MATHEMATICAL ECONOMICS
Lecture Notes[1]

Alexander W. Richter[2]

Research Department
Federal Reserve Bank of Dallas

Department of Economics
Southern Methodist University

August 2017

[1]I am especially grateful to Juergen Jung, Mike Treuden, and Nathaniel Throckmorton for their tremendous contributions to these notes. I also thank Michael Kaganovich and Eric Leeper for their guidance and for giving me the opportunity to develop and teach this course during graduate school. Several classes of Indiana University and Auburn University graduate students provided suggestions and corrections. Comments are welcome and all remaining errors are mine. The views expressed in these notes are my own and do not necessarily reflect the views of the Federal Reserve Bank of Dallas or the Federal Reserve System.

[2]Correspondence: Research Department, Federal Reserve Bank of Dallas, 2200 N Pearl St, Dallas, TX 75201, USA. Phone: +1(922)922-5360. E-mail: alex.richter@dal.frb.org or arichter@smu.edu.

Contents

List of Figures

Preface

These notes are intended for a one-semester course in mathematical economics. The goal of this course is to help prepare students for the mathematical rigor of graduate economics by providing a balance between theory and application. There are dozens of examples throughout the notes, which demonstrate many of the theoretical concepts and theorems. Below is a short list of the notation used thoughout these notes.

Symbol	Technical Description	Translation
$\forall$	universal qualification	for all
$\exists$	existential qualification	there exists
$\wedge$	logical conjunction	and
$\vee$	logical disjunction	or
$\Rightarrow$	material implication	implies/if, then
$\Leftrightarrow$	material equivalence	if and only if (iff)
$\equiv$ or :=	definition	is defined as/is equivalent to
$\in$	set membership	in/element of
$\mid$ or :	set restriction	such that/given that
$\subseteq$	subset	contained in
$\mathbb{N}$	natural numbers	set of positive integers
$\mathbb{Z}$	set of all integers	positive, negative or zero
$\mathbb{Q}$	set of all rational numbers	set of all proper and improper fractions
$\mathbb{R}$	set of all real numbers	all rational and irrational numbers

Chapter 1

Mathematical Preliminaries

1.1 Single-Variable Calculus

1.1.1 Limits of Functions

Definition 1.1.1 (Limit of a Function)**.** *Let f be a function defined on some open interval that contains the number a, except possibly a itself. Then we say that the limit of $f(x)$ as x approaches a is L, and we write*

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

That is, if for every number $\varepsilon > 0$ there is a number $\delta > 0$ such that

$$|f(x) - L| < \varepsilon \quad whenever \quad 0 < |x - a| < \delta.$$

Less formally, if anytime x is near a, the function f is near L, then we say that the limit of $f(x)$ as x approaches a is L (see figure 1.1). Note that logically the statement "q whenever p" is equivalent to "if p, then q", where it is customary to refer to p as the hypothesis and q as the conclusion. The related implication $\sim q \Rightarrow \sim p$ is called the *contrapositive*.

Example 1.1.1. Using the definition of a limit, prove the following:

(a) $\lim_{x \to 1}(2x^2 - 3x + 1)/(x - 1) = 1$

Solution: Let $\varepsilon > 0$, suppose $|x - 1| < \delta$, and choose $\delta = \varepsilon/2 > 0$. Then

$$|f(x) - 1| = \left| \frac{2x^2 - 3x + 1}{x - 1} - 1 \right| = 2|x - 1| < 2\delta = \varepsilon.$$

(b) $\lim_{x \to 5} x^2 - 3x + 1 = 11$

Solution: Let $\varepsilon > 0$, suppose $|x - 5| < \delta$, and choose $\delta = \min\{1, \varepsilon/8\}$. We can write

$$|f(x) - 11| = |x^2 - 3x - 10| = |(x - 5)(x + 2)|.$$

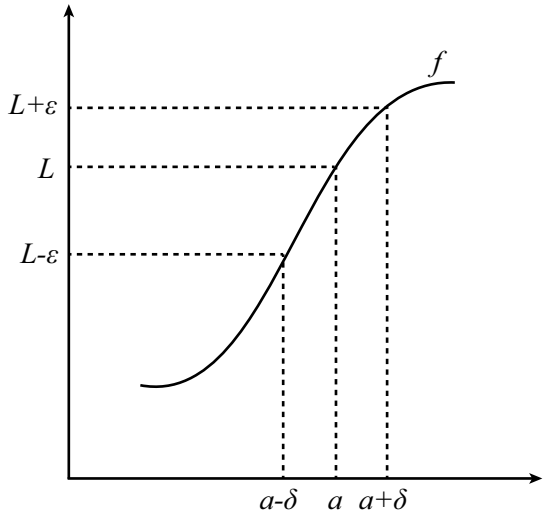
To make this small, we need a bound on the size of $x + 2$ when x is "close" to 5. For example, if we arbitrarily require that $|x - 5| < 1$, then

$$|x + 2| = |x - 5 + 7| \leq |x - 5| + 7 < 8.$$

To make $f(x)$ within ε units of 11, we shall want to have $|x + 2| < 8$ and $|x - 5| < \varepsilon/8$. Thus, under the above definition of δ

$$|f(x) - 11| = |(x - 5)(x + 2)| < 8\delta \leq \varepsilon.$$

Figure 1.1: Definition of a Limit



(c) $\lim_{x\to -2} x^2 + 2x + 7 = 7$

Solution: Let $\varepsilon > 0$, suppose $|x + 2| < \delta$, and choose $\delta = \min\{1, \varepsilon/3\}$. We can write

$$|f(x) - 7| = |x^2 + 2x| = |x(x + 2)|.$$

To make this small, we need a bound on the size of x when x is "close" to -2. For example, if we arbitrarily require that $|x + 2| < 1$, then

$$|x| - |2| \leq |x + 2| < 1,$$

since $|a| - |b| \leq ||a| - |b|| \leq |a \pm b|$ by the triangle inequality. Thus, $|x| < 3$, which implies

$$|f(x) - 7| = |x(x + 2)| < 3\delta \leq \varepsilon.$$

(d) $\lim_{x\to 2} x^3 = 8$

Solution: Let $\varepsilon > 0$, suppose $|x - 2| < \delta$, and choose $\delta = \min\{1, \frac{\varepsilon}{19}\}$. Then

$$\begin{aligned} |f(x) - 8| &= |x^3 - 8| = |x - 2||x^2 + 2x + 4| \\ &= |x - 2| \cdot |(x - 2)^2 + 6(x - 2) + 12| \\ &< \delta(\delta^2 + 6\delta + 12) \\ &\leq \frac{\varepsilon}{19}(1 + 6 + 12) \\ &= \varepsilon. \end{aligned}$$

1.1.2 Definition of a derivative and Tangent Lines

The derivative of y with respect to x at a is, geometrically, the slope of the tangent line to the graph of f at a. The slope of the tangent line is very close to the slope of the line through $(a, f(a))$ and a nearby point on the graph, for example $(a + h, f(a + h))$. These lines are called secant lines. Thus, a value of h close to zero will give a good approximation to the slope of the tangent line, and smaller values (in absolute value) of h will, in general, give better approximations. The slope of the secant line is the difference between the y values of these points divided by the difference between the x values. The following definition provides the more customary definition of a derivative.

Definition 1.1.2 (Derivative)**.** *The function $f\colon \mathbb{R} \to \mathbb{R}$ is differentiable at a if*

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

exists. This limit is called the derivative of f at a and is written $f'(a)$.

Definition 1.1.3 (Tangent Line)**.** *The tangent line to the curve $y = f(x)$ at the point $P\,(a, f(a))$ is the line through P with slope m, provided that the limit exists.*

Remark 1.1.1. In general, when a function f of one variable is differentiable at a point a, the equation of a tangent line to the graph of f at a is:

$$y = f(a) + f'(a)(x - a)$$

Example 1.1.2. Find an equation of the tangent line to the hyperbola $y = 3/x$ at the point $(3, 1)$ using the above definition of a derivative.

Solution: The slope is given by

$$\begin{aligned} m &= \lim_{h\to 0} \frac{f(3+h) - f(3)}{h} = \lim_{h\to 0} \frac{\frac{3}{3+h} - 1}{h} \\ &= \lim_{h\to 0} -\frac{1}{3+h} = -\frac{1}{3}. \end{aligned}$$

Therefore, an equation of the tangent line at the point $(3, 1)$ is

$$y = 1 - (x - 3)/3 \quad \to \quad x + 3y - 6 = 0.$$

Example 1.1.3. Find an equation of the tangent line to the curve $y = (x - 1)/(x - 2)$ at the point $(3, 2)$ using the above definition of a derivative.

Solution: The slope is given by

$$\begin{aligned} m &= \lim_{h\to 0} \frac{f(3+h) - f(3)}{h} = \lim_{h\to 0} \frac{\frac{h+2}{h+1} - 2}{h} \\ &= \lim_{h\to 0} -\frac{1}{h+1} = -1. \end{aligned}$$

Therefore, an equation of the tangent line at the point $(3, 2)$ is

$$y - 2 = -(x - 3) \quad \to \quad x + y = 5.$$

1.1.3 Properties of the Differential

Theorem 1.1.1 (Classic properties)**.** *Let I be an interval in $\mathbb{R}$ and suppose that $f : I \to \mathbb{R}$ and $g : I \to \mathbb{R}$ are differentiable at $a \in I$. Then*

(i) If $k \in \mathbb{R}$, then the function kf is differentiable at a and

$$(kf)'(a) = k \cdot f'(a)$$

(ii) The function $f + g$ is differentiable at a and

$$(f + g)'(a) = f'(a) + g'(a)$$

(iii) (Product Rule) The function fg is differentiable at a and

$$(fg)'(a) = f(a)g'(a) + g(a)f'(a)$$

(iv) (Quotient Rule) If $g(a) \neq 0$, then f/g is differentiable at a and

$$\left(\frac{f}{g}\right)'(a) = \frac{g(a)f'(a) - f(a)g'(a)}{[g(a)]^2}.$$

Theorem 1.1.2 (Chain Rule)**.** *Let I and J be intervals in $\mathbb{R}$, $f : I \to \mathbb{R}$, and $g : J \to \mathbb{R}$, where $f(I) \subseteq J$, and let $a \in I$. If f is differentiable at a and g is differentiable at $f(a)$, then the composite function $g \circ f$ is differentiable at a and*

$$(g \circ f)'(a) = g'(f(a)) \cdot f'(a).$$

Example 1.1.4. Using the properties of the differential, differentiate the following, where a, p, q, and b are constants

(a) $y = f(x) = \frac{1}{(x^2+x+1)^5}$

Solution: $f'(x) = \frac{-5(2x+1)}{(x^2+x+1)^6}$

(b) $y = f(x) = \sqrt{1+\sqrt{1+\sqrt{x}}}$

Solution: $f'(x) = 1/(8y\sqrt{x(1+\sqrt{x})})$

(c) $y = f(x) = x^a(px+q)^b$

Solution: $f'(x) = y[bp/(px+q) + a/x]$

Example 1.1.5. If $a(t)$ and $b(t)$ are positive-valued differentiable functions of t, and if A, α, β are constants, find expressions for $\frac{\dot{x}}{x} = \frac{dx/dt}{x}$, where

(a) $x = A\left\{[a(t)]^\alpha + [b(t)]^\beta\right\}^{\alpha+\beta}$

Solution: $\frac{\dot{x}}{x} = \frac{(\alpha+\beta)(\alpha a(t)^{\alpha-1}\dot{a}+\beta(t)^{\beta-1}\dot{b})}{[a(t)]^\alpha+[b(t)]^\beta}$

(b) $x = A[a(t)]^\alpha[b(t)]^\beta$

Solution: $\frac{\dot{x}}{x} = \frac{\beta\dot{b}}{b(t)} + \frac{\alpha\dot{a}}{a(t)}$

Example 1.1.6. If $F(x) = f(x^n g(x))$, find a formula for $F'(x)$.
Solution: $F'(x) = f'(x^n g(x))(x^n g'(x) + g(x)nx^{n-1})$

Theorem 1.1.3 (L'Hospital's Rule)**.** *Suppose f and g are differentiable and $g'(x) \neq 0$ near a (expect possibly at a). Suppose that*

$$\lim_{x\to a} f(x) = 0 \quad \textit{and} \quad \lim_{x\to a} g(x) = 0$$
$$\textit{or that } \lim_{x\to a} f(x) = \pm\infty \quad \textit{and} \quad \lim_{x\to a} g(x) = \pm\infty.$$

Then

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}$$

if the limit on the right side exists.

Example 1.1.7. Calculate $\lim_{x\to\infty} \frac{\ln x}{\sqrt[3]{x}}$.

Solution:

$$\lim_{x\to\infty} \frac{\ln x}{\sqrt[3]{x}} = \lim_{x\to\infty} \frac{1/x}{x^{-2/3}/3} = \lim_{x\to\infty} \frac{3}{\sqrt[3]{x}} = 0.$$

Example 1.1.8. Calculate $\lim_{x\to\infty} e^x/x^2$.

Solution:

$$\lim_{x\to\infty} e^x/x^2 = \lim_{x\to\infty} e^x/2x = \lim_{x\to\infty} e^x/2 = \infty.$$

Theorem 1.1.4 (Inverse Function Theorem)**.** *Suppose that f is differentiable on an interval I and $f'(x) \neq 0$ (no local maxima or minima) for all $x \in I$. Then f is injective, f^{-1} is differentiable on $f(I)$, and*

$$\frac{\partial x}{\partial y} = (f^{-1})'(y) = \frac{1}{f'(x)} = \frac{1}{\partial y/\partial x},$$

where $y = f(x)$.

Example 1.1.9. Let $n \in \mathbb{N}$ and $y = f(x) = x^{1/n}$ for $x > 0$. Then f is the inverse of the function $g(y) = y^n$. Use Theorem 1.1.4 to verify the familiar derivative formula for f : $f'(x) = (1/n)x^{1/n-1}$.

Solution:

$$f'(x) = \frac{1}{(f^{-1})'(y)} = \frac{1}{g'(y)} = \frac{1}{ny^{n-1}} = \frac{1}{n(x^{1/n})^{n-1}} = \frac{1}{n}x^{1/n-1}.$$

Example 1.1.10. Consider the following function:

$$y = f(\theta) = -\frac{\theta}{(1-\theta)\log(1-\theta)}.$$

Use Theorem 1.1.4, find $df^{-1}(y)/dy$.

Solution:

$$\frac{df^{-1}(y)}{dy} = \frac{d\theta}{dy} = \frac{1}{dy/d\theta} = -\frac{1}{\left[\frac{\log(1-\theta)+\theta}{(1-\theta)^2\log^2(1-\theta)}\right]} = -\frac{(1-\theta)^2\log^2(1-\theta)}{\log(1-\theta)+\theta}.$$

1.1.4 Single-Variable Maximization

Proposition 1.1.1 (First Derivative Test)**.** *Suppose c is a critical point of a continuous function f.*

(a) If f' changes from positive to negative at c, then f has a local maximum at c.

(b) If f' changes from negative to positive at c, then f has a local minimum at c.

(c) If f' does not change sign at c, then f has no local maximum or minimum at c.

If the sign of $f'(x)$ changes from positive to negative (negative to positive) at c, f is increasing (decreasing) to the left of c and decreasing (increasing) to the right of c. If follows that f has a local maximum (minimum) at c.

Proposition 1.1.2 (Second Derivative Test)**.** *Suppose f'' is continuous near c.*

(a) If $f'(c) = 0$ and $f''(c) > 0$, then f has local minimum at c.

(b) If $f'(c) = 0$ and $f''(c) < 0$, then f has local maximum at c.

If $f''(c) > 0 (< 0)$ near c, f is concave upward (downward) near c. Thus, the graph of f lies above (below) its horizonal tangent at c and so f has a local minimum (maximum) at c.

Example 1.1.11. The height of a plant after t months is given by $h(t) = \sqrt{t} - t/2$, $t \in [0, 3]$. At what time is the plant at its highest?

Solution: The first order condition is given by

$$h'(t) = \frac{1}{2\sqrt{t}} - \frac{1}{2} \overset{set}{=} 0 \Rightarrow t^* = 1,$$

where $h(1) = 1 - 1/2 = 1/2$. Since $h''(t) = -1/(4t^{3/2}) < 0$, $t^* = 1$ is a local maximum. Also, $h(0) = 0$, $h(3) = \sqrt{3} - 3/2 \approx 0.27$. Therefore, $t^* = 1$ is an absolute maximum.

Example 1.1.12. A sports club plans to charter a plane. The charge for 60 passengers is \$800 each. For each additional person above 60, all travelers get a discount of \$10. The plane can take at most 80 passengers.

(a) If $60 + x$ passengers fly, what is the total cost?

Solution: $TC(x) = (\$800 - \$10x)(60 + x)$

(b) Find the number of passengers that maximizes the total airfare paid by the club members.

Solution: $TC'(x) = 200 - 20x$ and $TC''(x) = -20$. Thus, $x^* = 10$ and airfare expenditures are maximized with 70 passengers ($TC(x^*) = \$49,000$).

Example 1.1.13. Let $C(Q)$ be the total cost function for a firm producing Q units of some commodity. $A(Q) = C(Q)/Q$ is then called the average cost function. If $C(Q)$ is differentiable, prove that $A(Q)$ has a stationary point (critical point) at $Q_0 > 0$ if and only if the marginal cost and the average cost functions are equal at Q_0. ($C'(Q_0) = A(Q_0)$)

Solution: By definition, $QA(Q) = C(Q)$. Differentiating with respect to Q yields

$$QA'(Q) + A(Q) = C'(Q).$$

Assume $A(Q)$ has a stationary point at $Q_0 > 0$. Then it is easy to see that $A(Q_0) = C'(Q_0)$ as desired. Now assume $A(Q_0) = C'(Q_0)$. Then $Q_0 A'(Q_0) = 0$, which implies that $A'(Q_0) = 0$ as desired since $Q_0 > 0$.

Example 1.1.14. With reference to the previous example, let $C(Q) = aQ^3 + bQ^2 + cQ + d$, where $a > 0$, $b \geq 0$, $c > 0$, and $d > 0$. Prove that $A(Q) = C(Q)/Q$ has a minimum in the interval $(0, \infty)$. Then let $b = 0$ and find the minimum point in this case.

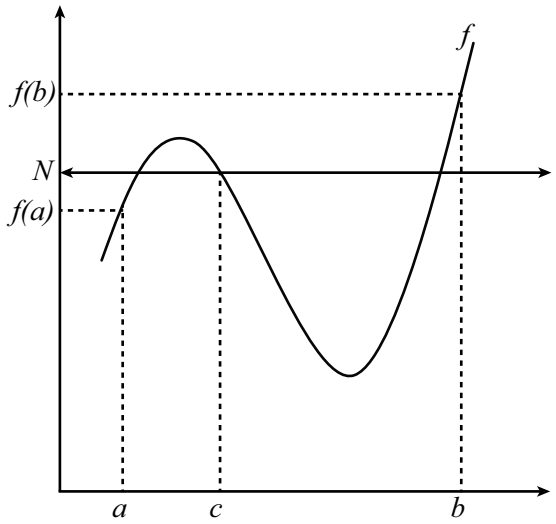
Solution: The average cost function is given by

$$A(Q) = \frac{C(Q)}{Q} = aQ^2 + bQ + c + \frac{d}{Q}.$$

Differentiating with respect to Q yields

$$A'(Q) = 2aQ + b - \frac{d}{Q^2} \overset{set}{=} 0.$$

Figure 1.2: Intermediate Value Theorem



Given the restrictions on the constants, as $Q \to 0$, $A'(Q) < 0$ and as $Q \to \infty$, $A'(Q) > 0$. Thus, there exists a $Q^* \in (0, \infty)$ that satisfies the first order condition. To determine whether this critical point is a minimum or maximum, differentiate $A'(Q)$ with respect to Q to obtain

$$A''(Q) = 2(a + d/Q^3) > 0,$$

given the restrictions on the parameters. Thus, $A(Q)$ has a minimum in the interval $(0, \infty)$ by the second derivative test. When $b = 0$

$$\frac{d}{(Q^*)^2} = 2aQ^* \quad \to \quad Q^* = \left(\frac{d}{2a}\right)^{1/3}.$$

1.1.5 Intermediate and Mean Value Theorems

Definition 1.1.4 (Intermediate Value Theorem)**.** *Suppose that f is continuous on the closed interval $[a, b]$ and let N be any number between $f(a)$ and $f(b)$, where $f(a) \neq f(b)$. Then there exists a number c in (a, b) such that $f(c) = N$.*

Simply put, the Intermediate Value Theorem says the graph of f must cross any horizontal line between $y = f(a)$ and $y = f(b)$ at one or more points in $[a, b]$ (see figure 1.2).

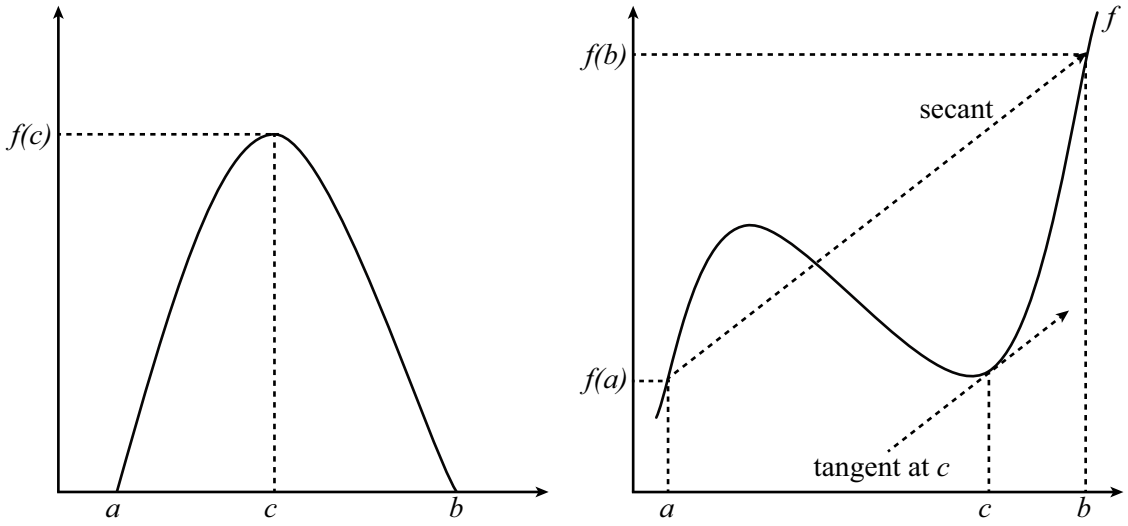
Example 1.1.15. Show that $2^x = 3x$ for some $x \in (0, 1)$.

Solution: Define $f(x) = 2^x - 3x$. Then f is continuous on $[0, 1]$ and $f(0) = 1$ and $f(1) = -1$. Thus, by the Intermediate Value Theorem $\exists\, x_0 \in (0, 1)$ such that $f(x_0) = 0$.

Theorem 1.1.5 (Rolle's Theorem)**.** *Let f be a continuous function on $[a, b]$ that is differentiable on (a, b) such that $f(a) = f(b) = 0$. Then there exists at least one point $c \in (a, b)$ such that $f'(c) = 0$.*

The geometric interpretation of Rolle's theorem is that, if the graph of a differentiable function touches the x-axis at arbitrary points a and b, where $b > a$, then for some point c between a and b there is a horizontal tangent (see figure 1.3a). If we allow the function to have different values at the endpoints, then we cannot be assured of a horizontal tangent, but there will be a point $c \in (a, b)$ such that the tangent to the graph at $x = c$ will be parallel to the chord between the endpoints of the graph. This is the essence of the Mean Value Theorem (see figure 1.3b).

Figure 1.3: Rolle's Theorem and Mean Value Theorem



(a) Rolle's Theorem

(b) Mean Value Theorem

Theorem 1.1.6 (Mean Value Theorem)**.** *Let f be a continuous function on $[a, b]$ that is differentiable on (a, b). Then there exists at least one point $c \in (a, b)$ such that*

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

Example 1.1.16. As a illustration of one use of the Mean Value Theorem (MVT), we will derive Bernoulli's inequality: for $x > 0$

$$(1 + x)^n \geq 1 + nx \quad \forall n \in \mathbb{N}.$$

Solution: Let $f(t) = (1 + t)^n$ on the interval $[0, x]$, which is clearly continuous and differentiable. Then, by the MVT, there exists a $c \in (0, x)$ such that

$$f(x) - f(0) = f'(c)(x - 0).$$

Thus, we have

$$(1 + x)^n - 1 = nx(1 + c)^{n-1} \geq nx,$$

since $f'(c) = n(1 + c)^{n-1}$, $1 + c > 1$, and $n - 1 \geq 0$.

Example 1.1.17. Use the Mean Value Theorem (MVT) to establish the following inequalities, assuming any relevant derivative formulas.

(a) $e^x > 1 + x$ for $x > 0$

Solution: Define $f(x) = e^x$ and recall from the MVT that $f(b) - f(a) = f'(c)(b - a)$ for some $c \in (0, x)$. Then

$$e^x - 1 = e^c(x - 0) = e^c x > x$$

since $e^c > 1$ for $c > 0$.

(b) $\frac{1}{8} < \sqrt{51} - 7 < \frac{1}{7}$

Solution: Define $f(x) = \sqrt{x}$ and consider the interval $[49, 51]$. Then by the MVT, $\exists c \in (49, 51)$ such that

$$\frac{\sqrt{51} - \sqrt{49}}{51 - 49} = f'(c) = \frac{1}{2\sqrt{c}}.$$

Consequently, we have

$$\frac{1}{\sqrt{c}} = \sqrt{51} - 7,$$

which implies

$$\frac{1}{8} = \frac{1}{\sqrt{64}} < \frac{1}{\sqrt{51}} < \frac{1}{\sqrt{c}} = \sqrt{51} - 7 < \frac{1}{\sqrt{49}} = \frac{1}{7}.$$

(c) $\sqrt{1+x} < 5 + \frac{x-24}{10}$ for $x > 24$

Solution: Define $f(x) = \sqrt{1+x}$ and consider the interval $[24, x]$. Then by the MVT, $\exists c \in (24, x)$ such that

$$f(x) - f(24) = \sqrt{1+x} - 5 = f'(c)(x - 24) = \frac{x - 24}{2\sqrt{1+c}} < \frac{x - 24}{10},$$

since $c > 24$, $\sqrt{1+c} > \sqrt{25} = 5$.

1.1.6 Taylor Approximations

Definition 1.1.5 (Taylor's Theorem)**.** *Let f and its first n derivatives be continuous on $[a, b]$ and differentiable on (a, b), and let $x_0 \in [a, b]$. Then for each $x \in [a, b]$ with $x \neq x_0$, there exists a point c between x and x_0 such that*

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}.$$

Taylor's theorem can be viewed as an extension of the MVT in the sense that taking $x = b$, $x_0 = a$, and $n = 0$ in Taylor's theorem yields the earlier result.

Example 1.1.18. As an illustration of the usefulness of Taylor's theorem in approximations, consider $f(x) = e^x$ for $x \in \mathbb{R}$. To find the nth Taylor polynomial for f at $x_0 = 0$, recall that $f^{(n)}(x) = e^x$ for all $n \in \mathbb{N}$. Thus, we have

$$p_n(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!}.$$

To find the error involved in approximating e^x by $p_n(x)$, we use the remainder term given by Taylor's theorem. That is,

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}x^{n+1} = \frac{e^c x^{n+1}}{(n+1)!},$$

where c is some number between 0 and x. For example, suppose that we take $n = 5$ and compute the error when $x \in [-1, 1]$. Since c is also in $[-1, 1]$, a simple calculation shows that $|R_5(x)| \leq e/6! < 0.0038$. Thus, for all $x \in [-1, 1]$, the polynomial

$$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!}$$

differs from e^x by less than 0.0038.

Remark 1.1.2. For the special case $x_0 = 0$, the Taylor Series is known as a Maclaurin series.

Example 1.1.19. Find an approximation to the following function about the given point

$$f(x) = (1+x)^5, \quad a = 0.$$

Solution: Applying the above formula

$$\begin{aligned} f(x) &\approx \left[(1+a)^5 + \frac{5(1+a)^4}{1!}(x-a) + \frac{20(1+a)^3}{2!}(x-a)^2 + \frac{60(1+a)^2}{3!}(x-a)^3 + \right. \\ &\quad \left. \frac{120(1+a)}{4!}(x-a)^4 + \frac{120}{5!}(x-a)^5\right]\Bigg|_{a=0} \\ &= 1 + 5x + 10x^2 + 10x^3 + 5x^4 + x^5. \end{aligned}$$

Example 1.1.20. Find quadratic approximations to the following functions about the given points:

(a) $F(K) = AK^\alpha, \quad K_0 = 1$
Solution: $F(K) \approx A[1 + \alpha(K-1) + \alpha(\alpha-1)(K-1)^2/2]$

(b) $f(\varepsilon) = \left(1 + \frac{3}{2}\varepsilon + \frac{1}{2}\varepsilon^2\right)^{1/2}, \quad \varepsilon_0 = 0$
Solution: $f(\varepsilon) \approx 1 + \frac{3}{4}\varepsilon - \frac{1}{32}\varepsilon^2$

(c) $H(x) = (1+x)^{-1}, \quad x_0 = 0$
Solution: $H(x) \approx 1 - x + x^2$

1.1.7 Laws of Logarithms

Consider a function, f with domain A and range B. If $a > 0$ and $a \neq 1$, the exponential function $f(x) = a^x$ is either increasing or decreasing, and thus injective. It therefore has an inverse function f^{-1}, which is called the logarithmic function with base a and is denoted $\log_a$. According to the definition of an inverse function,

$$f^{-1}(x) = y \quad \iff \quad a^y = x.$$

Thus, we have

$$\log_a x = y \quad \iff \quad a^y = x.$$

Moreover, since

$$\begin{aligned} f^{-1}(f(x)) &= x \quad \text{for every } x \text{ in } A \\ f(f^{-1}(x)) &= x \quad \text{for every } x \text{ in } B \end{aligned}$$

it is the case that

$$\begin{aligned} \log_a(a^x) &= x \quad \text{for every } x \in \mathbb{R} \\ a^{\log_a x} &= x \quad \text{for every } x > 0. \end{aligned}$$

The key properties of logarithmic functions are as follows:

(a) $\log_a(xy) = \log_a x + \log_a y$

(b) $\log_a(x/y) = \log_a x - \log_a y$

(c) $\log_a x^r = r \log_a x$ (where r is a real number)

Natural Logarithms

The mathematical constant, e, is the unique real number such that the value of the derivative (the slope of the tangent line) of the exponential function $f(x) = a^x$ at the point $x = 0$ is exactly 1. The logarithm with base e is called the natural logarithm and has a special notation

$$\log_e x \equiv \ln x.$$

Thus, the above properties generalize to

$$\begin{aligned} \ln x = y \quad &\iff \quad e^y = x \\ \ln(e^x) = x \quad &\text{for every } x \in \mathbb{R} \\ e^{\ln x} = x \quad &\text{for every } x > 0 \end{aligned}$$

In particular, if we set $x = 1$, we get

$$\ln e = 1.$$

Finally, when $y = \log_a x$, we have $a^y = x$. Thus, applying the natural logarithm to both sides of this equation, we get $y \ln a = \ln x$. Thus, the change of base formula is given by

$$y = \log_a x = \frac{\ln x}{\ln a}.$$

Example 1.1.21. Express $\ln a + \frac{1}{2} \ln b$ as a single logarithm.
Solution: $\ln a + \frac{1}{2} \ln b = \ln a + \ln b^{1/2} = \ln(a\sqrt{b})$

Example 1.1.22. Find the inverse function of the following: $m = f(t) = 24 \cdot 2^{-t/25}$.
Solution: $\frac{m}{24} = 2^{-t/25} \Rightarrow \ln m - \ln 24 = -\frac{t}{25} \ln 2 \Rightarrow t = f^{-1}(m) = \frac{25}{\ln 2}(\ln 24 - \ln m)$

Example 1.1.23. If $f(x) = 2x + \ln x$, find $f^{-1}(2)$
Solution: Define $y = f(x)$. Then $f^{-1}(y) = x$. Thus, at $y = 2$, we have $2x + \ln x = 2 \Rightarrow x = 1$. It immediately follows that $f^{-1}(2) = 1$

Example 1.1.24. Calculate $\lim_{x\to\infty}(1 + 1/x)^x$.
Solution: Define $y = (1 + 1/x)^x$. Then $\ln y = x \ln(1 + 1/x)$. We must first evaluate the limit of the right-hand-side as $x \to \infty$. Using L'Hospital's Rule, we obtain

$$\lim_{x\to\infty} x \ln(1 + 1/x) = \lim_{x\to\infty} \frac{\ln(1 + 1/x)}{1/x} = \lim_{x\to\infty} \frac{1}{1 + 1/x} = 1.$$

Thus, $\ln y \to 1$ as $x \to \infty$. Since e^x is a continuous function, we have $y = e^{\ln y} \to e$ as $x \to \infty$.

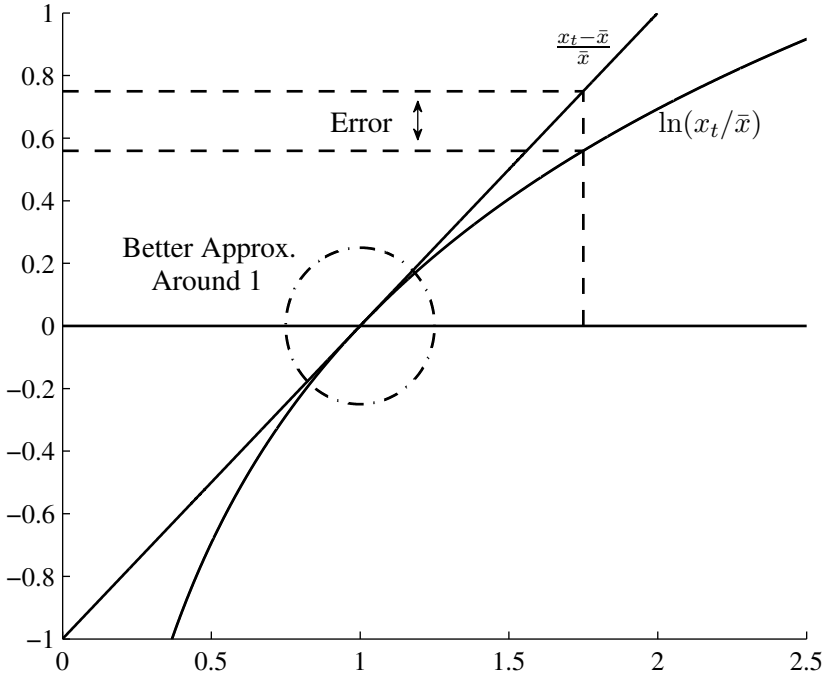
Example 1.1.25. Find a linear approximation to the following function about the given point:

$$x_t = \bar{x} e^{\ln(x_t/\bar{x})} = \bar{x} e^{\ln x_t - \ln \bar{x}} \equiv \bar{x} e^{\hat{x}_t} = f(\hat{x}_t), \quad a = \hat{x} = 0,$$

where $\bar{x}$ is the stationary value of x_t. Then show that percent changes are a good approximation for log deviations.
Solution:

$$\begin{aligned} x_t &\approx f(0) + f'(0)(\hat{x}_t - 0) \\ &= \bar{x} e^0 + \bar{x} e^0 (\hat{x}_t - 0) \\ &= \bar{x}[1 + \hat{x}_t] \end{aligned}$$

Figure 1.4: Log-Linear Approximation



which implies that

$$\hat{x}_t \equiv \ln x_t - \ln \bar{x} \approx \frac{x_t - \bar{x}}{\bar{x}}.$$

Since the above result is a first order approximation it includes an error term. $x_t/\bar{x}$ is interpreted as a gross deviation of an observation, x_t, from its stationary value (a value near one). Log linearization approximates the percent change, or the net deviation from the stationary value (a value near zero). We can see in the following graph that the approximation becomes less accurate as $x_t/\bar{x}$ moves away from one. For example, suppose $x_t/\bar{x} = 1.5$, a gross deviation of 150% (net deviation of 50%). Log linearization yields a net deviation of 40.55%. Thus, the approximation has an error of nearly 10 percentage points. Figure 1.4 illustrates that the approximation attains more accurate results when each observation is within a short distance of its stationary value.

1.1.8 Infinite Series

If we add the terms of an infinite sequence $\{a_n\}_{n=1}^{\infty}$ we get an expression of the form

$$a_1 + a_2 + a_3 + \cdots + a_n + \cdots,$$

which is called an *infinite series* (or just a series) and is denoted by

$$\sum_{n=1}^{\infty} a_n.$$

The logical question is whether it makes sense to talk about the sum of infinitely many terms? It would be impossible to find a finite sum for the series

$$1 + 2 + 3 + 4 + \cdots + n + \cdots$$

because if we start adding the terms we get the cumulative sums $1, 3, 6, 10, \ldots$ and after the n^{th} term, we get $n(n+1)/2$, which becomes very large as n increases. However, if we start to add the terms of the series

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots + \frac{1}{2^n} + \cdots$$

we get $\frac{1}{2}, \frac{3}{4}, \frac{7}{8}, \frac{15}{16}, \frac{31}{32}, \ldots, 1 - 1/2^n, \ldots$, which become closer and closer to 1. Using this idea, we can define a new sequence $\{s_n\}$ of *partial sums* given by

$$s_n = \sum_{k=1}^{n} a_k = a_1 + a_2 + \cdots + a_n.$$

If $\{s_n\}$ converges to a real number s, we say that the series is **convergent** and we write

$$\sum_{n=1}^{\infty} a_n = s.$$

We also refer to s as the *sum* of the series $\sum_{n=1}^{\infty} a_n$. A series that is not convergent is called *divergent*. If $\lim s_n = +\infty$, we say that the series $\sum_{n=1}^{\infty} a_n$ diverges to $+\infty$ and we write $\sum_{n=1}^{\infty} a_n = +\infty$.

Example 1.1.26. For the infinite series $\sum_{n=1}^{\infty} 1/[n(n+1)]$, we have the partial sums given by

$$\begin{aligned} s_n &= \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \cdots + \frac{1}{n(n+1)} \\ &= \left(\frac{1}{1} - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \cdots + \left(\frac{1}{n} - \frac{1}{n+1}\right) \\ &= 1 - \frac{1}{n+1}. \end{aligned}$$

This is an example of a *telescoping* series, so called because of the way in which the terms in the partial sums cancel. Since the sequence of partial sums converges to 1 for n large, we have

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = 1.$$

Example 1.1.27. Show that the *harmonic series*, given by,

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$$

is divergent.

Solution: The partial sums are given by

$$s_1 = 1$$

$$
\begin{aligned}
s_2 &= 1 + \frac{1}{2} \\
s_4 &= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) > 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) = 1 + \frac{2}{2} \\
s_8 &= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) \\
&> 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) \\
&= 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 1 + \frac{3}{2}
\end{aligned}
$$

Similarly, $s_{16} > 1 + \frac{4}{2}$, $s_{32} > 1 + \frac{5}{2}$, and in general

$$s_{2^n} > 1 + \frac{n}{2}.$$

This shows $s_{2^n} \to \infty$ as $n \to \infty$ and so $\{s_n\}$ is divergent. Therefore the harmonic series diverges.

Theorem 1.1.7. *If $\sum_{n=1}^{\infty} a_n$ is a convergent series, then $\lim_{n\to\infty} a_n = 0$.*

Proof. If $\sum_{n=1}^{\infty} a_n$ converges, then the sequence of partial sums $\{s_n\}$ must have a finite limit. But $a_n = s_n - s_{n-1}$, so $\lim_{n\to\infty} a_n = \lim_{n\to\infty} s_n - \lim_{n\to\infty} s_{n-1} = 0$. □

Remark 1.1.3. The converse of Theorem 1.1.7 is not true in general. If $\lim_{n\to\infty} a_n = 0$, we cannot conclude that $\sum_{n=1}^{\infty} a_n$ converges. Observe that for the harmonic series $\sum_{n=1}^{\infty} \frac{1}{n}$ we have $a_n = 1/n \to 0$ as $n \to \infty$, but we showed in 1.1.27 that $\sum_{n=1}^{\infty} \frac{1}{n}$ is divergent.

Remark 1.1.4 (Test for Divergence)**.** The contrapositive of Theorem 1.1.7, however, provides a useful test for divergence. If $\lim_{n\to\infty} a_n$ does not exist or if $\lim_{n\to\infty} a_n \neq 0$, then the series $\sum_{n=1}^{\infty} a_n$ is divergent.

Example 1.1.28. One of the most useful series in economics is the *geometric series*,

$$a + ar + ar^2 + ar^3 + \cdots + ar^{n-1} + \cdots = \sum_{n=1}^{\infty} ar^{n-1}.$$

Each term is obtained from the preceding one by multiplying it by the common ratio r. If $r = 1$, then $s_n = na \to \pm\infty$. Since the $\lim_{n\to\infty} s_n$ does not exist, the geometric series diverges in this case. If $r \neq 1$, we have

$$
\begin{aligned}
s_n &= a + ar + ar^2 + ar^3 + \cdots + ar^{n-1} \\
\text{and} \qquad rs_n &= ar + ar^2 + ar^3 + \cdots + ar^{n-1} + ar^n.
\end{aligned}
$$

Subtracting these equations, we obtain

$$s_n - rs_n = a - ar^n,$$

which implies

$$s_n = \frac{a(1 - r^n)}{1 - r}.$$

If $-1 < r < 1$, then $r^n \to 0$ as $n \to \infty$. Thus

$$\lim_{n\to\infty} s_n = \lim_{n\to\infty} \frac{a(1 - r^n)}{1 - r} = \frac{a}{1 - r}.$$

so a convergent geometric series equals the first term divided by one minus the common ratio.

Example 1.1.29. In the previous example, we found that for $|r| < 1$

$$\sum_{n=1}^{\infty} ar^{n-1} = \sum_{n=0}^{\infty} ar^n = \frac{a}{1-r}.$$

Differentiating the above equation gives

$$\sum_{n=1}^{\infty} anr^{n-1} = \frac{a}{(1-r)^2}.$$

This result is particularly useful for deriving a closed solution for the expected value of a discrete geometric random variable.

Example 1.1.30. Find the sum of each series

1. $\sum_{n=1}^{\infty} \left(\frac{1}{3}\right)^n$
   Solution: $\sum_{n=1}^{\infty} \left(\frac{1}{3}\right)^n = \frac{1/3}{1-(1/3)} = \frac{1}{2}$
2. $\sum_{n=3}^{\infty} \left(\frac{1}{2}\right)^n$
   Solution: $\sum_{n=3}^{\infty} \left(\frac{1}{2}\right)^n = \frac{1/8}{1-(1/2)} = \frac{1}{4}$

1.2 Multivariate Calculus

1.2.1 Level Surfaces

Even though many graphs have a three-dimensional representation, they are often drawn only in two dimensions for simplicity. The information about elevation can be captured by drawing in a set of contour lines, with each contour line representing all points with a specific elevation. The contour lines are not graphs of the function but are instead what are called *level curves* (level surfaces in higher dimensions). Level curves are so commonly used in economics that they often have special names. An indifference curve for a consumer is a level curve containing all bundles of goods that attain a certain level of utility. An isoquant in production theory is a level curve containing all bundles of inputs that attain a certain output level.

To see the difference between the graph of a function and the graph of a level curve, consider the function $z = f(x, y) = 25 - x^2 - y^2$. To find a level curve corresponding to $z = 16$, we take a plane at height 16 and intersect it with the graph and then project the intersection down to the x-y plane. A level curve is always in a figure with dimension one less than the graph (a two-dimensional plane versus three-dimensional space). More formally, we have the following definition.

Definition 1.2.1. *The* graph *of $f : \mathbb{R}^n \to \mathbb{R}^1$ is the set of points in $\mathbb{R}^{n+1}$ given by $\{x, f(x) | x \in \mathbb{R}^n\}$. The* level surface *corresponds to $f \equiv c$ (for some $c \in \mathbb{R}$) is $\{x \in \mathbb{R}^n | f(x) = c\}$.*

1.2.2 Projections

Definition 1.2.2 (Scalar Product)**.** *For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the scalar product (or dot product) of $\mathbf{x}$ and $\mathbf{y}$ is $\mathbf{x} \cdot \mathbf{y} := \sum_{i=1}^{n} x_i y_i$.*

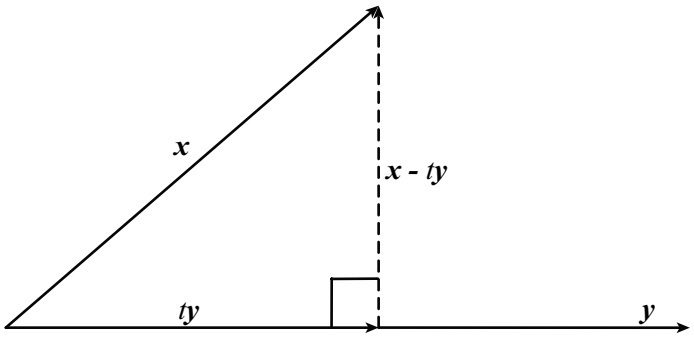
Definition 1.2.3 (Orthogonal Vectors)**.** *For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if $\mathbf{x} \cdot \mathbf{y} = 0$.*

Definition 1.2.4 (Vector Length)**.** *For $\mathbf{x} \in \mathbb{R}^n$, the length (or norm) of $\mathbf{x}$ is $\|\mathbf{x}\| := \sqrt{\mathbf{x} \cdot \mathbf{x}}$.*

Theorem 1.2.1. *If θ is the angle between vectors $\mathbf{x}$ and $\mathbf{y}$, then*

$$\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\| \|\mathbf{y}\| \cos\theta.$$

Figure 1.5: Projection of $\mathbf{x}$ onto $\mathbf{y}$



Definition 1.2.5 (Projections)**.**

- *Scalar Projection of* $\mathbf{x}$ *onto* $\mathbf{y}$*:* $\quad \text{comp}_{\mathbf{y}}\mathbf{x} = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{y}\|}$
- *Vector Projection of* $\mathbf{x}$ *onto* $\mathbf{y}$*:* $\quad \text{proj}_{\mathbf{y}}\mathbf{x} = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y}$

In order to see these formulas more clearly, note that $\cos\theta = \|t\mathbf{y}\|/\|\mathbf{x}\|$. Thus,

$$\|t\mathbf{y}\| = \|\mathbf{x}\|\cos\theta = \frac{\|\mathbf{x}\|\|\mathbf{y}\|\cos\theta}{\|\mathbf{y}\|} = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{y}\|},$$

which is the formula for the scalar projection of $\mathbf{x}$ onto $\mathbf{y}$. Moreover, using the fact that $t\mathbf{y} = \|t\mathbf{y}\|\mathbf{y}/\|\mathbf{y}\|$, simple algebra yields the vector projection of $\mathbf{x}$ onto $\mathbf{y}$.

For $t = (\mathbf{x}\cdot\mathbf{y})/\|\mathbf{y}\|^2$, $(\mathbf{x} - t\mathbf{y})\cdot\mathbf{y} = 0$ (See figure 1.5). Thus, we can decompose $\mathbf{x}$ into two parts, one a multiple of $\mathbf{y}$, $t\mathbf{y}$, and the other orthogonal to $\mathbf{y}$, $\mathbf{x} - t\mathbf{y}$.

Given a vector $\mathbf{y}$, which vector $\mathbf{x}$ with norm $c > 0$ maximizes $\mathbf{x}\cdot\mathbf{y}$? The set of vectors with $\|\mathbf{x}\| = c$ consists of all those vectors with heads on the "sphere" (could be a higher dimension) with radius c. To simplify the problem assume $\|\mathbf{y}\| = 1$. For any $\mathbf{x}$, if the projection of $\mathbf{x}$ on $\mathbf{y}$ is $t\mathbf{y}$, then

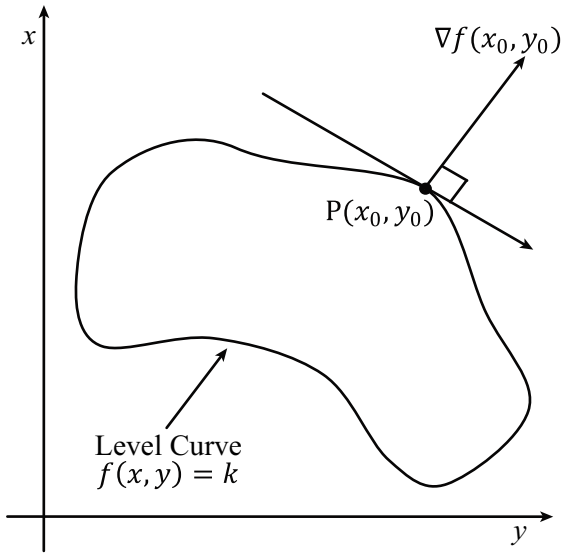
$$\text{proj}_{\mathbf{y}}\mathbf{x} \equiv t\mathbf{y} = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y}.$$

Thus equating coefficients, we obtain $\mathbf{x}\cdot\mathbf{y} = t\|\mathbf{y}\|^2 = t$. Since $\mathbf{x}\cdot\mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\|\cos\theta$, $\mathbf{x}\cdot\mathbf{y}$ is maximized when either $\|\mathbf{x}\|$ or $\|\mathbf{y}\|$ can be increased or when the angle θ between the two vectors is minimized so that $\cos(\theta) \to 1$, which implies the projection $t\mathbf{y}$ is maximized. Thus, to maximize $\mathbf{x}\cdot\mathbf{y}$ we must make the projection on $\mathbf{y}$ as large as possible subject to $\|\mathbf{x}\| = c$. By the Pythagorean theorem, $\|\mathbf{x}\|^2 = \|t\mathbf{y}\|^2 + \|\mathbf{x} - t\mathbf{y}\|^2$. But $\|\mathbf{x}\|^2 = c^2$ and $\|t\mathbf{y}\|^2 = t^2\|\mathbf{y}\|^2 = t^2$ so, rearranging, $t^2 = c^2 - \|\mathbf{x} - t\mathbf{y}\|^2$. Since the last term is nonnegative, $t \le c$. To make t as large as possible, set $\mathbf{x} - t\mathbf{y} = 0$, so $t = c$. Thus to maximize $\mathbf{x}\cdot\mathbf{y}$ subject to $\|\mathbf{x}\| = c$, set $\mathbf{x} = c\mathbf{y}$. [If $\|\mathbf{y}\| \neq 1$, we must adjust the solution to $\mathbf{x} = (c/\|\mathbf{y}\|)\mathbf{y}$.] Intuitively, we obtain the result that in order to maximize the dot product between $\mathbf{x}$ and $\mathbf{y}$, both vectors must point in the same direction.

1.2.3 Gradient Vector and its Relationship to the Level Surface

For a function f, consider the level surface S given by $f(x_1, x_2, \ldots, x_n) = k$ through a point $P(x_{1,0}, x_{2,0}, \ldots, x_{n,0})$. Let C be any curve that lies on the surface S and passes through the point P, where the curve is described by a continuous vector function, $\mathbf{r}(t) = \langle x_1(t), x_2(t), \ldots, x_n(t)\rangle$.

Figure 1.6: Gradient Vector Perpendicular to the Level Curve



Let t_0 be the parameter value corresponding to P; that is $\mathbf{r}(t_0) = \langle x_{1,0}, x_{2,0}, \ldots, x_{n,0}\rangle$. Since C lies on S, any point $(x_1(t), x_2(t), \ldots, x_n(t))$ must satisfy

$$f(x_1(t), x_2(t), \ldots, x_n(t)) = k.$$

By the Chain Rule, its total derivative is

$$\frac{\partial f}{\partial x_1}\frac{dx_1}{dt} + \frac{\partial f}{\partial x_2}\frac{dx_2}{dt} + \cdots + \frac{\partial f}{\partial x_n}\frac{dx_n}{dt} = \nabla f \cdot \mathbf{r}'(t) = 0.$$

Since $\nabla f = \langle f_{x_1}, f_{x_2}, \ldots, f_{x_n}\rangle$ and $\mathbf{r}'(t) = \langle x_1'(t), x_2'(t), \ldots, x_n'(t)\rangle$, at $t = t_0$ the above condition can be written

$$\nabla f(x_{1,0}, x_{2,0}, \ldots, x_{n,0}) \cdot \mathbf{r}'(t_0) = 0.$$

Thus, the gradient vector at P, $\nabla f(x_{1,0}, x_{2,0}, \ldots, x_{n,0})$, is perpendicular to the tangent vector $\mathbf{r}'(t)$ to any curve C on S that passes through P. To illustrate, consider a function f of two variables and a point $P(x_0, y_0)$ in its domain. The gradient vector $\nabla f(x_0, y_0)$ gives the direction of the fastest increase of f and is perpendicular to the level curve $f(x, y) = k$ that passes through P. This makes intuitive sense since the values of f remain constant as we move along the curve.

1.2.4 Gradients and Tangent Planes

Proposition 1.2.1. *Assume that $f : \mathbb{R}^n \to \mathbb{R}^1$ is differentiable at a with gradient vector $\nabla f(a)$. The following properties are consequences of differentiability:*

(i) f is continuous at a.

(ii) For all unit vectors, $\mathbf{u}$, the directional derivative in the direction of $\mathbf{u}$ is $f_{\mathbf{u}}(\mathbf{a}) = \nabla f(\mathbf{a}) \cdot \mathbf{u}$.

(iii) For all i the ith component of $\nabla f(\mathbf{a})$ is $\frac{\partial f(\mathbf{a})}{\partial x_i}$.

(iv) The equation of the tangent plane to the graph of f at $(a, f(a))$ is given by

$$f(\mathbf{x}) - f(\mathbf{a}) = \nabla f(\mathbf{a}) \cdot (\mathbf{x} - \mathbf{a}).$$

Note the similarity to the single-variable case.

(v) The equation of the tangent plane to the level curve corresponding to $f(\mathbf{x}) \equiv f(\mathbf{a})$ at the point $\mathbf{a}$ is given by

$$0 = \nabla f(\mathbf{a}) \cdot (\mathbf{x} - \mathbf{a}).$$

(vi) The marginal rate of substitution of x_i for x_j along the level curve corresponding to $f(\mathbf{x}) \equiv f(\mathbf{a})$ at the point $\mathbf{a}$ is the number of units of x_j, which must be removed in order to maintain a constant "output" f when a unit of x_i is added and all other "inputs" are unchanged. The change in input j (i) is v_j (v_i) and all other inputs are unchanged, so $v_k = 0$ for $k \neq i, j$. More formally we have

$$0 = \frac{\partial f}{\partial x_i}(\mathbf{a})v_i + \frac{\partial f}{\partial x_j}(\mathbf{a})v_j \quad \textit{or}$$

$$\left.\frac{\partial x_j}{\partial x_i}\right|_{\substack{\mathbf{x}=\mathbf{a} \\ f(\mathbf{x})\equiv f(\mathbf{a})}} = \frac{v_j}{v_i} = -\frac{\frac{\partial f(\mathbf{a})}{\partial x_i}}{\frac{\partial f(\mathbf{a})}{\partial x_j}}.$$

(vii) The direction of change of inputs $\mathbf{x}$ which most increases output $f(\mathbf{x})$ starting at $\mathbf{a}$ is the direction $\nabla f(\mathbf{a})$.

Example 1.2.1. $u(x, y, z) = x + y + z^2$ is differentiable at $(1, 1, 1)$.

(a) To find $\nabla u(1, 1, 1)$, $\nabla u(x, y, z) = \begin{pmatrix} 1 \\ 1 \\ 2z \end{pmatrix}$, so $\nabla u(1, 1, 1) = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$.

(b) To find the equation of the tangent plane to the graph of u at $(1, 1, 1)$, $u(1, 1, 1) = 3$, so

$$\big(u(x, y, z) - 3\big) = \nabla u(1, 1, 1) \cdot \begin{pmatrix} x - 1 \\ y - 1 \\ z - 1 \end{pmatrix} = (x - 1) + (y - 1) + 2(z - 1).$$

(c) The equation of the tangent plane to the level curve corresponding to $u \equiv 3$ at $(1, 1, 1)$ is

$$0 = \nabla u(1, 1, 1) \cdot \begin{pmatrix} x - 1 \\ y - 1 \\ z - 1 \end{pmatrix} = (x - 1) + (y - 1) + 2(z - 1).$$

(d) To find the marginal rate of substitution of x for z along the level curve $u \equiv 3$ at the point $(1, 1, 1)$,

$$0 = \frac{\partial u}{\partial x}(1, 1, 1)\Delta x + \frac{\partial u}{\partial z}(1, 1, 1)\Delta z$$

so

$$\left.\frac{\partial z}{\partial x}\right|_{\substack{(x,y,z)=(1,1,1) \\ u=3}} = -\frac{\frac{\partial u(1,1,1)}{\partial x}}{\frac{\partial u(1,1,1)}{\partial z}} = -\frac{1}{2}.$$

(e) To find the direction of change of inputs which yields the largest increase in output u at $(1,1,1)$, the direction is

$$\frac{1}{\|\nabla u(1,1,1)\|}\nabla u(1,1,1) = \frac{1}{\sqrt{6}}\begin{pmatrix}1\\1\\2\end{pmatrix},$$

assuming the magnitude of the total change in inputs is unity.

Example 1.2.2. Output, Y, is produced using two inputs, K and L, according to $Y = f(K,L) = 1 + (KL-1)^{1/3}$.

(a) What is the equation of the tangent plane to the production surface at the point corresponding to $K = 1$ and $L = 2$?

Solution:

$$\begin{aligned} Y &= f(1,2) + \nabla f(1,2)\cdot\begin{pmatrix}K-1\\L-2\end{pmatrix} \\ &= 2 + \frac{2}{3}(K-1) + \frac{1}{3}(L-2) \end{aligned}$$

or, equivalently, $Y = \frac{2}{3}K + \frac{1}{3}L + \frac{2}{3}$.

(b) What is the equation of the tangent plane to the isoquant (level curve) corresponding to $Y = 3$ at $(K,L) = (1,9)$?

Solution:

$$\begin{aligned} 3 &= 3 + \nabla f(1,9)\cdot\begin{pmatrix}K-1\\L-9\end{pmatrix} \\ \rightarrow \quad 0 &= \frac{3}{4}(K-1) + \frac{1}{12}(L-9) \end{aligned}$$

or, equivalently, $9K + L = 18$.

(c) What is the Marginal Rate of Substitution (MRS) of L for K along the isoquant corresponding to $Y = f(2,1)$ at $(K,L) = (2,1)$?

Solution: The Equation of the tangent plane to the level curve at $(K,L) = (2,1)$ is

$$0 = (\Delta K, \Delta L)\cdot(f_K(2,1), f_L(2,1)) = (\Delta K, \Delta L)\cdot(\frac{1}{3}, \frac{2}{3}),$$

which implies that $\text{MRS} = \Delta K/\Delta L = -2$.

(d) If starting at $(K,L) = (2,1)$, a tiny (marginal) amount of inputs could be added in any proportions with $||(\Delta K, \Delta L)|| = \varepsilon$, how many extra units of L should be added for each additional unit of K to maximize the increase in output (i.e., what is the ratio $\Delta L/\Delta K$)?

Solution: Equation of the tangent plane to the production surface at $(2,1)$ is

$$Y - f(2,1) = (\Delta K, \Delta L)\cdot(\frac{1}{3}, \frac{2}{3}).$$

An increase in output is maximized when the scalar product of $(\Delta K, \Delta L)$ and $\nabla f(2,1)$ is maximized. Thus,

$$(\Delta K, \Delta L) = \frac{\varepsilon}{\|\frac{1}{3}, \frac{2}{3}\|}(\frac{1}{3}, \frac{2}{3}) = \left(\frac{\varepsilon}{\sqrt{5}}, \frac{2\varepsilon}{\sqrt{5}}\right) \quad \rightarrow \quad \frac{\Delta L}{\Delta K} = 2.$$

Example 1.2.3. A firm uses capital, K, and labor, L, to produce output, Y, according to $Y = K^\alpha L^\beta$, where α and β are positive constants.

(a) What is the equation of the isoquant corresponding to a level of output equal to one?

Solution: When $Y = 1$, the corresponding level curve is $K^\alpha L^\beta = 1$.

(b) What is the equation of the tangent plane to the production surface at $K = L = 1$?

Solution:

$$\begin{aligned} Y &= Y(1,1) + \frac{\partial Y}{\partial K}(1,1)(K-1) + \frac{\partial Y}{\partial L}(1,1)(L-1) \\ &= 1 + \alpha(K-1) + \beta(L-1) \end{aligned}$$

(c) What is the equation of the tangent "plane" to the level curve corresponding to $Y = 1$ at $K = L = 1$?

Solution:

$$0 = \alpha(K-1) + \beta(L-1) \quad \rightarrow \quad \alpha K + \beta L = \alpha + \beta.$$

(d) For small changes along the level curve, starting at $K = L = 1$, how many units of labor are needed to replace each unit of capital (i.e. What is the $\text{MRS}_{L\to K}$)?

Solution: The Equation of the tangent plane to the level curve at $(K, L) = (1, 1)$ is

$$0 = (\Delta K, \Delta L) \cdot (f_K(1,1), f_L(1,1)) = (\Delta K, \Delta L) \cdot (\alpha, \beta),$$

which implies that $\Delta L/\Delta K = -\alpha/\beta$.

(e) If it were possible to increase K and L slightly in any proportion so that $||(\Delta K, \Delta L)|| = c$, where c is a very small positive number, what change in K and L would lead to the greatest increase in output?

Solution: Equation of the tangent plane to the production surface at $(1, 1)$ is

$$Y = 1 + (\Delta K, \Delta L) \cdot (\alpha, \beta).$$

Thus, the maximum value of Y is attained when the above dot product is maximized. This will occur when

$$(\Delta K, \Delta L) = \frac{c}{\sqrt{\alpha^2 + \beta^2}}(\alpha, \beta).$$

1.2.5 Chain Rule

Definition 1.2.6 (Chain Rule–Case 1)**.** *Suppose that $z = f(x, y)$ is a differentiable function of x and y, where $x = g(t)$ and $y = h(t)$ are both differentiable functions of t. Then z is a differentiable function of t and*

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt}.$$

Definition 1.2.7 (Chain Rule–Case 2)**.** *Suppose that $z = f(x, y)$ is a differentiable function of x and y, where $x = g(s, t)$ and $y = h(s, t)$ are both differentiable functions of s and t. Then*

$$\frac{\partial z}{\partial s} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial s} \qquad \textit{and} \qquad \frac{\partial z}{\partial t} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial t}.$$

Definition 1.2.8 (Chain Rule–General Version)**.** *Suppose that u is a differentiable function of the n variables $x_1, \ldots, x_n$ and each x_j is a differentiable function of the m variables $t_1, \ldots, t_m$. Then u is a function of $t_1, \ldots, t_m$ and*

$$\frac{\partial u}{\partial t_i} = \frac{\partial u}{\partial x_1}\frac{\partial x_1}{\partial t_i} + \frac{\partial u}{\partial x_2}\frac{\partial x_2}{\partial t_i} + \cdots + \frac{\partial u}{\partial x_n}\frac{\partial x_n}{\partial t_i}$$

for each $i = 1, 2, \ldots, m$.

Example 1.2.4. Given $z = f(x, y, t) = ye^x + t^2$ where $x = g(y, t) = \ln(y + t)$ and $y = h(t) = t^3 - 9$, use the chain rule to find the total effect of a change in t on z (dz/dt) at $t = 2$.

Solution:

$$\begin{aligned}\frac{dz}{dt} &= \frac{\partial z}{\partial x}\left(\frac{\partial x}{\partial y}\frac{dy}{dt} + \frac{\partial x}{\partial t}\right) + \frac{\partial z}{\partial y}\frac{dy}{dt} + \frac{\partial z}{\partial t} \\ &= ye^x\left(\frac{1}{y+t}3t^2 + \frac{1}{y+t}\right) + 3e^x t^2 + 2t.\end{aligned}$$

When $t = 2$, $y = -1$ and $x = 0$, so

$$\left.\frac{dz}{dt}\right|_{t=2} = -1(12 + 1) + 12 + 4 = 3.$$

Example 1.2.5. Assume the following functional forms:

$$z = x^2 + xy^3, \quad x = uv^2 + w^3, \quad y = u + ve^w.$$

Use the chain rule to find the indicated partial derivatives at the given point:

$$\frac{\partial z}{\partial u}, \ \frac{\partial z}{\partial v}, \ \frac{\partial z}{\partial w}, \ \text{when } u = 2, \ v = 1, \ w = 0.$$

Solution: First note that $u = 2, \ v = 1, \ w = 0$ implies $x = 2, \ y = 3$

$$\begin{aligned}\frac{\partial z}{\partial u} &= \frac{\partial z}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial u} \\ &= (2x + y^3)(v^2) + (3xy^2) \\ &= 85,\end{aligned} \qquad \begin{aligned}\frac{\partial z}{\partial v} &= \frac{\partial z}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial v} \\ &= (2x + y^3)(2uv) + (3xy^2)(e^w) \\ &= 178,\end{aligned}$$

$$\begin{aligned}\frac{\partial z}{\partial w} &= \frac{\partial z}{\partial x}\frac{\partial x}{\partial w} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial w} \\ &= (2x + y^3)(3w^2) + (3xy^2)(ve^w) \\ &= 54.\end{aligned}$$

Example 1.2.6. A function is called **homogeneous of degree n** if it satisfies the equation

$$f(tx, ty) = t^n f(x, y)$$

for all t, where n is a positive integer and f has continuous second order partial derivatives.

(a) Verify that $f(x, y) = x^2y + 2xy^2 + 5y^3$ is homogeneous of degree 3.

Solution: Applying the above definition

$$\begin{aligned}f(tx, ty) &= (tx)^2(ty) + 2(tx)(ty)^2 + 5(ty)^3 = t^3x^2y + 2t^3xy^2 + 5t^3y^3 \\ &= t^3[x^2y + 2xy^2 + 5y^3] = t^3 f(x, y).\end{aligned}$$

(b) Show that if f is homogeneous of degree n, then

$$x\frac{\partial f}{\partial x} + y\frac{\partial f}{\partial y} = nf(x,y).$$

Solution: To see this clearly, rewrite the above definition in the following way:

$$f(a,b) = t^n f(x,y),$$

where $a = tx$ and $b = ty$. Then differentiating with respect to t gives

$$\begin{aligned} & \frac{\partial f(tx,ty)}{\partial a}\frac{da}{dt} + \frac{\partial f(tx,ty)}{\partial b}\frac{db}{dt} = nt^{n-1}f(x,y) \\ \rightarrow \quad & \frac{\partial f(tx,ty)}{\partial a}x + \frac{\partial f(tx,ty)}{\partial b}y = nt^{n-1}f(x,y). \end{aligned}$$

Since this equation holds for all t, we can set $t = 1$ to obtain the desired result.

Theorem 1.2.2 (Young's Theorem)**.** *Suppose that $y = f(x_1, x_2, \ldots, x_n)$ is twice continuously differentiable (C^2) on an open region $J \in \mathbb{R}^n$. Then, for all $\mathbf{x} \in J$ and for each pair of indices i, j,*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}).$$

Example 1.2.7. Consider the general Cobb-Douglas production function $Q = kx^a y^b$. Then,

$$\begin{aligned} \frac{\partial Q}{\partial x} &= akx^{a-1}y^b, \qquad \frac{\partial Q}{\partial y} = bkx^a y^{b-1}, \\ \frac{\partial^2 Q}{\partial x \partial y} &= abkx^{a-1}y^{b-1} = \frac{\partial^2 Q}{\partial y \partial x}. \end{aligned}$$

We can continue taking higher order derivatives, and Young's theorem holds for these cases. For example, if we take an $x_1x_2x_4$ derivative of order three, then the order of differentiation does not matter for a C^3-function. We can keep going and define kth order partial derivatives and C^k functions. For C^k functions, the order you take the kth partial derivatives does not matter.

Example 1.2.8. If f is homogeneous of degree n, show that

$$x^2\frac{\partial^2 f(x,y)}{\partial x^2} + 2xy\frac{\partial^2 f(x,y)}{\partial x \partial y} + y^2\frac{\partial^2 f(x,y)}{\partial y^2} = n(n-1)f(x,y).$$

Solution: To obtain the desired result, differentiate the following result with respect to t

$$\frac{\partial f(tx,ty)}{\partial a}x + \frac{\partial f(tx,ty)}{\partial b}y = nt^{n-1}f(x,y).$$

Using the chain rule, we obtain

$$\begin{aligned} & x\frac{\partial^2 f(tx,ty)}{\partial a^2}\frac{da}{dt} + x\frac{\partial^2 f(tx,ty)}{\partial a \partial b}\frac{db}{dt} + y\frac{\partial^2 f(tx,ty)}{\partial b \partial a}\frac{da}{dt} + y\frac{\partial^2 f(tx,ty)}{\partial b^2}\frac{db}{dt} = n(n-1)t^{n-2}f(x,y) \\ \rightarrow \quad & x^2\frac{\partial^2 f(tx,ty)}{\partial a^2} + 2xy\frac{\partial^2 f(tx,ty)}{\partial a \partial b} + y^2\frac{\partial^2 f(tx,ty)}{\partial b^2} = n(n-1)t^{n-2}f(x,y). \end{aligned}$$

Again, setting $t = 1$ gives the desired result.

Example 1.2.9. If f is homogeneous of degree n, show that

$$f_x(tx, ty) = t^{n-1} f_x(x, y).$$

Solution: To obtain the desired result, differentiate

$$f(a, b) = t^n f(x, y)$$

with respect to x to obtain

$$\frac{\partial f(tx, ty)}{\partial a}\frac{da}{dx} = t^n \frac{\partial f(x, y)}{\partial x} \quad \rightarrow \quad \frac{\partial f(tx, ty)}{\partial x} = t^{n-1}\frac{\partial f(tx, ty)}{\partial x}.$$

Thus, if a function, f, is homogeneous of degree n, its derivative, f', is homogeneous of degree $n - 1$. Note that evaluating the function at (tx, ty) and subsequently taking the derivative of $f(tx, ty)$ with respect to the first argument, tx, is equivalent to taking the derivative of $f(x, y)$ at and evaluating at (tx, ty).

Example 1.2.10. Verify that partial derivative of $f(x, y) = x^2y + 2xy^2 + 5y^3$ with respect to x is homogeneous of degree 2.

Solution: The partial derivative is given by

$$f_x(x, y) = 2xy + 2y^2.$$

Evaluating at (tx, ty) gives

$$f_x(tx, ty) = 2(tx)(ty) + 2(ty)^2 = t^2[2xy + 2y^2] = t^2 f_x(x, y).$$

Thus partial derivative of the given function is homogeneous of degree 2.

1.2.6 Second Order Derivatives and Hessians

The *Hessian matrix* is a square matrix of second-order partial derivatives of a function. Let $x \in \mathbb{R}^n$ and let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a real-valued function having 2nd-order partial derivatives in an open set U containing x. Given the real-valued function $f(x_1, x_2, \ldots, x_n)$, the Hessian matrix of f is the matrix with elements

$$H(f)_{ij}(x) = D_i D_j f(x),$$

where $x = (x_1, x_2, \ldots, x_n)$ and $D_i D_j$ is the differentiation operator with respect to the ijth argument:

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2}. \end{bmatrix}$$

If all n^2 second-order partial derivatives of f exist and are continuous functions of $(x_1, x_2, \ldots, x_n)$, we say that f is twice continuously differentiable or C^2.

The *Bordered Hessian matrix* of f is a square matrix of second-order partial derivatives that is bordered by first-order partial derivatives. Given the real-valued function $f(x_1, x_2, \ldots, x_n)$, the bordered Hessian matrix of f is the matrix

$$\overline{H}(f) = \begin{bmatrix} 0 & \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \\ \frac{\partial f}{\partial x_1} & \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial f}{\partial x_2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_n} & \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

The importance of the Hessian and Bordered Hessian matrices will become clear in later sections.

1.3 Basic Analysis

1.3.1 Induction and Examples

Theorem 1.3.1 (Principle of Mathematical Induction)**.** *Let $P(n)$ be a statement that is either true or false for each $n \in \mathbb{N}$. Then $P(n)$ is true for all $n \in \mathbb{N}$, provided that*

(a) $P(1)$ is true, and

(b) for each $k \in \mathbb{N}$, if $P(k)$ is true, then $P(k+1)$ is true.

Example 1.3.1. Prove that $1 + 2 + 3 + \cdots + n = \frac{1}{2}n(n+1)$ for every natural number n.

Solution: Let $P(n)$ be the statement

$$1 + 2 + 3 + \cdots + n = \frac{1}{2}n(n+1)$$

Then $P(1)$ asserts that $1 = \frac{1}{2}(1)(1+1)$, $P(2)$ asserts that $1 + 2 = \frac{1}{2}(2)(2+1)$, and so on. In particular, we see that $P(1)$ is true, and this establishes the basis for induction. To verify the induction step, we suppose that $P(k)$ is true, where $k \in \mathbb{N}$. That is, we assume

$$1 + 2 + 3 + \cdots + k = \frac{1}{2}k(k+1).$$

Since we wish to conclude that $P(k+1)$ is true, we add $k+1$ to both sides to obtain

$$\begin{aligned} 1 + 2 + 3 + \cdots + k + (k+1) &= \frac{1}{2}k(k+1) + (k+1) \\ &= \frac{1}{2}[k(k+1) + 2(k+1)] \\ &= \frac{1}{2}(k+1)(k+2) \\ &= \frac{1}{2}(k+1)[(k+1)+1]. \end{aligned}$$

Thus $P(k+1)$ is true whenever $P(k)$ is true, and by principle of mathematical induction, we conclude that $P(n)$ is true for all n.

Since the format of a proof using mathematical induction always consists of the same two steps (establishing the basis for induction and verifying the induction step), it is common practice to reduce some of the formalism by omitting explicit reference to the statement $P(n)$. It is also acceptable to omit identifying the steps by name.

Example 1.3.2. Prove by induction that $7^n - 4^n$ is a multiple of 3, for all $n \in \mathbb{N}$.

Solution: This is true when $n = 1$, since $7^1 - 4^1 = 3$. Now let $k \in \mathbb{N}$ and suppose that $7^k - 4^k$ is a multiple of 3. That is, $7^k - 4^k = 3m$ for some $m \in \mathbb{N}$. It follows that

$$\begin{aligned} 7^{k+1} - 4^{k+1} &= 7^{k+1} - 7 \cdot 4^k + 7 \cdot 4^k - 4 \cdot 4^k \\ &= 7(7^k - 4^k) + 3 \cdot 4^k \\ &= 7(3m) + 3 \cdot 4^k \\ &= 3(7m + 4^k). \end{aligned}$$

Since m and k are natural numbers, so is $7m + 4^k$. Thus $7^{k+1} - 4^{k+1}$ is also a multiple of 3, and by induction we conclude that $7^n - 4^n$ is a multiple of 3 for all $n \in \mathbb{N}$.

In the above example, we have added and subtracted the term $7 \cdot 4^k$. Where did it come from? We want somehow to use the induction hypothesis $7^k - 4^k = 3m$, so we break 7^{k+1} apart into $7 \cdot 7^k$. We would like to have $7^k - 4^k = 3m$ as a factor instead of just 7^k, but to do this we must subtract (and add) the term $7 \cdot 4^k$.

Example 1.3.3. Prove that $1^2 + 2^2 + \cdots + n^2 = \frac{1}{6}n(n+1)(2n+1)$ for all $n \in \mathbb{N}$.

Solution: This is true when $n = 1$, since $\frac{1}{6}(1)(1+1)(2 \cdot 1 + 1) = 1$. Now let $k \in \mathbb{N}$ and suppose that $1^2 + 2^2 + \cdots + k^2 = \frac{1}{6}k(k+1)(2k+1)$. Adding $(k+1)^2$ to both sides, it follows that

$$\begin{aligned} 1^2 + 2^2 + \cdots + k^2 + (k+1)^2 &= \frac{1}{6}k(k+1)(2k+1) + (k+1)^2 \\ &= \frac{1}{6}[2k^3 + 3k^2 + k] + k^2 + 2k + 1 \\ &= \frac{1}{6}[2k^3 + 9k^2 + 13k + 6] \\ &= \frac{1}{6}[(k+1)(k+2)(2k+3)]. \end{aligned}$$

Thus, the above statement holds for $n = k + 1$ whenever it holds for $n = k$, and by principle of mathematical induction, we conclude that the statement is true for all n.

Example 1.3.4. Prove that

$$\frac{1}{3} + \frac{1}{15} + \frac{1}{35} + \cdots + \frac{1}{4n^2 - 1} = \frac{n}{2n+1}, \text{ for all } n \in \mathbb{N}$$

Solution: This is true when $n = 1$, since $\frac{1}{2 \cdot 1 + 1} = \frac{1}{3}$. Now let $k \in \mathbb{N}$ and suppose that

$$\frac{1}{3} + \frac{1}{15} + \frac{1}{35} + \cdots + \frac{1}{4k^2 - 1} = \frac{k}{2k+1}.$$

Adding $\frac{1}{4(k+1)^2-1}$ to both sides, it follows that

$$\begin{aligned}\frac{1}{3}+\frac{1}{15}+\frac{1}{35}+\cdots+\frac{1}{4k^2-1}+\frac{1}{4(k+1)^2-1} &= \frac{k}{2k+1}+\frac{1}{4(k+1)^2-1}\\ &= \frac{k}{2k+1}+\frac{1}{4k^2+8k+3}\\ &= \frac{k}{2k+1}+\frac{1}{(2k+1)(2k+3)}\\ &= \frac{2k^2+3k+1}{(2k+1)(2k+3)}\\ &= \frac{k+1}{2k+3}.\end{aligned}$$

Thus, the above statement holds for $n = k + 1$ whenever it holds for $n = k$, and by principle of mathematical induction, we conclude that the statement is true for all n.

1.3.2 Neighborhoods and Open and Closed Sets

Definition 1.3.1 (Neighborhood)**.** *Let $x \in \mathbb{R}$ and let $\varepsilon > 0$. A neighborhood of x (or an ε-neighborhood of x) is a set of the form*

$$N(x;\varepsilon) = \{y \in \mathbb{R} : |x - y| < \varepsilon\}.$$

The number ε is referred to as the radius of $N(x;\varepsilon)$.

A neighborhood of x of radius ε is the open interval $(x - \varepsilon, x + \varepsilon)$ of length 2ε centered at x.

Definition 1.3.2 (Deleted Neighborhood)**.** *Let $x \in \mathbb{R}$ and let $\varepsilon > 0$. A deleted neighborhood of x is a set of the form*

$$N^*(x;\varepsilon) = \{y \in \mathbb{R} : 0 < |x - y| < \varepsilon\}.$$

Clearly $N^*(x;\varepsilon) = N(x;\varepsilon)\backslash\{x\}$.

Neighborhoods give us a framework within which we can talk about "nearness".

Definition 1.3.3 (Interior and Boundary Points)**.** *Let S be a subset of $\mathbb{R}$. A point x in $\mathbb{R}$ is an interior point of S if there exists a neighborhood N of x such that $N \subseteq S$. If for every neighborhood N of x, $N \cap S \neq \varnothing$ and $N \cap (\mathbb{R}\backslash S) \neq \varnothing$, then x is called a boundary point of S. The set of all interior points of S is denoted int S, and the set of all boundary points of S is denoted by $\operatorname{bd} S$ (see figure 1.7).*

Example 1.3.5.

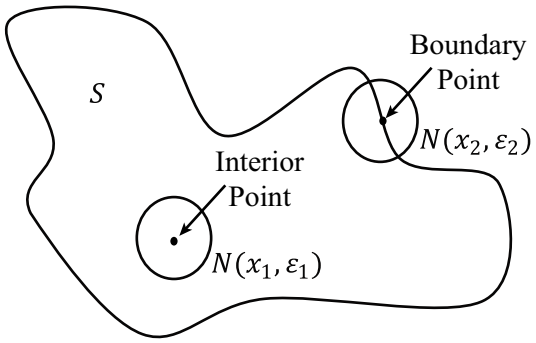
(a) Let S be the open interval $(0, 5)$ and let $x \in S$. If $\varepsilon = \min\{x, 5 - x\}$, then we claim that $N(x;\varepsilon) \subseteq S$. Indeed, for all $y \in N(x;\varepsilon)$ we have $|y - x| < \varepsilon$, so that

$$-x \leq -\varepsilon < y - x < \varepsilon \leq 5 - x$$

Thus $0 < y < 5$ and $y \in S$. That is, for some arbitrary point in S, there exists a neighborhood that is completely contained in S. It follows that every point in S is an interior point of S ($S \subseteq \operatorname{int} S$). Since the inclusion $\operatorname{int} S \subseteq S$ always holds, we have $S = \operatorname{int} S$.

The point 0 is not a member of S, but every neighborhood of 0 will contain positive numbers in S. Thus 0 is a boundary point of S. Similarly, $5 \in \operatorname{bd} S$ and, in fact, $\operatorname{bd} S = \{0, 5\}$. Note that none of the boundary of S is contained in S. Of course, there is nothing special about the open interval $(0, 5)$ in this example. Similar comments would apply to any open interval.

Figure 1.7: Interior and Boundary Points



(b) Let S be the closed interval $[0, 5]$. The point 0 is still a boundary point of S, since every neighborhood of x will contain negative numbers not in S. We have $\operatorname{int} S = (0, 5)$ and $\operatorname{bd} S = \{0, 5\}$. This time S contains all of its boundary points, and the same could be said of any other closed interval.

(c) Let S be the interval $[0, 5)$. Then again $\operatorname{int} S = (0, 5)$ and $\operatorname{bd} S = \{0, 5\}$. We see that S contains some of its boundary, but not all of it.

(d) Let S be the interval $[2, \infty)$. Then $\operatorname{int} S = (2, \infty)$ and $\operatorname{bd} S = \{2\}$. Note that there is no "point" at ∞ to be included as a boundary point at the right end.

(e) Let $S = \mathbb{R}$. Then $\operatorname{int} S = S$ and $\operatorname{bd} S = \varnothing$.

Example 1.3.6. Find the interior and boundary of the following set: $S = \left\{\frac{1}{n} : n \in \mathbb{N}\right\}$

Solution: The above set is

$$\left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \cdots\right\}$$

Given the distance between points, it is clear that there does not exist a neighborhood N around any point that is contained in S. Thus, the interior of $S = \varnothing$. The boundary of the set is $0 \cup S$

Example 1.3.7. Find the interior and boundary of each set.

1. $[0, 3] \cup (3, 5)$
   Solution: $(0, 5)$, $\{0, 5\}$
2. $\left\{r \in \mathbb{Q} : 0 < r < \sqrt{2}\right\}$
   Solution: $\varnothing$, $[0, \sqrt{2}]$
3. $\left\{r \in \mathbb{Q} : r \geq \sqrt{2}\right\}$
   Solution: $\varnothing$, $[\sqrt{2}, \infty)$
4. $[0, 2] \cap [2, 4]$
   Solution: $\varnothing$, $\{2\}$

Definition 1.3.4 (Open and Closed Sets)**.** *Let $S \subseteq \mathbb{R}$. If $\operatorname{bd} S \subseteq S$, then S is said to be closed. If $\operatorname{bd} S \subseteq \mathbb{R} \backslash S$, then S is said to be open.*

Theorem 1.3.2.

(a) A set S is open iff $S = \operatorname{int} S$. Thus, S is open iff every point in S is an interior point of S.

(b) A Set S is closed iff its complement $\mathbb{R} \backslash S$ is open.

Example 1.3.8. Classify each of the following sets, S, as open, closed, neither, or both.

1. $\mathbb{N}$

   Solution: Not open: $\operatorname{int} S = \varnothing \neq S$, Closed: $\operatorname{bd} S = S$

2. $\mathbb{Q}$

   Solution: Neither, $\operatorname{int} S = \varnothing \neq S$ and $\operatorname{bd} S = \mathbb{R} \not\subseteq S$

3. $\bigcap_{n=1}^{\infty} \left(0, \frac{1}{n}\right)$

   Solution: Both: $(0,1) \cap (0, \frac{1}{2}) \cap \cdots = \varnothing$ ($int\, \varnothing = \varnothing$ and $\operatorname{bd} \varnothing = \varnothing$)

4. $\left\{x : |x - 5| \leq \frac{1}{2}\right\}$

   Solution: Not open: $\operatorname{int} S = (4.5, 5.5) \neq S$, Closed: $\operatorname{bd} S = \{4.5, 5.5\} \in S$

5. $\{x : x^2 > 0\}$

   Solution: Open: $\operatorname{int} S = \mathbb{R} \backslash \{0\} = S$, Not Closed: $\operatorname{bd} S = \{0\} \notin S$

6. $\left\{\frac{1}{n} : n \in \mathbb{N}\right\}$

   Solution: Neither: $\operatorname{int} S = \varnothing \neq S$ and $\operatorname{bd} S \not\subseteq S$ ($0 \notin S$).

Example 1.3.9. True/False: If $S \subseteq \mathbb{R}^2$ is an open set, then $f(x, y) = x + y$ cannot have a global maximizer subject to $(x, y) \in S$.

Solution: True. Suppose (x^*, y^*) is a global maximizer of $x + y$ subject to $(x, y) \in S$. Since S is open, we can always find ε small enough so that an open ε-neighborhood around (x^*, y^*) is entirely contained in S. But then a point such as $(x^* + \varepsilon/2, y^* + \varepsilon/2)$ would lie in S with a value of the objective function $(x^* + \varepsilon/2) + (y^* + \varepsilon/2)$; i.e.,

$$f\left(x^* + \frac{\varepsilon}{2}, y^* + \frac{\varepsilon}{2}\right) = x^* + y^* + \varepsilon > x^* + y^* = f(x^*, y^*).$$

Thus, (x^*, y^*) cannot be a global maximizer, and our original assumption must be false. Specifically, $\nexists (x, y) \in S$ that is a global maximizer of $f(x, y) = x + y$.

Theorem 1.3.3.

(a) The union of any collection of open sets is an open set.

(b) The intersection of any finite collection of open sets is an open set.

Proof.

(a) Let G_i be an arbitrary open set and let $S = \bigcup_{i=1}^{\infty} G_i$, where $G = \{G_1, G_2, \ldots\}$ is an arbitrary collection of open sets. If $x \in S$, then $x \in G_i$ for some $G_i \in G$. Since G_i is open, x is an interior point of G_i. That is, there exists a neighborhood N of x such that $N \subseteq G_i$. But $G_i \subseteq S$, so $N \subseteq S$, implying that x is an interior point of S. Since we have shown that a neighborhood of an arbitrary point in S is completely contained in S, S is open.

(b) First note that this result does not hold for infinite collections of sets. To see why, notice that for each $n \in \mathbb{N}$, if we define $A_n = (-1/n, 1/n)$, then each A_n is an open set. However, $\bigcap_{n=1}^{\infty} A_n = \{0\}$, which is not open. Thus we see that we cannot generalize the above result to the intersection of an infinite collection of open sets.

Consider the finite case. Define $S := \bigcap_{i=1}^{n} G_i$, where $G = \{G_1, G_2, \ldots, G_n\}$ is an arbitrary collection of open sets. If $S = \varnothing$, we are done, since $\varnothing$ is open (and closed) ($\operatorname{int} \varnothing = \varnothing$). If $S \neq \varnothing$, let $x \in S$. Then $x \in G_i$ for all $i = 1, 2, \ldots, n$. Since each G_i is open, there exist neighborhoods $N_i(x; \varepsilon_i)$ of x such that $N_i(x; \varepsilon_i) \subseteq G_i$. Let $\varepsilon = \min\{\varepsilon_1, \ldots, \varepsilon_n\}$. Then $N(x; \varepsilon) \subseteq G_i$ for each $i = 1, \ldots, n$, so $N(x; \varepsilon) \subseteq S$. Thus x is an interior point of S, and S is open. □

Corollary 1.3.1.

(a) The intersection of any collection of closed sets is a closed set.

(b) The union of any finite collection of closed sets is a closed set.

Proof.

(a) To prove the above result, define $T := \bigcap_{i=1}^{\infty} F_i$, where $F = \{F_1, F_2, \ldots\}$ is an arbitrary infinite collection of closed sets. If $T = \varnothing$, we are done, since $\varnothing$ is closed (and open) ($\mathrm{bd}\,\varnothing = \varnothing \subseteq \varnothing$). $\mathbb{R} \setminus (\bigcap_{i=1}^{\infty} F_i) = \bigcup_{i=1}^{\infty} (\mathbb{R} \setminus F_i)$ (the complement of the intersection will be the union of the individual complements, which can be seen using a simple venn diagram). Thus, we have $\mathbb{R} \setminus (\bigcap_{i=1}^{\infty} F_i)$ equal to the union of open sets, since a set is closed if and only if its complement is open. Since we have shown above that the union of any collection of open sets is open, we have that $\mathbb{R} \setminus (\bigcap_{i=1}^{\infty} F_i)$ is open. Thus, $\bigcap_{i=1}^{\infty} F_i$ must be closed.

(b) To prove the above result, define $T := \bigcup_{i=1}^{n} F_i$, where $F = \{F_1, F_2, \ldots, F_n\}$ is an arbitrary finite collection of closed sets. $\mathbb{R} \setminus (\bigcup_{i=1}^{n} F_i) = \bigcap_{i=1}^{n} (\mathbb{R} \setminus F_i)$ (the complement of the union will be the intersection of the individual complements, which again can be seen using a simple venn diagram). Thus, we have $\mathbb{R} \setminus (\bigcup_{i=1}^{n} F_i)$ equal to the intersection of open sets. Since we have shown above that the intersection of any finite collection of open sets is open, we have that $\mathbb{R} \setminus (\bigcup_{i=1}^{n} F_i)$ is open. Thus, $\bigcup_{i=1}^{n} F_i$ must be closed by definition. □

1.3.3 Convergence and Boundedness

A sequence, $\{s_n\}$, is a function whose domain is the set $\mathbb{N}$ of natural numbers. If s is a sequence, denote its value at n by s_n.

Definition 1.3.5 (Convergence)**.** *A sequence $\{s_n\}$ is said to converge to the real number s provided that for each $\varepsilon > 0$, there exists $N \in \mathbb{R}$ such that for all $n \in \mathbb{N}$, $n > N$ implies that $|s_n - s| < \varepsilon$. If $\{s_n\}$ converges to s, then s is called the limit of the sequence $\{s_n\}$, and we write $\lim_{n\to\infty} s_n = s$ or simply $s_n \to s$. If a sequence does not converge to a real number, it is said to diverge.*

Example 1.3.10. Prove that $\lim 1/n = 0$.

Solution: Given $\varepsilon > 0$, let $N = 1/\varepsilon$. Then for any $n > N$, $|1/n - 0| = 1/n < 1/N = \varepsilon$.

Example 1.3.11. Prove that $\lim(n^2 + 2n)/(n^3 - 5) = 0$

Solution: Given $\varepsilon > 0$, let $N = \max\{3, 4/\varepsilon\}$. Then $n > N$ implies that $n > 3$ and $n > 4/\varepsilon$. Since $n > 3$, we have $n^2 + 2n < 2n^2$ and $n^3 - 5 > n^3/2$. Thus for $n > N$ we have

$$\left| \frac{n^2 + 2n}{n^3 - 5} - 0 \right| = \frac{n^2 + 2n}{n^3 - 5} < \frac{2n^2}{\frac{1}{2}n^3} = \frac{4}{n} < \frac{4}{N} = \varepsilon.$$

Definition 1.3.6 (Bounded Sequence)**.** *A sequence $\{s_n\}$ is said to be bounded if the range $\{s_n : n \in \mathbb{N}\}$ is a bounded set. That is if there exists an $M \geq 0$ such that $|s_n| \leq M$ for all $n \in \mathbb{N}$.*

Theorem 1.3.4. *Every convergent sequence in $\mathbb{R}^n$ is bounded.*

Proof. Let s_n be a convergent sequence and let $s_n \to s$. From the definition of convergence with $\varepsilon = 1$, we obtain $N \in \mathbb{R}$ such that $|s_n - s| < 1$ whenever $n > N$. Thus for $n > N$ the triangle inequality implies that $|s_n| < |s| + 1$. We know that s_n is bounded if the range $\{s_n : n \in \mathbb{N}\}$ is a bounded set, that is, if there exists an $M \geq 0$ such that $|s_n| \leq M$ for all $n \in \mathbb{N}$. Thus, let

$$M = \max\{|s_1|, \ldots, |s_N|, |s| + 1\},$$

(so that M could either be $|s| + 1$ or the largest absolute value among the first N terms) then we have $|s_n| \leq M$ for all $n \in \mathbb{N}$, so s_n is bounded. □

Theorem 1.3.5. *Suppose that $\{s_n\}$ and $\{t_n\}$ are convergent sequences with $\lim s_n = s$ and $\lim t_n = t$. Then*

(a) $\lim(s_n + t_n) = s + t$

(b) $\lim(ks_n) = ks$ *and* $\lim(k + s_n) = k + s$ *for any* $k \in \mathbb{R}$

(c) $\lim(s_n t_n) = st$

(d) $\lim(s_n/t_n) = s/t$, *provided that* $t_n \neq 0$ *for all* n *and* $t \neq 0$

Proof.

(a) To show that $s_n + t_n \to s + t$, we need to make the difference $|(s_n + t_n) - (s + t)|$ small. Using the triangle inequality, we have

$$|(s_n + t_n) - (s + t)| = |(s_n - s) + (t_n - t)| \leq |s_n - s| + |t_n - t|.$$

Now, given any $\varepsilon > 0$, since $s_n \to s$, there exists N_1 such that $n > N_1$ implies that $|s_n - s| < \frac{\varepsilon}{2}$. Similarly, since $t_n \to t$, there exists N_2 such that $n > N_2$ implies that $|t_n - t| < \frac{\varepsilon}{2}$. Thus if we let $N = \max\{N_1, N_2\}$, then $n > N$ implies that

$$|(s_n - s) + (t_n - t)| \leq |s_n - s| + |t_n - t| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

(b) To show that $ks_n \to ks$ for any $k \in \mathbb{R}$, we need to make the difference $|ks_n - ks|$ small. We have

$$\begin{aligned} |ks_n - ks| &= |k(s_n - s)| \\ &= |k||s_n - s|. \end{aligned}$$

Now, given any $\varepsilon > 0$, since $s_n \to s$, there exists N such that $n > N$ implies that $|s_n - s| < \varepsilon/|k|$. Thus for $n > N$

$$|ks_n - ks| = |k||s_n - s| < |k|\frac{\varepsilon}{|k|} = \varepsilon.$$

To show that $k + s_n \to k + s$, note that

$$|(k + s_n) - (k + s)| = |s_n - s|.$$

Thus, for $n > N$, $k + s_n \to k + s$.

(c) Using the triangle inequality

$$\begin{aligned} |s_n t_n - st| &= |(s_n t_n - s_n t) + (s_n t - st)| \\ &\leq |s_n t_n - s_n t| + |s_n t - st| \\ &= |s_n||t_n - t| + |t||s_n - s|. \end{aligned}$$

We know that every convergent sequence is bounded (Theorem 1.3.4). Thus, s_n is bounded, and there exists $M_1 > 0$ such that $|s_n| \leq M_1$ for all n. Letting $M = \max\{M_1, |t|\}$, we obtain the inequality

$$|s_n t_n - st| \leq M|t_n - t| + M|s_n - s|.$$

Now, given ant $\varepsilon > 0$, there exists N_1 and N_2 such that $|t_n - t| < \frac{\varepsilon}{2M}$ when $n > N_1$ and $|s_n - s| < \frac{\varepsilon}{2M}$ when $n > N_2$. Let $N = \max\{N_1, N_2\}$. Then $n > N$ implies that

$$\begin{aligned} |s_n t_n - st| &\leq M|t_n - t| + M|s_n - s| \\ &< M\left(\frac{\varepsilon}{2M}\right) + M\left(\frac{\varepsilon}{2M}\right) = \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

(d) Since $s_n/t_n = s_n(1/t_n)$, it suffices from part (c) to show that $\lim 1/t_n = 1/t$. That is, given $\varepsilon > 0$, we must show that

$$\left|\frac{1}{t_n} - \frac{1}{t}\right| = \left|\frac{t - t_n}{t_n t}\right| < \varepsilon$$

for all n sufficiently large. To get a lower bound on how small the denominator can be, we note that since $t \neq 0$, there exists an N_1 such that $n > N_1$ implies that $|t_n - t| < |t|/2$. Thus for $n > N_1$ we have

$$|t_n| = |t - (t - t_n)| \geq |t| - |t - t_n| > |t| - \frac{|t|}{2} = \frac{|t|}{2}$$

by the reverse triangle inequality. There also exists an N_2 such that $n > N_2$ implies that $|t_n - t| < \frac{1}{2}\varepsilon|t|^2$. Let $N = \max\{N_1, N_2\}$. Then $n > N$ implies that

$$\left|\frac{1}{t_n} - \frac{1}{t}\right| = \left|\frac{t - t_n}{t_n t}\right| < \frac{2}{|t|^2}|t_n - t| < \varepsilon.$$

Hence $\lim(1/t_n) = 1/t$.

□

Theorem 1.3.6. *A set $S \subseteq \mathbb{R}^n$ is closed if and only if every convergent sequence in S converges to a point in S.*

Proof. Suppose that S is closed and that x_n is a sequence in S converging to x. Suppose that $x \notin S$. Because S is closed, $\mathbb{R}^n \backslash S$ is open, so that for some $\varepsilon > 0$, there exists $N(x;\varepsilon) \subseteq \mathbb{R}^n \backslash S$. That is $N(x;\varepsilon) \bigcap S = \varnothing$. Since $\lim x_n = x$, there is an N such that $x_n \in N(x;\varepsilon)$, if $n > N$. But then $x_n \notin S$, whenever $n > N$, which is impossible since x_n is a sequence in S. Thus it must be the case that if a set S is closed, then every convergent sequence in S must converge to a point in S.

Suppose that S is not closed. Then $\mathbb{R}^n \backslash S$ is not open, implying that there exists an $x \in \mathbb{R}^n \backslash S$ such that $N(x;\varepsilon) \bigcap S \neq \varnothing$, for every $\varepsilon > 0$. Practically, this means that at least part of the neighborhood of x lies in S. In particular, for every positive integer n, there exists an $x_n \in S$, such that $|x_n - x| < 1/n$. This gives us a sequence $\{x_n\}$ in S converging to a point x not in S. Thus, by the contra-positive argument (not closed implies there exists a convergent sequence in S that converges to a point not in S), we have that if every convergent sequence in S converges to a point in S, then S must be closed. □

1.3.4 Compactness

Theorem 1.3.7 (Heine-Borel)**.** *A subset S of $\mathbb{R}^n$ is compact iff S is closed and bounded.*

Example 1.3.12. Let $S \subseteq \mathbb{R}^2$ be defined as follows:

$$S = \{(x, y) \in \mathbb{R}^2 : y = \sin(1/x), x > 0 \cup \{(0, 0)\}\}$$

Is S closed? Open? Bounded? Compact?

Solution: None. Let $s_k = (x_k, y_k)$. S is bounded if there exists an $\varepsilon > 0$ such that $|s_i - s_j| \leq \varepsilon$ for all $s_k \in S$. Since $x \in [0, +\infty)$, there does not exist a neighborhood of radius ε that contains S. Thus, S is not bounded, and hence not compact (Heine-Borel Theorem). In addition, S is not closed. To see this, recall that a set is closed if every convergent sequence in the set has its limit in the set. That is, a closed set is a set that is closed under limit operations. Now, consider the sequence $\{(x_k, y_k)\}_{k\geq 1} \in S$, given by $(x_k, y_k) = (\frac{2}{\pi(4k+1)}, 1)$, for $k \geq 1$. Then we have $(x_k, y_k) \in S$ since $\sin(1/x_k) = \sin[(4k + 1)(\pi/2)] = 1 = y_k$. We can now see that $(x_k, y_k) \to (0, 1)$ as $k \to \infty$. But $(0, 1) \notin S$, and as a consequence, S is not closed. Also, since any ε-neighborhood around any point in S is not contained in S, there are no interior points ($\text{int}\, S = \varnothing$). Thus, S is not open.

Chapter 2

Basic Matrix Properties and Operations

2.1 Determinants

2.1.1 Minors, Cofactors, and Evaluating Determinants

Definition 2.1.1 (Minor and Cofactor of a Matrix)**.** *Let A be an $n \times n$ matrix. Let A_{ij} be the $(n-1) \times (n-1)$ submatrix obtained by deleting row i and column j from A. Then, the scalar*

$$M_{ij} \equiv \det A_{ij}$$

is called the (i,j)th minor of A and the scalar

$$C_{ij} \equiv (-1)^{i+j} M_{ij}$$

is called the (i,j)th cofactor of A.

Definition 2.1.2 (Determinant)**.** *The determinant of an $n \times n$ matrix A is given by*

$$\begin{aligned} \det A &= a_{11}C_{11} + a_{12}C_{12} + \cdots + a_{1n}C_{1n} \\ &= a_{11}M_{11} - a_{12}M_{12} + \cdots + (-1)^{n+1} a_{1n}M_{1n}. \end{aligned}$$

Example 2.1.1. To calculate the determinant of a 2×2 matrix we have

$$\det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21}. \tag{2.1}$$

To calculate the determinant of a 3×3 matrix we have

$$\begin{aligned} \det(A_{3\times 3}) &= \det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \\ &= a_{11} \det \begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} - a_{12} \det \begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix} + a_{13} \det \begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}. \end{aligned} \tag{2.2}$$

2.1.2 Properties of Determinants

The following is a list of *some* useful properties of determinants:

(a) Rows and columns can be interchanged without affecting the value of a determinant. That is

$$|A| = |A^T|.$$

(b) If two rows (or columns) are interchanged, the sign of the determinant changes. For example

$$\begin{vmatrix} 3 & 4 \\ 1 & -2 \end{vmatrix} = - \begin{vmatrix} 1 & -2 \\ 3 & 4 \end{vmatrix}.$$

(c) If a row (or column) is changed by adding to or subtracting from its elements the corresponding elements of any other row (or column), the determinant remains unaltered. For example

$$\begin{vmatrix} 3 & 4 \\ 1 & -2 \end{vmatrix} \sim \begin{vmatrix} 3+1 & 4-2 \\ 1 & -2 \end{vmatrix} = \begin{vmatrix} 4 & 2 \\ 1 & -2 \end{vmatrix} = -10.$$

(d) If the elements in any row (or column) have a common factor α, then the determinant equals the determinant of the corresponding matrix in which $\alpha = 1$, multiplied by α. For example

$$\begin{vmatrix} 6 & 8 \\ 1 & -2 \end{vmatrix} = 2 \begin{vmatrix} 3 & 4 \\ 1 & -2 \end{vmatrix} = 2 \times (-10) = -20.$$

(e) When at least one row (or column) of a matrix is a linear combination of the other rows (or columns), the determinant is zero. Conversely, if the determinant is zero, then at least one row and/or one column are linearly dependent on the other rows and columns, respectively. For example the

$$\det \begin{pmatrix} 3 & 2 & 1 \\ 1 & 2 & -1 \\ 2 & -1 & 3 \end{pmatrix}$$

is zero because the first column is a linear combination of the second and third columns

$$\text{column } 1 = \text{column } 2 + \text{column } 3.$$

Similarly, there is a linear dependence between the rows, which is given by the relation

$$\text{row } 1 = \frac{7}{8} \text{ row } 2 + \frac{4}{5} \text{ row } 3.$$

(f) The determinant of an upper triangular or lower triangular matrix is the product of the main diagonal entries. For example:

$$\begin{vmatrix} 3 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & 4 \end{vmatrix} = 3 \times 2 \times 4 = 24.$$

(g) The determinant of the product of two square matrices is the product of the individual determinants. Fore example

$$|AB| = |A||B|.$$

This rule can be generalized to any number of factors. One immediate application is to matrix powers: $|A^2| = |A||A| = |A|^2$, and more generally $|A^n| = |A|^n$ for integer n.

2.1.3 Singular Matrices and Rank

If the determinant of a $n \times n$ square matrix is zero, then the matrix is said to be *singular*. This means that at least one row and/or one column are linearly dependent on the others. The rank r of a matrix A, written $r = \text{rank } A$, is the greatest number of linearly independent rows or columns that exist in the matrix A. Numerically, r is equal to the order of the largest non-vanishing determinant $|B|$ associated with any square matrix B, which can be constructed from A by a combination of r rows and r columns. If the determinant of A is nonzero, then A is said to be *nonsingular*. The rank of a nonsingular $n \times n$ matrix is equal to n. The rank of A^T is the same as that of A, since it is only necessary to swap "rows" and "columns" in the definition.

Example 2.1.2. The 3×3 matrix

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

has rank $r = 3$ because $|A| = -5 \neq 0$.

Example 2.1.3. The matrix

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

already used in an above section (2.1.2 on page 33) is singular because its first row and column may be expressed as linear combinations of the others. Removing the first row and column, we are left with a 2×2 matrix whose determinant is $2(3) - (-1)(-1) = 5 \neq 0$. Consequently, matrix A has rank $r = 2$.

2.2 Inverses of Matrices

Definition 2.2.1 (Matrix Inverse)**.** *Let $I = [e_1, e_2, \ldots, e_n]$ be an $n \times n$ identity matrix. Let A be an $n \times n$ matrix. An $n \times n$ matrix, B, is called an inverse of A if and only if $AB = I$ and $BA = I$. B is typically written as A^{-1}.*

The most important application of inverses is the solution of linear systems. Suppose

$$A\mathbf{x} = \mathbf{y},$$

where $\mathbf{x}$ and $\mathbf{y}$ are $n \times 1$ column vectors. Premultiplying both sides by A^{-1} we get the inverse relationship

$$\mathbf{x} = A^{-1}\mathbf{y}.$$

More generally, consider the matrix equation for multiple (m) right-hand sides:

$$\underset{n\times n}{A}\ \underset{n\times m}{X} = \underset{n\times m}{Y},$$

which reduces to $A\mathbf{x} = \mathbf{y}$ for $m = 1$. The inverse relation that gives X as a function of Y is

$$X = A^{-1}Y.$$

In particular, the solution of

$$AX = I$$

is $X = A^{-1}$. Practical methods for computing inverses are based on directly solving this equation.

2.2.1 Computation of Inverses

The explicit calculation of matrix inverses is seldom needed in large matrix computations. But, occasionally the need arises for calculating the explicit inverse of small matrices by hand. For example, the inversion of Jacobian matrices.

A general formula for elements of the inverse is as follows. Let $B = [b_{ij}] = A^{-1}$. Then

$$b_{ij} = \frac{C_{ji}}{|A|},$$

where b_{ij} denotes the entries of A^{-1}. C_{ij} is defined as (i,j)th cofactor of A or, more precisely, the determinant of the submatrix of order $(n-1)\times(n-1)$ obtained by deleting the i^{th} row and j^{th} column of A, multiplied by $(-1)^{i+j}$. The $n \times n$ matrix whose (i,j)th entry is C_{ji} is called the adjugate (or classical adjoint) of A and is written adj A.[1]

This direct inversion procedure is useful only for small matrix orders: 2 or 3. In the examples below the inversion formulas for second and third order matrices are listed.

Example 2.2.1. For order $n = 2$:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \text{implies} \quad A^{-1} = \frac{1}{|A|}\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix},$$

where $|A|$ is given by equation 2.1 on page 33.

Example 2.2.2. For order $n = 3$:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \qquad A^{-1} = \frac{1}{|A|}\begin{bmatrix} C_{11} & C_{21} & C_{31} \\ C_{12} & C_{22} & C_{32} \\ C_{13} & C_{23} & C_{33} \end{bmatrix}$$

where

$$C_{11} = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}, \quad C_{21} = -\begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix}, \quad C_{31} = \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix},$$

$$C_{12} = -\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix}, \quad C_{22} = \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}, \quad C_{32} = -\begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix},$$

$$C_{13} = \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}, \quad C_{23} = -\begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix}, \quad C_{33} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix},$$

and $|A|$ is given by equation 2.2 on page 33.

Example 2.2.3.

$$A = \begin{bmatrix} 2 & 4 & 2 \\ 3 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \qquad A^{-1} = -\frac{1}{8}\begin{bmatrix} 1 & -4 & 2 \\ -2 & 0 & 4 \\ -1 & 4 & -10 \end{bmatrix}.$$

If the order exceeds 3, the general inversion formula becomes rapidly useless as it displays combinatorial complexity.

[1] The adjugate has sometimes been called the "adjoint", but that terminology is ambiguous. Today, "adjoint" of a matrix normally refers to its conjugate transpose.

2.2.2 Properties of Inverses

The following is a list of *some* useful properties of inverses:

(a) The inverse of the transpose is equal to the transpose of the inverse. That is

$$(A^T)^{-1} = (A^{-1})^T,$$

because

$$AA^{-1} = (AA^{-1})^T = (A^{-1})^T A^T = I.$$

(b) The inverse of a symmetric matrix is also symmetric. Given the previous rule, $(A^T)^{-1} = A^{-1} = (A^{-1})^T$, hence A^{-1} is also symmetric.

(c) The inverse of a matrix product is the reverse product of the inverses of the factors. That is

$$(AB)^{-1} = B^{-1}A^{-1}.$$

This property generalizes to an arbitrary number of factors.

(d) For a diagonal matrix D in which all diagonal entries are nonzero, D^{-1} is again a diagonal matrix with entries $1/d_{ii}$.

(e) If $\mathbf{S}$ is a block diagonal matrix:

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{22} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S}_{33} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{S}_{nn} \end{bmatrix} = diag[\mathbf{S}_{ii}],$$

then the inverse matrix is also block diagonal and is given by

$$\mathbf{S}^{-1} = \begin{bmatrix} \mathbf{S}_{11}^{-1} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{22}^{-1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S}_{33}^{-1} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{S}_{nn}^{-1} \end{bmatrix} = diag[\mathbf{S}_{ii}^{-1}].$$

(f) The inverse of an upper (lower) triangular matrix is also an upper (lower) triangular matrix.

2.3 Quadratic Forms and Definiteness

2.3.1 Quadratic Forms

Definition 2.3.1 (Quadratic Form)**.** *Let A denote an $n \times n$ symmetric matrix with real entries and let $\mathbf{x}$ denote an $n \times 1$ column vector. Then*

$$Q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x}$$

is said to be of quadratic form.

Note that

$$\begin{aligned} Q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} &= \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \\ &= \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n a_{1i}x_i \\ \sum_{i=1}^n a_{2i}x_i \\ \vdots \\ \sum_{i=1}^n a_{ni}x_i \end{bmatrix} \\ &= \sum_{i,j} a_{ij}x_i x_j. \end{aligned}$$

For example, consider the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

and the vector $\mathbf{x}$. Q is given by

$$\begin{aligned} Q = \mathbf{x}^T A\mathbf{x} &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= \begin{bmatrix} x_1 + 2x_2 & 2x_1 + x_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= x_1^2 + 4x_1x_2 + x_2^2, \end{aligned}$$

which is clearly of quadratic form.

Every quadratic form has a critical point at $\mathbf{x} = 0$. Therefore, we can classify quadratic forms by whether $\mathbf{x} = 0$ is a maximum, minimum, or neither. This is what definiteness is about.

2.3.2 Definiteness of Quadratic Forms

Definition 2.3.2 (Positive/Negative Definiteness)**.** *Let A be an $n \times n$ symmetric matrix, then A is*

(a) positive definite if $\mathbf{x}^T A\mathbf{x} > 0$ for all $\mathbf{x} \neq 0$ in $\mathbb{R}^n$,

(b) positive semi-definite if $\mathbf{x}^T A\mathbf{x} \geq 0$ for all $\mathbf{x} \neq 0$ in $\mathbb{R}^n$,

(c) negative definite if $\mathbf{x}^T A\mathbf{x} < 0$ for all $\mathbf{x} \neq 0$ in $\mathbb{R}^n$,

(d) negative semi-definite if $\mathbf{x}^T A\mathbf{x} \leq 0$ for all $\mathbf{x} \neq 0$ in $\mathbb{R}^n$, and

(e) indefinite if $\mathbf{x}^T A\mathbf{x} > 0$ for some $\mathbf{x}$ in $\mathbb{R}^n$ and < 0 for some other $\mathbf{x}$ in $\mathbb{R}^n$.

Example 2.3.1. Consider a 3×3 diagonal matrix D given by

$$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

and a general 3-element vector $\mathbf{x}$. The general quadratic form is given by

$$\begin{aligned} Q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x} &= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= \begin{bmatrix} x_1 & 2x_2 & 4x_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= x_1^2 + 2x_2^2 + 4x_3^2. \end{aligned}$$

Note that for any real vector $\mathbf{x} \neq 0$, $Q(\mathbf{x})$ will be positive, since the square of any number is positive, the coefficients of the squared terms are positive, and the sum of positive numbers is always positive. Thus, A is positive definite.

Example 2.3.2. Now consider an alternative 3×3 matrix given by

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

The general quadratic form is given by

$$\begin{aligned} Q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x} &= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} -2 & 1 & 0 \\ 1 & -2 & 0 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= \begin{bmatrix} -2x_1 + x_2 & x_1 - 2x_2 & -2x_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= -2x_1^2 + 2x_1x_2 - 2x_2^2 - 2x_3^2 \\ &= -2x_1^2 - 2[x_2^2 - x_1x_2] - 2x_3^2. \end{aligned}$$

Note that $Q(\mathbf{x})$ will be negative if x_1 and x_2 are of opposite sign or equal to one another. Now consider the case where $|x_1| > |x_2|$. Write $Q(\mathbf{x})$ as

$$Q(\mathbf{x}) = -2x_1^2 + 2x_1x_2 - 2x_2^2 - 2x_3^2.$$

The first, third, and fourth terms are clearly negative. With $|x_1| > |x_2|$, $|2x_1^2| > |2x_1x_2|$, the first term is more negative than the second term is positive, and hence the whole expression is negative. Now consider the case where $|x_1| < |x_2|$. The first, third, and fourth terms are still negative. But, with $|x_1| < |x_2|$, $|2x_2^2| > |2x_1x_2|$ so that the third term is more negative than the second term is positive, and so the whole expression is negative. Thus, this quadratic form is negative definite for all real values of $\mathbf{x} \neq 0$.

Remark 2.3.1. A matrix that is positive (negative) definite is positive (negative) semi-definite.

The definiteness of a matrix plays an important role. For example, for a function $f(x)$ of one variable, the sign of the second derivative $f''(x_0)$ at a critical point x_0 gives a sufficient condition for determining whether x_0 is a maximum, minimum, or neither (proposition 1.1.2). This test generalizes to more than one variable using the definiteness of the Hessian matrix H (more on this when we get to optimization). There is a convenient way to test for the definiteness of a matrix, but before we can formulate this test we first need to define the concept of principal minors of a matrix.

Definition 2.3.3 (Principle Submatrix and Principle Minor)**.** *Let A be an $n \times n$ matrix. A $k \times k$ submatrix of A formed by deleting $(n-k)$ columns, say columns $i_1, i_2, ...i_{n-k}$ and the same $n-k$ rows, rows $i_1, i_2, ...i_{n-k}$, from A is called a k^{th} order principal submatrix of A. The determinant of a $k \times k$ principal submatrix denoted B_k is called a k^{th} order principal minor of A.*

Definition 2.3.4 (Leading Principle Submatrix and Leading Principle Minor)**.** *Let A be an $n \times n$ matrix. The kth order principle submatrix of A obtained by deleting the last $n-k$ rows and the last $n-k$ columns from A is called the kth order leading principle submatrix of A. Its determinant, denoted Δ_k, is called the kth order leading principle minor of A.*

Example 2.3.3. For a general 3×3 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

there is one third order principle minor: $B_3 = \det(A) = |A|$. There are three second order principle minors:

1. $B_2^{(1)} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$, formed by deleting column 3 and row 3 from A;
2. $B_2^{(2)} = \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}$, formed by deleting column 2 and row 2 from A;
3. $B_2^{(3)} = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}$, formed by deleting column 1 and row 1 from A.

There are three first order principle minors:

1. $B_1^{(1)} = a_{11}$, formed by deleting the last 2 rows and columns;
2. $B_1^{(2)} = a_{22}$, formed by deleting the first and third columns and rows, and
3. $B_1^{(3)} = a_{33}$, formed by deleting the first 2 rows and columns.

The leading principle minors of order k, denoted Δ_k, are:

$$\Delta_1 = a_{11},\ \Delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix},\ \text{and } \Delta_3 = \det(A) = |A|.$$

Theorem 2.3.1 (Positive/Negative Definiteness)**.** *Let A be an $n \times n$ symmetric matrix. Then,*

(a) A is positive definite if and only if $\Delta_k > 0$ for $k = 1, 2, \ldots, n$ (all of its leading principle minors are strictly positive);

(b) A is negative definite if and only if $(-1)^k \Delta_k > 0$ for $k = 1, 2, \ldots, n$ (every leading principle minor of odd order is strictly negative and every leading principle minor of even order is strictly positive);

(c) If some kth order principle minor of A (or some pair of them) is nonzero but does not fit either of the above two sign patterns, then A is indefinite.

Theorem 2.3.2 (Positive/Negative Semi-Definiteness)**.** *Let A be an $n \times n$ symmetric matrix. Then,*

(a) A is positive semi-definite if and only if $B_k \geq 0$ for $k = 1, 2, \ldots, n$ (every principle minor of A is nonnegative);

(b) A is negative semi-definite if and only if $(-1)^k B_k \geq 0$ for $k = 1, 2, , \ldots, n$ (every principle minor of odd order is nonpositive and every principle minor of even order is nonnegative).

Remark 2.3.2. Given an $n \times n$ symmetric matrix, the conditions $\Delta_k \geq 0$ for positive semi-definiteness and $(-1)^k \Delta_k \geq 0$ for negative semi-definiteness are necessary conditions but not sufficient conditions. To see this, consider the following 2×2 symmetric matrix.

$$A = \begin{pmatrix} 0 & 0 \\ 0 & -4 \end{pmatrix}.$$

For this matrix, $\Delta_1 = 0$ and $\Delta_2 = 0$. Thus, looking only at leading principle minors, one could falsely conclude that this matrix is positive semi-definite. However, we have to check all principle minors to deduce the correct form of semi-definiteness. If one checks all principle minors then

$$\begin{aligned} B_1^{(1)} &= 0 \leq 0, \\ B_1^{(2)} &= -4 \leq 0, \\ B_2 &= \begin{vmatrix} 0 & 0 \\ 0 & -4 \end{vmatrix} = 0 \geq 0, \end{aligned}$$

which violates the definition of positive semi-definiteness. In fact, this matrix is negative semi-definite.

Example 2.3.4. Suppose A is a 4×4 symmetric matrix. Then

(a) $\Delta_1 > 0, \Delta_2 > 0, \Delta_3 > 0, \Delta_4 > 0 \quad \rightarrow$ positive definite

(b) $\Delta_1 < 0, \Delta_2 > 0, \Delta_3 < 0, \Delta_4 > 0 \quad \rightarrow$ negative definite

(c) $\Delta_1 > 0, \Delta_2 > 0, \Delta_3 = 0, \Delta_4 < 0 \quad \rightarrow$ indefinite because of Δ_4

(d) $\Delta_1 < 0, \Delta_2 < 0, \Delta_3 < 0, \Delta_4 < 0 \quad \rightarrow$ indefinite because of Δ_2 and Δ_4

(e) $\Delta_1 = 0, \Delta_2 < 0, \Delta_3 > 0, \Delta_4 = 0 \quad \rightarrow$ indefinite because of Δ_2

(f) $\Delta_1 > 0, \Delta_2 = 0, \Delta_3 > 0, \Delta_4 > 0 \quad \rightarrow$ A is not definite, not negative semi-definite but might be positive semi-definite. However, to establish this, we must to check *all* 15 principle minors B_k, for $k = 1, 2, 3, 4$.

(g) $\Delta_1 = 0, \Delta_2 > 0, \Delta_3 = 0, \Delta_4 > 0 \quad \rightarrow$ A is not definite, but may be positive semi-definite or negative semi-definite. To determine semi-definiteness we must check all principle minors.

Example 2.3.5. Use the above theorems to check the following matrices for definiteness:

(a)

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}, \qquad \Delta_1 = 1 > 0, \ \Delta_2 = -3 < 0$$

Thus A is indefinite.

(b)

$$A = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}, \qquad B_1^{(1)} = B_1^{(2)} = -1 < 0,\ \Delta_2 = 0$$

Thus A is negative semi-definite.

(c)

$$B = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 3 & 0 \\ -1 & 0 & 4 \end{bmatrix}, \qquad \Delta_1 = 1 > 0,\ \Delta_2 = 3 > 0,\ \Delta_3 = 9 > 0$$

Thus, B is positive definite.

(d)

$$B = \begin{bmatrix} 1 & 3 & 2 & 3 \\ 3 & 5 & -2 & 4 \\ 2 & -2 & 2 & 1 \\ 3 & 4 & 1 & 2 \end{bmatrix}, \quad \Delta_1 = 1 > 0,\ \Delta_2 = -4 < 0,\ \Delta_3 = -56 < 0,\ \Delta_4 = 142 > 0$$

Thus B is indefinite.

2.4 Eigenvalues and Eigenvectors

Consider the special form of the linear system $A\mathbf{x} = \mathbf{y}$ in which the right-hand side vector $\mathbf{y}$ is a multiple of the solution vector $\mathbf{x}$:

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \tag{2.3}$$

or, written in full,

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & \lambda x_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & \lambda x_2 \\ \vdots & & \vdots & & \cdots & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \cdots & + & a_{nn}x_n & = & \lambda x_n. \end{array}$$

This is called the standard (or classical) *algebraic eigenproblem*.

Definition 2.4.1 (Eigenvalues and Eigenvectors)**.** *The number λ is said to be an eigenvalue of the $n \times n$ matrix A provided that there exists a nonzero vector $\mathbf{x}$ such that*

$$A\mathbf{x} = \lambda\mathbf{x},$$

in which case the vector $\mathbf{x}$ is called an eigenvector of the matrix A.

The system (2.3) can be rearranged into homogeneous form

$$(A - \lambda I)\mathbf{x} = \mathbf{0}.$$

A nontrivial solution of this equation is possible if and only if the coefficient matrix $A - \lambda I$ is singular. Such a condition can be expressed as the vanishing of the determinant

$$|A - \lambda I| = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = \mathbf{0}.$$

When this determinant is expanded, we obtain an algebraic polynomial equation in λ of degree n:

$$P(\lambda) = \lambda^n + \alpha_1\lambda^{n-1} + \cdots + \alpha_n = 0.$$

This is known as the *characteristic equation* of the matrix A. The left-hand side is known as the *characteristic polynomial*. We know that a polynomial of degree n has n (generally complex) roots $\lambda_1, \lambda_2, \ldots, \lambda_n$. These n numbers are called *eigenvalues*, *eigenroots*, or *characteristic values* of the matrix A. The following theorem summarizes the results.

Theorem 2.4.1. *The number λ is an eigenvalue of the $n \times n$ matrix A if and only if λ satisfies the characteristic equation*

$$|A - \lambda I| = 0.$$

With each eigenvalue λ_i, there is an associated vector $\mathbf{x}_i$ that satisfies

$$A\mathbf{x}_i = \lambda\mathbf{x}_i.$$

This $\mathbf{x}_i$ is called an *eigenvector* or *characteristic vector* of the matrix A.

An eigenvector is unique only up to a scale factor since if $\mathbf{x}_i$ is an eigenvector, so is $\beta\mathbf{x}_i$, where β is an arbitrary nonzero number. For a general 2×2 matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

the characteristic polynomial is given by

$$|A - \lambda I| = \lambda^2 - (a_{11} + a_{22})\lambda + a_{11}a_{22} - a_{12}a_{21},$$

and we can solve for the associated eigenvalues by setting the above characteristic equation to zero and solving for λ.

Example 2.4.1. Find the eigenvalues and associated eigenvectors of the matrix

$$A = \begin{bmatrix} 5 & 7 \\ -2 & -4 \end{bmatrix}.$$

Solution:

$$A - \lambda I = \begin{bmatrix} 5-\lambda & 7 \\ -2 & -4-\lambda \end{bmatrix} \tag{2.4}$$

so the characteristic equation of A is

$$\begin{aligned} 0 &\stackrel{\text{set}}{=} \det(A - \lambda I) \\ &= \begin{vmatrix} 5-\lambda & 7 \\ -2 & -4-\lambda \end{vmatrix} \\ &= (5-\lambda)(-4-\lambda) - (-2)(7) \\ &= \lambda^2 - \lambda - 6 \\ &= (\lambda + 2)(\lambda - 3). \end{aligned}$$

Thus, the matrix A has two eigenvalues -2 and 3. To distinguish them, we write $\lambda_1 = -2$ and $\lambda_2 = 3$. To find the associated eigenvectors, we must separately substitute each eigenvalue into (2.4) and then solve the resulting system $(A - \lambda I)\mathbf{x} = \mathbf{0}$.

Case 1: $\lambda_1 = -2$. With $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$, the system $(A - \lambda I)\mathbf{x} = 0$ is

$$\begin{bmatrix} 7 & 7 \\ -2 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Each of the two scalar equations here is a multiple of the equation $x_1 + x_2 = 0$, and any nontrivial solution $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$ of this equation is a nonzero multiple of $\begin{bmatrix} 1 & -1 \end{bmatrix}^T$. Hence, to within a constant multiple, the only eigenvector associated with $\lambda_1 = -2$ is $\mathbf{x}_1 = \begin{bmatrix} 1 & -1 \end{bmatrix}^T$.

Case 2: $\lambda_1 = 3$. With $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$, the system $(A - \lambda I)\mathbf{x} = 0$ is

$$\begin{bmatrix} 2 & 7 \\ -2 & -7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Again, we have only a single equation, $2x_1 + 7x_2 = 0$, and any nontrivial solution of this equation will suffice. The choice $x_2 = -2$ yields $x_1 = 7$ so (to within a constant multiple) the only eigenvector associated with $\lambda_2 = 3$ is $\mathbf{x}_2 = \begin{bmatrix} 7 & -2 \end{bmatrix}^T$.

Example 2.4.2. Find the eigenvalues and eigenvectors of the 2×2 matrix

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

Solution: Given A, $|A - \lambda I| = \lambda^2 + \lambda - 6 \stackrel{\text{set}}{=} 0$. It follows that $\lambda_1 = -3$ and $\lambda_2 = 2$. We can then set up the equation $A\mathbf{x}_i = \lambda_i \mathbf{x}_i$ or $(A - \lambda_i I)\mathbf{x}_i = 0$ for $i \in \{1, 2\}$. Specifically, we have

$$\begin{bmatrix} 2 & 3 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -3 & 3 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which results in

$$\mathbf{x}_1 = \begin{bmatrix} 3 \\ -2 \end{bmatrix} \quad \text{and} \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

or a multiple thereof.

Theorem 2.4.2. *Let A be a $k \times k$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_k$. Then,*

(a) $\lambda_1 + \lambda_2 + ... + \lambda_k = tr(A)$ *and*

(b) $\lambda_1 \cdot \lambda_2 \cdots \lambda_k = \det(A)$.

Proof. Consider the case for 2×2 matrices. Assume $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, then $|A - \lambda I| = \lambda^2 - (a_{11} + a_{22})\lambda + (a_{11}a_{22} - a_{12}a_{21})$, so that the characteristic equation is

$$\begin{aligned} p(\lambda) &= \lambda^2 - (a_{11} + a_{22})\lambda + (a_{11}a_{22} - a_{12}a_{21}) \\ &= \lambda^2 - tr(A)\lambda + \det(A). \end{aligned}$$

If λ_1 and λ_2 are the roots of this polynomial, we can rewrite the solution as

$$\begin{aligned} p(\lambda) &= \beta(\lambda_1 - \lambda)(\lambda_2 - \lambda) \\ &= \beta\lambda^2 - \beta(\lambda_1 + \lambda_2)\lambda + \beta\lambda_1\lambda_2, \end{aligned}$$

for some constant β. Comparing the two expressions, we find that $\beta = 1$ and

$$\begin{aligned} tr(A) &= \beta(\lambda_1 + \lambda_2), \\ \det(A) &= \beta\lambda_1\lambda_2. \end{aligned}$$

The theorem naturally extends to higher dimensional cases. □

2.4.1 Properties of Eigenvalues and Eigenvectors

(a) Eigenvalues and eigenvectors are defined only for square matrices.

(b) A square matrix A is singular if and only if 0 is an eigenvalue. However, the zero vector cannot be an eigenvector.

(c) A matrix is invertible if and only if none of its eigenvalues is equal to zero.

(d) Every eigenvalue has an infinite number of eigenvectors associated with it, as any nonzero scalar multiple of an eigenvector is also an eigenvector.

(e) The entries of a diagonal matrix D are eigenvalues of D.

(f) The eigenvalues of A and A^T are the same (as their characteristic polynomials are the same), but there is no simple relationship between their eigenvectors.

(g) The eigenvalues of a shifted matrix $A - \alpha I$ are $\lambda - \alpha$ and the eigenvectors are the same as those of A since

$$A\mathbf{x} = \lambda\mathbf{x} \Rightarrow (A - \alpha I)\mathbf{x} = (\lambda - \alpha)\mathbf{x}.$$

(h) The eigenvalues of A^{-1} are $1/\lambda$ and the eigenvectors are the same as those of A since

$$A\mathbf{x} = \lambda\mathbf{x} \Rightarrow A^{-1}\mathbf{x} = \lambda^{-1}\mathbf{x}.$$

(i) The eigenvalues of A^2 are λ^2, and the eigenvectors are the same as those of A since

$$A\mathbf{x} = \lambda\mathbf{x} \Rightarrow A^2\mathbf{x} = A(\lambda\mathbf{x}) = \lambda(A\mathbf{x}) = \lambda^2\mathbf{x}.$$

This property naturally generalizes to high-order powers (i.e. the eigenvalues of A^k are λ^k).

(j) Let $\mathbf{x}$ be the eigenvector of an $n \times n$ matrix A with eigenvalue λ. Then, the eigenvalue of

$$\alpha_k A^k + \alpha_{k-1}A^{k-1} + \cdots + \alpha_1 A + \alpha_0 I$$

associated with eigenvector $\mathbf{x}$ is

$$\alpha_k \lambda^k + \alpha_{k-1}\lambda^{k-1} + \cdots + \alpha_1\lambda + \alpha_0,$$

where $\alpha_k, \alpha_{k-1}, \ldots, \alpha_1, \alpha_0$ are real numbers and k is a positive integer.

2.4.2 Definiteness and Eigenvalues

Definition 2.4.2. *Let A be an $n \times n$ symmetric matrix. Then,*

(a) A is positive definite if and only if all of its eigenvalues are strictly positive ($\lambda_i > 0\ \forall i$)

(b) A is negative definite if and only if all of its eigenvalues are strictly negative ($\lambda_i < 0\ \forall i$)

(c) A is positive semi-definite if and only if all of its eigenvalues are nonnegative ($\lambda_i \geq 0\ \forall i$)

(d) A is negative semi-definite if and only if all of its eigenvalues are nonpositive ($\lambda_i \leq 0\ \forall i$)

Example 2.4.3. If $A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$, then $\Delta_1 = 1 > 0$, $\Delta_2 = -3 < 0$, which implies that A is indefinite. Or, using eigenvalues, we can solve

$$\det\left(\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} - \lambda I\right) = \lambda^2 - 2\lambda - 3 = (\lambda - 3)(\lambda + 1) = 0$$

which implies $\lambda_1 = -1$ and $\lambda_2 = 3$, and hence indefiniteness.

Example 2.4.4. If $A = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$, then $\Delta_1 = -1 < 0, \Delta_2 = 0 \leq 0$ and also $B_1^{(2)} = -1 < 0$. Thus, A is negative semi-definite. Alternatively, using eigenvalues

$$\det\left(\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} - \lambda I\right) = \lambda(2 + \lambda) = 0,$$

which implies $\lambda_1 = 0$ and $\lambda_2 = -2$ and fulfills the condition for negative semi-definiteness.

Example 2.4.5. If $A = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 3 & 0 \\ -1 & 0 & 4 \end{bmatrix}$, then $\Delta_1 = 1 > 0, \Delta_2 = \begin{vmatrix} 1 & 0 \\ 0 & 3 \end{vmatrix} = 3 > 0$ and $\Delta_3 = 9 > 0$, which implies A is positive definite. Alternatively, using eigenvalues,

$$\det\left(\begin{bmatrix} 1-\lambda & 0 & -1 \\ 0 & 3-\lambda & 0 \\ -1 & 0 & 4-\lambda \end{bmatrix}\right) = (3 - \lambda)(\lambda^2 - 5\lambda + 3) = 0,$$

which implies $\lambda_{1,2} = (5 \pm \sqrt{13})/2$ and fulfills the condition for positive definiteness.

Chapter 3

Advanced Topics in Linear Algebra

In this chapter, we will look at linear algebra more from a transformational viewpoint than system of equations viewpoint (e.g., function from one vector space to another).

3.1 Vector Spaces and Subspaces

Definition 3.1.1 (Vector Space)**.** *A vector space (or a linear space) V over a field F consists of a set on which two operations (called addition and scalar multiplication) are defined so that for each pair of elements $\mathbf{x}$, $\mathbf{y}$ in V there is a unique element $\mathbf{x}+\mathbf{y}$ in V, and for each element a in F and each element $\mathbf{x}$ in V there is a unique element $a\mathbf{x}$ in V, such that the following conditions hold:*

(VS 1) For all $\mathbf{x}, \mathbf{y}$ in V, $\mathbf{x}+\mathbf{y}=\mathbf{y}+\mathbf{z}$ (commutativity of addition).

(VS 2) For all $\mathbf{x}, \mathbf{y}, \mathbf{z}$ in V, $(\mathbf{x}+\mathbf{y})+\mathbf{z}=\mathbf{x}+(\mathbf{y}+\mathbf{z})$ (associativity of addition).

(VS 3) There exists an element in V denoted by $\mathbf{0}$ such that $\mathbf{x}+\mathbf{0}=\mathbf{x}$ for each $\mathbf{x}$ in V.

(VS 4) For each element $\mathbf{x}$ in V there exists an element $\mathbf{y}$ in V such that $\mathbf{x}+\mathbf{y}=\mathbf{0}$.

(VS 5) For each element $\mathbf{x}$ in V, $1\mathbf{x}=\mathbf{x}$.

(VS 6) For each pair of elements a, b in F and each element $\mathbf{x}$ in V, $(ab)\mathbf{x}=a(b\mathbf{x})$.

(VS 7) For each element a in F and each pair of elements $\mathbf{x}, \mathbf{y}$ in V, $a(\mathbf{x}+\mathbf{y})=a\mathbf{x}+ay$.

(VS 8) For each pair of elements a, b in F and each element $\mathbf{x}$ in V, $(a+b)\mathbf{x}=a\mathbf{x}+b\mathbf{x}$

The elements of the field F are called *scalars* and the elements of the vector space V are called *vectors*. The following are examples of vector spaces:

1. An object of the form $(a_1, a_2, \ldots, a_n)$, where the entries a_i, $i = \{1, \ldots, n\}$ are elements from a field F, is called an *n-tuple* with entries from F. The set of all n-tuples with entries from F is a vector space, denoted F^n, under the operations of coordinate-wise addition and scalar multiplication; that is, if $\mathbf{u} = (a_1, a_2, \ldots, a_n) \in F^n$, $\mathbf{v} = (b_1, b_2, \ldots, b_n) \in F^n$ and $c \in F$, then

$$\mathbf{u}+\mathbf{v} = (a_1+b_1, a_2+b_2, \ldots, a_n+b_n) \quad \text{and} \quad cu = (ca_1, ca_2, \ldots, ca_n).$$

2. The set of all $m \times n$ matrices with entries from a field F is a vector space, denoted $M_{m\times n}(F)$, under the following operations of addition and scalar multiplication: For $A, B \in M_{m\times n}(F)$ and $c \in F$

$$(A + B)_{ij} = A_{ij} + B_{ij} \quad \text{and} \quad (cA)_{ij} = cA_{ij}$$

for $1 \leq i \leq m$ and $1 \leq j \leq n$.

3. Let S be any nonempty set and F be any field, and let $\mathcal{F}(S, F)$ denote the set of functions from S to F. The set $\mathcal{F}(S, F)$ is a vector space under the operations of addition and scalar multiplication defined for $f, g \in \mathcal{F}(S, F)$ and $c \in F$ by

$$(f + g)(s) = f(s) + g(s) \quad \text{and} \quad (cf)(s) = c[f(s)]$$

for each $s \in S$. Note that these are the familiar operations of addition and scalar multiplication for the functions used in algebra and calculus.

4. Let $S = \{(a_1, a_2) : a_1, a_2 \in \mathbb{R}\}$. For $(a_1, a_2), (b_1, b_2) \in S$ and $c \in \mathbb{R}$, define

$$(a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 - b_2) \quad \text{and} \quad c(a_1, a_2) = (ca_1, ca_2).$$

Since (VS 1), (VS 2), and (VS 8) fail to hold, S is not a vector space under these operations.

Theorem 3.1.1 (Cancelation Law for Vector Addition)**.** *If* $\mathbf{x}$*,* $\mathbf{y}$*, and* $\mathbf{z}$ *are elements of a vector space* V *such that* $\mathbf{x} + \mathbf{z} = \mathbf{y} + \mathbf{z}$*, then* $\mathbf{x} = \mathbf{y}$.

Proof. There exists an element $\mathbf{v}$ in V such that $\mathbf{z} + \mathbf{v} = \mathbf{0}$ by (VS 4). Thus,

$$\begin{aligned}\mathbf{x} &= \mathbf{x} + \mathbf{0} = \mathbf{x} + (\mathbf{z} + \mathbf{v}) = (\mathbf{x} + \mathbf{z}) + \mathbf{v} \\ &= (\mathbf{y} + \mathbf{z}) + \mathbf{v} = \mathbf{y} + (\mathbf{z} + \mathbf{v}) \\ &= \mathbf{y} + \mathbf{0} = \mathbf{y}\end{aligned}$$

by (VS 2) and (VS 3). □

Corollary 3.1.1. *V has exactly one zero vector*

Proof. By (VS 3), there exists $\mathbf{0} \in V$ such that $\mathbf{x} + \mathbf{0} = \mathbf{x}$, for all $\mathbf{x} \in V$. Now suppose that $\mathbf{x} + \mathbf{z} = \mathbf{x}$, for all $\mathbf{x} \in V$. Then $\mathbf{x} + \mathbf{0} = \mathbf{x} + \mathbf{z}$, which implies $\mathbf{0} = \mathbf{z}$ by Theorem 3.1.1. □

Corollary 3.1.2. *For each* $\mathbf{x} \in V$*, there is exactly one vector* $\mathbf{y}$ *so that* $\mathbf{x} + \mathbf{y} = \mathbf{0}$.

Proof. Suppose $\mathbf{x} + \mathbf{z} = \mathbf{0} = \mathbf{x} + \mathbf{y}$. Then $\mathbf{z} = \mathbf{y}$ by Theorem 3.1.1. This means there is only one additive inverse to $\mathbf{x}$. Thus, we say $\mathbf{y} = -\mathbf{x}$. □

Theorem 3.1.2. *In any vector space V the following statements are true:*

(a) $0\mathbf{x} = \mathbf{0}$ *for each* $\mathbf{x} \in V$.

(b) $(-a)\mathbf{x} = -(a\mathbf{x}) = a(-\mathbf{x})$ *for each* $a \in F$ *and each* $\mathbf{x} \in V$.

(c) $a\mathbf{0} = \mathbf{0}$ *for each* $a \in F$.

Theorem 3.1.3 (Subspace)**.** *Let V be a vector space and W a subset of V. Then W is a subspace of V if and only if the following three conditions hold for the operations defined in V:*

(a) $\mathbf{0} \in W$.

(b) $\mathbf{x} + \mathbf{y} \in W$ *whenever* $x \in W$ *and* $y \in W$.

(c) $c\mathbf{x} \in W$ *whenever* $c \in F$ *and* $\mathbf{c} \in W$.

This theorem provides a simple method for determining whether or not a given subset of a vector space is a subspace. The following are examples of subspaces:

1. An $n \times n$ matrix M is called a *diagonal matrix* if $M_{ij} = 0$ whenever $i \neq j$. Define $V = M_{n\times n}(\mathbb{R})$ and $W = \{A \in M_{n\times n} : A_{ij} = 0 \text{ whenever } i \neq j\}$. Clearly, the zero matrix is a diagonal matrix because all of its entries are 0 and hence belongs to W. Moreover, if A and B are diagonal $n \times n$ matrices, then whenever $i \neq j$

$$(A + B)_{ij} = A_{ij} + B_{ij} = 0 + 0 = 0 \quad \text{and} \quad (cA)_{ij} = cA_{ij} = c0 = 0$$

for any scalar c. Hence $A + B$ and cA are diagonal matrices for any scalar c and also belong to W. Therefore the set of diagonal matrices, W, is a subspace of $V = M_{n\times n}(F)$.

2. The *transpose* A^t of an $m \times n$ matrix A is the $n \times m$ matrix obtained from A by interchanging the rows with the columns; that is, $(A^t)_{ij} = A_{ji}$. A *symmetric matrix* is a matrix A such that $A^t = A$. Define $V = M_{n\times n}(\mathbb{R})$ and $W = \{A \in M_{n\times n} : A^t = A\}$. The zero matrix is equal to its transpose and hence belongs to W. Let $A, B \in W$ and $c \in \mathbb{R}$, then

$$(cA + B)^t = (cA)^t + B^t = cA^t + B^t = cA + B.$$

Thus, $cA + B \in W$ and W is a subspace of V.

3. Let $A \in M_{n\times n}(\mathbb{R})$. Define $W = \{\mathbf{x} \in M_{n\times 1}(\mathbb{R}) : A\mathbf{x} = \mathbf{0}\} \subseteq M_{n\times 1}(\mathbb{R}) = V$. Since $A\mathbf{0} = \mathbf{0}$, $\mathbf{0} \in W$. Also, for any $\mathbf{x}_1, \mathbf{x}_2 \in W$ and $c \in \mathbb{R}$

$$A(c\mathbf{x}_1 + \mathbf{x}_2) = cA\mathbf{x}_1 + A\mathbf{x}_2 = \mathbf{0}.$$

Thus, $c\mathbf{x}_1 + \mathbf{x}_2 \in W$ and W is a subspace of V. This W is referred to as the *null space* of A.

4. The set of matrices in $M_{n\times n}(\mathbb{R})$ having nonnegative entries is not a subspace of $M_{n\times n}(\mathbb{R})$ because it is not closed under scalar multiplication.

Theorem 3.1.4. *Any intersection of subspaces of a vector space V is a subspace of V*

Proof. Let $\mathcal{C}$ be a collection of subspaces of V, and let $W = \bigcap U$ denote the intersection of the subspaces in $\mathcal{C}$. Since every subspace contains the zero vector, $\mathbf{0} \in W$. Let $a \in F$ and $\mathbf{x}, \mathbf{y} \in W$. Then $\mathbf{x}, \mathbf{y} \in U$ for all $U \in \mathcal{C}$. Thus, $c\mathbf{x} + \mathbf{y} \in U$ for all $U \in \mathcal{C}$. Hence, $a\mathbf{x} + \mathbf{y} \in W$ and W is a subspace of V. □

3.2 Linear Combinations and Spanning Conditions

Definition 3.2.1 (Linear Combination)**.** *Let V be a vector space and S a nonempty subset of V. A vector $\mathbf{v} \in V$ is called a linear combination of elements of S if there exist a finite number of elements $u_1, u_2, \ldots, u_n$ in S and scalars $a_1, a_2, \ldots, a_n$ in F such that $v = a_1u_1 + a_2u_2 + \cdots + a_nu_n$.*

Example 3.2.1. Claim: $2x^3 - 2x^2 + 12x - 6$ is a linear combination of

$$x^3 - 2x^2 - 5x - 3 \quad \text{and} \quad 3x^3 - 5x^2 - 4x - 9$$

in $P_3(\mathbb{R})$. To show this, find scalars a and b such that

$$\begin{aligned} 2x^3 - 2x^2 + 12x - 6 &= a(x^3 - 2x^2 - 5x - 3) + b(3x^3 - 5x^2 - 4x - 9) \\ &= (a + 3b)x^3 + (-2a - 5b)x^2 + (-5a - 4b)x + (3a - 9b). \end{aligned}$$

Equating coefficients

$$\begin{bmatrix} 1 & 3 & 2 \\ -2 & -5 & -2 \\ -5 & -4 & 12 \\ -3 & -9 & -6 \end{bmatrix} \overset{\text{rref}}{\sim} \begin{bmatrix} 1 & 0 & -4 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Thus $a = -4$ and $b = 2$, which proves that $2x^3 - 2x^2 + 12x - 6$ is a linear combination of $x^3 - 2x^2 - 5x - 3$ and $3x^3 - 5x^2 - 4x - 9$.

Definition 3.2.2 (Span)**.** *Let S be a nonempty subset of a vector space V. The span of S, denoted* $\operatorname{span}(S)$*, is the set consisting of all linear combinations of the elements of S. For convenience we define* $\operatorname{span}(\varnothing) = \{\mathbf{0}\}$.

In $\mathbb{R}^3$, for instance, the span of the set $\{(1,0,0),(0,1,0)\}$ consists of all vectors in $\mathbb{R}^3$ that have the form $a(1,0,0) + b(0,1,0) = (a,b,0)$ for some scalars a and b. Thus the span of $\{(1,0,0),(0,1,0)\}$ contains all the points in the xy-plane. In this case, the span of the set is a subspace of $\mathbb{R}^3$. This fact is true in general.

Theorem 3.2.1. *The span of any subset S of a vector space V is a subspace of V. Moreover, any subspace of V that contains S must also contain the span of S.*

Proof. This result is immediate if $S = \varnothing$ because $\operatorname{span}(\varnothing) = \{\mathbf{0}\}$, which is a subspace that is contained in any subspace of V.

If $S \neq \varnothing$, then S contains an element $\mathbf{z}$ and $0\mathbf{z} = \mathbf{0}$ is an element of $\operatorname{span}(S)$. Let $\mathbf{x}, \mathbf{y} \in \operatorname{span}(S)$. Then there exists elements $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m, \mathbf{v}_1, \mathbf{v}_2 \ldots, \mathbf{v}_n$ in S and scalars $a_1, a_2, \ldots, a_m$ and $b_1, b_2, \ldots, b_n$ such that

$$\mathbf{x} = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \cdots + a_m\mathbf{u}_m \quad \text{and} \quad \mathbf{y} = b_1\mathbf{v}_1 + b_2\mathbf{v}_2 + \cdots + b_m\mathbf{v}_m.$$

Then

$$\mathbf{x} + \mathbf{y} = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \cdots + a_m\mathbf{u}_m + b_1\mathbf{v}_1 + b_2\mathbf{v}_2 + \cdots + b_m\mathbf{v}_m$$

and, for any scalar c,

$$c\mathbf{x} = (ca_1)\mathbf{u}_1 + (ca_2)\mathbf{u}_2 + \cdots + (ca_m)\mathbf{u}_m$$

are clearly linear combinations of the elements of S; so $\mathbf{x} + \mathbf{y}$ and $c\mathbf{x}$ are elements of $\operatorname{span}(S)$. Thus, $\operatorname{span}(S)$ is a subspace of V.

Now let W denote any subspace of V that contains S. If $\mathbf{w} \in \operatorname{span}(S)$, then $\mathbf{w}$ has the form $\mathbf{w} = c_1\mathbf{w}_1 + c_2\mathbf{w}_2 + \cdots + c_k\mathbf{w}_k$ for some elements $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k \in W$ and some scalars $c_1, c_2, \ldots, c_k$. Since $S \subseteq W$, we have $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k \in W$ in W. Therefore $\mathbf{w} = c_1\mathbf{w}_1 + c_2\mathbf{w}_2 + \cdots + c_k\mathbf{w}_k$ is an element of W (since W is a subspace of V, it is closed under addition and scalar multiplication). Since $\mathbf{w}$, an arbitrary element of $\operatorname{span}(S)$, belongs to W, it follows that $\operatorname{span}(S) \subseteq W$. □

Example 3.2.2. Suppose $V = \mathbb{R}^3$ and $S = \{(0,-2,2),(1,3,-1)\}$. Is $(3,1,5) \in \text{span}(S)$. To answer this question, try to find constants a and b so that

$$a(0,-2,2) + b(1,3,-1) = (3,1,5).$$

Equating coefficients

$$\begin{bmatrix} 0 & 1 & 3 \\ -2 & 3 & 1 \\ 2 & -1 & 5 \end{bmatrix} \overset{\text{rref}}{\sim} \begin{bmatrix} 1 & 0 & 4 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix}$$

Thus $a = 4$ and $b = 3$, which proves that $(3,1,5) \in \text{span}(S)$.

Definition 3.2.3. *A subset S of a vector space V generates (or spans) V if* $\text{span}(S) = V$. *In this situation we may also say that the elements of S generate (or span) V.*

Example 3.2.3.

1. Let $V = \mathbb{R}^n$ and $S = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$, where $\mathbf{e}_j$ denotes a vector whose jth coordinate is 1 and whose other coordinates are 0. Since $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ can be written

$$\begin{aligned} \mathbf{x} &= x_1(1,0,0,\ldots,0) + x_2(0,1,0,\ldots,0) + \cdots + x_n(0,0,\ldots,0,1) \\ &= \sum_{k=1}^{n} x_k e_k \in \text{span}(S). \end{aligned}$$

   Hence $S = \mathbb{R}^n$ and S generates $\mathbb{R}^n$.

2. Let $V = P_n(\mathbb{R}) = \{a_0 + a_1 x + \cdots + a_n x^n : \textit{ each } a_k \in \mathbb{R}\}$ and $S = \{1, x, x^2, \ldots, x^n\}$. Clearly $\text{span}(S) \in P_n(\mathbb{R})$. Also, for any $p(x) \in P_n(\mathbb{R})$

$$\begin{aligned} p(x) &= a_0 + a_1 x + \cdots + a_n x^n \\ &= a_0(1) + a_1(x) + \cdots + a_n(x^n) \in \text{span}(S). \end{aligned}$$

   Hence $S = P_n(\mathbb{R})$ and S generates $P_n(\mathbb{R})$.

3. Let $V = M_{2\times 2}(\mathbb{R})$ and

$$S = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\}.$$

   Since

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + a_{12} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + a_{21} \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + a_{22} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \in \text{span}(S)$$

   and $A \in M_{2\times 2}(\mathbb{R})$, S generates $M_{2\times 2}(\mathbb{R})$.

4. Let $V = \mathbb{R}^3$ and $S = \{(1,1,0),(1,0,1),(0,1,1)\}$. Let $\mathbf{x} \in \text{span}(S)$. Then for $a, b, c \in \mathbb{R}$

$$\mathbf{x} = a(1,1,0) + b(1,0,1) + c(0,1,1) = (a+b, a+c, b+c) \in \mathbb{R}^3.$$

   Thus, $\text{span}(S) \subseteq \mathbb{R}^3$. Now let $\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$, then we must find $a, b, c \in \mathbb{R}$ satisfying

$$a(1,1,0) + b(1,0,1) + c(0,1,1) = (x_1, x_2, x_3).$$

Equating coefficients

$$\begin{bmatrix} 1 & 1 & 0 & x_1 \\ 1 & 0 & 1 & x_2 \\ 0 & 1 & 1 & x_3 \end{bmatrix} \overset{\text{rref}}{\sim} \begin{bmatrix} 1 & 0 & 0 & (x_1+x_2-x_3)/2 \\ 0 & 1 & 0 & (x_1-x_2+x_3)/2 \\ 0 & 0 & 1 & (x_2+x_1+x_3)/2 \end{bmatrix},$$

which is consistent. Hence $\mathbf{x} \in \text{span}(S)$, $\mathbb{R}^3 \subseteq \text{span}(S)$, and S generates $\mathbb{R}^3$.

3.3 Linear Independence and Linear Dependence

Definition 3.3.1 (Linear Dependence)**.** *A subset S of a vector space V is called linearly dependent if there exist a finite number of distinct vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ in S and scalars $a_1, a_2, \ldots, a_n$, not all zero, such that*

$$a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \cdots + a_n\mathbf{u}_n = \mathbf{0}.$$

In this case we say that the elements of S are linearly dependent.

For any vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ we have $a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \cdots + a_n\mathbf{u}_n = \mathbf{0}$ if $a_1 = a_2 = \cdots = a_n = 0$. We call this the *trivial representation* of $\mathbf{0}$ as a linear combination of $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$. Thus for a set to be linearly dependent means that there is a nontrivial representation of $\mathbf{0}$ as a linear combination of vectors in the set. Consequently, any subset of a vector space that contains the zero vector is linearly dependent, because $\mathbf{0} = 1 \cdot \mathbf{0}$ is a nontrivial representation of $\mathbf{0}$ as a linear combination of vectors in the set.

Definition 3.3.2 (Linear Independence)**.** *A subset S of a vector space that is not linearly dependent is called linearly independent. As before we also say that the elements of S are linearly independent.*

The following facts about linearly independent sets are true in any vector space.

1. The empty set is linearly independent, for linearly dependent sets must be nonempty.
2. A set consisting of a single nonzero vector is linearly independent. For if $\{\mathbf{u}\}$ is linearly dependent, the $a\mathbf{u} = \mathbf{0}$ for some nonzero scalar a. Thus
$$\mathbf{u} = a^{-1}(a\mathbf{u}) = a^{-1}\mathbf{0} = \mathbf{0}.$$
3. A set is linearly independent if and only if the only representations of $\mathbf{0}$ as linear combinations of its elements are trivial representations.

The condition in 3 provides a useful method for determining if a finite set is linearly independent. This technique is illustrated in the following example.

Example 3.3.1. Determine whether the following sets are linearly dependent or linearly independent.

1. In $P_2(\mathbb{R})$, let $S = \{3 + x + x^2, 2 - x + 5x^2, 4 - 3x^2\}$. Consider the equation:
$$a(3 + x + x^2) + b(2 - x + 5x^2) + c(4 - 3x^2) = (3a + 2b + 4c) + (a - b)x + (a + 5b - 3c)x^2 = 0$$
for all x. Equating coefficients
$$\begin{bmatrix} 3 & 2 & 4 \\ 1 & -1 & 0 \\ 1 & 5 & -3 \end{bmatrix} \overset{\text{rref}}{\sim} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$
Thus, the only solution is $a = b = c = 0$, which implies that S is linearly independent in $P_2(\mathbb{R})$.

2. In $M_{2\times 2}(\mathbb{R})$, let

$$S = \left\{ \begin{pmatrix} 1 & -3 \\ -2 & 4 \end{pmatrix}, \begin{pmatrix} -2 & 6 \\ 4 & 8 \end{pmatrix} \right\}.$$

Consider the equation:

$$a \begin{pmatrix} 1 & -3 \\ -2 & 4 \end{pmatrix} + b \begin{pmatrix} -2 & 6 \\ 4 & 8 \end{pmatrix} = \begin{pmatrix} a_1 - 2a_2 & -3a_1 + 6a_2 \\ -2a_1 + 4a_2 & 4a_1 - 8a_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Equating coefficients

$$\begin{bmatrix} 1 & -2 \\ -3 & 6 \\ -2 & 4 \\ 4 & -8 \end{bmatrix} \overset{\text{rref}}{\sim} \begin{bmatrix} 1 & -2 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

which implies that $a = 2b$. Therefore, there are infinitely many nontrivial solutions and so S is linearly dependent in $M_{2\times 2}(\mathbb{R})$.

3.4 Bases and Dimension

Definition 3.4.1 (Basis)**.** *A basis β for a vector space V is a linearly independent subset of V that generates V (i.e. β is linearly independent and $\operatorname{span}(\beta) = V$).*

Theorem 3.4.1. *let V be a vector space and $\beta = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ be a subset of V. Then β is a basis for V if and only if each vector $\mathbf{v}$ can be uniquely expressed as a linear combination of vectors in β, that is, can be expressed in the form*

$$\mathbf{v} = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \cdots + a_n\mathbf{u}_n$$

for unique scalars $a_1, a_2, \ldots, a_n$.

Proof. First let β be a basis for V. If $\mathbf{v} \in V$, then $\mathbf{v} \in \operatorname{span}(\beta)$ because $\operatorname{span}(\beta) = V$. Thus $\mathbf{v}$ is a linear combination of the elements in β. Suppose that

$$\mathbf{v} = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \cdots + a_n\mathbf{u}_n \quad \text{and} \quad \mathbf{v} = b_1\mathbf{u}_1 + b_2\mathbf{u}_2 + \cdots + b_n\mathbf{u}_n$$

are two such representations of $\mathbf{v}$. Subtracting the second equation from the first gives

$$(a_1 - b_1)\mathbf{u}_1 + (a_2 - b_2)\mathbf{u}_2 + \cdots + (a_n - b_n)\mathbf{u}_n.$$

Since β is linearly independent, it follows that $a_1 - b_1 = a_2 - b_2 = \cdots = a_n - b_n = 0$. Hence $a_1 = b_1, a_2 = b_2, \ldots, a_n = b_n$ and so $\mathbf{v}$ is uniquely expressible as a linear combination of the elements in β.

Now let every $\mathbf{v} \in V$ can be uniquely expressed as a linear combination of vectors in β. Note first that β is trivially a generating set for V, since we are told each $\mathbf{v} \in V$ is also in $\operatorname{span}(\beta)$. Since $\mathbf{v} = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \cdots + a_n\mathbf{u}_n = b_1\mathbf{u}_1 + b_2\mathbf{u}_2 + \cdots + b_n\mathbf{u}_n$, $a_i = b_i$ for each i such that $1 \leq i \leq n$, otherwise each $\mathbf{v} \in V$ would not be uniquely expressible as a linear combination of vectors from β. This implies that $a_i - b_i = 0$ for each i, proving that β is linearly independent. β is therefore a basis for V. ∎

Example 3.4.1.

1. Since $\text{span}(\varnothing) = \{0\}$ and $\varnothing$ is linearly independent, $\varnothing$ is a basis for the vector space $\{0\}$.
2. In F^n, $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$, where $\mathbf{e}_j$ denotes a vector whose jth coordinate is 1 and whose other coordinates are 0 is a basis for F^n and is called the *standard basis*.
3. In $M_{m\times n}(F)$, let M^{ij} denote the matrix whose only nonzero entry is a 1 in the ith row and jth column. Then $\{M^{ij} : 1 \leq i \leq, 1 \leq j \leq n\}$ is a basis for $M_{m\times n}(F)$.
4. In $P_n(F)$ the set $\{1, x, x^2, \ldots, x^n\}$ is a basis. We call this basis the *standard basis* for $P_n(F)$.
5. In $P(F)$ the set $\{1, x, x^2, \ldots\}$ is a basis.

Theorem 3.4.2 (Extraction)**.** *If a vector space V is generated by a finite set S, then some subset of S is a basis for V. Hence V has a finite basis.*

Proof. If $S = \varnothing$ or $S = \{\mathbf{0}\}$, then $V = \{\mathbf{0}\}$ and $\varnothing$ is a subset of S that is a basis for V. Otherwise suppose there exists $\mathbf{u}_1 \in S$ and $\mathbf{u}_1 \neq \mathbf{0}$. Define $S_1 = \{\mathbf{u}_1\}$. Then $S_1 \subseteq S$ and S_1 is independent. If $\text{span}(S_1) = V$, we are done. If not, then there exists a $\mathbf{u}_2 \in S$ where $\mathbf{u}_2 \notin \text{span}(S_1)$. Define $S_2 = S_1 \cup \{\mathbf{u}_2\}$. Then S_2 is independent. If $\text{span}(S_2) = V$, we are done. If not,..., then there a $\mathbf{u}_m \in S$ where $\mathbf{u}_m \notin \text{span}(S_{m-1})$. Define $S_m = S_{m-1} \cup \{\mathbf{u}_m\}$. Then S_m is independent. If $k > m$ and $\mathbf{u}_k \in S$, then $\mathbf{u} \in \text{span}(S_m)$. Thus $\text{span}(S_m) = V$ and $S_m \subseteq S$ is a basis for V. □

Example 3.4.2. Find a basis for the following sets:

1. Let $V = \mathbb{R}^3$ and $S = \{(2, -1, 4), (1, -1, 3), (1, 1, -1), (1, -2, 1)\}$. Let $\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$, then we must find $a, b, c, d \in \mathbb{R}$ satisfying

$$a(2, -1, 4) + b(1, -1, 3) + c(1, 1, -1) + d(1, -1, 1) = (x_1, x_2, x_3).$$

Equating coefficients

$$\begin{bmatrix} 2 & 1 & 1 & 1 \\ -1 & -1 & 1 & -2 \\ 4 & 3 & -1 & 1 \end{bmatrix} \overset{\text{rref}}{\sim} \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & -3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

which is consistent for all $\mathbf{x}$. Hence $\mathbf{x} \in \text{span}(S)$ and S generates $\mathbb{R}^3$. By Theorem 3.4.2, $S \setminus \{(1, 1, -1)\}$ is a basis for $\mathbb{R}^3$.

2. Let $V = M_{2\times 2}(\mathbb{R})$ and

$$S = \left\{ \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \begin{bmatrix} 1 & 5 \\ 5 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 3 \\ 3 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix} \right\}.$$

Form the coefficient matrix

$$\begin{bmatrix} 1 & 2 & 1 & 2 & 1 \\ 2 & 1 & 5 & 3 & 2 \\ 2 & 1 & 5 & 3 & 3 \\ 1 & 2 & 1 & 1 & 2 \end{bmatrix} \overset{\text{rref}}{\sim} \begin{bmatrix} 1 & 0 & 3 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

which is consistent. Hence $\text{span}(S) = M_{2\times 2}(\mathbb{R})$. By Theorem 3.4.2, $S \setminus \{S_3\}$ is a basis for $M_{2\times 2}(\mathbb{R})$.

3. Let $V = P_2(\mathbb{R})$ and $S = \{1+x, 1-x, 1+x+x^2, 1-x+x^2, x+x^2\}$. Form the coefficient matrix

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & -1 & 1 & -1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \overset{\text{rref}}{\sim} \begin{bmatrix} 1 & 0 & 0 & -1 & -1/2 \\ 0 & 1 & 0 & 1 & -1/2 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix},$$

which is consistent. Hence $\text{span}(S) = P_2(\mathbb{R})$. By Theorem 3.4.2, $S \setminus \{1-x+x^2, x+x^2\}$ is a basis for $P_2(\mathbb{R})$.

Theorem 3.4.3. *Let $\beta = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be a basis for V and suppose $S = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\} \subseteq V$. If $m = |S| > n = |\beta|$, then S is dependent.*[1].

Proof. For each $\mathbf{u}_j$, $1 \leq j \leq m$, there exist unique scalars a_{ij}, $1 \leq i \leq n$ so that

$$\mathbf{u}_j = a_{1j}\mathbf{v}_1 + a_{2j}\mathbf{v}_2 + \cdots + a_{nj}\mathbf{v}_n = \sum_{i=1}^{n} a_{ij}\mathbf{v}_i$$

If we form $\sum_{j=1}^{m} x_j \mathbf{u}_j = \mathbf{0}$, then

$$\sum_{j=1}^{m} x_j \left(\sum_{i=1}^{n} a_{ij}\mathbf{v}_i \right) = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} a_{ij}x_j \right) \mathbf{v}_i = \mathbf{0}$$

Since β is independent $\sum_j = 1^m a_{ij}x_j = 0$ for each $1 \leq i \leq n$. Set

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \quad X = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}.$$

Then $AX = 0$. Since $m > n$, we will always have at least one free variable and so some $\mathbf{x}_j \neq 0$. Thus S is dependent. □

Definition 3.4.2. *A vector space is called finite-dimensional if it has a basis consisting of a finite number of elements. The unique number of elements in each basis for V is called the dimension of V and is denoted $\dim(V)$. A vector space that is not finite-dimensional is call infinite-dimensional.*

Example 3.4.3.

1. $\dim(F^n) = n$
2. $\dim(P_n) = n + 1$ (note the constant term)
3. $\dim(M_{m\times n}) = mn$
4. $\dim(\{\mathbf{0}\}) = 0$
5. $\dim(P) = \infty$

Corollary 3.4.1. *Let V be a vector space, $\dim(V) = n < \infty$, and $S \subseteq V$. If $|S| < n$, then $\text{span}(S) \neq V$.*

[1] For any set S, $|S|$ corresponds to the number of vectors in S.

Proof. Suppose $\text{span}(S) = V$. and $|s| < n$. By Theorem 3.4.2, there exists $\beta \subseteq S$ so that β is a basis for V. Then $n = |\beta| \leq |s| < n$, which is a contradiction. □

Remark 3.4.1. Let V be a vector space, $\dim(V) = n < \infty$, and $S \subseteq V$. Then the these results directly follow from earlier results:

1. If S is independent, then $|S| \leq \dim(V)$ (contrapositive of Theorem 3.4.3).
2. If S generates V ($\text{span}(S) = V$), then $|S| \geq \dim(V)$ (contrapositive of Corollary 3.4.1).

Corollary 3.4.2. *Let V be a vector space, $\dim(V) = n < \infty$, $S \subseteq V$, and $|S| = \dim(V)$. Then*

1. *If $\text{span}(S) = V$, then S is a basis for V.*
2. *If S is linearly independent, then S is a basis for V.*

Proof.

1. Since $\text{span}(S) = V$, by Theorem 3.4.2 there exists $\beta \subseteq S$ so β is a basis for V. Thus, $|\beta| = \dim(V) = |S|$ and so $\beta = S$.
2. On the contrary, suppose $\text{span}(S) \neq V$. Then there exists $\mathbf{v} \in V$ so $\mathbf{v} \notin \text{span}(S)$. Hence $S \cup \{\mathbf{v}\}$ is independent. This, however is a contradiction to Theorem 3.4.3, which says that if $|S \cup \{\mathbf{v}\}| > \dim(V)$, then $S \cup \{\mathbf{v}\}$ is dependent. Therefore $\text{span}(S) = V$ and so S is a basis.

□

Example 3.4.4. Do the polynomials $x^3 2x^2 + 1$, $4x^2 x + 3$, and $3x2$ generate $P_3(\mathbb{R})$? No, since $|S| < \dim(\mathbb{R}^3)$, $\text{span}(S) \neq P_3$ (Corollary 3.4.1).

Example 3.4.5. Determine whether $S = \{(1, 0, -1), (2, 5, 1), (0, -4, 3)\}$ form a basis for $\mathbb{R}^3$. Since $|S| = 3 = \dim(\mathbb{R}^3)$, it is sufficient to show that S is linearly independent (Corollary 3.4.2). Equating coefficients

$$\begin{bmatrix} 1 & 2 & 0 \\ 0 & 5 & -4 \\ -1 & 1 & 3 \end{bmatrix} \overset{\text{rref}}{\sim} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Hence S is linearly independent and forms a basis for $\mathbb{R}^3$.

Example 3.4.6. Find a basis for the following subspaces of F^5:

$$W = \{(a_1, a_2, a_3, a_4, a_5) \in F^5 : a_1 - a_3 - a_4 = 0\}.$$

What is the dimensions of W? $(a_1, a_2, a_3, a_4, a_5) \in W_1$ if and only if $(a_1, a_2, a_3, a_4, a_5) = (s, t, r, t, 2s)$ for some $r, s, t \in \mathbb{R}$. Thus, a spanning set for W_1 is

$$\beta = \{(0, 0, 1, 0, 0), (1, 0, 0, 0, 2), (0, 1, 0, 1, 0)\}.$$

Since

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 2 & 0 \end{bmatrix} \overset{\text{rref}}{\sim} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

the spanning set, β, is linearly independent and thus forms a basis for W. The $\dim(W) = 4$.

3.5 Linear Transformations

In the previous sections, we developed the theory of abstract vector spaces in detail. It is now natural to consider those functions defined on vector spaces that in some sense "preserve" the structure. These special functions are called linear transformations.

Definition 3.5.1 (Linear Transformation)**.** *Let V and W be vector spaces over F. We call a function $T : V \to W$ a Linear Transformation from V into W if for all $\mathbf{x}, \mathbf{y} \in V$ and $c \in F$ we have*

(a) $T(\mathbf{x} + \mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y})$

(b) $T(c\mathbf{x}) = cT(\mathbf{x})$

Theorem 3.5.1. *The following are basic facts about the function $T : V \to W$:*

1. *If T is linear, then $T(\mathbf{0}) = \mathbf{0}$.*
2. *T is linear if and only if $T(\mathbf{x} - \mathbf{y}) = T(\mathbf{x}) - T(\mathbf{y})$.*
3. *T is linear if and only if $T(c\mathbf{x} + \mathbf{y}) = cT(\mathbf{x}) + T(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in V$ and $c \in F$.*
4. *T is linear if and only if for $x_1, \ldots, x_n \in V$ and $a_1 \ldots, a_n \in F$ we have*

$$T\left(\sum_{i=1}^{n} a_i x_i\right) = \sum_{i=1}^{n} a_i T(x_i)$$

Proof. Let $T : V \to W$ be linear. Clearly $T(\mathbf{0}) = \mathbf{0}$, for otherwise by linearity $T(\mathbf{0}) = T(\mathbf{x} + (\mathbf{x})) = T(\mathbf{x}) + T(\mathbf{x}) = T(\mathbf{x}) + (T(\mathbf{x})) \neq 0$, which is absurd. Also note that since T is linear, $T(c\mathbf{x}+\mathbf{y}) = T(c\mathbf{x})+T(\mathbf{y}) = cT(\mathbf{x})+T(\mathbf{y})$, and $T(\mathbf{x}-\mathbf{y}) = T(\mathbf{x}+(-\mathbf{y})) = T(\mathbf{x})+T(-\mathbf{y}) = T(\mathbf{x})T(\mathbf{y})$. To prove property 4, note that if T is linear, an inductive argument can be used to show that for all $\mathbf{x}_1, \ldots, \mathbf{x}_n \in V$ and $a_1, \ldots, a_n \in F$, we have

$$T\left(\sum_{i=1}^{n} a_i x_i\right) = \sum_{i=1}^{n} T(a_i x_i) = \sum_{i=1}^{n} a_i T(x_i)$$

Now, assume $T(c\mathbf{x} + \mathbf{y}) = cT(\mathbf{x}) + T(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in V$ and $c \in F$. Let $\mathbf{x}, \mathbf{y} \in V$. Then we obtain $T(\mathbf{x} + \mathbf{y}) = T(1\mathbf{x} + \mathbf{y}) = 1 \cdot T(\mathbf{x}) + T(\mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y})$. Next, let $\mathbf{x} \in V$ and $c \in F$. Then we obtain $T(c\mathbf{x}) = T(c\mathbf{x} + \mathbf{0}) = c \cdot T(\mathbf{x}) + T(\mathbf{0}) = c \cdot T(\mathbf{x})$. This proves T is linear. The same type of reasoning can be used to show that if T satisfies

$$T\left(\sum_{i=1}^{n} a_i x_i\right) = \sum_{i=1}^{n} a_i T(x_i),$$

then T must be linear. □

Example 3.5.1. Given $A \in M_{m\times n}(F)$, define $L_A : M_{n\times 1} \to M_{m\times 1}(F)$ by $L_A(\mathbf{x}) = A\mathbf{x}$. To see that this is a linear transformation, note that for any $\mathbf{x}, \mathbf{y} \in M_{n\times 1}$ and any $c \in \mathbb{R}$

$$L_A(c\mathbf{x} + \mathbf{y}) = A(c\mathbf{x} + \mathbf{y}) = c(A\mathbf{x}) + A\mathbf{y} = cL_A(\mathbf{x}) + L_A(\mathbf{y}).$$

Example 3.5.2. Define $T : \mathbb{R}^2 \to \mathbb{R}^2$ by $T(a_1, a_2) = (2a_1 + a_2, a_1)$. To show that T is linear, let $c \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$, where $\mathbf{x} = (b_1, b_2)$ and $\mathbf{y} = (d_1, d_2)$. Since

$$c\mathbf{x} + \mathbf{y} = (cb_1 + d_1, cb_2 + d_2),$$

we have

$$T(c\mathbf{x} + \mathbf{y}) = (2(cb_1 + d_1) + cb_2 + d_2, cb_1 + d_1).$$

Also

$$\begin{aligned} cT(\mathbf{x}) + T(\mathbf{y}) &= c(2b_1 + b_2, b_1) + (2d_1 + d_2, d_1) \\ &= (2cb_1 + cb_2 + 2d_1 + d_2, cb_1 + d_1) \\ &= (2(cb_1 + d_1) + cb_2 + d_2, cb_1 + d_1). \end{aligned}$$

Thus T is linear.

Example 3.5.3. Define $T : P_n(\mathbb{R}) \to P_{n-1}(\mathbb{R})$ by $T(f) = f'$, where f' denotes the derivative of f. To show that T is linear, let g and h be vectors in $P_n(\mathbb{R})$ and $a \in \mathbb{R}$. Then

$$T(ag + h) = (ag + h)' = ag' + h' = aT(g) + T(h).$$

Thus T is linear.

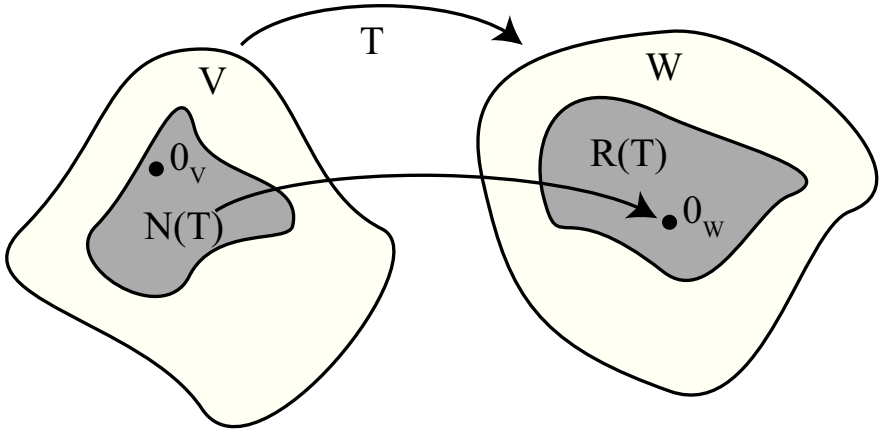
We now turn our attention to two very important sets associated with linear transformations: the *range* and *null space*. The determination of these sets allows us to examine more closely the intrinsic properties of linear transformations.

Definition 3.5.2 (Null Space and Range)**.** *Let V and W be vector spaces and let $T : V \to W$ be linear. We define the null space (or kernel) $N(T)$ of T to be the set of all vectors $\mathbf{x}$ in V such that $T(\mathbf{x}) = \mathbf{0}$; that is, $N(T) = \{\mathbf{x} \in V : T(\mathbf{x} = \mathbf{0})\}$.*

We define the range (or image) $\mathbb{R}(T)$ of T to be the subset of W consisting of all images (under T) of elements of V; that is $\mathbb{R}(T) = \{T(\mathbf{x}) : \mathbf{x} \in V\}$.

Definition 3.5.3 (Nullity and Rank)**.** *Let V and W be vector spaces and let $T : V \to W$ be linear. If $N(T)$ and $\mathbb{R}(T)$ are finite-dimensional, then we define the nullity of T, denoted* $\text{nullity}(T)$*, and the rank of T, denoted* $\text{rank}(T)$*, to be the dimensions of $N(T)$ and $\mathbb{R}(T)$, respectively.*

As a illustration of these definitions, consider the following figure:



The next theorem provides a method for finding a spanning set for the range of a linear transformation. With this accomplished, a basis for the range is easy to discover.

Theorem 3.5.2. *Let V and W be vector spaces, and let $T : V- \to W$ be linear. If $\beta = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a finite basis for V, then*

$$\mathbb{R}(T) = \text{span}(\{T(\mathbf{v}_1), \ldots, T(\mathbf{v}_n)\}).$$

Proof. Clearly $T(\mathbf{v}_i) \in \mathbb{R}(T)$ for each i. Because $\mathbb{R}(T)$ is a subspace of V that contains the set $\{T(\mathbf{v}_1), \ldots, T(\mathbf{v}_n)\}$, by Theorem 3.2.1 $\mathbb{R}(T)$ contains $\text{span}(\{T(\mathbf{v}_1), \ldots, T(\mathbf{v}_n)\})$. Now suppose that $\mathbf{w} \in \mathbb{R}(T)$. Then $\mathbf{w} = T(\mathbf{v})$ for some $\mathbf{v} \in V$. Because β is a basis for V, we have

$$\mathbf{v} = \sum_{i=1}^{n} a_i \mathbf{v}_i \quad \text{for some } a_1, \ldots, a_n \in F.$$

Since T is linear, it follows that

$$\mathbf{w} = T(\mathbf{v}) = \sum_{i=1}^{n} a_i T(\mathbf{v}_i) \in \text{span}(T(\beta)).$$

□

Reflecting on the action of a linear transformation, we see intuitively that the larger the nullity, the smaller the rank. In other words, the more vectors that are carried into $\mathbf{0}$, the smaller the range. The same heuristic reasoning tells us that the larger the rank, the smaller the nullity. The balance between rank and nullity is made precise in the next theorem.

Theorem 3.5.3 (Dimension Theorem)**.** *Let V and W be vector spaces, and let $T : V \to W$ be linear. If V is finite-dimensional, then*

$$\text{nullity}(T) + \text{rank}(T) = \dim(V).$$

Proof. Suppose that $\dim(V) = n$, $\dim(N(T)) = k$, and $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is a basis for $N(T)$. We may extend $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ to a basis $\beta = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ for V. We claim that $S = \{T(\mathbf{v}_{k+1}), \ldots, T(\mathbf{v}_n)\}$ is a basis for $\mathbb{R}(T)$.

First we prove that S generates $\mathbb{R}(T)$. Using Theorem 3.5.2 and the fact that $T(\mathbf{v}_i) = \mathbf{0}$ for $1 \leq i \leq k$, we have

$$\mathbb{R}(T) = \text{span}(\{T(\mathbf{v}_{k+1}), \ldots, T(\mathbf{v}_n)\})$$

Now we prove that S is linearly independent. Form

$$\sum_{i=k+1}^{n} b_i T(\mathbf{v}_i) = \mathbf{0} \quad \text{for } b_{k+1}, \ldots, b_n \in F.$$

Using the fact that T is linear, we have

$$T\left(\sum_{i=k+1}^{n} b_i v_i\right) = \mathbf{0} \quad \text{which implies} \quad \sum_{i=k+1}^{n} b_i v_i \in N(T).$$

Hence there exist $c_1, \ldots, c_k \in F$ such that

$$\sum_{i=k+1}^{n} b_i \mathbf{v}_i = \sum_{i=1}^{k} c_i \mathbf{v}_i \quad \text{which implies} \quad \sum_{i=1}^{k} (-c_i)\mathbf{v}_i + \sum_{i=k+1}^{n} b_i \mathbf{v}_i = \mathbf{0}$$

Since β is a basis for V, we have $b_i = 0$ for all i. Hence S is linearly independent. Notice that this argument also shows that $T(\mathbf{v}_{k+1}), \ldots, T(\mathbf{v}_n)$ are distinct, and hence $\text{rank}(T) = n - k$. □

Theorem 3.5.4. *Let V and W be vector spaces, and let $T : V \to W$ be linear. Then T is one-for-one if and only if $N(T) = \{\mathbf{0}\}$.*

Proof. First note that T is one-to-one if and only if $T(\mathbf{x}) = T(\mathbf{y})$ implies $\mathbf{x} = \mathbf{y}$. Suppose $N(T) = \{\mathbf{0}\}$ and $T(\mathbf{x}) = T(\mathbf{y})$. Since T is linear, $T(\mathbf{x} - \mathbf{y}) = T(\mathbf{x}) - T(\mathbf{y}) = \mathbf{0}$. Thus, $\mathbf{x} - \mathbf{y} \in N(T)$. By assumption, $N(T) = \{\mathbf{0}\}$. Hence, $\mathbf{x} - \mathbf{y} = \mathbf{0}$, which implies $\mathbf{x} = \mathbf{y}$. Now assume that T is injective. Let $\mathbf{x} \in N(T)$, then $T(\mathbf{x}) = \mathbf{0} = T(\mathbf{0})$. Hence $\mathbf{x} = \mathbf{0}$, since T is injective. □

Example 3.5.4. Let $M : M_{2\times 3}(\mathbb{R}) \to M_{2\times 2}(\mathbb{R})$ defined by

$$T\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} = \begin{pmatrix} 2a_{11} - a_{12} & a_{13} + 2a_{12} \\ 0 & 0 \end{pmatrix}.$$

For any $\mathbf{x}, \mathbf{y} \in M_{2\times 3}(\mathbb{R})$ and any $c \in \mathbb{R}$

$$\begin{aligned} T\begin{pmatrix} cx_{11} + y_{11} & cx_{12} + y_{12} & cx_{13} + y_{13} \\ cx_{21} + y_{21} & cx_{22} + y_{22} & cx_{23} + y_{23} \end{pmatrix} &= \begin{pmatrix} 2(cx_{11} + y_{11}) - (cx_{12} + y_{12}) & cx_{13} + y_{13} + 2(cx_{12} + y_{12}) \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} c(2x_{11} - x_{12}) & c(x_{13} + 2x_{12}) \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 2y_{11} - y_{12} & y_{13} + 2y_{12} \\ 0 & 0 \end{pmatrix} \\ &= cT\begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{bmatrix} + T\begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \end{pmatrix}. \end{aligned}$$

To find the null space, we must find an $\mathbf{x} \in M_{2\times 3}(\mathbb{R})$ such that

$$2x_{11} - x_{12} = 0 \quad \text{and} \quad x_{13} + 2x_{12} = 0.$$

Equating coefficients

$$\begin{bmatrix} 2 & -1 & 0 \\ 0 & 2 & 1 \end{bmatrix} \overset{\text{rref}}{\sim} \begin{bmatrix} 1 & 0 & 1/4 \\ 0 & 1 & 1/2 \end{bmatrix}.$$

Thus, $N(T) = \{(-r/4, -r/2, r, s, t, u)\}$ for $r, s, t, u \in \mathbb{R}$. Hence, a basis for the null space can be written

$$\beta_N = \left\{ \begin{bmatrix} -1/4 & -1/2 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right\}$$

and $\text{nullity}(T) = \dim(N(T)) = 4$. Since

$$\begin{aligned} \begin{bmatrix} 2a_{11} - a_{12} & a_{13} + 2a_{12} \\ 0 & 0 \end{bmatrix} &= \begin{bmatrix} 2a_{11} - a_{12} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & a_{13} + 2a_{12} \\ 0 & 0 \end{bmatrix} \\ &= (2a_{11} - a_{12}) \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + (a_{13} + 2a_{12} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \end{aligned}$$

a basis for the range is given by

$$\beta_{\mathbb{R}} = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \right\}$$

and $\text{rank}(T) = \dim(\mathbb{R}(T)) = 2$. This verifies the dimension theorem, since $\dim(M_{2\times 3}(\mathbb{R})) = 6 = 2 + 4$. Since $\text{nullity}(T) \neq 0$, T is not one-to-one. Since $\text{rank}(T) = 2 \neq \dim(M_{2\times 3}(\mathbb{R}))$, T is not onto.

Example 3.5.5. Let

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

and define $L_A : M_{5\times 1} \to M_{3\times 1}$ by $L_A(\mathbf{x}) = A\mathbf{x}$. To find a basis for $N(L_A)$, solve $A\mathbf{x} = \mathbf{0}$. We have

$$A = \begin{bmatrix} 1 & -1 & -1 & 2 & 1 \\ 2 & -2 & -1 & 3 & 3 \\ -1 & 1 & -1 & 0 & -3 \end{bmatrix} \overset{\text{rref}}{\sim} \begin{bmatrix} 1 & -1 & 0 & 1 & 2 \\ 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Thus, $N(T) = (r - s - 2t, r, s - t, s, t)$ for $r, s, t \in \mathbb{R}$. Hence, a basis for the null space can be written

$$\beta_N = \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -2 \\ 0 \\ -1 \\ 0 \\ 1 \end{bmatrix} \right\}$$

and $\text{nullity}(T) = \dim(N(T)) = 3$. In general $A\mathbf{x} = \mathbf{b}$ if and only if $\mathbf{b}$ is a linear combination of the columns of A (the column space). Hence

$$\beta_{\mathbb{R}} = \left\{ \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix} \right\}.$$

Chapter 4

Concavity, Convexity, Quasi-Concavity, and Quasi-Convexity

4.1 Convex Sets

Definition 4.1.1 (Convex Set)**.** *A set $X \in \mathbb{R}^n$ is (strictly) convex if given any two points $\mathbf{x}'$ and $\mathbf{x}''$ in X, the point*

$$\mathbf{x}^\lambda = (1-\lambda)\mathbf{x}' + \lambda\mathbf{x}''$$

is also in X (int X) for every $\lambda \in [0,1]$ ($\lambda \in (0,1)$).

Remark 4.1.1. A vector of the form $\mathbf{x}^\lambda = (1-\lambda)\mathbf{x}' + \lambda\mathbf{x}''$, with $\lambda \in [0,1]$ is called a *convex combination* of $\mathbf{x}'$ and $\mathbf{x}''$.

Theorem 4.1.1. *Let X and Y be convex sets in $\mathbb{R}^n$, and let α be a real number. Then the sets*

$$\alpha X = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{z} = \alpha\mathbf{x} \text{ for some } \mathbf{x} \in X\} \quad \text{and}$$
$$X + Y = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{z} = \mathbf{x} + \mathbf{y} \text{ for some } \mathbf{x} \in X \text{ and } \mathbf{y} \in Y\}$$

are convex.

Proof. Let X and Y be any two convex subsets of $\mathbb{R}^n$. Suppose $\alpha\mathbf{x}'$ and $\alpha\mathbf{x}''$ are any two points in αX, naturally with $\mathbf{x}'$ and $\mathbf{x}''$ in X. Given $\lambda \in [0,1]$, since X is convex, we know $\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'' \in X$. Thus,

$$\lambda(\alpha\mathbf{x}') + (1-\lambda)(\alpha\mathbf{x}'') = \alpha[\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}''] \in \alpha X.$$

Therefore αX is shown to be convex. Now suppose $\mathbf{x}' + \mathbf{y}'$ and $\mathbf{x}'' + \mathbf{y}''$ are any two points in $X + Y$, naturally with $\mathbf{x}'$ and $\mathbf{x}''$ in X and $\mathbf{y}'$ and $\mathbf{y}''$ in Y. Given $\lambda \in [0,1]$, since X and Y are both convex, we know $\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'' \in X$ and $\lambda\mathbf{y}' + (1-\lambda)\mathbf{y}'' \in Y$. Thus,

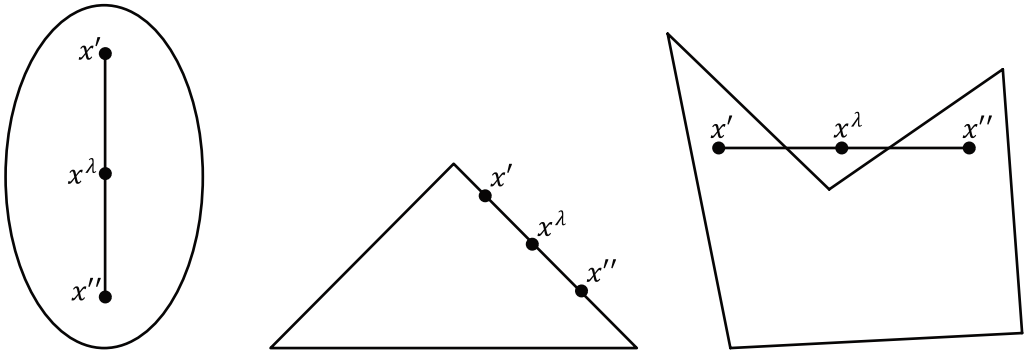
$$\lambda(\mathbf{x}' + \mathbf{y}') + (1-\lambda)(\mathbf{x}'' + \mathbf{y}'') = \lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'' + \lambda\mathbf{y}' + (1-\lambda)\mathbf{y}'' \in X + Y$$

Therefore $X + Y$ is shown to be convex. □

Example 4.1.1. Is $\{(x,y) \in \mathbb{R}^2 | y \geq x \wedge xy \geq 1\}$ open, closed, compact or convex?

Solution: Not open, since $S \neq \operatorname{int} S$. For example, $(1,1)$ is in the set, but every neighborhood around $(1,1)$ contains points which are not in the set. Closed since the set contains all of its boundary points. Not compact since the set is not bounded [(n,n) is in the set for $n \in \mathbb{N}$]. Not convex since since $(1,1)$ and $(-1,-1)$ are in the set, but $(1/2)(1,1) + (1/2)(-1,-1) = (0,0)$ is not.

Figure 4.1: Convex and Non-Convex Sets



(a) Strictly Convex (b) Convex, but *not* Strictly Convex (c) Not Convex

4.2 Concave and Convex Functions

Definition 4.2.1 (Concave Function)**.** *The function $f : \mathbb{R}^n \supseteq X \to \mathbb{R}$, where X is a convex set, is concave if given any two points $\mathbf{x}'$ and $\mathbf{x}''$ in X we have*

$$(1-\lambda)f(\mathbf{x}') + \lambda f(\mathbf{x}'') \leq f[(1-\lambda)\mathbf{x}' + \lambda\mathbf{x}''] \equiv f(\mathbf{x}^\lambda) \quad \forall\lambda \in [0,1]$$

and is strictly concave if the inequality holds strictly for $\lambda \in (0,1)$, that is, if

$$\forall\mathbf{x}', \mathbf{x}'' \in X \text{ and } \lambda \in (0,1), \;\; (1-\lambda)f(\mathbf{x}') + \lambda f(\mathbf{x}'') < f[(1-\lambda)\mathbf{x}' + \lambda\mathbf{x}''] \equiv f(\mathbf{x}^\lambda).$$

Remark 4.2.1. Reversing the direction of the inequalities in the theorem, we obtain the definitions of convexity and strict convexity.

Many introductory calculus texts call convex functions "concave up" and concave functions "concave down", as we did in section 1.1.4. Henceforth, we will stick with the more classical terms: "convex" and "concave".

Theorem 4.2.1. *Let $f : \mathbb{R}^n \supseteq X \to \mathbb{R}$ be a C^1 function defined on an open and convex set X. Then f is concave if and only if given any two points $\mathbf{x}$ and $\mathbf{x}_0$ in X, we have*

$$f(\mathbf{x}) \leq f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0).$$

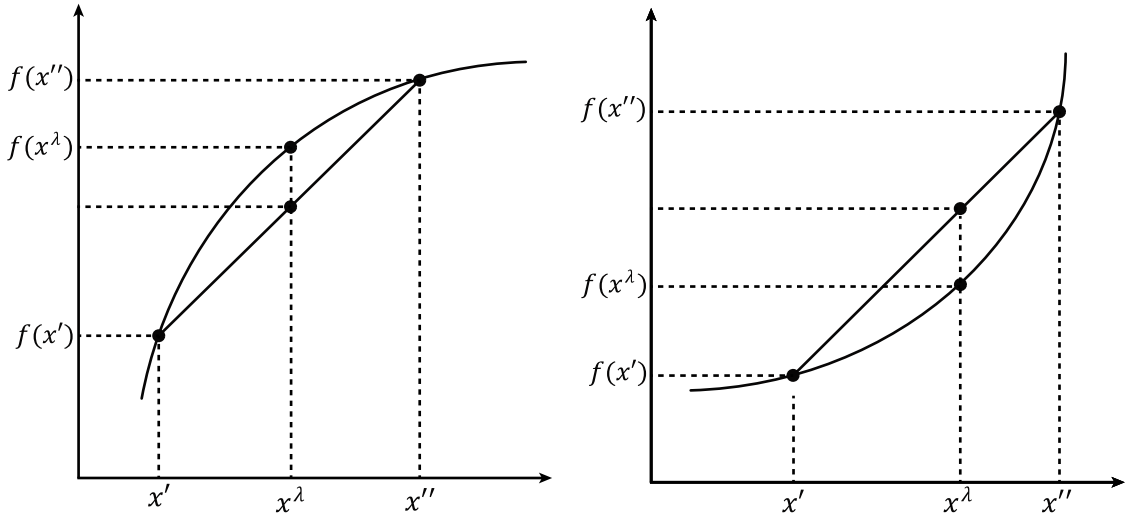
Moreover, f is strictly concave if and only if the inequality holds strictly, that is, if and only if

$$f(\mathbf{x}) < f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$$

for all pairs of distinct points $\mathbf{x}_0$ and $\mathbf{x}$ in X.

Remark 4.2.2. This theorem says that a function f is concave if and only if the graph of f lies everywhere on or below any tangent plane. Equivalently, it says that a function is concave if and only if the slope of the function at some arbitrary point, say $\mathbf{x}_0 < \mathbf{x}$, is greater than the slope of the secant line between points $\mathbf{x}$ and $\mathbf{x}_0$. Reversing the direction of the inequalities in the theorem, we obtain a theorem corresponding to convexity and strict convexity.

Figure 4.2: Concave and Convex Functions



(a) Concave Function (b) Convex Function

Theorem 4.2.2. *Let $f : \mathbb{R}^n \supseteq X \to \mathbb{R}$ be a concave function and $g : \mathbb{R} \to \mathbb{R}$ be an increasing and concave function defined on an interval I containing $f(X)$. Then the function $h : X \to \mathbb{R}$ defined by $h(\mathbf{x}) = g[f(\mathbf{x})]$ is concave. Moreover, if f is strictly concave and g is strictly increasing then h is strictly concave. Analogous claims hold for f convex (again with g increasing).*

Proof. Consider any $\mathbf{x}', \mathbf{x}'' \in X$ and $\lambda \in [0, 1]$. Since f is concave, we have

$$f(\mathbf{x}^\lambda) \equiv f(\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'') \geq \lambda f(\mathbf{x}') + (1-\lambda)f(\mathbf{x}'').$$

Since g is increasing and concave

$$\begin{aligned} h(\mathbf{x}^\lambda) = g[f(\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'')] &\geq g[\lambda f(\mathbf{x}') + (1-\lambda)f(\mathbf{x}'')] \\ &\geq \lambda g[f(\mathbf{x}')] + (1-\lambda)g[f(\mathbf{x}'')] \\ &= \lambda h(\mathbf{x}') + (1-\lambda)h(\mathbf{x}''). \end{aligned}$$

This establishes that h is concave. If f is strictly concave and g is strictly increasing, then the inequality is strict and hence h is strictly concave. □

Example 4.2.1. Let the domain be $\mathbb{R}_{++}$. Consider $h(x) = e^{1/x}$. Let $f(x) = 1/x$ and let $g(y) = e^y$. Then $h(x) = g[f(x)]$. Function f is strictly convex and g is (strictly) convex and strictly increasing. Therefore, by Theorem 4.2.2, h is strictly convex.

Remark 4.2.3. It is important in Theorem 4.2.2 that g is increasing. To see this, let the domain be $\mathbb{R}_{++}$. Consider $h(x) = e^{-x^2}$, which just the standard normal density except that it is off by a factor of $1/\sqrt{2\pi}$. Let $f(x) = e^{x^2}$ and let $g(y) = 1/y$. Then $h(x) = g[f(x)]$. Now, f is convex on $\mathbb{R}_{++}$ and g is also convex on $\mathbb{R}_{++}$. The function h is not, however, convex. While it is strictly convex for $|x|$ sufficiently large, for x near zero it is strictly concave. This does not contradict Theorem 4.2.2 because g here is decreasing.

Theorem 4.2.3. *Suppose $f_1, \ldots, f_n$ are concave functions where $f_i : \mathbb{R}^n \supseteq X \to \mathbb{R}$. Then for any $\alpha_1, \ldots, \alpha_n$ for which each $\alpha_i \geq 0$, $f \equiv \sum_{i=1}^n \alpha_i f_i$ is also a concave function. If, in addition, at least one f_j is strictly concave and $\alpha_j > 0$, then f is strictly concave.*

Proof. Consider any $\mathbf{x}', \mathbf{x}'' \in X$ and $\lambda \in [0, 1]$. If each f_i is concave, we have

$$f_i(\mathbf{x}^\lambda) \equiv f_i(\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'') \geq \lambda f_i(\mathbf{x}') + (1-\lambda) f_i(\mathbf{x}'')$$

Therefore,

$$f(\mathbf{x}^\lambda) \equiv f(\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'') = \sum_{i=1}^n \alpha_i f_i(\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'') \geq \sum_{i=1}^n \alpha_i [\lambda f_i(\mathbf{x}') + (1-\lambda) f_i(\mathbf{x}'')]$$
$$= \lambda \sum_{i=1}^n \alpha_i f_i(\mathbf{x}') + (1-\lambda) \sum_{i=1}^n \alpha_i f_i(\mathbf{x}'') \equiv \lambda f(\mathbf{x}') + (1-\lambda) f(\mathbf{x}'').$$

This establishes that f is concave. If some f_j is strictly concave and $\alpha_j > 0$, then the inequality is strict. □

4.3 Concavity, Convexity, and Definiteness

The following result says that a function is concave if and only if its Hessian is negative semi-definite everywhere. A twice-differentiable function of a *single* variable is concave (convex) if and only if $f''(x) \leq (\geq) 0$ everywhere.

Theorem 4.3.1. *Let f be a C^2 function on an open convex set X of $\mathbb{R}^n$. Then*

(a) If $H(f)$ is negative definite for every $\mathbf{x} \in X$, then f is strictly concave.

(b) If $H(f)$ is negative semi-definite for every $\mathbf{x} \in X$, then f is concave.

(c) If f is concave, then $H(f)$ is negative semi-definite for every $\mathbf{x} \in X$.

(d) If $H(f)$ is positive definite for every $\mathbf{x} \in X$, then f is strictly convex.

(e) If $H(f)$ is positive semi-definite for every $\mathbf{x} \in X$, then f is convex.

(f) If f is convex, then $H(f)$ is positive semi-definite for every $\mathbf{x} \in X$.

Remark 4.3.1. Note that if f is strictly concave, the Hessian can either be negative semi-definite or negative definite. Thus, if you show that the Hessian is *not* negative definite, but only negative semi-definite, you *cannot* conclude that f *is not* strictly concave.

Example 4.3.1. The Hessian of the function $f(x, y) = x^4 + x^2y^2 + y^4 - 3x - 8y$ is

$$H(f) = \begin{bmatrix} 12x^2 + 2y^2 & 4xy \\ 4xy & 2x^2 + 12y^2 \end{bmatrix}.$$

The principle minors, $B_1^{(1)} = 12x^2 + 2y^2$, $B_1^{(2)} = 2x^2 + 12y^2$, and $B_2 = 24x^4 + 132x^2y^2 + 24y^4$ are all weakly positive for all values of x and y, so f is a convex function on all $\mathbb{R}^n$.

Example 4.3.2. A simple utility or production function is $f(x, y) = xy$. Its Hessian is

$$H(f) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

whose second order principle minor is $\det H(f) = -1$. Since this second order principle minor is negative, $H(f)$ is indefinite and f is neither concave nor convex.

Example 4.3.3. Consider the monotonic transformation of the function f in the previous example by the function $g(z) = z^{1/4} : g[f(x, y)] = x^{1/4}y^{1/4}$, defined only on the positive quadrant $\mathbb{R}^2_+$. The hessian of g is

$$H(g) = \begin{bmatrix} -\frac{3}{16}x^{-7/4}y^{1/4} & \frac{1}{16}x^{-3/4}y^{-3/4} \\ \frac{1}{16}x^{-3/4}y^{-3/4} & -\frac{3}{16}x^{1/4}y^{-7/4} \end{bmatrix}.$$

The first order principle minors are both non-positive and the second order principle minor, $x^{-3/2}y^{-3/2}/32$, is non-negative. Therefore, $H(g)$ is negative semi-definite on $\mathbb{R}^2_+$ and G is a concave function on $\mathbb{R}^2_+$.

4.4 Quasi-concave and Quasi-convex Functions

Definition 4.4.1 (Quasi-concavity and Quasi-convexity)**.** *Let $f : \mathbb{R}^n \supseteq X \to \mathbb{R}$ be a real-valued function defined on a convex set X. We say that f is quasi-concave (quasi-convex) if for all $\mathbf{x}'$ and $\mathbf{x}''$ in X and all $\lambda \in [0, 1]$ we have*

$$f[(1-\lambda)\mathbf{x}' + \lambda\mathbf{x}''] \geq \min\{f(\mathbf{x}'), f(\mathbf{x}'')\}$$
$$(f[(1-\lambda)\mathbf{x}' + \lambda\mathbf{x}''] \leq \max\{f(\mathbf{x}'), f(\mathbf{x}'')\}).$$

We say that f is strictly quasi-concave (quasi-convex) if for all $\mathbf{x}'$ and $\mathbf{x}''$ in X and all $\lambda \in (0, 1)$ we have

$$f[(1-\lambda)\mathbf{x}' + \lambda\mathbf{x}''] > \min\{f(\mathbf{x}'), f(\mathbf{x}'')\}$$
$$(f[(1-\lambda)\mathbf{x}' + \lambda\mathbf{x}''] < \max\{f(\mathbf{x}'), f(\mathbf{x}'')\}).$$

Theorem 4.4.1. *Let f be a real-valued function defined on a convex set $X \subseteq \mathbb{R}^n$. Then f is quasi-concave (quasi-convex) if and only if the upper contour sets (lower contour sets) of f are all convex, that is, if for any $a \in \mathbb{R}$ the set*

$$U_a = \{\mathbf{x} \in X : f(\mathbf{x}) \geq a\}$$
$$(L_a = \{\mathbf{x} \in X : f(\mathbf{x}) \leq a\})$$

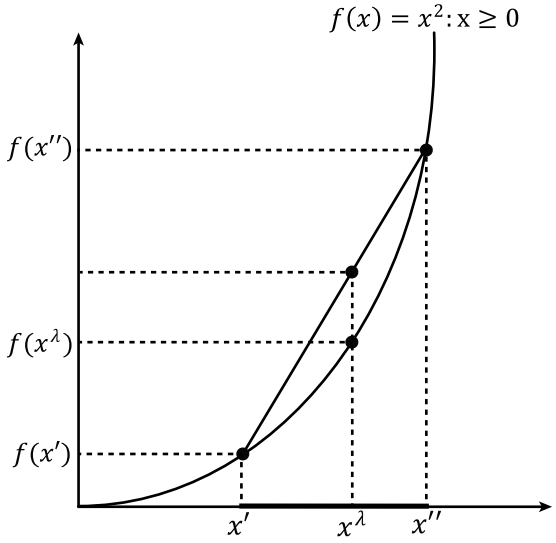
is convex.

Proof. Assume f is quasi-concave. Fix a and let $\mathbf{x}', \mathbf{x}'' \in U_a$. Then for all $\lambda \in [0, 1]$,

$$f(\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'') \geq \min\{f(\mathbf{x}'), f(\mathbf{x}'')\}$$

Since $\mathbf{x}', \mathbf{x}'' \in U_a$, $f(\mathbf{x}') \geq a$ and $f(\mathbf{x}'') \geq a$, which implies the $\min\{f(\mathbf{x}'), f(\mathbf{x}'')\} \geq a$. Therefore, $f(\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'') \geq a$ and thus $\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'' \in U_a$. Hence, the upper contour set is convex. Now assume the upper contour set is convex. Then for all $\lambda \in [0, 1]$ and for $\mathbf{x}', \mathbf{x}'' \in U_a$, we have $\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'' \in U_a$. This implies that $f(\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'') \geq a$. Since this result must hold for any a, it must hold for $a = \min\{f(\mathbf{x}'), f(\mathbf{x}'')\}$. Thus, f is quasi-concave. A similar argument holds for quasi-convexity. □

Figure 4.3: Quasi-concave but not concave



Theorem 4.4.2. *All concave (convex) functions are quasi-concave (quasi-convex) and all strictly concave (strictly convex) functions are strictly quasi-concave (strictly quasi-convex).*

Proof. Without loss of generality assume $f(\mathbf{x}') \geq f(\mathbf{x}'')$. Since f is concave, for $0 \leq \lambda \leq 1$

$$\begin{aligned} f(\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'') &\geq \lambda f(\mathbf{x}') + (1-\lambda) f(\mathbf{x}'') \\ &\geq \lambda f(\mathbf{x}'') + (1-\lambda) f(\mathbf{x}'') \\ &= \min\{f(\mathbf{x}'), f(\mathbf{x}'')\}. \end{aligned}$$

Thus, f is also quasi-concave. If f is strictly concave then the inequalities become strict and hence f is strictly quasi-concave. The convex case can be proved in a similar fashion. □

It is important to note that the converse of the above theorem is not valid in general (see figure 4.3). The function f defined on $X = \{x : x \geq 0\}$ by $f(x) = x^2$ is quasi-concave (UCS is convex) but not concave on X, actually it is strictly convex on X.

Theorem 4.4.3. *Suppose $f : \mathbb{R}^n \supseteq X \to \mathbb{R}$ is quasi-concave (quasi-convex) and $\phi : f(X) \to \mathbb{R}$ is increasing. Then $\phi \circ f : X \to \mathbb{R}$ is quasi-concave (quasi-convex). If f is strictly quasi-concave (quasi-convex) and ϕ is strictly increasing, then $\phi \circ f$ is strictly quasi-concave (quasi-convex).*

Proof. Consider any $\mathbf{x}', \mathbf{x}'' \in X$. If f is quasi-concave, then

$$f(\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'') \geq \min\{f(\mathbf{x}'), f(\mathbf{x}'')\}.$$

Therefore, ϕ increasing implies

$$\phi[f(\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'')] \geq \phi[\min\{f(\mathbf{x}'), f(\mathbf{x}'')\}] = \min\{\phi[f(\mathbf{x}')], \phi[f(\mathbf{x}'')]\}.$$

Thus, $\phi \circ f = \phi[f(\mathbf{x})]$ is quasi-concave. If f is strictly quasi-concave and ϕ is strictly increasing, the inequalities are strict and thus $\phi \circ f$ is strictly quasi-concave. A similar argument holds for quasi-convexity. □

Corollary 4.4.1. *Suppose $f : \mathbb{R}^n \supseteq X \to \mathbb{R}$ is quasi-concave and $\phi : f(X) \to \mathbb{R}$ is decreasing. Then $\phi \circ f : X \to \mathbb{R}$ is quasi-convex. Suppose $f : X \to \mathbb{R}$ is quasi-convex and $\phi : f(X) \to \mathbb{R}$ is decreasing. Then $\phi \circ f : X \to \mathbb{R}$ is quasi-concave.*

Proof. Consider any $\mathbf{x}', \mathbf{x}'' \in X$. If f is quasi-concave, then

$$f(\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'') \geq \min\{f(\mathbf{x}'), f(\mathbf{x}'')\}.$$

Therefore, ϕ decreasing implies

$$\phi[f(\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'')] \leq \phi[\min\{f(\mathbf{x}'), f(\mathbf{x}'')\}] = \max\{\phi[f(\mathbf{x}')], \phi[f(\mathbf{x}'')]\}.$$

Thus, $\phi \circ f = \phi[f(\mathbf{x})]$ is quasi-convex. Similarly if f is quasi-convex. □

Remark 4.4.1. The sum of quasi-concave functions need not be quasi-concave unlike the sum of concave functions which is concave. For instance $f_1(x) = x^3$ and $f_2(x) = -x$ are both quasi-concave, but the sum $f_3(x) = f_1(x) + f_2(x) = x^3 - x$ is neither quasi-concave nor convex.

Example 4.4.1. Show that the Cobb-Douglas function

$$f(\mathbf{x}) = \prod_{i=1}^{n} x_i^{\alpha_i},$$

where $\alpha_i > 0$ for $i = 1, \ldots, n$ is quasi-concave for $\mathbf{x} \gg 0$.

Solution: Consider the natural log of Cobb-Douglas function,

$$\ln f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \ln x_i,$$

which is concave since for all i, $\ln x_i$ is concave and the sum of concave functions is concave (Theorem 4.2.3). Thus, since concavity implies quasi-concavity, $\ln f(\mathbf{x})$ is also quasi-concave. The exponent e^t is strictly increasing function: $\mathbb{R} \to \mathbb{R}$, hence $f(\mathbf{x}) = \exp(\ln f(\mathbf{x}))$ is quasi-concave by Theorem 4.4.3.

4.5 Quasi-concavity, Quasi-convexity, and Definiteness

Theorem 4.5.1. *Let $f : \mathbb{R}^n \supseteq X \to \mathbb{R}$ be a C^2 function defined on an open and convex set $X \subseteq \mathbb{R}^n$, and let Δ_k be the leading principle minor of the bordered Hessian of f.*

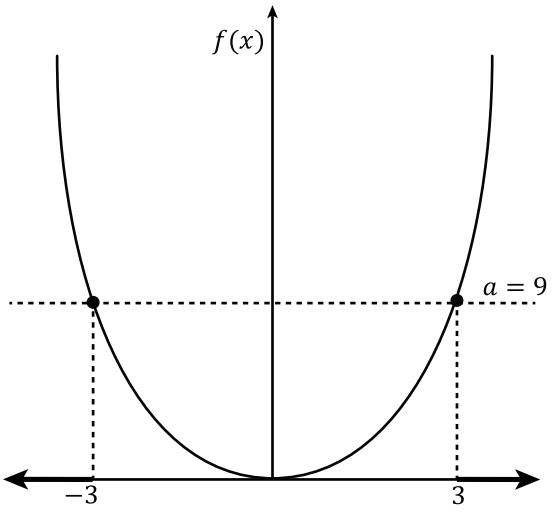
(a) A necessary condition for f to be quasi-concave (quasi-convex) is that

$$(-1)^{k+1}\Delta_k \geq 0 \; \forall k = 2, \ldots, n+1 \; \forall \mathbf{x} \in X$$
$$(\Delta_k \leq 0 \; \forall k = 2, \ldots, n+1 \; \forall \mathbf{x} \in X)$$

(b) A sufficient condition for f to be quasi-concave (quasi-convex) is that

$$(-1)^{k+1}\Delta_k > 0 \; \forall k = 2, \ldots, n+1 \; \forall \mathbf{x} \in \mathbb{R}^n_+$$
$$(\Delta_k < 0 \; \forall k = 2, \ldots, n+1 \; \forall \mathbf{x} \in \mathbb{R}^n_+)$$

Figure 4.4: Not Quasi-concave



(c) If $X \subseteq \mathbb{R}^n_{++}$, f is monotonically increasing (decreasing), and

$$(-1)^{k+1}\Delta_k > 0 \; \forall k = 2, \ldots, n+1 \; \forall \mathbf{x} \in \mathbb{R}^n_+$$
$$(\Delta_k < 0 \; \forall k = 2, \ldots, n+1 \; \forall \mathbf{x} \in \mathbb{R}^n_+),$$

then f is strictly quasi-concave (strictly quasi-convex).

Remark 4.5.1. Be very careful with the direction of the above definitions. In figure 4.4, the function $f(x) = x^2$ is not quasi-concave since the upper contour set: $\{x \in \mathbb{R} | f(x) \geq 3\} = (-\infty, -3] \cup [3, \infty)$ is not a convex set. However,

$$\overline{H}(f) = \begin{bmatrix} 0 & 2x \\ 2x & 2 \end{bmatrix}$$

so $(-1)^{2+1}\Delta_2 = 4x^2 \geq 0$ for all x with strict inequality everywhere except at $x = 0$. Although this function fulfills the necessary condition for quasi-concavity, the function is not quasi-concave.

Example 4.5.1. Prove or give a counterexample: If f is a strictly convex function, then f cannot be quasi-concave.

Solution: False. For $f(x) = 1/x$ defined on $\mathbb{R}_{++}$, the upper contour set $\{x : f(x) \geq c\}$ is convex for all c; For example, for $c = 1$, the upper contour set is the interval (0,1], which is clearly convex. Thus, this function is quasiconcave. However, the function is also strictly convex since ($f''(x) = 2/x^3 > 0$ for all $x > 0$).

Example 4.5.2. Consider $f(x) = x^3 + x$. For $x \in \mathbb{R}$, the second order condition, $f_{xx} = 6x$, is not always nonpositive. Thus, this function is not concave. The bordered Hessian is given by

$$\overline{H}(f) = \begin{bmatrix} 0 & 3x^2 + 1 \\ 3x^2 + 1 & 6x \end{bmatrix}$$

so $(-1)^{2+1}\Delta_2 = (3x^2 + 1)^2 > 0$. Therefore, this function is quasi-concave.

Example 4.5.3.

For the region with $x > 0$ and $-x < y < x$, define $f(x, y) = x^2 - y^2$. Is f concave where defined? Is f quasiconcave where defined?

Solution: Given $f(x, y) = x^2 - y^2$,

(a) $f_x = 2x, f_{xx} = 2, f_y = -2y, f_{xy} = 0, f_{yy} = -2$. The Hessian is

$$H = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}.$$

Since

$$\Delta_1 = 2 > 0 \text{ and } \Delta_2 = \begin{vmatrix} 2 & 0 \\ 0 & -2 \end{vmatrix} = -4 < 0,$$

H is not negative semidefinite so f is not concave.

(b) Bordering the Hessian with first partials, we obtain

$$\overline{H}(x, y) = \begin{bmatrix} 0 & 2x & -2y \\ 2x & 2 & 0 \\ -2y & 0 & -2 \end{bmatrix},$$

$$(-1)^{2+1}\Delta_2 = (-1)^3(-4x^2) > 0 \ \ \forall x \neq 0,$$
$$(-1)^{2+2}\Delta_3 = (-1)^4 8(x^2 - y^2) > 0 \text{ as } |y| < x.$$

Since $(-1)^{k+1}|\Delta_k| > 0, k = 2, 3$, f is quasi-concave.

Example 4.5.4. Is the utility function $u(x_1, x_2) = \sqrt{x_1} + \sqrt{x_2}$ quasi-concave for $x_1 > 0$, $x_2 > 0$? Is it concave for $x_1 > 0$, $x_2 > 0$?

Solution: Given $U(x_1, x_2) = \sqrt{x_1} + \sqrt{x_2}$, the Hessian is given by

$$H = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} = \begin{bmatrix} -x_1^{-3/2}/4 & 0 \\ 0 & -x_2^{-3/2}/4 \end{bmatrix}$$

For all $(x_1, x_2) \in \mathbb{R}^2_{++}$, $\Delta_1 = -x_1^{-3/2}/4 < 0$ and $\Delta_2 = |H| = x_1^{-3/2}x_2^{-3/2}/16 > 0$. Hence H is negative definite, which implies that $U(x_1, x_2)$ is strictly concave on $\mathbb{R}^2_{++}$. Since concavity implies quasi-concavity, $U(x_1, x_2)$ is also quasi-concave on $\mathbb{R}^2_{++}$.

Example 4.5.5. Is $f(x, y) = x^2y^2$ concave and/or quasi-concave on $\{(x, y) \in \mathbb{R}^2 | x \geq 0, y \geq 0\}$? Is f concave and/or quasi-concave on $\mathbb{R}^2$?

Solution: The Hessian is given by

$$H = \begin{bmatrix} 2y^2 & 4xy \\ 4xy & 2x^2 \end{bmatrix}.$$

For all $x, y \geq 0$, $\Delta_1 = 2y^2 \geq 0$ for all $x, y \geq 0$ and $\Delta_2 = |H| = -12x^2y^2 \leq 0$. Thus, H is not negative semi-definite, which implies that f is not concave on $\mathbb{R}^2_+$ (and also not concave on $\mathbb{R}^2$). Checking quasi-concavity:

$$\overline{H} = \begin{bmatrix} 0 & 2xy^2 & 2x^2y \\ 2xy^2 & 2y^2 & 4xy \\ 2x^2y & 4xy & 2x^2 \end{bmatrix}.$$

For all $x, y > 0$, $(-1)^{2+1}\Delta_2 = 4x^2y^4 > 0$ and $(-1)^{3+1}\Delta_3 = 16x^4y^4 > 0$. Thus, for $(x, y) \in \mathbb{R}^2_{++}$, the sufficient conditions for quasi-concavity hold since the inequalities are both strict. It remains to check whether the function f is quasiconcave on $\mathbb{R}^2_+$. Since $\{(x, y) \in \mathbb{R}^2_+ | f(x, y) \geq 0\} = \mathbb{R}^2_+$, all upper contour sets are convex, and f is quasi-concave on $\mathbb{R}^2_+$. Note f is not quasi-concave on $\mathbb{R}^2$ since

$$f((1/2)(1, 1) + (1/2)(-1, -1)) = f(0, 0) = 0 \ngeq \min\{f(1, 1), f(-1, -1)\} = 1.$$

Example 4.5.6. Is $f(x, y) = \ln(x + y)$ quasi-concave on the set of strictly positive x and y values?

Solution: The Hessian is given by

$$H = \begin{bmatrix} -\frac{1}{(x+y)^2} & -\frac{1}{(x+y)^2} \\ -\frac{1}{(x+y)^2} & -\frac{1}{(x+y)^2} \end{bmatrix}.$$

For all $x, y > 0$, $B_1^{(1)} = B_1^{(2)} = -\frac{1}{(x+y)^2} < 0$ and $\Delta_2 = |H| = 0$. Thus, H is negative semidefinite on $\mathbb{R}^2_{++}$, implying that f is concave, and hence quasi-concave.

Example 4.5.7. Is $f(x, y, z) = \sqrt{x} + \sqrt{y} + z^2$ concave and/or quasi-concave on $\mathbb{R}^3_{++}$?

Solution: The Hessian is given by

$$\begin{bmatrix} -\frac{1}{4}x^{-3/2} & 0 & 0 \\ 0 & -\frac{1}{4}y^{-3/2} & 0 \\ 0 & 0 & 2 \end{bmatrix},$$

which is not negative semi-definite ($\Delta_1 < 0$, $\Delta_2 > 0$, and $\Delta_3 > 0$), so the function is not concave. The bordered Hessian

$$\begin{bmatrix} 0 & \frac{1}{2}x^{-1/2} & \frac{1}{2}y^{-1/2} & 2z \\ \frac{1}{2}x^{-1/2} & -\frac{1}{4}x^{-3/2} & 0 & 0 \\ \frac{1}{2}y^{-1/2} & 0 & -\frac{1}{4}y^{-3/2} & 0 \\ 2z & 0 & 0 & 2 \end{bmatrix}$$

has determinant $(2z^2x^3 - x^{5/2} - x^{3/2}y)/8$, which is positive for some $(x, y, z) \in \mathbb{R}^3_{++}$ and negative for others. Thus the function is not quasi-concave on $\mathbb{R}^3_{++}$.

Example 4.5.8. Is $f(x, y) = x^2y$ concave and/or quasi-concave on $\mathbb{R}^2_{++}$? Explain.

Solution: The gradient and Hessian are respectively given by

$$\nabla f = \begin{pmatrix} 2xy \\ x^2 \end{pmatrix} \qquad H = \begin{bmatrix} 2y & 2x \\ 2x & 0 \end{bmatrix}.$$

Thus, f is not negative semidefinite since $\Delta_2 = -4x^2 < 0$. Hence f is not concave. The bordered Hessian is given by

$$\overline{H} \begin{bmatrix} 0 & 2xy & x^2 \\ 2xy & 2y & 2x \\ x^2 & 2x & 0 \end{bmatrix}$$

Thus, $(-1)^3\Delta_2 = 4x^2y^2 > 0$ for all $(x, y) \in \mathbb{R}^2_{++}$ and $(-1)^4\Delta_3 = 6x^4y > 0$ for all $(x, y) \in \mathbb{R}^2_{++}$. Thus f is quasi-concave.

Chapter 5

Optimization

5.1 Unconstrained Optimization

Theorem 5.1.1. *If $f : \mathbb{R}^n \to \mathbb{R}$ and all its first partial derivatives are continuously differentiable on a set which contains input vector $\mathbf{x}^*$ in its interior, then*

(a) (Necessary Condition) f has a local maximum (minimum) at $\mathbf{x}^*$ only if $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $H(f(\mathbf{x}^*))$ is negative (positive) semi-definite.

(b) (Sufficient Condition) f has a strict local maximum (minimum) at $\mathbf{x}^*$ if $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $H(f(\mathbf{x}^*))$ is negative (positive) definite.

(c) If $H(f(\mathbf{x}^*))$ is indefinite, then $\mathbf{x}^*$ is neither a local maximum nor a local minimum.

Remark 5.1.1. Note that in the univariate case $\nabla f = \mathbf{0}$ is replaced by $f' = 0$ and $H(f)$ NSD (PSD, ND, and PD, respectively) is replaced by $f'' \leq (\geq, <, >)\, 0$.

Example 5.1.1. Let $f(x, y) = -3x^2 + xy - 2x + y - y^2 + 1$. Then

$$\nabla f = \begin{pmatrix} -6x + y - 2 \\ x + 1 - 2y \end{pmatrix}, \qquad H(f) = \begin{bmatrix} -6 & 1 \\ 1 & -2 \end{bmatrix},$$

$$\Delta_1 = -6 < 0, \quad \text{and} \quad \Delta_2 = \begin{vmatrix} -6 & 1 \\ 1 & -2 \end{vmatrix} = 11 > 0,$$

so $H(f)$ is negative definite.

$$\nabla f = \begin{bmatrix} -6 & 1 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

so $\nabla f(x, y) = \mathbf{0}$ if

$$\begin{bmatrix} -6 & 1 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} -2 \\ 1 \end{bmatrix} = \mathbf{0}$$

or

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -6 & 1 \\ 1 & -2 \end{bmatrix}^{-1} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \frac{1}{11} \begin{bmatrix} -2 & -1 \\ -1 & -6 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} -3/11 \\ 4/11 \end{bmatrix}.$$

Thus, f has a strict local maximum at $(x, y) = (-3/11, 4/11)$.

Definition 5.1.1 (Saddle Point)**.** *A critical point $\mathbf{x}^*$ of f for which $H(f(\mathbf{x}^*))$ is indefinite is called a saddle point of f. A saddle point is a minimum of f in some directions and a maximum of f in other directions.*

Example 5.1.2. Let $f(x,y) = x^3 + x^2y + 2y^2$. Then

$$\nabla f = \begin{pmatrix} 3x^2 + 2xy \\ x^2 + 4y \end{pmatrix}; \qquad H(f) = \begin{bmatrix} 6x + 2y & 2x \\ 2x & 4 \end{bmatrix}.$$

Here $H(f)$ depends on the (x,y) at which it is evaluated. $\nabla f(x,y) = \mathbf{0}$ if $3x^2 + 2xy = 0$ and $x^2 + 4y = 0$. From the last equation $y = -x^2/4$. Substituting this value into the previous equation, $(x^2/2)(6 - x) = 0$. Thus, $(x,y) = (0,0)$ or $(x,y) = (6,-9)$.

$$H[f(0,0)] = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix},$$

which is PSD ($B_1^{(1)} = 0 = 0$, $B_1^{(2)} = 4$, and $B_2 = 0$).

$$H[f(6,-9)] = \begin{bmatrix} 18 & 12 \\ 12 & 4 \end{bmatrix}$$

and its leading principle minors are

$$\Delta_1 = 18 > 0 \quad \text{and} \quad \Delta_2 = \begin{vmatrix} 18 & 12 \\ 12 & 4 \end{vmatrix} = -72 < 0.$$

Thus, since the Hessian is indefinite at $(6,-9)$, this point is a saddle point for the function. The point $(0,0)$ satisfies the necessary conditions for a local minimum, but not the sufficient conditions. However, $f(x,0) = x^3$ so f cannot attain either a local maximum or a local minimum at $(0,0)$. This function has no local maxima or local minima.

Theorem 5.1.2 (Global Maxima/Minima)**.** *Let $f : X \to \mathbb{R}$ be a C^2 function whose domain is a convex open subset X of $\mathbb{R}^n$. If f is a concave (convex) function on X and $\nabla f(\mathbf{x}^*) = \mathbf{0}$ for some $\mathbf{x}^* \in U$, then $\mathbf{x}^*$ is a global maximum (minimum) of f on X.*

Remark 5.1.2. It is interesting to compare Theorems 5.1.1 and 5.1.2. In order to guarantee that a critical point $\mathbf{x}^*$ of a C^2 function f is a strict local maximum (minimum), we need to show that $H(f(\mathbf{x}^*))$ is negative (positive) definite; showing that $H(f(\mathbf{x}^*))$ is negative (positive) semi-definite is not strong enough. However, if we can show that $H(f(\mathbf{y}))$ is negative (positive) semi-definite not just at $\mathbf{x}^*$ but for all $\mathbf{y}$ in a neighborhood about $\mathbf{x}^*$, then by Theorem 5.1.2, we can conclude that $\mathbf{x}^*$ is a maximum (minimum) of f.

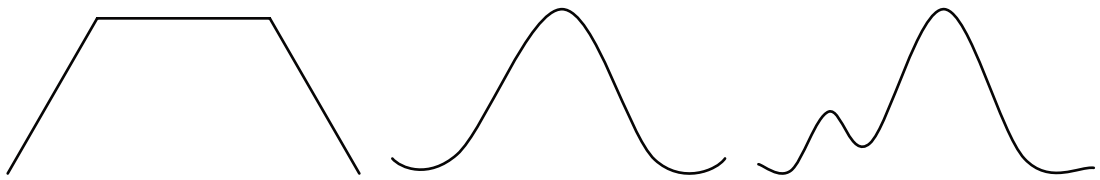
Remark 5.1.3. A global maximum (or minimum) does not necessarily have to be a strict maximum (or minimum). Moreover a strict maximum (or minimum) does not necessarily have to be a global maximum (or minimum). To see this consider figure 5.1.

Example 5.1.3. Find all local maxima and minima of $f(x,y,z) = x^2 + x(z-2) + 3(y-1)^2 + z^2$

Solution: The associated first order conditions are

$$\begin{aligned} \frac{\partial f}{\partial x} &= 2x + z - 2 \overset{\text{set}}{=} 0 \\ \frac{\partial f}{\partial y} &= 6(y-1) \overset{\text{set}}{=} 0 \\ \frac{\partial f}{\partial z} &= x + 2z \overset{\text{set}}{=} 0 \end{aligned}$$

Figure 5.1: Strict and/or Global Extrema



(a) Global not Strict (b) Global and Strict (c) Strict not Global

solve the above equations to get $(x, y, z) = (\frac{4}{3}, 1, -\frac{2}{3})$. The Hessian matrix is

$$H = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 6 & 0 \\ 1 & 0 & 2 \end{bmatrix}$$

and the corresponding leading principle minors are

$$\begin{aligned} \Delta_1 &= 2 > 0 \\ \Delta_2 &= \begin{vmatrix} 2 & 0 \\ 0 & 6 \end{vmatrix} = 12 > 0 \\ \Delta_3 &= |H| = 18 > 0 \end{aligned}$$

Thus, H is PD, and therefore f is strictly convex. From the previous theorem, we can conclude that the point $(x, y, z) = (\frac{4}{3}, 1, -\frac{2}{3})$ is the unique global minimizer (and a strict local minimizer).

5.2 Constrained Optimization I: Equality Constraints

Consider the problem of maximizing a function $f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n)$ of n variables constrained by $m < n$ equality constraints. Let the functions $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_m(\mathbf{x})$ define the constraint set. Thus, our problem is to

$$\begin{aligned} \text{maximize or minimize} \quad & f(x_1, x_2, \ldots, x_n) \\ \text{subject to} \quad & g_1(x_1, x_2, \ldots, x_n) = 0 \\ & g_2(x_1, x_2, \ldots, x_n) = 0 \\ & \vdots \\ & g_m(x_1, x_2, \ldots, x_n) = 0 \end{aligned}$$

or, equivalently, in a more compact form

$$\begin{aligned} &\text{maximize or minimize } f(\mathbf{x}) \\ &\text{subject to } g(\mathbf{x}) = 0, \end{aligned} \tag{5.1}$$

where $g(\mathbf{x}) = 0$ denotes an $m \times 1$ vector of constraints, $m < n$. The condition $m < n$ is needed to ensure a proper degree of freedom. Without this condition, there would not be a way for the variables to adjust toward the optimum.

The most intuitive solution method for problem (5.1) involves the elimination of m variables from the problem by use of the constraint equations, thereby converting the problem into an equivalent unconstrained optimization problem. The actual solution of the constraint equations for m variables in terms of the remaining $n-m$ can often prove a difficult, if not impossible, task. Moreover, the elimination of variables is seldom applicable to economic problems, as economic theory rarely allows for the specification of particular functional forms. Nevertheless, the theory underlying the method of elimination of variables can be used to obtain analytically useful characterizations of solutions to equality constrained problems.

Alternatively, the solution can be obtained using the *Lagrangian* function defined as

$$\begin{aligned}\mathcal{L}(\mathbf{x};\lambda) &= f(\mathbf{x}) - \lambda_1 g_1(\mathbf{x}) - \lambda_2 g_2(\mathbf{x}) - \cdots - \lambda_m g_m(\mathbf{x}) \\ &= f(\mathbf{x}) - \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}),\end{aligned} \tag{5.2}$$

where $\lambda_1, \lambda_2 \ldots, \lambda_m$ multiply the constraints and are known as *Lagrange multipliers*. In order to solve the problem, we find the critical points of the Lagrangian by solving the equations

$$\begin{aligned}&\frac{\partial \mathcal{L}}{\partial x_1} = 0; \frac{\partial \mathcal{L}}{\partial x_2} = 0; \cdots ; \frac{\partial \mathcal{L}}{\partial x_n} = 0; \\ &\frac{\partial \mathcal{L}}{\partial \lambda_1} = 0; \frac{\partial \mathcal{L}}{\partial \lambda_2} = 0; \cdots ; \frac{\partial \mathcal{L}}{\partial \lambda_m} = 0,\end{aligned} \tag{5.3}$$

which represent $n+m$ equations for the $n+m$ variables $x_1, x_2, \ldots, x_n, \lambda_1, \lambda_2 \ldots, \lambda_m$. Thus, we have transformed what was a constrained problem of n variables into an unconstrained problem of $n+m$ variables. Note that since $\lambda_1, \lambda_2 \ldots, \lambda_m$ simply multiply the constraints, $\frac{\partial \mathcal{L}}{\partial \lambda_i}$, for $i = 1, 2, \ldots, m$, is equivalent to each multipliers' respective constraint. Thus, the system of equations, (5.3), can be written more compactly as

$$\begin{aligned}\nabla f(\mathbf{x}) &= \lambda \nabla g(\mathbf{x}), \\ g(\mathbf{x}) &= 0\end{aligned} \tag{5.4}$$

where λ is a $1 \times m$ vector of Lagrange multipliers and $\nabla g(\mathbf{x})$ is an $m \times n$ Jacobian matrix of the constraint set.

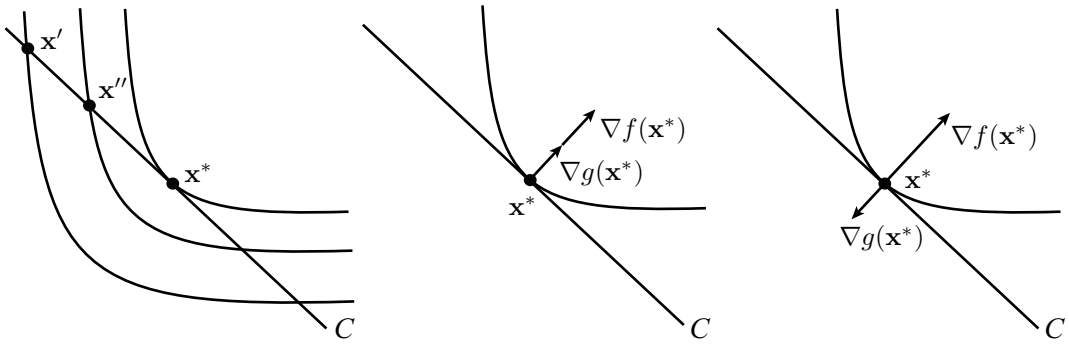
To understand this more clearly, consider the problem of maximizing $f(x_1, x_2)$ subject to $g(x_1, x_2) = 0$. Geometrically, our goal is to find the highest valued level-curve of f, which meets the constraint set C (see figure 5.2). The highest level-curve of f cannot cross the constraint curve C (see point $\mathbf{x}'$); if it did, nearby higher level sets would also cross (see point $\mathbf{x}''$). Thus, the highest level set of f must be tangent to C at the constrained max, $\mathbf{x}^*$.

The gradient vector of the objective function and constraint set is given by

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} \quad \text{and} \quad \nabla g(\mathbf{x}) = \begin{bmatrix} \frac{\partial g}{\partial x_1} \\ \frac{\partial g}{\partial x_2} \end{bmatrix},$$

which are perpendicular to the level sets of f and g. Since the level sets of f and g have the same slope at $\mathbf{x}^*$, the gradient vectors $\nabla f(\mathbf{x})$ and $\nabla g(\mathbf{x})$ must line up at $\mathbf{x}^*$. Thus they point in the same direction or opposite directions (see figure 5.2). In either case, the gradients are scalar multiples of each other. If the corresponding Lagrange multiplier is λ^*, then $\nabla f(\mathbf{x}^*) = \lambda^* \nabla g(\mathbf{x}^*)$ as the Lagrange formulation, given in (5.4), suggests.

Figure 5.2: Constrained Optimization and the Lagrange Principle



Remark 5.2.1. In order for this transformation to remain valid, we must place a restriction on the constraint set known as *constraint qualification*, which requires:

(a) $\nabla g(\mathbf{x}^*) \neq \mathbf{0}$ if the problem defined in (5.1) has only one constraint; and

(b) the rank of the Jacobian matrix

$$J(\mathbf{x}^*) = \begin{bmatrix} \nabla g_1(\mathbf{x}^*) \\ \nabla g_2(\mathbf{x}^*) \\ \vdots \\ \nabla g_m(\mathbf{x}^*) \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1(\mathbf{x}^*)}{\partial x_1} & \cdots & \frac{\partial g_1(\mathbf{x}^*)}{\partial x_n} \\ \frac{\partial g_2(\mathbf{x}^*)}{\partial x_1} & \cdots & \frac{\partial g_2(\mathbf{x}^*)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m(\mathbf{x}^*)}{\partial x_1} & \cdots & \frac{\partial g_m(\mathbf{x}^*)}{\partial x_n} \end{bmatrix}$$

equals m (full rank) if the problem defined in (5.1) has m constraints, $m > 1$.

Example 5.2.1. This example illustrates why the rank m condition is required for the transformation given in (5.2). Suppose our problem is to

$$\begin{aligned} \text{maximize} \quad & f(x_1, x_2, x_3) = x_1 \\ \text{subject to} \quad & g_1(x_1, x_2, x_3) = (x_1 - 1)^2 - x_3 = -1 \\ & g_2(x_1, x_2, x_3) = (x_1 - 1)^2 + x_3 = 1. \end{aligned}$$

Then the gradient vectors are

$$\nabla f = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \nabla g_1 = \begin{bmatrix} 2(x_1 - 1) \\ 0 \\ -1 \end{bmatrix}, \quad \text{and} \quad \nabla g_2 = \begin{bmatrix} 2(x_1 - 1) \\ 0 \\ 1 \end{bmatrix}.$$

The set of points satisfying both constraints is $\{(1, y, 1) | y \in \mathbb{R}\}$. If the transformation in (5.2) is valid, (5.4) implies

$$(1, 0, 0) = \lambda_1(0, 0, -1) + \lambda_2(0, 0, 1),$$

which is not possible since the gradient vectors are linearly dependent. The problem here is that the transformation is not valid, since constraint qualification is not satisfied ($\text{rank } J(\mathbf{x}^*) = 1 < m$).

Remark 5.2.2. Constraint qualification says that for transformation (5.2) to be valid, no point satisfying the constraint set can be a critical point of the constraint set. This means that if the constraint set is linear, constraint qualification will automatically be satisfied.

Theorem 5.2.1 (Necessary and Sufficient Conditions for an Extremum)**.** *Let $f, g_1, g_2, \ldots, g_k$ be C^2 real-valued functions on $\mathbb{R}^n$. Consider the problem of maximizing (minimizing) f on the constraint set $g(\mathbf{x}) = 0$, where $g(\mathbf{x}) = 0$ denotes an $m \times 1$ vector of constraints, $m < n$. Then*

(a) (necessary condition)

$$\nabla \mathcal{L}(\mathbf{x}^*, \lambda^*) = 0$$

(b) (Sufficient Condition) If there exist vectors $\mathbf{x}^* \in \mathbb{R}^n, \lambda^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_m^*) \in \mathbb{R}^m$ such that

$$\nabla \mathcal{L}(\mathbf{x}^*, \lambda^*) = 0$$

and for every non-zero vector $\mathbf{z} \in \mathbb{R}^n$ satisfying

$$\nabla g_i(\mathbf{x}^*) \cdot \mathbf{z} = 0, \quad i = 1, 2, \ldots, m$$

it follows that

$$\mathbf{z}^T \nabla_x^2 \mathcal{L}(\mathbf{x}^*, \lambda^*) \mathbf{z} < (>) 0, \qquad \text{(Hessian is negative (positive) definite)}$$

then f has a strict local maximum (minimum) at $\mathbf{x}^*$ □

Conveniently, these conditions for a maximum or minimum can be stated in terms of the Hessian of the Lagrangian function, which turns out to be a bordered Hessian. The following rules work with the bordered Hessian of a constrained optimization problem of the form:

$$\overline{H} = \begin{bmatrix} 0 & \vdots & B^T \\ \cdots & \cdots & \cdots \\ B & \vdots & A \end{bmatrix} = \begin{bmatrix} \mathcal{L}_{\lambda_1\lambda_1} & \cdots & \mathcal{L}_{\lambda_1\lambda_m} & \vdots & \mathcal{L}_{\lambda_1 x_1} & \cdots & \mathcal{L}_{\lambda_1 x_n} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathcal{L}_{\lambda_m\lambda_1} & \cdots & \mathcal{L}_{\lambda_m\lambda_m} & \vdots & \mathcal{L}_{\lambda_m x_1} & \cdots & \mathcal{L}_{\lambda_m x_n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathcal{L}_{x_1\lambda_1} & \cdots & \mathcal{L}_{x_1\lambda_m} & \vdots & \mathcal{L}_{x_1 x_1} & \cdots & \mathcal{L}_{x_1 x_n} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathcal{L}_{x_n\lambda_1} & \cdots & \mathcal{L}_{x_n\lambda_m} & \vdots & \mathcal{L}_{x_n x_1} & \cdots & \mathcal{L}_{x_n x_n} \end{bmatrix}$$

$$= \begin{bmatrix} 0 & \cdots & 0 & \vdots & \partial g_1/\partial x_1 & \cdots & \partial g_1/\partial x_n \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \vdots & \partial g_m/\partial x_1 & \cdots & \partial g_m/\partial x_n \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \partial g_1/\partial x_1 & \cdots & \partial g_m/\partial x_1 & \vdots & \mathcal{L}_{x_1 x_1} & \cdots & \mathcal{L}_{x_1 x_n} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \partial g_1/\partial x_n & \cdots & \partial g_m/\partial x_n & \vdots & \mathcal{L}_{x_n x_1} & \cdots & \mathcal{L}_{x_n x_n} \end{bmatrix}.$$

Criterion 5.2.1 (Sufficient Conditions for strict local Maximum with constraints)**.** *Let f and constraints $g_1, g_2, \ldots, g_m$ be twice continuously differentiable real-valued functions. If there exist vectors $\mathbf{x}^* \in \mathbb{R}^n, \lambda^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_m^*) \in \mathbb{R}^m$ such that*

$$\nabla \mathcal{L}(\mathbf{x}^*, \lambda^*) = 0$$

and if $\overline{H}$ is negative definite on the constraint set, which is the case if $(-1)^k \Delta_{m+k} > 0$ for $k = m+1, ..., n$, where m is the number of constraints that hold with equality, n is the number of endogenous variables, and Δ_k is the leading principle minor of order k (Note: Lagrange multipliers do not count as endogenous variables), then f has a strict local maximum at $\mathbf{x}^*$.

Criterion 5.2.2 (Sufficient Conditions for strict local Minimum with constraints)**.** *Let f and constraints $g_1, g_2, \ldots, g_m$ be twice continuously differentiable real-valued functions. If there exist vectors $\mathbf{x}^* \in \mathbb{R}^n, \lambda^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_m^*) \in \mathbb{R}^m$ such that*

$$\nabla \mathcal{L}(\mathbf{x}^*, \lambda^*) = 0$$

and if $\overline{H}$ is positive definite on the constraint set, which is the case if $(-1)^m \Delta_{m+k} > 0$ for $k = m+1, ..., n$, then f has a strict local minimum at $\mathbf{x}^*$.

Remark 5.2.3. In short, we must check $n - m$ leading principle minors starting with the principle minor of highest order and working backwards. For example, if a problem contains 5 variables and 3 constraints, it will be necessary to check the signs of two principle minors: Δ_7 and Δ_8.

Example 5.2.2 (Minimizing Cost subject to an Output Constraint)**.** Consider a production function given by

$$y = 20x_1 - x_1^2 + 15x_2 - x_2^2.$$

Let the prices of x_1 and x_2 be 10 and 5 respectively with an output constraint of 55. Then to minimize the cost of producing 55 units of output given these prices, we set up the following Lagrangian

$$\mathcal{L}(x_1, x_2, \lambda) = (10x_1 + 5x_2) - \lambda(20x_1 - x_1^2 + 15x_2 - x_2^2 - 55),$$

which has first order conditions

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial x_1} &= 10 - \lambda(20 - 2x_1) \stackrel{\text{set}}{=} 0 \\
\frac{\partial \mathcal{L}}{\partial x_2} &= 5 - \lambda(15 - 2x_2) \stackrel{\text{set}}{=} 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= 20x_1 - x_1^2 + 15x_2 - x_2^2 - 55 \stackrel{\text{set}}{=} 0.
\end{aligned}$$

If we take the ratio of the first two first order conditions, we obtain

$$\begin{aligned}
2 = \frac{20 - 2x_1}{15 - 2x_2} \quad &\rightarrow \quad 30 - 4x_2 = 20 - 2x_1 \\
\rightarrow \quad 10 - 4x_2 = -2x_1 \quad &\rightarrow \quad x_1 = 2x_2 - 5.
\end{aligned}$$

Now plug this into the last first order condition to obtain

$$20(2x_2 - 5) - (2x_2 - 5)^2 + 15x_2 - x_2^2 - 55 = 0.$$

Multiplying out and solving for x_2 will give

$$\begin{aligned} 40x_2 - 100 - 4x_2^2 + 20x_2 - 25 + 15x_2 - x_2^2 - 55 &= 0 \\ \rightarrow x_2^2 - 15x_2 + 36 &= 0 \\ \rightarrow (x_2 - 12)(x_2 - 3) &= 0. \end{aligned}$$

Therefore, we have two potential solutions $(x_1, x_2) = (19, 12)$ and $(x_1, x_2) = (1, 3)$. The Lagrange multiplier λ is obtained by plugging the solutions into the above first order conditions to obtain

$$\begin{aligned} 10 - \lambda(20 - 2(19)) = 0 \quad &\rightarrow \quad \lambda = -\frac{5}{9} \\ 10 - \lambda(20 - 2(1)) = 0 \quad &\rightarrow \quad \lambda = \frac{5}{9}. \end{aligned}$$

Given that $\nabla y(19, 12) = (-18, -9) \neq 0$ and $\nabla y(1, 3) = (18, 9) \neq 0$, constraint qualification holds. To check for a maximum or minimum, we set up the bordered Hessian. Consider first the point $(19, 12, -5/9)$. The bordered Hessian in this case is

$$\begin{aligned} \overline{H} &= \begin{bmatrix} \frac{\partial^2 L(\mathbf{x}^*, \lambda^*)}{\partial x_1^2} & \frac{\partial^2 L(\mathbf{x}^*, \lambda^*)}{\partial x_1 \partial x_2} & \frac{\partial g(x^*)}{\partial x_1} \\ \frac{\partial^2 L(\mathbf{x}^*, \lambda^*)}{\partial x_2 \partial x_1} & \frac{\partial^2 L(\mathbf{x}^*, \lambda^*)}{\partial x_2^2} & \frac{\partial g(x^*)}{\partial x_2} \\ \frac{\partial g(x^*)}{\partial x_1} & \frac{\partial g(x^*)}{\partial x_2} & 0 \end{bmatrix} = \begin{bmatrix} 2\lambda & 0 & 20 - 2x_1 \\ 0 & 2\lambda & 15 - 2x_2 \\ 20 - 2x_1 & 15 - 2x_2 & 0 \end{bmatrix} \\ &= \begin{bmatrix} -\frac{10}{9} & 0 & -18 \\ 0 & -\frac{10}{9} & -9 \\ -18 & -9 & 0 \end{bmatrix}. \end{aligned}$$

Since we have only two endogenous variables $(n = 2)$ and one constraint $(m = 1)$, it is sufficient to check only the principle minor of highest magnitude (Δ_3 or, more precisely, the determinant of the bordered Hessian) in order to determine definiteness.

$$\begin{aligned} |\overline{H}| &= (-1)^2 \left(-\frac{10}{9}\right) \begin{vmatrix} -\frac{10}{9} & -9 \\ -9 & 0 \end{vmatrix} + (-1)^4(-18) \begin{vmatrix} 0 & -\frac{10}{9} \\ -18 & -9 \end{vmatrix} \\ &= -\frac{10}{9}(-81) + (-18)(-20) \\ &= 450. \end{aligned}$$

Since $k = 2$, $(-1)^2\Delta_3 = (-1)^2(450) > 0$. Therefore, $\overline{H}$ is negative definite on the constraint set, and thus this point is a strict local maximum.

Now consider the other point, $(1, 3, 5/9)$. The bordered Hessian is given by

$$\begin{bmatrix} 2\lambda & 0 & 20 - 2x_1 \\ 0 & 2\lambda & 15 - 2x_2 \\ 20 - 2x_1 & 15 - 2x_2 & 0 \end{bmatrix} = \begin{bmatrix} \frac{10}{9} & 0 & 18 \\ 0 & \frac{10}{9} & 9 \\ 18 & 9 & 0 \end{bmatrix}.$$

Again, it is sufficient to check only the determinant of the bordered Hessian in order to determine definiteness. In this case, $\det \overline{H} = -450$ and $(-1)\Delta_3 = (-1)(-450) > 0$. Therefore, $\overline{H}$ is positive definite on the constraint set, and thus this point is a strict local minimum. The minimum cost is obtained by substituting this point into the cost expression (objective function) to obtain

$$C = 10(1) + 5(3) = 25.$$

Example 5.2.3. Consider the problem of maximizing $x^2y^2z^2$ subject to the constraint $g(x,y,z) = x^2+y^2+z^2 = 3$. The Lagrangian function is given by

$$\mathcal{L} = x^2y^2z^2 + \lambda(3 - x^2 - y^2 - z^2)$$

and the corresponding first order conditions are

$$\frac{\partial \mathcal{L}}{\partial x} = 2xy^2z^2 - 2\lambda x \stackrel{set}{=} 0 \qquad \frac{\partial \mathcal{L}}{\partial y} = 2x^2yz^2 - 2\lambda y \stackrel{set}{=} 0$$
$$\frac{\partial \mathcal{L}}{\partial z} = 2x^2y^2z - 2\lambda z \stackrel{set}{=} 0 \qquad \frac{\partial \mathcal{L}}{\partial \lambda} = x^2+y^2+z^2-3 \stackrel{set}{=} 0,$$

with solution $x^2 = y^2 = z^2 = \lambda = 1$. Since $\nabla g(\pm 1, \pm 1, \pm 1) = (\pm 2, \pm 2, \pm 2) \neq (0,0,0)$, constraint qualification holds. At $x = y = z = \lambda = 1$, the bordered Hessian for this problem is

$$\overline{H} = \begin{bmatrix} 0 & 2x & 2y & 2z \\ 2x & 2y^2z^2 - 2\lambda & 4xyz^2 & 4xy^2z \\ 2y & 4xyz^2 & 2x^2z^2 - 2\lambda & 4x^2yz \\ 2z & 4xy^2z & 4x^2yz & 2x^2y^2 - 2\lambda \end{bmatrix} = \begin{bmatrix} 0 & 2 & 2 & 2 \\ 2 & 0 & 4 & 4 \\ 2 & 4 & 0 & 4 \\ 2 & 4 & 4 & 0 \end{bmatrix}.$$

Since $n = 3$ and $m = 1$, we have to check the signs of the two leading principle minors of highest order, Δ_3 and Δ_4. After computation, we find $\Delta_3 = 32$ and $\Delta_4 = -192$. For $k = 2$, $(-1)^2\Delta_3 = (-1)^2(32) > 0$ and for $k = 3$, $(-1)^3\Delta_4 = (-1)^3(-192) > 0$. Therefore, $\overline{H}$ is negative definite on the constraint set, and thus this point is a strict local maximum. By the properties of determinant, the remaining seven critical points also satisfy the sufficiency condition conditions and are classified as local maxima. This is an example of a situation where the solution is globally optimal, but not unique.

5.3 Constrained Optimization II: Non-negative Variables

To map our problem in section 5.2 into one that makes greater economic sense, consider the following problem, which postulates that the variables $x_1, x_2, \ldots, x_j$ are subject to inequality constraints.

$$\begin{aligned} \text{maximize or minimize } & f(x_1, x_2, \ldots, x_n) \\ \text{subject to } & g_1(x_1, x_2, \ldots, x_n) = 0 \\ & g_2(x_1, x_2, \ldots, x_n) = 0 \\ & \vdots \\ & g_m(x_1, x_2, \ldots, x_n) = 0 \\ & x_1, x_2, \ldots, x_j \geq 0 \;\; x_{j+1}, \ldots, x_n > 0. \end{aligned}$$

If the optimum, $\mathbf{x}^*$, happens to be that these requirements are not binding, that is, $x_1, x_2, \ldots, x_j$ are in fact strictly positive, then the procedure outlined in the preceding section for determining optimal points remains unaltered. That is, assigning a Lagrange multiplier $\lambda_i,\ i = 1, 2, \ldots, m$ to each constraint, the Lagrangian function can once again be written as

$$\begin{aligned} L(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m) &= f(x_1, x_2, \ldots, x_n) - \lambda_1 g_1(x_1, \ldots, x_n) - \cdots - \lambda_m g_m(x_1, \ldots, x_n) \\ &= f(x_1, x_2, \ldots, x_n) - \sum_{k=1}^{m} \lambda_k g_k(x_1, \ldots, x_n), \end{aligned}$$

and the first order necessary conditions satisfied at the optimum $\mathbf{x}^*$ are

$$\frac{\partial \mathcal{L}}{\partial x_j} = 0, \qquad j = 1, 2, \ldots, n \tag{5.5}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_k} = 0, \qquad k = 1, 2, \ldots, m, \tag{5.6}$$

so long as constraint qualification is satisfied. However, it is often the case that some components are positive while others are zero. Thus, an equation like (5.5) should hold for the partial derivative of the Lagrangian with respect to every component that is strictly positive and an inequality with respect to every component that is zero. In other words for $x_1, x_2, \ldots, x_n$, we should have

$$\frac{\partial \mathcal{L}}{\partial x_j} \leq 0, \qquad x_j \geq 0, \tag{5.7}$$

with at least one of these holding with equality. The requirement that at least one inequality in (5.7) should hold as an equality is sometimes stated more compactly as

$$x_j \frac{\partial \mathcal{L}}{\partial x_j} = 0.$$

The point is that the product is zero only if at least one of the factors is zero. A pair of inequalities like (5.7), not both of which can be strict, is said to show *Complementary Slackness*, which we will denote "CS". A single inequality, say $x_j \geq 0$, is *binding* if it holds as an equality, that is, if x_j is at the extreme limit of its permitted range; the inequality is said to be *slack* if x_j is positive, meaning it has some room to maneuver before hitting its extreme. Each one of the pair of inequalities in (5.7) therefore complements the slackness of the other; if one is slack the other is binding.

The intuition is as follows: if $x_j^* > 0$, the constraint is not binding and it is possible to adjust x_j until the marginal benefit of further adjustments is zero ($\partial \mathcal{L}/\partial x_j = 0$), given the other constraints. If, on the other hand, $x_j^* = 0$, then the constraint is binding and the marginal benefit of increasing x_j is negative ($\partial \mathcal{L}/\partial x_j \leq 0$). In this case, it is possible that the objective value could be improved if negative x_j values were permitted.

Example 5.3.1. Consider the following problem:

$$\begin{aligned} &\text{Maximize } \ f(x, y) = 1 - 8x + 10y - 2x^2 - 3y^2 + 4xy \\ &\text{subject to } \ x \geq 0 \text{ and } y \geq 0. \end{aligned}$$

Begin by calculating the gradient and Hessian:

$$\nabla f(x, y) = \begin{bmatrix} -8 - 4x + 4y \\ 10 - 6y + 4x \end{bmatrix}, \qquad H(f) = \begin{bmatrix} -4 & 4 \\ 4 & -6 \end{bmatrix}.$$

For all (x, y), the Hessian matrix is negative definite, since $\Delta_1 = -4 < 0$ and $\Delta_2 = 8 > 0$. Hence f is a concave function, and any (x, y) that satisfies complementary slackness is a constrained global maximum. It remains to locate such a point.

Our first pass at the problem is to look for a solution with $x > 0, y > 0$ so that $\nabla f(x, y) = \mathbf{0}$. Equating the gradient to the zero-vector gives

$$x - y = -2 \quad \text{and} \quad 2x - 3y = -5.$$

These equations are satisfied only when $x = -1$ and $y = 1$, which obviously violates the fact that $x > 0$. Thus our problem is not yet solved. However, the calculation just made was not a wasted effort, for we have in fact found the unconstrained maximum. And, since this has $x < 0$, it is likely that $x = 0$ at the constrained maximum. We therefore look for a solution with $x = 0$ and $y > 0$, so that $\partial f/\partial y = 0$. Equating x and $\partial f/\partial y$ to zero we see that $10 - 6y = 0$, so $y = 5/3$. Thus the point $(0, 5/3)$ satisfies the conditions

$$x = 0, \quad y > 0, \quad \partial f/\partial y = 0,$$

and it remains to show that this point satisfies the remaining condition for a constrained maximum, namely $\partial f/\partial x < 0$. At $(0, 5/3)$,

$$\partial f/\partial x = -4(2 + 0 - 5/3) = -4/3 < 0,$$

so the condition is satisfied. Thus the constrained maximum is attained where $x = 0$ and $y = 5/3$; the constrained maximum value of f is therefore $28/3$. This is of course less than the value taken by f at the unconstrained maximum $(-1, 1)$, which is in fact 10.

Example 5.3.2. Define the profit function as

$$\Pi(x, y) = (-80 + 24x + 78y - 3x^2 - 3xy - 4y^2)/10,$$

where x and y are the quantities of two different goods. We wish to choose x and y to maximize $\Pi(x, y)$ subject to the constraints $x \geq 0$ and $y \geq 0$.

In this case, $\partial\Pi/\partial x = \frac{1}{10}(24 - 6x - 3y)$ and $\partial\Pi/\partial y = \frac{1}{10}(78 - 3x - 8y)$. The Hessian Matrix

$$H = \begin{bmatrix} -\frac{3}{5} & -\frac{3}{10} \\ -\frac{3}{10} & -\frac{4}{5} \end{bmatrix}$$

has principle minors $\Delta_1 < 0$ and $\Delta_2 > 0$. Therefore $\Pi(x, y)$ is negative definite. Hence the profit function is concave, and the first order conditions give a global constrained maximum. It is not hard to see that the only values of x and y for which $\partial\Pi/\partial x = \partial\Pi/\partial y = 0$ are $x = -14/13$ and $y = 132/13$, which clearly violates the constraints. We therefore look for a solution (x, y) such that $x = 0$, $y > 0$, $\partial\Pi/\partial x \leq 0$, and $\partial\Pi/\partial y = 0$. The first, second, and fourth conditions are satisfied where $x = 0$ and $y = 9.75$. Since $9.75 > 8$, the third condition is also satisfied. Hence the solution is $x = 0$, $y = 9.75$ and the maximal profit is 30.025.

5.4 Constrained Optimization III: Inequality Constraints

To find the constrained maximum or minimum of a function, we simply constructed the Lagrangian, set its $(m+n)$ first order conditions equal to zero, and then solved these $(m+n)$ equations in $(m+n)$ unknowns. However, the vast majority of constrained optimization problems that arise in economics have their constraints defined by inequalities:

$$g_1(x_1, x_2, \ldots, x_n) \leq 0, \; g_2(x_1, x_2, \ldots, x_n) \leq 0, \; \ldots, \; g_m(x_1, x_2, \ldots, x_n) \leq 0.$$

Unfortunately, the method for finding the constrained maxima in problems with inequality constraints is a bit more complex than the method we used for equality constraints. The first order

conditions involve both equalities and inequalities and their solution eentaiils the investigations of a number of cases. To see this more clearly, consider the following problem:

$$\begin{aligned} \text{maximize} \quad & f(x_1, x_2, \ldots, x_n) \\ \text{subject to} \quad & g_1(x_1, x_2, \ldots, x_n) \leq 0 \\ & g_2(x_1, x_2, \ldots, x_n) \leq 0 \\ & \vdots \\ & g_m(x_1, x_2, \ldots, x_n) \leq 0 \\ & x_1, x_2, \ldots, x_n \geq 0. \end{aligned} \tag{5.8}$$

As an alternative to the procedure outlined in the previous section, we could introduce n new constraints in addition to the m original ones:

$$g_{m+1}(\mathbf{x}) = -x_1 \leq 0, \quad \ldots, \quad g_{m+n}(\mathbf{x}) = -x_n \leq 0. \tag{5.9}$$

Then, if we introduce Lagrange multipliers $\lambda_1, \ldots, \lambda_m$ that are associated with the constraints and $\mu_1, \ldots, \mu_n$ to go with the non-negativity constraints, our Lagrangian function is of the form

$$\mathcal{L}(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) - \sum_{j=1}^{m} \lambda_j g_j(\mathbf{x}) - \sum_{i=1}^{n} \mu_i(-\mathbf{x}_i), \tag{5.10}$$

where $\lambda = \{\lambda_1, \ldots, \lambda_m\}$ and $\mu = \{\mu_1, \ldots, \mu_n\}$. The necessary conditions for $\mathbf{x}^*$ to solve this problem are

$$\frac{\partial f(\mathbf{x}^*)}{\partial x_i} - \sum_{j=1}^{m} \lambda_j \frac{\partial g_j(\mathbf{x}^*)}{\partial x_i} + \mu_i = 0, \qquad i = 1, \ldots, n \tag{i}$$

$$g_j(\mathbf{x}^*) \leq 0, \lambda_j \geq 0 \ (\lambda_j = 0 \text{ if } g_j(\mathbf{x}^*) < 0), \qquad j = 1, \ldots, m \tag{ii}$$

$$x_i \geq 0, \mu_i \geq 0 \ (\mu_i = 0 \text{ if } x_i > 0), \qquad i = 1, \ldots, n. \tag{iii}$$

To reduce this collection of $m + n$ constraints and $m + n$ Lagrange multipliers, the necessary conditions for problem (5.8) are often formatted slightly differently. In fact, it follows from (i) that $\frac{\partial f(\mathbf{x}^*)}{\partial x_i} - \sum_{j=1}^{m} \lambda_j \frac{\partial g_j(\mathbf{x}^*)}{\partial x_i} = -\mu_i$. Since $\mu_i \geq 0$ and $-\mu_i = 0$ if $x_i > 0$, we see that (i) and (iii) together are equivalent to the condition

$$\frac{\partial f(\mathbf{x}^*)}{\partial x_i} - \sum_{j=1}^{m} \lambda_j \frac{\partial g_j(\mathbf{x}^*)}{\partial x_i} \leq 0 \ (= 0 \text{ if } x_i^* > 0), \quad i = 1, \ldots, n.$$

With the possibility of inequality constraints, there are now two kinds of possible solutions: one where the constrained optimum lies in the region where $g_j(\mathbf{x}) < 0$, in which case constraint j is slack, and one where the constrained optimum lies on the boundary $g_j(\mathbf{x}) = 0$, in which case constraint j is binding. In the former case, the function $g_j(\mathbf{x})$ plays no role. $\mathbf{x}^*$ still corresponds to the optimum of the Lagrangian given in (5.10), but this time with $\lambda_j = 0$. The latter case, where the optimum lies on the boundary of each constraint, is analogous to the equality constraint discussed previously and corresponds to the optimum of the Lagrangian with $\lambda_j \neq 0$ for all j. In this case, however, the sign of the Lagrange multiplier is crucial, because the objective function $f(\mathbf{x})$ will only be at a *maximum* if its gradient is oriented away from the region $g(\mathbf{x}) < 0$ (i.e. $\nabla g(\mathbf{x})$ and $\nabla f(\mathbf{x})$ point in the same direction). We therefore have $\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x})$ for $\lambda_j \geq 0$ for all j (if the constraint was written as $g(\mathbf{x}) \geq 0$, the gradient vectors would point in opposite directions and $\nabla f(\mathbf{x}) = -\lambda \nabla g(\mathbf{x})$ for $\lambda_j \geq 0$ for all j).

Remark 5.4.1 (Constraint Qualification)**.** In order for the transformation in (5.10) to be valid, the gradient vectors $\nabla g_j(\mathbf{x}^*)$ $(j = 1, \ldots, m)$ corresponding to those constraints that are binding at $\mathbf{x}^*$ must be linearly independent. In other words, the corresponding Jacobian matrix must be full rank.

Remark 5.4.2. When solving optimization problems subject to inequality constraints, it is helpful to map the problem into the standard form given in (5.8). If the problem is one of minimizing $f(\mathbf{x})$, the equivalent problem of maximizing $-f(\mathbf{x})$ should be solved. Also, all inequality constraints should be written as $g_j(\mathbf{x}) \leq 0$ (i.e. if the original constraint was $r_j(\mathbf{x}) \leq b_j$, then $g_j(\mathbf{x}) = r_j(\mathbf{x}) - b_j$, while if the original was $r_j(\mathbf{x}) \geq b_j$, then $g_j(\mathbf{x}) = b_j - r_j(\mathbf{x})$).

Theorem 5.4.1 (Kuhn-Tucker Necessary Conditions)**.** *Suppose that $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$ solves (5.8). Suppose further that the constraint qualification is satisfied. The there exist unique numbers $\lambda_1^*, \ldots, \lambda_m^*$ such that*

(a) $\frac{\partial f(\mathbf{x}^*)}{\partial x_i} - \sum_{j=1}^m \lambda_j \frac{\partial g_j(\mathbf{x}^*)}{\partial x_i} \leq 0 \ (= 0 \text{ if } x_j^* > 0), \quad i = 1, \ldots, n$

(b) $g_j(\mathbf{x}^*) \leq 0, \lambda_j \geq 0 \ (= 0 \text{ if } g_j(\mathbf{x}^*) < 0), \quad j = 1, \ldots, m.$

Theorem 5.4.2 (Kuhn-Tucker Sufficient Conditions)**.** *Consider problem (5.8) and suppose that $\mathbf{x}^*$ and $\lambda_1^*, \ldots \lambda_m^*$ satisfy conditions (a) and (b) in theorem (5.4.1). If the Lagrangian $\mathcal{L} = f(\mathbf{x}^*) - \sum_{j=1}^m \lambda_j^* g_j(\mathbf{x}^*)$ is concave, then $\mathbf{x}^*$ is optimal.*

Note that in this formulation of the necessary/sufficient conditions we use the ordinary Lagrangian, not the extended Lagrangian used earlier. The exhaustive procedure for finding a solution using this theorem involves searching among all 2^{m+n} patterns that are possible from the $(m + n)$ complementary slackness conditions. Fortunately, short-cuts are usually available.

Theorem 5.4.3 (Conditions for Globality)**.**

(a) If the feasible set is compact and the objective function is continuous, then the best of the local solutions is the global solution.

(b) If the feasible set is convex and the objective function is concave, then any point satisfying the first-order conditions is a global maximizer. If the feasible set is convex and the objective function is strictly concave, then any point satisfying the first-order conditions is the unique global maximizer. (Similar conclusions hold for convex objective functions and minimizers.)

(c) If the feasible set is convex and the objective function is quasi-concave, then any point satisfying the first-order conditions [with $\nabla f \neq \mathbf{0}$] is a global maximizer. If, in addition, the feasible set is strictly convex or the objective function is strictly quasi-concave, then any point satisfying the first-order conditions (with $\nabla f \neq \mathbf{0}$) is the unique global maximizer. [To see why we need $\nabla f \neq \mathbf{0}$, consider the problem of maximizing $f(x, y) = xy$ subject to $x \geq 0, y \geq 0$, and $x + y \leq 2$. The feasible set is convex and f is quasi-concave on $\mathbb{R}^2_+$. The first-order conditions hold at (0,0) with $\nabla f(0, 0) = \mathbf{0}$, but (0,0) is clearly not even a local maximizer.]

Example 5.4.1. Now let us apply the above rules to the following maximization problem

$$\begin{aligned} \max \ & xy \\ \text{subject to } & x + y \geq -1 \\ & x + y \leq 2 \end{aligned}$$

The associated Lagrangian function is

$$\mathcal{L}(x, y, \lambda, \mu) = xy + \lambda(x + y + 1) + \mu(2 - x - y).$$

and the first order conditions are as follows:

$$\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial x} = y + \lambda - \mu \overset{\text{set}}{=} 0 \\
&\frac{\partial \mathcal{L}}{\partial y} = x + \lambda - \mu \overset{\text{set}}{=} 0 \\
&\frac{\partial \mathcal{L}}{\partial \lambda} = x + y + 1 \geq 0 \qquad \lambda \geq 0 \qquad \text{with "CS"} \\
&\frac{\partial \mathcal{L}}{\partial \mu} = 2 - x - y \geq 0 \qquad \mu \geq 0 \qquad \text{with "CS"}
\end{aligned}$$

The bordered Hessian is

$$\overline{H} = \begin{bmatrix} \mathcal{L}_{\lambda\lambda} & \mathcal{L}_{\lambda\mu} & \mathcal{L}_{\lambda x} & \mathcal{L}_{\lambda y} \\ \mathcal{L}_{\mu\lambda} & \mathcal{L}_{\mu\mu} & \mathcal{L}_{\mu x} & \mathcal{L}_{\mu y} \\ \mathcal{L}_{x\lambda} & \mathcal{L}_{x\mu} & \mathcal{L}_{xx} & \mathcal{L}_{xy} \\ \mathcal{L}_{y\lambda} & \mathcal{L}_{y\mu} & \mathcal{L}_{yx} & \mathcal{L}_{yy} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & -1 & -1 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 1 & 0 \end{bmatrix}.$$

Then we have these solutions for the following cases:

Case 1: $\lambda = 0 = \mu$ results in $s_1^* = (0, 0, 0, 0)$. Thus, we no longer have any binding constraints, that is, they drop out of the Lagrangian. Therefore, $m = 0$ and $n = 2$ and we must check the two leading principle minors of highest magnitude. With the constraint dropping out, the bordered Hessian becomes

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

We then see that for $k = 2$, $(-1)^2\Delta_2 = -1 < 0$, which violates the rule for negative definiteness. Thus, this point is not a local maximum.

Case 2: $\lambda = 0, \mu > 0$ results in $s_2^* = (1, 1, 0, 1)$. Then we have one binding constraint, so that $m = 1$ and $n = 2$. Thus, we must check only the last leading principle minor. The bordered Hessian is

$$\overline{H} = \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix}.$$

Since $(-1)^2\Delta_3 = \det \overline{H} = 2 > 0$, this point is a local maximum.

Case 3: $\lambda > 0, \mu = 0$ results in $s_3^* = (-1/2, -1/2, 1/2, 0)$. Again, we have one binding constraint, so that $m = 1$ and $n = 2$, and we must check only the last leading principle minor. The bordered Hessian is

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

Since $(-1)^2\Delta_3 = \det \overline{H} = 2 > 0$, this point is a local maximum.

Case 4: $\lambda > 0, \mu > 0$ results in no solution.

For all $x, y \in \mathbb{R}$, the feasible set is not compact. However, for $x, y > 0$ or $x, y < 0$, it is, which is the relevant set since $xy < 0$ when x and y have opposite signs. Comparing the output for $f(1,1)$ and $f(-1/2, -1/2)$ and noting that the feasible set is compact when x and y have the same sign implies that $(x^*, y^*) = (1,1)$ is the unique global maximum.

Example 5.4.2. Consider the following problem:

$$\begin{aligned} &\min\ x^2 + 2y^2 + 3z^2, \\ &\text{subject to}\ 3x + 2y + z \geq 17. \end{aligned}$$

The Lagrangian function is (note the negative sign, so we in fact minimize once we find the maximizer of the negative of the objective function):

$$\mathcal{L} = -(x^2 + 2y^2 + 3z^2) + \lambda(3x + 2y + z - 17).$$

The first order conditions are:

$$\frac{\partial \mathcal{L}}{\partial x} = -2x + 3\lambda \overset{\text{set}}{=} 0, \tag{i}$$

$$\frac{\partial \mathcal{L}}{\partial y} = -4y + 2\lambda \overset{\text{set}}{=} 0, \tag{ii}$$

$$\frac{\partial \mathcal{L}}{\partial z} = -6z + \lambda \overset{\text{set}}{=} 0, \tag{iii}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 3x + 2y + z - 17 \geq 0, \quad \lambda \geq 0 \qquad \text{with "CS".} \tag{iv}$$

Case 1 ($\lambda = 0$): From (i)-(iii) $x = 0, y = 0, z = 0$, which violates (iv).

Case 2 ($\lambda > 0$): From (iv), $3x + 2y + z - 17 = 0$. Solving (i)-(iii) along with this equation, results in $(x, y, z, \lambda) = (9/2, 3/2, 1/2, 3)$. Now check the second order conditions. The bordered Hessian of the Lagrangian is

$$\overline{H} = \begin{bmatrix} \mathcal{L}_{\lambda\lambda} & \mathcal{L}_{\lambda x} & \mathcal{L}_{\lambda y} & \mathcal{L}_{\lambda z} \\ \mathcal{L}_{x\lambda} & \mathcal{L}_{xx} & \mathcal{L}_{xy} & \mathcal{L}_{xz} \\ \mathcal{L}_{y\lambda} & \mathcal{L}_{yx} & \mathcal{L}_{yy} & \mathcal{L}_{yz} \\ \mathcal{L}_{z\lambda} & \mathcal{L}_{zx} & \mathcal{L}_{zy} & \mathcal{L}_{zz} \end{bmatrix} = \begin{bmatrix} 0 & 3 & 2 & 1 \\ 3 & -2 & 0 & 0 \\ 2 & 0 & -4 & 0 \\ 1 & 0 & 0 & -6 \end{bmatrix}.$$

In this case, $n = 3$ and $m = 1$, and we must check whether $(-1)^2\Delta_3 > 0$ and $(-1)^3\Delta_4 > 0$. Since $(-1)^2\Delta_{1+2} = 44 > 0$ and $(-1)^3\Delta_{1+3} = 272 > 0$, we have found a local maximum.

The original objective function is strictly convex (positive definite), and the feasible set is convex (linear), so $(x^*, y^*) = (9/2, 3/2)$ is the unique global minimizer.

Example 5.4.3. Find all local maximizers for the function $f(x, y) = -\sqrt{x + y}$ subject to $x \geq 0$ and $x^2 y \geq 108$. Then find all global maximizers for the problem or show none exist.

Solution: The Lagrangian function is

$$\mathcal{L} = -\sqrt{x + y} - \lambda(108 - x^2 y)$$

and the associated first order conditions are as follows

$$\mathcal{L}_x = -1/2(x + y)^{-1/2} + 2\lambda xy \leq 0, \qquad x \geq 0 \qquad \text{with "CS"}$$

$$\mathcal{L}_y = -1/2(x+y)^{-1/2} + \lambda x^2 = 0$$
$$\mathcal{L}_\lambda = x^2 y - 108 \geq 0, \qquad \lambda \geq 0 \qquad \text{with "CS"}$$

Both $\lambda = 0$ and $x = 0$ are inconsistent with $\mathcal{L}_y = 0$. Thus, $\lambda > 0$ and $x > 0$ is the only possible case, and the unique potential solution is $(6, 3, 1/216)$. The associated bordered Hessian is

$$\overline{H} = \begin{bmatrix} 2\lambda y + (x+y)^{-3/2}/4 & 2\lambda x + (x+y)^{-3/2}/4 & 2xy \\ 2\lambda x + (x+y)^{-3/2}/4 & (x+y)^{-3/2}/4 & x^2 \\ 2xy & x^2 & 0 \end{bmatrix} = \begin{bmatrix} 1/27 & 7/108 & 36 \\ 7/108 & 1/108 & 36 \\ 36 & 36 & 0 \end{bmatrix}.$$

In this case, $n - m = 1$, so we only have one condition to check: $(-1)^2\Delta_3 = 108 > 0$. Thus, $(x^*, y^*) = (6, 3)$ is a strict local maximizer and the only local maximizer. In order to assess globality, first note that both x and y must be strictly positive to be feasible, so the feasible set turns out to be the subset of $\mathbb{R}^2_{++}$, where $g(x, y) = x^2 y \geq 108$. Checking the bordered Hessian for g, we find g is strictly quasi-concave on $\mathbb{R}^2_{++}$. Thus, the feasible set, an upper contour set for g, is convex. The objective function is quasi-concave: for any $c \leq 0$, the set of (x, y) such that $-\sqrt{x+y} \geq c$ is $\{(x, y) \in \mathbb{R}^2 | 0 \leq x + y \leq c^2\}$, which is convex (for $c > 0$ the set is empty). With a quasi-concave objective function and a convex feasible set, the unique local maximizer is the unique global maximizer.

Chapter 6

Comparative Statics

In many economic problems we need to know how an optimal solution or an equilibrium solution changes when a parameter in the problem changes. For example, how does the utility-maximizing bundle for a competitive consumer change when a price changes, or how does a market equilibrium price change when a tax on the good changes? These are examples of comparative statics questions. In each case we are interested in how changes in exogenous variables (the parameters determined outside the model) affect the endogenous variables (those determined within the model).

For the consumer choice problem, the endogenous variables are the quantities demanded (chosen by the consumer), while the exogenous variables are prices (outside the control of the competitive consumer). For the market example, the endogenous variable is the market equilibrium price (determined by supply and demand in the market), while the exogenous variable is the tax rate (determined outside the market in some political process). In this section, you will find two extremely helpful tools for evaluating comparative statics.

6.1 Cramer's Rule

Cramer's Rule provides a recipe for solving linear algebraic equations in terms of determinants. Denote the simultaneous equations by

$$A\mathbf{x} = \mathbf{y}, \tag{6.1}$$

where A is a given $n \times n$ matrix and $\mathbf{y}$ is a given $n \times 1$ vector of unknowns.

The explicit solutions of the components $x_1, x_2, \ldots, x_n$ of $\mathbf{x}$ in terms of determinants are

$$x_1 = \frac{\begin{vmatrix} y_1 & a_{12} & a_{13} & \ldots & a_{1n} \\ y_2 & a_{22} & a_{23} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_n & a_{n2} & a_{n3} & \ldots & a_{nn} \end{vmatrix}}{|A|}, \qquad x_2 = \frac{\begin{vmatrix} a_{11} & y_1 & a_{13} & \ldots & a_{1n} \\ a_{22} & y_2 & a_{23} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n2} & y_n & a_{n3} & \ldots & a_{nn} \end{vmatrix}}{|A|}, \ldots \tag{6.2}$$

Theorem 6.1.1 (Cramer's Rule)**.** *Let A be a nonsingular matrix. Then the unique solution $\mathbf{x} = (x_1, \ldots, x_n)$ of the $n \times n$ system $A\mathbf{x} = \mathbf{y}$ is*

$$x_i = \frac{\det B_i}{\det A}, \qquad \text{for } i = 1, \ldots, n,$$

where B_i is the matrix A with the right-hand side $\mathbf{y}$ replacing the ith column of A.

Example 6.1.1. Solve the following 3×3 linear system:

$$\begin{bmatrix} 5 & 2 & 1 \\ 3 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 8 \\ 5 \\ 3 \end{bmatrix}.$$

Using Cramer's rule:

$$x_1 = \frac{\begin{vmatrix} 8 & 2 & 1 \\ 5 & 2 & 0 \\ 3 & 0 & 2 \end{vmatrix}}{\begin{vmatrix} 5 & 2 & 1 \\ 3 & 2 & 0 \\ 1 & 0 & 2 \end{vmatrix}} = \frac{6}{6} = 1, \quad x_2 = \frac{\begin{vmatrix} 5 & 8 & 1 \\ 3 & 5 & 0 \\ 1 & 3 & 2 \end{vmatrix}}{\begin{vmatrix} 5 & 2 & 1 \\ 3 & 2 & 0 \\ 1 & 0 & 2 \end{vmatrix}} = \frac{6}{6} = 1, \quad x_3 = \frac{\begin{vmatrix} 5 & 2 & 8 \\ 3 & 2 & 5 \\ 1 & 0 & 3 \end{vmatrix}}{\begin{vmatrix} 5 & 2 & 1 \\ 3 & 2 & 0 \\ 1 & 0 & 2 \end{vmatrix}} = \frac{6}{6} = 1.$$

Example 6.1.2. Solve the following 2×2 linear algebraic system:

$$\begin{bmatrix} 2+\beta & -\beta \\ -\beta & 1+\beta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \end{bmatrix}.$$

Using Cramer's rule:

$$x_1 = \frac{\begin{vmatrix} 5 & -\beta \\ 0 & 1+\beta \end{vmatrix}}{\begin{vmatrix} 2+\beta & -\beta \\ -\beta & 1+\beta \end{vmatrix}} = \frac{5+5\beta}{2+3\beta}, \quad x_2 = \frac{\begin{vmatrix} 2+\beta & 5 \\ -\beta & 0 \end{vmatrix}}{\begin{vmatrix} 2+\beta & -\beta \\ -\beta & 1+\beta \end{vmatrix}} = \frac{5\beta}{2+3\beta}.$$

6.2 Implicit Function Theorem

To understand the Implicit Function Theorem (IFT), first consider the simplest case: one equation with one endogenous and one exogenous variable of the form

$$f(x, y) = 0. \tag{6.3}$$

Assuming that f is C^1 and (6.3) defines y as a differentiable function of x, implicit differentiation yields

$$\frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy = 0.$$

If $\partial f/\partial y \neq 0$, then

$$\frac{dy}{dx} = -\left(\frac{\partial f}{\partial x}\right) \Big/ \left(\frac{\partial f}{\partial y}\right).$$

Now, carry out this computation more generally for the implicit function $G(x, y) = c$ around the specific point $x = x_0, y = y_0$. We suppose that there is a C^1 solution $y = y(x)$ to the equation $G(x, y) = c$, that is,

$$G(x, y(x)) = c. \tag{6.4}$$

We will use the Chain Rule (section 1.2.5) to differentiate (6.4) with respect to x at x_0:

$$\frac{\partial G}{\partial x}(x_0, y(x_0)) \cdot \frac{dx}{dx} + \frac{\partial G}{\partial y}(x_0, y(x_0)) \cdot \frac{dy}{dx}(x_0) = 0,$$

$$\frac{\partial G}{\partial x}(x_0, y_0) + \frac{\partial G}{\partial y}(x_0, y_0) \cdot y'(x_0) = 0.$$

Solving for $y'(x_0)$ yields

$$y'(x_0) = \frac{dy(x_0)}{dx} = -\frac{\frac{\partial G}{\partial x}(x_0, y_0)}{\frac{\partial G}{\partial y}(x_0, y_0)}.$$

Theorem 6.2.1 (Implicit Function Theorem-One Exogenous Variable)**.** *Let $G(x, y)$ be a C^1 function on an ε-neighborhood about (x_0, y_0) in $\mathbb{R}^2$. Suppose that $G(x_0, y_0) = c$ and consider the expression*

$$G(x, y) = c.$$

If $(\partial G/\partial y)(x_0, y_0) \neq 0$, then there exists a C^1 function $y = y(x)$ defined on an interval I about the point x_0 such that:

(a) $G(x, y(x)) \equiv c$ for all x in I,

(b) $y(x_0) = y_0$, and

(c) $y'(x_0) = -\frac{\frac{\partial G}{\partial x}(x_0, y_0)}{\frac{\partial G}{\partial y}(x_0, y_0)}$.

Example 6.2.1. Show that the equation $x^2e^y - 2y + x = 0$ defines y as a function of x in an interval around the point $(-1, 0)$. Find the derivative of this function at $x = -1$.

Solution: Define $f(x, y) = x^2e^y - 2y + x$. Then $f_1(x, y) = 2xe^y + 1$, $f_2(x, y) = x^2e^y - 2$, and f is C^1 everywhere. Also, $f(-1, 0) = 0$ and $f_2(-1, 0) = -1 \neq 0$. By Theorem 6.2.1, the equation defines y as a C^1 function of x in an interval around $(-1, 0)$. Moreover, we have

$$y'(-1, 0) = -\left.\frac{f_1(x, y)}{f_2(x, y)}\right|_{\substack{x=-1 \\ y=0}} = -\left.\frac{2xe^y + 1}{x^2e^y - 2}\right|_{\substack{x=-1 \\ y=0}} = -1.$$

Example 6.2.2. Consider the equation

$$f(x, y) \equiv x^2 - 3xy + y^3 - 7 = 0 \tag{6.5}$$

about the point $(x_0, y_0) = (4, 3)$. Notice that $f(4, 3)$ satisfies (6.5). The first order partials are

$$\frac{\partial f}{\partial x} = 2x - 3y \quad \text{and} \quad \frac{\partial f}{\partial y} = -3x + 3y^2.$$

Since $(\partial f/\partial y)(4, 3) = 15 \neq 0$, Theorem (6.2.1) tells us that (6.5) does indeed define y as a C^1 function of x around $x_0 = 4$, $y_0 = 3$. Furthermore,

$$y'(4, 3) = -\left.\frac{2x - 3y}{3y^2 - 3x}\right|_{\substack{x=4 \\ y=3}} = \frac{1}{15}.$$

6.2.1 Several Exogenous Variables

Now consider a case where there exists one equation with one endogenous and several exogenous variables of the form

$$G(x_1, x_2, \ldots, x_k, y) = c. \tag{6.6}$$

Around a given point $(x_1^*, \ldots, x_k^*, y^*)$, we want to vary $\mathbf{x} = (x_1, \ldots, x_k)$ and then find a y-value which corresponds to each such $(x_1, \ldots, x_k)$. In this case, we say that equation (6.6) defines y as a *implicit function* of $(x_1, \ldots, x_k)$. Once again, given G and $(\mathbf{x}^*, y^*)$, we want to know whether this functional relationship exists and, if it does, how does y change if any of the x_i's change from x_i^*. Since we are working with a function of several variables $(x_1, \ldots, x_k)$, we will hold all but one of the x_i's constant and vary one exogenous variable at a time. However, this puts us right back in the two-variable case that we have been discussing. The natural extension of Theorem (6.2.1) to this setting is the following.

Theorem 6.2.2 (Implicit Function Theorem-Several Exogenous Variables)**.** *Let $G(x_1, \ldots, x_k, y)$ be a C^1 function around the point $(x_1^*, \ldots, x_k^*, y^*)$. Suppose further that $(x_1^*, \ldots, x_k^*, y^*)$ satisfies*

$$\begin{aligned} G(x_1^*, \ldots, x_k^*, y^*) &= c \\ \text{and } \frac{\partial G}{\partial y}(x_1^*, \ldots, x_k^*, y^*) &\neq 0. \end{aligned}$$

Then there exists a C^1 function $y = y(x_1, \ldots, x_k)$ defined on an open neighborhood N about $(x_1^*, \ldots, x_k^*, y^*)$ so that:

(a) $G(x_1, \ldots, x_k, y(x_1, \ldots, x_k)) = c$ for all $x_1, \ldots, x_k \in N$,

(b) $y^* = y(x_1^*, \ldots, x_k^*)$, and

(c) for each index i

$$\frac{\partial y}{\partial x_i}(x_1^*, \ldots, x_k^*) = -\frac{\frac{\partial G}{\partial x_i}(x_1^*, \ldots, x_k^*, y^*)}{\frac{\partial G}{\partial y}(x_1^*, \ldots, x_k^*, y^*)}. \tag{6.7}$$

6.2.2 The General Case

Consider the following nonlinear system of m equations and $m + n$ unknowns defined as

$$\begin{aligned} F_1(y_1, \ldots, y_m, x_1, \ldots, x_n) &= c_1 \\ F_2(y_1, \ldots, y_m, x_1, \ldots, x_n) &= c_2 \\ \vdots \qquad\qquad & \qquad \vdots \\ F_m(y_1, \ldots, y_m, x_1, \ldots, x_n) &= c_m, \end{aligned} \tag{6.8}$$

where $y_1, \ldots, y_m$ are endogenous and $x_1, \ldots, x_n$ are exogenous. Totally differentiating (6.8) the above system of equations about the point $(\mathbf{y}^*, \mathbf{x}^*)$, we obtain

$$\begin{aligned} \frac{\partial F_1}{\partial y_1} dy_1 + \cdots + \frac{\partial F_1}{\partial y_m} dy_m + \frac{\partial F_1}{\partial x_1} dx_1 + \cdots + \frac{\partial F_1}{\partial x_n} dx_n &= 0 \\ \vdots \qquad\qquad & \qquad \vdots \\ \frac{\partial F_m}{\partial y_1} dy_1 + \cdots + \frac{\partial F_m}{\partial y_m} dy_m + \frac{\partial F_m}{\partial x_1} dx_1 + \cdots + \frac{\partial F_m}{\partial x_n} dx_n &= 0, \end{aligned} \tag{6.9}$$

where all the partial derivatives are evaluated at the point $(\mathbf{y}^*, \mathbf{x}^*)$. By the Implicit Function Theorem, the linear system (6.9) can be solved for $dy_1, \ldots, dy_m$ in terms of $dx_1, \ldots, dx_n$ if and only if

the coefficient (Jacobian) matrix of the dy_i's,

$$\frac{\partial(F_1,\ldots,F_m)}{\partial(y_1,\ldots,y_m)} \equiv \begin{pmatrix} \frac{\partial F_1}{\partial y_1} & \cdots & \frac{\partial F_1}{\partial y_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial y_1} & \cdots & \frac{\partial F_m}{\partial y_m} \end{pmatrix} \tag{6.10}$$

is nonsingular at $(\mathbf{y}^*, \mathbf{x}^*)$. Since this system is linear, when the coefficient matrix (6.10) is nonsingular, we can use the inverse of (6.10) to solve the system (6.9) for the dy_i's in terms of the dx_j's and everything else. Thus, in matrix notation we obtain

$$\begin{pmatrix} \frac{\partial F_1}{\partial y_1} & \cdots & \frac{\partial F_1}{\partial y_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial y_1} & \cdots & \frac{\partial F_m}{\partial y_m} \end{pmatrix} \begin{pmatrix} dy_1 \\ \vdots \\ dy_m \end{pmatrix} = - \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial x_1} & \cdots & \frac{\partial F_m}{\partial x_n} \end{pmatrix} \begin{pmatrix} dx_1 \\ \vdots \\ dx_n \end{pmatrix},$$

which implies

$$\begin{pmatrix} dy_1 \\ \vdots \\ dy_m \end{pmatrix} = - \begin{pmatrix} \frac{\partial F_1}{\partial y_1} & \cdots & \frac{\partial F_1}{\partial y_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial y_1} & \cdots & \frac{\partial F_m}{\partial y_m} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n \frac{\partial F_1}{\partial x_i} dx_i \\ \vdots \\ \sum_{i=1}^n \frac{\partial F_m}{\partial x_i} dx_i \end{pmatrix}. \tag{6.11}$$

Since the linear approximation (6.9) of the original system (6.8) is a implicit function of the dy_i's in terms of the dx_j's, the nonlinear system (6.8) defines the y_i's as implicit functions of the x_j's in a neighborhood of $(\mathbf{y}^*, \mathbf{x}^*)$. Furthermore, we can use the linear solution of the dy_i's in terms of the dx_j's, (6.11), to find the derivatives of the y_i's with respect to the x_j's at $(\mathbf{x}^*, \mathbf{y}^*)$. To compute $\partial y_k/\partial x_h$ for some fixed indices h and k, recall that this derivative estimates the effect on y_k of a one unit increase in x_h $(dx_h = 1)$. So, we set all the dx_j's equal to zero in (6.9) or (6.11) except dx_h, and then we solve (6.9) or (6.11) for the corresponding dy_i's. Thus, (6.11)) reduces to

$$\begin{pmatrix} \frac{dy_1}{dx_h} \\ \vdots \\ \frac{dy_m}{dx_h} \end{pmatrix} = - \begin{pmatrix} \frac{\partial F_1}{\partial y_1} & \cdots & \frac{\partial F_1}{\partial y_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial y_1} & \cdots & \frac{\partial F_m}{\partial y_m} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial F_1}{\partial x_h} \\ \vdots \\ \frac{\partial F_m}{\partial x_h} \end{pmatrix} \tag{6.12}$$

Alternatively, we can apply Cramer's rule to (6.9) and compute

$$\frac{dy_k}{dx_h} = - \frac{\det \begin{pmatrix} \frac{\partial F_1}{\partial y_1} & \cdots & \frac{\partial F_1}{\partial x_h} & \cdots & \frac{\partial F_1}{\partial y_m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial y_1} & \cdots & \frac{\partial F_m}{\partial x_h} & \cdots & \frac{\partial F_m}{\partial y_m} \end{pmatrix}}{\det \begin{pmatrix} \frac{\partial F_1}{\partial y_1} & \cdots & \frac{\partial F_1}{\partial y_k} & \cdots & \frac{\partial F_1}{\partial y_m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial y_1} & \cdots & \frac{\partial F_m}{\partial y_k} & \cdots & \frac{\partial F_m}{\partial y_m}. \end{pmatrix}} \tag{6.13}$$

The following theorem—the most general form of the Implicit Function Theorem—summarizes these conclusions.

Theorem 6.2.3. *Let $F_1,\ldots,F_m : \mathbb{R}^{m+n} \to \mathbb{R}^1$ be C^1 functions. Consider the system of equations*

$$\begin{aligned} F_1(y_1,\ldots,y_m,x_1,\ldots,x_n) &= c_1 \\ F_2(y_1,\ldots,y_m,x_1,\ldots,x_n) &= c_2 \\ \vdots \qquad\qquad & \qquad \vdots \\ F_m(y_1,\ldots,y_m,x_1,\ldots,x_n) &= c_m \end{aligned} \tag{6.14}$$

as possibly defining $y_1, \ldots, y_m$ as implicit functions of $x_1, \ldots, x_n$. Suppose that $(\mathbf{y}^*, \mathbf{x}^*)$ is a solution of (6.14). If the determinant of the $m \times m$ matrix

$$\begin{pmatrix} \frac{\partial F_1}{\partial y_1} & \cdots & \frac{\partial F_1}{\partial y_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial y_1} & \cdots & \frac{\partial F_m}{\partial y_m} \end{pmatrix}$$

evaluated at $(\mathbf{y}^*, \mathbf{x}^*)$ is nonzero, then there exist C^1 functions

$$\begin{aligned} y_1 &= f_1(x_1, \ldots, x_n) \\ &\vdots \quad \vdots \\ y_m &= f_m(x_1, \ldots, x_n) \end{aligned}$$

defined on a neighborhood N about $\mathbf{x}^*$ such that

$$\begin{aligned} F_1(f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}), x_1, \ldots, x_n) &= c_1 \\ F_2(f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}), x_1, \ldots, x_n) &= c_1 \\ \vdots \qquad\qquad & \qquad \vdots \\ F_m(f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}), x_1, \ldots, x_n) &= c_1 \end{aligned}$$

for all $\mathbf{x} = (x_1, \ldots, x_n)$ in N and

$$\begin{aligned} y_1^* &= f_1(x_1^*, \ldots, x_n^*) \\ &\vdots \quad \vdots \\ y_m^* &= f_m(x_1^*, \ldots, x_n^*). \end{aligned}$$

Furthermore, one can compute $(\partial f_k/\partial x_h)(\mathbf{y}^*, \mathbf{x}^*) = (\partial y_k/\partial x_h)(\mathbf{y}^*, \mathbf{x}^*)$ by setting $dx_h = 1$ and $dx_j = 0$ for $j \neq h$ in (6.9) and solving the resulting system for dy_k. This can be accomplished:

(a) by inverting the nonsingular matrix (6.10) to obtain the solution (6.12) or

(b) by applying Cramer's rule to (6.9) to obtain the solution (6.13).

Example 6.2.3. Consider the system of equations

$$\begin{aligned} F_1(x, y, a) &\equiv x^2 + axy + y^2 - 1 = 0 \\ F_2(x, y, a) &\equiv x^2 + y^2 - a^2 + 3 = 0 \end{aligned} \tag{6.15}$$

around the point $x = 0$, $y = 1$, $a = 2$. If we change a a little to a' near $a = 2$, can we find an (x', y') near $(0, 1)$ so that (x', y', a') satisfies these two equations? To answer this question, we need to find the Jacobian of (F_1, F_2) (the matrix of partial derivatives with respect to the endogenous variables x and y) at the point $x = 0$, $y = 1$, $a = 2$

$$\det \begin{pmatrix} \frac{\partial F_1}{\partial x} & \frac{\partial F_1}{\partial y} \\ \frac{\partial F_2}{\partial x} & \frac{\partial F_2}{\partial y} \end{pmatrix} (0, 1, 2) = \det \begin{pmatrix} 2 & 2 \\ 0 & 2 \end{pmatrix} = 4 \neq 0.$$

Thus, we can solve system (6.15) for x and y as functions of a near $(0,1,2)$. Furthermore, using Cramer's rule, we obtain

$$\frac{dy}{da} = -\frac{\det\begin{pmatrix} \frac{\partial F_1}{\partial x} & \frac{\partial F_1}{\partial a} \\ \frac{\partial F_2}{\partial x} & \frac{\partial F_2}{\partial a} \end{pmatrix}}{\det\begin{pmatrix} \frac{\partial F_1}{\partial x} & \frac{\partial F_1}{\partial y} \\ \frac{\partial F_2}{\partial x} & \frac{\partial F_2}{\partial y} \end{pmatrix}} = -\frac{\det\begin{pmatrix} 2x+ay & xy \\ 2x & -2a \end{pmatrix}}{\det\begin{pmatrix} 2x+ay & ax+2y \\ 2x & 2y \end{pmatrix}}.$$

Evaluating at $x=0$, $y=1$, $a=2$, gives

$$\frac{dy}{da}(0,1,2) = -\frac{\det\begin{pmatrix} 2 & 0 \\ 0 & -4 \end{pmatrix}}{\det\begin{pmatrix} 2 & 2 \\ 0 & 2 \end{pmatrix}} = \frac{8}{4} = 2 > 0.$$

Therefore, if a increases to 2.1, y will increase to 1.2. Let us now use the method of total differentiation to compute the effect on x. Total differentiating the non-linear system (6.15), we obtain

$$\begin{aligned} (2x+ay)dx + (ax+2y)dy + xy\,da &= 0 \\ 2x\,dx + 2y\,dy - 2a\,da &= 0. \end{aligned}$$

Evaluating at $x=0$, $y=1$, $a=2$:

$$\begin{aligned} 2\,dx + 2\,dy &= 0\,da \\ 0\,dx + 2\,dy &= 4\,da. \end{aligned}$$

Clearly, $dy = 2\,da$ (as we just computed above) and $dx = -dy = -2da$. Thus, if a increases to 2.1, x will decrease to $-.2$.

Example 6.2.4. Consider the following system

$$\begin{aligned} Y &= C + I + G \\ C &= C(Y-T) \\ I &= I(r) \\ M^s &= M(Y,r), \end{aligned} \tag{6.16}$$

where the nonlinear functions $x \mapsto C(x)$, $r \mapsto I(r)$, and $(Y,r) \mapsto M(Y,r)$ satisfy

$$0 < C'(x) < 1, \quad I'(r) < 0, \quad \frac{\partial M}{\partial Y} > 0, \quad \frac{\partial M}{\partial r} < 0. \tag{6.17}$$

System (6.16) can be reduced to

$$\begin{aligned} Y - C(Y-T) - I(r) &= G \\ M(Y,r) &= M^s, \end{aligned}$$

where we have defined Y and r as implicit functions of G, M^s, and T. Suppose that the current (G, M^s, T) is (G^*, M^{s*}, T^*) and that the corresponding (Y,r)-equilibrium is (Y^*, r^*). If we vary

(G, M^s, T) a little, is there a corresponding equilibrium (Y, r) and how does it change? Totally differentiating system (6.16), we obtain

$$\left(1 - \frac{\partial C}{\partial Y}\right) dY - \frac{\partial I}{\partial r} dr = dG - \frac{\partial C}{\partial T} dT$$
$$\frac{\partial M}{\partial Y} dY + \frac{\partial M}{\partial r} dr = dM^s$$

or, in matrix notation,

$$\begin{pmatrix} 1 - \frac{\partial C}{\partial Y} & -\frac{\partial I}{\partial r} \\ \frac{\partial M}{\partial Y} & \frac{\partial M}{\partial r} \end{pmatrix} \begin{pmatrix} dY \\ dr \end{pmatrix} = \begin{pmatrix} dG - \frac{\partial C}{\partial T} dT \\ dM^s \end{pmatrix} \tag{6.18}$$

all evaluated at $(Y^*, r^*, G^*, M^{s*}, T^*)$. The determinant of the coefficient matrix in (6.18),

$$D \equiv \left(1 - \frac{\partial C}{\partial Y}\right) \frac{\partial M}{\partial r} + \frac{\partial I}{\partial r} \frac{\partial M}{\partial Y}$$

is negative by (6.17) and therefore nonzero. By Theorem 6.2.3, the system (6.16) does indeed define Y and r as implicit functions of G, M^s, and T around $(Y^*, r^*, G^*, M^{s*}, T^*)$. Inverting (6.18), we compute

$$\begin{pmatrix} \partial Y \\ \partial r \end{pmatrix} = \frac{1}{D} \begin{pmatrix} \frac{\partial M}{\partial r} & \frac{\partial I}{\partial r} \\ -\frac{\partial M}{\partial Y} & 1 - \frac{\partial C}{\partial Y} \end{pmatrix} \begin{pmatrix} dG - \frac{\partial C}{\partial T} dT \\ dM^s \end{pmatrix}.$$

If we increase government spending G, keeping M^s and T fixed, we find

$$\frac{dY}{dG} = \frac{1}{D} \frac{\partial M}{\partial r} \quad \text{and} \quad \frac{dr}{dG} = -\frac{1}{D} \frac{\partial M}{\partial Y},$$

so both Y and r increase. Notice that using Cramer's rule on system (6.18) (keeping M^s and T fixed), we obtain

$$\frac{dY}{dG} = \frac{1}{D} \det \begin{pmatrix} 1 & \frac{\partial I}{\partial r} \\ 0 & \frac{\partial M}{\partial r} \end{pmatrix} = \frac{1}{D} \frac{\partial M}{\partial r}$$
$$\frac{dr}{dG} = \frac{1}{D} \det \begin{pmatrix} 1 - \frac{\partial C}{\partial Y} & 1 \\ \frac{\partial M}{\partial Y} & 0 \end{pmatrix} = -\frac{1}{D} \frac{\partial M}{\partial Y},$$

validating the results we obtained using total differentiation.

Example 6.2.5. Consider the follow problem

$$\begin{aligned} &\text{Minimize } x + 2y + 4z \\ &\text{subject to } x^2 + y^2 + z^2 = 21 \end{aligned} \tag{6.19}$$

The Lagrangian function is

$$\mathcal{L}(x, y, z, \mu) = -(x + 2y + 4z) - \mu(21 - x^2 - y^2 - z^2)$$

and the associated first order conditions are

$$\frac{\partial \mathcal{L}}{\partial x} = -1 + 2\mu x \overset{\text{set}}{=} 0$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial y} &= -2 + 2\mu y \overset{\text{set}}{=} 0\\ \frac{\partial \mathcal{L}}{\partial z} &= -4 + 2\mu z \overset{\text{set}}{=} 0\\ \frac{\partial \mathcal{L}}{\partial \mu} &= -21 + x^2 + y^2 + z^2 \overset{\text{set}}{=} 0.\end{aligned}$$

The first three equations can be solved to obtain $x = 1/2\mu$, $y = 1/\mu$, and $z = 2/\mu$. Substituting these results into the fourth equation yields $\mu^2 = 1/4$. Thus, there are two potential solutions: $(1, 2, 4, 1/2)$ and $(-1, -2, -4, -1/2)$. The associated bordered Hessian is

$$\begin{bmatrix}\mathcal{L}_{xx} & \mathcal{L}_{xy} & \mathcal{L}_{xz} & \mathcal{L}_{x\mu}\\ \mathcal{L}_{yx} & \mathcal{L}_{yy} & \mathcal{L}_{yz} & \mathcal{L}_{y\mu}\\ \mathcal{L}_{zx} & \mathcal{L}_{zy} & \mathcal{L}_{zz} & \mathcal{L}_{z\mu}\\ \mathcal{L}_{\mu x} & \mathcal{L}_{\mu y} & \mathcal{L}_{\mu z} & \mathcal{L}_{\mu\mu}\end{bmatrix} = \begin{bmatrix}2\mu & 0 & 0 & 2x\\ 0 & 2\mu & 0 & 2y\\ 0 & 0 & 2\mu & 2z\\ 2x & 2y & 2z & 0\end{bmatrix}.$$

Evaluated at $(1, 2, 4, 1/2)$, $\Delta_3 = -8\mu(y^2 + z^2) = -80$ and $\Delta_4 = -16\mu^2(x^2 + y^2 + z^2) = -84$. Since $(-1)^2\Delta_3 < 0$, this potential solution fails the second order condition. Evaluated at $(-1, -2, -4, -1/2)$, $\Delta_3 = 80$ and $\Delta_4 = -84$. This passes the second order test since $(-1)^2\Delta_3 = 80 > 0$ and $(-1)^3\Delta_4 = 84 > 0$. Thus, $(x^*, y^*, z^*) = (-1, -2, -4)$ is the global maximizer of f with value -21 (the constraint set is compact).

If we slightly alter the constraint, so that the problem (6.19) becomes

$$\begin{aligned}&\text{Minimize } \ x + 2y + 4z\\ &\text{subject to } \ x^2 + y^2 + z^2 = 21 + w,\end{aligned}$$

how does the optimal choice of x change as w changes from $w = 0$? The endogenous variables are x, y, z, and μ. The exogenous variable is w. The first order conditions remain the same except for the constraint $(\partial\mathcal{L}/\partial\mu)$, which becomes

$$\frac{\partial \mathcal{L}}{\partial \mu} = -21 - w + x^2 + y^2 + z^2 = 0.$$

When $w = 0$, we already know $(-1, -2, -4, -1/2)$ is the global maximizer (minimizer for the original problem), since it satisfies the first order conditions. Moreover, the bordered Hessian at this point,

$$\bar{H} \equiv \begin{bmatrix}-1 & 0 & 0 & -2\\ 0 & -1 & 0 & -4\\ 0 & 0 & -1 & -8\\ -2 & -4 & -8 & 0\end{bmatrix},$$

remains the same and its determinant: $\Delta_4 = -84 \neq 0$. Thus, we can apply Theorem 6.2.3 to obtain

$$\begin{bmatrix}\frac{dx}{dw}\\ \frac{dy}{dw}\\ \frac{dz}{dw}\\ \frac{d\mu}{dw}\end{bmatrix} = -\bar{H}^{-1}\begin{bmatrix}\frac{\partial^2 L}{\partial x \partial w}\\ \frac{\partial^2 L}{\partial y \partial w}\\ \frac{\partial^2 L}{\partial z \partial w}\\ \frac{\partial^2 L}{\partial \mu \partial w}\end{bmatrix} = -\bar{H}^{-1}\begin{bmatrix}0\\ 0\\ 0\\ -1\end{bmatrix} = \begin{bmatrix}-1/42\\ -1/21\\ -2/21\\ 1/84\end{bmatrix}.$$

Although we are not interested in $d\mu$, μ was one of the endogenous variables in the first order conditions, so it must be included in this stage, making $\bar{H}$ a 4×4 matrix. Also, note that almost all terms can be read off from the problem where $w = 0$.

Alternatively we could apply Cramer's rule to obtain

$$\frac{dx}{dw} = -\frac{\begin{vmatrix} 0 & 0 & 0 & -2 \\ 0 & -1 & 0 & -4 \\ 0 & 0 & -1 & -8 \\ -1 & -4 & -8 & 0 \end{vmatrix}}{\begin{vmatrix} -1 & 0 & 0 & -2 \\ 0 & -1 & 0 & -4 \\ 0 & 0 & -1 & -8 \\ -2 & -4 & -8 & 0 \end{vmatrix}} = -\frac{-2}{-84} = -\frac{1}{42}.$$

Thus, a one unit increase in w requires a reduction in x by $1/42$ units.

Chapter 7

Introduction to Complex Numbers

7.1 Basic Operations

7.1.1 Sums and Products

It is customary to denote a complex number (x, y) by z so that

$$z = (x, y).$$

The real numbers x and y are, moreover, known as the real and imaginary parts of z, respectively; and we write

$$\operatorname{Re} z = x \quad \text{and} \quad \operatorname{Im} z = y.$$

Two complex numbers $z_1 = (x_1, y_1)$ and $z_2 = (x_2, y_2)$ are equal whenever they have the same real and imaginary parts. That is, when $x_1 = x_2$ and $y_1 = y_2$. The sum $z_1 + z_2$ and product $z_1 z_2$ of two complex numbers is defined in the following manner:

$$(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$$
$$(x_1, y_1)(x_2, y_2) = (x_1 x_2 - y_1 y_2, y_1 x_2 + x_1 y_2).$$

Note that the operations defined in the above equations become the usual operations of addition and multiplication when restricted to the real numbers:

$$(x_1, 0) + (x_2, 0) = (x_1 + x_2, 0)$$
$$(x_1, 0)(x_2, 0) = (x_1 x_2, 0).$$

Thus, the complex number system is a natural extension of the real number system. Any complex number $z = (x, y)$ can be written $z = (x, 0) + (0, y)$ and it is easy to see that $(0, 1)(y, 0) = (0, y)$. Hence,

$$z = (x, 0) + (0, 1)(y, 0)$$

and, if we think of a real number as either x or $(x, 0)$ and let i denote the pure imaginary number $(0, 1)$, it is clear that

$$z = x + iy.$$

Also, we find that

$$i^2 = (0, 1)(0, 1) = (-1, 0) \quad \text{or} \quad i^2 = -1.$$

7.1.2 Moduli

The modulus or absolute value of a complex number, z, is defined as the nonnegative real number $\sqrt{x^2+y^2}$ and is denoted by $|z|$; that is,

$$|z| = \sqrt{x^2+y^2}.$$

Geometrically, the number $|z|$ is the distance between the point (x, y) and the origin, or the length of the vector representing z. It reduces to the absolute value in the real number system when $y = 0$. Note that while the inequality $z_1 < z_2$ is meaningless unless both z_1 and z_2 are real, the statement $|z_1| < |z_2|$ means that the point z_1 is closer to the origin than the point z_2.

7.1.3 Complex Conjugates

The complex conjugate, or simply the conjugate, of a complex number $z = x + iy$ is defined as the complex number $x - iy$ and is denoted $\overline{z}$; that is

$$\overline{z} = x - iy.$$

The number $\overline{z}$ is represented by the point $(x, -y)$, which is the reflection about the real axis of the point (x, y) representing z. Note the following properties of conjugates and moduli

$$\overline{\overline{z}} = z \quad \text{and} \quad |\overline{z}| = |z|.$$

If $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$, then

$$\overline{z_1 + z_2} = (x_1 + x_2) - i(y_1 + y_2) = (x_1 - iy_1) + (x_2 - iy_2) = \overline{z_1} + \overline{z_2}.$$

Thus, the conjugate of the sum is the sum of the conjugates. Similar conclusions can be drawn for differences, products, and quotients. That is

$$\overline{z_1 - z_2} = \overline{z_1} - \overline{z_2}$$
$$\overline{z_1 z_2} = \overline{z_1}\,\overline{z_2}$$
$$\overline{\left(\frac{z_1}{z_2}\right)} = \frac{\overline{z_1}}{\overline{z_2}} \quad (z_2 \neq 0).$$

The sum $z + \overline{z}$ of a complex number $z = x + iy$ and its conjugate $\overline{z} = x - iy$ is the real number $2x$, and the difference $z - \overline{z}$ is the pure imaginary number $2iy$. Hence

$$\operatorname{Re} z = \frac{z + \overline{z}}{2} \quad \text{and} \quad \operatorname{Im} z = \frac{z - \overline{z}}{2i}.$$

An important identity relating the conjugate of a complex number $z = x + iy$ to its modulus is

$$z\overline{z} = |z|^2,$$

where each side is equal to $x^2 + y^2$. Using the above property, we can establish the fact that the modulus of the product is the product of the modulus. That is

$$|z_1 z_2| = |z_1||z_2|,$$

which can be seen by noting that

$$|z_1 z_2|^2 = (z_1 z_2)(\overline{z_1 z_2}) = (z_1\overline{z_1})(z_2\overline{z_2}) = |z_1|^2|z_2|^2 = (|z_1||z_2|)^2$$

and recalling that the modulus is never negative.

Example 7.1.1. Reduce each of the following quantities to a real number

(a) $\frac{1+2i}{3-4i} + \frac{2-i}{5i}$

Solution: Multiply each term by the conjugate of its denominator

$$\frac{(1+2i)}{(3-4i)}\frac{(3+4i)}{(3+4i)} + \frac{(2-i)}{(5i)}\frac{(-5i)}{(-5i)} = \frac{10i-5}{25} - \frac{10i+5}{25} = -\frac{2}{5}.$$

(b) $\frac{5i}{(1-i)(2-i)(3-i)}$

Solution: Multiply by the conjugate of each of the terms in the denominator

$$\frac{5i}{(1-i)(2-i)(3-i)}\frac{(1+i)(2+i)(3+i)}{(1+i)(2+i)(3+i)} = \frac{25(i-1)(i+1)}{100} = -\frac{1}{2}.$$

(c) $(1-i)^4$

Solution: Note that $(1-i)^2 = -2i$. Thus,

$$(1-i)^4 = (-2i)^2 = -4.$$

7.2 Exponential Form

Notice that any complex number, z, can be written as

$$z \equiv \alpha + i\beta = \sqrt{\alpha^2+\beta^2}\left(\frac{\alpha}{\sqrt{\alpha^2+\beta^2}} + i\frac{\beta}{\sqrt{\alpha^2+\beta^2}}\right).$$

Thus, redefining z in terms of polar coordinates, we obtain

$$z = \alpha + i\beta = r[\cos(\theta) + i\sin(\theta)], \tag{7.1}$$

where $\cos(\theta) = \alpha/r$ and $\sin(\theta) = \beta/r$ as illustrated in figure 7.1. In section 7.1.2, we saw that the real number r is not allowed to be negative and is the length of the radius vector for z; that is $r = |z|$. The real number θ represents the angle, measured in radians, that z makes with the positive real axis. As in calculus, θ has an infinite number of possible values, including negative ones, that differ by multiples of 2π. Those values can be determined from the equation $\tan(\theta) = \beta/\alpha$, where the quadrant containing the point corresponding to z must be specified. Each value of θ is called an *argument* of z, and the set of all such values is denoted by $\arg z$. The *principle value* of $\arg z$, denoted by $\mathrm{Arg}\, z$, is the unique value Θ such that $-\pi < \Theta \leq \pi$. Note that

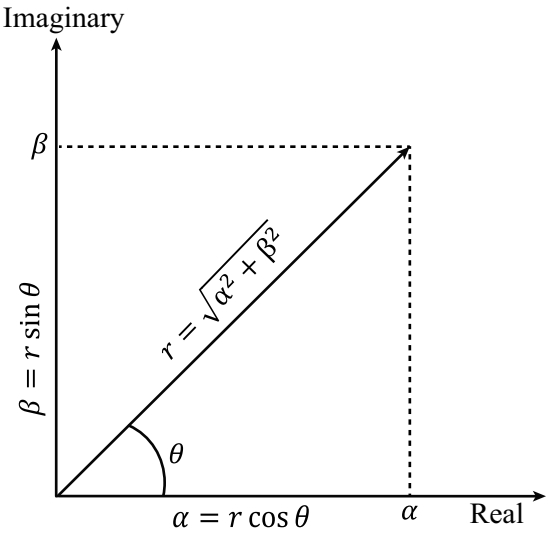
$$\arg z = \mathrm{Arg}\, z + 2\pi n. \qquad (n = 0, \pm 1, \pm 2, \ldots)$$

Example 7.2.1. The complex number $z = -1 - i$, which lies in the third quadrant, has principle argument $-3\pi/4$. That is,

$$\mathrm{Arg}(-1-i) = -\frac{3\pi}{4}.$$

It must be emphasized that, because of the restriction $-\pi < \Theta \leq \pi$ of the principle argument Θ, it is *not* true that $\mathrm{Arg}(-1-i) = 5\pi/4$.

Figure 7.1: Complex Number in Polar Form



The symbol $e^{i\theta}$, or $\exp(i\theta)$, is defined by means of Euler's formula as

$$e^{i\theta} = \cos\theta + i\sin\theta,$$

where θ is measured in radians.

Proof. In order to derive this result, first recall the Maclaurin expansions of the functions e^x, $\cos x$, and $\sin x$ given by:

$$\begin{aligned} e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots, \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots, \\ \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots. \end{aligned}$$

For a complex number z, define each of the functions by the above series, replacing the real variable x with the complex variable z. We then find that

$$\begin{aligned} e^{iz} &= 1 + iz + \frac{(iz)^2}{2!} + \frac{(iz)^3}{3!} + \frac{(iz)^4}{4!} + \frac{(iz)^5}{5!} + \frac{(iz)^6}{6!} + \frac{(iz)^7}{7!} + \frac{(iz)^8}{8!} + \cdots \\ &= 1 + iz - \frac{z^2}{2!} - \frac{iz^3}{3!} + \frac{z^4}{4!} + \frac{iz^5}{5!} - \frac{z^6}{6!} - \frac{iz^7}{7!} + \frac{z^8}{8!} + \cdots \\ &= \left(1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \frac{z^8}{8!} - \cdots\right) + i\left(z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots\right) \\ &= \cos z + i\sin z. \end{aligned}$$

□

This result enables us to write the polar form (7.1) more compactly in exponential form as

$$z = re^{i\theta}.$$

Example 7.2.2. The complex number $z = -1 - i$ in the previous example has exponential form

$$-1 - i = \sqrt{2}\exp\left[i(-3\pi/4)\right].$$

This expression is, of course, only one of an infinite number of possibilities for the exponential form of z.

It is geometrically obvious that

$$e^{i\pi} = -1 \qquad e^{-i\pi/2} = -i \qquad e^{-i4\pi} = 1$$

Note too that the equation

$$z = Re^{i\theta}$$

is a parametric representation of the circle $|z| = R$ centered at the origin with radius R.

Another nice aspect of complex numbers written in polar form is that their powers are easily computed. For example

$$\begin{aligned} z^2 = (\alpha + i\beta)^2 &= r^2\left[\cos(\theta) + i\sin(\theta)\right]^2 \\ &= r^2\left[\cos^2(\theta) + 2\cos(\theta)i\sin(\theta) - \sin^2(\theta)\right] \\ &= r^2\left[\cos(2\theta) + i\sin(2\theta)\right], \end{aligned}$$

using the double angle formulas, which state that $\cos(2a) = \cos^2(a) - \sin^2(a)$ and $\sin(2a) = 2\sin(a)\cos(a)$. Continuing in this manner, we obtain the following result.

Definition 7.2.1 (DeMoivre's formula)**.** *For complex number $z = \alpha + i\beta$ with polar representation $r\left[\cos(\theta) + i\sin(\theta)\right]$ and any positive integer n,*

$$z^n = (\alpha + i\beta)^n = r^n\left[\cos(n\theta) + i\sin(n\theta)\right]. \tag{7.2}$$

This result can alternatively be derived in a much simpler manner by noting that $(re^{i\theta})^n = r^n e^{in\theta}$.

Example 7.2.3. In order to put $(\sqrt{3} + i)^7$ is rectangular form, one need only write

$$(\sqrt{3} + i)^7 = (2e^{i\pi/6})^7 = 2^7 e^{i7\pi/6} = (2^6 e^{i\pi})(2e^{i\pi/6}) = -64(\sqrt{3} + i).$$

7.3 Complex Eigenvalues

Complex eigenvalues of real matrices occur in complex conjugate pairs. That is, if $z = \alpha + i\beta$ is a root of the characteristic polynomial, so is $\bar{z} = \alpha - i\beta$. To see this, consider a general 2×2 matrix A given by

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

Then the eigenvalues of the matrix A can be found by solving the following equation

$$\det(A - \lambda I) = \det\begin{pmatrix} a - \lambda & b \\ c & d - \lambda \end{pmatrix} = (a - \lambda)(d - \lambda) - cb \stackrel{\text{set}}{=} 0.$$

The associated roots are then given by

$$\lambda = \frac{a + d}{2} \pm \frac{1}{2}\sqrt{(a + d)^2 - 4(ad - bc)}$$

or, after simplification,

$$\lambda = \frac{a+d}{2} \pm \frac{1}{2}\sqrt{(a-d)^2 + 4bc}.$$

If the discriminant (the expression under the square root) is negative, we will have complex roots. That is, if $(a-d)^2 + 4bc < 0$, then the roots are complex (conjugate) pairs of the form

$$\lambda_1 = \frac{a+d}{2} + \frac{1}{2}i\sqrt{|(a-d)^2 + 4bc|}$$
$$\lambda_2 = \overline{\lambda}_1 = \frac{a+d}{2} - \frac{1}{2}i\sqrt{|(a-d)^2 + 4bc|}.$$

The fact that complex eigenvalues come in complex conjugate pairs and the fact that $\lambda \neq \overline{\lambda}$, imply that complex eigenvalues of a 2×2 system are always distinct. It is only with 4×4 matrices that the possibility of repeated complex eigenvalues arises. We next show that complex eigenvectors also come in conjugate pairs.

Suppose the general form an eigenvalue of matrix A is given by $\lambda = \alpha + i\beta$. Then the corresponding eigenvector, $\mathbf{w}$, is a non-zero solution to

$$[A - (\alpha + i\beta)\, I]\, \mathbf{w} = 0,$$

or reformulated

$$A\mathbf{w} = (\alpha + i\beta)\mathbf{w}. \tag{7.3}$$

Now write the complex vector $\mathbf{w}$ in its general form: $\mathbf{w} = \mathbf{u} + i\mathbf{v}$, where $\mathbf{u}$ and $\mathbf{v}$ are real vectors. Then (7.3) becomes

$$A\,(\mathbf{u} + i\mathbf{v}) = (\alpha + i\beta)\,(\mathbf{u} + i\mathbf{v})\,. \tag{7.4}$$

Applying the conjugate to both sides and recalling that for any two complex numbers z_1 and z_2 it holds that $\overline{z_1 z_2} = \overline{z}_1\,\overline{z}_2$, we obtain

$$\begin{aligned} & \overline{A\,(\mathbf{u} + i\mathbf{v})} = \overline{(\alpha + i\beta)\,(\mathbf{u} + i\mathbf{v})} \\ \rightarrow \quad & A\,(\mathbf{u} - i\mathbf{v}) = (\alpha - i\beta)\,(\mathbf{u} - i\mathbf{v}) \\ \rightarrow \quad & A\overline{\mathbf{w}} = (\alpha - i\beta)\,\overline{\mathbf{w}}. \end{aligned} \tag{7.5}$$

Then from (7.4) and (7.5) we see that complex eigenvectors also come in conjugate pairs. More specifically, if $\mathbf{u} + i\mathbf{v}$ is an eigenvector for $\alpha + i\beta$, then $\mathbf{u} - i\mathbf{v}$ is an eigenvector for $\alpha - i\beta$. The following theorem summarizes the discussion thus far.

Theorem 7.3.1. *Let A be a $k \times k$ matrix with real entries. If $\lambda = \alpha + i\beta$ is an eigenvalue of A, so is its conjugate $\overline{\lambda} = \alpha - i\beta$. If $(\mathbf{u} + i\mathbf{v})$ is an eigenvector for eigenvalue $\lambda = \alpha + i\beta$, then $(\mathbf{u} - i\mathbf{v})$ is an eigenvector for eigenvalue $\bar{\lambda} = \alpha - i\beta$. If k is an odd number, then A must have at least one real eigenvalue.*

Example 7.3.1. Consider the 2×2 matrix

$$A = \begin{bmatrix} 1 & 1 \\ -9 & 1 \end{bmatrix}.$$

Its characteristic polynomial is given by

$$p(\lambda) = \lambda^2 - 2\lambda + 10,$$

and its corresponding roots are

$$\lambda_{1,2} = 1 \pm \frac{1}{2}\sqrt{4 - 4(10)} = 1 \pm \frac{1}{2}\sqrt{-36} = 1 \pm 3i.$$

Thus, we have two complex eigenvalues that are complex conjugates. Use $\lambda_1 = 1 + 3i$ to calculate the first eigenvector for matrix A.

$$[A - (1 + 3i)I]\mathbf{w} = \begin{bmatrix} -3i & 1 \\ -9 & -3i \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Row-reducing the coefficient matrix in the above equation implies $-3iw_1 + w_2 = 0$. Normalizing w_1 to 1, we get $w_2 = 3i$. Thus, the first eigenvector is

$$\mathbf{w}_1 = \begin{bmatrix} 1 \\ 3i \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + i \begin{bmatrix} 0 \\ 3 \end{bmatrix}.$$

Since we have already seen that eigenvectors or complex eigenvalues come in conjugate pairs, it must be the case that the second eigenvector is given by

$$\mathbf{w}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} - i \begin{bmatrix} 0 \\ 3 \end{bmatrix}.$$

Note that we could have also found the second eigenvector using the second eigenvalue $\lambda_2 = 1 - 3i$.

Chapter 8

Linear Difference Equations and Lag Operators

8.1 Lag Operators

The backshift or *lag operator* is defined by

$$\begin{aligned} Lx_t &= x_{t-1} \\ L^n x_t &= x_{t-n} \quad \text{for} \quad n = \ldots, -2, -1, 0, 1, 2, \ldots . \end{aligned}$$

Multiplying a variable x_t by L^n thus gives the value of x shifted back n periods. Notice that if $n < 0$, the effect of multiplying x_t by L^n is to shift x *forward* in time by n periods.

Consider polynomials in the lag operator given by

$$A(L) = a_0 + a_1 L + a_2 L^2 + \cdots = \sum_{j=0}^{\infty} a_j L^j,$$

where the a_j's are constant and $L^0 \equiv 1$. Operating on x_t with $A(L)$ yields a moving sum of $x's$:

$$\begin{aligned} A(L)x_t &= (a_0 + a_1 L + a_2 L^2 + \cdots)x_t \\ &= a_0 x_t + a_1 x_{t-1} + a_2 x_{t-2} + \cdots \\ &= \sum_{j=0}^{\infty} a_j x_{t-j}. \end{aligned}$$

To take a simple example of a rational polynomial in L, consider

$$A(L) = \frac{1}{1 - \lambda L} = 1 + \lambda L + \lambda^2 L^2 + \cdots , \tag{8.1}$$

which holds assuming the normal expansion for an infinite geometric series applies. Note that this can be verified by multiplying both sides of the equality by $(1 - \lambda L)$. However, this result is sometimes only of practical use when $|\lambda| < 1$. To see why, consider the following infinite sum

$$\frac{1}{1 - \lambda L} x_t = \sum_{j=0}^{\infty} \lambda^j x_{t-j},$$

for the case in which the path of x_t is constant at $\overline{x}$. Then,

$$\frac{1}{1-\lambda L}x_t = \overline{x}\sum_{j=0}^{\infty}\lambda^j.$$

and if $|\lambda| > 1$, we get that the above sum is unbounded. In some instances, we will want a solution infinitely far back in time and where the infinite sum is bounded. Thus, we sometimes require $|\lambda| < 1$. It is useful to realize that we can use an alternative expansion for the geometric polynomial $1/(1-\lambda L)$. Formally, if the normal expansion for this infinite geometric series applies, then

$$\begin{aligned}\frac{1}{1-\lambda L} &= \frac{(-\lambda L)^{-1}}{1-(\lambda L)^{-1}} = -\frac{1}{\lambda L}\left(1+\frac{1}{\lambda L}+\frac{1}{(\lambda L)^2}+\cdots\right)\\ &= -\frac{1}{\lambda}L^{-1}-\left(\frac{1}{\lambda}\right)^2 L^{-2}-\left(\frac{1}{\lambda}\right)^3 L^{-3}-\cdots,\end{aligned} \tag{8.2}$$

which is particularly useful when $|\lambda| > 1$. In this case, operating on x_t gives

$$\frac{1}{1-\lambda L}x_t = -\frac{1}{\lambda}x_{t+1}-\left(\frac{1}{\lambda}\right)^2 x_{t+2}-\cdots = -\sum_{j=1}^{\infty}\left(\frac{1}{\lambda}\right)^j x_{t+j},$$

which shows $[1/(1-\lambda L)]x_t$ to be a geometrically declining weighted sum of future values of x. Notice that for this infinite sum to be finite for a constant time path $x_{t+j} = \overline{x}$ for all j and t, the series $-\sum_{j=1}^{\infty}(1/\lambda)^j$ must be convergent, which requires that $|1/\lambda| < 1$ or, equivalently, $|\lambda| > 1$.

8.2 First-Order Difference Equations

A linear difference equation can be defined as an equation that relates the endogenous (determined within the model) variable y_t to its previous values linearly. The simplest one is a first-order scalar linear difference equation such as

$$y_t = \lambda y_{t-1} + bx_t + a, \tag{8.3}$$

where x_t is an exogenous (determined outside the model) variable and a is a constant. This is a first-order difference equation, since y_t is dependent on only its first lag y_{t-1}. Here, we are interested in finding a solution of y_t in terms of current, past, or future values of the exogenous variable x_t and (less importantly) the constant a. Put it different, we want to characterize the endogenous sequence y_t in terms of the exogenous sequence x_t.

Using the lag operator defined above, we can rewrite (8.3) as follows:

$$(1-\lambda L)y_t = bx_t + a. \tag{8.4}$$

Operating on both sides of this equation by $(1-\lambda L)^{-1}$, we can obtain a *particular* solution for (8.4), denoted by $\hat{y}_t$, as follows:

$$\hat{y}_t = \frac{bx_t}{1-\lambda L} + \frac{a}{1-\lambda}. \tag{8.5}$$

Note that since a is a constant, $a/(1-\lambda L) = a/(1-\lambda)$ irrespective of the size of $|\lambda|$ (This can be verified by considering the expansion given in (8.1) when $|\lambda| < 1$ and the expansion given in (8.2) when $|\lambda| > 1$). In order to obtain the *general* solution, we need to add a term to (8.5). For this

purpose, suppose that $\tilde{y} = \hat{y} + w_t$ is also a solution to (8.4). Then, using the particular solution to (8.4), we obtain

$$\begin{aligned}(1-\lambda L)\tilde{y} &= (1-\lambda L)\hat{y} + w_t - \lambda w_{t-1} \\ &= bx_t + a + w_t - \lambda w_{t-1}.\end{aligned}$$

Therefore, as long as $w_t = \lambda w_{t-1}$, $\tilde{y}_t$ is also a solution. Note that we can iterate on this condition to obtain

$$w_t = \lambda w_{t-1} = \lambda^2 w_{t-2} = \lambda^3 w_{t-3} = \cdots = \lambda^t w_0,$$

where $w_0 \equiv c$, an arbitrary initial value. Hence, the general solution to (8.3) is given by

$$\begin{aligned}y_t &= \frac{bx_t}{1-\lambda L} + \frac{a}{1-\lambda} + \lambda^t c \\ &= b\sum_{j=0}^{\infty} \lambda^j x_{t-j} + \frac{a}{1-\lambda} + \lambda^t c, \end{aligned} \tag{8.6}$$

where c is an arbitrary constant. Notice that for y_t, defined by (8.6), to be finite $\lambda^j x_{t-j}$ must be small for large j. That is, we require

$$\lim_{n\to\infty} \sum_{j=n}^{\infty} \lambda^j x_{t-j} = 0 \qquad \text{for all } t.$$

For the case of $x_{t-j} = \overline{x}$ for all j and t, the above condition requires $|\lambda| < 1$. Notice also that the infinite sum $a\sum_{i=0}^{\infty} \lambda^j$ in (8.6) is also finite only if $|\lambda| < 1$, in which case it equals $a/(1-\lambda)$ for $a \neq 0$ and 0 otherwise. Tentatively, assume that $|\lambda| < 1$.

In order to analyze (8.6), rewrite the equation for $t \geq 1$ as

$$\begin{aligned}y_t &= a\sum_{j=0}^{t-1} \lambda^j + a\sum_{j=t}^{\infty} \lambda^j + b\sum_{j=0}^{t-1} \lambda^j x_{t-j} + b\sum_{j=t}^{\infty} \lambda^j x_{t-j} + \lambda^t c \\ &= \frac{a(1-\lambda^t)}{1-\lambda} + \frac{a\lambda^t}{1-\lambda} + b\sum_{j=0}^{t-1} \lambda^j x_{t-j} + b\lambda^t \sum_{j=0}^{\infty} \lambda^j x_{0-j} + \lambda^t c \\ &= \frac{a(1-\lambda^t)}{1-\lambda} + b\sum_{j=0}^{t-1} \lambda^j x_{t-j} + \lambda^t \left[\frac{a}{1-\lambda} + b\sum_{j=0}^{\infty} \lambda^j x_{0-j} + \lambda^0 c\right] \\ &= \frac{a(1-\lambda^t)}{1-\lambda} + b\sum_{j=0}^{t-1} \lambda^j x_{t-j} + \lambda^t y_0 \qquad \text{(using (8.6))} \\ &= \frac{a}{1-\lambda} + \lambda^t \left[y_0 - \frac{a}{1-\lambda}\right] + b\sum_{j=0}^{t-1} \lambda^j x_{t-j}, \quad t \geq 1.\end{aligned}$$

Consider the special case in which $x_t = 0$ for all t. Under this condition, we obtain

$$y_t = \frac{a}{1-\lambda} + \lambda^t \left[y_0 - \frac{a}{1-\lambda}\right]. \tag{8.7}$$

Notice that if the initial condition $y_0 = a/(1-\lambda)$, then $y_t = y_0$. In the case, y is constant across all future time periods and $a/(1-\lambda)$ is known as a *stationary point*. Moreover, it is easy to see that for $|\lambda| < 1$, the second term in (8.7) tends to zero and thus

$$\lim_{t\to\infty} y_t = \frac{a}{1-\lambda}.$$

This shows that the system is *stable*, tending to approach the stationary value as time passes.

The difference equation (8.4) can also be solved using the alternative representation of $(1-\lambda L)^{-1}$ given in (8.2). Using this this result, the general solution is given by

$$\begin{aligned} y_t &= \frac{(-\lambda L)^{-1}}{1-(\lambda L)^{-1}}a + \frac{(-\lambda L)^{-1}}{1-(\lambda L)^{-1}}bx_t + \lambda^t c \\ &= \frac{a}{1-\lambda} - b\sum_{j=1}^{\infty}\lambda^{-j}x_{t+j} + \lambda^t c. \end{aligned} \tag{8.8}$$

The equivalence of the solutions (8.6) and (8.8) will hold whenever

$$\frac{b}{1-\lambda L}x_t \quad \text{and} \quad \frac{(\lambda L)^{-1}}{1-(\lambda L)^{-1}}bx_t$$

are both finite. However, it is often the case that one of these two conditions fails to hold. For example, if the sequence $\{x_t\}$ is bounded, this is sufficient to imply that $\{[b/(1-\lambda L)]x_t\}$ is a bounded sequence if $|\lambda| < 1$, but not sufficient to imply that

$$\frac{(\lambda L)^{-1}}{1-(\lambda L)^{-1}}bx_t$$

is a convergent sum for all t. Similarly, if $|\lambda| > 1$, boundedness of the sequence $\{x_t\}$ is sufficient to imply that

$$\left\{\frac{(\lambda L)^{-1}}{1-(\lambda L)^{-1}}bx_t\right\}$$

is a bounded sequence, but fails to guarantee finiteness of $b/(1-\lambda L)x_t$. In instances where one of

$$\frac{b}{1-\lambda L}x_t \quad \text{or} \quad \frac{(\lambda L)^{-1}}{1-(\lambda L)^{-1}}bx_t$$

is always finite and the other is not, we shall take our solution to the first-order difference equation (8.4) as either (8.6), where the backward sum in x_t is finite, or (8.8), where the forward sum in x_t is finite. This procedure assures us that we shall find the unique solution of (8.4) that is finite for all t, provided that such a solution exists.

If we want to guarantee that the sequence $\{y_t\}$ given by (8.6) or (8.8) is bounded for all t, it is evident that we must set $c = 0$. This is necessary since if $\lambda > 1$ and $c > 0$,

$$\lim_{t\to\infty} c\lambda^t = \infty,$$

while if $\lambda < 1$ and $c > 0$,

$$\lim_{t\to-\infty} c\lambda^t = \infty.$$

Thus, when $|\lambda| < 1$, the bounded sequence y_t from (8.6) can be obtained by following the backward representation with the initial condition $c = 0$ and is given by

$$\begin{aligned} y_t &= b\left(1 + \lambda L + (\lambda L)^2 + (\lambda L)^3 + \cdots\right) x_t + \frac{a}{1-\lambda} \\ &= b\sum_{j=0}^{\infty} \lambda^j x_{t-j} + \frac{a}{1-\lambda}. \end{aligned}$$

On the other hand, when $|\lambda| > 1$, we need to use the forward representation in order to get the bounded sequence y_t as follows:

$$\begin{aligned} y_t &= \frac{(-\lambda L)^{-1}}{1-(\lambda L)^{-1}} b x_t + \frac{(-\lambda L)^{-1}}{1-(\lambda L)^{-1}} a \\ &= -b\sum_{j=1}^{\infty} \lambda^{-j} x_{t+j} + \frac{a}{1-\lambda}, \end{aligned}$$

again setting $c = 0$. In general, the convention is to solve *stable* roots ($|\lambda| < 1$) backward and *unstable* roots ($|\lambda| > 1$) forward.

8.3 Second-Order Difference Equations

A second-order difference equation relates the endogenous variable y_t to its previous two values, y_{t-1} and y_{t-2}, linearly. Consider following second-order difference equation given by

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + b x_t + a, \tag{8.9}$$

where x_t is again an exogenous sequence of real numbers for $t = \ldots, -1, 0, 1, \ldots$. Using the lag operator, we can write (8.9) as follows:

$$(1 - \phi_1 L - \phi_2 L^2) y_t = b x_t + a.$$

It is convenient to write the polynomial $1 - \phi_1 L - \phi_2 L^2$ in an alternative way, given by the factorization

$$\begin{aligned} 1 - \phi_1 L - \phi_2 L^2 &= (1 - \lambda_1 L)(1 - \lambda_2 L) \\ &= 1 - (\lambda_1 + \lambda_2) L + \lambda_1 \lambda_2 L^2, \end{aligned}$$

where $\lambda_1 \lambda_2 = -\phi_2$ and $\lambda_1 + \lambda_2 = \phi_1$. To see how λ_1 and λ_2 are related to the roots or zeros of the polynomial $A(z) = 1 - \phi_1 z - \phi_2 z^2$, notice that

$$(1 - \lambda_1 z)(1 - \lambda_2 z) = \lambda_1 \lambda_2 \left(\frac{1}{\lambda_1} - z\right)\left(\frac{1}{\lambda_2} - z\right).$$

Note that we use a function of a number z (possibly complex) instead of the lag operator L since it does not really make much to talk about roots or zeros of a polynomial that is a function of a lag operator. If we set the above equation to zero in order to solve for its roots, it is clear that the equation is satisfied at the two roots $z_1 = 1/\lambda_1$ and $z_2 = 1/\lambda_2$. Given the polynomial $A(z) = 1 - \phi_1 z - \phi_2 z^2$, the roots $1/\lambda_1$ and $1/\lambda_2$ are found by solving the characteristic equation

$$1 - \phi_1 z - \phi_2 z^2 = 0$$

for two values of z. Given that $\lambda_i = z_i^{-1}$ for $i = 1, 2$, multiplying the above equation by z^{-2} yields

$$z^{-2} - z^{-1}\phi_1 - \phi_2 = \lambda^2 - \phi_1\lambda - \phi_2 = 0.$$

Applying the quadratic formula then gives

$$\lambda_i = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2},$$

which enables us to obtain the reciprocals of λ_1 and λ_2 for given values of ϕ_1 and ϕ_2.

8.3.1 Distinct Real Eigenvalues

General Solution

When $\lambda_1 \neq \lambda_2$ and $\lambda_i \neq 1$ for all i, the second order difference equation (8.9) can be written as

$$(1 - \lambda_1 L)(1 - \lambda_2 L)y_t = bx_t + a. \tag{8.10}$$

Thus, the general solution to (8.10) is given by

$$y_t = \frac{1}{(1 - \lambda_1 L)(1 - \lambda_2 L)}bx_t + \frac{a}{(1 - \lambda_1)(1 - \lambda_2)} + \lambda_1^t c_1 + \lambda_2^t c_2, \tag{8.11}$$

where c_1 and c_2 are any constants that can be verified by noticing

$$(1 - \lambda_1 L)(1 - \lambda_2 L)c_1\lambda_1^t = 0 \quad \text{and} \quad (1 - \lambda_1 L)(1 - \lambda_2 L)c_2\lambda_2^t = 0.$$

Particular Solution

If both eigenvalues are distinct as in (8.11), then the above coefficient can be written as

$$\frac{1}{(1 - \lambda_1 L)(1 - \lambda_2 L)} = \frac{1}{\lambda_1 - \lambda_2}\left(\frac{\lambda_1}{1 - \lambda_1 L} - \frac{\lambda_2}{1 - \lambda_2 L}\right). \tag{8.12}$$

Thus, if either $a = 0$ or the magnitude of both eigenvalues is strictly less than unity (that is, if $|\lambda_1| < 1$ and $|\lambda_2| < 1$), (8.11) can be written as

$$\begin{aligned} y_t &= \frac{a}{(1 - \lambda_1)(1 - \lambda_2)} + \frac{1}{\lambda_1 - \lambda_2}\left(\frac{\lambda_1}{1 - \lambda_1 L} - \frac{\lambda_2}{1 - \lambda_2 L}\right)bx_t + \lambda_1^t c_1 + \lambda_2^t c_2 \\ &= a\sum_{j=0}^{\infty}\lambda_1^j \sum_{j=0}^{\infty}\lambda_2^j + \frac{\lambda_1 b}{\lambda_1 - \lambda_2}\sum_{j=0}^{\infty}\lambda_1^j x_{t-j} - \frac{\lambda_2 b}{\lambda_1 - \lambda_2}\sum_{j=0}^{\infty}\lambda_2^j x_{t-j} + \lambda_1^t c_1 + \lambda_2^t c_2 \end{aligned} \tag{8.13}$$

provided that

$$\lim_{n\to\infty}\sum_{j=n}^{\infty}\lambda_i^j x_{t-j} = 0, \qquad \text{for all } t$$

for $i = 1, 2$. Note that this stipulation is needed so that the corresponding geometric sums are finite.

In order to analyze (8.13), let's first consider the special case where $a = 0$, so that this equation holds regardless of the magnitude of λ_i. Then rewriting (8.13) for $t \geq 1$ gives

$$y_t = \frac{\lambda_1 b}{\lambda_1 - \lambda_2}\sum_{j=0}^{t-1}\lambda_1^j x_{t-j} - \frac{\lambda_2 b}{\lambda_1 - \lambda_2}\sum_{j=0}^{t-1}\lambda_2^j x_{t-j} + \frac{\lambda_1 b}{\lambda_1 - \lambda_2}\sum_{j=t}^{\infty}\lambda_1^j x_{t-j}$$

$$
\begin{aligned}
&- \frac{\lambda_2 b}{\lambda_1 - \lambda_2} \sum_{j=t}^{\infty} \lambda_2^j x_{t-j} + \lambda_1^t c_1 + \lambda_2^t c_2 \\
&= \frac{\lambda_1 b}{\lambda_1 - \lambda_2} \sum_{j=0}^{t-1} \lambda_1^j x_{t-j} - \frac{\lambda_2 b}{\lambda_1 - \lambda_2} \sum_{j=0}^{t-1} \lambda_2^j x_{t-j} + \lambda_1^t \theta_0 + \lambda_2^t \eta_0,
\end{aligned}
$$

where

$$
\theta_0 \equiv \left\{ c_1 + \frac{\lambda_1 b}{\lambda_1 - \lambda_2} \sum_{j=0}^{\infty} \lambda_1^j x_{0-j} \right\} \quad \text{and} \quad \eta_0 \equiv \left\{ c_2 - \frac{\lambda_2 b}{\lambda_1 - \lambda_2} \sum_{j=0}^{\infty} \lambda_2^j x_{0-j} \right\}.
$$

Thus, for the case in which $x_t = 0$ for $t \geq 1$, we obtain

$$
y_t = \lambda_1^t \theta_0 + \lambda_2^t \eta_0. \tag{8.14}
$$

If $\theta_0 = \eta_0 = 0$, then $y_t = 0$ for all $t \geq 1$, regardless of the values of λ_1 and λ_2. Thus, $y = 0$ is the stationary point or long-run equilibrium value of (8.14). If λ_1 and λ_2 are real then $\lim_{t \to \infty}$ will equal its stationary point if and only if both $|\lambda_1| < 1$ and $|\lambda_2| < 1$. If one or both or the λ's exceed one in absolute value, the behavior of y will eventually be dominated by the term in (8.14) associated with the λ that is larger in absolute value.

Now let's return to the more general case. If we are interested in a bounded sequence $\{y_t\}$ mapped from a bounded sequence $\{x_t\}$, then we need to set both of the constants c_1 and c_2 to zero, and focus on the associated particular solution. If the magnitudes of both eigenvalues are strictly less than unity, that is, if $|\lambda_1| < 1$ and $|\lambda_2| < 1$, then the bounded solution to (8.9) is given by

$$
y_t = \frac{a}{(1 - \lambda_1)(1 - \lambda_2)} + \frac{\lambda_1 b}{\lambda_1 - \lambda_2} \sum_{j=0}^{\infty} \lambda_1^j x_{t-j} - \frac{\lambda_2 b}{\lambda_1 - \lambda_2} \sum_{j=0}^{\infty} \lambda_2^j x_{t-j}.
$$

If, without loss of generality, $|\lambda_1| < 1$ and $|\lambda_2| > 1$, then we can write

$$
\begin{aligned}
\frac{1}{(1 - \lambda_1 L)(1 - \lambda_2 L)} &= \frac{1}{\lambda_1 - \lambda_2} \left(\frac{\lambda_1}{1 - \lambda_1 L} + \frac{L^{-1}}{1 - (\lambda_2 L)^{-1}} \right) \\
&= \frac{\lambda_1}{\lambda_1 - \lambda_2} \sum_{j=0}^{\infty} (\lambda_1 L)^j + \frac{\lambda_2}{\lambda_1 - \lambda_2} \sum_{j=1}^{\infty} (\lambda_2 L)^{-j}.
\end{aligned}
$$

Thus, in this case, the bounded solution to (8.9) is given by

$$
y_t = \frac{a}{(1 - \lambda_1)(1 - \lambda_2)} + \frac{\lambda_1 b}{\lambda_1 - \lambda_2} \sum_{j=0}^{\infty} \lambda_1^j x_{t-j} + \frac{\lambda_2 b}{\lambda_1 - \lambda_2} \sum_{j=1}^{\infty} \lambda_2^{-j} x_{t+j}.
$$

Finally, if $|\lambda_1| > 1$ and $|\lambda_2| > 1$, then we can write

$$
\begin{aligned}
\frac{1}{(1 - \lambda_1 L)(1 - \lambda_2 L)} &= \frac{1}{\lambda_1 - \lambda_2} \left(\frac{-L^{-1}}{1 - (\lambda_1 L)^{-1}} + \frac{L^{-1}}{1 - (\lambda_2 L)^{-1}} \right) \\
&= -\frac{\lambda_1}{\lambda_1 - \lambda_2} \sum_{j=1}^{\infty} (\lambda_1 L)^{-j} + \frac{\lambda_2}{\lambda_1 - \lambda_2} \sum_{j=1}^{\infty} (\lambda_2 L)^{-j}.
\end{aligned}
$$

Therefore, in this final case, the bounded solution to (8.9) is given by

$$
y_t = \frac{a}{(1 - \lambda_1)(1 - \lambda_2)} - \frac{\lambda_1 b}{\lambda_1 - \lambda_2} \sum_{j=1}^{\infty} \lambda_1^{-j} x_{t+j} + \frac{\lambda_2 b}{\lambda_1 - \lambda_2} \sum_{j=1}^{\infty} \lambda_2^{-j} x_{t+j}.
$$

8.3.2 Complex Eigenvalues

Recall that if one eigenvalue turns out to be a complex number, then the other eigenvalue is its complex conjugate. That is, if $\lambda_1 = \alpha + i\beta$, then $\lambda_2 = \alpha - i\beta$, where

$$\lambda_1 + \lambda_2 = 2\alpha = \phi_1, \quad \lambda_1\lambda_2 = \alpha^2 + \beta^2 = -\phi_2, \quad |\lambda_i| = \sqrt{\alpha^2 + \beta^2} \ \forall\, i.$$

Moreover, using the useful polar representation defined in section 7.2, we have

$$\lambda_1 = re^{iw} = r(\cos w + i \sin w), \quad \lambda_2 = re^{-iw} = r(\cos w - i \sin w),$$

where $r = \sqrt{\alpha^2 + \beta^2}$ and $\tan w = \beta/\alpha$. Finally notice that

$$\lambda_1 + \lambda_2 = r(e^{iw} + e^{-iw}) = 2r \cos w \quad \text{and} \quad \lambda_1 - \lambda_2 = r(e^{iw} - e^{-iw}) = 2ri \sin w.$$

Returning to the special case where $a = 0$ and $x_t = 0$, when the eigenvalues are complex, (8.14), becomes

$$\begin{aligned} y_t &= \theta_0(re^{iw})^t + \eta_0(re^{-iw})^t \\ &= \theta_0(r^t e^{iwt}) + \eta_0(r^t e^{-iwt}) \\ &= \theta_0 r^t[\cos wt + i \sin wt] + \eta_0 r^t[\cos wt - i \sin wt] \\ &= (\theta_0 + \eta_0) r^t \cos wt + i(\theta_0 - \eta_0) r^t \sin wt. \end{aligned}$$

Since y_t must be a real number for all t, it follows that $\theta_0 + \eta_0$ must be real and $\theta_0 - \eta_0$ must be imaginary. Therefore, θ_0 and η_0 must be complex conjugates, say $\theta_0 = pe^{i\theta}$ and $\eta_0 = pe^{-i\theta}$. Thus, we can write

$$\begin{aligned} y_t = pe^{i\theta} r^t e^{iwt} + pe^{-i\theta} r^t e^{-iwt} &= pr^t[e^{i(wt+\theta)} + e^{-i(wt+\theta)}] \\ &= pr^t[\cos(wt + \theta) + i \sin(wt + \theta) \\ &\quad + \cos(-(wt + \theta)) + i \sin(-(wt + \theta))] \\ &= 2pr^t \cos(wt + \theta), \end{aligned}$$

where we have made use of the fact that $\cos$ is an even function and $\sin$ is an odd function (i.e. for any input x, $\cos(x) = \cos(-x)$ and $\sin(-x) = -\sin(x)$). The path of y_t oscillates with a frequency determined by w. The dampening factor, r^t, is determined by the amplitude, r, of the complex roots. When $r < 1$, the stationary point of the difference equation, $y_t = 0$, is approached as $t \to \infty$. Moreover, as long as $w \neq 0$, the system displays damped oscillations. If $r = 1$, y_t displays repeated oscillations of unchanging amplitude and the solution is periodic. If $r > 1$ the path of y_t displays explosive oscillations, unless the initial conditions are say, $y_0 = 0$ and $y_1 = 0$ so that y starts out at the stationary point for two successive values.

Now if we once again consider the more general case where we are interested in a bounded sequence $\{y_t\}$ mapped from a bounded sequence $\{x_t\}$, we need to set both of the constants c_1 and c_2 to zero, and focus on the associated particular solution. If we note that moduli of complex eigenvalues are same, then when $|\lambda| < 1$, we can write

$$\begin{aligned} \frac{1}{(1 - \lambda_1 L)(1 - \lambda_2 L)} &= \frac{\lambda_1}{\lambda_1 - \lambda_2} \sum_{j=0}^{\infty} (\lambda_1 L)^j - \frac{\lambda_2}{\lambda_1 - \lambda_2} \sum_{j=0}^{\infty} (\lambda_2 L)^j \\ &= \frac{1}{\lambda_1 - \lambda_2} \sum_{j=0}^{\infty} \left(\lambda_1^{j+1} - \lambda_2^{j+1}\right) L^j \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{re^{iw} - re^{-iw}} \sum_{j=0}^{\infty} \left[(re^{iw})^{j+1} - (re^{-iw})^{j+1} \right] L^j \\
&= \frac{1}{2ri \sin w} \sum_{j=0}^{\infty} 2r^{j+1} i \sin[w(j+1)] L^j \\
&= \sum_{j=0}^{\infty} r^j \frac{\sin[w(j+1)]}{\sin w} L^j.
\end{aligned}
$$

Thus, the bounded solution to (8.9) is given by

$$
y_t = a \sum_{j=0}^{\infty} r^j \frac{\sin[w(j+1)]}{\sin w} + b \sum_{j=0}^{\infty} r^j \frac{\sin[w(j+1)]}{\sin w} x_{t-j}.
$$

If, on the other hand, $|\lambda| > 1$, we can write

$$
\begin{aligned}
\frac{1}{(1-\lambda_1 L)(1-\lambda_2 L)} &= -\frac{\lambda_1}{\lambda_1 - \lambda_2} \sum_{j=1}^{\infty} (\lambda_1 L)^{-j} + \frac{\lambda_2}{\lambda_1 - \lambda_2} \sum_{j=1}^{\infty} (\lambda_2 L)^{-j} \\
&= -\frac{1}{\lambda_1 - \lambda_2} \sum_{j=0}^{\infty} \left(\lambda_1^{-j} - \lambda_2^{-j} \right) L^{-(j+1)} \\
&= -\sum_{j=0}^{\infty} r^{-(j+1)} \frac{\sin(wj)}{\sin w} L^{-(j+1)}.
\end{aligned}
$$

Thus, in this case, the bounded solution to (8.9) is given by

$$
y_t = -a \sum_{j=0}^{\infty} r^{-(j+1)} \frac{\sin(wj)}{\sin w} - b \sum_{j=0}^{\infty} r^{-(j+1)} \frac{\sin(wj)}{\sin w} x_{t+j+1}.
$$

8.3.3 Stability Conditions for Distinct Eigenvalues

Recall that the roots of (8.9) are given by

$$
\lambda_i = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2}.
$$

For the roots to be complex, the discriminant must be negative, i.e.,

$$
\phi_1^2 + 4\phi_2 < 0,
$$

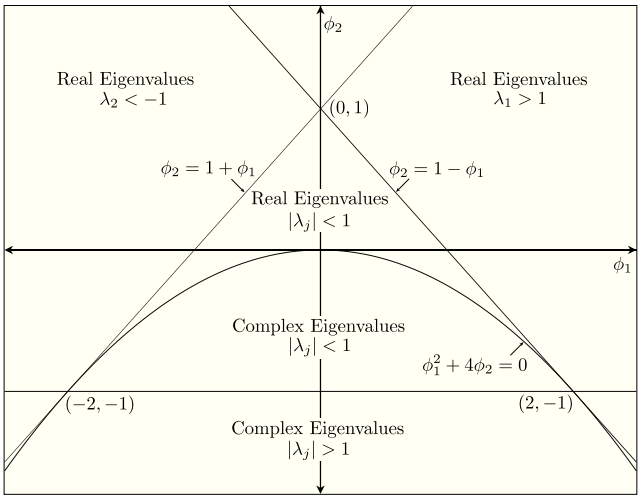
which implies that ϕ_2 is negative. When the above condition is satisfied, the roots are given by

$$
\lambda_1 = \frac{\phi_1}{2} + i \frac{\sqrt{-(\phi_1^2 + 4\phi_2)}}{2} \equiv a + ib, \qquad \lambda_2 = \frac{\phi_1}{2} - i \frac{\sqrt{-(\phi_1^2 + 4\phi_2)}}{2} \equiv a - ib.
$$

Once again, recall that in polar form

$$
a + ib = r[\cos w + i \sin w] = re^{iw},
$$

Figure 8.1: Second Order Difference Equation: Regions of Stability



where $r \equiv \sqrt{a^2 + b^2}$ and $\tan w = \beta/\alpha$. Thus we have that

$$r = \sqrt{\left(\frac{\phi_1}{2}\right)^2 - \frac{(\phi_1^2 + 4\phi_2)}{4}} = \sqrt{-\phi_2}.$$

For the oscillations to be damped, meaning that in the long-run the difference equation will be stable, we require that $r = \sqrt{-\phi_2} < 1$, which requires that $\phi_2 > -1$.

If the roots are real, the difference equation will be stable if both roots are less that one in magnitude. This requires

$$-1 < \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{2} < 1 \quad \text{and} \quad -1 < \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{2} < 1.$$

Note that it is sufficient to find conditions such that statement on the left hand side is less than unity while the condition on the right hand side is greater than minus one. The former condition requires

$$\begin{aligned} \frac{1}{2}\left(\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}\right) < 1 \quad &\rightarrow \quad \sqrt{\phi_1^2 + 4\phi_2} < 2 - \phi_1 \\ \rightarrow \quad \phi_1^2 + 4\phi_2 < 4 + \phi_1^2 - 4\phi_1 \quad &\rightarrow \quad \phi_1 + \phi_2 < 1. \end{aligned}$$

The latter condition requires

$$\begin{aligned} \frac{1}{2}\left(\sqrt{\phi_1^2 + 4\phi_2} - \phi_1\right) < 1 \quad &\rightarrow \quad \sqrt{\phi_1^2 + 4\phi_2} < 2 + \phi_1 \\ \rightarrow \quad \phi_1^2 + 4\phi_2 < 4 + \phi_1^2 + 4\phi_1 \quad &\rightarrow \quad \phi_2 - \phi_1 < 1. \end{aligned}$$

Therefore, when $\phi_2 > -1$, $\phi_1 + \phi_2 < 1$, and $\phi_2 - \phi_1 < 1$ hold, the roots, regardless of whether they are real ($\phi_1^2 + 4\phi_2 \geq 0$) or complex ($\phi_1^2 + 4\phi_2 < 0$), will yield a stable second order difference equation. The following figure summarizes these results.

8.3.4 Repeated Real Eigenvalues

General Solution

When $\lambda_1 = \lambda_2 \equiv \lambda$ and $\lambda_i \neq 1$ for all i, the second order difference equation (8.9) becomes

$$(1 - \lambda L)^2 y_t = bx_t + a. \tag{8.15}$$

Thus, the general solution to (8.15) is given by

$$y_t = \frac{1}{(1-\lambda L)^2} b x_t + \frac{a}{(1-\lambda L)^2} + \lambda^t c_1 + t\lambda^t c_2, \tag{8.16}$$

where c_1 and c_2 are constants. To see this note that

$$(1-\lambda L)^2(\lambda^t c_1 + t\lambda^t c_2) = 0.$$

Particular Solution

If we are interested in a bounded sequence $\{y_t\}$ mapped from a bounded sequence $\{x_t\}$, then we need to set both of the constants c_1 and c_2 to zero, and focus on the associated particular solution. When $\lambda_1 = \lambda_2 \equiv \lambda$ and $|\lambda| < 1$, we can show that

$$\sum_{j=0}^{\infty} (\lambda L)^{j+1} = \frac{\lambda L}{1-\lambda L}.$$

Applying the derivative trick from Example 1.1.29 gives

$$\frac{1}{(1-\lambda_1 L)(1-\lambda_2 L)} = \frac{1}{(1-\lambda L)^2} = \sum_{j=0}^{\infty} (j+1)(\lambda L)^j.$$

Thus, in this case, the bounded solution to (8.9) is given by

$$y_t = \frac{a}{(1-\lambda L)^2} + b\sum_{j=0}^{\infty} (j+1)\lambda^j x_{t-j}.$$

If, on the other hand, $|\lambda| > 1$, we can show that

$$\frac{1}{1-\lambda L} = \frac{-(\lambda L)}{1-(\lambda L)^{-1}} = -\sum_{j=0}^{\infty} (\lambda L)^{-(j+1)}.$$

Applying the derivative trick yields

$$\frac{1}{(1-\lambda L)^2} = \sum_{j=0}^{\infty} (j+1)(\lambda L)^{-(j+2)}.$$

Thus, in this case, the bounded solution to (8.9) is given by

$$y_t = \frac{a}{(1-\lambda L)^2} + b\sum_{j=0}^{\infty} (j+1)(\lambda)^{-(j+2)} x_{t+j+2}.$$

8.4 Systems of Linear Difference Equations

8.4.1 Solution Technique with Real Eigenvalues

Consider the following k-dimensional system of equations:

$$\mathbf{z}_{t+1} = A\mathbf{z}_t,$$

with $\mathbf{z} \in \mathbb{R}^k$ and $A \in \mathbb{R}^{k \times k}$. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the eigenvalues of A and let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k \in \mathbb{R}^k$ be the corresponding eigenvectors. Then the projection matrix $P = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k]$ and

$$\begin{aligned} AP &= [A\mathbf{v}_1, A\mathbf{v}_2, \ldots, A\mathbf{v}_k] \\ &= [\lambda_1\mathbf{v}_1, \lambda_2\mathbf{v}_2, \ldots, \lambda_k\mathbf{v}_k] \\ &= [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k] \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} \\ &= P \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} \end{aligned}$$

If P is invertible we can obtain:

$$P^{-1}AP = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} \tag{8.17}$$

with $\lambda_1, \lambda_2, \ldots, \lambda_k$ be distinct. Now we want to solve:

$$\mathbf{z}_{t+1} = A\mathbf{z}_t.$$

We can go through the following steps: First multiply the above equation by P^{-1} and use equation (8.17) to obtain

$$\begin{aligned} P^{-1}\mathbf{z}_{t+1} &= P^{-1}A\mathbf{z}_t \\ &= \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} P^{-1}\mathbf{z}_t. \end{aligned}$$

Then, if we define $Z = P^{-1}\mathbf{z}$, we get the following *decoupled* system:

$$Z_{t+1} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} Z_t,$$

which, expanded out, can be written as:

$$\begin{bmatrix} Z_{1,t+1} \\ Z_{2,t+1} \\ \vdots \\ Z_{k,t+1} \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} \begin{bmatrix} Z_{1,t} \\ Z_{2,t} \\ \vdots \\ Z_{k,t} \end{bmatrix}. \tag{8.18}$$

The above form represents k separate difference equations that can be recursively solved to obtain

$$\begin{aligned}
Z_{1,t+1} &= \lambda_1 Z_{1,t} = \lambda_1^2 Z_{1,t-1} = \cdots = \lambda_1^{t+1} Z_{1,0} \\
Z_{2,t+1} &= \lambda_2 Z_{2,t} = \lambda_2^2 Z_{2,t-1} = \cdots = \lambda_2^{t+1} Z_{2,0} \\
&\vdots \\
Z_{k,t+1} &= \lambda_k Z_{k,t} = \lambda_k^2 Z_{k,t-1} = \cdots = \lambda_k^{t+1} Z_{k,0},
\end{aligned}$$

which, in matrix form, can be represented as

$$\begin{bmatrix} Z_{1,t+1} \\ Z_{2,t+1} \\ \vdots \\ Z_{k,t+1} \end{bmatrix} = \begin{bmatrix} \lambda_1^{t+1} & 0 & \cdots & 0 \\ 0 & \lambda_2^{t+1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k^{t+1} \end{bmatrix} \begin{bmatrix} Z_{1,0} \\ Z_{2,0} \\ \vdots \\ Z_{k,0} \end{bmatrix}. \tag{8.19}$$

Moreover, since $\mathbf{z} \equiv PZ$, the above system can be written as

$$\begin{aligned}
P \begin{bmatrix} Z_{1,t+1} \\ Z_{2,t+1} \\ \vdots \\ Z_{k,t+1} \end{bmatrix} &= P \begin{bmatrix} \lambda_1^{t+1} & 0 & \cdots & 0 \\ 0 & \lambda_2^{t+1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k^{t+1} \end{bmatrix} P^{-1} P \begin{bmatrix} Z_{1,0} \\ Z_{2,0} \\ \vdots \\ Z_{k,0} \end{bmatrix} \\
\begin{bmatrix} z_{1,t+1} \\ z_{2,t+1} \\ \vdots \\ z_{k,t+1} \end{bmatrix} &= P \begin{bmatrix} \lambda_1^{t+1} & 0 & \cdots & 0 \\ 0 & \lambda_2^{t+1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k^{t+1} \end{bmatrix} P^{-1} \begin{bmatrix} z_{1,0} \\ z_{2,0} \\ \vdots \\ z_{k,0} \end{bmatrix}.
\end{aligned}$$

If we define $\text{diag}(\lambda_1, \ldots, \lambda_t) = D$, we obtain

$$\mathbf{z}_{t+1} = PD^{t+1}P^{-1}\mathbf{z}_0.$$

Alternatively we can use (8.19) to express the solution as

$$\begin{bmatrix} Z_{1,t+1} \\ Z_{2,t+1} \\ \vdots \\ Z_{k,t+1} \end{bmatrix} = \begin{bmatrix} \lambda_1^{t+1} c_1 \\ \lambda_2^{t+1} c_2 \\ \vdots \\ \lambda_k^{t+1} c_k \end{bmatrix},$$

which, if we multiply both sides of the above result by the projection matrix P, implies

$$\begin{aligned}
\mathbf{z}_{t+1} &= [\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_k] \begin{bmatrix} \lambda_1^{t+1} c_1 \\ \lambda_2^{t+1} c_2 \\ \vdots \\ \lambda_k^{t+1} c_k \end{bmatrix} \\
&= \mathbf{v}_1 \lambda_1^{t+1} c_1 + \mathbf{v}_2 \lambda_2^{t+1} c_2 + \cdots + \mathbf{v}_k \lambda_k^{t+1} c_k,
\end{aligned}$$

where $\mathbf{v}_k \in \mathbb{R}^k$ represents the kth eigenvector and the constant $c_k \equiv Z_{k,0} = P^{-1}\mathbf{z}_{k,0}$. The following theorem summarizes this alternative approach.

Theorem 8.4.1 (Difference Equation with Real Roots)**.** *Let A be a $k \times k$ matrix with k distinct real eigenvalues $\lambda_1, \ldots, \lambda_k$ and corresponding eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$. The general solution of the system of difference equations is*

$$\mathbf{z}_{t+1} = \mathbf{v}_1 \lambda_1^{t+1} c_1 + \mathbf{v}_2 \lambda_2^{t+1} c_2 + \cdots + \mathbf{v}_k \lambda_k^{t+1} c_k. \tag{8.20}$$

Remark 8.4.1. Consider a system of two linear difference equations

$$\begin{aligned} x_{t+1} &= ax_t + by_t \\ y_{t+1} &= cx_t + dy_t \end{aligned}$$

or, in matrix form,

$$\mathbf{z}_{t+1} \equiv \begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} \equiv A\mathbf{z}_t.$$

If $b = c = 0$ in these equations, they are uncoupled:

$$\begin{aligned} x_{t+1} &= ax_t \\ y_{t+1} &= dy_t \end{aligned}$$

and are easily solved as two separate one-dimensional problems:

$$x_t = a^t x_0 \quad \text{and} \quad y_t = d^t y_0.$$

When the equations are coupled ($b \neq 0$ or $c \neq 0$), the technique for solving the system is to find a change of variables that *decouples* these equations. This is precisely the role of eigenvalues and eigenvectors.

Example 8.4.1. Consider the following coupled system of difference equations

$$\begin{aligned} x_{t+1} &= x_t + 4y_t \\ y_{t+1} &= 0.5x_t \end{aligned}$$

or, in matrix form,

$$\mathbf{z}_{t+1} \equiv \begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 4 \\ 0.5 & 0 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} \equiv A\mathbf{z}_t.$$

To find the eigenvalues solve the following

$$|A - \lambda I| = (\lambda - 2)(\lambda + 1) \stackrel{\text{set}}{=} 0,$$

which implies that $\lambda_1 = 2$ and $\lambda_2 = -1$ are the the eigenvalues to this system. To find the corresponding eigenvectors, row-reduce $A - 2I$ and $A + I$ to obtain the following equations

$$\begin{aligned} x_t &= 4y_t \\ 2x_t &= -4y_t. \end{aligned}$$

Normalizing y_t to 1, we get the following basis vectors that form the relevant projection matrix

$$P = \begin{bmatrix} 4 & -2 \\ 1 & 1 \end{bmatrix}.$$

Applying the change of variables $Z = P^{-1}\mathbf{z}$ to the original difference equation, we obtain

$$\begin{aligned} Z_{t+1} &= (P^{-1}AP)Z_t \\ &= \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} Z_t = \begin{bmatrix} \lambda_1^t & 0 \\ 0 & \lambda_2^t \end{bmatrix} Z_0 \\ &= \begin{bmatrix} 2^t & 0 \\ 0 & (-1)^t \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}. \end{aligned}$$

Thus, if we reapply the change of variables, we have

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = c_1 2^t \begin{bmatrix} 4 \\ 1 \end{bmatrix} + c_2(-1)^t \begin{bmatrix} -2 \\ 1 \end{bmatrix},$$

which is the same equation that we would have arrived at had we applied Theorem 8.4.1.

8.4.2 Solution Technique with Complex Eigenvalues

Consider the following two-dimensional system of equations

$$\mathbf{z}_{t+1} = A\mathbf{z}_t,$$

where A is a 2×2 matrix with complex eigenvalues $\alpha \pm i\beta$. Applying the change of variables $\mathbf{z} = PZ$ to the above difference equation yields

$$PZ_{t+1} = APZ_t \quad \rightarrow \quad Z_{t+1} = P^{-1}APZ_t.$$

Since A is assumed to have complex eigenvalues, it has corresponding complex eigenvectors $\mathbf{w}_1 = \mathbf{u} + i\mathbf{v}$ and $\mathbf{w}_2 = \mathbf{u} - i\mathbf{v}$. Thus, the projection matrix is given by

$$P = [\mathbf{w}_1, \mathbf{w}_2] = [\mathbf{u} + i\mathbf{v}, \mathbf{u} - i\mathbf{v}].$$

Moreover, using equation (8.17), it then follows that

$$P^{-1}AP = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} \alpha + i\beta & 0 \\ 0 & \alpha - i\beta \end{bmatrix}.$$

Thus, the decoupled system becomes

$$Z_{t+1} \equiv \begin{bmatrix} X_{t+1} \\ Y_{t+1} \end{bmatrix} = \begin{bmatrix} \alpha + i\beta & 0 \\ 0 & \alpha - i\beta \end{bmatrix} \begin{bmatrix} X_t \\ Y_t \end{bmatrix}.$$

Recursive substitution then yields

$$\begin{aligned} X_t &= k_1 (\alpha + i\beta)^t \\ Y_t &= k_2 (\alpha - i\beta)^t, \end{aligned}$$

where the constant $K \equiv [k_1, k_2]^T = Z_0 = P^{-1}\mathbf{z}_0$ could be real or complex. Using the fact that $\mathbf{z}_t = PZ_t$, we can transforming transform the variables into their original form to obtain

$$\begin{aligned} \mathbf{z}_t = \begin{bmatrix} x_t \\ y_t \end{bmatrix} &= [\mathbf{u} + i\mathbf{v}, \mathbf{u} - i\mathbf{v}] \begin{bmatrix} k_1 (\alpha + i\beta)^t \\ k_2 (\alpha - i\beta)^t \end{bmatrix} \\ &= k_1 (\alpha + i\beta)^t (\mathbf{u} + i\mathbf{v}) + k_2 (\alpha - i\beta)^t (\mathbf{u} - i\mathbf{v}). \end{aligned} \tag{8.21}$$

Notice that this solution takes the same form as (8.20), except with complex eigenvalues and eigenvectors replacing the real eigenvectors and eigenvalues.

Since the original problem contained only real numbers, we would like to find a solution that contains only real numbers. Since every solution of the system is contained in equation (8.21) for different choices of $K = [k_1, k_2]^T$, we want to know if we can find parameters k_1 and k_2 so that equation (8.21) is real.

Notice that except for the constant factors, the first term in equation (8.21) is the complex conjugate of the second. Since the sum of any complex number and its conjugate is the real number 2α, we want to choose the first constant, k_1, to be any complex constant $c_1 + ic_2$ and let the second constant, k_2, be its conjugate pair, $c_1 - ic_2$. Then the first and second term in (8.21) turn out to be complex conjugates and the sum of them will be a real solution given by

$$\begin{aligned}\mathbf{z}_t &= (c_1 + ic_2)(\alpha + i\beta)^t(\mathbf{u} + i\mathbf{v}) + (c_1 - ic_2)(\alpha - i\beta)^t(\mathbf{u} - i\mathbf{v}) \\ &= 2\,\mathrm{Re}\left\{(c_1 + ic_2)(\alpha + i\beta)^t(\mathbf{u} + i\mathbf{v})\right\}. \end{aligned} \tag{8.22}$$

Applying Demoivre's Formula (7.2), the above result can be written as

$$\begin{aligned}\mathbf{z} &= 2\,\mathrm{Re}\left\{(c_1 + ic_2)\, r^t\left[\cos(t\theta) + i\sin(t\theta)\right](\mathbf{u} + i\mathbf{v})\right\} \\ &= 2r^t\,\mathrm{Re}\left\{\left[(c_1\cos(t\theta) - c_2\sin(t\theta)) + i\,(c_2\cos(t\theta) + c_1\sin(t\theta))\right](\mathbf{u} + i\mathbf{v})\right\} \\ &= 2r^t\left[(c_1\cos(t\theta) - c_2\sin(t\theta))\,\mathbf{u} - (c_2\cos(t\theta) + c_1\sin(t\theta))\,\mathbf{v}\right], \end{aligned}$$

which is now a real solution.

Theorem 8.4.2. *Let A be a 2×2 matrix with complex eigenvalues $\alpha^* \pm i\beta^*$ and corresponding eigenvectors $\mathbf{u}^* \pm i\mathbf{v}^*$. Write eigenvalues in polar coordinates as $r^*\left[\cos(\theta^*) + i\sin(\theta^*)\right]$, where*

$$r^* = \sqrt{(\alpha^*)^2 + (\beta^*)^2} \quad \textit{and} \quad (\cos(\theta^*), \sin(\theta^*)) = \left(\frac{\alpha^*}{r^*}, \frac{\beta^*}{r^*}\right).$$

Then the general solution of the difference equation $\mathbf{z}_{t+1} = A\mathbf{z}_t$ is

$$\mathbf{z}_t = (r^*)^t\left[(c_1\cos(t\theta^*) - c_2\sin(t\theta^*))\,\mathbf{u}^* - (c_2\cos(t\theta^*) + c_1\sin(t\theta^*))\,\mathbf{v}^*\right].$$

Example 8.4.2. In Example 7.3.1, we found that the eigenvalues of

$$A = \begin{bmatrix} 1 & 1 \\ -9 & 1 \end{bmatrix}$$

are $1 \pm 3i$ with corresponding eigenvectors

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \pm i \begin{bmatrix} 0 \\ 3 \end{bmatrix}.$$

In polar coordinates, the eigenvalues become

$$1 + 3i = \sqrt{10}\left(\frac{1}{\sqrt{10}} + i\frac{3}{\sqrt{10}}\right) = \sqrt{10}\,(\cos\theta^* + i\sin\theta^*),$$

where $\theta^* = \arccos\left(\frac{1}{\sqrt{10}}\right) \approx 71.565°$ or 1.249 radians. The general solution for the system

$$\mathbf{z}_{t+1} \equiv \begin{bmatrix} x_{t+1} \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -9 & 1 \end{bmatrix}\begin{bmatrix} x_t \\ x_t \end{bmatrix}$$

is given by

$$\begin{aligned}\begin{bmatrix} x_t \\ y_t \end{bmatrix} &= (\sqrt{10})^t \left[(c_1 \cos(t\theta^*) - c_2 \sin(t\theta^*)) \begin{bmatrix} 1 \\ 0 \end{bmatrix} - (c_2 \cos(t\theta^*) + c_1 \sin(t\theta^*)) \begin{bmatrix} 0 \\ 3 \end{bmatrix} \right] \\ &= (\sqrt{10})^t \begin{bmatrix} c_1 \cos(t\theta^*) - c_2 \sin(t\theta^*) \\ -3c_2 \cos(t\theta^*) - 3c_1 \sin(t\theta^*) \end{bmatrix}.\end{aligned}$$

Remark 8.4.2. In higher dimensions, a given matrix can have both real and complex eigenvalues. The solution of the corresponding system of difference equations is the obvious combination of the solutions described in Equation 8.4.1 and Theorem 8.4.2.

Bibliography

ADDA, J. AND R. COOPER (2003): *Dynamic Economnics*, Cambridge, MA: The MIT Press.

BARRO, R. J. AND X. SALA-I-MARTIN (2003): *Economic Growth*, Cambridge, MA: The MIT Press, 2nd ed.

BEAVIS, B. AND I. M. DOBBS (1990): *Optimization and Stability Theory for Economic Analysis*, Cambridge, MA: Cambridge University Press.

BLANCHARD, O. AND S. FISCHER (1989): *Lectures on Macroeconomics*, Cambridge, MA: The MIT Press.

BROWN, J. AND R. CHURCHILL (2008): *Complex Variables and Applications*, New York, NY: McGraw-Hill, 8th ed.

CHIANG, A. C. AND K. WAINWRIGHT (2004): *Fundamental Methods of Mathematical Economics*, New York, NY: McGraw-Hill, 4th ed.

DE LA FUENTE, A. (2000): *Mathematical Methods and Models for Economists*, Cambridge, MA: Cambridge University Press.

DIXIT, A. K. (1990): *Optimization in Economic Theory*, New York, NY: Oxford University Press, 2nd ed.

EDWARDS, C. H. AND D. E. PENNEY (2008): *Differential Equations and Linear Algebra*, Upper Saddle River, NJ: Prentice Hall, 3rd ed.

FRIEDBERG, S. H., A. J. INSEL, AND L. E. SPENCE (2002): *Linear Algebra*, Upper Saddle River, NJ: Prentice Hall, 4th ed.

FRYER, M. J. AND J. V. GREENMAN (1987): *Optimisation Theory: Applications in OR and Economics*, Baltimore, MD: Edward Arnold.

HAMILTON, J. D. (1994): *Time Series Analysis*, Princeton, NJ: Princeton University Press.

K. SYDSÆTER, K., A. STROM, AND P. BERCK (2005): *Economists Mathematical Manual*, New York, NY: Springer, 4th ed.

LAY, S. R. (2004): *Analysis with an Introduction to Proof*, Upper Saddle River, NJ: Prentice Hall, 4rd ed.

LJUNGQVIST, L. AND T. J. SARGENT (2004): *Recirsive Macroeconomic Theory*, Cambridge, MA: The M.I.T. Press, 2nd ed.

MCCANDLESS, G. (2008): *The ABCs of RBCs: An Introduction to Dynamic Macroeconomic Models*, Cambridge, MA: Harvard University Press.

NOVSHEK, W. (1993): *Mathematics for Economists*, Cambridge, MA: Emerald Group Publishing Limited.

PEMBERTON, M. AND N. RAU (2007): *Mathematics for Economists: An introductory Textbook*, Manchester, UK: Manchester University Press, 2nd ed.

SARGENT, T. J. (1987): *Macroeconomic Theory*, Cambridge, MA: Emerald Group Publishing Limited, 2nd ed.

SIMON, C. P. AND L. BLUME (1994): *Mathematics for Economists*, New York, NY: W. W. Norton and Company.

STEWART, J. (2011): *Calculus: Early Transcendentals*, Belmont, CA: Thomson Learning, 7th ed.

STOKY, N., R. E. L. JR., AND E. C. PRESCOTT (1989): *Recursive Methods in Economic Dynamics*, Cambridge, MA: Harvard University Press.

SUNDARAM, R. K. (1996): *A First Course in Optimization Theory*, Cambridge, MA: Cambridge University Press.

SYDSÆTER, K. AND P. HAMMOND (2002): *Essential Mathematics for Economic Analysis*, Upper Saddle River, NJ: Prentice Hall.

SYDSÆTER, K., P. HAMMOND, A. SEIERSTAD, AND A. K. STROM (2005): *Further Mathematics for Economic Analysis*, Upper Saddle River, NJ: Prentice Hall.

TURKINGTON, D. A. (2007): *Mathematical Tools for Economists*, Malden, MA: Blackwell Publishing.

WICKENS, M. (2012): *Macroeconomic Theory: A Dynamic General Equilibrium Approach*, Princeton, NJ: Princeton University Press, 2nd ed.