Econ 8000: Mathematics for Economists

Maria Titova, Vanderbilt University

Fall 2026

Contents

Notation

Sets and logic

Notation	Meaning
$\mathbb{N} = \{1, 2, 3, \dots\}$	Natural numbers.
$\mathbb{Z}$	Integers.
$\mathbb{Q}$	Rational numbers.
$\mathbb{R}$	Real numbers.
$x \in A$, $x \notin A$	Membership and nonmembership.
$\emptyset$	Empty set.
$\{x \in X : P(x)\}$	Elements of X satisfying P.
2^S	Power set of S.
$A \subseteq B$	A is a subset of B, possibly $A = B$.
$A \subset B$	A is a proper subset of B.
$A \cup B$, $\bigcup_{i \in I} A_i$	Union and indexed union.
$A \cap B$, $\bigcap_{i \in I} A_i$	Intersection and indexed intersection.
$A \setminus B$	Set difference.
A^c	Complement of A relative to the stated universe.
$A \times B$	Cartesian product.
A^n	n-fold Cartesian product of A.
$\neg P$	Negation of P.
$P \wedge Q$	Conjunction: P and Q.
$P \vee Q$	Disjunction: P or Q.
$P \Rightarrow Q$	Implication.
$P \Leftrightarrow Q$	Logical equivalence.
$\forall$	Universal quantifier: for every.
$\exists$	Existential quantifier: there exists.

Functions and sequences

Notation	Meaning
$f : X \to Y$	Function with domain X and codomain Y.
$f(A)$	Image of $A \subseteq X$ under f.
$f(X)$	Range of f.
$f^{-1}(B)$	Preimage of $B \subseteq Y$.
$f^{-1} : Y \to X$	Inverse of a bijective function $f : X \to Y$.
$g \circ f$	Composition, with f applied first.
$(x_k)_{k=1}^{\infty}$, (x_k)	Sequence.
$(x_{k_j})_{j=1}^{\infty}$	Subsequence of (x_k).
$x_k \to x$	Convergence of (x_k) to x.

Functions and sequences (continued)

Notation	Meaning
$\lim_{x \to a} f(x)$	Limit of $f(x)$ as x approaches a.
$x \downarrow a$, $x \uparrow a$	Approach to a from the right and left, respectively.
$\sup A$, $\inf A$	Supremum and infimum of $A \subseteq \mathbb{R}$.
$\max A$, $\min A$	Maximum and minimum of A, when attained.

Euclidean space and order

Notation	Meaning
$\mathbb{R}^n$	Real column vectors of dimension n.
$x = (x_1, \dots, x_n)^\top$	Column vector and its coordinates.
e_i	ith standard basis vector.
$x \cdot y = \sum_{i=1}^{n} x_i y_i$	Euclidean inner product.
$\lvert a \rvert$	Absolute value of a scalar.
$\lVert x \rVert$	Euclidean norm of a vector.
$d(x, y) = \lVert x - y \rVert$	Euclidean distance.
$B(x, r)$	Open ball with center x and radius r.
$\operatorname{int} D$	Interior of D.
$x \geq y$	$x_i \geq y_i$ for every i.
$x > y$	$x \geq y$ and $x \neq y$.
$x \gg y$	$x_i > y_i$ for every i.
$\mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x \geq 0\}$	Nonnegative orthant.
$\mathbb{R}^n_{++} = \{x \in \mathbb{R}^n : x \gg 0\}$	Positive orthant.

Linear algebra

Notation	Meaning
$A = (a_{ij}) \in \mathbb{R}^{m \times n}$	Real $m \times n$ matrix; i indexes rows and j columns.
$A^\top$, $x^\top$	Matrix transpose; transpose of a column vector.
$x^\top y = x \cdot y$	Matrix form of the inner product.
I_n	$n \times n$ identity matrix; write I when the dimension is clear.
$\operatorname{diag}(d_1, \dots, d_n)$	Diagonal matrix with the displayed diagonal entries.
AB	Matrix product; $(AB)_{ij} = \sum_k a_{ik} b_{kj}$.
$\operatorname{span}\{v_1, \dots, v_k\}$	Span of the displayed vectors.
$\dim V$	Dimension of a vector space V.
$\operatorname{Row}(A)$	Row space of A.
$\operatorname{Col}(A)$	Column space of A.
$\operatorname{rank} A$	Rank of A.
$N(A) = \{x : Ax = 0\}$	Null space, or kernel, of A.
A^{-1}	Matrix inverse.
$\det A$	Determinant of a square matrix.
$\operatorname{tr} A$	Trace of a square matrix.
$Av = \lambda v$, $v \neq 0$	Eigenvalue λ and associated eigenvector v.

Calculus and curvature

Notation	Meaning
x_0, $h = x - x_0$	Base point and increment.
$f'(x_0)$	Ordinary derivative at x_0.

Calculus and curvature (continued)

Notation	Meaning
$\dfrac{\partial f}{\partial x_i}(x_0)$	ith partial derivative at x_0.
$\nabla f(x_0)$	Gradient of a scalar-valued function; an $n \times 1$ column.
$\mathrm{d}f_{x_0}(h)$	Differential of f at x_0, applied to h.
$\mathrm{D}F(x_0)$	Jacobian of $F : \mathbb{R}^n \to \mathbb{R}^m$; an $m \times n$ matrix.
$\mathrm{D}f(x_0) = \nabla f(x_0)^\top$	Jacobian–gradient relation for scalar-valued f.
$\mathrm{D}_x F$, $\mathrm{D}_\theta F$	Partial Jacobians with respect to x and θ.
C^1, C^2	Once and twice continuously differentiable.
$\partial_v f(x_0)$	Directional derivative in v; v need not be normalized.
$H_f(x) = \mathrm{D}(\nabla f)(x)$	Hessian of a scalar-valued function.
$Q_A(x) = x^\top A x$	Quadratic form associated with A.
A_I	Principal submatrix using rows and columns indexed by I.
$\Delta_k = \det A_{\{1,\dots,k\}}$	kth leading principal minor.
$(1-t)x + ty$, $t \in [0,1]$	Convex combination of x and y.
$S_\alpha(f) = \{x : f(x) \geq \alpha\}$	Superlevel set of f at level α.
$H(a,c) = \{x : a \cdot x = c\}$	Hyperplane with normal a at level c.

Optimization and constraints

Notation	Meaning
$f(x;\theta)$	Objective at choice x and parameter θ.
x, θ	Choice or endogenous variable; parameter.
$D(\theta)$	Feasible set at parameter θ.
$\max_{x \in D} f(x)$	Maximum, used when attainment is established.
$\sup_{x \in D} f(x)$	Supremum, without presuming attainment.
$\arg\max_{x \in D} f(x)$	Set of maximizers; arg min denotes the set of minimizers.
x^*	Candidate or optimizer, as context indicates.
$X^*(\theta)$	Solution set $\arg\max_{x \in D(\theta)} f(x;\theta)$.
$V(\theta)$	Optimized value, when a maximum is attained.
$h(x) = 0$	Equality constraints.
$g(x) \geq 0$	Inequality constraints in a maximization problem, componentwise.
μ	Unrestricted equality-constraint multipliers.
$\lambda \geq 0$	Inequality-constraint multipliers.
$L(x,\mu,\lambda;\theta)$	Lagrangian $f + \mu \cdot h + \lambda \cdot g$.
$I(x) = \{i : g_i(x) = 0\}$	Active set of inequality constraints.
$\lambda_i g_i(x) = 0$	Complementary slackness.
LICQ	Linear independence constraint qualification.
KKT	Karush–Kuhn–Tucker conditions.
$T = N(\mathrm{D}h(x^*))$	Tangent space at a regular equality-constrained point.

Fixed points and correspondences

Notation	Meaning
$f(x^*) = x^*$	Fixed point of a function $f : X \to X$.
$\Gamma : X \rightrightarrows Y$	Correspondence, or set-valued map.
$\Gamma(x) \subseteq Y$	Value of the correspondence at x.
$x^* \in \Gamma(x^*)$	Fixed point of a correspondence $\Gamma : X \rightrightarrows X$.
$\beta \in [0,1)$	Contraction modulus.

Dynamic programming

Notation	**Meaning**
$t = 0, 1, 2, \ldots$	Time index.
k_t	Capital, or state, at date t.
c_t	Consumption at date t.
k_{t+1}, k'	Next-period capital; the recursive choice.
u	One-period utility.
f	Production function.
$\beta \in (0, 1)$	Discount factor.
$V(k)$	Value function at state k.
$g(k)$	Policy function at state k.
$V_t(k)$	Finite-horizon value at date t.
$s \in S$	State in a general stationary problem.
$a \in \Gamma(s)$	Feasible action at state s.
$r(s, a)$	Current reward.
$s' = F(s, a)$	State transition.
$B(S)$	Bounded real-valued functions on S.
$\|v\|_\infty = \sup_{s \in S} \|v(s)\|$	Sup norm on $B(S)$.
$(\mathcal{T}v)(s)$	$\sup_{a \in \Gamma(s)} \{r(s, a) + \beta v(F(s, a))\}$; Bellman operator.
$\mathcal{T}V = V$	Bellman equation as a fixed-point equation.

Lecture 1

Sets, Functions, Logic, and Proofs

1.1 Sets

1.1.1 Definition and Operations

A *set* is a collection of objects which we call *elements*. We typically denote sets by capital letters and elements by lowercase letters. If x is an element of a set A, we write $x \in A$. If x is not an element of A, we write $x \notin A$. For example, if $A = \{1, 2, 3\}$, then $1 \in A$ and $4 \notin A$.

There are multiple ways to describe a set. One is to list its elements, as in $A = \{1, 2, 3\}$. Another is to describe a property that characterizes its elements, as in $B = \{x \in \mathbb{R} : x > 0\}$, which is the set of positive real numbers.

Some examples of sets include:

- The empty set $\emptyset$.
- The set of integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$.
- The set of real numbers $\mathbb{R}$.
- The set of natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$.
- The set of rational numbers $\mathbb{Q} = \{p/q : p \in \mathbb{Z}, q \in \mathbb{Z} \setminus \{0\}\}$.
- The set of all subsets of a finite set S, called the power set and denoted 2^S.

Definition 1.1.1 (Inclusion and equality)**.** For sets A and B, write $A \subseteq B$ (A is a *subset* of B) if every element of A belongs to B. The sets are equal if $A \subseteq B$ and $B \subseteq A$. Write $A \subset B$ (A is a *proper subset* of B) if $A \subseteq B$ and $A \neq B$.

For example, the statements

$$\{1, 2\} \subseteq \{1, 2, 3\}, \qquad \{1, 2, 3\} = \{3, 2, 1\}, \qquad \mathbb{Z} \subset \mathbb{R}, \qquad \mathbb{N} \subset \mathbb{Z}$$

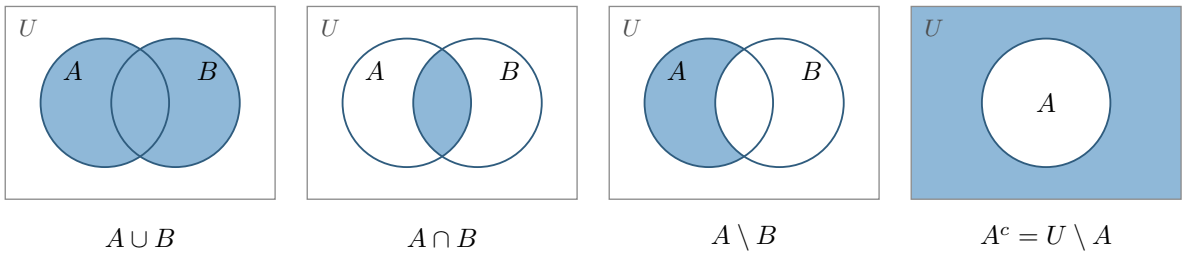
are all true. Furthermore, for any set S, we have $\emptyset \subseteq S$ and $S \subseteq S$.

Definition 1.1.2 (Set operations)**.** For sets A and B, the *union*, *intersection*, and *set difference* are

$$\begin{aligned} A \cup B &= \{x : x \in A \text{ or } x \in B\}, \\ A \cap B &= \{x : x \in A \text{ and } x \in B\}, \\ A \setminus B &= \{x \in A : x \notin B\}. \end{aligned}$$

The *complement* of a set A relative to a universal set U is $A^c = \{x : x \in U \text{ and } x \notin A\} = U \setminus A$.

These operations are called union, intersection, and set difference. They are typically illustrated by Venn diagrams:



$A \cup B$ $A \cap B$ $A \setminus B$ $A^c = U \setminus A$

Figure 1.1.1: Union, intersection, set difference, and complement.

Of course, we can also define the union and intersection of more than two sets. Given a collection of sets $\{A_i\}_{i \in I}$ indexed by a set I, we define

$$\bigcup_{i \in I} A_i = \{x : x \in A_i \text{ for some } i \in I\}, \qquad \bigcap_{i \in I} A_i = \{x : x \in A_i \text{ for every } i \in I\}.$$

The following theorem summarizes some properties of aforementioned set operations. Its proof is left as an exercise.

Theorem 1.1.3 (Set algebra)**.** *For sets A, B, and C,*

1. $A \cup B = B \cup A$, $A \cap B = B \cap A$ *(commutativity),*
2. $A \cup (B \cup C) = (A \cup B) \cup C$, $A \cap (B \cap C) = (A \cap B) \cap C$ *(associativity),*
3. $A \cup (A \cap B) = A$, $A \cap (A \cup B) = A$ *(absorption),*
4. $A \setminus B = A \cap B^c$ *(set difference),*
5. $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$ *(De Morgan's laws).*

We say that sets are *disjoint* if their intersection is empty. For example, the sets of even and odd integers are disjoint. A *partition* of a set S is a collection of nonempty, pairwise disjoint subsets of S whose union is S. For example, the sets of even and odd integers form a partition of $\mathbb{Z}$.

1.1.2 Cartesian Product and $\mathbb{R}^n$

Definition 1.1.4. A *Cartesian product* of two sets A and B is the set

$$A \times B = \{(a, b) : a \in A, b \in B\}.$$

For example, if $A = \{1, 2\}$ and $B = \{x, y\}$, then

$$A \times B = \{(1, x), (1, y), (2, x), (2, y)\}.$$

The product $A_1 \times \cdots \times A_n$ consists of tuples $(a_1, \ldots, a_n)$ satisfying $a_i \in A_i$ for every i. An n-fold product of a set A with itself is denoted $A^n = A \times \cdots \times A$.

The most commonly used in economics example of an n-fold product is $\mathbb{R}^n$. In this course, we treat elements of $\mathbb{R}^n$ as column vectors:

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n,$$

but that convention is not important until we discuss linear algebra.

Sometimes we need to compare vectors in $\mathbb{R}^n$. While not all vectors are comparable, we can compare them coordinatewise.

Definition 1.1.5 (Coordinatewise comparisons)**.** For $x, y \in \mathbb{R}^n$, write

$$\begin{aligned} x \geq y &\text{ if } x_i \geq y_i \text{ for every } i, \\ x > y &\text{ if } x \geq y \text{ and } x \neq y, \\ x \gg y &\text{ if } x_i > y_i \text{ for every } i. \end{aligned}$$

The nonnegative and positive orthants are

$$\mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x \geq 0\}, \qquad \mathbb{R}^n_{++} = \{x \in \mathbb{R}^n : x \gg 0\}.$$

1.2 Functions

1.2.1 Vocabulary and properties

Definition 1.2.1 (Function)**.** Let X and Y be two sets. A *function* $f : X \to Y$ is a rule that assigns to each $x \in X$ exactly one element $f(x) \in Y$. We also call f a *map*, or a *mapping*, from X to Y.

For a function $f : X \to Y$, we call X the *domain* and Y the *codomain*. For a subset $A \subseteq X$ of the domain, the *image* of A is the set

$$f(A) = \{f(x) : x \in A\} \subseteq Y$$

of values taken by f on A. In particular, we call $f(X)$ the *range* of f. The *preimage* of a subset $B \subseteq Y$ of the codomain is the set

$$f^{-1}(B) = \{x \in X : f(x) \in B\} \subseteq X$$

of points in the domain that map into B. The *graph* of f is the set of pairs

$$\{(x, f(x)) : x \in X\} \subseteq X \times Y.$$

Definition 1.2.2 (Injective, surjective, bijective)**.** A function $f : X \to Y$ is

1. *injective*, or *one-to-one*, if $f(x) = f(x')$ implies $x = x'$ for all $x, x' \in X$;
2. *surjective*, or *onto*, if its range is all of Y, that is, $f(X) = Y$;

3. *bijective* if it is both injective and surjective.

Injectivity says that distinct points of the domain have distinct values, and surjectivity says that every $y \in Y$ equals $f(x)$ for at least one $x \in X$. The two conditions are checked in opposite directions: injectivity starts from two points of X and compares their values, while surjectivity starts from a point of Y and asks for a point of X that reaches it.

Note that whether a function is surjective depends on the declared domain and codomain. For example, the function $f(x) = x^2$ is not surjective if $f : \mathbb{R} \to \mathbb{R}$ but surjective if $f : \mathbb{R} \to \mathbb{R}_+$. On the other hand, the function $g(x) = x^3$ is surjective if $g : \mathbb{R} \to \mathbb{R}$ or $g : \mathbb{R}_+ \to \mathbb{R}_+$.

Definition 1.2.3 (Composition)**.** Let $f : X \to Y$ and $g : Y \to Z$, so that the codomain of f is the domain of g. The *composition* of f and g is the function $g \circ f : X \to Z$ defined by

$$(g \circ f)(x) = g(f(x)) \quad \text{for every } x \in X.$$

1.2.2 Inverse Function

Let $f : X \to Y$ be a function and let $y \in Y$. Recalling the definition of a preimage, let us consider the set $f^{-1}(\{y\})$, which contains all elements of X that map to the point y. In general, this set may be empty, contain one point, or contain multiple points. By Definition 1.2.2, if f is surjective, then $f^{-1}(\{y\})$ contains at least one point for every $y \in Y$, and if f is injective, then it contains at most one point. Consequently, if f is bijective, then it contains exactly one point. This allows us to slightly abuse notation and define $f^{-1}(y)$ to be the unique point of X that maps to y, so that $f^{-1}(\{y\}) = \{f^{-1}(y)\}$.

Definition 1.2.4 (Inverse)**.** Let $f : X \to Y$ be a bijective function. The *inverse* of f is the function $f^{-1} : Y \to X$ that assigns to each $y \in Y$ the unique $x \in X$ with $f(x) = y$.

The function f and its inverse f^{-1} undo each other, in the sense that

$$f^{-1}(f(x)) = x \ \text{ for every } x \in X, \qquad f(f^{-1}(y)) = y \ \text{ for every } y \in Y.$$

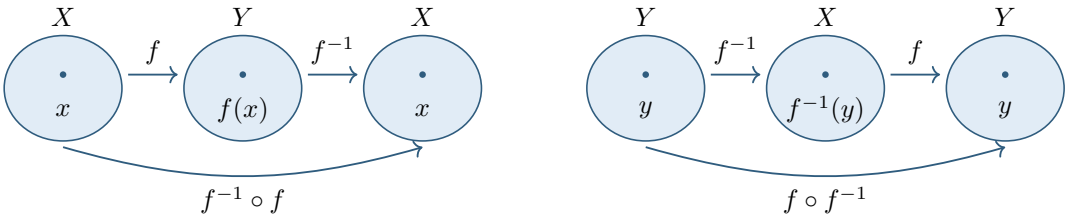


Figure 1.2.1: Applying f and then f^{-1} returns every point of X to itself; applying f^{-1} and then f returns every point of Y to itself.

The theorem below shows that if we tried to define an "inverse" for a function that is not bijective, we would fail.

Theorem 1.2.5 (Existence of an inverse)**.** *For a function $f : X \to Y$, the following are equivalent.*

1. *f is bijective.*

2. *There is a function $g : Y \to X$ with $(g \circ f)(x) = x$ for every $x \in X$ and $(f \circ g)(y) = y$ for every $y \in Y$.*

Moreover, the function g in (2) is unique and equals f^{-1}.

Proof (optional). An equivalence asserts two implications, so we prove that (1) implies (2) and that (2) implies (1). The uniqueness claim is proved last.

(1) implies (2). Assume f is bijective. Then $f^{-1} : Y \to X$ is defined by Definition 1.2.4, and we take $g = f^{-1}$. Let $x \in X$ and put $y = f(x)$. By Definition 1.2.4, $f^{-1}(y)$ is the unique point of X that f maps to y; since $f(x) = y$, that point is x. Using Definition 1.2.3 to expand the composition,

$$(f^{-1} \circ f)(x) = f^{-1}(f(x)) = f^{-1}(y) = x.$$

Now let $y \in Y$ and put $x = f^{-1}(y)$, so that $f(x) = y$ by Definition 1.2.4. Then

$$(f \circ f^{-1})(y) = f(f^{-1}(y)) = f(x) = y.$$

Both equations hold for arbitrary $x \in X$ and $y \in Y$, so $g = f^{-1}$ is a function of the kind required in (2).

(2) implies (1). Assume such a g exists. To show that f is injective, let $x, x' \in X$ satisfy $f(x) = f(x')$. Applying g to both sides and using Definition 1.2.3,

$$x = (g \circ f)(x) = g(f(x)) = g(f(x')) = (g \circ f)(x') = x',$$

where the first and last equalities are the hypothesis on g. So f is injective in the sense of Definition 1.2.2. To show that f is surjective, let $y \in Y$ and put $x = g(y)$, which lies in X because $g : Y \to X$. Then

$$f(x) = f(g(y)) = (f \circ g)(y) = y,$$

so $y \in f(X)$. Since y was an arbitrary point of Y, this gives $Y \subseteq f(X)$, and the reverse inclusion $f(X) \subseteq Y$ holds because f takes its values in Y. Hence $f(X) = Y$, which is surjectivity in Definition 1.2.2. Being both injective and surjective, f is bijective.

Uniqueness. Let g be any function satisfying (2). By the previous paragraph f is bijective, so f^{-1} is defined. Let $y \in Y$ and put $x = f^{-1}(y)$, so that $f(x) = y$ by Definition 1.2.4. Then

$$g(y) = g(f(x)) = (g \circ f)(x) = x = f^{-1}(y).$$

The functions g and f^{-1} have the same domain Y and the same codomain X, and we have just shown that they agree at every $y \in Y$. Hence $g = f^{-1}$. □

1.3 Logic

1.3.1 Logical Statements and Operations

A *statement* (also referred to as a *proposition* or an *assertion*) is a sentence that is either true or false. For example, "$2 + 2 = 4$" is a true statement and "$2 + 2 = 5$" is a false one.

Given a statement P, its *negation* $\neg P$ is true if and only if P is false. Given two statements P and Q, their *conjunction* $P \wedge Q$ is true if and only if both P and Q are true, while their

disjunction $P \vee Q$ is true if and only if at least one of P or Q is true. We typically illustrate these logical operations with truth tables:

P	$\neg P$
T	F
F	T

P	Q	$P \wedge Q$
T	T	T
T	F	F
F	T	F
F	F	F

P	Q	$P \vee Q$
T	T	T
T	F	T
F	T	T
F	F	F

1.3.2 Implication and equivalence

Perhaps the most important logical operation is implication. We define it via a truth table as follows:

P	Q	$P \Rightarrow Q$
T	T	T
T	F	F
F	T	T
F	F	T

The implication $P \Rightarrow Q$ is read "P implies Q," or "if P, then Q," or "P is sufficient for Q," or "Q is necessary for P." Its *converse* is $Q \Rightarrow P$, and its *contrapositive* is $\neg Q \Rightarrow \neg P$. If $P \Rightarrow Q$ and $Q \Rightarrow P$ both hold, we write $P \Leftrightarrow Q$ and say that P and Q are *equivalent*, or P if and only if (IFF) Q.

Our first proof involves showing that an implication and its contrapositive are equivalent.

Theorem 1.3.1 (Contrapositive law)**.**

$$(P \Rightarrow Q) \iff (\neg Q \Rightarrow \neg P).$$

Proof. The truth tables for $P \Rightarrow Q$ and $\neg Q \Rightarrow \neg P$ are

P	Q	$P \Rightarrow Q$
T	T	T
T	F	F
F	T	T
F	F	T

P	Q	$\neg Q$	$\neg P$	$\neg Q \Rightarrow \neg P$
T	T	F	F	T
T	F	T	F	F
F	T	F	T	T
F	F	T	T	T

Since the truth tables are the same, the statements are equivalent. □

Note that an implication $P \Rightarrow Q$ and its converse $Q \Rightarrow P$ are not equivalent. Showing that is left as an exercise, along with proving the following negation rules for statements.

Theorem 1.3.2 (Negation rules for statements)**.** *For statements P and Q,*

$$\begin{aligned} \neg(P \wedge Q) &\iff \neg P \vee \neg Q, \\ \neg(P \vee Q) &\iff \neg P \wedge \neg Q, \\ \neg(P \Rightarrow Q) &\iff P \wedge \neg Q. \end{aligned}$$

1.3.3 Properties and quantifiers

Given a set X, a *property* on X is a sentence $P(x)$ that becomes either true or false (hence a statement) once a particular $x \in X$ is specified. For instance, if $X = \mathbb{Z}$ and $P(x)$ is "x is even," then $P(2)$ is true and $P(3)$ is false.

To turn a property into a statement, we use quantifiers. The *universal quantifier* $\forall$ means "for every" and the *existential quantifier* $\exists$ means "there exists." Continuing the example above, the statement

$$\forall x \in \mathbb{Z},\ P(x)$$

is false, since 3 is not even, while the statement

$$\exists x \in \mathbb{Z},\ P(x)$$

is true, since 2 is even.

Theorem 1.3.3 (Negation rules for quantifiers)**.** *For a property $P(x)$ on a set X,*

$$\neg(\forall x \in X,\ P(x)) \iff \exists x \in X,\ \neg P(x),$$
$$\neg(\exists x \in X,\ P(x)) \iff \forall x \in X,\ \neg P(x).$$

For example, $A \subseteq B$ means $\forall x\,(x \in A \Rightarrow x \in B)$. Its negation is

$$\exists x\,(x \in A \text{ and } x \notin B).$$

1.4 Proofs

1.4.1 What a proof is

A *theorem* is a statement that has been shown to be true, and a *proof* is the argument that shows it. Hammack's *Book of Proof* describes a proof as follows:

> "A proof of a theorem is a written verification that shows that the theorem is definitely and unequivocally true. A proof should be understandable and convincing to anyone who has the requisite background and knowledge."

A proof is typically written in ordinary prose plus mathematical notation. However, every step in a proof must be justified by something already accepted: a hypothesis of the theorem, a definition, a result established earlier, or a rule of logic. Proofs are often accompanied by pictures, but a picture usually does not replace a formal argument.

1.4.2 Direct proof

In a *direct proof* of an implication $P \Rightarrow Q$, we assume P and derive Q. In a direct proof of a universal statement $\forall x \in X,\ P(x)$, we let $x \in X$ be *arbitrary*, meaning that nothing is assumed about x beyond $x \in X$, and derive $P(x)$; since the argument used no property of x, it applies to every element of X.

Below is one example of a direct proof. It uses the following vocabulary: an integer n is *even* if $n = 2a$ for some $a \in \mathbb{Z}$, and *odd* if $n = 2a + 1$ for some $a \in \mathbb{Z}$. Every integer is even or odd, and no integer is both; we take that as given.

Lemma 1.4.1. *If n is an even integer, then $-5n-3$ is an odd integer.*

Proof. Let n be an arbitrary even integer. Then $n = 2a$ for some $a \in \mathbb{Z}$, so

$$-5n - 3 = -5(2a) - 3 = -10a - 3 = 2(-5a - 2) + 1.$$

Now set $b = -5a - 2$. Since a is an integer, so is b, and we obtain $-5n - 3 = 2b + 1$. Hence $-5n - 3$ is odd. □

A *proof by contrapositive* of $P \Rightarrow Q$ is a direct proof of $\neg Q \Rightarrow \neg P$: we assume $\neg Q$ and derive $\neg P$. By Theorem 1.3.1 the two implications are equivalent, so proving either one proves the other. It is worth switching when the negations are the easier statements to work with.

Here is a second proof of Lemma 1.4.1, this time by contrapositive. Its hypothesis is that $-5n - 3$ is not odd, hence even, and its conclusion is that n is not even, hence odd.

Second proof of Lemma 1.4.1. We prove the contrapositive: if $-5n - 3$ is an even integer, then n is an odd integer. Let n be an integer for which $-5n - 3$ is even, so that $-5n - 3 = 2b$ for some $b \in \mathbb{Z}$. Adding $6n + 3$ to both sides gives

$$n = (-5n - 3) + 6n + 3 = 2b + 6n + 3 = 2(b + 3n + 1) + 1.$$

Now set $c = b + 3n + 1$. Since b and n are integers, so is c, and the display reads $n = 2c + 1$. So n is odd, and therefore not even. By Theorem 1.3.1, this proves Lemma 1.4.1. □

1.4.3 Proof by contradiction

A proof by contrapositive requires the statement to be an implication. A *proof by contradiction* applies to a statement of any shape. To prove a statement S, we assume that S is false, and from that assumption we derive some statement C together with its negation $\neg C$. Whatever C may be, the statement $C \wedge \neg C$ is false, so the assumption cannot hold and S is true. The method is also called *reductio ad absurdum*: supposing the opposite of what we want reduces us to an absurdity.

The proof by contradiction is justified by the equivalence

$$S \iff \big(\neg S \Rightarrow (C \wedge \neg C)\big),$$

which holds for every statement C: showing that $\neg S$ leads to a contradiction is the same as showing that S is true. Verifying the equivalence with a truth table, as in the proof of Theorem 1.3.1, is left as an exercise.

When the statement to be proved is an implication $P \Rightarrow Q$, its negation is $P \wedge \neg Q$ by Theorem 1.3.2, so the proof begins by assuming both P and $\neg Q$. Which statement C will do the job is usually not known at the outset: we assume the negation, derive consequences, and stop once some statement and its negation are both on the page. Often the statement contradicted is one assumed at the start, a hypothesis or the negated conclusion, and sometimes it is a fact known independently.

Lemma 1.4.2. *Let $x \in \mathbb{R}$. If $x \leq \varepsilon$ for every real number $\varepsilon > 0$, then $x \leq 0$.*

Proof. The hypothesis P is that $x \le \varepsilon$ for every $\varepsilon > 0$, and the conclusion Q is that $x \le 0$. By Theorem 1.3.2, to assume the implication false is to assume P and $\neg Q$. Therefore, we start by assuming that

$$x \le \varepsilon \ \text{ for every } \varepsilon > 0, \qquad \text{and} \qquad x > 0.$$

By the assumption on the right, $x/2 > 0$. The assumption on the left holds for every $\varepsilon > 0$, so it holds for $\varepsilon = x/2$, giving

$$x \le \frac{x}{2}.$$

Subtracting $x/2$ from both sides gives $x/2 \le 0$, and multiplying by 2 gives $x \le 0$. The statement C is $x \le 0$: we have just derived it, and its negation $x > 0$ is one of our assumptions. So C is both true and false, a contradiction. □

1.4.4 Proof by induction

Some statements come indexed by the natural numbers: for each $n \in \mathbb{N}$ there is a statement $P(n)$, and the claim is that $P(n)$ holds for every n. The following theorem reduces that claim, which involves infinitely many statements, to two checks.

Theorem 1.4.3 (Principle of mathematical induction)**.** *For each $n \in \mathbb{N}$, let $P(n)$ be a statement. Suppose that*

1. *$P(1)$ is true, and*
2. *for every $n \in \mathbb{N}$, $P(n) \Rightarrow P(n+1)$.*

Then $P(n)$ is true for every $n \in \mathbb{N}$.

Condition (1) is the *base step* and condition (2) is the *induction step*; the assumption $P(n)$ made while proving the induction step is the *induction hypothesis*. The two conditions act like a line of dominoes: the first one falls by (1), and each one that falls knocks over the next by (2), so all of them fall.

The induction step is itself a universal implication, so it is proved directly: let $n \in \mathbb{N}$ be arbitrary, assume $P(n)$, and derive $P(n+1)$.

Lemma 1.4.4 (Bernoulli's inequality)**.** *Let $x \in \mathbb{R}$ with $x > -1$. Then for every $n \in \mathbb{N}$,*

$$(1+x)^n \ge 1 + nx.$$

Proof. Fix $x > -1$ and let $P(n)$ be the displayed inequality.

Base step. For $n = 1$ both sides equal $1 + x$, so $P(1)$ holds.

Induction step. Let $n \in \mathbb{N}$ be arbitrary and assume $P(n)$, that is, $(1+x)^n \ge 1+nx$ (*induction hypothesis*). Multiply the induction hypothesis by $1 + x$, which is positive because $x > -1$, to obtain

$$(1+x)^{n+1} = (1+x)^n(1+x) \ge (1+nx)(1+x) = 1 + (n+1)x + nx^2.$$

Since $nx^2 \ge 0$, we have $(1+x)^{n+1} \ge 1 + (n+1)x$, which is $P(n+1)$. Thus $P(n) \Rightarrow P(n+1)$ for every $n \in \mathbb{N}$.

By Theorem 1.4.3, $P(n)$ holds for every $n \in \mathbb{N}$. □

Lecture 2

Real Analysis in $\mathbb{R}^n$

In economics, we often want to maximize a function subject to constraints:

$$\max_{x \in D} f(x).$$

Here f is the objective function and D is the *feasible set*: the set of choices satisfying the constraints. There is little point in trying to solve this problem if no solution exists. This lecture develops the machinery to guarantee that a solution exists, in $\mathbb{R}^n$.

2.1 Distance and sequences

We begin by defining Euclidean distance in $\mathbb{R}^n$. Note that there are other ways to measure and define distance, even in $\mathbb{R}^n$. We opt to only use Euclidean distance for most of this class.

Definition 2.1.1 (Inner product, norm, and distance)**.** For $x, y \in \mathbb{R}^n$,

- the *inner product* (also known as the *dot product*) of x and y is

$$x \cdot y = \sum_{i=1}^{n} x_i y_i,$$

- the (Euclidean) *norm* (or *length*) of x is

$$\|x\| = \sqrt{x \cdot x} = \sqrt{\sum_{i=1}^{n} x_i^2},$$

- the *distance* between x and y is

$$d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

For example, $x = (3, 4) \in \mathbb{R}^2$ has $x \cdot x = 25$ and $\|x\| = 5$, and the distance between $(1, 2)$ and $(4, 6)$ is 5 as well. In $\mathbb{R}^1$ the norm is the absolute value, so $d(x, y) = |x - y|$.

The inner product and the norm are related by the Cauchy–Schwarz inequality:

Theorem 2.1.2 (Cauchy–Schwarz)**.** *For every $x, y \in \mathbb{R}^n$,*

$$(x \cdot y)^2 \leq (x \cdot x)(y \cdot y),$$

and hence $|x \cdot y| \leq \|x\|\|y\|$.

Proof (optional). If $y = 0$, both sides of the first inequality vanish.

Next suppose $y \neq 0$, so that $y \cdot y = \sum_{i=1}^n y_i^2 > 0$. For every $\lambda \in \mathbb{R}$, consider a vector $z = x - \lambda y$. By Definition 2.1.1, $z \cdot z = \sum_{i=1}^n z_i^2 \geq 0$. Hence, we have

$$0 \leq z \cdot z = (x - \lambda y) \cdot (x - \lambda y) = x \cdot x - 2\lambda\,(x \cdot y) + \lambda^2\,(y \cdot y).$$

Next, choose $\lambda = (x \cdot y)/(y \cdot y)$ so that

$$0 \leq x \cdot x - 2\frac{(x \cdot y)}{(y \cdot y)}\,(x \cdot y) + \left(\frac{(x \cdot y)}{(y \cdot y)}\right)^2\,(y \cdot y) = x \cdot x - \frac{(x \cdot y)^2}{y \cdot y}.$$

Multiplying it by $y \cdot y > 0$ gives the first inequality in Theorem 2.1.2. Taking square roots gives the second inequality in Theorem 2.1.2, since $(x \cdot x)^{1/2} = \|x\|$ by Definition 2.1.1. □

Here are some useful properties of the norm.

Theorem 2.1.3 (Norm properties)**.** *For $x, y \in \mathbb{R}^n$ and $a \in \mathbb{R}$,*

1. *$\|x\| \geq 0$, and $\|x\| = 0$ if and only if $x = 0$ (positivity),*
2. *$\|ax\| = |a|\|x\|$ (homogeneity),*
3. *$\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality).*

Proof. Parts (1) and (2) follow from Definition 2.1.1. The triangle inequality is shown as follows:

$$\|x + y\|^2 = \|x\|^2 + 2\,x \cdot y + \|y\|^2 \leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2,$$

where the first inequality uses the Cauchy–Schwarz inequality. Taking square roots gives the triangle inequality. □

Read in terms of the distance function d, the three norm properties say that the distance behaves as a distance should: for all $x, y, z \in \mathbb{R}^n$,

$$\begin{aligned} &d(x, y) \geq 0, \quad \text{with equality exactly when } x = y, \\ &d(x, y) = d(y, x), \qquad d(x, z) \leq d(x, y) + d(y, z). \end{aligned}$$

Here, nonnegativity is part (1) applied to $x - y$; symmetry is part (2) with $a = -1$; and the triangle inequality is part (3) applied to $x - z = (x - y) + (y - z)$. A function satisfying these three conditions is called a *metric*.

Equipped with a concept of distance, we can name the set of points lying within a given distance of a point.

Definition 2.1.4 (Open ball)**.** For $x \in \mathbb{R}^n$ and $r > 0$, the *open ball* with center x and radius r is

$$B(x, r) = \{y \in \mathbb{R}^n : \|y - x\| < r\}.$$

Definition 2.1.5 (Interior points, open sets, and limit points)**.** Let $D \subseteq \mathbb{R}^n$.

- A point $x \in D$ is an *interior point* of D if $B(x, r) \subseteq D$ for some $r > 0$. The set of all interior points of D is its *interior*, written $\operatorname{int} D$.
- The set D is *open* if every point of D is an interior point.
- A point $a \in \mathbb{R}^n$ is a *limit point* of D if every ball $B(a, r)$ contains a point of D different from a; equivalently,

$$B(a, r) \cap \big(D \setminus \{a\}\big) \neq \emptyset \qquad \text{for every } r > 0.$$

For example, let $a < b$ and take $x \in (a, b)$. The number $r = \min\{x - a, b - x\}$ is positive and $B(x, r) \subseteq (a, b)$, so (a, b) is open. The endpoints of $[a, b]$ are not interior points, and therefore

$$\operatorname{int}[a, b] = (a, b).$$

The limit-point condition says that points of D different from a can be found arbitrarily close to a; the point a need not itself belong to D.

Geometrically, in $\mathbb{R}$ the ball $B(x, r)$ is the interval $(x - r, x + r)$, and in $\mathbb{R}^2$ it is the disk of radius r around x without its bounding circle.

Definition 2.1.6 (Sequence and convergence)**.** A *sequence* in $\mathbb{R}^n$ is a list $(x_k)_{k=1}^{\infty}$ with $x_k \in \mathbb{R}^n$. It *converges* to $x \in \mathbb{R}^n$, and x is its *limit*, written $x_k \to x$, if for every $\varepsilon > 0$ there is $K \in \mathbb{N}$ such that

$$k \geq K \quad \Longrightarrow \quad \|x_k - x\| < \varepsilon.$$

A sequence that has no limit in $\mathbb{R}^n$ *diverges*.

In the notation of Definition 2.1.4, the condition says that every ball $B(x, \varepsilon)$, however small its radius, contains all terms of the sequence from some index on.

We check the definition on two sequences in $\mathbb{R}$.

Example 2.1.7 (Convergence and divergence)**.**

1. The sequence $x_k = 1/k$ converges to 0. Let $\varepsilon > 0$ and take an integer $K > 1/\varepsilon$. Then every $k \geq K$ satisfies

$$|x_k - 0| = \frac{1}{k} \leq \frac{1}{K} < \varepsilon,$$

which is the implication Definition 2.1.6 requires of K. Since $\varepsilon > 0$ was arbitrary, $x_k \to 0$.

2. The sequence $x_k = (-1)^k$ diverges. Let $x \in \mathbb{R}$ be a candidate limit and take $\varepsilon = 1$. By part (3) of Theorem 2.1.3,

$$|1 - x| + |-1 - x| \geq |(1 - x) - (-1 - x)| = 2,$$

so at least one of the two numbers is at least 1. The values 1 and -1 both occur at arbitrarily large indices, so every $K \in \mathbb{N}$ admits some $k \geq K$ with $|x_k - x| \geq 1$, and the implication in Definition 2.1.6 fails for $\varepsilon = 1$. Since x was arbitrary, no limit exists.

A subsequence is exactly what is sounds like: a sequence drawn from another sequence by skipping some terms.

Definition 2.1.8 (Subsequence)**.** If $k_1 < k_2 < \cdots$ is a strictly increasing sequence of indices, then $(x_{k_j})_{j=1}^{\infty}$ is a *subsequence* of (x_k).

Here are some useful properties of sequences and their limits.

Theorem 2.1.9 (Basic limit facts)**.** *Let (x_k) be a sequence in $\mathbb{R}^n$.*

1. *If $x_k \to x$ and $x_k \to y$, then $x = y$.*
2. *If $x_k \to x$, then (x_k) is bounded: there is $M > 0$ with $\|x_k\| \le M$ for every k.*
3. *The convergence $x_k \to x$ holds if and only if $x_{k,i} \to x_i$ in $\mathbb{R}$ for every coordinate $i = 1, \dots, n$.*
4. *If $x_k \to x$, then every subsequence of (x_k) also converges to x.*

Proof (optional). Uniqueness. Let $\varepsilon > 0$. By Definition 2.1.6 there is K_1 with $\|x_k - x\| < \varepsilon/2$ for $k \ge K_1$, and K_2 with $\|x_k - y\| < \varepsilon/2$ for $k \ge K_2$. Fix one $k \ge \max\{K_1, K_2\}$ and write $x - y = (x - x_k) + (x_k - y)$; parts (2) and (3) of Theorem 2.1.3 give

$$\|x - y\| \le \|x_k - x\| + \|x_k - y\| < \varepsilon.$$

So the nonnegative number $\|x - y\|$ lies below every positive ε, hence equals 0, and part (1) of Theorem 2.1.3 gives $x = y$.

Boundedness. Left as an exercise: Definition 2.1.6 with $\varepsilon = 1$ leaves all but finitely many terms in $B(x, 1)$, and a bound is the largest of $\|x\| + 1$ and the norms of those remaining terms.

Coordinates. Write $a_j = x_{k,j} - x_j$. Since a_i^2 is one of the squares summed in Definition 2.1.1, and since that sum is at most $\left(\sum_{j=1}^n |a_j|\right)^2$, taking square roots gives

$$|x_{k,i} - x_i| \le \|x_k - x\| \le \sum_{j=1}^{n} |x_{k,j} - x_j|.$$

Let $\varepsilon > 0$. If $x_k \to x$, the left inequality lets any K that Definition 2.1.6 supplies for ε serve the coordinate sequence $(x_{k,i})$ as well. Conversely, if every coordinate converges, choose for each j an index K_j beyond which $|x_{k,j} - x_j| < \varepsilon/n$; then $k \ge \max\{K_1, \dots, K_n\}$ makes the right-hand sum, and hence $\|x_k - x\|$, smaller than ε.

Subsequences. The indices of a subsequence satisfy $k_j \ge j$: this holds at $j = 1$ because indices are natural numbers, and $k_{j+1} > k_j \ge j$ forces $k_{j+1} \ge j + 1$. Now let $\varepsilon > 0$ and take the K that Definition 2.1.6 supplies for (x_k). Every $j \ge K$ has $k_j \ge j \ge K$, and therefore $\|x_{k_j} - x\| < \varepsilon$. □

Limits also respect the arithmetic of $\mathbb{R}^n$, one operation at a time.

Theorem 2.1.10 (Algebra of limits)**.** *Suppose $x_k \to x$ and $y_k \to y$ in $\mathbb{R}^n$, and $a_k \to a$ and $b_k \to b$ in $\mathbb{R}$. Then*

1. *$x_k + y_k \to x + y$,*
2. *$a_k x_k \to ax$,*
3. *$a_k b_k \to ab$,*

4. $x_k \cdot y_k \to x \cdot y$,

5. $\|x_k\| \to \|x\|$,

6. $a_k/b_k \to a/b$, *provided* $b \neq 0$ *and* $b_k \neq 0$ *for every* k,

7. $x \geq y$, *provided* $x_k \geq y_k$ *coordinatewise for every* k.

Its proof is left as an exercise; the claims follow from part (3) of Theorem 2.1.3, which gives $|\|x_k\| - \|x\|| \leq \|x_k - x\|$, from Theorem 2.1.2, and from the coordinatewise criterion in part (3) of Theorem 2.1.9, together with the boundedness supplied by part (2).

2.2 Suprema and compactness

We first develop two ingredients for existence arguments: a way to describe the limiting upper value of a set, and a condition that makes sequences behave well inside a feasible set.

2.2.1 Suprema and maximizing sequences

Definition 2.2.1 (Bounds and extrema)**.** Let $A \subseteq \mathbb{R}$ be nonempty.

- A number u is an *upper bound* of A if
$$a \leq u \qquad \text{for every } a \in A,$$
and A is *bounded above* if it has an upper bound.
- The *supremum*, or *least upper bound*, of A, written $\sup A$, is an upper bound s of A satisfying
$$s \leq u \qquad \text{for every upper bound } u \text{ of } A.$$
- A number z is the *maximum* of A, written $\max A$, if
$$z \in A \qquad \text{and} \qquad a \leq z \ \text{ for every } a \in A.$$

Lower bounds, *bounded below*, the *infimum*, or *greatest lower bound*, $\inf A$, and the *minimum* $\min A$ are defined by reversing every inequality above.

If a maximum exists, it is automatically the supremum. Conversely, if the supremum exists and belongs to A, it is the maximum. Thus
$$\max A \text{ exists} \iff \sup A \text{ exists and } \sup A \in A,$$
and then the two numbers are equal.

For example, the upper bounds of $(0, 1)$ are precisely the numbers at least 1, so
$$\sup(0, 1) = 1, \qquad \inf(0, 1) = 0.$$
Neither bound belongs to the interval, so it has neither a maximum nor a minimum. The interval $[0, 1]$ has the same supremum and infimum, but it contains both; hence $\max[0, 1] = 1$ and $\min[0, 1] = 0$.

Definition 2.2.1 does not guarantee that a supremum exists; it is possible for a set to have upper bounds but no least one. For instance, the set of rational numbers q with $q^2 < 2$ is bounded above by 2, but it has no least upper bound in $\mathbb{Q}$. When dealing with $\mathbb{R}$, we make a fundamental assumption that every nonempty set that is bounded above has a supremum. This is known as the completeness property of the real numbers.

Axiom 2.2.2 (Completeness of $\mathbb{R}$)**.** Every nonempty subset of $\mathbb{R}$ that is bounded above has a supremum in $\mathbb{R}$.

Note the role of both hypotheses of the axiom. First, the set $\mathbb{R}$ is nonempty but not bounded above, so it has no supremum. Second, the empty set $\emptyset$ is bounded above by every real number, but it has no least upper bound.

2.2.2 Closed, bounded, and compact sets

Our existence argument will turn a sequence of feasible points into a candidate solution. For this to work, some subsequence must converge and its limit must remain in the feasible set. Compactness guarantees both.

Definition 2.2.3 (Closed, bounded, and compact sets)**.** A set $D \subseteq \mathbb{R}^n$ is

- *closed* if $x_k \in D$ and $x_k \to x$ imply $x \in D$;
- *bounded* if there is $M > 0$ such that $\|x\| \le M$ for every $x \in D$;
- *compact* if every sequence (x_k) in D has a subsequence converging to some point of D.

There are two failures to remember: $x_k = k$ runs off to infinity in $\mathbb{R}$, while $x_k = 1/k$ converges to a point missing from $(0, 1)$. Compactness rules out both.

In Euclidean space these conditions are related by a fundamental result that we use without proof.

Theorem 2.2.4 (Heine–Borel)**.** *A set $D \subseteq \mathbb{R}^n$ is compact if and only if it is closed and bounded.*

Thus every closed interval $[a, b]$ is compact, whereas $(0, 1)$ and $\mathbb{R}$ are not. The empty set is compact, so compactness alone does not supply a feasible point.

2.3 Continuity and existence

2.3.1 Continuity

So far, a limit has described the behavior of a sequence. We will also need to describe the behavior of a function as its input approaches a point.

Definition 2.3.1 (Limit of a function)**.** Let $D \subseteq \mathbb{R}^n$, let $f : D \to \mathbb{R}^m$, let a be a limit point of D, and let $L \in \mathbb{R}^m$. We write

$$\lim_{x \to a} f(x) = L$$

if, for every sequence (x_k) in $D \setminus \{a\}$ such that $x_k \to a$, we have $f(x_k) \to L$.

In Euclidean space this sequential definition is equivalent to the familiar ε–δ condition: for every $\varepsilon > 0$ there is $\delta > 0$ such that

$$x \in D, \quad 0 < \|x - a\| < \delta \quad \Longrightarrow \quad \|f(x) - L\| < \varepsilon.$$

The restriction $x \neq a$ is deliberate: a function limit depends on values near a, not on the value at a. The point a need not be in the domain. For example, if $q(h) = h^2/|h| = |h|$ for $h \neq 0$, then Theorem 2.1.10 gives $\lim_{h\to 0} q(h) = 0$, even though $q(0)$ is not defined.

For functions of one real variable, suppose a is approached by points of D from the indicated side. The notation

$$\lim_{x \downarrow a} f(x) = L \qquad \text{and} \qquad \lim_{x \uparrow a} f(x) = L$$

denotes the *right-hand limit* and *left-hand limit*, respectively. The first applies Definition 2.3.1 to $D \cap (a, \infty)$; the second applies it to $D \cap (-\infty, a)$. If the two-sided limit exists, then every one-sided limit that is defined exists and equals it.

Definition 2.3.2 (Continuity)**.** Let $D \subseteq \mathbb{R}^n$, let $f : D \to \mathbb{R}^m$, and let $x \in D$. The function f is *continuous at* x if, whenever $x_k \in D$ and $x_k \to x$, we have

$$f(x_k) \to f(x).$$

The function is *continuous on* D, or simply *continuous*, if it is continuous at every point of D.

At a limit point $x \in D$, Definition 2.3.2 is equivalent to

$$\lim_{y \to x} f(y) = f(x).$$

Equivalently, for every $\varepsilon > 0$ there is $\delta > 0$ such that, for every $y \in D$,

$$\|y - x\| < \delta \quad \Longrightarrow \quad \|f(y) - f(x)\| < \varepsilon.$$

We will continue to use the sequential formulation because it fits our existence argument. In Lecture 4, the same function-limit notation will define derivatives: a difference quotient is evaluated only for $h \neq 0$, while h approaches 0.

Here are the rules that let us build continuous functions out of simpler ones.

Theorem 2.3.3 (Operations preserving continuity)**.** *Let $D \subseteq \mathbb{R}^n$.*

1. *If $f, g : D \to \mathbb{R}$ are continuous and $a \in \mathbb{R}$, then $f + g$, af, and fg are continuous on D, and the quotient f/g is continuous on the subset of D on which $g \neq 0$.*
2. *A composition of continuous functions is continuous.*

Its proof is left as an exercise. Part (1) follows from Definition 2.3.2 applied to an arbitrary sequence $x_k \to x$, together with Theorem 2.1.10; part (2) follows from Definition 2.3.2 applied twice.

For example, constant functions and the identity function $f(x) = x$ are continuous on $\mathbb{R}$ by Definition 2.3.2. Part (1) then makes every polynomial continuous on $\mathbb{R}$, and every rational function continuous off the zeros of its denominator, such as $(x^2+1)/(x-2)$ on $\mathbb{R} \setminus \{2\}$. Taking as known that exp is continuous on $\mathbb{R}$ and log on $(0, \infty)$, part (2) gives combinations: e^{-x^2} is continuous on $\mathbb{R}$, and so is $\log(1 + x^2)$, since $1 + x^2$ stays positive.

2.3.2 The Weierstrass theorem

We are now ready to show existence. First, let us formalize what a solution to an optimization problem is.

Definition 2.3.4 (Maximizers and minimizers)**.** Let $D \subseteq \mathbb{R}^n$ be nonempty and let $f : D \to \mathbb{R}$. The set of maximizers of f on D is

$$\arg\max_{x \in D} f(x) = \{x^* \in D : f(x^*) \geq f(x) \text{ for every } x \in D\}.$$

The set $\arg\min_{x \in D} f(x)$ of minimizers is defined by reversing the inequality.

The main existence result is the Weierstrass theorem, which we can now state and prove.

Theorem 2.3.5 (Weierstrass)**.** *Let $D \subseteq \mathbb{R}^n$ be nonempty and compact, and let $f : D \to \mathbb{R}$ be continuous. Then f attains a maximum and a minimum on D; equivalently,*

$$\arg\max_{x \in D} f(x) \neq \emptyset, \qquad \arg\min_{x \in D} f(x) \neq \emptyset.$$

Proof (optional). We first prove that f has a maximizer; the minimizer will follow at the end.

Step 1: show that the attainable values $f(D)$ are bounded above. Suppose they were not. Then for each k there would be $y_k \in D$ with $f(y_k) > k$. By Definition 2.2.3, some subsequence satisfies $y_{k_j} \to y \in D$, and Definition 2.3.2 gives $f(y_{k_j}) \to f(y)$. Thus $(f(y_{k_j}))$ is bounded by part (2) of Theorem 2.1.9. But $k_j \geq j$, so $f(y_{k_j}) > k_j \geq j$ for every j, making the same sequence unbounded. This contradiction proves that $f(D)$ is bounded above.

Step 2: construct feasible points whose values approach the best possible value. The set $f(D)$ is nonempty because D is nonempty. By Step 1 it is also bounded above, so completeness gives the real number $S = \sup f(D)$. For each k, the smaller number $S - 1/k$ cannot be an upper bound of $f(D)$. Hence there is $x_k \in D$ with

$$S - \frac{1}{k} < f(x_k) \leq S, \qquad \text{that is,} \qquad |f(x_k) - S| < \frac{1}{k}.$$

Since $1/k \to 0$ by part (1) of Example 2.1.7, this proves $f(x_k) \to S$. The sequence (x_k) is called a *maximizing sequence*.

Step 3: obtain a feasible candidate for the maximizer. Compactness of D gives a subsequence x_{k_j} and a point $x^* \in D$ such that $x_{k_j} \to x^*$. The membership $x^* \in D$ is important: the candidate remains feasible.

Step 4: prove that the candidate is a maximizer. Continuity gives $f(x_{k_j}) \to f(x^*)$. On the other hand, because $f(x_k) \to S$, its subsequence also satisfies $f(x_{k_j}) \to S$ by part (4) of Theorem 2.1.9. Limits are unique, so $f(x^*) = S$. Since S is an upper bound of $f(D)$, we have $f(x) \leq S = f(x^*)$ for every $x \in D$. Therefore x^* is a maximizer in the sense of Definition 2.3.4.

Step 5: obtain a minimizer. The function $-f$ is continuous by part (1) of Theorem 2.3.3. Applying Steps 1–4 to $-f$ gives a point that maximizes $-f$, which is exactly a point that minimizes f. □

Note that failure of any one of the three hypotheses — nonemptiness, compactness, or continuity — can lead to a function that does not attain its maximum or minimum.

Example 2.3.6 (One failed hypothesis at a time)**.** None of the following problems has a maximizer, and in each one exactly one hypothesis of Theorem 2.3.5 fails.

1. In $\mathbb{R}^2$, let $f(x) = x_1 + x_2$ on

$$D = \{x \in \mathbb{R}^2 : x_1 \geq 0,\ x_2 \geq 0,\ x_1 + x_2 \leq 1,\ x_1 x_2 \geq 1\}.$$

   The feasible set D is empty, so f has no maximizer.

2. $D = (0, 1)$ and $f(x) = x$. The domain is bounded but not closed, since $x_k = 1/(k+1)$ lies in D and converges to $0 \notin D$, so again it is not compact. Here $\sup f(D) = 1$, while $f(x) < 1$ for every $x \in D$.

3. $D = [0, 1]$, which is compact, and

$$f(x) = \begin{cases} x, & 0 \leq x < 1, \\ 0, & x = 1. \end{cases}$$

   What fails is continuity, and only at 1: the sequence $x_k = 1 - 1/k$ lies in D and converges to 1, while $f(x_k) = 1 - 1/k \to 1 \neq 0 = f(1)$, which Definition 2.3.2 forbids. Again $\sup f(D) = 1$ is not attained.

That said, a maximum might exist even if one of the hypotheses of Theorem 2.3.5 fails. In that case, one must prove existence by other means, such as by finding a candidate and verifying that it is indeed a maximizer.

2.3.3 Intermediate values and fixed points

While we need linear algebra to meaningfully study functions in $\mathbb{R}^n$, we can already prove some useful results about continuous functions in $\mathbb{R}$. The intermediate value theorem says that a continuous function on an interval takes on every value between its values at the endpoints.

Theorem 2.3.7 (Intermediate value theorem)**.** *Let $f : [a, b] \to \mathbb{R}$ be continuous. If*

$$\min\{f(a), f(b)\} \leq c \leq \max\{f(a), f(b)\},$$

then there is $x \in [a, b]$ such that $f(x) = c$.

Proof (optional). Assume first that $f(a) \leq c \leq f(b)$, and let

$$A = \{x \in [a, b] : f(x) \leq c\}.$$

Then A contains a and is bounded above by b, so $s = \sup A$ exists by Axiom 2.2.2 and satisfies $a \leq s \leq b$. We show that $f(s) = c$.

$f(s) \leq c$. For each k, the number $s - 1/k$ is smaller than s and therefore not an upper bound of A, so we may pick $a_k \in A$ with

$$s - \frac{1}{k} < a_k \leq s.$$

Then $a_k \to s$, since $1/k \to 0$ by part (1) of Example 2.1.7, and Definition 2.3.2 gives $f(a_k) \to f(s)$. Each a_k lies in A, so $f(a_k) \leq c$ for every k, and part (7) of Theorem 2.1.10 carries the inequality to the limit: $f(s) \leq c$.

$f(s) \geq c$. If $s = b$, this is the hypothesis $f(b) \geq c$. If $s < b$, no point of $(s, b]$ lies in A, so $f(x) > c$ on $(s, b]$. The points $x_k = s + 1/k$ lie in $(s, b]$ as soon as $1/k \leq b - s$, and $x_k \to s$, so

Definition 2.3.2 gives $f(x_k) \to f(s)$. Part (7) of Theorem 2.1.10 again carries $f(x_k) > c$ to the limit, giving $f(s) \geq c$.

This proves $f(s) = c$. In the remaining case $f(b) \leq c \leq f(a)$, apply what we have just proved to $-f$, continuous by part (1) of Theorem 2.3.3, and to the value $-c$: this produces a point x with $-f(x) = -c$, that is, $f(x) = c$. □

From that, we obtain our first *fixed point theorem* in $\mathbb{R}$ for free.

Theorem 2.3.8 (Brouwer in $\mathbb{R}$)**.** *Let $a \leq b$. Every continuous function $f : [a, b] \to [a, b]$ has a fixed point: there is $x^* \in [a, b]$ such that $f(x^*) = x^*$.*

Proof. Let $g(x) = f(x) - x$, continuous on $[a, b]$ by part (1) of Theorem 2.3.3. Since f takes its values in $[a, b]$, we have $f(a) \geq a$ and $f(b) \leq b$, so

$$g(a) = f(a) - a \geq 0, \qquad g(b) = f(b) - b \leq 0.$$

Hence $g(b) \leq 0 \leq g(a)$, and Theorem 2.3.7 applied to g with $c = 0$ gives $x^* \in [a, b]$ with $g(x^*) = 0$, which means $f(x^*) = x^*$. □

Lecture 3

Linear Algebra

Linear algebra studies *linear systems*: systems of equations in which every equation is linear. It is the foundation of optimization and also a powerful tool for analyzing nonlinear systems. In this lecture we introduce the basic vocabulary and operations of linear algebra, and we study the solution sets of linear systems. Throughout this lecture, vectors are columns.

3.1 Matrix algebra

3.1.1 Matrix vocabulary and basic arithmetic

Definition 3.1.1 (Matrices and transpose)**.** An $m \times n$ *matrix* $A = (a_{ij})$ is a rectangular array with m rows and n columns, with a_{ij} in row i and column j. Two matrices are equal when they have the same dimensions and all corresponding entries are equal. The *transpose* of A is the $n \times m$ matrix $A^\top$ defined by

$$(A^\top)_{ij} = a_{ji}.$$

A matrix is *square* if it has the same number of rows and columns, *diagonal* if every off-diagonal entry is zero, and *symmetric* if $A = A^\top$.

Matrices of the same dimensions are added entrywise, and scalar multiplication is entrywise:

$$(A + B)_{ij} = a_{ij} + b_{ij}, \qquad (\alpha A)_{ij} = \alpha a_{ij}.$$

The $m \times n$ *zero matrix* has every entry zero. The $n \times n$ *identity matrix* I_n has ones on the diagonal and zeros elsewhere. The diagonal matrix with diagonal entries $d_1, \ldots, d_n$ is written $\operatorname{diag}(d_1, \ldots, d_n)$.

Example 3.1.2 (Matrix arithmetic)**.** Let

$$A = \begin{pmatrix} 1 & 2 & 0 \\ -1 & 0 & 3 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 1 & 1 \\ 2 & -1 & 0 \end{pmatrix}.$$

Then

$$A + B = \begin{pmatrix} 1 & 3 & 1 \\ 1 & -1 & 3 \end{pmatrix}, \qquad 2A = \begin{pmatrix} 2 & 4 & 0 \\ -2 & 0 & 6 \end{pmatrix}, \qquad A^\top = \begin{pmatrix} 1 & -1 \\ 2 & 0 \\ 0 & 3 \end{pmatrix}.$$

Addition requires matching dimensions. Multiplication uses a different compatibility rule.

3.1.2 Matrix multiplication

Definition 3.1.3 (Matrix multiplication)**.** If A is $m \times n$ and B is $n \times p$, then their product AB is the $m \times p$ matrix with

$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

The product is defined only when the inner dimensions agree. In that case the matrices are *conformable*:

$$(m \times n)(n \times p) = (m \times p).$$

Each entry of AB is a row of A multiplied against a column of B.

Example 3.1.4 (Matrix multiplication)**.** Let

$$C = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \qquad D = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

Then C is 2×3 and D is 3×2, so

$$CD = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}$$

is 2×2. The reverse product is also defined in this example, but it is different in both size and entries:

$$DC = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 1 & 3 & 1 \end{pmatrix}.$$

Since $CD \neq DC$, this example also shows that matrix multiplication is not generally commutative.

Whenever the products are conformable,

$$(AB)C = A(BC), \qquad A(B + C) = AB + AC, \qquad (AB)^\top = B^\top A^\top,$$

and the identity matrices satisfy $AI_n = A$ and $I_m A = A$ for an $m \times n$ matrix A.

Vectors can be treated as matrices with one column. For $x, y \in \mathbb{R}^n$, $x^\top y$ is a 1×1 matrix whose entry is the inner product $x \cdot y$, whereas $xy^\top$ is an $n \times n$ matrix.

3.1.3 Linear combinations, span, and independence

Definition 3.1.5 (Linear combination)**.** Let $v_1, \ldots, v_k \in \mathbb{R}^n$ be a collection of vectors. A *linear combination* of $v_1, \ldots, v_k$ is a vector $\sum_{i=1}^{k} \alpha_i v_i$ with $\alpha_i \in \mathbb{R}$.

Below is a collection of vectors we will use to illustrate concepts in this lecture.

Example 3.1.6 (A linear combination)**.** Let $v_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $v_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$, and $v_3 = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$. Here $v_3 = v_1 + v_2$, so v_3 is a linear combination of v_1 and v_2, with $\alpha_1 = \alpha_2 = 1$.

Definition 3.1.7 (Subspace and span)**.** Let $V \subseteq \mathbb{R}^n$ and let $v_1, \ldots, v_k \in \mathbb{R}^n$.

- The set V is a *subspace* of $\mathbb{R}^n$ if it is nonempty and

$$x, y \in V, \quad \alpha, \beta \in \mathbb{R} \quad \Longrightarrow \quad \alpha x + \beta y \in V.$$

- Their *span* is the set of all such combinations:

$$\operatorname{span}\{v_1, \ldots, v_k\} = \left\{ \sum_{i=1}^{k} \alpha_i v_i : \alpha_1, \ldots, \alpha_k \in \mathbb{R} \right\}.$$

Every span is a subspace. Indeed, adding two linear combinations of $v_1, \ldots, v_k$, or multiplying one by a scalar, produces another linear combination of the same vectors.

The span of one nonzero vector v is the line $\{\alpha v : \alpha \in \mathbb{R}\}$ through the origin. A line that does not go through the origin cannot be a subspace, because every subspace contains the zero vector.

Definition 3.1.8 (Independence, basis, and dimension)**.** Let $V \subseteq \mathbb{R}^n$ be a subspace.

- The collection $v_1, \ldots, v_k$ of vectors is *linearly independent* if

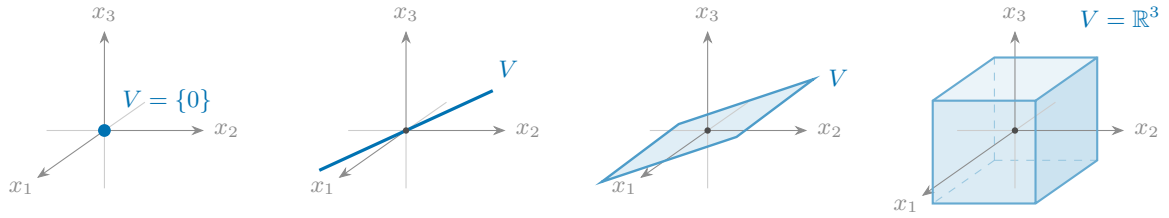
$$\alpha_1 v_1 + \cdots + \alpha_k v_k = 0 \quad \Longrightarrow \quad \alpha_1 = \cdots = \alpha_k = 0$$

  for all $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$. Otherwise it is *linearly dependent*.

- A *basis* of V is a linearly independent collection of vectors that spans V.

- The *dimension* $\dim V$ of V is the number of vectors in a basis of V.

In $\mathbb{R}^3$ we can illustrate all four possible dimensions of subspaces — 0, 1, 2, and 3. That is, every subspace of $\mathbb{R}^3$ is one of the following: the origin; a line through the origin; a plane through the origin; or all of $\mathbb{R}^3$. Figure 3.1.1 shows one subspace of each dimension. This is the picture to keep in mind for any subspace of $\mathbb{R}^n$: a subspace is flat, passes through the origin, and has dimension at most n. Unless it is $\{0\}$, it extends without bound in every direction that it contains.



(a) $\dim V = 0$: the origin

(b) $\dim V = 1$: a line through 0

(c) $\dim V = 2$: a plane through 0

(d) $\dim V = 3$: all of $\mathbb{R}^3$

Figure 3.1.1: The four kinds of subspaces of $\mathbb{R}^3$, one for each dimension: the origin alone, a line through the origin, a plane through the origin, and the whole space. Every subspace of $\mathbb{R}^3$ is one of these. Each is flat and contains the origin. The line and plane panels show representative viewing windows: those subspaces continue without bound. There is one dimension-0 subspace and one dimension-3 subspace, but infinitely many dimension-1 and dimension-2 subspaces.

One could think of linear dependence as redundancy: if vectors are dependent, then at least one of them can be expressed as a linear combination of the others.

Example 3.1.9. Let

$$v_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \qquad v_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \qquad v_3 = \begin{pmatrix} 3 \\ 4 \end{pmatrix}.$$

Because $v_3 = v_1 + v_2$, the three-vector collection is dependent and v_3 may be deleted without changing its span. The remaining vectors v_1 and v_2 are independent: if $\alpha_1 v_1 + \alpha_2 v_2 = 0$, then

$$\alpha_1 + 2\alpha_2 = 0, \qquad 2\alpha_1 + 2\alpha_2 = 0,$$

which forces $\alpha_1 = \alpha_2 = 0$. Thus v_1, v_2 form a basis of $\mathbb{R}^2$, and the three vectors span $\mathbb{R}^2$. Figure 3.1.2 illustrates these relationships.

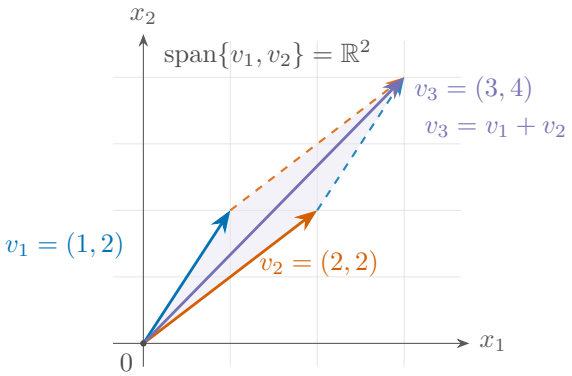


Figure 3.1.2: The vectors v_1, v_2, v_3 in $\mathbb{R}^2$. The solid arrows all begin at the origin. The dashed translated copies complete the shaded parallelogram and show $v_3 = v_1 + v_2$. Thus v_3 contributes no new direction: v_1, v_2 are independent and already span $\mathbb{R}^2$, while the three-vector collection is dependent. The span is the entire plane, not only the shaded parallelogram.

We use the standard finite-dimensional facts that every subspace of $\mathbb{R}^n$ has a basis, any two bases of the same subspace have the same number of vectors, and every independent collection in $\mathbb{R}^n$ has at most n vectors. The *standard basis* of $\mathbb{R}^n$ is the collection $e_1, \ldots, e_n$, where e_i is the vector with 1 in coordinate i and 0 in every other coordinate.

3.2 Linear maps and linear systems

So far we have treated matrices as arrays of numbers. We now show that matrices represent linear functions from $\mathbb{R}^n$ to $\mathbb{R}^m$, and we use this perspective to study the solution sets of systems of linear equations. The same product Ax also has two useful interpretations. We begin by naming the rows and columns of A, and then develop each interpretation.

3.2.1 Two readings of a matrix-vector product

Let

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

Write the rows of A as $r_1, r_2, \ldots, r_m$ and its columns as $c_1, c_2, \ldots, c_n$. Thus

$$A = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{pmatrix} = \begin{pmatrix} c_1 & c_2 & \cdots & c_n \end{pmatrix}.$$

Suppose $x \in \mathbb{R}^n$. We call x the *input* and $Ax \in \mathbb{R}^m$ the *output*. The product can be represented as

$$Ax = \begin{pmatrix} a_{11} \ldots a_{1n} \\ \vdots \\ a_{m1} \ldots a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{pmatrix} = x_1c_1 + \cdots + x_nc_n. \tag{3.2.1}$$

The product Ax is thus a linear combination of the columns of A, with entries of x as weights. For the row interpretation of matrix A, we need to consider an equation $Ax = b$.

Definition 3.2.1 (Linear system)**.** For an $m \times n$ matrix A and $b \in \mathbb{R}^m$, the equation

$$Ax = b, \qquad x \in \mathbb{R}^n,$$

is a *linear system* with *coefficient matrix* A and *right-hand side* b. It is *consistent* if it has a solution and *inconsistent* otherwise.

The row interpretation writes the system as one scalar equation for each row of A:

$$Ax = b \iff r_ix = a_{i1}x_1 + \cdots + a_{in}x_n = b_i, \qquad \text{for each row } i. \tag{3.2.2}$$

Each row is thus a restriction on what the input x could be to reach output b using matrix A. At the same time, the column representation essentially asks which linear combination of the columns of A produces the target b.

Make sure you understand the representations in (3.2.1) and (3.2.2), as they will be crucial for understanding just about everything that comes next. Here is the running example.

Example 3.2.2. Let

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 4 \end{pmatrix}, \qquad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

The rows and columns of A are

$$r_1 = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}, \qquad r_2 = \begin{pmatrix} 2 & 2 & 4 \end{pmatrix},$$

and

$$c_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \qquad c_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \qquad c_3 = \begin{pmatrix} 3 \\ 4 \end{pmatrix}.$$

The column interpretation gives

$$Ax = x_1c_1 + x_2c_2 + x_3c_3 = x_1 \begin{pmatrix} 1 \\ 2 \end{pmatrix} + x_2 \begin{pmatrix} 2 \\ 2 \end{pmatrix} + x_3 \begin{pmatrix} 3 \\ 4 \end{pmatrix}.$$

For

$$b = \begin{pmatrix} 3 \\ 4 \end{pmatrix},$$

the row interpretation writes $Ax = b$ as

$$x_1 + 2x_2 + 3x_3 = 3, \qquad 2x_1 + 2x_2 + 4x_3 = 4.$$

The theorem below shows that a function from $\mathbb{R}^n$ to $\mathbb{R}^m$ is linear if and only if it can be represented as an $m \times n$ matrix. We use the word map or transformation instead of the word function to emphasize that the domain and codomain are vector spaces, and that the function preserves the vector space structure.

Definition 3.2.3 (Linear map)**.** A function $T : \mathbb{R}^n \to \mathbb{R}^m$ is a *linear map*, or a *linear transformation*, if

$$T(x + y) = T(x) + T(y), \qquad T(\alpha x) = \alpha T(x)$$

for all $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$.

Here x is the *input* and $T(x)$ is the *output*.

Theorem 3.2.4 (Matrices represent linear maps)**.** *A function $T : \mathbb{R}^n \to \mathbb{R}^m$ is linear if and only if there is an $m \times n$ matrix A such that $T(x) = Ax$ for every $x \in \mathbb{R}^n$. In that case A is unique, and its jth column is $T(e_j)$. If $S : \mathbb{R}^m \to \mathbb{R}^p$ is linear with matrix B, then the composition $S \circ T$ is linear with matrix BA.*

Proof (optional). Linear implies matrix. Let T be linear. Every $x \in \mathbb{R}^n$ is the linear combination $x = x_1 e_1 + \cdots + x_n e_n$ of the standard basis, so linearity applied to each term gives

$$T(x) = T\left(\sum_{j=1}^{n} x_j e_j\right) = \sum_{j=1}^{n} x_j T(e_j).$$

Let $A = [\,T(e_1) \;\cdots\; T(e_n)\,]$, which is $m \times n$ because each $T(e_j)$ lies in $\mathbb{R}^m$. Since Ax is the combination of the columns of A with weights $x_1, \ldots, x_n$, the display says exactly that $T(x) = Ax$.

Matrix implies linear. If $T(x) = Ax$ for an $m \times n$ matrix A, then $A(x + y) = Ax + Ay$ and $A(\alpha x) = \alpha(Ax)$, so T is linear.

Uniqueness. Suppose $Ax = Bx$ for every $x \in \mathbb{R}^n$. Taking $x = e_j$ makes Ae_j and Be_j the jth columns of A and B, so the two matrices agree column by column and $A = B$. The same substitution in $T(x) = Ax$ identifies the jth column of A as $T(e_j)$.

Composition. Let S be linear with matrix B. For every $x \in \mathbb{R}^n$, associativity of the matrix product gives

$$(S \circ T)(x) = S(T(x)) = B(Ax) = (BA)x,$$

so $S \circ T$ is represented by the $p \times n$ matrix BA, and it is linear by the second part. □

3.2.2 Row space, column space, and rank

Definition 3.2.5 (Row and column spaces)**.** Let A be an $m \times n$ matrix. Its *row space* and *column space* are

$$\mathrm{Row}(A) = \mathrm{span}\{r_1, \dots, r_m\} \subseteq \mathbb{R}^n, \qquad \mathrm{Col}(A) = \mathrm{span}\{c_1, \dots, c_n\} \subseteq \mathbb{R}^m.$$

By (3.2.1),

$$\mathrm{Col}(A) = \{Ax : x \in \mathbb{R}^n\},$$

so the column space is the range of the linear map $x \mapsto Ax$.

The row and column spaces live in different ambient spaces, but they have the same dimension. We use this fundamental result without proof.

Theorem 3.2.6 (Equality of row and column rank)**.** *For every matrix A,*

$$\dim \mathrm{Row}(A) = \dim \mathrm{Col}(A).$$

Definition 3.2.7 (Rank)**.** The *rank* of A is their common dimension:

$$\mathrm{rank}\, A = \dim \mathrm{Row}(A) = \dim \mathrm{Col}(A).$$

The row reading interprets rank as the number of independent restrictions. The column reading interprets it as the dimension of the set of outputs that A can produce.

Example 3.2.8 (Rank of a matrix)**.** Let

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 4 \end{pmatrix}.$$

The first two columns of A form a basis of $\mathbb{R}^2$. Thus

$$\mathrm{Col}(A) = \mathbb{R}^2, \qquad \mathrm{rank}\, A = 2.$$

The two rows are consequently independent: the system contains two independent restrictions on three variables.

3.2.3 Null space and changes that preserve the restrictions

Suppose the input x produces the output b, so $Ax = b$. Let $h \in \mathbb{R}^n$ be a proposed change to the input. After the change, the new input is $x + h$. It produces the same output exactly when

$$A(x + h) = b \quad \Longleftrightarrow \quad Ah = 0.$$

Thus h leaves Ax unchanged if and only if $Ah = 0$. The collection of all such changes is the null space.

Definition 3.2.9 (Null space)**.** For an $m \times n$ matrix A, its *null space*, or *kernel*, is

$$N(A) = \{h \in \mathbb{R}^n : Ah = 0\}.$$

The dimension of $N(A)$ is called the *nullity* of A.

Every null space is a subspace. Indeed, if $u, v \in N(A)$ and $\alpha, \beta \in \mathbb{R}$, then

$$A(\alpha u + \beta v) = \alpha Au + \beta Av = 0.$$

We have now seen two ways to specify a subspace: a span specifies one using generating vectors, whereas a null space specifies one using equations of the form $Ax = 0$.

Its elements have two interpretations. In the row interpretation,

$$Ah = 0 \quad \Longleftrightarrow \quad a_{i1}h_1 + \cdots + a_{in}h_n = 0 \quad \text{for every row } i,$$

so adding h to x changes the left-hand side of every constraint by zero. In the column reading,

$$Ah = 0 \quad \Longleftrightarrow \quad h_1c_1 + \cdots + h_nc_n = 0.$$

Here h_j is the change in the weight on column c_j. Thus $Ah = 0$ means that these changes cancel, leaving the output unchanged. A nonzero such h exists exactly when the columns are linearly dependent.

Rank and nullity measure two different things. The rank is the dimension of the set of outputs that A can produce. The nullity is the dimension of the set of changes that leave Ax unchanged. The rank–nullity theorem relates these two counts. We use it without proof.

Theorem 3.2.10 (Rank–nullity theorem)**.** *If A is an $m \times n$ matrix, then*

$$\operatorname{rank} A + \dim N(A) = n.$$

Example 3.2.11 (Null space of a matrix)**.** Let

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 4 \end{pmatrix}.$$

Here $n = 3$ and $\operatorname{rank} A = 2$, so $\dim N(A) = 1$. To identify the null space, solve $Ah = 0$:

$$h_1 + 2h_2 + 3h_3 = 0, \qquad 2h_1 + 2h_2 + 4h_3 = 0.$$

Subtracting twice the first equation from the second gives $h_2 = -h_3$. Substitution into the first gives $h_1 = -h_3$. Letting $t = -h_3$,

$$N(A) = \left\{ \begin{pmatrix} t \\ t \\ -t \end{pmatrix} : t \in \mathbb{R} \right\} = \operatorname{span} \left\{ \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \right\}.$$

This illustrates rank–nullity:

$$\operatorname{rank} A + \dim N(A) = 2 + 1 = 3 = n.$$

Here the nullity 1 means that one scalar, t, can be chosen freely when describing changes that preserve all restrictions. More generally, if $\operatorname{rank} A = r$, then $n - r$ coefficients can be chosen freely; these are the $n - r$ degrees of freedom.

3.2.4 The solution set

The column interpretation answers the existence question immediately:

$$Ax = b \text{ is consistent} \quad \Longleftrightarrow \quad b \in \mathrm{Col}(A). \tag{3.2.3}$$

The following theorem describes the solution set of a consistent linear system $Ax = b$ in terms of a particular solution and the null space (which is the solution set for $Ax = 0$).

Theorem 3.2.12 (Solution set)**.** *If x_0 is one solution of $Ax = b$, then the full solution set is*

$$\{x \in \mathbb{R}^n : Ax = b\} = x_0 + N(A), \qquad \textit{where } x_0 + N(A) = \{x_0 + h : h \in N(A)\}.$$

Consequently, a consistent system has a unique solution if and only if $N(A) = \{0\}$.

Proof. If $Ax = b = Ax_0$, then $A(x - x_0) = 0$, so $x - x_0 \in N(A)$. Conversely, if $h \in N(A)$, then $A(x_0 + h) = Ax_0 + Ah = b$. □

Together with the column-space criterion, the theorem gives the complete description

$$\{x \in \mathbb{R}^n : Ax = b\} = \begin{cases} \varnothing, & b \notin \mathrm{Col}(A), \\ x_0 + N(A), & b \in \mathrm{Col}(A). \end{cases}$$

Example 3.2.13 (Solution set of a linear system)**.** Let

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 4 \end{pmatrix}, \qquad b = \begin{pmatrix} 3 \\ 4 \end{pmatrix}.$$

The null space is

$$N(A) = \operatorname{span}\left\{ \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \right\},$$

and one particular solution is $x_0 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$. Therefore

$$\{x : Ax = b\} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + \operatorname{span}\left\{ \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \right\}.$$

Adding the displayed null vector gives another solution:

$$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix},$$

reflecting that the third column of A is the sum of its first two columns.

3.2.5 Rank, existence, and uniqueness

The preceding results give the two full-rank criteria.

Theorem 3.2.14 (Full row and column rank)**.** *Let A be $m \times n$.*

1. $\operatorname{rank} A = m$ *if and only if* $Ax = b$ *has at least one solution for every* $b \in \mathbb{R}^m$.
2. $\operatorname{rank} A = n$ *if and only if* $Ax = b$ *has at most one solution for every* $b \in \mathbb{R}^m$.

The first clause follows because $\operatorname{rank} A = m$ exactly when $\operatorname{Col}(A) = \mathbb{R}^m$, which means every $b \in \mathbb{R}^m$ belongs to the column space. For the second clause, rank–nullity gives

$$\operatorname{rank} A = n \quad \Longleftrightarrow \quad N(A) = \{0\}.$$

The solution-set theorem then says that a consistent system has exactly one solution.

Together, rank and the column-space criterion classify every system:

Rank condition	Solutions
$\operatorname{rank} A = m = n$	every b has exactly one
$\operatorname{rank} A = m < n$	every b has infinitely many
$\operatorname{rank} A = n < m$	none or exactly one, depending on b
$\operatorname{rank} A < \min\{m, n\}$	none or infinitely many, depending on b

3.3 Square matrices and invertibility

Definition 3.3.1 (Inverse)**.** An $n \times n$ matrix A is *invertible*, or *nonsingular*, if there is an $n \times n$ matrix A^{-1} satisfying

$$A^{-1}A = AA^{-1} = I_n.$$

If no inverse exists, A is *singular*.

An inverse is unique: if B and C are both inverses of A, then $B = B(AC) = (BA)C = C$.

Theorem 3.3.2 (Invertible-matrix equivalences)**.** *For an $n \times n$ matrix A, the following are equivalent:*

1. *A is invertible;*
2. $\operatorname{rank} A = n$*;*
3. *the columns of A are linearly independent;*
4. $N(A) = \{0\}$*;*
5. *for every $b \in \mathbb{R}^n$, the system $Ax = b$ has exactly one solution.*

When these conditions hold, the solution is $x = A^{-1}b$.

The equivalence of clauses (2)–(5) follows from full row and column rank and the affine solution-set theorem. If A is invertible, multiplying $Ax = b$ by A^{-1} gives the unique solution $x = A^{-1}b$. Conversely, if every $Ax = e_j$ has a unique solution x^j, put those solutions into the columns of $B = [\, x^1 \ \cdots \ x^n \,]$. Then $AB = I_n$, and the triviality of $N(A)$ forces $BA = I_n$, so $B = A^{-1}$.

The determinant is a function that turns a square matrix into a scalar.

Definition 3.3.3 (Determinant)**.** For a 1×1 matrix, $\det(a) = a$. For $n \geq 2$, let A_{1j} denote the $(n-1) \times (n-1)$ matrix obtained from A by deleting row 1 and column j. The *determinant* of A is

$$\det A = \sum_{j=1}^{n} (-1)^{1+j} a_{1j} \det A_{1j}.$$

This is called *cofactor expansion along the first row*: each entry of the first row is multiplied by the determinant of the matrix left after deleting that entry's row and column, and the signs alternate. Expanding along any other row or column gives the same number.

Example 3.3.4 (2×2 determinant)**.** Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Here $A_{11} = (d)$ and $A_{12} = (c)$, so

$$\det A = ad - bc.$$

Example 3.3.5 (3×3 determinant)**.** Let $A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$. Each of the three terms now needs a 2×2 determinant:

$$\det A = a(ei - fh) - b(di - fg) + c(dh - eg).$$

The determinant supplies a scalar test for the equivalent conditions above.

Theorem 3.3.6 (Determinant criterion)**.** *A square matrix A is invertible if and only if* $\det A \neq 0$.

Example 3.3.7 (The inverse in the 2×2 case)**.** Let

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad \det A = ad - bc.$$

If $\det A \neq 0$, multiplying in either order gives

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = (ad - bc) I_2.$$

As long as the determinant $ad - bc \neq 0$, we can divide by it to get the inverse:

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

3.4 Eigenvalues and eigenvectors

Definition 3.4.1 (Eigenvalues and eigenvectors)**.** Let A be a real $n \times n$ matrix. A scalar $\lambda \in \mathbb{R}$ is an *eigenvalue* of A if there is a nonzero vector $v \in \mathbb{R}^n$ such that

$$Av = \lambda v.$$

Such a vector v is an *eigenvector* associated with λ.

The eigenvalue equation can be rewritten as

$$Av = \lambda v \quad \Longleftrightarrow \quad (A - \lambda I)v = 0. \tag{3.4.1}$$

As such, we immediately have the following equivalences, which are useful for computing eigenvalues and eigenvectors.

Theorem 3.4.2 (Eigenvalue equivalences)**.** *For a real $n \times n$ matrix A and $\lambda \in \mathbb{R}$, the following are equivalent:*

1. *λ is an eigenvalue of A;*
2. *$N(A - \lambda I) \neq \{0\}$;*
3. *$A - \lambda I$ is singular;*
4. *$\det(A - \lambda I) = 0$.*

Proof (optional). (1) $\Leftrightarrow$ (2). By Definition 3.4.1, λ is an eigenvalue exactly when $Av = \lambda v$ for some $v \neq 0$. By (3.4.1) such a v satisfies $(A - \lambda I)v = 0$, and conversely. So an eigenvalue exists precisely when $A - \lambda I$ has a nonzero null vector, which is the statement $N(A - \lambda I) \neq \{0\}$.

(2) $\Leftrightarrow$ (3). Apply Theorem 3.3.2 to the square matrix $A - \lambda I$: it is invertible if and only if its null space is trivial. Negating both sides, $A - \lambda I$ is singular if and only if $N(A - \lambda I) \neq \{0\}$.

(3) $\Leftrightarrow$ (4). Apply Theorem 3.3.6 to $A - \lambda I$: it is invertible if and only if $\det(A - \lambda I) \neq 0$. Negating both sides gives the claim. □

If one of the eigenvalues of a matrix is zero, then the matrix is singular; that can be seen by taking $\lambda = 0$ in Theorem 3.4.2.

Corollary 3.4.3 (Zero eigenvalue)**.** *A square matrix A is singular if and only if 0 is an eigenvalue of A. Equivalently, A is invertible if and only if all of its eigenvalues are nonzero.*

The equation

$$\det(A - \lambda I) = 0$$

is the *characteristic equation*. Its left-hand side is the *characteristic polynomial*. The roots of the characteristic polynomial are the eigenvalues of A. For each eigenvalue λ, the associated eigenvectors are the nonzero elements of $N(A - \lambda I)$.

Example 3.4.4 (Eigenvalues and eigenvectors of a 2×2 matrix)**.** Let

$$A = \begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix}.$$

Subtracting λ from each diagonal entry and taking the determinant gives the characteristic polynomial

$$\det(A - \lambda I) = \det \begin{pmatrix} 1 - \lambda & 2 \\ 4 & 3 - \lambda \end{pmatrix} = (1 - \lambda)(3 - \lambda) - 8 = \lambda^2 - 4\lambda - 5 = (\lambda - 5)(\lambda + 1).$$

Its roots are the eigenvalues $\lambda_1 = 5$ and $\lambda_2 = -1$.

To find the eigenvectors belonging to $\lambda_1 = 5$, solve $(A - 5I)v = 0$:

$$A - 5I = \begin{pmatrix} -4 & 2 \\ 4 & -2 \end{pmatrix},$$

so both equations reduce to $v_2 = 2v_1$, and $N(A - 5I) = \operatorname{span}\{(1, 2)^\top\}$. For $\lambda_2 = -1$, solve $(A + I)v = 0$:

$$A + I = \begin{pmatrix} 2 & 2 \\ 4 & 4 \end{pmatrix},$$

so both equations reduce to $v_2 = -v_1$, and $N(A + I) = \operatorname{span}\{(1, -1)^\top\}$.

Note that in each case the null space is a line rather than a single vector — any nonzero multiple of an eigenvector is again an eigenvector for the same eigenvalue.

Remark 3.4.5 (How many eigenvalues). The characteristic polynomial of an $n \times n$ matrix has degree n, so A has at most n distinct real eigenvalues. It can have fewer for two separate reasons: a root may repeat, or a root may fail to be real.

Example 3.4.6 (A repeated eigenvalue)**.** Let $A = \operatorname{diag}(2, 2)$. Its characteristic polynomial is $(2 - \lambda)^2$, which has the single root 2. Thus A has one distinct eigenvalue rather than two.

Example 3.4.7 (No real eigenvalues)**.** Let

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \qquad \det(A - \lambda I) = \lambda^2 + 1.$$

Since Definition 3.4.1 requires $\lambda \in \mathbb{R}$, this matrix has no eigenvalues and no nonzero eigenvectors in $\mathbb{R}^2$. It does have complex eigenvalues and eigenvectors, but we do not consider those here.

Below we list some useful results featuring eigenvalues without proving them.

Definition 3.4.8 (Trace)**.** The *trace* of an $n \times n$ matrix A is the sum of its diagonal entries,

$$\operatorname{tr} A = \sum_{i=1}^{n} a_{ii}.$$

Theorem 3.4.9 (Trace and determinant)**.** *Let A be an $n \times n$ matrix whose characteristic polynomial factors into real linear terms,*

$$\det(A - \lambda I) = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda), \qquad \lambda_1, \ldots, \lambda_n \in \mathbb{R}.$$

Then

$$\sum_{i=1}^{n} \lambda_i = \operatorname{tr} A, \qquad \prod_{i=1}^{n} \lambda_i = \det A.$$

Definition 3.4.10 (Diagonalizable matrix)**.** An $n \times n$ matrix A is *diagonalizable* if there are an invertible $n \times n$ matrix P and a diagonal matrix D with

$$A = PDP^{-1}.$$

Theorem 3.4.11 (Diagonalization)**.** *An $n \times n$ matrix A is diagonalizable if and only if it has n linearly independent eigenvectors $v^1, \ldots, v^n$. In that case one may take*

$$P = [\, v^1 \ \cdots \ v^n \,], \qquad D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

where λ_i is the eigenvalue belonging to v^i. In particular, A is diagonalizable whenever it has n distinct real eigenvalues.

This result is particularly useful for real symmetric matrices, which always have n linearly independent eigenvectors and are therefore diagonalizable.

Lecture 4

Differential Calculus

4.1 Functions of one variable

4.1.1 Definition

Let $I \subseteq \mathbb{R}$ be an open interval and let $f : I \to \mathbb{R}$. If the input moves from x_0 to $x_0 + h$, then the input changes by h and the value of the function changes by

$$f(x_0 + h) - f(x_0).$$

For $h \neq 0$, the ratio

$$\frac{f(x_0 + h) - f(x_0)}{h} \tag{4.1.1}$$

is the *average rate of change* between the two points. Geometrically, it is the slope of the line that goes through $(x_0, f(x_0))$ and $(x_0 + h, f(x_0 + h))$. The derivative is the limiting slope as $h \to 0$. Thus x_0 is the fixed point at which the derivative is evaluated, while $x_0 + h$ is the moving input.

Definition 4.1.1 (Derivative)**.** The function f is *differentiable at* x_0 if there is $a \in \mathbb{R}$ such that

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} = a.$$

The number a is the *derivative* of f at x_0, written $f'(x_0)$. If f is differentiable at every point of I, then f' denotes the *derivative function* $x \mapsto f'(x)$.

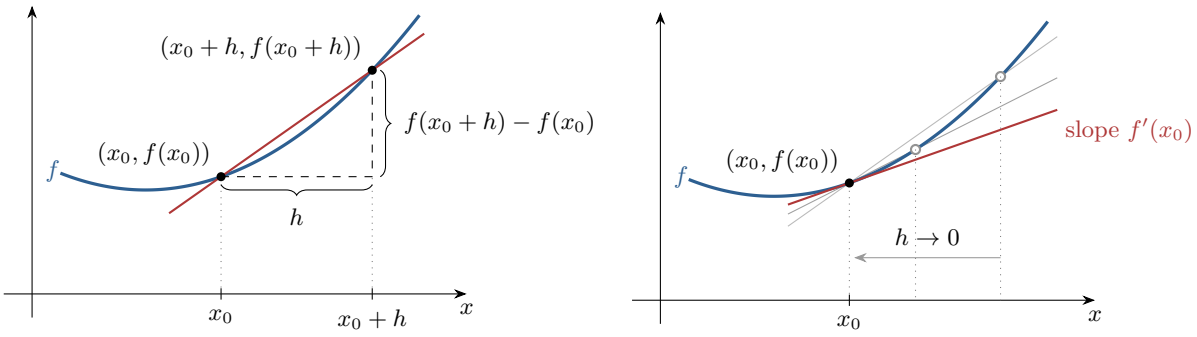


Figure 4.1.1: Left: for $h \neq 0$ the difference quotient (4.1.1) is the rise $f(x_0+h)-f(x_0)$ over the run h, that is, the slope of the secant line through $(x_0, f(x_0))$ and $(x_0+h, f(x_0+h))$. Right: as $h \to 0$ the second point slides along the graph toward the first and the secants rotate toward a single limiting line. Its slope is the derivative $f'(x_0)$, and the line itself is the tangent to the graph at $(x_0, f(x_0))$.

We begin with three direct computations of the derivative using the definition.

Example 4.1.2 (Constant, affine, and quadratic functions)**.** Fix $x_0 \in \mathbb{R}$, and let $a, b, c \in \mathbb{R}$.

1. If $f(x) = c$, then $[f(x_0+h) - f(x_0)]/h = 0$ for every $h \neq 0$, so $f'(x_0) = 0$.
2. If $f(x) = ax + b$, then
$$\frac{f(x_0+h) - f(x_0)}{h} = \frac{a(x_0+h) + b - (ax_0 + b)}{h} = a,$$
so $f'(x_0) = a$. The slope of an affine function is the same at every point.
3. If $f(x) = x^2$, then
$$\frac{f(x_0+h) - f(x_0)}{h} = \frac{(x_0+h)^2 - x_0^2}{h} = 2x_0 + h \longrightarrow 2x_0.$$
Thus $f'(x) = 2x$. In particular, $f'(3) = 6$.

4.1.2 Derivative as a linear approximation

Definition 4.1.1 has an equivalent form that will extend cleanly to several variables. Because I is open, $f(x_0+h)$ is defined for every sufficiently small change h, so the limit below makes sense.

Theorem 4.1.3 (Derivative as a linear approximation)**.** *Let $I \subseteq \mathbb{R}$ be an open interval, let $f : I \to \mathbb{R}$, and let $x_0 \in I$. Then f is differentiable at x_0 if and only if there is $a \in \mathbb{R}$ such that*

$$\lim_{h \to 0} \frac{|f(x_0+h) - f(x_0) - ah|}{|h|} = 0. \tag{4.1.2}$$

In that case a is unique, and $a = f'(x_0)$.

Proof. The proof relies on two facts. First, for any real-valued function u of h, $|u(h)| \to 0$ if and only if $u(h) \to 0$ as $h \to 0$ (see Lecture 2). Second, $|b/c| = |b|/|c|$ for all $b \in \mathbb{R}$ and $c \in \mathbb{R} \setminus \{0\}$. We have:

$$\begin{aligned} f \text{ is differentiable at } x_0 &\iff \exists a \in \mathbb{R} \text{ such that } \lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h} = a. \\ &\iff \exists a \in \mathbb{R} \text{ such that } \lim_{h\to 0} \left(\frac{f(x_0+h)-f(x_0)-ah}{h} \right) = 0. \\ &\iff \exists a \in \mathbb{R} \text{ such that } \lim_{h\to 0} \left| \frac{f(x_0+h)-f(x_0)-ah}{h} \right| = 0. \\ &\iff \exists a \in \mathbb{R} \text{ such that } \lim_{h\to 0} \frac{|f(x_0+h)-f(x_0)-ah|}{|h|} = 0, \end{aligned}$$

where the first equivalence is by the definition of differentiability, the second equivalence is by subtracting a from both sides, the third equivalence is the first fact, and the fourth equivalence is the second fact. Furthermore, the a in Definition 4.1.1 is unique and equals $f'(x_0)$, so the a in (4.1.2) is also unique and equals $f'(x_0)$.

□

Setting $x = x_0 + h$ gives an equivalent moving-point form of (4.1.2):

$$\lim_{x\to x_0} \frac{|f(x)-f(x_0)-a(x-x_0)|}{|x-x_0|} = 0. \tag{4.1.3}$$

In both forms x_0 is fixed. In (4.1.2), $h \to 0$; in (4.1.3), the input $x = x_0 + h$ tends to x_0.

The reason why eq. (4.1.2) is called derivative as a linear approximation is as follows. Let $r(h) = f(x_0+h) - f(x_0) - ah$ in the numerator of (4.1.2). Then, $f(x_0+h) = f(x_0) + ah + r(h)$, where $f(x_0)+ah$ is the *linear approximation* of f evaluated at x_0+h, while $r(h)$ is the *remainder* term. The condition in (4.1.2), now written as $|r(h)|/|h| \to 0$ as $h \to 0$ says that the remainder term is negligible compared to h as $h \to 0$.

Example 4.1.4 (Reading the derivative from an expansion)**.** Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = x^2$, and let $x_0 = 3$. The function evaluated at $x_0 + h$ is

$$f(3+h) = (3+h)^2 = \underbrace{9}_{f(3)} + \underbrace{6h}_{\text{linear term}} + \underbrace{h^2}_{r(h)}.$$

The remainder satisfies

$$\lim_{h\to 0} \frac{|r(h)|}{|h|} = \lim_{h\to 0} \frac{|h^2|}{|h|} = \lim_{h\to 0} |h| = 0.$$

Therefore the linear coefficient is the derivative: $f'(3) = 6$.

We now have two ways of calculating derivatives for functions from $\mathbb{R}$ to $\mathbb{R}$: the difference-quotient Definition 4.1.1 and the linear-expansion method of Theorem 4.1.3. The latter method is generalizable to functions of several variables, so we will use it in the next section.

4.1.3 Properties of differentiability

Theorem 4.1.5 (Differentiability implies continuity)**.** *Let $I \subseteq \mathbb{R}$ be an open interval, let $f : I \to \mathbb{R}$, and let $x_0 \in I$. If f is differentiable at x_0, then f is continuous at x_0.*

Proof (optional). For $h \neq 0$ small enough that $x_0 + h \in I$,

$$\begin{aligned}\lim_{h\to 0}\big[f(x_0+h)-f(x_0)\big] &= \lim_{h\to 0}\left(\frac{f(x_0+h)-f(x_0)}{h}\cdot h\right)\\ &= \left(\lim_{h\to 0}\frac{f(x_0+h)-f(x_0)}{h}\right)\left(\lim_{h\to 0} h\right)\\ &= f'(x_0)\cdot 0 = 0,\end{aligned}$$

where the last line uses Definition 4.1.1. Hence

$$\lim_{h\to 0} f(x_0+h) = f(x_0) + \lim_{h\to 0}\big[f(x_0+h)-f(x_0)\big] = f(x_0),$$

which is continuity of f at x_0. □

The converse fails: a continuous function need not be differentiable.

Example 4.1.6 (Continuity without differentiability)**.** Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = |x|$. At the origin, $|f(x) - f(0)| = |x| \to 0$, so f is continuous. But for $h \neq 0$, the difference quotient $[f(h) - f(0)]/h = |h|/h$ equals 1 when $h > 0$ and -1 when $h < 0$. It has no limit as $h \to 0$, so f is not differentiable at 0.

Theorem 4.1.7 (Rules of differentiation)**.** *Let f and g be differentiable at x_0, and let $\alpha, \beta \in \mathbb{R}$. Then*

1. $(\alpha f + \beta g)'(x_0) = \alpha f'(x_0) + \beta g'(x_0)$;
2. $(fg)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0)$;
3. *if $g(x_0) \neq 0$, then*
$$(f/g)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{g(x_0)^2}.$$

Its proof uses Definition 4.1.1 and is left as an exercise.

Theorem 4.1.8 (Chain rule)**.** *Let $I, J \subseteq \mathbb{R}$ be open intervals, let $g : I \to \mathbb{R}$ satisfy $g(I) \subseteq J$, and let $f : J \to \mathbb{R}$. If g is differentiable at $x_0 \in I$ and f is differentiable at $g(x_0)$, then $f \circ g$ is differentiable at x_0 and*

$$(f \circ g)'(x_0) = f'(g(x_0))\, g'(x_0).$$

Proof (optional). Write $y_0 = g(x_0)$. Because J is open, $y_0 + t \in J$ for every sufficiently small t, and for those t define

$$\varphi(t) = \begin{cases} \dfrac{f(y_0+t)-f(y_0)}{t}, & t \neq 0,\\[2mm] f'(y_0), & t = 0.\end{cases}$$

By Definition 4.1.1, $\varphi(t) \to f'(y_0) = \varphi(0)$ as $t \to 0$, so φ is continuous at 0. Multiplying by t,

$$f(y_0+t) - f(y_0) = t\,\varphi(t), \tag{4.1.4}$$

which holds for every t in the domain of φ, including $t = 0$.

Next, let $h \neq 0$ tend to zero with $x_0 + h \in I$, and set $t = g(x_0 + h) - g(x_0)$. Since g is continuous at x_0 by Theorem 4.1.5, $t \to 0$, so for every sufficiently small h, (4.1.4) reads

$$f(g(x_0 + h)) - f(g(x_0)) = \big[g(x_0 + h) - g(x_0)\big]\, \varphi\big(g(x_0 + h) - g(x_0)\big),$$

Dividing by h,

$$\frac{f(g(x_0 + h)) - f(g(x_0))}{h} = \frac{g(x_0 + h) - g(x_0)}{h} \cdot \varphi\big(g(x_0 + h) - g(x_0)\big).$$

As $h \to 0$, the first factor tends to $g'(x_0)$ by Definition 4.1.1. For the second, $t \to 0$ as noted above, and φ is continuous at 0, so $\varphi(g(x_0 + h) - g(x_0)) \to \varphi(0) = f'(y_0)$. Hence

$$\begin{aligned}\lim_{h\to 0} \frac{f(g(x_0 + h)) - f(g(x_0))}{h} &= \left(\lim_{h\to 0} \frac{g(x_0 + h) - g(x_0)}{h}\right) \left(\lim_{h\to 0} \varphi\big(g(x_0 + h) - g(x_0)\big)\right) \\ &= g'(x_0)\, f'(y_0).\end{aligned}$$

□

4.1.4 Useful theorems for functions of one variable

Theorem 4.1.9 (Interior extremum)**.** *Let $a < b$, let $f : [a, b] \to \mathbb{R}$, and let $x^* \in (a, b)$. Suppose f is differentiable at x^* and that for some $\varepsilon > 0$,*

$$f(x^*) \geq f(x) \qquad \text{for every } x \in [a, b] \text{ with } |x - x^*| < \varepsilon.$$

Then $f'(x^*) = 0$. The same conclusion holds if instead $f(x^*) \leq f(x)$ for every such x.

Proof (optional). Because $x^* \in (a, b)$, we have $x^* + h \in [a, b]$ for every sufficiently small $|h|$. For such h with $0 < |h| < \varepsilon$ the hypothesis gives $f(x^* + h) - f(x^*) \leq 0$, and dividing by h preserves the inequality when $h > 0$ and reverses it when $h < 0$:

$$\frac{f(x^* + h) - f(x^*)}{h} \leq 0 \quad \text{if } h > 0, \qquad \frac{f(x^* + h) - f(x^*)}{h} \geq 0 \quad \text{if } h < 0.$$

Since f is differentiable at x^*, the two-sided limit of the difference quotient exists, so both one-sided limits exist and equal $f'(x^*)$. Letting $h \downarrow 0$ gives $f'(x^*) \leq 0$, and letting $h \uparrow 0$ gives $f'(x^*) \geq 0$, so $f'(x^*) = 0$. For the reversed inequality, apply this to $-f$. □

Theorem 4.1.10 (Rolle's theorem)**.** *Let $a < b$ and let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). If $f(a) = f(b)$, then there is $c \in (a, b)$ such that $f'(c) = 0$.*

Proof (optional). If f is constant, then $f'(c) = 0$ for every $c \in (a, b)$.

Otherwise $f(t) \neq f(a)$ for some $t \in [a, b]$. Suppose $f(t) > f(a)$. Since $[a, b]$ is compact and f is continuous, Theorem 2.3.5 from Lecture 2 gives a maximizer $c \in [a, b]$ of f on $[a, b]$. We thus have

$$f(c) \geq f(t) > f(a) = f(b),$$

in particular, $c \in (a, b)$. Then f is differentiable at c, and $f(c) \geq f(x)$ holds for every $x \in [a, b]$, including those x with $|x - c| < \varepsilon$ for some $\varepsilon > 0$. By Theorem 4.1.9, we conclude that $f'(c) = 0$.

If instead $f(t) < f(a)$, run the same argument with c being a minimizer. □

Theorem 4.1.11 (Mean value theorem)**.** *Let $a < b$ and let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Then there is $c \in (a, b)$ such that*

$$f(b) - f(a) = (b - a)f'(c).$$

Proof (optional). Let

$$g(t) = f(t) - f(a) - \frac{f(b) - f(a)}{b - a}(t - a), \qquad t \in [a, b].$$

Then g is continuous on $[a, b]$, differentiable on (a, b), and $g(a) = g(b) = 0$. By Theorem 4.1.10 there is $c \in (a, b)$ with

$$0 = g'(c) = f'(c) - \frac{f(b) - f(a)}{b - a},$$

yielding the result. □

4.2 Real-valued functions of several variables

In this section, we consider functions from $\mathbb{R}^n$ to $\mathbb{R}$. Input vectors are column vectors. We use x for a generic or moving input and x_0 for a fixed point at which a derivative is evaluated. Thus $h = x - x_0$ is the change from x_0 to x.

4.2.1 Partial derivatives

Recall from Lecture 3 that e_i is the ith standard basis vector of $\mathbb{R}^n$, with 1 in coordinate i and 0 in every other coordinate. For a real number t, the vector te_i has t in coordinate i and 0 elsewhere, and $x_0 + te_i$ adds t to coordinate i of x_0 while leaving every other coordinate of x_0 unchanged.

Definition 4.2.1 (Partial derivative)**.** Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}$, and let $x_0 \in U$. The *ith partial derivative* of f at x_0, where $i \in \{1, \dots, n\}$, is

$$\frac{\partial f}{\partial x_i}(x_0) = \lim_{t \to 0} \frac{f(x_0 + te_i) - f(x_0)}{t},$$

provided the limit exists.

This is an ordinary one-variable derivative: hold all variables except x_i fixed and differentiate with respect to x_i.

Example 4.2.2 (A function of two variables)**.** Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by

$$f(x_1, x_2) = x_1^2 x_2.$$

Holding x_2 fixed and differentiating in x_1, and then reversing their roles, gives

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = 2x_1 x_2, \qquad \frac{\partial f}{\partial x_2}(x_1, x_2) = x_1^2.$$

4.2.2 Differentiability and C^1 functions

Partial derivatives describe a function only along the coordinate lines through a point. Differentiability requires one linear approximation to work for small changes in all directions at once. A linear map from $\mathbb{R}^n$ to $\mathbb{R}$ has the form $h \mapsto a \cdot h$ for some $a \in \mathbb{R}^n$, so a and h, which were scalars in the one-variable case, are now both vectors.

Definition 4.2.3 (Differentiability)**.** Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}$, and let $x_0 \in U$. The function f is *differentiable at* x_0 if there is $a \in \mathbb{R}^n$ such that

$$\lim_{h \to 0} \frac{|f(x_0 + h) - f(x_0) - a \cdot h|}{\|h\|} = 0,$$

where the limit is over $h \in \mathbb{R}^n$ with $h \neq 0$. It is *differentiable on* U if it is differentiable at every point of U.

When $n = 1$, the vector h and the coefficient a are scalars, $a \cdot h = ah$, and $\|h\| = |h|$. Thus Definition 4.2.3 is exactly the linear-approximation form of the one-variable derivative in Theorem 4.1.3. Only the form of the linear term has changed.

Equivalently,

$$f(x_0 + h) = f(x_0) + a \cdot h + r(h), \qquad \frac{|r(h)|}{\|h\|} \longrightarrow 0 \quad \text{as } h \to 0. \tag{4.2.1}$$

The same rule $h \mapsto a \cdot h$ must work for every way in which h can approach zero.

Alternatively, set $x = x_0 + h$. Then f is differentiable at x_0 if and only if there is $a \in \mathbb{R}^n$ such that

$$\lim_{x \to x_0} \frac{|f(x) - f(x_0) - a \cdot (x - x_0)|}{\|x - x_0\|} = 0. \tag{4.2.2}$$

The base point x_0 is fixed in this limit; x is the moving input.

Definition 4.2.4 (Gradient)**.** If all partial derivatives of f exist at x_0, the *gradient* of f at x_0 is the column vector

$$\nabla f(x_0) = \begin{pmatrix} \partial f/\partial x_1(x_0) \\ \vdots \\ \partial f/\partial x_n(x_0) \end{pmatrix} \in \mathbb{R}^n.$$

When f is differentiable, the vector in its linear approximation is exactly the gradient.

Theorem 4.2.5 (Consequences of differentiability)**.** *Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be differentiable at $x_0 \in U$, with vector a as in Definition 4.2.3. Then every partial derivative of f exists at x_0 and*

$$a = \nabla f(x_0).$$

In particular, the vector a is unique. Moreover, f is continuous at x_0.

Proof (optional). Fix i and set $h = te_i$. Then $\|h\| = |t|$ and $a \cdot h = ta_i$, so Definition 4.2.3 gives

$$\frac{|f(x_0 + te_i) - f(x_0) - ta_i|}{|t|} \longrightarrow 0.$$

Equivalently,

$$\left| \frac{f(x_0 + te_i) - f(x_0)}{t} - a_i \right| \longrightarrow 0,$$

so the difference quotient tends to a_i. Thus the ith partial derivative exists and equals a_i. Repeating the argument for every i gives $a = \nabla f(x_0)$, and because the partial derivatives are defined independently of a, this also proves uniqueness.

For continuity, the Cauchy–Schwarz inequality gives

$$|a \cdot h| \leq \|a\|\|h\| \longrightarrow 0,$$

and

$$|r(h)| = \frac{|r(h)|}{\|h\|}\|h\| \longrightarrow 0.$$

Thus (4.2.1) gives $f(x_0 + h) \to f(x_0)$. □

Example 4.2.6 (A linear approximation in two variables)**.** Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x_1, x_2) = x_1^2 x_2$, and let $x_0 = (1, 2)^\top$. Expanding around x_0 gives

$$\begin{aligned} f(x_0 + h) &= (1 + h_1)^2(2 + h_2) \\ &= 2 + 4h_1 + h_2 + \big(2h_1^2 + 2h_1h_2 + h_1^2h_2\big). \end{aligned}$$

The constant is $f(x_0)$ and the linear term is $(4, 1)^\top \cdot h$. If $\|h\| \leq 1$, the absolute value of the remainder is at most

$$2|h_1|^2 + 2|h_1h_2| + |h_1|^2|h_2| \leq 5\|h\|^2.$$

After division by $\|h\|$, this bound tends to zero. Thus f is differentiable at $x_0 = (1, 2)^\top$ and

$$\nabla f(x_0) = \begin{pmatrix} 4 \\ 1 \end{pmatrix}.$$

The converse of the first claim in Theorem 4.2.5 is false: partial derivatives can exist without a valid linear approximation. A convenient condition rules out this problem.

Definition 4.2.7 (C^1 function)**.** Let $U \subseteq \mathbb{R}^n$ be open. A function $f : U \to \mathbb{R}$ is C^1 *on* U, or *continuously differentiable on* U, if every partial derivative $\partial f/\partial x_i$ exists on U and is continuous as a function from U to $\mathbb{R}$.

Theorem 4.2.8 (C^1 implies differentiable)**.** *If $f : U \to \mathbb{R}$ is C^1 on the open set $U \subseteq \mathbb{R}^n$, then f is differentiable on U and the gradient $\nabla f : U \to \mathbb{R}^n$ is continuous.*

We use this theorem without proof. It provides the usual way to verify differentiability: compute the partial derivatives and check that they are continuous. In particular, polynomials are C^1 on $\mathbb{R}^n$, and rational functions are C^1 wherever their denominators are nonzero.

The implications established in this section are

$$\begin{aligned} C^1 \quad &\Longrightarrow \quad \text{differentiable} \quad \Longrightarrow \quad \text{continuous}, \\ \text{differentiable} \quad &\Longrightarrow \quad \text{all partial derivatives exist}. \end{aligned}$$

None of the converses holds in general.

The linear approximation can also be written using differential notation.

Remark 4.2.9 (Differential notation). If f is differentiable at x_0, its *differential at* x_0 is the linear map

$$\mathrm{d}f_{x_0}(h) := \nabla f(x_0) \cdot h = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0) h_i.$$

Thus $\mathrm{d}f_{x_0}(h) = \nabla f(x_0) \cdot h$ is exactly the linear term in

$$f(x_0 + h) = f(x_0) + \nabla f(x_0) \cdot h + r(h).$$

Economics texts often write the same expression as

$$\mathrm{d}f = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0)\, \mathrm{d}x_i.$$

Here $\mathrm{d}x_i$ labels a coordinate change; it is not the product of a number d and x_i. We use the actual changes h_i when applying the differential.

4.2.3 Directional derivatives

A partial derivative permits only one coordinate to change. A directional derivative allows the coordinates to change together in fixed proportions: for $v = (v_1, \ldots, v_n)^\top$, varying t traces the line $x_0 + tv$ through x_0.

Definition 4.2.10 (Directional derivative)**.** Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}$, and let $x_0 \in U$. The *directional derivative* of f at x_0 in the direction $v \in \mathbb{R}^n$ is

$$\partial_v f(x_0) = \lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t},$$

provided the limit exists.

Taking $v = e_i$ gives the ith partial derivative. We do not require $\|v\| = 1$, so v specifies both the relative changes in the coordinates and their scale. When f is differentiable, its gradient gives every directional derivative.

Theorem 4.2.11 (Directional derivatives from the gradient)**.** *Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be differentiable at $x_0 \in U$. Then, for every $v \in \mathbb{R}^n$, the directional derivative exists and*

$$\partial_v f(x_0) = \nabla f(x_0) \cdot v = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0) v_i.$$

Proof (optional). If $v = 0$, the difference quotient is zero for every $t \neq 0$, so the result holds. If $v \neq 0$, substitute $h = tv$ into (4.2.1) and use $a = \nabla f(x_0)$:

$$\frac{f(x_0 + tv) - f(x_0)}{t} = \nabla f(x_0) \cdot v + \frac{r(tv)}{t}.$$

Moreover,

$$\left| \frac{r(tv)}{t} \right| = \|v\| \frac{|r(tv)|}{\|tv\|} \longrightarrow 0.$$

Taking the limit proves the first equality; the second writes out the inner product coordinate by coordinate. □

4.2.4 Steepest increase

To compare directions independently of scale, we restrict v to unit vectors and ask which one gives the largest directional derivative.

Theorem 4.2.12 (Steepest increase)**.** *Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}$ be differentiable at $x_0 \in U$, and suppose $\nabla f(x_0) \neq 0$. Define the unit vector*

$$v^* = \frac{\nabla f(x_0)}{\|\nabla f(x_0)\|}.$$

Among all unit vectors $v \in \mathbb{R}^n$, the directional derivative $\partial_v f(x_0)$ has its unique maximum at $v = v^*$ and its unique minimum at $v = -v^*$. The corresponding values are

$$\partial_{v^*} f(x_0) = \|\nabla f(x_0)\|, \qquad \partial_{-v^*} f(x_0) = -\|\nabla f(x_0)\|.$$

Proof (optional). Write $g = \nabla f(x_0)$. For every unit vector v, Theorem 4.2.11 and the Cauchy–Schwarz inequality give

$$\partial_v f(x_0) = g \cdot v \leq |g \cdot v| \leq \|g\|\|v\| = \|g\|.$$

The unit vector $v^* = g/\|g\|$ satisfies $g \cdot v^* = \|g\|$, so it attains the bound. If a unit vector v also attains it, then $g \cdot v = \|g\|$ and

$$\|g - \|g\|v\|^2 = 2\|g\|\big(\|g\| - g \cdot v\big) = 0.$$

Hence $v = g/\|g\| = v^*$. Applying the maximum result to $-f$ gives the minimum and its unique minimizer. □

Thus the gradient points in the direction of steepest increase, and its norm is the largest rate of increase per unit change. By Theorem 4.2.11, $\partial_v f(x_0) = 0$ exactly when $\nabla f(x_0) \cdot v = 0$. In this case, v and $\nabla f(x_0)$ are *orthogonal*. If $\nabla f(x_0) = 0$, every directional derivative is zero, so first-order information does not select a direction of increase or decrease.

This result underlies *gradient descent*, an algorithm used to train many machine-learning and AI models. If f is the loss function and x_k is the current parameter vector, gradient descent chooses a step size $\alpha_k > 0$ and updates

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$

The negative gradient gives the direction of steepest local decrease, while α_k determines the size of the step.

4.3 Vector-valued functions

4.3.1 Jacobians

A map $F : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ has m outputs. Its linear approximation is therefore a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$, represented by an $m \times n$ matrix.

Definition 4.3.1 (Jacobian)**.** Let $U \subseteq \mathbb{R}^n$ be open, let $F : U \to \mathbb{R}^m$, and let $x_0 \in U$. The map F is *differentiable at* x_0 if there is an $m \times n$ matrix A such that

$$\lim_{h \to 0} \frac{\|F(x_0 + h) - F(x_0) - Ah\|}{\|h\|} = 0,$$

where the limit is over $h \in \mathbb{R}^n$ with $h \neq 0$. The matrix A is the *derivative*, or *Jacobian*, of F at x_0, written $\mathrm{D}F(x_0)$. If $F = (F_1, \ldots, F_m)^\top$, the scalar-valued maps $F_i : U \to \mathbb{R}$ are the *component functions* of F. The map F is C^1 *on* U if every component F_i is C^1 on U.

Equivalently,

$$F(x_0 + h) = F(x_0) + \mathrm{D}F(x_0)h + r(h), \qquad \frac{\|r(h)\|}{\|h\|} \longrightarrow 0 \quad \text{as } h \to 0. \tag{4.3.1}$$

The linear term has dimensions

$$\mathrm{D}F(x_0)h : \qquad (m \times n)(n \times 1) = m \times 1,$$

so each row corresponds to an output and each column to an input.

With the moving input $x = x_0 + h$, the same definition is

$$\lim_{x \to x_0} \frac{\|F(x) - F(x_0) - \mathrm{D}F(x_0)(x - x_0)\|}{\|x - x_0\|} = 0. \tag{4.3.2}$$

Theorem 4.3.2 (Entries of the Jacobian)**.** *The map $F : U \to \mathbb{R}^m$ is differentiable at x_0 if and only if every component F_i is differentiable at x_0. In that case the Jacobian is unique and*

$$\mathrm{D}F(x_0) = \begin{pmatrix} \dfrac{\partial F_1}{\partial x_1}(x_0) & \cdots & \dfrac{\partial F_1}{\partial x_n}(x_0) \\ \vdots & \ddots & \vdots \\ \dfrac{\partial F_m}{\partial x_1}(x_0) & \cdots & \dfrac{\partial F_m}{\partial x_n}(x_0) \end{pmatrix}.$$

Thus the ith row of $\mathrm{D}F(x_0)$ is $\nabla F_i(x_0)^\top$.

Proof (optional). Suppose first that F is differentiable with derivative A, and let $r(h) = F(x_0 + h) - F(x_0) - Ah$. Since the absolute value of each coordinate of $r(h)$ is at most $\|r(h)\|$,

$$\frac{|F_i(x_0 + h) - F_i(x_0) - \sum_{j=1}^n A_{ij} h_j|}{\|h\|} \longrightarrow 0.$$

Thus F_i is differentiable, and Theorem 4.2.5 gives

$$(A_{i1}, \ldots, A_{in})^\top = \nabla F_i(x_0).$$

Conversely, suppose each F_i is differentiable, and define $A = (A_{ij})$ by

$$A_{ij} = \frac{\partial F_i}{\partial x_j}(x_0).$$

If $r_i(h)$ is the approximation error for component i, then $|r_i(h)|/\|h\| \to 0$. Because there are only finitely many components,

$$\frac{\|F(x_0+h) - F(x_0) - Ah\|}{\|h\|} = \left(\sum_{i=1}^{m} \left(\frac{r_i(h)}{\|h\|} \right)^2 \right)^{1/2} \longrightarrow 0.$$

Thus F is differentiable with derivative A. The first part identifies every entry of any possible derivative, so the Jacobian is unique. □

It follows from Theorems 4.2.8 and 4.3.2 that every C^1 map is differentiable.

Example 4.3.3 (A 2×3 Jacobian)**.** Let $F : \mathbb{R}^3 \to \mathbb{R}^2$ be given by

$$F(x, y, z) = \begin{pmatrix} x^2 y \\ y + 3z \end{pmatrix}.$$

Then

$$\mathrm{D}F(x, y, z) = \begin{pmatrix} 2xy & x^2 & 0 \\ 0 & 1 & 3 \end{pmatrix}, \qquad \mathrm{D}F(x_0) = \begin{pmatrix} 4 & 1 & 0 \\ 0 & 1 & 3 \end{pmatrix} \quad \text{at } x_0 = (1, 2, 0)^\top.$$

The two rows correspond to the two outputs, and the three columns correspond to the inputs x, y, z.

For a scalar-valued function $f : U \to \mathbb{R}$, the Jacobian is a $1 \times n$ row, whereas the gradient is an $n \times 1$ column:

$$\mathrm{D}f(x_0) = \nabla f(x_0)^\top. \tag{4.3.3}$$

Consequently, $\mathrm{D}f(x_0)h = \nabla f(x_0) \cdot h$. The gradient is convenient for geometric statements, while the Jacobian is convenient for composition. When $n = 1$, both contain the ordinary derivative $f'(x_0)$.

4.3.2 The multivariable chain rule

The one-variable chain rule says that derivatives multiply. The same rule holds for functions between Euclidean spaces, with matrix multiplication in place of scalar multiplication.

Theorem 4.3.4 (Chain rule)**.** *Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^m$ be open. Let $G : U \to \mathbb{R}^m$ satisfy $G(U) \subseteq V$, and let $F : V \to \mathbb{R}^p$. If G is differentiable at $x_0 \in U$ and F is differentiable at $G(x_0)$, then $F \circ G$ is differentiable at x_0 and*

$$\mathrm{D}(F \circ G)(x_0) = \mathrm{D}F(G(x_0))\, \mathrm{D}G(x_0).$$

We use this theorem without proof. Using the row description from Theorem 4.3.2, the product is

$$\mathrm{D}(F \circ G)(x_0) = \underbrace{\begin{pmatrix} \nabla F_1(G(x_0))^\top \\ \vdots \\ \nabla F_p(G(x_0))^\top \end{pmatrix}}_{p \times m} \underbrace{\begin{pmatrix} \nabla G_1(x_0)^\top \\ \vdots \\ \nabla G_m(x_0)^\top \end{pmatrix}}_{m \times n}.$$

The product is $p \times n$, the required size for the derivative of $F \circ G : U \to \mathbb{R}^p$. The outer rows are evaluated at $G(x_0)$, while the inner rows are evaluated at x_0. When $n = m = p = 1$, the formula reduces to the one-variable chain rule.

A useful special case follows a curve through the domain of a real-valued function.

Corollary 4.3.5 (Derivative along a curve)**.** *Let $V \subseteq \mathbb{R}^m$ be open, let $\gamma : I \to V$ be differentiable at $t_0 \in I$, where $I \subseteq \mathbb{R}$ is an open interval, and let $f : V \to \mathbb{R}$ be differentiable at $\gamma(t_0)$. Then*

$$(f \circ \gamma)'(t_0) = \nabla f(\gamma(t_0)) \cdot \gamma'(t_0) = \sum_{i=1}^{m} \frac{\partial f}{\partial x_i}(\gamma(t_0))\gamma_i'(t_0).$$

Indeed, (4.3.3) and the chain rule give $\mathrm{D}f(\gamma(t_0))\mathrm{D}\gamma(t_0) = \nabla f(\gamma(t_0))^\top \gamma'(t_0)$.

4.3.3 A worked chain rule

Example 4.3.6 (A vector-valued composition)**.** Define

$$G : \mathbb{R} \to \mathbb{R}^3, \qquad G(t) = \begin{pmatrix} 1+t \\ 2 \\ 3+2t \end{pmatrix}, \qquad F : \mathbb{R}^3 \to \mathbb{R}^2, \qquad F(u,v,w) = \begin{pmatrix} uw \\ v+w^2 \end{pmatrix}.$$

To find the derivative of $F \circ G$ at $t = 0$, compute

$$\mathrm{D}G(t) = G'(t) = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \qquad \mathrm{D}F(u,v,w) = \begin{pmatrix} w & 0 & u \\ 0 & 1 & 2w \end{pmatrix}.$$

Because $G(0) = (1,2,3)^\top$, the chain rule gives

$$\begin{aligned} \mathrm{D}(F \circ G)(0) &= \mathrm{D}F(G(0))\,\mathrm{D}G(0) \\ &= \begin{pmatrix} 3 & 0 & 1 \\ 0 & 1 & 6 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 5 \\ 12 \end{pmatrix}. \end{aligned}$$

The dimensions are $(2 \times 3)(3 \times 1) = 2 \times 1$, as required for a map from $\mathbb{R}$ to $\mathbb{R}^2$.

For a direct check,

$$(F \circ G)(t) = \begin{pmatrix} (1+t)(3+2t) \\ 2 + (3+2t)^2 \end{pmatrix} = \begin{pmatrix} 3 + 5t + 2t^2 \\ 11 + 12t + 4t^2 \end{pmatrix}.$$

Differentiating the components at 0 again gives $(5,12)^\top$.

Lecture 5

Curvature, Convexity, and Separation

Lecture 4 used a linear map to approximate a differentiable function near a point. In the one-variable case, we saw that

$$f(x_0 + h) \approx f(x_0) + f'(x_0)h.$$

In this lecture we are interested in the curvature of a function, which is described by a second-order approximation. In the one-variable case, if f is twice continuously differentiable near x_0, the second-order Taylor approximation adds a quadratic term:

$$f(x_0 + h) \approx f(x_0) + f'(x_0)h + \frac{1}{2}f''(x_0)h^2.$$

When $f'(x_0) = 0$, the linear term vanishes, so first-order information does not tell us whether the graph bends upward or downward. The sign of $f''(x_0)$ does. In several variables, the Hessian H_f generalizes $f''(x_0)$, and the quadratic term becomes $\frac{1}{2}h^\top H_f(x_0)h$.

5.1 Second-order approximation and curvature

5.1.1 Hessians and local quadratic models

Definition 5.1.1 (C^2 function)**.** Let $U \subseteq \mathbb{R}^n$ be open. A function $f : U \to \mathbb{R}$ is C^2 *on* U, or *twice continuously differentiable on* U, if each first partial derivative $\partial f/\partial x_i$ is C^1 on U. Equivalently, all second partial derivatives of f exist and are continuous on U.

Definition 5.1.2 (Hessian)**.** Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be differentiable. If the gradient $\nabla f : U \to \mathbb{R}^n$ is differentiable at $x \in U$, the *Hessian* of f at x is the Jacobian of its gradient,

$$H_f(x) = \mathrm{D}(\nabla f)(x).$$

It is the $n \times n$ matrix whose (i, j) entry is

$$\big(H_f(x)\big)_{ij} = \frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right)(x).$$

Example 5.1.3 (Computing a Hessian)**.** Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by

$$f(x_1, x_2) = x_1^2 x_2.$$

In Lecture 4 we computed

$$\nabla f(x) = \begin{pmatrix} 2x_1x_2 \\ x_1^2 \end{pmatrix}.$$

The Hessian is the Jacobian of the gradient, so

$$H_f(x) = \begin{pmatrix} 2x_2 & 2x_1 \\ 2x_1 & 0 \end{pmatrix}.$$

Theorem 5.1.4 (Equality of mixed partials)**.** *Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be C^2. Then*

$$\frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right)(x) = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right)(x)$$

for every $x \in U$ and every i, j. Thus $H_f(x)$ is symmetric.

We use this result without proof. Importantly, it requires continuity of the second partial derivatives.

Theorem 5.1.5 (Second-order Taylor expansion)**.** *Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}$ be C^2, and let $x_0 \in U$. For changes h such that $x_0 + th \in U$ for every $t \in [0, 1]$,*

$$f(x_0 + h) = f(x_0) + \nabla f(x_0) \cdot h + \frac{1}{2} h^\top H_f(x_0) h + r(h), \qquad \frac{|r(h)|}{\|h\|^2} \longrightarrow 0 \quad \text{as } h \to 0. \tag{5.1.1}$$

We use this result without proof. It has exactly the form of the linear approximation in Lecture 4, with one additional term. The expression

$$f(x_0) + \nabla f(x_0) \cdot h + \frac{1}{2} h^\top H_f(x_0) h$$

is the *quadratic approximation* to $f(x_0 + h)$. The first-order approximation stops after the linear term and has a remainder negligible relative to $\|h\|$. The second-order approximation keeps the quadratic term and has a remainder negligible relative to $\|h\|^2$.

5.1.2 Quadratic forms and their shapes

The quadratic term of the Taylor approximation has the form $x^\top Ax$. We next study expressions of this form, called quadratic forms. They are essentially quadratic polynomials in several variables. Like quadratic functions in one dimension, quadratic forms have a constant Hessian. That allows us to characterize all possible types of curvature in terms of the sign of the quadratic form.

Definition 5.1.6 (Quadratic form)**.** Let A be a symmetric $n \times n$ matrix. The function

$$Q_A(x) = x^\top Ax, \qquad x \in \mathbb{R}^n,$$

is the *quadratic form* associated with A.

The restriction to symmetric matrices is without loss of generality. That is, if a quadratic form is associated with a nonsymmetric matrix B, then the same quadratic form is associated with a symmetric matrix $\widetilde{B} = (B + B^\top)/2$.

Next, the Hessian of $Q_A = x^\top Ax$ equals $2A$. To see this, observe that

$$Q_A(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j,$$

so, using the symmetry $a_{ij} = a_{ji}$, the kth component of the gradient and the (k, l) second partial are

$$\frac{\partial Q_A}{\partial x_k}(x) = 2 \sum_{j=1}^{n} a_{kj} x_j, \qquad \frac{\partial}{\partial x_l} \left(\frac{\partial Q_A}{\partial x_k} \right)(x) = 2a_{kl}.$$

In matrix form, $\nabla Q_A(x) = 2Ax$ and

$$H_{Q_A}(x) = 2A,$$

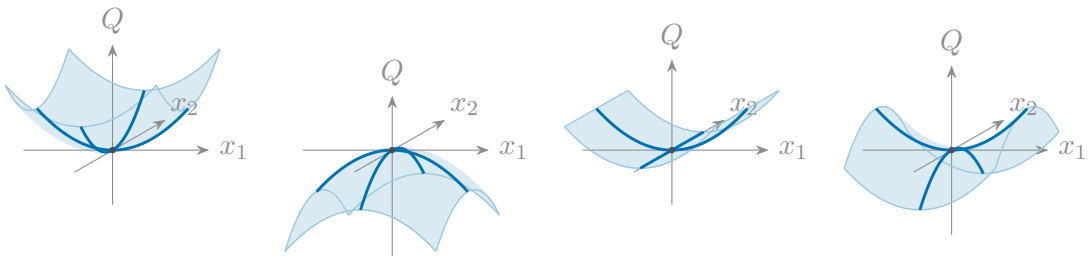
which indeed does not depend on x.

The sign of $x^\top Ax$ (which is a scalar) describes the possible shapes of the curvature of the quadratic form in $\mathbb{R}^n$.

Definition 5.1.7 (Definiteness)**.** Let A be a symmetric $n \times n$ matrix.

$$\begin{aligned}
A \text{ is } \textit{positive definite} &\iff x^\top Ax > 0 && \text{for every } x \neq 0, \\
A \text{ is } \textit{positive semidefinite} &\iff x^\top Ax \geq 0 && \text{for every } x, \\
A \text{ is } \textit{negative definite} &\iff x^\top Ax < 0 && \text{for every } x \neq 0, \\
A \text{ is } \textit{negative semidefinite} &\iff x^\top Ax \leq 0 && \text{for every } x.
\end{aligned}$$

The matrix is *indefinite* if its quadratic form takes both positive and negative values.



(a) positive definite: $x_1^2 + x_2^2$ (b) negative definite: $-x_1^2 - x_2^2$ (c) positive semidefinite: x_1^2 (d) indefinite: $x_1^2 - x_2^2$

Figure 5.1.1: Four representative quadratic forms in two variables. The light blue surface is the graph of Q over a square around the origin; the two thick curves show the values of Q along the x_1-axis and along the x_2-axis.

To read the pictures, fix a direction $v \neq 0$ and move through the origin along the line tv. Along this line, the quadratic form becomes the one-variable function $\phi_v(t) = Q_A(tv) = t^2\, v^\top Av$, a parabola through the origin with second derivative

$$\phi_v''(t) = 2v^\top Av.$$

The sign of $v^\top Av$ therefore tells us whether this slice bends upward, bends downward, or is flat. Thus definiteness describes every one-dimensional slice through the origin. Positive definite means that every nonzero-direction slice bends upward, so the origin is the unique global minimum; negative definite means that every such slice bends downward, so the origin is the unique global maximum. A positive semidefinite form may be flat in some directions, while an indefinite form bends upward in some directions and downward in others, so the origin is neither a maximum nor a minimum. These are exactly the shapes in Figure 5.1.1.

5.1.3 Computational tests for definiteness

Checking definiteness of a matrix A using the definition in Definition 5.1.7 requires evaluating $x^\top Ax$ for every nonzero $x \in \mathbb{R}^n$; that is not very practical.

Instead, this section introduces two types of tests that are easier to apply. The first uses eigenvalues, and the second uses determinants of submatrices.

For a diagonal matrix

$$D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix},$$

the check is immediate:

$$x^\top Dx = \lambda_1 x_1^2 + \lambda_2 x_2^2 + \cdots + \lambda_n x_n^2,$$

a sum of squares weighted by the diagonal entries. If every $\lambda_i > 0$, the sum is positive for every $x \neq 0$, so D is positive definite; if every $\lambda_i \geq 0$, the sum is nonnegative, so D is positive semidefinite; the negative cases flip the inequalities. If some $\lambda_i > 0 > \lambda_j$, then $e_i^\top De_i = \lambda_i > 0$ and $e_j^\top De_j = \lambda_j < 0$, so D is indefinite. The definiteness of a diagonal matrix is thus decided by the signs of its diagonal entries.

Recall the diagonalization theorem of Lecture 3: a matrix A with n independent eigenvectors can be written as $A = PDP^{-1}$, where the columns of P are the eigenvectors and D is the diagonal matrix of the corresponding eigenvalues; we also noted, without proof, that symmetric matrices always have n independent eigenvectors. For symmetric matrices slightly more is true: the eigenvectors can be chosen so that $P^{-1} = P^\top$ — such a P is called *orthogonal* — and the factorization becomes $A = PDP^\top$. This refinement is the *spectral theorem*, which we again use without proof. It reduces any symmetric matrix to the diagonal case: changing variables to $z = P^\top x$ turns $x^\top Ax$ into $z^\top Dz$, and the diagonal computation above applies. This gives the first test.

Theorem 5.1.8 (Eigenvalue test)**.** *Let A be a symmetric $n \times n$ matrix. Then*

$$\begin{aligned} A \text{ is positive definite} &\iff \text{every eigenvalue of } A \text{ is positive}, \\ A \text{ is positive semidefinite} &\iff \text{every eigenvalue of } A \text{ is nonnegative}, \\ A \text{ is negative definite} &\iff \text{every eigenvalue of } A \text{ is negative}, \\ A \text{ is negative semidefinite} &\iff \text{every eigenvalue of } A \text{ is nonpositive}. \end{aligned}$$

The matrix is indefinite if and only if it has eigenvalues of both signs.

Proof (optional). By the spectral theorem, write $A = PDP^\top$ with $D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ and P orthogonal, and put $z = P^\top x$. Then

$$x^\top A x = x^\top P D P^\top x = z^\top D z = \sum_{i=1}^{n} \lambda_i z_i^2.$$

Since P is orthogonal, $x = Pz$, so $x = 0$ if and only if $z = 0$, and as x ranges over $\mathbb{R}^n$ so does z. The quantity $x^\top A x$ therefore has the same range of signs as $\sum_i \lambda_i z_i^2$ over all z. That sum is positive for every $z \neq 0$ exactly when every $\lambda_i > 0$: the "if" direction is immediate, and taking $z = e_i$ gives λ_i itself, which forces the "only if" direction. The other three cases are the same argument with the inequality changed, and the indefinite case follows by taking $z = e_i$ for eigenvalues of each sign. □

The second test requires no eigenvalues. Instead, it looks at the signs of the determinants of certain submatrices of A.

Definition 5.1.9 (Principal minors)**.** Let A be an $n \times n$ matrix. For a nonempty index set $I \subseteq \{1, \dots, n\}$, let A_I be the square matrix obtained by retaining the rows and columns whose indices belong to I. The determinant $\det A_I$ is a *principal minor* of order k when I contains k indices. The minors

$$\Delta_k = \det A_{\{1,\dots,k\}}, \qquad k = 1, \dots, n,$$

are the *leading principal minors* of A.

In general, a *minor* of A is the determinant of a submatrix obtained by retaining some rows and some columns; a minor is *principal* when the retained row and column index sets are the same set I. For example, for a 3×3 matrix $A = (a_{ij})$, the index set $I = \{1, 3\}$ retains rows $1, 3$ and columns $1, 3$:

$$A_{\{1,3\}} = \begin{pmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{pmatrix},$$

and $\det A_{\{1,3\}}$ is a principal minor of order 2. Counting nonempty index sets, A has $2^3 - 1 = 7$ principal minors. The three leading ones come from the sets $\{1\}$, $\{1, 2\}$, and $\{1, 2, 3\}$, so they grow from the top-left corner:

$$\Delta_1 = a_{11}, \qquad \Delta_2 = \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \qquad \Delta_3 = \det A.$$

Definiteness can be read off the signs of these minors.

Theorem 5.1.10 (Principal-minor tests)**.** *Let A be a symmetric $n \times n$ matrix, and let $\Delta_1, \dots, \Delta_n$ be its leading principal minors.*

1. *A is positive definite if and only if $\Delta_k > 0$ for every $k = 1, \dots, n$.*
2. *A is negative definite if and only if $(-1)^k \Delta_k > 0$ for every $k = 1, \dots, n$.*
3. *A is positive semidefinite if and only if every principal minor of A is nonnegative.*

4. A is negative semidefinite if and only if every principal minor of order k has sign satisfying

$$(-1)^k \det A_I \geq 0.$$

We use this result without proof. The first two parts are commonly called *Sylvester's criterion*. Notice the difference between the definite and semidefinite cases: definite matrices require only the n leading principal minors, while semidefinite matrices require all $2^n - 1$ nonempty principal minors.

Example 5.1.11 (The 2×2 tests)**.** Let

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

Its leading principal minors are a and $ac - b^2$, while all of its principal minors are a, c, and $ac - b^2$. Therefore

$$\begin{aligned} A \text{ is positive definite} &\iff a > 0 \text{ and } ac - b^2 > 0, \\ A \text{ is negative definite} &\iff a < 0 \text{ and } ac - b^2 > 0, \\ A \text{ is positive semidefinite} &\iff a \geq 0,\ c \geq 0, \text{ and } ac - b^2 \geq 0, \\ A \text{ is negative semidefinite} &\iff a \leq 0,\ c \leq 0, \text{ and } ac - b^2 \geq 0. \end{aligned}$$

For the positive-definite case, the first line can also be seen directly by completing the square (possible whenever $a \neq 0$):

$$x^\top A x = a\left(x_1 + \frac{b}{a} x_2\right)^2 + \frac{ac - b^2}{a}\, x_2^2.$$

If $a > 0$ and $ac - b^2 > 0$, both coefficients are positive. Conversely, if A is positive definite, evaluating at $x = e_1$ gives $a > 0$, and evaluating at $x = (-b/a, 1)^\top$ gives $(ac - b^2)/a > 0$.

The next example shows why the semidefinite tests need more than just the leading minors.

Example 5.1.12 (Weak leading-minor inequalities are insufficient)**.** Let $A = \operatorname{diag}(0, -1)$. Its leading minors satisfy $\Delta_1 = \Delta_2 = 0$, but the principal minor indexed by $I = \{2\}$ is -1. Thus weak inequalities on the leading minors alone do not establish positive semidefiniteness. The missing principal minor detects the negative direction e_2, along which $e_2^\top A e_2 = -1$.

Example 5.1.13 (A three-variable quadratic form)**.** Let $Q : \mathbb{R}^3 \to \mathbb{R}$ be given by

$$Q(x_1, x_2, x_3) = -x_1^2 - x_2^2 - x_3^2 + x_1 x_2 + x_2 x_3.$$

Its Hessian is

$$H_Q = \begin{pmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix},$$

whose leading principal minors are

$$\Delta_1 = -2, \qquad \Delta_2 = 3, \qquad \Delta_3 = -4.$$

Thus $(-1)^k \Delta_k$ equals 2, 3, and 4, respectively, so by Theorem 5.1.10 H_Q is negative definite. This establishes negative definiteness without computing any eigenvalues.

The tests apply equally when the Hessian entries are not constant.

Example 5.1.14 (A varying Hessian)**.** Let $f : \mathbb{R}^2_{++} \to \mathbb{R}$ be given by $f(x_1, x_2) = \log x_1 + \log x_2$. Differentiating twice,

$$H_f(x) = \begin{pmatrix} -1/x_1^2 & 0 \\ 0 & -1/x_2^2 \end{pmatrix},$$

which is diagonal with negative entries and hence negative definite at every x, by Theorem 5.1.8. The Hessian varies with x, but its definiteness does not.

In the last example of this section, definiteness depends on a parameter.

Example 5.1.15 (An interaction term)**.** Fix $b \in \mathbb{R}^2$ and $\gamma \in \mathbb{R}$, and define $q : \mathbb{R}^2 \to \mathbb{R}$ by

$$q(x) = b \cdot x - \left(x_1^2 + \gamma x_1 x_2 + x_2^2\right).$$

Here H_q is constant. Direct differentiation gives

$$H_q = \begin{pmatrix} -2 & -\gamma \\ -\gamma & -2 \end{pmatrix}.$$

The leading principal minors are -2 and $4 - \gamma^2$. Therefore Theorem 5.1.10 shows that H_q is negative definite when $|\gamma| < 2$. When $|\gamma| = 2$, its two order-one principal minors are negative and its determinant is zero, so it is negative semidefinite. When $|\gamma| > 2$, its determinant is negative, so it is indefinite: symmetry makes both eigenvalues real, and their negative product forces them to have opposite signs. The interaction term can therefore preserve, flatten, or overturn the downward curvature of the two squared terms.

At this point two exact tests are available: definiteness can be read off the signs of the eigenvalues (Theorem 5.1.8) or off the signs of the principal minors (Theorem 5.1.10). Often a full classification is unnecessary, and a cheaper observation settles the question. Table 5.1.1 collects these quick checks for the negative case; reversing every sign gives the positive case. A necessary condition can rule the property out but never establishes it.

Check	Verdict
$a_{ii} \le 0$ for every i	necessary for negative semidefiniteness, not sufficient
$a_{ii} < 0$ for every i	necessary for negative definiteness, not sufficient
$(-1)^k \Delta_k \ge 0$ for every k	necessary for negative semidefiniteness, not sufficient
$(-1)^k \Delta_k > 0$ for every k	equivalent to negative definiteness
some x has $x^\top A x > 0$	counterexample to negative semidefiniteness
some $x \neq 0$ has $x^\top A x = 0$	counterexample to negative definiteness
negative semidefinite and $\det A \neq 0$	sufficient for negative definiteness

Table 5.1.1: Quick checks for a symmetric matrix A with entries a_{ij} and leading principal minors Δ_k.

The diagonal checks follow by evaluating the quadratic form at the coordinate vectors, since $e_i^\top A e_i = a_{ii}$. The reverse is not true in the first three rows: for example, $\left(\begin{smallmatrix} -1 & 2 \\ 2 & -1 \end{smallmatrix}\right)$ passes both diagonal checks yet is indefinite. The last row follows from the eigenvalue test: the eigenvalues are nonpositive and their product $\det A$ is nonzero, so all of them are negative.

5.2 Convex sets and concave functions

5.2.1 Convex sets and preservation

The Hessian and Taylor expansion describe a function near one point. Turning their curvature information into a statement about the function everywhere requires moving between arbitrary points along the segment joining them. The domain must contain that segment.

Definition 5.2.1 (Convex combination and convex set)**.** Let $x, y \in \mathbb{R}^n$.

- A *convex combination* of x and y is a point

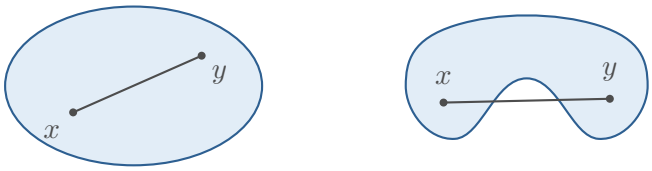
$$(1-t)x + ty, \qquad t \in [0,1],$$

  and the set of all of them is the *line segment* joining x and y.

- A set $C \subseteq \mathbb{R}^n$ is *convex* if

$$(1-t)x + ty \in C \qquad \text{for every } x, y \in C \text{ and every } t \in [0,1].$$

At $t = 0$ the combination is x and at $t = 1$ it is y, so a set is convex exactly when the segment joining any two of its points stays inside it. For example, $\mathbb{R}^n$ itself, every affine subspace, every halfspace $\{x : a \cdot x \leq c\}$, and the orthants $\mathbb{R}^n_+$ and $\mathbb{R}^n_{++}$ are convex. The unit circle $\{x \in \mathbb{R}^2 : \|x\| = 1\}$ is not: it contains e_1 and $-e_1$ but not their midpoint 0. Figure 5.2.1 shows the definition in $\mathbb{R}^2$.



(a) convex (b) not convex

Figure 5.2.1: The set on the left is convex: for any two points in the set, the segment joining them is inside the set. The right is not convex: the segment joining the marked points x and y is not entirely in the set.

Rather than verify Definition 5.2.1 for each new set, it is usually quicker to build the set out of convex pieces. An *affine map* $T : \mathbb{R}^n \to \mathbb{R}^m$ has the form $T(x) = Ax + b$ for an $m \times n$ matrix A and a vector $b \in \mathbb{R}^m$.

Theorem 5.2.2 (Preservation rules)**.** *The following constructions preserve convexity.*

1. *The intersection of any collection of convex sets is convex.*
2. *The Cartesian product of convex sets is convex.*
3. *The image of a convex set under an affine map is convex.*
4. *The inverse image of a convex set under an affine map is convex.*

Proof (optional). Intersections. Let $C = \bigcap_\alpha C_\alpha$ with each C_α convex, and let $x, y \in C$ and $t \in [0,1]$. For each α both x and y lie in C_α, so Definition 5.2.1 puts $(1-t)x + ty$ in C_α. Being in every C_α, it lies in C.

Products. Convex combinations in a product are taken coordinate block by coordinate block, so the claim follows by applying Definition 5.2.1 in each factor.

Affine maps. Let $T(x) = Ax + b$. Expanding, and recombining the two copies of b as $(1-t)b + tb$,

$$T\big((1-t)x + ty\big) = A\big((1-t)x + ty\big) + b = (1-t)T(x) + tT(y), \tag{5.2.1}$$

so T carries the segment joining x and y onto the segment joining $T(x)$ and $T(y)$. If C is convex, (5.2.1) exhibits every point between $T(x)$ and $T(y)$ as the image of a point of C, so $T(C)$ is convex. If D is convex and $x, y \in T^{-1}(D)$, then (5.2.1) puts $T((1-t)x + ty)$ between $T(x)$ and $T(y)$, hence in D. □

A useful consequence of the preservation rules: any finite list of linear equalities and inequalities defines a convex set. The example below shows why.

Example 5.2.3 (Linear restrictions define a convex set)**.** Let A be $m \times n$, let B be $p \times n$, and let $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^p$. The set

$$C = \{x \in \mathbb{R}^n : Ax \le b,\ Bx = c\}$$

is convex, where the inequality is coordinatewise. Each row of these restrictions has the form $a \cdot x = d$ or $a \cdot x \le d$ for some $a \in \mathbb{R}^n$ and $d \in \mathbb{R}$. The set each restriction defines is the inverse image of the convex set $\{d\}$ or $(-\infty, d]$ under the affine map $x \mapsto a \cdot x$. Part (4) of Theorem 5.2.2 makes each restriction convex, and part (1) makes their intersection C convex.

5.2.2 Concavity and superlevel sets

Convexity of a set says that its line segments stay in the domain. Concavity then compares the height of a function along those same segments.

Definition 5.2.4 (Concave and convex functions)**.** Let $C \subseteq \mathbb{R}^n$ be convex and let $f : C \to \mathbb{R}$.

- f is *concave* if

$$f\big((1-t)x + ty\big) \ge (1-t)f(x) + tf(y) \qquad \text{for every } x, y \in C \text{ and every } t \in [0,1].$$

- f is *strictly concave* if that inequality is strict whenever $x \ne y$ and $t \in (0,1)$.
- f is *convex* if $-f$ is concave, and *strictly convex* if $-f$ is strictly concave.

The right-hand side is the height at $(1-t)x + ty$ of the chord joining the two points $(x, f(x))$ and $(y, f(y))$ on the graph, so a concave function lies on or above each of its chords. This is the global version of bending downward: it compares any two points, not only nearby ones. Strict concavity additionally forbids the graph from containing a line segment. If $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$, the affine function $f(x) = a \cdot x + b$ satisfies Definition 5.2.4 with equality, by the computation (5.2.1). Thus it is both concave and convex, and neither strictly.

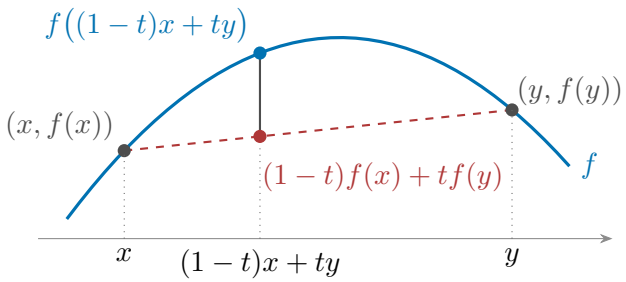


Figure 5.2.2: The chord inequality of Definition 5.2.4. At each point $(1-t)x+ty$ of the segment joining x and y, the graph of a concave function lies on or above the chord joining $(x, f(x))$ and $(y, f(y))$. Here f is strictly concave, but establishing that requires checking the inequality for every pair of points $x \neq y$.

Concavity is a statement about the graph. It also imposes structure on the sets of points at which the function is at least as large as a given level.

Definition 5.2.5 (Superlevel set)**.** Let $C \subseteq \mathbb{R}^n$, let $f : C \to \mathbb{R}$, and let $\alpha \in \mathbb{R}$. The *superlevel set* of f at level α is

$$S_\alpha(f) = \{x \in C : f(x) \geq \alpha\}.$$

Theorem 5.2.6 (Superlevel sets of a concave function)**.** *If C is convex and $f : C \to \mathbb{R}$ is concave, then $S_\alpha(f)$ is convex for every $\alpha \in \mathbb{R}$.*

Proof. Let $x, y \in S_\alpha(f)$ and $t \in [0,1]$, and put $z = (1-t)x + ty$. Convexity of C puts z in C by Definition 5.2.1, and Definition 5.2.4 gives

$$f(z) \geq (1-t)f(x) + tf(y) \geq (1-t)\alpha + t\alpha = \alpha,$$

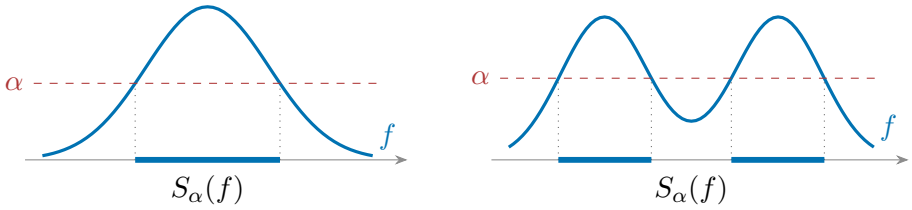
the second inequality because $f(x) \geq \alpha$ and $f(y) \geq \alpha$ by Definition 5.2.5. Hence $z \in S_\alpha(f)$. □

5.2.3 Quasiconcavity

The proof of Theorem 5.2.6 used concavity only to keep $f(z)$ above α, never to keep it above the weighted average. Demanding just the conclusion gives a weaker property, and it is the one that many results actually need.

Definition 5.2.7 (Quasiconcavity)**.** Let $C \subseteq \mathbb{R}^n$ be convex. A function $f : C \to \mathbb{R}$ is *quasiconcave* if $S_\alpha(f)$ is convex for every $\alpha \in \mathbb{R}$.

In one variable, the convex sets are the intervals, so quasiconcavity requires each superlevel set to be an interval. Figure 5.2.3 shows a single-peaked function that passes this test and a two-peaked function that fails it.



(a) quasiconcave (b) not quasiconcave

Figure 5.2.3: Superlevel sets in one variable. The single-peaked function on the left is quasiconcave: at each level α, the superlevel set $S_\alpha(f)$ is an interval. The two-peaked function on the right is not: at the level shown, its superlevel set is a union of two disjoint intervals, which is not convex.

Quasiconcavity can also be stated as an inequality along segments, parallel to Definition 5.2.4 but with the minimum in place of the weighted average.

Theorem 5.2.8 (Minimum characterization)**.** *Let $C \subseteq \mathbb{R}^n$ be convex. A function $f : C \to \mathbb{R}$ is quasiconcave if and only if*

$$f\big((1-t)x + ty\big) \geq \min\{f(x), f(y)\} \qquad \text{for every } x, y \in C \text{ and every } t \in [0,1].$$

Proof. Convex superlevel sets imply the inequality. Let $x, y \in C$ and $t \in [0,1]$, and put $\alpha = \min\{f(x), f(y)\}$. Then $x, y \in S_\alpha(f)$ by Definition 5.2.5, and that set is convex by Definition 5.2.7, so $(1-t)x + ty \in S_\alpha(f)$, which is the inequality.

The inequality implies convex superlevel sets. Fix α and let $x, y \in S_\alpha(f)$ and $t \in [0,1]$. Then $(1-t)x + ty \in C$ by Definition 5.2.1, and

$$f\big((1-t)x + ty\big) \geq \min\{f(x), f(y)\} \geq \alpha,$$

so the point lies in $S_\alpha(f)$. □

Concavity implies quasiconcavity, since a weighted average of $f(x)$ and $f(y)$ is at least their minimum, so Definition 5.2.4 gives the inequality of Theorem 5.2.8. The converse fails.

Example 5.2.9 (Quasiconcave but not concave)**.** Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = x^3$. Since f is increasing, each superlevel set is the interval $[\alpha^{1/3}, \infty)$, which is convex, so f is quasiconcave by Definition 5.2.7. It is not concave: taking the midpoint of 0 and 2,

$$f(1) = 1 < 4 = \tfrac{1}{2}f(0) + \tfrac{1}{2}f(2),$$

which violates Definition 5.2.4.

What made Example 5.2.9 work is that x^3 is an increasing relabelling of the values of x, and quasiconcavity, unlike concavity, survives any such relabelling.

Theorem 5.2.10 (Increasing transformations)**.** *Let $C \subseteq \mathbb{R}^n$ be convex, let $f : C \to \mathbb{R}$ be quasiconcave, and let $\varphi : \mathbb{R} \to \mathbb{R}$ be nondecreasing. Then $\varphi \circ f$ is quasiconcave.*

Proof. Let $x, y \in C$ and $t \in [0,1]$. Applying φ to the inequality of Theorem 5.2.8 preserves it, because φ is nondecreasing, so

$$\varphi\Big(f\big((1-t)x+ty\big)\Big) \geq \varphi\big(\min\{f(x),f(y)\}\big) = \min\big\{\varphi(f(x)),\varphi(f(y))\big\},$$

the equality because a nondecreasing φ sends the smaller of two numbers to the smaller of their images. By Theorem 5.2.8 again, $\varphi \circ f$ is quasiconcave. □

5.3 Concavity criteria, global optimality, and separation

5.3.1 Restriction to line segments

To decide whether a function of n variables is concave, we reduce the question to one variable. The reduction rests on three results: concavity of f is equivalent to concavity of its restriction to every line segment of the domain (Lemma 5.3.1); such a restriction has second derivative $v^{\top} H_f(x+tv)v$ (Theorem 5.3.2); and a one-variable function is concave exactly when its second derivative is nonpositive (Lemma 5.3.3).

Lemma 5.3.1 (Line-segment restriction)**.** *Let $U \subseteq \mathbb{R}^n$ be convex and let $f : U \to \mathbb{R}$. For $x, y \in U$ let $v = y - x$ and*

$$\phi(t) = f(x+tv), \qquad t \in [0,1].$$

Then f is concave if and only if ϕ is concave on $[0,1]$ for every choice of $x, y \in U$, and f is strictly concave if and only if ϕ is strictly concave for every choice of $x, y \in U$ with $x \neq y$.

Proof (optional). Concavity of f implies concavity of every ϕ. Fix $x, y \in U$ and let $s, s' \in [0,1]$ and $t \in [0,1]$. The point $x + \big((1-t)s + ts'\big)v$ is the convex combination $(1-t)(x+sv) + t(x+s'v)$ of two points of U, so Definition 5.2.4 gives

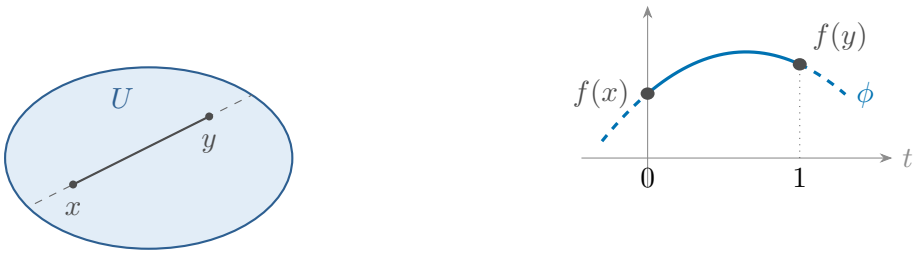
$$\phi\big((1-t)s+ts'\big) = f\big((1-t)(x+sv)+t(x+s'v)\big) \geq (1-t)\phi(s) + t\phi(s').$$

Concavity of every ϕ implies concavity of f. Given $x, y \in U$ and $t \in [0,1]$, take $s = 0$ and $s' = 1$ above. Since $\phi(0) = f(x)$ and $\phi(1) = f(y)$,

$$f\big((1-t)x+ty\big) = \phi(t) \geq (1-t)f(x) + tf(y).$$

For the strict statements, note that $x \neq y$ makes $v \neq 0$, so distinct values of s give distinct points $x + sv$; both displays above then hold with strict inequalities under the corresponding strict hypothesis. □

For a function of one variable, the second derivative tells us whether the graph bends upward or downward. For a function of several variables, the input can move in many directions, and the curvature may depend on the direction. We therefore restrict the function to the line $x + tv$ and study it as a function of the single variable t. Figure 5.3.1 shows the construction.



(a) the segment $x + tv$ in the domain

(b) the restriction $\phi(t) = f(x + tv)$

Figure 5.3.1: Restricting f to a line segment. The segment joining x and y in the n-dimensional domain $U \subseteq \mathbb{R}^n$ (left) is parameterized as $x+tv$ with $v = y-x$; along it, the n-variable function f collapses to the function $\phi(t) = f(x+tv)$ of the single real variable t (right), with $\phi(0) = f(x)$ and $\phi(1) = f(y)$.

Theorem 5.3.2 (Directional curvature)**.** *Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}$ be C^2, let $x \in U$, and let $v \in \mathbb{R}^n$. Put $\phi(t) = f(x + tv)$ for those t with $x + tv \in U$. Then*

$$\phi'(t) = \nabla f(x + tv) \cdot v, \qquad \phi''(t) = v^{\top} H_f(x + tv)\, v.$$

Proof (optional). Define the curve $\gamma : \mathbb{R} \to \mathbb{R}^n$ by

$$\gamma(t) = x + tv.$$

Its ith coordinate is the affine function $\gamma_i(t) = x_i + tv_i$, whose derivative is $\gamma_i'(t) = v_i$. Therefore $\gamma'(t) = v$.

Fix any t such that $\gamma(t) \in U$. Because U is open, there is $\varepsilon > 0$ such that $B(\gamma(t), \varepsilon) \subseteq U$. If $v \neq 0$, then

$$\|\gamma(s) - \gamma(t)\| = |s - t|\|v\| < \varepsilon$$

whenever $|s - t| < \varepsilon/\|v\|$. If $v = 0$, then $\gamma(s) = \gamma(t)$ for every s. In either case, $f \circ \gamma$ is defined on an open interval containing t. Since $\phi = f \circ \gamma$, the derivative-along-a-curve formula from Lecture 4 gives

$$\phi'(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\gamma(t))\gamma_i'(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x + tv)v_i = \nabla f(x + tv) \cdot v.$$

For each i, define $g_i : U \to \mathbb{R}$ by

$$g_i(y) = \frac{\partial f}{\partial x_i}(y).$$

Because f is C^2, each g_i is C^1 and hence differentiable. Applying the same derivative-along-a-curve formula to $g_i \circ \gamma$ gives

$$\begin{aligned}(g_i \circ \gamma)'(t) &= \sum_{j=1}^{n} \frac{\partial g_i}{\partial x_j}(\gamma(t))\gamma_j'(t) \\ &= \sum_{j=1}^{n} \big(H_f(x + tv)\big)_{ij} v_j,\end{aligned}$$

where the second equality follows from the definition of the Hessian. For every s in the open interval above, the formula for the first derivative can be written as

$$\phi'(s) = \sum_{i=1}^{n} v_i (g_i \circ \gamma)(s).$$

Since each v_i is constant, differentiating this equality at $s = t$ gives

$$\phi''(t) = \sum_{i=1}^{n} v_i (g_i \circ \gamma)'(t) = \sum_{i=1}^{n} \sum_{j=1}^{n} v_i \big(H_f(x + tv)\big)_{ij} v_j,$$

which equals $v^\top H_f(x + tv)v$ by the definition of matrix multiplication. Since t was arbitrary, both formulas hold for every t in the domain of ϕ. □

Thus $v^\top H_f(x)v$ is the second derivative of f along the line through x in direction v: the Hessian measures curvature one direction at a time, just as the gradient measures slope one direction at a time. We use this result below to relate the signs of $v^\top H_f(x)v$ across directions v to concavity, and in Lecture 6 to derive necessary second-order conditions for a local optimum.

Finally, we state the one-variable second-derivative test.

Lemma 5.3.3 (One-variable second-derivative test)**.** *Let $I \subseteq \mathbb{R}$ be an interval, and suppose ϕ is twice continuously differentiable on an open interval containing I. Then ϕ is concave on I if and only if $\phi''(t) \le 0$ for every $t \in I$. If $\phi''(t) < 0$ for every $t \in I$, then ϕ is strictly concave.*

We use this result without proof. The second half of the statement does not reverse: a strictly concave ϕ may have $\phi'' = 0$ somewhere, as $\phi(t) = -t^4$ does at $t = 0$.

5.3.2 Tangent and Hessian criteria

The first criterion replaces the chord inequality by a comparison with the tangent plane at one point.

Theorem 5.3.4 (Tangent characterization)**.** *Let $U \subseteq \mathbb{R}^n$ be open and convex, and let $f : U \to \mathbb{R}$ be differentiable. Then f is concave if and only if*

$$f(y) \le f(x) + \nabla f(x) \cdot (y - x) \qquad \text{for every } x, y \in U. \tag{5.3.1}$$

Proof (optional). Concavity implies (5.3.1). Let $x, y \in U$ and $t \in (0, 1)$. By Definition 5.2.4 applied to x and y,

$$f\big(x + t(y - x)\big) \ge (1 - t) f(x) + t f(y),$$

and subtracting $f(x)$ and dividing by $t > 0$ gives

$$\frac{f\big(x + t(y - x)\big) - f(x)}{t} \ge f(y) - f(x).$$

As t decreases to zero the left-hand side tends to $\phi'(0)$, where $\phi(t) = f(x + t(y - x))$, and the chain rule gives $\phi'(0) = \nabla f(x) \cdot (y - x)$. The inequality passes to the limit, which is (5.3.1).

(5.3.1) *implies concavity.* Let $x, y \in U$, let $t \in [0,1]$, and put $z = (1-t)x + ty$, which lies in U because U is convex. Applying (5.3.1) at the point z to each of x and y,

$$f(x) \le f(z) + \nabla f(z) \cdot (x - z),$$
$$f(y) \le f(z) + \nabla f(z) \cdot (y - z).$$

Multiply the first by $1-t$ and the second by t and add. The gradient terms combine into $\nabla f(z) \cdot \big((1-t)x + ty - z\big)$, which is zero by the definition of z, leaving $(1-t)f(x) + tf(y) \le f(z)$. $\square$

A differentiable concave function thus lies on or below each of its tangent planes. Note that the right-hand side of (5.3.1) uses the value and gradient of f at the point x only, yet it bounds f everywhere. Figure 5.3.2 shows the inequality in one variable.

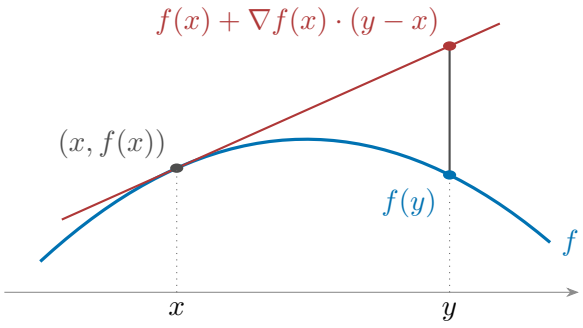


Figure 5.3.2: The tangent characterization of Theorem 5.3.4. The graph of a differentiable concave function lies on or below the tangent line at x, so the value $f(x) + \nabla f(x) \cdot (y - x)$ computed from data at x alone bounds $f(y)$ at every other point y.

The second criterion uses the Hessian. Negative semidefiniteness means that every line restriction bends downward or remains flat.

Theorem 5.3.5 (Hessian characterization)**.** *Let $U \subseteq \mathbb{R}^n$ be open and convex, and let $f : U \to \mathbb{R}$ be C^2.*

1. *f is concave if and only if $H_f(x)$ is negative semidefinite for every $x \in U$.*
2. *If $H_f(x)$ is negative definite for every $x \in U$, then f is strictly concave.*

The corresponding statements for convexity reverse every sign.

Proof (optional). Negative semidefinite implies concave. Fix $x, y \in U$, put $v = y - x$, and let $\phi(t) = f(x + tv)$ as in Lemma 5.3.1. By openness of U, this restriction is C^2 on an open interval containing $[0,1]$. By Theorem 5.3.2, $\phi''(t) = v^{\top} H_f(x + tv) v$, which is at most 0 by Definition 5.1.7. So ϕ is concave by Lemma 5.3.3, and since x and y were arbitrary, f is concave by Lemma 5.3.1.

Concave implies negative semidefinite. Fix $x \in U$ and $v \in \mathbb{R}^n$. Since U is open, $x + tv \in U$ for every t in some open interval containing 0. The restriction $\phi(t) = f(x + tv)$ is concave by Lemma 5.3.1, so Lemma 5.3.3 gives $\phi''(0) \le 0$. By Theorem 5.3.2,

$$\phi''(0) = v^{\top} H_f(x) v \le 0.$$

Since v was arbitrary, $H_f(x)$ is negative semidefinite.

Negative definite implies strictly concave. Fix distinct $x, y \in U$, put $v = y - x$, and let $\phi(t) = f(x + tv)$. The direction v is nonzero, so $\phi''(t) < 0$ throughout by Theorem 5.3.2 and Definition 5.1.7. Hence Lemma 5.3.3 makes ϕ strictly concave. Since x and y were arbitrary, Lemma 5.3.1 makes f strictly concave. □

Part (2) does not reverse, for the same reason Lemma 5.3.3 does not: $f(x) = -x^4$ is strictly concave on $\mathbb{R}$ while $f''(0) = 0$. Negative definiteness is sufficient for strict concavity, never necessary.

5.3.3 Global optimality and uniqueness

For concave functions, a first-order condition at a single point is sufficient for a global maximum.

Corollary 5.3.6 (First-order sufficiency)**.** *Let $U \subseteq \mathbb{R}^n$ be open and convex, let $C \subseteq U$ be convex, and let $f : U \to \mathbb{R}$ be differentiable and concave. If $x^* \in C$ satisfies*

$$\nabla f(x^*) \cdot (x - x^*) \leq 0 \qquad \text{for every } x \in C, \tag{5.3.2}$$

then x^* maximizes f on C. If f is strictly concave, then x^* is the only maximizer.

Proof. For $x \in C$, Theorem 5.3.4 at the point x^* and then (5.3.2) give

$$f(x) \leq f(x^*) + \nabla f(x^*) \cdot (x - x^*) \leq f(x^*).$$

Uniqueness. Suppose f is strictly concave and $x \neq y$ both maximize f on C, with common value m. Their midpoint lies in C by Definition 5.2.1, and Definition 5.2.4 gives $f\big(\frac{1}{2}x + \frac{1}{2}y\big) > \frac{1}{2}m + \frac{1}{2}m = m$, contradicting that m is the maximum. □

In particular, any $x^* \in C$ with $\nabla f(x^*) = 0$ satisfies (5.3.2). It is therefore a global maximizer of a concave function, and under strict concavity it is the only one. The more general inequality (5.3.2) also allows x^* to lie on the boundary of C, where the gradient need not vanish.

Example 5.3.7 (Maximizing a strictly concave quadratic)**.** Fix $b \in \mathbb{R}^2$ and $\gamma \in \mathbb{R}$ with $|\gamma| < 2$, and define $q : \mathbb{R}^2 \to \mathbb{R}$ by

$$q(x) = b \cdot x - \big(x_1^2 + \gamma x_1 x_2 + x_2^2\big).$$

Its Hessian has leading principal minors -2 and $4 - \gamma^2 > 0$, so it is negative definite by Theorem 5.1.10. Hence q is strictly concave on $\mathbb{R}^2$ by Theorem 5.3.5. Setting $\nabla q(x) = 0$ gives

$$\begin{pmatrix} 2 & \gamma \\ \gamma & 2 \end{pmatrix} x = b, \qquad \text{so} \qquad x^* = \frac{1}{4 - \gamma^2} \begin{pmatrix} 2 & -\gamma \\ -\gamma & 2 \end{pmatrix} b.$$

By Corollary 5.3.6, x^* is the unique global maximizer of q over $\mathbb{R}^2$. The first-order equations produce the candidate; strict concavity is what makes it the answer.

5.3.4 Separation

We end with a geometric consequence of convexity. Separating hyperplanes place two convex sets in opposite halfspaces.

Definition 5.3.8 (Hyperplane)**.** Let $a \in \mathbb{R}^n \setminus \{0\}$ and $c \in \mathbb{R}$. The *hyperplane* with normal a at level c is

$$H(a, c) = \{x \in \mathbb{R}^n : a \cdot x = c\},$$

and its two *closed halfspaces* are $\{x : a \cdot x \leq c\}$ and $\{x : a \cdot x \geq c\}$.

Theorem 5.3.9 (Separating hyperplane)**.** *Let $A, B \subseteq \mathbb{R}^n$ be nonempty, disjoint, and convex. There exist $a \neq 0$ and $c \in \mathbb{R}$ such that*

$$a \cdot x \leq c \leq a \cdot y \qquad \text{for every } x \in A \text{ and every } y \in B.$$

We use this result without proof. The two sets lie in opposite closed halfspaces of $H(a, c)$. The inequalities are weak: the common level c may be approached, or even attained, by points of the two sets. Figure 5.3.3 shows the configuration in $\mathbb{R}^2$.

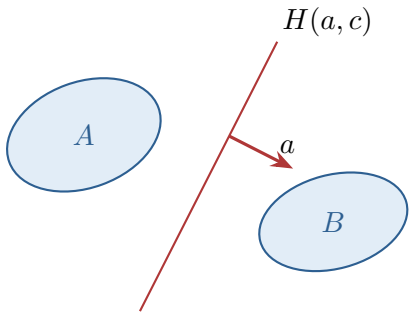


Figure 5.3.3: A hyperplane separating two disjoint convex sets. Every point of A satisfies $a \cdot x \leq c$ and every point of B satisfies $a \cdot y \geq c$.

Example 5.3.10 (No gap between the sets)**.** Let $A = \{x \in \mathbb{R}^2 : x_2 < 0\}$ and $B = \{x \in \mathbb{R}^2 : x_2 \geq 0\}$, which are nonempty, disjoint, and convex. Taking $a = e_2$ and $c = 0$ gives

$$a \cdot x = x_2 < 0 \leq y_2 = a \cdot y \qquad (x \in A,\ y \in B).$$

Thus the sets are separated by the level 0, but there is no positive gap between their separating levels: points of A have second coordinate arbitrarily close to 0.

A positive gap therefore requires hypotheses beyond disjointness.

Lecture 6

General Optimization

Solving an optimization problem requires three separate steps. First, we show that a solution exists. Second, we find the points that satisfy the necessary conditions for a solution. Third, we determine whether any of those points is a global solution. Setting a derivative equal to zero addresses only the second step.

6.1 The problem and existence

6.1.1 The problem and the candidate list

Many optimization problems belong to a family indexed by a parameter. We hold the parameter fixed while solving a single problem, but keep it in the notation so that the solution set and value can later be compared across problems.

Definition 6.1.1 (Optimization problem)**.** Let Θ be a parameter set, let $U \subseteq \mathbb{R}^n$, and let $f : U \times \Theta \to \mathbb{R}$. For a fixed $\theta \in \Theta$, the problem

$$\max_{x \in D(\theta)} f(x;\theta), \qquad D(\theta) \subseteq U,$$

has the following parts.

- The vector x is the *choice variable* and θ is the *parameter*.
- The set $D(\theta)$ is the *feasible set* and $f(\,\cdot\,;\theta)$ is the *objective*.
- The *solution set* is
$$X^*(\theta) = \arg\max_{x \in D(\theta)} f(x;\theta),$$
and when it is nonempty its elements all share one objective value, the *value* $V(\theta) = f(x^*;\theta)$ for any $x^* \in X^*(\theta)$.

Example 6.1.2 (Utility maximization)**.** A consumer with utility $u : \mathbb{R}^n_+ \to \mathbb{R}$ faces prices $p \in \mathbb{R}^n_{++}$ and wealth $w > 0$:

$$\max_{x \in B(p,w)} u(x), \qquad B(p,w) = \{x \in \mathbb{R}^n_+ : p \cdot x \le w\}.$$

Here the bundle x is the choice variable, (p, w) is the parameter, the budget set is the feasible set, demand is the solution set, and indirect utility is the value.

Note that writing max instead of sup presumes that the supremum is attained. A supremum can be finite without being attained, and then $X^*(\theta)$ is empty and $V(\theta)$ is undefined. Before using the notation, we must establish that a solution exists.

Definition 6.1.3 (Local and global maximizers)**.** Let $D \subseteq \mathbb{R}^n$ and $f : D \to \mathbb{R}$, and let $x^* \in D$.

- x^* is a *global maximizer* if $f(x^*) \geq f(x)$ for every $x \in D$.
- x^* is a *local maximizer* if there is $\varepsilon > 0$ with

  $$f(x^*) \geq f(x) \qquad \text{for every } x \in D \cap B(x^*, \varepsilon),$$

  and a *strict* local maximizer if that inequality is strict for every such $x \neq x^*$.
- x^* is *interior* if $x^* \in \operatorname{int} D$, meaning that $B(x^*, \varepsilon) \subseteq D$ for some $\varepsilon > 0$, and a *boundary point* if every ball $B(x^*, \varepsilon)$ meets both D and its complement.

Minimizers are defined by reversing every inequality.

Every global maximizer is a local maximizer, since $D \cap B(x^*, \varepsilon) \subseteq D$; the converse holds under concavity, as we show in Section 6.3. The interior–boundary distinction determines which first-order condition applies: at an interior point every small perturbation is feasible, so the objective can be compared with x^* in both directions along every line, while at a boundary point some directions leave the feasible set and the comparison is one-sided.

Recall the existence result of Lecture 2, which we use without proof.

Theorem 6.1.4 (Weierstrass)**.** *If $D \subseteq \mathbb{R}^n$ is nonempty and compact and $f : D \to \mathbb{R}$ is continuous, then f attains both a maximum and a minimum on D.*

In the next example, we first establish existence and then generate candidates.

6.1.2 Existence on an unbounded set

The following condition is sufficient for existence on an unbounded feasible set.

Definition 6.1.5 (Coercivity for maximization)**.** Let $D \subseteq \mathbb{R}^n$ be unbounded. A function $f : D \to \mathbb{R}$ is *coercive for maximization* on D if

$$x_k \in D, \quad \|x_k\| \to \infty \quad \Longrightarrow \quad f(x_k) \to -\infty$$

for every sequence (x_k); equivalently, if for every $M \in \mathbb{R}$ there is $R > 0$ such that $f(x) < M$ whenever $x \in D$ and $\|x\| > R$.

Theorem 6.1.6 (Coercive existence)**.** *Let $D \subseteq \mathbb{R}^n$ be nonempty, closed, and unbounded. If $f : D \to \mathbb{R}$ is continuous and coercive for maximization, then f attains a maximum on D.*

Proof (optional). Choose any $x_0 \in D$. Applying Definition 6.1.5 with $M = f(x_0)$ gives $R > 0$, which we may take larger than $\|x_0\|$, such that

$$x \in D, \ \|x\| > R \quad \Longrightarrow \quad f(x) < f(x_0).$$

Put $K = D \cap \{x : \|x\| \leq R\}$. It contains x_0, so it is nonempty; it is bounded by R; and it is closed, being the intersection of two closed sets. By the Heine–Borel theorem it is therefore compact, so Theorem 6.1.4 supplies a maximizer x^* of f on K. Any $x \in D \setminus K$ has $\|x\| > R$ and hence $f(x) < f(x_0) \leq f(x^*)$, the second inequality because $x_0 \in K$. So $f(x) \leq f(x^*)$ for every $x \in D$, which is Definition 6.1.3. $\square$

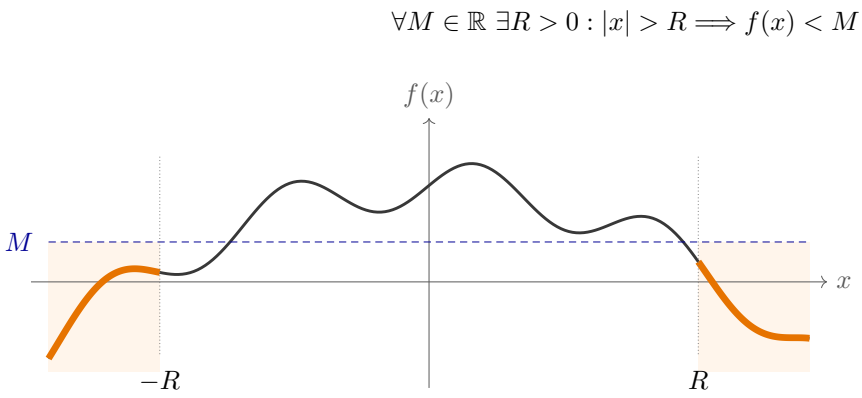


Figure 6.1.1: Coercivity reduces maximization on an unbounded closed feasible set to maximization on a compact subset that contains every candidate with value at least $f(x_0)$.

Notice the structure of the proof: coercivity allows us to replace the unbounded feasible set with a compact one that contains a maximizer, and then apply the Weierstrass theorem. Like compactness, coercivity is sufficient for existence but not necessary.

Example 6.1.7 (Coercivity is not necessary)**.** On $D = [0, \infty)$, the function $f(x) = xe^{-x}$ is not coercive for maximization, since $f(x) \to 0$ rather than $-\infty$; yet $f'(x) = (1 - x)e^{-x}$ is positive before 1 and negative after 1, so $x = 1$ is the unique global maximizer. On $D = \mathbb{R}$, the function $f(x) = \arctan x$ is likewise not coercive, but here there is no maximizer: f increases toward the supremum $\pi/2$, which it never attains.

6.2 Necessary local conditions

6.2.1 The interior first-order condition

Both first-order conditions below are proved the same way: we restrict f to a line through the candidate point and apply one-variable arguments to the restriction. We collect the one-variable facts first.

Lemma 6.2.1 (One-variable optimality)**.** *Let $\delta > 0$ and let $g : (-\delta, \delta) \to \mathbb{R}$ have a local maximum at 0.*

1. *If g is differentiable at 0, then $g'(0) = 0$.*
2. *If g is twice continuously differentiable, then $g''(0) \leq 0$.*

Proof. For part (1), take t small enough that $g(t) \leq g(0)$. When $t > 0$, dividing $g(t) - g(0) \leq 0$ by t preserves the inequality, so

$$\frac{g(t) - g(0)}{t} \leq 0 \qquad \text{and letting } t \downarrow 0, \qquad g'(0) \leq 0.$$

When $t < 0$, dividing by t reverses it, so the same quotient is nonnegative and letting $t \uparrow 0$ gives $g'(0) \geq 0$. The derivative exists, so both one-sided limits equal it, and $g'(0) = 0$.

For part (2), part (1) gives $g'(0) = 0$. The one-dimensional case of the second-order Taylor expansion from Lecture 5 gives

$$g(t) - g(0) = \frac{1}{2}g''(0)t^2 + r(t), \qquad \frac{|r(t)|}{t^2} \longrightarrow 0.$$

If $g''(0) > 0$, then for every sufficiently small nonzero t the positive quadratic term dominates the remainder, so $g(t) > g(0)$. This contradicts the local maximum. Hence $g''(0) \leq 0$. □

Theorem 6.2.2 (Interior first-order condition)**.** *Let $U \subseteq \mathbb{R}^n$ be open, let $D \subseteq U$, and let $f : U \to \mathbb{R}$ be differentiable at $x^* \in \operatorname{int} D$. If x^* is a local maximizer or a local minimizer of f on D, then $\nabla f(x^*) = 0$.*

Proof. Fix $v \in \mathbb{R}^n$ and put $g(t) = f(x^* + tv)$. Since $x^* \in \operatorname{int} D$, there is $\varepsilon > 0$ with $B(x^*, \varepsilon) \subseteq D$, so g is defined for $|t| < \varepsilon/\|v\|$ when $v \neq 0$. If x^* is a local maximizer of f on D, then $g(t) \leq g(0)$ for all small $|t|$ by Definition 6.1.3, so 0 is a local maximum of g. By part (1) of Lemma 6.2.1 and the chain rule for a line restriction,

$$0 = g'(0) = \nabla f(x^*) \cdot v.$$

This holds for every $v \in \mathbb{R}^n$; taking $v = \nabla f(x^*)$ gives $\|\nabla f(x^*)\|^2 = 0$, so $\nabla f(x^*) = 0$. For a local minimizer, apply the same argument to $-f$. □

Definition 6.2.3 (Critical point)**.** A point x at which f is differentiable and $\nabla f(x) = 0$ is a *critical point*, or *stationary point*, of f.

Theorem 6.2.2 is a necessary condition, not a sufficient one. Its converse fails even at a differentiable interior point. Moreover, the condition itself fails if we drop interiority, and it does not apply without differentiability.

Example 6.2.4 (Three counterexamples)**.**

1. Criticality does not imply optimality. The origin is critical for $f(x) = x^3$ on $\mathbb{R}$ by Definition 6.2.3, yet $f(t) > 0 > f(-t)$ for every $t > 0$, so it is neither a local maximizer nor a local minimizer.

2. Optimality at a non-interior point does not imply criticality. For $f(x) = x$ on $D = [0, 1]$, the maximizer is $x^* = 1$, which is a boundary point of D, and $f'(1) = 1 \neq 0$.

3. The condition does not apply without differentiability. The function $f(x) = -|x|$ on $\mathbb{R}$ has a strict global maximizer at 0, an interior point, but is not differentiable there, so Theorem 6.2.2 does not apply.

6.2.2 Second-order conditions

A critical point can be a local maximizer, a local minimizer, or neither. The Hessian classifies it whenever the associated quadratic form is definite. The tool is the second-order Taylor expansion of Lecture 5: if f is C^2 on an open set containing x^*, then

$$f(x^* + h) - f(x^*) = \nabla f(x^*) \cdot h + \frac{1}{2}h^\top H_f(x^*)h + r(h), \qquad \frac{|r(h)|}{\|h\|^2} \to 0, \tag{6.2.1}$$

and the second derivative of a line restriction $g(t) = f(x^* + tv)$ is the quadratic form $g''(t) = v^\top H_f(x^* + tv)v$.

Theorem 6.2.5 (Second-order conditions)**.** *Let f be C^2 on an open set containing x^*.*

1. *If x^* is a local maximizer, then $H_f(x^*)$ is negative semidefinite.*
2. *If $\nabla f(x^*) = 0$ and $H_f(x^*)$ is negative definite, then x^* is a strict local maximizer.*
3. *If $\nabla f(x^*) = 0$ and $H_f(x^*)$ is indefinite, then x^* is a* saddle point*, which is neither a local maximizer nor a local minimizer.*

For local minimizers, reverse the signs in the first two statements.

Proof (optional). (1) Necessity. Let $v \in \mathbb{R}^n$ and put $g(t) = f(x^* + tv)$, which has a local maximum at 0. Part (2) of Lemma 6.2.1 and the line-restriction formula give $v^\top H_f(x^*)v = g''(0) \le 0$. Since v was arbitrary, $H_f(x^*)$ is negative semidefinite.

(2) Sufficiency. The function $u \mapsto u^\top H_f(x^*)u$ is continuous and the unit sphere $\{u : \|u\| = 1\}$ is nonempty and compact, so Theorem 6.1.4 gives a maximum value $-c$ on it, and $c > 0$ because negative definiteness makes every such value negative. For $h \neq 0$, applying this to $u = h/\|h\|$ and multiplying by $\|h\|^2$ gives

$$h^\top H_f(x^*)h \le -c\|h\|^2.$$

Since $|r(h)|/\|h\|^2 \to 0$, there is $\rho > 0$ with $|r(h)| \le (c/4)\|h\|^2$ whenever $0 < \|h\| < \rho$. For such h, (6.2.1) with $\nabla f(x^*) = 0$ gives

$$f(x^* + h) - f(x^*) \le -\frac{c}{2}\|h\|^2 + \frac{c}{4}\|h\|^2 = -\frac{c}{4}\|h\|^2 < 0,$$

so x^* is a strict local maximizer.

(3) Indefiniteness. Choose v and w with $v^\top H_f(x^*)v > 0$ and $w^\top H_f(x^*)w < 0$. Because the gradient vanishes, substituting $h = tv$ into (6.2.1) gives

$$f(x^* + tv) - f(x^*) = \frac{t^2}{2}v^\top H_f(x^*)v + r(tv) > 0$$

for every sufficiently small nonzero t: after division by t^2, the remainder tends to zero while the quadratic coefficient is positive. The same argument with $h = tw$ makes the difference negative for every sufficiently small nonzero t. Thus every ball around x^* contains points with larger and smaller values, so x^* is neither a local maximizer nor a local minimizer.

The minimum statements follow by applying the above to $-f$, whose Hessian is $-H_f$. □

When the Hessian is semidefinite but not definite, neither part (2) nor part (3) applies, and no conclusion follows. The next example shows that every outcome is possible in that case.

Example 6.2.6 (A semidefinite Hessian is inconclusive)**.** Each of $-x^4$, x^4, and x^3 has $f'(0) = f''(0) = 0$, so their first- and second-order data at the origin are identical. Yet the origin is a strict maximizer of the first, a strict minimizer of the second, and neither for the third. A zero Hessian therefore supplies no classification.

In practice, we check definiteness of the Hessian with the tests of Lecture 5 instead of the definition.

Example 6.2.7 (A saddle point classified)**.** Let $f(x_1, x_2) = x_1^2 - x_2^2$. Then $\nabla f(0) = 0$ and $H_f = \operatorname{diag}(2, -2)$, whose eigenvalues 2 and -2 have opposite signs, so the Hessian is indefinite and part (3) of Theorem 6.2.5 makes the origin neither a local maximizer nor a local minimizer.

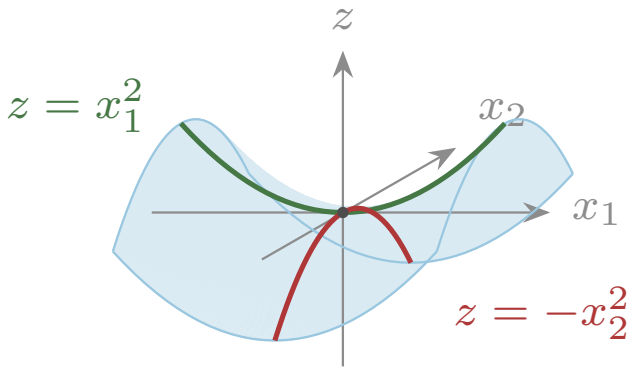


Figure 6.2.1: The graph of $f(x_1, x_2) = x_1^2 - x_2^2$ near its saddle point at the origin.

6.2.3 Boundary first-order conditions

At a boundary maximizer the gradient need not vanish, as Example 6.2.4 showed. A one-sided statement remains true, and to state it we first name the directions that stay inside the feasible set.

Definition 6.2.8 (Feasible direction)**.** Let $x \in D \subseteq \mathbb{R}^n$. A vector $v \in \mathbb{R}^n$ is a *feasible direction* from x if there is $\delta > 0$ with

$$x + tv \in D \qquad \text{for every } t \in (0, \delta).$$

If D is convex, then $y - x$ is a feasible direction from x for every $y \in D$, since Definition 6.2.8 is satisfied with $\delta = 1$.

Theorem 6.2.9 (Boundary first-order condition)**.** *Let $U \subseteq \mathbb{R}^n$ be open, let $D \subseteq U$, and let $f : U \to \mathbb{R}$ be differentiable. If $x^* \in D$ is a local maximizer of f on D, then*

$$\nabla f(x^*) \cdot v \le 0 \qquad \textit{for every feasible direction } v \textit{ from } x^*.$$

In particular, if D is convex, then

$$\nabla f(x^*) \cdot (y - x^*) \le 0 \qquad \textit{for every } y \in D. \tag{6.2.2}$$

Proof (optional). Fix a feasible direction v and put $g(t) = f(x^* + tv)$. By Definition 6.2.8, this restriction is feasible for all small $t > 0$. Since x^* is a local maximizer, $g(t) \le g(0)$ for all sufficiently small $t > 0$, so the difference quotients satisfy

$$\frac{g(t) - g(0)}{t} \le 0 \qquad \text{for all small } t > 0.$$

Letting $t \downarrow 0$ and evaluating the derivative of the line restriction by the chain rule,

$$0 \ge g'(0) = \nabla f(x^*) \cdot v.$$

If D is convex and $y \in D$, then $v = y - x^*$ is feasible, which gives (6.2.2). □

At an interior point the condition collapses back to Theorem 6.2.2: for any v, both $y = x^* + \varepsilon v$ and $y = x^* - \varepsilon v$ lie in D for small $\varepsilon > 0$, so (6.2.2) gives $\nabla f(x^*) \cdot v \le 0$ and $-\nabla f(x^*) \cdot v \le 0$, forcing $\nabla f(x^*) \cdot v = 0$ for every v.

If x^* is a boundary local maximizer, then the gradient is allowed to not be zero, and if it's not, then it must point out of the feasible set. To see this, note that $\nabla f(x^*) \cdot v \le 0$ means that the angle between the gradient and any feasible direction v is at least ninety degrees. That is equivalent to saying that $f(x^* + tv) \le f(x^*)$, meaning that moving in the direction v decreases the value of f relative to $f(x^*)$. That is, at a boundary point, the function is allowed to increase in some directions (those that point out of the feasible set), but it cannot increase in any direction that points into the feasible set.

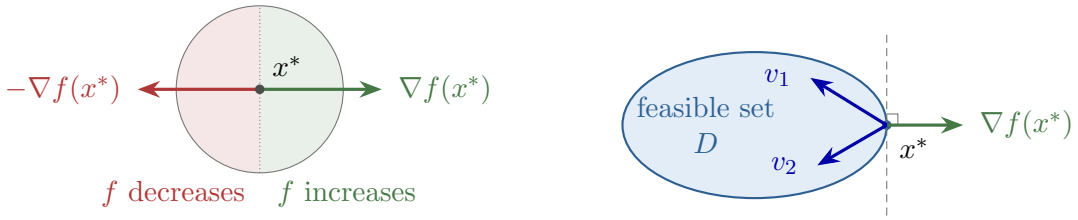


Figure 6.2.2: Left: the directions in which f increases and decreases relative to $f(x^*)$. f cannot rise if the angle between the gradient and a feasible direction is at least ninety degrees (the red area). Right: the first-order condition (6.2.2) at a local maximizer x^* at the boundary of D. The gradient points out of D and makes an angle of at least ninety degrees with every feasible direction v. Moving in a feasible direction therefore cannot increase f.

On a nonnegative orthant the condition becomes one sign restriction and one complementarity relation per coordinate.

Theorem 6.2.10 (Componentwise form on the orthant)**.** *Let $U \subseteq \mathbb{R}^n$ be open with $\mathbb{R}^n_+ \subseteq U$, let $f : U \to \mathbb{R}$ be differentiable, and let $x^* \in \mathbb{R}^n_+$. Then* (6.2.2) *with $D = \mathbb{R}^n_+$ holds if and only if*

$$\frac{\partial f}{\partial x_i}(x^*) \le 0, \qquad x_i^* \ge 0, \qquad x_i^* \frac{\partial f}{\partial x_i}(x^*) = 0 \qquad (i = 1, \dots, n). \tag{6.2.3}$$

Proof (optional). Necessity. Taking $y = x^* + e_i \in D$ in (6.2.2) gives $\partial f/\partial x_i(x^*) \le 0$. If in addition $x_i^* > 0$, then $y = x^* - x_i^* e_i$ is also in D, and (6.2.2) applied to it gives $-x_i^* \, \partial f/\partial x_i(x^*) \le 0$, hence $\partial f/\partial x_i(x^*) \ge 0$ and therefore $= 0$. So the product $x_i^* \, \partial f/\partial x_i(x^*)$ vanishes whether x_i^* is zero or positive.

Sufficiency. Let $y \in D$, so $y_i \ge 0$ for every i. Expanding the inner product and using (6.2.3) twice,

$$\nabla f(x^*) \cdot (y - x^*) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^*)\, y_i - \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^*)\, x_i^* = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^*)\, y_i \le 0,$$

the second sum vanishing term by term and each remaining term being a product of a nonpositive number with a nonnegative one. □

The three conditions in (6.2.3) say that increasing any one coordinate cannot raise f to first order, that feasibility holds, and that a partial derivative can be strictly negative only where the corresponding coordinate is already zero and cannot be reduced. The last identity is a complementarity relation; in the parameterized example of Section 6.3.2, it decides between an interior point and a corner.

6.3 Global certification

6.3.1 Concavity, globality, and uniqueness

Everything so far has been about local maximizers, and a local maximizer can fail to be a global one. However, if a function is concave, then every local maximizer is also a global maximizer.

Theorem 6.3.1 (Concave maximization)**.** *Let $D \subseteq \mathbb{R}^n$ be convex and let $f : D \to \mathbb{R}$ be concave.*

1. *Every local maximizer of f on D is a global maximizer.*
2. *The set $X^* = \arg\max_{x \in D} f(x)$ is either empty or convex.*
3. *If f is strictly concave, then X^* contains at most one point.*

Proof (optional). Local implies global. Suppose x^* is a local maximizer that is not global, so some $y \in D$ has $f(y) > f(x^*)$. By Definition 6.1.3 there is $\varepsilon > 0$ with $f(x^*) \geq f(x)$ for every feasible $x \in B(x^*, \varepsilon)$. The points $x_t = (1-t)x^* + ty$ are feasible by convexity of D, and $\|x_t - x^*\| = t\|y - x^*\|$, so $x_t \in B(x^*, \varepsilon)$ once t is small enough. For such $t \in (0, 1)$, concavity gives

$$f(x_t) \geq (1-t)f(x^*) + tf(y) > (1-t)f(x^*) + tf(x^*) = f(x^*),$$

contradicting local maximality.

Convexity of X^*. Suppose X^* is nonempty and let $x_0, x_1 \in X^*$ have common maximum value M. For $t \in [0, 1]$, the point $(1-t)x_0 + tx_1$ is feasible, and concavity gives

$$f\big((1-t)x_0 + tx_1\big) \geq (1-t)M + tM = M.$$

Since M is the maximum, the value is M and the point lies in X^*. So X^* contains the segment joining any two of its points, and is therefore convex.

Strict-concavity uniqueness. If distinct $x, y \in X^*$ had common value M, their midpoint would be feasible and strict concavity would give

$$f\left(\frac{x+y}{2}\right) > \frac{1}{2}f(x) + \frac{1}{2}f(y) = M,$$

contradicting maximality. □

Under concavity, the necessary condition of Theorem 6.2.9 is also sufficient, so a single first-order check certifies a global maximizer.

Theorem 6.3.2 (First-order characterization)**.** *Let $D \subseteq \mathbb{R}^n$ be convex and let f be differentiable and concave on an open convex set containing D. A point $x^* \in D$ is a global maximizer of f on D if and only if*

$$\nabla f(x^*) \cdot (y - x^*) \leq 0 \qquad \text{for every } y \in D.$$

Proof. Necessity. A global maximizer is a local maximizer, so Theorem 6.2.9 applies.

Sufficiency. For $y \in D$, the tangent inequality from Lecture 5 at x^* and then the hypothesis give

$$f(y) \leq f(x^*) + \nabla f(x^*) \cdot (y - x^*) \leq f(x^*),$$

so x^* is a global maximizer by Definition 6.1.3. □

Part (3) of Theorem 6.3.1 bounds the number of maximizers without producing one. The function $f(x) = -e^x$ is strictly concave on $\mathbb{R}$, and its supremum 0 is approached as $x \to -\infty$ but never attained, so the solution set is empty. Existence remains a separate step. However, if there exists a point satisfying the first-order condition of Theorem 6.3.2, then existence is guaranteed by Theorem 6.3.2.

6.3.2 A complete parameterized problem

We now apply the three steps, in order, to a parameterized problem.

Example 6.3.3. For $\theta \in \mathbb{R}$, consider

$$\max_{x \in \mathbb{R}_+} f(x;\theta), \qquad f(x;\theta) = \theta x - \frac{1}{2}x^2.$$

Find the solution set $X^*(\theta)$ and value $V(\theta)$ for every θ.

Existence. For fixed θ the objective is continuous, and $D = \mathbb{R}_+$ is nonempty, closed, and unbounded. Since $f(x;\theta) \to -\infty$ as $x \to \infty$, the objective is coercive for maximization in the sense of Definition 6.1.5, so Theorem 6.1.6 gives a maximizer.

Candidates. Here $n = 1$ and

$$\frac{\partial f}{\partial x}(x;\theta) = \theta - x.$$

By Theorem 6.2.10 the condition (6.2.3) reads

$$\theta - x^* \le 0, \qquad x^* \ge 0, \qquad x^*(\theta - x^*) = 0.$$

If $x^* > 0$ the last equation forces $x^* = \theta$, which is admissible only when $\theta > 0$; if $x^* = 0$ the first inequality requires $\theta \le 0$. Thus exactly one point satisfies the condition: $x^* = \max\{\theta, 0\}$.

Globality and uniqueness. Since $\partial^2 f/\partial x^2 = -1 < 0$, the function is strictly concave on the convex set $\mathbb{R}_+$. The first-order condition is therefore sufficient by Theorem 6.3.2, and uniqueness follows from part (3) of Theorem 6.3.1. Hence

$$X^*(\theta) = \{\max\{\theta, 0\}\}, \qquad V(\theta) = \begin{cases} 0, & \theta \le 0, \\ \frac{1}{2}\theta^2, & \theta > 0. \end{cases}$$

The two regimes are the two ways (6.2.3) can hold: at $\theta \le 0$ the solution is the corner 0, whereas at $\theta > 0$ it is the interior point θ.

Removing the quadratic term separates the roles of concavity, strict concavity, and coercivity.

Example 6.3.4 (Removing the curvature)**.** For $\beta \ge 0$, let $f_\beta(x;\theta) = \theta x - (\beta/2)x^2$ on $\mathbb{R}_+$. When $\beta > 0$, the argument above gives the unique maximizer $x^*_\beta(\theta) = \max\{\theta/\beta, 0\}$. When $\beta = 0$, the objective is the affine function θx, and the three regimes are

$$X^*(\theta) = \{0\} \text{ if } \theta < 0, \qquad X^*(\theta) = \mathbb{R}_+ \text{ if } \theta = 0, \qquad X^*(\theta) = \emptyset \text{ if } \theta > 0.$$

At $\theta = 0$, existence holds but uniqueness fails because the objective is no longer strictly concave. At $\theta > 0$, existence fails because the affine objective is unbounded above.

Example 6.3.5 (A non-concave problem in $\mathbb{R}^2$)**.** Find the solution set X^* and value V of

$$\max_{x \in \mathbb{R}^2} \{-x_1^4 - x_2^4 + 4x_1^2 + 2x_2^2\}.$$

Example 6.3.6 (A concave problem in $\mathbb{R}^2$)**.** Find the solution set X^* and value V of

$$\max_{x \in \mathbb{R}^2} \{3x_1 + 3x_2 - x_1^2 - x_1x_2 - x_2^2\}.$$

6.3.3 A workflow

The steps below summarize the workflow for solving a parameterized maximization problem.

1. Specify the objective, its domain, the feasible set, and the parameter, as in Definition 6.1.1.
2. Establish existence, by compactness (Theorem 6.1.4), coercivity (Theorem 6.1.6), or a direct argument.
3. Generate every candidate: interior critical points (Theorem 6.2.2), boundary points satisfying the feasible-direction condition of Theorem 6.2.9, and points where f fails to be differentiable, where no first-order condition applies.
4. Classify interior candidates locally with Theorem 6.2.5 when the classification is needed and the Hessian is definite or indefinite. Discard the minimizers and saddle points, and keep the local maximizers and the inconclusive points.
5. Certify globality, either by concavity (Theorem 6.3.1 and Theorem 6.3.2) or by comparing the values of f across all remaining candidate points. If f is concave, then any critical point in D is a global maximizer, and under strict concavity the unique one.
6. Report $X^*(\theta)$ and $V(\theta)$.

Lecture 7

Constrained Optimization

7.1 Overview: constraints, tangent directions, and multipliers

Lecture 6 solved an optimization problem in three separate steps: establish that a solution exists, generate candidates from first-order conditions, and certify a candidate as global, by concavity or by comparing values. Its first-order condition at a boundary point was one-sided: at a local maximizer x^*,

$$\nabla f(x^*) \cdot v \leq 0 \qquad \text{for every feasible direction } v \text{ from } x^*, \tag{7.1.1}$$

where v is a feasible direction if the segment $x^* + tv$ stays in the feasible set for all small $t > 0$.

In economic problems the feasible set is typically described by equations and inequalities. For example, in a consumer problem, the budget constraint $p \cdot x \leq m$ is a linear inequality. In a planner's problem, the requirement $x_1 + x_2 = m$ that a resource be split between two uses is a linear equality. In a producer's problem, the technology constraint $q = F(z)$, linking output q to inputs z, is a nonlinear equality whenever the production function F is nonlinear.

Depending on the shape of the feasible set, the condition (7.1.1) may be vacuous, i.e., it may rule out nothing. Suppose first that there is just one *equality* constraint $h(x) = 0$. If h is affine, as in the planner's problem, the feasible set is a line: segments along the line stay in the set, so nonzero feasible directions exist and (7.1.1) is not vacuous (Figure 7.1.1, left). If instead the feasible set is curved, such as the circle $\{x \in \mathbb{R}^2 : x_1^2 + x_2^2 = 2\}$, it contains no segments: every straight move away from a point of the circle leaves the circle. The only feasible direction is then $v = 0$, so (7.1.1) holds at every point of the circle and rules out nothing (Figure 7.1.1, right).

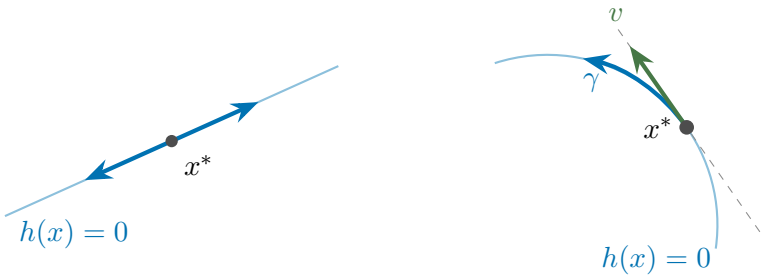


Figure 7.1.1: Feasible movement under an equality constraint. Left: an affine constraint. The feasible set is the line $h(x) = 0$ with $h(x) = m - x_1 - x_2$; the two blue arrows from x^* are feasible directions, since segments along the line stay in the set. Right: a curved constraint. The feasible set is the circle $h(x) = 0$ with $h(x) = 2 - x_1^2 - x_2^2$; every straight move from x^* leaves the circle. The feasible movement is the curve γ, which bends with the circle; the green vector v is its velocity at x^*, tangent to the circle (the dashed tangent line).

For a curved feasible set, directions of movement come from curves rather than segments. A feasible movement through x^* is a curve $\gamma(t)$ that stays in the feasible set, with $\gamma(0) = x^*$: think of $\gamma(t)$ as the position of a moving point at time t. Its *velocity* $v = \gamma'(0)$ is the vector of the instantaneous rates of change of its coordinates — the direction in which the moving point passes through x^*. The path bends with the set, but the velocity is an ordinary vector: on the circle it is tangent to the circle, and on the line the curve can be the line itself, whose velocity is one of the old feasible directions.

At points satisfying a regularity condition (LICQ), these velocities — the *tangent directions* — are exactly the vectors orthogonal to every constraint gradient $\nabla h_j(x^*)$, and the first-order condition becomes

$$\nabla f(x^*) + \mu_1^* \nabla h_1(x^*) + \cdots + \mu_k^* \nabla h_k(x^*) = 0$$

for some coefficients μ_j^*, which are the *Lagrange multipliers*. The function $L(x, \mu) = f(x) + \mu \cdot h(x)$ is the *Lagrangian*; its critical points are precisely the pairs (x, μ) that satisfy this equation together with $h(x) = 0$.

Section 7.3 then adds inequality constraints, written $g_i(x) \geq 0$. At a candidate x^*, a constraint is *active* if $g_i(x^*) = 0$ and *slack* if $g_i(x^*) > 0$; a slack constraint does not restrict small movements, so it drops out of the first-order condition. Stationarity, feasibility, nonnegative multipliers on the inequalities, and the rule that each multiplier or its constraint is zero (*complementary slackness*) together form the *Karush–Kuhn–Tucker (KKT) conditions*.

Section 7.4 then provides the global certification, specialized to the cases of Sections 7.2 and 7.3. As in Lecture 6, existence must be shown separately, using Weierstrass, coercivity, or a direct argument.

7.2 Equality constraints and the Lagrangian

7.2.1 Tangent directions and constraint normals

Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}$ and $h : U \to \mathbb{R}^k$ be C^1, and consider

$$\max_{x \in U} \; f(x) \qquad \text{subject to} \qquad h(x) = 0.$$

Here, the feasible set $\{x \in U : h(x) = 0\}$ is described by a system of k equations. When one of these equations is nonlinear, the feasible set may be curved. We thus work with feasible curves and their velocities.

We call γ a *feasible curve through* x^* if γ is differentiable, $\gamma(0) = x^*$, and $h(\gamma(t)) = 0$ for all t near 0. Its velocity at x^* is $v = \gamma'(0)$. Such a velocity is called a *tangent direction* to the constraint set at x^*.

Lemma 7.2.1 (Tangent directions lie in the null space)**.** *Let γ be a feasible curve through x^* with velocity v. Then*

$$\mathrm{D}h(x^*)\, v = 0.$$

Proof. $h(\gamma(t)) = 0$ for all t near 0, and differentiating using the chain rule from Lecture 4 gives

$$\mathrm{D}h(\gamma(t))\, \gamma'(t) = 0 \quad \text{at } t = 0, \quad \text{so} \quad \mathrm{D}h(x^*)\, v = 0.$$

□

Every tangent direction thus lies in the null space $N(\mathrm{D}h(x^*))$ from Lecture 3. Row by row, the display says

$$\nabla h_j(x^*) \cdot v = 0 \qquad \text{for each } j.$$

In other words, each constraint gradient is *orthogonal* (perpendicular) to every tangent direction. A vector perpendicular to the tangent directions is called a *normal* to the constraint set. Thus the constraint gradients are normals to the constraint set.

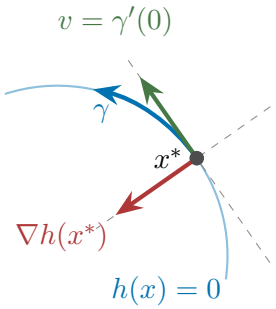


Figure 7.2.1: The local geometry at x^* for one constraint ($k = 1$). The feasible curve γ stays on the constraint set. Its velocity $v = \gamma'(0)$ is tangent to the constraint set, while $\nabla h(x^*)$ is normal to it; hence $\nabla h(x^*) \cdot v = 0$. The dashed lines are the tangent and normal lines at x^*.

The converse is generally not true. That is, if $\mathrm{D}h(x^*)\, v = 0$ for some vector v, there may be no feasible curve through x^* with velocity v. For example, if $h(x) = x^2$, then the feasible set is the single point $\{0\}$, so the only feasible velocity at 0 is 0. Yet, the null space of $h'(0) = 0$ is all of $\mathbb{R}$, so any vector $v \in \mathbb{R}$ satisfies $\mathrm{D}h(0)\, v = 0$.

The following condition ensures that all vectors in the null space of $\mathrm{D}h(x^*)$ are velocities of feasible curves through x^*.

Definition 7.2.2 (LICQ for equality constraints)**.** The *linear independence constraint qualification* holds at a feasible point x^* if the constraint gradients $\nabla h_1(x^*), \ldots, \nabla h_k(x^*)$ are linearly independent; equivalently, if $\operatorname{rank} \mathrm{D}h(x^*) = k$.

Remark 7.2.3 (How many constraints?). The number k of equations need not equal the number of independent restrictions they place near x^*; that number is $r = \operatorname{rank} \mathrm{D}h(x^*)$. When the constraints have a locally equivalent description by r independent equations, the feasible set is locally a surface of dimension $n - r$. Thus $r < n$ leaves feasible movement, while $r = n$ makes the feasible point locally isolated. LICQ says $r = k$, so it requires $k \leq n$: more than n vectors in $\mathbb{R}^n$ cannot be linearly independent. When $k > n$, no description with that many equations satisfies LICQ; when possible, re-describe the same feasible set by independent equations, and otherwise analyze it directly.

Lemma 7.2.4 (Tangent directions under LICQ)**.** *Let h be C^1 and let LICQ hold at the feasible point x^*. If $\mathrm{D}h(x^*)\, v = 0$, then v is the velocity of a differentiable feasible curve through x^*.*

Intuitively, if a vector v is orthogonal to the constraint normals, then it is tangent to the constraint set, and a curve can be drawn along the constraint set in the direction of v. LICQ thus ensures that the constraints form a smooth surface locally around x^*, so that every vector of the null space is realized as the velocity of a feasible curve.

Combining Lemmas 7.2.1 and 7.2.4, we conclude that under LICQ, the set of feasible velocities at x^* is exactly $N(\mathrm{D}h(x^*))$.

7.2.2 First-order conditions and the Lagrangian

Now suppose x^* is a local maximizer at which LICQ holds. For any feasible curve through x^*, the restriction $t \mapsto f(\gamma(t))$ has a local maximum at the interior point $t = 0$, so its derivative at $t = 0$ is zero. Chain rule and Lemma 7.2.4 thus yield

$$\nabla f(x^*) \cdot v = 0 \qquad \text{for every } v \in N(\mathrm{D}h(x^*)).$$

This is the equality-constraint analogue of (7.1.1). It holds with equality because both directions along the constraint set are available: if v is a tangent direction, then so is $-v$, the velocity of the reversed curve $\gamma(-t)$. Simply put, the gradient of the objective f must be orthogonal to every tangent direction v, or else moving along a feasible curve in one of the directions v and $-v$ would increase f. Under LICQ, the tangent directions are exactly the vectors orthogonal to the constraint gradients.[1] Since $\nabla f(x^*)$ is orthogonal to every tangent direction, it must be a linear combination of the constraint gradients; that is, there are coefficients $\mu_1^*, \ldots, \mu_k^*$ such that

$$\nabla f(x^*) + \mu_1^* \nabla h_1(x^*) + \cdots + \mu_k^* \nabla h_k(x^*) = 0.$$

These coefficients are called the *Lagrange multipliers*. The first-order condition is summarized by the following theorem.

Theorem 7.2.5 (Lagrange multiplier theorem)**.** *Let f and h be C^1, let x^* be a local extremum (maximizer or minimizer) of f subject to $h(x) = 0$, and let LICQ hold at x^*. Then there is a unique vector $\mu^* \in \mathbb{R}^k$ such that*

$$\nabla f(x^*) + \mathrm{D}h(x^*)^\top \mu^* = 0.$$

[1] The row interpretation of the null space from Lecture 3 says that $Av = 0$ if and only if v is orthogonal to every row of A. We also use the following row-space/null-space fact without proof: every vector orthogonal to all of $N(A)$ is a linear combination of the rows of A. For $A = \mathrm{D}h(x^*)$, those rows are the transposed constraint gradients.

Proof (optional, using Lecture 8's implicit function theorem). Stack the objective gradient on top of the constraint gradients:

$$M = \begin{pmatrix} \nabla f(x^*)^\top \\ \nabla h_1(x^*)^\top \\ \vdots \\ \nabla h_k(x^*)^\top \end{pmatrix}.$$

Suppose, for contradiction, that $\operatorname{rank} M = k+1$. Then some $k+1$ columns of M form an invertible matrix. Hold the remaining coordinates fixed at their x^*-values and apply the implicit function theorem to the selected coordinates in the system

$$f(x) = c_0, \qquad h(x) = c.$$

It follows that every right-hand side (c_0, c) sufficiently close to $(f(x^*), 0)$ has a solution near x^*. In particular, take $c = 0$ and $c_0 = f(x^*) + \varepsilon$. If x^* is a local maximizer, choose $\varepsilon > 0$; if it is a local minimizer, choose $\varepsilon < 0$. Either choice produces a nearby feasible point that contradicts the optimality of x^*. Therefore, $\operatorname{rank} M \le k$.

By LICQ, the k constraint gradients are linearly independent. Since adding $\nabla f(x^*)$ does not increase their rank, $\nabla f(x^*)$ must be a linear combination of them. Thus there is a vector $\mu^* \in \mathbb{R}^k$ such that

$$\nabla f(x^*) + \mathrm{D}h(x^*)^\top \mu^* = 0.$$

Finally, if both μ^* and $\tilde{\mu}$ satisfy this equation, then $\mathrm{D}h(x^*)^\top(\mu^* - \tilde{\mu}) = 0$. LICQ makes the constraint gradients linearly independent, so $\mu^* = \tilde{\mu}$. □

As in Lecture 6, the theorem gives only a necessary condition for a local extremum. Note that the Lagrange multipliers are not restricted in sign, i.e., they can be positive or negative. For a fixed objective, constraint description, and candidate x^*, LICQ makes the multiplier vector unique. Rescaling or rewriting a constraint (e.g., multiplying both sides by 2) changes its multiplier. Finally, the theorem does not guarantee that an optimizer exists; existence must be established separately.

LICQ cannot be dropped from Theorem 7.2.5, and all points in the feasible set that do not satisfy LICQ must be treated separately.

Example 7.2.6 (A singular description loses the multiplier)**.** Consider $\max_{x \in \mathbb{R}} x$ subject to $h(x) = x^2 = 0$. We have $f'(x) = 1$ and $h'(x) = 2x$. If we were to apply the Lagrange multiplier theorem, we would get $1 + 2x\mu = 0$. However, the feasible set is the single point $\{0\}$, meaning that $x^* = 0$ is the global maximizer. Clearly, at $x^* = 0$ there does not exist a μ that satisfies the multiplier condition $1 + 2x^*\mu = 0$.

The Lagrange multiplier condition and the feasibility requirement $h(x) = 0$ form a system of $n+k$ equations in the $n+k$ unknowns (x, μ). These equations can be written as the critical-point conditions of a single function of $n + k$ variables, the *Lagrangian*.

Definition 7.2.7 (Lagrangian)**.** The *Lagrangian* of the problem $\max f(x)$ subject to $h(x) = 0$ is the function of x and μ given by

$$L(x, \mu) = f(x) + \mu \cdot h(x).$$

Its partial gradients are:

$$\nabla_x L(x,\mu) = \nabla f(x) + \mathrm{D}h(x)^\top \mu, \qquad \nabla_\mu L(x,\mu) = h(x).$$

Consequently, the critical points of L are exactly the pairs (x,μ) that satisfy the first-order condition and the feasibility requirement. Note also that the Lagrangian is not being maximized: a constrained maximizer is a critical point of $L(\,\cdot\,,\mu^*)$, and can thus be a maximum, a minimum, or a saddle point of the Lagrangian.

Example 7.2.8 (Solving directly and with a Lagrangian)**.** Consider max x_1x_2 subject to $x_1 + x_2 = 2$.

Direct solution. The constraint gives $x_2 = 2 - x_1$, so the problem reduces to

$$\max_{x_1\in\mathbb{R}} \; x_1(2-x_1) = \max_{x_1\in\mathbb{R}} \big[1 - (x_1-1)^2\big].$$

The unique global maximizer is $x_1^* = 1$, and the constraint then gives $x_2^* = 1$. Thus $x^* = (1,1)^\top$.

Lagrangian solution. Write the constraint as $h(x) = 2 - x_1 - x_2 = 0$. Since $\nabla h(x) = (-1,-1)^\top \neq 0$, LICQ holds. The Lagrangian is

$$L(x,\mu) = x_1x_2 + \mu\,(2 - x_1 - x_2),$$

so every constrained optimum must be among the solutions to

$$\frac{\partial L}{\partial x_1} = x_2 - \mu = 0, \qquad \frac{\partial L}{\partial x_2} = x_1 - \mu = 0, \qquad \frac{\partial L}{\partial \mu} = 2 - x_1 - x_2 = 0.$$

The first two equations give $x_1 = x_2 = \mu$. Substituting into the third gives $\mu^* = 1$ and hence $x^* = (1,1)^\top$, the same point found above. The direct solution established that this point is the global maximizer; the Lagrangian equations by themselves identified it only as a candidate.

In particular, we did not maximize the Lagrangian. As a function of x with $\mu = \mu^*$ fixed,

$$L(x,\mu^*) = x_1x_2 + 2 - x_1 - x_2 \qquad \text{has Hessian} \qquad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

whose characteristic equation $\lambda^2 - 1 = 0$ gives eigenvalues 1 and -1. The Hessian is indefinite, so by the second-order classification of Lecture 6 the point x^* is a saddle point of $L(\,\cdot\,,\mu^*)$. On the feasible set, $h(x) = 0$ makes $L(x,\mu^*) = f(x)$, so x^* does maximize the Lagrangian among feasible points. The Lagrangian collects the first-order conditions of the constrained problem; it does not turn the problem into an unconstrained maximization of L.

7.3 Inequalities and the KKT conditions

7.3.1 Standard form and active sets

We fix one convention for inequality constraints. Every inequality is written as

$$g_i(x) \geq 0,$$

so $a(x) \leq b$ becomes $b - a(x) \geq 0$.

The optimization problem is

$$\max_{x\in U} f(x) \qquad \text{subject to} \qquad h(x)=0, \quad g(x)\geq 0,$$

where $h : U \to \mathbb{R}^k$ and $g : U \to \mathbb{R}^\ell$. Its Lagrangian is

$$L(x,\mu,\lambda) = f(x) + \mu\cdot h(x) + \lambda\cdot g(x), \qquad \mu\in\mathbb{R}^k, \quad \lambda\geq 0.$$

The requirement $\lambda \geq 0$ is forced by the convention $g \geq 0$, and the one-dimensional calculation after Theorem 7.3.2 shows why.

For inequality constraints, we will use the following classification.

Definition 7.3.1 (Active and slack constraints, mixed LICQ)**.** At a feasible point x, the *active set* is

$$I(x) = \{i : g_i(x) = 0\},$$

and a constraint with $g_i(x) > 0$ is *slack*. LICQ holds at x for the mixed problem if the vectors

$$\{\nabla h_j(x) : j = 1,\dots,k\} \cup \{\nabla g_i(x) : i \in I(x)\}$$

are linearly independent.

A slack constraint stays satisfied throughout a neighborhood of x (by continuity of g_i), so it places no restriction on local movement and its gradient does not enter the constraint qualification.

7.3.2 The KKT theorem

Theorem 7.3.2 (Karush–Kuhn–Tucker)**.** *Let f, h, and g be C^1, let x^* be a local maximizer subject to $h(x) = 0$ and $g(x) \geq 0$, and let LICQ hold at x^*. Then there are $\mu^* \in \mathbb{R}^k$ and $\lambda^* \in \mathbb{R}^\ell$ such that*

$$\begin{aligned}
\nabla f(x^*) + \mathrm{D}h(x^*)^\top\mu^* + \mathrm{D}g(x^*)^\top\lambda^* &= 0 && && \textit{(stationarity)}\\
h(x^*) = 0, \qquad g(x^*) &\geq 0 && && \textit{(feasibility)}\\
\lambda^* &\geq 0 && && \textit{(multiplier signs)}\\
\lambda_i^*\, g_i(x^*) &= 0 && (i=1,\dots,\ell) && \textit{(complementary slackness)}
\end{aligned}$$

We use this result without proof.

Complementary slackness forces $\lambda_i^* = 0$ whenever $g_i(x^*) > 0$, and permits $\lambda_i^* > 0$ only where $g_i(x^*) = 0$: slack constraints have zero multipliers. The converse implication fails: an *active* constraint can also have a zero multiplier (as in Example 7.3.3 below).

The sign restriction $\lambda \geq 0$ can be read off a one-dimensional picture. Suppose $n = \ell = 1$, the constraint $g(x^*) = 0$ is active, and $g'(x^*) > 0$, so $g(x^* + t) > 0$ for every small $t > 0$ and $v = 1$ is a feasible direction. If x^* is a local maximizer, the objective cannot increase into the feasible side, which is (7.1.1) with $v = 1$: $f'(x^*) \leq 0$. Stationarity reads $f'(x^*) + \lambda^* g'(x^*) = 0$, so

$$\lambda^* = -\frac{f'(x^*)}{g'(x^*)} \geq 0.$$

With the constraint written as $g \leq 0$ instead, the same calculation would produce the opposite sign, so the convention must be fixed once and recorded.

Like every first-order condition in this course, the KKT conditions generate candidates but do not certify them. On $\max_{x\geq 0} x^3$, the point $x^* = 0$ with $\lambda^* = 0$ satisfies every condition of Theorem 7.3.2, yet 0 is not even a local maximizer, since $x^3 > 0$ for every $x > 0$. Existence and certification remain separate steps, and here existence in fact fails.

7.3.3 Active-set reasoning in a capacity problem

With ℓ inequalities, the practical method is to reason/guess about which constraints bind. Each guess of an active set of constraints turns the KKT conditions into equalities that can be solved. The remaining KKT conditions — feasibility and multiplier signs — then confirm that the guess is a solution, or reject it.

Example 7.3.3 (KKT by guess-and-verify)**.** Let $a > 0$ and $K > 0$, and consider

$$\max_q \; aq - \tfrac{1}{2}q^2 \qquad \text{subject to} \qquad q \geq 0, \quad K - q \geq 0.$$

The feasible set $[0, K]$ is nonempty and compact and the objective is continuous, so a maximizer exists by Weierstrass. With $g_1(q) = q$ and $g_2(q) = K - q$, the Lagrangian is

$$L(q, \lambda) = aq - \tfrac{1}{2}q^2 + \lambda_1 q + \lambda_2(K - q),$$

and the KKT conditions are

$$a - q + \lambda_1 - \lambda_2 = 0, \qquad 0 \leq q \leq K, \qquad \lambda_1, \lambda_2 \geq 0, \qquad \lambda_1 q = 0, \quad \lambda_2(K - q) = 0.$$

There are four cases to check.

1. $\lambda_1 = \lambda_2 = 0$. Then stationarity gives $q = a$. To satisfy the second constraint, we must have $K - q \geq 0 \iff K \geq a$.
2. $\lambda_1 = 0$, $K - q = 0$. Then stationarity gives $\lambda_2 = a - K$, and $\lambda_2 \geq 0$ is satisfied if and only if $K \leq a$.
3. $\lambda_2 = 0$, $q = 0$. Then stationarity gives $\lambda_1 = -a < 0$, which violates the sign restriction; this guess is rejected for every $a > 0$.
4. $\lambda_1, \lambda_2 > 0$. This case is impossible because we cannot have $q = 0$ and $q = K$ for $K > 0$.

The solution is thus

$$q^*(a, K) = \min\{a, K\}, \qquad \lambda_2^*(a, K) = \max\{a - K, 0\}.$$

Note that when $a = K$, the solution is $q^* = K$ and the multiplier is $\lambda_2^* = 0$, meaning that the capacity constraint is active and the multiplier is zero.

To certify that q^* is a global maximizer, observe that the objective can be rewritten as follows:

$$aq - \tfrac{1}{2}q^2 = -\tfrac{1}{2}(q - a)^2 + \tfrac{1}{2}a^2,$$

which shows that the objective is maximized by the point q^* that minimizes $(q - a)^2$, i.e., the point closest to a. Since the point of $[0, K]$ closest to a is $\min\{a, K\} = q^*$, it is indeed a global maximizer.

7.4 Global certification

7.4.1 Concave programs

In Lecture 6, if the objective function f is concave on a convex set, then the first-order condition is sufficient for a global maximum. The same logic extends to the following class of constrained problems.

Definition 7.4.1 (Concave program)**.** Let $U \subseteq \mathbb{R}^n$ be convex. The problem

$$\max_{x \in U} \; f(x) \qquad \text{subject to} \qquad h(x) = 0, \quad g(x) \geq 0,$$

is a *concave program* if f and every g_i are concave on U and h is affine.

The sufficiency proof again relies on the tangent inequality from Lecture 5, $f(y) \leq f(x) + \nabla f(x) \cdot (y - x)$.

Theorem 7.4.2 (Sufficiency in a concave program)**.** *Consider a concave program on an open convex set U, and suppose that f and every g_i are differentiable. If a feasible x^* and multipliers (μ^*, λ^*) satisfy the KKT conditions of Theorem 7.3.2, then x^* is a global maximizer.*

Proof (optional). Let y be feasible and put $v = y - x^*$. Concavity of f gives

$$f(y) - f(x^*) \leq \nabla f(x^*) \cdot v,$$

and stationarity rewrites the right-hand side as

$$\nabla f(x^*) \cdot v = -\mu^* \cdot \mathrm{D}h(x^*)v - \lambda^* \cdot \mathrm{D}g(x^*)v.$$

Because h is affine and both points are feasible, $h(y) - h(x^*) = \mathrm{D}h(x^*)v = 0$, so the μ^* term vanishes. For each i, concavity of g_i gives $\nabla g_i(x^*) \cdot v \geq g_i(y) - g_i(x^*)$, and since $\lambda^* \geq 0$,

$$-\lambda^* \cdot \mathrm{D}g(x^*)v \leq -\lambda^* \cdot \big(g(y) - g(x^*)\big) = \lambda^* \cdot g(x^*) - \lambda^* \cdot g(y) \leq 0,$$

where the final step uses complementary slackness, $\lambda^* \cdot g(x^*) = 0$, and then $\lambda^* \geq 0$ with $g(y) \geq 0$. Chaining the displays gives $f(y) \leq f(x^*)$. □

Recall from Lecture 6 that the set of maximizers of a strictly concave function on a convex set contains at most one point (see Theorem 3.1, part 3 in Lecture 6). The constraints of a concave program define a convex feasible set (possibly empty): it is the intersection of superlevel sets of concave functions and an affine equality set, each convex (see Lecture 5). Thus, if f is strictly concave, the program has at most one maximizer.

The following condition requires the set of strictly feasible points (points that satisfy the equality constraints and satisfy every inequality strictly) to be nonempty. In a concave program, it makes the KKT conditions necessary; they are already sufficient by Theorem 7.4.2.

Definition 7.4.3 (Slater's condition)**.** A concave program with affine equalities satisfies *Slater's condition* if there is $\bar{x} \in U$ with

$$h(\bar{x}) = 0 \qquad \text{and} \qquad g_i(\bar{x}) > 0 \quad \text{for every } i.$$

Theorem 7.4.4 (Necessity under Slater's condition)**.** *Consider a concave program on an open convex set U, and suppose that f and every g_i are differentiable. If Slater's condition holds, then every global maximizer satisfies the KKT conditions with some multipliers (μ^*, λ^*).*

Slater and LICQ are both constraint qualifications, but they are different: LICQ is local and point-specific, while Slater is global and specific to concave programs. However, both are sufficient to guarantee the existence of Lagrange multipliers.

Here is another example of what goes wrong when a constraint qualification fails.

Example 7.4.5 (A concave program with no multipliers)**.** Consider

$$\max_{(x_1,x_2)\in\mathbb{R}^2} x_2 \qquad \text{subject to} \qquad g_1(x) = x_1 \geq 0, \quad g_2(x) = -x_1 - x_2^2 \geq 0.$$

This is a concave program: the objective and g_1 are affine, and g_2 is concave. Feasibility requires $0 \leq x_1 \leq -x_2^2 \leq 0$, so the feasible set is $\{(0,0)\}$ and $x^* = (0,0)$ is the global maximizer. Slater's condition fails because $g_1(x) > 0$ requires $x_1 > 0$, whereas $g_2(x) > 0$ requires $x_1 < -x_2^2 \leq 0$.

At x^*, stationarity would require

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \lambda_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} -1 \\ 0 \end{pmatrix} = 0,$$

which is impossible for any multipliers because the second component is always 1. Thus the global maximizer has no KKT multipliers. Yet, the global maximizer exists and is unique — it is the only feasible point $(0,0)$.

7.4.2 Second-order conditions on the tangent space

For equality-constrained problems there is also a second-order local test. The relevant matrix is the Hessian of the *Lagrangian*, not of the objective: curvature only matters along the tangent directions, and along those directions both the objective and the constraints curve. The multiplier terms in L account for the curvature of the constraints.

Theorem 7.4.6 (Second-order conditions on the tangent space)**.** *Let f and h be C^2, let x^* be feasible with LICQ, and let μ^* satisfy the stationarity condition from the Lagrange multiplier Theorem 7.2.5:*

$$\nabla f(x^*) + \mathrm{D}h(x^*)^\top \mu^* = 0.$$

Write

$$T = N(\mathrm{D}h(x^*)), \qquad H = H_f(x^*) + \sum_{j=1}^{k} \mu_j^* H_{h_j}(x^*),$$

so H is the Hessian of $L(\,\cdot\,, \mu^*)$ at x^*.

1. *If x^* is a local constrained maximizer, then $v^\top H v \leq 0$ for every $v \in T$.*
2. *If $v^\top H v < 0$ for every nonzero $v \in T$, then x^* is a strict local constrained maximizer.*

Example 7.4.7 (Classifying two candidates on a circle)**.** Consider $\max x_1 + x_2$ subject to $h(x) = 2 - x_1^2 - x_2^2 = 0$.

Candidates. Since $\nabla h(x) = (-2x_1, -2x_2)^\top \neq 0$ at every point of the circle, LICQ holds at every feasible point. The stationarity system of $L = x_1 + x_2 + \mu(2 - x_1^2 - x_2^2)$ is

$$1 - 2\mu x_1 = 0, \qquad 1 - 2\mu x_2 = 0, \qquad x_1^2 + x_2^2 = 2,$$

with the two solutions $(x^*, \mu^*) = \big((1,1)^\top, \tfrac{1}{2}\big)$ and $\big((-1,-1)^\top, -\tfrac{1}{2}\big)$.

Classification. The objective is linear, so $H_f = 0$ and the curvature comes entirely from the constraint:

$$H = \mu^* H_h = -2\mu^* I,$$

which is $-I$ at the first candidate and I at the second. Both tangent spaces are $T = \{v : v_1 + v_2 = 0\}$. On T, the first candidate has $v^\top(-I)v = -\|v\|^2 < 0$ for $v \neq 0$, a strict local maximum, and the second has $v^\top I v > 0$, a strict local minimum. The Hessian of the objective alone is zero at both points, so it cannot distinguish them; the classification comes from the constraint curvature term $\mu^* H_h$.

Certification. The circle is compact, so Weierstrass gives a global maximizer and a global minimizer, and by Theorem 7.2.5 both are among the two candidates. Comparing values, $f(1,1) = 2$ and $f(-1,-1) = -2$, so the local classification is also the global one.

There is also a second-order local test for inequality constraints, but it goes beyond the scope of this course due to the fact that the relevant set of directions is a cone rather than a subspace.

Lecture 8

Comparative Statics, Envelopes, and Fixed Points

Lectures 6 and 7 solve one optimization problem at a time. Economics rarely stops there: the interesting question is usually how the solution and its value respond when a parameter moves. What does a tax do to the equilibrium price? What does an extra unit of capacity add to profit? The solved problem almost never has a closed form, so the answers must be read from the equations that characterize the solution. This method is called *comparative statics*. The main tool of comparative statics is the *implicit function theorem*, which differentiates through a system of equations.

8.1 The implicit function theorem

8.1.1 Scalar implicit differentiation

Let $F : U \to \mathbb{R}$ be C^1 on an open set $U \subseteq \mathbb{R}^2$, with the arguments written $F(x;\theta)$: an equation in the endogenous variable x, indexed by the parameter θ. Suppose a differentiable function $x(\theta)$ solves the equation identically,

$$F\big(x(\theta);\theta\big) = 0,$$

on some interval of parameters. Such a function is a *solution branch* of the equation. Differentiating the identity using the chain rule from Lecture 4, we get

$$F_x\big(x(\theta);\theta\big)\,x'(\theta) + F_\theta\big(x(\theta);\theta\big) = 0,$$

and at any point where $F_x \neq 0$,

$$x'(\theta) = -\frac{F_\theta\big(x(\theta);\theta\big)}{F_x\big(x(\theta);\theta\big)}.$$

The calculation is conditional: *if* a differentiable solution branch runs through the point, its slope is determined by the equation above.

8.1.2 The system theorem

The same logic runs for systems. Let $X \subseteq \mathbb{R}^n$ and $\Theta \subseteq \mathbb{R}^r$ be open, and let $F : X \times \Theta \to \mathbb{R}^n$ collect n equations in the n endogenous variables x, with r parameters θ. The scalar condition $F_x \neq 0$ becomes invertibility of the endogenous Jacobian.

Theorem 8.1.1 (Implicit function theorem)**.** *Let F be C^1, let $F(x^0;\theta^0)=0$, and let $\mathrm{D}_x F(x^0;\theta^0)$ be invertible. Then there are neighborhoods $\mathcal{U}$ of θ^0 and $\mathcal{V}$ of x^0 and a unique C^1 function $x:\mathcal{U}\to\mathcal{V}$ such that*

$$x(\theta^0)=x^0 \qquad \text{and} \qquad F\big(x(\theta);\theta\big)=0 \quad \text{for every } \theta\in\mathcal{U},$$

and for each $\theta\in\mathcal{U}$ the point $x(\theta)$ is the only solution of $F(x;\theta)=0$ in $\mathcal{V}$. The derivative satisfies

$$\mathrm{D}_x F\big(x(\theta);\theta\big)\,\mathrm{D}_\theta x(\theta) = -\mathrm{D}_\theta F\big(x(\theta);\theta\big). \tag{8.1.1}$$

We use the system statement without proof; the scalar case can be proved with the tools of Lectures 2 and 4.

Proof (optional, for one equation and one parameter). The two signs are symmetric, so suppose $F_x(x^0;\theta^0)>0$. Since F_x is continuous, there is a rectangle $[x^0-\delta,x^0+\delta]\times[\theta^0-\eta,\theta^0+\eta]$ on which $F_x>0$, so on this rectangle F is strictly increasing in x for each fixed θ. Since $F(\,\cdot\,;\theta^0)$ increases through its zero at x^0,

$$F(x^0+\delta;\theta^0)>0 \qquad \text{and} \qquad F(x^0-\delta;\theta^0)<0.$$

Both signs survive small moves of the parameter: F is continuous in θ, so there is $\varepsilon\in(0,\eta]$ such that $F(x^0+\delta;\theta)>0$ and $F(x^0-\delta;\theta)<0$ for every θ within ε of θ^0. For each such θ, the Theorem 2.3.7 from Lecture 2 gives a solution $x(\theta)\in(x^0-\delta,x^0+\delta)$ of $F(x;\theta)=0$, and strict monotonicity in x makes it the only one in the interval: this is the branch. Repeating the argument with a smaller δ traps the branch in a smaller rectangle, so $x(\theta)\to x^0$ as $\theta\to\theta^0$: the branch is continuous. We take the differentiability of the branch without proof; its derivative is then the chain-rule formula of Section 8.1.1. □

Everything in the theorem is local: invertibility at one solution gives a branch near that solution and carries no information about solutions elsewhere. Equation (8.1.1) is the scalar calculation written with matrices: differentiate the identity $F(x(\theta);\theta)=0$ by the chain rule and solve. For each of the r parameters, the corresponding column of $\mathrm{D}_\theta x$ solves one linear system with the same coefficient matrix $\mathrm{D}_x F$, as in Lecture 3; in practice, the matrix is inverted once, or each system is solved directly.

Example 8.1.2 (Incidence of a per-unit tax)**.** Fix demand and supply primitives $a,c\in\mathbb{R}$ and $b,s>0$. The parameter is the per-unit tax t. The endogenous variables are the producer price p and quantity q; consumers pay $p+t$. Demand and supply are

$$q=a-b\,(p+t), \qquad q=c+s\,p.$$

At a reference tax t^0, suppose an equilibrium $x^0=(p^0,q^0)^\top$ exists. How do the local equilibrium price $p(t)$, quantity $q(t)$, and consumer price $p(t)+t$ respond to the tax?

Solution. With endogenous vector $x=(p,q)^\top$, equilibrium is the system

$$F(p,q;t)=\begin{pmatrix} q-a+b\,(p+t)\\ q-c-s\,p\end{pmatrix}=0, \qquad \mathrm{D}_{(p,q)}F=\begin{pmatrix} b & 1\\ -s & 1\end{pmatrix},$$

with $\det \mathrm{D}_{(p,q)}F=b+s>0$, so Theorem 8.1.1 applies at any equilibrium: the equilibrium price and quantity are differentiable functions of the tax. Differentiating with respect to t as in (8.1.1),

$$\begin{pmatrix} b & 1\\ -s & 1\end{pmatrix}\begin{pmatrix} p_t\\ q_t\end{pmatrix}=\begin{pmatrix} -b\\ 0\end{pmatrix} \qquad\Longrightarrow\qquad p_t=-\frac{b}{b+s}, \quad q_t=-\frac{bs}{b+s}.$$

The consumer price moves by

$$\frac{\mathrm{d}(p+t)}{\mathrm{d}t} = 1 + p_t = \frac{s}{b+s} \in (0,1).$$

The tax splits between the two sides of the market in proportion to the slopes: the producer price falls by $b/(b+s)$ of the tax, the consumer price rises by the complementary share $s/(b+s)$, and quantity falls. Note that we never computed the equilibrium.

8.2 Comparative statics of optimizers and values

8.2.1 Optimizing systems

When the equations being differentiated are optimality conditions, the implicit function theorem turns into the comparative statics of choice. Let $f : X \times \Theta \to \mathbb{R}$ be C^2 with $X \subseteq \mathbb{R}^n$ and $\Theta \subseteq \mathbb{R}^r$ open. At a parameter θ^0, suppose x^0 is a critical point satisfying Lecture 6's first-order condition,

$$\nabla_x f(x^0; \theta^0) = 0.$$

This is a system $F(x;\theta) = \nabla_x f(x;\theta)$, and its endogenous Jacobian is the Hessian:

$$\mathrm{D}_x F(x;\theta) = H_f(x;\theta).$$

If the Hessian is invertible at $(x^0;\theta^0)$, Theorem 8.1.1 yields a C^1 branch of critical points with

$$H_f\big(x(\theta);\theta\big)\,\mathrm{D}_\theta x(\theta) = -\mathrm{D}_\theta\,\nabla_x f\big(x(\theta);\theta\big).$$

For constrained problems, we differentiate the Lagrange system. For $\max_x f(x;\theta)$ subject to $h(x;\theta) = 0$, Lecture 7's Lagrangian $L(x,\mu;\theta) = f(x;\theta) + \mu \cdot h(x;\theta)$ produces the system

$$G(x,\mu;\theta) = \begin{pmatrix} \nabla_x L(x,\mu;\theta) \\ h(x;\theta) \end{pmatrix} = 0$$

in the stacked endogenous vector (x,μ). Its endogenous Jacobian is written in blocks, each block a matrix whose dimensions are read off the grouping of equations and variables:

$$\mathrm{D}_{(x,\mu)} G = \begin{pmatrix} H_L(x,\mu;\theta) & \mathrm{D}_x h(x;\theta)^\top \\ \mathrm{D}_x h(x;\theta) & 0 \end{pmatrix},$$

where H_L is the Hessian of L in x alone. When this matrix is invertible at a stationary feasible pair, the implicit function theorem gives a differentiable branch $(x(\theta), \mu(\theta))$ of such pairs. Once the branch is established to consist of optimizers and their multipliers, we write it $(x^*(\theta), \mu^*(\theta))$.

Inequality constraints use the same technique. On a parameter region where the same constraints bind and every slack constraint stays slack, the binding constraints can be carried as equalities and everything above applies; the derivatives are valid strictly within that region, and Section 8.2.5 discusses what happens at its boundary.

8.2.2 Envelope theorems

So far we differentiated a critical point of an optimization problem. The second object of comparative statics is the optimized value

$$V(\theta) = f\big(x^*(\theta);\theta\big),$$

whose derivative is even simpler.

Theorem 8.2.1 (Envelope theorems)**.** *Assume the functions appearing below are C^1 in (x,θ).*

1. *If $x^*(\theta)$ is a differentiable branch of interior optimizers, so that $\nabla_x f(x^*(\theta);\theta) = 0$, then*

$$\mathrm{D}_\theta V(\theta) = \mathrm{D}_\theta f\big(x^*(\theta);\theta\big).$$

2. *If $x^*(\theta)$ is a differentiable branch of optimizers and $\mu^*(\theta)$ is its differentiable multiplier branch satisfying Lecture 7's stationarity and feasibility conditions, $\nabla_x L(x^*(\theta),\mu^*(\theta);\theta) = 0$ and $h(x^*(\theta);\theta) = 0$, then*

$$\mathrm{D}_\theta V(\theta) = \mathrm{D}_\theta L\big(x^*(\theta),\mu^*(\theta);\theta\big) = \mathrm{D}_\theta f + \mu^*(\theta)^\top \mathrm{D}_\theta h,$$

   with every derivative on the right taken in θ alone and evaluated along the branch.

3. *For a problem with equalities $h(x;\theta) = 0$ and inequalities $g(x;\theta) \geq 0$, suppose $x^*(\theta)$ is a differentiable branch of optimizers, $\mu^*(\theta)$ and $\lambda^*(\theta)$ are its differentiable multiplier branches satisfying the KKT conditions, and the active set is constant near the parameter under consideration. Then*

$$\mathrm{D}_\theta V(\theta) = \mathrm{D}_\theta L\big(x^*(\theta),\mu^*(\theta),\lambda^*(\theta);\theta\big) = \mathrm{D}_\theta f + \mu^*(\theta)^\top \mathrm{D}_\theta h + \lambda^*(\theta)^\top \mathrm{D}_\theta g.$$

Proof. For the interior case, the chain rule gives

$$\mathrm{D}_\theta V = \nabla_x f^\top \, \mathrm{D}_\theta x^* + \mathrm{D}_\theta f,$$

and the first term vanishes by stationarity. For the constrained case, the same chain rule step and stationarity $\nabla_x f^\top = -(\mu^*)^\top \mathrm{D}_x h$ give

$$\mathrm{D}_\theta V = \mathrm{D}_\theta f + \nabla_x f^\top \, \mathrm{D}_\theta x^* = \mathrm{D}_\theta f - (\mu^*)^\top \mathrm{D}_x h \, \mathrm{D}_\theta x^*.$$

Differentiating the feasibility identity $h(x^*(\theta);\theta) = 0$ gives $\mathrm{D}_x h \, \mathrm{D}_\theta x^* = -\mathrm{D}_\theta h$, and substituting finishes the proof of part 2.

For part 3, let I be the active set, which is constant by hypothesis. The equations $h = 0$ and $g_I = 0$ form one equality system, so part 2 applies with multipliers (μ^*, λ_I^*). Every inactive multiplier is zero by complementary slackness. Adding those zero terms gives the displayed formula with the full vectors λ^* and g. □

Notice that $\mathrm{D}_\theta x^*$ does not appear on any right-hand side. To first order, the response of the choice does not affect the value, because the first-order condition makes the objective flat in the choice direction; only the direct effect of the parameter remains.

Example 8.2.2 (The envelope formula on the quadratic family)**.** For each parameter $\theta \in \mathbb{R}$, the scalar choice variable x solves

$$\max_{x \geq 0} \ \theta x - \tfrac{1}{2}x^2.$$

The optimizer and value, respectively, are known to be $x^*(\theta) = \max\{\theta, 0\}$ and

$$V(\theta) = \begin{cases} \tfrac{1}{2}\theta^2, & \theta > 0, \\ 0, & \theta \leq 0. \end{cases}$$

(see Example 6.3.3 in Lecture 6). What does the envelope theorem give for $V'(\theta)$ on the interior and corner regimes, and does it agree with the known value function?

Solution.

On the interior regime $\theta > 0$, part 1 of Theorem 8.2.1 predicts

$$V'(\theta) = \frac{\partial}{\partial \theta}\Big(\theta x - \tfrac{1}{2}x^2\Big)\Big|_{x = x^*(\theta)} = x^*(\theta) = \theta,$$

and differentiating $\frac{1}{2}\theta^2$ directly confirms it. On the corner regime $\theta < 0$, write the constraint as $g(x) = x \geq 0$. The active set is constant, $x^* = 0$, and part 3 gives $V'(\theta) = \partial f/\partial \theta = x^*(\theta) = 0$, again matching the derivative of the constant value.

8.2.3 Interpretation of Lagrange multipliers

Lecture 7 introduced a Lagrange multiplier as an additional unknown in the first-order conditions. We can now interpret that unknown. Start with the equality-constrained problem

$$\max_{x} \ f(x) \qquad \text{subject to} \qquad r(x) = b.$$

Here x is the vector of choice variables, f and r are fixed primitives, and b is the constraint's right-hand side. Depending on the application, b could be a budget, capacity, time limit, or available material. Changing b shifts the constraint and therefore changes the best attainable value.

Using the course's sign convention, write the constraint and Lagrangian as

$$h(x; b) = b - r(x) = 0, \qquad L(x, \mu; b) = f(x) + \mu h(x; b).$$

Let

$$V(b) = f\big(x^*(b)\big)$$

be the optimized value. Suppose the optimizer $x^*(b)$ and its multiplier $\mu^*(b)$ form differentiable local branches. The equality-constrained envelope formula, restated for this problem, is

$$V'(b) = \frac{\partial L}{\partial b}\big(x^*(b), \mu^*(b); b\big) = \mu^*(b).$$

The last equality follows immediately from the Lagrangian: b enters L with coefficient μ. Thus the multiplier is the *shadow price* of the constraint's right-hand side—the marginal increase in the optimized value from increasing b. For a small change Δb, the first-order approximation is

$$\Delta V \equiv V(b + \Delta b) - V(b) \approx \mu^*(b)\, \Delta b.$$

Its units are units of optimized value per unit of b.

For an upper-bound inequality $r(x) \le b$, use the course's KKT convention

$$g(x;b) = b - r(x) \ge 0, \qquad L(x,\lambda;b) = f(x) + \lambda g(x;b), \qquad \lambda \ge 0.$$

On a differentiable regime with a fixed active set, a binding constraint has

$$V'(b) = \frac{\partial L}{\partial b} = \lambda^*(b).$$

A slack constraint has $\lambda^*(b) = 0$ by complementary slackness and therefore no marginal value on that regime. More generally, if constraint i is written as $g_i(x) + c_i \ge 0$, a positive c_i relaxes it and

$$\frac{\partial V}{\partial c_i} = \lambda_i^*.$$

The sign of a multiplier cannot be separated from the way its constraint is written. If the equality constraint is instead $r(x) - b = 0$ with multiplier $\widetilde{\mu}^*$, then $\widetilde{\mu}^* = -\mu^*$ and $V'(b) = -\widetilde{\mu}^*(b)$. Writing the constraint as $b - r(x) = 0$ makes the multiplier itself equal to the shadow price.

Example 8.2.3 (Capacity valued by the envelope formula)**.** Fix parameters $a > 0$ and $K > 0$. The scalar choice variable q solves

$$\max_q \; aq - \tfrac{1}{2}q^2 \qquad \text{subject to} \qquad q \ge 0, \quad g_2(q;K) = K - q \ge 0.$$

The optimizer and capacity multiplier are known to be $q^*(a,K) = \min\{a,K\}$ and $\lambda_2^*(a,K) = \max\{a-K,0\}$ (see Example 7.3.3 in Lecture 7). What do the envelope formulas give for the effects of capacity K and productivity a on the value, and do those effects agree with the known closed-form solution?

Solution.

The value function is available in closed form:

$$V(a,K) = \begin{cases} \tfrac{1}{2}a^2, & a \le K, \\ aK - \tfrac{1}{2}K^2, & a > K, \end{cases} \qquad \text{so} \qquad \frac{\partial V}{\partial K} = \begin{cases} 0, & a < K, \\ a - K, & a > K, \end{cases}$$

which is $\lambda_2^*(a,K)$, as part 3 of Theorem 8.2.1 predicts with $\mathrm{D}_K f = 0$ and $\mathrm{D}_K g_2 = 1$: the marginal value of capacity is zero while the capacity constraint is slack and positive once it binds. Differentiating in a instead gives $\partial V/\partial a = \min\{a,K\} = q^*(a,K)$ on both regimes, which is the direct effect $\mathrm{D}_a f = q$ evaluated at the optimum. On each regime the active set is constant and the branch $q^*(a,K)$ is differentiable, so the formulas apply regime by regime; the switch points belong to Section 8.2.5.

8.2.4 Behavior from optimized values

One classical corollary applies the envelope theorem to the theory of the firm.

Example 8.2.4 (Hotelling's lemma)**.** A firm chooses an input vector $x \in \mathbb{R}^n$ and produces output $y(x)$. The parameters are the output price p and input-price vector $w \in \mathbb{R}^n$, and profit is

$$\pi(x;p,w) = p\,y(x) - w \cdot x.$$

Suppose $x^*(p, w)$ is a differentiable branch of interior profit maximizers, with maximized profit $\Pi(p, w)$. What are the derivatives of Π with respect to p and each w_i?

Solution.

Part 1 of Theorem 8.2.1 gives

$$\frac{\partial \Pi}{\partial p} = y\big(x^*(p, w)\big), \qquad \frac{\partial \Pi}{\partial w_i} = -x_i^*(p, w) :$$

the derivative of maximized profit with respect to the output price is the optimal output, and with respect to an input price it is the negative of the optimal input use. Differentiating the value function recovers the firm's behavior without re-solving its problem. The same calculation with a budget constraint produces Roy's identity in consumer theory, and with a cost-minimization problem produces Shephard's lemma; both appear early in the first-year sequence, and both are Theorem 8.2.1 applied to a particular value function.

8.2.5 Limits of the formulas

Every formula in this section is local and conditional: it needs a differentiable branch through a regular solution, and for inequality problems it needs the active set to stay constant. The following example shows how the formulas fail where the active set changes.

Example 8.2.5 (No single marginal value at an active-set change)**.** The scalar choice variable is x, and the scalar parameter is $b > 0$. Consider

$$\max_x \; x \qquad \text{subject to} \qquad x \geq 0, \quad b - x \geq 0, \quad 1 - x \geq 0.$$

What are the optimizer and value, and how does the value derivative compare with the multiplier on the constraint $b - x \geq 0$ for $b < 1$, $b > 1$, and at the switch point $b = 1$?

Solution.

The feasible set is $[0, \min\{b, 1\}]$, and the objective is increasing in x, so the optimizer and value are $x^*(b) = V(b) = \min\{b, 1\}$. For $b < 1$ the constraint $b - x \geq 0$ binds alone and carries multiplier 1; for $b > 1$ it is slack and carries multiplier 0. At $b = 1$ both upper constraints are active, stationarity only requires the two multipliers to sum to one,

$$\lambda_b^* \in [0, 1], \qquad \lambda_{\text{cap}}^* = 1 - \lambda_b^*,$$

and the value function has left derivative 1 and right derivative 0. At the switch point the value function is not differentiable, so there is no single marginal value, and correspondingly the multiplier is not unique.

On either side of the switch, the formulas hold regime by regime. Example 8.2.2 shows the benign case: the active set changes at $\theta = 0$, yet the two regime values $\frac{1}{2}\theta^2$ and 0 join with matching derivative, so the value is differentiable even at $\theta = 0$. After an active-set change, the value may or may not remain differentiable, and this must be checked.

8.3 Fixed-point theorems

A fixed-point argument looks for a point that a map sends to itself. This section introduces several theorems that give conditions under which such points exist.

8.3.1 Fixed points and Brouwer's theorem

Definition 8.3.1 (Fixed point)**.** Let X be a set and $f : X \to X$. A point $x^* \in X$ is a *fixed point* of f if $f(x^*) = x^*$.

For a function of one variable, a fixed point is a crossing of the graph of f with the 45-degree line. Lecture 2 closed with our first fixed point theorem, Brouwer in $\mathbb{R}$: every continuous $f : [a, b] \to [a, b]$ has a fixed point. The proof applied the intermediate value theorem to $g(x) = f(x) - x$, which is nonnegative at a and nonpositive at b. Figure 8.3.1 shows the picture: a continuous graph that starts on or above the diagonal and ends on or below it must cross it.

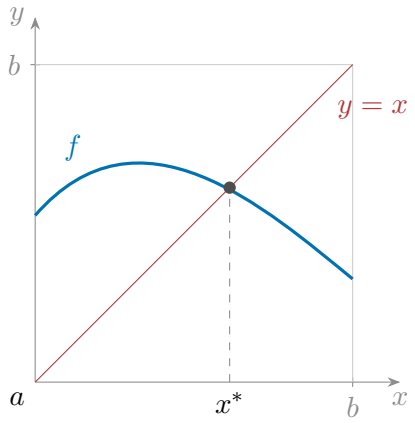


Figure 8.3.1: A continuous function from $[a, b]$ to itself. At x^* the graph of f meets the line $y = x$, so $f(x^*) = x^*$.

In $\mathbb{R}^n$ the interval is replaced by a compact convex set.

Theorem 8.3.2 (Brouwer)**.** *Let $D \subseteq \mathbb{R}^n$ be nonempty, compact, and convex, and let $f : D \to D$ be continuous. Then f has a fixed point.*

8.3.2 Set-valued maps and Kakutani's theorem

Definition 8.3.3 (Set-valued map)**.** A *set-valued map* (*correspondence*) $\Gamma : X \rightrightarrows Y$ assigns to every $x \in X$ a nonempty set $\Gamma(x) \subseteq Y$. A *fixed point* of $\Gamma : X \rightrightarrows X$ is a point x^* with $x^* \in \Gamma(x^*)$.

For correspondences, the continuity hypothesis of Brouwer's theorem is replaced by a condition on the set of pairs $\{(x, y) : x \in X,\ y \in \Gamma(x)\}$, the *graph* of Γ: the graph must be a closed set, in the sequential sense of Lecture 2. In words, if $x_k \to x$, $y_k \to y$, and $y_k \in \Gamma(x_k)$ for every k, then $y \in \Gamma(x)$.

Theorem 8.3.4 (Kakutani)**.** *Let $D \subseteq \mathbb{R}^n$ be nonempty, compact, and convex, and let $\Gamma : D \rightrightarrows D$ have convex values and a closed graph. Then Γ has a fixed point.*

Example 8.3.5 (Ties close the gap)**.** The variable is $x \in D = [0, 1]$; there are no parameters. Let $f : D \to D$ be the jump map $f(x) = 1$ for $x < \frac{1}{2}$ and $f(x) = 0$ for $x \geq \frac{1}{2}$, which has no fixed

point. Replace its value at the jump by the whole interval, defining $\Gamma : D \rightrightarrows D$ by

$$\Gamma(x) = \begin{cases} \{1\} & x < \tfrac{1}{2}, \\ [0,1] & x = \tfrac{1}{2}, \\ \{0\} & x > \tfrac{1}{2}. \end{cases}$$

Does Kakutani's theorem apply, and what is the fixed point? Why would replacing $\Gamma(\tfrac{1}{2}) = [0,1]$ by the two-point set $\{0,1\}$ destroy the conclusion?

Solution.

The jump map f has singleton values but not a closed graph: the pairs $(\tfrac{1}{2} - \tfrac{1}{k}, 1)$ lie in the graph and converge to $(\tfrac{1}{2}, 1)$, yet $f(\tfrac{1}{2}) = 0$. Under Γ, the graph gains the vertical segment $\{\tfrac{1}{2}\} \times [0,1]$ and becomes closed, every value is convex, and Theorem 8.3.4 applies. The fixed point is visible in Figure 8.3.2: the segment meets the diagonal at $x^* = \tfrac{1}{2}$, and indeed $\tfrac{1}{2} \in \Gamma(\tfrac{1}{2})$. If the value at the jump were the two-point set $\{0,1\}$ instead, the graph would still be closed, but the value would not be convex and no fixed point would exist. In a game, the interval is the set of mixtures available to a player who is indifferent; allowing the tie restores equilibrium existence.

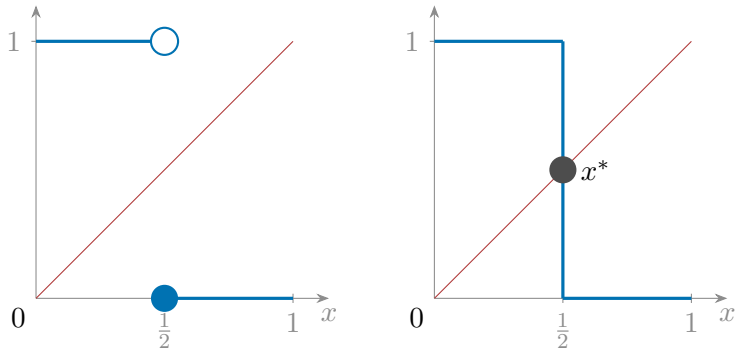


Figure 8.3.2: Left: the graph of the jump map passes over the diagonal without touching it. Right: the set-valued map whose value at $x = \tfrac{1}{2}$ is $[0,1]$; the added segment meets the diagonal at x^*.

8.3.3 The contraction mapping theorem

Brouwer and Kakutani establish existence only: they do not say whether the fixed point is unique, and they do not provide a way to compute it. The third theorem strengthens the hypothesis on the map and gives both uniqueness and an algorithm.

Definition 8.3.6 (Contraction)**.** Let $S \subseteq \mathbb{R}^n$. A map $f : S \to S$ is a *contraction with modulus* β if $0 \le \beta < 1$ and

$$\|f(x) - f(y)\| \le \beta \, \|x - y\| \qquad \text{for all } x, y \in S.$$

A contraction shrinks the distance between any two points to at most β times its original value. Every contraction is continuous: if $x_k \to x$, then $\|f(x_k) - f(x)\| \le \beta \|x_k - x\| \to 0$.

Theorem 8.3.7 (Contraction mapping theorem)**.** *Let $S \subseteq \mathbb{R}^n$ be nonempty and closed, and let $f : S \to S$ be a contraction with modulus β. Then f has exactly one fixed point $x^* \in S$. Moreover, from any starting point $x_0 \in S$, the iterates $x_{k+1} = f(x_k)$ converge to x^*, and*

$$\|x_k - x^*\| \leq \beta^k \|x_0 - x^*\|.$$

Remark 8.3.8 (Why the fixed point is unique). Uniqueness follows in two lines, and the argument uses $\beta < 1$ directly. If x^* and y^* are both fixed points, then

$$\|x^* - y^*\| = \|f(x^*) - f(y^*)\| \leq \beta \|x^* - y^*\|,$$

so $(1 - \beta)\|x^* - y^*\| \leq 0$. Since $\beta < 1$, this forces $\|x^* - y^*\| = 0$, that is, $x^* = y^*$. The error bound is the same inequality applied k times: $\|x_k - x^*\| = \|f(x_{k-1}) - f(x^*)\| \leq \beta\|x_{k-1} - x^*\|$.

Note that in Theorem 8.3.7 the set S need not be bounded or convex; the strong hypothesis on f replaces both.

Example 8.3.9 (A discounted sum as a fixed point)**.** Fix parameters $a \in \mathbb{R}$ and $\beta \in (0, 1)$. The fixed-point variable is $x \in S = \mathbb{R}$, and the map is $T : S \to S$ defined by $T(x) = a + \beta x$. Starting from $x_0 = 0$, does T have a unique fixed point, what are the iterates $x_{k+1} = T(x_k)$, and what economic object does their limit represent?

Solution.

We have $|T(x) - T(y)| = \beta|x - y|$, so T is a contraction with modulus β, and its fixed point solves $x = a + \beta x$:

$$x^* = \frac{a}{1 - \beta}.$$

Starting the iteration at $x_0 = 0$,

$$x_k = a\left(1 + \beta + \cdots + \beta^{k-1}\right) = a\,\frac{1 - \beta^k}{1 - \beta} \longrightarrow \frac{a}{1 - \beta},$$

since $\beta^k \to 0$ for $0 < \beta < 1$. The fixed point of "today's reward plus β times the continuation" is the value of receiving a in every period, discounted by β. In Lecture 9, the central operator has this form, with a function in place of the number x.

Theorem	The set	The map	Conclusion
Brouwer (Theorem 8.3.2)	nonempty, compact, convex	continuous $f : X \to X$	a fixed point exists
Kakutani (Theorem 8.3.4)	nonempty, compact, convex	convex values, closed graph	a fixed point exists
Contraction (Theorem 8.3.7)	nonempty, closed	contraction with modulus $\beta < 1$	exactly one fixed point

Table 8.3.1: The fixed-point theorems used in the first-year sequence. The first two assert existence only; the third requires the contraction property and also gives uniqueness and convergence of iteration.

Lecture 9

Dynamic Programming and the Bellman Equation

In many economic problems, the decision maker faces a sequence of choices, each affecting the next. A household that saves today has more to spend tomorrow; a firm that invests today produces more next year. Today's choice earns a payoff now and changes the situation for every later choice, so the choices at different dates cannot be studied separately. Dynamic programming is a tool for solving such problems.

9.1 Sequential problems and their recursive form

9.1.1 The growth problem

Time is discrete, $t = 0, 1, 2, \ldots$. An agent (decision-maker) starts with a stock of capital $k_0 > 0$. Capital k_t produces output $f(k_t)$ at date t, and output is either consumed, c_t, or carried forward as next period's capital, k_{t+1}. The agent values consumption streams by a discounted sum of one-period utilities and solves

$$\max_{\{c_t,\, k_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t) \qquad \text{subject to} \qquad c_t + k_{t+1} = f(k_t), \quad c_t \geq 0, \quad k_{t+1} \geq 0, \quad k_0 \text{ given.} \tag{9.1.1}$$

The number $\beta \in (0, 1)$ is the *discount factor*. The utility function $u : \mathbb{R}_+ \to \mathbb{R}$ is continuous, strictly increasing, and strictly concave, and $f : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous, increasing, and concave with $f(0) = 0$; both are differentiable on $(0, \infty)$. When we differentiate a value function we assume in addition that $u'(c) \to \infty$ as $c \to 0$, so that an agent never chooses zero consumption.

Problem (9.1.1) is written as a *sequential problem*: the unknown is the whole sequence of consumptions and capital stocks, $c_0, k_1, c_1, k_2, \ldots$, chosen at once. Since $c_t = f(k_t) - k_{t+1}$, the capital sequence determines the consumption sequence. We can therefore use $k_{t+1} \in [0, f(k_t)]$ as the choice at date t. Every date has the same structure, drawn in Figure 9.1.1:

$$\underbrace{k_t}_{\text{state}} \longrightarrow \underbrace{k_{t+1}}_{\text{choice}} \longrightarrow \underbrace{k_{t+1}}_{\text{tomorrow's state}} .$$

The agent inherits the *state* k_t at date t and chooses k_{t+1}, earning the current payoff $u(f(k_t) - k_{t+1})$. The *transition* makes today's choice tomorrow's state.

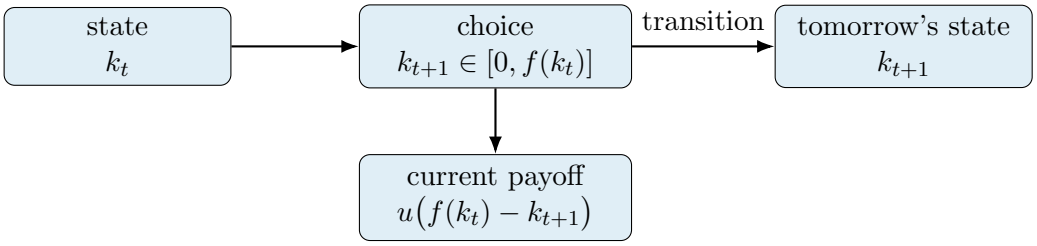


Figure 9.1.1: One date of the growth problem. The state k_t fixes the feasible choices $[0, f(k_t)]$; the choice k_{t+1} produces the current payoff and becomes the state at date $t+1$.

The state is not simply "a variable indexed by t." It is the information inherited from the past that matters for the choices still to be made. In the growth problem, once k_t is known, the feasible choices at date t and everything after depend on the past only through k_t: two histories that arrive at the same k_t face the same future. Which sequence of earlier consumptions produced k_t is irrelevant, so it is not part of the state. If instead the payoff at date t were $u(c_t, c_{t-1})$, so that yesterday's consumption changed the value of today's, then c_{t-1} would have to be carried in the state along with k_t. A state must therefore contain all information from the history that affects future feasible choices, payoffs, or transitions.

9.1.2 The principle of optimality and the Bellman equation

Suppose the agent begins a date with capital k and chooses k'. From tomorrow on the agent faces the same continuation problem as (9.1.1), with k' in place of k_0: the same technology, preferences, and infinite horizon. The highest utility attainable from tomorrow on therefore depends only on k'.

Definition 9.1.1 (Value function)**.** The *value function* $V : \mathbb{R}_+ \to \mathbb{R}$ assigns to each initial stock k the largest lifetime utility attainable from it,

$$V(k) = \max \sum_{t=0}^{\infty} \beta^t u(c_t) \qquad \text{subject to the constraints of (9.1.1) with } k_0 = k.$$

Whenever we write a maximum, we assume that the objective is finite and the maximum is attained. Bounded one-period utility is sufficient for convergence: if $|u(c)| \le M$ for every $c \ge 0$, every feasible stream has discounted utility of absolute value at most $M/(1-\beta)$ by the geometric-sum calculation of Lecture 8's Example 8.3.9. Attainment requires additional conditions supplied by the macroeconomics sequence.

Now split lifetime utility from k into today's payoff and everything after. If today's choice is k', then today yields $u(f(k) - k')$, and the most that tomorrow and later can yield is $V(k')$, discounted once because it starts one period later. An optimal choice of k' makes the total as large as possible:

$$V(k) = \max_{0 \le k' \le f(k)} \big\{ u\big(f(k) - k'\big) + \beta V(k') \big\} \qquad \text{for every } k \ge 0. \tag{9.1.2}$$

This is the *Bellman equation* of the growth problem. Its terms correspond to the pieces of Section 9.1.1:

- k is the current state;

- k' is the choice, and $[0, f(k)]$ is the set of feasible choices;
- $u(f(k) - k')$ is the current payoff;
- today's choice k' becomes tomorrow's state, which is why the same letter appears inside $V(\cdot)$;
- $V(k')$ is the *continuation value*, the value of behaving optimally from tomorrow on;
- β discounts the continuation value back to today.

The reasoning that produced (9.1.2) is Bellman's *principle of optimality*: after any feasible first choice, the best continuation is an optimal plan for the problem that starts from the resulting state. In particular, every tail of an optimal plan is optimal from the state reached at that date. Otherwise, replacing the tail by an optimal continuation would raise lifetime utility without changing any earlier payoff.

The unknown in (9.1.2) is a *function*. The equation must hold at every $k \geq 0$, and the same V appears on both sides, once evaluated at k and once at k'. Thus solving the Bellman equation means finding a function V that satisfies it at every state. For a fixed k and a fixed candidate V, the maximization inside the braces is a one-variable problem of the kind solved in Lecture 6. The recursive formulation repeats this one-period choice at every state.

The recursive formulation gives more than an equation for V. Under our assumptions, the value function of Definition 9.1.1 is the unique bounded solution of the Bellman equation when u is bounded. Moreover, a feasible capital sequence is optimal for (9.1.1) if and only if, at every date, its k_{t+1} attains the maximum in (9.1.2) at the state k_t. We use these facts without proof; Section 9.3.2 returns to the uniqueness claim.

9.1.3 Value function and policy function

Solving (9.1.2) produces two objects, and it is important to keep them apart. The first is the value function V. The second is the choice that attains the maximum at each state.

Definition 9.1.2 (Policy function)**.** The *policy function* $g : \mathbb{R}_+ \to \mathbb{R}_+$ assigns to each state the choice that attains the maximum in the Bellman equation,

$$g(k) \in \operatorname*{arg\,max}_{0 \leq k' \leq f(k)} \big\{ u\big(f(k) - k'\big) + \beta V(k') \big\}.$$

Under our assumptions the maximizer is unique at every k, so g is a function; we take this without proof. Consumption then follows from the constraint, $c = f(k) - g(k)$. The two functions answer different questions:

$$k \overset{V}{\longmapsto} \text{ lifetime utility from } k, \qquad\qquad k \overset{g}{\longmapsto} \text{ tomorrow's capital } k'.$$

The policy function describes the agents's choice at each state and generates the whole optimal path from k_0 alone:

$$k_1 = g(k_0), \qquad k_2 = g(k_1), \qquad k_3 = g(k_2), \ \ldots$$

The Bellman equation determines the value function. Its continuation term $\beta V(k')$ values tomorrow's capital in today's utility. The sequential problem (9.1.1) and the recursive problem (9.1.2) generate the same optimal paths; the recursive formulation describes them with a rule g rather than a list of numbers.

9.2 First-order and envelope conditions

The maximization inside the Bellman equation is a one-variable problem. Lecture 6 supplies its first-order condition, and Lecture 8's envelope theorem gives the derivative of the optimized value. In this section we assume, for $k > 0$, that V is differentiable and concave and that the maximizer $g(k)$ is interior, $0 < g(k) < f(k)$, and differentiable in k. Standard dynamic-programming results establish these properties under suitable hypotheses; we impose them rather than prove them.

9.2.1 The first-order condition

Fix k and let $c = f(k) - k'$. The objective in (9.1.2) is $u(f(k) - k') + \beta V(k')$, and Lecture 6's interior first-order condition, differentiating in k', gives

$$-u'(c) + \beta V'(k') = 0, \qquad \text{that is,} \qquad u'(c) = \beta V'(k'). \tag{9.2.1}$$

Since the objective is concave in k', the first-order condition is sufficient by Lecture 6's Theorem 6.3.2: it locates the maximizer $k' = g(k)$. In words, one more unit of capital carried into tomorrow costs $u'(c)$ in utility today and is worth $\beta V'(k')$, the discounted marginal value of capital tomorrow. At the optimum the two are equal.

To eliminate the unknown derivative $V'(k')$ from (9.2.1), we use the envelope condition.

9.2.2 The envelope condition and the Euler equation

The Bellman equation defines $V(k)$ as an optimized value with parameter k. By Lecture 8's Theorem 8.2.1, part 1, we hold the choice at its optimum and differentiate the objective only with respect to k. The parameter appears only in $u(f(k) - k')$, so

$$V'(k) = u'\big(f(k) - g(k)\big)\, f'(k) = u'(c)\, f'(k). \tag{9.2.2}$$

This is the *envelope condition*. One more unit of capital today raises output by $f'(k)$, and at the margin the extra output is worth $u'(c)$ whether it is consumed or saved, because the first-order condition has equated the two uses; the response of $g(k)$ contributes nothing to first order.

Equation (9.2.2) holds at every state, so it holds at tomorrow's state k' with tomorrow's consumption $c' = f(k') - g(k')$: $V'(k') = u'(c')f'(k')$. Substituting into (9.2.1),

$$u'(c_t) = \beta\, u'(c_{t+1})\, f'(k_{t+1}), \tag{9.2.3}$$

where we have restored time subscripts, $c = c_t$, $k' = k_{t+1}$, and $c' = c_{t+1}$. This is the *Euler equation* of the growth problem. Its left side is the utility cost of saving one more unit today. Its right side is the benefit: the unit becomes $f'(k_{t+1})$ units of output tomorrow, each worth $u'(c_{t+1})$, discounted once. Substitution has eliminated V' and left a relation between consumption at adjacent dates and the model's primitives.

The calculation has three steps that recur throughout macroeconomics:

Bellman equation $\longrightarrow$ first-order condition and envelope condition $\longrightarrow$ Euler equation

9.2.3 The same Euler equation from the sequential problem

We can also derive (9.2.3) directly from the sequential formulation. Substitute $c_t = f(k_t) - k_{t+1}$ into (9.1.1), so that the objective becomes

$$\sum_{t=0}^{\infty} \beta^t u\big(f(k_t) - k_{t+1}\big),$$

a function of the capital sequence alone. The variable k_{t+1} appears in exactly two terms of the sum, dated t and $t+1$. At an interior optimum, Lecture 6's first-order condition in the single variable k_{t+1}, all other capital stocks held fixed, gives

$$-\beta^t u'(c_t) + \beta^{t+1} u'(c_{t+1}) f'(k_{t+1}) = 0,$$

which is (9.2.3) after dividing by β^t. The economics is a one-period perturbation: consume one unit less at t, carry it forward, and consume the proceeds at $t+1$; at an optimum the change in lifetime utility is zero to first order.

The two derivations produce the same condition because the recursive and sequential formulations describe the same problem. The Bellman equation, its first-order condition, and its envelope condition reproduce the intertemporal optimality condition obtained directly from the sequential problem, using a one-period problem with the same form at every state. The Euler equation is a relation between three consecutive capital stocks (k_t, k_{t+1}, k_{t+2}); the initial condition k_0 pins down one end of the path, and the macroeconomics sequence supplies the condition that pins down the other, the *transversality condition*. We do not develop the transversality condition in this course.

9.2.4 Writing down a Bellman equation

Given a new dynamic problem, the procedure is: identify the state, the choice, the payoff, and the transition; then write today's payoff plus β times the value at tomorrow's state, maximized over today's choice. The consumption–saving problem also shows how two choices can parameterize the same recursion.

Example 9.2.1 (Consumption and saving at a fixed interest rate)**.** A consumer holds wealth $a_t \geq 0$ at the start of date t, consumes $c_t \in [0, a_t]$, and invests the rest at gross return $R > 0$, so that

$$a_{t+1} = R\,(a_t - c_t).$$

Lifetime utility is $\sum_{t=0}^{\infty} \beta^t u(c_t)$, and a_0 is given.

- *State.* Wealth a. Once a_t is known, nothing about the past matters for the future.
- *Choice.* Consumption $c \in [0, a]$.
- *Payoff.* $u(c)$.
- *Transition.* $a' = R(a - c)$.

The Bellman equation is

$$V(a) = \max_{0 \leq c \leq a} \big\{ u(c) + \beta V\big(R(a-c)\big) \big\}.$$

There is more than one convenient way to parameterize the choice. Since $c = a - a'/R$, the consumer may instead choose tomorrow's wealth $a' \in [0, Ra]$:

$$V(a) = \max_{0 \le a' \le Ra} \Big\{ u\Big(a - \frac{a'}{R}\Big) + \beta V(a') \Big\}.$$

The two equations have the same solution V; the economic problem determines the recursion, and the notation is a matter of convenience. In the growth problem we chose k' for the same reason: it keeps the continuation value $V(k')$ simple.

Under the assumptions of Section 9.2, the first-order condition in c from the first formulation is

$$u'(c) = \beta R\, V'(a'),$$

and the envelope condition, differentiating $u(c) + \beta V(R(a-c))$ in the parameter a at the optimal c, is

$$V'(a) = \beta R\, V'(a') = u'(c).$$

Applying the envelope condition at tomorrow's state, $V'(a') = u'(c')$, and substituting into the first-order condition gives the Euler equation

$$u'(c_t) = \beta R\, u'(c_{t+1}).$$

The consumer equates the marginal utility of a unit consumed today with that of the R units it would become tomorrow, discounted. If $\beta R = 1$, consumption is constant over time; if $\beta R > 1$, marginal utility falls and consumption rises. The first-year consumption theory sequence uses this equation as its starting point.

9.3 Why the recursion works: finite horizons and the fixed point

Finite horizons make the recursive logic explicit because they provide a last date from which to work backward. For an infinite horizon, the contraction mapping theorem determines the value function without a last date.

9.3.1 Finite horizons and backward induction

Suppose the world ends after date T: the agent solves

$$\max \sum_{t=0}^{T} \beta^t u(c_t) \qquad \text{subject to} \qquad c_t + k_{t+1} = f(k_t), \quad c_t, k_{t+1} \ge 0, \quad k_0 \text{ given}, \tag{9.3.1}$$

with no use for capital after date T. The value function now depends on the date as well as on the state, because the agent at date t has $T - t + 1$ decision dates remaining. Write $V_t(k)$ for the largest utility from date t on when the date-t state is k.

At the last date the problem is trivial. Capital left over is wasted and u is increasing, so the agent consumes all output:

$$V_T(k) = u\big(f(k)\big).$$

At date $T-1$ the agent chooses k_T knowing its continuation value, $V_T(k_T)$:

$$V_{T-1}(k) = \max_{0 \le k' \le f(k)} \big\{ u\big(f(k) - k'\big) + \beta V_T(k') \big\}.$$

Once V_T is known, this is a one-variable problem of Lecture 6, and solving it at every k gives the function V_{T-1}. Then V_{T-2} is obtained from V_{T-1} in the same way, and so on down to V_0:

$$V_T \longrightarrow V_{T-1} \longrightarrow V_{T-2} \longrightarrow \cdots \longrightarrow V_0.$$

Proposition 9.3.1 (Finite-horizon Bellman recursion)**.** *For $t = T-1, T-2, \dots, 0$ and every $k \geq 0$,*

$$V_t(k) = \max_{0 \leq k' \leq f(k)} \bigl\{ u\bigl(f(k) - k'\bigr) + \beta V_{t+1}(k') \bigr\}, \qquad V_T(k) = u\bigl(f(k)\bigr),$$

and a feasible capital sequence solves (9.3.1) *if and only if at every date $t \leq T-1$ its k_{t+1} attains the maximum at the state k_t.*

Proof (optional). Fix t and k, and write $W(k)$ for the right-hand side of the display. Every feasible plan from date t at state k begins with some $k' \in [0, f(k)]$ and continues with a feasible plan from date $t+1$ at state k', whose utility from $t+1$ on is at most $V_{t+1}(k')$. So every plan earns at most $u(f(k) - k') + \beta V_{t+1}(k') \leq W(k)$, and $V_t(k) \leq W(k)$. Conversely, let k' attain the maximum and follow it with an optimal plan from date $t+1$ at state k'; that plan exists by induction downward from date T, where the optimal plan is to consume everything. The combined plan is feasible and earns $W(k)$, so $V_t(k) \geq W(k)$. The two inequalities give the recursion. The same comparison shows that a plan achieves $V_t(k)$ exactly when its first choice attains the maximum and its continuation is optimal. This proves the biconditional. □

Computing $V_T, V_{T-1}, \dots, V_0$ in that order is called *backward induction*: it is Lecture 1's induction, run from the last date toward the first. Each step is a static problem, and each maximum is attained by the Weierstrass theorem of Lecture 2 when the objective is continuous, since $[0, f(k)]$ is compact. The principle of optimality is also established by the argument: an optimal plan's continuation from any date is optimal for the problem starting from the state reached on that date.

Example 9.3.2 (Two dates of cake eating)**.** Let $u(c) = \sqrt{c}$ and $f(k) = k$. Capital does not reproduce in this special case: k is a stock of cake, and the agent divides it between consumption now and cake carried to the next date. Fix an upper bound $\bar{k} > 0$ and take the state space to be $S = [0, \bar{k}]$, which is invariant because every feasible choice satisfies $0 \leq k' \leq k$. Let $T = 1$, so there are two dates. At date 1 the agent consumes everything:

$$V_1(k) = \sqrt{k}.$$

At date 0 the recursion gives

$$V_0(k) = \max_{0 \leq k' \leq k} \bigl\{ \sqrt{k - k'} + \beta \sqrt{k'} \bigr\}.$$

For $k > 0$ the objective is strictly concave and its one-sided slopes at the two endpoints point toward the interior. Its first-order condition is

$$\frac{1}{2\sqrt{k - k'}} = \frac{\beta}{2\sqrt{k'}} \qquad \text{and hence} \qquad k' = \frac{\beta^2}{1 + \beta^2}\, k,$$

the unique maximizer by Lecture 6's Theorem 6.3.1. The agent carries the fraction $\beta^2/(1+\beta^2)$ of the cake to date 1 and consumes the rest. Substituting back gives

$$V_0(k) = \sqrt{1 + \beta^2}\, \sqrt{k}.$$

Thus V_1 and V_0 are both a constant multiple of $\sqrt{k}$.

The pattern continues. If $V_{t+1}(k) = B_{t+1}\sqrt{k}$ with $B_{t+1} > 0$, the same calculation gives

$$k' = \frac{\beta^2 B_{t+1}^2}{1+\beta^2 B_{t+1}^2}\,k, \qquad V_t(k) = B_t\sqrt{k}, \qquad B_t^2 = 1+\beta^2 B_{t+1}^2, \qquad B_T = 1.$$

Consequently,

$$B_t^2 = 1+\beta^2+\beta^4+\cdots+\beta^{2(T-t)}.$$

As the number of remaining dates grows, B_t increases to $1/\sqrt{1-\beta^2}$, and the fraction carried forward increases to β^2. Section 9.4 obtains these limits by solving the infinite-horizon problem directly.

9.3.2 The infinite horizon as a fixed point

With no last date there is no V_T from which to start. Example 9.3.2 shows the coefficient on $\sqrt{k}$ and the fraction carried forward approaching limits as dates are added. In a *stationary* problem, one whose payoff, feasible set, and transition are the same at every date, the infinite-horizon value does not depend on the calendar. The recursion of Proposition 9.3.1 with $V_t = V_{t+1} = V$ is the Bellman equation (9.1.2). Whether an infinite-horizon V exists, and whether it is unique, is a fixed-point question. Under bounded rewards, a contraction argument on the space of bounded functions answers it.

To state the contraction result, write a stationary problem in general notation. It has a state s in a state space S, a set $\Gamma(s)$ of feasible actions at s, a reward $r(s,a)$, a transition $s' = F(s,a)$, and a discount factor $\beta \in (0,1)$; the growth problem is the case $s = k$, $\Gamma(k) = [0, f(k)]$, $a = k'$, $r(k,k') = u(f(k)-k')$, and $F(k,k') = k'$. Table 9.3.1 lists the correspondence. The Bellman equation is

$$V(s) = \sup_{a\in\Gamma(s)} \big\{ r(s,a) + \beta V\big(F(s,a)\big) \big\} \qquad \text{for every } s \in S,$$

written with a supremum so that it makes sense before anything is known about attainment.

Object	Growth problem	General stationary problem
State	capital k	$s \in S$
Feasible choices	$k' \in [0, f(k)]$	$a \in \Gamma(s)$
Current payoff	$u(f(k)-k')$	$r(s,a)$
Transition	k'	$s' = F(s,a)$
Bellman equation	$V(k) = \max\{u(f(k)-k') + \beta V(k')\}$	$V(s) = \sup\{r(s,a) + \beta V(F(s,a))\}$
Policy function	$k' = g(k)$	$a = g(s)$

Table 9.3.1: The pieces of the growth problem and their names in a general stationary dynamic problem. In the growth problem the transition is the identity in the choice, which is why k' serves as both the choice and tomorrow's state.

Definition 9.3.3 (Bellman operator)**.** Let $B(S)$ be the set of bounded functions $v : S \to \mathbb{R}$ with the sup norm $\|v\|_\infty = \sup_{s\in S}|v(s)|$. The *Bellman operator* $\mathcal{T}$ sends a candidate value function v to the function

$$(\mathcal{T}v)(s) = \sup_{a\in\Gamma(s)} \big\{ r(s,a) + \beta\, v\big(F(s,a)\big) \big\}.$$

The number $(\mathcal{T}v)(s)$ is the supremum value of a one-period choice at s when v values tomorrow's state. The Bellman equation says $\mathcal{T}V = V$: the value function is a *fixed point* of $\mathcal{T}$. Backward induction is iteration of $\mathcal{T}$, since $V_t = \mathcal{T}V_{t+1}$ in Proposition 9.3.1. The operator $\mathcal{T}$ is the function-valued version of the map $T(x) = a + \beta x$ in Lecture 8's Example 8.3.9: today's reward plus β times the continuation.

Theorem 9.3.4 (The Bellman operator is a contraction)**.** *Suppose $\Gamma(s)$ is nonempty for every $s \in S$ and there is $M < \infty$ with $|r(s,a)| \le M$ whenever $a \in \Gamma(s)$. Then $\mathcal{T}$ maps $B(S)$ into $B(S)$, and for all $v, w \in B(S)$,*

$$\|\mathcal{T}v - \mathcal{T}w\|_\infty \le \beta \, \|v - w\|_\infty.$$

Proof (optional). Every number whose supremum defines $(\mathcal{T}v)(s)$ lies between $-M - \beta\|v\|_\infty$ and $M + \beta\|v\|_\infty$, so $|(\mathcal{T}v)(s)| \le M + \beta\|v\|_\infty$ for every s and $\mathcal{T}v \in B(S)$. Now fix s. For every $a \in \Gamma(s)$ the definition of the sup norm gives $v(F(s,a)) \le w(F(s,a)) + \|v - w\|_\infty$, hence

$$r(s,a) + \beta\, v\big(F(s,a)\big) \le r(s,a) + \beta\, w\big(F(s,a)\big) + \beta\|v - w\|_\infty.$$

The inequality holds action by action, so it survives the supremum over $a \in \Gamma(s)$: $(\mathcal{T}v)(s) \le (\mathcal{T}w)(s) + \beta\|v - w\|_\infty$. Exchanging v and w gives the reverse bound, so $|(\mathcal{T}v)(s) - (\mathcal{T}w)(s)| \le \beta\|v - w\|_\infty$ at every s, and taking the supremum over s finishes the proof. □

The space $B(S)$ is complete in the sup norm, a fact we use without proof. The contraction mapping theorem therefore gives three conclusions. The Bellman equation has *exactly one* bounded solution. Starting from any bounded v_0, even $v_0 = 0$, the iterates $v_{n+1} = \mathcal{T}v_n$ converge to it in sup norm, with the worst-case error shrinking by the factor β each time. This procedure is called *value function iteration*. For $v_0 = 0$, the iterate v_n is the value of the problem with n dates remaining and no terminal value. Thus the infinite-horizon value is the limit of finite-horizon values, as in Example 9.3.2. More generally, value iteration can start from any bounded function, and the effect of that starting function vanishes at the geometric rate β.

Discounting supplies the contraction modulus β. As $\beta \to 1$, the worst-case convergence bound becomes arbitrarily slow; at $\beta = 1$, Lecture 8's map $T(x) = a + x$ has no fixed point when $a \neq 0$: an undiscounted constant stream of nonzero rewards has no finite value. The theorem also requires a bounded reward. In Example 9.3.2, $S = [0, \bar{k}]$ and $0 \le \sqrt{k - k'} \le \sqrt{\bar{k}}$, so the hypothesis holds and the finite-horizon values converge uniformly to the unique bounded fixed point. If instead the state space were all of $\mathbb{R}_+$, the same reward would be unbounded above and this theorem would no longer apply.

9.4 A Bellman equation solved by hand

Value function iteration is useful for numerical work. When the primitives are special enough, the Bellman equation can be solved analytically by *guess and verify*: conjecture that V has a particular functional form with unknown constants, substitute the conjecture into the Bellman equation, carry out the maximization, and check whether the result has the conjectured form again; if it does, matching constants yields equations for them. The guess proposes a candidate; substitution verifies that the candidate is a fixed point. A separate theorem is still needed to show that this fixed point is the value function and is unique in the relevant class.

Example 9.4.1 (Square-root cake eating)**.** Continue the cake-eating problem of Example 9.3.2, with $u(c) = \sqrt{c}$, $f(k) = k$, and $S = [0, \bar{k}]$. Given $k_0 = k \in S$, the agent solves

$$\max_{\{c_t, k_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \sqrt{c_t} \quad \text{subject to} \quad c_t + k_{t+1} = k_t, \qquad c_t, k_{t+1} \geq 0.$$

Its Bellman equation is

$$V(k) = \max_{0 \leq k' \leq k} \{\sqrt{k - k'} + \beta V(k')\}.$$

The finite-horizon value functions were multiples of $\sqrt{k}$, so we conjecture

$$V(k) = B\sqrt{k}$$

with $B > 0$ and look for a constant B that makes the equation hold at every $k \in [0, \bar{k}]$.

The maximization. With the conjecture substituted, the objective $\sqrt{k - k'} + \beta B\sqrt{k'}$ is strictly concave in k'. For $k > 0$ its unique maximizer is interior and satisfies

$$\frac{1}{2\sqrt{k - k'}} = \frac{\beta B}{2\sqrt{k'}},$$

so

$$k' = \frac{\beta^2 B^2}{1 + \beta^2 B^2}\, k, \qquad k - k' = \frac{k}{1 + \beta^2 B^2}.$$

For $k = 0$, the only feasible choice is $k' = 0$, which the same formula gives.

The verification. Substituting the maximizer back, the right-hand side of the Bellman equation becomes

$$\frac{\sqrt{k}}{\sqrt{1 + \beta^2 B^2}} + \frac{\beta^2 B^2 \sqrt{k}}{\sqrt{1 + \beta^2 B^2}} = \sqrt{1 + \beta^2 B^2}\, \sqrt{k}.$$

The conjectured form is preserved. Matching the coefficient on $\sqrt{k}$ gives

$$B = \sqrt{1 + \beta^2 B^2} \qquad \Longleftrightarrow \qquad B = \frac{1}{\sqrt{1 - \beta^2}}.$$

The value and policy. The value function and its implied policy and consumption are therefore

$$V(k) = \frac{\sqrt{k}}{\sqrt{1 - \beta^2}}, \qquad g(k) = \beta^2 k, \qquad c = (1 - \beta^2)k.$$

The agent carries the fixed fraction β^2 of the remaining cake to the next date. These are the limits found in Example 9.3.2. The optimal stock path is $k_t = \beta^{2t} k_0$, so the remaining cake converges to zero.

Substituting the policy into the Bellman equation verifies the solution directly:

$$\sqrt{k - g(k)} + \beta V\big(g(k)\big) = \sqrt{(1 - \beta^2)k} + \frac{\beta\sqrt{\beta^2 k}}{\sqrt{1 - \beta^2}} = \frac{\sqrt{k}}{\sqrt{1 - \beta^2}} = V(k).$$

The Euler check. Since $f'(k) = 1$ and $c_{t+1} = \beta^2 c_t$,

$$\beta u'(c_{t+1}) f'(k_{t+1}) = \frac{\beta}{2\sqrt{\beta^2 c_t}} = \frac{1}{2\sqrt{c_t}} = u'(c_t).$$

Finally, V is bounded on $[0, \bar{k}]$ and the reward is bounded by $\sqrt{\bar{k}}$. The contraction theorem therefore shows that this fixed point is the unique bounded solution of the Bellman equation, and value function iteration from the finite-horizon problems converges to it.