Ax$ is injective.

1.6. Let A, B, C be real symmetric matrices of the same size.

1. Prove that $A \succeq A$ (reflexivity).

2. Prove that $A \succeq B$ and $B \succeq A$ imply $A = B$ (antisymmetry).

3. Prove that $A \succeq B$ and $B \succeq C$ imply $A \succeq C$ (transitivity).

Hence $\succeq$ is a partial order for real symmetric matrices.

1.7. Let P be a matrix such that $P^2 = P$. Show that the eigenvalues of P are either 0 or 1.

1.8. Let A be real symmetric. Show that A is positive definite if and only if all eigenvalues of A are positive.

1.9. Let A be real symmetric and positive semidefinite. Show that there exists a real symmetric and positive semidefinite matrix B such that $A = B^2$.

1.10. Let A be real symmetric with eigenvalues $\alpha_1, \ldots, \alpha_N$, where $|\alpha_1| \leq \cdots \leq |\alpha_N|$. Let $\|\cdot\|$ be the Euclidean norm. Show that for any nonzero vector $x \in \mathbb{R}^N$, we have $|\alpha_1| \leq \|Ax\| / \|x\| \leq |\alpha_N|$.

1.11. Let A be an $M \times N$ matrix. Let a_n be the n-th column vector of A, so $A = [a_1, \ldots, a_N]$. Show that the matrix $A'A$ is positive definite if and only if the set of vectors $\{a_1, \ldots, a_N\}$ is linearly independent.

1.12. 1. Let $A = (a_{mn})$ be an upper triangular matrix, i.e., $a_{mn} = 0$ whenever $m > n$. Prove that the eigenvalues of A are the diagonal entries $\{a_{nn}\}_{n=1}^N$.

2. Prove that the matrix $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ is not diagonalizable.

1.13. Let $\|\cdot\|$ be any norm on $\mathbb{R}^N$. For any $N \times N$ matrix A, define

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

Show that $\|A\|$ is a matrix norm.

1.14. Let A be a square nonnegative matrix. Show that if $z > \rho(A)$, then the matrix $zI - A$ is regular and $(zI - A)^{-1}$ is nonnegative. (Hint: let $B = \frac{1}{z}A$ and consider the identity $(I - B)(I + \cdots + B^{k-1}) = I - B^k$.)

1.15. 1. Let A, B be square nonnegative matrices such that $0 \leq A \leq B$ entry-wise. Show that $\rho(A) \leq \rho(B)$.

2. Show that if A is a square positive matrix, then $\rho(A) > 0$.

1.16. Let A be a square nonnegative matrix. If there exists a positive eigenvector $x \gg 0$ with eigenvalue α, so $Ax = \alpha x$, show that $\alpha = \rho(A)$. (Hint: let $y > 0$ be a left eigenvector corresponding to $\rho(A)$, and multiply y' from left to $Ax = \alpha x$.)

1.17. If $\alpha_1, \ldots, \alpha_N$ are eigenvalues of a square matrix A, for any scalar z, show that the eigenvalues of $A + zI$ are $\alpha_1 + z, \ldots, \alpha_N + z$. Use this property to fill in the details of the proof of Theorem 1.8.

1.18. Let $A = (a_{nn'})$ be a Metzler matrix such that

$$a_{nn} = -\frac{1}{d_n} \sum_{n' \neq n} a_{nn'} d_{n'},$$

where $d_n > 0$ for all n.

1. Define the vector $d = (d_1, \ldots, d_N)'$. Show that $Ad = 0$.

2. Show that the spectral abscissa of A is $\zeta(A) = 0$. (Hint: let $y > 0$ be a left eigenvector corresponding to $\zeta(A)$, and multiply y' from left to the identity $(A - \zeta(A)I)d = -\zeta(A)d$.)

Chapter 2

Topology in Euclidean Spaces

2.1 Convergence of sequences

By the triangle inequality, the (Euclidean) norm $\|\cdot\|$ on $\mathbb{R}^N$ can be used to define a distance. For $x, y \in \mathbb{R}^N$, we define the distance between these two points by

$$\operatorname{dist}(x, y) = \|x - y\| .$$

Let $\{x_k\}_{k=1}^{\infty}$ be a sequence in $\mathbb{R}^N$. (Here each $x_k = (x_{1k}, \ldots, x_{Nk})$ is a vector in $\mathbb{R}^N$.) We say that $\{x_k\}_{k=1}^{\infty}$ *converges* to $x \in \mathbb{R}^N$ if

$$(\forall \epsilon > 0)(\exists K > 0) k > K \Longrightarrow \|x_k - x\| < \epsilon,$$

that is, for any small error tolerance $\epsilon > 0$, we can find a large enough number K such that the distance between x_k and x can be made smaller than the error tolerance ϵ, provided that the index satisfies $k > K$. When $\{x_k\}_{k=1}^{\infty}$ converges to x, we write $\lim_{k\to\infty} x_k = x$ or $x_k \to x$ $(k \to \infty)$. Sometimes we are sloppy and write $\lim x_k = x$ or $x_k \to x$. A sequence $\{x_k\}_{k=1}^{\infty}$ is *convergent* if it converges to some point.

An acute reader may notice that we have defined the convergence of a sequence using the Euclidean norm, and thus may be worried that a sequence $\{x_k\}_{k=1}^{\infty}$ may converge to x with respect to the Euclidean norm but not with respect to another norm. A remarkable property of finite dimensional spaces $(\mathbb{R}^N)$ is that it does not matter which norm we use to define convergence.

Theorem 2.1 (Equivalence of norms in $\mathbb{R}^N$)**.** *Let $\|\cdot\|_1$, $\|\cdot\|_2$ be two norms on $\mathbb{R}^N$. Then there exist constants $0 < c \le C$ such that*

$$c \|x\|_1 \le \|x\|_2 \le C \|x\|_1 \tag{2.1}$$

for all $x \in \mathbb{R}^N$. Consequently, a sequence that is convergent with respect to one norm is convergent with respect to any other norm.

The proof of Theorem 2.1 is in Problem 2.5. In general, two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ are said to be equivalent if (2.1) holds. Here let us show that the Euclidean

and sup norms

$$\|x\|_2 \coloneqq \sqrt{\sum_{n=1}^{N} x_n^2} \quad \text{and} \quad \|x\|_\infty \coloneqq \max_n |x_n|$$

are equivalent. Clearly

$$\|x\|_2 = \sqrt{\sum_{n=1}^{N} x_n^2} \geq |x_n|$$

for any n, so taking the maximum over n, we get $\|x\|_2 \geq \|x\|_\infty$. Furthermore, since by definition $|x_n| \leq \|x\|_\infty$ for all n, we get

$$\|x\|_2 = \sqrt{\sum_{n=1}^{N} x_n^2} \leq \sqrt{N \|x\|_\infty^2} = \sqrt{N}\, \|x\|_\infty .$$

Therefore

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{N}\, \|x\|_\infty ,$$

so we can take $c = 1$ and $C = \sqrt{N}$ in (2.1).

A sequence $\{x_k\}_{k=1}^\infty$ is *bounded* if there exists $b > 0$ such that $\|x_k\| \leq b$ for all k. More generally, a set $A \subset \mathbb{R}^N$ is bounded if there exists $b > 0$ such that $\|x\| \leq b$ for all $x \in A$. By Theorem 2.1, it does not matter which norm we use to define bounded sequences or sets.

Proposition 2.2. *A convergent sequence is bounded.*

Proof. Suppose that $x_k \to x$. Setting $\epsilon = 1$ in the definition of convergence, we can take $K > 0$ such that $\|x_k - x\| < 1$ for all $k > K$. By the triangle inequality, we have $\|x_k\| \leq \|x\| + 1$ for $k > K$. Therefore

$$\|x_k\| \leq b \coloneqq \max\{\|x_1\|, \ldots, \|x_K\|, \|x\| + 1\} . \quad \square$$

The sequence $\{x_{k_l}\}_{l=1}^\infty$ is called a *subsequence* of $\{x_k\}_{k=1}^\infty$ if $k_1 < k_2 < \cdots < k_l < \cdots$. The following proposition shows that the subsequence of a convergent sequence converges to the same limit.

Proposition 2.3. *If $x_k \to x$ and $\{x_{k_l}\}_{l=1}^\infty$ is a subsequence of $\{x_k\}_{k=1}^\infty$, then $x_{k_l} \to x$.*

Proof. By the definition of convergence, for any $\epsilon > 0$ we can take $K > 0$ such that $\|x_k - x\| < \epsilon$ whenever $k > K$. Since $k_l \geq l$, it follows that $\|x_{k_l} - x\| < \epsilon$ whenever $l > K$, so $x_{k_l} \to x$. $\square$

Let $\{x_k\} \subset \mathbb{R}$ be a real sequence. Define $\alpha_l = \sup_{k \geq l} x_k$ and $\beta_l = \inf_{k \geq l} x_k$, possibly $\pm\infty$. Clearly $\{\alpha_l\}$ is decreasing and $\{\beta_l\}$ is increasing, so they have limits α, β in $[-\infty, \infty]$ by the continuity property of the real numbers. We write

$$\limsup_{k\to\infty} x_k \coloneqq \alpha = \lim_{l\to\infty} \sup_{k\geq l} x_k,$$
$$\liminf_{k\to\infty} x_k \coloneqq \beta = \lim_{l\to\infty} \inf_{k\geq l} x_k,$$

and call them the *limit superior* and *limit inferior* of $\{x_k\}$, respectively.

2.2 Topological properties

Let $F \subset \mathbb{R}^N$ be a set. F is *closed* if for any convergent sequence $\{x_k\}_{k=1}^{\infty}$ in F (meaning that $x_k \in F$ for all k and $x_k \to x$ for some $x \in \mathbb{R}^N$), the limit point belongs to F (meaning that $x \in F$).[1] Intuitively, a closed set is one that includes its own boundary. Thus the set $[0,1] = \{x \,|\, 0 \le x \le 1\}$ is closed but $(0,1) = \{x \,|\, 0 < x < 1\}$ is not.

Let $U \subset \mathbb{R}^N$ be a set. The *complement* of U, denoted by U^c, is defined by $U^c = \left\{x \in \mathbb{R}^N \,\middle|\, x \notin U\right\}$. That is, the complement of a set consists of those points that do not belong to the original set. U is said to be *open* if U^c is closed.[2] Thus $(0,1) = \{x \,|\, 0 < x < 1\}$ is open, because its complement $(0,1)^c = (-\infty, 0] \cup [1, \infty)$ is closed. Intuitively, an open set is one that does not include its own boundary.

A set $K \subset \mathbb{R}^N$ is said to be *compact*[3] if any sequence in K has a convergent subsequence with a limit in K.[4] That is, K is compact if for any sequence $\{x_k\}_{k=1}^{\infty} \subset K$, we can find a subsequence $\{x_{k_l}\}_{l=1}^{\infty}$ and a point $x \in K$ such that $x_{k_l} \to x$ as $l \to \infty$. Compact sets are important because as we see in Theorem 2.5 below, any continuous function attains a maximum and a minimum on a compact set. The following Heine-Borel theorem characterizes compact sets in $\mathbb{R}^N$.

Theorem 2.4 (Heine-Borel)**.** *$K \subset \mathbb{R}^N$ is compact if and only if it is closed and bounded.*

Proof. Suppose that K is compact. Take any convergent sequence $\{x_k\} \subset K$ with $\lim x_k = x$. Since K is compact, we can take a subsequence $\{x_{k_l}\}$ such that $x_{k_l} \to y$ for some $y \in K$. But by Proposition 2.3 we get $x = y \in K$, so K is closed. To prove that K is bounded, suppose that it is not. Then for any k we can find $x_k \in K$ such that $\|x_k\| > k$. For any subsequence $\{x_{k_l}\}$, since $\|x_{k_l}\| > k_l \to \infty$ as $l \to \infty$, $\{x_{k_l}\}$ is not bounded. Hence by Proposition 2.2 $\{x_{k_l}\}$ is not convergent. Since $\{x_k\}$ has no convergent subsequence, K is not compact, which is a contradiction. Hence K is bounded.

Conversely, suppose that K is closed and bounded. Let us show by induction on the dimension N that any bounded sequence in $\mathbb{R}^N$ has a convergent subsequence. By the remark after Theorem 2.1, we may use the sup norm $\|x\| = \max_n |x_n|$ instead of the Euclidean norm to define convergence. For $N = 1$, let $\{x_k\}_{k=1}^{\infty} \subset [-b, b]$ be a bounded sequence, where $b > 0$. Define $\alpha_l = \sup_{k \ge l} x_k$. Since $x_k \in [-b, b]$, it follows that $\alpha_l \in [-b, b]$. Clearly $\{\alpha_l\}$ is a decreasing sequence, so it has a limit $\alpha \in [-b, b]$. For each l, choose $k_l \ge l$ such that $|x_{k_l} - \alpha_l| < 1/l$, which is possible by the definition of α_l. Then

$$|x_{k_l} - \alpha| \le |x_{k_l} - \alpha_l| + |\alpha_l - \alpha| < \frac{1}{l} + |\alpha_l - \alpha| \to 0$$

as $l \to \infty$, so $x_{k_l} \to \alpha$. Therefore $\{x_k\}_{k=1}^{\infty}$ has a convergent subsequence $\{x_{k_l}\}_{l=1}^{\infty}$.

[1] The letter F is often used for a closed set since the French word for "closed" is *fermé*.

[2] The letters U, V are often used for an open set since the French word for "open" is *ouvert* but the letter O is confusing due to the resemblance to 0.

[3] The letter K is often used for a compact set since the German word for "compact" is *kompakt*.

[4] Strictly speaking, this is the definition of a *sequentially compact* set, but in $\mathbb{R}^N$ (or more generally metric spaces) the two concepts are equivalent.

Suppose that the claim is true up to dimension $N-1$. Let $\{x_k\}_{k=1}^{\infty} \in [-b,b]^N$ be a bounded sequence, where $x_k = (x_{1k}, \dots, x_{Nk})$. Since $\{x_{1k}\}_{k=1}^{\infty} \subset [-b,b]$, it has a convergent subsequence $\{x_{1k'}\}$. By the induction hypothesis, the sequence of $(N-1)$-vectors $\{(x_{2k'}, \dots, x_{Nk'})\} \subset [-b,b]^{N-1}$ has a convergent subsequence $\{(x_{2k''}, \dots, x_{Nk''})\}$. Since $\{x_{1k''}\}$ is a subsequence of $\{x_{1k'}\}$, by Proposition 2.3 it is also convergent. Therefore $\{x_{k''}\} = \{(x_{1k''}, \dots, x_{Nk''})\} \subset [-b,b]^N$ also converges, so $\{x_k\}_{k=1}^{\infty}$ has a convergent subsequence.

We have shown that any $\{x_k\} \subset K \subset [-b,b]^N$ has a convergent subsequence $\{x_{k_l}\}$. Since K is closed, the limit belongs to K. Therefore K is compact. □

2.3 Continuous functions

Let $U \subset \mathbb{R}^N$ be a set and $f : U \to \mathbb{R}$ be a function. We say that f is *continuous at* $x \in U$ if $f(x_k) \to f(x)$ for any sequence $\{x_k\} \subset U$ such that $x_k \to x$. We say that f *continuous* on U if it is continuous at every point of U. Intuitively, f is continuous if its graph has no gaps. The following theorem is important because it gives a sufficient condition for the existence of a solution to an optimization problem.

Theorem 2.5 (Extreme Value Theorem)**.** *If $K \subset \mathbb{R}^N$ is nonempty and compact and $f : K \to \mathbb{R}$ is continuous, then $f(K)$ is compact. In particular, f attains its maximum and minimum over K.*

Proof. Let $\{y_k\} \subset f(K)$. Then we can take $\{x_k\} \subset K$ such that $y_k = f(x_k)$ for all k. Since K is compact, we can take a subsequence $\{x_{k_l}\}_{l=1}^{\infty}$ such that $x_{k_l} \to x \in K$. Then by the continuity of f, we have $y_{k_l} = f(x_{k_l}) \to f(x) \in f(K)$, so $f(K)$ is compact.

Since $f(K)$ is compact, by Theorem 2.4 it is bounded. Hence $M \coloneqq \sup f(K) < \infty$. Repeating the above argument with $\{y_k\}$ such that $y_k \to M$, it follows that $M = \lim y_{k_l} = \lim f(x_{k_l}) = f(x)$, so f attains its maximum. The case for the minimum is similar. □

With applications in mind, it is useful to allow some discontinuous functions and functions that take values $\pm\infty$. We say that $f : \mathbb{R}^N \to [-\infty, \infty]$ is *lower semi-continuous* at x if for any $x_k \to x$ we have $f(x) \le \liminf_{k\to\infty} f(x_k)$. We say that f is *upper semi-continuous* if $f(x) \ge \limsup_{k\to\infty} f(x_k)$. Clearly f is upper semi-continuous if and only if $-f$ is lower semi-continuous.

The following theorem generalizes Theorem 2.5.

Theorem 2.6 (Extreme Value Theorem for semi-continuous functions)**.** *Let K be compact and $f : K \to [-\infty, \infty]$ be lower (upper) semi-continuous. Then f attains a minimum (maximum) over K.*

Proof. We show only for the case f is lower semi-continuous. If $f(x) = -\infty$ for some $x \in K$ or $f(x) = \infty$ for all $x \in K$, there is nothing to prove. Hence assume that $f(x) > -\infty$ for all $x \in K$ and $f(x) < \infty$ for some $x \in K$. Let $m = \inf_{x\in K} f(x)$. Take a sequence $\{x_k\} \subset K$ such that $f(x_k) \to m$. Since K is compact, there is a subsequence such that $x_{k_l} \to x$ for some $x \in K$. Since f is lower semi-continuous, we get

$$m \le f(x) \le \liminf_{l\to\infty} f(x_{k_l}) = m,$$

so $-\infty < f(x) = m$. □

Problems

2.1. 1. Let $\{F_i\}_{i\in I} \subset \mathbb{R}^N$ be a collection of closed sets. Prove that $\bigcap_{i\in I} F_i$ is closed.

2. Let $A \subset \mathbb{R}^N$ be any set. Prove that there exists a smallest closed set that includes A. (We denote this set by $\operatorname{cl} A$ and call it the *closure* of A.)

3. Prove that there exists a largest open subset of A. (We denote this set by $\operatorname{int} A$ and call it the *interior* of A.)

2.2. Let $B(x,\epsilon) = \left\{y \in \mathbb{R}^N \,\middle|\, \|y-x\| < \epsilon\right\}$ be the *open ball* with center x and radius ϵ.

1. Prove that U is open if and only if for any $x \in U$, there exists $\epsilon > 0$ such that $B(x,\epsilon) \subset U$.

2. Prove that if U_1, U_2 are open, so is $U_1 \cap U_2$. Prove that if F_1, F_2 are closed, so is $F_1 \cup F_2$.

3. Let A, B be any set. Prove that $\operatorname{int}(A\cap B) = \operatorname{int} A \cap \operatorname{int} B$ and $\operatorname{cl}(A\cup B) = \operatorname{cl} A \cup \operatorname{cl} B$.

2.3. In the proof of the Heine-Borel theorem (Theorem 2.4), we used the fact that any bounded set $A \subset \mathbb{R}$ has a supremum $\alpha = \sup A$, known as the *axiom of continuity*. Using this axiom, prove that any bounded monotone sequence is convergent (which we also used in the proof).

2.4. A sequence $\{x_k\}_{k=1}^{\infty} \subset \mathbb{R}^N$ is said to be *Cauchy* if

$$\forall \epsilon > 0, \exists K > 0, k, l > K \Longrightarrow \|x_k - x_l\| < \epsilon,$$

that is, the terms with sufficiently large indices are arbitrarily close to each other.

1. Prove that a Cauchy sequence is bounded.

2. Prove that a Cauchy sequence converges. (Hint: Heine-Borel theorem. This property is called the *completeness* of $\mathbb{R}^N$.)

2.5. This problem asks you to prove Theorem 2.1.

1. For any norm $\|\cdot\|$ on $\mathbb{R}^N$, define $f : \mathbb{R}^N \to \mathbb{R}$ by $f(x) = \|x\|$. Show that f is continuous, where we define convergence of sequences using the sup norm $\|\cdot\|_\infty$. (Hint: express $x = (x_1, \ldots, x_N)'$ as $x = \sum_{n=1}^N x_n e_n$, where e_n is a unit vector, and use the triangle inequality.)

2. Define the set $K = \left\{x \in \mathbb{R}^N \,\middle|\, \|x\|_\infty = 1\right\}$. Show that K is nonempty and compact.

3. Define $g : K \to \mathbb{R}$ by $g(x) = f(x)/\|x\|_\infty$. Show that g is continuous.

4. Show that there exist constants $0 < c \le C$ such that

$$c\|x\|_\infty \le \|x\| \le C\|x\|_\infty$$

for all $x \in \mathbb{R}^N$.

5. Prove Theorem 2.1.

2.6. Let $f : (a, b) \to \mathbb{R}$ be increasing, so $a < x_1 \leq x_2 < b$ implies $f(x_1) \leq f(x_2)$.

1. Show that for each $x \in (a, b)$,

$$g^{\pm}(x) \coloneqq \lim_{h \to \pm 0} f(x + h)$$

exist, and that $g^-(x) \leq g^+(x)$ for all $x \in (a, b)$.

2. Show that f is continuous on (a, b) except at at most countably many points. (Hint: if $g^-(x) < g^+(x)$, then there is a rational number in between.)

Chapter 3

One-Variable Optimization

3.1 A motivating example

Suppose you are the owner of a firm that produces a good. It costs $c \geq 0$ dollars per unit of good produced. If you produce more, in general you will need to charge a lower price in order to sell everything, so assume the price at which you can sell x units of good is $p(x) = a - bx$, with $a, b > 0$. Then what is the optimal production plan? This is the type of problems you will learn to solve in this course.

This particular problem can be solved using only the mathematical knowledge up to high school. If you produce x, by assumption the *revenue* is price times quantity, so

$$p(x)x = (a - bx)x.$$

Also, the cost is cx. Therefore the *profit* is

$$f(x) = p(x)x - cx = (a - bx)x - cx = -bx^2 + (a - c)x.$$

One way to maximize this *objective function* is to complete the squares:

$$f(x) = -bx^2 + (a - c)x = -b\left(x - \frac{a - c}{2b}\right)^2 + \frac{(a - c)^2}{4b}.$$

Since the first term is nonpositive and the second term does not depend on x, assuming $a > c$ the optimal production level is $\bar{x} = \frac{a-c}{2b}$ (which makes the first term exactly zero) and the maximum profit is $f(\bar{x}) = \frac{(a-c)^2}{4b}$.

This profit maximization problem has a few typical features. First, the objective function

$$f(x) = -bx^2 + (a - c)x$$

is a *nonlinear* function of the variable x. (In this case, it is a *quadratic* function.) Second, since you cannot produce a negative amount of good, implicitly there is the *constraint* $x \geq 0$.

In this course we will learn how to solve these types of problems—*nonlinear constrained optimization problems*. Since the objective function is nonlinear, the technique of linear programming does not apply. We will go step-by-step. First we will consider unconstrained optimization problems with a single variable and proceed to constrained optimization problems with a single or multiple variables.

3.2 One-variable calculus

A powerful tool for solving nonlinear optimization problems is *calculus*. In the motivating example above, we were able to solve the problem without using calculus because the objective function was quadratic and we could complete the squares. Such a trick does not apply in general.

3.2.1 Differentiation

Differentiation (taking derivatives) is basically approximating a nonlinear function by a linear one. Suppose we want to approximate a function $f(x)$ by a linear function around the point $x = a$, so

$$f(x) \approx p(x - a) + q$$

for some numbers p, q. The approximation should be exact at $x = a$, so substituting $x = a$ we must have $q = f(a)$. Subtracting q and dividing by $x - a$ (when $x \neq a$), we get

$$p \approx \frac{f(x) - f(a)}{x - a}.$$

Since the approximation is for x close to a, it makes sense to define p by the limit of $\frac{f(x)-f(a)}{x-a}$ as x approaches to a (we write this as $x \to a$).

$$p = f'(a) := \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

is called the *derivative* of $f(x)$ at $x = a$. Letting $x = a + h$ with $h \neq 0$, we can also write

$$f'(a) = \lim_{h \to 0} \frac{f(a + h) - f(a)}{h}.$$

Example 3.1. Let $f(x) = x$. Then

$$f'(a) = \lim_{h \to 0} \frac{(a + h) - a}{h} = \lim_{h \to 0} \frac{h}{h} = \lim_{h \to 0} 1 = 1.$$

Example 3.2. Let $f(x) = x^2$. Then

$$f'(a) = \lim_{h \to 0} \frac{(a + h)^2 - a^2}{h} = \lim_{h \to 0} \frac{2ah + h^2}{h} = \lim_{h \to 0} (2a + h) = 2a.$$

If the derivative of f exists at every point, then the function f is called *differentiable*. The derivative of f at x is denoted by $f'(x)$. The derivative $f'(x)$ is itself another function. If f' is continuous, then f is called *continuously differentiable*, or simply a C^1 function. If $f'(x)$ is again differentiable, then its derivative is denoted by $f''(x)$ and is called the second derivative of f. You can define $f'''(x)$, $f''''(x)$, etc. in the same way. The n-th derivative of f is usually denoted by $f^{(n)}(x)$. If f is n times differentiable and $f^{(n)}(x)$ is continuous ("n times continuously differentiable"), then f is called a C^n function.

The following proposition shows why calculus is a powerful tool.

Proposition 3.1. *Consider the optimization problem*

$$\text{maximize } f(x),$$

where f is differentiable. If $\bar{x}$ is a solution, then $f'(\bar{x}) = 0$.

Proof. Take any $h > 0$. Since $\bar{x}$ attains the maximum of f, we have

$$f(\bar{x} + h) \leq f(\bar{x}).$$

Subtracting $f(\bar{x})$ and dividing by $h > 0$, we get

$$\frac{f(\bar{x} + h) - f(\bar{x})}{h} \leq 0.$$

Letting $h \to 0$ and using the definition of the derivative, we get $f'(\bar{x}) \leq 0$. By considering the case $h < 0$, we can show $f'(\bar{x}) \geq 0$. Therefore $f'(\bar{x}) = 0$. □

Proposition 3.1 says that in order to maximize a differentiable function (with no constraints), it is *necessary* that the derivative is zero.

Example 3.3. Consider the motivating example above. The objective function is $f(x) = -bx^2 + (a - c)x$. The derivative is

$$f'(x) = -2bx + a - c.$$

By Proposition 3.1, the solution $\bar{x}$ must satisfy

$$f'(\bar{x}) = -2b\bar{x} + a - c = 0 \iff \bar{x} = \frac{a - c}{2b}.$$

However, setting the derivative to zero is not *sufficient* in general, as the following example shows.

Example 3.4. Let $f(x) = x^3 - 12x$. Then

$$f'(x) = 3x^2 - 12 = 3(x - 2)(x + 2),$$

so $f'(x) = 0 \iff x = \pm 2$. Now $f(\pm 2) = \mp 16$. But $f(\pm 5) = \mp 65$, so $x = \pm 2$ are neither the minimum nor the maximum of f.

3.2.2 Mean value theorem and Taylor's theorem

Let f be a differentiable function. By definition, $f'(a)$ is the limit of $\frac{f(b)-f(a)}{b-a}$—the slope between the points $(a, f(a))$ and $(b, f(b))$—as b approaches a. Is there an exact relationship between f' and arbitrary b? The mean value theorem gives an answer.

Proposition 3.2 (Mean value theorem)**.** *Let f be continuous on $[a, b]$ and differentiable on (a, b). Then there exists $c \in (a, b)$ such that*

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

Proof. Let

$$\phi(x) = f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a).$$

By direct substitution, we can show $\phi(a) = \phi(b) = 0$. If $\phi \equiv 0$ on $[a, b]$, then

$$0 = \phi'(x) = f'(x) - \frac{f(b) - f(a)}{b - a}$$

on (a, b), so we can take any $c \in (a, b)$. Suppose there exists $x \in [a, b]$ such that $\phi(x) > 0$. Since ϕ is continuous, by the extreme value theorem it attains the maximum at some point $c \in [a, b]$. Since $\phi(a) = \phi(b) = 0$ and ϕ takes a positive value, it must be $c \in (a, b)$. By Proposition 3.1, we have

$$0 = \phi'(c) = f'(c) - \frac{f(b) - f(a)}{b - a} \iff \frac{f(b) - f(a)}{b - a} = f'(c).$$

The proof if ϕ takes a negative value is similar. □

Remember that differentiation is basically a linear approximation:

$$f(x) \approx f(a) + f'(a)(x - a).$$

Changing the notation in the mean value theorem such that $b = x$ and $c = \xi$, we obtain

$$f'(\xi) = \frac{f(x) - f(a)}{x - a} \iff f(x) = f(a) + f'(\xi)(x - a).$$

There is no reason to stop at a linear (first-order) approximation. If, for example, we continue to a quadratic (second-order) approximation, we can show that for each x, there exists a number ξ between a and x such that

$$f(x) = f(a) + f'(a)(x - a) + \frac{1}{2} f''(\xi)(x - a)^2.$$

In general, by increasing the order of the polynomial approximation, we can prove the following Taylor's theorem (proof in Problem 3.7):

Proposition 3.3 (Taylor's theorem)**.** *Let f be n times differentiable. Then for each x, there exists a number ξ between a and x such that*

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{1}{(n-1)!} f^{(n-1)}(a)(x-a)^{n-1} + \frac{1}{n!} f^{(n)}(\xi)(x-a)^n.$$

Here $n! = n \times (n-1) \times \cdots \times 2 \times 1$ is the n factorial.

The proof is in Problem 3.7.

3.3 Convex functions

Proposition 3.1 tells us that if a function is differentiable, the derivative is zero at the optimum (maximum or minimum). Therefore, setting the derivative to zero (first-order condition) is a *necessary* condition for optimality. Is there a *sufficient* condition for optimality? The answer is yes: there is a special but large enough class of functions such that the first-order condition is also sufficient.

A function f is said to be *convex* if for any x_1, x_2 and $0 \le \alpha \le 1$ we have

$$f((1 - \alpha)x_1 + \alpha x_2) \le (1 - \alpha) f(x_1) + \alpha f(x_2).$$

Graphically, a function is convex if the segment joining the points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ lies above the graph of f (Figure 3.1). f is *strictly convex* if the inequality is strict for $0 < \alpha < 1$. f is *concave* if $-f$ is convex.

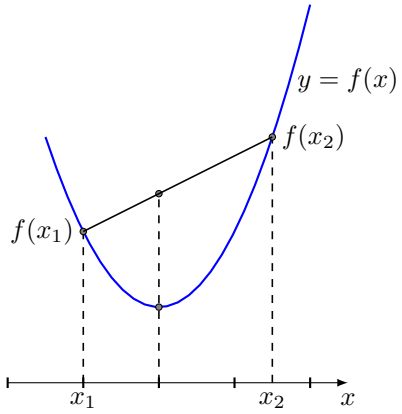

Figure 3.1. Convex function.

As shown in Problems 3.10 and 3.11, a twice continuously differentiable function f is convex if and only if the second derivative is nonnegative, so $f''(x) \geq 0$. The intuitive explanation is as follows. When $f''(x) \geq 0$, then $f'(x)$—the derivative or the slope of f—is increasing. Therefore if you imagine flying along the graph of f, you will be constantly turning upwards. Therefore the segment that joins arbitrary two points on the trajectory must line above the actual trajectory.

The following proposition shows that setting the derivative to zero is sufficient for optimization when the objective function is convex or concave.

Proposition 3.4. *Let f be twice differentiable and convex (concave). If $f'(\bar{x}) = 0$, then $\bar{x}$ is the minimum (maximum) of f.*

Proof. Suppose that f is convex, so $f''(x) \geq 0$. Applying Taylor's theorem (Proposition 3.3) for $a = \bar{x}$ and $n = 2$, for any x there exists ξ such that

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + \frac{1}{2} f''(\xi)(x - \bar{x})^2.$$

Since by assumption $f'(\bar{x}) = 0$ and $f''(\xi) \geq 0$, we obtain $f(x) \geq f(\bar{x})$. Therefore $\bar{x}$ is the minimum of f. A similar argument holds when f is concave. □

Example 3.5. Consider the motivating example above. The objective function is $f(x) = -bx^2 + (a - c)x$ and the first derivative is $f'(x) = -2bx + a - c$. Since the second derivative is

$$f''(x) = -2b < 0,$$

f is concave. Therefore

$$f'(\bar{x}) = -2b\bar{x} + a - c = 0 \iff \bar{x} = \frac{a - c}{2b}$$

is the maximum of f.

Problems

3.1. Using the definition, compute the derivative of the following functions.

1. $f(x) = x^3$.
2. $f(x) = x^4$.
3. $f(x) = x^n$, where n is a natural number. (Hint: binomial theorem.)
4. $f(x) = 1/x$.
5. $f(x) = \sqrt{x}$.

3.2. Let f, g be differentiable and α be a real number. Show that

1. $(f(x) + g(x))' = f'(x) + g'(x)$,
2. $(\alpha f(x))' = \alpha f'(x)$,
3. $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$,
4. $(g(f(x)))' = g'(f(x))f'(x)$ (chain rule).

3.3. The exponential function is defined by

$$\mathrm{e}^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \cdots = \sum_{n=0}^{\infty} \frac{1}{n!}x^n,$$

where $\mathrm{e} = 2.718281828\ldots$. It satisfies $\mathrm{e}^{x+y} = \mathrm{e}^x\mathrm{e}^y$ and $(\mathrm{e}^x)' = \mathrm{e}^x$. The logarithmic function is the inverse function of the exponential, so $\mathrm{e}^{\log x} = x$ and $\log \mathrm{e}^x = x$. Using the chain rule, show that

1. $(\log x)' = 1/x$,
2. $(x^\alpha)' = \alpha x^{\alpha-1}$.

3.4. Define $f(x)$ by $f(x) = x^2 \sin(1/x)$ if $x \neq 0$ and $f(0) = 0$.

1. Compute $f'(x)$ when $x \neq 0$.
2. Using the definition, compute $f'(0)$.
3. Show that f is differentiable but not continuously differentiable.

3.5.
1. Fill in the details of the proof of Proposition 3.1.
2. Show that Proposition 3.1 also holds for minimization.

3.6.
1. Let $f : (a, b) \to \mathbb{R}$ be differentiable and $f' > 0$. Show that f is strictly increasing, i.e., $x_1 < x_2$ implies $f(x_1) < f(x_2)$.
2. Let $f : [a, b] \to \mathbb{R}$ be continuous and differentiable on (a, b). Let $\bar{x} \in [a, b]$ be a maximum of f, which exists by Theorem 2.5. If $f'(x) > 0$ for x sufficiently close to a, show that $\bar{x} \neq a$.

3.7. This problem asks you to prove Taylor's theorem. Let $a \neq b$ and $f : [a, b] \to \mathbb{R}$. Suppose that f is n times differentiable and $f^{(k)}(x)$ is continuous on $[a, b]$ for $k = 1, \ldots, n-1$.

1. Define the polynomial $P(x) = \sum_{k=0}^{n-1} \frac{f^{(k)}(a)}{k!}(x-a)^k$, let $M = \frac{f(b)-P(b)}{(b-a)^n}$, and
$$\phi(x) = f(x) - P(x) - M(x-a)^n.$$
Show that $\phi(a) = \phi'(a) = \cdots = \phi^{(n-1)}(a) = 0$ and $\phi^{(n)}(x) = f^{(n)}(x) - n!M$.

2. Prove Taylor's theorem.

3.8. For each of the following functions, show whether it is convex, concave, or neither.

1. $f(x) = 10x - x^2$,
2. $f(x) = x^4 + 6x^2 + 12x$,
3. $f(x) = 2x^3 - 3x^2$,
4. $f(x) = x^4 + x^2$,
5. $f(x) = x^3 + x^4$,
6. $f(x) = \mathrm{e}^x$,
7. $f(x) = \log x$ $(x > 0)$,
8. $f(x) = x \log x$ $(x > 0)$,
9. $f(x) = x^\alpha$, where $\alpha \neq 0$ and $x > 0$. (Hint: there are a few cases to consider.)

3.9. Suppose that you are running a firm that produces an output good using an input good. When the input is x, the output is $2\sqrt{x}$. Suppose that the price of the input is c and the price of the output is p. Compute the input level that maximizes the profit.

The following two exercises ask you to show that a twice differentiable function is convex if and only if the second derivative is nonnegative.

3.10. Let f be differentiable.

1. Fix $x \neq y$ and let $g(t) = \frac{f((1-t)x+ty)-f(x)}{t}$, where $t > 0$. For $0 < s < t$, show that
$$g(s) \leq g(t) \iff f((1-s)x + sy) \leq \left(1 - \frac{s}{t}\right) f(x) + \frac{s}{t} f((1-t)x + ty).$$

2. Show that the function g is increasing if and only if f is convex.

3. Compute $g(1)$ and $\lim_{t\to 0} g(t)$.

4. Show that f is convex if and only if
$$f(y) - f(x) \geq f'(x)(y-x)$$
for all x, y.

3.11. Using Taylor's theorem and the previous exercise, show that a twice continuously differentiable function f is convex if and only if $f''(x) \geq 0$ for all x.

3.12. Prove Proposition 3.4 assuming only that f is differentiable (but not necessarily twice differentiable). (Hint: Problem 3.10.)

3.13. Let f be strictly convex. If f has a minimum, show that it is unique. (Hint: assume there are two minima x_1, x_2 and derive a contradiction using the definition of convexity.)

Chapter 4

Multi-Variable Calculus

4.1 A motivating example

Suppose you are the owner of a firm that produces two goods. The unit price of good 1 and 2 are p_1 and p_2, respectively. To produce x_1 units of good 1 and x_2 units of good 2, it costs

$$c(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2).$$

What is the optimal production plan?

This problem can be solved using only high school algebra. If you produce (x_1, x_2), the profit is

$$f(x_1, x_2) = p_1 x_1 + p_2 x_2 - c(x_1, x_2) = p_1 x_1 + p_2 x_2 - \frac{1}{2}(x_1^2 + x_2^2).$$

Since f is a quadratic function, you can complete the squares:

$$f(x_1, x_2) = -\frac{1}{2}(x_1 - p_1)^2 - \frac{1}{2}(x_2 - p_2)^2 + \frac{1}{2}(p_1^2 + p_2^2),$$

so the optimal plan is $(x_1, x_2) = (p_1, p_2)$, with maximum profit $\frac{1}{2}(p_1^2 + p_2^2)$.

Many practical problems are optimization problems that involve two or more variables, as in this example. In the previous chapter, we saw that calculus is a powerful tool for solving one-variable optimization problems. The same is true for the multi-variable case. This chapter introduces the basics of multi-variable calculus.

4.2 Differentiation

Consider a function of two variables, $f(x_1, x_2)$. In Chapter 3, we motivated differentiation by a linear approximation. The same is true for functions of two or more variables. Suppose we want to approximate $f(x_1, x_2)$ by a linear function around the point $(x_1, x_2) = (a_1, a_2)$, so

$$f(x_1, x_2) \approx p_1(x_1 - a_1) + p_2(x_2 - a_2) + q \tag{4.1}$$

for some numbers p_1, p_2, q. The approximation should be exact at $(x_1, x_2) = (a_1, a_2)$, so substituting $(x_1, x_2) = (a_1, a_2)$ we must have $q = f(a_1, a_2)$. The values of p_1, p_2 should be such that as (x_1, x_2) approaches (a_1, a_2), the approximation should get better and better. Therefore subtracting $f(a_1, a_2)$ from both sides of (4.1) and letting $x_2 = a_2$ and $x_1 \to a_1$, it must be

$$p_1 = \frac{\partial f}{\partial x_1}(a_1, a_2) \coloneqq \lim_{x_1 \to a_1} \frac{f(x_1, a_2) - f(a_1, a_2)}{x_1 - a_1}.$$

This quantity is called the *partial derivative* of f with respect to x_1 (evaluated at (a_1, a_2)). A partial derivative, as the name suggests, is just a derivative of a function with respect to one variable, fixing all other variables. Intuitively, the partial derivative is the rate of change (slope) of the function in the direction of one particular coordinate. By a similar argument, we obtain

$$p_2 = \frac{\partial f}{\partial x_2}(a_1, a_2) \coloneqq \lim_{x_2 \to a_2} \frac{f(a_1, x_2) - f(a_1, a_2)}{x_2 - a_2},$$

the partial derivative of f with respect to x_2.

If you know how to take the derivative of a one-variable function, computing partial derivatives of a multi-variable function is straightforward: you just pretend that all variables except one are constants.

Example 4.1. Let $f(x_1, x_2) = x_1 + 2x_2 + 3x_1^2 + 4x_1x_2 + 5x_2^2$. Then

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = 1 + 6x_1 + 4x_2,$$
$$\frac{\partial f}{\partial x_2}(x_1, x_2) = 2 + 4x_1 + 10x_2.$$

A function is said to be *partially differentiable* if the partial derivatives exist. If a function is partially differentiable and the partial derivatives are continuous, we call it a C^1 function. In general, a C^r function means that you can partially differentiate r times (with an arbitrary choice of variables) and the resulting function is continuous. A function is said to be *differentiable* if the linear approximation (4.1) becomes exact as the point (x_1, x_2) gets closer to (a_1, a_2), so formally

$$f(a_1 + h_1, a_2 + h_2) - f(a_1, a_2) = p_1h_1 + p_2h_2 + \epsilon(h_1, h_2) \tag{4.2}$$

with $\epsilon(h_1, h_2)/\sqrt{h_1^2 + h_2^2} \to 0$ as $(h_1, h_2) \to (0, 0)$, where p_1, p_2 are partial derivatives. It is known that a C^1 function is differentiable (Problem 4.4).

In the one-variable case, the derivative of a function at a minimum or a maximum is zero. The same is true for partial derivatives of a multi-variable function. We omit the proof because it is essentially the same as the one-variable case.

Proposition 4.1. *Consider the optimization problem*

$$\text{maximize } f(x_1, x_2),$$

where f is partially differentiable. If $(\bar{x}_1, \bar{x}_2)$ is a solution, then $\frac{\partial f}{\partial x_1}(\bar{x}_1, \bar{x}_2) = \frac{\partial f}{\partial x_2}(\bar{x}_1, \bar{x}_2) = 0$.

Example 4.2. Consider the motivating example. Then

$$\frac{\partial f}{\partial x_1} = p_1 - x_1,$$
$$\frac{\partial f}{\partial x_2} = p_2 - x_2,$$

so $\partial f/\partial x_1 = \partial f/\partial x_2 = 0$ implies $(x_1, x_2) = (p_1, p_2)$, which maximizes the profit.

4.3 Vector notation and gradient

Equation (4.2) shows that the difference in f is approximately a linear function of the differences in the coordinates, $p_1 h_1 + p_2 h_2$. Define the vectors a, p, h by $a = \left[\begin{smallmatrix} a_1 \\ a_2 \end{smallmatrix}\right]$, $p = \left[\begin{smallmatrix} p_1 \\ p_2 \end{smallmatrix}\right]$, and $h = \left[\begin{smallmatrix} h_1 \\ h_2 \end{smallmatrix}\right]$. If you remember the definition of the inner product,[1] (4.2) can be compactly written as

$$f(a + h) - f(a) = p \cdot h + \epsilon(h)$$

with $\epsilon(h)/\|h\| \to 0$ as $h \to 0$, where $\|h\| = \sqrt{h \cdot h} = \sqrt{h_1^2 + h_2^2}$ is the (Euclidean) norm of the vector h. The vector of partial derivatives,

$$\nabla f(a) := \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x_1}(a_1, a_2) \\ \frac{\partial f}{\partial x_2}(a_1, a_2) \end{bmatrix},$$

is called the *gradient* of f at (a_1, a_2). (You read the symbol ∇ "nabla".) The above equation then becomes

$$f(a + h) - f(a) = \nabla f(a) \cdot h + \epsilon(h).$$

Example 4.3. Let $f(x_1, x_2) = x_1^2 x_2^3$. Then

$$\nabla f(x_1, x_2) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 x_2^3 \\ 3x_1^2 x_2^2 \end{bmatrix}.$$

Using the gradient, Proposition 4.1 simplifies as follows: if $\bar{x}$ is a solution of the optimization problem

$$\text{maximize } f(x),$$

where f is partially differentiable, then $\nabla f(\bar{x}) = 0$.[2] The same is true for minimization.

My experience tells that when students are first introduced to the vector notation, they are overwhelmed by the "abstract" nature. It is true that imagining a vector requires more mental effort than imagining a real number. However, the vector notation has two important advantages over the component-wise notation. First, since you don't need to write down all the components, it saves space and you can focus on the substantive content. Second, since the vector notation applies to any dimension $(1, 2, \dots)$, you can develop a single theory that applies to all cases. For these reasons, you should get used to the vector

[1] The inner product is also called the dot product, although inner product is more common.
[2] We use the letter 0 to denote the zero vector $\left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$.

notation. Whenever you think it is too abstract, consider the two-dimensional case for concreteness.

Intuitively, the gradient $\nabla f(a)$ is the direction at which the function increases fastest at the point a. To see this, take any vector d and evaluate the value of f along the straight line $x = a + td$ that passes through the point a and points to the direction d, where t is a parameter. The value is then $f(a + td)$. The slope of f along this line is

$$\lim_{t \to 0} \frac{f(a + td) - f(a)}{t} = \nabla f(a) \cdot d,$$

which can be shown using the chain rule (Proposition 4.3). This quantity is known as the *directional derivative* of f (with direction d). In particular, if d is a unit vector (say $d = \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ or $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ in the two-dimensional case), then the directional derivative is a partial derivative. Assuming d has length 1 (so $\|d\| = 1$) and applying the Cauchy-Schwarz inequality $x \cdot y \leq \|x\| \, \|y\|$ to $x = \nabla f(a)$ and $y = d$, it follows that

$$\nabla f(a) \cdot d \leq \|\nabla f(a)\| \, \|d\| = \|\nabla f(a)\| \,, \tag{4.3}$$

with equality if d is parallel to $\nabla f(a)$, so $d = \nabla f(a) / \|\nabla f(a)\|$. This inequality shows that the directional derivative (the rate of change of the function) is maximum when the direction is that of the gradient. In other words, an interpretation of the gradient $\nabla f(a)$ is the direction of steepest ascent of f at $x = a$. Similarly, $-\nabla f(a)$ is the direction of steepest descent of f at $x = a$.

Example 4.4. Let $f(x_1, x_2) = \sqrt{x_1^2 + x_2^2}$. Letting $x_1 = r \cos\theta$ and $x_2 = r \sin\theta$, we have $f(x_1, x_2) = r$, so f is constant along a circle and increases away from the circle. Therefore at any point $x = (x_1, x_2)$, f increases fastest along the radius joining the origin and x. In fact, the gradient is

$$\nabla f(x_1, x_2) = \begin{bmatrix} \frac{x_1}{\sqrt{x_1^2 + x_2^2}} \\ \frac{x_2}{\sqrt{x_1^2 + x_2^2}} \end{bmatrix} = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix},$$

which points to the direction of the radius.

4.4 Mean value theorem and Taylor's theorem

In one-variable optimization problems, the mean value theorem (Proposition 3.2) and Taylor's theorem (Proposition 3.3) are useful to characterize the solution. The same is true with multiple variables.

Proposition 4.2 (Mean value theorem)**.** *Let f be differentiable. For any vectors a, b, there exists a number $0 < \theta < 1$ such that*

$$f(b) - f(a) = \langle \nabla f((1 - \theta)a + \theta b), b - a \rangle \,.$$

Here $\langle x, y \rangle = x \cdot y = x_1 y_1 + \cdots + x_N y_N$ is another notation for the inner product. The proof is in Problem 4.3. The mean value theorem for the one-variable case says that there exists a number c between a and b such that

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

Multiplying both sides by $b-a$ and choosing $0 < \theta < 1$ such that $c = (1-\theta)a+\theta b$ (which is possible because c is between a and b), we get

$$f(b) - f(a) = f'((1-\theta)a + \theta b)(b-a).$$

Therefore the multi-variable version of the mean value theorem is a generalization of the one-variable case.

Taylor's theorem also generalizes to the multi-variable case. Suppose you want to approximate $f(x)$ around $x = a$. Let $h = x - a$, and consider the one-variable function $g(t) = f(a+th)$. Then $g(0) = f(a)$ and $g(1) = f(x)$. Now apply Taylor's theorem to the one-variable function $g(t)$ and set $t = 1$. The result is Taylor's theorem for the multi-variable function $f(x)$. The multi-variable version of Taylor's theorem is most useful in the second-order approximation. The result is

$$f(x) = f(a) + \langle \nabla f(a), x-a \rangle + \frac{1}{2} \left\langle x-a, \nabla^2 f(\xi)(x-a) \right\rangle,$$

where $\xi = (1-\theta)a + \theta x$ for some $0 < \theta < 1$. Here $\nabla^2 f$ is the matrix of second partial derivatives of f, which is known as the *Hessian*:

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix}.$$

In general, the (m,n) element of the Hessian $\nabla^2 f$ is $\frac{\partial^2 f}{\partial x_m \partial x_n}$. Although we do not prove it, for C^2 functions, the order of the partial derivatives can be exchanged: $\frac{\partial^2 f}{\partial x_2 \partial x_1} = \frac{\partial^2 f}{\partial x_1 \partial x_2}$.

Example 4.5. Consider the motivating example. Then

$$\frac{\partial^2 f}{\partial x_1^2} = -1, \ \frac{\partial^2 f}{\partial x_1 \partial x_2} = 0, \ \frac{\partial^2 f}{\partial x_2^2} = -1,$$

so the Hessian is

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

4.5 Chain rule

Let me convince you that the vector and matrix notation is quite useful by proving the chain rule. Instead of a real valued function of several variables, consider a vector valued function, for example

$$f(x) = \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix}.$$

Here the variables are x_1, x_2. $f(x)$ is a two-dimensional vector with first component $f_1(x_1, x_2)$ and second component $f_2(x_1, x_2)$. More generally, you can consider $f : \mathbb{R}^N \to \mathbb{R}^M$, where N is the dimension of the domain (variables) and M is the dimension of the range (value). In the example above, we have

$M = N = 2$. Such a function f is *differentiable* at the point $a = (a_1, \ldots, a_N)$ if there exists an $M \times N$ matrix A and a function $\epsilon(h)$ such that

$$f(a+h) - f(a) = Ah + \epsilon(h)$$

with $\epsilon(h)/\|h\| \to 0$ as $h \to 0$, where $h = (h_1, \ldots, h_N)$. Setting $h_n = 0$ for all but one n, dividing by $h_n \neq 0$, taking the limit as $h_n \to 0$, and comparing the m-th component of both sides, you can show that the (m, n) component of the matrix A is the partial derivative $\frac{\partial f_m}{\partial x_n}(a)$. The matrix A is called the *Jacobian* of f at a, and is often denoted by $Df(a)$. In particular, if the dimension of the range is $M = 1$ (so f is a real valued function), then

$$Df(a) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(a) & \cdots & \frac{\partial f}{\partial x_N}(a) \end{bmatrix},$$

the $1 \times N$ matrix obtained by transposing the gradient $\nabla f(a)$.

With these notations, we can now state and prove the chain rule.

Proposition 4.3. *Let $f : \mathbb{R}^N \to \mathbb{R}^M$ be differentiable at a and $g : \mathbb{R}^M \to \mathbb{R}^L$ be differentiable at $b = f(a)$. Then $g \circ f : \mathbb{R}^M \to \mathbb{R}^L$ defined by $(g \circ f)(x) = g(f(x))$ is differentiable at a and*

$$D(g \circ f)(a) = Dg(b)Df(a). \tag{4.4}$$

Proof. By definition, there exists an $M \times N$ matrix A and a function $\epsilon(h)$ such that

$$f(a+h) - f(a) = Ah + \epsilon(h)$$

with $\epsilon(h)/\|h\| \to 0$ as $h \to 0$. Similarly, there exists an $L \times M$ matrix B and a function $\delta(k)$ such that

$$g(b+k) - g(b) = Bk + \delta(k)$$

with $\delta(k)/\|k\| \to 0$ as $k \to 0$. Consider the function obtained by composing g and f. Letting $k = f(a+h) - f(a)$, we obtain

$$\begin{aligned} g(f(a+h)) - g(f(a)) &= g(b+k) - g(b) = Bk + \delta(k) \\ &= B(f(a+h) - f(a)) + \delta(f(a+h) - f(a)) \\ &= B(Ah + \epsilon(h)) + \delta(Ah + \epsilon(h)) \\ &= BAh + B\epsilon(h) + \delta(Ah + \epsilon(h)). \end{aligned}$$

Since ϵ and δ are negligible compared with their arguments, it follows that $g(f(x))$ is differentiable at $x = a$ and

$$D(g \circ f)(a) = BA = Dg(b)Df(a),$$

which is (4.4). □

Proposition 4.3 generalizes the familiar formula $(g(f(x)))' = g'(f(x))f'(x)$ to the multi-dimensional case.

What does (4.4) mean? For example, let g be a real valued function of two variables, say $g(x_1, x_2)$. Let f be a vector valued function of one variable, say $f(t) = \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}$. Since

$$Dg = \begin{bmatrix} \frac{\partial g}{\partial x_1} & \frac{\partial g}{\partial x_2} \end{bmatrix} \quad \text{and} \quad Df = \begin{bmatrix} f_1'(t) \\ f_2'(t) \end{bmatrix},$$

it follows that

$$\frac{d}{dt}g(f_1(t), f_2(t)) = D(g \circ f) = DgDf$$
$$= \begin{bmatrix} \frac{\partial g}{\partial x_1} & \frac{\partial g}{\partial x_2} \end{bmatrix} \begin{bmatrix} f_1'(t) \\ f_2'(t) \end{bmatrix} = \frac{\partial g}{\partial x_1} f_1'(t) + \frac{\partial g}{\partial x_2} f_2'(t).$$

In general, if f is an M-dimensional function of $x_1, \ldots, x_N$ and g is a real valued function of $y_1, \ldots, y_M$, we have

$$\frac{\partial(g \circ f)}{\partial x_n} = \sum_{m=1}^{M} \frac{\partial g}{\partial y_m} \frac{\partial f_m}{\partial x_n}.$$

Problems

4.1. Compute the partial derivatives, the gradient, and the Hessian of the following functions.

1. $f(x_1, x_2) = a_1 x_1 + a_2 x_2$, where a_1, a_2 are constants.
2. $f(x_1, x_2) = a x_1^2 + 2b x_1 x_2 + c x_2^2$, where a, b, c are constants.
3. $f(x_1, x_2) = x_1 x_2$.
4. $f(x_1, x_2) = x_1 \log x_2$, where $x_2 > 0$.

4.2. Compute the gradient and the Hessian of the following functions.

1. $f(x) = \langle a, x \rangle$, where a, x are vectors of the same dimensions and $\langle a, x \rangle = a \cdot x$ is the inner product of a and x.
2. $f(x) = \langle x, Ax \rangle$, where A is a square matrix of the same dimension as the vector x.

4.3. This problem asks you to prove the multi-variable mean value theorem (Proposition 4.2). Let f be differentiable.

1. Let $g(t) = f(a + t(b - a))$. Using the chain rule, compute $g'(t)$.
2. Using the one-variable mean value theorem, prove the multi-variable mean value theorem.

4.4. This problem asks you to prove that a C^1 function is differentiable. Let $f(x_1, x_2)$ be a C^1 function (i.e., partially differentiable and the partial derivatives are continuous). Fix (a_1, a_2).

1. Using the one-variable mean value theorem, show that there exist numbers $0 < \theta_1, \theta_2 < 1$ such that

   $$f(a_1+h_1, a_2+h_2) - f(a_1, a_2) = \frac{\partial f}{\partial x_1}(a_1+\theta_1 h_1, a_2+h_2)h_1 + \frac{\partial f}{\partial x_2}(a_1, a_2+\theta_2 h_2)h_2.$$

   (Hint: subtract and add $f(a_1, a_2 + h_2)$ to the left-hand side.)

2. Let

   $$\epsilon(h) = f(a + h) - f(a) - \nabla f(a) \cdot h,$$

   where $a = (a_1, a_2)$ and $h = (h_1, h_2)$. Prove that $\lim_{h \to 0} \epsilon(h) / \|h\| = 0$.

Chapter 5

Multi-Variable Unconstrained Optimization

5.1 First and second-order conditions

Consider the unconstrained optimization problem

$$\text{minimize } f(x), \tag{5.1}$$

where f is a (one- or multi-variable) differentiable function. Recall from Proposition 4.1 that if $\bar{x}$ is a solution, then $\nabla f(\bar{x}) = 0$, where

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \end{bmatrix}$$

is the gradient (the vector of partial derivatives) in the two-dimensional case but the general case is similar. The condition $\nabla f(\bar{x}) = 0$ is called the *first-order* condition for optimality. It is necessary, but not sufficient, as the following example shows.

Example 5.1. Let $f(x) = x^3 - 3x$. Since

$$f'(x) = 3x^2 - 3 = 3(x-1)(x+1) \begin{cases} > 0, & (x < -1, x > 1) \\ = 0, & (x = \pm 1) \\ < 0, & (-1 < x < 1) \end{cases}$$

$x = \pm 1$ are stationary points. $x = 1$ is a local minimum and $x = -1$ is a local maximum. However, since $f(x) \to \pm\infty$ as $x \to \pm\infty$, they are neither global minimum nor global maximum.

Can we derive a sufficient condition for optimality? The answer is yes. To this end we need to introduce a few notations. We say that $\bar{x}$ is a *(global) solution* to the unconstrained minimization problem (5.1) if $f(x) \geq f(\bar{x})$ for all

x. We say that $\bar{x}$ is a *local solution* if $f(x) \geq f(\bar{x})$ for all x close enough to $\bar{x}$. Finally, $\bar{x}$ is called a *stationary point* if $\nabla f(\bar{x}) = 0$.

Example 5.1 shows that in general, we can only expect that a stationary point is a local optimum, not a global optimum. We can use Taylor's theorem to derive conditions under which this is indeed true. Suppose that f is a C^2 (twice continuously differentiable) function, and $\bar{x}$ is a stationary point. Take any $x = \bar{x} + h$. By Taylor's theorem, there exists $0 < \alpha < 1$ such that

$$f(\bar{x} + h) = f(\bar{x}) + \langle \nabla f(\bar{x}), h \rangle + \frac{1}{2} \left\langle h, \nabla^2 f(\bar{x} + \alpha h) h \right\rangle, \tag{5.2}$$

where $\langle a, b \rangle$ is the inner product between vectors a, b and $\nabla^2 f$ is the Hessian (matrix of second derivatives) of f:

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix}$$

in the two-variable case. Since $\bar{x}$ is a stationary point, we have $\nabla f(\bar{x}) = 0$, so (5.2) implies

$$f(\bar{x} + h) = f(\bar{x}) + \frac{1}{2} \left\langle h, \nabla^2 f(\bar{x} + \alpha h) h \right\rangle.$$

Therefore whether $f(x) = f(\bar{x} + h)$ is greater than or less than $f(\bar{x})$ depends on whether the quantity $\left\langle h, \nabla^2 f(\bar{x} + \alpha h) h \right\rangle$ is positive or negative. Letting $Q = \nabla^2 f(\bar{x})$ be the Hessian of f evaluated at the stationary point $\bar{x}$, if $h = x - \bar{x}$ is small, then $\nabla^2 f(\bar{x} + \alpha h)$ is close to Q. Therefore

$$f(x) \gtrless f(\bar{x}) \iff \langle h, Qh \rangle \gtrless 0. \tag{5.3}$$

Recall from Chapter 1 that a symmetric matrix A is positive (negative) definite if $\langle h, Ah \rangle > 0$ (< 0) for all vector $h \neq 0$, and that A is positive (negative) semidefinite if $\langle h, Ah \rangle \geq 0$ (≤ 0) for all h. The equivalence (5.3) says that $\bar{x}$ is a local minimum (maximum) if $Q = \nabla^2 f(\bar{x})$ is positive (negative) definite. Thus we obtain the following sufficient condition for local optimality, called the *second-order condition.*

Proposition 5.1. *Let f be a twice continuously differentiable function and $\nabla f(\bar{x}) = 0$. If $\bar{x}$ is a local minimum (maximum), then $\nabla^2 f(\bar{x})$ is positive (negative semidefinite. Conversely, if $\nabla^2 f(\bar{x})$ is positive (negative) definite, then $\bar{x}$ is a local minimum (maximum).*

The proof of Proposition 5.1 is straightforward using the second-order Taylor approximation (5.2), so it is left as an exercise (Problem 5.3).

Example 5.2. Let $f(x) = x^3 - 3x$. Since $f'(x) = 3x^2 - 3$ and $f''(x) = 6x$, we have $f'(\pm 1) = 0$ and $f''(\pm 1) = \pm 6$. Therefore $x = 1$ is a local minimum and $x = -1$ is a local maximum.

Example 5.3. Let $f(x_1, x_2) = x_1^2 + x_1 x_2 + x_2^2$. Then the gradient is $\nabla f(x) = \begin{bmatrix} 2x_1 + x_2 \\ x_1 + 2x_2 \end{bmatrix}$ and the Hessian is

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

Since $\nabla f(0) = 0$, $(x_1, x_2) = (0, 0)$ is a stationary point. Now

$$\begin{aligned}\langle h, \nabla^2 f(0) h\rangle &= \begin{bmatrix} h_1 & h_2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \\ &= 2h_1^2 + 2h_1h_2 + 2h_2^2 = 2\left(h_1 + \frac{1}{2}h_2\right)^2 + \frac{3}{2}h_2^2 \geq 0,\end{aligned}$$

with strict inequality if $(h_1, h_2) \neq (0, 0)$, so $\nabla^2 f(0)$ is positive definite. Therefore $(x_1, x_2) = (0, 0)$ is a local minimum (indeed, a global minimum).

Example 5.4. Let $f(x_1, x_2) = x_1^2 - x_2^2$. Since $\nabla f(x) = \left[\begin{smallmatrix} 2x_1 \\ -2x_2 \end{smallmatrix}\right]$, $(x_1, x_2) = (0, 0)$ is a stationary point. However, since $f(x_1, 0) = x_1^2$ attains the minimum at $x_1 = 0$ and $f(0, x_2) = -x_2^2$ attains the maximum at $x_2 = 0$, $(x_1, x_2) = (0, 0)$ is neither a local minimum nor a local maximum (it is a *saddle point*). Indeed, the Hessian

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

is neither positive nor negative definite.

In order to determine whether a stationary point is a local minimum, maximum, or a saddle point, we need to determine whether the Hessian is positive definite, negative definite, or neither. Although there are a few ways to do so (see Proposition 1.2), usually the easiest way is to complete the squares. If $h = (h_1, h_2)$ and Q is a symmetric matrix, $\langle h, Qh\rangle$ is a quadratic function of h_1, h_2, so you can complete the squares as in the example above. If the result is the sum of two positive terms (N positive terms if there are N variables), then Q is positive (semi)definite. If the result is the sum of negative terms, the Q is negative (semi) definite. If the result is the sum of positive and negative terms, then Q is neither positive nor negative definite. The order you complete the squares doesn't matter—a property known as Sylvester's law of inertia.[1]

5.2 Convex optimization

Recall that in the one-variable case, a twice differentiable function f is convex if $f''(x) \geq 0$ for all x, and if f is convex, $f(\bar{x}) = 0$ implies that $\bar{x}$ is the (global) minimum (Proposition 3.4). That is, the first order condition is necessary and sufficient. The same holds for the multi-variable case.

5.2.1 General case

As in the one-variable case, a function f is said to be *convex* if for any x_1, x_2 and $0 \leq \alpha \leq 1$ we have

$$f((1-\alpha)x_1 + \alpha x_2) \leq (1-\alpha)f(x_1) + \alpha f(x_2).$$

A function is called *concave* if the reverse inequality holds (so $-f$ is convex). Proposition 10.4 shows that a twice continuously differentiable function is convex (concave) if and only if the Hessian is positive (negative) semidefinite.

[1] http://en.wikipedia.org/wiki/Sylvester's_law_of_inertia

Let f be a twice continuously differentiable convex function, and $\bar{x}$ be a stationary point (so $\nabla f(\bar{x}) = 0$). By (5.2) and using the definition of the positive definiteness, we obtain

$$\begin{aligned} f(\bar{x}+h) &= f(\bar{x}) + \langle \nabla f(\bar{x}), h\rangle + \frac{1}{2}\left\langle h, \nabla^2 f(\bar{x}+\alpha h)h\right\rangle \\ &= f(\bar{x}) + \frac{1}{2}\left\langle h, \nabla^2 f(\bar{x}+\alpha h)h\right\rangle \geq f(\bar{x}) \end{aligned}$$

for all h, so $\bar{x}$ is a global minimum. This important result is summarized in the following theorem.

Theorem 5.2. *Let f be a differentiable convex (concave) function. Then $\bar{x}$ is a minimum (maximum) of f if and only if $\nabla f(\bar{x}) = 0$.*

Although the above discussion requires that f is twice continuously differentiable, actually this is not necessary, as we see in Proposition 11.1.

5.2.2 Quadratic case

A special but important class of convex and concave functions are *quadratic* functions, because we can solve for the optimum in closed-form. A general quadratic function with two variables has the following form:

$$f(x_1, x_2) = a + b_1 x_1 + b_2 x_2 + c_1 x_1^2 + c_2 x_1 x_2 + c_3 x_2^2,$$

where a, b, c's are constants. It turns out that it is useful to change the notation such that $c_1 = \frac{1}{2}q_{11}$, $c_2 = q_{12}$, and $c_3 = \frac{1}{2}q_{22}$. Then

$$\begin{aligned} f(x_1, x_2) &= a + b_1 x_1 + b_2 x_2 + \frac{1}{2}q_{11}x_1^2 + q_{12}x_1 x_2 + \frac{1}{2}q_{22}x_2^2 \\ &= a + \langle b, x\rangle + \frac{1}{2}\langle x, Qx\rangle, \end{aligned}$$

where $b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ and $Q = \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix}$. The gradient is

$$\nabla f(x) = \begin{bmatrix} b_1 + q_{11}x_1 + q_{12}x_2 \\ b_2 + q_{12}x_1 + q_{22}x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = b + Qx,$$

and the Hessian is

$$\nabla^2 f(x) = \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix} = Q.$$

The vector and matrix notation is valid with an arbitrary number of variables.

Since the Hessian of a quadratic function is constant, f is convex (concave) if Q is positive (negative) semidefinite. Since

$$0 = \nabla f(x) = b + Qx \iff x = -Q^{-1}b,$$

(Q^{-1} is the inverse matrix of Q) if Q is positive (negative) definite, then $\bar{x} = -Q^{-1}b$ is the minimum (maximum) of $f(x) = a + \langle b, x\rangle + \frac{1}{2}\langle x, Qx\rangle$.

Problems

5.1. Let $f(x) = 10x^3 - 15x^2 - 60x$. Find the local maxima and minima of f. Does f have global maximum and minimum?

5.2. Let $f(x) = 180x - 15x^2 - 10x^3$. Solve

$$\text{maximize } f(x) \text{ subject to } x \geq 0.$$

5.3. Prove Proposition 5.1.

5.4. Let $f(x_1, x_2) = x_1^2 - x_1x_2 + 2x_2^2 - x_1 - 3x_2$.

1. Compute the gradient and the Hessian of f.
2. Determine whether f is convex, concave, or neither.
3. Find the stationary point(s) of f.
4. Determine whether each stationary point is a maximum, minimum, or neither.

5.5. Let $f(x_1, x_2) = x_1^2 - x_1x_2 - 6x_1 + x_2^3 - 3x_2$.

1. Find the stationary point(s) of f.
2. Determine whether each stationary point is a local maximum, local minimum, or a saddle point.

5.6. Let $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ be a 2×2 symmetric matrix. Show that A is positive definite if and only if $a > 0$ and $ac - b^2 > 0$.

5.7. For each of the following symmetric matrices, show whether it is positive (semi)definite, negative (semi)definite, or neither.

1. $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$,
2. $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$,
3. $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$,
4. $A = \begin{bmatrix} 2 & 1 \\ 1 & -2 \end{bmatrix}$,
5. $A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$,
6. $A = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$,
7. $A = \begin{bmatrix} -3 & 1 \\ 1 & -4 \end{bmatrix}$,

5.8. For each of the following functions, show whether it is convex, concave, or neither.

1. $f(x_1, x_2) = x_1 x_2 - x_1^2 - x_2^2$,
2. $f(x_1, x_2) = 3x_1 + 2x_1^2 + 4x_2 + x_2^2 - 2x_1 x_2$,
3. $f(x_1, x_2) = x_1^2 + 3x_1 x_2 + 2x_2^2$,
4. $f(x_1, x_2) = 20x_1 + 10x_2$,
5. $f(x_1, x_2) = x_1 x_2$,
6. $f(x_1, x_2) = \mathrm{e}^{x_1} + \mathrm{e}^{x_2}$,
7. $f(x_1, x_2) = \log x_1 + \log x_2$, where $x_1, x_2 > 0$,
8. $f(x_1, x_2) = \log(\mathrm{e}^{x_1} + \mathrm{e}^{x_2})$,
9. $f(x_1, x_2) = (x_1^p + x_2^p)^{\frac{1}{p}}$, where $x_1, x_2 > 0$ and $p \neq 0$ is a constant.

5.9. Let $p \geq 1$. For $x \in \mathbb{R}^N$, let $\|x\|_p = \left(\sum_{n=1}^N |x_n|^p\right)^{1/p}$. Show Minkowski's inequality

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p$$

and that $\|\cdot\|_p$ is a norm on $\mathbb{R}^N$.

5.10. Let $K \geq 2$. Prove that f is convex if and only if

$$f\left(\sum_{k=1}^K \alpha_k x_k\right) \leq \sum_{k=1}^K \alpha_k f(x_k)$$

for all $\{x_k\}_{k=1}^K \subset \mathbb{R}^N$ and $\{\alpha_k\} \geq 0$ such that $\sum_{k=1}^K \alpha_k = 1$.

5.11.

1. Prove that $f(x) = \log x$ $(x > 0)$ is strictly concave.
2. Prove the following *inequality of arithmetic and geometric means*: for any $x_1, \ldots, x_K > 0$ and $\alpha_1, \ldots, \alpha_K > 0$ such that $\sum \alpha_k = 1$, we have

$$\sum_{k=1}^K \alpha_k x_k \geq \prod_{k=1}^K x_k^{\alpha_k},$$

with equality if and only if $x_1 = \cdots = x_K$.

5.12. Let $p, q > 0$ be numbers such that $1/p + 1/q = 1$.

1. Fixing $b \geq 0$, define $f(x) = \frac{1}{p}x^p - bx + \frac{1}{q}b^q$ for $x \geq 0$. Show that f is convex.
2. Show Young's inequality

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

for all $a, b \geq 0$.

3. Let $x, y \in \mathbb{R}^N$. Define $\|x\|_p = \left(\sum_{n=1}^N |x_n|^p\right)^{1/p}$. Show Hölder's inequality

$$\sum_{n=1}^N |x_n y_n| \leq \|x\|_p \|y\|_q .$$

(Hint: set $a = |x_n| / \|x\|_p$, $b = |y_n| / \|y\|_q$ and use Young.)

5.13. 1. Show that if $\{f_i(x)\}_{i=1}^I$ are convex, so is $f(x) = \sum_{i=1}^I \alpha_i f_i(x)$ for any $\alpha_1, \ldots, \alpha_I \geq 0$.

2. Show that if $\{f_i(x)\}_{i=1}^I$ are convex, so is $f(x) = \max_{1 \leq i \leq I} f_i(x)$.

3. Suppose that $h : \mathbb{R}^M \to \mathbb{R}$ is increasing (meaning that h is increasing in each coordinate $x_1, \ldots, x_M$) and convex and $g_m : \mathbb{R}^N \to \mathbb{R}$ is convex for $m = 1, \ldots, M$. Prove that $f(x) = h(g_1(x), \ldots, g_M(x))$ is convex.

5.14. Let $f : \mathbb{R}^N \to (-\infty, \infty]$ be convex.

1. Show that the set of solutions to $\min_{x \in \mathbb{R}^N} f(x)$ is a convex set.

2. If f is strictly convex, show that the solution (if it exists) is unique.

5.15. Suppose you collect some two-dimensional data $\{(x_n, y_n)\}_{n=1}^N$, where N is the sample size. You wish to fit a straight line $y = a + bx$ to the data. Suppose you do so by making the observed value y_n as close as possible to the theoretical value $a + bx_n$ by minimizing the sum of squares

$$f(a, b) = \sum_{n=1}^N (y_n - a - bx_n)^2.$$

1. Is f convex, concave, or neither?

2. Compute the gradient of f.

3. Express a, b that minimize f using the following quantities:

$$\mathrm{E}[X] = \frac{1}{N}\sum_{n=1}^N x_n, \qquad \mathrm{Var}[X] = \frac{1}{N}\sum_{n=1}^N (x_n - \mathrm{E}[X])^2,$$
$$\mathrm{E}[Y] = \frac{1}{N}\sum_{n=1}^N y_n, \qquad \mathrm{Cov}[X, Y] = \frac{1}{N}\sum_{n=1}^N (x_n - \mathrm{E}[X])(y_n - \mathrm{E}[Y]).$$

5.16. In the previous problem, the variable y is explained by two variables, $1, x$. Generalize the problem when y is explained by K variables, $x = (x_1, \ldots, x_K)$. It will be useful to define the $N \times K$ matrix $\mathbf{X} = (x_{nk})$ and N-vector $\mathbf{y} = (y_1, \ldots, y_N)$, where x_{nk} is the n-th observation of the k-th variable. The equation you want to fit is

$$y_n = \beta_1 x_{n1} + \cdots + \beta_K x_{nK} + \text{error term},$$

and $\beta = (\beta_1, \ldots, \beta_K)$ is the vector of coefficients.

5.17. Let A be a symmetric positive definite matrix.

1. Let $f(x) = \langle y, x\rangle - \frac{1}{2}\langle x, Ax\rangle$, where y is a fixed vector. Compute the gradient and Hessian of f and show that f is concave.

2. Find the maximum of f and its value.

3. Let A, B be symmetric positive definite matrices. We write $A \succeq B$ if the matrix $C = A - B$ is positive semidefinite. Show that $A \succeq B$ if and only if $B^{-1} \succeq A^{-1}$.

This problem is motivated by Toda (2011).

Chapter 6

Multi-Variable Constrained Optimization

In the real world, optimization problems come with *constraints*. Most of us have a budget and cannot spend more money than we have, so we have to choose what to buy or not. Our stomach has a finite capacity and we cannot eat more than a certain amount, so we must choose what to eat or not. This chapter provides an intuitive introduction to the optimization of a multi-variable function subject to constraints. The rigorous theory is developed in Chapter 12.

6.1 A motivating example

6.1.1 The problem

Suppose there are two goods (say apples and bananas), and your satisfaction is represented by the function (called *utility function*)

$$u(c_1, c_2) = \log c_1 + \log c_2,$$

where c_1, c_2 are the amounts of good 1 and 2 that you consume. Suppose that the unit price of goods are p_1 and p_2, and your budget is w. If you buy c_1 and c_2 units of good each, your expenditure is

$$p_1 c_1 + p_2 c_2.$$

Since your budget is w, your *budget constraint* is

$$p_1 c_1 + p_2 c_2 \leq w.$$

So the problem of attaining maximum satisfaction within your budget can be mathematically expressed as:

$$\begin{array}{ll} \text{maximize} & \log c_1 + \log c_2 \\ \text{subject to} & p_1 c_1 + p_2 c_2 \leq w. \end{array}$$

Here $u(c_1, c_2) = \log c_1 + \log c_2$ is called the *objective function*, and $p_1 c_1 + p_2 c_2 \leq w$ is the *constraint*.[1]

[1] Strictly speaking, there are other constraints $c_1 \geq 0$ and $c_2 \geq 0$, since you cannot consume a negative amount.

6.1.2 A solution

How can we solve this problem? Some of you might find a trick that turns this *constrained* optimization problem into an *unconstrained* one, as follows. First, since the objective function $\log c_1 + \log c_2$ is increasing in both c_1 and c_2, you will always exhaust your budget. That is, you will always want to consume in a way such that the budget constraint holds with equality, i.e.,

$$p_1 c_1 + p_2 c_2 = w.$$

Solving this for c_2, we get

$$c_2 = \frac{w - p_1 c_1}{p_2}.$$

Substituting this into the objective function, the problem is equivalent to finding the maximum of

$$f(c_1) = \log c_1 + \log \frac{w - p_1 c_1}{p_2} = \log c_1 + \log(w - p_1 c_1) - \log p_2.$$

Setting the derivative equal to zero, we get

$$f'(c_1) = \frac{1}{c_1} - \frac{p_1}{w - p_1 c_1} = 0 \iff c_1 = \frac{w}{2p_1}.$$

In this case, f tends to $-\infty$ when c_1 approaches the boundaries $c_1 = 0$ and $c_1 = w/p_1$, so we need not worry about the boundaries. The value of c_2 corresponding to $c_1 = \frac{w}{2p_1}$ is

$$c_2 = \frac{w - p_1 \frac{w}{2p_1}}{p_2} = \frac{w}{2p_2}.$$

Therefore the solution is

$$(c_1, c_2) = \left(\frac{w}{2p_1}, \frac{w}{2p_2} \right).$$

6.1.3 Why study the general theory?

The above solution is mathematically correct, but too special to be useful. The reasons it worked are

1. the inequality constraint $p_1 c_1 + p_2 c_2 \leq w$ could be turned into an equality constraint $p_1 c_1 + p_2 c_2 = w$,
2. the equality constraint $p_1 c_1 + p_2 c_2 = w$ could be solved for one variable c_2, and
3. after substitution the optimization problem became unconstrained, which is why we could apply calculus.

But in general we cannot hope that any of these steps work. As an exercise, try to solve the following problem:

$$\begin{aligned} &\text{maximize} && x_1 + 2x_2 + 3x_3 \\ &\text{subject to} && x_1^2 + x_2^2 + x_3^2 \leq 1. \end{aligned}$$

Some of you might be able to solve this problem using an ingenious trick, but such tricks are generally inapplicable. That is why we need a general theory for solving constrained optimization problems.

6.2 Optimization with linear constraints

To build intuition, we start the discussion of constrained optimization from the simplest cases, namely when constraints are linear.

6.2.1 One linear constraint

Consider the two-variable optimization problem with one linear constraint,

$$\begin{aligned} &\text{minimize} && f(x_1, x_2) \\ &\text{subject to} && a_1 x_1 + a_2 x_2 \le c, \end{aligned}$$

where f is differentiable and a_1, a_2, c are constants. Suppose a solution $\bar{x} = (\bar{x}_1, \bar{x}_2)$ exists. The goal is to derive a necessary condition for $\bar{x}$.

If $a_1\bar{x}_1 + a_2\bar{x}_2 < c$, so the constraint does not *bind* or is *inactive*, since x can move freely around $\bar{x}$, the point $\bar{x}$ must be a local minimum of f. Therefore $\nabla f(\bar{x}) = 0$. If $a_1\bar{x}_1 + a_2\bar{x}_2 = c$, so the constraint *binds* or is *active*, then the situation is more complicated. Let

$$\Omega = \{x = (x_1, x_2) \,|\, a_1 x_1 + a_2 x_2 \le c\}$$

be the constraint set. The boundary of Ω is the straight line $a_1x_1 + a_2x_2 = c$, which has the normal vector $a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$. Figure 6.1 shows the constraint set Ω (with the boundary and the normal vector), the solution $\bar{x}$, and the negative of the gradient $-\nabla f(\bar{x})$. (Here we draw the negative of the gradient because that is the direction at which the function f decreases fastest, and we are solving a minimization problem.)

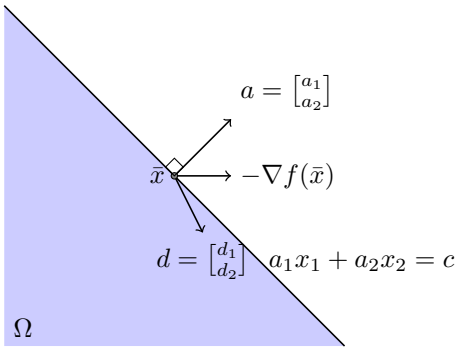

Figure 6.1. Gradient and feasible direction.

Consider moving towards the direction d from the solution $\bar{x}$. Since $\bar{x}$ is on the boundary, we have $\langle a, \bar{x}\rangle = c$. The point $x = \bar{x} + td$ (where $t > 0$ is small) is feasible if and only if

$$\langle a, \bar{x} + td\rangle \le c = \langle a, \bar{x}\rangle \iff \langle a, d\rangle \le 0,$$

which is when the vectors a, d make an obtuse angle as in the picture. Since $\bar{x}$ is a solution, we have $f(\bar{x} + td) \ge f(\bar{x})$ for small enough $t > 0$. Therefore

$$0 \le \lim_{t \downarrow 0} \frac{f(\bar{x} + td) - f(\bar{x})}{t} = \langle \nabla f(\bar{x}), d\rangle \iff \langle -\nabla f(\bar{x}), d\rangle \le 0,$$

so the vectors $-\nabla f(\bar{x})$ and d make an obtuse angle. Therefore we obtain the following necessary condition for optimality:

If a and d make an obtuse angle, then so do $-\nabla f(\bar{x})$ and d.

In Figure 6.1, the angle between $-\nabla f(\bar{x})$ and d is acute, so f decreases towards the direction d. But this is a contradiction because $\bar{x}$ is by assumption a solution to the constrained minimization problem, so f cannot decrease towards any feasible direction. Thus Figure 6.1 is false.

The only case that $-\nabla f(\bar{x})$ and d make an obtuse angle whenever a and d do so is when $-\nabla f(\bar{x})$ and a point to the same direction, as in Figure 6.2. Therefore if $\bar{x}$ is a solution, there must be a number $\lambda \geq 0$ such that

$$\nabla f(\bar{x}) = -\lambda a \iff \nabla f(\bar{x}) + \lambda a = 0.$$

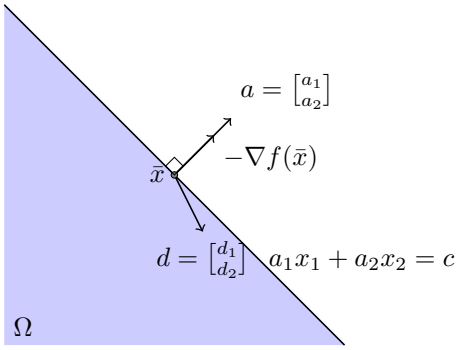

Figure 6.2. Necessary condition for optimality.

In the discussion above, we considered two cases depending on whether the constraint is binding (active) or not binding (inactive), but the inactive case ($\nabla f(\bar{x}) = 0$) is a special case of the active case ($\nabla f(\bar{x}) + \lambda a = 0$) by setting $\lambda = 0$. Furthermore, although we explained the intuition in two-dimension, the result clearly holds in arbitrary dimensions. Therefore we can summarize the necessary condition for optimality as in the following proposition.

Proposition 6.1. *Consider the optimization problem*

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \langle a, x \rangle \leq c, \end{array}$$

where f is differentiable, a is a nonzero vector, and c is a constant. If $\bar{x}$ is a solution, then there exists a number $\lambda \geq 0$ such that

$$\nabla f(\bar{x}) + \lambda a = 0.$$

Example 6.1. Consider the motivating example

$$\begin{array}{ll} \text{maximize} & \log x_1 + \log x_2 \\ \text{subject to} & p_1 x_1 + p_2 x_2 \leq w. \end{array}$$

Maximizing $\log x_1 + \log x_2$ is the same as minimizing $f(x_1, x_2) = -\log x_1 - \log x_2$. The gradient is $\nabla f(x) = -\left[\begin{smallmatrix} 1/x_1 \\ 1/x_2 \end{smallmatrix}\right]$. Let $p = \left[\begin{smallmatrix} p_1 \\ p_2 \end{smallmatrix}\right]$. By Proposition 6.1, there is a number $\lambda \geq 0$ such that

$$\nabla f(\bar{x}) + \lambda p = 0 \iff -\begin{bmatrix} 1/x_1 \\ 1/x_2 \end{bmatrix} + \lambda \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Therefore it must be $x_1 = \frac{1}{\lambda p_1}$ and $x_2 = \frac{1}{\lambda p_2}$.

Since the objective function is increasing, at the solution the constraint $p_1 x_1 + p_2 x_2 \leq w$ must bind. Therefore

$$p_1 x_1 + p_2 x_2 = w \iff p_1 \frac{1}{\lambda p_1} + p_2 \frac{1}{\lambda p_2} = w \iff \lambda = \frac{2}{w}.$$

Substituting again, we get the solution $(x_1, x_2) = (\frac{w}{2p_1}, \frac{w}{2p_2})$.

6.2.2 Multiple linear constraints

Now consider the optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \langle a_1, x \rangle \leq c_1, \\ & \langle a_2, x \rangle \leq c_2, \end{array}$$

where f is differentiable, a_1, a_2 are nonzero vectors, and c_1, c_2 are constants. Let $\bar{x}$ be a solution and Ω be the constraint set, i.e., the set

$$\Omega = \{x \,|\, g_1(x) \leq 0, g_2(x) \leq 0\},$$

where $g_i(x) = \langle a_i, x \rangle - c_i$ for $i = 1, 2$. Assume that both constraints are active at the solution. Figure 6.3 shows the situation.

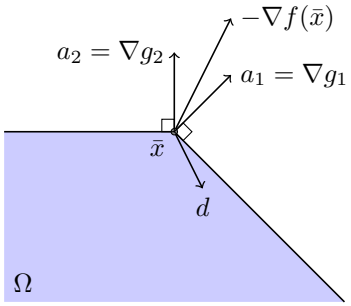

Figure 6.3. Gradient and feasible direction.

In general, the vector d is called a *feasible direction* if you can move a little bit towards the direction d from the point $\bar{x}$, so $\bar{x} + td \in \Omega$ for small enough $t > 0$. If d is a feasible direction and $\bar{x}$ is the minimum of f, then f cannot decrease towards the direction d, so we must have

$$0 \leq \lim_{t \downarrow 0} \frac{f(\bar{x} + td) - f(\bar{x})}{t} = \langle \nabla f(\bar{x}), d \rangle \iff \langle -\nabla f(\bar{x}), d \rangle \leq 0.$$

Therefore the negative of the gradient $-\nabla f(\bar{x})$ and any feasible direction d must make an obtuse angle. Recall that d is a feasible direction if d and $a_i = \nabla g_i$ make an obtuse angle for $i = 1, 2$. By looking at Figure 6.3, in order for $\bar{x}$ to be the minimum, it is necessary that $-\nabla f(\bar{x})$ lies between the vectors a_1 and a_2. This is true if and only if there are numbers $\lambda_1, \lambda_2 \geq 0$ such that

$$-\nabla f(\bar{x}) = \lambda_1 a_1 + \lambda_2 a_2 \iff \nabla f(\bar{x}) + \lambda_1 \nabla g_1(\bar{x}) + \lambda_2 \nabla g_2(\bar{x}) = 0.$$

Although we have assumed that both constraints bind, this equation is true even if one (or both) of them does not bind by setting $\lambda_1 = 0$ and/or $\lambda_2 = 0$. Also, it is clear that this argument holds for an arbitrary number of linear constraints. Therefore we obtain the following general theorem.

Theorem 6.2 (Karush-Kuhn-Tucker theorem with linear constraints)**.** *Consider the optimization problem*

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & g_i(x) \leq 0 & (i = 1, \ldots, I), \end{array}$$

where f is differentiable and $g_i(x) = \langle a_i, x \rangle - c_i$ is linear with $a_i \neq 0$. If $\bar{x}$ is a solution, then there exist numbers (called Lagrange multipliers*) $\lambda_1, \ldots, \lambda_I$ such that*

$$\nabla f(\bar{x}) + \sum_{i=1}^{I} \lambda_i \nabla g_i(\bar{x}) = 0, \tag{6.1a}$$

$$(\forall i)\ \lambda_i \geq 0,\ g_i(\bar{x}) \leq 0,\ \lambda_i g_i(\bar{x}) = 0. \tag{6.1b}$$

Condition (6.1a) is called the *first-order condition*. Its interpretation is that at the minimum $\bar{x}$, the negative of the gradient $-\nabla f(\bar{x})$ must lie between all the normal vectors $a_i = \nabla g_i(\bar{x})$ corresponding to the active constraints. Condition (6.1b) is called the *complementary slackness condition*. The first-order condition and the complementary slackness condition are jointly called the *Karush-Kuhn-Tucker (KKT) conditions*.[2] The condition $\lambda_i \geq 0$ says that the Lagrange multiplier is nonnegative, and $g_i(\bar{x}) \leq 0$ says that the constraint is satisfied, which are not new. The condition $\lambda_i g_i(\bar{x}) = 0$ takes care of both the active (binding) and inactive (non-binding) cases. If the constraint i is active, then $g_i(\bar{x}) = 0$, so we have $\lambda_i g_i(\bar{x}) = 0$ automatically. If the constraint i is inactive, we have $\lambda_i = 0$, so again $\lambda_i g_i(\bar{x}) = 0$ automatically.

An easy way to remember (6.1a) is as follows. Given the objective function $f(x)$ and the constraints $g_i(x) \leq 0$, define the *Lagrangian*

$$L(x, \lambda) = f(x) + \sum_{i=1}^{I} \lambda_i g_i(x),$$

where $\lambda = (\lambda_1, \ldots, \lambda_I)$. The Lagrangian is the sum of the objective function $f(x)$ and the weighted sum of the constraint functions $g_i(x)$ weighted by the

[2] A version of this theorem appeared in the 1939 Master's thesis of William Karush (1917-1997) but did not get much attention. (Applied Mathematics gained respect only after proving its usefulness during World War II.) The theorem became widely known after the rediscovery by Harold Kuhn (1925-2014) and Albert Tucker (1905-1995) in a conference paper in 1951. However, the paper (http://projecteuclid.org/download/pdf_1/euclid.bsmsp/1200500249) does not cite Karush.

Lagrange multipliers λ_i. Pretend that λ is constant and you want to minimize $L(x, \lambda)$ with respect to x. Then the first-order condition is

$$0 = \nabla L(x, \lambda) = \nabla f(x) + \sum_{i=1}^{I} \lambda_i \nabla g_i(x),$$

which is the same as (6.1a).

6.2.3 Linear inequality and equality constraints

So far we have considered the case when all constraints are inequalities, but what if there are also equalities? For example, consider the optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \langle a, x\rangle \le c, \\ & \langle b, x\rangle = d, \end{array}$$

where f is differentiable, a, b are nonzero vectors, and c, d are constants. We derive a necessary condition for optimality by turning this problem into one with only inequality constraints. Note that $\langle b, x\rangle = d$ is equivalent to $\langle b, x\rangle \le d$ and $\langle b, x\rangle \ge d$. Furthermore, $\langle b, x\rangle \ge d$ is equivalent to $\langle -b, x\rangle \le -d$. Therefore the problem is equivalent to:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \langle a, x\rangle - c \le 0, \\ & \langle b, x\rangle - d \le 0, \\ & \langle -b, x\rangle + d \le 0. \end{array}$$

Setting $g_1(x) = \langle a, x\rangle - c$, $g_2(x) = \langle b, x\rangle - d$, and $g_3(x) = \langle -b, x\rangle + d$, this problem is exactly the form in Theorem 6.2. Therefore there exist Lagrange multipliers $\lambda_1, \lambda_2, \lambda_3 \ge 0$ such that

$$\nabla f(\bar{x}) + \lambda_1 \nabla g_1(\bar{x}) + \lambda_2 \nabla g_2(\bar{x}) + \lambda_3 \nabla g_3(\bar{x}) = 0.$$

Substituting $\nabla g_i(\bar{x})$'s, we get

$$\nabla f(\bar{x}) + \lambda_1 a + \lambda_2 b + \lambda_3(-b) = 0.$$

Letting $\lambda = \lambda_1$ and $\mu = \lambda_2 - \lambda_3$, we get

$$\nabla f(\bar{x}) + \lambda a + \mu b = 0.$$

This equation is similar to the KKT condition (6.1), except that μ can be positive or negative. In general, we obtain the following theorem.

Theorem 6.3 (Karush-Kuhn-Tucker theorem with linear constraints)**.** *Consider the optimization problem*

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & g_i(x) \le 0 & (i = 1, \dots, I), \\ & h_j(x) = 0 & (j = 1, \dots, J), \end{array}$$

where f is differentiable and $g_i(x) = \langle a_i, x\rangle - c_i$ and $h_j(x) = \langle b_j, x\rangle - d_j$ are linear with $a_i, b_j \neq 0$. If $\bar{x}$ is a solution, then there exist Lagrange multipliers $\lambda_1, \ldots, \lambda_I$ and $\mu_1, \ldots, \mu_J$ such that

$$\nabla f(\bar{x}) + \sum_{i=1}^{I} \lambda_i \nabla g_i(\bar{x}) + \sum_{j=1}^{J} \mu_j \nabla h_j(\bar{x}) = 0, \tag{6.2a}$$

$$(\forall i)\ \lambda_i \geq 0,\ g_i(\bar{x}) \leq 0,\ \lambda_i g_i(\bar{x}) = 0, \tag{6.2b}$$

$$(\forall j)\ h_j(\bar{x}) = 0. \tag{6.2c}$$

An easy way to remember the KKT conditions (6.2) is as follows. As in the case with only inequality constraints, define the Lagrangian

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{I} \lambda_i g_i(x) + \sum_{j=1}^{J} \mu_j h_j(x).$$

Pretend that λ, μ are constants and you want to minimize L with respect to x. The first-order condition is $\nabla_x L(x, \lambda, \mu) = 0$, which is exactly (6.2a). The complementary slackness condition (6.2b) is the same as in the case with only inequality constraints. The new condition (6.2c) merely says that the solution $\bar{x}$ must satisfy the equality constraints $h_j(x) = 0$.

6.3 Optimization with nonlinear constraints

We now consider the general case, the optimization of a nonlinear multi-variable function subject to nonlinear constraints.

6.3.1 Karush-Kuhn-Tucker theorem

Consider the optimization problem

$$\begin{aligned} &\text{minimize} && f(x) && && (6.3a)\\ &\text{subject to} && g_i(x) \leq 0 && (i = 1, \ldots, I). && (6.3b) \end{aligned}$$

Let $\bar{x}$ be a solution. By Taylor's theorem, we have

$$g_i(x) \approx g_i(\bar{x}) + \langle \nabla g_i(\bar{x}), x - \bar{x}\rangle .$$

The gradient of both sides at $x = \bar{x}$ is $\nabla g_i(\bar{x})$. A natural idea is to approximate the nonlinear constraints by linear ones around the solution this way, and derive a necessary condition corresponding to these linear constraints. This idea indeed works, subject to some caveats.

Theorem 6.4 (Karush-Kuhn-Tucker theorem with nonlinear constraints)**.** *Consider the optimization problem (6.3), where f and g_i's are differentiable. If $\bar{x}$ is a solution and a regularity condition (called* constraint qualification, CQ*) holds, then there exist Lagrange multipliers $\lambda_1, \ldots, \lambda_I$ such that*

$$\nabla f(\bar{x}) + \sum_{i=1}^{I} \lambda_i \nabla g_i(\bar{x}) = 0, \tag{6.4a}$$

$$(\forall i)\ \lambda_i \geq 0,\ g_i(\bar{x}) \leq 0,\ \lambda_i g_i(\bar{x}) = 0. \tag{6.4b}$$

Proving the Karush-Kuhn-Tucker theorem in the full generality is beyond the scope of this chapter, and the rigorous discussion is deferred to Chapter 12. The conclusion of Theorem 6.4 is the same as that of Theorem 6.2. What is different is the assumption. While in the linear case the conclusion holds without any qualification, in the nonlinear case we need to verify certain "constraint qualifications". The following trivial example shows the need for such conditions.

Example 6.2. Consider the minimization problem

$$\begin{aligned} &\text{minimize} && x \\ &\text{subject to} && -x^3 \leq 0. \end{aligned}$$

Since $-x^3 \leq 0 \iff x \geq 0$, the solution is obviously $\bar{x} = 0$. Now let $f(x) = x$ and $g(x) = -x^3$. If the KKT theorem holds, there must be a Lagrange multiplier $\lambda \geq 0$ such that

$$f'(\bar{x}) + \lambda g'(\bar{x}) = 0.$$

But since $f'(x) = 1$ and $g'(x) = -3x^2$, we have $f'(0) = 1$ and $g'(0) = 0$, so there is no number $\lambda \geq 0$ such that $f'(\bar{x}) + \lambda g'(\bar{x}) = 0$.

Below are a few examples of the constraint qualifications. In general, let $I(\bar{x})$ be the set of indices such that $g_i(\bar{x}) = 0$—that is, i's for which the constraint $g_i(x) \leq 0$ binds at $x = \bar{x}$.

Linear independence (LICQ) The gradients of the active constraints,

$$\{\nabla g_i(\bar{x}) \,|\, i \in I(\bar{x})\},$$

are linearly independent.

Slater (SCQ) g_i's are convex, and there exists a point x_0 such that the constraints are satisfied with strict inequalities: $g_i(x_0) < 0$ for all i.

There are other weaker constraint qualifications, which are deferred to Chapter 12.

In practice, many optimization problems have only linear constraints, in which case there is no need to check any constraint qualifications. If there are equality constraints as well as inequality constraints, then the conclusion of Theorem 6.3 holds under certain constraint qualifications.

6.3.2 Convex optimization

We previously learned (Theorem 5.2) that for unconstrained optimization problems, the first-order condition is not just necessary but also sufficient for optimality when the objective function is convex or concave. The same holds for constrained optimization problems. Indeed, we can prove the following theorem.

Theorem 6.5 (Karush-Kuhn-Tucker theorem for convex optimization)**.** *Consider the optimization problem (6.3), where f and g_i's are differentiable and convex.*

1. *If $\bar{x}$ is a solution and there exists a point x_0 such that $g_i(x_0) < 0$ for all i, then there exist Lagrange multipliers $\lambda_1, \ldots, \lambda_I$ such that (6.4) holds.*

2. *If there exist Lagrange multipliers $\lambda_1, \ldots, \lambda_I$ such that (6.4) holds, then $\bar{x}$ is a solution.*

The first part of Theorem 6.5 is just the KKT theorem with the Slater constraint qualification. The second part says that the first-order conditions are sufficient for optimality.

Proof. We only prove the second part. Suppose that there exist Lagrange multipliers $\lambda_1, \ldots, \lambda_I$ such that (6.4) holds. Let

$$L(x, \lambda) = f(x) + \sum_{i=1}^{I} \lambda_i g_i(x)$$

be the Lagrangian. Since $\lambda_i \geq 0$ and f, g_i's are convex, L is convex as a function of x. By the first-order condition (6.4a), we have

$$0 = \nabla f(\bar{x}) + \sum_{i=1}^{I} \lambda_i \nabla g_i(\bar{x}) = \nabla_x L(\bar{x}, \lambda),$$

so L attains the minimum at $\bar{x}$. Therefore

$$\begin{aligned}
f(\bar{x}) &= f(\bar{x}) + \sum_{i=1}^{I} \lambda_i g_i(\bar{x}) && (\because \lambda_i g_i(\bar{x}) = 0 \text{ for all } i \text{ by (6.4b)}) \\
&= L(\bar{x}, \lambda) \leq L(x, \lambda) && (\because \nabla_x L(\bar{x}, \lambda) = 0) \\
&= f(x) + \sum_{i=1}^{I} \lambda_i g_i(x) && (\because \text{Definition of } L) \\
&\leq f(x), && (\because \lambda_i \geq 0 \text{ and } g_i(x) \leq 0 \text{ for all } i)
\end{aligned}$$

so $\bar{x}$ is a solution. □

Theorem 6.5 makes our life easy for solving nonlinear constrained optimization problems. A general approach is as follows.

Step 1. Verify that f and g_i's are differentiable and convex, and that the Slater constraint qualification holds.

Step 2. Set the Lagrangian $L(x, \lambda) = f(x) + \sum_{i=1}^{I} \lambda_i g_i(x)$ and derive the KKT conditions (6.4).

Step 3. Solve for x and λ. The solution of the original problem is x.

Example 6.3. Consider the problem

$$\begin{aligned}
&\text{minimize} && \frac{1}{x_1} + \frac{1}{x_2} \\
&\text{subject to} && x_1 + x_2 \leq 2,
\end{aligned}$$

where $x_1, x_2 > 0$. Let us solve this problem step by step.

Step 1. Let $f(x_1, x_2) = \frac{1}{x_1} + \frac{1}{x_2}$ be the objective function. Since

$$(1/x)'' = (-x^{-2})' = 2x^{-3} > 0,$$

the Hessian of f,

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

is positive definite. Therefore f is convex. Let $g(x_1, x_2) = x_1 + x_2 - 2$. Since g is linear, it is convex. For $(x_1, x_2) = (\frac{1}{2}, \frac{1}{2})$ we have $g(x_1, x_2) = -1 < 0$, so the Slater condition holds.

Step 2. Let

$$L(x_1, x_2, \lambda) = \frac{1}{x_1} + \frac{1}{x_2} + \lambda(x_1 + x_2 - 2)$$

be the Lagrangian. The first-order condition is

$$0 = \frac{\partial L}{\partial x_1} = -\frac{1}{x_1^2} + \lambda \iff x_1 = \frac{1}{\sqrt{\lambda}},$$
$$0 = \frac{\partial L}{\partial x_2} = -\frac{1}{x_2^2} + \lambda \iff x_2 = \frac{1}{\sqrt{\lambda}}.$$

The complementary slackness condition is

$$\lambda(x_1 + x_2 - 2) = 0.$$

Step 3. From these equations it must be $\lambda > 0$ and

$$x_1 + x_2 - 2 = 0 \iff \frac{2}{\sqrt{\lambda}} - 2 = 0 \iff \lambda = 1$$

and $x_1 = x_2 = 1/\sqrt{\lambda} = 1$. Therefore $(x_1, x_2) = (1, 1)$ is the (only) solution.

In practice, in convex optimization problems (meaning that both the objective function and the constraints are convex) we often skip verifying the Slater constraint qualification. The reason is that if you set up the Lagrangian and find a point that satisfies the KKT conditions, then it is automatically a solution by the second part of Theorem 6.5. If you cannot find any point satisfying the KKT conditions, then either there is no solution or the Slater constraint qualification is violated.

If the problem is *not* a convex optimization problem, then the procedure is slightly more complicated.

Step 1. Show that a solution exists (e.g., by showing that the objective function is continuous and the constraint set is compact).

Step 2. Verify that f and g_i's are differentiable, and some constraint qualification holds.

Step 3. Set up the Lagrangian $L(x, \lambda) = f(x) + \sum_{i=1}^{I} \lambda_i g_i(x)$ and derive the KKT conditions (6.4).

Step 4. Solve for x and λ. If you get a unique x, it is the solution. If you get multiple x's, compute $f(x)$ for each and pick the minimum.

6.3.3 Constrained maximization

Next, we briefly discuss maximization. Although maximization is equivalent to minimization by flipping the sign of the objective function, doing so every time is awkward. So consider the maximization problem

$$\begin{array}{lll} \text{maximize} & f(x) & \\ \text{subject to} & g_i(x) \geq 0 & (i = 1, \dots, I), \end{array} \tag{6.5}$$

where f, g_i's are differentiable. (6.5) is equivalent to the minimization problem

$$\begin{array}{lll} \text{minimize} & -f(x) & \\ \text{subject to} & -g_i(x) \leq 0 & (i = 1, \dots, I). \end{array}$$

Assuming that Theorem 6.4 applies, the necessary condition for optimality is

$$-\nabla f(\bar{x}) - \sum_{i=1}^{I} \lambda_i \nabla g_i(\bar{x}) = 0, \tag{6.6a}$$

$$(\forall i)\ \lambda_i(-g_i(\bar{x})) = 0. \tag{6.6b}$$

But (6.6) is equivalent to (6.4) by multiplying everything by (-1). For this reason, it is customary to formulate a maximization problem as in (6.5) so that the inequality constraints are always "greater than or equal to zero".

Example 6.4. Consider a consumer with utility function

$$u(x) = \alpha \log x_1 + (1 - \alpha) \log x_2,$$

where $0 < \alpha < 1$ and $x_1, x_2 \geq 0$ are consumption of good 1 and 2 (such a function is called *Cobb-Douglas*). Let $p_1, p_2 > 0$ the price of each good and $w > 0$ be the wealth.

Then the consumer's utility maximization problem (UMP) is

$$\begin{array}{ll} \text{maximize} & \alpha \log x_1 + (1 - \alpha) \log x_2 \\ \text{subject to} & x_1 \geq 0, x_2 \geq 0, p_1 x_1 + p_2 x_2 \leq w. \end{array}$$

Step 1. Clearly the objective function is concave and the constraints are linear (hence concave). The point $(x_1, x_2) = (\epsilon, \epsilon)$ strictly satisfies the inequalities for small enough $\epsilon > 0$, so the Slater condition holds. Therefore we can apply Theorem 6.5.

Step 2. Let

$$L(x, \lambda, \mu) = \alpha \log x_1 + (1 - \alpha) \log x_2 + \lambda(w - p_1 x_1 - p_2 x_2) + \mu_1 x_1 + \mu_2 x_2,$$

where $\lambda \geq 0$ is the Lagrange multiplier corresponding to the budget constraint

$$p_1 x_1 + p_2 x_2 \leq w \iff w - p_1 x_1 - p_2 x_2 \geq 0$$

and μ_n is the Lagrange multiplier corresponding to $x_n \geq 0$ for $n = 1, 2$. By the first-order condition, we get

$$0 = \frac{\partial L}{\partial x_1} = \frac{\alpha}{x_1} - \lambda p_1 + \mu_1,$$

$$0 = \frac{\partial L}{\partial x_2} = \frac{1 - \alpha}{x_2} - \lambda p_2 + \mu_2.$$

By the complementary slackness condition, we have $\lambda(w-p_1x_1-p_2x_2) = 0$, $\mu_1 x_1 = 0$, $\mu_2 x_2 = 0$.

Step 3. Since $\log 0 = -\infty$, $x_1 = 0$ or $x_2 = 0$ cannot be an optimal solution. Hence $x_1, x_2 > 0$, so by complementary slackness we get $\mu_1 = \mu_2 = 0$. Then by the first order condition we get $x_1 = \frac{\alpha}{\lambda p_1}$, $x_2 = \frac{1-\alpha}{\lambda p_2}$, so $\lambda > 0$. Substituting these into the budget constraint $p_1x_1 + p_2x_2 = w$, we get

$$\frac{\alpha}{\lambda} + \frac{1-\alpha}{\lambda} = w \iff \lambda = \frac{1}{w},$$

so the solution is

$$(x_1, x_2) = \left(\frac{\alpha w}{p_1}, \frac{(1-\alpha)w}{p_2}\right).$$

Remark. In the above example, you notice that the constraints $x_1 \geq 0$ and $x_2 \geq 0$ never bind because the objective function is increasing in both arguments. (We can use an argument similar to Problem 3.6 to be more precise.) Therefore a quicker way to solve the problem is to ignore these constraints and set the Lagrangian

$$L(x_1, x_2, \lambda) = \alpha \log x_1 + (1-\alpha)\log x_2 + \lambda(w - p_1x_1 - p_2x_2).$$

Problems

6.1. This problem asks you to show that in general you cannot omit the constraint qualification. Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & x \\ \text{subject to} & x^2 \leq 0. \end{array}$$

1. Find the solution.
2. Show that both the objective function and the constraint function are convex, but the Slater constraint qualification does not hold.
3. Show that the Karush-Kuhn-Tucker condition does not hold.

6.2. Consider the problem

$$\begin{array}{ll} \text{minimize} & \dfrac{1}{x_1} + \dfrac{4}{x_2} \\ \text{subject to} & x_1 + x_2 \leq 3, \end{array}$$

where $x_1, x_2 > 0$.

1. Prove that the objective function is convex.
2. Write down the Lagrangian.
3. Explain why the Karush-Kuhn-Tucker conditions are both necessary and sufficient for a solution.

4. Compute the solution.

6.3. Consider the problem

$$\begin{aligned} &\text{maximize} && \frac{4}{3}x_1^3 + \frac{1}{3}x_2^3 \\ &\text{subject to} && x_1 + x_2 \leq 1, \\ & && x_1 \geq 0, x_2 \geq 0. \end{aligned}$$

1. Are the Karush-Kuhn-Tucker conditions necessary for a solution? Answer yes or no, then explain why.
2. Are the Karush-Kuhn-Tucker conditions sufficient for a solution? Answer yes or no, then explain why.
3. Write down the Lagrangian.
4. Compute the solution.

6.4. Let $f(x_1, x_2) = x_1^2 + x_1x_2 + x_2^2$.

1. Compute the gradient and the Hessian of f.
2. Show that f is convex.
3. Solve

$$\begin{aligned} &\text{minimize} && x_1^2 + x_1x_2 + x_2^2 \\ &\text{subject to} && x_1 + x_2 \geq 2. \end{aligned}$$

6.5. Consider the problem

$$\begin{aligned} &\text{maximize} && x_1 + \log x_2 - \frac{1}{2x_3^2} \\ &\text{subject to} && x_1 + p_2x_2 + p_3x_3 \leq w, \end{aligned}$$

where $x_2, x_3 > 0$ but x_1 is unconstrained and $p_2, p_3, w > 0$ are constants.

1. Show that the objective function is concave.
2. Write down the Lagrangian.
3. Find the solution.

6.6. Let A be an $N \times N$ symmetric positive definite matrix. Let B be an $M \times N$ matrix with $M \leq N$ such that the M row vectors of B are linearly independent. Let $c \in \mathbb{R}^M$ be a vector. Solve

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\langle x, Ax \rangle \\ &\text{subject to} && Bx = c. \end{aligned}$$

6.7. Solve

$$\begin{aligned} &\text{maximize} && x_1 + x_2 \\ &\text{subject to} && x_1^2 + x_1x_2 + x_2^2 \leq 1. \end{aligned}$$

6.8. Solve

$$\begin{array}{ll} \text{maximize} & \langle b, x\rangle \\ \text{subject to} & \langle x, Ax\rangle \leq r^2, \end{array}$$

where $0 \neq b \in \mathbb{R}^N$, A is an $N \times N$ symmetric positive definite matrix, and $r > 0$.

6.9. Consider a consumer with utility function

$$u(x) = \sum_{n=1}^{N} \alpha_n \log x_n,$$

where $\alpha_n > 0$, $\sum_{n=1}^{N} \alpha_n = 1$, and $x_n \geq 0$ is the consumption of good n. Let $p = (p_1, \ldots, p_N)$ be a price vector with $p_n > 0$ and $w > 0$ be the wealth.

1. Formulate the consumer's utility maximization problem.
2. Compute the solution.

6.10. Solve the same problem as above for the case

$$u(x) = \sum_{n=1}^{N} \alpha_n \frac{x_n^{1-\sigma}}{1-\sigma},$$

where $1 \neq \sigma > 0$.

Chapter 7

Introduction to Dynamic Programming

7.1 Introduction

So far, we have only considered the maximization or minimization of a given function, subject to some constraints. Such a problem is (sometimes) called a *static* optimization problem because there is only one decision to make, namely choosing the variables that optimize the objective function. In some cases, writing down or evaluating the objective function itself may be complicated. Furthermore, in many problems the decision maker makes multiple decisions over time instead of a single decision.

Dynamic programming (DP) is a mathematical programming (optimization) technique that exploits the sequential structure of the problem. It is easier to understand the logic by examples instead of the abstract formulation. Suppose that you want to minimize the function

$$f(x_1, x_2) = 2x_1^2 - 2x_1x_2 + x_2^2 - 2x_1 - 4x_2.$$

One way to solve this is to compute the gradient and set it equal to zero, so

$$\nabla f(x_1, x_2) = \begin{bmatrix} 4x_1 - 2x_2 - 2 \\ -2x_1 + 2x_2 - 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \iff \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}.$$

(This is only a necessary condition for optimality, but since the objective function is convex because the Hessian is positive definite, so it is also sufficient.)

Another way to solve this problem is in two steps. First, assume that we have already determined the value of x_1, so treat x_1 as a constant. Then the objective function is a (convex) quadratic function in x_2. Taking the partial derivative with respect to x_2 and setting it equal to zero, we get

$$\frac{\partial f}{\partial x_2} = -2x_1 + 2x_2 - 4 = 0 \iff x_2 = x_1 + 2.$$

Then the function value becomes

$$\begin{aligned} g(x_1) &\coloneqq f(x_1, x_1 + 2) \\ &= 2x_1^2 - 2x_1(x_1 + 2) + (x_1 + 2)^2 - 2x_1 - 4(x_1 + 2) \\ &= x_1^2 - 6x_1 - 4. \end{aligned}$$

Here $g(x)$ is the minimum value that we can attain if we choose x_2 optimally, given $x_1 = x$. Clearly we can solve the original problem by choosing x_1 so as to minimize g. Since g is a convex quadratic function, setting the derivative equal to zero, we get

$$g'(x_1) = 2x_1 - 6 = 0 \iff x = 3.$$

Therefore the solution is $(x_1, x_2) = (x_1, x_1 + 2) = (3, 5)$, as it should be.

Essentially, dynamic programming amounts to breaking a single optimization problem with many variables into multiple optimization problems with fewer variables. By doing so, the problem sometimes becomes easier to handle, especially when the problem is *stochastic* (probabilistic). In the above example, we have solved the single problem with two variables

$$\min_{x_1, x_2} f(x_1, x_2)$$

by breaking it into two problems with one variable each,

$$g(x_1) \coloneqq \min_{x_2} f(x_1, x_2) \text{ and } \min_{x_1} g(x_1).$$

7.2 Examples

We now discuss several concrete examples.

7.2.1 Knapsack problem

Suppose you are a thief who has broken into a jewelry store. You have a knapsack of size S (an integer) to pack what you have stolen. There are I types of jewelries indexed by $i = 1, 2, \ldots, I$, and a type i jewelry has integer size s_i and value v_i. You want to pack your knapsack so as to maximize the value of jewelries that you have stolen.

Formulating this problem as a constrained optimization problem is not particularly hard. Letting n_i be the number of type i jewelries that you pack, the total value is $\sum_{i=1}^{I} n_i v_i$ and the total size is $\sum_{i=1}^{I} n_i s_i$. Therefore the problem is equivalent to

$$\begin{aligned} &\text{maximize} && \sum_{i=1}^{I} n_i v_i \\ &\text{subject to} && \sum_{i=1}^{I} n_i s_i \le S, \\ & && n_i\text{: nonnegative integer.} \end{aligned}$$

One way to solve this problem is to use the theory on integer linear programming (which I do not discuss further).

Another way to solve is to use dynamic programming. Let $V(S)$ be the maximum value of jewelries that can be packed in a size S knapsack. (This is called a *value function*.) Clearly $V(S) = 0$ if $S < \min_i s_i$ since you cannot pack anything in this case. If you put anything at all in your knapsack (so $S \geq \min_i s_i$), clearly you start packing with some type of jewelry. If you put object i, then you get value v_i and you are left with remaining size $S - s_i$. By the definition of the value function, if you continue packing optimally, you get total value $V(S - s_i)$ from the remaining space. Therefore if you first pack object i, the maximum value that you can get is

$$v_i + V(S - s_i).$$

Since you want to pick the first object optimally, you want to maximize this value with respect to i, which will give you the total maximum value $V(S)$ of the original problem. Therefore

$$V(S) = \max_i [v_i + V(S - s_i)].$$

You can iterate this equation (called the *Bellman equation*) backward starting from $V(S) = 0$ for $S < \min_i s_i$ to find out the maximum value.

For example, let $I = 3$ (three types), $(s_1, s_2, s_3) = (1, 2, 5)$, and $(v_1, v_2, v_3) = (1, 3, 8)$. Then

$$\begin{aligned}
&V(0) = 0,\\
&V(1) = v_1 + V(0) = 1,\\
&V(2) = \max_i [v_i + V(2 - s_i)] = \max\{1 + V(1), 3 + V(0)\} = \max\{2, 3\} = 3,\\
&V(3) = \max_i [v_i + V(3 - s_i)] = \max\{1 + V(2), 3 + V(1)\} = \max\{4, 4\} = 4,\\
&V(4) = \max\{1 + V(3), 3 + V(2)\} = \max\{5, 6\} = 6,\\
&V(5) = \max\{1 + V(4), 3 + V(3), 8 + V(0)\} = \max\{7, 7, 8\} = 8,
\end{aligned}$$

and so on.

7.2.2 Shortest path problem

Suppose that there are locations indexed by $i = 1, \ldots, I$. Traveling directly from i to j costs $c_{ij} \geq 0$, with $c_{ii} = 0$. (If there is no direct route from i to j, simply define $c_{ij} = \infty$.) You want to find the cheapest way to travel from any point i to any other point j.

To solve this problem, let $V_N(i, j)$ be the minimum cost to travel from i to j in at most N steps. Let k be the first connection (including possibly $k = i$). Traveling from i to k costs c_{ik}, and now you need to travel from k to j in at most $N - 1$ steps. If you continue optimally, the cost from k to j is (by the definition of the value function) $V_{N-1}(k, j)$. Therefore the Bellman equation is

$$V_N(i, j) = \min_k \{c_{ik} + V_{N-1}(k, j)\}.$$

Since $0 \leq V_N(i, j) \leq V_{N-1}(i, j)$ (because $c_{ii} = 0$), the limit $\lim_{N\to\infty} V_N(i, j)$ exists.[1] Therefore the cheapest path can be found by iterating backwards from $V_1(i, j) = c_{ij}$.

[1] In fact, it converges in finite steps. This is because since you visit each point at most once, the number of connections is at most $I - 1$, so $V_N = V_{N-1}$ for $N \geq I$.

7.2.3 Optimal saving problem

Suppose that you live for $T+1$ years indexed by $t = 0, 1, \ldots, T$. You have initial wealth w_0. At each point in time, you can either consume some of your wealth or save it at gross interest rate $R > 0$. That is, if you save 1 dollar this year, it will grow to R dollars next year.

To solve this problem, let w_t be your wealth at the beginning of year t. If you consume c_t in year t, the next year's wealth will be $w_{t+1} = R(w_t - c_t)$. For concreteness, assume that the utility function is

$$U_T(c_0, \ldots, c_T) = \sum_{t=0}^{T} \beta^t \log c_t.$$

(The subscript T in U_T means that there are T years to go in the future.) Clearly we have

$$U_T(c_0, \ldots, c_T) = \log c_0 + \beta U_{T-1}(c_1, \ldots, c_T).$$

Let $V_T(w)$ be the maximum utility you get when you start with capital w and there are T years to go. If $T = 0$, you have no choice but to consume everything, so $V_0(w) = \log w$. If $T > 0$ and you consume c this year, by the budget equation you will have capital $w' = R(w - c)$ next year and there will be $T - 1$ years to go. Therefore the Bellman equation is

$$V_T(w) = \max_{0 \le c \le w} [\log c + \beta V_{T-1}(R(w - c))].$$

In principle you can compute $V_T(w)$ by iterating backwards from $T = 0$ using $V_0(w) = \log w$. Let us compute $V_1(w)$, for example. By the Bellman equation and $V_0(w) = \log w$, we have

$$\begin{aligned} V_1(w) &= \max_{0 \le c \le w} [\log c + \beta V_0(R(w - c))] \\ &= \max_{0 \le c \le w} [\log c + \beta \log(R(w - c))]. \end{aligned}$$

The right-hand side inside the brackets is concave in c, so we can maximize it by setting the derivative equal to zero. The first-order condition is

$$\frac{1}{c} + \beta \frac{-1}{w - c} = 0 \iff w - c = \beta c \iff c = \frac{w}{1 + \beta}.$$

Therefore the value function is

$$V_1(w) = \log \frac{w}{1 + \beta} + \beta \log \left(R \frac{\beta w}{1 + \beta} \right) = (1 + \beta) \log w + \text{constant},$$

Where "constant" is some constant that depends only on the given parameters β and R.

7.2.4 Drawing cards

Suppose there are equal numbers of black and red cards (say N each), and you draw one card at a time. You have the option to stop at any time. The score you get when you stop is

"number of black cards drawn" − "number of red cards drawn".

You want to maximize the expected score. What is the optimal strategy?

Let b, r be the number of black and red cards that remain in the stack. Then you have already drawn $N - b$ black cards and $N - r$ red cards, so your current score is $(N - b) - (N - r) = r - b$. If you stop, you get $r - b$. If you continue, on the next draw you draw a black card with probability $\frac{b}{b+r}$ (and b decreases by 1) and a red card with probability $\frac{r}{b+r}$ (and r decreases by 1). Let $V(b, r)$ be the expected score when b black cards and r red cards remain. Then the Bellman equation is

$$V(b,r) = \max\left\{ \underbrace{r-b}_{\text{stop}}, \underbrace{\frac{b}{b+r}V(b-1,r) + \frac{r}{b+r}V(b,r-1)}_{\text{continue}} \right\}.$$

You can find the optimal strategy by iterating backwards from $V(0, 0) = 0$.

7.2.5 Optimal proposal

Suppose you know you are going to meet N persons one at a time that you may want to marry. You can propose only once (possibly because your proposal will be accepted for sure and the cost of divorce is prohibitive). The value of your potential partner is independently distributed uniformly over the interval $0 \le v \le 1$. Having observed a candidate, you can either propose or wait to see the next candidate (but you cannot go back once forgone). You want to maximize the expected value of your marriage. What is the best strategy to propose?

Let $V_n(v)$ be the maximum expected value when faced with a candidate with value v and there are n candidates to go. Clearly $V_0(v) = v$. The Bellman equation is

$$V_n(v) = \max\{v, \mathrm{E}[V_{n-1}(v')]\} = \max\left\{v, \int_0^1 V_{n-1}(v')\,\mathrm{d}v'\right\},$$

where the expectation is taken with respect to v', the value of the next candidate.

In this case we can do more than writing down the Bellman equation. Since $\mathrm{E}[V_{n-1}(v')]$ is just a constant depending on n, say a_n, it follows that $V_n(v) = \max\{v, a_n\}$. Therefore the optimal strategy is to propose if $v \ge a_n$ and wait otherwise. Using the definition of a_n and the Bellman equation, it follows that

$$\begin{aligned} a_n &= \mathrm{E}[V_{n-1}(v')] = \int_0^1 V_{n-1}(v')\,\mathrm{d}v' \\ &= \int_0^{a_{n-1}} a_{n-1}\,\mathrm{d}v' + \int_{a_{n-1}}^1 v'\,\mathrm{d}v' \\ &= a_{n-1}^2 + \frac{1}{2}(1 - a_{n-1}^2) = \frac{1}{2}(1 + a_{n-1}^2). \end{aligned}$$

Starting from $a_0 = 0$, we can compute the threshold for proposing a_n.

7.3 General formulation

In general, we can formulate a dynamic programming problem as follows. At each stage, there are variables that define your current situation, called *state*

variables. Let x_n be the state variable when there are n stages to go. (The state variable may be a number, a vector, or whatever is relevant for decision making. The dimension of x_n may depend on n.) The state variable x_n determines your constraint set, denoted by $\Gamma_n(x_n)$. A feasible action is an element of the set $\Gamma_n(x_n)$, which is called a *control variable*. By choosing a control y_n, the next stage's state variable is determined by the *law of motion* $x_{n-1} = g_n(x_n, y_n)$. (Here I am indexing the state and control variables by the number of stages to go, so x's are counted backwards.)

A sequence of state variables $(x_n, \dots, x_0)$ and control variables $(y_n, \dots, y_0)$ are said to be *feasible* if they satisfy the constraint and the law of motion. That is, $y_k \in \Gamma_k(x_k)$ and $x_{k-1} = g_k(x_k, y_k)$ for all $k = 0, \dots, n$. Given a feasible sequence of state and control variables up to n stages from the last, there corresponds a value (real number) $U_n(\{(x_k, y_k)\}_{k=0}^n)$. ($U_n$ is a function that takes as argument all present and future state and control variables.) We want to maximize or minimize U_n depending on the context, but for concreteness assume that we want to maximize U_n and let us call it the utility function.

In order to apply dynamic programming, the utility function U_n must admit a special recursive structure. That is, today's utility must be a function of today's state and control and tomorrow's utility. Thus we require

$$U_n = f_n(x_n, y_n, U_{n-1}), \tag{7.1}$$

where the function f_n is called the *aggregator*, assumed to be continuous and increasing in the third argument. The supremum of the feasible utility,

$$V_n(x_n) = \sup \left\{ U_n(\{(x_k, y_k)\}_{k=0}^n) \,|\, (\forall k) y_k \in \Gamma_k(x_k), x_{k-1} = g_k(x_k, y_k) \right\}, \tag{7.2}$$

is called the *value function*. The following *principle of optimality* is extremely important.

Theorem 7.1 (Principle of Optimality)**.** *Suppose that the aggregator $f_n(x, y, v)$ is continuous and increasing in v. Then*

$$V_n(x_n) = \sup_{y_n \in \Gamma_n(x_n)} f_n(x_n, y_n, V_{n-1}(g_n(x_n, y_n))). \tag{7.3}$$

The relation (7.3) is called the *Bellman equation*.

Proof. For any feasible $\{(x_k, y_k)_{k=0}^n\}$, we have

$$\begin{aligned}
&U_n(\{(x_k, y_k)_{k=0}^n\}) \\
&= f_n(x_n, y_n, U_{n-1}(\{(x_k, y_k)_{k=0}^{n-1}\})) && (\because (7.1)) \\
&\le f_n(x_n, y_n, V_{n-1}(x_{n-1})) && (\because (7.2), f_n \text{ monotone}) \\
&= f_n(x_n, y_n, V_{n-1}(g_n(x_n, y_n))) && (\because x_{n-1} \text{ feasible}) \\
&\le \sup_{y_n \in \Gamma_n(x_n)} f_n(x_n, y_n, V_{n-1}(g_n(x_n, y_n))). && (\because y_n \text{ feasible})
\end{aligned}$$

Taking the supremum of the left-hand side over all feasible paths, we get

$$V_n(x_n) \le \sup_{y_n \in \Gamma_n(x_n)} f_n(x_n, y_n, V_{n-1}(g_n(x_n, y_n))).$$

To show the reverse inequality, pick any $y_n \in \Gamma_n(x_n)$ and let $x_{n-1} = g_n(x_n, y_n)$. By the definition of V_{n-1}, for any $v < V_{n-1}(x_{n-1})$ there exists a feasible sequence $\{(x_k, y_k)_{k=0}^{n-1}\}$ such that $v < U_{n-1}(\{(x_k, y_k)_{k=0}^{n-1}\})$. Therefore

$$\begin{aligned} V_n(x_n) &\geq U_n(\{(x_k, y_k)_{k=0}^{n}\}) && (\because (7.2)) \\ &= f_n(x_n, y_n, U_{n-1}(\{(x_k, y_k)_{k=0}^{n-1}\})) && (\because (7.1)) \\ &\geq f_n(x_n, y_n, v). && (\because f_n \text{ monotone}) \end{aligned}$$

Since f_n is continuous in v, letting $v \uparrow V_{n-1}(x_{n-1})$ we get

$$V_n(x_n) \geq f_n(x_n, y_n, V_{n-1}(x_{n-1})) = f_n(x_n, y_n, V_{n-1}(g_n(x_n, y_n))).$$

Taking the supremum of the right-hand side with respect to $y_n \in \Gamma_n(x_n)$, we get

$$V_n(x_n) \geq \sup_{y_n \in \Gamma_n(x_n)} f_n(x_n, y_n, V_{n-1}(g_n(x_n, y_n))). \quad \square$$

Remark. The power of dynamic programming is to break a single optimization problem with many variables into multiple optimization problems with fewer variables. Without dynamic programming, the evaluation of the objective function alone might be a nightmare. For example, in principle we get the utility function $U_n(\{(x_k, y_k)\}_{k=0}^n)$ by iterating (7.1) backwards, but this function may be extremely complicated.

Remark. In the above formulation, we implicitly assumed that the optimization problem is deterministic, but the stochastic case is similar. In the stochastic case, the number of control variables increases exponentially with the number of stages. (For example, flipping a coin n times has 2^n potential outcomes.) Then solving the optimization problem in one shot would be impossible when the number of stages is large. Dynamic programming would be the only practical way to solve the problem.

7.4 Solving dynamic programming problems

There are a few ways to solve dynamic programming problems.

7.4.1 Value function iteration

The most basic way to solve a dynamic programming problem is by *value function iteration*, also called *backward induction*. Under mild conditions, we know that the Bellman equation (7.3) holds. Starting from $n = 0$, which is merely

$$V_0(x_0) = \sup_{y_0 \in \Gamma_0(x_0)} U_0(x_0, y_0),$$

in principle we can compute $V_n(x_n)$ by iterating the Bellman equation (7.3) from backwards.

The knapsack problem, shortest path problem, and drawing cards problem can all be solved this way using a computer, which are left as an exercise.

7.4.2 Guess and verify

Sometimes we can guess the functional form of the value function from the structure of the problem. For example, in the optimal proposal problem, we know that the value function must be of the form

$$V_n(v) = \max\{v, a_n\}$$

for some constant a_n, with $a_0 = 0$. Then we derived a difference equation that a_n satisfies,

$$a_n = \frac{1}{2}(1 + a_{n-1}^2).$$

Thus the original problem of finding the value *function* $V_n(v)$ reduced to finding the *number* a_n.

The optimal saving problem can also be solved by guess and verify. We know that $V_0(w) = \log w$. We might guess in general that $V_T(w) = a_T + b_T \log w$, where a_T and b_T are some constants with $b_T > 0$. Assuming that this is correct, substituting into the Bellman equation we get

$$a_T + b_T \log w = \max_{0 \le c \le w} [\log c + \beta(a_{T-1} + b_{T-1} \log(R(w-c)))].$$

Taking the derivative of the expression inside the brackets with respect to c and setting it equal to zero, we get

$$\frac{1}{c} - \frac{\beta b_{T-1}}{w-c} = 0 \iff c = \frac{w}{1 + \beta b_{T-1}}.$$

Substituting this into the Bellman equation, we get

$$\begin{aligned} a_T + b_T \log w &= \log c + \beta(a_{T-1} + b_{T-1} \log(R(w-c))) \\ &= (1 + \beta b_{T-1}) \log w + \text{constant}. \end{aligned}$$

In order for this to be an identity, it must be $b_T = 1 + \beta b_{T-1}$, which is a first order linear difference equation (so can be solved). Since $b_0 = 1$, the general term is

$$b_T = 1 + \beta + \cdots + \beta^T = \frac{1 - \beta^{T+1}}{1 - \beta}.$$

There will also be a difference equation for a_T, which is unimportant because the value of a_T does not affect the behavior (it only affects the utility level). Therefore the optimal consumption is

$$c = \frac{w}{1 + \beta b_{T-1}} = \frac{w}{b_T} = \frac{1 - \beta}{1 - \beta^{T+1}} w$$

when there are T periods to go. This formula means that you should consume a fraction $\frac{1-\beta}{1-\beta^{T+1}}$ when there are T periods to go, independent of the interest rate.

Problems

7.1. What are the state variables of the knapsack problem, optimal saving problem, drawing cards, and optimal proposal? What are the control variables? What are the aggregators?

7.2. There are N types of coins. A coin of type n has integer value v_n. You want to find the minimum number of coins needed for the value of the coins to sum to S, where $S \geq 0$ is an integer.

1. What is (are) the state variable(s)?
2. Write down the Bellman equation.
3. Solve the problem for $S = 10$ when $N = 3$ and $(v_1, v_2, v_3) = (1, 2, 4)$.

7.3. Suppose you live for $1 + T$ years and your utility function is

$$\mathrm{E} \sum_{t=0}^{T} \beta^t u(c_t),$$

where c_t is consumption. At each time t you get a job offer (income) y drawn from some distribution. If you accept the job, you receive y each year for the rest of your life. If you reject the job, you get unemployment benefit b today and you can search for a job next period. Assume that you cannot save or borrow, so you spend all your income every period. Write down the Bellman equation.

7.4. You have a call option on a stock with strike price K and time to expiration T. This means that if you exercise the option at time $t \leq T$ when the stock price is S_t, you will get $S_t - K$ at t. If you don't exercise the option, you will get nothing. You want to exercise the option so as to maximize the expected discounted payoff

$$\mathrm{E}\left[\frac{1}{(1+r)^t}(S_t - K)\right],$$

where t is the exercise date and r is the interest rate. Assume that the gross return of the stock is

$$\frac{S_{t+1}}{S_t} = \begin{cases} 1 + \mu + \sigma, & \text{(with probability 1/2)} \\ 1 + \mu - \sigma, & \text{(with probability 1/2)} \end{cases}$$

where $\mu > 0$ is the expected return and $\sigma > \mu$ is the volatility.

1. What is (are) the state variable(s)?
2. Write down the Bellman equation that the option value satisfies.
3. Compute the option value when $T = 1$ and the current stock price is S, where $S < K < (1 + \mu + \sigma)S$.

7.5. Suppose you currently have a T year mortgage at fixed interest rate r, which you can keep or refinance once. Assume that there are J types of mortgage in the market and S states of the world. The term of mortgage j is T_j years and the interest rate is r_{js} in state s. The transition probability from state s to s' is $p_{ss'}$, so $\sum_{s'=1}^{S} p_{ss'} = 1$. Letting the mortgage payment at time t be m_t, your objective is to minimize the discounted expected payments

$$\mathrm{E} \sum_{t \geq 1} \beta^t m_t,$$

where $\beta > 0$ is your discount factor.

1. What are the state variables?

2. Write down the Bellman equation. (Hint: the objective function is linear in mortgage payments, so consider the value function *per dollar borrowed*.)

7.6. Set up concrete numbers for the knapsack problem, the shortest path problem, and the drawing cards problem. Solve the problems using your favorite programming language (Matlab, Python, etc.).

7.7. You are a potato farmer. You start with some stock of potatoes. At each time, you can eat some of them and plant the rest. If you plant x potatoes, you will harvest Ax^α potatoes at the beginning of the next period, where $A, \alpha > 0$. You want to maximize your utility from consuming potatoes

$$\sum_{t=0}^{T} \beta^t \log c_t,$$

where $0 < \beta < 1$ is the discount factor, $c_t > 0$ is consumption of potatoes at time t, and T is the number of periods you live.

1. If you have k potatoes now and consume c out of it, how many potatoes can you harvest next period?

2. Let $V_T(k)$ be the maximum utility you get when you start with k potatoes. Write down the Bellman equation.

3. Solve for the optimal consumption when $T = 1$.

4. Guess that $V_T(k) = a_T + b_T \log k$ for some constants a_T, b_T. Assuming that this guess is correct, derive a relation between b_T and b_{T-1}.

7.8. Consider the optimal saving problem with the utility function

$$\sum_{t=0}^{T} \beta^t \frac{c_t^{1-\gamma}}{1-\gamma},$$

where $\gamma > 0$ and $\gamma \neq 1$.

1. Write down the Bellman equation.

2. Show that the value function must be of the form $V_T(w) = a_T \frac{w^{1-\gamma}}{1-\gamma}$ for some $a_T > 0$ with $a_0 = 1$.

3. Take the first-order condition and express the optimal consumption as a function of a_{T-1}.

4. Substitute the optimal consumption into the Bellman equation and derive a relation between a_T and a_{T-1}.

5. Solve for a_T and the optimal consumption rule.

7.9. Consider the optimal saving problem with stochastic interest rates. Let R_s be the gross interest rate in state $s \in \{1, \dots, S\}$, and let $p_{ss'}$ be the probability of moving from state s to s'.

1. Write down the Bellman equation.

2. Show that the value function must be of the form $V_T(w, s) = a_{s,T}\frac{w^{1-\gamma}}{1-\gamma}$ for some $a_{s,T} > 0$ with $a_{s,0} = 1$.

3. By solving for the optimal consumption rule, derive a relation between $a_{s,T}$ and $\{a_{s',T-1}\}_{s'=1}^{S}$.

Part II

Advanced Topics

Chapter 8

Contraction Mapping Theorem and Applications

8.1 Contraction Mapping Theorem

In economics, we often apply fixed point theorems. Let X be a set and $T : X \to X$ a mapping that maps X into itself. (Such a mapping is called a *self map*.) Then $x \in X$ is called a *fixed point* if $T(x) = x$, i.e., the point x stays fixed by applying the mapping T.

One of the most useful (and easiest to prove) fixed point theorems is the *contraction mapping theorem*, also known as the *Banach fixed point theorem* (named after the Polish mathematician who proved it first).

Before stating the theorem, we need to introduce some definitions. Let X be a set. Then the function $d : X \times X \to \mathbb{R}$ is called a *metric* (or *distance*) if

1. (positivity) $d(x, y) \geq 0$ for all $x, y \in X$, and $d(x, y) = 0$ if and only if $x = y$,
2. (symmetry) $d(x, y) = d(y, x)$ for all $x, y \in X$, and
3. (triangle inequality) $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in X$.

The pair (X, d) is called a *metric space* if d is a metric on X. When the metric d is clear from the context, we often call X a metric space. A typical example is $X = \mathbb{R}^N$ and

$$d(x, y) = \sqrt{\sum_{n=1}^{N} (x_n - y_n)^2}$$

(the Euclidean distance), or more generally

$$d(x, y) = \left(\sum_{n=1}^{N} |x_n - y_n|^p \right)^{\frac{1}{p}}$$

for $p \geq 1$ (l^p distance). When $p = \infty$, it becomes $d(x, y) = \max_n |x_n - y_n|$ (sup norm).

Example 8.1. If X is a *normed space*, meaning that X is a vector space with a norm $\|\cdot\|$, then (X, d) becomes a metric space by setting $d(x, y) = \|x - y\|$.

Proof. Take any $x, y, z \in X$. We have $d(x, y) = \|x - y\| \geq 0$, with equality if and only if $x = y$. Furthermore, $d(x, y) = \|x - y\| = \|y - x\| = d(y, x)$. Finally, by the triangle inequality for the norm, we obtain

$$d(x, z) = \|x - z\| = \|x - y + y - z\| \leq \|x - y\| + \|y - z\| = d(x, y) + d(y, z). \quad \square$$

A metric space (X, d) is called *complete* if any Cauchy sequence is convergent, that is, $x = \lim_{n\to\infty} x_n$ exists whenever

$$(\forall \epsilon > 0)(\exists N > 0)(\forall m, n \geq N)\ d(x_m, x_n) < \epsilon.$$

A complete normed space is called a *Banach space*. We discuss a few examples.

Example 8.2. Let Ω be a set and $b\Omega$ be the set of all bounded functions from Ω to $\mathbb{R}$. Then $b\Omega$ is a Banach space with the sup norm $\|f\| = \sup_{x\in\Omega} |f(x)|$.

Proof. Let us first show that $\|\cdot\|$ is a norm. Since f is bounded for $f \in b\Omega$, $\|f\| = \sup_{x\in\Omega} |f(x)| < \infty$ is well-defined. Clearly $\|f\| = \sup_{x\in\Omega} |f(x)| \geq 0$, with equality if and only if $f = 0$. For any $\alpha \in \mathbb{R}$, we have

$$\|\alpha f\| = \sup_{x\in\Omega} |\alpha f(x)| = |\alpha| \sup_{x\in\Omega} |f(x)| = |\alpha| \, \|f\| .$$

Let $f, g \in b\Omega$. Then for any $x \in \Omega$ we have

$$|f(x) + g(x)| \leq |f(x)| + |g(x)| \leq \|f\| + \|g\| .$$

Taking the supremum of the left-hand side over $x \in \Omega$, we obtain $\|f + g\| \leq \|f\| + \|g\|$, so the triangle inequality holds. Hence $\|\cdot\|$ is a norm.

To show that $b\Omega$ is a Banach space, it suffices to show that $b\Omega$ is complete. Let $\{f_n\}_{n=1}^{\infty}$ be a Cauchy sequence in $b\Omega$. Then for all $\epsilon > 0$, there exists N such that $m, n \geq N$ implies $\|f_m - f_n\| < \epsilon$. Hence $|f_m(x) - f_n(x)| < \epsilon$ for all $m, n \geq N$ and $x \in \Omega$. Since for each $x \in \Omega$ the real sequence $\{f_n(x)\}_{n=1}^{\infty}$ is Cauchy and $\mathbb{R}$ is complete, there exists $f(x) = \lim_{n\to\infty} f_n(x)$. Letting $m \to \infty$ in $|f_m(x) - f_n(x)| < \epsilon$, we obtain $|f(x) - f_n(x)| \leq \epsilon$. Using the triangle inequality and noting that $f_n \in b\Omega$, we obtain $|f(x)| \leq |f_n(x)| + \epsilon \leq \|f_n\| + \epsilon < \infty$, so f is bounded with $\|f\| \leq \|f_n\| + \epsilon$. Therefore $f \in b\Omega$. Taking the supremum of $|f(x) - f_n(x)| \leq \epsilon$ over $x \in \Omega$, we obtain $\|f - f_n\| \leq \epsilon$. Therefore $f_n \to f$, and $b\Omega$ is a Banach space. $\square$

Example 8.3. Let Ω be a topological space and $bc\Omega$ be the set of all bounded continuous functions from Ω to $\mathbb{R}$. Then $bc\Omega$ is a Banach space with the sup norm $\|f\| = \sup_{x\in\Omega} |f(x)|$.

Proof. Since $bc\Omega \subset b\Omega$, by Example 8.2, $\|\cdot\|$ is a norm on $bc\Omega$.

To show completeness, let $\{f_n\}_{n=1}^{\infty}$ be a Cauchy sequence in $bc\Omega$. Then by Example 8.2, we have $f = \lim_{n\to\infty} f_n \in b\Omega$. Therefore to show that $bc\Omega$ is a Banach space, it suffices to show that f is continuous. Take any $\epsilon > 0$. Since $f_n \to f$ in $b\Omega$, we can take N such that $\|f - f_n\| < \epsilon/3$ for $n > N$. Fix such n

and take any $x \in \Omega$. Since f_n is continuous, we can take a neighborhood U of x such that $|f_n(y) - f_n(x)| < \epsilon/3$ for $y \in U$. Then

$$\begin{aligned} |f(y) - f(x)| &\le |f(y) - f_n(y)| + |f_n(y) - f_n(x)| + |f_n(x) - f(x)| \\ &\le \|f - f_n\| + \frac{\epsilon}{3} + \|f - f_n\| < \epsilon, \end{aligned}$$

so f is continuous. □

Let (X, d) be a metric space. A mapping $T : X \to X$ is called a *contraction mapping* (or simply a *contraction*) if there exists a constant $\beta \in [0, 1)$ such that

$$d(T(x), T(y)) \le \beta d(x, y) \tag{8.1}$$

for all $x, y \in X$. Intuitively, the condition (8.1) means that when we apply T, the distance between two points shrinks by a factor at most $\beta < 1$. The following contraction mapping theorem (also called the Banach fixed point theorem) is elementary but has many important applications.

Theorem 8.1 (Contraction Mapping Theorem)**.** *Let (X, d) be a complete metric space and $T : X \to X$ be a contraction, so (8.1) holds for all $x, y \in X$. Then*

1. *T has a unique fixed point $x \in X$,*
2. *for any $x_0 \in X$, we have $x = \lim_{n\to\infty} T^n(x_0)$, and*
3. *the approximation error $d(T^n(x_0), x)$ has order of magnitude β^n.*

Proof. First note that a contraction is continuous (indeed, uniformly continuous) because for any $\epsilon > 0$, if $d(x, y) < \epsilon$ then $d(T(x), T(y)) \le \beta d(x, y) \le \beta\epsilon \le \epsilon$.

Take any $x_0 \in X$ and define $x_n = T(x_{n-1})$ for $n \ge 1$. Then $x_n = T^n(x_0)$. Since T is a contraction, we have

$$d(x_n, x_{n-1}) = d(T(x_{n-1}), T(x_{n-2})) \le \beta d(x_{n-1}, x_{n-2}) \le \cdots \le \beta^{n-1} d(x_1, x_0).$$

If $m > n \ge N$, then by the triangle inequality we have

$$\begin{aligned} d(x_m, x_n) &\le d(x_m, x_{m-1}) + \cdots + d(x_{n+1}, x_n) \\ &\le (\beta^{m-1} + \cdots + \beta^n) d(x_1, x_0) \\ &= \frac{\beta^n - \beta^m}{1 - \beta} d(x_1, x_0) \le \frac{\beta^n}{1 - \beta} d(x_1, x_0) \le \frac{\beta^N}{1 - \beta} d(x_1, x_0). \end{aligned}$$

Since $0 \le \beta < 1$, β^N gets smaller as N gets large, so $\{x_n\}$ is a Cauchy sequence. Since X is complete, $x = \lim_{n\to\infty} x_n$ exists. Since

$$d(T(x_n), x_n) = d(x_{n+1}, x_n) \le \beta^n d(x_1, x_0),$$

letting $n \to \infty$ and using the continuity of T, we get $d(T(x), x) = 0$. Since d is a metric, we have $T(x) = x$, so x is a fixed point of T.

To show uniqueness, suppose that x, y are fixed points of T, so $T(x) = x$ and $T(y) = y$. Since T is a contraction, we have

$$0 \le d(x, y) = d(T(x), T(y)) \le \beta d(x, y) \implies (\beta - 1) d(x, y) \ge 0.$$

Since $\beta < 1$, it must be $d(x, y) = 0$ and hence $x = y$. Therefore the fixed point is unique.

Finally, let x be the fixed point of T, x_0 be any point, and $x_n = T^n(x_0)$. Then

$$d(x_n, x) = d(T(x_{n-1}), T(x)) \leq \beta d(x_{n-1}, x) \leq \cdots \leq \beta^n d(x_0, x),$$

so letting $n \to \infty$ we have $x_n \to x$, and the error has order of magnitude β^n. □

Sometimes, we need to work with mappings T such that T^k is a contraction for some $k \in \mathbb{N}$, although T itself may not be a contraction. The following theorem extends Theorem 8.1 to such cases. (See Problem 8.1 for an example.)

Theorem 8.2. *Let (X, d) be a complete metric space and $T : X \to X$ be such that T^k is a contraction for some $k \in \mathbb{N}$. Then T has a unique fixed point $x \in X$ and we have $x = \lim_{n\to\infty} T^n(x_0)$ for any $x_0 \in X$.*

Proof. By the contraction mapping theorem, T^k has a unique fixed point $x \in X$, so $T^k(x) = x$. Since

$$T(x) = T(T^k(x)) = T^{k+1}(x) = T^k(T(x)),$$

$T(x)$ is also a fixed point of T^k. Since the fixed point of T^k is unique, it must be $T(x) = x$, so x is a fixed point of T. If x, y are fixed points of T, then $x = T(x) = \cdots = T^k(x)$ and $y = T(y) = \cdots = T^k(y)$, so x, y are also fixed points of T^k. Since T^k is a contraction, it must be $x = y$. Therefore the fixed point of T is unique.

To show $T^n(x_0) \to x$ for any $x_0 \in X$, express any $n \in \mathbb{N}$ uniquely as $n = km_n + r_n$, where $m_n \in \mathbb{Z}_+$ and $r_n \in \{0, \ldots, k-1\}$. Then for each fixed $r \in \{0, \ldots, k-1\}$, applying Theorem 8.1 to the initial value $T^r(x_0)$ we have $x = \lim_{m\to\infty}(T^k)^m T^r(x_0) = \lim_{m\to\infty} T^{km+r}(x_0)$. Since $r \in \{0, \ldots, k-1\}$ is arbitrary, we obtain $T^n(x_0) \to x$. □

8.2 Blackwell's condition for contraction

It would be convenient if there is a sufficient condition for contraction that is easily verifiable. Blackwell (1965) provides one such example.

Let X be a set. We say that a binary relation $\leq$ on X is a *partial order* if

1. (reflexivity) $x \leq x$ for all $x \in X$,
2. (antisymmetry) if $x \leq y$ and $y \leq x$, then $x = y$,
3. (transitivity) if $x \leq y$ and $y \leq z$, then $x \leq z$.

A set with a partial order is called a *partially ordered set*, or *poset* for short. $X = \mathbb{R}^N$ is a partially ordered Banach space by letting $x \leq y$ whenever $x_n \leq y_n$ for all n.

Example 8.4. Let Ω be a topological space and $bc\Omega$ be the set of all continuous bounded functions from Ω to $\mathbb{R}$. Then $bc\Omega$ is a partially ordered Banach space if we define $f \leq g$ whenever $f(x) \leq g(x)$ for all $x \in \Omega$ and $\|f\| = \sup_{x\in\Omega} |f(x)|$.

The following theorem provides a sufficient condition for a contraction.

Theorem 8.3 (Blackwell, 1965)**.** *Let Ω be a set and $X \subset b\Omega$ be a subset of bounded functions from Ω to $\mathbb{R}$. Suppose that for all $f \in X$ and $c \in \mathbb{R}_+$, we have $f + c \in X$, and $T : X \to X$ satisfies*

1. *(monotonicity) $f \leq g$ implies $Tf \leq Tg$,*
2. *(discounting) there exists $\beta \in [0, 1)$ such that $T(f + c) \leq Tf + \beta c$ for all constant $c \geq 0$.*

Then T is a contraction.

Proof. Let $\|\cdot\|$ be the sup norm. Take any $f, g \in X$ and $x \in \Omega$. Since

$$f(x) = f(x) - g(x) + g(x) \leq g(x) + \|f - g\|,$$

we have $f \leq g + \|f - g\|$. Hence

$$\begin{aligned} Tf &\leq T(g + \|f - g\|) && (\because \text{monotonicity}) \\ &\leq Tg + \beta \|f - g\| && (\because \text{discounting for } c = \|f - g\|) \\ \implies Tf - Tg &\leq \beta \|f - g\|. \end{aligned}$$

Interchanging the role of f, g, we obtain $Tg - Tf \leq \beta \|f - g\|$. This shows that $|(Tf)(x) - (Tg)(x)| \leq \beta \|f - g\|$ for any $x \in \Omega$. Taking the supremum over x, we obtain $\|Tf - Tg\| \leq \beta \|f - g\|$, so T is a contraction. □

8.3 Markov chain and Perron's theorem

When a random variable is indexed by time, it is called a stochastic process. Let $\{X_t\}$ be a stochastic process, where $t = 0, 1, 2 \ldots$. When the distribution of X_t conditional on the past information $X_{t-1}, X_{t-2}, \ldots$ depends only on the most recent past (i.e., X_{t-1}), $\{X_t\}$ is called a *Markov* process. For example, an AR(1) process

$$X_t = aX_{t-1} + \epsilon_t$$

(where a is a number and ϵ_t is independent and identically distributed over time) is a Markov process. When the Markov process $\{X_t\}$ takes on finitely many values, it is called a *finite-state Markov chain*. Let $\{X_t\}$ be a (finite-state) Markov chain and $n = 1, 2, \ldots, N$ be an index of the values the process can take. (We say that $X_t = x_n$ when the state at t is n.) Since there are finitely many states, the distribution of X_t conditional on X_{t-1} is just a multinomial distribution. Therefore the Markov chain is completely characterized by the *transition probability (stochastic) matrix* $P = (p_{nn'})$, where $p_{nn'}$ is the probability of moving from state n to n'. Clearly, we have $p_{nn'} \geq 0$ and $\sum_{n'=1}^{N} p_{nn'} = 1$.

Suppose that X_0 is distributed according to the distribution $\mu = [\mu_1, \ldots, \mu_N]$ (μ_n is the probability of being in state n). Then what is the distribution of X_1? Using transition probabilities, the probability of being in state n' at $t = 1$ is $\sum_{n=1}^{N} \mu_n p_{nn'}$, because the process must be in some state (say n) at $t = 0$ (which happens with probability μ_n) and conditional on being in state n at $t = 0$, the probability of moving to state n' at $t = 1$ is $p_{nn'}$. By the definition of matrix multiplication,

$$\sum_{n=1}^{N} \mu_n p_{nn'} = (\mu P)_{n'}$$

is the n'-th element of the row vector μP. Therefore μP is the distribution of X_1. Similarly, the distribution of X_2 is $(\mu P)P = \mu P^2$, and in general the distribution of X_t is μP^t.

As we let the system run for a long time, does the distribution settle down to some fixed distribution? That is, does $\lim_{t\to\infty} \mu P^t$ exist, and if so, is it unique? We can answer this question by using the contraction mapping theorem.

Theorem 8.4. *Let $P = (p_{nn'})$ be a stochastic matrix such that $p_{nn'} > 0$ for all n, n'. Then there exists a unique invariant distribution π such that $\pi = \pi P$, and $\lim_{t\to\infty} \mu P^t = \pi$ for all initial distribution μ.*

Proof. Let $\Delta = \left\{ x \in \mathbb{R}_+^N \,\middle|\, \sum_{n=1}^N x_n = 1 \right\}$ be the set of all multinomial distributions. Since $\Delta \subset \mathbb{R}^N$ is closed and $\mathbb{R}^N$ is a complete metric space with the L^1 norm (that is, $d(x,y) = \|x - y\|$ for $\|x\| = \sum_{n=1}^N |x_n|$), Δ is also a complete metric space.

Define $T : \Delta \to \Delta$ by $T(x) = xP$. To show that $T(x) \in \Delta$, note that if $x \in \Delta$, since $p_{nn'} \geq 0$ for all n, n', we have $xP \geq 0$, and since $\sum_{n'=1}^N p_{nn'} = 1$, we have

$$\sum_{n'=1}^N (xP)_{n'} = \sum_{n'=1}^N \sum_{n=1}^N x_n p_{nn'} = \sum_{n=1}^N x_n \sum_{n'=1}^N p_{nn'} = \sum_{n=1}^N x_n = 1.$$

Therefore $T(x) = xP \in \Delta$.

Next, let us show that T is a contraction mapping. Since $p_{nn'} > 0$ and the number of states is finite, there exists $\epsilon > 0$ such that $p_{nn'} > \epsilon$ for all n, n'. Without loss of generality, we may assume $N\epsilon < 1$. Let $q_{nn'} = \frac{p_{nn'} - \epsilon}{1 - N\epsilon} > 0$ and $Q = (q_{nn'})$. Since $\sum_{n'} p_{nn'} = 1$, we obtain $\sum_{n'} q_{nn'} = 1$, so Q is also a stochastic matrix. Letting J be the matrix with all elements equal to 1, we have $P = (1 - N\epsilon)Q + \epsilon J$.

Now let $\mu, \nu \in \Delta$. Then

$$\mu P - \nu P = (1 - N\epsilon)(\mu Q - \nu Q) + \epsilon(\mu J - \nu J).$$

Since all elements of J are 1 and the vectors μ, ν sum up to 1, $\mu J, \nu J$ are both vectors of ones. Therefore $\mu J = \nu J$, so letting $0 < \beta = 1 - N\epsilon < 1$, we get

$$\begin{aligned}
\|T(\mu) - T(\nu)\| &= \|\mu P - \nu P\| = \beta \|\mu Q - \nu Q\| \\
&= \beta \sum_{n'=1}^N |(\mu Q)_{n'} - (\nu Q)_{n'}| = \beta \sum_{n'=1}^N \left| \sum_{n=1}^N (\mu_n - \nu_n) q_{nn'} \right| \\
&\leq \beta \sum_{n'=1}^N \sum_{n=1}^N |\mu_n - \nu_n| \, q_{nn'} = \beta \sum_{n=1}^N |\mu_n - \nu_n| \sum_{n'=1}^N q_{nn'} \\
&= \beta \sum_{n=1}^N |\mu_n - \nu_n| = \beta \|\mu - \nu\| .
\end{aligned}$$

Therefore T is a contraction. By the contraction mapping theorem, there exists a unique $\pi \in \Delta$ such that $\pi P = \pi$, and $\lim_{t\to\infty} \mu P^t = \pi$ for all $\mu \in \Delta$. □

Remark. The same conclusion holds if there exists a number k such that P^k is a positive matrix. Just apply Theorem 8.2.

We can prove Perron's theorem (Theorem 1.6) using Theorem 8.4.

Proof of Theorem 1.6. Let $\alpha = \rho(A)$ be the spectral radius of A. Let us first show Parts 1 and 2. Since A is positive, we can take $d > 0$ such that $A \geq dI$. Then $\alpha = \rho(A) \geq \rho(dI) = d > 0$ by Problem 1.15. Let λ be an eigenvalue of A with $|\lambda| = \alpha > 0$ and $u \neq 0$ be a corresponding eigenvector. Let $v = (|u|_1, \dots, |u|_N)'$ be the vector of absolute values. Since $Au = \lambda u$, taking the absolute value of each entry and noting that A is positive, we obtain

$$\alpha |u_m| = \left| \sum_{n=1}^N a_{mn} u_n \right| \leq \sum_{n=1}^N a_{mn} |u_n| \iff \alpha v \leq Av.$$

Let us show that $Av = \alpha v$. Suppose to the contrary that $Av > \alpha v$. Then $Av - \alpha v > 0$, so multiplying A from the left and noting that A is positive, we obtain

$$A(Av - \alpha v) \gg 0 \iff A^2 v \gg \alpha A v.$$

Since A is finite dimensional, we can take $\epsilon > 0$ such that $A^2 v \geq (1+\epsilon)\alpha A v$. Multiplying both sides from left by A^{k-1}, we obtain

$$A^{k+1} v \geq (1+\epsilon)\alpha A^k v \geq \cdots \geq [(1+\epsilon)\alpha]^k A v.$$

Taking the norm of both sides, we obtain

$$\left\|A^k\right\| \|Av\| \geq \left\|A^{k+1} v\right\| \geq [(1+\epsilon)\alpha]^k \|Av\| \implies \left\|A^k\right\|^{1/k} \geq (1+\epsilon)\alpha.$$

Letting $k \to \infty$, by the Gelfand spectral radius formula (Proposition 1.5), we obtain $\alpha \geq (1+\epsilon)\alpha$, which is a contradiction since $\alpha > 0$. Therefore $Av = \alpha v$. Since $v > 0$, we have $Av \gg 0$, so $v = \frac{1}{\alpha} Av \gg 0$.

To show Part 3, let x be a right Perron vector of A, so $Ax = \alpha x$. Suppose there exists a (complex) vector u such that $Au = \alpha u$. Since A, α are both real, by taking the real and imaginary parts, $v = \operatorname{Re} u, \operatorname{Im} u$ both satisfy $Av = \alpha v$. If v is not collinear with x, then $v \neq 0$. Without loss of generality, we may assume v has a positive entry, so (since $x \gg 0$) we can take $c > 0$ such that $x - cv > 0$ and at least one entry of $x - cv$ is zero. But then

$$0 \ll A(x - cv) = \alpha(x - cv),$$

a contradiction. Therefore $v = \operatorname{Re} u, \operatorname{Im} u$ are both collinear with x, and so is u. Hence the eigenvalue $\alpha = \rho(A)$ is geometrically simple.

To show Part 4, let $x, y \gg 0$ be the (unique) right and left Perron vectors of A. Then for each m we have $\sum_{n=1}^N a_{mn} x_n = \alpha x_m$. Define the diagonal matrix $D = \operatorname{diag}(x_1, \dots, x_N)$, which is regular. Let $P = \frac{1}{\alpha} D^{-1} A D$. Comparing the (m, n) entry, we obtain $p_{mn} = \frac{a_{mn} x_n}{\alpha x_m}$, so P is positive and

$$\sum_{n=1}^N p_{mn} = \sum_{n=1}^N \frac{a_{mn} x_n}{\alpha x_m} = 1.$$

Thus P is a positive stochastic matrix. By Theorem 8.4, there exists a unique vector $\theta \gg 0$ with $\sum_{n=1}^N \theta_n = 1$ such that

$$\theta' P = \theta' \iff \theta' \frac{1}{\alpha} D^{-1} A D = \theta' \iff \theta' D^{-1} A = \alpha \theta' D^{-1}.$$

This shows that $\theta' D^{-1}$ is a positive left eigenvector of A corresponding to the eigenvalue $\alpha = \rho(A)$, so it must be $\theta' D^{-1} = y'$ (up to a multiplicative constant). Multiplying x from the right and noting that $\sum_{n=1}^{N} \theta_n = 1$ and $D = \operatorname{diag}(x_1, \ldots, x_N)$, it must be $y'x = 1$. Setting μ to be the unit vectors $e_1, \ldots, e_N$ in Theorem 8.4 and letting 1 be the vector of ones, we obtain

$$1\theta' = \lim_{k\to\infty} P^k = \lim_{k\to\infty} D^{-1} \left[\frac{1}{\alpha} A\right]^k D$$

$$\iff \lim_{k\to\infty} \left[\frac{1}{\rho(A)} A\right]^k = D1\theta' D^{-1} = xy'. \qquad \square$$

8.4 Implicit function theorem

When solving economic models, we often encounter equations like $f(x, y) = 0$, where y is an endogenous variable and x is an exogenous variable. Oftentimes y does not have an explicit expression, but nevertheless we might be interested in $\mathrm{d}y/\,\mathrm{d}x$. The implicit function theorem lets you compute this derivative as follows. Let $y = g(x)$. Then $f(x, g(x)) = 0$. Differentiating both sides with respect to x and using the chain rule, we get $f_x + f_y g'(x) = 0$, so $g'(x) = -f_x/f_y$. (f_x is the shorthand for $\partial f/\partial x$.)

The same argument holds for multi dimensions. If x is M-dimensional, y is N-dimensional, and $f : \mathbb{R}^M \times \mathbb{R}^N \to \mathbb{R}^N$, then differentiating both sides of $f(x, g(x)) = 0$ we get $D_x f + D_y f D_x g = 0$, so $D_x g = -[D_y f]^{-1} D_x f$. ($D_x f$ is the Jacobian (matrix of partial derivatives) of f with respect to x: the (m, n) element of $D_x f$ is $\partial f_m/\partial x_n$.) The goal of this section is to prove the following implicit function theorem.

Theorem 8.5 (Implicit Function Theorem)**.** *Let $f : \mathbb{R}^M \times \mathbb{R}^N \to \mathbb{R}^N$ be a continuously differentiable function. If $f(x_0, y_0) = 0$ and $D_y f(x_0, y_0)$ is regular, then there exist neighborhoods U of x_0 and V of y_0 and a function $g : U \to V$ such that*

1. *for all $x \in U$, $f(x, y) = 0 \iff y = g(x)$,*
2. *g is continuously differentiable, and*
3. *$D_x g(x) = -[D_y f(x, y)]^{-1} D_x f(x, y)$, where $y = g(x)$.*

The last claim of the implicit function theorem follows from the first two and the chain rule. The first two claims follow from the inverse function theorem:

Theorem 8.6 (Inverse Function Theorem)**.** *Let $f : \mathbb{R}^N \to \mathbb{R}^N$ be a continuously differentiable function. If $Df(x_0)$ is regular, then there exists a neighborhood V of $y_0 = f(x_0)$ such that*

1. *$f : U = f^{-1}(V) \to V$ is bijective (one-to-one and onto),*
2. *$g = f^{-1}$ is continuously differentiable, and*
3. *$Dg(y) = [Df(g(y))]^{-1}$ on V.*

Proof of Implicit Function Theorem. Define $F : \mathbb{R}^{M+N} \to \mathbb{R}^{M+N}$ by

$$F(x, y) = \begin{bmatrix} x \\ f(x, y) \end{bmatrix}.$$

Then F is continuously differentiable. Furthermore, since

$$DF(x, y) = \begin{bmatrix} I_M & O_{M,N} \\ D_x f(x, y) & D_y f(x, y) \end{bmatrix},$$

we have $\det DF(x_0, y_0) = \det D_y f(x_0, y_0) \neq 0$, so $DF(x_0, y_0)$ is regular. Since $F(x_0, y_0) = (x_0, 0)$, by the inverse function theorem there exists a neighborhood V of $(x_0, 0)$ such that $F : F^{-1}(V) \to V$ is bijective. Let G be the inverse function of F. Then for any $(z, w) \in V$, we have $F(x, y) = (z, w) \iff (x, y) = G(z, w) = (G_1(z, w), G_2(z, w))$. Since by definition $F(x, y) = (x, f(x, y))$, we have $x = z$, so $f(x, y) = w \iff y = G_2(x, w)$. Letting $w = 0$, we have $f(x, y) = 0 \iff y = g(x) \coloneqq G_2(x, 0)$. Since G is continuously differentiable, so is g. Therefore the implicit function theorem is proved. □

Next, we prove the inverse function theorem. If f is linear, so $f(x) = y_0 + A(x - x_0)$ for some matrix A, then we can find the inverse function by solving $y = y_0 + A(x - x_0) \iff x = x_0 + A^{-1}(y - y_0)$ (provided that A is regular). The idea to prove the nonlinear case is to linearize f around x_0.

Since f is differentiable, we have $y = f(x) \approx f(x_0) + Df(x_0)(x - x_0)$. Solving this equation for x, we obtain

$$x \approx x_0 + Df(x_0)^{-1}(y - f(x_0)).$$

This equation shows that given an approximate solution x_0 of $f(x) = y$, the solution can be approximated further by $x_0 + Df(x_0)^{-1}(y - f(x_0))$ (which is essentially the Newton-Raphson algorithm for solving the nonlinear equation $f(x) = y$). This intuition is helpful for understanding the proof below.

To prove the inverse function theorem, we need the following result.

Proposition 8.7 (Mean value inequality)**.** *Let $f : \mathbb{R}^N \to \mathbb{R}^M$ be differentiable and $\|\cdot\|$ be the Euclidean norm (as well as the operator norm induced by $\|\cdot\|$). Then*

$$\|f(x_2) - f(x_1)\| \leq \sup_{t \in [0,1]} \|Df(x_1 + t(x_2 - x_1))\| \, \|x_2 - x_1\|.$$

Proof. The claim is trivial if $f(x_1) = f(x_2)$, so assume $f(x_1) \neq f(x_2)$. Take any $d \in \mathbb{R}^M$ with $\|d\| = 1$. Define $\phi : [0, 1] \to \mathbb{R}$ by

$$\phi(t) = d'(f(x_1 + t(x_2 - x_1)) - f(x_1)).$$

Then $\phi(0) = 0$ and $\phi(1) = d'(f(x_2) - f(x_1))$. By the mean value theorem, there exists $t \in [0, 1]$ such that

$$d'(f(x_2) - f(x_1)) = \phi(1) - \phi(0) = \phi'(t) = d' Df(x_1 + t(x_2 - x_1))(x_2 - x_1),$$

where the last equality follows from the chain rule. Taking the norm of both sides, we obtain

$$\begin{aligned} |d'(f(x_2) - f(x_1))| &\leq \|d' Df(x_1 + t(x_2 - x_1))(x_2 - x_1)\| \\ &\leq \|Df(x_1 + t(x_2 - x_1))\| \, \|x_2 - x_1\| \end{aligned}$$

because $\|d\| = 1$. Taking the supremum over $t \in [0, 1]$ and setting $d = (f(x_2) - f(x_1))/\|f(x_2) - f(x_1)\|$, we obtain the desired inequality. □

Proof of Inverse Function Theorem. Fix $y \in \mathbb{R}^N$ and define $T : \mathbb{R}^N \to \mathbb{R}^N$ by

$$T(x) = x + Df(x_0)^{-1}(y - f(x)),$$

which is well-defined because $Df(x_0)$ is regular. For notational simplicity, let $A = Df(x_0)^{-1}$. Applying the mean value inequality to $T(x)$, for any x_1, x_2 we have

$$\|T(x_2) - T(x_1)\| \leq \sup_{t\in[0,1]} \|I - ADf(x(t))\| \, \|x_2 - x_1\| ,$$

where $x(t) = x_1 + t(x_2 - x_1)$. Since f is continuously differentiable and $A = Df(x_0)^{-1}$, we can take $\epsilon > 0$ such that $\|I - ADf(x)\| \leq 1/2$ whenever $\|x - x_0\| \leq \epsilon$. Note that since y cancel out in $T(x_2) - T(x_1)$, so ϵ does not depend on y. Let $B = \{x \,|\, \|x - x_0\| \leq \epsilon\}$ be the closed ball with center x_0 and radius ϵ. If $x_1, x_2 \in B$, then

$$\|x(t) - x_0\| \leq (1 - t) \|x_1 - x_0\| + t \|x_2 - x_0\| \leq \epsilon,$$

so $x(t) \in B$. Then by assumption $\|T(x_2) - T(x_1)\| \leq \frac{1}{2} \|x_2 - x_1\|$ on B. Let us show that $T(B) \subset B$ if y is sufficiently close to $y_0 = f(x_0)$. To see this, note that

$$T(x) - x_0 = x - x_0 + A(f(x_0) - f(x)) + A(y - y_0).$$

Using the mean value inequality to $h(x) = x - Af(x)$, we obtain

$$\|T(x) - x_0\| \leq \sup_{t\in[0,1]} \|I - ADf(x_0 + t(x - x_0))\| \, \|x - x_0\| + \|A(y - y_0)\| .$$

Take a neighborhood V of y_0 such that $\|A(y - y_0)\| \leq \frac{1}{2}\epsilon$ for all $y \in V$. If $y \in V$ and $x \in B$, then we have

$$\|T(x) - x_0\| \leq \frac{1}{2} \|x - x_0\| + \frac{1}{2}\epsilon \leq \epsilon,$$

so $T(x) \in B$. This shows that $T(B) \subset B$.

Since $T : B \to B$, B is closed, and T is a contraction on B, there exists a unique fixed point x of T. Since

$$x = T(x) = x + A(y - f(x)),$$

we have $y = f(x)$.

Let the unique $x \in B$ such that $y = f(x)$ for any $y \in V$ be denoted by $x = g(y)$. Since $f(x) = y$, we have $f(g(y)) = y$. First let us show that g is continuous on V. Suppose $y_n \to y$ and $x_n = g(y_n)$ but $x_n \not\to x$. Since B is compact, by taking a subsequence if necessary, we may assume that $x_n \to x'$, where $x \neq x' \in B$. Since $f(x_n) = f(g(y_n)) = y_n$, letting $n \to \infty$, since f is continuous we get $f(x') = y$, which is a contradiction because $f(x) = y$ has a unique solution on B. Therefore g is continuous.

To show the differentiability of g, for small enough $h \in \mathbb{R}^N$, let $k(h) \coloneqq g(y+h)-g(y)$. Since g is continuous, so is k. Noting that $g(y+h) = g(y)+k(h) = x+k$ and f is differentiable, we obtain

$$y + h = f(g(y + h)) = f(x + k) = f(x) + Dfk + o(k).$$

Since $y = f(x)$, we get $h = Dfk + o(k)$, so h and k have the same order of magnitude. Finally,

$$g(y+h) = g(y) + k = g(y) + [Df]^{-1}(h - o(k)) = g(y) + [Df]^{-1}h + o(h),$$

so g is differentiable and $Dg(y) = [Df(g(y))]^{-1}$. □

As an application of the implicit function theorem, let us study how an investor's asset allocation is related to wealth. Consider an agent with von Neumann-Morgenstern utility function u and initial wealth $w > 0$. Suppose that there are two assets, one risky (stock) with gross return $R > 0$ and the other risk-free (bond) with gross risk-free rate $R_f > 0$. If the investor invests x in the risky asset, the total wealth after investment is

$$R(x) \coloneqq Rx + R_f(w - x).$$

Therefore the utility maximization problem is

$$\operatorname*{maximize}_x \mathrm{E}[u(R(x))],$$

where E denotes the expectation. Suppose that $u' > 0$ and $u'' < 0$, so utility is strictly increasing and concave.

The following lemma shows that if the expected excess return of the risky asset is positive, then the investor always holds a positive amount of stocks.

Lemma 8.8. *Suppose that* $\mathrm{E}[R] > R_f$ *and a solution* x *to the utility maximization problem exists. Then* $x > 0$.

Proof. Let $f(x) = \mathrm{E}[u(R(x))]$ be the expected utility. Then by the chain rule we have

$$f'(x) = \mathrm{E}[u'(R(x))(R - R_f)],$$
$$f''(x) = \mathrm{E}[u''(R(x))(R - R_f)^2].$$

Since $u'' < 0$, and $R \neq R_f$ with positive probability because $\mathrm{E}[R] > R_f$, it follows that $f''(x) < 0$. Therefore f is strictly concave.

If x solves the utility maximization problem, by the first-order condition we have $f'(x) = 0$. Since

$$f'(0) = \mathrm{E}[u'(R_f w)(R - R_f)] = u'(R_f w)(\mathrm{E}[R] - R_f) > 0$$

because $u' > 0$ and $\mathrm{E}[R] > R_f$ and f' is strictly decreasing because $f'' < 0$, $f'(x) = 0 < f'(0)$ implies $x > 0$. □

A measure of risk aversion known as the *absolute risk aversion* coefficient at wealth w is defined by the quantity $\alpha(w) = -u''(w)/u'(w)$. The following proposition shows that if an investor's absolute risk aversion is decreasing, then the investor holds more stocks as he gets richer. The proof is an application of the implicit function theorem.

Proposition 8.9. *Suppose that* $\alpha(w) = -u''(w)/u'(w)$ *is decreasing,* $\mathrm{E}[R] > R_f$*, and let* $x > 0$ *be the optimal stock holdings. Then* $\partial x/\partial w \geq 0$.

Proof. By the first-order condition, we have

$$F(w, x) \coloneqq \mathrm{E}[u'(R(x))(R - R_f)] = 0.$$

Assuming that the implicit function theorem is applicable, we have $\partial x/\partial w = -(\partial F/\partial w)/(\partial F/\partial x)$. The denominator is

$$\frac{\partial F}{\partial x} = \mathrm{E}[u''(R(x))(R - R_f)^2] < 0,$$

so we can apply the implicit function theorem. Using the definition of α, the numerator is

$$\frac{\partial F}{\partial w} = \mathrm{E}[u''(R(x))(R - R_f)R_f] = -\mathrm{E}[\alpha(R(x))u'(R(x))(R - R_f)R_f].$$

If $R \geq R_f$, since $x > 0$ we get $R(x) = Rx + R_f(w - x) = R_f w + (R - R_f)x \geq R_f w$. Since α is decreasing, we get $\alpha(R(x)) \leq \alpha(R_f w)$. Multiplying both sides by $R - R_f \geq 0$, we obtain

$$\alpha(R(x))(R - R_f) \leq \alpha(R_f w)(R - R_f).$$

If $R \leq R_f$, by a similar argument $R(x) \leq R_f$ and $\alpha(R(x)) \geq \alpha(R_f w)$, so again

$$\alpha(R(x))(R - R_f) \leq \alpha(R_f w)(R - R_f).$$

Since $u' > 0$, we obtain

$$\begin{aligned}\mathrm{E}[\alpha(R(x))u'(R(x))(R - R_f)R_f] &\leq \mathrm{E}[\alpha(R_f w)u'(R(x))(R - R_f)R_f] \\ &= \alpha(R_f w)R_f \,\mathrm{E}[u'(R(x))(R - R_f)] = 0\end{aligned}$$

by the first order condition, so $\partial F/\partial w \geq 0$. Hence by the implicit function theorem $\partial x/\partial w = -(\partial F/\partial w)/(\partial F/\partial x) \geq 0$. □

Problems

8.1. Consider the integral equation

$$f(x) = \lambda \int_a^x K(x, y)f(y)\,\mathrm{d}y + \phi(x), \tag{8.2}$$

where $\phi : [a, b] \to \mathbb{R}$ and $K : [a, b]^2 \to \mathbb{R}$ are given continuous functions and $\lambda \in \mathbb{R}$.

1. Let X be the space of continuous functions $f : [a, b] \to \mathbb{R}$ equipped with the supremum norm $\|\cdot\|$. For each $x \in [a, b]$, define

   $$(Tf)(x) = \lambda \int_a^x K(x, y)f(y)\,\mathrm{d}y + \phi(x).$$

   Show that $T : X \to X$.

2. Show that

   $$|(Tf)(x) - (Tg)(x)| \leq |\lambda|\, M(x - a)\, \|f - g\|,$$

   where $M \coloneqq \sup_{(x,y)\in[a,b]^2} |K(x, y)|$.

3. Show that there exists a unique solution f to (8.2) by showing that T^k is a contraction for some $k \in \mathbb{N}$.

8.2. Let A be a square nonnegative matrix and suppose that there exists $k \in \mathbb{N}$ such that A^k is positive. Show that the conclusions of Theorem 1.6 remain true. (Hint: use Theorem 8.2.)

8.3. Consider the function $F : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$F(x_1, x_2, y_1, y_2) = \begin{bmatrix} x_1^2 - x_2^2 - y_1^3 + y_2^2 + 4 \\ 2x_1x_2 + x_2^2 - 2y_1^2 + 3y_2^4 + 8 \end{bmatrix}.$$

1. Compute the Jacobian $DF(x_1, x_2, y_1, y_2)$.
2. Show $F(2, -1, 2, 1) = (0, 0)$.
3. Let $x = (x_1, x_2)$ and $y = (y_1, y_2)$. Show that if x is sufficiently close to $(2, -1)$, then there exists a function $G(x)$ such that $F(x, y) = 0 \iff y = G(x)$.

8.4. For a utility function u satisfying $u' > 0$ and $u'' < 0$, the quantity

$$\gamma(w) = -\frac{wu''(w)}{u'(w)}$$

is called the *relative risk aversion* at wealth w. Consider an optimal portfolio problem, where an investor chooses the optimal fraction of wealth invested in stocks. If an investor has initial wealth w and invests fraction θ in stocks, then the final wealth becomes $R(\theta)w$, where $R(\theta) \coloneqq R\theta + R_f(1 - \theta)$ and R, R_f are the gross returns on stocks and bonds.

Prove that if an investor has decreasing relative risk aversion (so $\gamma(w)$ is decreasing in w), then a rational investor that solves

$$\underset{\theta}{\text{maximize}}\, \mathrm{E}[u(R(\theta)w)]$$

invests a higher fraction of wealth θ in stocks as he gets richer.

Chapter 9

Convex Sets

9.1 Convex sets

A set $C \subset \mathbb{R}^N$ is said to be *convex* if the line segment generated by any two points in C is entirely contained in C. Formally, C is convex if $x, y \in C$ implies $(1-\alpha)x + \alpha y \in C$ for all $\alpha \in [0, 1]$ (Figure 9.1). So a circle, triangle, and square are convex but a star-shape is not (Figure 9.2). One of my favorite jokes is that the Chinese character for "convex" is not convex (Figure 9.3).

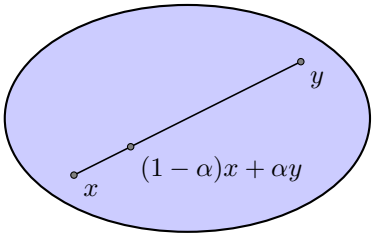

Figure 9.1. Definition of a convex set.

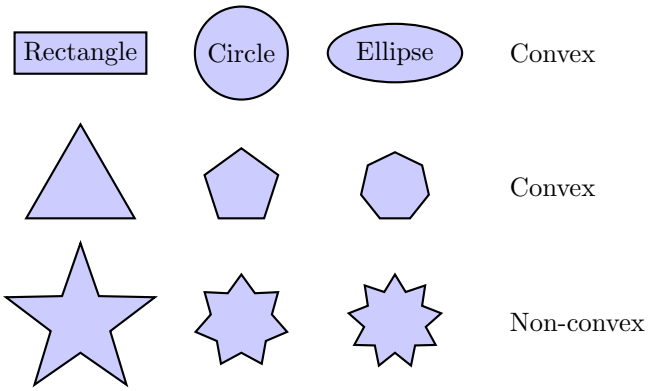

Figure 9.2. Examples of convex and non-convex sets.

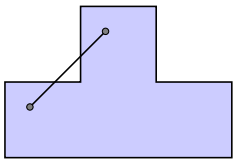

Figure 9.3. Chinese character for "convex" is not convex.

Let $A \subset \mathbb{R}^N$ be any set. The smallest convex set that includes A is called the *convex hull* of A and is denoted by $\operatorname{co} A$. (Its existence is proved in Problem 9.1.) For example, in Figure 9.4, the convex hull of the set A consisting of two circles is the entire region in between.

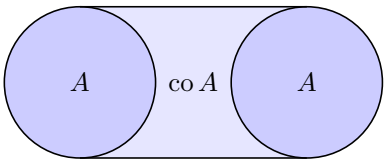

Figure 9.4. Convex hull.

Let $x_k \in \mathbb{R}^N$ for $k = 1, \dots, K$. A point of the form

$$x = \sum_{k=1}^{K} \alpha_k x_k,$$

where $\alpha_k \geq 0$ and $\sum_{k=1}^{K} \alpha_k = 1$, is called a *convex combination* of the points $\{x_k\}_{k=1}^{K}$. The following lemma provides a constructive way to obtain the convex hull of a set. Its proof is in Problem 9.2.

Lemma 9.1. *Let $A \subset \mathbb{R}^N$ be any set. Then $\operatorname{co} A$ consists of all convex combinations of points of A.*

9.2 Hyperplanes and half spaces

You should know from high school that the equation of a line in $\mathbb{R}^2$ is

$$a_1 x_1 + a_2 x_2 = c$$

for some real numbers a_1, a_2, c, and that the equation of a plane in $\mathbb{R}^3$ is

$$a_1 x_1 + a_2 x_2 + a_3 x_3 = c.$$

Letting $a = (a_1, \dots, a_N)$ and $x = (x_1, \dots, x_N)$ be vectors in $\mathbb{R}^N$, the equation $\langle a, x \rangle = c$ is a line if $N = 2$ and a plane if $N = 3$, where

$$\langle a, x \rangle = a_1 x_1 + \cdots + a_N x_N$$

is the inner product of the vectors a and x.[1] In general, we say that the set

$$\left\{ x \in \mathbb{R}^N \,\middle|\, \langle a, x \rangle = c \right\}$$

[1]The inner product is sometimes called the *vector product* or the *dot product*. Common notations for the inner product are $\langle a, x \rangle$, (a, x), $a \cdot x$, etc.

is a *hyperplane* if $a \neq 0$. The vector a is orthogonal to this hyperplane (is a *normal vector*). To see this, let x_0 be a point in the hyperplane. Since $\langle a, x_0 \rangle = c$, by subtraction and linearity of inner product we get $\langle a, x - x_0 \rangle = 0$. This means that the vector a is orthogonal to the vector $x - x_0$, which can point to any direction in the plane by moving x. So it makes sense to say that a is orthogonal to the hyperplane $\langle a, x \rangle = c$. The sets

$$H^+ = \left\{ x \in \mathbb{R}^N \,\middle|\, \langle a, x \rangle \geq c \right\},$$
$$H^- = \left\{ x \in \mathbb{R}^N \,\middle|\, \langle a, x \rangle \leq c \right\}$$

are called *half spaces*, since H^+ (H^-) is the portion of $\mathbb{R}^N$ separated by the hyperplane $\langle a, x \rangle = c$ towards the direction of a ($-a$). Hyperplanes and half spaces are convex sets (Problem 9.3).

9.3 Separation of convex sets

Let A, B be two sets. We say that the hyperplane $\langle a, x \rangle = c$ *separates* A, B if $A \subset H^-$ and $B \subset H^+$ (Figure 9.5), that is,

$$x \in A \Longrightarrow \langle a, x \rangle \leq c,$$
$$x \in B \Longrightarrow \langle a, x \rangle \geq c.$$

(The inequalities may be reversed.)

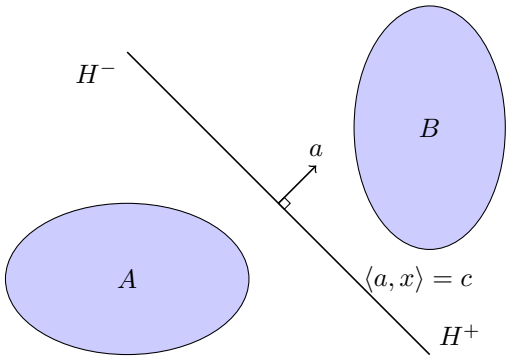

Figure 9.5. Separation of convex sets.

Clearly A, B can be separated if and only if

$$\sup_{x \in A} \langle a, x \rangle \leq \inf_{x \in B} \langle a, x \rangle,$$

since we can take c between these two numbers. We say that A, B can be *strictly separated* if the inequality is strict, so

$$\sup_{x \in A} \langle a, x \rangle < \inf_{x \in B} \langle a, x \rangle.$$

The remarkable property of convex sets is the following separation property.

Theorem 9.2 (Separating Hyperplane Theorem)**.** *Let $C, D \subset \mathbb{R}^N$ be nonempty and convex. If $C \cap D = \emptyset$, then there exists a hyperplane that separates C, D. If C, D are closed and one of them is compact, then they can be strictly separated.*

We need the following lemma to prove Theorem 9.2.

Lemma 9.3. *Let C be nonempty and convex. Then any $x \in \mathbb{R}^N$ has a unique closest point $P_C(x) \in \operatorname{cl} C$, called the* projection of x on $\operatorname{cl} C$. *Furthermore, for any $z \in C$ we have*

$$\langle x - P_C(x), z - P_C(x) \rangle \leq 0.$$

Proof. Let $\delta = \inf \{\|x - y\| \,|\, y \in C\} \geq 0$ be the distance from x to C (Figure 9.6).

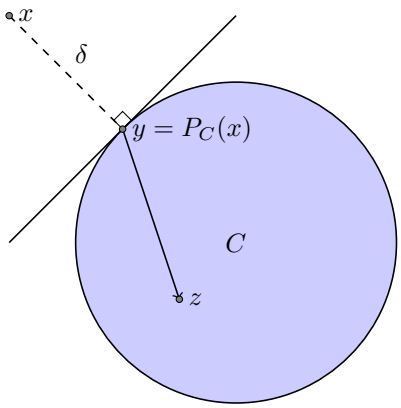

Figure 9.6. Projection on a convex set.

Take a sequence $\{y_k\} \subset C$ such that $\|x - y_k\| \to \delta$. Then by simple algebra we get

$$\|y_k - y_l\|^2 = 2\|x - y_k\|^2 + 2\|x - y_l\|^2 - 4\left\|x - \frac{1}{2}(y_k + y_l)\right\|^2. \tag{9.1}$$

Since C is convex, we have $\frac{1}{2}(y_k + y_l) \in C$, so by the definition of δ we get

$$\|y_k - y_l\|^2 \leq 2\|x - y_k\|^2 + 2\|x - y_l\|^2 - 4\delta^2 \to 2\delta^2 + 2\delta^2 - 4\delta^2 = 0$$

as $k, l \to \infty$. Since $\{y_k\} \subset C$ is Cauchy, it converges to some point $y \in \operatorname{cl} C$. Then

$$\|x - y\| \leq \|x - y_k\| + \|y_k - y\| \to \delta + 0 = \delta,$$

so y is the closest point to x in $\operatorname{cl} C$. If y_1, y_2 are two closest points, then by the same argument we get

$$0 \leq \|y_1 - y_2\|^2 \leq 2\|x - y_1\|^2 + 2\|x - y_2\|^2 - 4\delta^2 \leq 0,$$

so $y_1 = y_2$. Thus $y = P_C(x)$ is unique.

Finally, let $z \in C$ be any point. Take $\{y_k\} \subset C$ such that $y_k \to y = P_C(x)$. Since C is convex, for any $0 < \alpha \leq 1$ we have $(1 - \alpha)y_k + \alpha z \in C$. Therefore

$$\delta^2 = \|x - y\|^2 \leq \|x - (1 - \alpha)y_k - \alpha z\|^2.$$

Letting $k \to \infty$ we get $\|x - y\|^2 \leq \|x - y - \alpha(z - y)\|^2$. Expanding both sides, dividing by $\alpha > 0$, and letting $\alpha \to 0$, we get $\langle x - y, z - y \rangle \leq 0$, which is the desired inequality. □

The following proposition shows that a point that is not an interior point of a convex C can be separated from C.

Proposition 9.4. *Let $C \subset \mathbb{R}^N$ be nonempty and convex and $\bar{x} \notin \operatorname{int} C$. Then there exists a hyperplane $\langle a, x\rangle = c$ that separates $\bar{x}$ and C, i.e.,*

$$\langle a, \bar{x}\rangle \geq c \geq \langle a, z\rangle$$

for any $z \in C$. If $\bar{x} \notin \operatorname{cl} C$, then the above inequalities can be made strict.

Proof. Suppose that $\bar{x} \notin \operatorname{cl} C$. Let $y = P_C(\bar{x})$ be the projection of $\bar{x}$ on $\operatorname{cl} C$. Then $\bar{x} \neq y$ because $y \in \operatorname{cl} C$ and $\bar{x} \notin \operatorname{cl} C$. Let $a = \bar{x} - y \neq 0$ and $c = \langle a, y\rangle + \frac{1}{2}\|a\|^2$. Then for any $z \in C$ we have

$$\langle \bar{x} - y, z - y\rangle \leq 0 \Longrightarrow \langle a, z\rangle \leq \langle a, y\rangle < \langle a, y\rangle + \frac{1}{2}\|a\|^2 = c,$$
$$\langle a, \bar{x}\rangle - c = \langle \bar{x} - y, \bar{x} - y\rangle - \frac{1}{2}\|a\|^2 = \frac{1}{2}\|a\|^2 > 0 \iff \langle a, \bar{x}\rangle > c.$$

Therefore the hyperplane $\langle a, x\rangle = c$ strictly separates $\bar{x}$ and C.

If $\bar{x} \in \operatorname{cl} C$, since by assumption $\bar{x} \notin \operatorname{int} C$, we can take a sequence $\{x_k\}$ such that $x_k \notin \operatorname{cl} C$ and $x_k \to \bar{x}$. Then we can find a vector $a_k \neq 0$ and a number $c_k \in \mathbb{R}$ such that

$$\langle a_k, x_k\rangle \geq c_k \geq \langle a_k, z\rangle$$

for all $z \in C$. By dividing both sides by $\|a_k\| \neq 0$, without loss of generality we may assume $\|a_k\| = 1$. Since $x_k \to \bar{x}$, the sequence $\{c_k\}$ is bounded. Therefore we can find a convergent subsequence $(a_{k_l}, c_{k_l}) \to (a, c)$. Letting $l \to \infty$, we get

$$\langle a, \bar{x}\rangle \geq c \geq \langle a, z\rangle$$

for any $z \in C$. Therefore the hyperplane $\langle a, x\rangle = c$ separates $\bar{x}$ and C. □

Proof of Theorem 9.2. Let $E = C - D \coloneqq \{x - y \mid x \in C, y \in D\}$. Since C, D are nonempty and convex, so is E. Since $C \cap D = \emptyset$, we have $0 \notin E$. In particular, $0 \notin \operatorname{int} E$. By Proposition 9.4, there exists $a \neq 0$ such that $\langle a, 0\rangle = 0 \geq \langle a, z\rangle$ for all $z \in E$. By the definition of E, we have

$$\langle a, x - y\rangle \leq 0 \iff \langle a, x\rangle \leq \langle a, y\rangle$$

for any $x \in C$ and $y \in D$. Letting $\sup_{x\in C}\langle a, x\rangle \leq c \leq \inf_{y\in D}\langle a, y\rangle$, it follows that the hyperplane $\langle a, x\rangle = c$ separates C and D.

Suppose that C is closed and D is compact. Let us show that $E = C - D$ is closed. For this purpose, suppose that $\{z_k\} \subset E$ and $z_k \to z$. Then we can take $\{x_k\} \subset C$, $\{y_k\} \subset D$ such that $z_k = x_k - y_k$. Since D is compact, there is a subsequence such that $y_{k_l} \to y \in D$. Then $x_{k_l} = y_{k_l} + z_{k_l} \to y + z$, but since C is closed, $x = y + z \in C$. Therefore $z = x - y \in E$, so E is closed.

Since $E = C - D$ is closed and $0 \notin E$, by Proposition 9.4 there exists $a \neq 0$ such that $\langle a, 0\rangle = 0 > \langle a, z\rangle$ for all $z \in E$. The rest of the proof is similar. □

9.4 Application: asset pricing

Consider an economy with two periods, denoted by $t = 0, 1$. Suppose that at $t = 1$ the state of the economy can be one of $s = 1, \ldots, S$. There are J assets in the economy, indexed by $j = 1, \ldots, J$. One share of asset j trades for price q_j at time 0 and pays A_{sj} in state s. (It can be $A_{sj} < 0$, in which case the holder of one share of asset j must deliver $-A_{sj} > 0$ in state s.) Let $q = (q_1, \ldots, q_J)$ the vector of asset prices and $A = (A_{sj})$ be the matrix of asset payoffs. Define

$$W = W(q, A) = \begin{bmatrix} -q' \\ A \end{bmatrix}$$

be the $(1 + S) \times J$ matrix of net payments of one share of each asset in each state. Here, state 0 is defined by time 0 and the presence of $-q = (-q_1, \ldots, -q_J)$ means that in order to receive A_{sj} in state s one must purchase one share of asset j at time 0, thus paying q_j (receiving $-q_j$).

Let $\theta \in \mathbb{R}^J$ be a *portfolio*. (θ_j is the number of shares of asset j an investor buys. $\theta_j < 0$ corresponds to shortselling.) The net payments of the portfolio θ is the vector

$$W\theta = \begin{bmatrix} -q'\theta \\ A\theta \end{bmatrix} \in \mathbb{R}^{1+S}.$$

Here the investor pays $q'\theta$ at $t = 0$ for buying the portfolio θ, and receives $(A\theta)_s$ in state s at $t = 1$.

Let $\langle W \rangle = \left\{ W\theta \,\middle|\, \theta \in \mathbb{R}^J \right\} \subset \mathbb{R}^{1+S}$ be the set of payoffs generated by all portfolios, called the *asset span*. We say that the asset span $\langle W \rangle$ exhibits *no-arbitrage* if

$$\langle W \rangle \cap \mathbb{R}^{1+S}_+ = \{0\}.$$

That is, it is impossible to find a portfolio that pays a non-negative amount in every state and a positive amount in at least one state. Then we can show the following theorem, due to Harrison and Kreps (1979).

Theorem 9.5 (Fundamental Theorem of Asset Pricing)**.** *The asset span $\langle W \rangle$ exhibits no-arbitrage if and only if there exists $p \in \mathbb{R}^S_{++}$ such that $[1, p']W = 0$.*

In this case, the asset prices are given by

$$q_j = \sum_{s=1}^{S} p_s A_{sj}.$$

$p_s > 0$ is called the state price *in state s.*

Proof. Suppose that such a p exists. If $0 \neq w = (w_0, \ldots, w_S) \in \mathbb{R}^{1+S}_+$, then

$$[1, p']w = w_0 + \sum_{s=1}^{S} p_s w_s > 0,$$

so $w \notin \langle W \rangle$ because $[1, p']W = 0$. This shows $\langle W \rangle \cap \mathbb{R}^{1+S}_+ = \{0\}$.

Conversely, suppose that there is no arbitrage. Then $\langle W \rangle \cap \Delta = \emptyset$, where $\Delta = \left\{ w \in \mathbb{R}^{1+S}_+ \,\middle|\, \sum_{s=0}^{S} w_s = 1 \right\}$ is the unit simplex. Clearly $\langle W \rangle, \Delta$ are convex and nonempty, and Δ is compact. By the (strong version of) separating hyperplane theorem, we can find $0 \neq \lambda \in \mathbb{R}^{1+S}$ such that

$$\langle \lambda, w \rangle < \langle \lambda, d \rangle \tag{9.2}$$

for any $w \in \langle W \rangle$ and $d \in \Delta$. Let us show that $\lambda' W = 0$. Suppose not. Consider the portfolio $\theta = \alpha W' \lambda \in \mathbb{R}^J$, where $\alpha > 0$. Then by (9.2), for $w = W\theta$ we obtain

$$\langle \lambda, d \rangle > \langle \lambda, w \rangle = \langle \lambda, W(\alpha W' \lambda) \rangle = \alpha \lambda' W W' \lambda = \alpha \left\| \lambda' W \right\|^2 \to \infty$$

as $\alpha \to \infty$ because $\lambda' W \neq 0$, which is a contradiction. Therefore $\lambda' W = 0$, so $\langle \lambda, w \rangle = 0$ for all $w \in \langle W \rangle$. Then (9.2) becomes

$$0 < \langle \lambda, d \rangle$$

for all $d \in \Delta$. Letting $d = e_s$ (unit vector) for $s = 0, 1, \ldots, S$, we get $\lambda_s > 0$. Dividing both sides of $\lambda' W = 0$ by $\lambda_0 > 0$ and letting $p_s = \lambda_s / \lambda_0$ for $s = 1, \ldots, S$, the vector $p = (p_1, \ldots, p_S)$ satisfies $p \gg 0$ and $[1, p'] W = 0$. Writing down this equation component-wise, we get $q_j = \sum_{s=1}^{S} p_s A_{sj}$. □

Notes

Rockafellar (1970) is a classic reference for convex analysis. Much of the theory of separation of convex sets can be generalized to infinite-dimensional spaces. The proof in this chapter using the projection generalizes to Hilbert spaces, but for more general spaces (topological vector spaces) we need the Hahn-Banach theorem. See Berge (1959) and Luenberger (1969).

Problems

9.1. 1. Let $\{C_i\}_{i \in I} \subset \mathbb{R}^N$ be a collection of convex sets. Prove that $\bigcap_{i \in I} C_i$ is convex.

2. Let $A \subset \mathbb{R}^N$ be any set. Prove that there exists a smallest convex set that includes A (convex hull of A).

9.2. Let $A \subset \mathbb{R}^N$ be any set. Prove that $\operatorname{co} A$ consists of all convex combinations of points of A.

9.3. 1. Let $0 \neq a \in \mathbb{R}^N$ and $c \in \mathbb{R}$. Show that the hyperplane $H = \left\{ x \in \mathbb{R}^N \,\middle|\, \langle a, x \rangle = c \right\}$ and the half space $H^+ = \left\{ x \in \mathbb{R}^N \,\middle|\, \langle a, x \rangle \geq c \right\}$ are convex sets.

2. Let A be an $M \times N$ matrix and $b \in \mathbb{R}^M$. The set of the form

$$P = \left\{ x \in \mathbb{R}^N \,\middle|\, Ax \leq b \right\}$$

is called a *polytope*. Show that a polytope is convex.

9.4. Let $A \subset \mathbb{R}^N$ be any nonempty set.

1. Show that $\operatorname{cl} \operatorname{co} A$ (the closure of the convex hull of A) is a closed convex set.

2. Show by example that $\operatorname{co} \operatorname{cl} A$ (the convex hull of the closure of A) need not be closed.

9.5. 1. Let $a, b \in \mathbb{R}^N$. Prove the following *parallelogram law*:

$$\|a+b\|^2 + \|a-b\|^2 = 2\|a\|^2 + 2\|b\|^2 .$$

2. Using the parallelogram law, prove (9.1).

9.6. Let $A = \{(x,y) \in \mathbb{R}^2 \mid y > x^3\}$ and $B = \{(x,y) \in \mathbb{R}^2 \mid x \geq 1, y \leq 1\}$.

1. Draw a picture of the sets A, B on the xy plane.
2. Can A, B be separated? If so, provide an equation of a straight line that separates them. If not, explain why.

9.7. Let $C = \{(x,y) \in \mathbb{R}^2 \mid y > \mathrm{e}^x\}$ and $D = \{(x,y) \in \mathbb{R}^2 \mid y \leq 0\}$.

1. Draw a picture of the sets C, D on the xy plane.
2. Provide an equation of a straight line that separates C, D.
3. Can C, D be strictly separated? Answer yes or no, then explain why.

9.8. This problem asks you to prove Stiemke's theorem: if A is an $M \times N$ matrix, then exactly one of the following statements is true:

(a) There exists $x \in \mathbb{R}^N_{++}$ such that $Ax = 0$.

(b) There exists $y \in \mathbb{R}^M$ such that $A'y > 0$.

Prove Stiemke's theorem using the following hints.

1. Show that statements (a) and (b) cannot both be true.
2. Define the sets $C, D \subset \mathbb{R}^N$ by

$$C = \{A'y \mid y \in \mathbb{R}^M\},$$
$$D = \left\{ x \in \mathbb{R}^N \,\middle|\, x \geq 0, \sum_{n=1}^{N} x_n = 1 \right\}.$$

   Show that C, D are nonempty, closed, convex, and D is compact.
3. Show that if statement (b) does not hold, then statement (a) holds.

9.9. A typical linear programming problem is

$$\begin{array}{ll} \text{minimize} & \langle c, x \rangle \\ \text{subject to} & Ax \geq b, \end{array}$$

where $x \in \mathbb{R}^N$, $0 \neq c \in \mathbb{R}^N$, $b \in \mathbb{R}^M$, and A is an $M \times N$ matrix with $M \geq N$. A standard algorithm for solving a linear programming problem is the *simplex method*. The idea is that you keep moving from one vertex of the polytope

$$P = \{x \in \mathbb{R}^N \mid Ax \geq b\}$$

to a neighboring vertex as long as the function value decreases, and if there are no neighboring vertex with smaller function value, you stop.

1. Prove that the simplex method terminates in finite steps.
2. Prove that when the algorithm stops, you are at a solution of the original problem.

Chapter 10

Convex Functions

10.1 Convex functions

Let $f : \mathbb{R}^N \to [-\infty, \infty]$ be a function. The set

$$\operatorname{epi} f \coloneqq \big\{(x, y) \in \mathbb{R}^N \times \mathbb{R} \,\big|\, f(x) \le y\big\}$$

is called the *epigraph* of f (Figure 10.1), for the obvious reason that $\operatorname{epi} f$ is the set of points that lie on or above the graph of f. A function f is said to be *convex* if $\operatorname{epi} f$ is a convex set. It is straightforward to show (Problem 10.1) that a function f is convex if and only if for any $x_1, x_2 \in \mathbb{R}^N$ and $\alpha \in [0, 1]$, we have

$$f((1-\alpha)x_1 + \alpha x_2) \le (1-\alpha)f(x_1) + \alpha f(x_2). \tag{10.1}$$

This inequality is often used as the definition of a convex function.

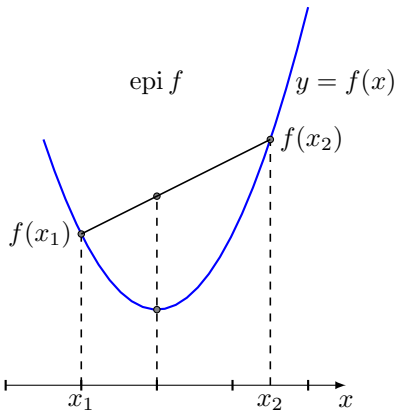

Figure 10.1. Convex function and its epigraph.

When the inequality (10.1) is strict whenever $x_1 \neq x_2$ and $\alpha \in (0, 1)$, we say that f is *strictly convex*. A convex function is *proper* if $f(x) > -\infty$ for all x and $f(x) < \infty$ for some x. If f is a convex function, then the set

$$\operatorname{dom} f \coloneqq \big\{x \in \mathbb{R}^N \,\big|\, f(x) < \infty\big\}$$

is a convex set, since $x_1, x_2 \in \operatorname{dom} f$ implies

$$f((1-\alpha)x_1 + \alpha x_2) \leq (1-\alpha)f(x_1) + \alpha f(x_2) < \infty.$$

$\operatorname{dom} f$ is called the *effective domain* of f.

Another useful but weaker concept is quasi-convexity. The set

$$L_f(y) = \left\{x \in \mathbb{R}^N \,\middle|\, f(x) \leq y\right\}$$

is called the *lower contour set* of f at level y. f is said to be *quasi-convex* if all lower contour sets are convex. It is straightforward to show (Problem 10.2) that f is quasi-convex if and only if for any $x_1, x_2 \in \mathbb{R}^N$ and $\alpha \in [0,1]$, we have

$$f((1-\alpha)x_1 + \alpha x_2) \leq \max\left\{f(x_1), f(x_2)\right\}.$$

Again if the inequality is strict whenever $x_1 \neq x_2$ and $0 < \alpha < 1$, then f is said to be *strictly quasi-convex*.

A function f is said to be *concave* if $-f$ is convex, that is, f is a convex function flipped upside down. The definition for strict concavity or quasi-concavity is similar.

10.2 Continuity of convex functions

A nice property of convex functions is that they are continuous except at boundary points of the domain.

Theorem 10.1. *Let $U \subset \mathbb{R}^N$ be an open convex set and $f : U \to \mathbb{R}$ be convex. Then f is continuous.*

Proof. Equip $\mathbb{R}^N$ with the supremum norm defined by $\|x\| = \max_n |x_n|$ for a vector $x = (x_1, \ldots, x_N) \in \mathbb{R}^N$. For any $x \in \mathbb{R}^N$ and $r > 0$, define the closed ball with center x and radius r by

$$\bar{B}(x, r) \coloneqq \left\{y \in \mathbb{R}^N \,\middle|\, \|y - x\| \leq r\right\}.$$

By the definition of the supremum norm, $\bar{B}(x, r)$ is actually the hypercube

$$[x_1 - r, x_1 + r] \times \cdots \times [x_N - r, x_N + r]$$

with 2^N vertices $(x_1 \pm r, \ldots, x_N \pm r)$.

Take any $x \in U$. Since U is open, we can take $r > 0$ such that $\bar{B}(x, r) \subset U$. Let the vertices of $\bar{B}(x, r)$ be denoted by $\{\bar{x}_k\}_{k=1}^K$, where $K = 2^N$. Define $M \coloneqq \max_k f(\bar{x}_k) < \infty$. Since clearly any point of $\bar{B}(x, r)$ can be expressed as a convex combination of $\{\bar{x}_k\}_{k=1}^K$ (the proof is by induction on N), we have

$$f(z) \leq M \text{ for all } z \in \bar{B}(x, r). \tag{10.2}$$

Now take any $y \in \bar{B}(x, r) \setminus \{x\}$, let $0 \neq d = y - x$, $\epsilon = \|d\| / r \in (0, 1]$, and define the points z_1, z_2 by $z_1 = x + d/\epsilon$ and $z_2 = x - d/\epsilon$ (Figure 10.2). Then clearly $\|z_j - x\| = \|d\| / \epsilon = r$ for $j = 1, 2$, so $z_j \in \bar{B}(x, r)$.

By the definition of z_j, we have

$$\begin{aligned} y - x = d = \epsilon(z_1 - x) &\iff y = (1-\epsilon)x + \epsilon z_1, \\ y - x = d = -\epsilon(z_2 - x) &\iff x = \frac{1}{1+\epsilon} y + \frac{\epsilon}{1+\epsilon} z_2. \end{aligned}$$

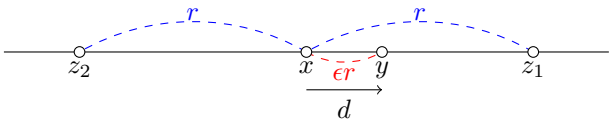

Figure 10.2. Definition of z_1 and z_2.

Hence by the convex inequality (10.1) and the upper bound (10.2), we obtain

$$f(y) \le (1-\epsilon)f(x) + \epsilon f(z_1) \implies f(y) - f(x) \le \epsilon(M - f(x)),$$
$$f(x) \le \frac{1}{1+\epsilon} f(y) + \frac{\epsilon}{1+\epsilon} f(z_2) \implies f(x) - f(y) \le \epsilon(M - f(x)).$$

Combining these two inequalities, we obtain

$$|f(y) - f(x)| \le \epsilon(M - f(x)) = \frac{M - f(x)}{r} \|y - x\|. \tag{10.3}$$

Therefore $f(y) \to f(x)$ as $y \to x$, so f is continuous. □

A convex function need not be continuous at boundary points of the domain. For example, define $f : [0,1] \to \mathbb{R}$ by $f(x) = 0$ if $x < 1$ and $f(1) = 1$. Then clearly f is convex but not continuous at $x = 1$.

A corollary of the proof of Theorem 10.1 is that convex functions are actually locally Lipschitz continuous. Recall that $f : U \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant $L \ge 0$ if for all $x, y \in U$, we have $|f(x) - f(y)| \le L\|x - y\|$.

Corollary 10.2. *Let $U \subset \mathbb{R}^N$ be a nonempty open convex set and $f : U \to \mathbb{R}$ be convex. Then f is locally Lipschitz.*

Proof. Take any $x \in U$ and $r > 0$ such that $\bar{B}(x, r) \subset U$, and define $V = \bar{B}(x, r/3)$. Let us show that f is Lipschitz on V. Since by Theorem 10.1 f is continuous on the compact set $\bar{B}(x, r)$, it attains a minimum m and a maximum M. Take any $x_1, x_2 \in V$. Then

$$\|x_1 - x_2\| \le \|x_1 - x\| + \|x - x_2\| \le \frac{2r}{3},$$

so $x_1 \in \bar{B}(x_2, 2r/3)$. If $y \in \bar{B}(x_2, 2r/3)$, then

$$\|y - x\| \le \|y - x_2\| + \|x_2 - x\| \le \frac{2r}{3} + \frac{r}{3} = r,$$

so $\bar{B}(x_2, 2r/3) \subset \bar{B}(x, r)$. Applying (10.3) to $y = x_1$ and $x = x_2$, we obtain

$$|f(x_1) - f(x_2)| \le \frac{M - m}{2r/3} \|x_1 - x_2\|,$$

which shows that f is Lipschitz on V with Lipschitz constant $L := \frac{3(M-m)}{2r}$. □

Unlike convex functions, quasi-convex functions need not be continuous. For example, any strictly increasing function $f : \mathbb{R} \to \mathbb{R}$ is quasi-convex, but there are many of them that are discontinuous.

10.3 Characterization of convex functions

When f is differentiable, there are simple ways to establish convexity.

Proposition 10.3. *Let $U \subset \mathbb{R}^N$ be an open convex set and $f : U \to \mathbb{R}$ be differentiable. Then f is (strictly) convex if and only if*

$$f(y) - f(x) \geq (>) \langle \nabla f(x), y - x \rangle$$

for all $x \neq y$.

Proof. Suppose that f is (strictly) convex. Let $x \neq y \in U$ and define $g : (0, 1] \to \mathbb{R}$ by

$$g(t) = \frac{f((1-t)x + ty) - f(x)}{t}.$$

Then g is (strictly) increasing, for if $0 < s < t \leq 1$ we have

$$\begin{aligned} & g(s) \leq (<) g(t) \\ \iff & \frac{f((1-s)x + sy) - f(x)}{s} \leq (<) \frac{f((1-t)x + ty) - f(x)}{t} \\ \iff & f((1-s)x + sy) \leq (<) \left(1 - \frac{s}{t}\right) f(x) + \frac{s}{t} f((1-t)x + ty), \end{aligned}$$

but the last inequality holds by letting $\alpha = s/t$, $x_1 = x$, $x_2 = (1-t)x + ty$, and using the definition of convexity. Therefore

$$f(y) - f(x) = g(1) \geq (>) \lim_{t \to 0} g(t) = \langle \nabla f(x), y - x \rangle .$$

Conversely, suppose that

$$f(y) - f(x) \geq (>) \langle \nabla f(x), y - x \rangle$$

for all $x \neq y$. Take any $x_1 \neq x_2$ and $\alpha \in (0, 1)$. Setting $y = x_1, x_2$ and $x = (1-\alpha)x_1 + \alpha x_2$, we get

$$\begin{aligned} f(x_1) - f((1-\alpha)x_1 + \alpha x_2) &\geq (>) \langle \nabla f(x), x_1 - x \rangle \\ f(x_2) - f((1-\alpha)x_1 + \alpha x_2) &\geq (>) \langle \nabla f(x), x_2 - x \rangle . \end{aligned}$$

Multiplying each by $1 - \alpha$ and α respectively and adding the two inequalities, we get

$$(1-\alpha) f(x_1) + \alpha f(x_2) - f((1-\alpha)x_1 + \alpha x_2) \geq (>) 0,$$

so f is (strictly) convex. □

Figure 10.3 shows the geometric intuition of Proposition 10.3. Since QR $= f(y) - f(x)$ and SR $= \langle \nabla f(x), y - x \rangle$, we have $f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle$.

A twice differentiable function $f : \mathbb{R} \to \mathbb{R}$ is convex if and only if $f''(x) \geq 0$ for all x. The following proposition is the generalization for $\mathbb{R}^N$.

Proposition 10.4. *Let $U \subset \mathbb{R}^N$ be an open convex set and $f : U \to \mathbb{R}$ be C^2. Then f is convex if and only if the Hessian (matrix of second derivatives)*

$$\nabla^2 f(x) = \left[\frac{\partial^2 f(x)}{\partial x_m \partial x_n} \right]$$

is positive semidefinite.

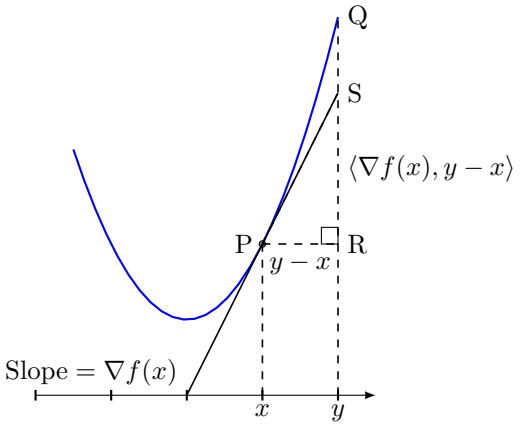

Figure 10.3. Characterization of a convex function.

Proof. Suppose that f is a C^2 function on U. Take any $x \neq y \in U$. Applying Taylor's theorem to $g(t) = f((1-t)x + ty)$ for $t \in [0,1]$, there exists $\alpha \in (0,1)$ such that

$$\begin{aligned} f(y) - f(x) = g(1) - g(0) &= g'(0) + \frac{1}{2} g''(\alpha) \\ &= \langle \nabla f(x), y - x \rangle + \frac{1}{2} \left\langle y - x, \nabla^2 f(x + \alpha(y-x))(y-x) \right\rangle. \end{aligned}$$

If f is convex, by Proposition 10.3

$$\frac{1}{2} \left\langle y - x, \nabla^2 f(x + \alpha(y-x))(y-x) \right\rangle = f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geq 0.$$

Since y is arbitrary, take any $d \in \mathbb{R}^N$ and let $y = x + \epsilon d$ for small enough $\epsilon > 0$ such that $y \in U$. Dividing the above inequality by $\frac{1}{2}\epsilon^2 > 0$ and letting $\epsilon \to 0$, we get

$$0 \leq \left\langle d, \nabla^2 f(x + \epsilon\alpha d) d \right\rangle \to \left\langle d, \nabla^2 f(x) d \right\rangle,$$

so $\nabla^2 f(x)$ is positive semidefinite. Conversely, if $\nabla^2 f(x)$ is positive semidefinite, then

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \frac{1}{2} \left\langle y - x, \nabla^2 f(x + \alpha(y-x))(y-x) \right\rangle \geq 0,$$

so by Proposition 10.3 f is convex. □

10.4 Characterization of quasi-convex functions

As in the case with convex functions, there are simple ways to establish quasi-convexity if f is differentiable or C^2.

Proposition 10.5. *Let f be differentiable. Then f is quasi-convex if and only if for all x, y we have*

$$f(y) \leq f(x) \implies \langle \nabla f(x), y - x \rangle \leq 0. \tag{10.4}$$

Proof. Suppose that f is quasi-convex and $f(y) \le f(x)$. Then for any $0 < t \le 1$ we have

$$f((1-t)x+ty) \le \max\{f(x), f(y)\} = f(x) \implies \frac{1}{t}(f(x+t(y-x)) - f(x)) \le 0.$$

Letting $t \to 0$, we obtain $\langle \nabla f(x), y-x\rangle \le 0$, so (10.4) holds.

Conversely, suppose that (10.4) holds. If f is not quasi-convex, there exist $\bar{x}, \bar{y}$, and $0 \le t \le 1$ such that

$$f((1-t)\bar{x}+t\bar{y}) > \max\{f(\bar{x}), f(\bar{y})\}. \tag{10.5}$$

Without loss of generality, we may assume $f(\bar{x}) \ge f(\bar{y})$. Define $g : [0,1] \to \mathbb{R}$ by $g(t) = f((1-t)\bar{x}+t\bar{y})$ and $T = \{t \in [0,1] \,|\, g(t) > g(0)\}$. Since $g(0) = f(\bar{x}) \ge f(\bar{y}) = g(1)$, (10.5) implies $T \ne \emptyset$ and $T \subset (0,1)$.

Let us show that $t \in T$ implies $g'(t) \ge 0$. To see this, take any $t \in T$ and let $x = (1-t)\bar{x}+t\bar{y}$ and $y = \bar{x}$. Since $f(x) = g(t) > g(0) = f(\bar{x}) = f(y)$, it follows from (10.4) that

$$\langle \nabla f(x), y-x\rangle \le 0 \iff 0 \le \langle \nabla f((1-t)\bar{x}+t\bar{y}), \bar{y}-\bar{x}\rangle = g'(t).$$

Since g is continuous, T is open. Let $(t_1, t_2) \subset T$ be a connected component of T. By continuity, we have $g(t_1) = g(t_2) = g(0)$. Since $g(t) > g(0)$ on T, we can take $t_3 \in (t_1, t_2)$ such that $g(t_3) > g(0) = g(t_2)$. By the mean value theorem, there exists $t_4 \in (t_3, t_2)$ such that

$$g'(t_4) = \frac{g(t_2) - g(t_3)}{t_2 - t_3} < 0,$$

which contradicts $g'(t_4) \ge 0$. □

Proposition 10.6. *Let f be C^2. If f is quasi-convex, then*

$$\langle \nabla f(x), d\rangle = 0 \implies \left\langle d, \nabla^2 f(x) d\right\rangle \ge 0$$

for all x and $d \ne 0$. Conversely, if

$$\langle \nabla f(x), d\rangle = 0 \implies \left\langle d, \nabla^2 f(x) d\right\rangle > 0$$

for all x and $d \ne 0$, then f is quasi-convex.

Proof. Let $g(t) = f(x+td)$. If f is quasi-convex, so is g. Suppose $\langle \nabla f(x), d\rangle = 0$ but $\left\langle d, \nabla^2 f(x) d\right\rangle < 0$. Since $g'(0) = \langle \nabla f(x), d\rangle = 0$ and $g''(0) = \left\langle d, \nabla^2 f(x) d\right\rangle < 0$, $t = 0$ is a strict local maximum of g, which is a contradiction.

Conversely, suppose

$$\langle \nabla f(x), d\rangle = 0 \implies \left\langle d, \nabla^2 f(x) d\right\rangle > 0$$

for all x and $d \ne 0$. By Proposition 10.5, it suffices to show

$$f(y) \le f(x) \implies \langle \nabla f(x), y-x\rangle \le 0.$$

If $x = y$, the claim is trivial. Suppose that $x \ne y$, $f(y) \le f(x)$, and let $d = y - x \ne 0$. Suppose $\langle \nabla f(x), y-x\rangle > 0$. Define $g : [0,1] \to \mathbb{R}$ by $g(t) = f((1-t)x+ty)$. Since g is continuous, it attains a maximum $t \in [0,1]$. Since

$g(0) = f(x) \geq f(y) = g(1)$ and $g'(0) = \langle \nabla f(x), y - x \rangle > 0$, we have $0 < t < 1$. Since t is an interior maximum, we have

$$0 = g'(t) = \langle \nabla f((1-t)x + ty), d \rangle = 0,$$

so by assumption

$$0 < \left\langle d, \nabla^2 f((1-t)x + ty)d \right\rangle = g''(t).$$

However, this shows that t is a strict local minimum, which is a contradiction. □

10.5 Subgradient of convex functions

Suppose that f is a proper convex function and $x \in \operatorname{int} \operatorname{dom} f$. Since $(x, f(x)) \in \operatorname{epi} f$ but $(x, f(x) - \epsilon) \notin \operatorname{epi} f$ for all $\epsilon > 0$, we have $(x, f(x)) \notin \operatorname{int} \operatorname{epi} f$. Hence by the separating hyperplane theorem, there exists a vector $0 \neq (\eta, \beta) \in \mathbb{R}^N \times \mathbb{R}$ such that

$$\langle \eta, x \rangle + \beta f(x) \leq \langle \eta, y \rangle + \beta z$$

for any $(y, z) \in \operatorname{epi} f$. Letting $z \to \infty$, we get $\beta \geq 0$. Letting $z = f(y)$ we get

$$\beta(f(y) - f(x)) \geq \langle -\eta, y - x \rangle$$

for all y. If $\beta = 0$, since $x \in \operatorname{int} \operatorname{dom} f$, the vector $y - x$ can point to any direction. Then $\eta = 0$, which contradicts $(\eta, \beta) \neq 0$. Therefore $\beta > 0$. Letting $\xi = -\eta/\beta$, we get

$$f(y) - f(x) \geq \langle \xi, y - x \rangle \tag{10.6}$$

for any y.

A vector ξ that satisfies (10.6) is called a *subgradient* (subdifferential) of f at $x \in \operatorname{int} \operatorname{dom} f$. The set of all subgradients at x is denoted by

$$\partial f(x) = \left\{ \xi \in \mathbb{R}^N \,\middle|\, (\forall y) f(y) - f(x) \geq \langle \xi, y - x \rangle \right\}.$$

$\partial f(x)$ is a closed convex set (exercise).

If f is partially differentiable, letting $y = x + td$ and letting $t \to 0$, we get

$$\langle \xi, d \rangle \leq \frac{f(x + td) - f(x)}{t} \to f'(x; d) = \langle \nabla f(x), d \rangle$$

for all d, so $\xi = \nabla f(x)$ and $\partial f(x) = \{\nabla f(x)\}$.

Problems

10.1.

Prove that $\operatorname{epi} f$ is a convex set if and only if

$$f((1-\alpha)x_1 + \alpha x_2) \leq (1-\alpha) f(x_1) + \alpha f(x_2)$$

for all $x_1, x_2 \in \mathbb{R}^N$ and $\alpha \in [0, 1]$.

10.2. Prove that f is quasi-convex if and only if

$$f((1-\alpha)x_1 + \alpha x_2) \leq \max\{f(x_1), f(x_2)\}$$

for all $x_1, x_2 \in \mathbb{R}^N$ and $\alpha \in [0,1]$.

10.3. 1. Show that if $\{f_i\}_{i=1}^I$ are convex, so is $f = \sum_{i=1}^I \beta_i f_i$ for any $\beta_1, \ldots, \beta_I \geq 0$.

2. Show that if $\{f_i\}_{i\in I}$ are (quasi-)convex, so is $f = \sup_{i\in I} f_i$.

3. Suppose that $h : \mathbb{R}^M \to \mathbb{R}$ is increasing (meaning $x \leq y$ implies $h(x) \leq h(y)$) and (quasi-)convex and $g_m : \mathbb{R}^N \to \mathbb{R}$ is convex for $m = 1, \ldots, M$. Prove that $f(x) = h(g_1(x), \ldots, g_M(x))$ is (quasi-)convex.

10.4. Let X, Y be vector spaces, $f : X \times Y \to (-\infty, \infty]$ be (quasi-)convex, and define $g : Y \to [-\infty, \infty]$ by $g(y) = \inf_{x\in X} f(x,y)$. Show that g is (quasi-)convex.

10.5. Let X be a vector space, Y a set, $\Gamma : X \twoheadrightarrow Y$ a correspondence (so for each $x \in X$, $\Gamma(x)$ is a subset of Y), and $f : Y \to [-\infty, \infty]$. Suppose that Γ satisfies

$$\Gamma((1-\alpha)x_1 + \alpha x_2) \subset \Gamma(x_1) \cup \Gamma(x_2)$$

for all $x_1, x_2 \in X$ and $\alpha \in [0,1]$. Define

$$\begin{aligned} \bar{g}(x) &= \sup_{y\in\Gamma(x)} f(y), \\ \underline{g}(x) &= \inf_{y\in\Gamma(x)} f(y). \end{aligned}$$

Prove that $\bar{g}$ is quasi-convex and $\underline{g}$ is quasi-concave.

10.6. Let C be a convex set of a vector space X. We say that $x \in C$ is an *extreme point* if there exist no $x_1 \neq x_2 \in C$ and $\alpha \in (0,1)$ such that $x = (1-\alpha)x_1 + \alpha x_2$. If $f : C \to \mathbb{R}$ is strictly quasi-convex and $\bar{x} \in C$ achieves the maximum of f over C, prove that $\bar{x}$ is an extreme point of C.

10.7. Let $f : [a,b] \to \mathbb{R}$ be convex, continuous, and $f(a) < 0 < f(b)$. Show that there exists a unique $x \in (a,b)$ such that $f(x) = 0$.

10.8. Let $f : (a,b) \to \mathbb{R}$ be convex.

1. Show that for each $x \in (a,b)$,

$$g^{\pm}(x) \coloneqq \lim_{h\to\pm 0} \frac{f(x+h) - f(x)}{h}$$

exist.

2. Show that $g^-(x) \leq g^+(x)$ for each $x \in (a,b)$.

3. Show that f is differentiable on (a,b) except at at most countably many points.

10.9. Let $\emptyset \neq X \subset \mathbb{R}^N$ and $u : X \to \mathbb{R}$. Define $e : \mathbb{R}^N \times \mathbb{R} \to [-\infty, \infty]$ by

$$e(p, u) = \inf \{p \cdot x \,|\, x \in X, u(x) \geq u\},$$

where by convention we define $\inf \emptyset = \infty$. (Economically, X is a consumption set, u is a utility function, p is a price vector, and e is the minimum expenditure to achieve utility level u given the price vector p, which is called the *expenditure function.*) Prove that $e(p, u)$ is concave in p.

10.10. Let $\emptyset \neq X \subset \mathbb{R}^N$ and $u : X \to \mathbb{R}$. Define $v : \mathbb{R}^N \times \mathbb{R} \to [-\infty, \infty]$ by

$$v(p, w) = \sup \{u(x) \,|\, x \in X, p \cdot x \leq w\},$$

where by convention we define $\sup \emptyset = -\infty$. (Economically, X is a consumption set, u is a utility function, p is a price vector, w is wealth, and v is the maximum utility given the price vector p and wealth w, which is called the *indirect utility function.*)

1. Take any $(p_j, w_j) \in \mathbb{R}^N \times \mathbb{R}$ and define $p = (1 - \alpha)p_1 + \alpha p_2$, $w = (1 - \alpha)w_1 + \alpha w_2$, where $\alpha \in [0, 1]$. Show that if $x \in X$ satisfies $p \cdot x \leq w$, then $p_j \cdot x \leq w_j$ for at least one j.

2. Prove that $v(p, w)$ is quasi-convex in (p, w).

10.11. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = |x|$. Compute the subdifferential $\partial f(x)$.

10.12. Let f be a proper convex function and $x \in \operatorname{int} \operatorname{dom} f$. Prove that $\partial f(x)$ is a closed convex set.

10.13. Define $f : \mathbb{R} \to (-\infty, \infty]$ by

$$f(x) = \begin{cases} \infty, & (x < 0) \\ -\sqrt{x}. & (x \geq 0) \end{cases}$$

1. Show that f is a proper convex function.

2. Compute the subdifferential $\partial f(x)$ for $x > 0$. Does $\partial f(0)$ exist?

Chapter 11

Convex Programming

11.1 Convex programming

Consider the minimization problem

$$\text{minimize } f(x) \text{ subject to } x \in C. \tag{11.1}$$

The minimization problem (11.1) is called a *convex programming problem* if $f(x)$ is a convex function and C is a convex set. By redefining f such that $f(x) = \infty$ for $x \notin C$, the constrained optimization problem (11.1) becomes the unconstrained optimization problem

$$\text{minimize } f(x).$$

By Proposition 4.1, if f is differentiable and $\bar{x}$ minimizes f, then $\nabla f(\bar{x}) = 0$, which is called the first-order necessary condition. In general, the first-order condition is not sufficient. For instance, for $f(x) = x^3$ we have $f'(0) = 0$, but $x = 0$ does not minimize f since $f(-1) = -1 < 0 = f(0)$.

For a convex programming problem, however, the first-order necessary condition is also sufficient, as the following proposition shows.

Proposition 11.1. *Let $f : \mathbb{R}^N \to (-\infty, \infty]$ be a proper convex function. Then $\bar{x} \in \operatorname{int} \operatorname{dom} f$ is a solution to (11.1) if and only if $0 \in \partial f(\bar{x})$. In particular, if f is differentiable, $f(\bar{x}) = \min f(x)$ if and only if $\nabla f(\bar{x}) = 0$.*

Proof. If $f(\bar{x}) = \min f(x)$, then for any x

$$f(x) - f(\bar{x}) \geq 0 = \langle 0, x - \bar{x} \rangle = 0,$$

so $0 \in \partial f(\bar{x})$. If $0 \in \partial f(\bar{x})$, then

$$f(x) - f(\bar{x}) \geq \langle 0, x - \bar{x} \rangle = 0,$$

so $f(\bar{x}) = \min f(x)$. □

In applications, the constraint set is often given by inequalities and equa-

tions. Consider the constrained minimization problem

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & g_i(x) \le 0 & (i = 1, \dots, I), \\ & h_j(x) = 0 & (j = 1, \dots, J), \\ & x \in \Omega, & \end{array} \tag{11.2}$$

where f, g_i's, and h_j's are functions and $\Omega \subset \mathbb{R}^N$ is a set. The constraints $g_i(x) \le 0$ and $h_j(x) = 0$ are called *inequality* and *equality* constraints, respectively. The condition $x \in \Omega$ is called a *side constraint*.

When f, g_i's are convex, h_j's are affine (so $h_j(x) = \langle a_j, x\rangle - b_j$ for some $a_j \in \mathbb{R}^N \setminus \{0\}$, $b_j \in \mathbb{R}$ for all j), and Ω is convex, then (11.2) is a convex programming problem. Without loss of generality, we may assume that $\{a_j\}$ is linearly independent, for otherwise either the constraint set is empty or some constraints are redundant (Problem 11.2). Letting $A = (a_1, \dots, a_J)'$ (an $J \times N$ matrix) and $b = (b_1, \dots, b_J)'$, the equality constraints can be compactly written as $Ax - b = 0$.

To characterize the solution of (11.2), define the *Lagrangian* by

$$L(x, \lambda, \mu) = \begin{cases} f(x) + \sum_{i=1}^I \lambda_i g_i(x) + \sum_{j=1}^J \mu_j h_j(x), & (\lambda \in \mathbb{R}^I_+) \\ -\infty, & (\lambda \notin \mathbb{R}^I_+) \end{cases}$$

where $\lambda = (\lambda_1, \dots, \lambda_I) \in \mathbb{R}^I$ and $\mu = (\mu_1, \dots, \mu_J) \in \mathbb{R}^J$ are called *Lagrange multipliers*. (Defining L to be $-\infty$ when $\lambda \notin \mathbb{R}^I_+$ is useful later when explaining duality.) A point $(\bar{x}, \bar{\lambda}, \bar{\mu}) \in \Omega \times \mathbb{R}^I \times \mathbb{R}^J$ is called a *saddle point* if it achieves the minimum with respect to x and maximum with respect to (λ, μ). Formally, $(\bar{x}, \bar{\lambda}, \bar{\mu})$ is a saddle point if

$$L(\bar{x}, \lambda, \mu) \le L(\bar{x}, \bar{\lambda}, \bar{\mu}) \le L(x, \bar{\lambda}, \bar{\mu}) \tag{11.3}$$

for all $(x, \lambda, \mu) \in \Omega \times \mathbb{R}^{I+J}$.

The following theorem gives necessary and sufficient conditions for optimality.

Theorem 11.2 (Saddle Point Theorem)**.** *Let $\Omega \subset \mathbb{R}^N$ be a convex set. Suppose that $f, g_i : \Omega \to (-\infty, \infty]$'s are convex and h_j's are affine in the minimization problem (11.2). Let*

$$(h_1(x), \dots, h_J(x))' = Ax - b,$$

where A is an $J \times N$ matrix and $b \in \mathbb{R}^J$.

1. *If (i) $\bar{x}$ is a solution to the minimization problem (11.2), (ii) there exists $x_0 \in \mathbb{R}^N$ such that $g_i(x_0) < 0$ for all i and $Ax_0 - b = 0$, and (iii) $0 \in \operatorname{int}(A\Omega - b)$, then there exist Lagrange multipliers $\bar{\lambda} \in \mathbb{R}^I_+$ and $\bar{\mu} \in \mathbb{R}^J$ such that $(\bar{x}, \bar{\lambda}, \bar{\mu})$ is a saddle point of L.*

2. *If there exist Lagrange multipliers $\bar{\lambda} \in \mathbb{R}^I_+$ and $\bar{\mu} \in \mathbb{R}^J$ such that $(\bar{x}, \bar{\lambda}, \bar{\mu})$ is a saddle point of L, then $\bar{x}$ is a solution to the minimization problem (11.2).*

Remark. Condition (1ii) is called the *Slater constraint qualification* and will be discussed in more detail in Chapter 12. In applications, we often have $\Omega = \mathbb{R}^N$.

Then condition (1iii) holds automatically since $A\mathbb{R}^N - b = \mathbb{R}^J$ when the row vectors of A are linearly independent, which we may assume without loss of generality. Condition (1iii) also holds when there are no equality constraints ($J = 0$).

Proof of Theorem 11.2.

Necessity (Claim 1). Assume that $\bar{x} \in \Omega$ is a solution to (11.2). Define the sets $C, D \subset \mathbb{R}^{1+I+J}$ by

$$C = \left\{(u, v, w) \in \mathbb{R}^{1+I+J} \,\middle|\, (\exists x \in \Omega) u \geq f(x), (\forall i) v_i \geq g_i(x), w = Ax - b\right\},$$
$$D = \left\{(u, v, w) \in \mathbb{R}^{1+I+J} \,\middle|\, u < f(\bar{x}), (\forall i) v_i < 0, (\forall j) w_j = 0\right\}.$$

(See Figure 11.1.)

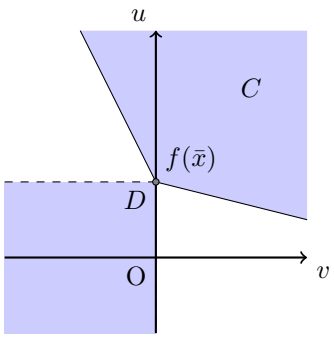

Figure 11.1. Saddle point theorem.

Clearly C, D are convex since f, g_i's are convex and Ω is convex. Since

$$(f(\bar{x}), \bar{v}, A\bar{x} - b) \in C$$

for $\bar{v}_i = g_i(\bar{x})$, C is nonempty. Letting $v_{0i} = g_i(x_0) < 0$, $v_0 = (v_{01}, \ldots, v_{0I})$, and $u_0 < f(\bar{x})$, we have $(u_0, v_0, 0) \in D$, so D is nonempty. If $(u, v, w) \in C \cap D$, since $(u, v, w) \in D$ we have $u < f(\bar{x})$, $v \ll 0$, and $w = 0$. Then since $(u, v, 0) \in C$ there exists $x \in \Omega$ such that $f(x) \leq u < f(\bar{x})$, $g_i(x) < 0$ for all i, and $Ax - b = 0$, contradicting the optimality of $\bar{x}$. Therefore $C \cap D = \emptyset$. By the separating hyperplane theorem, there exists $0 \neq (\alpha, \beta, \gamma) \in \mathbb{R} \times \mathbb{R}^I \times \mathbb{R}^J$ such that

$$\sup_{(u,v,w)\in D} \alpha u + \langle \beta, v \rangle + \langle \gamma, w \rangle \leq \inf_{(u,v,w)\in C} \alpha u + \langle \beta, v \rangle + \langle \gamma, w \rangle.$$

Taking $(u, v, w) \in D$ and letting $u \to -\infty$, it must be $\alpha \geq 0$. Similarly, letting $v_i \to -\infty$, it must be $\beta_i \geq 0$ for all i.

Let us show that $\alpha > 0$. For any $\epsilon > 0$, we have $(f(\bar{x}) - \epsilon, -\epsilon\mathbf{1}, 0) \in D$, so

$$\alpha(f(\bar{x}) - \epsilon) - \epsilon \langle \beta, \mathbf{1} \rangle \leq \alpha f(x) + \sum_{i=1}^{I} \beta_i g_i(x) + \sum_{j=1}^{J} \gamma_j h_j(x) \tag{11.4}$$

for any x. Letting $x = x_0$ and $\epsilon \to 0$, we get

$$\alpha f(\bar{x}) \leq \alpha f(x_0) + \sum_{i=1}^{I} \beta_i g_i(x_0).$$

If $\alpha = 0$, since by assumption $g_i(x_0) < 0$, it must be $\beta_i = 0$ for all i. If $J = 0$ (no equality constraints), then $(\alpha, \beta) = 0$, a contradiction. Hence $\alpha > 0$. If $J > 0$, then

$$\sup_{(u,v,w)\in D} \langle \gamma, w\rangle \le \inf_{(u,v,w)\in C} \langle \gamma, w\rangle \iff (\forall x \in \Omega)\gamma'(Ax - b) \ge 0.$$

Since by assumption 0 is an interior point of $A\Omega - b$, for small enough $\delta > 0$ there exists $x \in \Omega$ with $-\delta\gamma = Ax - b$. Therefore $-\delta \|\gamma\|^2 \ge 0$ implies $\gamma = 0$. Then $(\alpha, \beta, \gamma) = 0$, a contradiction. Hence $\alpha > 0$.

Since $\alpha > 0$, define $\bar\lambda = \beta/\alpha$ and $\bar\mu = \gamma/\alpha$. Then letting $\epsilon \to 0$ in (11.4), we get

$$f(\bar x) \le f(x) + \sum_{i=1}^{I} \bar\lambda_i g_i(x) + \sum_{j=1}^{J} \bar\mu_j h_j(x) =: L(x, \bar\lambda, \bar\mu)$$

for all x. Since $\bar\lambda_i \ge 0$ and $g_i(\bar x) \le 0$ for all i and $h_j(\bar x) = 0$ for all j, it follows that

$$\begin{aligned} L(\bar x, \bar\lambda, \bar\mu) &= f(\bar x) + \sum_{i=1}^{I} \bar\lambda_i g_i(\bar x) + \sum_{j=1}^{J} \bar\mu_j h_j(\bar x) \\ &\le f(\bar x) \le f(x) + \sum_{i=1}^{I} \bar\lambda_i g_i(x) + \sum_{j=1}^{J} \bar\mu_j h_j(x) = L(x, \bar\lambda, \bar\mu), \end{aligned}$$

which is the right inequality of (11.3) and it must be $\bar\lambda_i g_i(\bar x) = 0$ for all i. It remains to show the left inequality of (11.3). If $\lambda \notin \mathbb{R}^I_+$, by the definition of the Lagrangian we have $L(x, \lambda, \mu) = -\infty$, so it is trivial. If $\lambda \in \mathbb{R}^I_+$, then since $g_i(\bar x) \le 0$ and $h_j(\bar x) = 0$, we obtain

$$\begin{aligned} L(\bar x, \lambda, \mu) &= f(\bar x) + \sum_{i=1}^{I} \lambda_i g_i(\bar x) + \sum_{j=1}^{J} \mu_j h_j(\bar x) \\ &\le f(\bar x) = L(\bar x, \bar\lambda, \bar\mu), \end{aligned}$$

which is the left inequality of (11.3).

Sufficiency (Claim 2). Assume that $(\bar x, \bar\lambda, \bar\mu) \in \Omega \times \mathbb{R}^I_+ \times \mathbb{R}^J$ is a saddle point of L. By the left inequality of (11.3), for any $\lambda \in \mathbb{R}^I_+$ and $\mu \in \mathbb{R}^J$ we obtain

$$\begin{aligned} &f(\bar x) + \sum_{i=1}^{I} \lambda_i g_i(\bar x) + \sum_{j=1}^{J} \mu_j h_j(\bar x) \le f(\bar x) + \sum_{i=1}^{I} \bar\lambda_i g_i(\bar x) + \sum_{j=1}^{J} \bar\mu_j h_j(\bar x) \\ \Longrightarrow\ &\sum_{i=1}^{I} \lambda_i g_i(\bar x) + \sum_{j=1}^{J} \mu_j h_j(\bar x) \le \sum_{i=1}^{I} \bar\lambda_i g_i(\bar x) + \sum_{j=1}^{J} \bar\mu_j h_j(\bar x). \end{aligned}$$

Letting $\mu_j \to \pm\infty$, it must be $h_j(\bar x) = 0$ for all j. Letting $\lambda_i \to \infty$, we get $g_i(\bar x) \le 0$ for all i. Letting $\lambda = 0$, we get $0 \le \sum_{i=1}^{I} \bar\lambda_i g_i(\bar x)$, so it must be $\bar\lambda_i g_i(\bar x) = 0$ for all i. Then by the right inequality of (11.3), for any $x \in \Omega$ we

obtain

$$f(\bar{x}) + \sum_{i=1}^{I} \bar{\lambda}_i g_i(\bar{x}) + \sum_{j=1}^{J} \bar{\mu}_j h_j(\bar{x}) \leq f(x) + \sum_{i=1}^{I} \bar{\lambda}_i g_i(x) + \sum_{j=1}^{J} \bar{\mu}_j h_j(x)$$

$$\Longrightarrow f(\bar{x}) \leq f(x) + \sum_{i=1}^{I} \bar{\lambda}_i g_i(x) + \sum_{j=1}^{J} \bar{\mu}_j h_j(x).$$

Since $\bar{\lambda}_i \geq 0$, if $g_i(x) \leq 0$ and $h_j(x) = 0$ it follows that $f(\bar{x}) \leq f(x)$, so $\bar{x}$ is a solution to the constrained minimization problem (11.2). □

The following corollary is useful for computing the solution of a convex programming problem.

Corollary 11.3. *Let $\Omega \subset \mathbb{R}^N$ be a convex set, $f, g_i : \Omega \to (-\infty, \infty]$ be convex and differentiable on Ω, and h_j be affine.*

1. *If (i) $\bar{x}$ is a solution to the minimization problem (11.2), (ii) there exists $x_0 \in \mathbb{R}^N$ such that $g_i(x_0) < 0$ for all i and $Ax_0 - b = 0$, and (iii) $0 \in \operatorname{int}(A\Omega - b)$, then there exist Lagrange multipliers $\bar{\lambda} \in \mathbb{R}^I_+$ and $\bar{\mu} \in \mathbb{R}^J$ such that*

$$\nabla f(\bar{x}) + \sum_{i=1}^{I} \bar{\lambda}_i \nabla g_i(\bar{x}) + \sum_{j=1}^{J} \bar{\mu}_j \nabla h_j(\bar{x}) = 0, \tag{11.5a}$$

$$(\forall i)\bar{\lambda}_i \geq 0,\ g_i(\bar{x}) \leq 0,\ \bar{\lambda}_i g_i(\bar{x}) = 0, \tag{11.5b}$$

$$(\forall j) h_j(\bar{x}) = 0. \tag{11.5c}$$

2. *If there exist Lagrange multipliers $\bar{\lambda} \in \mathbb{R}^I_+$ and $\bar{\mu} \in \mathbb{R}^J$ such that (11.5) holds, then $\bar{x}$ is a solution to the minimization problem (11.2).*

Proof. If $\bar{x}$ is a solution, by the saddle point theorem and its proof, there exists Lagrange multipliers such that (11.5b) and (11.5c) hold and $\bar{x}$ minimizes $L(x, \bar{\lambda}, \bar{\mu})$. Then by Proposition 11.1, (11.5a) holds.

If (11.5) holds, then by condition (11.5a) and Proposition 11.1, $\bar{x}$ minimizes $L(x, \bar{\lambda}, \bar{\mu})$. By conditions (11.5b) and (11.5c), we can show that $\bar{x}$ solves (11.4) by imitating the sufficiency proof of Theorem 11.2. □

Condition (11.5a) is called the *first-order condition*. Condition (11.5b) is called the *complementary slackness condition*.

By Corollary 11.3, we can solve a constrained minimization problem (11.2) as follows.

Step 1. Verify that the functions f, g_i's are convex, h_j's are affine, and the constraint set Ω is convex.

Step 2. Verify the Slater condition (there exists $x_0 \in \Omega$ such that $g_i(x_0) < 0$ for all i and $h_j(x_0) = 0$ for all j) and $0 \in \operatorname{int} A\Omega - b$.

Step 3. Form the Lagrangian

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{I} \lambda_i g_i(x) + \sum_{j=1}^{J} \mu_j h_j(x).$$

Derive the first-order condition and complementary slackness condition (11.5).

Step 4. Solve (11.5). If there is a solution $\bar{x}$, it is a solution to the minimization problem 11.2. Otherwise, there are no solutions.

As an example, consider the problem

$$\begin{aligned} &\text{minimize} && \frac{1}{x_1} + \frac{1}{x_2} \\ &\text{subject to} && x_1 + x_2 \leq 2, \\ & && x_1, x_2 > 0. \end{aligned}$$

Let us solve this problem step by step.

Step 1. Let $f(x_1, x_2) = \frac{1}{x_1} + \frac{1}{x_2}$ be the objective function. Since

$$(1/x)'' = (-x^{-2})' = 2x^{-3} > 0,$$

the Hessian of f,

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

is positive definite. Therefore f is convex. Let $g(x_1, x_2) = x_1 + x_2 - 2$. Since g is affine, it is convex. Clearly the set

$$\Omega = \mathbb{R}^2_{++} = \{(x_1, x_2) \,|\, x_1, x_2 > 0\}$$

is convex.

Step 2. For $(x_1, x_2) = (\frac{1}{2}, \frac{1}{2})$ we have $g(x_1, x_2) = -1 < 0$, so the Slater condition holds.

Step 3. Let

$$L(x_1, x_2, \lambda) = \frac{1}{x_1} + \frac{1}{x_2} + \lambda(x_1 + x_2 - 2)$$

be the Lagrangian. The first-order condition is

$$\begin{aligned} 0 = \frac{\partial L}{\partial x_1} = -\frac{1}{x_1^2} + \lambda \iff x_1 = \frac{1}{\sqrt{\lambda}}, \\ 0 = \frac{\partial L}{\partial x_2} = -\frac{1}{x_2^2} + \lambda \iff x_2 = \frac{1}{\sqrt{\lambda}}. \end{aligned}$$

The complementary slackness condition is

$$\lambda(x_1 + x_2 - 2) = 0.$$

Step 4. From these equations it must be $\lambda > 0$ and

$$x_1 + x_2 - 2 = 0 \iff \frac{2}{\sqrt{\lambda}} - 2 = 0 \iff \lambda = 1$$

and $x_1 = x_2 = 1/\sqrt{\lambda} = 1$. Therefore $(x_1, x_2) = (1, 1)$ is the (only) solution.

When f, g_i are not convex, but only quasi-convex, we can still give a sufficient condition for optimality.

Theorem 11.4. *Consider the constrained optimization problem*

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & g_i(x) \leq 0 & (i = 1, \dots, I), \end{array}$$

where f, g_i's are differentiable and quasi-convex. Suppose that the Slater constraint qualification holds and that $\bar{x}$ and λ satisfy the KKT conditions. If $\nabla f(\bar{x}) \neq 0$ then $\bar{x}$ is a solution.

Proof. First, let us show that $\langle \nabla f(\bar{x}), x - \bar{x} \rangle \geq 0$ for all feasible x. Multiplying $x - \bar{x}$ as an inner product to the first-order condition

$$\nabla f(\bar{x}) + \sum_{i=1}^{I} \lambda_i \nabla g_i(\bar{x}) = 0,$$

we obtain

$$\langle \nabla f(\bar{x}), x - \bar{x} \rangle = -\sum_{i=1}^{I} \lambda_i \langle \nabla g_i(\bar{x}), x - \bar{x} \rangle .$$

Therefore it suffices to show $\lambda_i \langle \nabla g_i(\bar{x}), x - \bar{x} \rangle \leq 0$ for all i. If $g_i(\bar{x}) < 0$, by complementary slackness we have $\lambda_i = 0$, so there is nothing to prove. If $g_i(\bar{x}) = 0$, since x is feasible, we have $g_i(x) \leq 0 = g_i(\bar{x})$. Hence by Proposition 10.5, we have $\langle \nabla g_i(\bar{x}), x - \bar{x} \rangle \leq 0$. Since $\lambda_i \geq 0$, we have $\lambda_i \langle \nabla g_i(\bar{x}), x - \bar{x} \rangle \leq 0$.

Next, let us show that there exists a feasible point x_1 such that

$$\langle \nabla f(\bar{x}), x_1 - \bar{x} \rangle > 0.$$

Consider the point x_0 in the Slater condition. By the previous result we have $\langle \nabla f(\bar{x}), x_0 - \bar{x} \rangle \geq 0$. Since $\nabla f(\bar{x}) \neq 0$, letting $x_1 = x_0 + \epsilon \nabla f(\bar{x})$ with $\epsilon > 0$ sufficiently small, then by the Slater condition x_1 is feasible and

$$\langle \nabla f(\bar{x}), x_1 - \bar{x} \rangle \geq \epsilon \left\| \nabla f(\bar{x}) \right\|^2 > 0.$$

Finally, take any feasible x and x_1 as above. Since $\langle \nabla f(\bar{x}), x - \bar{x} \rangle \geq 0$, for any $0 < t < 1$ we have

$$\begin{aligned} \langle \nabla f(\bar{x}), (1-t)x + tx_1 - \bar{x} \rangle &= \langle \nabla f(\bar{x}), (1-t)(x - \bar{x}) + t(x_1 - \bar{x}) \rangle \\ &= (1-t) \langle \nabla f(\bar{x}), x - \bar{x} \rangle + t \langle \nabla f(\bar{x}), x_1 - \bar{x} \rangle > 0. \end{aligned}$$

Since f is quasi-convex, by Proposition 10.5 this inequality implies that

$$f((1-t)x + tx_1) > f(\bar{x}).$$

Letting $t \to 0$, we get $f(x) \geq f(\bar{x})$, so $\bar{x}$ is a solution. □

The rest of this chapter contains applications to portfolio selection and capital asset pricing model, which can be skipped by uninterested readers.

11.2 Portfolio selection

11.2.1 The problem

Suppose you live in a world with two periods, $t = 0, 1$. There are J assets indexed by $j = 1, \ldots, J$. Asset j trades at price P_j per share and it pays off X_j per share at $t = 1$, where X_j is a random variable. You have some wealth w_0 to invest at $t = 0$ and let w_1 be your wealth at $t = 1$. Assume that (as most of you do) you like money but dislike risk. So suppose you want to minimize the variance $\mathrm{Var}[w_1]$ while keeping the expected wealth $\mathrm{E}[w_1]$ at some value $\bar{w}$.

This is the classic portfolio selection problem studied by Harry Markowitz[1] in his dissertation at University of Chicago, which was eventually published in *Journal of Finance* Markowitz (1952) and made him win the Nobel Prize in 1990.

11.2.2 Mathematical formulation

To mathematically formulate the problem, let n_j be the number of shares you buy. (If $n_j > 0$, you buy asset j; if $n_j < 0$, you shortsell asset j. I assume that shortselling is allowed.) Then the problem is

$$\begin{aligned} &\text{minimize} && \mathrm{Var}\left[\sum_{j=1}^{J} X_j n_j\right] \\ &\text{subject to} && \mathrm{E}\left[\sum_{j=1}^{J} X_j n_j\right] = \bar{w}, \ \sum_{j=1}^{J} P_j n_j = w_0. \end{aligned}$$

$R_j = X_j/P_j$ be the *gross return* of asset j and let $\theta_j = P_j n_j / w_0$ be the fraction of wealth invested in asset j. Since by definition $\sum_{j=1}^{J} \theta_j = 1$ and

$$X_j n_j = \frac{X_j}{P_j} \frac{P_j n_j}{w_0} w_0 = w_0 R_j \theta_j,$$

the problem is equivalent to

$$\begin{aligned} &\text{minimize} && \mathrm{Var}[R(\theta)] \\ &\text{subject to} && \mathrm{E}[R(\theta)] = \bar{\mu}, \ \sum_{j} \theta_j = 1, \end{aligned}$$

where $\theta = (\theta_1, \ldots, \theta_J)$, $R(\theta) = \sum_j R_j \theta_j$, and $\bar{\mu} = \bar{w}/w_0$.

11.2.3 Solution

Let $\mu_j = \mathrm{E}[R_j]$ be the expected return of asset j and $\Sigma = (\sigma_{ij})$ be the variance-covariance matrix of $(R_1, \ldots, R_J)$, where

$$\sigma_{ij} = \mathrm{Cov}[R_i, R_j] = \mathrm{E}[(R_i - \mu_i)(R_j - \mu_j)].$$

[1] Harry Markowitz is currently Professor of Finance at Rady School of Management, UCSD.

Then the problem can be stated as

$$\begin{array}{ll} \text{minimize} & \dfrac{1}{2}\langle\theta, \Sigma\theta\rangle \\ \text{subject to} & \langle\mu, \theta\rangle = \bar{\mu},\ \ \langle 1, \theta\rangle = 1, \end{array}$$

where $\mu = (\mu_1, \ldots, \mu_J)$ is the vector of expected returns and $1 = (1, \ldots, 1)$ is the vector of ones. (The $\frac{1}{2}$ is there for making the first derivative nice.) Assume that Σ is positive definite. Then the objective function is strictly convex. The Lagrangian is

$$L(\theta, \lambda_1, \lambda_2) = \frac{1}{2}\langle\theta, \Sigma\theta\rangle + \lambda_1(\bar{\mu} - \langle\mu, \theta\rangle) + \lambda_2(1 - \langle 1, \theta\rangle).$$

The first-order condition is

$$\Sigma\theta - \lambda_1\mu - \lambda_2 1 = 0 \iff \theta = \lambda_1\Sigma^{-1}\mu + \lambda_2\Sigma^{-1}1.$$

The first interesting observation is that the optimal portfolio θ is spanned by two vectors, $\Sigma^{-1}\mu$ and $\Sigma^{-1}1$ (*mutual fund theorem*). Substituting θ into the constraints, we obtain

$$\begin{aligned} \lambda_1\left\langle\mu, \Sigma^{-1}\mu\right\rangle + \lambda_2\left\langle\mu, \Sigma^{-1}1\right\rangle &= \bar{\mu}, \\ \lambda_1\left\langle 1, \Sigma^{-1}\mu\right\rangle + \lambda_2\left\langle 1, \Sigma^{-1}1\right\rangle &= 1. \end{aligned}$$

Noting that $\left\langle\mu, \Sigma^{-1}1\right\rangle = \left\langle 1, \Sigma^{-1}\mu\right\rangle$ because Σ is symmetric and letting

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix} = \begin{bmatrix} \left\langle\mu, \Sigma^{-1}\mu\right\rangle & \left\langle\mu, \Sigma^{-1}1\right\rangle \\ \left\langle 1, \Sigma^{-1}\mu\right\rangle & \left\langle 1, \Sigma^{-1}1\right\rangle \end{bmatrix}^{-1},$$

we get

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}\begin{bmatrix} \bar{\mu} \\ 1 \end{bmatrix}.$$

Using the first-order condition and the constraints, the minimum variance is

$$\begin{aligned} \sigma^2 = \langle\theta, \Sigma\theta\rangle &= \langle\theta, \lambda_1\mu + \lambda_2 1\rangle = \lambda_1\bar{\mu} + \lambda_2 1 \\ &= [\bar{\mu}, 1]\begin{bmatrix} a & b \\ b & c \end{bmatrix}\begin{bmatrix} \bar{\mu} \\ 1 \end{bmatrix} = a\bar{\mu}^2 + 2b\bar{\mu} + c \\ &= a\left(\bar{\mu} + \frac{b}{a}\right)^2 + c - \frac{b^2}{a} = a(\bar{\mu} - m)^2 + s^2. \end{aligned}$$

Using

$$\frac{1}{ac - b^2}\begin{bmatrix} c & -b \\ -b & a \end{bmatrix} = \begin{bmatrix} \left\langle\mu, \Sigma^{-1}\mu\right\rangle & \left\langle\mu, \Sigma^{-1}1\right\rangle \\ \left\langle 1, \Sigma^{-1}\mu\right\rangle & \left\langle 1, \Sigma^{-1}1\right\rangle \end{bmatrix},$$

we can compute

$$\begin{aligned} m &= -\frac{b}{a} = \frac{\left\langle 1, \Sigma^{-1}\mu\right\rangle}{\left\langle 1, \Sigma^{-1}1\right\rangle}, \\ s^2 &= \frac{ac - b^2}{a} = \frac{1}{\left\langle 1, \Sigma^{-1}1\right\rangle}. \end{aligned}$$

The relationship between the target return $\bar{\mu}$ and standard deviation σ,

$$\sigma^2 = a(\bar{\mu} - m)^2 + s^2 \iff \bar{\mu} = m \pm \frac{1}{\sqrt{a}}\sqrt{\sigma^2 - s^2},$$

is a hyperbola (Figure 11.2). A portfolio satisfying this relation is called an *mean-variance efficient portfolio.*

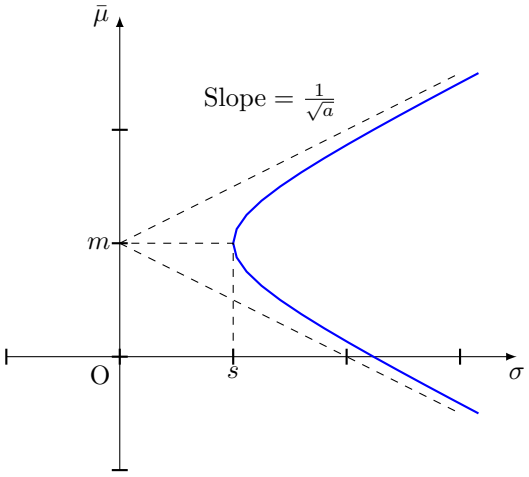

Figure 11.2. Mean-variance efficient frontier.

11.3 Capital asset pricing model (CAPM)

11.3.1 The model

Now consider an economy consisting of I agents indexed by $i = 1, \ldots, I$. Let $w_i > 0$ be the initial wealth of agent i. Instead of minimizing the variance of portfolio returns subject to a target expected portfolio returns, suppose that agent i wants to maximize

$$v_i(\theta) = \mathrm{E}[R(\theta)] - \frac{1}{2\tau_i}\mathrm{Var}[R(\theta)],$$

where $R(\theta)$ is the portfolio return and $\tau_i > 0$ is the "risk tolerance". Assume that in addition to the risky J assets, agents can trade a risk-free asset in zero net supply, where the risk-free rate R_f is determined in equilibrium. Letting θ_j the fraction of wealth invested in asset j, the fraction of wealth invested in the risk-free asset is $1 - \sum_j \theta_j$. Therefore the portfolio return is

$$R(\theta) = \sum_j R_j\theta_j + R_f\left(1 - \sum_j \theta_j\right) = R_f + \sum_j (R_j - R_f)\theta_j.$$

The expected return and variance of the portfolio are

$$\mathrm{E}[R(\theta)] = R_f + \langle \mu - R_f 1, \theta \rangle, \quad \mathrm{Var}[R(\theta)] = \langle \theta, \Sigma\theta \rangle,$$

respectively. Thus the optimal portfolio problem of agent i reduces to

$$\text{maximize } R_f + \langle \mu - R_f 1, \theta \rangle - \frac{1}{2\tau_i} \langle \theta, \Sigma\theta \rangle ,$$

where $\theta \in \mathbb{R}^J$ is unconstrained. The first-order condition is

$$\mu - R_f 1 - \frac{1}{\tau_i}\Sigma\theta = 0 \iff \theta_i = \tau_i \Sigma^{-1}(\mu - R_f 1), \tag{11.6}$$

where the subscript i indicates that it refers to agent i.

11.3.2 Equilibrium

Mathematics ends and economics starts here. Since every asset must be held by someone and the risk-free asset is in zero net supply by definition, the average portfolio weighted by individual wealth, $\sum w_i\theta_i / \sum_i w_i$, must be the *market portfolio* (value-weighted average portfolio), denoted by θ_m. Letting $\bar{\tau} = \sum_{i=1}^I w_i\tau_i / \sum_{i=1}^I w_i$ be the "average risk tolerance", it follows from the first-order condition (11.6) that

$$\theta_m = \bar{\tau}\Sigma^{-1}(\mu - R_f 1). \tag{11.7}$$

Comparing (11.6) and (11.7), we obtain

$$\theta_i = \frac{\tau_i}{\bar{\tau}}\theta_m. \tag{11.8}$$

This means that you should hold risky assets in the same proportion as in the market portfolio, where the fraction of the market portfolio is $\tau_i/\bar{\tau}$ and the fraction of the risk-free asset is $1 - \tau_i/\bar{\tau}$. Thus the mutual fund theorem holds with the market portfolio and the risk-free asset. This strong implication had an enormous impact on investment practice, and led to the creation of the first index fund in 1975 by Vanguard.[2]

11.3.3 Asset pricing

Since by definition the market portfolio invests only in risky assets, we have

$$1 = \langle 1, \theta_m \rangle = \bar{\tau} \left\langle 1, \Sigma^{-1}(\mu - R_f 1) \right\rangle \iff R_f = \frac{\left\langle 1, \Sigma^{-1}\mu \right\rangle - \frac{1}{\bar{\tau}}}{\langle 1, \Sigma^{-1} 1 \rangle},$$

which determines the risk-free rate. Since the market portfolio does not invest in the risk-free asset, it must lie on the hyperbola corresponding to no risk-free asset (Figure 11.4).

Letting $R_m = \sum_j R_j\theta_{m,j}$ be the market return, we have

$$\mathrm{Cov}[R_m, R_i] = \mathrm{Cov}\left[R_i, \sum_j R_j\theta_{m,j}\right] = (\Sigma\theta_m)_i.$$

Hence multiplying both sides of (11.7) by Σ and taking the i-th element, we obtain

$$\mathrm{Cov}[R_m, R_i] = \bar{\tau}(\mathrm{E}[R_i] - R_f). \tag{11.9}$$

[2] See http://en.wikipedia.org/wiki/Index_fund.

Multiplying both sides of (11.9) by $\theta_{m,i}$ and summing of i, we obtain

$$\mathrm{Var}[R_m] = \mathrm{Cov}[R_m, R_m] = \bar{\tau}(\mathrm{E}[R_m] - R_f). \tag{11.10}$$

Dividing (11.9) by (11.10), we obtain the *covariance pricing formula*

$$\mathrm{E}[R_i] - R_f = \frac{\mathrm{Cov}[R_m, R_i]}{\mathrm{Var}[R_m]}(\mathrm{E}[R_m] - R_f). \tag{11.11}$$

Thus, the *excess return* of an asset $\mathrm{E}[R_i] - R_f$ is proportional to the covariance of the asset with the market, $\mathrm{Cov}[R_m, R_i]$. The quantity

$$\beta_i = \frac{\mathrm{Cov}[R_m, R_i]}{\mathrm{Var}[R_m]}$$

is called the *beta* of the asset. By definition, the market beta is $\beta_m = 1$. Beta measures the *market risk* of an asset.

Rewriting (11.11), we obtain

$$\mathrm{E}[R_i] = R_f + \beta_i(\mathrm{E}[R_m] - R_f).$$

The theoretical linear relationship between β_i and $\mathrm{E}[R_i]$ is called the *security market line* (SML) (Figure 11.3). An asset above (below) the security market line, that is,

$$\mathrm{E}[R_i] > (<) R_f + \beta_i(\mathrm{E}[R_m] - R_f)$$

is undervalued (overvalued) because the expected return is higher (lower) than predicted.

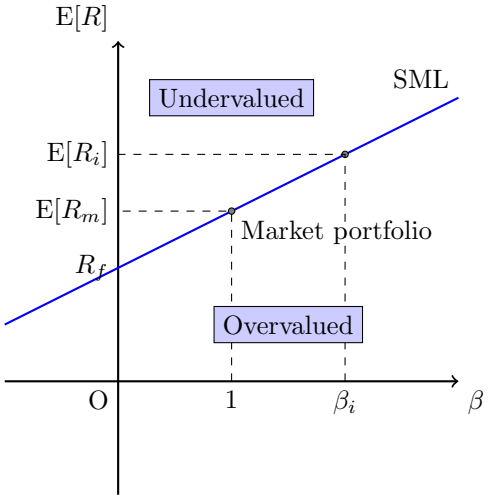

Figure 11.3. Security market line.

Since both sides of (11.11) are linear in R_i, (11.11) also holds for any linear combination (hence portfolio) of assets. In particular, letting R_i be the optimal portfolio return of agent i, $R(\theta_i)$, it follows from (11.8) that

$$\sigma_{\theta_i} = \sqrt{\mathrm{Var}[R(\theta_i)]} = \frac{\tau_i}{\bar{\tau}}\sqrt{\mathrm{Var}[R_m]},$$
$$\mathrm{Cov}[R_m, R(\theta_i)] = \frac{\tau_i}{\bar{\tau}}\,\mathrm{Var}[R_m].$$

Substituting these equations into (11.11), eliminating $\tau_i/\bar{\tau}$, and letting $\sigma_m = \sqrt{\mathrm{Var}[R_m]}$, we get

$$\mathrm{E}[R(\theta_i)] - R_f = \frac{\mathrm{E}[R_m] - R_f}{\sigma_m}\sigma_{\theta_i}.$$

This linear relationship between the standard deviation of the optimal portfolio σ_{θ_i} and the excess return $\mathrm{E}[R(\theta_i)] - R_f$ is called the *capital market line*, denoted by CML in Figure 11.4. Agents that are relatively risk tolerant ($\tau_i > \bar{\tau}$) borrow and choose a portfolio on the capital market line to the right of the market portfolio. Agents that are relatively risk averse ($\tau_i < \bar{\tau}$) lend and choose a portfolio on the capital market line to the left of the market portfolio. The slope,

$$\frac{\mathrm{E}[R_m] - R_f}{\sigma_m},$$

is called the *Sharpe ratio*, named after William Sharpe who invented it (Sharpe, 1964). The capital market line is tangent to the efficient frontier with no risk-free asset at the market portfolio. Clearly the market portfolio (and any optimal portfolio) attains the maximum Sharpe ratio.

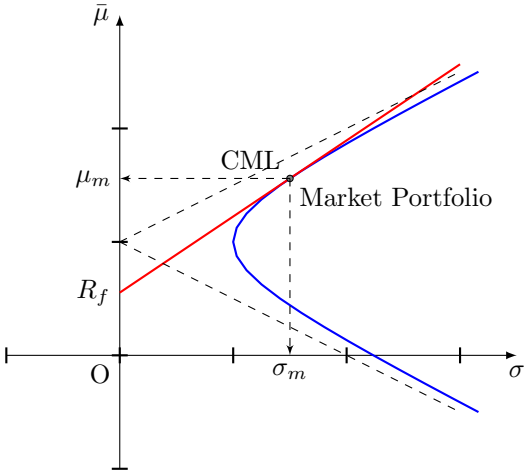

Figure 11.4. Capital market line.

Problems

11.1. Let Ω be a convex set and $f : \Omega \to (-\infty, \infty]$ be quasi-convex.

1. Show that the set of solutions to $\min_{x\in\Omega} f(x)$ is a convex set.
2. If f is strictly quasi-convex, show that the solution (if it exists) is unique.

11.2. Let $\{a_j\}_{j=1}^J$ be vectors in $\mathbb{R}^N$ and $\{b_j\}_{j=1}^J$ be scalars. Define the set

$$C = \left\{x \in \mathbb{R}^N \,\middle|\, (\forall j)\, \langle a_j, x\rangle = b_j\right\}.$$

Show that if $\{a_j\}_{j=1}^J$ are linearly dependent, then either $C = \emptyset$ or some constraints are redundant (i.e., we may drop some j without affecting the set C).

Chapter 12

Nonlinear Programming

12.1 The problem and the solution concept

We are interested in solving a general constrained optimization problem

$$\text{minimize } f(x) \text{ subject to } x \in C, \tag{12.1}$$

where f is the *objective function* and $C \subset \mathbb{R}^N$ is the *constraint set*. Such an optimization problem is called a *linear programming problem* when both the objective function and the constraints are linear. Otherwise, it is called a *nonlinear programming problem*. If both the objective function and the constraints happen to be convex, it is called a *convex programming problem*. Thus

Linear Programming $\subset$ Convex Programming $\subset$ Nonlinear Programming.

We focus on minimization because maximizing f is the same as minimizing $-f$. If f is continuous and C is compact, by the extreme value theorem (Theorem 2.5) we know there is a solution, but the theorem does not tell you how to compute it. In this chapter you will learn how to derive necessary conditions for optimality for general nonlinear programming problems. Oftentimes, the necessary conditions alone will pin down the solution to a few candidates, so you only need to compare these candidates.

We call the point $\bar{x} \in C$ a *global* solution if $f(x) \geq f(\bar{x})$ for all $x \in C$ and a *local solution* if there exists $\epsilon > 0$ such that $f(x) \geq f(\bar{x})$ whenever $x \in C$ and $\|x - \bar{x}\| < \epsilon$. If $\bar{x}$ is a global solution, clearly it is also a local solution.

12.2 Cone and dual cone

We first introduce some mathematical concepts, cones and dual cones.

A set $C \subset \mathbb{R}^N$ is said to be a *cone* if it contains a ray originating from 0 and passing through any point of C. Formally, C is a cone if $x \in C$ and $\alpha \geq 0$ implies $\alpha x \in C$. An example of a cone is the nonnegative orthant

$$\mathbb{R}^N_+ = \left\{ x = (x_1, \ldots, x_N) \in \mathbb{R}^N \,\middle|\, (\forall n) x_n \geq 0 \right\}.$$

Another example is the set

$$\left\{ x = \sum_{k=1}^{K} \alpha_k a_k \,\middle|\, (\forall k) \alpha_k \geq 0 \right\}, \tag{12.2}$$

where $a_1, \ldots, a_K$ are fixed vectors. The set (12.2) is called the *polyhedral cone* generated by vectors $a_1, \ldots, a_K$, and is denoted by $\operatorname{cone}[a_1, \ldots, a_K]$ (Figure 12.1). Clearly $\mathbb{R}^N_+ = \operatorname{cone}[e_1, \ldots, e_N]$, where $e_1, \ldots, e_N$ are unit vectors of $\mathbb{R}^N$. A polyhedral cone is a closed convex cone (Problem 12.1).

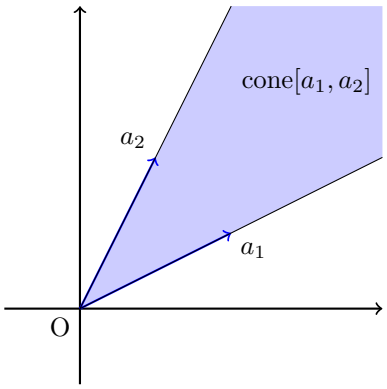

Figure 12.1. Cone generated by vectors.

Let $C \subset \mathbb{R}^N$ be any nonempty set. The set

$$C^* = \left\{ y \in \mathbb{R}^N \,\middle|\, (\forall x \in C)\, \langle x, y \rangle \leq 0 \right\} \tag{12.3}$$

is called the *dual cone* of C. Thus the dual cone C^* consists of all vectors that make an obtuse angle with any vector in C (Figure 12.2).

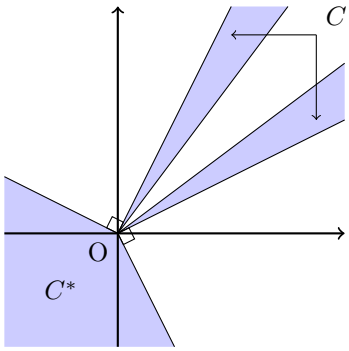

Figure 12.2. Cone and its dual.

Note that in the definition of the dual cone (12.3), the set C is arbitrary (not necessarily a cone). Yet, C^* is called the dual cone, which suggests that C^* is always a cone. In fact this is the case, and the following proposition proves some basic properties of the dual cone.

Proposition 12.1. *Let $\emptyset \neq C \subset D$. Then (i) the dual cone C^* is a nonempty, closed, convex cone, (ii) $C^* = (\operatorname{co} C)^*$, and (iii) $C^* \supset D^*$.*

Proof. C^* is nonempty since $0 \in C^*$. If $y \in C^*$, then by definition $\langle x, y \rangle \leq 0$ for all $x \in C$. Then for any $\alpha \geq 0$ and $x \in C$, we have $\langle x, \alpha y \rangle = \alpha \langle x, y \rangle \leq 0$, so $\alpha y \in C^*$. Therefore C^* is a cone. To show that C^* is closed, take any sequence

$\{y_k\}_{k=1}^{\infty} \subset C^*$ and $y_k \to y$. Since $y_k \in C^*$, by definition $\langle x, y_k \rangle \leq 0$ for all $x \in C$, so letting $k \to \infty$ we obtain $\langle x, y \rangle \leq 0$ for all $x \in C$. Therefore $y \in C^*$ and hence C^* is closed.

If $y \in D^*$, then $\langle x, y \rangle \leq 0$ for all $x \in D$. Since $C \subset D$, we have $\langle x, y \rangle \leq 0$ for all $x \in C$. Therefore $y \in C^*$, which proves $D^* \subset C^*$. Finally, since $C \subset \operatorname{co} C$, we have $C^* \supset (\operatorname{co} C)^*$ by letting $D = \operatorname{co} C$. To prove the reverse inclusion, take any $x \in \operatorname{co} C$. By Lemma 9.1, there exists a convex combination $x = \sum_{k=1}^{K} \alpha_k x_k$ such that $x_k \in C$ for all k. If $y \in C^*$, it follows that

$$\langle x, y \rangle = \left\langle \sum \alpha_k x_k, y \right\rangle = \sum \alpha_k \langle x_k, y \rangle \leq 0,$$

so $y \in (\operatorname{co} C)^*$. Therefore $C^* \subset (\operatorname{co} C)^*$. □

Proposition 12.2. *Let $C \subset \mathbb{R}^N$ be a nonempty cone. Then $C^{**} = \operatorname{cl} \operatorname{co} C$. ($C^{**}$ is the dual cone of C^*, so it is the dual cone of the dual cone of C.)*

Proof. Let $x \in C$. For any $y \in C^*$ we have $\langle x, y \rangle \leq 0$. This implies $x \in C^{**}$. Hence $C \subset C^{**}$. Since by Proposition 12.1 the dual cone is closed and convex, we have $\operatorname{cl} \operatorname{co} C \subset C^{**}$.

To show the reverse inclusion, suppose that $x \notin \operatorname{cl} \operatorname{co} C$. Then by Proposition 9.4 there exists a nonzero vector a such that

$$\sup_{z \in \operatorname{cl} \operatorname{co} C} \langle a, z \rangle < c < \langle a, x \rangle .$$

In particular, $\langle a, z \rangle < c$ for all $z \in C$. Since C is a cone, it must be $\langle a, z \rangle \leq 0$ for all z. To see this, suppose there exists $z_0 \in C$ with $\langle a, z_0 \rangle > 0$. Then for any $\beta > 0$ we have $\beta z_0 \in C$, so letting β large enough we have $\langle a, \beta z_0 \rangle = \beta \langle a, z_0 \rangle > c$, a contradiction. Since $\langle a, z \rangle \leq 0$ for all $z \in C$, we have $a \in C^*$. Again since C is a cone, we have $0 \in C$, so $c > \langle a, 0 \rangle = 0$. Since $\langle a, x \rangle > c > 0$ and $a \in C^*$, it follows that $x \notin C^{**}$. Therefore $C^{**} \subset \operatorname{cl} \operatorname{co} C$. □

The following corollary plays an important role in optimization theory.

Corollary 12.3 (Farkas)**.** *Let $C = \operatorname{cone}[a_1, \ldots, a_K]$ be the polyhedral cone generated by the vectors $a_1, \ldots, a_K$. Let $D = \left\{ y \in \mathbb{R}^N \,\middle|\, (\forall k) \langle a_k, y \rangle \leq 0 \right\}$. Then $D = C^*$ and $C = D^*$ (Figure 12.3).*

Proof. Let $y \in D$. For any $x \in C$, we can take $\{\alpha_k\}_{k=1}^{K} \subset \mathbb{R}_+$ such that $x = \sum_k \alpha_k a_k$. Then

$$\langle x, y \rangle = \sum_k \alpha_k \langle a_k, y \rangle \leq 0,$$

so $y \in C^*$. Conversely, let $y \in C^*$. Since $a_k \in C$, we get $\langle a_k, y \rangle \leq 0$ for all k, so $y \in D$. Therefore $D = C^*$.

Since C is a closed convex cone, by Propositions 12.2 and 12.1 (iii) we get

$$C = \operatorname{cl} \operatorname{co} C = C^{**} = (C^*)^* = D^*. \quad \square$$

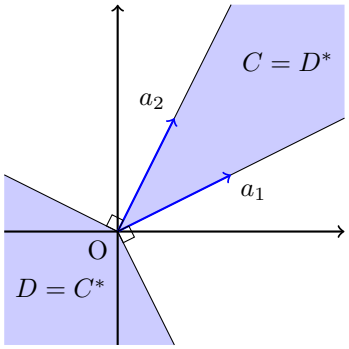

Figure 12.3. Farkas' lemma.

12.3 Necessary condition

In this section we derive the first-order necessary condition for optimality using the tangent cone of the constraint set.

Let $C \subset \mathbb{R}^N$ be any nonempty set and $\bar{x} \in C$ be any point. The *tangent cone* of C at $\bar{x}$ is defined by

$$T_C(\bar{x}) = \left\{ y \in \mathbb{R}^N \;\middle|\; (\exists)\, \{\alpha_k\} \geq 0, \{x_k\} \subset C, \lim_{k\to\infty} x_k = \bar{x}, y = \lim_{k\to\infty} \alpha_k(x_k - \bar{x}) \right\}.$$

That is, $y \in T_C(\bar{x})$ if y points to the same direction as the limiting direction of $\{x_k - \bar{x}\}$ as x_k approaches to $\bar{x}$. Intuitively, the tangent cone of C at $\bar{x}$ consists of all directions that can be approximated by that from $\bar{x}$ to another point in C. Figure 12.4 shows an example. Here C is the region in between the two curves, and the tangent cone is the shaded area.

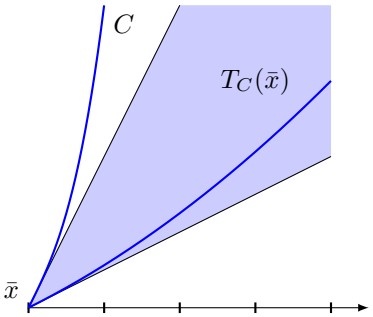

Figure 12.4. Tangent cone.

Lemma 12.4. *$T_C(\bar{x})$ is a nonempty closed cone.*

Proof. Setting $\alpha_k = 0$ for all k we get $0 \in T_C(\bar{x})$, so $T_C(\bar{x}) \neq \emptyset$. If $y \in T_C(\bar{x})$, then $y = \lim \alpha_k(x_k - \bar{x})$ for some $\{\alpha_k\} \geq 0$ and $\{x_k\} \subset C$ such that $\lim x_k = \bar{x}$. Then for $\beta \geq 0$ we have $\beta y = \lim \beta\alpha_k(x_k - \bar{x}) \in T_C(\bar{x})$, so $T_C(\bar{x})$ is a cone. To show that $T_C(\bar{x})$ is closed, let $\{y_l\} \subset T_C(\bar{x})$ and $y_l \to \bar{y}$. For each l we can take a sequence such that $\alpha_{k,l} \geq 0$, $\lim_{k\to\infty} x_{k,l} = \bar{x}$, and $y_l = \lim_{k\to\infty} \alpha_{k,l}(x_{k,l} - \bar{x})$.

Hence we can take k_l such that $\|x_{k_l,l} - \bar{x}\| < 1/l$ and $\|y_l - \alpha_{k_l,l}(x_{k_l,l} - \bar{x})\| < 1/l$. Then $x_{k_l,l} \to \bar{x}$ and

$$\|\bar{y} - \alpha_{k_l,l}(x_{k_l,l} - \bar{x})\| \le \|\bar{y} - y_l\| + \|y_l - \alpha_{k_l,l}(x_{k_l,l} - \bar{x})\| \to 0,$$

so $\bar{y} \in T_C(\bar{x})$. □

The dual cone of $T_C(\bar{x})$ is called the *normal cone at* $\bar{x}$ and is denoted by $N_C(\bar{x})$ (Figure 12.5). By the definition of the dual cone, we have

$$N_C(\bar{x}) = (T_C(\bar{x}))^* = \left\{ z \in \mathbb{R}^N \,\middle|\, (\forall y \in T_C(\bar{x}))\, \langle y, z \rangle \le 0 \right\}.$$

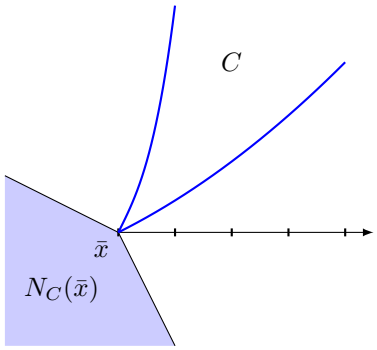

Figure 12.5. Normal cone.

The following theorem is fundamental for constrained optimization.

Theorem 12.5. *If f is differentiable and $\bar{x}$ is a local solution to the problem*

$$\text{minimize } f(x) \text{ subject to } x \in C,$$

then $-\nabla f(\bar{x}) \in N_C(\bar{x})$.

Proof. By the definition of the normal cone, it suffices to show that

$$\langle -\nabla f(\bar{x}), y \rangle \le 0 \iff \langle \nabla f(\bar{x}), y \rangle \ge 0$$

for all $y \in T_C(\bar{x})$. Let $y \in T_C(\bar{x})$ and take a sequence such that $\alpha_k \ge 0$, $x_k \to \bar{x}$, and $\alpha_k(x_k - \bar{x}) \to y$. Since $\bar{x}$ is a local solution, for sufficiently large k we have $f(x_k) \ge f(\bar{x})$. Since f is differentiable, we have

$$0 \le f(x_k) - f(\bar{x}) = \langle \nabla f(\bar{x}), x_k - \bar{x} \rangle + o(\|x_k - \bar{x}\|).^{1}$$

Multiplying both sides by $\alpha_k \ge 0$ and letting $k \to \infty$, we get

$$\begin{aligned} 0 &\le \langle \nabla f(\bar{x}), \alpha_k(x_k - \bar{x}) \rangle + \|\alpha_k(x_k - \bar{x})\| \cdot \frac{o(\|x_k - \bar{x}\|)}{\|x_k - \bar{x}\|} \\ &\to \langle \nabla f(\bar{x}), y \rangle + \|y\| \cdot 0 = \langle \nabla f(\bar{x}), y \rangle. \end{aligned}$$

□

The geometric interpretation of Theorem 12.5 is the following. By the discussion around (4.3), $-\nabla f(\bar{x})$ is the direction towards which f decreases fastest around the point $\bar{x}$. The tangent cone $T_C(\bar{x})$ consists of directions towards which x can move around $\bar{x}$ without violating the constraint $x \in C$. Hence in order for $\bar{x}$ to be a local minimum, $-\nabla f(\bar{x})$ must make an obtuse angle with any vector in the tangent cone, for otherwise f can be decreased further. This is the same as $-\nabla f(\bar{x})$ belonging to the normal cone.

[1] $o(h)$ represents any quantity $q(h)$ such that $q(h)/h \to 0$ as $h \to 0$.

12.4 Karush-Kuhn-Tucker theorem

Theorem 12.5 is very general. Usually, we are interested in the cases where the constraint set C is given parametrically. Consider the minimization problem

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & g_i(x) \leq 0 & (i = 1, \dots, I), \\ & h_j(x) = 0 & (j = 1, \dots, J). \end{array} \tag{12.4}$$

This problem is a special case of problem (12.1) by setting

$$C = \left\{ x \in \mathbb{R}^N \,\middle|\, (\forall i) g_i(x) \leq 0, (\forall j) h_j(x) = 0 \right\}.$$

$g_i(x) \leq 0$ is called an *inequality constraint*. $h_j(x) = 0$ is an *equality constraint*. Let $\bar{x} \in C$ be a local solution. To study the shape of C around $\bar{x}$, we define as follows. The set of indices for which the inequality constraints are binding,

$$I(\bar{x}) = \{ i \,|\, g_i(\bar{x}) = 0 \},$$

is called the *active set*. Assume that g_i's and h_j's are differentiable. The set

$$L_C(\bar{x}) = \left\{ y \in \mathbb{R}^N \,\middle|\, (\forall i \in I(\bar{x})) \langle \nabla g_i(\bar{x}), y \rangle \leq 0, (\forall j) \langle \nabla h_j(\bar{x}), y \rangle = 0 \right\} \tag{12.5}$$

is called the *linearizing cone* of the constraints g_i's and h_j's. The reason why $L_C(\bar{x})$ is called the linearizing cone is the following. Since

$$g_i(\bar{x} + ty) - g_i(\bar{x}) = t \langle \nabla g_i(\bar{x}), y \rangle + o(t),$$

the point $x = \bar{x} + ty$ almost satisfies the constraint $g_i(x) \leq 0$ if $g_i(\bar{x}) = 0$ (i is an active constraint) and $\langle \nabla g_i(\bar{x}), y \rangle \leq 0$. The same holds for h_j's. Thus $y \in L_C(\bar{x})$ implies that from $\bar{x}$ we can move slightly towards the direction of y and still (approximately) satisfy the constraints. Thus we can expect that the linearizing cone is approximately equal to the tangent cone. The following proposition make this statement precise.

Proposition 12.6. *Suppose that $\bar{x} \in C$. Then $\operatorname{co} T_C(\bar{x}) \subset L_C(\bar{x})$.*

Proof. Clearly the linearizing cone (12.5) is a closed convex cone, so it suffices to prove $T_C(\bar{x}) \subset L_C(\bar{x})$. Let $y \in T_C(\bar{x})$. Take $\{x_k\} \subset C$ and $\{\alpha_k\} \subset \mathbb{R}_+$ such that $x_k \to \bar{x}$ and $\alpha_k(x_k - \bar{x}) \to y$. Since $g_i(\bar{x}) = 0$ for $i \in I(\bar{x})$ and g_i is differentiable, we get

$$0 \geq g_i(x_k) = g_i(x_k) - g_i(\bar{x}) = \langle \nabla g_i(\bar{x}), x_k - \bar{x} \rangle + o(\|x_k - \bar{x}\|).$$

Multiplying both sides by $\alpha_k \geq 0$ and letting $k \to \infty$, we get

$$\begin{aligned} 0 &\geq \langle \nabla g_i(\bar{x}), \alpha_k(x_k - \bar{x}) \rangle + \|\alpha_k(x_k - \bar{x})\| \cdot \frac{o(\|x_k - \bar{x}\|)}{\|x_k - \bar{x}\|} \\ &\to \langle \nabla g_i(\bar{x}), y \rangle + \|y\| \cdot 0 = \langle \nabla g_i(\bar{x}), y \rangle. \end{aligned}$$

A similar argument applies to h_j. Hence $y \in L_C(\bar{x})$. □

Note that while the tangent cone is directly defined by the constraint set C, the linearizing cone is defined through the functions that define the set C. Therefore different parametrizations of the same set C may lead to different linearizing cones (Problem 12.3).

The main result in static optimization is the following.

Theorem 12.7 (Karush-Kuhn-Tucker)**.** *Suppose that f, g_i, h_j are differentiable and $\bar{x}$ is a local solution to the minimization problem (12.4). If $L_C(\bar{x}) \subset \operatorname{co} T_C(\bar{x})$, then there exist vectors (called* Lagrange multipliers*) $\lambda \in \mathbb{R}^I_+$ and $\mu \in \mathbb{R}^J$ such that*

$$\nabla f(\bar{x}) + \sum_{i=1}^{I} \lambda_i \nabla g_i(\bar{x}) + \sum_{j=1}^{J} \mu_j \nabla h_j(\bar{x}) = 0, \tag{12.6a}$$

$$(\forall i)\ \lambda_i g_i(\bar{x}) = 0. \tag{12.6b}$$

Proof. By Theorem 12.5, $-\nabla f(\bar{x}) \in N_C(\bar{x}) = (T_C(\bar{x}))^*$. By Proposition 12.6 and the assumption $L_C(\bar{x}) \subset \operatorname{co} T_C(\bar{x})$, we get $L_C(\bar{x}) = \operatorname{co} T_C(\bar{x})$. Hence by the property of dual cones, we get $(T_C(\bar{x}))^* = (\operatorname{co} T_C(\bar{x}))^* = (L_C(\bar{x}))^*$. Now let K be the polyhedral cone generated by $\{\nabla g_i(\bar{x})\}_{i \in I(\bar{x})}$ and $\{\pm \nabla h_j(\bar{x})\}_{j=1}^J$. By Farkas's lemma (Corollary 12.3), K^* is equal to

$$\left\{ y \in \mathbb{R}^N \,\middle|\, (\forall i \in I(\bar{x}))\, \langle \nabla g_i(\bar{x}), y \rangle \leq 0, (\forall j)\, \langle \pm \nabla h_j(\bar{x}), y \rangle \leq 0 \right\},$$

which is precisely the linearizing cone $L_C(\bar{x})$. Again by Farkas's lemma, we have $(L_C(\bar{x}))^* = K$. Therefore $-\nabla f(\bar{x}) \in K$, so there exist numbers $\lambda_i \geq 0$ $(i \in I(\bar{x}))$ and $\alpha_j, \beta_j \geq 0$ such that

$$-\nabla f(\bar{x}) = \sum_{i \in I(\bar{x})} \lambda_i \nabla g_i(\bar{x}) + \sum_{j=1}^{J} (\alpha_j - \beta_j) \nabla h_j(\bar{x}).$$

Letting $\lambda_i = 0$ for $i \notin I(\bar{x})$ and $\mu_j = \alpha_j - \beta_j$, we get (12.6a). Finally, (12.6b) holds for $i \in I(\bar{x})$ since $g_i(\bar{x}) = 0$. It also holds for $i \in I(\bar{x})$ since we defined $\lambda_i = 0$ for such i. □

Here is an easy way to remember the conditions in (12.6). Define the *Lagrangian* of the minimization problem 12.4 by

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{I} \lambda_i g_i(x) + \sum_{j=1}^{J} \mu_j h_j(x),$$

which is the sum of the objective function $f(x)$ and the constraint functions $g_i(x)$, $h_j(x)$ weighted by the Lagrange multipliers λ_i, μ_j. Then (12.6a) implies that the derivative of $L(\cdot, \lambda, \mu)$ at $\bar{x}$ is zero. (12.6a) is called the *first-order condition*. (12.6b) is called the *complementary slackness condition*. Together, (12.6a) and (12.6b) are called *Karush-Kuhn-Tucker (KKT) conditions*.

12.5 Constraint qualifications

Conditions of the form $L_C(\bar{x}) \subset \operatorname{co} T_C(\bar{x})$ in Theorem 12.7 are called *constraint qualifications* (CQ). These are necessary conditions in order for the KKT conditions to hold. There are many constraint qualifications in the literature:

Guignard (GCQ) $L_C(\bar{x}) \subset \operatorname{co} T_C(\bar{x})$.

Abadie (ACQ) $L_C(\bar{x}) \subset T_C(\bar{x})$.

Mangasarian-Fromovitz (MFCQ) $\{\nabla h_j(\bar{x})\}_{j=1}^J$ are linearly independent, and there exists $y \in \mathbb{R}^N$ such that $\langle \nabla g_i(\bar{x}), y\rangle < 0$ for all $i \in I(\bar{x})$ and $\langle \nabla h_j(\bar{x}), y\rangle = 0$ for all j.

Slater (SCQ) g_i's are convex, $h_j(x) = \langle a_j, x\rangle - c_j$ where $\{a_j\}_{j=1}^J$ are linearly independent, and there exists $x_0 \in \mathbb{R}^N$ such that $g_i(x_0) < 0$ for all i and $h_j(x_0) = 0$ for all j.

Linear independence (LICQ) $\{\nabla g_i(\bar{x})\}_{i\in I(\bar{x})}$ and $\{\nabla h_j(\bar{x})\}_{j=1}^J$ are linearly independent.

The point of listing these constraint qualifications is that some of them are general but hard to verify (GCQ and ACQ), while others are special but easy to verify (SCQ and LICQ). Users of the KKT theorem need to pick the appropriate constraint qualification for the problem under consideration. The following theorem shows the relation between these constraint qualifications.

Theorem 12.8. *The following is true for constraint qualifications.*

$$\textit{LICQ or SCQ} \implies \text{MFCQ} \implies \text{ACQ} \implies \text{GCQ}.$$

Proof.

ACQ $\implies$ GCQ. Trivial because $T_C(\bar{x}) \subset \operatorname{co} T_C(\bar{x})$.

MFCQ $\implies$ ACQ. By dropping non-binding constraints, without loss of generality we may assume all the constraints bind, so $I(\bar{x}) = \{1, \dots, I\}$. Define $G : \mathbb{R}^N \to \mathbb{R}^I$ by $G(x) = (g_1(x), \dots, g_I(x))'$ and $H : \mathbb{R}^N \to \mathbb{R}^J$ by $H(x) = (h_1(x), \dots, h_J(x))'$. Then MFCQ holds if and only if the $J \times N$ Jacobian $DH(\bar{x})$ has full row rank and there exists $y \in \mathbb{R}^N$ such that $[DG(\bar{x})]y \ll 0$ and $[DH(\bar{x})]y = 0$, where $v \ll 0$ means that all entries of v are strictly negative.

Define the set

$$\tilde{L}_C(\bar{x}) = \left\{ y \in \mathbb{R}^N \,\middle|\, [DG(\bar{x})]y \ll 0, [DH(\bar{x})]y = 0 \right\}.$$

Since MFCQ holds, by definition we have $\tilde{L}_C(\bar{x}) \neq \emptyset$. Since $\operatorname{cl} \tilde{L}_C(\bar{x}) = L_C(\bar{x})$ by the definition of the linearizing cone, and since $T_C(\bar{x})$ is closed, it suffices to show $\tilde{L}_C(\bar{x}) \subset T_C(\bar{x})$.

Since $DH(\bar{x})$ has full row rank, by relabeling the variables if necessary, we may assume that we can split the variables as $x = (x_1, x_2) \in \mathbb{R}^{N-J} \times \mathbb{R}^J$ and write $DH(\bar{x}) = [D_{x_1}H, D_{x_2}H]$, where $D_{x_2}H = D_{x_2}H(\bar{x})$ is regular. By the implicit function theorem, for x close enough to $\bar{x}$, we can write

$$0 = H(x) = H(x_1, x_2) \iff x_2 = \phi(x_1),$$

where ϕ is C^1 and $D\phi = -[D_{x_2}H]^{-1} D_{x_1}H$.

Take any $y = (y_1, y_2) \in \tilde{L}_C(\bar{x})$, where $y_1 \in \mathbb{R}^{N-J}$ and $y_2 \in R^J$. For small $t > 0$, define

$$x(t) = (x_1(t), x_2(t)) = (\bar{x}_1 + ty_1, \phi(\bar{x}_1 + ty_1)).$$

Let us show that $x(0) = \bar{x}$, $x(t) \in C$ for small $t > 0$, and $x'(0) = y$, which implies that $y \in T_C(\bar{x})$.

Since $H(\bar{x}) = 0$, by the implicit function theorem we have $x(0) = (\bar{x}_1, \phi(\bar{x}_1)) = (\bar{x}_1, \bar{x}_2) = \bar{x}$.

Using the chain rule, we obtain $x'(0) = (y_1, [D\phi]y_1)$. Since $y \in \tilde{L}_C(\bar{x})$, it follows that

$$0 = [DH(\bar{x})]y = [DH_{x_1}]y_1 + [DH_{x_2}]y_2 \iff y_2 = -[DH_{x_2}]^{-1}[DH_{x_1}]y_1 = [D\phi]y_1$$

by the implicit function theorem. Therefore $x'(0) = (y_1, y_2) = y$.

Finally, by the chain rule and the definition of $\tilde{L}_C(\bar{x})$, at $t = 0$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t} G(x(0)) = [DG(\bar{x})]x'(0) = [DG(\bar{x})]y \ll 0.$$

Therefore for small enough $t > 0$, we have

$$\frac{G(x(t))}{t} = \frac{G(x(t)) - G(\bar{x})}{t} \ll 0$$

because $G(\bar{x}) = 0$, so $G(x(t)) \ll 0$. Since $H(x(t)) = H(x_1(t), \phi(x_1(t))) = 0$, it follows that $x(t) \in C$ for small enough $t > 0$.

SCQ $\implies$ MFCQ. Suppose that g_i's are convex, $h_j(x) = \langle a_j, x\rangle - c_j$ where $\{a_j\}_{j=1}^J$ are linearly independent, and there exists $x_0 \in \mathbb{R}^N$ such that $g_i(x_0) < 0$ for all i and $h_j(x_0) = 0$ for all j.

Since $\nabla h_j = a_j$ and $\{a_j\}_{j=1}^J$ are linearly independent, $\{\nabla h_j(\bar{x})\}_{j=1}^J$ are linearly independent. If $i \in I(\bar{x})$, since g_i is convex, by Proposition 10.3 we have

$$0 > g_i(x_0) = g_i(x_0) - g_i(\bar{x}) \geq \langle \nabla g_i(\bar{x}), x_0 - \bar{x}\rangle .$$

Setting $y = x_0 - \bar{x}$, we have $\langle \nabla g_i(\bar{x}), y\rangle < 0$ for all $i \in I(\bar{x})$. Since $\bar{x}, x_0$ are feasible, we have $\langle a_j, \bar{x}\rangle - c_j = 0$ and $\langle a_j, x_0\rangle - c_j = 0$, so taking the difference $\langle \nabla h_j(\bar{x}), y\rangle = \langle a_j, x_0 - \bar{x}\rangle = 0$. Therefore MFCQ holds.

LICQ $\implies$ MFCQ. As in the previous case we may assume $I(\bar{x}) = \{1, \ldots, I\}$. Suppose on the contrary that MFCQ does not hold. Then there exist no y such that $\langle \nabla g_i(\bar{x}), y\rangle < 0$ for all i and $\langle \nabla h_j(\bar{x}), y\rangle = 0$ for all j. Let $G(x) = (g_1(x), \ldots, g_I(x))'$, $H(x) = (h_1(x), \ldots, h_J(x))'$, and define the $(I + J) \times N$ matrix M by $M = \begin{bmatrix} DG(\bar{x}) \\ DH(\bar{x}) \end{bmatrix}$. Define the sets $A, B \subset \mathbb{R}^{I+J}$ by

$$A = -\mathbb{R}^I_{++} \times \{0\} \subset \mathbb{R}^I \times \mathbb{R}^J,$$
$$B = \left\{ z \in \mathbb{R}^{I+J} \,\middle|\, (\exists y \in \mathbb{R}^N) z = My \right\}.$$

Since MFCQ does not hold, we have $A \cap B = \emptyset$. Clearly A, B are nonempty and convex. By the separating hyperplane theorem, there exists $0 \neq a \in \mathbb{R}^{I+J}$ such that

$$\sup_{z \in A} \langle a, z\rangle \leq \inf_{z \in B} \langle a, z\rangle = \inf_{y \in \mathbb{R}^N} a'My.$$

Since $y \mapsto a'My$ is linear and $\sup_{z \in A} \langle a, z\rangle > -\infty$ because $A \neq \emptyset$, in order for the above inequality to hold, it is necessary that $a'M = 0$. Letting $a = (\lambda, \mu) \in \mathbb{R}^I \times \mathbb{R}^J$, then

$$0 = M'a = \sum_{i=1}^I \lambda_i \nabla g_j(\bar{x}) + \sum_{j=1}^J \mu_j \nabla h_j(\bar{x}).$$

Since $a = (\lambda, \mu) \neq 0$, $\{\nabla g_i(\bar{x})\}_{i=1}^I$ and $\{\nabla h_j(\bar{x})\}_{j=1}^J$ are not linearly independent. Hence LICQ does not hold. □

In applications, oftentimes constraints are linear. In that case GCQ is automatically satisfied, so there is no need to check it (Problem 12.4). It is known that the GCQ is the weakest possible condition (Gould and Tolle, 1971).

12.6 Sufficient condition

The Karush-Kuhn-Tucker theorem provides necessary conditions for optimality: if the constraint qualification holds, then a local solution must satisfy the Karush-Kuhn-Tucker conditions (first-order conditions and complementary slackness conditions). Note that the KKT conditions are equivalent to

$$\nabla_x L(\bar{x}, \lambda, \mu) = 0, \tag{12.7}$$

where $L(x, \lambda, \mu)$ is the Lagrangian. (12.7) is the first-order necessary condition of the unconstrained minimization problem

$$\min_{x \in \mathbb{R}^N} L(x, \lambda, \mu). \tag{12.8}$$

Below I give a sufficient condition for optimality.

Proposition 12.9. *Suppose that $\bar{x}$ is a solution to the unconstrained minimization problem (12.8) for some $\lambda \in \mathbb{R}^I_+$ and $\mu \in \mathbb{R}^J$. If $g_i(\bar{x}) \leq 0$ and $\lambda_i g_i(\bar{x}) = 0$ for all i and $h_j(\bar{x}) = 0$ for all j, then $\bar{x}$ is a solution to the constrained minimization problem (12.1).*

Proof. Take any x such that $g_i(x) \leq 0$ for all i and $h_j(x) = 0$ for all j. Then

$$\begin{aligned} f(\bar{x}) &= f(\bar{x}) + \sum_{i=1}^I \lambda_i g_i(\bar{x}) + \sum_{j=1}^J \mu_j h_j(\bar{x}) \\ &= L(\bar{x}, \lambda, \mu) \leq L(x, \lambda, \mu) \\ &= f(x) + \sum_{i=1}^I \lambda_i g_i(x) + \sum_{j=1}^J \mu_j h_j(x) \leq f(x). \end{aligned}$$

The first line is due to $\lambda_i g_i(\bar{x}) = 0$ for all i and $h_j(\bar{x}) = 0$ for all j. The second line is the assumption that $\bar{x}$ minimizes $L(\cdot, \lambda, \mu)$. The third line is due to $\lambda_i \geq 0$ and $g_i(x) \leq 0$ for all i and $h_j(x) = 0$ for all j. □

Corollary 12.10. *If f, g_i are all convex, then the KKT conditions are sufficient for optimality.*

Next I give a second order sufficient condition that will be useful later. Consider the minimization problem (12.4). Assume that the KKT conditions (12.6) hold at $\bar{x}$ with corresponding Lagrange multipliers $\lambda \in \mathbb{R}^I_+$ and $\mu \in \mathbb{R}^J$. Remember that the *active set* of the inequality constraints is $I(\bar{x}) = \{i \,|\, g_i(\bar{x}) = 0\}$.

Let $\tilde{I}(\bar{x}) = \{i \,|\, \lambda_i > 0\}$ be the set of constraints such that the Lagrange multiplier is positive. Since $\lambda_i g_i(\bar{x}) = 0$ by complementary slackness, $\lambda_i > 0$ implies $g_i(\bar{x}) = 0$, so necessarily $\tilde{I}(\bar{x}) \subset I(\bar{x})$. Define the cone

$$\tilde{L}_C(\bar{x}) = \Big\{ y \in \mathbb{R}^N \Big| (\forall i \in I(\bar{x}) \backslash \tilde{I}(\bar{x}))\, \langle \nabla g_i(\bar{x}), y \rangle \le 0, \\ (\forall i \in \tilde{I}(\bar{x}))\, \langle \nabla g_i(\bar{x}), y \rangle = 0, (\forall j)\, \langle \nabla h_j(\bar{x}), y \rangle = 0 \Big\}.$$

Clearly $\tilde{L}_C(\bar{x}) \subset L_C(\bar{x})$. The following theorem gives a second order sufficient condition for local optimality.

Theorem 12.11. *Suppose that f, g_i's, and h_j's are twice differentiable, the KKT conditions (12.6) hold at $x = \bar{x}$, and*

$$\left\langle y, \nabla_x^2 L(\bar{x}, \lambda, \mu) y \right\rangle > 0$$

for all $0 \neq y \in \tilde{L}_C(\bar{x})$. Then $\bar{x}$ is a strict local solution to the minimization problem (12.4), i.e., there exists a neighborhood Ω of $\bar{x}$ such that $f(\bar{x}) < f(x)$ whenever $x \in \Omega$ satisfies the constraints in (12.4).

Proof. Suppose that $\bar{x}$ is not a strict local solution. Then we can take a sequence $C \ni x^k \to \bar{x}$ such that $f(x^k) \le f(\bar{x})$. Let $\alpha_k = 1/\left\|x^k - \bar{x}\right\| > 0$. Then $\left\|\alpha_k (x^k - \bar{x})\right\| = 1$, so by taking a subsequence if necessary we may assume $\alpha_k (x^k - \bar{x}) \to y$ with $\|y\| = 1$. Let us show that $y \in \tilde{L}_C(\bar{x})$.

Multiplying both sides of

$$f(x^k) - f(\bar{x}) \le 0,\ g_i(x^k) - g_i(\bar{x}) \le 0\ (i \in I(\bar{x})),\ h_j(x^k) - h_j(\bar{x}) = 0$$

by α_k and letting $k \to \infty$, we get

$$\langle \nabla f(\bar{x}), y \rangle \le 0,\ \langle \nabla g_i(\bar{x}), y \rangle \le 0\ (i \in I(\bar{x})),\ \langle \nabla h_j(\bar{x}), y \rangle = 0. \tag{12.9}$$

Multiplying both sides of the first-order condition (12.6a) by y as an inner product, noting that $\lambda_i = 0$ if $i \notin I(\bar{x})$ by complementary slackness, and using (12.9), we get

$$\langle \nabla f(\bar{x}), y \rangle + \sum_{i \in I(\bar{x})} \lambda_i \langle \nabla g_i(\bar{x}), y \rangle = 0.$$

Again by (12.9) it must be $\langle \nabla f(\bar{x}), y \rangle = 0$ and $\lambda_i \langle \nabla g_i(\bar{x}), y \rangle = 0$ for all $i \in I(\bar{x})$. Therefore if $i \in \tilde{I}(\bar{x})$, so $\lambda_i > 0$, it must be $\langle \nabla g_i(\bar{x}), y \rangle = 0$. Hence by definition we have $y \in \tilde{L}_C(\bar{x})$.

Since $f(x^k) \le f(\bar{x})$, $\lambda_i \ge 0$, $g_i(x^k) \le 0$, and $\lambda_i g_i(\bar{x}) = 0$, it follows that

$$L(x^k, \lambda, \mu) = f(x^k) + \sum_{i \in I(\bar{x})} \lambda_i g_i(x^k) \le f(\bar{x}) = L(\bar{x}, \lambda, \mu).$$

By Taylor's theorem

$$0 \ge L(x^k, \lambda, \mu) - L(\bar{x}, \lambda, \mu) \\ = \left\langle \nabla_x L(\bar{x}, \lambda, \mu), x^k - \bar{x} \right\rangle + \frac{1}{2} \left\langle x^k - \bar{x}, \nabla_x^2 L(\bar{x}, \lambda, \mu)(x^k - \bar{x}) \right\rangle + o(\left\|x^k - \bar{x}\right\|^2).$$

By the KKT conditions, the first term in the right-hand side is zero. Multiplying both sides by α_k^2 and letting $k \to \infty$, we get

$$0 \ge \frac{1}{2} \left\langle y, \nabla_x^2 L(\bar{x}, \lambda, \mu) y \right\rangle,$$

which is a contradiction. Therefore $\bar{x}$ is a strict local solution. □

Problems

12.1. Let $a_1, \ldots, a_K$ be vectors. This problem asks you to prove that the polyhedral cone $C = \operatorname{cone}[a_1, \ldots, a_K]$ is a closed convex cone.

1. Prove that C is a nonempty convex cone.
2. Prove that if $x \in C$, then x can be expressed as $x = \sum_{j=1}^{J} \alpha_j a_{k_j}$, where $\alpha_j \geq 0$ and $a_{k_1}, \ldots, a_{k_J}$ are linearly independent.
3. Prove that C is closed.

12.2. Let C be any set and suppose $\bar{x} \in \operatorname{int} C$ (interior point of C).

1. Compute the tangent cone $T_C(\bar{x})$ and the normal cone $N_C(\bar{x})$.
2. Interpret Theorem 12.5.

12.3. Let $x \in \mathbb{R}$ and consider the constraints (i) $x \leq 0$ and (ii) $x^3 \leq 0$.

1. Show that the constraints (i) and (ii) are equivalent, and compute the tangent cone at $x = 0$.
2. Compute the linearizing cones corresponding to constraints (i) and (ii) at $x = 0$, respectively. Are they the same?
3. Construct an example such that the Slater condition $g_i(x_0) < 0$ holds and g_i is quasi-convex (but not convex) but the KKT conditions do not hold.

12.4. Suppose that $g_i(x) = \langle a_i, x \rangle - c_i$ and $h_j(x) = \langle b_j, x \rangle - d_j$ in the minimization problem (12.4). Show that the Guignard constraint qualification is satisfied.

Chapter 13

Maximum and Envelope Theorems

13.1 A motivating example

Suppose you are managing a firm that produces a final product by using labor and raw materials. The production function is

$$y = Al^{\alpha}x^{1-\alpha},$$

where y is the quantity of output, $A > 0$ is a productivity parameter, $l > 0$ is labor input, $x > 0$ is the input of raw materials, and $0 < \alpha < 1$ is a parameter.

Assume that you cannot hire or fire workers in the short run and therefore you see labor input l as constant, but can choose the input of raw materials x freely. The wage rate is $w > 0$ and the unit price of the raw material is $p > 0$. The unit price of the final product is normalized to 1. You are interested in two questions:

1. What is the optimal amount of input of raw materials?
2. What would happen to the firm's profit if parameter values change?

We can answer the first question by solving the optimization problem. We can also answer the second question once we have solved the optimization problem, but the topic of this chapter is how to (partly) answer the second question without solving the optimization problem.

Mathematically, the problem is

$$\begin{array}{ll} \text{maximize} & Al^{\alpha}x^{1-\alpha} - wl - px \\ \text{subject to} & l \text{ fixed}, x \geq 0. \end{array}$$

It is not hard to see that the objective function is concave in x. Clearly the constraint is linear. Therefore the KKT conditions are necessary and sufficient for optimality. The Lagrangian is

$$L(x, \lambda) = Al^{\alpha}x^{1-\alpha} - wl - px + \lambda x.$$

The KKT conditions are

$$Al^{\alpha}(1-\alpha)x^{-\alpha} - p + \lambda = 0, \tag{13.1a}$$
$$\lambda \geq 0,\ x \geq 0,\ \lambda x = 0. \tag{13.1b}$$

(13.1a) is the first-order condition. (13.1b) is the complementary slackness condition.

If $x = 0$, then the first-order condition (13.1a) will be $\infty - p + \lambda = 0$, a contradiction. Therefore it must be $x > 0$. By the complementary slackness condition (13.1b), we get $\lambda = 0$. Substituting this into (13.1a) and solving for x, we get

$$Al^{\alpha}(1-\alpha)x^{-\alpha} - p = 0 \iff x = \left(\frac{A(1-\alpha)}{p}\right)^{\frac{1}{\alpha}} l. \tag{13.2}$$

Substituting this into the objective function, after some algebra the maximized profit is

$$\pi(\alpha, A, p, w, l) \coloneqq \alpha(1-\alpha)^{\frac{1}{\alpha}-1} A^{\frac{1}{\alpha}} p^{1-\frac{1}{\alpha}} l - wl.$$

Regarding the second question, suppose that we are interested in how the maximized profit π change when the price of the raw materials p changes. Then we compute

$$\frac{\partial \pi}{\partial p} = \alpha\left(1 - \frac{1}{\alpha}\right)(1-\alpha)^{\frac{1}{\alpha}-1} A^{\frac{1}{\alpha}} p^{-\frac{1}{\alpha}} l = -\left(\frac{A(1-\alpha)}{p}\right)^{\frac{1}{\alpha}} l.$$

This is simply the negative of the optimal input of raw materials computed in (13.2). If we had partially differentiated the profit function

$$Al^{\alpha}x^{1-\alpha} - wl - px$$

with respect to p, we get the same answer $-x$ (evaluated at the optimal solution, though)! Is this a coincidence? The answer is no. The Maximum Theorem tells that the optimal value and the solution are continuous in the parameter. The Envelope Theorem tells that the optimal value is differentiable in parameters and the derivatives are related to the Lagrange multipliers.

13.2 Maximum Theorem

Let $X \subset \mathbb{R}^N$ and $Y \subset \mathbb{R}^M$ be sets. $\Gamma : X \twoheadrightarrow Y$ is a *correspondence* (or *multi-valued function*) if for each $x \in X$ we have $\Gamma(x) \subset Y$, a subset of Y. Note that we use an arrow with two heads "$\twoheadrightarrow$" for a correspondence, while we use the usual arrow "$\to$" for a function. Another common notation is $\Gamma : X \rightrightarrows Y$. Γ is said to be *compact (convex) valued* if for each $x \in X$, the set $\Gamma(x)$ is compact (convex). Γ is said to be *uniformly bounded* if for each $\bar{x} \in X$, there exists a neighborhood U of $\bar{x}$ such that $\bigcup_{x \in U} \Gamma(x)$ is bounded. Of course, $\Gamma(x)$ need not be uniformly bounded just because $\Gamma(x)$ is bounded. For instance, let $X = \mathbb{R}$ and

$$\Gamma(x) = \begin{cases} [0,1], & (x \leq 0) \\ [0, 1/x]. & (x > 0) \end{cases}$$

Then $\Gamma(x)$ is bounded but not uniformly bounded at $\bar{x} = 0$ (draw a picture).

Remember that a function $f : X \to Y$ is continuous if $x_n \to x$ implies $f(x_n) \to f(x)$. "$x_n \to x$" is a shorthand notation for "$\lim_{n\to\infty} x_n = x$". We can define continuity for correspondences.

Definition 13.1 (Upper hemicontinuity)**.** $\Gamma : X \twoheadrightarrow Y$ is *upper hemicontinuous* if it is uniformly bounded and $x_n \to x$, $y_n \in \Gamma(x_n)$, and $y_n \to y$ implies $y \in \Gamma(x)$.

Upper hemicontinuity is also called upper semi-continuity (I often use semi-continuity). Perhaps hemicontinuity is less confusing since there is a separate semi-continuity concept for functions, introduced below. When the requirement that Γ is uniformly bounded is dropped, then Γ is called *closed*. When Y is itself bounded, upper hemicontinuity is the same as closedness. Upper hemicontinuity says that if a sequence in the image of a convergent sequence is convergent, then the limit belongs to the image of the limit. There is also a concept called lower hemicontinuity, which is roughly the converse. If you take a point in the image of the limit, then you can take a sequence in the image of the sequence that converges to that point.

Definition 13.2 (Lower hemicontinuity)**.** $\Gamma : X \twoheadrightarrow Y$ is *lower hemicontinuous* if for any $x_n \to x$ and $y \in \Gamma(x)$, there exists a number N and a sequence $y_n \to y$ such that $y_n \in \Gamma(x_n)$ for $n > N$.

A correspondence that is both upper and lower hemicontinuous is called *continuous*.

The next Maximum Theorem guarantees that the maximum value of a parametric maximization problem is continuous and the solution set is upper hemicontinuous.

Theorem 13.3 (Maximum Theorem)**.** *Let $f : X \times Y \to \mathbb{R}$ and $\Gamma : X \twoheadrightarrow Y$ be continuous. Assume*

$$\Gamma^*(x) = \arg\max_{y\in\Gamma(x)} f(x, y) \neq \emptyset$$

and let $f^*(x) = \max_{y\in\Gamma(x)} f(x, y)$. Then f^* is continuous and $\Gamma^* : X \twoheadrightarrow Y$ is upper hemicontinuous.

The proof of the maximum theorem is not so difficult, but it is clearer to weaken the assumptions and prove several weaker statements. To do so I define semi-continuity for functions.

Definition 13.4 (Semi-continuity of functions)**.** $f : X \to [-\infty, \infty]$ is *upper semi-continuous* if $x_n \to x$ implies $\limsup_{n\to\infty} f(x_n) \leq f(x)$. f is *lower semi-continuous* if $x_n \to x$ implies $\liminf_{n\to\infty} f(x_n) \geq f(x)$.

Clearly, f is upper semi-continuous if $-f$ is lower semi-continuous, and f is continuous if it is both upper and lower semi-continuous. The extreme value theorem (Theorem 2.5) guarantees that a continuous function attains the maximum on a compact set. Indeed all we need is upper semi-continuity, as the following theorem shows.

Theorem 13.5. *Let X be nonempty and compact and $f : X \to [-\infty, \infty)$ upper semi-continuous. Then f attains the maximum on X.*

Proof. Let $M = \sup_{x \in X} f(x)$. Take $\{x_n\}$ such that $f(x_n) \to M$. Since X is compact, $\{x_n\}$ has a convergent subsequence. Assume $x_{n_k} \to x$. Since f is upper semi-continuous, we have

$$M \geq f(x) \geq \limsup_{k \to \infty} f(x_{n_k}) = M,$$

so $f(x) = M = \max_{x \in X} f(x)$. □

By the same argument, a lower semi-continuous function attains the minimum on a compact set. Theorem 13.5 is useful. For example, the Cobb-Douglas utility function

$$u(x_1, x_2) = \alpha_1 \log x_1 + \alpha_2 \log x_2$$

is not continuous at $x_1 = 0$ or $x_2 = 0$ in the usual sense. But if we define $\log 0 = -\infty$, u becomes upper semi-continuous. Therefore if the budget set is compact, we know a priori that a solution to the utility maximization problem exists.

We prove two lemmas to prove the maximum theorem.

Lemma 13.6. *Let $f : X \times Y \to \mathbb{R}$ be upper semi-continuous and $\Gamma : X \twoheadrightarrow Y$ upper hemicontinuous. Then $f^*(x) = \sup_{y \in \Gamma(x)} f(x, y)$ is upper semi-continuous.*

Proof. Take any $x_n \to x$ and $\epsilon > 0$. Take a subsequence $\{x_{n_k}\}$ such that $f^*(x_{n_k}) \to \limsup_{n \to \infty} f^*(x_n)$. For each k, take $y_{n_k} \in \Gamma(x_{n_k})$ such that $f(x_{n_k}, y_{n_k}) > f^*(x_{n_k}) - \epsilon$. Since Γ is upper hemicontinuous, it is uniformly bounded. Therefore there exists a neighborhood U of x such that $\bigcup_{x' \in U} \Gamma(x')$ is bounded. Since $x_{n_k} \to x$, there exists K such that $\bigcup_{k > K} \Gamma(x_{n_k})$ is bounded. Hence $\{y_{n_k}\}$ is bounded. By taking a subsequence if necessary, we may assume $y_{n_k} \to y$. Since Γ is upper hemicontinuous, we have $y \in \Gamma(x)$. Since f is upper semi-continuous,

$$f^*(x) \geq f(x, y) \geq \limsup_{k \to \infty} f(x_{n_k}, y_{n_k}) \geq \lim_{k \to \infty} f^*(x_{n_k}) - \epsilon = \limsup_{n \to \infty} f^*(x_n) - \epsilon.$$

Letting $\epsilon \to 0$, it follows that f^* is upper semi-continuous. □

Lemma 13.7. *Let $f : X \times Y \to \mathbb{R}$ be lower semi-continuous and $\Gamma : X \twoheadrightarrow Y$ lower hemicontinuous. Then $f^*(x) = \sup_{y \in \Gamma(x)} f(x, y)$ is lower semi-continuous.*

Proof. Take any $x_n \to x$ and $\epsilon > 0$. Take $y \in \Gamma(x)$ such that $f(x, y) > f^*(x) - \epsilon$. Since Γ is lower hemicontinuous, there exist N and $y_n \to y$ such that $y_n \in \Gamma(x_n)$ for all $n > N$. Then $f^*(x_n) \geq f(x_n, y_n)$. Since f is lower semi-continuous,

$$\liminf_{n \to \infty} f^*(x_n) \geq \liminf_{n \to \infty} f(x_n, y_n) \geq f(x, y) > f^*(x) - \epsilon.$$

Letting $\epsilon \to 0$, it follows that f^* is lower semi-continuous. □

Proof of the maximum theorem. By Lemmas (13.6) and (13.7), f^* is continuous. Since $\Gamma^*(x) \subset \Gamma(x)$ and Γ is uniformly bounded, so is Γ^*. Take any $x_n \to x$, $y_n \in \Gamma^*(x_n)$, and assume $y_n \to y$. Since f and f^* are continuous, we have

$$f(x, y) = \lim_{n \to \infty} f(x_n, y_n) = \lim_{n \to \infty} f^*(x_n) = f^*(x),$$

so $y \in \Gamma^*(x)$. Hence Γ^* is upper hemicontinuous. □

13.3 Envelope Theorem

Consider the parametric minimization problem

$$\begin{array}{lll} \text{minimize} & f(x,u) & \\ \text{subject to} & g_i(x,u) \leq 0 & (i=1,\ldots,I). \end{array} \tag{13.3}$$

Here $x \in \mathbb{R}^N$ is the control variable and $u \in \mathbb{R}^p$ is a parameter. Let

$$\phi(u) = \inf_x \{f(x,u) \,|\, (\forall i) g_i(x,u) \leq 0\}$$

be the minimum value function. We are interested in how $\phi(u)$ changes when u does.

Recall the second-order sufficient condition for optimality (Theorem 12.11). The *active set* of the inequality constraints is

$$I(\bar{x}) = \{i \,|\, g_i(\bar{x}) = 0\}.$$

Let

$$\tilde{I}(\bar{x}) = \{i \,|\, \lambda_i > 0\}$$

be the set of constraints such that the Lagrange multiplier is positive. Since $\lambda_i g_i(\bar{x}) = 0$ by complementary slackness, $\lambda_i > 0$ implies $g_i(\bar{x}) = 0$, so necessarily $\tilde{I}(\bar{x}) \subset I(\bar{x})$. Define the cone

$$\begin{aligned} &\tilde{L}_C(\bar{x}) \\ &= \left\{ y \in \mathbb{R}^N \,\middle|\, (\forall i \in I(\bar{x}) \backslash \tilde{I}(\bar{x}))\, \langle \nabla g_i(\bar{x}), y \rangle \leq 0, (\forall i \in \tilde{I}(\bar{x}))\, \langle \nabla g_i(\bar{x}), y \rangle = 0 \right\}. \end{aligned}$$

Then the second-order sufficient condition for local optimality is

$$\left\langle y, \nabla_x^2 L(\bar{x}, \lambda, \mu) y \right\rangle > 0 \quad (\forall 0 \neq y \in \tilde{L}_C(\bar{x})). \tag{13.4}$$

Theorem 13.8 (Sensitivity Analysis)**.** *Suppose that $\bar{x} \in \mathbb{R}^N$ is a local solution to the parametric optimization problem (13.3) corresponding to $\bar{u} \in \mathbb{R}^p$. Assume that*

1. *The vectors $\{\nabla_x g_i(\bar{x}, \bar{u})\}_{i \in I(\bar{x})}$ are linearly independent, so the Karush-Kuhn-Tucker theorem holds with Lagrange multiplier $\bar{\lambda} \in \mathbb{R}^I_+$,*
2. *Strict complementary slackness condition holds, so $g_i(\bar{x}, \bar{u}) = 0$ implies $\bar{\lambda}_i > 0$ and therefore $I(\bar{x}) = \tilde{I}(\bar{x})$,*
3. *The second order condition (13.4) holds.*

Then there exists a neighborhood U of $\bar{u}$ and C^1 functions $x(u), \lambda(u)$ such that for any $u \in U$, $x(u)$ is the local solution to the parametric optimization problem (13.3) and $\lambda(u)$ is the corresponding Lagrange multiplier.

In our case, since strict complementary slackness holds and there are no equality constraints (h_j's), condition (13.4) reduces to

$$y \neq 0,\ (\forall i \in I(\bar{x}))\, \langle \nabla_x g_i(\bar{x}, \bar{u}), y \rangle = 0 \Longrightarrow \left\langle y, \nabla_x^2 L(\bar{x}, \bar{\lambda}, \bar{u}) y \right\rangle > 0. \tag{13.5}$$

We need the following lemma in order to prove the theorem.

Lemma 13.9. *Let everything be as in Theorem 13.8. Define the* $(N+I)\times(N+I)$ *matrix* A *by*

$$A = \begin{bmatrix} \nabla_x^2 L & \nabla_x g_1 & \cdots & \nabla_x g_I \\ \bar{\lambda}_1 \nabla_x g_1' & g_1 & & 0 \\ \vdots & & \ddots & \\ \bar{\lambda}_I \nabla_x g_I' & 0 & & g_I \end{bmatrix},$$

where all functions are evaluated at $(\bar{x}, \bar{\lambda}, \bar{u})$*. Then* A *is regular.*

Proof. Suppose that $A\begin{bmatrix} v \\ w \end{bmatrix} = 0$, where $v \in \mathbb{R}^N$ and $w \in \mathbb{R}^I$. It suffices to show that $v = 0$ and $w = 0$. By the definition of A, we get

$$\nabla_x^2 L v + \sum_{i=1}^{I} w_i \nabla_x g_i = 0, \tag{13.6a}$$

$$(\forall i)\ \bar{\lambda}_i \langle \nabla_x g_i, v \rangle + w_i g_i = 0. \tag{13.6b}$$

For $i \in I(\bar{x})$ (hence $g_i(\bar{x}, \bar{u}) = 0$), by (13.6b) and strict complementary slackness we have $\bar{\lambda}_i > 0$ and therefore $\langle \nabla_x g_i, v \rangle = 0$. For $i \notin I(\bar{x})$ (hence $g_i(\bar{x}, \bar{u}) < 0$), again by (13.6b) and strict complementary slackness we have $\bar{\lambda}_i = 0$ and therefore $w_i = 0$. Therefore (13.6a) becomes

$$\nabla_x^2 L v + \sum_{i \in I(\bar{x})} w_i \nabla_x g_i = 0. \tag{13.7}$$

Multiplying (13.7) by v as an inner product and using $\langle \nabla_x g_i, v \rangle = 0$ for $i \in I(\bar{x})$, we obtain

$$\left\langle v, \nabla_x^2 L v \right\rangle = 0.$$

By condition (13.5), it must be $v = 0$. Then by (13.7) we obtain

$$\sum_{i \in I(\bar{x})} w_i \nabla_x g_i = 0,$$

but since $\{\nabla_x g_i\}_{i \in I(\bar{x})}$ are linearly independent, it must be $w_i = 0$ for all i. Therefore $v = 0$ and $w = 0$. □

Proof of Theorem 13.8. Define $f : \mathbb{R}^N \times \mathbb{R}^I \times \mathbb{R}^p \to \mathbb{R}^N \times \mathbb{R}^I$ by

$$f(x, \lambda, u) = \begin{bmatrix} \nabla_x L(x, \lambda, u) \\ \lambda_1 g_i(x, u) \\ \vdots \\ \lambda_I g_I(x, u) \end{bmatrix}.$$

Then the Jacobian of f with respect to (x, λ) evaluated at $(\bar{x}, \bar{\lambda}, \bar{u})$ is A, which is regular. Furthermore, since $\bar{x}$ is a local solution corresponding to $u = \bar{u}$, by the KKT Theorem we have $f(\bar{x}, \bar{\lambda}, \bar{u}) = 0$. Therefore by the Implicit Function Theorem, there exists a neighborhood U of $\bar{u}$ and C^1 functions $x(u), \lambda(u)$ such that

$$f(x(u), \lambda(u), u) = 0$$

for $u \in U$. By the second-order sufficient condition (Theorem 12.11), $x(u)$ is the local solution to the parametric optimization problem (13.3) and $\lambda(u)$ is the corresponding Lagrange multiplier. □

The following theorem is extremely important.

Theorem 13.10 (Envelope Theorem)**.** *Let everything be as in Theorem 13.8 and*

$$L(x, \lambda, u) = f(x, u) + \sum_{i=1}^{I} \lambda_i g_i(x, u)$$

be the Lagrangian. Assume that the parametric optimization problem (13.3) has a solution $x(u)$ with Lagrange multiplier $\lambda(u)$. Let

$$\phi(u) = \min_x \{f(x, u) \,|\, (\forall i) g_i(x, u) \leq 0\}$$

be the minimum value function. Then ϕ is differentiable and

$$\nabla\phi(u) = \nabla_u L(x(u), \lambda(u), u),$$

i.e., the derivative of ϕ can be computed by differentiating the Lagrangian with respect to the parameter u alone, treating x and λ as constants.

Proof. By the definition of ϕ and complementary slackness, we obtain

$$\phi(u) = f(x(u), u) = L(x(u), \lambda(u), u).$$

Differentiating both sides with respect to u, we get

$$\underbrace{D_u\phi(u)}_{1\times p} = \underbrace{D_x L}_{1\times N}\underbrace{D_u x(u)}_{N\times p} + \underbrace{D_\lambda L}_{1\times I}\underbrace{D_u\lambda(u)}_{I\times p} + \underbrace{D_u L}_{1\times p}.$$

By the KKT theorem, we have $D_x L = 0$. By the strict complementary slackness, we have $\lambda_i(u) = 0$ for $i \notin I(\bar{x})$ and

$$D_{\lambda_i} L(x(u), \lambda(u), u) = g_i(x(u), u) = 0$$

for $i \in I(\bar{x})$, so $D_\lambda L D_u \lambda(u) = 0$. Therefore $\nabla\phi(u) = \nabla_u L(x(u), \lambda(u), u)$. □

Corollary 13.11. *Consider the special case*

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & g_i(x) \leq u_i & (i = 1, \ldots, I). \end{array}$$

Then $\nabla\phi(u) = -\lambda(u)$.

Proof. The Lagrangian is

$$L(x, \lambda, u) = f(x) + \sum_{i=1}^{I} \lambda_i [g_i(x) - u_i].$$

By the envelope theorem, $\nabla\phi(u) = \nabla_u L(x(u), \lambda(u), u) = -\lambda(u)$. □

Problems

13.1. Consider the *utility maximization problem* (UMP)

$$\begin{array}{ll} \text{maximize} & \alpha \log x_1 + (1-\alpha)\log x_2 \\ \text{subject to} & p_1 x_1 + p_2 x_2 \leq w, \end{array}$$

where $0 < \alpha < 1$ is a parameter, $p_1, p_2 > 0$ are prices, and $w > 0$ is wealth.

1. Solve UMP.
2. Let $v(p_1, p_2, w)$ be the value function. Compute the partial derivatives of v with respect to each of p_1, p_2, and w.
3. Verify *Roy's identity* $x_n = -\frac{\partial v}{\partial p_n} / \frac{\partial v}{\partial w}$ for $n = 1, 2$, where x_n is the optimal demand of good n.

13.2. Consider the UMP

$$\begin{array}{ll} \text{maximize} & u(x) \\ \text{subject to} & x \in \mathbb{R}^L_+, \ \ \langle p, x \rangle \leq w, \end{array}$$

where $w > 0$ is the wealth of the consumer, $p \in \mathbb{R}^L_{++}$ is the price vector, $x \in \mathbb{R}^L_+$ is the demand, and $u : \mathbb{R}^L_+ \to \mathbb{R}$ is differentiable and strictly quasi-concave. Let $x(p, w)$ be the solution to UMP (called *Marshallian demand*) and $v(p, w)$ the value function. If $x(p, w) \gg 0$, prove *Roy's identity*

$$x(p, w) = -\frac{\nabla_p v(p, w)}{\nabla_w v(p, w)}.$$

(Hint: envelope theorem.)

13.3. Consider the *expenditure minimization problem* (EMP)

$$\begin{array}{ll} \text{minimize} & p_1 x_1 + p_2 x_2 \\ \text{subject to} & \alpha \log x_1 + (1-\alpha)\log x_2 \geq u \end{array}$$

where $0 < \alpha < 1$ is a parameter, $p_1, p_2 > 0$ are prices, and $u \in \mathbb{R}$ is the desired utility level.

1. Solve EMP.
2. Let $e(p_1, p_2, u)$ be the value function. Compute the partial derivatives of e with respect to p_1 and p_2.
3. Verify *Shephard's lemma* $x_n = \frac{\partial e}{\partial p_n}$ for $n = 1, 2$, where x_n is the optimal demand of good n.

13.4. Consider the EMP

$$\begin{array}{ll} \text{minimize} & \langle p, x \rangle \\ \text{subject to} & u(x) \geq u, \end{array}$$

where $p \in \mathbb{R}^L_{++}$ is the price vector, $x \in \mathbb{R}^L$ is the demand, $u(x)$ is a strictly quasi-concave differentiable utility function, and $u \in \mathbb{R}$ is the desired utility

level. Let $h(p,u)$ be the solution to EMP (called *Hicksian demand*) and $e(p,u)$ be the minimum expenditure (called *expenditure function*). Prove *Shephard's lemma*

$$h(p,u) = \nabla_p e(p,u).$$

13.5. Prove the following *Slutsky equation*:

$$\underbrace{D_p x(p,w)}_{L\times L} = \underbrace{D_p^2 e(p,u)}_{L\times L} - \underbrace{[D_w x(p,w)]}_{L\times 1}\underbrace{[x(p,w)]'}_{1\times L},$$

where $x(p,w)$ is the Marshallian demand, $u = u(x(p,w))$ is the utility level evaluated at the demand, and $e(p,u)$ is the expenditure function.

Chapter 14

Duality Theory

14.1 Motivation

Consider the following constrained minimization problem

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & g_i(x) \leq 0 & (i = 1, \ldots, I) \end{array} \tag{14.1}$$

with Lagrangian

$$L(x, \lambda) = \begin{cases} f(x) + \sum_{i=1}^{I} \lambda_i g_i(x), & (\lambda \in \mathbb{R}^I_+) \\ -\infty. & (\lambda \notin \mathbb{R}^I_+) \end{cases}$$

(The following discussion can easily accommodate equality constraints as well.) Recall the saddle point theorem (Theorem 11.2): if f, g_i's are all convex and there is a point $x_0 \in \mathbb{R}^N$ such that $g_i(x_0) < 0$ for all i (Slater constraint qualification), then $\bar{x}$ is a solution to (14.1) if and only if there is $\bar{\lambda} \in \mathbb{R}^I_+$ such that $(\bar{x}, \bar{\lambda})$ is a saddle point of L, that is,

$$L(\bar{x}, \lambda) \leq L(\bar{x}, \bar{\lambda}) \leq L(x, \bar{\lambda}) \tag{14.2}$$

for all $x \in \mathbb{R}^N$ and $\lambda \in \mathbb{R}^I$. By (14.2), we get

$$\sup_{\lambda} L(\bar{x}, \lambda) \leq L(\bar{x}, \bar{\lambda}) \leq \inf_{x} L(x, \bar{\lambda}).$$

Therefore

$$\inf_{x} \sup_{\lambda} L(x, \lambda) \leq L(\bar{x}, \bar{\lambda}) \leq \sup_{\lambda} \inf_{x} L(x, \lambda). \tag{14.3}$$

On the other hand, $L(x, \lambda) \leq \sup_{\lambda} L(x, \lambda)$ always, so taking the infimum with respect to x, we get

$$\inf_{x} L(x, \lambda) \leq \inf_{x} \sup_{\lambda} L(x, \lambda).$$

Noting that the right-hand side is just a constant, taking the supremum of the left-hand side with respect to λ, we get

$$\sup_{\lambda} \inf_{x} L(x, \lambda) \leq \inf_{x} \sup_{\lambda} L(x, \lambda). \tag{14.4}$$

Combining (14.3) and (14.4), it follows that

$$L(\bar{x}, \bar{\lambda}) = \inf_x \sup_\lambda L(x, \lambda) = \sup_\lambda \inf_x L(x, \lambda). \tag{14.5}$$

Define

$$\theta(x) = \sup_\lambda L(x, \lambda),$$
$$\omega(\lambda) = \inf_x L(x, \lambda).$$

Then (14.5) is equivalent to

$$L(\bar{x}, \bar{\lambda}) = \inf_x \theta(x) = \sup_\lambda \omega(\lambda). \tag{14.6}$$

Note that by the definition of the Lagrangian,

$$\theta(x) = \sup_\lambda L(x, \lambda) = \begin{cases} f(x), & (\forall i, g_i(x) \le 0) \\ \infty. & (\exists i, g_i(x) > 0) \end{cases}$$

Therefore the constrained minimization problem (14.1) is equivalent to the unconstrained minimization problem

$$\text{minimize } \theta(x). \tag{P}$$

For this reason the problem (P) is called the *primal problem*. In view of (14.6), define the *dual problem* by

$$\text{maximize } \omega(\lambda). \tag{D}$$

Then (14.6) implies that the primal and dual values coincide.

The above discussion suggests that in order to solve the constrained minimization problem (14.1) and hence the primal problem (P), it might be sufficient to solve the dual problem (D). Since $L(x, \lambda)$ is linear in λ, $\omega(\lambda) = \inf_x L(x, \lambda)$ is always a concave function of λ no matter what f or g_i's are. Therefore we can expect that solving the dual problem is much easier than solving the primal problem.

14.2 Example

14.2.1 Linear programming

A typical linear programming problem is

$$\begin{array}{ll} \text{minimize} & \langle c, x \rangle \\ \text{subject to} & Ax \ge b, \end{array}$$

where $x \in \mathbb{R}^N$ is the vector of decision variables, $c \in \mathbb{R}^N$ is the vector of the coefficients, A is an $M \times N$ matrix, and $b \in \mathbb{R}^M$ is a vector. The Lagrangian is

$$L(x, \lambda) = \langle c, x \rangle + \langle \lambda, b - Ax \rangle ,$$

where $\lambda \in \mathbb{R}_+^M$ is the vector of Lagrange multipliers. Since

$$\begin{aligned}\omega(\lambda) = \inf_x L(x,\lambda) &= \inf_x [\langle c - A'\lambda, x\rangle + \langle b, \lambda\rangle] \\ &= \begin{cases} \langle b, \lambda\rangle, & (A'\lambda = c) \\ -\infty. & (A'\lambda \neq c) \end{cases}\end{aligned}$$

The dual problem is

$$\begin{array}{ll} \text{maximize} & \langle b, \lambda\rangle \\ \text{subject to} & A'\lambda = c,\ \lambda \geq 0. \end{array}$$

14.2.2 Entropy maximization

Let $\mathbf{p} = (p_1, \ldots, p_N)$ be a multinomial distribution, so $p_n \geq 0$ and $\sum_{n=1}^N p_n = 1$. The quantity

$$H(\mathbf{p}) = -\sum_{n=1}^N p_n \log p_n$$

is called the *entropy* of $\mathbf{p}$.

In practice we often want to find the distribution $\mathbf{p}$ that has the maximum entropy satisfying some moment constraints. Suppose the constraints are given by

$$\sum_{n=1}^N a_{in} p_n = b_i. \ (i = 1, \ldots, I)$$

Since maximizing H is equivalent to minimizing $-H$, the problem is

$$\begin{array}{lll} \text{minimize} & \displaystyle\sum_{n=1}^N p_n \log p_n & \\ \text{subject to} & \displaystyle\sum_{n=1}^N a_{in} p_n = b_i, & (i = 0, \ldots, I) \end{array} \tag{14.7}$$

where a_{in}'s and b_i's are given and I define $a_{0n} = 1$ and $b_0 = 1$ to accommodate the constraint $\sum_n p_n = 1$ (accounting of probability).

If the number of unknown variables N is large (say $N \sim 10000$), then it would be very hard to solve the problem even using a computer since the objective function $p_n \log p_n$ is nonlinear. However, it turns out that the dual problem is very simple.

Although $p \log p$ is not well-defined when $p \leq 0$, define

$$p \log p = \begin{cases} 0, & (p = 0) \\ \infty. & (p < 0) \end{cases}$$

Then the constraint $p_n \geq 0$ is built in the problem. The Lagrangian is

$$\begin{aligned} L(\mathbf{p}, \lambda) &= \sum_{n=1}^N p_n \log p_n + \sum_{i=0}^I \lambda_i \left(b_i - \sum_{n=1}^N a_{in} p_n \right) \\ &= \langle b, \lambda\rangle + \sum_{n=1}^N \left(p_n \log p_n - \langle a_n, \lambda\rangle\, p_n \right), \end{aligned}$$

where $b = (b_0, \ldots, b_I)$ and $a_n = (a_{0n}, \ldots, a_{In})$. To derive the dual problem, we need to compute $\inf_{\mathbf{p}} L(\mathbf{p}, \lambda)$, which reduces to computing

$$\inf_p [p \log p - cp]$$

for $c = \langle a_n, \lambda \rangle$ above. But this problem is easy! Differentiating with respect to p, the first-order condition is

$$\log p + 1 - c = 0 \iff p = \mathrm{e}^{c-1},$$

with the minimum value

$$p \log p - cp = \mathrm{e}^{c-1}(c-1) - c\mathrm{e}^{c-1} = -\mathrm{e}^{c-1}.$$

Substituting $p_n = \mathrm{e}^{\langle a_n, \lambda \rangle - 1}$ in the Lagrangian, after some algebra the objective function of the dual problem becomes

$$\omega(\lambda) = \inf_{\mathbf{p}} L(\mathbf{p}, \lambda) = \langle b, \lambda \rangle - \sum_{n=1}^{N} \mathrm{e}^{\langle a_n, \lambda \rangle - 1}.$$

Hence the dual problem of (14.7) is

$$\text{maximize} \quad \langle b, \lambda \rangle - \sum_{n=1}^{N} \mathrm{e}^{\langle a_n, \lambda \rangle - 1}. \tag{14.8}$$

Numerically solving (14.8) is much easier than (14.7) because the dual problem (14.8) is *unconstrained* and the number of unknown variables $1 + I$ is typically *much smaller* than N.

14.3 Convex conjugate function

In the above example we needed to compute

$$\inf_p [p \log p - cp] = -\sup_p [cp - p \log p].$$

In general, for any function $f : \mathbb{R}^N \to [-\infty, \infty]$ we define

$$f^*(\xi) = \sup_{x \in \mathbb{R}^N} [\langle \xi, x \rangle - f(x)],$$

which is called the *convex conjugate function* of f. Fixing x, since $\langle \xi, x \rangle - f(x)$ is an affine (hence closed and convex) function of ξ, f^* is a closed convex function. (A function is closed if its epigraph is closed. Closedness of a function is the same as lower semi-continuity.)

For any function f, the largest closed convex function that is less than or equal to f is called the *closed convex hull* of f, and is denoted by $\operatorname{cl}\operatorname{co} f$. Formally, we define

$$\operatorname{cl}\operatorname{co} f(x) = \sup \{ g(x) \,|\, g \text{ is closed convex and } f(x) \geq g(x) \}.$$

(At least one such g exists—$g(x) = -\infty$.) Clearly $\phi = \operatorname{cl}\operatorname{co} f$ satisfies $\operatorname{epi}\phi = \operatorname{cl}\operatorname{co}\operatorname{epi} f$, so

$$\operatorname{epi}\operatorname{cl}\operatorname{co} f = \operatorname{cl}\operatorname{co}\operatorname{epi} f$$

is also the definition of $\operatorname{cl}\operatorname{co} f$.

The convex conjugate function of the convex conjugate function of $f(x)$ is called the *biconjugate* function. Formally,

$$f^{**}(x) = \sup_{\xi\in\mathbb{R}^N} [\langle x, \xi\rangle - f^*(\xi)].$$

In the case of cones, we had $C^{**} = \operatorname{cl}\operatorname{co} C$ for any cone C. The same holds for functions, under some conditions.

Theorem 14.1. *Let f be a function. If $\operatorname{cl}\operatorname{co} f$ is proper (so $\operatorname{cl}\operatorname{co} f(x) > -\infty$ for all x), then*

$$f^{**}(x) = \operatorname{cl}\operatorname{co} f(x).$$

In particular, $f^{**}(x) = f(x)$ if f is a proper closed convex function.

We need a lemma in order to prove Theorem 14.1.

Lemma 14.2. *For any function f, let $\mathcal{L}(f)$ be the set of affine functions that are less than or equal to f, so*

$$\mathcal{L}(f) = \{h \mid h \text{ is affine and } f(x) \geq h(x) \text{ for all } x\}.$$

If f is a proper closed convex function, then $\mathcal{L}(f) \neq \emptyset$.

Proof. Let f be a proper closed convex function. The claim is obvious if $f = \infty$, so we may assume $\operatorname{dom} f \neq \emptyset$. Take any vector $\bar{x} \in \operatorname{dom} f$ and real number $\bar{y}$ such that $\bar{y} < f(\bar{x})$. Then $(\bar{x}, \bar{y}) \notin \operatorname{epi} f$. Since f is a closed convex function, $\operatorname{epi} f$ is a closed convex set. Therefore by the separating hyperplane theorem we can take a vector $(0,0) \neq (\eta, \beta) \in \mathbb{R}^N \times \mathbb{R}$ and a constant $\gamma \in \mathbb{R}$ such that

$$\langle \eta, \bar{x}\rangle + \beta\bar{y} < \gamma < \langle \eta, x\rangle + \beta y$$

for any $(x, y) \in \operatorname{epi} f$. Letting $x = \bar{x}$ and $y \to \infty$, it must be $\beta > 0$. Dividing both sides by $\beta > 0$ and letting $a = -\eta/\beta$ and $c = \gamma/\beta$, we get

$$-\langle a, \bar{x}\rangle + \bar{y} < c < -\langle a, x\rangle + y$$

for all $(x, y) \in \operatorname{epi} f$. Let $h(x) = \langle a, x\rangle + c$. If $x \notin \operatorname{dom} f$, then clearly $\infty = f(x) > h(x)$. If $x \in \operatorname{dom} f$, letting $y = f(x)$ in the right inequality we get $f(x) > \langle a, x\rangle + c = h(x)$. Therefore $f(x) \geq h(x)$ for all x, so $h \in \mathcal{L}(f) \neq \emptyset$. □

Lemma 14.3. *Let f be a function. If $\operatorname{cl}\operatorname{co} f$ is proper, then*

$$\operatorname{cl}\operatorname{co} f(x) = \sup\{h(x) \mid h \in \mathcal{L}(f)\}.$$

Proof. Since $\phi(x) = \operatorname{cl}\operatorname{co} f$ is the largest closed convex function that is less than equal to f, and any $h \in \mathcal{L}(f)$ is an affine (hence closed convex) function that is less than equal to f, clearly

$$\phi(x) \geq \sup\{h(x) \mid h \in \mathcal{L}(f)\}.$$

To prove that equality holds, suppose that

$$\phi(\bar{x}) > \sup\{h(\bar{x}) \,|\, h \in \mathcal{L}(f)\}$$

for some $\bar{x}$. Then we can take a real number $\bar{y}$ such that

$$\phi(\bar{x}) > \bar{y} > \sup\{h(\bar{x}) \,|\, h \in \mathcal{L}(f)\}. \tag{14.9}$$

By the left inequality, $(\bar{x}, \bar{y}) \notin \operatorname{epi}\phi$. Since $\operatorname{epi}\phi = \operatorname{epi}\operatorname{cl}\operatorname{co} f = \operatorname{cl}\operatorname{co}\operatorname{epi} f$ is a closed convex set, by the separating hyperplane theorem we can take a vector $(0,0) \neq (\eta, \beta) \in \mathbb{R}^N \times \mathbb{R}$ and a constant $\gamma \in \mathbb{R}$ such that

$$\langle \eta, \bar{x} \rangle + \beta\bar{y} < \gamma < \langle \eta, x \rangle + \beta y \tag{14.10}$$

for any $(x, y) \in \operatorname{epi}\phi$. Letting $y \to \infty$, we get $\beta \geq 0$. There are two cases to consider.

Case 1: $\boldsymbol{\beta > 0}$. If $\beta > 0$, as in the proof of Lemma 14.2 dividing both sides of (14.10) by $\beta > 0$ and letting $a = -\eta/\beta$ and $c = \gamma/\beta$, we get

$$-\langle a, \bar{x} \rangle + \bar{y} < c < -\langle a, x \rangle + y$$

for all $(x, y) \in \operatorname{epi}\phi$. Since $f(x) \geq \operatorname{cl}\operatorname{co} f(x) = \phi(x)$, letting $y = f(x)$ we get $f(x) > \langle a, x \rangle + c$ and $\bar{y} < \langle a, \bar{x} \rangle + c$. The first inequality implies that $h_1(x) := \langle a, x \rangle + c$ satisfies $h_1 \in \mathcal{L}(f)$. Hence by the second inequality we get

$$\bar{y} < h_1(\bar{x}) \leq \sup\{h(\bar{x}) \,|\, h \in \mathcal{L}(f)\},$$

which contradicts (14.9)

Case 2: $\boldsymbol{\beta = 0}$. If $\beta = 0$, let $h_1(x) = -\langle \eta, x \rangle + \gamma$. Then by (14.10) we get $h_1(x) < 0 < h_1(\bar{x})$ for any $x \in \operatorname{dom}\phi$. Since by assumption ϕ is a proper closed convex function, be Lemma 14.2 we can take an affine function $h_2(x)$ such that $\phi(x) \geq h_2(x)$. Hence for any $\lambda \geq 0$ we have

$$\phi(x) > \lambda h_1(x) + h_2(x)$$

for any x, so $h(x) = \lambda h_1(x) + h_2(x)$ satisfies $h \in \mathcal{L}(f)$. But since $h_1(\bar{x}) > 0$, for large enough λ we have $h(\bar{x}) = \lambda h_1(\bar{x}) + h_2(\bar{x}) > \bar{y}$, which contradicts (14.9). □

Proof of Theorem 14.1. For $h(x) = \langle \xi, x \rangle - \beta$, by the definition of $\mathcal{L}(f)$ and f^* we have

$$\begin{aligned}
h \in \mathcal{L}(f) &\iff (\forall x)\ f(x) \geq \langle \xi, x \rangle - \beta \\
&\iff (\forall x)\ \beta \geq \langle \xi, x \rangle - f(x) \\
&\iff \beta \geq \sup_x [\langle \xi, x \rangle - f(x)] = f^*(\xi) \\
&\iff (\xi, \beta) \in \operatorname{epi} f^*.
\end{aligned}$$

Hence by the definition of the biconjugate function and Lemma 14.3,

$$\begin{aligned}
f^{**}(x) &= \sup\left\{\langle x, \xi \rangle - f^*(\xi) \,\middle|\, \xi \in \mathbb{R}^N\right\} \\
&= \sup\{\langle x, \xi \rangle - \beta \,|\, (\xi, \beta) \in \operatorname{epi} f^*\} \\
&= \sup\{h(x) \,|\, h \in \mathcal{L}(f)\} = \operatorname{cl}\operatorname{co} f(x).
\end{aligned}$$

□

14.4 Duality theory

Looking at the entropy maximization example, the key to simplifying the dual problem is to reduce it to the calculation of the convex conjugate function. However, unless the functions g_i in the constraints (14.1) are all affine, the Lagrangian $L(x, \lambda)$ will not contain an affine function and therefore we cannot reduce it to a convex conjugate function.

To circumvent this issue, instead of the original constrained minimization problem (14.1) consider the perturbed parametric minimization problem

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & g_i(x) \le u_i, & (i = 1, \dots, I) \end{array} \tag{14.11}$$

where $u = (u_1, \dots, u_I)$ is a parameter. Using the Lagrangian of (14.1), the Lagrangian of (14.11) is

$$L(x, \lambda) - \langle \lambda, u \rangle .$$

Let

$$F(x, u) = \begin{cases} f(x), & (\forall i, g_i(x) \le u_i) \\ \infty, & (\exists i, g_i(x) > u_i) \end{cases}$$

be the value of f restricted to the feasible set and

$$\phi(u) = \inf_x F(x, u)$$

be the value function of the minimization problem 14.11. By the definition of $\theta(x)$, we obtain $\theta(x) = F(x, 0)$.

Lemma 14.4. *Let everything be as above. Then*

$$L(x, \lambda) = \inf_u [F(x, u) + \langle \lambda, u \rangle], \tag{14.12a}$$

$$F(x, u) = \sup_\lambda [L(x, \lambda) - \langle \lambda, u \rangle]. \tag{14.12b}$$

Proof. Since $F(x, u) = \infty$ if $g_i(x) > u_i$ for some i, in taking the infimum of (14.12a) we may assume $g_i(x) \le u_i$ for all i. If $\lambda_i < 0$ for some i, then letting $u_i \to \infty$ we get

$$\inf_u [F(x, u) + \langle \lambda, u \rangle] = -\infty = L(x, \lambda).$$

If $\lambda_i \ge 0$ for all i, since $F(x, u) = f(x)$ the infimum is attained when $u_i = g_i(x)$ for all i, so

$$\inf_u [F(x, u) + \langle \lambda, u \rangle] = f(x) + \sum_{i=1}^{I} \lambda_i g_i(x) = L(x, \lambda).$$

Since $L(x, \lambda) = -\infty$ if $\lambda_i < 0$ for some i, in taking the supremum of (14.12b) we may assume $\lambda_i \ge 0$ for all i. Then

$$\begin{aligned} \sup_\lambda [L(x, \lambda) - \langle \lambda, u \rangle] &= \sup_{\lambda \ge 0} \left[f(x) + \sum_{i=1}^{I} \lambda_i (g_i(x) - u_i) \right] \\ &= \begin{cases} f(x), & (\forall i, g_i(x) \le u_i) \\ \infty, & (\exists i, g_i(x) > u_i) \end{cases} \\ &= F(x, u). \end{aligned}$$

□

We need one more lemma.

Lemma 14.5. *Let everything be as above. Then*

$$\begin{aligned} \omega(\lambda) &= -\phi^*(-\lambda), \\ \sup_\lambda \omega(\lambda) &= \phi^{**}(0). \end{aligned}$$

Proof. By the definition of ω, ϕ, and the convex conjugate function, we get

$$\begin{aligned} \omega(\lambda) &= \inf_x L(x,\lambda) \\ &= \inf_{x,u}[F(x,u) + \langle \lambda, u\rangle] && (\because (14.12a)) \\ &= \inf_u [\phi(u) + \langle \lambda, u\rangle] \\ &= -\sup_u [\langle -\lambda, u\rangle - \phi(u)] = -\phi^*(-\lambda). \end{aligned}$$

Therefore

$$\sup_\lambda \omega(\lambda) = \sup_\lambda [-\phi^*(-\lambda)] = \sup_\lambda [\langle 0, -\lambda\rangle - \phi^*(-\lambda)] = \phi^{**}(0). \quad \square$$

We immediately obtain the main result.

Theorem 14.6 (Duality theorem)**.** *The primal value* $\inf_x \theta(x)$ *and the dual value* $\sup_\lambda \omega(\lambda)$ *coincide if and only if* $\phi(0) = \phi^{**}(0)$*. In particular, this is the case if* ϕ *is a proper closed convex function.*

Proof. Clearly $\inf_x \theta(x) = \inf_x F(x,0) = \phi(0)$. By the previous lemma we have $\sup_\lambda \omega(\lambda) = \phi^{**}(0)$. If ϕ is a proper closed convex function, then $\phi(u) = \phi^{**}(u)$ for all u, so the primal and dual values coincide. $\square$

Problems

14.1. Compute the convex conjugate functions of the following functions.

1. $f(x) = \frac{1}{p}|x|^p$, where $p > 1$. (Express the solution using $q > 1$ such that $1/p + 1/q = 1$.)
2. $f(x) = \begin{cases} \infty, & (x < 0) \\ 0, & (x = 0) \\ x \log \frac{x}{a}, & (x > 0) \end{cases}$ where $a > 0$.
3. $f(x) = \begin{cases} \infty, & (x \le 0) \\ -\log x. & (x > 0) \end{cases}$
4. $f(x) = \langle a, x\rangle$, where $a \in \mathbb{R}^N$.
5. $f(x) = \delta_a(x) := \begin{cases} 0, & (x = a) \\ \infty, & (x \ne a) \end{cases}$ where $a \in \mathbb{R}^N$.
6. $f(x) = \frac{1}{2}\langle x, Ax\rangle$, where A is an $N \times N$ symmetric positive definite matrix.

14.2. Let

$$f(x) = \begin{cases} \infty, & (x < 0) \\ -x^2. & (x \geq 0) \end{cases}$$

1. Compute $f^*(\xi)$, $f^{**}(x)$, $\operatorname{cl}\operatorname{co} f(x)$.
2. Does $f^{**}(x) = \operatorname{cl}\operatorname{co} f(x)$ hold? If not, is it a contradiction?

14.3. Derive the dual problem of

$$\begin{array}{ll} \text{minimize} & \langle c, x \rangle \\ \text{subject to} & Ax \geq b,\ x \geq 0. \end{array}$$

14.4. Derive the dual problem of

$$\begin{array}{ll} \text{minimize} & \langle c, x \rangle + \dfrac{1}{2} \langle x, Qx \rangle \\ \text{subject to} & Ax \geq b, \end{array}$$

where Q is a symmetric positive definite matrix.

14.5. Consider the entropy maximization problem (14.7). Noting that $a_{0n} = 1$ and $b_0 = 1$, let $a_n = (1, T_n) \in \mathbb{R}^{1+I}$, $b = (1, \bar{T}) \in \mathbb{R}^{1+I}$, and $\lambda = (\lambda_0, \lambda_1) \in \mathbb{R} \times \mathbb{R}^I$. Carry out the maximization in (14.8) with respect to λ_0 alone and show that (14.8) is equivalent to

$$\text{maximize} \quad \langle \bar{T}, \lambda_1 \rangle - \log \left(\sum_{n=1}^{N} \mathrm{e}^{\langle T_n, \lambda_1 \rangle} \right),$$

and also to

$$\text{minimize} \quad \log \left(\sum_{n=1}^{N} \mathrm{e}^{\langle T_n - \bar{T}, \lambda_1 \rangle} \right) \iff \text{minimize} \quad \sum_{n=1}^{N} \mathrm{e}^{\langle T_n - \bar{T}, \lambda_1 \rangle}.$$

14.6. Let $\mathbf{p} = (p_1, \ldots, p_N)$ and $\mathbf{q} = (q_1, \ldots, q_N)$ be multinomial distributions. A concept closely related to entropy is the *Kullback-Leibler information* of $\mathbf{p}$ with respect to $\mathbf{q}$, defined by

$$H(\mathbf{p}; \mathbf{q}) = \sum_{n=1}^{N} p_n \log \frac{p_n}{q_n}.$$

Derive the dual problem of the minimum information problem

$$\begin{array}{lll} \text{minimize} & \displaystyle\sum_{n=1}^{N} p_n \log \frac{p_n}{q_n} & \\ \text{subject to} & \displaystyle\sum_{n=1}^{N} a_{in} p_n = b_i, & (i = 0, \ldots, I) \end{array}$$

where the minimization is over $\mathbf{p}$ given $\mathbf{q}$.

14.7. Derive the dual problem of

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b, \end{array}$$

where $f : \mathbb{R}^N \to (-\infty, \infty]$.

(Hint: define the Lagrangian by $L(x, \lambda) = f(x) + \langle \lambda, b - Ax \rangle$.)

Chapter 15

Dynamic Programming in Infinite Horizon

15.1 A motivating example

Suppose you live forever and need to finance your consumption from savings. Let $x_0 > 0$ be your initial wealth, $R > 0$ be the gross risk-free rate, and the flow utility from consumption y is $u(y) = \frac{y^{1-\gamma}}{1-\gamma}$, where $0 \leq \gamma \neq 1$ is the relative risk aversion coefficient, and you discount future utility with discount factor $\beta > 0$. Then what is the optimal way to consume and save?

Mathematically, the problem is

$$\text{maximize} \qquad \sum_{t=0}^{\infty} \beta^t \frac{y_t^{1-\gamma}}{1-\gamma} \tag{15.1a}$$

$$\text{subject to} \qquad 0 \leq y_t \leq x_t,\ x_{t+1} = R(x_t - y_t), \tag{15.1b}$$

where $x_0 > 0$ is given, $y_t > 0$ is consumption at time t, and x_t is the financial wealth at the beginning of time t.

15.2 General formulation

The previous example can be generalized as follows. Let X, Y be nonempty sets. At each stage $t = 0, 1, \ldots$ (which we call "time" for concreteness), given the state variable $x_t \in X$ at time t, the decision maker can choose the control variable $y_t \in \Gamma(x_t)$, where $\Gamma : X \twoheadrightarrow Y$ is a correspondence with $\Gamma(x) \neq \emptyset$. Given the state x_t and control y_t, the decision maker receives a flow utility $u(x_t, y_t)$ and the next period's state is determined by $x_{t+1} = g(x_t, y_t)$, where $u : X \times Y \to [-\infty, \infty)$ and $g : X \times Y \to X$. The decision maker discounts future utility using discount factor $\beta \in [0, 1)$.

Mathematically, the problem is

$$\begin{aligned} &\text{maximize} && \sum_{t=0}^{\infty} \beta^t u(x_t, y_t) && \text{(15.2a)} \\ &\text{subject to} && (\forall t) y_t \in \Gamma(x_t),\ x_{t+1} = g(x_t, y_t), && \text{(15.2b)} \\ & && x_0 \in X \text{ given.} && \text{(15.2c)} \end{aligned}$$

The motivating example (15.1) is a special case by setting $X = \mathbb{R}_+$, $Y = \mathbb{R}_+$, $u(x, y) = \frac{y^{1-\gamma}}{1-\gamma}$, $\Gamma(x) = [0, x]$, and $g(x, y) = R(x - y)$.

Given $x_0 \in X$, we say that the sequence of state and control variables $\{(x_t, y_t)\}_{t=0}^{\infty}$ is *feasible* if $y_t \in \Gamma(x_t)$ and $x_{t+1} = g(x_t, y_t)$ for all t. Thus the goal is to maximize the discounted sum of utility $\sum_{t=0}^{\infty} \beta^t u(x_t, y_t)$ among all feasible sequences.[1]

To solve the problem (15.2), define the (supremum) *value function* $V^* : X \to [-\infty, \infty]$ by

$$V^*(x) = \sup \left\{ \sum_{t=0}^{\infty} \beta^t u(x_t, y_t) \,\middle|\, x_0 = x,\ \{(x_t, y_t)\}_{t=0}^{\infty} \text{ feasible} \right\}. \tag{15.3}$$

Then by the principle of optimality (Theorem 7.1), the Bellman equation

$$V^*(x) = \sup_{y \in \Gamma(x)} \{u(x, y) + \beta V^*(g(x, y))\} \tag{15.4}$$

holds. Therefore V^* is a fixed point of the operator T defined by

$$(TV)(x) = \sup_{y \in \Gamma(x)} \{u(x, y) + \beta V(g(x, y))\}, \tag{15.5}$$

where $V : X \to (-\infty, \infty]$ is any function.

15.3 Verification theorem

The above argument suggests that we may solve the problem (15.2) by computing a fixed point of T. The following lemma provides a simple sufficient condition.

Lemma 15.1 (Verification Theorem)**.** *Let X, Y be nonempty sets, $u : X \times Y \to [-\infty, \infty)$, $\Gamma : X \twoheadrightarrow Y$ be such that $\Gamma(x) \neq \emptyset$ for all $x \in X$, and $g : X \times Y \to X$. Suppose that V is a fixed point of the Bellman operator (15.5). Then the followings are true.*

1. *If for any feasible $\{(x_t, y_t)\}_{t=0}^{\infty}$ we have*

$$\limsup_{t \to \infty} \beta^t V(x_t) \geq 0, \tag{15.6}$$

then $V(x_0) \geq V^*(x_0)$.

[1] For this purpose, the discounted sum of utility $\sum_{t=0}^{\infty} \beta^t u(x_t, y_t)$ needs to be well defined in the first place. This needs to be verified in particular applications. Here we simply assume that the objective function is well defined. If the existence of a limit is a priori unclear, we may write (15.2a) as $\liminf_{T \to \infty} \sum_{t=0}^{T} \beta^t u(x_t, y_t)$.

2. *Suppose that for any $x \in X$, the set*

$$\Gamma^*(x) := \arg\max_{y \in \Gamma(x)} \{u(x,y) + \beta V(g(x,y))\} \tag{15.7}$$

is nonempty. If $y_t \in \Gamma^*(x_t)$ for all t and

$$\liminf_{t\to\infty} \beta^t V(x_t) \leq 0, \tag{15.8}$$

then $V(x_0) \leq V^*(x_0)$.

In particular, if both (15.6) and (15.8) hold, then $V = V^*$ and any feasible $\{(x_t, y_t)\}_{t=0}^{\infty}$ with $y_t \in \Gamma^*(x_t)$ is a solution to (15.2).

Proof. Suppose that (15.6) holds for any feasible $\{(x_t, y_t)\}_{t=0}^{\infty}$. Since by assumption V satisfies the Bellman equation (15.4), for any $x_t \in X$ we have

$$V(x_t) = \sup_{y_t \in \Gamma(x_t)} \{u(x_t, y_t) + \beta V(g(x_t, y_t))\} \geq u(x_t, y_t) + \beta V(x_{t+1}),$$

where we have used $x_{t+1} = g(x_t, y_t)$ by feasibility. Iterating this inequality, we obtain

$$V(x_0) \geq \sum_{t=0}^{T-1} \beta^t u(x_t, y_t) + \beta^T V(x_T).$$

Taking the lim sup as $T \to \infty$ and noting that $\sum_{t=0}^{\infty} \beta^t u(x_t, y_t)$ exists by assumption, it follows from (15.6) that

$$V(x_0) \geq \sum_{t=0}^{\infty} \beta^t u(x_t, y_t) + \limsup_{T\to\infty} \beta^T V(x_T) \geq \sum_{t=0}^{\infty} \beta^t u(x_t, y_t).$$

Taking the supremum over all feasible sequences, we obtain $V(x_0) \geq V^*(x_0)$.

Next, suppose that Γ^* in (15.7) is nonempty and a feasible sequence $\{(x_t, y_t)\}_{t=0}^{\infty}$ satisfies $y_t \in \Gamma^*(x_t)$ and (15.8). Then by (15.7), we obtain

$$V(x_t) = u(x_t, y_t) + \beta V(x_{t+1}).$$

Iterating this equation, we obtain

$$V(x_0) = \sum_{t=0}^{T-1} \beta^t u(x_t, y_t) + \beta^T V(x_T).$$

Taking the lim inf as $T \to \infty$ and noting that $\sum_{t=0}^{\infty} \beta^t u(x_t, y_t)$ exists by assumption, it follows from (15.8) that

$$V(x_0) = \sum_{t=0}^{\infty} \beta^t u(x_t, y_t) + \liminf_{T\to\infty} \beta^T V(x_T) \leq \sum_{t=0}^{\infty} \beta^t u(x_t, y_t) \leq V^*(x_0).$$

In particular, if both (15.6) and (15.8) hold, then $V = V^*$ and any feasible $\{(x_t, y_t)\}_{t=0}^{\infty}$ with $y_t \in \Gamma^*(x_t)$ is a solution to (15.2). □

The conditions (15.6) and (15.8) are called *transversality conditions*. In general, transversality conditions are boundary conditions at infinity that are necessary or sufficient for optimality, which take various forms. In the case of Lemma 15.1, these conditions are sufficient but not necessary. For more discussion, see Kamihigashi (2002, 2014).

Example 15.1. Let us apply Lemma 15.1 to solve the optimal consumption-saving problem (15.1) for the case $0 < \gamma < 1$. Since $u(y) = \frac{y^{1-\gamma}}{1-\gamma}$ is homogeneous of degree $1-\gamma$ and $g(x,y) = R(x-y)$ is homogeneous of degree 1, we can conjecture that the value function V is also homogeneous of degree $1-\gamma$. Hence conjecture a solution to the Bellman equation of the form

$$V(x) = a\frac{x^{1-\gamma}}{1-\gamma}$$

for some $a > 0$. Then the Bellman equation (15.4) becomes

$$a\frac{x^{1-\gamma}}{1-\gamma} = \max_{0 \le y \le x} \left\{ \frac{y^{1-\gamma}}{1-\gamma} + \beta a \frac{[R(x-y)]^{1-\gamma}}{1-\gamma} \right\}.$$

Since $R, a > 0$, it is straightforward to show that the expression inside the braces is a concave function of y. Therefore by Proposition 11.1, the first-order condition for optimality is sufficient for the maximum, which is

$$y^{-\gamma} - \beta a R^{1-\gamma}(x-y)^{-\gamma} = 0 \iff y = \frac{x}{1 + (\beta R^{1-\gamma} a)^{1/\gamma}}.$$

Substituting this into the Bellman equation, after some algebra we obtain

$$a\frac{x^{1-\gamma}}{1-\gamma} = \left[1 + (\beta R^{1-\gamma} a)^{1/\gamma}\right]^{\gamma} \frac{x^{1-\gamma}}{1-\gamma}.$$

Comparing the coefficients of $\frac{x^{1-\gamma}}{1-\gamma}$, we obtain

$$\begin{aligned} a = \left[1 + (\beta R^{1-\gamma} a)^{1/\gamma}\right]^{\gamma} &\iff a^{1/\gamma} = 1 + (\beta R^{1-\gamma} a)^{1/\gamma} \\ &\iff a^{1/\gamma} = [1 - (\beta R^{1-\gamma})^{1/\gamma}]^{-1}, \end{aligned}$$

provided that $\beta R^{1-\gamma} < 1$. Suppose that this is the case. Then the consumption function must be

$$y = \frac{x}{1 + (\beta R^{1-\gamma} a)^{1/\gamma}} = a^{-1/\gamma} x = [1 - (\beta R^{1-\gamma})^{1/\gamma}]x.$$

In the context of Lemma 15.1, we have verified that $V(x) = a\frac{x^{1-\gamma}}{1-\gamma}$ with $a = [1-(\beta R^{1-\gamma})^{1/\gamma}]^{-\gamma}$ satisfies the Bellman equation and $\Gamma^*(x) = \left\{[1 - (\beta R^{1-\gamma})^{1/\gamma}]x\right\}$. To show that this is the solution to the optimal consumption-saving problem (15.1), it remains to show the transversality conditions (15.6) and (15.8).

Since $\gamma > 0$, we have $V(x) \ge 0$, so (15.6) is trivial. Under the consumption policy $y = [1-(\beta R^{1-\gamma})^{1/\gamma}]x$, by the budget constraint the next period's wealth is

$$x' = R(x-y) = (\beta R)^{1/\gamma} x.$$

Therefore $x_t = (\beta R)^{t/\gamma} x_0$, and

$$\beta^t V(x_t) = \beta^t a \frac{[(\beta R)^{t/\gamma} x_0]^{1-\gamma}}{1-\gamma} = a\frac{x_0^{1-\gamma}}{1-\gamma}(\beta R^{1-\gamma})^{t/\gamma} \to 0$$

as $t \to \infty$ because $\beta R^{1-\gamma} < 1$ by assumption. Therefore (15.8) holds.

Remark. When $\gamma > 1$, the condition (15.6) does not hold, so we cannot apply Lemma 15.1 to show that the consumption rule $y = [1-(\beta R^{1-\gamma})^{1/\gamma}]x$ is optimal. It turns out that this is indeed optimal, but a different argument is required, as we shall see below. See, for example, Toda (2014, 2019).

15.4 Contraction argument

Returning to the dynamic optimization problem (15.2), since V^* is a fixed point of the Bellman equation (15.4), one way to compute V^* is to show that the Bellman operator T in (15.5) is a contraction and then apply the contraction mapping theorem (Theorem 8.1). For any set X, let bX denote the space of bounded functions from X to $\mathbb{R}$. Then by Example 8.2, bX is a Banach space with the supremum norm $\|f\| = \sup_{x \in X} |f(x)|$. The following theorem provides an algorithm for computing the value function V^* in (15.3).

Theorem 15.2. *Let X, Y be nonempty sets, $u \in b(X \times Y)$, $\Gamma : X \twoheadrightarrow Y$ be such that $\Gamma(x) \neq \emptyset$ for all $x \in X$, and $g : X \times Y \to X$. Then the followings are true.*

1. *$T : bX \to bX$ is a contraction mapping.*
2. *The value function V^* in (15.3) is the unique fixed point of T.*
3. *For any $V^{(0)} \in bX$, letting $V^{(n)} = T^n V^{(0)}$, we have $\left\|V^* - V^{(n)}\right\| \to 0$ as $n \to \infty$.*

Proof. We know from the principle of optimality (Theorem 7.1) that V^* is a fixed point of T. Therefore by the contraction mapping theorem (Theorem 8.1), it suffices to show that (i) $V^* \in bX$, and (ii) $T : bX \to bX$ is a contraction. Since by assumption $u \in b(X \times Y)$, we have

$$\left|\sum_{t=0}^{\infty} \beta^t u(x_t, y_t)\right| \leq \sum_{t=0}^{\infty} \beta^t |u(x_t, y_t)| \leq \sum_{t=0}^{\infty} \beta^t \|u\| = \frac{1}{1-\beta} \|u\| < \infty.$$

Therefore the sum of discounted utility is always finite. In particular, the value function V^* exists and $\|V^*\| \leq \frac{1}{1-\beta} \|u\|$, so $V^* \in bX$.

To show that $T : bX \to bX$ is a contraction, we verify Blackwell's sufficient conditions (Theorem 8.3). If $V_1, V_2 \in bX$ and $V_1 \leq V_2$, then by the definition of T in (15.5), we have

$$\begin{aligned}(TV_1)(x) &= \sup_{y \in \Gamma(x)} \{u(x, y) + \beta V_1(g(x, y))\} \\ &\leq \sup_{y \in \Gamma(x)} \{u(x, y) + \beta V_2(g(x, y))\} = (TV_2)(x),\end{aligned}$$

so $TV_1 \leq TV_2$. Therefore monotonicity holds. If $V \in bX$ and $c \geq 0$, then

$$\begin{aligned}(T(V + c))(x) &= \sup_{y \in \Gamma(x)} \{u(x, y) + \beta(V(g(x, y)) + c)\} \\ &= \sup_{y \in \Gamma(x)} \{u(x, y) + \beta V(g(x, y))\} + \beta c = (TV)(x) + \beta c,\end{aligned}$$

so $T(V + c) \leq TV + \beta c$. Therefore discounting holds. Hence by Theorem 8.3, $T : bX \to bX$ is a contraction. □

The last part of Theorem 15.2 shows that to compute the value function V^*, it suffices to start from any bounded function $V^{(0)}$ and iterate the Bellman equation, which is called *value function iteration*.

Theorem 15.2 does not say anything about the existence of a solution. By combining Theorem 15.2 and Lemma 15.1 and putting more topological structure, we can prove the existence of a solution to the dynamic programming

problem (15.2). Recall that for a topological space X, we denote the space of bounded continuous functions $f : X \to \mathbb{R}$ by bcX.

Theorem 15.3. *Let X, Y be topological spaces, $u \in bc(X \times Y)$, $\Gamma : X \twoheadrightarrow Y$ continuous, and $g : X \times Y \to X$ continuous. Then the followings are true.*

1. *$T : bcX \to bcX$ is a contraction mapping.*
2. *The value function V^* in (15.3) is the unique fixed point of T. In particular, $V^* \in bcX$.*
3. *For any $V^{(0)} \in bcX$, letting $V^{(n)} = T^n V^{(0)}$, we have $\left\|V^* - V^{(n)}\right\| \to 0$ as $n \to \infty$.*
4. *If in addition $\Gamma^*(x)$ in (15.7) is nonempty for each $x \in X$, then any feasible sequence $\{(x_t, y_t)\}_{t=0}^{\infty}$ with $y_t \in \Gamma^*(x_t)$ is a solution to the problem (15.2).*

Proof. We know from Theorem 15.2 that $T : bX \to bX$ is a contraction. Since $bcX \subset bX$, to show that $T : bcX \to bcX$ is a contraction, it suffices to show that for any $V \in bcX$, the function $TV : X \to \mathbb{R}$ is continuous. But since by assumption $u : X \times Y \to \mathbb{R}$, $\Gamma : X \twoheadrightarrow Y$, and $g : X \times Y \to X$ are all continuous, the claim immediately follows from the maximum theorem (Theorem 13.3).

By the contraction mapping theorem, $T : bcX \to bcX$ has a unique fixed point $V \in bcX$. Since $bcX \subset bX$, V is also a fixed point of $T : bX \to bX$. Since by Theorem 15.2 V^* is the unique fixed point of $T : bX \to bX$, it must be $V^* = V$. Therefore V^* is the unique fixed point of $T : bcX \to bcX$, and in particular $V^* \in bcX$.

By Theorem 15.2, for any $V^{(0)} \in bX$ (in particular, $V^{(0)} \in bcX$), letting $V^{(n)} = TV^{(0)}$, we have $\left\|V^* - V^{(n)}\right\| \to 0$.

Since $V^* \in bcX$, for any feasible $\{(x_t, y_t)\}_{t=0}^{\infty}$, we have

$$\left|\beta^t V^*(x_t)\right| \leq \beta^t \left\|V^*\right\| \to 0$$

as $t \to \infty$, so the transversality conditions (15.6) and (15.8) hold. Therefore by the verification theorem (Lemma 15.1), any feasible sequence $\{(x_t, y_t)\}_{t=0}^{\infty}$ with $y_t \in \Gamma^*(x_t)$ is a solution to the problem (15.2). □

Corollary 15.4. *If $\Gamma(x)$ in Theorem 15.3 is nonempty and compact, then $\Gamma^*(x) \neq \emptyset$. Consequently, a solution to the problem 15.2 exists.*

Proof. By Theorem 15.3, the value function V^* is continuous. Since by assumption $\Gamma(x)$ is nonempty and compact, by the extreme value theorem (Theorem 2.5) the right-hand side of (15.7) attains a maximum and $\Gamma^*(x) \neq \emptyset$. □

Almost all dynamic programming problems that appear in applications do not admit closed-form solutions and thus need to be solved numerically. Theorem 15.3 shows that under its assumptions, the solution can be computed by value function iteration.

Although the transversality conditions (15.6) and (15.8) do not explicitly appear in Theorem 15.3 (because they automatically hold due to the boundedness of the value function), in general they cannot be omitted, as the following example shows.

Example 15.2. Consider the optimal consumption-saving problem (15.1) with $\gamma = 0$. The Bellman equation is then

$$V(x) = \max_{0 \le y \le x} \{y + \beta V(R(x-y))\}.$$

Since the utility function is linear, we can conjecture that the value function is also linear, so $V(x) = ax$ for some $a > 0$. Then

$$ax = \max_{0 \le y \le x} \{y + \beta a R(x-y)\}.$$

If $\beta R = 1$, then $V(x) = x$ $(a = 1)$ clearly satisfies the Bellman equation, and any $y \in [0, x]$ is a maximizer. Take $y_t = 0$ for all t. Then $x_t = R^t x_0$, and

$$\beta^t V(x_t) = \beta^t R^t x_0 = (\beta R)^t x_0 = x_0 > 0,$$

so the transversality condition (15.8) does not hold. Indeed, choosing $(0, 0, \dots)$ (consuming zero forever) gives lifetime utility 0, whereas choosing $(x_0, 0, \dots)$ (consuming everything now and zero in the future) gives lifetime utility $x_0 > 0$, so $(0, 0, \dots)$ is not optimal.

15.5 Non-contraction argument

Theorem 15.3 is quite elegant in that it proves the existence, uniqueness, and a computational algorithm for the solution to a dynamic programming problem. Unfortunately, Theorem 15.3 is quite unsatisfactory from an applied perspective because its assumptions are often not satisfied in applications. For example, in many applications it is common to use the constant relative risk aversion (CRRA) utility function

$$u(y) = \begin{cases} \frac{y^{1-\gamma}}{1-\gamma}, & (0 < \gamma \neq 1) \\ \log y. & (\gamma = 1) \end{cases}$$

It is clear that $u(y)$ is unbounded above if $\gamma \le 1$ and unbounded below if $\gamma \ge 1$, so $u(y)$ is always unbounded. Therefore unless we a priori know that y takes values in a bounded set (and bounded away from 0), we cannot apply Theorem 15.3.

Sometimes, we can prove the existence of a solution using value function iteration starting from the zero function.

Theorem 15.5. *Let X, Y be nonempty sets, $u : X \times Y \to [-\infty, \infty)$, $\Gamma : X \twoheadrightarrow Y$ be such that $\Gamma(x) \neq \emptyset$ for all $x \in X$, and $g : X \times Y \to X$. Define $V^{(0)}(x) \equiv 0$, $V^{(n)}(x) = T^n V^{(0)}$, and*

$$V(x) = \liminf_{n \to \infty} V^{(n)}(x). \tag{15.9}$$

Then $V \ge V^*$ and $V \ge TV$. If in addition $V \le TV$, Γ^* in (15.7) is nonempty, and the transversality condition (15.8) holds, then $V = V^*$ and any feasible $\{(x_t, y_t)\}_{t=0}^{\infty}$ with $y_t \in \Gamma^*(x_t)$ is a solution to (15.2).

Proof. Take any feasible sequence $\{(x_t, y_t)\}_{t=0}^{\infty}$. Then by definition

$$V^{(n)}(x_t) = \sup_{y_t \in \Gamma(x_t)} \left\{ u(x_t, y_t) + \beta V^{(n-1)}(g(x_t, y_t)) \right\} \ge u(x_t, y_t) + \beta V^{(n-1)}(x_{t+1}).$$

Iterating this for $n = T, T-1, \ldots, 1$, we obtain

$$V^{(T)}(x_0) \geq \sum_{t=0}^{T-1} \beta^t u(x_t, y_t) + \beta^T V^{(0)}(x_T) = \sum_{t=0}^{T-1} \beta^t u(x_t, y_t),$$

where we have used $V^{(0)} \equiv 0$. Taking the lim inf as $T \to \infty$ and noting that $\sum_{t=0}^{\infty} \beta^t u(x_t, y_t)$ exists by assumption, we obtain

$$V(x_0) \geq \sum_{t=0}^{\infty} \beta^t u(x_t, y_t).$$

Taking the supremum over all feasible sequences, we obtain $V(x_0) \geq V^*(x_0)$.

Similarly, by definition

$$V^{(n)}(x) = (TV^{(n-1)})(x) \geq u(x, y) + \beta V^{(n-1)}(g(x, y)).$$

Taking the lim inf as $n \to \infty$, we obtain

$$V(x) \geq u(x, y) + \beta V(g(x, y)).$$

Taking the supremum over $y \in \Gamma(x)$, we obtain $V \geq TV$.

If $V \leq TV$, then (since $V \geq TV$) V is a fixed point of T. Therefore if Γ^* in (15.7) is nonempty and the transversality condition (15.8) holds, then by the verification theorem (Lemma 15.1) we have $V \leq V^*$, so it must be $V = V^*$. □

Corollary 15.6 (Positive utility)**.** *Let everything be as in Theorem 15.5 and suppose $u \geq 0$. Then $TV = V$. Consequently, if Γ^* in (15.7) is nonempty and the transversality condition (15.8) holds, then $V = V^*$ and any feasible $\{(x_t, y_t)\}_{t=0}^{\infty}$ with $y_t \in \Gamma^*(x_t)$ is a solution to (15.2).*

Proof. Since $u(x, y) \geq 0$ and $V^{(0)} \equiv 0$, we have

$$V^{(1)}(x) = \sup_{y \in \Gamma(x)} u(x, y) \geq 0 = V^{(0)}(x).$$

Using the monotonicity of T, by induction we obtain

$$V^{(n)} \geq V^{(n-1)} \geq \cdots \geq V^{(1)} \geq 0.$$

Therefore $V^{(n)} \uparrow V$ as $n \to \infty$, so

$$V^{(n)}(x) = \sup_{y \in \Gamma(x)} \left\{ u(x, y) + \beta V^{(n-1)}(g(x, y)) \right\} \leq \sup_{y \in \Gamma(x)} \left\{ u(x, y) + \beta V(g(x, y)) \right\}.$$

Letting $n \to \infty$, we obtain $V \leq TV$. The conclusion follows from Theorem 15.5. □

Problems

Part III

Introduction to Numerical Analysis

Chapter 16

Solving Nonlinear Equations

So far, we have studied optimization problems from a theoretical perspective. If the objective function happens to be convex or concave, to minimize or maximize it, all we need to do is to find a point at which the derivative is zero. This is easier said than done. In practice, almost all problems have no closed-form solutions, and therefore we need to use some kind of numerical algorithms to find the (approximate) solution. Note that if a (one-variable) function f is differentiable, the first-order condition for optimality is $f'(x) = 0$. Letting $g(x) = f'(x)$, it thus suffices to solve the nonlinear equation $g(x) = 0$. This chapter discusses algorithms for solving such nonlinear equations.

16.1 Bisection method

Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function and suppose that

$$g(x) \begin{cases} < 0, & (x < x^*) \\ = 0, & (x = x^*) \\ > 0. & (x > x^*) \end{cases}$$

These inequalities show that we know exactly whether the current approximate solution x is greater or less than the true solution x^* according as $g(x) \gtrless 0$. The idea of the bisection method is to decrease x if $g(x) > 0$ and increase x if $g(x) < 0$, until we find a value such that $g(x) \approx 0$.

To describe the algorithm, let

$$\begin{aligned} x &= \text{current approximate solution}, \\ \underline{x} &= \text{current lower bound of } x^*, \\ \bar{x} &= \text{current upper bound of } x^*, \\ \varepsilon &= \text{error tolerance for } x^*. \end{aligned}$$

The following is the bisection algorithm.

Initialization Verify that $g(x) = 0$ has a unique solution and crosses 0 from below (e.g., g is increasing). Select the error tolerance $\varepsilon > 0$. Find $\underline{x}$ and $\bar{x}$ such that $g(\underline{x}) < 0$ and $g(\bar{x}) > 0$.

Iteration 1. Compute $x = \frac{\underline{x}+\bar{x}}{2}$ and $g(x)$.

2. If $g(x) < 0$, set $\underline{x} = x$. If $g(x) > 0$, set $\bar{x} = x$.

Stopping rule If $\bar{x} - \underline{x} < \varepsilon$, stop. The approximate solution is x. Otherwise, repeat the iteration step.

The bisection method also works when g crosses 0 from above, but the updating rule of the lower and upper bounds must be interchanged in the obvious way.

The bisection method is a sure way to obtain a solution but is slow. Since the interval gets halved at each iteration, after n iterations the length of the interval is of the order 2^{-n}. Therefore convergence is (only) exponentially fast.

16.2 Order of convergence

At this point it is useful to define how fast an algorithm converges. Let $\{x_n\}_{n=0}^{\infty}$ be the sequence of approximate solutions generated by some algorithm. Let x^* be the true solution. We say that the *order of convergence* of the algorithm is α if there exist constants $\alpha \geq 1$ and $\beta > 0$ such that

$$|x_{n+1} - x^*| \leq \beta \, |x_n - x^*|^{\alpha} \tag{16.1}$$

for sufficiently large n.

If $\alpha = 1$, we also require $\beta < 1$ to guarantee convergence. In that case, by iteration (and assuming that the inequality (16.1) holds for all n) we get

$$|x_n - x^*| \leq \beta^n \, |x_0 - x^*| \, ,$$

so $\{x_n\}$ converges to x^* exponentially fast. Therefore the bisection method has order of convergence 1. If $\alpha > 1$, then $\{x_n\}$ converges to x^* *double exponentially*. To see this, let us find a constant $C > 0$ such that

$$C \, |x_{n+1} - x^*| \leq (C \, |x_n - x^*|)^{\alpha}.$$

Comparing with the definition of the order of convergence (16.1), it suffices to choose C such that $\beta = C^{\alpha-1} \iff C = \beta^{\frac{1}{\alpha-1}}$. Iterating (16.1) over n, we obtain

$$C \, |x_n - x^*| \leq (C \, |x_0 - x^*|)^{\alpha^n},$$

so provided that $|x_0 - x^*| < 1/C$, we get

$$|x_n - x^*| \leq C^{-1} (C \, |x_0 - x^*|)^{\alpha^n} \to 0,$$

and the speed of convergence is double exponentially fast.

How many iterations are needed to compute the solution up to some decimal place, say d? When the order of convergence is 1 (exponential), the number of iterations required is approximately

$$\beta^n = 10^{-d} \iff n = -\frac{d}{\log_{10} \beta} \implies n = \text{constant} \times d.$$

On the other hand, when the order of convergence is $\alpha > 1$, the number of iterations required is approximately

$$(C\,|x_0 - x^*|)^{\alpha^n} = 10^{-d} \iff \alpha^n = -\frac{d}{\log_{10} C\,|x_0 - x^*|}$$
$$\implies n = \frac{1}{\log \alpha}\left(\log d + \text{constant}\right).$$

Since $\log d$ is much smaller than d, in practice it is important to use algorithms that have order $\alpha > 1$.

16.3 Newton method

The bisection method is inefficient in the sense that the only information of $g(x)$ the algorithm uses is its sign. Unsurprisingly, the order of convergence is 1, which is slow. The Newton method, which is based on Taylor's theorem, uses the function value and the derivative and achieves a much faster convergence.

The idea of the Newton method is as follows. Let g be continuously differentiable. Suppose that you have an approximate solution at $x = a$. By Taylor's theorem, we have

$$g(x) \approx g(a) + g'(a)(x - a).$$

Since the right-hand side is linear in x, we can just solve it to obtain

$$0 = g(a) + g'(a)(x - a) \iff x = a - \frac{g(a)}{g'(a)}.$$

The formal algorithm of the Newton method is as follows.

1. Pick an initial value x_0 and error tolerance $\varepsilon > 0$.
2. For $n = 1, 2, \ldots$, compute
$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}. \tag{16.2}$$
3. Stop if $|x_{n+1} - x_n| < \varepsilon$. The approximate solution is x_{n+1}.

The following theorem shows that the Newton method has order of convergence 2.

Theorem 16.1. *Let $g : \mathbb{R} \to \mathbb{R}$ be twice continuously differentiable. Suppose that $x^* \in \mathbb{R}$ satisfies $g(x^*) = 0$ and $g'(x^*) \neq 0$. Then there exists a constant $C > 0$ and a neighborhood U of x^* such that $x_n \in U$ implies*

$$|x_{n+1} - x^*| \leq C\,|x_n - x^*|^2.$$

Proof. Let x_n be the current approximate solution. Subtracting x^* from both sides of (16.2), we get

$$x_{n+1} - x^* = x_n - x^* - \frac{g(x_n)}{g'(x_n)} = -\frac{g(x_n) + g'(x_n)(x^* - x_n)}{g'(x_n)}.$$

Since g is twice continuously differentiable, applying Taylor's theorem to $g(x^*)$ around x_n, there exists $t \in [0,1]$ such that $\xi \coloneqq (1-t)x^* + tx_n$ satisfies

$$0 = g(x^*) = g(x_n) + g'(x_n)(x^* - x_n) + \frac{1}{2}g''(\xi)(x^* - x_n)^2.$$

Substituting into the expression for $x_{n+1} - x^*$ and assuming $g'(x_n) \neq 0$, we get

$$x_{n+1} - x^* = \frac{g''(\xi)}{2g'(x_n)}(x_n - x^*)^2.$$

Since by assumption g is twice continuously differentiable and $g'(x^*) \neq 0$, we can take a neighborhood U of x^* such that $g'(x) \neq 0$ for $x \in U$ and

$$\beta \coloneqq \sup_{t\in[0,1]} \sup_{x\in U} \left| \frac{g''((1-t)x + tx^*)}{2g'(x)} \right| < \infty.$$

Then $|x_{n+1} - x^*| \leq \beta |x_n - x^*|^2$ whenever $x_n \in U$, so the order of convergence of the Newton method is (at least) 2. □

The Newton method can be applied to solve a system of nonlinear equations. For example, let $g : \mathbb{R}^N \to \mathbb{R}^N$ and we would like to solve $g(x) = 0$. By Taylor's theorem, we have

$$g(x) \approx g(a) + Dg(a)(x - a) \iff x \approx a - [Dg(a)]^{-1}g(a),$$

where Dg denotes the $N \times N$ Jacobian of g. Thus if x_0 is close to a true solution x^* and $Dg(x^*)$ is regular, we can expect that iterating

$$x_{n+1} = x_n - [Dg(x_n)]^{-1}g(x_n)$$

converges to x^* (Problem 16.4).

16.4 Linear interpolation

The Newton method requires the function value $g(x)$ and its derivative $g'(x)$ to implement it. Oftentimes, the derivative $g'(x)$ has a complicated form. In some cases (e.g., the objective function is defined only numerically, not analytically), it is impossible to compute the derivative. In such cases, we can use linear interpolation to solve for the solution.

Let x_n and x_{n-1} be the two most recent approximate solutions to $g(x) = 0$. Approximating g by the linear function that agrees with g at these two points, we obtain

$$g(x) \approx \frac{g(x_n) - g(x_{n-1})}{x_n - x_{n-1}}(x - x_n) + g(x_n).$$

Setting the right-hand side equal to 0, we obtain

$$\begin{aligned} &\frac{g(x_n) - g(x_{n-1})}{x_n - x_{n-1}}(x - x_n) + g(x_n) = 0 \\ \iff &x_{n+1} = x_n - g(x_n)\frac{x_n - x_{n-1}}{g(x_n) - g(x_{n-1})}. \end{aligned} \tag{16.3}$$

Problem 16.3 asks you to show that the order of convergence of the linear interpolation method is the golden ratio $\alpha = \frac{1+\sqrt{5}}{2} = 1.618\ldots$.

16.5 Quadratic interpolation

The linear interpolation method approximates a nonlinear function by a linear one by interpolating between two points. This way, we can solve for the new approximate solution explicitly by solving a linear equation. However, we can also solve quadratic equations explicitly. The quadratic interpolation method fits a quadratic function to three points.

Suppose that you have three approximate solutions $a < b < c$ to the the nonlinear equation $g(x) = 0$, with $g(a)g(c) < 0$. The quadratic interpolation method constructs a quadratic function that agrees with g at these three points, and then finds the root. By direct substitution, we can show that the quadratic function

$$\begin{aligned} q(x) &= g(a)\frac{(x-b)(x-c)}{(a-b)(a-c)} + g(b)\frac{(x-c)(x-a)}{(b-c)(b-a)} + g(c)\frac{(x-a)(x-b)}{(c-a)(c-b)} \\ &= Ax^2 + Bx + C \end{aligned}$$

satisfies $q(x) = g(x)$ for $x = a, b, c$. Comparing the coefficients, we obtain

$$\begin{aligned} A &= \frac{g(a)}{(a-b)(a-c)} + \frac{g(b)}{(b-c)(b-a)} + \frac{g(c)}{(c-a)(c-b)}, \\ B &= -\frac{g(a)(b+c)}{(a-b)(a-c)} - \frac{g(b)(c+a)}{(b-c)(b-a)} - \frac{g(c)(a+b)}{(c-a)(c-b)}, \\ C &= \frac{g(a)bc}{(a-b)(a-c)} + \frac{g(b)ca}{(b-c)(b-a)} + \frac{g(c)ab}{(c-a)(c-b)}. \end{aligned}$$

Using the formula for the solution to a quadratic equation, we obtain

$$x = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A},$$

where we should pick the sign $\pm$ such that $a < x < c$.

The quadratic interpolation algorithm is defined as follows.

1. Pick initial values $a_0 < b_0 < c_0$ and error tolerance $\varepsilon > 0$.
2. For each n, compute $d = x_{n+1}$ given current a_n, b_n, c_n. Stop if $|x_{n+1} - x_n| < \varepsilon$. Otherwise, set

$$(a_{n+1}, b_{n+1}, c_{n+1}) = \begin{cases} (a_n, d, b_n), & (a_n < d < b_n) \\ (b_n, d, c_n). & (b_n < d < c_n) \end{cases}$$

The order of convergence of the quadratic interpolation method is the root $\alpha > 1$ of the equation

$$x^3 - x^2 - x - 1 = 0,$$

which is $\alpha = 1.8393\ldots$.

16.6 Robustifying the algorithms

The linear interpolation, quadratic interpolation, and Newton methods are usually much faster than the bisection method because the order of convergence

exceeds 1. However, this does not mean that we should always use such algorithms. There are at least two reasons. First, algorithms other than the bisection method are only *local*, meaning that they are guaranteed to converge only if the initial value is close enough to the true solution. If the initial value is far away from the true solution, the algorithm may not converge at all. On the other hand, the bisection method is a *global* algorithm (if the function has a single solution), so convergence is guaranteed (although slow). One way to make the algorithm robust is to initially use a global algorithm such the bisection method (or grid search), and then switch to a faster local algorithm.

Second, because the linear interpolation and Newton methods approximate the function by a linear function and the quadratic interpolation method by a quadratic function, the approximation may be poor when the function is highly nonlinear (or highly non-quadratic). For example, consider the equation

$$g(x) = x^{100} - 2.$$

Then the function value is of order 1 on $[0, 1]$ but is huge when $x > 1$. Then the convergence of linear and quadratic interpolation can be slow. In that case it may be useful to "robustify" the algorithm by considering the equation $h(x) = 0$ instead, where

$$h(x) = \max\{-1, \min\{g(x), 1\}\}.$$

Problems

16.1. Let $f(x) = \sqrt{x^2 + 1}$.

1. Compute $f'(x)$, $f''(x)$, and show that f is convex.
2. Find the minimum of f.
3. Using your favorite programming language, implement the Newton method for finding the minimum (solving $f'(x) = 0$). Experiment what happens when the initial values are $x_0 = 0.9$, 1, and 1.1.

16.2. Consider the nonlinear equation

$$g(x) = x^3 - 2 = 0,$$

where $x > 0$. Clearly the solution is $x = 2^{1/3} \in (1, 2)$. Using your favorite programming language, implement the bisection, linear interpolation, quadratic interpolation, and Newton methods and compare the speed of convergence. What if $g(x) = x^{100} - 2$?

16.3. This problem asks you to derive the order of convergence of the linear interpolation method. Let g be a twice continuously differentiable function with $g(x^*) = 0$ and $g'(x^*) \neq 0$. Consider the linear interpolation algorithm (16.3).

1. Let $\phi(x; a) = \frac{g(x)-g(a)}{x-a}$ for $x \neq a$. Show that

$$x_{n+1} - x^* = (x_n - x^*)\frac{\phi(x_{n-1}; x_n) - \phi(x^*; x_n)}{\phi(x_{n-1}; x_n)}.$$

2. Using the mean value theorem, show that there exists ξ_n between x_n and x_{n-1} such that $\phi(x_{n-1}; x_n) = g'(\xi_n)$.

3. Regard $\phi(x; a)$ as a function of x. Using the mean value theorem, show that there exists a number η_n between x_{n-1} and x^* such that

$$\phi(x_{n-1}; x_n) - \phi(x^*; x_n) = \phi'(\eta_n; x_n)(x_{n-1} - x^*).$$

4. Compute $\phi'(x; a)$ explicitly.

5. Using Taylor's theorem, show that there exists a number ζ_n such that

$$\phi'(\eta_n; x_n) = \frac{1}{2} g''(\zeta_n).$$

6. Show that if x_n, x_{n-1} are sufficiently close to x^*, there exists a constant $C > 0$ such that

$$|x_{n+1} - x^*| \leq C\,|x_n - x^*|\,|x_{n-1} - x^*|\,.$$

7. Show that the order of convergence of the linear interpolation method is at least $\alpha = \frac{1+\sqrt{5}}{2} = 1.618\ldots$.

16.4. Let $g : \mathbb{R}^N \to \mathbb{R}^N$ be twice continuously differentiable. Assume that $g(x^*) = 0$ and the Jacobian $Dg(x^*)$ is regular. Show that the Newton algorithm converges to x^* double exponentially fast if the initial value is close enough to x^*. (Hint: use the mean value inequality (Proposition 8.7).)

Chapter 17

Polynomial approximation

Polynomials are useful for approximating smooth functions because they can be differentiated and integrated analytically. This chapter studies the polynomial approximation of a one-variable function.

17.1 Lagrange interpolation

Since a degree $n-1$ polynomial is determined by n coefficients, once we specify n points on the xy plane, there exists (at most) one polynomial that passes through these points. Lagrange interpolation gives an explicit formula for the interpolating polynomial.

Proposition 17.1. *Let $x_1 < \cdots < x_n$ and define the k-th Lagrange polynomial*

$$L_k(x) = \frac{\prod_{l \neq k}(x - x_l)}{\prod_{l \neq k}(x_k - x_l)}$$

for $k = 1, \ldots, n$. Then

$$p(x) = \sum_{k=1}^{n} y_k L_k(x)$$

is the unique polynomial of degree up to $n-1$ satisfying $p(x_k) = y_k$ for $k = 1, \ldots, n$.

Proof. By the definition of $L_k(x)$, we have $L_k(x_l) = \delta_{kl}$, where $\delta_{kl} = 1$ if $k = l$ and $\delta_{kl} = 0$ if $k \neq l$, which is called Kronecker's delta.[1] Therefore for all l, we have

$$p(x_l) = \sum_{k=1}^{n} y_k L_k(x_l) = \sum_{k=1}^{n} y_k \delta_{kl} = y_l.$$

Clearly $L_k(x)$ is a polynomial of degree $n-1$, so $p(x)$ is a polynomial of degree up to $n-1$. □

If we interpolate a function $f(x)$ at the points $x_1 < \cdots < x_n$ by a degree $n-1$ polynomial, what is the approximation error? The following proposition gives an error bound if f is sufficiently smooth.

[1] https://en.wikipedia.org/wiki/Kronecker_delta

Proposition 17.2. *Let $f : \mathbb{R} \to \mathbb{R}$ be C^n and p_{n-1} be the interpolating polynomial of f at $x_1 < \cdots < x_n$. Then for any x, there exists ξ in the convex hull of $\{x, x_1, \ldots, x_n\}$ such that*

$$f(x) - p_{n-1}(x) = \frac{f^{(n)}(\xi)}{n!} \prod_{k=1}^{n} (x - x_k). \tag{17.1}$$

Proof. If $x = x_k$ for some k, then $f(x_k) - p_{n-1}(x_k) = 0$, so (17.1) is trivial. Suppose $x \neq x_k$ for all k and let $I = \operatorname{co}\{x, x_1, \ldots, x_n\}$. For any $t \in I$, let $R(t) = f(t) - p_{n-1}(t)$ be the error term and define

$$g(t) = R(t)S(x) - R(x)S(t),$$

where $S(t) = \prod_{k=1}^{n}(t - x_k)$. Clearly $g(x) = 0$. Furthermore, since $R(x_k) = S(x_k) = 0$, we have $g(x_k) = 0$ for $k = 1, \ldots, n$. In general, if g is differentiable and $g(a) = g(b) = 0$, by the mean value theorem (Proposition 3.2) there exists $c \in (a, b)$ such that $g'(c) = 0$. Applying this to the n non-overlapping intervals with endpoints $x, x_1, \ldots, x_n$, there exist n distinct points $y_1, \ldots, y_n$ between $x, x_1, \ldots, x_n$ such that $g'(y_k) = 0$ for $k = 1, \ldots, n$. Continuing this argument, there exists $\xi \in I$ such that $g^{(n)}(\xi) = 0$. But since S is a degree n polynomial with leading coefficient 1, we have $S^{(n)} = n!$, so

$$0 = g^{(n)}(\xi) = R^{(n)}(\xi)S(x) - R(x)n!.$$

Since $R(t) = f(t) - p_{n-1}(t)$ and $\deg p_{n-1} \leq n - 1$, we obtain $R^{(n)}(\xi) = f^{(n)}(\xi)$. Therefore

$$f(x) - p_{n-1}(x) = R(x) = \frac{1}{n!} f^{(n)}(\xi)S(x) = \frac{f^{(n)}(\xi)}{n!} \prod_{k=1}^{n} (x - x_k). \quad \square$$

17.2 Chebyshev polynomials

If we want to interpolate a function on an interval by a polynomial but we are free to choose the interpolation nodes $x_1, \ldots, x_n$, how should we choose them? By mapping the interval with an affine function, without loss of generality we may assume that the interval is $[-1, 1]$. Since $f^{(n)}(\xi)$ in (17.1) depends on the particular function f but $\prod_{k=1}^{n}(x - x_k)$ does not, it is natural to find $x_1, \ldots, x_n$ so as to minimize

$$\max_{x \in [-1,1]} \left| \prod_{k=1}^{n} (x - x_k) \right|.$$

Chebyshev[2] has solved this problem a long time ago.

The degree n Chebyshev polynomial $T_n(x)$ is obtained by expanding $\cos n\theta$ as a degree n polynomial of $\cos\theta$ and setting $x = \cos\theta$. For instance,

$$\begin{aligned} \cos 0\theta &= 1 && \implies T_0(x) = 1, \\ \cos\theta &= \cos\theta && \implies T_1(x) = x, \\ \cos 2\theta &= 2\cos^2\theta - 1 && \implies T_2(x) = 2x^2 - 1, \end{aligned}$$

[2] https://en.wikipedia.org/wiki/Pafnuty_Chebyshev

and so on. In general, adding

$$\cos(n+1)\theta = \cos n\theta \cos\theta - \sin n\theta \sin\theta,$$
$$\cos(n-1)\theta = \cos n\theta \cos\theta + \sin n\theta \sin\theta,$$

and setting $x = \cos n\theta$, we obtain

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{17.2}$$

The coefficients of Chebyshev polynomials can be easily computed by iterating (17.2).

Theorem 17.3. *The solution to*

$$\min_{x_1 \geq \cdots \geq x_n} \max_{x \in [-1,1]} \left| \prod_{k=1}^{n} (x - x_k) \right|$$

is given by $x_k = \cos \frac{2k-1}{2n}\pi$*, in which case* $\prod_{k=1}^{n}(x - x_k) = 2^{1-n}T_n(x)$.

Proof. Let $p(x) = 2^{1-n}T_n(x)$. By the recursive formula (17.2), the leading coefficient of $T_n(x)$ is 2^{n-1}. Therefore the leading coefficient of $p(x)$ is 1. Since $p(\cos\theta) = 2^{1-n}\cos n\theta$, clearly

$$\sup_{x \in [-1,1]} |p(x)| = \sup_{\theta \in [-\pi,\pi]} 2^{1-n} |\cos n\theta| = 2^{1-n}.$$

Suppose that there exists a degree n polynomial $q(x)$ with leading coefficient 1 such that $\sup_{x\in[-1,1]} |q(x)| < 2^{1-n}$. Again since $p(\cos\theta) = 2^{1-n}\cos n\theta$, we have $p(x) = (-1)^k 2^{1-n}$ at $x = y_k = \cos k\pi/n$, where $k = 0, 1, \ldots, n$. Since $|q(x)| < 2^{1-n}$ for all $x \in [-1, 1]$, by the intermediate value theorem there exist $z_1, \ldots, z_n$ between $y_0, \ldots, y_n$ such that $p(z_k) - q(z_k) = 0$. But since p, q are polynomials of degree n with leading coefficient 1, $r(x) \coloneqq p(x) - q(x)$ is a polynomial of degree up to $n-1$. Since $r(z_k) = 0$ for $k = 1, \ldots, n$, it must be $r(x) \equiv 0$ or $p \equiv q$, which is a contradiction. Therefore $\prod_{k=1}^{n}(x - x_k) = 2^{1-n}T_n(x)$, so $x_k = \cos\frac{2k-1}{2n}\pi$ for $k = 1, \ldots, n$. □

17.3 Projection

In economics, we often need to solve functional equations. For instance, in a dynamic programming problem, we need to solve for either the value function or the policy function. The projection method (a standard reference is Judd, 1992) approximates the policy function (here whatever you want to solve for) on some compact set by a polynomial.

By Theorem 17.3, if you want to approximate a smooth function f on $[-1, 1]$ by a degree $N-1$ polynomial, it is "optimal" to interpolate f at the roots of the degree N Chebyshev polynomial. The idea of the projection method is to approximate the policy function f by a linear combination of Chebyshev polynomials,

$$f(x) \approx \widehat{f}(x) = \sum_{n=0}^{N-1} a_n T_n(x),$$

and determine the coefficients $\{a_n\}_{n=0}^{N-1}$ to make the functional equation (that you want to solve) true at the Chebyshev nodes.

It is easier to see how things work by looking at a simple example. Suppose you want to solve the ordinary differential equation (ODE)

$$y'(t) = y(t) \tag{17.3}$$

with initial condition $y(0) = 1$. Of course the solution is $y(t) = \mathrm{e}^t$, but let us pretend that we do not know the solution and solve it numerically. Suppose we want to compute a numerical solution for $t \in [0, T]$, where $T > 0$ is some upper bound. We can do as follows.

1. Map $[0, T]$ to $[-1, 1]$ by the affine transformation $t \mapsto \frac{2t-T}{T}$.
2. Approximate $y(t)$ by

$$\widehat{y}(t) = \sum_{n=0}^{N-1} a_n T_n\left(\frac{2t-T}{T}\right),$$

   where $\{a_n\}_{n=0}^{N-1}$ are unknown coefficients.
3. Determine $\{a_n\}_{n=0}^{N-1}$ by setting $\widehat{y}'(t) = \widehat{y}(t)$ at t corresponding to the roots of the degree N Chebyshev polynomial; more precisely, find t_n by solving $\frac{2t_n-T}{T} = \cos\left(\frac{2n-1}{2N}\pi\right)$ for $n = 1, \ldots, N$.
4. In this example, we must also impose the initial condition $\widehat{y}(0) = 1$, so for example we can minimize the sum of squared residuals at Chebyshev nodes:

$$\underset{\{a_n\}_{n=0}^{N-1}}{\text{minimize}} \quad \sum_{n=0}^{N-1} (\widehat{y}'(t_n) - \widehat{y}(t_n))^2 \tag{17.4a}$$

$$\text{subject to} \quad \widehat{y}(0) = 1. \tag{17.4b}$$

In general, when we solve a minimization problem such as (17.4) numerically, for numerical stability it is a good idea to start from a low order approximation (say $N - 1 = 2$) and compute the solution for progressively higher order using the previous solution as an initial guess. (Implementing numerical methods is an art and often requires a lot of trial and error.) Figure 17.1 shows the $\log_{10}$ relative errors when $T = 4$ and $N - 1 = 3, 4, \ldots$. We can see that the relative errors become smaller as we increase the degree of polynomial approximation.

For more details on the projection method, see Pohl et al. (2018), which has a nice application to solving asset pricing models.

Problems

17.1. Using your favorite programming language, write a code that computes the coefficients of a Chebyshev polynomial of a given degree.

17.2. Using your favorite programming language, implement the projection method for solving the ordinary differential equation (17.3) and replicate Figure 17.1.

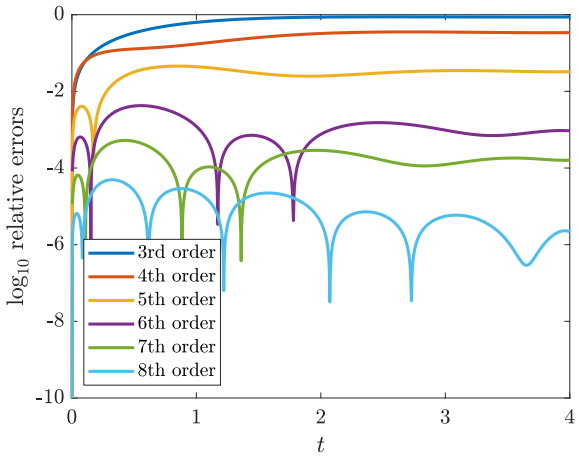

Figure 17.1. $\log_{10}$ relative errors for solving the ODE (17.3).

Chapter 18

Quadrature and Discretization

Many economic problems involve maximizing the expected value. Unless the distribution is discrete, expectations become integrals, and we cannot compute them explicitly except for special cases. Therefore we need numerical methods to evaluate integrals, which are called *quadrature* (or numerical integration).

A typical quadrature formula has the form

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{n=1}^{N} w_n f(x_n), \tag{18.1}$$

where f is a general integrand, $\{x_n\}_{n=1}^N$ are nodes, and $\{w_n\}_{n=1}^N$ are weights of the quadrature rule. In this chapter we cover the most basic theory of quadrature. See Davis and Rabinowitz (1984) for a more complete textbook treatment.

18.1 Newton-Cotes quadrature

The simplest quadrature rule is to divide the interval $[a, b]$ into $N - 1$ evenly spaced subintervals (so $x_n = a + \frac{n-1}{N-1}(b - a)$ for $n = 1, \ldots, N$) and choose the weights $\{w_n\}_{n=1}^N$ so that one can integrate all polynomials of degree $N - 1$ or less exactly. This quadrature rule is known as the N-point Newton-Cotes rule. Since we can map the interval $[0, 1]$ to $[a, b]$ through the linear transformation $x \mapsto a + (b - a)x$, without loss of generality let us assume $a = 0$ and $b = 1$. We consider several examples.

18.1.1 Trapezoidal rule ($N = 2$)

The 2-point Newton-Cotes rule is known as the *trapezoidal rule*. In this case we have $x_n = 0, 1$, and we choose w_1, w_2 to integrate a linear function exactly.

Therefore requiring that (18.1) holds exactly for $f(x) = 1, x$, we obtain

$$1 = \int_0^1 1 \, \mathrm{d}x = w_1 + w_2,$$
$$\frac{1}{2} = \int_0^1 x \, \mathrm{d}x = w_2.$$

Solving these equations, we obtain $w_1 = w_2 = \frac{1}{2}$. Changing the interval from $[0, 1]$ to $[a, b]$, the trapezoidal rule becomes

$$\int_a^b f(x) \, \mathrm{d}x \approx \frac{b-a}{2}(f(a) + f(b)). \tag{18.2}$$

Let us estimate the error of this approximation. Let $p(x)$ be the degree 1 interpolating polynomial of f at $x = a, b$. Since p agrees with f at a, b, clearly

$$\int_a^b p(x) \, \mathrm{d}x = \frac{b-a}{2}(f(a) + f(b)).$$

Therefore by Proposition 17.2, we obtain

$$\begin{aligned}\int_a^b f(x) \, \mathrm{d}x - \frac{b-a}{2}(f(a) + f(b)) &= \int_a^b (f(x) - p(x)) \, \mathrm{d}x \\ &= \int_a^b \frac{f''(\xi(x))}{2}(x-a)(x-b) \, \mathrm{d}x,\end{aligned}$$

where $\xi(x) \in (a, b)$. Since $(x-a)(x-b) < 0$ on (a, b), by the mean value theorem for Riemann-Stieltjes integrals, there exists $c \in (a, b)$ such that

$$\int_a^b \frac{f''(\xi(x))}{2}(x-a)(x-b) \, \mathrm{d}x = \frac{f''(c)}{2} \int_a^b (x-a)(x-b) \, \mathrm{d}x = -\frac{f''(c)}{12}(b-a)^3.$$

Therefore we can estimate the error in (18.2) as

$$\left| \int_a^b f(x) \, \mathrm{d}x - \frac{b-a}{2}(f(a) + f(b)) \right| \leq \frac{\|f''\|}{12}(b-a)^3, \tag{18.3}$$

where $\|\cdot\|$ denotes the sup norm on $[a, b]$.

18.1.2 Simpson's rule ($N = 3$)

The 3-point Newton-Cotes rule is known as *Simpson's rule*. In this case the quadrature nodes are $x_n = 0, 1/2, 1$, and we choose the weights w_1, w_2, w_3 so as to integrate a quadratic function exactly. Therefore requiring that (18.1) holds exactly for $f(x) = 1, x, x^2$, we obtain

$$1 = \int_0^1 1 \, \mathrm{d}x = w_1 + w_2 + w_3,$$
$$\frac{1}{2} = \int_0^1 x \, \mathrm{d}x = \frac{1}{2}w_2 + w_3,$$
$$\frac{1}{3} = \int_0^1 x^2 \, \mathrm{d}x = \frac{1}{4}w_2 + w_3.$$

Solving these equations, we obtain $w_1 = w_3 = \frac{1}{6}$ and $w_2 = \frac{2}{3}$. Changing the interval from $[0,1]$ to $[a,b]$, Simpson's rule becomes

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right). \tag{18.4}$$

Interestingly, since

$$\frac{1}{4} = \int_0^1 x^3\,\mathrm{d}x = \frac{1}{8}w_2 + w_3,$$

Simpson's rule actually integrates polynomials of degree 3 exactly, even though it is not designed to do so.

To estimate the error of Simpson's rule (18.4), take any point $d \in (a,b)$ and let $p(x)$ be a degree 3 interpolating polynomial of f at $x = a, \frac{a+b}{2}, b, d$. Since Simpson's rule integrates degree 3 polynomials exactly, by Proposition 17.2 we have

$$\begin{aligned}&\int_a^b f(x)\,\mathrm{d}x - \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)\\ &= \int_a^b (f(x) - p(x))\,\mathrm{d}x = \int_a^b \frac{f^{(4)}(\xi(x))}{4!}(x-a)\left(x - \frac{a+b}{2}\right)(x-b)(x-d)\,\mathrm{d}x.\end{aligned}$$

Since $d \in (a,b)$ is arbitrary, we can take $d = \frac{a+b}{2}$. Since

$$(x-a)\left(x - \frac{a+b}{2}\right)^2(x-b) < 0$$

on (a,b) almost everywhere, as before we can apply the mean value theorem. Using the change of variable $x = \frac{a+b}{2} + \frac{b-a}{2}t$, we can compute

$$\begin{aligned}&\int_a^b (x-a)\left(x - \frac{a+b}{2}\right)^2(x-b)\,\mathrm{d}x\\ &= \left(\frac{b-a}{2}\right)^5 \int_{-1}^1 (t+1)t^2(t-1)\,\mathrm{d}t = -\frac{1}{120}(b-a)^5.\end{aligned}$$

Since $4! = 24$ and $24 \times 120 = 2880$, the integration error of (18.4) is

$$\left|\int_a^b f(x)\,\mathrm{d}x - \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)\right| \le \frac{\left\|f^{(4)}\right\|}{2880}(b-a)^5. \tag{18.5}$$

18.1.3 Compound rule

Newton-Cotes rule with $N \ge 4$ are almost never used because beyond some order N, some of the weights $\{w_n\}_{n=1}^N$ become negative, which introduces rounding errors. One way to avoid this problem is to divide the interval $[a,b]$ into N evenly spaced subintervals and apply the trapezoidal rule or the Simpson's rule to each subinterval. This method is known as the compound (or composite) rule.

If you apply the trapezoidal rule to N subintervals, then there are $N+1$ endpoints. Letting $x_n = n/N$ for $n = 0, 1, \ldots, N$, the formula for $[0,1]$ is

$$\begin{aligned}\int_0^1 f(x)\,\mathrm{d}x &\approx \sum_{n=1}^{N} \frac{1}{2N}(f(x_{n-1}) + f(x_n)) \\ &= \frac{1}{2N}\left(f(x_0) + 2f(x_1) + \cdots + 2f(x_{N-1}) + f(x_N)\right).\end{aligned}$$

(Just remember that the relative weights are 1 at endpoints and 2 in between.) Since $b - a = 1/N$ and there are N subintervals, the error of the $(N+1)$-point trapezoidal rule is of order $\frac{\|f''\|}{12}N^{-2}$.

If you apply Simpson's rule, then there are 3 points on each subinterval, of which there are N, and $N-1$ endpoints are counted twice. Therefore the total number of points is $3N - (N-1) = 2N+1$. Letting $x_n = n/(2N)$ for $n = 0, 1, \ldots, 2N$, the formula for $[0,1]$ is

$$\begin{aligned}\int_0^1 f(x)\,\mathrm{d}x &\approx \sum_{n=1}^{N} \frac{1}{6N}(f(x_{2n-2}) + 4f(x_{2n-1}) + f(x_{2n})) \\ &= \frac{1}{6N}\left(f(x_0) + 4f(x_1) + 2f(x_2) + \cdots + 4f(x_{2N-1}) + f(x_{2N})\right).\end{aligned}$$

(Just remember that the relative weights are 1 at endpoints, and they alternate like $4, 2, 4, 2, \ldots, 4, 2, 4$ in between.) Since $b - a = 1/N$ and there are N subintervals, the error of the $(2N+1)$-point Simpson's rule is of order $\frac{\|f^{(4)}\|}{2880}N^{-4}$.

Since the quadrature weights are given explicitly for trapezoidal and Simpson's rules, it is straightforward to write codes for computing numerical integrals. Tables 18.1 and 18.2 show the $\log_{10}$ relative errors of integrals over the interval $[0,1]$ ($\log_{10}\left|\widehat{I}/I - 1\right|$, where I is the true integral and $\widehat{I}$ is the numerical one) for several functions when we use the N-point compound trapezoidal and Simpson's rule. As the above error analysis suggests, errors tend to be smaller when the integrand is smoother (has higher order derivatives). Furthermore, Simpson's rule is more accurate than the trapezoidal rule.

Table 18.1. $\log_{10}$ relative errors of compound trapezoidal rule.

# points	$x^{1/2}$	$x^{3/2}$	$x^{5/2}$	$x^{7/2}$	$x^{9/2}$	e^x
3	-1.0238	-1.1743	-0.7343	-0.4896	-0.3041	-1.6830
5	-1.4550	-1.7558	-1.3394	-1.0875	-0.8937	-2.2838
9	-1.8926	-2.3438	-1.9427	-1.6885	-1.4928	-2.8855
17	-2.3346	-2.9361	-2.5452	-2.2902	-2.0941	-3.4874
33	-2.7795	-3.5314	-3.1474	-2.8922	-2.6960	-4.0895
65	-3.2264	-4.1287	-3.7495	-3.4943	-3.2980	-4.6915

18.2 Gaussian quadrature

In the Newton-Cotes quadrature, we assume that the nodes are evenly spaced, but of course there is no particular reason to do so. Can we do better by choosing

Table 18.2. $\log_{10}$ relative errors of compound Simpson's rule.

# points	$x^{1/2}$	$x^{3/2}$	$x^{5/2}$	$x^{7/2}$	$x^{9/2}$	e^x
3	-1.3676	-2.2275	-2.3780	-1.8192	-1.1040	-3.4722
5	-1.8179	-2.9667	-3.3705	-2.9823	-2.3199	-4.6667
9	-2.2691	-3.7142	-4.3841	-4.1584	-3.5289	-5.8684
17	-2.7206	-4.4649	-5.4112	-5.3435	-4.7350	-7.0720
33	-3.1722	-5.2168	-6.4470	-6.5346	-5.9399	-8.2759
65	-3.6237	-5.9692	-7.4884	-7.7297	-7.1443	-9.4800

the quadrature nodes optimally? In general, consider the integral

$$\int_a^b w(x)f(x)\,\mathrm{d}x, \tag{18.6}$$

where $-\infty \le a < b \le \infty$ are endpoints of integration, $w(x) > 0$ is some (fixed) weighting function, and f is a general integrand. A typical example is $a = -\infty$, $b = \infty$, and $w(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-(x-\mu)^2/2\sigma^2}$, in which case we want to compute the expectation $\mathrm{E}[f(X)]$ when the random variable X is normally distributed as $X \sim N(\mu, \sigma^2)$.

Let us prove some properties of the Gaussian quadrature. Assume that $\int_a^b w(x)x^n\,\mathrm{d}x$ exists for all $n \ge 0$, where $-\infty \le a < b \le \infty$ are fixed. For functions f, g, define the inner product (f, g) by

$$(f, g) = \int_a^b w(x)f(x)g(x)\,\mathrm{d}x. \tag{18.7}$$

As usual, define the norm of f by $\|f\| = \sqrt{(f, f)}$. For notational simplicity, let us omit a, b, so $\int$ means $\int_a^b$.

The first step is to construct orthogonal polynomials $\{p_n(x)\}_{n=0}^N$ corresponding to the inner product (18.7).

Definition 18.1 (Orthogonal polynomial)**.** The polynomials $\{p_n(x)\}_{n=0}^N$ are called *orthogonal* if (i) $\deg p_n = n$ and the leading coefficient of p_n is 1, and (ii) for all $m \ne n$, we have $(p_m, p_n) = 0$.

Some authors require that the polynomials are orthonormal, so $(p_n, p_n) = 1$. Here we normalize the polynomials by requiring that the leading coefficient is 1, which is useful for computation. The following three-term recurrence relation (TTRR) shows the existence of orthogonal polynomials and provides an explicit algorithm for computing them.

Proposition 18.2 (Three-term recurrence relation, TTRR)**.** *Let* $p_0(x) = 1$, $p_1(x) = x - \frac{(xp_0, p_0)}{\|p_0\|^2}$, *and for* $n \ge 1$ *define*

$$p_{n+1}(x) = \left(x - \frac{(xp_n, p_n)}{\|p_n\|^2}\right)p_n(x) - \frac{\|p_n\|^2}{\|p_{n-1}\|^2}p_{n-1}(x). \tag{18.8}$$

Then $p_n(x)$ *is the degree* n *orthogonal polynomial.*

Proof. Let us show by induction on n that (i) p_n is an degree n polynomial with leading coefficient 1, and (ii) $(p_n, p_m) = 0$ for all $m < n$. The claim is trivial for $n = 0$. For $n = 1$, by construction p_1 is a degree 1 polynomial with leading coefficient 1, and since $p_0(x) = 1$, we obtain

$$(p_1, p_0) = \left(\left(x - \frac{(xp_0, p_0)}{\|p_0\|^2} \right) p_0, p_0 \right) = (xp_0, p_0) - (xp_0, p_0) = 0.$$

Suppose the claim holds up to n. Then for $n + 1$, by (18.8) the leading coefficient of p_{n+1} is the same as that of xp_n, which is 1. If $m = n$, then

$$\begin{aligned}(p_{n+1}, p_n) &= \left(\left(x - \frac{(xp_n, p_n)}{\|p_n\|^2} \right) p_n - \frac{\|p_n\|^2}{\|p_{n-1}\|^2} p_{n-1}, p_n \right) \\ &= (xp_n, p_n) - (xp_n, p_n) - \frac{\|p_n\|^2}{\|p_{n-1}\|^2} (p_{n-1}, p_n) = 0.\end{aligned}$$

If $m = n - 1$, then

$$\begin{aligned}(p_{n+1}, p_{n-1}) &= \left(\left(x - \frac{(xp_n, p_n)}{\|p_n\|^2} \right) p_n - \frac{\|p_n\|^2}{\|p_{n-1}\|^2} p_{n-1}, p_{n-1} \right) \\ &= (xp_n, p_{n-1}) - \frac{(xp_n, p_n)}{\|p_n\|^2} (p_n, p_{n-1}) - \|p_n\|^2 \\ &= (p_n, xp_{n-1}) - \|p_n\|^2 .\end{aligned}$$

Since the leading coefficients of p_n, p_{n-1} are 1, we can write $xp_{n-1}(x) = p_n(x) + q(x)$, where $q(x)$ is a polynomial of degree at most $n - 1$. Clearly q can be expressed as a linear combination of $p_0, p_1, \ldots, p_{n-1}$, so $(p_n, q) = 0$. Therefore

$$(p_{n+1}, p_{n-1}) = (p_n, p_n + q) - \|p_n\|^2 = \|p_n\|^2 + (p_n, q) - \|p_n\|^2 = 0.$$

Finally, if $m < n - 1$, then

$$\begin{aligned}(p_{n+1}, p_m) &= \left(\left(x - \frac{(xp_n, p_n)}{\|p_n\|^2} \right) p_n - \frac{\|p_n\|^2}{\|p_{n-1}\|^2} p_{n-1}, p_m \right) \\ &= (xp_n, p_m) - \frac{(xp_n, p_n)}{\|p_n\|^2} (p_n, p_m) - \frac{\|p_n\|^2}{\|p_{n-1}\|^2} (p_{n-1}, p_m) \\ &= (p_n, xp_m) = 0\end{aligned}$$

because xp_m is a polynomial of degree $1 + m < n$. □

The following lemma shows that an degree n orthogonal polynomial has exactly n real roots (so they are all simple).

Lemma 18.3. *$p_n(x)$ has exactly n real roots on (a, b).*

Proof. By the fundamental theorem of algebra, $p_n(x)$ has exactly n roots in $\mathbb{C}$. Suppose to the contrary that $p_n(x)$ has fewer than n real roots on (a, b). Let $x_1, \ldots, x_k$ $(k < n)$ be those roots at which $p_n(x)$ changes its sign and

$q(x) = (x - x_1) \cdots (x - x_k)$. Since $p_n(x)q(x)$ has a constant sign but is not identically equal to zero, we have

$$(p_n, q) = \int w(x) p_n(x) q(x) \, \mathrm{d}x \neq 0$$

because $w(x) > 0$. On the other hand, since $\deg q = k < n$, we have $(p_n, q) = 0$, which is a contradiction. □

The following theorem shows that using the N roots of the degree N orthogonal polynomial $p_N(x)$ as quadrature nodes and choosing specific weights, we can integrate all polynomials of degree up to $2N - 1$ exactly. Thus Gaussian quadrature always exists.

Theorem 18.4 (Gaussian quadrature)**.** *Let $a < x_1 < \cdots < x_N < b$ be the N roots of the degree N orthogonal polynomial p_N and define*

$$w_n = \int w(x) L_n(x) \, \mathrm{d}x \tag{18.9}$$

for $n = 1, \ldots, N$, where

$$L_n(x) = \prod_{m \neq n} \frac{x - x_m}{x_n - x_m}$$

is the degree $N - 1$ polynomial that takes value 1 at x_n and 0 at x_m ($m \in \{1, \ldots, N\} \setminus n$). Then

$$\int w(x) p(x) \, \mathrm{d}x = \sum_{n=1}^{N} w_n p(x_n) \tag{18.10}$$

for all polynomials $p(x)$ of degree up to $2N - 1$.

Proof. Since $\deg p \leq 2N - 1$ and $\deg p_N = N$, we can write

$$p(x) = p_N(x) q(x) + r(x),$$

where $\deg q, \deg r \leq N - 1$. Since q can be expressed as a linear combination of orthogonal polynomials of degree up to $N - 1$, we have $(p_N, q) = 0$. Hence

$$\int w(x) p(x) \, \mathrm{d}x = (p_N, q) + \int w(x) r(x) \, \mathrm{d}x = \int w(x) r(x) \, \mathrm{d}x.$$

On the other hand, since $\{x_n\}_{n=1}^N$ are roots of p_N, we have

$$p(x_n) = p_N(x_n) q(x_n) + r(x_n) = r(x_n)$$

for all n, so in particular

$$\sum_{n=1}^{N} w_n p(x_n) = \sum_{n=1}^{N} w_n r(x_n).$$

Therefore it suffices to show (18.10) for polynomials r of degree up to $N - 1$. Since $\deg r \leq N - 1$ and $\deg L_n = N - 1$, by Proposition 17.1 we have

$$r(x) = \sum_{n=1}^{N} r(x_n) L_n(x)$$

identically. Since r can be represented as a linear combination of L_n's, it suffices to show (18.10) for all L_n's. But since by (18.9) we have

$$\int w(x)L_n(x)\,\mathrm{d}x = w_n = \sum_{m=1}^{N} w_m\delta_{mn} = \sum_{m=1}^{N} w_m L_n(x_m),$$

the claim is true. □

In practice, how can we compute the nodes $\{x_n\}_{n=1}^N$ and weights $\{w_n\}_{n=1}^N$ of the N-point Gaussian quadrature established in Theorem 18.4? The solution is given by the following Golub-Welsch algorithm.

Theorem 18.5 (Golub and Welsch, 1969)**.** *For each $n \geq 1$, define α_n, β_n by*

$$\alpha_n = \frac{(xp_{n-1}, p_{n-1})}{\|p_{n-1}\|^2}, \quad \beta_n = \frac{\|p_n\|}{\|p_{n-1}\|} > 0.$$

Define the $N \times N$ symmetric tridiagonal matrix

$$T_N = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & \cdots & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \ddots & \vdots \\ 0 & \beta_2 & \alpha_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \beta_{N-1} \\ 0 & \cdots & 0 & \beta_{N-1} & \alpha_N \end{bmatrix}. \tag{18.11}$$

Then the Gaussian quadrature nodes $\{x_n\}_{n=1}^N$ are eigenvalues of T_N. Letting $v_n = (v_{n1}, \ldots, v_{nn})'$ be an eigenvector of T_N corresponding to the eigenvalue x_n, the weights $\{w_n\}_{n=1}^N$ in (18.9) are equal to

$$w_n = \frac{v_{n1}^2}{\|v_n\|^2} \int w(x)\,\mathrm{d}x > 0. \tag{18.12}$$

Proof. By (18.8) and the definition of α_n, β_n, for all $n \geq 0$ we have

$$p_{n+1}(x) = (x - \alpha_{n+1})p_n(x) - \beta_n^2 p_{n-1}(x).$$

Note that this is true for $n = 0$ by defining $p_{-1}(x) = 0$ and $\beta_0 = 0$. For each n, let $p_n^*(x) = p_n(x)/\|p_n\|$ be the normalized orthogonal polynomial. Then the above equation becomes

$$\|p_{n+1}\|\, p_{n+1}^*(x) = \|p_n\|\,(x - \alpha_{n+1})p_n^*(x) - \|p_{n-1}\|\,\beta_n^2 p_{n-1}^*(x).$$

Dividing both sides by $\|p_n\| > 0$, using the definition of β_n, β_{n+1}, and rearranging terms, we obtain

$$\beta_n p_{n-1}^*(x) + \alpha_{n+1}p_n^*(x) + \beta_{n+1}p_{n+1}^*(x) = xp_n^*(x).$$

In particular, setting $x = x_k$ (where x_k is a root of p_N), we obtain

$$\beta_n p_{n-1}^*(x_k) + \alpha_{n+1}p_n^*(x_k) + \beta_{n+1}p_{n+1}^*(x_k) = x_k p_n^*(x_k).$$

for all n and $k = 1, \dots, N$. Since $\beta_0 = 0$ by definition and $p_N^*(x_k) = 0$ (since x_k is a root of p_N and hence $p_N^* = p_N / \|p_N\|$), letting $P(x) = (p_0^*(x), \dots, p_{N-1}^*(x))'$ and collecting the above equation into a vector, we obtain

$$T_N P(x_k) = x_k P(x_k)$$

for $k = 1, \dots, N$. Define the $N \times N$ matrix P by $P = (P(x_1), \dots, P(x_N))$. Then $T_N P = \operatorname{diag}(x_1, \dots, x_N) P$, so $x_1, \dots, x_N$ are eigenvalues of T_N provided that P is invertible. Now since $\{p_n^*\}_{n=0}^{N-1}$ are normalized and Gaussian quadrature integrates all polynomials of degree up to $2N - 1$ exactly, we have

$$\delta_{mn} = (p_m^*, p_n^*) = \int w(x) p_m^*(x) p_n^*(x) \, \mathrm{d}x = \sum_{k=1}^{N} w_k p_m^*(x_k) p_n^*(x_k)$$

for $m, n \le N - 1$. Letting $W = \operatorname{diag}(w_1, \dots, w_N)$, this equation becomes $PWP' = I$. Therefore P, W are invertible and $x_1, \dots, x_N$ are eigenvalues of T_N. Solving for W and taking the inverse, we obtain

$$W^{-1} = P'P \iff \frac{1}{w_n} = \sum_{k=0}^{N-1} p_k^*(x_n)^2 > 0$$

for all n. To show (18.12), let v_n be an eigenvector of T_N corresponding to the eigenvalue x_n. Then $v_n = cP(x_n)$ for some constant $c \neq 0$. Taking the norm, we obtain

$$\|v_n\|^2 = c^2 \|P(x_n)\|^2 = c^2 \sum_{k=0}^{N-1} p_k^*(x_n)^2 = \frac{c^2}{w_n} \iff w_n = \frac{c^2}{\|v_n\|^2}.$$

Comparing the first element of $v_n = cP(x_n)$, noting that $p_0(x) = 1$ and hence $p_0^* = p_0 / \|p_0\| = 1 / \|p_0\|$, we obtain

$$c^2 = v_{n1}^2 \|p_0\|^2 = v_{n1}^2 \int w(x) p_0(x)^2 \, \mathrm{d}x = v_{n1}^2 \int w(x) \, \mathrm{d}x,$$

which implies (18.12). □

Below are a few examples of the Gaussian quadrature. By doing a Google search, you can find subroutines in Matlab or whatever programming language that compute the nodes and weights of these quadratures.

Example 18.1. The case $(a, b) = (-1, 1)$, $w(x) = 1$ is known as the Gauss-Legendre quadrature.

Example 18.2. The case $(a, b) = (-1, 1)$, $w(x) = 1/\sqrt{1 - x^2}$ is known as the Gauss-Chebyshev quadrature. It is useful for computing Fourier coefficients (through the change of variable $x = \cos\theta$).

Example 18.3. The case $(a, b) = (-\infty, \infty)$, $w(x) = \mathrm{e}^{-x^2}$ is known as the Gauss-Hermite quadrature, which is useful for computing the expectation with respect to the normal distribution.

Example 18.4. The case $(a, b) = (0, \infty)$, $w(x) = \mathrm{e}^{-x}$ is known as the Gauss-Laguerre quadrature, which is useful for computing the expectation with respect to the exponential distribution.

Table 18.3 shows the $\log_{10}$ relative errors when using the N-point Gauss-Legendre quadrature. Comparing to Tables 18.1 and 18.2, we can see that Gaussian quadrature is overwhelmingly more accurate than Newton-Cotes.

Table 18.3. $\log_{10}$ relative errors of Gauss-Legendre.

# points	$x^{1/2}$	$x^{3/2}$	$x^{5/2}$	$x^{7/2}$	$x^{9/2}$	e^x
3	-2.4237	-3.3289	-3.8570	-4.0525	-3.8824	-6.3191
5	-3.0245	-4.3578	-5.3560	-6.0948	-6.6082	-12.4194
9	-3.7418	-5.5649	-7.0688	-8.3362	-9.4106	-15.9546
17	-4.5396	-6.8986	-8.9436	-10.7592	-12.3913	-15.9546
33	-5.3862	-8.3108	-10.9229	-13.3092	-15.3525	$-\infty$

If $(a, b) = (-\infty, \infty)$ and $\int_{-\infty}^{\infty} w(x)\,\mathrm{d}x = 1$ in (18.6), then $w(x)$ can be viewed as a probability density and (18.6) becomes an expectation. After a suitable transformation, the Gauss-Legendre, Gauss-Hermite, and Gauss-Laguerre quadratures can then be viewed as approximations to the uniform, normal, and exponential distributions. The same idea can be applied to a wider class of distributions. Since by Theorem 18.5 all we need for implementing the Gaussian quadrature are the polynomial moments $\int w(x)x^n\,\mathrm{d}x$ of the weighting functions w, Gaussian quadrature can be used for approximating any distribution that has explicit moments. Toda (2020) uses this idea to discretize nonparametric distributions from data.

18.3 Discretization

If the goal is to solve a single optimization problem that involves expectations (e.g., static optimal portfolio problem), a highly accurate Gaussian quadrature is a natural choice. However, many economic problems are dynamic, in which case one needs to compute conditional expectations. Furthermore, to reduce the computational complexity of the problem, it is desirable that the quadrature nodes are preassigned instead of being dependent on the particular state of the model. Discretization is a useful tool for solving such problems.

This section explains the Farmer and Toda (2017) method of discretizing Markov processes, which is based on the maximum entropy discretization method of distributions introduced in Tanaka and Toda (2013, 2015). Matlab codes are available at `https://github.com/alexisakira/discretization`.

18.3.1 Earlier methods

For concreteness, consider the Gaussian AR(1) process

$$x_t = \rho x_{t-1} + \varepsilon_t, \quad \epsilon_t \sim N(0, \sigma^2).$$

Then the conditional distribution of x_t given x_{t-1} is $N(\rho x_{t-1}, \sigma^2)$. How can we discretize (find a finite-state Markov chain approximation) of this stochastic process?

A classic method is Tauchen (1986) but it should not be used it because it is inaccurate (so we will not explain it further). Similarly, the quantile method in Adda and Cooper (2003) is poor, as documented in the accuracy comparison

in Farmer and Toda (2017). For Gaussian AR(1) processes, the Rouwenhorst (1995) method is good because the conditional moments are matched exactly up to order 2 and the method is constructive (does not involve optimization). It is especially useful when $\rho \geq 0.99$.

The Tauchen and Hussey (1991) method is based on the Gauss-Hermite quadrature (Example 18.3). First consider discretizing $N(0, \sigma^2)$. Letting $\{x_n\}_{n=1}^N$ and $\{w_n\}_{n=1}^N$ be the nodes and weights of the N-point Gauss-Hermite quadrature, since for any integrand g we have

$$\begin{aligned}
\mathrm{E}[g(X)] &= \int_{-\infty}^{\infty} g(x) \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{x^2}{2\sigma^2}} \,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} g(\sqrt{2}\sigma y) \frac{1}{\sqrt{\pi}} \mathrm{e}^{-y^2} \,\mathrm{d}y \\
&\approx \sum_{n=1}^{N} \frac{w_n}{\sqrt{\pi}} g(\sqrt{2}\sigma x_n),
\end{aligned}$$

we can use the nodes $x'_n = \sqrt{2}\sigma x_n$ and weights $w'_n = w_n/\sqrt{\pi}$ to discretize $N(0, \sigma^2)$.

The same idea can be used to discretize the Gaussian AR(1) process. Let us fix the nodes $\{x'_n\}_{n=1}^N$ as constructed above. Since for any integrand g, letting $\mu = \rho x'_m$ we have

$$\begin{aligned}
\mathrm{E}\left[g(x_t) \,|\, x_{t-1} = x'_m\right] &= \int_{-\infty}^{\infty} g(x) \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} g(x) \mathrm{e}^{-\frac{\mu^2 - 2x\mu}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{x^2}{2\sigma^2}} \,\mathrm{d}x \\
&\approx \sum_{n=1}^{N} w'_n \mathrm{e}^{-\frac{\mu^2 - 2x'_n\mu}{2\sigma^2}} g(x'_n),
\end{aligned}$$

so we can construct the transition probability matrix $P = (p_{mn})$ by

$$p_{mn} \propto w'_n \mathrm{e}^{-\frac{\mu^2 - 2x'_n\mu}{2\sigma^2}},$$

where $\mu = \rho x'_m$ and the constant of proportionality is determined such that $\sum_{n=1}^N p_{mn} = 1$. The Tauchen-Hussey method is relatively accurate if $\rho \leq 0.5$, although a drawback is that it assumes Gaussian shocks. Furthermore, the performance deteriorates quickly when ρ becomes larger.

18.3.2 Farmer-Tanaka-Toda maximum entropy method

Several papers by me and my coauthors (Tanaka and Toda, 2013, 2015; Farmer and Toda, 2017) provide a more accurate and generally applicable discretization method (so it should be the first choice!). Below I briefly explain the method, but see Farmer and Toda (2017) for more details.

Discretizing probability distributions

Suppose that we are given a continuous probability density function $f : \mathbb{R}^K \to \mathbb{R}$, which we want to discretize. Let X be a random vector with density f, and

$g : \mathbb{R}^K \to \mathbb{R}$ be any bounded continuous function. The first step is to pick a quadrature formula

$$\mathrm{E}[g(X)] = \int_{\mathbb{R}^K} g(x) f(x)\,\mathrm{d}x \approx \sum_{n=1}^{N} w_n g(x_n) f(x_n), \tag{18.13}$$

where N is the number of integration points, $\{x_n\}_{n=1}^N$, and $w_n > 0$ is the weight on the integration point x_n.

For now, we do not take a stance on the choice of the initial quadrature formula, but take it as given. Given the quadrature formula (18.13), a coarse but valid discrete approximation of the density f would be to assign probability q_n to the point x_n proportional to $w_n f(x_n)$, so

$$q_n = \frac{w_n f(x_n)}{\sum_{n=1}^{N} w_n f(x_n)}. \tag{18.14}$$

However, this is not necessarily a good approximation because the moments of the discrete distribution $\{q_n\}$ do not generally match those of f.

Tanaka and Toda (2015) propose exactly matching a finite set of moments by updating the probabilities $\{q_n\}$ in a particular way. Let $T : \mathbb{R}^K \to \mathbb{R}^L$ be a function that defines the moments that we wish to match and let $\bar{T} = \int_{\mathbb{R}^K} T(x) f(x)\,\mathrm{d}x$ be the vector of exact moments. For example, if we want to match the first and second moments in the one dimensional case ($K = 1$), then $T(x) = (x, x^2)'$. Tanaka and Toda (2015) update the probabilities $\{q_n\}$ by solving the optimization problem

$$\begin{aligned} \underset{\{p_n\}}{\text{minimize}} \quad & \sum_{n=1}^{N} p_n \log \frac{p_n}{q_n} \\ \text{subject to} \quad & \sum_{n=1}^{N} p_n T(x_n) = \bar{T}, \ \sum_{n=1}^{N} p_n = 1, \ p_n \geq 0. \end{aligned} \tag{P}$$

The objective function in the primal problem (P) is the Kullback and Leibler (1951) information of $\{p_n\}$ relative to $\{q_n\}$, which is also known as the relative entropy. This method matches the given moments exactly while keeping the probabilities $\{p_n\}$ as close to the initial approximation $\{q_n\}$ as possible in the sense of the Kullback-Leibler information. Note that since (P) is a convex minimization problem, the solution (if one exists) is unique.

The optimization problem (P) is a constrained minimization problem with a large number (N) of unknowns ($\{p_n\}$) with $L + 1$ equality constraints and N inequality constraints, which is in general computationally intensive to solve. However, it is well-known that entropy-like minimization problems are computationally tractable by using duality theory (Borwein and Lewis, 1991). Tanaka and Toda (2015) convert the primal problem (P) to the dual problem

$$\max_{\lambda \in \mathbb{R}^L} \left[\lambda' \bar{T} - \log \left(\sum_{n=1}^{N} q_n \mathrm{e}^{\lambda' T(x_n)} \right) \right], \tag{D}$$

which is a *low dimensional* (L unknowns) *unconstrained* concave maximization problem and hence computationally tractable. The following theorem shows how the solutions to the two problems (P) and (D) are related. Below, the symbols "int" and "co" denote the interior and the convex hull of sets.

Theorem 18.6. *1. The primal problem (P) has a solution if and only if $\bar{T} \in \operatorname{co} T(D_N)$. If a solution exists, it is unique.*

2. The dual problem (D) has a solution if and only if $\bar{T} \in \operatorname{int} \operatorname{co} T(D_N)$. If a solution exists, it is unique.

3. If the dual problem (D) has a (unique) solution λ_N, then the (unique) solution to the primal problem (P) is given by

$$p_n = \frac{q_n \mathrm{e}^{\lambda_N' T(x_n)}}{\sum_{n=1}^N q_n \mathrm{e}^{\lambda_N' T(x_n)}} = \frac{q_n \mathrm{e}^{\lambda_N' (T(x_n) - \bar{T})}}{\sum_{n=1}^N q_n \mathrm{e}^{\lambda_N' (T(x_n) - \bar{T})}}. \tag{18.15}$$

Theorem 18.6 provides a practical way to implement the Tanaka-Toda method. After choosing the initial discretization $Q = \{q_n\}$ and the moment defining function T, one can numerically solve the unconstrained optimization problem (D). To this end, we can instead solve

$$\min_{\lambda \in \mathbb{R}^L} \sum_{n=1}^N q_n \mathrm{e}^{\lambda' (T(x_n) - \bar{T})} \tag{D$'$}$$

because the objective function in (D$'$) is a monotonic transformation (-1 times the exponential) of that in (D). Since (D$'$) is an unconstrained convex minimization problem with a (relatively) small number (L) of unknowns (λ), solving it is computationally simple. Letting $J_N(\lambda)$ be the objective function in (D$'$), its gradient and Hessian can be analytically computed as

$$\nabla J_N(\lambda) = \sum_{n=1}^N q_n \mathrm{e}^{\lambda' (T(x_n) - \bar{T})} (T(x_n) - \bar{T}), \tag{18.16a}$$

$$\nabla^2 J_N(\lambda) = \sum_{n=1}^N q_n \mathrm{e}^{\lambda' (T(x_n) - \bar{T})} (T(x_n) - \bar{T})(T(x_n) - \bar{T})', \tag{18.16b}$$

respectively. In practice, we can quickly solve (D$'$) numerically using optimization routines by supplying the analytical gradient and Hessian.[1]

If a solution to (D$'$) exists, it is unique, and we can compute the updated discretization $P = \{p_n\}$ by (18.15). If a solution does not exist, it means that the regularity condition $\bar{T} \in \operatorname{int} \operatorname{co} T(D_N)$ does not hold and we cannot match moments. Then one needs to select a smaller set of moments. Numerically checking whether moments are matched is straightforward: by (18.15), (D$'$), and (18.16a), the error is

$$\sum_{n=1}^N p_n T(x_n) - \bar{T} = \frac{\sum_{n=1}^N q_n \mathrm{e}^{\lambda_N' (T(x_n) - \bar{T})} (T(x_n) - \bar{T})}{\sum_{n=1}^N q_n \mathrm{e}^{\lambda_N' (T(x_n) - \bar{T})}} = \frac{\nabla J_N(\lambda_N)}{J_N(\lambda_N)}. \tag{18.17}$$

[1] Since the dual problem (D) is a concave maximization problem, one may also solve it directly. However, according to our experience, solving (D$'$) is numerically more stable. This is because the objective function in (D) is close to linear when $\|\lambda\|$ is large, so the Hessian is close to singular and not well-behaved. On the other hand, since the objective function in (D$'$) is the sum of exponential functions, it is well-behaved.

Discretizing general Markov processes

Next we show how to extend the Tanaka-Toda method to the case of time-homogeneous Markov processes.

Consider the time-homogeneous first-order Markov process

$$P(x_t \le x' | x_{t-1} = x) = F(x', x),$$

where x_t is the vector of state variables and $F(\cdot, x)$ is a cumulative distribution function (CDF) that determines the distribution of $x_t = x'$ given $x_{t-1} = x$. The dynamics of any Markov process are completely characterized by its Markov transition kernel. In the case of a discrete state space, this transition kernel is simply a matrix of transition probabilities, where each row corresponds to a conditional distribution. We can discretize the continuous process x by applying the Tanaka-Toda method to each conditional distribution separately.

More concretely, suppose that we have a set of grid points $D_N = \{x_n\}_{n=1}^N$ and an initial coarse approximation $Q = (q_{nn'})$, which is an $N \times N$ probability transition matrix. Suppose we want to match some conditional moments of x, represented by the moment defining function $T(x)$. The exact conditional moments when the current state is $x_{t-1} = x_n$ are

$$\bar{T}_n = \mathrm{E}\left[T(x_t) \,|\, x_n\right] = \int T(x)\, \mathrm{d}F(x, x_n),$$

where the integral is over x, fixing x_n. (If these moments do not have explicit expressions, we can use highly accurate quadrature formulas to compute them.) By Theorem 18.6, we can match these moments exactly by solving the optimization problem

$$\begin{aligned} \underset{\{p_{nn'}\}_{n'=1}^N}{\text{minimize}} \quad & \sum_{n'=1}^N p_{nn'} \log \frac{p_{nn'}}{q_{nn'}} \\ \text{subject to} \quad & \sum_{n'=1}^N p_{nn'} T(x_{n'}) = \bar{T}_n, \;\; \sum_{n'=1}^N p_{nn'} = 1, \;\; p_{nn'} \ge 0 \end{aligned} \tag{P_n}$$

for each $n = 1, 2, \ldots, N$, or equivalently the dual problem

$$\min_{\lambda \in \mathbb{R}^L} \sum_{n'=1}^N q_{nn'} \mathrm{e}^{\lambda'(T(x_{n'}) - \bar{T}_n)}. \tag{D'_n}$$

(D'_n) has a unique solution if and only if the regularity condition

$$\bar{T}_n \in \operatorname{int} \operatorname{co} T(D_N) \tag{18.18}$$

holds. We summarize our procedure in Algorithm 1 below.

Algorithm 1 (Discretization of Markov processes)**.**

1. Select a discrete set of points $D_N = \{x_n\}_{n=1}^N$ and an initial approximation $Q = (q_{nn'})$.
2. Select a moment defining function $T(x)$ and corresponding exact conditional moments $\{\bar{T}_n\}_{n=1}^N$. If necessary, approximate the exact conditional moments with a highly accurate numerical integral.

3. For each $n = 1, \ldots, N$, solve minimization problem (D'_n) for λ_n. Check whether moments are matched using formula (18.17), and if not, select a smaller set of moments. Compute the conditional probabilities corresponding to row n of $P = (p_{nn'})$ using (18.15).

The resulting discretization of the process is given by the transition probability matrix $P = (p_{nn'})$. Since the dual problem (D'_n) is an unconstrained convex minimization problem with a typically small number of variables, standard Newton type algorithms can be applied. Furthermore, since the probabilities (18.15) are strictly positive by construction, the transition probability matrix $P = (p_{nn'})$ is a strictly positive matrix, so the resulting Markov chain is stationary and ergodic.

Bibliography

Jérôme Adda and Russel W. Cooper. *Dynamic Economics: Quantitative Methods and Applications*. MIT Press, Cambridge, MA, 2003.

R. B. Bapat and T. E. S. Raghavan. *Nonnegative Matrices and Applications*. Number 64 in Encyclopedia of Mathematics and Its Applications. Cambridge University Press, 1997.

Claude Berge. *Espaces Topologiques: Fonctions Multivoques*. Dunod, Paris, 1959. English translation: Translated by E. M. Patterson. *Topological Spaces*, New York: MacMillan, 1963. Reprinted: Mineola, NY: Dover, 1997.

Abraham Berman and Robert J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Number 9 in Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1994. doi:10.1137/1.9781611971262.

David Blackwell. Discounted dynamic programming. *Annals of Mathematical Statistics*, 36(1):226–235, February 1965. doi:10.1214/aoms/1177700285.

Jonathan M. Borwein and Adrian S. Lewis. Duality relationships for entropy-like minimization problems. *SIAM Journal on Control and Optimization*, 29 (2):325–338, March 1991. doi:10.1137/0329017.

Philip J. Davis and Philip Rabinowitz. *Methods of Numerical Integration*. Academic Press, Orlando, FL, second edition, 1984.

Leland E. Farmer and Alexis Akira Toda. Discretizing nonlinear, non-Gaussian Markov processes with exact conditional moments. *Quantitative Economics*, 8(2):651–683, July 2017. doi:10.3982/QE737.

Gene H. Golub and John H. Welsch. Calculation of Gauss quadrature rules. *Mathematics of Computation*, 23(106):221–230, may 1969. doi:10.1090/S0025-5718-69-99647-1.

F. J. Gould and Jon W. Tolle. A necessary and sufficient qualification for constrained optimization. *SIAM Journal of Applied Mathematics*, 20(2):164–172, March 1971. doi:10.1137/0120021.

Gary Harris and Clyde Martin. The roots of a polynomial vary continuously as a function of the coefficients. *Proceedings of the American Mathematical Society*, 100(2):390–392, June 1987. doi:10.2307/2045978.

J. Michael Harrison and David M. Kreps. Martingales and arbitrage in multiperiod securities market. *Journal of Economic Theory*, 20(3):381–408, June 1979. doi:10.1016/0022-0531(79)90043-7.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, second edition, 2013.

Kenneth L. Judd. Projection methods for solving aggregate growth models. *Journal of Economic Theory*, 58(2):410–452, December 1992. doi:10.1016/0022-0531(92)90061-L.

Takashi Kamihigashi. A simple proof of the necessity of the transversality condition. *Economic Theory*, 20(2):427–433, September 2002. doi:10.1007/s001990100198.

Takashi Kamihigashi. Elementary results on solutions to the Bellman equation of dynamic programming: Existence, uniqueness, and convergence. *Economic Theory*, 56(2):251–273, 2014. doi:10.1007/s00199-013-0789-4.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, March 1951. doi:10.1214/aoms/1177729694.

Peter D. Lax. *Linear Algebra and Its Applications*. John Wiley & Sons, Hoboken, NJ, second edition, 2007.

David G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, New York, 1969.

Harry Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, March 1952. doi:10.1111/j.1540-6261.1952.tb01525.x.

Walter Pohl, Karl Schmedders, and Ole Wilms. Higher-order effects in asset pricing models with long-run risks. *Journal of Finance*, 73(3):1061–1111, June 2018. doi:10.1111/jofi.12615.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

K. Geert Rouwenhorst. Asset pricing implications of equilibrium business cycle models. In Thomas F. Cooley, editor, *Frontiers of Business Cycle Research*, chapter 10, pages 294–330. Princeton University Press, 1995.

William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3):425–442, September 1964. doi:10.1111/j.1540-6261.1964.tb02865.x.

Ken'ichiro Tanaka and Alexis Akira Toda. Discrete approximations of continuous distributions by maximum entropy. *Economics Letters*, 118(3):445–450, March 2013. doi:10.1016/j.econlet.2012.12.020.

Ken'ichiro Tanaka and Alexis Akira Toda. Discretizing distributions with exact moments: Error estimate and convergence analysis. *SIAM Journal on Numerical Analysis*, 53(5):2158–2177, 2015. doi:10.1137/140971269.

George Tauchen. Finite state Markov-chain approximations to univariate and vector autoregressions. *Economics Letters*, 20(2):177–181, 1986. doi:10.1016/0165-1765(86)90168-0.

George Tauchen and Robert Hussey. Quadrature-based methods for obtaining approximate solutions to nonlinear asset pricing models. *Econometrica*, 59 (2):371–396, March 1991. doi:10.2307/2938261.

Alexis Akira Toda. Operator reverse monotonicity of the inverse. *American Mathematical Monthly*, 118(1):82–83, January 2011. doi:10.4169/amer.math.monthly.118.01.082.

Alexis Akira Toda. Incomplete market dynamics and cross-sectional distributions. *Journal of Economic Theory*, 154:310–348, November 2014. doi:10.1016/j.jet.2014.09.015.

Alexis Akira Toda. Wealth distribution with random discount factors. *Journal of Monetary Economics*, 104:101–113, June 2019. doi:10.1016/j.jmoneco.2018.09.006.

Alexis Akira Toda. Data-based automatic discretization of nonparametric distributions. *Computational Economics*, 2020. doi:10.1007/s10614-020-10012-6.