Lecture Notes in Calculus & Optimisation

September 2020

Preparatory Course in Mathematics and Statistics

Aytek Erdil (nae25@cam.ac.uk)

These lecture notes are also available on the course moodle site. Please let me know if you spot errors, typos, etc. In addition to these notes you can consult a range of textbooks which cover this material. For example:

- Carl P. Simon & Lawrence Blume, *Mathematics for Economists*, W.W. Norton & Company.
- Alpha C. Chiang, *Fundamental Methods of Mathematical Economics*, McGraw-Hill.
- Andreu Mas-Colell, Michael Whinston & Jeremy Green, "Mathematical Appendix" in *Microeconomic Theory*, OUP.

An introduction to some of this material as well as the assumed background can be found in:

- Sydsaeter, K. & P. Hammond, *Essential Mathematics for Economic Analysis*, Prentice Hall.
- Pemberton, M. & N. Rau, *Mathematics for Economists*, Manchester University Press.

Contents

1 A few preliminaries

This section is not meant to be complete in any sense. Rather, it serves as a reminder of some of the notation we will use extensively.

1.1 Real numbers and the least upper bound property

We denote the set of real numbers by $\mathbb{R}$, and often visualise this set as a line which we will refer to as the **real line**.

A subset of $\mathbb{R}$ is called **bounded above** if there exists a number (an **upper bound**) which is greater than or equal to all elements of the set. Likewise, we call the set **bounded below** if there is a number (a **lower bound**) which is smaller than or equal to all elements of the set. If the set is bounded both above and below, we simply call the set **bounded**. Written formally, a set $S \subset \mathbb{R}$ is bounded if and only if there exists a number B such that $|x| \leq B$ for all $x \in S$.

$\mathbb{R}$ satisfies the so-called **least upper bound property** which states that:

> If a subset S of $\mathbb{R}$ is bounded above, then it has an upper bound which is smaller than all other bounds. That is, for every $S \subset \mathbb{R}$ which is bounded above, there exists a number b such that
>
> - $x \leq b$ for all $x \in S$, and
> - if $x \leq B$ for all $x \in S$, then $b \leq B$
>
> This upper bound b is called the least upper bound of S.

The least upper bound (**lub**) of S is also called its **supremum**. If the lub of S belongs to S, then it is necessarily the maximum element of S. It is not hard to see that every set bounded below necessarily has a **greatest lower bound** (also called **infimum**).

1.2 The set of ordered real pairs: $\mathbb{R}^2$

The set of ordered real pairs is the **Cartesian product** of $\mathbb{R}$ with itself:

$$\mathbb{R}^2 \quad = \quad \mathbb{R} \times \mathbb{R} \quad = \quad = \{(x_1, x_2) \mid x_1, x_2 \in \mathbb{R}\}$$

In words, this is the set of all ordered pairs (x_1, x_2) such that x_1 and x_2 are real numbers. The term "ordered" is to remind us that $(3, 5)$ is not the same as $(5, 3)$.

Similar to our visualisation of $\mathbb{R}$ as a line, we will visualise $\mathbb{R}^2$ as a plane, and on this plane we will often draw two lines explicitly:

- $\{(x_1, 0) \mid x_1 \in \mathbb{R}\}$ called the horizontal axis
- $\{(0, x_2) \mid x_2 \in \mathbb{R}\}$ called the vertical axis

We can appeal to plane geometry to extend our notion of distance from the real line (a visualisation of the set $\mathbb{R}$) to the real plane (a visualisation of the set $\mathbb{R}^2$). Using the geometric

notion of distance between two points $\boldsymbol{x} = (x_1, x_2)$ and $\boldsymbol{z} = (z_1, z_2)$, we define the distance between two points in $\mathbb{R}^2$ as:

$$\text{dist}((x_1, x_2), (z_1, z_2)) = \left((x_1 - z_1)^2 + (x_2 - z_2)^2\right)^{1/2}$$

1.3 The set of ordered n-tuples

If we can talk about pairs, why not also define n-tuples of real numbers? That is, ordered strings of n reals:

$$\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n) \mid x_i \in \mathbb{R} \text{ for all } i = 1, \ldots, n\}$$

A convention is to treat the cartesian product $\mathbb{R}^m \times \mathbb{R}^n$ as $\mathbb{R}^{m+n}$. This is quite intuitive since the first set($\mathbb{R}^m$) is the set of ordered m-tuples, and the second set ($\mathbb{R}^n$) is the set of ordered n-tuples. An ordered pair of an m-tuple and an n-tuple can be seen as an $m + n$-tuple:

$$\begin{aligned} ((x_1, \ldots, x_m), (\tilde{x}_1, \ldots, \tilde{x}_n)) &= (x_1, \ldots, x_m, \tilde{x}_1, \ldots, \tilde{x}_n) \\ &= (x_1, \ldots, x_m, x_{m+1}, \ldots, x_{m+n}) \qquad \text{where } x_{m+i} = \tilde{x}_i \text{ for all } i = 1, \ldots n \end{aligned}$$

1.4 $\mathbb{R}^n$ as a vector space

When we talk about bundles of n goods, we often index them as good 1, good 2, good 3, and so on. For example, apple is good 1, banana is good 2, etc. Then we denote the amount of goods in a bundle with a string of n numbers:

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$$

which indicates that the bundle $\boldsymbol{x}$ contains x_1 apples, x_2 bananas, and so on. We can refer to this n-tuple as an n-dimensional point which is an element of the n-dimensional real space $\mathbb{R}^n$. Now we would like to develop a bit more machinery to use these n-dimensional points effectively in various applications.

In order to take advantage of visual intuition, we will first introduce concepts for two-dimensional vectors, and then extend them to n dimensions.

Special case: the two-dimensional vector space $\mathbb{R}^2$

When we are dealing with two types of goods only, the bundles represented by pairs (x_1, x_2) enjoy a convenient geometric representation: points on the coordinate plane that depicts $\mathbb{R}^2$

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x_1, x_2) \mid x_1, x_2 \in \mathbb{R}\}$$

Another word for points in $\mathbb{R}^2$ is **vector**, and sometimes it is useful to visualise $\boldsymbol{x} = (x_1, x_2)$ as an arrow which begins at the origin $(0, 0)$ and ends at (x_1, x_2). While we can use the words point and vector interchangeably, the word vector can be particularly useful in highlighting the fact that we are talking about a tuple, not a single number.

For vectors $\boldsymbol{x}, \boldsymbol{y}$ in $\mathbb{R}^2$, and numbers c in $\mathbb{R}$ we define the following operations

- **vector addition:** $(x_1, x_2) + (y_1, y_2) = (x_1 + y_1, x_2 + y_2)$
- **scalar multiplication:** $c(x_1, x_2) = (cx_1, cx_2)$

The symbol c stands for an arbitrary number in $\mathbb{R}$ and in order to emphasise that it is a number, not a vector, we sometimes use the word **scalar** when we refer to numbers which we multiply with vectors.

With these operations of **vector addition** and **scalar multiplication**, the set $\mathbb{R}^2$ gets a structure, called **vector space**, which allows us not only to add but also subtract one vector from another. After all, subtracting $\boldsymbol{y}$ from $\boldsymbol{x}$ can be achieved by first multiplying $\boldsymbol{y}$ with the scalar -1 to obtain $-\boldsymbol{y}$, and then add $-\boldsymbol{y}$ with $\boldsymbol{x}$. The vector $(0, 0)$ is the identity element of vector addition, denoted $\mathbf{0}$, or simply 0 if we are lazy with notation.

More generally: vectors in $\mathbb{R}^n$

In the same fashion as above, we can define the following algebraic operations for vectors.

Vector addition. Given two vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$,

$$\boldsymbol{x} + \boldsymbol{y} \quad = \quad (x_1 + y_1, \ldots, x_n + y_n)$$

Scalar multiplication. Given a vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ and a scalar $c \in \mathbb{R}$,

$$c\boldsymbol{x} \quad = \quad \boldsymbol{x}c \quad = \quad (cx_1, \ldots, cx_n)$$

It is straightforward to verify that these operations satisfy the following properties:

- $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$ (Commutative Law)
- $\boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z}) = (\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z}$ (Associative Law)
- $\boldsymbol{x} + \mathbf{0} = \boldsymbol{x}$, where $\mathbf{0} = (0, \ldots, 0)$ is the zero vector
- $c(\boldsymbol{x} + \boldsymbol{y}) = c\boldsymbol{x} + c\boldsymbol{y}$ (Distributive Law)

The inner product (dot product) of vectors

The vector space $\mathbb{R}^n$ admits a third operation, whose geometric interpretation is less obvious, but is very convenient in capturing widely used operations in applications. Given two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, we define their **inner product** (also called the **dot product**) as

$$\boldsymbol{x} \cdot \boldsymbol{y} \quad = \quad x_1y_2 + x_2y_2 + \cdots + x_ny_n = \sum_{i=1}^{n} x_iy_i$$

Note the "dot" sign between $\boldsymbol{x}$ and $\boldsymbol{y}$. This is another reminder that $\boldsymbol{x}$ and $\boldsymbol{y}$ refer to vectors (not scalars). Secondly, the outcome of a dot product is not another vector in $\mathbb{R}^n$, but instead a scalar, i.e., a number in $\mathbb{R}$.

What is this good for? Well, if nothing, to express concisely the cost of a bundle. If $\boldsymbol{x} = (x_1, \ldots, x_n)$ stands for a bundle, and if $\boldsymbol{p} = (p_1, \ldots, p_n)$ is the vector of prices for the n goods we have in mind, then the cost of the bundle is nothing but the dot product of $\boldsymbol{p}$ with $\boldsymbol{x}$.

The following properties of the dot product are not hard to establish:

For all vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n$ and scalars $c \in \mathbb{R}$

- $\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{y} \cdot \boldsymbol{x}$
- $\boldsymbol{x} \cdot (\boldsymbol{y} + \boldsymbol{z}) = \boldsymbol{x} \cdot \boldsymbol{y} + \boldsymbol{x} \cdot \boldsymbol{z}$
- $(c\boldsymbol{x}) \cdot \boldsymbol{y} = c(\boldsymbol{x} \cdot \boldsymbol{y}) = \boldsymbol{x} \cdot (c\boldsymbol{y})$

Distance in $\mathbb{R}^n$. We simply adapt the two-dimensional geometric distance notion to the world of n-dimensional points and define the distance between two points $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathbb{R}^n$ as

$$\|\boldsymbol{a} - \boldsymbol{b}\| = \sqrt{(a_1 - b_1)^2 + \cdots + (a_i - b_i)^2 + \cdots + (a_n - b_n)^2}.$$

Note that if $\boldsymbol{a}$ and $\boldsymbol{b}$ are not the same points, the distance between them will be positive.

The origin. The point $\mathbf{0} = (0, 0, \ldots, 0)$ in $\mathbb{R}^n$ is the standard reference point of the n-dimensional real space, and as such deserves the special name: **the origin**. For brevity, we might occasionally denote this point as 0 with the understanding that it really corresponds to $\mathbf{0} = (0, \ldots, 0) \in \mathbb{R}^n$.

The norm of a point (vector) in $\mathbb{R}^n$. We refer to the distance between $\boldsymbol{x}$ and the origin as the **norm** of $\boldsymbol{x}$, and denote it by $\|\boldsymbol{x}\|$.

$$\|\boldsymbol{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}$$

1.5 Open and closed sets in $\mathbb{R}$

An interval is a particular kind of a subset of $\mathbb{R}$. Namely it is the set of numbers which lie between two specified points. An **open interval** (a, b) is the set of all real numbers greater than a and less than b. A **closed interval** $[a, b]$ is the set of all real numbers greater than or equal to a and less than or equal to b.

$$(a, b) = \{x \mid a < x < b\} \qquad [a, b] = \{x \mid a \leq x \leq b\}.$$

The following are also called intervals:

$$[a, b) = \{x \mid a \leq x < b\} \qquad (a, b] = \{x \mid a < x \leq b\}.$$

- For a given positive real number ϵ and a point a, we define the **ϵ-neighbourhood of a in $\mathbb{R}$** as the set of all points in $\mathbb{R}$ which are less than ϵ away from a. Another word for this set is the **open ball around a with radius ϵ**. Formally we can express this set as $\{x \in \mathbb{R} \mid |x - a| < \epsilon\}$. Equivalently this is nothing but the open interval $(a - \epsilon, a + \epsilon)$.
- A set $S \subseteq \mathbb{R}$ is called **open** if for every point in S, there exists an open ball around s all of which is also contained in S. That is, for every $s \in S$, there exists a positive number ϵ such that $(s - \epsilon, s + \epsilon) \subseteq S$.
- A set $C \subseteq \mathbb{R}$ is called **closed** if its complement in $\mathbb{R}$ is open. That is, C is closed if $\mathbb{R} \backslash C$ is open.
- A set $C \subset \mathbb{R}$ is called **compact** if it is closed and bounded.

An alternative way of defining a *closed* set is via the notion of *boundary points*. We say a point s is a **boundary point** of S if every ball around the point s contains some element of S as well as some element which does not belong to S. A set is called **closed** if it contains all of its boundary points. It is a good exercise to verify that this definition of a closed set of equivalent to the one given earlier.

1.6 Open and closed sets in $\mathbb{R}^n$

We can generalise some of the above concepts to higher dimensions. In fact the definitions of open and closed sets will be the same as they were for sets in $\mathbb{R}$.

The r-neighbourhood of a point in $\mathbb{R}^n$. The r-neighbourhood of a point a in $\mathbb{R}$ is the open interval $(a - r, a + r)$. Described in the terminology of distance, this is the set of all numbers whose distance from a is less than r. This latter distance formulation makes it really easy to define the notion of the **r-neighbourhood** of a point $\boldsymbol{a}$ in $\mathbb{R}^n$. That is, the set of all points in $\mathbb{R}^n$ whose distance from $\boldsymbol{a}$ is less than r. Denoting it by $N(\boldsymbol{a}, r)$, we can also write it as

$$\begin{aligned} N(\boldsymbol{a}, r) &= \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x} - \boldsymbol{a}\| < r\} \\ &= \left\{\boldsymbol{x} \in \mathbb{R}^n \mid \sqrt{(x_1 - a_1)^2 + \cdots + (x_n - a_n)^2} < r\right\} \\ &= \{\boldsymbol{x} \in \mathbb{R}^n \mid (x_1 - a_1)^2 + \cdots + (x_n - a_n)^2 < r^2\} \end{aligned}$$

Some people call this set **the open r-ball centred at $\boldsymbol{a}$**. The qualifier *open* refers to the fact that the surface of the ball (i.e., the set of points which are exactly r away from $\boldsymbol{a}$) is not included in the set. So, the r-ball centred at $a \in \mathbb{R}$ is the open interval $(a - r, a + r)$. The r-ball centred at $\boldsymbol{a} \in \mathbb{R}^2$ is the open disk in the real plane with the centre $\boldsymbol{a}$ and radius r.

A set $S \subseteq \mathbb{R}^n$ is called **open** if for every point in S, there exists an open ball around $\boldsymbol{s}$ all of which is also contained in S. That is, for every $\boldsymbol{s} \in S$, there exists a positive number ϵ such that $N(\boldsymbol{s}, \epsilon) \subseteq S$. A set $C \subseteq \mathbb{R}^n$ is called **closed** if its complement in $\mathbb{R}^n$, i.e., $\mathbb{R}^n \backslash C$ is open.

A set $S \in \mathbb{R}^n$ is **bounded** if it is contained in the "n-dimensional cube" $[-B, B]^n$ for some $B > 0$. In other words, there exists a number $B > 0$ such that for every $\boldsymbol{s} = (s_1, \ldots, s_n) \in S$, we have $|s_i| \leq B$ for all $i = 1, \ldots, n$.

Finally, a set $S \in \mathbb{R}^n$ is **compact** if it is closed and bounded.

We say a point $\boldsymbol{s}$ is an **interior** point of S if there exists open ball around $\boldsymbol{s}$ all of which is contained in S. The set of all interior points of S is called the **interior of** S, denoted by $\text{int}(S)$.

While it is not obvious how one can generalise the notion of an interval to sets in $\mathbb{R}^n$, one useful definition which allows some results for intervals in $\mathbb{R}$ to generalise to results in $\mathbb{R}^n$ is the following:

A set $K \subseteq \mathbb{R}^n$ is **convex** if for all $\lambda \in [0, 1]$,

$$\boldsymbol{a}, \boldsymbol{b} \in K \quad \Longrightarrow \quad \lambda \boldsymbol{a} + (1 - \lambda)\boldsymbol{b} \in K.$$

K is **strictly convex** if

$$\lambda \in (0, 1) \quad \text{and} \quad \boldsymbol{a} \neq \boldsymbol{b} \in K \quad \Longrightarrow \quad \lambda \boldsymbol{a} + (1 - \lambda)\boldsymbol{b} \in \text{int}(K).$$

2 Functions

Functions summarise/capture/report relationships between two or more "variables". Writing down a function which specifies/describes a relationship between one variable and another does not imply or assume causality; nor does it necessarily provide some structural explanation. A function simply keeps track of the relationship.

When we speak of a function, we have two sets and a "rule/mapping" in mind. That is, a function is not only the "rule" that describes what is mapped to what (i.e., the relationship), but also the specification of the domain and the codomain.

$$\begin{array}{rcccc} f & : & A & \longrightarrow & B \\ & & a & \longmapsto & b \end{array}$$

This notation means: for every element a of A, there is an element in B which is called the **image** of a under f. We denote by $f(a)$ the image of a under f. If $b = f(a)$, one can also say "f sends a to b" or "f maps a to b", etc.

The set A is called the **domain** of the function f. And the set B is called the **codomain**.

We can also talk about the **image of a set S under f**, which is nothing but the collection of the images of all elements in S. I.e., if S is a subset of the domain, then its image under f is

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

Given a function f, the image of its domain under f is called the **range** of f.

A function $f : A \to B$ is called **one-to-one** (or **1-1**, or **injective**) if for every $a, a' \in A$

$$f(a) = f(a') \quad \Longrightarrow \quad a = a'.$$

In other words, a function is one-to-one if and only distinct elements of its domain have distcint images.

A function $f : A \to B$ is called **onto** (or **surjective**) if for each $b \in B$, there exists $a \in A$ such that $f(a) = b$.

The **inverse image** of b under f is the set of elements in A which are mapped to b:

$$f^{-1}(b) = \{a \mid f(a) = b\}$$

If $f : A \to B$ is 1-1 and onto, we can talk about its **inverse function**, denoted f^{-1}, which is the unique function from B to A pinned down by

$$f(f^{-1}(b)) = b \quad \text{for all } b \in B \quad \text{and,} \quad f^{-1}(f(a)) = a \quad \text{for all } a \in A.$$

We say $f : A \to B$ and $f^{-1} : B \to A$ are the inverses of each other.

We say a set S has **infinitely many** elements (or more briefly S is an **infinite set**) if there exists a one-to-one function $f : \mathbb{N} \to S$, where $\mathbb{N}$ is the set of natural numbers.

Simple algebra of functions. If a set B admits basic algebraic operations such as addition and multiplication (e.g., if B is the set of real numbers), we can extend such rules to those functions from A to B. Given two functions $f : A \to B$ and $g : A \to B$,

- the sum of f and g is $f+g$, which is another function from A to B defined by $(f+g)(a) = f(a) + g(a)$ for all $a \in A$.

Likewise,

- the product of f and g is fg, which is another function from A to B defined by $(fg)(a) = f(a)g(a)$ for all $a \in A$
- the product of f with a scalar b in B is bf, which is another function from A to B defined by $(bf)(a) = bf(a)$ for all $b \in B$ and all $a \in A$.

Composition of functions. Given functions $f : A \to B$ and $g : B \to C$ we define the function $g \circ f : A \to C$ by setting

$$(g \circ f)(a) = g(f(a)) \qquad \text{for all } a \in A$$

For example, given a one-to-one and onto function $f : A \to B$, composing f with its inverse f^{-1} would yield an **identity function**:

$$f \circ f^{-1} = \mathrm{id}_B : B \to B \qquad \text{where } \mathrm{id}_B(b) = b \quad \text{for all } b \in B$$

and

$$f^{-1} \circ f = \mathrm{id}_A : A \to A \qquad \text{where } \mathrm{id}_A(a) = a \quad \text{for all } a \in A$$

Note that unless $A = B$, the functions $f \circ f^{-1}$ and $f^{-1} \circ f$ are not the same.

2.1 Sequences

A **sequence** with values in set A is nothing but a function $f : \mathbb{Z}_{>0} \to A$, where $\mathbb{Z}_{>0}$ stands for the set of positive integers. The custom is to use a different notation though. Instead of using parentheses, we will use subindices. For example, instead of writing $f(n)$, we can write a_n. (In this particular occasion, the choice of the letter a is motivated by the fact that f takes values in A. We could have used any other letter we wish.) Then the sequence will be denoted (a_n). We will refer to a_n as the n-th term of the sequence (a_n).

We will be dealing mainly with sequences with values in real numbers (or later in the course with values in $\mathbb{R}^k$ for $k > 1$).

If (n_k) is an increasing sequence of positive integers, then a_{n_k} is called a **subsequence** of (a_n). For example, if we set $b_n = a_{2n}$, then (b_n) would be a subsequence of (a_n).

A real-valued sequence is **non-increasing** if $a_{n+1} \leq a_n$ for all n. Likewise, we say a sequence is **non-decreasing** if $a_{n+1} \geq a_n$ for all n. Of course, a sequence can be neither. If a sequence is non-increasing or non-decreasing, then it is called **monotone**.

A real-valued sequence (a_n) is said to be **bounded** if there exists B such that $|a_n| \leq B$ for all n. Analogously, a sequence (a_n) with terms in $\mathbb{R}^n$ is called bounded if there exists B such that $\|a_n\| \leq B$ for all n.

2.2 $\mathbb{R} \to \mathbb{R}$ functions

Real-valued functions of one real variable is the class of functions for which we will develop most of our analysis. When we talk about these functions, we actually have larger class in mind, namely those real-valued functions whose domains are typically an interval or a union of intervals in $\mathbb{R}$. For example the function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ given by $f(x) = 1/x$.

Given a function $f : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$, we will say

- f is **non-decreasing** on D if for every $a, b \in D$, we have $a > b$ implies $f(a) \geq f(b)$
- f is **increasing** on D if for every $a, b \in D$, we have $a > b$ implies $f(a) > f(b)$
- f is **non-increasing** on D if for every $a, b \in D$, we have $a > b$ implies $f(a) \leq f(b)$
- f is **decreasing** on D if for every $a, b \in D$, we have $a > b$ implies $f(a) < f(b)$
- If f satisfies any of the above, it is called **monotonic**

We say $f : D \to \mathbb{R}$ is **bounded above** on $E \subseteq D$ if there exists a number $\bar{b}$ such that $f(x) \leq \bar{b}$ for all $x \in E$. Likewise, we say f is **bounded below** on $E \subseteq D$ if there exists a number $\underline{b}$ such that $f(x) \geq \underline{b}$ for all $x \in E$. If f is both bounded above and below, it is simply called bounded. For example, $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = -x^2$ is bounded above on $\mathbb{R}$, is bounded below on every finite interval $I \subset \mathbb{R}$, but is not bounded below on $\mathbb{R}$ or on $\mathbb{R}_-$ or on $\mathbb{N}$.

We say f is **single-peaked** on D if there exists p such that f is increasing on the left hand side of p, and decreasing on the right hand side of p. That is, for every $a, b \in D$, if $a < b \leq p$ then $f(a) \leq f(b)$; and if $p \leq a < b$, then $f(a) \geq f(b)$.

2.3 Graphs of $\mathbb{R} \to \mathbb{R}$ functions

Plotting the graph of a function from $\mathbb{R}$ to $\mathbb{R}$ amounts to creating a visualisation (a picture) of the following subset of $\mathbb{R}^2$

$$\text{Graph}(f) = \{(x, f(x)) \mid x \in \mathbb{R}\}$$

on the real plane.

Obviously, if the domain of f is $D \subset \mathbb{R}$, then its graph is the picture of

$$\text{Graph}(f) = \{(x, f(x)) \mid x \in D\}$$

on the real plane.

A function which has the form

$$f(x) = ax + b \qquad \text{where } a \neq 0$$

is also called a **linear function**, because if we plot its graph, that is, if we draw the picture of the following set in $\mathbb{R}^2$

$$\text{Graph}(f) = \{(x, ax + b) \mid x \in \mathbb{R}\}$$

in the real plane, we get a line that crosses the vertical axis at $(0, b)$ and the horizontal axis at $(-b/a, 0)$.

A **quadratic function** has the form

$$g(x) = mx^2 + nx + p \qquad \text{where } m \neq 0$$

and the graph of a quadratic function is called a **parabola**.

Note on graphs of inverse functions: If a function f maps a to b, and if f has an inverse function f^{-1}, then by definition f^{-1} maps b to a. Bringing this fact into the graphs, we can see that the graph of f^{-1} can be obtained by "reflecting" the graph of f about the 45^o line that goes through the origin. In other words, in the x-y coordinate plane, the graph of f is the mirror image of the graph of of f^{-1} if we think of the $x = y$ line as the mirror.

2.4 Polynomials and polynomial functions

An nth degree polynomial with real coefficients is the following object

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

where $a_0, a_1, \ldots, a_n \in \mathbb{R}$ such that $a_n \neq 0$. These numbers $a_n, \ldots, a_0$ are called the **coefficients** of the polynomial. a_n is called the **leading coefficient**.

Each polynomial $P = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ corresponds to a $\mathbb{R} \to \mathbb{R}$ polynomial function $P(x) : \mathbb{R} \to \mathbb{R}$ in an obvious manner which maps

$$x \longmapsto P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

A number $r \in \mathbb{R}$ is called a **real root** of a polynomial P if $P(r) = 0$.

If r is a real root of a polynomial P of degree n, then P is divisible by $(x - r)$, meaning it can be written as

$$P(x) = (x - r)Q(x)$$

where Q is an $(n-1)$st degree polynomial.

An nth degree polynomial P has at most n roots. If P has n distinct real roots $r_1, \ldots, r_n$ and its leading coefficient is a_n, then

$$P(x) = a_n(x - r_1)(x - r_2)\cdots(x - r_n)$$

The expression on the right hand side above is called the **factorisation** of polynomial P. Each $x - r_i$ is called a factor of P.

The nth binomial $(x + 1)^n$ is

$$x^n + \binom{n}{n-1} x^{n-1} + \binom{n}{n-2} x^{n-2} + \cdots + x^2 \binom{n}{2} + x \binom{n}{1} + 1$$

where

$$\binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k(k-1)(k-2)\cdots 1}$$

We read the above expression as "n choose k". Using the **factorial** notation

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

we can rewrite "n choose k" as

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

2.5 $\mathbb{R}^n \to \mathbb{R}$ functions

Our ability to visualise multi-variable functions is far more limited compared with the convenience of a coordinate plane where we could depict $\mathbb{R} \to \mathbb{R}$ functions (or at least the nicely behaving ones). Take for example a function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x, y) = x^2 y^3$$

The graph of this function is the following subset of $\mathbb{R}^3$:

$$\text{Graph}(f) = \{(x, y, z) \in \mathbb{R}^3 \mid z = f(x, y)\}$$

requiring three-dimensional imagery, which obviously is not as easy to achieve on a two-dimensional paper or board. And once we allow more than two variables, then the graph of the function will be an object which lives in an at least four dimensional world (three dimensions for three variables, and one dimension for the values of the function). As a result, our geometric intuition is likely (or perhaps certainly) to be limited. However, some qualities or properties satisfied by single-variable functions do hold for multi-variable functions, too, and we can extend some of our intuition from the single-variable domain to the multi-variable domain.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **homogeneous of degree** k if

$$f(\lambda x_1, \lambda x_2, \ldots, \lambda x_n) = \lambda^k f(x_1, x_2, \ldots, x_n)$$

for all $\lambda > 0$ and $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$.

The definition of a $\mathbb{R}^n \to \mathbb{R}$ function being bounded is similar to that of $\mathbb{R} \to \mathbb{R}$ functions. Given $D \subseteq \mathbb{R}^n$, we say $f : D \to \mathbb{R}$ is **bounded above** on $E \subseteq D$ if there exists a number $\bar{b}$ such that $f(\boldsymbol{x}) \leq \bar{b}$ for all $\boldsymbol{x} \in E$. Likewise, we say f is **bounded below** on $E \subseteq D$ if there exists a number $\underline{b}$ such that $f(\boldsymbol{x}) \geq \underline{b}$ for all $\boldsymbol{x} \in E$. If f is both bounded above and below, it is simply called bounded.

3 Limits

It is probably easier to introduce the formal notion of a limit in the context of $\mathbb{R}$-valued sequences before discussing limits of $\mathbb{R} \to \mathbb{R}$ functions.

3.1 Limit of a real-valued sequence

A real valued sequence (a_n) is said to **converge** to L if for each $\epsilon > 0$,

$$\text{there exists } K \text{ such that} \quad n > K \Longrightarrow |a_n - L| < \epsilon.$$

That is, however small the given ϵ is, eventually all terms of the sequence are within the ϵ-neighbourhood of L. Pay attention to the wording: after which term the sequence is trapped in to the ϵ-neighbourhood of L can (and will typically) depend on ϵ. The smaller the ϵ, the bigger K might need to be. But, the key is, for every $\epsilon > 0$, we can indeed find a K that works.

For example, suppose $a_0 = 1$ and $a_n = 1/\sqrt{n}$ for $n > 0$. It is not hard to guess that $a_n \to 0$. Let's prove this. Given any ϵ, can we find a corresponding K such that all terms of the sequence after the K-th term are in the ϵ-neighbourhood of 0? Well, what does it mean for a_n to be in the ϵ-neighbourhood of 0? It means

$$|a_n| < \epsilon$$

that is,

$$\frac{1}{\sqrt{n}} < \epsilon$$

which is equivalent to

$$n > \frac{1}{\epsilon^2}$$

If we choose K to be an integer larger than $\frac{1}{\epsilon^2}$, then any n larger than K is automatically larger than $\frac{1}{\epsilon^2}$, and hence $|a_n| < \epsilon$. Done!

We say L is the limit of sequence (a_n) as n approaches to infinity. There are two common notations to express this, and you will see both, sometimes in the same text:

$$\lim_{n\to\infty} a_n = L \qquad \text{and} \qquad a_n \to L \qquad \text{both mean the same thing.}$$

If a sequence does not converge to any L in $\mathbb{R}$, we say that the sequence **diverges**. But it is useful to distinguish the two types of divergence:

Diverging to ∞ or to $-\infty$.

We say that (a_n) diverges to **infinity** if

$$\text{for every bound } M > 0 \text{ there exists an index } K > 0, \text{ such that } n > K \Longrightarrow a_n > M.$$

Similarly, we say that (a_n) diverges to **minus infinity** if

$$\text{for every bound } M > 0 \text{ there exists an index } K > 0, \text{ such that } n > K \Longrightarrow a_n < -M.$$

Diverging without an eventual "trend". If a sequence diverges, but not to ∞ nor $-\infty$, then we might have even less to say about its "long-run trend". In order to illustrate this sort of divergence, consider for instance the sequence defined by $a_n = (-1)^n$ which alternates between 1 and -1.

3.2 Limit points of sets in $\mathbb{R}$ and sets in $\mathbb{R}^n$

Suppose S is a subset of $\mathbb{R}$. We say a number a is a **limit point** of set S if S has elements that are arbitrarily close to a. To put it more precisely: given any $\epsilon > 0$, S contains an element $s \neq a$ which is less than ϵ away from a. In fact, we can write it a bit more concisely: given any $\epsilon > 0$, there exists an $s \in S$ such that $0 < |s - a| < \epsilon$. Another word for a limit point is **cluster point**.[1]

The set of points which are less than ϵ away from a is called the **ϵ-neighbourhood of** a. Let's express the definition of a limit point in one more wording: a is a limit point of set S if for every $\epsilon > 0$, the ϵ-neighbourhood of a includes an element of S other than a. (*Exercise:* an equivalent definition of a limit point states that a is a limit point of S if and only if every neighbourhood of a includes infinitely many elements of S.)

Is it possible for a set S to have a limit point which does not belong to S?

Take, for example, the set of all positive real numbers, denoted $\mathbb{R}_{++}$ or $\mathbb{R}_{>0}$. Note that $1/n$ is in this set for every possible positive integer n. Secondly, note that 0 is not in this set. And finally, note that whatever $\epsilon > 0$ we are given, we can look at $1/\epsilon$. Since natural numbers grow without bound, there must exist a natural number n such that $n > 1/\epsilon$. But then, $\epsilon > 1/n$, which means $1/n \in \mathbb{R}_{>0}$ is less than ϵ away from 0. Hence 0 is a limit point of $\mathbb{R}_{>0}$ even though it is not an element of $\mathbb{R}_{>0}$.

In fact, we can extend this idea to the real number line. Note that the set of rational numbers (denoted $\mathbb{Q}$) is "dense" on this line in the sense that however narrow is an interval, we can always find rationals in there. That means, whatever point P given on the line, and whatever ϵ-neighbourhood of P we consider, there will always be rational numbers in that interval. So, P is a limit point of the set of rational numbers. Thus the limit points of the set $\mathbb{Q}$ is $\mathbb{R}$.

How about limit points of sets in $\mathbb{R}^n$? Using the standard notion of a neighbourhood in $\mathbb{R}^n$, the definition of limit points in $\mathbb{R}^n$ is the same as above.

Theorem. A set S in $\mathbb{R}^n$ is closed if and only if it contains all its limit points.

The above theorem is sometimes stated as the definition of "closed" in $\mathbb{R}^n$. If we were to work with this definition, then the statement "a set in $\mathbb{R}^n$ is closed if and only if it is the complement of an open set in $\mathbb{R}^n$" could be derived as a theorem. Recall that we had introduced this latter statement as the original definition.

[1] Note that we expressed the same thing three times in this paragraph. Which one sounds closest to daily language? Which one sounds least likely to cause multiple understandings?

3.3 Limit of a real-valued function of a single real variable

If we evaluate a function in smaller and smaller neighbourhoods of a point a, is it the case that the values are trapped in smaller and smaller neighbourhoods of a particular number? The existence of a limit of a function is concerned with this (intuitive but not not entirely precise) question.

We say the limit of a single-real-variable, real-valued function f as x approaches to a is L if for every given neighbourhood of L, there exists a corresponding neighbourhood of a such that every $x \neq a$ in the latter neighbourhood is mapped to the given neighbourhood of L.

In other words the **limit of a function f as x approaches to a is L** if the following statement holds:

Given whatever $\epsilon > 0$, there exists a $\delta > 0$ such that

$$0 < |x - a| < \delta \quad \Longrightarrow \quad |f(x) - L| < \epsilon$$

When the above statement holds, we write $\lim_{x \to a} f(x) = L$.

Note that the existence of this limit as x approaches to L makes no reference to the value of the function at a. It only requires those points "close to a" to be mapped "close to L". In fact, it is possible that f is not even defined at the point a.

The concept of limit is critical in formalising a key notion, used extensively in economics: that is, the notion of *marginal*. The language of the marginal requires us to look at what happens "at the margin" of a point of interest. For example, how does a consumer's taste change when she is "very close" to the point of spending all of her income? Note, however, that the adjectives "close" or "small" have no descriptive meaning on their own. They can only make sense in relative terms.

What do we mean, then, when we say "at the margin" or "for small changes"?

3.4 Limit of a function $f : \mathbb{R}^n \to \mathbb{R}^m$

Note that the definition can be phrased in the language of neighbourhoods, and thus can be extended to $\mathbb{R}^n$.

Suppose f is a function defined in a neighbourhood of $\boldsymbol{a} \in \mathbb{R}^n$ which takes values in $\mathbb{R}^m$. We say the limit of f as $\boldsymbol{x}$ approaches to $\boldsymbol{a}$ exists and is equal to $\boldsymbol{L}$ if:

For every given $\epsilon > 0$, there exists a $\delta > 0$ such that

$$0 < \|\boldsymbol{x} - \boldsymbol{a}\| < \delta \quad \Longrightarrow \quad \|f(\boldsymbol{x}) - \boldsymbol{L}\| < \epsilon$$

When the above statement holds, we write

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

Continuity

Limits allow us to make formal what it means to say an $\mathbb{R}^n \to \mathbb{R}^m$ function is continuous. We say such a function f is **continuous at the point $\boldsymbol{a}$** if $\lim_{\boldsymbol{x}\to\boldsymbol{a}} f(\boldsymbol{x}) = f(\boldsymbol{a})$.

3.5 Limits of a few familiar functions

Let's begin with everyone's favourite family of functions from $\mathbb{R}$ to $\mathbb{R}$, namely the *constant functions*. If a function takes the same value at every element of its domain, then it is called a **constant function**. When talking about functions from $\mathbb{R}$ to $\mathbb{R}$, a function which takes the value 5 at every point is also denoted by 5 even though this notation occasionally leads to confusion, because 5 happens to be the symbol for a number as well. Having said that this double usage of the same symbol both for a function and a number is more often convenient than confusing.

For example, if $f : \mathbb{R} \to \mathbb{R}$ is given by $f(x) = 0$, we will simply refer to this function as 0. Zero everywhere!

Fine. Do these functions have limits as x approaches to this or that point. Well, you might say obviously. The constant function 7 takes the value 7 at every x, and therefore it approaches to 7 at every point. Hard to disagree with that. However, if you insist on verifying this from the formal definition of a limit of a function, don't let me stop you.

OK, here is a more complicated function, the so-called **identity function** from $\mathbb{R}$ to $\mathbb{R}$ defined by $f(x) = x$. Once again, we might intuit that

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

and thus conclude that the identity function on $\mathbb{R}$ is continuous everywhere!

3.6 Basic algebra of limits (when the codomain is a subset of $\mathbb{R}$)

If $\lim_{x\to a} f(x) = L$ and $\lim_{x\to a} f(x) = M$, then

- $\lim_{x\to a}(f+g)(x) = L + M$.
- $\lim_{x\to a}(fg)(x) = LM$

If we also have $M \neq 0$, then

- $\lim_{x\to a}(f/g)(x) = L/M$

Combining these with just the knowledge of the limit of the identity function and constant functions, we can argue:

- Polynomial functions are continuous at every point of $\mathbb{R}$.
- Rational functions are continuous at every point where they are defined.

Another useful fact is the following: if (a_n) and (b_n) are two sequences such that $a_n \geq b_n$ and $a_n \to a$ and $b_n \to b$, then $a \geq b$.

3.7 Limits of composition of functions

Limits are all about getting close! They talk about tendencies. Say g tends to L as its argument tends to a. And say f tends to M as its argument tends to L.

Well, then what would you expect from $f \circ g$ as its argument tends to a:

$$f(\underbrace{g(\overbrace{\square}^{\text{goes to } a})}_{\text{goes to } L}) \longrightarrow M$$

In other words:

$$\lim_{x \to a} g(x) = L \quad \text{and} \quad \lim_{x \to L} f(x) = M \quad \Longrightarrow \quad \lim_{x \to a} f(g(x)) = M$$

Proof. Given $\epsilon > 0$, we'd like to show that there exists a δ such that

$$0 < |x - a| < \delta \implies |f(g(x)) - M| < \epsilon$$

First, we know that for the same ϵ given above, there exists a $\Delta > 0$ such that

$$0 < |x - L| < \Delta \implies |f(x) - M| < \epsilon$$

So what we need from δ is to make sure that $|g(x) - L| < \Delta$ for $0 < |x - a| < \delta$. But since $g(x) \to L$ as $x \to a$, it is indeed possible to find such δ. And therefore

$$0 < |x - a| < \delta \implies |g(x) - L| < \Delta \implies |f(g(x)) - M| < \epsilon$$

□

3.8 A quick detour around ∞

We have already formalised what it means for the argument of a function to go to ∞. However, we haven't given a clear definition of what it means to say a function is approaching to ∞ as its argument approaches to this or that.[2]

What is ∞? Is it a number? Does it really exist? No, and yes, respectively.

∞ is not a real number. That is, it is not a member of $\mathbb{R}$. Can we just add ∞ to this set and talk about a new and larger set of numbers? If we really want, of course we can, but we need to be careful, because ∞ does not behave like other numbers in algebraic operations. And, therefore, we can't really treat it as if it is a number as far as algebra goes in the way we have been used to. And so much of our analysis relies on our ability to carry out algebra as we know it so far.

Since we can't treat ∞ as we treat real numbers, let's just not force "numberhood" on ∞ and carry on. One way to think about ∞ is to observe that it summarises a **particular behaviour of a set of numbers**. As such ∞ is a property which really describes something about a set. What kind of set? What kind of behaviour? For example, if we say a sequence

[2]Likewise, we have formalised what it means for the terms of a sequence to go to ∞, but we haven't given a clear definition of what it means to say a sequence approaching to ∞

goes (diverges) to ∞, what we mean is that given whatever bound, the terms of the sequence will eventually exceed that bound. To make it more precise, given any $B > 0$, there exists k such that $a_n > B$ for all $n \geq k$. All terms of the sequence will be larger than that bound B, once we look at the kth term and beyond. What is k? That will depend on the sequence and the bound B. For example we might need a bigger k as B gets bigger. The key is that, for every given B, there exists such a k.

Adapting this formalism to limits of functions, we say f **goes (diverges) to infinity as** x **approaches to** a, and write

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

if

$$\text{for every given } B, \text{ there exists } \delta > 0 \text{ such that } 0 < |x - a| < \delta \Longrightarrow f(x) > B.$$

In other words, if you look at points close enough to a, the values of the function at those points will be guaranteed to exceed B. How about the point a? Well, remember again that the limit of f as x goes to a is a concept unrelated to what happens at the point a. Note that the requirement $f(x) > B$ is for those points which satisfy $0 < |x - a| < \delta$.

While we are at it, let's also make clear what it means to say "as the variable (or the argument) of the function goes to infinity, such and such happens". We say $f(x)$ **approaches to** L **as** x **approaches to** ∞, and write

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

if

$$\text{given any } \epsilon > 0, \text{ there exists } B \text{ such that } x > B \Longrightarrow |f(x) - L| < \epsilon.$$

Note. We never say $\frac{1}{0}$ is equal to ∞. Instead we would say "$\frac{1}{0}$ is not defined". Likewise, we would not (at least formally) write things like $\frac{1}{\infty} = 0$. We would say: as x goes to ∞, the expression $\frac{1}{x}$ goes to 0.

Limits of rational functions $P(x)/Q(x)$ as $x \to \infty$

Say $P(x)$ and $Q(x)$ are polynomial functions of degree m and n, respectively, so they have the form:

$$P(x) = a_m x^m + a_{m-1} x^{m-1} + \cdots + a_1 x + a_0 \quad \text{and} \quad Q(x) = b_n x^n + b_{n-1} x^{n-1} + \cdots + b_1 x + b_0$$

where a_m and b_n are non-zero.

The limit of $P(x)/Q(x)$ as $x \to \infty$ is

$$\lim_{x \to \infty} \frac{P(x)}{Q(x)} = \begin{cases} a_m/b_n & \text{if } m = n \\ 0 & \text{if } m < n \\ \infty & \text{if } m > n \text{ and } a_m/b_n > 0 \\ -\infty & \text{if } m > n \text{ and } a_m/b_n < 0 \end{cases}$$

By the way, you may treat all of these results as exercises and try to verify them.

3.9 Some key results about $\mathbb{R}$-valued continuous functions

Extreme value theorem in $\mathbb{R}$. A continuous real-valued function f defined on a closed interval $[a, b]$ is bounded. Moreover, it takes its minimum and maximum values over the interval.

In order to appreciate the above statement, note that it does not hold for an open interval. Consider (and draw the graphs for) the functions $f : (0, 1) \to \mathbb{R}$ with $f(x) = 1/x$; and $g : (0, 1) \to \mathbb{R}$ with $g(x) = x$. Now evaluate how the conclusion of the theorem fails for each function f and g.

More generally:

Extreme value theorem. Suppose K is a compact subset of $\mathbb{R}^n$ and $f : K \to \mathbb{R}$ is continuous. Then f is bounded, and it takes its minimum and maximum values over that compact domain. I.e., there exist $\boldsymbol{m}, \boldsymbol{M} \in K$ such that $f(\boldsymbol{m}) \leq f(\boldsymbol{x}) \leq f(\boldsymbol{M})$ for all $\boldsymbol{x} \in K$.

Intermediate Value Theorem. If f is continuous on $[a, b]$, then f takes all the values between $f(a)$ and $f(b)$.

For this theorem, the interval being closed is not of importance. One implication is that if f is defined over an interval I, and its maximum and minimum values are given by M and m, respectively, then f takes all the values between m and M. In other words, for every $\mu \in [m, M]$, there must be $c \in I$ such that $f(c) = \mu$.

A fixed point theorem. If $f : [0, 1] \to [0, 1]$ is a continuous function, then it has a fixed point, i.e., there exists a point $c \in [0, 1]$ such that $f(c) = c$.

More generally:

Brouwer's fixed point theorem. Let S be a non-empty, compact, convex subset of $\mathbb{R}^n$. If $f : S \to S$ is continuous, then f has a fixed point, i.e., there must be a point $\boldsymbol{c} \in S$ such that $f(\boldsymbol{c}) = \boldsymbol{c}$.

This theorem (which is a lot harder to prove) has important applications: used in showing the existence of equilibrium prices in an exchange market; Nash equilibrium in a non-cooperative game, etc.[3]

The algebra of continuity

Following the so-called algebra of limits, it is immediate to conclude that continuity is a property preserved by addition, multiplication, division, powers, inverses, and composition. To be more precise, if f and g are functions continuous at x, then so is $f \pm g$, fg and f^g. If, moreover, $g(x) \neq 0$, then f/g too is continuous at x. If f has an inverse over a neighbourhood of x, then the inverse of f is continuous at $f^{-1}(x)$. Finally, if f is continuous at x and g is continuous at $f(x)$, then $g \circ f$ is continuous at x.

[3]Sometimes a variant of Brouwer's fixed point theorem called Kakutani's fixed point theorem is used in economics, which allows f to be set-valued, for example when $f(x)$ describes the set of best responses to x.

3.10 Left/right limits

Recall that for an $\mathbb{R} \to \mathbb{R}$, we say the limit of f as x approaches to a is L if

$$\text{for any } \epsilon > 0, \text{ there exists } \delta > 0 \text{ such that } \quad 0 < |x - a| < \delta \quad \Longrightarrow \quad |f(x) - L| < \epsilon.$$

We write

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

and we also say **the limit of f as x goes to a is L.**

The above definition formalises the somewhat vague idea of $f(x)$ getting closer and closer to L as x gets closer and closer to a.

There is a less demanding property than having a limit, namely having a **left limit**, which captures the requirement that $f(x)$ should get closer and closer to L as x gets closer and closer to a while $x < a$. This is less demanding, because it doesn't put any discipline on those x greater than a, i.e., it doesn't require anything of $f(x)$ for those x greater than a.

Formally, we say that **the limit from left of f as x approaches to a is L** if

$$\text{for any } \epsilon > 0, \text{ there exists } \delta > 0 \text{ such that } \quad a - \delta < x < a \quad \Longrightarrow \quad |f(x) - L| < \epsilon.$$

We write

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

and we say, as x approaches to a from the left, $f(x)$ approaches to L.

Now, you write a formal definition for what the concept of **right limit** must be in the space below:

And, here's your exercise. Prove the following statement

Theorem. *$\lim_{x \to a} f(x)$ exists if and only if both $\lim_{x \to a^-} f(x)$ and $\lim_{x \to a^+} f(x)$ exist and are equal to each other.*

Example. *Illustrate a case in which both left and right limits exist at point a, but the limit as x approaches to a does not exist.*

Now, the natural thing to do would be to extend these concepts to continuity. Fill in the blanks in what follows:

Definition. We say that f is **left continuous** at a if

Definition. We say that f is **right continuous** at a if

Theorem. *Suppose that f is defined over an open interval I, and $a \in I$. f is continuous at a if and only if f is both left and right continuous at a.*

Example. *Illustrate a function which is defined over a neighbourhood of a, but is neither left, nor right continuous at a*

Example. *Illustrate a function which is defined over a neighbourhood of a, left continuous at a, but not right continuous at a.*

4 Differentiation (for $\mathbb{R} \to \mathbb{R}$ functions)

Assuming prior understanding of an intuitive notion of differentiation (of $\mathbb{R} \to \mathbb{R}$ functions) as a measure of *rate of change*, its connection with the graph and various properties of the function, we will quickly review the formal definition of derivatives in two different ways.

4.1 The derivative and linear approximations

The derivative of a function f with respect to its argument (variable) is a measure of the **rate of change** of $f(x)$ in relation to changes in x. For example, if the argument of the function increases from a to $a + h$, then the change in the value of the function is $f(a + h) - f(a)$. Looking at the rate of change is to have a measure of this change $f(a+h)-f(a)$ with respect (in proportion) to the change h in the argument of the function. That is, the derivative of the function aims to formalise the following ratio

$$\frac{f(a+h)-f(a)}{h}$$

In order to capture what the rate of change is exactly at the point a, we look at the above ratio for smaller and smaller changes denoted by h. And if by looking at smaller and smaller h, a unique number emerges, that is,

if

$$\lim_{h\to 0} \frac{f(a+h)-f(a)}{h} \quad \text{exists}$$

then we say this limit is the **derivative of** f **at** a and denote it with $f'(a)$.

Note that we can also write this limit as

$$\lim_{x\to a} \frac{f(x)-f(a)}{x-a}$$

Differentiability implies continuity. That is, if f is differentiable at a, then it must be continuous at a. Why is that? Well just have look at the definition. The denominator of the fraction goes to 0 as $x \to a$. If the fraction has any chance of having a limit, the numerator must also be going to 0. That is we must have $f(x) \to f(a)$ as $x \to a$, which is nothing but the definition of f being continuous at a.

Linear approximations. Here's another definition (or important interpretation) of derivatives. The function f having a derivative at the point a can also be interpreted as f having a "reasonable linear approximation" around the point a. What does a linear approximation mean, and what does it mean for it to be reasonable? We would like a linear function, that is, something of the form $Mx + N$ which is

"sufficiently close" to $f(x)$ when x is "close" to a

But what does that really mean? Surely, we should demand something better than "$f(x) - (Mx + N)$ goes to 0 as $x - a$ goes to 0", because even the constant function with $M = 0$ and $N = f(a)$ would satisfy that since f is continuous.

So we need the **error term** of our approximation

$$f(x) - (Mx + N)$$

not only to go to zero as x approaches a, but to do so at a speed an order of magnitude faster than $h = x - a$ does.

More specifically, we want

$$\frac{\text{the error term}}{h} \to 0 \quad \text{as } h \to 0$$

And this "speed of approximation" is possible if f is differentiable at a. The function f has a linear approximation around a, that is

$$f(a + h) \approx f(a) + f'(a)h$$

in the sense that

$$f(a + h) = f(a) + f'(a)h + \text{Error}(h) \quad \text{where} \quad \lim_{h \to 0} \frac{\text{Error}(h)}{h} = 0$$

An alternative definition of derivative. We say the derivative of f at the point a exists and is equal to α if

$$\frac{f(a + h) - f(a) - \alpha h}{h} \quad \to \quad 0 \qquad \text{as} \qquad h \to 0$$

The derivative of f: If we have the above limit for every x in a domain D of the function f, we say f is differentiable over the domain D. We write f' for the function which associates $f'(x)$ to x, and call this function f' the derivative of f.

Leibniz notation for derivatives: Derivative of f with respect to x can also be denoted as

$$\frac{df}{dx}$$

Sometimes, we'd like to keep track of the name of the variable x separately from the particular points at which we are analysing the function. This notation is useful:

$$\left.\frac{df}{dx}\right|_{x=a}$$

to stand for the derivative of f with respect to x evaluated at the point a. In other words: $f'(a)$. The advantage of the Leibniz notation is in reminding us what the differentiation variable is when lots of symbols are floating around; a reminder especially handy when we are dealing with functions of multiple variables.

4.2 The derivatives of two simple functions

The simplest of all $\mathbb{R} \to \mathbb{R}$ functions is a constant function. Is this function differentiable anywhere? Well, let's remember the definition of what it means to be differentiable at a point a, and rephrase that last question. Does the following limit exist:

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

The fact that f is a constant function implies, in particular, that $f(a+h) = f(a)$ whatever h is. Thus the above expression is nothing but

$$\lim_{h \to 0} \frac{0}{h} \quad = \quad \lim_{h \to 0} 0 \quad = \quad 0$$

So, yes, constant functions are differentiable everywhere, and their derivative is 0 everywhere. Makes sense: constant means to change. Hence zero rate of change!

Our next function of interest is the identity function on $\mathbb{R}$, that is $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x$. Is it differentiable anywhere? Just apply the definition, and ask if the following exists:

$$\lim_{h \to 0} \frac{a+h-a}{h} = \lim_{h \to 0} \frac{h}{h} = \lim_{h \to 0} 1$$

which obviously exists and is equal to 1 at every a. Thus, $f(x) = x$ is a differentiable function with the derivative being 1 everywhere.

4.3 An important result

If the derivative of a function f is 0 at every point of an interval $[a, b]$, then f must be constant over that interval. (This result can be proven using the Mean Value Theorem which we will mention later.)

4.4 Basic algebra of derivatives

We have already established that the derivative of a constant function is 0, and the derivative of x is 1.

Given f and g differentiable at a, we have the following convenient results:

- $(f+g)'(a) = f'(a) + g'(a)$
- if c is a constant, then $(cf)'(a) = cf'(a)$
- $(fg)'(a) = f'(a)g(a) + f(a)g'(a)$
- $\left(\frac{f}{g}\right)' = \frac{f'(a)g(a) - f(a)g'(a)}{(g(a))^2}$ as long as $g(a) \neq 0$.

Now, use these rules with the knowledge of the derivatives of constants and the identity function, and you can calculate the derivative of all polynomials (and rational functions)!

In particular

- $(x^n)' = nx^{n-1}$ for integers $n \geq 0$
- $\left(\frac{1}{x}\right)' = (x^{-1})' = -\frac{1}{x^2}$
- $\left(\frac{1}{x^n}\right)' = (x^{-n})' = -n\frac{1}{x^{n+1}}$ for integer $n < 0$

Actually, it looks like we can summarise all of the above in one expression

$$(x^n)' = nx^{n-1} \qquad \text{for all integers } n$$

Note: We must acknowledge that we have been a bit lazy in our notation above. In particular, when we were talking about the function $f(x) = x^n$, we just wrote x^n. In doing so, we are keeping in mind that x stands for a variable. These kinds of shortcuts in notation are convenient, but sometimes potentially confusing.

4.5 Higher order derivatives

If f is differentiable everywhere in a domain, then we can talk about its derivative f' as another function on this domain. Perhaps f' is also differentiable. If so, the derivative of f' is called the second derivative of f. In general, if exists, we can talk about the nth derivative of f and denote it by $f^{(n)}$, so

$$f^{(n)}(x) = (f^{(n-1)}(x))'$$

i.e., the nth derivative of f is the derivative of the $(n-1)$st derivative of f.

Note, for example, that the kth derivative of the function $f(x) = x^n$ with $n > 0$ exists:

$$\text{If} \quad f(x) = x^n, \quad \text{then} \quad f^{(k)}(x) = \begin{cases} n(n-1)\cdots(n-k+1)x^{n-k} & \text{if } 0 < k \leq n \\ 0 & \text{if } 0 < n < k \end{cases}$$

4.6 The chain rule

Given functions $f : A \to B$ and $g : B \to C$, remember that the composed function $g \circ f : A \to C$ is defined as

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

The Chain Rule. If f is differentiable at x, and g is differentiable at $f(x)$, then $g \circ f$ is differentiable at x and

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

The idea behind the chain rule. How does this come about? Since this is a rule we will be using over and over and over, it is worth developing a feel for why it indeed holds. It is actually fairly intuitive.

For small δ, we have

$$f(x+\delta) - f(x) \approx f'(x)\delta \qquad \text{and therefore} \qquad f(x+\delta) \approx f(x) + f'(x)\delta \qquad (\star)$$

Likewise, for small Δ we have

$$g(y+\Delta) - g(y) \approx g'(y)\Delta \qquad \text{and therefore} \qquad g(y+\Delta) \approx g(y) + g'(y)\Delta \qquad (\star\star)$$

First, g is continuous, so when δ is small, we can appeal to $(\star)$ to infer

$$g(f(x+\delta)) \approx g(f(x) + f'(x)\delta)$$

If δ is small, we can treat $f'(x)\delta$ as the small Δ in $(\star\star)$, and set $y = f(x)$, so $(\star\star)$ becomes

$$g(f(x) + f'(x)\delta) \approx g(f(x)) + g'(f(x))f'(x)\delta$$

Combining the last two approximate equations:

$$g(f(x+\delta)) \approx g(f(x)) + g'(f(x))f'(x)\delta$$

Rearranging this yields

$$g(f(x+\delta)) - g(f(x)) \approx g'(f(x))f'(x)\delta$$

and therefore we must have $(g \circ f)'(x) = g'(f(x))f'(x)$. ◇

4.7 L'hôpital's rule and variations to compute limits

Given two functions f and g such that

- $\lim_{x\to a} f(x) = L$, where $L \in \mathbb{R}$ or $L = \pm\infty$
- $\lim_{x\to a} g(x) = M$, where $M \in \mathbb{R}$ or $M = \pm\infty$

what can we say about $\lim\limits_{x\to a} \dfrac{f(x)}{g(x)}$ if $L = M = 0$? Or if $L = \pm\infty$ and $M = \pm\infty$.

L'Hôpital's rule. Suppose two functions f and g are differentiable on an open interval I except possibly at a contained in I. Suppose also that $g'(x) \neq 0$ for all $x \in I$ with $x \neq a$. If $\lim\limits_{x\to a} f(x) = \lim\limits_{x\to a} g(x) = 0$ or $\lim\limits_{x\to a} f(x) = \lim\limits_{x\to a} g(x) = \infty$, and if $\lim\limits_{x\to a} \dfrac{f'(x)}{g'(x)}$ exists, then

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

In the same fashion, here's a list of strategies to deal with limits which resemble one of

$$\infty^0 \quad , \quad \infty - \infty \quad , \quad 0^0 \quad , \quad 1^\infty$$

- For fg where $f \to \infty$ and $g \to 0$, look at $\dfrac{f}{1/g}$ or $\dfrac{g}{1/f}$ instead.
- For f^g where $f \to \infty$ and $g \to 0$, look at $\ln(f^g) = g \ln f$ (can be viewed as $\frac{g}{1/\ln f}$ or $\frac{\ln f}{1/g}$).
- For $f - g$ where $f \to \infty$ and $g \to \infty$, first try to simplify and see if you can arrive at an expression of the form $u(x)/v(x)$. Having exhausted options, have a look at $\exp(f-g) = \exp f / \exp g$.
- For f^g where $f \to 0$ and $g \to 0$, look at $\ln(f^g) = g \ln f$ (can be viewed as $\frac{g}{1/\ln f}$ or $\frac{\ln f}{1/g}$).
- For f^g where $f \to 1$ and $g \to \infty$, look at $\ln(f^g) = g \ln f$ (can be viewed as $\frac{g}{1/\ln f}$ or $\frac{\ln f}{1/g}$).

In other words, try to convert the limit question to one that looks like $0/0$ or ∞/∞ to be able to apply L'hôpital's rule.

4.8 Differential of an $\mathbb{R} \to \mathbb{R}$ function

The differential of a single-variable, real-valued function f is a real-valued function of two real variables given by

$$df(x,h) \quad = \quad f'(x)h$$

Note that the expression df is stands for an $\mathbb{R}^2 \to \mathbb{R}$ function. It is common to use the notation Δx for the second variable, so the defining equation can be rewritten as

$$df(x,\Delta x) \quad = \quad f'(x)\Delta x$$

The most common intuitive interpretation treats Δx to be a "small" quantity, whatever "small" means in the mind of the interpreter.

The above definition is useful in understanding the formal foundation of differentials, but the usage often reduces the notation to

$$df(x) \quad = \quad f'(x)dx$$

even though this last expression obscures the fact that df is a two-variable function. For application purposes, since the second variable (denoted by h in the original definition) is usually treated to be an "infinitesimal" value (again, whatever infinitesimal might mean), this last notation (where we write dx instead of h) prevails. When we see this notation of dx, we must remember that this two-letter symbol stands for a quantity which often captures a notion of a "small change" in the argument of the function f.

Hence, the differential $f'(x)dx$ is meant to capture an approximation to the change in the value of $f(x)$ as a result of a change dx in the argument of f. That is:

$$f(x+dx) - f(x) \quad \approx \quad df(x) \quad = \quad f'(x)dx$$

When we do algebraic operations with differentials, we think of dx as "infinitesimally small" and evaluate the rate of "infinitesimally small" changes in the value of function with respect to the "infinitesimally small" changes in x.

Rules for differentials

Suppose a and b are constants, whereas f and g are $\mathbb{R} \to \mathbb{R}$ functions. Then

- $d(af + bg) = a\ df + b\ dg$
- $d(fg) = g\ df + f\ dg$
- $d\left(\dfrac{f}{g}\right) = \dfrac{g\ df - f\ dg}{g^2}$ whenever $g \neq 0$.

5 Some things we learn from the first derivative

5.1 The first order condition for optimisation

Recall the extreme value theorem which states that if f is a continuous function on $[a, b]$, then f attains its maximum and minimum values over this interval. That is, there exist points c and d in $[a, b]$ such that $f(c) \leq f(x) \leq f(d)$ for all $x \in [a, b]$.

Often the point d is referred to as "the maximum" of f over the interval $[a, b]$. Clearly, d is not the maximum value attained by f, but rather the point at which the function $f : [a, b] \to \mathbb{R}$ attains its maximum value. Likewise, c is referred to as "the minimum" of f.

In addition to these two points, also of interest are the notions of the *local maxima* and *local minima*. Formally, a point m is called a **local minimum** of f if there exists a neighbourhood of m within which the function f takes its minimum value at m. Likewise, a point M is called a **local maximum** of f if there exists a neighbourhood of M within which the function f takes its maximum value at M.

Now suppose f is also differentiable on (a, b). If those points c and d are not the end points, that is, if they are in (a, b), let's look at the rate of change of f around c and d.

To begin with, look at c first. The fact that f is minimised at c when x varies over $[a, b]$ implies that

$$f(c + h) - f(c) \geq 0 \qquad \text{for all } h \neq 0$$

Dividing this expression by $h \neq 0$:

$$(\star) \qquad \frac{f(c + h) - f(c)}{h} \geq 0 \quad \text{for } h > 0$$

$$(\star\star) \qquad \frac{f(c + h) - f(c)}{h} \leq 0 \quad \text{for } h < 0$$

The fact that

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

exists means both $(\star)$ and $(\star\star)$ approach to that same limit as h approaches to 0. (In the case of $(\star)$, we are talking about positive values of h approaching to 0. In the case of $(\star\star)$, we are talking about negative values of h approaching to 0.)

But $(\star)$ always stays ≥ 0 because the numerator is always nonnegative, and the denominator is always positive. Therefore it cannot approach to something negative. Likewise, $(\star\star)$ always stays ≤ 0 because the numerator is always non-positive, and the denominator is always positive. Thus it cannot approach to something positive. So, given that both approach to the same limit (which is $f'(c)$), that limit has to be 0.

We can tell a similar story for the point $d \in (a, b)$ where the function is maximised as x varies over $[a, b]$. The only difference with the above analysis would be that $(\star)$ would now be ≤ 0, whereas $(\star\star)$ would be ≥ 0. But the same logic will apply: if these expressions converge to the same number, i.e., $f'(d)$, that number has to be 0.

In fact we can reach this conclusion at any point $e \in (a, b)$ as long as e minimises or maximises a differentiable function in a neighbourhood around e. In other words, e does not

have to be the point at which f is minimised or maximised for the above analysis to hold. It is sufficient for e to be an **interior local minimum** or an **interior local maximum**. How so? Well, suppose f takes its minimum value over $(e-\delta, e+\delta)$ at the point e. Then we can carry out all of the above reasoning, because as we look at smaller and smaller h, the points $e-h$ and $e+h$ are eventually in the δ-neighbourhood of e. Thus we have the conclusion which is known as the **first order condition** for a function's minima and maxima.

Theorem. Suppose a function is differentiable on some open interval containing the point d. If d is a local minimum or a local maximum, then $f'(d) = 0$.

Rolle's Theorem. If f is continuous on $[a, b]$, differentiable on (a, b), and $f(a) = f(b)$, then there exists a point $c \in (a, b)$ such that $f'(c) = 0$.

This can be seen as an application of the EVT and the FOC. If f is constant over $[a, b]$, then its derivative is zero for all $x \in (a, b)$. If it is not constant, then there exist points in (a, b), where f takes greater or smaller values than $f(a) = f(b)$. Say it takes higher values. Since f is continuous, it attains its maximum value over this domain at some point, and that point cannot be a or b since f takes higher values somewhere between a and b. Say $c \neq a, b$ is where f takes its maximum value. The FOC implies that $f'(c) = 0$ □

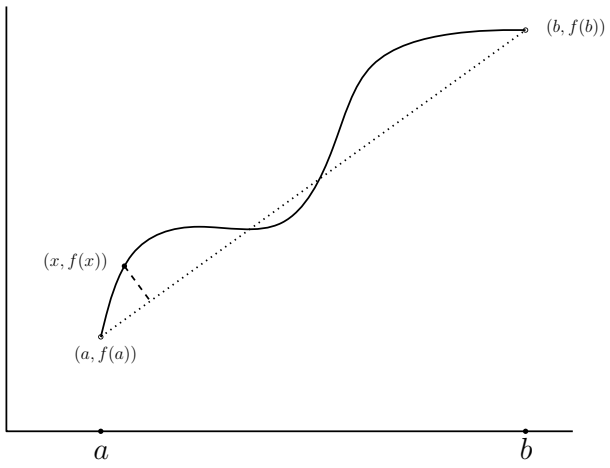
The mean value theorem

Suppose we drove from Cambridge to London. The journey was exactly 60 miles long and it took us exactly 1 hour to complete this journey. The speedometer must have shown 60 mph at some point of the journey. That's what the mean value theorem (MVT) says. Denoting an hour with the unit interval $[0, 1]$, let $x \in [0, 1]$ stand for time since departure, and let $f(x)$ stand for the distance travelled in the first x hours of the journey. So we have $f(0) = 0$ and $f(1) = 60$. The speed of the car at time x is the rate of change of the distance travelled with respect to time. That is $f'(x)$. The average speed over the whole journey is $(f(1) - f(0))/1 = 60$. To paraphrase the conclusion of the MVT, the average (i.e., mean) speed of the car is attained at some point during the journey: there exists c such that $f'(c) = 60$.

Mean Value Theorem. Let f be continuous over $[a, b]$, and differentiable over (a, b). There exists $c \in (a, b)$ such that $f'(c) = \frac{f(b)-f(a)}{b-a}$.

The theorem suggests a point where the slope of the tangent to the graph is the same as the slope of the dotted line that connects $(a, f(a))$ and $(b, f(b))$. It's not hard to convince yourself that the points on the graph whose distance from that dotted line is maximised (or even locally maximised) will be the points where the tangent will be parallel to the dotted line. (The dashed line segment marked on the graph represents the distance from the point $(x, f(x))$ on the graph to the dotted line. his is clearly a point which does not maximise the distance to the dotted line. The tangent at this point being steeper than the dotted line suggests that moving northeast further along the curve will increase the distance to the dotted line.)

Exercise: If the derivative of a function f is 0 at every point of an interval $[a, b]$, then f must be constant over that interval.



5.2 The monotonicity of a function and the sign of its derivative

The derivative, remember, captures the rate of change of f with respect to its argument. So, it should not be surprising that it carries some information regarding whether f is increasing or decreasing. The following conclusions are fairly intuitive (though they should ideally be verified):

- If $f' > 0$ on an interval (a, b), then f is increasing on (a, b).
- If $f' < 0$ on an interval (a, b), then f is decreasing on (a, b).
- If f is non-decreasing on an interval, then $f' \geq 0$ on that interval.
- If f is non-increasing on an interval, then $f' \leq 0$ on that interval.
- f is constant on (a, b) if and only if $f' = 0$ on (a, b).

5.3 Concavity and convexity of $\mathbb{R} \to \mathbb{R}$ functions

An $\mathbb{R} \to \mathbb{R}$ function f is said to be **concave** over an interval I if for every $a, b \in I$ such that $a \neq b$, and any $\lambda \in (0, 1)$:

$$f(\lambda a + (1 - \lambda)b) \quad \geq \quad \lambda f(a) + (1 - \lambda) f(b).$$

Intuitively, concavity of f on interval I is equivalent to the following geometric property of its graph: whichever two points you pick on the graph of $f : I \to \mathbb{R}$, and connect them with a line segment, the graph of the function will lie *above* that line segment.

If the above inequality is strict for all $a \neq b \in I$ and all $\lambda \in (0, 1)$, we say f is strictly concave over I.

A function $f(x)$ is said to be **convex** over an interval I if for every $a, b \in I$ such that $a \neq b$, and any $\lambda \in (0, 1)$:

$$f(\lambda a + (1 - \lambda)b) \quad \leq \quad \lambda f(a) + (1 - \lambda) f(b).$$

Intuitively, the graph of the function between $x = a$ and $x = b$ lies *below* the line segment connecting $(a, f(a))$ and $(b, f(b))$.

If the above inequality is strict for all $a \neq b \in I$ and all $\lambda \in (0, 1)$, we say f is strictly convex over I.

It is not hard to verify that f is concave if and only if $-f$ is convex. (Treat this as an exercise in algebra to practice your skills in writing the conditions for concavity and convexity.) Hence we can formulate our statements for concave functions, and appropriate versions for convex functions should follow by changing the signs in the right places, replacing maximums with minimums, and so on.

Theorem. Let f be a continuously differentiable function defined on an interval I. f is concave on I if and only if

$$f(b) - f(a) \quad \leq \quad f'(a)(b - a) \qquad \text{for all } a \text{ and } b \text{ in } I. \tag{1}$$

Theorem. Let I be an open interval and $f : I \to \mathbb{R}$ be twice differentiable such that f'' is continuous. f is concave on I if and only if $f''(x) \leq 0$ for all $x \in I$.

So, we learn something simply from the sign of the second derivative:

- If $f'' < 0$ on an interval, then f is strictly concave over that interval.
- If $f'' > 0$ on an interval, then f is strictly convex over that interval.

5.4 Classifying stationary points of $\mathbb{R} \to \mathbb{R}$ functions

Stationary points of a single-variable real-valued differentiable function are those points at which the derivative is equal to zero. Each such point corresponds to one of the following:

- A local minimum (At this point $f'' \geq 0$.)
- A local maximum (At this point $f'' \leq 0$.)
- An inflection point, if it is neither a local min or a local max.

Knowing the second derivative at a stationary point can help identify the nature of that stationary point. Suppose $f'(a) = 0$.

- If $f''(a) > 0$, then a is a local minimum.
- If $f''(a) < 0$, then a is a local maximum.
- If $f''(a) = 0$, then no conclusion can be drawn on the basis of f'' alone.

These conditions are also referred to as the **second-order conditions (SOC)** for a single-variable function.

Theorem. Let f be a differentiable, concave function on the whole real line. Then $f'(m) = 0$ if and only if m is a global maximum.

6 Series

Recall that a real-valued sequence is nothing but a function $a : \mathbb{N} \to \mathbb{R}$, where $\mathbb{N}$ stands for the set of natural numbers. Since the standard notation for sequences uses subindices instead of parentheses, we will write a_n instead of $a(n)$. Moreover, we will refer to a_n as the nth term of the sequence (a_n) instead of calling it the value of the function a at the point n.

Associated with the notion of a real-valued sequence (a_n), we sometimes would like to talk about the sum of its terms. But there are infinitely many terms, and it is not obvious how we add up infinitely many numbers in a meaningful way. In fact an attempt to add up all terms of a sequence will certainly fail at times. Take, for example, a constant sequence whose terms are all equal to 1. How can we talk about adding up infinitely many ones and expect to get a number as an answer?

Before we jump to any major conclusion, let's think how we may try to conceptualise a notion of adding up infinitely many terms of a sequence. Remember that *infinity*, to begin with, was a property of a set. Namely, the property of not being finite, not ever ending when we tried to count the elements of a set. But then we talked about the limit of a sequence (s_n) as n approaches to infinity. There is no such thing as s_∞, but if the sequence (s_n) is convergent, then there is such a thing as $\lim_{n\to\infty} s_n$.

Ah, maybe we should formalise what we mean by "adding up infinitely many numbers" as a "limit of adding up finitely many numbers". The terms of the sequence are neatly indexed and it feels only natural to list them beginning from the lowest index to higher indices

$$a_0, a_1, a_2, a_3, \ldots$$

so perhaps we have in mind a summation which looks like

$$a_0 + a_1 + a_2 + a_3 + \cdots$$

Of course, this last expression doesn't make sense yet since it suggests adding up infinitely many terms. On the other hand, it also suggests which "finite groups of numbers" we might begin thinking about when we are trying to make sense of an "infinite sum". Let's do the summation step-by-step, adding one term at a time, going from left to right. And let's denote the answer to step k by S_k so we have

$$\begin{aligned}
S_0 &= a_0 \\
S_1 &= a_0 + a_1 \\
S_2 &= a_0 + a_1 + a_2 \\
S_3 &= a_0 + a_1 + a_2 + a_3 \\
&\vdots \quad \vdots \quad \vdots \\
S_n &= a_0 + a_1 + a_2 + a_3 + \cdots + a_n
\end{aligned}$$

As n gets bigger and bigger, we have more and more terms added up, and we will think of the desired summation of all terms of (a_n) as the limit of these finite sums. We will refer to

$$\sum_{i=0}^{\infty} a_i = a_0 + a_1 + a_2 + a_3 + \cdots$$

as the **series** whose nth term is a_n, and whose nth partial sum is S_n.

Clearly (S_n) is a sequence on its own account, and we can talk about the limit of S_n as n approaches to infinity. This is what the series is meant to capture. If S_n converges to L, then we say the series $\sum_{i=0}^{\infty} a_i$ converges to L. If S_n diverges to ∞, we say the series $\sum_{i=0}^{\infty} a_i$ diverges to ∞. And so on.

For example, if $a_i = 1$ for all $i = 1, 2, \ldots$, then $\sum a_i$ diverges to ∞. If $b_i = (-1)^i$, then $\sum b_i$ diverges because its partial sums alternate between 1 and 0. Does a series ever converge? Well, if $z_i = 0$, then $\sum z_i$ converges to 0. Sure, but is there any series which involves adding up non-zero terms and still converges?

6.1 Geometric sequences and series

A geometric sequence is one where the consecutive terms have a fixed ratio, that is, when there exists a constant c such that

$$\text{for every } n \in \mathbb{N}, \text{ we have } \quad \frac{a_{n+1}}{a_n} = c$$

The sort of expressions which relate the $n+1$st term of a sequence to its nth (and perhaps a few earlier terms) are sometimes referred to as a *recursive expression*. For example, we say the expression $a_{n+1} = ca_n$ describes the sequence (a_n) recursively. If we also know the value of a_0, then we can figure out every term of the sequence.[4] In order to see why, simply observe that

$$\begin{aligned} a_1 &= c \times a_0 \\ a_2 &= c \times a_1 = c \times c \times a_0 = c^2 \times a_0 \\ a_3 &= c \times a_2 = c \times c^2 \times a_0 = c^3 \times a_0 \\ &\vdots \quad\quad \vdots \\ a_n &= c^n \times a_0 \end{aligned}$$

One conclusion of the above argument also provides us with an alternative definition of what it means for a sequence to be a geometric sequence. A geometric sequence is one whose terms can be listed as:

$$a, ac, ac^2, ac^3, \ldots$$

Convergence properties of a geometric sequence are fairly straightforward to explore, and they depend on the value of c. Take for example the geometric sequence given by $a_n = ac^n$, where $a \neq 0$.

- $a_n \to 0$ if $|c| < 1$
- $a_n \to a$ if $c = 1$

[4] In general, one might have more complicated recursive expressions. For example, a very famous example is the Fibonacci sequence which begins as $F_0 = 0$ and $F_1 = 1$, and then described by the recursive expression $F_{n+2} = F_{n+1} + F_n$.

- a_n diverges if $c = -1$. More specifically, it fluctuates between a and $-a$.
- a_n diverges to ∞ if $c > 1$
- a_n diverges if $c < -1$, with its magnitude growing without bound, but its sign fluctuating between positive and negative.

A **geometric series** is a series associated with a geometric sequence, and hence looks like

$$\sum_{i=0}^{\infty} ac^i$$

Here are two quick observations regarding the partial sums of this geometric series:

$$S_{n+1} = S_n + ac^{n+1} \qquad \text{and} \qquad S_{n+1} = cS_n + a$$

which imply

$$S_n + ac^{n+1} \quad = \quad cS_n + a$$

When $c \neq 1$, this last equation allows us to solve for S_n:

$$S_n = a\frac{1 - c^{n+1}}{1 - c}$$

That is

If $c \neq 1$, then $\displaystyle\sum_{i=0}^{n} ac^i = a + ac + ac^2 + \cdots + ac^n = a\frac{1 - c^{n+1}}{1 - c}$.

But we know that if $|c| < 1$, then $c^n \to 0$ as $n \to \infty$. Thus we have the following result:

If $|c| < 1$, then the geometric series $\displaystyle\sum_{i=0}^{\infty} ac^i$ converges to $\dfrac{a}{1 - c}$.

6.2 Trigonometric functions and basic properties

We can talk about sequences of functions as well. If f_n is a function on $\mathbb{R}$, then

$$\text{we say } f_n \to f \text{ if for every } x, \text{ we have } f_n(x) \to f(x).$$

For different values of x, the convergence of $f_n(x)$ to $f(x)$ might have different "speeds". If they converge at a speed not so different from each other, then we say the convergence is **uniform**. (More precisely speaking, given ϵ if we can find a δ that works independent of x in some closed and bounded domain D, then we say convergence is uniform over the domain D.)

If f_n is continuous for each n, and if $f_n \to f$ uniformly, then f is continuous. Here's an example of non-uniform convergence: $g_n(x) = x^n$ over domain $x \in [0, 1]$. Note that for every

$x \in [0,1)$ we necessarily have $x^n \to 0$. For $x = 1$, obviously $x^n \to 1$. Note that $g_n \to g$, where $g(x) = 0$ for $x \in [0,1)$, and $g(1) = 1$. In particular $g(x)$ is not continuous.

If f_n is differentiable on D, and the series

$$\lim_{n\to\infty} \sum_{k=1}^{n} f_n(x) \to G(x) \qquad \text{uniformly}$$

then we can differentiate the series term-by-term:

$$G'(x) = \sum_{k=1}^{\infty} f_n'(x)$$

Sine and cosine

We define the functions $\sin x$ and $\cos x$ as follows

$$\begin{aligned} \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!} \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!} \end{aligned}$$

These series converge uniformly. Using this fact show that

(i) $(\sin x)' = \cos x$

(ii) $(\cos x)' = -\sin x$

(iii) $(\sin x)^2 + (\cos x)^2 = 1$

7 Integration

We first provide a quick review of integration which involves a mechanical description of integration (of $\mathbb{R} \to \mathbb{R}$ functions) as the "inverse operator" of differentiation, and the geometric interpretation of definite integrals as areas associated with the graphs of $\mathbb{R} \to \mathbb{R}$ functions. Then we will discuss the foundations of integration to develop a more coherent view.

7.1 A quick review

If the derivative of F is f, then we say an **indefinite integral** of f is F. Note that if F is an indefinite integral of f, so is $F + c$ for every constant c, because $F' = (F + c)'$ (because the derivative of any constant function is 0).

If F is an indefinite integral of f, then the **definite integral of f from a to b** is denoted

$$\int_a^b f(x)dx \quad \text{and is equal to} \quad F(b) - F(a).$$

Note that in the last expression it does not matter if we replace F with G where $G = F + c$ where c is a constant.

We can think of *indefinite integration* as an operator which takes as its input an integrable function f, and returns as its output a differentiable function F whose derivative is f

$$f \xrightarrow{\text{integrated}} F \qquad \text{such that } F'(x) = f(x).$$

(Note, once more, that the above description of the operator is not entirely complete in the sense that the output of the indefinite integration operator is not uniquely determined. If F is a function whose derivative is f, then so is the function $G(x) = F(x) + 5$ or $H(x) = F(x) - 3$, and so on.)

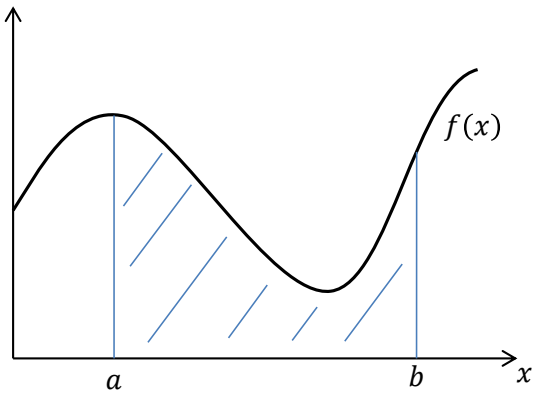


Figure 1: The area of the shaded region is $\int_a^b f(x)dx$.

If f is nonnegative on $[a, b]$, then $\int_a^b f(x)dx$ will capture the area between the graph of f, the horizontal axis, and the vertical lines at $x = a$ and $x = b$. More generally, if f is at times positive, and at times negative over the interval $[a, b]$, then $\int_a^b f(x)dx$ will capture the sum of

areas between the graph of f and the horizontal axis, with a minus sign in front of the areas that lie below the x-axis.

Basic rules

Being the inverse operation of differentiation, it is not surprising that integration will readily inherit the basic algebraic properties of differentiation: namely, addition, subtraction and scalar multiplication: given integrable functions f and g,

$$\int \alpha f + \beta g = \alpha \int f + \beta \int g$$

where α and β are constants (i.e., scalars).

Again, the anti-derivative nature of integration implies that

- $\int x^n dx = \frac{1}{n+1} x^{n+1}$ for $n \neq -1$
- $\int x^{-1} dx = \ln x$
- $\int \exp x dx = \exp x$

7.2 Foundations of Riemann Integration

Let f be a continuous function on interval $[a, b]$. We'd like to partition this interval, beginning with halving it, then halving both half intervals, and then halving all four quarter intervals, and so on. So, if we denote by ℓ the length of the original interval, i.e., setting $\ell = b - a$, first we create two intervals of length $\ell/2$ each. Then four intervals of length $\ell/4$ each, and then eight intervals of length $\ell/8$ each, etc. After n iterations, we have divided the interval into 2^n equal sub-intervals of length $\ell/2^n$ each.

Now, for this partition, let m_i be the minimum value of f over ith subinterval, and M_i be the maximum value of f over that same subinterval. So, we have $m_i \leq M_i$ for every subinterval $i = 1, \ldots, 2^n$.

Next, define the nth lower sum s_n and nth higher sum S_n as follows

$$s_n = \frac{1}{2^n} \sum_{i=1}^{2^n} m_i \qquad S_n = \frac{1}{2^n} \sum_{i=1}^{2^n} M_i$$

Clearly $s_n \leq S_n$. (Geometrically speaking, if f is nonnegative over $[a, b]$, then s_n is meant to approximate, from the inside, the area between the graph of f and the x-axis, whereas S_n is meant to approximate, from the outside.)

One more critical observation.[5] If we make the partition finer, that is, if we divide the subintervals into half one more time, the lower sum gets higher, and the upper sum gets lower. That is

[5]This is something you should verify. To get an idea, simply compare, for an arbitrary function g, the following two quantities: $\min_{x\in[0,2]} g(x)$ versus $\frac{1}{2}\min_{x\in[0,1]} g(x) + \frac{1}{2}\min_{x\in[1,2]} g(x)$.

$$s_n \leq s_{n+1} \leq S_{n+1} \leq S_n.$$

Note that s_n is a non-decreasing sequence which is bounded above. Hence it is convergent. Say it converges to s. Likewise, S_n is a non-increasing sequence bounded below, and hence is convergent as well, say, with limit S. The above observations tell us that $s \leq S$.

If $s = S$, then we say that the **integral of f from a to b** is this number s, and we write

$$\int_a^b f(t)dt = s$$

Remark. These numbers s and S are not equal for all functions. Thus, there are non-Riemann-integrable functions. For example think about the limits of lower and higher sums for the function

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is rational} \\ 0 & \text{otherwise} \end{cases}$$

over the interval $[0, 1]$. What do you get for s? What do you get for S?

Remark. We have discussed one particular way subdividing the interval $[a, b]$, and then computing the associated lower and upper sums. As the subintervals get smaller and smaller, if the lower sums and upper sums converge to the same limit, we defined that limit as the integral. But in fact, we could have carried out this exercise with partitions of arbitrary subdivisions as long as the length of the longest subinterval at each iteration goes to zero as we keep subdividing.

Notational convention. For practical purposes, it will be useful to define what we mean by an integral from b to a when f is a function on $[a, b]$. And we will define it as

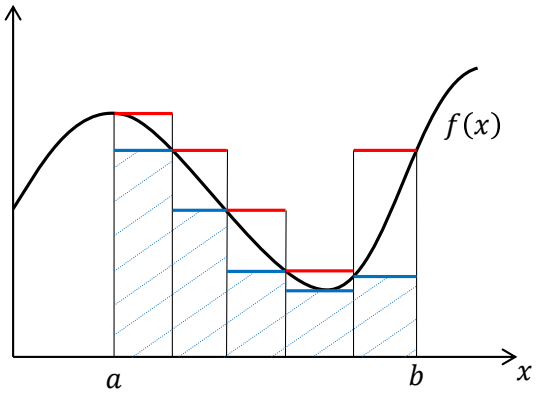
$$\int_b^a f(x)dx = -\int_a^b f(x)dx$$

Remark. The interpretation of integral as the area between the graph of the function and the horizontal axis makes sense if the function is nonnegative as can be seen in the figure. The calculation of lower and upper sums involves multiplying values of the function with the lengths of subintervals. When those values are positive, the products can be seen as the areas if rectangles whose heights are given by the values of the function. If those values, however, are negative, then the products will be negative, and we can think of rectangles below the horizontal axis having negative areas.

7.3 A few basic properties of integrals

1. If $a < b < c$, and if f is continuous on $[a, c]$, then

$$\int_a^b f(x)dx + \int_b^c f(x)dx = \int_a^c f(x)dx$$



2. If f and g are continuous on $[a,b]$, and if $f(x) \leq g(x)$ for all $x \in [a,b]$, then
$$\int_a^b f(x)dx \leq \int_a^b g(x)dx$$

3. If f and g are continuous on $[a,b]$, and if α, β are real numbers, then
$$\int (\alpha f(x) + \beta g(x))dx = \alpha \int f(x)dx + \beta \int g(x)dx$$

4. **[Integral Mean Value Theorem]** Let f be continuous over $[a,b]$. There exists $c \in [a,b]$ such that
$$(b-a)f(c) = \int_a^b f(x)dx.$$

   Proof. Since f is continuous over $[a,b]$, it takes its minimum and maximum values over this interval. Call these values m and M, respectively so $m \leq f(x) \leq M$, for every $x \in [a,b]$. Therefore,
$$m(b-a) \quad \leq \quad \int_a^b f(x)dx \quad \leq \quad M(b-a).$$
   Hence
$$m \quad \leq \quad \frac{1}{b-a}\int_a^b f(x)dx \quad \leq \quad M$$
   On the other hand, by the Intermediate Value Theorem, any value between m and M is attained by f on $[a,b]$. Hence there exists $c \in [a,b]$, such that $\frac{1}{b-a}\int_a^b f(x)dx = f(c)$. □

7.4 Integral as antiderivative

Now that we have defined definite integrals of a function f, we can look at an associated function F defined as
$$F(x) = \int_a^x f(t)dt$$
(provided that f is integrable of course).

Fundamental Theorem of Calculus. Let $f : [a, b] \to \mathbb{R}$ be continuous function, and $F : [a, b] \to \mathbb{R}$ be defined by $F(x) = \int_a^x f(t)dt$. Then

- $F'(x) = f(x)$ for all $x \in (a, b)$.
- $\int_c^d f(t)dt = F(d) - F(c)$ for every c and d in $[a, b]$.

Proof. By definition, $F'(x)$ is:

$$\begin{aligned}\lim_{h\to 0} \frac{F(x+h) - F(x)}{h} &= \lim_{h\to 0} \frac{\int_a^{x+h} f(t)dt - \int_a^x f(t)dt}{h} \\ &= \lim_{h\to 0} \frac{\int_x^{x+h} f(t)dt}{h} \\ &= \lim_{h\to 0} f(c_h) \qquad \text{for some } c_h \in [x, x+h] \text{ by IMVT.}\end{aligned}$$

Since $c_h \in [x, x + h]$, and f is continuous, we know that $f(c_h) \to f(x)$ as $h \to 0$.

As for the second part, we have by definition $F(d) = \int_a^d f(t)dt$ and $F(c) = \int_a^c f(t)dt$. Writing

$$\int_a^d f(t)dt = \int_a^c f(t)dt + \int_c^d f(t)dt$$

and rearranging, we get

$$\int_c^d f(t)dt = \int_a^d f(t)dt - \int_a^c f(t)dt = F(d) - F(c)$$

□

Thus, the Fundamental Theorem of Calculus lets us think of integrals as sort of "antiderivatives". This antiderivative of f is *unique up to a constant*. It is customary to call this the **indefinite integral** of f, though it is important to keep in mind that it is really *an* antiderivative, and not *the*.

7.5 Integration by parts

The so-called "integration by parts" is often expressed concisely as

$$\int udv = uv - \int vdu$$

In this notation, both u and v stand for functions of a variable, say x. Moreover

$$dv \quad \text{stands for} \quad v'(x)dx$$

and

$$du \quad \text{stands for} \quad u'(x)dx$$

So, what's really meant by the above formula is

$$\int u(x)v'(x)dx = u(x)v(x) - \int v(x)u'(x)dx$$

In order to see where this comes from, simply rearrange it to obtain

$$\begin{aligned} \int u(x)v'(x)dx + \int v(x)u'(x)dx &= u(x)v(x) \\ \int (u(x)v'(x) + v(x)u'(x))dx &= u(x)v(x) \\ \int (u(x)v(x))'dx &= u(x)v(x) \end{aligned}$$

which is nothing but the fundamental theorem of calculus applied to the function uv.

7.6 Change of variables

Recall the chain rule: if u is defined around t and differentiable at t, and if f is defined around $u(t)$, and differentiable at $u(t)$, then $f \circ u$ is differentiable at t with the derivative

$$(f \circ u)'(t) = f'(u(t))u'(t)$$

By the fundamental theorem of calculus, the integral of the left hand side is $f(u(t)) + C$, so we can write

$$f(u(t)) + C = \int f'(u(t))u'(t)dt$$

But, let us express the right hand side a bit differently. In particular, let's "forget about t" for a moment, and adopt the notation

$$du = u'(t)dt$$

so that the integral can be rewritten as

$$\int f'(u)du.$$

This last expression is integration with respect to u. By the fundamental theorem of calculus, it is equal to

$$f(u).$$

But recall that $u = u(t)$, so indeed the answer is $f(u(t))$.

The natural logarithm

The **natural logarithm** is a function $\ln x : \mathbb{R}_{>0} \to \mathbb{R}$ specified as

$$\ln x = \int_1^x \frac{1}{t}dt$$

Note that

- $\ln 1 = 0$
- $\ln x$ is strictly increasing
- If $x \in (0, 1)$, then $\ln x < 0$
- By the fundamental theorem of calculus:

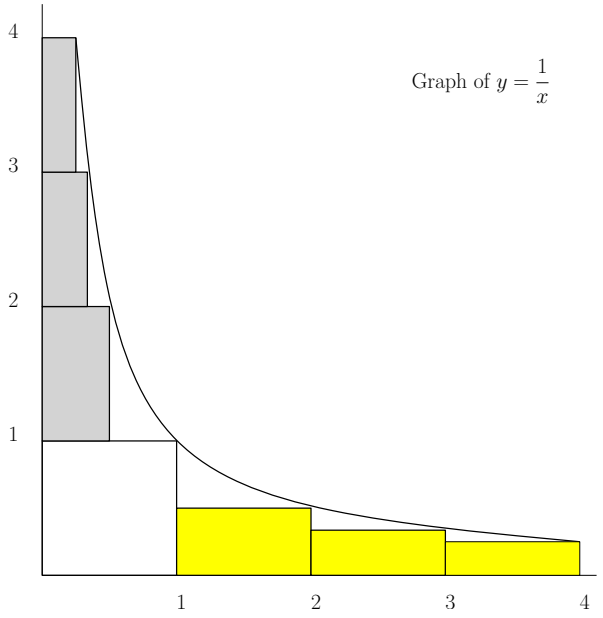
$$(\ln x)' = \frac{1}{x}$$

What is the range of ln**?**

Or put differently what are the limits $\lim_{x\to 0} \ln x$ and $\lim_{x\to\infty} \ln x$?

Using the interpretation of integral as area, we will have a closer look at the question of how $\ln x$ behaves as x goes to ∞ or to 0. First let's observe that the boxes that lie on the x-axis and below the curve $y = 1/x$ in the following figure all have the same width 1, and decreasing heights given by

$$\frac{1}{2} \quad \frac{1}{3} \quad \frac{1}{4} \quad \text{and so on}$$



Our next observation is that adding up the areas of the n shaded boxes lying on the x-axis from left to right would amount to

$$\frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n}$$

and this area must be smaller than the area underneath the curve from $x = 1$ to $x = n$ given by

$$\int_1^n \frac{1}{x} dx$$

Thus we can conclude that

$$\ln n \quad > \quad \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n}$$

Now, if we can show that the series

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

diverges, then we can conclude that $\ln n \to \infty$ as $n \to \infty$.

So, now let's turn out attention to that series and observe that

$$\begin{aligned} & \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \frac{1}{9} + \frac{1}{10} + \cdots + \frac{1}{16} + \cdots \\ > \quad & \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{16} + \frac{1}{16} + \cdots + \frac{1}{16} + \cdots \end{aligned}$$

because comparing the two series term-by-term, the bottom series has equal or smaller terms at every term.

And finally, observe that the latter series

$$\frac{1}{2} + \underbrace{\frac{1}{4} + \frac{1}{4}}_{\frac{1}{2}} + \underbrace{\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}}_{\frac{1}{2}} + \underbrace{\frac{1}{16} + \frac{1}{16} + \cdots + \frac{1}{16}}_{\frac{1}{2}} + \cdots$$

will involve infinitely many groups each of which adds up to $1/2$. Thus this last series diverges to ∞, therefore the top series diverges, and therefore $\ln n$ diverges to ∞.

How about $\lim_{x \to 0} \ln x$? For this one, note that

$$\lim_{x \to 0} \ln x = -(\text{Area underneath the curve from } x = 0 \text{ to } x = 1)$$

But that area clearly is bigger than the same series above, and thus diverges to ∞, which means $\ln x$ diverges to $-\infty$ as $x \to 0$.

Thus we can conclude that the range of the natural logarithm functions is the whole real line.

So far, we have established that $\ln : \mathbb{R}_{>0} \to \mathbb{R}$ is a strictly increasing function, differentiable function whose range is $\mathbb{R}$. (Hence it has an inverse whose domain is $\mathbb{R}$ and range is $\mathbb{R}_{>0}$.) Moreover, $\ln 1 = 0$.

Here's a fundamental and less obvious property

$$\text{For any } a, b > 0, \quad \ln(ab) = \ln a + \ln b$$

Proof. Let $f : \mathbb{R}_{>0} \to \mathbb{R}$ be defined by $f(x) = \ln(ax)$. Since f is the composition of two differentiable functions (namely ln and multiplication by a), it is differentiable too with the derivative

$$f'(x) \quad = \quad \frac{1}{ax} \times a \quad = \quad \frac{1}{x}$$

Since f and ln has the same derivatives, it must be that they differ by a constant, i.e.,

$$f(x) \quad = \quad C + \ln x \qquad \text{where } C \text{ is a constant.}$$

Evaluating f at $x = 1$ yields $f(1) = \ln a$. Evaluting the above equation at $x = 1$ yields

$$f(1) \quad = \quad C$$

Thus $\ln(ax) = f(x) = \ln a + \ln x$. Since $a > 0$ is an arbitrary number, the equation $\ln(ab) = \ln a + \ln b$ holds for all $a, b > 0$. □

Other logarithm functions

Given $a > 0$, we will define **logarithm with base** a as

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

The exponential function

Since the natural logarithm

$$\ln : \mathbb{R}_{>0} \to \mathbb{R}$$

is one-to-one and onto, it admits an inverse function from $\mathbb{R}$ to $\mathbb{R}_{>0}$ which we will call the **exponential function**, denoted by exp.

The fact that exp is the inverse of ln can be used in a fairly straightforward fashion to obtain the following properties:

- $\exp x > 0$ for all $x \in \mathbb{R}$
- $\lim_{x \to -\infty} \exp x = 0 \quad$ and $\quad \lim_{x \to \infty} \exp x = \infty$
- $\exp 0 = 1$
- $\exp(x + y) = \exp x \exp y$

  Hint: Differentiate $\exp(x + y)/\exp x$ *with respect to* x*.*
- The derivative of exp is itself, that is, $(\exp x)' = \exp x$

  Hint: Differentiate the function $\ln \circ \exp$ *via the chain rule.*

Raising a number to an arbitrary power x

Now, given a real number $a > 0$, let us define the function a^x. We sometimes talk about raising a to a power, but if that power x is not a rational number, it is not as easy to make sense of this with the interpretation of multiplying a with itself x times. But the following definition does make sense

$$a^x = \exp(x \ln a)$$

since we know what $\ln a$ means (the integral of $1/x$ from 1 to a), and we know what the exponential of $x \ln a$ means (the inverse image of $x \ln a$ under the natural logarithm function).

Now, it is not hard to establish the following:

- $(\exp x)^y = \exp(xy)$
- $a^{x+y} = a^x a^y$
- $(a^x)^y = a^{xy}$
- $a^x b^x = (ab)^x$
- $\ln x^y = y \ln x \qquad$ for all $x > 0$ and y
- $\log_a x$ is the inverse of a^x
- $\log_a(xy) = \log_a x + \log_a y$ for all $a, x, y > 0$
- $\log_a x^y = y \log_a x \qquad$ for all $x > 0$ and y

Exercise. Evaluate the derivative and the integral of a^x.

8 Taylor series

8.1 Taylor polynomials

Suppose f is differentiable at least n times around the point a. Define the n-th degree **Taylor polynomial** of f around the point a as

$$P_n(x) \quad = \quad f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$

Observation 0. The function P_n agrees with f at a. That is, $P_n(a) = f(a)$.

Observation 1. The derivative of the function P_n agrees with the derivative of f at a. That is, $P_n'(a) = f'(a)$, because

$$P_n'(x) \quad = \quad f'(a) + f''(a)(x-a) + \frac{f'''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{(n-1)!}(x-a)^{n-1}$$

Observation 2. Differentiating P_n', we get

$$P_n''(x) \quad = \quad f''(a) + f'''(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{(n-2)!}(x-a)^{n-2}$$

and evaluating this at the point a, we end up with $f''(a)$.

$\vdots$

Observation k. Continuing in the same fashion, we will find that

$$P_n^{(k)}(a) = f^{(k)}(a) \qquad \text{for each} \quad k = 1, \ldots, n.$$

If we know the values $f(a), f'(a), f''(a), \ldots, f^{(n)}(a)$, we can construct an nth degree polynomial P_n which has this same "local information" as f in the sense that P_n and f have the same value at a, and have the same 1st to nth order derivatives at a.

QUESTION. Is using Taylor polynomials a good way to approximate a function?

Sometimes, but not always!

8.2 Taylor series

If f is infinitely differentiable at a point a, then we can construct Taylor polynomials of all degrees $n > 0$, and talk about their limit:

When f has derivatives of all orders at $x = a$, its **Taylor series** about the point $x = a$ is

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k = f(a)+f'(a)(x-a)+\frac{f''(a)}{2!}(x-a)^2+\frac{f'''(a)}{3!}(x-a)^3+\cdots+\frac{f^{(k)}(a)}{k!}(x-a)^k+\cdots$$

A few possibilities for the above series:

1. The series diverges for some x. If so, the Taylor polynomials P_n do not provide a good approximation to f at x.
2. The series converges for x, but to a value other than $f(x)$. If so, once again, the Taylor polynomials P_n do not provide a good approximation to f at x.
3. The series at x converges to $f(x)$. If so, the Taylor polynomials provide a method to approximate the function at x.

We will mostly deal with Case 3: when dealing with a Taylor series expansion around a point a, this series will indeed converge to $f(x)$ for those x which are in some neighbourhood of a. Such functions are called **analytic**.

How good is such an approximation? That is, can we say anything about how far off the Taylor polynomial is from the original function?

That is,

$$\text{How large is the error term} \quad R_n(x) = f(x) - P_n(x) \quad ?$$

8.3 Taylor's Remainder Theorem

Taylor's Remainder Theorem.

If f is n times continuously differentiable on $[a, x]$, and $n + 1$ times differentiable on (a, x), then

$$R_n(x) = \int_a^x \frac{f^{(n+1)}(t)}{n!}(x-t)^n dt = \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1} \quad \text{for some } c \in [a, x]$$

An example on Taylor series and Taylor's remainder theorem

1. What is the Taylor series expansion of $f(x) = \ln(1 + x)$ around the point 0?
2. Provide a numerical estimate for the value $\ln(1.01)$ which you know is correct up to maximum error of 10^{-6}.

Solution.

1. In order to explore its Taylor series around $x = 0$, we will evaluate $f(0), f'(0), f''(0), \ldots$.

   First, note that

$$\begin{aligned} f'(x) &= \frac{1}{1+x} \\ f''(x) &= \frac{-1}{(1+x)^2} \\ f'''(0) &= \frac{2}{(1+x)^3} \\ \vdots &\ \vdots\ \vdots \\ f^{(k)}(x) &= \frac{(-1)^{k+1}(k-1)!}{(1+x)^k} \end{aligned}$$

   Evaluating the above expressions at $x = 0$, and plugging them into the Taylor series

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!}x^k = \sum_{k=1}^{\infty} \frac{(-1)^k(k-1)!}{k!}x^k = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}x^k = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots$$

2. Using Taylor polynomials for $\ln(1+x)$, we can estimate $\ln(1.01)$.

   Note that $\ln(1.01) = \ln(1 + 0.01) = f(0.01)$. So, using the second degree Taylor polynomial of the function $\ln(1+x)$, the remainder theorem tells us that

$$\begin{aligned} |f(0.01) - P_2(0.01)| &= \left|\frac{f^{(3)}(c)}{3!}(0.01)^3\right| \quad \text{for some } c \in [0, 0.01] \\ &= \frac{1}{3(1+c)^3}(0.01)^3 \\ &\leq 0.34 \times 0.000001 = 0.00000034 \end{aligned}$$

   $P_2(x) = x - \frac{x^2}{2}$, so we can easily compute $P_2(0.01)$:

$$P_2(0.01) = 0.01 - \frac{0.0001}{2} = 0.01 - 0.00005 = 0.00995$$

   Thus we can conclude

$$\ln(1.01) \approx 0.00995 \qquad \text{with an accuracy level of } 0.00000034$$

9 Multi-variable functions

We have already introduced functions from $\mathbb{R}^n$ to $\mathbb{R}$ in Section 2.5. Now, we have a closer look with an aim towards generalising the analysis we have so far done for single-variable functions.

9.1 Partial derivatives

When addressing questions of optimisation and rate of change in the context of single-variable functions, differentiation proved to be useful. So, we may want to adapt the definition of differentiability to multi-variable functions. For example, we can ask:

$$\text{Given } f : \mathbb{R}^n \to \mathbb{R} \quad \text{does} \quad \lim_{\boldsymbol{x}\to\boldsymbol{a}} \frac{f(\boldsymbol{x}) - f(\boldsymbol{a})}{\boldsymbol{x} - \boldsymbol{a}} \quad \text{exist?}$$

One problem with the above formulation is that while $f(\boldsymbol{x}) - f(\boldsymbol{a})$ is a real number, $\boldsymbol{x} - \boldsymbol{a}$ is a point in $\mathbb{R}^n$, and it doesn't make sense to divide an element of $\mathbb{R}$ with an element of $\mathbb{R}^n$ when $n > 1$. Instead of looking for a way of getting around this, we will now turn to a more manageable approach as far as differentiating multi-variable functions go.

Given a multi-variable function $f(x_1, \ldots, x_n)$, viewing it as a function of x_1 alone means analysing the behaviour of the single-variable function given by

$$x_1 \quad \longmapsto \quad f(x_1, x_2, \ldots, x_n)$$

Note that, for every specification of $x_2, \ldots, x_n$, we'd get another function. For example

$$x_1 \quad \longmapsto \quad f(x_1, a_2, \ldots, a_n)$$

is yet another single-variable function, and the derivative of this function will be called the **partial derivative of f with respect to its first variable at** $(x_1, a_2, \ldots, a_n)$.

In this fashion, we can define the **partial derivative of f with respect to its first variable** as the following function from $\mathbb{R}^n$ to $\mathbb{R}$:

$$\frac{\partial f}{\partial x_1} \;:\; (x_1, \ldots, x_n) \quad \longmapsto \quad \lim_{h\to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)}{h}$$

A careful notation for the evaluation of this function at a specific point $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ is

$$\left.\frac{\partial f}{\partial x_1}\right|_{\boldsymbol{x}=\boldsymbol{a}}$$

Likewise we define the **partial derivative of f with respect to its kth variable** x_k as:

$$\frac{\partial f}{\partial x_k} \quad = \quad \lim_{h\to 0} \frac{f(x_1, \ldots, x_k + h, \ldots, x_n) - f(\boldsymbol{x})}{h}$$

For brevity, we sometimes write f_k instead of $\partial f/\partial x_k$ when we talk about the partial derivative of f with respect to its kth variable.

An important property about second partial derivatives: Young's theorem

If a function $f : \mathbb{R}^n \to \mathbb{R}$ has continuous second partial derivatives at $\boldsymbol{a}$, then

$$f_{ij}(\boldsymbol{a}) = f_{ji}(\boldsymbol{a})$$

Written in an alternative notation

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(a_1, \ldots, a_n) = \frac{\partial^2 f}{\partial x_i \partial x_j}(a_1, \ldots, a_n)$$

This property (Young's theorem) is sometimes quoted as *mixed partials commuting*. We will mostly deal with multi-variable functions whose second partial derivatives are continuous, and therefore it won't matter in which order we take partial derivatives.

9.2 The chain rule with multi-variable functions

Remember that the chain rule is about relating the derivative of composition of functions to the derivatives of the functions composed. For example, when we have two functions f and g from $\mathbb{R}$ to $\mathbb{R}$, say differentiable everywhere for ease of exposition, the composition $f \circ g$ is differentiable with the derivative

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

Now, we'd like to generalise this result to be able to relate the derivatives of the three functions

$$f : \mathbb{R}^2 \to \mathbb{R} \qquad x : \mathbb{R} \to \mathbb{R} \qquad y : \mathbb{R} \to \mathbb{R}$$

to the derivative of the function from $\mathbb{R}$ to $\mathbb{R}$ defined by

$$t \quad \longmapsto \quad f(x(t), y(t)) \tag{$\heartsuit$}$$

We were being very careful in not referring to $f(x(t), y(t))$ as a function because without any clarification as to what the variable is, it is not obvious how this expression describes a function. In contrast, expressing it as in ($\heartsuit$) above makes it clear that we are talking about a single-variable function.

The Chain Rule. Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ has continuous partial derivatives with respect to both its variables. If x and y are differentiable functions from $\mathbb{R}$ to $\mathbb{R}$, then

$$\frac{d}{dt} f(x(t), y(t)) = \frac{\partial f}{\partial x}(x(t), y(t))x'(t) + \frac{\partial f}{\partial y}(x(t), y(t))y'(t)$$

The notation on the right hand side is prone to confusion, because when we write $\frac{\partial f}{\partial x}$ we are not really referring to the function $x : \mathbb{R} \to \mathbb{R}$. Rather, we are talking about partially differentiating $f : \mathbb{R}^2 \to \mathbb{R}$ with respect to its first variable. The whole thing has some feel of consistency in that we are thinking of the first variable and the second variable of f as

functions of yet another variable t. As t changes, the variables of f change, and those changes in the variables of f are described by $x(t)$ and $y(t)$. The expression of the left hand side is really the derivative (with respect to t) of the function defined by ($\heartsuit$).

Sometimes, you see the chain rule expressed in Leibniz notation

$$\frac{df}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}$$

which is potentially confusing from a notational perspective (since what stands for variables of which function is implicit), but is more concise.

Yet another alternative notation for partial derivatives is to write f_i for the partial derivative of f with respect to its ith variable. So, the chain rule can also be expressed as

$$\frac{d}{dt}f(x_1(t), x_2(t)) = f_1(x_1(t), x_2(t))x_1'(t) + f_2(x_1(t), x_2(t))x_2'(t)$$

Or we can stick to referring to the variables of f as x and y, and denote the corresponding partial derivatives by f_x and f_y. Then the chain rule will look like

$$\frac{d}{dt}f(x(t), y(t)) = f_x(x(t), y(t))x'(t) + f_y(x(t), y(t))y'(t)$$

Challenge. In order to give an idea as to how the chain rule comes about for single-variable functions, we gave an intuitive explanation based on the interpretation of the derivative as the slope of a linear approximation. Can you extend that idea to the more general form of the chain rule given above?

9.3 Differentials for multi-variable functions

Suppose f is a real-valued function with n real variables, i.e., an $\mathbb{R}^n \to \mathbb{R}$ function. Its **partial differential** with respect to its ith variable is an $\mathbb{R}^{n+1} \to \mathbb{R}$ function with the following mapping

$$(\boldsymbol{x}, \Delta x_i) \quad \longmapsto \quad f_i(\boldsymbol{x})\ \Delta x_i \quad \text{where} \quad f_i = \frac{\partial f}{\partial x_i}(\boldsymbol{x})$$

The common shorthand notation for this function, however, is

$$f_i\ dx_i$$

Sum of all its partial differentials is called the **total differential** (or **total derivative**) of f, denoted df:

$$df \quad = \quad f_1 dx_i + \cdots + f_n dx_n$$

If each x_i were to be a function of a variable t, then by plugging $dx_i = x_i' dt$, we obtain

$$df \quad = \quad f_1 x_1' dt + \cdots + f_n x_n' dt \quad = \quad (f_1 x_1' + \cdots + f_n x_n')dt$$

which is another way of stating the chain rule for multi-variable functions.

Example

A macro model of a closed economy consists of the following equations

$$\begin{aligned} Y &= C + I + G \\ M &= M^D \end{aligned}$$

where Y is income, C is consumption expenditure, I is investment, G is net government spending, M is the money supply, and M^D is the demand for money. In addition, you are told that consumption expenditure is a function of income, investment is a function of the interest rate r, and the demand for money is a function of both income and the interest rate.

When the system is very close to equilibrium, how are changes in the endogenous variables Y and r related to changes in the exogenous variables G and M?

The exogenous variable G features in the first equilibrium equation. Also we are told that Y and r are the endogenous variables. Using the chain rule, let's differentiate the first equation with respect to G to obtain

$$\frac{dY}{dG} = C_Y \frac{dY}{dG} + C_r \frac{dr}{dG} + I_Y \frac{dY}{dG} + I_r \frac{dr}{dG} + \frac{dG}{dG}$$

where the subscripts stand for partial derivatives. The partial derivatives C_r and I_Y are zero, so the above expression simplifies to

$$\frac{dY}{dG} = C_Y \frac{dY}{dG} + I_r \frac{dr}{dG} + \frac{dG}{dG}$$

which we can rewrite as

$$dG = dY - C_Y dY - I_r dr$$

The exogenous variable M features in the second equilibrium equation. Differentiating that equation with respect to M

$$\begin{aligned} \frac{dM}{dM} &= \frac{dM^D}{dM} \\ &= M_Y^D \frac{dY}{dM} + M_r^D \frac{dr}{dM} \end{aligned}$$

which we rewrite as

$$dM = M_Y^D dY + M_r^D dr$$

Combining the two equations we derived above in matrix notation

$$\begin{pmatrix} dG \\ dM \end{pmatrix} = \begin{pmatrix} 1 - C_Y & -I_r \\ M_Y^D & M_r^D \end{pmatrix} \begin{pmatrix} dY \\ dr \end{pmatrix}$$

Solving the above matrix equation for dY and dr, we obtain

$$\begin{aligned} dY &= \frac{M_r^D dG + I_r dM}{M_r^D - M_r^D C_Y + M_Y^D I_r} \\ dr &= \frac{-m_Y^D dG + (1 - C_Y) dM}{M_r^D - M_r^D C_Y + M_Y^D I_r} \end{aligned}$$

If we want to express dY/dG and dY/dM separately, we can compute them from dY as

$$\begin{aligned}\frac{dY}{dG} &= \frac{M_r^D}{M_r^D - M_r^D C_Y + M_Y^D I_r} \\ \frac{dY}{dM} &= \frac{I_r}{M_r^D - M_r^D C_Y + M_Y^D I_r}\end{aligned}$$

9.4 The derivative of an $\mathbb{R}^n \to \mathbb{R}^m$ function

We have already talked about the partial derivatives of a function $f : \mathbb{R}^n \to \mathbb{R}$. Differentiability of such a function, however, is more demanding than the existence of partial derivatives. Even more generally than $\mathbb{R}^n \to \mathbb{R}$ functions, we can talk about the derivative of a multi-variable function with vector values as in $f : \mathbb{R}^n \to \mathbb{R}^m$. As it was for $\mathbb{R} \to \mathbb{R}$ functions, differentiability of f at a point $\boldsymbol{a}$ is related to whether it has a linear approximation at that point. If the linear approximation is

$$f(\boldsymbol{x}) \approx f(\boldsymbol{a}) + \boldsymbol{\alpha}(\boldsymbol{x} - \boldsymbol{a}),$$

where $\boldsymbol{\alpha}$ is an $m \times n$ matrix, then f has derivative $\boldsymbol{\alpha}$. Note that $f(\boldsymbol{x})$ and $f(\boldsymbol{a})$ are vectors in $\mathbb{R}^m$, whereas $(\boldsymbol{x} - \boldsymbol{a})$ is a vector in $\mathbb{R}^n$. Formally,

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is **differentiable at $\boldsymbol{a}$** with **derivative $\boldsymbol{\alpha}$** if

$$\frac{f(\boldsymbol{a} + \boldsymbol{\delta}) - f(\boldsymbol{a}) - \boldsymbol{\alpha}\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|} \to 0 \quad \text{as} \quad \boldsymbol{\delta} \to 0.$$

The $m \times n$ matrix $\boldsymbol{\alpha}$ is called the derivative of f at $\boldsymbol{a}$ and is denoted by $f'(\boldsymbol{a})$ or $Df(\boldsymbol{a})$.

Equivalently, using the little-oh notation,[6]

$$f(\boldsymbol{a} + \boldsymbol{\delta}) = f(\boldsymbol{a}) + \boldsymbol{\alpha}\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|).$$

This gives us the **first order Taylor series approximation** around $\boldsymbol{a}$:

$$f(\boldsymbol{a} + \boldsymbol{\delta}) \approx f(\boldsymbol{a}) + f'(\boldsymbol{a})\boldsymbol{\delta}.$$

This all looks very much like the expressions for functions from $\mathbb{R}$ to $\mathbb{R}$, but keep in mind that $\boldsymbol{a}$ is a point in $\mathbb{R}^n$ and f is a function from $\mathbb{R}^n$ to $\mathbb{R}^m$. The derivative $f'(\boldsymbol{a})$ is an $m \times n$ matrix, and $\boldsymbol{\delta}$ is a vector in $\mathbb{R}^n$. When we write $\boldsymbol{\alpha}\boldsymbol{\delta}$, we plug in $\boldsymbol{\delta}$ as an $n \times 1$ column matrix. The approximation gets better and better as $\boldsymbol{\delta}$ gets closer to 0 in the sense that the error term $f(\boldsymbol{a} + \boldsymbol{\delta}) - f(\boldsymbol{a}) - f'(\boldsymbol{a})\boldsymbol{\delta}$ approaches to 0 much faster than $\|\boldsymbol{\delta}\|$.

If f is differentiable at every point $\boldsymbol{a}$ of a domain $A \subseteq \mathbb{R}^n$, then we say f is **differentiable** in A. If the derivative is continuous, f is **continuously differentiable**.

The matrix $f'(\boldsymbol{a})$ is also known as the **Jacobian** of f at $\boldsymbol{a}$, and in order to highlight its multi-dimensional nature some writers prefer the notation $Df(\boldsymbol{a})$. If $f : \mathbb{R}^n \to \mathbb{R}^m$, we can write

$$f(x_1, \ldots, x_n) = (f^1(x_1, \ldots, x_n), \ldots, f^m(x_1, \ldots, x_n)),$$

[6]For a function $g(z)$ to be of $o(z)$ means $g(z)/z \to 0$ as $z \to 0$.

and the derivative of f at $\boldsymbol{x}$ is given by:

$$Df(\boldsymbol{x}) \;=\; \begin{pmatrix} \frac{\partial f^1}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial f^1}{\partial x_n}(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f^m}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial f^m}{\partial x_n}(\boldsymbol{x}) \end{pmatrix}$$

9.5 Special case of $\mathbb{R}^n \to \mathbb{R}$ functions

If the multivariable function of interest is real valued (as is common in many of the applications we will study), a widely used notation for the derivative is

$$\nabla f(\boldsymbol{x}) \;=\; (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_n(\boldsymbol{x})),$$

So the symbol ∇ is used in stead of D (which was used in higher dimensions). $\nabla f(\boldsymbol{x})$ is also called the **gradient** of f.

Differentiability of f means for small $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n) \in \mathbb{R}^n$, we have the following approximation

$$\begin{aligned} f(\boldsymbol{x}+\boldsymbol{\delta}) &\approx f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}) \cdot \boldsymbol{\delta} \\ &= f(\boldsymbol{x}) + f_1(\boldsymbol{x})\delta_1 + f_2(\boldsymbol{x})\delta_2 + \cdots + f_n(\boldsymbol{x})\delta_n \end{aligned}$$

The second derivative of f is the rate of change of its first derivative, i.e., it is the rate of change of the vector $\nabla f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_n(\boldsymbol{x}))$. The rate of change of entry $f_i(\boldsymbol{x})$ is itself a vector, namely $\nabla f_i(\boldsymbol{x}) = (f_{i1}(\boldsymbol{x}), f_{i2}(\boldsymbol{x}), \ldots, f_{in}(\boldsymbol{x}))$. Thus, we get the second derivative of f expressed as an $n \times n$ matrix called the Hessian matrix of mixed partial derivatives.

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable at $\boldsymbol{x}$. The **Hessian** of f is the $n \times n$ symmetric matrix

$$f''(\boldsymbol{x}) \;=\; D^2 f(\boldsymbol{x}) \;=\; \begin{bmatrix} \nabla f_1(\boldsymbol{x}) \\ \vdots \\ \nabla f_n(\boldsymbol{x}) \end{bmatrix} \;=\; \begin{bmatrix} f_{11}(\boldsymbol{x}) & \ldots & f_{1n}(\boldsymbol{x}) \\ \vdots & & \vdots \\ f_{n1}(\boldsymbol{x}) & \ldots & f_{nn}(\boldsymbol{x}) \end{bmatrix}$$

where $f_{ij}(\boldsymbol{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\boldsymbol{x})$.

We can now write the second-degree Taylor series approximation for f:

$$f(\boldsymbol{x}+\boldsymbol{\delta}) \;=\; f(\boldsymbol{x}) + f'(\boldsymbol{x}) \cdot \boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta} f''(\boldsymbol{x}) \boldsymbol{\delta}^{\intercal} + o(\|\boldsymbol{\delta}\|^2)$$

Remember the little-oh notation: for a function $R(z)$ to be of $o(z)$ means $R(z)/z \to 0$ as $z \to 0$. Hence, the term denoted by $o(\|\boldsymbol{\delta}\|^2)$ is a quantity Q such that $Q/\|\boldsymbol{\delta}\|^2 \to 0$ as $\boldsymbol{\delta} \to 0$.

9.6 Implicit differentiation

Consider a production function $f(K, L)$ where K stands for capital, and L stands for labour. Suppose the production target is fixed, say, at ω. We want to understand how the amounts of K and L delivering this output relate to each other. Assuming production goes up with both K and L, we know that we need to increase one input while another decreases in order to keep the level of production the same as before. But by how much? For example, if we decrease capital by some amount, by how much should labour be increased so that we keep the level of production at ω?

If we could solve for L in the equation

$$f(K, L) = \omega$$

for every value of K in some interval I, then we can express L as a function of K:

$$K \quad \mapsto \quad L = L(K; \omega)$$

If this function "behaves well", we can take the derivative of this expression to get the rate of substitution between K and L while we are ensuring the production level ω. (Note that this will tell us something about the slope of the tangent to the isoquant curve.)

Example. Let $f(K, L) = 3K^2L + KL^2$. When $K = L = 1$, the total production is 4. If we want to keep the production at that level, how much change in K would compensate a small change in L? In other words we would like to compute

$$\frac{\partial K}{\partial L}$$

Solving for K from the equation $3K^2L + KL^2 = 4$ might be messy. But we don't really need to solve for K. We can simply differentiate the equation with respect to L assuming that K can indeed be expressed as a function of L. So we have

$$\begin{aligned} \frac{\partial}{\partial L}(3K^2L + KL^2) &= \frac{\partial 4}{\partial L} \\ 6K\frac{\partial K}{\partial L}L + 3K^2\frac{\partial L}{\partial L} + \frac{\partial K}{\partial L}L^2 + K \cdot 2L\frac{\partial L}{\partial L} &= 0 \\ (6KL + L^2)\frac{\partial K}{\partial L} + 3K^2 + 2KL &= 0 \end{aligned}$$

Now we can evaluate these at $K = L = 1$ to get

$$7\left.\frac{\partial K}{\partial L}\right|_{K=L=1} + 5 = 0$$

and therefore

$$\left.\frac{\partial K}{\partial L}\right|_{K=L=1} = -\frac{5}{7}$$

$\diamond$

What we have done in the above example is called **implicit differentiation**. We were interested in $\partial K/\partial L$ evaluated at $K = L = 1$. Instead of expressing L as a function of K and differentiating it with respect to L, we differentiated the equation which relates K to L, used the fact that $\partial L/\partial L = 1$, and then plugged in $K = L = 1$ to get what we were looking for.

Why did we use the partial differentiation symbol ∂? In this particular case, we did not have to, but in principle, we could have ω as another variable, and think of K as a function of L and ω. The above exercise was keeping ω constant (namely at $\omega = 4$), while differentiating with respect to L.

Another example. Consider the unit circle, i.e., the set of points (x, y) in the real plane which satisfy the equation $x^2 + y^2 = 1$. What is the rate of change in y with respect to a change in x around the point $(1, 0)$? Again, we can avoid solving the equation for y in terms of x, and go straight into differentiating the equation with respect to x:

$$2x + 2y\frac{dy}{dx} = 0$$

However, evaluating this at $x = 1$ and $y = 0$ gives $2 = 0$, a contradiction!

What went wrong? Now going back to the derivative of the equation, note that if $y \neq 0$, we get

$$\frac{dy}{dx} = -\frac{x}{y}$$

and therefore $\frac{dy}{dx}$ diverges as $y \to 0$. What is going on around the point $(1, 0)$?

Note that however close we get to the point $(1, 0)$, that is, in *every* neighbourhood of the point $(1, 0)$, there are two different values of y which solves the equation $x^2 + y^2 = 1$. When we are trying to express y as a function x, should we pick $g(x) = \sqrt{1 - x^2}$ or $h(x) = -\sqrt{1 - x^2}$? The fact that we cannot solve uniquely turns out to be intimately connected with the fact that we did not have a well defined $\frac{dy}{dx}$ at $(1, 0)$.

By the way, if y can be expressed as a function of x with an inverse, then we have x as a function of y. If $y = g(x)$ and if g^{-1} is its inverse, then $x = g^{-1}(y)$, and the derivatives satisfy

$$\frac{dy}{dx} = g'(x) = \frac{1}{(g^{-1})'(y)} = \frac{1}{\frac{dx}{dy}}$$

The issue of $\frac{dy}{dx}$ diverging can be reinterpreted as $\frac{dx}{dy}$ being 0 at $(1, 0)$. In fact the theorem below says this derivative being nonzero allows us to solve y as a uniquely defined function of x around a given point satisfying the equation. $\diamond$

Implicit Function Theorem in 2×1 dimensions. Suppose $F(\cdot, \cdot)$ is continuously differentiable in some neighbourhood of (a, b). If $F(a, b) = 0$, and if $F_2(a, b) \neq 0$, then there exist $N(a, \delta_1)$ and $N(b, \delta_2)$, and there exists a unique function $g : N(a, \delta_1) \to N(b, \delta_2)$ such that

- For every $x \in N(a, \delta_1)$, $F(x, g(x)) = 0$.
- The function g is continuously differentiable in $N(a, \delta_1)$.

When the theorem holds, differentiating the equation $F(x, g(x)) = 0$ with respect to x yields

$$F_1(x, g(x)) + F_2(x, g(x))g'(x) = 0$$

which then implies for every $x \in N(a, \delta_1)$,

$$g'(x) = -\frac{F_1(x, g(x))}{F_2(x, g(x))}.$$

The theorem has a higher dimensional version which rests on a similar intuition:

Implicit Function Theorem. Given a function $F : \mathbb{R}^{m+n} \to \mathbb{R}^n$, let us denote the first m variables of F as $x_1, \ldots, x_m$, and the last n variables as $y_1, \ldots, y_n$. Since F takes values in $\mathbb{R}^n$, we can write it as $F = (f^1, \ldots, f^n)$, with the functions $f^i : \mathbb{R}^{m+n} \to \mathbb{R}$ for $i = 1, \ldots n$. Suppose all f^i are continuously differentiable in some neighbourhood of $(\boldsymbol{a}, \boldsymbol{b}) = (a_1, \ldots, a_m; b_1, \ldots, b_n)$. If $F(\boldsymbol{a}, \boldsymbol{b}) = 0$, and if the Jacobian of F with respect to $\boldsymbol{y}$ evaluated at $(\boldsymbol{a}, \boldsymbol{b})$ is non-singular, i.e., if

$$\begin{vmatrix} \frac{\partial f^1}{\partial y_1}(\boldsymbol{a}; \boldsymbol{b}) & \cdots & \frac{\partial f^1}{\partial y_n}(\boldsymbol{a}; \boldsymbol{b}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f^n}{\partial y_1}(\boldsymbol{a}; \boldsymbol{b}) & \cdots & \frac{\partial f^n}{\partial y_n}(\boldsymbol{a}; \boldsymbol{b}) \end{vmatrix} \neq 0,$$

then there exist neighbourhoods $N(\boldsymbol{a}, \delta_1) \subset \mathbb{R}^m$ and $N(\boldsymbol{b}, \delta_2) \subset \mathbb{R}^n$; and for each $i = 1, \ldots, n$ there exists a unique function $g_i : N(\boldsymbol{a}, \delta_1) \to N(\boldsymbol{b}, \delta_2)$, such that

$$F(\boldsymbol{x}; g_1(\boldsymbol{x}), \ldots, g_n(\boldsymbol{x})) = 0 \quad \text{for every } x \in N(\boldsymbol{a}, \delta_1),$$

and

$$g_i(\boldsymbol{a}) = b_i \quad \text{for every } i.$$

Revisiting the last example with the help of IFT

Consider the graph $\{(x, y) \mid x^2 + y^2 = 1\}$. Around what points can we define y as a function of x locally? What is $\frac{dy}{dx}$ around such a point?

Formulating it like in the IFT:

$$\begin{aligned} F : \mathbb{R}^{1+1} &\to \mathbb{R} \\ F(x, y) &= x^2 + y^2 - 1 \end{aligned}$$

We want to be able to parametrise y smoothly as a function of x. IFT says we can do that around those points where F_y is non-zero. I.e., as long as $2y \neq 0$. Note that problematic points are where $\frac{dy}{dx}$ is not defined. Everywhere else we have $y = \sqrt{1-x^2}$ or $y = -\sqrt{1-x^2}$. $\diamond$

Another example on the IFT

The point $(\alpha, u, v) = (2, -1, 2)$ satisfies the following two equations:

$$\alpha^2 + uv - v^2 + 2 = 0$$
$$\alpha + u^2 + uv - 1 = 0$$

(a) Can u and v be defined as implicit functions of α around the point $(2, -1, 2)$?

(b) Compute $du/d\alpha$ and $dv/d\alpha$ at this point.

Solution. Two equations, three variables. In order to use the IFT, we will rephrase the ingredients in the format of the statement of the IFT. Let

$$\begin{aligned} f : \mathbb{R}^{2+1} &\to \mathbb{R}^2 \\ f(u, v, \alpha) &= (\alpha^2 + uv - v^2 + 2, \alpha + u^2 + uv - 1) \end{aligned}$$

Denoting

$$\begin{aligned} f^1(u, v, \alpha) &= \alpha^2 + uv - v^2 + 2 \\ f^2(u, v, \alpha) &= \alpha + u^2 + uv - 1 \end{aligned}$$

we need a non-singular Jacobian, i.e.

$$\det \begin{pmatrix} f^1_u & f^1_v \\ f^2_u & f^2_v \end{pmatrix} \neq 0$$

Note that

$$\begin{pmatrix} f^1_u & f^1_v \\ f^2_u & f^2_v \end{pmatrix} = \begin{pmatrix} v & u - 2v \\ 2u + v & u \end{pmatrix} = \begin{pmatrix} 2 & -5 \\ 0 & -1 \end{pmatrix} \qquad \text{at} \quad (u, v, \alpha) = (-1, 2, 2)$$

For comparative statics, remember that α is related to u and v via the equations:

$$\begin{aligned} f^1(\alpha, u, v) &= \alpha^2 + uv - v^2 + 2 &= 0 \\ f^2(\alpha, u, v) &= \alpha + u^2 + uv - 1 &= 0 \end{aligned}$$

Differentiating both equations with respect to α gives

$$\begin{aligned} f^1_u \frac{du}{d\alpha} + f^1_v \frac{dv}{d\alpha} &= 0 \\ f^2_u \frac{du}{d\alpha} + f^2_v \frac{dv}{d\alpha} &= 0 \end{aligned}$$

which we can write as

$$\begin{pmatrix} f^1_u & f^1_v \\ f^2_u & f^2_v \end{pmatrix} \begin{pmatrix} \frac{du}{d\alpha} \\ \frac{dv}{d\alpha} \end{pmatrix} = 0$$

Hence

$$\begin{pmatrix} \frac{du}{d\alpha} \\ \frac{dv}{d\alpha} \end{pmatrix} = - \begin{pmatrix} 2 & -5 \\ 0 & -1 \end{pmatrix}^{-1} \begin{pmatrix} f^1_\alpha \\ f^2_\alpha \end{pmatrix} = - \begin{pmatrix} 2 & -5 \\ 0 & -1 \end{pmatrix}^{-1} \begin{pmatrix} 4 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/2 \\ 1 \end{pmatrix}$$

$\diamond$

9.7 Concavity and convexity of $\mathbb{R}^n \to \mathbb{R}$ functions

The notions of concavity and convexity can be extended to $\mathbb{R}^n \to \mathbb{R}$ functions in a fairly straightforward fashion. For our purposes, the multi-dimensional analogue of an interval is a convex set in $\mathbb{R}^n$. Recall that a set K in $\mathbb{R}^n$ is called **convex** if for every $\boldsymbol{x}, \boldsymbol{y} \in K$ and for every $\lambda \in (0,1)$, the point $\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}$ is also in K. Any such point is called a convex combination of $\boldsymbol{x}$ and $\boldsymbol{y}$. It is a good exercise to verify that if we take any k points $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k$ in a convex set K, and arbitrary real numbers $0 < \lambda_1, \ldots, \lambda_k < 1$ such that $\lambda_1 + \cdots + \lambda_k = 1$, then the point $\lambda_1\boldsymbol{x}^1 + \cdots + \lambda_k\boldsymbol{x}^k$ is also in K. Any such point is called a convex combination of the k points $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k$.

Suppose f is a real-valued function defined on a convex subset K of $\mathbb{R}^n$. We say f is **concave** over K if for every $\boldsymbol{a}, \boldsymbol{b} \in K$ and every $\lambda \in (0,1)$:

$$f(\lambda\boldsymbol{a} + (1-\lambda)\boldsymbol{b}) \quad \geq \quad \lambda f(\boldsymbol{a}) + (1-\lambda) f(\boldsymbol{b}).$$

If the inequality is strict for all $\boldsymbol{a} \neq \boldsymbol{b} \in K$ and all $\lambda \in (0,1)$, then we say f is **strictly concave** over K.

Theorem. Suppose K is a convex set. If $f : K \to \mathbb{R}$ is strictly concave over K, then f takes its maximum value at most once in K.

Proof. Suppose, for a contradiction, that f attained its maximum value at two different points $\boldsymbol{a}$ and $\boldsymbol{b}$ in K. The fact that K is convex implies $\boldsymbol{a}/2 + \boldsymbol{b}/2$ is in K. The fact that f is strictly concave over K implies $f(\boldsymbol{a}/2 + \boldsymbol{b}/2) > f(\boldsymbol{a})/2 + f(\boldsymbol{b})/2 = f(\boldsymbol{a}) = f(\boldsymbol{b})$. That is, f takes an even higher value at $\boldsymbol{a}/2 + \boldsymbol{b}/2 \in K$, a contradiction with its taking its maximum value at $\boldsymbol{a}$ and $\boldsymbol{b}$. □

f is called [***strictly***] **convex** over a convex set K if $-f$ is [*strictly*] concave over K.

We would like to obtain tests of concavity/convexity for multi-variable real valued function akin to the second derivative test for single variable functions. However f'' is a more complicated object when f is a multi-variable function. Namely, it is an $n \times n$ matrix. Before we state the relevant testing conditions on this matrix, we will remind a few concepts from linear algebra.

Quadratic forms

A symmetric $n \times n$ matrix A gives rise to a so-called quadratic form:

$$\boldsymbol{x} \quad \longmapsto \quad \boldsymbol{x}A\boldsymbol{x}^\intercal$$

where $\boldsymbol{x}$ is an n-dimensional vector, and $\boldsymbol{x}^\intercal$ is its transpose (and hence an n-dimensional column vector).

For example a general 1 dimensional form is

$$x \quad \longmapsto \quad ax^2$$

A general 2-D quadratic form is

$$\begin{pmatrix} x_1 & x_2 \end{pmatrix} \quad \longmapsto \quad \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad = \quad a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2$$

The matrix A (and the associated quadratic form) is called

- **negative semi-definite** if $\boldsymbol{x}A\boldsymbol{x}^\intercal \leq 0$ for all $\boldsymbol{x}$,
- **positive semi-definite** if $\boldsymbol{x}A\boldsymbol{x}^\intercal \geq 0$ for all $\boldsymbol{x}$,
- **negative definite** if $\boldsymbol{x}A\boldsymbol{x}^\intercal < 0$ for all $\boldsymbol{x} \neq 0$,
- **positive definite** if $\boldsymbol{x}A\boldsymbol{x}^\intercal > 0$ for all $\boldsymbol{x} \neq 0$

Given an $n \times n$ matrix A, its k^{th}-order **leading principal minor** is the determinant of the matrix obtained by deleting the last $(n-k)$ rows and columns of A. For example, for the 3×3 matrix A given by

$$A \quad = \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}$$

its leading first-order principal minor is $|a_{11}|$, leading second-order principal minor is $\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$, and its leading third-order principal minor is $|A|$.

Theorem. A symmetric $n \times n$ matrix is:

- negative definite iff the leading principal minors alternate in sign: $|A_1| < 0, |A_2| > 0, |A_3| < 0$, etc., with the k^{th}-order leading principal minor having sign $(-1)^k$.
- positive definite iff all the leading principal minors are > 0.
- indefinite if (but not only if) the leading principal minors are $\neq 0$ and neither condition above is satisfied

Second derivative conditions for concavity

Remember that for $\mathbb{R} \to \mathbb{R}$ functions, the second derivative of the function helped us identify concavity/convexity of a function (and provided us the SOC to classify critical points).

Remember that the second derivative of a function $f : \mathbb{R}^n \to \mathbb{R}$, also called its Hessian, is its "matrix of second partial derivatives" given by

$$D^2f(\boldsymbol{x}) \quad = \quad \begin{pmatrix} f_{11}(\boldsymbol{x}) & f_{12}(\boldsymbol{x}) & \cdots & f_{1n}(\boldsymbol{x}) \\ f_{21}(\boldsymbol{x}) & f_{22}(\boldsymbol{x}) & \cdots & f_{2n}(\boldsymbol{x}) \\ \cdots & \vdots & \cdots & \vdots \\ f_{n1}(\boldsymbol{x}) & f_{n2}(\boldsymbol{x}) & \cdots & f_{nn}(\boldsymbol{x}) \end{pmatrix} \qquad \text{where} \quad f_{ij}(\boldsymbol{x}) \quad = \quad \frac{\partial^2 f}{\partial x_j \partial x_i}(\boldsymbol{x})$$

Theorem. Suppose f is twice partial differentiable with continuous second partial derivatives.

- f is concave [*convex*] iff $D^2f(x)$ is negative [*positive*] semi-definite for all x.
- If $D^2f(x)$ is negative [*positive*] definite for all x, then f is strictly concave [*convex*].

A special case: two-variables

The **Hessian** of $f : \mathbb{R}^2 \to \mathbb{R}$ at the point $\boldsymbol{c}$ is

$$\begin{bmatrix} f_{11}(\boldsymbol{c}) & f_{12}(\boldsymbol{c}) \\ f_{21}(\boldsymbol{c}) & f_{22}(\boldsymbol{c}) \end{bmatrix}$$

If f has continuous second partial derivatives, we know that $f_{12}(\boldsymbol{c}) = f_{21}(\boldsymbol{c})$.

In that case we can write the Hessian as

$$\begin{bmatrix} f_{11}(\boldsymbol{c}) & f_{12}(\boldsymbol{c}) \\ f_{12}(\boldsymbol{c}) & f_{22}(\boldsymbol{c}) \end{bmatrix}$$

Let $f : D \to \mathbb{R}$ have continuous second partial derivatives, where D is a convex subset of $\mathbb{R}^2$.

- If $f_{11}(\boldsymbol{x})f_{22}(\boldsymbol{x}) - (f_{12}(\boldsymbol{x}))^2 > 0$ and $f_{11}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in D$, then f is strictly convex on D.
- If $f_{11}(\boldsymbol{x})f_{22}(\boldsymbol{x}) - (f_{12}(\boldsymbol{x}))^2 > 0$ and $f_{11}(\boldsymbol{x}) < 0$ for all $\boldsymbol{x} \in D$, then f is strictly concave on D.

9.8 Quasiconcave and quasiconvex functions

A common assumption for utility functions is that of *quasiconcavity* which captures the idea that a mixture of two choices cannot be worse than both of those choices. To define it formally,

Let X be a convex domain. We say a function $f : X \to \mathbb{R}$ **quasiconcave** iff for any $\boldsymbol{x}$ and $\boldsymbol{y} \neq \boldsymbol{x}$ in X and $\theta \in (0, 1)$:

$$f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \quad \geq \quad \min\{f(\boldsymbol{x}), f(\boldsymbol{y})\}$$

f is **strictly quasiconcave** iff the inequalities above are strict.

Quasiconvexity can be defined by reversing the inequality and replacing min with max. Thus, it is not hard to see $f(x)$ is [*strictly*] quasiconcave iff $-f(x)$ is [*strictly*] quasiconvex.

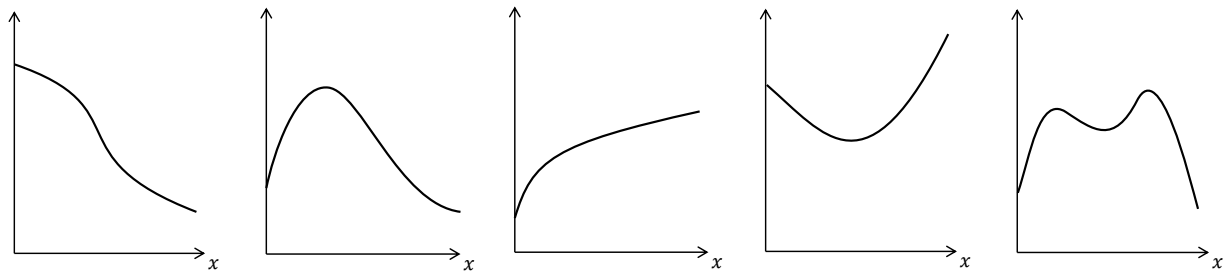
Strict quasiconcavity of utility functions (or profit functions) is convenient because it ensures the uniqueness of the utility maximising choice: if f is strictly quasiconcave and has attains a maximum at $\boldsymbol{x}^*$, then $\boldsymbol{x}^*$ is the unique maximum.

A few observations on quasiconcavity [quasiconvexity]

Concavity implies quasiconcavity. Likewise, convexity implies quasiconvexity. But not vice versa. In fact the family of quasiconcave functions are much bigger than those of concave functions. In the same vein, strictly concave [*convex*] functions are strictly quasiconcave [*quasiconvex*].

While concavity [*convexity*] is a cardinal property, quasiconcavity [*quasiconvexity*] is an ordinal property, preserved under a transformation by any increasing function. That is, if $f : X \to \mathbb{R}$ is [*strictly*] quasiconcave and $g : \mathbb{R} \to \mathbb{R}$ is [*strictly*] increasing, then $g(f(x))$ is [*strictly*] quasiconcave.

In 1 dimension, [*strict*] quasiconcavity means that a function is [*strictly*] increasing up to some point and then [*strictly*] decreasing, i.e. is **single peaked**. In more dimensions, this must hold along any line. This peak may be at the far right or far left, as in the following figures:



(a) Quasiconcave (b) QC (c) QC (d) Not QC (e) Not QC

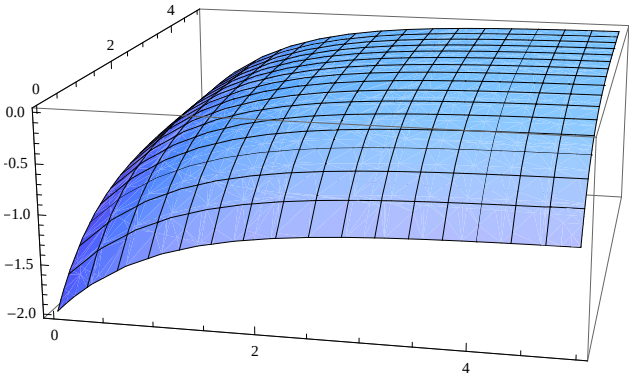
Figure 2: Quasiconcavity in 1 dimension

Upper contour sets

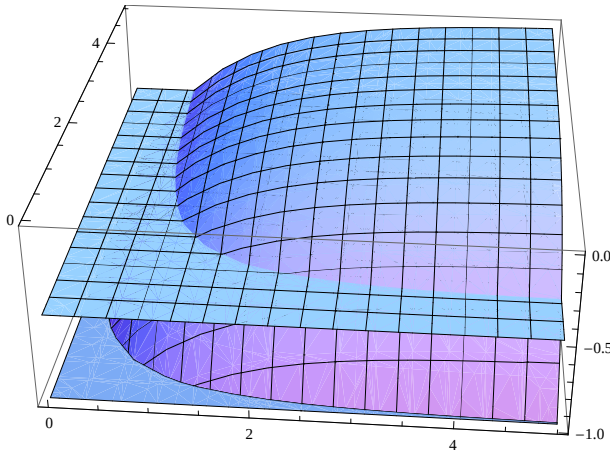
The **upper contour set** of f at level k is the set $\text{UCS}_k = \{\boldsymbol{x} \in X : f(\boldsymbol{x}) \geq k\}$. The **lower contour set** of f at level k is the set $\text{LCS}_k = \{\boldsymbol{x} \in X : f(\boldsymbol{x}) \leq k\}$.

Thinking of f as a utility function, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are two choices each of which with a payoff of at least k, then any intermediate choice $\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}$ also gives a payoff of at least k. Thus, we have the following:

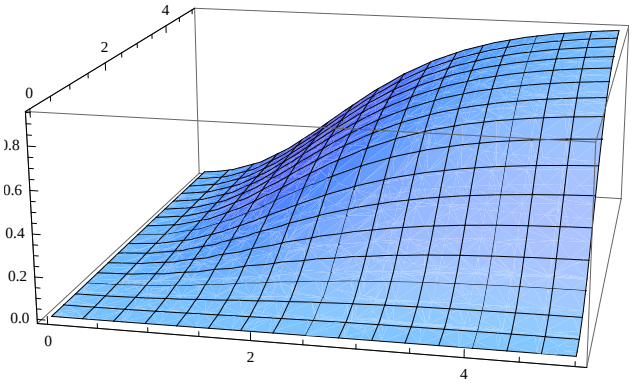
f is quasiconcave [*quasiconvex*] iff all its upper [*lower*] contour sets are convex.



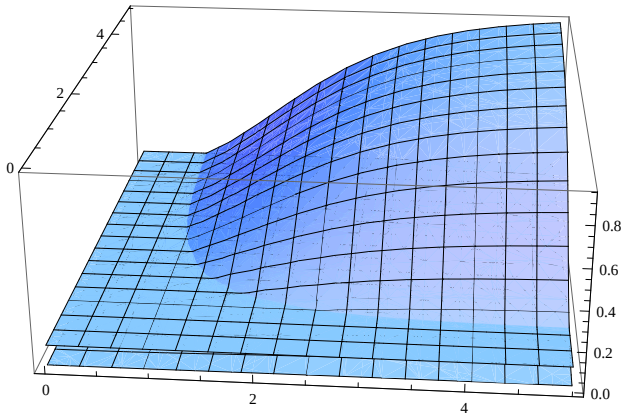
(a) The function $f(x, y) = -e^x - e^y$ is concave.



(b) It is therefore also quasi-concave: convex upper contour sets. E.g. the points giving values above $-.5$ form a convex set.

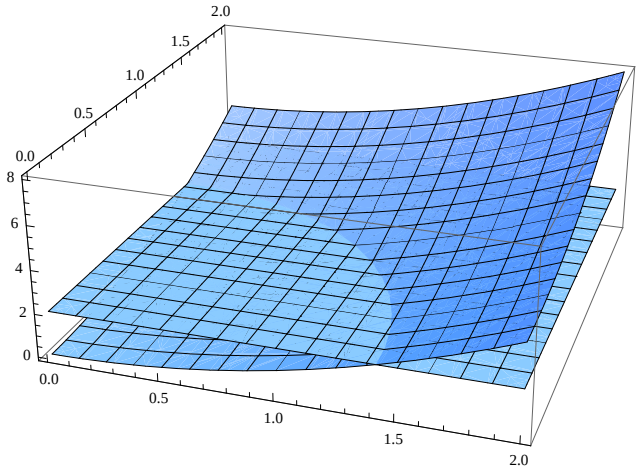


(c) The monotonic transform $e^{5f(x,y)}$ is no longer concave.

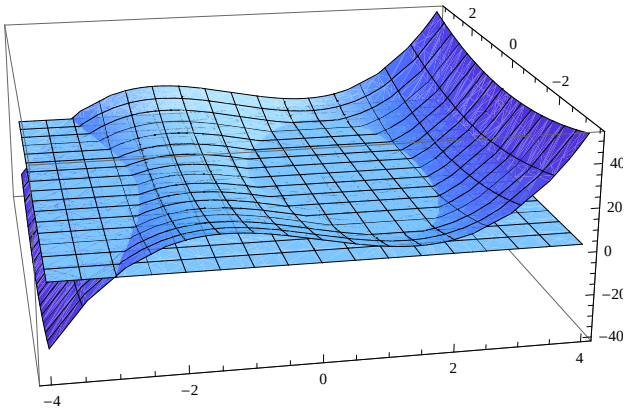


(d) But it is still quasiconcave, with convex upper contour sets.

Figure 3: Concavity and quasiconcavity



(a) UCS_2 for $g(x,y) = x^2 + y^2$ is not convex, therefore g is not quasiconcave. On the other hand, g is convex, and hence is also quasiconvex.



(b) The function $h(x,y) = x^3 - 6x + y^2$ is neither quasiconcave nor convex. We can see both using the same value of $k = 2$: neither UCS_2 nor LCS_2 are convex.

Figure 4: Concavity and quasiconcavity

10 Optimisation

Given $f : \mathbb{R}^n \to \mathbb{R}$, suppose we are interested in the maximum (or the minimum) value f takes over a subset S of the domain. We can write the problem as

$$\max_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{x} \in S$$

We refer to f as the **objective function**, and $\boldsymbol{x}$ as the optimisation variable. The set S is the **constraint set**, i.e., the set over which we are optimising, that is the set of feasible points among which we will look for the solution to the optimisation problem. For example for a standard consumer choice problem, the objective function will be a utility function which represents the consumer's preferences over possible bundles. The optimisation variable is a bundle, and the constraint set will typically be the budget set (i.e., the set of all bundles which are affordable for the consumer). If there are additional constraints, e.g., the consumer must consume at least 5 bananas, then the constraint set will be a strict subset of the set of affordable bundles which contain at least five bananas.

When we write $\arg\max_S f(\boldsymbol{x})$, we refer to the set of points $\boldsymbol{x}^* \in S$ which satisfy the property that $f(\boldsymbol{x}^*) \geq f(\boldsymbol{y})$ for all $\boldsymbol{y} \in S$. Likewise, a point $\boldsymbol{x}^* \in S$ is in $\arg\min_S f(\boldsymbol{x})$ iff $f(\boldsymbol{x}^*) \leq f(\boldsymbol{y})$ for all $\boldsymbol{y} \in S$.

In the context of single-variable optimisation, we have already established a few useful facts. If x is an interior point of the constraint set and if x a local min or a local max of a differentiable function, then $f'(x) = 0$. This is what we called the first order condition (FOC) for optimisation. FOC is a necessary condition for an interior optimum. If the optimal point is on the boundary of the constraint set, the FOC need not be satisfied. The points at which FOC is satisfied are called **critical points** or **stationary points**, and can be classified into local min, local max or saddle points. A global min (if exists) is necessarily a local min. Likewise a global max (if exists) is a local max. Let's first look at the following figure for a classification of extrema.

The function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ depicted above has:

- **Stationary points** $[a, b] \cup \{c, d\}$
- Two (strict) **local maxima** at 0 and c.
  - In an interval around each strict local maximum, f takes a strictly lower value.
- A whole interval $[a, b]$ of (non-strict) local minima.[7]
  - In an area around each local minimum, in this interval, f takes a weakly higher value.
- A **saddle point** at d

[7]Note that all the points in (a, b) are also (non-strict) local maxima. The points a and b, however, are not local maxima.

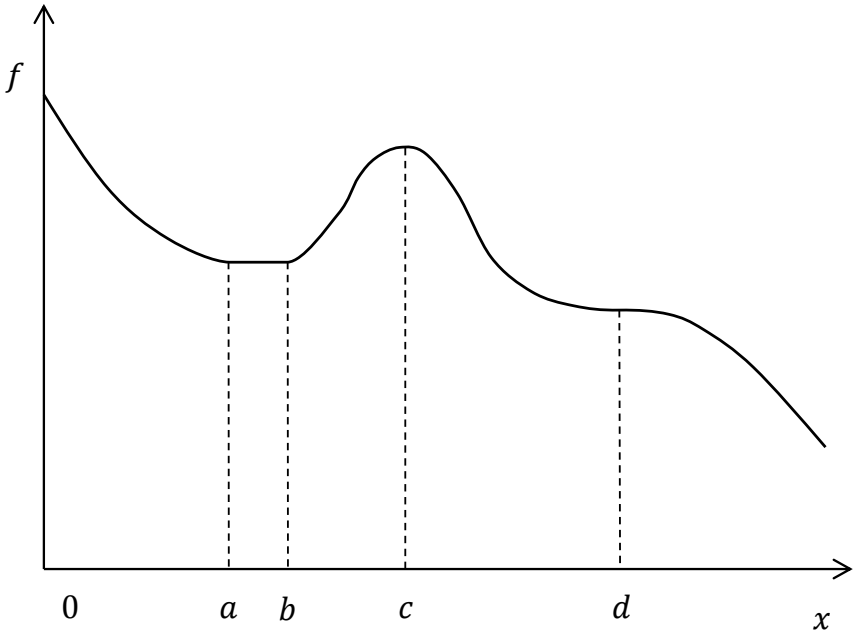


Figure 5: An objective function

– The gradient at d is zero, but there are points arbitrarily close to d where f takes a strictly higher value, and points arbitrarily close where f takes a strictly lower value. So d is neither a local min, nor a local max.

- f has a **global maximum** at 0, i.e., $0 \in \arg\max f$.

10.1 Unconstrained optimisation of $\mathbb{R}^n \to \mathbb{R}$ functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ have all its partial derivatives. Suppose the constraint set for the problem of maximising f is $\mathbb{R}^n$, then we call the problem **unconstrained**.

First, observe that if $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)$ is a solution to our problem, i.e., if it maximises [*minimises*] $f : \mathbb{R}^n \to \mathbb{R}$, then it must be that x_1^* maximises [*minimises*] the function

$$g(\cdot) : \mathbb{R} \to \mathbb{R} \quad \text{defined as} \quad g(u) = g(u, x_2^*, \ldots, x_n^*).$$

This implies that at $u = x_1^*$, the FOC for single-variable optimisation holds, that is,

$$g'(x_1^*) \;=\; 0$$

Rewriting this last equation

$$\frac{\partial f}{\partial x_1}(x_1^*, \ldots, x_n^*) \;=\; 0$$

Repeating the above argument for $x_2, \ldots, x_n$, we conclude that at $\boldsymbol{x} = (x_1^*, \ldots, x_n^*)$, the first order condition holds for f with respect to each variable separately:

FOC for unconstrained optimisation. Suppose $\boldsymbol{x}^* \in \mathbb{R}^n$ is a local max or a local min of a function $f : \mathbb{R}^n \to \mathbb{R}$. If f is differentiable at $\boldsymbol{x}^*$, then

$$\frac{\partial f}{\partial x_i}(x_1^*, \ldots, x_n^*) = 0 \quad \text{for all } i.$$

Note that the FOC holds at saddle points as well. In order to determine whether a critical point is indeed a local max or a local min we will turn to the *second order conditions*.

10.2 Second order conditions (SOC) for extremum points

While the FOC is necessary for an interior point $\boldsymbol{x}^*$ to be a local extremum, it doesn't tell us whether this point is a local max, local min or a saddle point. The second derivatives come in handy at this point.

Suppose f is twice differentiable around an interior critical point $\boldsymbol{x}^*$.

Second Order Sufficient Condition: $\boldsymbol{x}^*$ is a strict local max if the Hessian matrix of f at $\boldsymbol{x}^*$, that is, $D^2 f(\boldsymbol{x})$ is negative definite.

(In this case, x^* is called a **regular** maximum, which has the nice property that the solution is continuous and differentiable in the parameters.)

Second Order Necessary Condition: If $\boldsymbol{x}^*$ is a local max, then the Hessian of f at $\boldsymbol{x}^*$, that is, $D^2 f(\boldsymbol{x})$ is negative semi-definite.

The logic behind the SOC. In order to see how the second order conditions work, look at the second order Taylor approximation of f around the point $\boldsymbol{x}^*$:

$$f(\boldsymbol{x}^* + \boldsymbol{\delta}) \approx f(\boldsymbol{x}^*) + f'(\boldsymbol{x}^*) \cdot \boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta} f''(\boldsymbol{x}^*)\boldsymbol{\delta}^\intercal$$

Since the FOC implies $f'(\boldsymbol{x}^*) = 0$, the above Taylor approximation reduces to

$$f(\boldsymbol{x}^* + \boldsymbol{\delta}) \approx f(\boldsymbol{x}^*) + \frac{1}{2}\boldsymbol{\delta} f''(\boldsymbol{x}^*)\boldsymbol{\delta}^\intercal$$

For local min, replace "negative" with "positive" in the above statements.

A few conclusions

Suppose the domain of f is a convex set K.

- If f is concave [*convex*] on K, then the FOC implies a global max [*min*] over K.
- If SONC is satisfied globally, i.e., if f'' is negative semi-definite everywhere, then FOC implies a global max.

- If f'' is positive semi-definite everywhere, then FOC implies a global min.
- If f is strictly concave [*convex*], FOC implies a unique global max [*min*].
- If f is [*strictly*] quasiconcave, and satisfies FOC and SOSC at $\boldsymbol{x}^*$, then $\boldsymbol{x}^*$ is the [*unique*] global max.

10.3 The envelope theorem for unconstrained optimisation

Given $f : \mathbb{R}^n \to \mathbb{R}$, consider the unconstrained optimisation problem

$$\max_{\boldsymbol{x}} f(\boldsymbol{x}, \alpha)$$

where α is an exogenous parameter, i.e., a variable of f, but not a choice variable of the optimisation problem. Therefore, solving the above problem is solving for optimal $\boldsymbol{x}$ for a given α. Hence, we will denote the solution by $\boldsymbol{x}^*(\alpha)$ in order to keep track of the fact that this optimal choice indeed depends on the value of the parameter α. For simplicity let us denote by $v(\alpha)$ the achieved value of the function when the maximisation problem is solved for a given α. That is, we define the so-called **value function** as $v(\alpha) = f(\boldsymbol{x}^*(\alpha), \alpha)$.

If the maximiser $\boldsymbol{x}^*(\alpha)$ is differentiable, then:

$$\frac{dv}{d\alpha} = \left.\frac{\partial f}{\partial \alpha}\right|_{\boldsymbol{x}=\boldsymbol{x}^*}$$

That is,

$$\begin{array}{c}\text{The rate of change}\\ \text{of the optimal value}\\ \text{with respect to the parameter}\end{array} = \begin{array}{c}\text{The rate of change}\\ \text{of the objective function}\\ \text{with respect to the parameter,}\\ \text{evaluated at the optimal solution}\end{array}$$

Proof. Given parameter α, the optimal value of the objective function is

$$v(\alpha) = f(\boldsymbol{x}^*(\alpha), \alpha)$$

where $\boldsymbol{x}^*(\alpha)$ is the solution to

$$\max_{\boldsymbol{x}} f(\boldsymbol{x}, \alpha)$$

Differentiating $v(\alpha)$ using the Chain Rule

$$\frac{dv}{d\alpha} = \left.\frac{\partial f}{\partial x_1}\right|_{\boldsymbol{x}=\boldsymbol{x}^*} \frac{dx_1^*}{d\alpha} + \cdots + \left.\frac{\partial f}{\partial x_n}\right|_{\boldsymbol{x}=\boldsymbol{x}^*} \frac{dx_n^*}{d\alpha} + \left.\frac{\partial f}{\partial \alpha}\right|_{\boldsymbol{x}=\boldsymbol{x}^*} \frac{d\alpha}{d\alpha}$$

We know that the FOC is satisfied at the solution, that is,

$$\left.\frac{\partial f}{\partial x_1}\right|_{\boldsymbol{x}=\boldsymbol{x}^*} = \cdots = \left.\frac{\partial f}{\partial x_n}\right|_{\boldsymbol{x}=\boldsymbol{x}^*} = 0,$$

and $d\alpha/d\alpha = 1$, hence the expression for $dv/d\alpha$ simplifies to

$$\frac{dv}{d\alpha} = \left.\frac{\partial f}{\partial \alpha}\right|_{\boldsymbol{x}=\boldsymbol{x}^*}$$

□

It is common to refer to $\frac{\partial f}{\partial \alpha}$ as the **direct effect** of changing α on the value of $f(\boldsymbol{x}, \alpha)$. This term ignores the effect of change in $\boldsymbol{x}^*(\alpha)$.

What's often called an **indirect effect** has to do with the fact that a change in α leads to a change in the optimal choice $\boldsymbol{x}^*$, and through that change we have a change in the value of $f(\boldsymbol{x}^*, \alpha)$. This is captured by $\sum f_i \frac{dx_i^*}{\partial \alpha}$.

The envelope theorem rests on the observation that marginal changes in the exogenous parameter have zero indirect effect on the value function. (Note the distinction between the objective function and the value function.)

10.4 Using IFT to do comparative statics

Consider the problem $\max f(x, \alpha)$ over $x \in \mathbb{R}$ (or more generally $\boldsymbol{x} \in \mathbb{R}^n$). Suppose there is a unique maximiser $x^*(\alpha)$. That is, given parameter α, the optimal x^* is unique. We'd like to see how x^* varies as a function of α.

The maximiser $x^*(\alpha)$ must satisfy the FOC:

$$\frac{\partial f}{\partial x}(x^*(\alpha), \alpha) \quad = \quad 0 \qquad \text{(FOC)}$$

Given that there is a unique solution x^* for each α, the FOC defines x implicitly in terms of α. Solving this equation for $x^*(\alpha)$, and then differentiating the solution with respect to α should deliver the desired comparative static exercise. However, solving such an equation might be quite cumbersome. Luckily, computing $dx^*/d\alpha$ often does not require solving for a functional expression for $x^*(\alpha)$.

Instead, let us begin by differentiating (FOC) with respect to α. Using the chain rule:

$$\frac{d}{d\alpha}\left(\frac{\partial f}{\partial x}(x^*(\alpha), \alpha)\right) \quad = \quad f_{xx}\frac{dx^*}{d\alpha} + f_{x\alpha} \quad = \quad 0, \qquad (\tfrac{d}{d\alpha}\text{FOC})$$

so

$$\boxed{\frac{dx^*}{d\alpha} \quad = \quad -\frac{f_{x\alpha}}{f_{xx}}}$$

When x is one-dimensional, $x^*(\alpha)$ being the unique maximum means the second derivative of f at that point (i.e., f_{xx}) is negative. Hence, if $f_{x\alpha} > 0$, then $\frac{dx^*}{d\alpha} > 0$. Likewise, $f_{x\alpha} < 0$ implies $\frac{dx^*}{d\alpha} < 0$. Thus the sign of $f_{x\alpha} > 0$ is sufficient for us to know the direction of change in the maximiser as a result of a change in α. (Often this direction of change is what we want to know.)

If f is from $\mathbb{R}^{n+1}$ to $\mathbb{R}$ with $n > 1$ (i.e., $\boldsymbol{x}$ is $n > 1$ dimensional), then f_{xx} in the expression $\frac{d}{d\alpha}$FOC becomes the Hessian matrix:

$$H = \begin{bmatrix} f_{11} & \cdots & f_{1n} \\ \vdots & \ddots & \vdots \\ f_{n1} & \cdots & f_{nn} \end{bmatrix} \quad \text{evaluated at } \boldsymbol{x}^*(\alpha) = (x_1^*(\alpha), \ldots, x_n^*(\alpha))$$

and

$$\boxed{\left(\frac{d\boldsymbol{x}^*}{d\alpha}\right)^\intercal = \left(\frac{dx_1^*}{d\alpha}, \ldots, \frac{dx_n^*}{d\alpha}\right) = -H^{-1}\left(f_{x_1\alpha}, \ldots, f_{x_n\alpha}\right)^\intercal}$$

Note: For multiple parameters (as in α, β, γ, etc. in $f(x, \alpha, \beta, \gamma, \ldots)$), we would work with the partial derivatives $\frac{\partial x^*}{\partial \alpha}$, $\frac{\partial x^*}{\partial \beta}$, $\frac{\partial x^*}{\partial \gamma}$, etc.

Example. A monopolist has constant marginal cost C of production. The demand for the monopolist's product is such that when it supplies the market with q units, the resulting market price $A - Bq$. Hence its profit as a function of its supply is $\pi(q) = q(A - Bq - C)$.

Assume an interior optimum with $q > 0$. $\pi_{qq} = -2B$, so the SOC is satisfied and we can apply implicit function theorem results. Let's find the effect of changing the parameters without even solving the problem.

The optimal quantity varies as a function of A, and this dependence is captured by dq^*/dA, which is equal to

$$\frac{dq^*}{dA} = \frac{\pi_{qA}}{\pi_{qq}} = \frac{1}{2B}$$

Thus q^* increases as A increases.

Likewise, we have

$$\frac{dq^*}{dB} = \frac{\pi_{qB}}{\pi_{qq}} = \frac{-2q^*}{2B} = -\frac{q^*}{B},$$

so q^* decreases in B.

Finally,

$$\frac{dq^*}{dC} = \frac{\pi_{qC}}{\pi_{qq}} = \frac{-1}{2B},$$

and therefore q^* is decreasing in C. ◇

Example. Given parameters p, q, a firm chooses $x \geq 0$ and $y \geq 0$ to maximise $\pi = px + qy - x^3 - (y - 2x)^2$.

1. Suppose $p = 8, q = 2$. Calculate the firm's optimal choice of (x, y).
2. How do firm's choices of x and y respond to marginal changes in p when $p = 8, q = 2$?

Solution. The FOC

$$\begin{aligned} \pi_x &= p - 3x^2 + 4(y - 2x) = 0 \\ \pi_y &= q - 2(y - 2x) \end{aligned}$$

Can verify that $x = 2$, $y = 5$ satisfies this when $p = 8$ and $q = 2$.

The Hessian of π with respect to (x, y)

$$\begin{pmatrix} \pi_{xx} & \pi_{xy} \\ \pi_{xy} & \pi_{yy} \end{pmatrix} = \begin{pmatrix} -6x - 8 & 4 \\ 4 & -2 \end{pmatrix}$$

is positive definite everywhere, so the FOC yields the unique maximum on the domain $x, y \geq 0$.

The rate of change of x^* and y^* with respect to p is given by

$$\begin{pmatrix} \frac{\partial x^*}{\partial p} \\ \frac{\partial y^*}{\partial p} \end{pmatrix} = -H^{-1} \begin{pmatrix} \pi_{xp} \\ \pi_{yp} \end{pmatrix}$$

The Hessian at $p = 8, q = 2$ is

$$H = \begin{pmatrix} -20 & 4 \\ 4 & -2 \end{pmatrix}$$

and its inverse is

$$H^{-1} = -\frac{1}{12} \begin{pmatrix} 1 & 2 \\ 2 & 10 \end{pmatrix}$$

Hence, at $p = 8, q = 2$

$$\begin{pmatrix} \frac{\partial x^*}{\partial p} \\ \frac{\partial y^*}{\partial p} \end{pmatrix} = \frac{1}{12} \begin{pmatrix} 1 & 2 \\ 2 & 10 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1/12 \\ 1/6 \end{pmatrix}$$

$\diamond$

Example. A monopolist produces Q units of a good, at total cost $c(Q)$. The price is then $P(Q)$, where P is the inverse demand function. Assume $P' < 0$, $P'' \leq 0$, $c' \geq 0$, and $c'' \geq 0$.

1. What is the profit function? What is the FOC for an optimum Q^*? Is a point satisfying the FOC necessarily an optimum?

2. Assume the FOC is satisfied. State, with justification, whether the following changes increase or decrease the optimal Q^*:

   (a) An everywhere increase in marginal cost

   (b) A constant increase in the inverse demand function

   (c) An everywhere decrease in P' (i.e. it becomes more negative), which keeps $P(0)$ fixed.

Solution.

1. Profit is $\pi(Q) = QP(Q) - c(Q)$. Thus the first order condition is

$$\pi'(Q^*) = P(Q) + QP'(Q) - c'(Q) = 0$$

2. The second derivative is $\pi''(Q) = 2P'(Q) + QP''(Q) - c''(Q) < 0$, so π is strictly concave, and therefore a solution to the FOC give the unique maximiser.

3. Say an exonegous change increases $\pi'(Q)$ at Q^*.

   π' is strictly decreasing in Q^*, therefore for the FOC to still hold, Q^* must rise.

   (Using the implicit function theorem: for any parameter α,

   $$\frac{\partial Q^*}{\partial \alpha} = -\frac{\pi_{q\alpha}}{\pi_{qq}}$$

   Since $\pi_{qq} < 0$, the sign of $\frac{\partial Q^*}{\partial \alpha}$ is the same as the sign of $\pi_{q\alpha}$, evaluated at Q^*.

   (a) An everywhere increse in c' leads to a decrease in

   $$\pi'(Q^*) = P(Q) + QP'(Q) - c'(Q)$$

   Since π' decreasing in Q, for the FOC to continue to be satisfies, Q^* must decrease.

   (b) A constant increase in $P(Q)$ does not affect $P'(Q)$, and so increases $\pi'(\cdot)$.
   For the FOC to continue to hold, Q^* must increase.

   (c) An everywhere decrease in P' which keeps $P(0)$ fixed decreases $P(\cdot)$ and decreases $P'(\cdot)$, so decreases $\pi'(\text{cot})$.
   Q^* must decrease for the FOC to continue to hold.

$\diamond$

11 Equality-constrained optimisation

Suppose we have an objective function $f : \mathbb{R}^n \to \mathbb{R}$, meaning we are interested in identifying those x which maximise or minimise f. In most applications, we would have an additional requirement that solutions to this optimisation problem satisfy some extra conditions. For example, f might be a utility function, and we might require the solutions to also satisfy a budget constraint. When the constraints that we impose on the solutions can be expressed via equations (for example budget identities in the case of a standard consumer choice problem), we refer to the problem as one of equality-constrained optimisation.

Say the constraints we want the solutions to satisfy are given by $m \leq n$ equations:

$$\begin{aligned} g^1(\boldsymbol{x}) &= 0 \\ &\vdots \\ g^m(\boldsymbol{x}) &= 0 \end{aligned}$$

The **constraint set** is the set of points in $\mathbb{R}^n$ which satisfy these equations. Written concisely:

$$C = \{\boldsymbol{x} \in \mathbb{R}^n \ : \ g(\boldsymbol{x}) = 0\}$$

where $g : \mathbb{R}^n \to \mathbb{R}^m$ describe the constraints.

We refer to those point in C as **feasible**, that is, $\boldsymbol{x}$ is called feasible if it satisfies the constraints specified for the optimisation problem.

11.1 Constraint Qualification

Now suppose $\boldsymbol{a}$ is a feasible point, i.e., $g(\boldsymbol{a}) = 0$. If g satisfies a particular condition which we will explain below, the set C of feasible points has an $(n-m)$-dimensional shape around the point $\boldsymbol{a}$. (Formally, there exists a small enough neighbourhood of $\boldsymbol{a}$ whose intersection with C is an $(n-m)$-dimensional manifold.)

This condition, sometimes described as the constraints being locally linearly independent around $\boldsymbol{a}$ is called the Constraint Qualification (CQ) at $\boldsymbol{a}$. It requires each g^i to be differentiable at $\boldsymbol{a}$, and the m vectors $\nabla g^1(\boldsymbol{a}), \ldots, \nabla g^m(\boldsymbol{a})$ to be linearly independent.

In other words, CQ is satisfied at $\boldsymbol{a} \in C$ if the $m \times n$ matrix

$$\begin{bmatrix} g_1^1(\boldsymbol{a}) & \ldots & g_n^1(\boldsymbol{a}) \\ \vdots & & \vdots \\ g_1^m(\boldsymbol{a}) & \ldots & g_n^m(\boldsymbol{a}) \end{bmatrix}$$

has full rank m, where g_i^j stands for the partial derivative of g^j with respect to its ith variable.

Equivalently, the above matrix has an $m \times m$ submatrix with non-zero determinant. When this condition holds, we will say that the **Non Degenerate Constraint Qualification (NDCQ)** is satisfied. This is a mild condition, and when $m < n$, it is generically[8] satisfied.

[8]The word "generically" has a technical definition which is beyond the material covered in these notes. Roughly speaking, it captures the idea that if the functions g^j are "randomly" picked according to some natural notion of randomness, then "generically" means "with probability one".

In particular, for most economic applications it will hold. For example, when the constraints $g^1 = \cdots = g^m = 0$ are linear, the constraint qualification becomes an easy condition, and we don't even need to consider it.

11.2 Lagrange's theorem

Given an objective function

$$f : \mathbb{R}^n \to \mathbb{R}$$

and constraints

$$\begin{array}{lll} g^1 : & \mathbb{R}^n \to \mathbb{R} & \\ \vdots & \vdots & g(\boldsymbol{x}) = \mathbf{0} \\ g^m : & \mathbb{R}^n \to \mathbb{R} & \end{array}$$

the associated Lagrangian is a function from $\mathbb{R}^{n+m}$ to $\mathbb{R}$ defined as

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) \quad = \quad f(\boldsymbol{x}) + \sum_{j=1}^{m} \lambda_j g^j(\boldsymbol{x})$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$ is called the vector of **Lagrange multipliers**.

Note that if $\boldsymbol{x}$ is a feasible point (i.e., $g(\boldsymbol{x}) = \mathbf{0}$), then $L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x})$.

The idea behind the Lagrangian approach to optimisation is to study the behaviour of L without any restrictions on $\boldsymbol{x}$, instead of focusing on the behaviour of f on the constraint set.

So, instead of imposing $g(\boldsymbol{x}) = 0$ to be satisfied, the objective function incorporates marginal costs $\lambda_1, \ldots, \lambda_m$ for violating the constraints $g^1 = 0, \ldots, g^m = 0$, respectively. These marginal costs are sometimes referred to as **shadow prices**.[9]

If $\boldsymbol{x}^*$ is a local max of L and if it satisfies the constraint, then it is automatically a local max of L on the constraint set. But, since $L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x})$ for all $\boldsymbol{x}$ on the constraint set, $\boldsymbol{x}^*$ is also a local max of f on the constraint. Clearly, the same argument applies for a local min. So we have:

Lagrange's theorem. Suppose $\boldsymbol{x}^*$ is a local max or a local min of f subject to the m equality constraints given by $g(\boldsymbol{x}) = 0$. Suppose also f and each g^i are differentiable at $\boldsymbol{x}^*$, and CQ is satisfied at $\boldsymbol{x}^*$. Then there are numbers $\lambda_1^*, \ldots, \lambda_m^*$ such that

$$\begin{array}{rcll} \dfrac{\partial L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)}{\partial x_i} & = & 0 & \text{for } i = 1, \ldots, n \\ \dfrac{\partial L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)}{\partial \lambda_j} & = & 0 & \text{for } j = 1, \ldots, m. \end{array}$$

Note that the second row of equations above, i.e., the one involving the partial derivatives with respect to λ_j is redundant, because the fact that $\boldsymbol{x}^*$ is feasible implies $\frac{\partial L}{\partial \lambda_j}(\boldsymbol{x}^*, \boldsymbol{\lambda}) = 0$

[9] The cost-minimisation exercise in consumer/producer theory provides a intuitive interpretation for the term *shadow prices*.

for all $\boldsymbol{\lambda}$. We wrote those m equations as part of the conclusion anyway, because in order to solve for $\boldsymbol{x}^*$ we will need to jointly solve the n equations coming from $\frac{\partial L}{\partial x_i}$ along with the m feasibility equations (contraints).

11.3 Applying Lagrange's theorem

The spirit of the Lagrangian methods is to convert a constrained optimisation problem into an unconstrained optimisation problems, which in turn, reduces to the set of equations implied by the FOC.

In order to solve a constrained optimisation problem with equality constraints:

1. Find all the solutions of the Lagrangian FOC which involves solving $n+m$ equations in $n+m$ unknowns.
2. Verify whether NDCQ is satisfied everywhere (in most economic problems it is indeed satisfied), and make a note of those points where NDCQ fails.
3. Evaluate the function at all those points you identified (solutions to the FOC set of equations as well as the points where NDCQ fails), and thus find which ones are indeed the global maxima/minima.

Many economic problems are nice as in maximising a quasiconcave function over a convex set, and there will be only one solution of the FOC, which must be the global maximum.

Example. Consider the problem of

$$\text{maximising } xyz \text{ subject to the conditions } x^2+y^2=1 \text{ and } x+z=1.$$

If we want this problem to look like the one we explained above, we can set $f(x,y,z)=xyz$ as the objective function, and represent the two equality constraints by setting the function $g:\mathbb{R}^3\to\mathbb{R}^2$ to $\mathbf{0}$, where

$$\begin{aligned} g^1(x,y,z) &= x^2+y^2-1 \\ g^2(x,y,z) &= x+z-1 \end{aligned}$$

Hence, we can rewrite the problem as

$$\max f(x,y,z) \qquad \text{subject to} \quad g(x,y,z)=\mathbf{0}$$

In order to check NDCQ, we should look at the rank of the matrix

$$Dg\,(x,y,z) = \begin{pmatrix} 2x & 2y & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

This has rank less than 2 only if $x=y=0$. But $x=y=0$ does not satisfy the first constraint (which requires $x^2+y^2=1$), so NDCQ holds on the whole constraint set. The Lagrangian is

$$L = xyz+\lambda_1\left(1-x^2-y^2\right)+\lambda_2\left(1-x-z\right)$$

FOC (which can be summarised as $\nabla L = 0$) are

$$\begin{aligned} L_x &= yz - 2\lambda_1 x - \lambda_2 = 0 \\ L_y &= xz - 2\lambda_1 y = 0 \\ L_z &= xy - \lambda_2 = 0 \\ L_{\lambda_1} &= 1 - x^2 - y^2 = 0 \\ L_{\lambda_2} &= 1 - x - z = 0 \end{aligned}$$

Using the second equation to find λ_1 and λ_2 in terms of x, y, z, and substituting these into the first equation gives:

$$\begin{aligned} yz - 2\left(\frac{xz}{2y}\right)x - xy &= 0 \\ y^2 z - x^2 z - xy^2 &= 0 \end{aligned}$$

Using the fourth equation for y^2 in terms of x^2, and the last equation for z in terms of x gives:

$$\begin{aligned} \left(1 - x^2\right)(1 - x) - x^2(1 - x) - x\left(1 - x^2\right) &= 0 \\ (1 - x)\left(1 - x - 3x^2\right) &= 0 \end{aligned}$$

The solutions to this equation are $x = 1$, $x = \frac{-1 \pm \sqrt{13}}{6}$.

So we have five candidates, which are approximately:

$$\begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$$
$$\begin{pmatrix} .43 & \pm .90 & .57 \end{pmatrix}$$
$$\begin{pmatrix} -.77 & \pm .64 & 1.77 \end{pmatrix}$$

We know that a maximiser exists (Extreme Value Theorem) and every maximum (and minimum) must be among the five candidates above (Lagrange's theorem). So the maximiser is $\begin{pmatrix} -.77 & -.64 & 1.77 \end{pmatrix}$.

11.4 The envelope theorem for constrained optimisation

Remember that the envelope theorem concerns how the optimised value changes with respect to a parameter $\alpha \in \mathbb{R}$ which is an exogenous variable in the sense that the problem faced by the agent changes as this exogenous variable changes. When we use the word parameter (i.e., the exogenous variable) for α, we mean to distinguish this number from the agent's choice variables. For example α might be the temperature or the interest rate which can change from one time to another, but the agent has no effect on this variable (hence it is exogenously given to the agent).

Formally, denoting by x the agent's choice variable, suppose the agent's problem is

$$\text{choose } x \text{ to maximise } f(x, \alpha) \text{ subject to } x \in C(\alpha)$$

Note that in the above formulation, we allow both the objective function and the constraint set to depend on the exogenous variable (i.e., the parameter) α. As α changes, the agent's

optimisation problem changes. As a result, the agent's optimal choice, and therefore the optimal value of the objective function will change. How does this optimal value vary with α?

We'll assume two things about the problem:

- There is a unique maximiser $x^*(\alpha) = \arg\max\{f(x,\alpha) \mid x \in C(\alpha)\}$ satisfying the FOC
- It is differentiable w.r.t. α

With the above assumptions, we rephrase the above question:

$$\text{How does } f(x^*(\alpha), \alpha) \text{ change as } \alpha \text{ changes?}$$

For notational simplicity, the optimal value for parameter α is often denoted by $v(\alpha)$ and $v(\cdot)$ is called the agent's **value function**. Note that the agent's choice variable does not feature as a variable of this function, because this being the optimal value really means that the choice is actually determined (of course optimally) for every given α. At the risk of repeating ourselves, we can rewrite $v(\cdot)$ as

$$v(\alpha) \quad = \quad \max_x\{f(x,\alpha) \mid x \in C(\alpha)\} \quad = \quad f(x^*(\alpha))$$

To make it even more concrete, suppose we have a constrained optimisation problem: $\max f(\boldsymbol{x}, \alpha)$ subject to $g^1(\boldsymbol{x}, \alpha) = \cdots = g^m(\boldsymbol{x}, \alpha) = 0$. The associated Lagrangian is

$$L(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m, \alpha) \quad = \quad f(\boldsymbol{x}, \alpha) + \sum_{j=1}^{m} \lambda_j g^j(\boldsymbol{x}, \alpha),$$

Given α, if $\boldsymbol{x}^*(\alpha)$ is a solution to the above constrained optimisation problem, then by Lagrange's theorem, there must exist $\boldsymbol{\lambda}^*(\alpha)$ such that $(\boldsymbol{x}^*(\alpha), \boldsymbol{\lambda}^*(\alpha))$ satisfies the Lagrangian FOC.

We might remember from our discussion of the envelope theorem for unconstrained optimisation (i.e., in the absence of constraints $g^i = 0$) that changing α marginally has zero indirect effect (captured by $\sum \frac{\partial f}{\partial x_i}\frac{\partial x_i^*}{\partial \alpha}$) on the optimised value due to $\frac{\partial f}{\partial x_i}$ being zero at the optimal solution. These partial derivatives are not necessarily zero any more, because the FOC is satisfied for the Lagrangian function $L(\boldsymbol{x}, \boldsymbol{\lambda})$, not for he objective function f. However, we can take advantage of the fact that $\boldsymbol{x}^*$ being a solution implies $g(\boldsymbol{x}^*)$ is zero. (After all, we were searching for a point which satisfied the constraints.) More generally, for every $\boldsymbol{x}$ which satisfies the constraint $g(\boldsymbol{x})$, and for every $\boldsymbol{\lambda}$, we have $f(\boldsymbol{x}, \alpha) = L(\boldsymbol{x}, \boldsymbol{\lambda}, \alpha)$. But then

$$v(\alpha) \quad = \quad f(\boldsymbol{x}^*(\alpha), \alpha) \quad = \quad L(\boldsymbol{x^*(\alpha)}, \boldsymbol{\lambda^*(\alpha)}, \alpha)$$

Now, differentiating the above equation with respect to α:

$$v'(\alpha) \quad = \quad \sum_{i=1}^{n} \frac{\partial L}{\partial x_i}\frac{dx_i^*}{d\alpha} + \sum_{i=j}^{m} \frac{\partial L}{\partial \lambda_j}\frac{d\lambda_j^*}{d\alpha} + \frac{\partial L}{\partial \alpha}\frac{d\alpha}{d\alpha}\Bigg|_{\boldsymbol{x}=\boldsymbol{x}^*(\alpha),\ \boldsymbol{\lambda}=\boldsymbol{\lambda}^*(\alpha)}$$

Since we know that $(\boldsymbol{x}^*(\alpha), \boldsymbol{\lambda}^*(\alpha))$ solves the FOC for $L(\boldsymbol{x}, \boldsymbol{\lambda}, \alpha)$, we have

$$\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial \lambda_j} = 0 \Bigg|_{\boldsymbol{x}=\boldsymbol{x}^*(\alpha),\ \boldsymbol{\lambda}=\boldsymbol{\lambda}^*(\alpha)} \qquad \text{for all } \ i = 1, \ldots, n \text{ and } j = 1, \ldots, m$$

Finally, noting that $\frac{d\alpha}{d\alpha} = 1$, we can conclude

The envelope theorem. If the maximiser $\boldsymbol{x}^*(\alpha)$ is differentiable, then:

$$\frac{dv}{d\alpha} = \left.\frac{\partial L}{\partial \alpha}\right|_{\boldsymbol{x}=\boldsymbol{x}^*(\alpha),\ \boldsymbol{\lambda}=\boldsymbol{\lambda}^*(\alpha)}$$

Written in plain words:

The rate of change of the optimal value with respect to the exogenous parameter	=	The rate of change of the Lagrangian with respect to the parameter, evaluated at the optimal solution

If there are multiple parameters $\alpha_1, \alpha_2, \ldots$, then work with $\frac{\partial}{\partial \alpha_i}$ instead of $\frac{d}{d\alpha}$.

Applications to consumer choice

Some key results of consumer theory can be obtained by using the envelope theorem.

The standard choice problem in consumer theory is nothing but maximising a utility function subject to a budget constraint and nonnegativity constraints.

$$\max u(\boldsymbol{x}) \qquad \text{s.t.} \qquad \boldsymbol{p}\cdot\boldsymbol{x} \leq m \quad \text{and} \quad x_i \geq 0 \quad \text{for all } i = 1, \ldots, n$$

Assume increasing u, so the optimal solution involves spending all of m. That is, the budget constraint binds. For simplicity, let us also assume that the preferences are such that the consumer would always choose positive amounts of each good (which, for example, is the case for Cobb-Douglas agents). So the problem can be rewritten as

$$\max u(\boldsymbol{x}) \qquad \text{s.t.} \qquad \boldsymbol{p}\cdot\boldsymbol{x} = m \quad \text{and} \quad x_i > 0 \quad \text{for all } i = 1, \ldots, n$$

Finally, assume that given any m and $\boldsymbol{p}$, there is a unique solution to this problem (which, for example, is the case for strictly quasi-convex preferences).

The Marshallian demand function $\boldsymbol{x}^M(\boldsymbol{p}, m)$ maps given prices and budget to the unique solution of the consumer's problem.

The value function $v(\boldsymbol{p}, m)$ for this problem is called the indirect utility function. It basically maps given prices and budget to the maximum utility the consumer can achieve:

$$v(\boldsymbol{p}, m) = u(\boldsymbol{x}^M(\boldsymbol{p}, m))$$

Since the non-negativity constraints do not bind (i.e., we know $x_i > 0$ for all i), the associated Lagrangian is

$$L = u(\boldsymbol{x}) + \lambda(m - \sum p_i x_i)$$

Assuming differentiability of the demand function, the envelope theorem says

$$\frac{\partial v}{\partial p_i} = \left.\frac{\partial L}{\partial p_i}\right|_{\boldsymbol{x}=\boldsymbol{x}^M(\boldsymbol{p},m),\ \lambda=\lambda^*} = -\lambda^* x_i^M$$

and

$$\frac{\partial v}{\partial m} = \left.\frac{\partial L}{\partial m}\right|_{\boldsymbol{x}=\boldsymbol{x}^M(\boldsymbol{p},m),\ \lambda=\lambda^*} = \lambda^*$$

Dividing the first equation by the second equation gives us the so-called

Roy's identity: $$x_i^M = -\frac{\frac{\partial v}{\partial p_i}}{\frac{\partial v}{\partial m}}$$

Now, let's turn the problem of cost minimisation whose solution is referred to as the Hicksian (or compensated) demand. Again, let us assume that given a utility level $\bar{u}$ (i.e., an indifference curve) and a price vector $\boldsymbol{p}$, there is a unique cheapest bundle which achieves utility $\bar{u}$. That is, the problem

$$\min \boldsymbol{p} \cdot \boldsymbol{x} \quad \text{subject to} \quad u(\boldsymbol{x}) \geq \bar{u} \quad \text{and} \quad x_i \geq 0$$

has a unique solution which we denote by $\boldsymbol{x}^H(\boldsymbol{p}, \bar{u})$.

The value function for this problem is called the **expenditure function**

$$e(\boldsymbol{p}, \bar{u}) = \boldsymbol{p} \cdot \boldsymbol{x}^H(\boldsymbol{p}, \bar{u})$$

Increasing utility implies the constraint $u(\boldsymbol{x}) \geq \bar{u}$ binds. Positive consumption of each good assumption means the non-negativity constraints do not bind, and therefore the relevant Lagrangian is

$$L(\boldsymbol{x}, \mu, \boldsymbol{p}, \bar{u}) = \boldsymbol{p} \cdot \boldsymbol{x} + \mu(\bar{u} - u(\boldsymbol{x}))$$

Assuming $\boldsymbol{x}^H$ is differentiable, we can apply the envelope theorem to conclude

Shephard's lemma: $$\frac{\partial e}{\partial p_i} = \left.\frac{\partial L}{\partial p_i}\right|_{\boldsymbol{x}=\boldsymbol{x}^H(\boldsymbol{p},\bar{u}),\ \mu=\mu^*} = x_i^H$$

and

$$\frac{\partial e}{\partial \bar{u}} = \left.\frac{\partial L}{\partial \bar{u}}\right|_{\boldsymbol{x}=\boldsymbol{x}^H(\boldsymbol{p},\bar{u}),\ \mu=\mu^*} = \mu^*$$

Note that the two optimisation exercises above are dual problems of each other in the sense that given prices and a utility level, if we compute the cheapest bundle achieving that utility (i.e., $\boldsymbol{x}^H(\boldsymbol{p}, \bar{u})$), and then treat the cost of that bundle as a monetary budget (i.e., setting $m = e(\boldsymbol{p}, \bar{u})$), and then compute the most preferred bundle (i.e., $x_i^M(\boldsymbol{p}, e(\boldsymbol{p}, \bar{u}))$) then we must necessarily have the same bundle. That is:

$$x_i^H(\boldsymbol{p}, \bar{u}) = x_i^M(\boldsymbol{p}, e(\boldsymbol{p}, \bar{u})) \qquad (\heartsuit)$$

Alternatively, starting with the Marshallian exercise, we can conclude

$$x_i^M(\boldsymbol{p}, m) = x_i^H(\boldsymbol{p}, v(\boldsymbol{p}, m))$$

Using the Chain Rule to differentiate ($\heartsuit$) with respect to p_j gives

$$\frac{\partial x_i^H}{\partial p_j} = \frac{\partial x_i^M}{\partial p_j} + \frac{\partial x_i^M}{\partial m}\frac{\partial e}{\partial p_j}$$

Since Shephard's lemma gives $\frac{\partial e}{\partial p_j} = x_j^H$, subbing this into the above equation and rearranging yields the so-called

Slutsky equation:	$\dfrac{\partial x_i^M}{\partial p_j} = \dfrac{\partial x_i^H}{\partial p_j} - \dfrac{\partial x_i^M}{\partial m} x_j^H$

Shadow prices

Consider the following "cost minimisation problem"

$$\min_{\boldsymbol{x}} C(\boldsymbol{x}) \quad \text{subject to} \quad \begin{cases} g^1(\boldsymbol{x}) = d^1 \\ \vdots \\ g^m(\boldsymbol{x}) = d^m \end{cases}$$

Once again, for the minimisation problem, $x_1, \ldots, x_n$ are the choice variables, whereas the parameters $d^1, \ldots, d^m$ are exogenously given and treated as fixed constants. As these parameters change, the problem changes, and as a result the solution $\boldsymbol{x}^*(d^1, \ldots, d^m)$ and the optimal achieved value $v(d^1, \ldots, d^m) = C(\boldsymbol{x}^*(d^1, \ldots, d^m))$ of the problem will change. For example in the Marshallian demand example above, there is a single constraint which is the budget identity and the exogenous parameter is the income. As the income changes, the consumer's budget set and choices will change. In the case of Hicksian demand, the single constraint is achieving the utility target. As this target changes, the consumer's cost-minising choices (and therefore expenditure) will change.

The Lagrangian for the above problem is:

$$L = c(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i \left[d^i - g^i(\boldsymbol{x}) \right]$$

By the envelope theorem:

$$\frac{\partial v}{\partial d^i} = \left. \frac{\partial L}{\partial d^i} \right|_{\boldsymbol{x}=\boldsymbol{x}^*,\ \boldsymbol{\lambda}=\boldsymbol{\lambda}^*} = \lambda_i^*$$

In words, we can interpret λ_i^* as the "unit cost associated with increasing the parameter d_i". This is because a marginal increase of δ in d_i leads to an increase of $\lambda_i^* \Delta$ in the optimal value of c. For example if g^i is a production function, and therefore $g^i(\boldsymbol{x}) = d^i$ describes a production target, then λ_i^* is the "shadow price" of producing one extra unit.

11.5 Second Order Conditions for Constrained Optimisation

When the critical points identified in our optimisation problem are interior points of the domain of the objective function, the second order conditions expressed via the Hessian matrix can be used to classify these critical points.

In many contrained optimisation problem, however, the points of interest are typically on the boundary. The relevant second order conditions, in that case, require the use of the so-called **Bordered Hessian**.

Suppose we are given the problem

$$\max_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}) \quad \text{subject to} \quad g^1(\boldsymbol{x}) = g^2(\boldsymbol{x}) = \cdots = g^m(\boldsymbol{x}) = 0$$

That is, a problem with n choice variables and m constraints which lead to the Lagrangian

$$L \quad = \quad f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g^i(\boldsymbol{x})$$

Let $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ be such that the $n+m$ first order conditions for L hold:

$$\frac{\partial L}{\partial x_1} = \cdots = \frac{\partial L}{\partial x_n} = 0, \quad \frac{\partial L}{\partial \lambda_1} = \cdots = \frac{\partial L}{\partial \lambda_m} = 0 \qquad \text{at } (\boldsymbol{x}^*, \lambda^*)$$

Now, consider the Hessian matrix of the Lagrangian at $(\boldsymbol{x}^*, \lambda^*)$, i.e.,

$$D^2L(\boldsymbol{x}^*, \lambda^*) \quad = \quad \left(\begin{array}{ccc|ccc} 0 & \cdots & 0 & -\frac{\partial g^1}{\partial x_1} & \cdots & -\frac{\partial g^1}{\partial x_n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & -\frac{\partial g^m}{\partial x_1} & \cdots & -\frac{\partial g^m}{\partial x_n} \\ -- & -- & -- & -- & -- & -- \\ -\frac{\partial g^1}{\partial x_1} & \cdots & -\frac{\partial g^m}{\partial x_1} & \frac{\partial^2 L}{\partial x_1^2} & \cdots & \frac{\partial^2 L}{\partial x_n \partial x_1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -\frac{\partial g^1}{\partial x_n} & \cdots & -\frac{\partial g^m}{\partial x_n} & \frac{\partial^2 L}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 L}{\partial x_n^2} \end{array}\right) \text{ evaluated at } (\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$$

Given the above optimisation problem, its **Bordered Hessian** matrix is obtained by multiplying each of the last m rows and first m columns of $D^2L(\boldsymbol{x}^*, \lambda^*)$ by -1 to obtain

$$BH \quad = \quad \left(\begin{array}{ccc|ccc} 0 & \cdots & 0 & \frac{\partial g^1}{\partial x_1} & \cdots & \frac{\partial g^1}{\partial x_n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{\partial g^m}{\partial x_1} & \cdots & \frac{\partial g^m}{\partial x_n} \\ -- & -- & -- & -- & -- & -- \\ \frac{\partial g^1}{\partial x_1} & \cdots & -\frac{\partial g^m}{\partial x_1} & \frac{\partial^2 L}{\partial x_1^2} & \cdots & \frac{\partial^2 L}{\partial x_n \partial x_1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g^1}{\partial x_n} & \cdots & \frac{\partial g^m}{\partial x_n} & \frac{\partial^2 L}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 L}{\partial x_n^2} \end{array}\right) \text{ evaluated at } (\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$$

SOC via the Bordered Hessian.

(a) If the last $n-k$ leading principal minors of BH at $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ alternate in sign, where the determinant of BH (i.e., the final leading principal minor) is of the same sign as $(-1)^n$, then $\boldsymbol{x}^*$) is a local maximum of $f(\boldsymbol{x})$ subject to $g(\boldsymbol{x}) = 0$.

(b) If the last $n-k$ leading principal minors of BH at $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ are of the same sign as $(-1)^m$, then $\boldsymbol{x}^*$ is a local minimum of $f(\boldsymbol{x})$ subject to $g(\boldsymbol{x}) = 0$.

(c) If both conditions (a) and (b) are violated by nonzero leading principal minors, then $\boldsymbol{x}^*$ is neither a local max nor a local min of $f(\boldsymbol{x})$ subject to $g(\boldsymbol{x}) = 0$.

12 Optimisation with inequality constraints

An optimisation problem with m inequality constraints looks like

$$\max f(\boldsymbol{x}) \quad \text{subject to} \quad g^j(\boldsymbol{x}) \geq 0, \quad j = 1 \dots m$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function, and $g^j(\boldsymbol{x}) \geq 0$ is the jth constraint for $j = 1, \dots, m$.

If $\boldsymbol{x}^*$ is a solution to this problem such that $g^j(\boldsymbol{x}) > 0$, then we say "for this particular solution $\boldsymbol{x}$, the jth constraint does not bind."

12.1 When none of the constraints bind

If $\boldsymbol{x}^*$ is a solution to the above problem such that none of the constraints bind (i.e., $g^j(\boldsymbol{x}) > 0$ for all $j = 1, \dots, m$), then we say $\boldsymbol{x}^*$ is an interior solution. All such solutions must satisfy the standard the FOC for interior optima:

FOC for interior optima. Suppose $\boldsymbol{x}^* \in \mathbb{R}^n$ is a local max or a local min of a function f whose constraint set is $C \subseteq \mathbb{R}^n$. Assume also that $\boldsymbol{x}^*$ in an interior point, that is, there exists a ball centred at $\boldsymbol{x}^*$ all of which is contained in C. If f is differentiable at $\boldsymbol{x}^*$, then

$$\frac{\partial f}{\partial x_i}(x_1^*, \dots, x_n^*) = 0 \quad \text{for all } i.$$

Not knowing beforehand whether the constraints will bind or not, we need to develop a method to think about how to find solutions for which some of the constraints bind. We begin with some special cases.

12.2 Complementary slackness for a nonnegativity constraint in 1 dimension

We first look at a special case of inequality constraints, namely the so-called nonnegativity constraints which require some (or all) of the choice variables to be nonnegative.

There is a strict local maximum at $x^* = 0$, where the $x \geq 0$ constraint *binds*. That means, if we were allowed to violate the constraint to let x move below 0, we would obtain a higher value than $f(0)$, at least on the margin. This is because $f'(0) < 0$.

The nonnegativity constraint requires us to write the FOC for local maxima more carefully:

$$\underbrace{x^* > 0 \ \text{ and } \ f'(x^*) = 0}_{\text{for an interior local max}}, \quad \text{or} \quad \underbrace{x^* = 0 \ \text{ and } \ f'(x^*) \leq 0}_{\text{for a local max where the } x \geq 0 \text{ constraint binds}}$$

Alternatively, we can write these (more compact but less practical) as

$$x^* \geq 0, \quad f'(x^*) \leq 0, \quad x^* f'(x^*) = 0$$

The last requirement, namely $x^* f'(x^*) = 0$ is called the **complementary slackness** condition. It basically says that if one of the first two requirements holds as a strict inequality (is

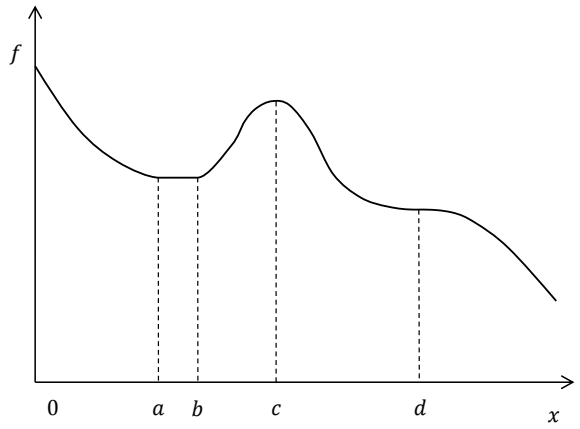


Figure 6: An objective function with nonnegativity constraint.

slack), then the other must hold as an equality (*bind*). Note that it is possible for both of the first two requirements to bind.

12.3 Complementary slackness for nonnegativity constraints in n dimensions

It is not hard to generalise the above FOC to the case of multi-variable optimisation with nonnegativity constraints.

The FOC for a local max. If $\boldsymbol{x}^* \in \mathbb{R}^n$ is a local max of a differentiable f subject to $\boldsymbol{x} \geq \mathbf{0}$, then for each $i = 1, \ldots, n$,

$$\left[x_i^* > 0 \text{ and } \frac{\partial f}{\partial x_i}(\boldsymbol{x}^*) = 0\right] \quad \text{or} \quad \left[x_i^* = 0 \text{ and } \frac{\partial f}{\partial x_i}(\boldsymbol{x}^*) \leq 0\right]$$

The FOC for a local min. If $\boldsymbol{x}^* \in \mathbb{R}^n$ is a local min of a differentiable f subject to $\boldsymbol{x} \geq \mathbf{0}$, then for each $i = 1, \ldots, n$,

$$\left[x_i^* > 0 \text{ and } \frac{\partial f}{\partial x_i}(\boldsymbol{x}^*) = 0\right] \quad \text{or} \quad \left[x_i^* = 0 \text{ and } \frac{\partial f}{\partial x_i}(\boldsymbol{x}^*) \geq 0\right]$$

In our search for local optima we need to check all possible combinations in which the nonnegativity constraints can bind. For example in 2 dimensions, suppose we have the two constraints: $x_1 \geq 0$ and $x_2 \geq 0$. At the optimal solution, there are four possibilities: neither constraint binding, or only the first constraint binding, or only the second constraint binding, or both constraints binding. In our search for the optimal solution, we need to examine each case:

- Test for an interior solution, i.e., $x_1 > 0$, $x_2 > 0$. We have two equations $f_1 = f_2 = 0$. See if these have a solution with $x_1 > 0, x_2 > 0$.

- Test $x_1 = 0$, $x_2 > 0$. We have one equation $f_2 = 0$ in one unknown x_2. See if it has a solution with $x_2 > 0$, and see if this satisfies the requirement $f_1 \leq 0$.
- Test $x_1 > 0$, $x_2 = 0$. Similarly to the case above.
- Test $x_1 = x_2 = 0$. Test $(0, 0)$ to see if $f_1 \leq 0$ and $f_2 \leq 0$ there.

In general with n-variables all of which are required to be nonnegative, we will have 2^n different cases.

12.4 Single variable optimisation with one inequality constraint

Now we turn to inequality constraints that are more general than nonnegativity constraints. In order to explain the basic principle, let's begin with $n = 1$ and a single constraint. To be concrete, suppose we are maximising $u(x) : \mathbb{R} \to \mathbb{R}$ subject to $x \leq 1$.

Suppose x^* is a local max of u subject to $x \leq 1$. If the constraint doesn't bind for this local max, that is, if $x^* < 1$, then the usual FOC holds, because we can move around x^* without hitting the constraint. The fact that x^* is a local max then implies $u'(x^*) = 0$.

If, however, $x^* = 1$, then the constraint binds, and this point might be a local max even though the derivative at this point is non-zero. In particular, perhaps u would increase if we could push x^* beyond 1, but we can't due to the constraint $x \leq 1$. If that's the case, we would expect $u'(1) > 0$. Unlike the nonnegativity constraints, we don't always have a compact way to write the conditions for the solutions for which the constraints bind. So, we will treat them within the more general Lagrangian approach.

So, for the above example, we will convert the 1-variable optimisation problem with 1 constraint into a 2-variable unconstrained optimisation problem. Namely we will study the Lagrangian

$$L(x, \lambda) \quad = \quad u(x) + \lambda(1 - x)$$

Around the point where the constraint binds (i.e., around $x = 1$) the maximiser's instinct is to raise x beyond 1. Instead of imposing the constraint $x \leq 1$, the Lagrangian function allows x to be greater 1, but incorporates a **shadow price** of λ for x going above 1.

And the spirit of the Lagrangian approach is about identifying the correct shadow price λ^* (Lagrangian multiplier) which ensures that a critical point x^* of u subject to the constraint corresponds to a critical point (x^*, λ^*) of L.

A solution $x^* < 1$ is simply an interior solution, and can be identified by checking the FOC for $u(x)$. That, of course, corresponds to the solution of the first order condition $L_x = 0$ for the Lagrangian $L = u(x) + \lambda(1 - x)$, where $\lambda^* = 0$. The constraint does not bind, and the "shadow price of relaxing the constraint is 0". (No need to pay for relaxing a constraint which does not bind.)

If, however, $x^* = 1$ is a solution, then the constraint binds, and allowing for $x > 1$, that is, relaxing the constraint can yield a bigger value of u. That means there will be a shadow price $\lambda^* \geq 0$ to relax the constraint. In other words, the solution $x^* = 1$ satisfies the Lagrangian FOC $L_x = 0$ with some $\lambda^* \geq 0$.

To summarise, the problem $\max u(x)$ s.t. $x \leq 1$ can be solved by setting $L(x, \lambda) = u(x) + \lambda(1 - x)$, and considering both cases below:

- Solve $L_x = 0$ with $\lambda^* = 0$, and see if this gives any $x^* < 1$.
- Set $x^* = 1$, and see if $L_x = 0$ is met for some $\lambda^* \geq 0$.

12.5 Inequality constraints with multiple variables

For ease of exposition, let's first discuss the case of a single constraint.

Single constraint

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function. Consider the problem

$$\max f(\boldsymbol{x}) \quad \text{subject to} \quad g(\boldsymbol{x}) \geq 0$$

Assume also that g is differentiable, and that NDCQ is satisfied for those $\boldsymbol{x}$ where the constraint binds. That is, we also assume that if $g(\boldsymbol{x}) = c$, then $\nabla g(\boldsymbol{x}) = (g_1(\boldsymbol{x}), g_2(\boldsymbol{x}), \ldots, g_n(\boldsymbol{x})) \neq \mathbf{0}$.

The Lagrangian $L : \mathbb{R}^{n+1} \to \mathbb{R}$ is given by

$$L(\boldsymbol{x}, \lambda) \quad = \quad f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$$

If $\boldsymbol{x}^*$ is a solution, then there exists λ^* such that

- First Order Conditions for $\boldsymbol{x}$ hold: $\frac{\partial L}{\partial x_i}(\boldsymbol{x}^*, \lambda^*) = 0$ for each $i = 1, \ldots, n$,
- Complementary Slackness condition holds:

$$\underbrace{g(\boldsymbol{x}^*) > 0 \ \text{ and } \ \lambda^* = 0}_{\text{the constraint not binding}} \quad \text{or} \quad \underbrace{g(\boldsymbol{x}^*) = 0 \ \text{ and } \ \lambda^* \geq 0}_{\text{the constraint binding}}$$

Multiple constraints

Given a differentiable $f : \mathbb{R}^n \to \mathbb{R}$, consider the problem

$$\max f(\boldsymbol{x}) \quad \text{subject to} \quad g^j(\boldsymbol{x}) \geq 0, \ \ j = 1, \ldots, m$$

The Lagrangian is the same as for equality constraints:

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) \quad = \quad f(\boldsymbol{x}) + \sum_{j=1}^{m} \lambda_j g^j(\boldsymbol{x})$$

Assume that g^j is differentiable for each $j = 1, \ldots, m$.

Assume also that NDCQ is satisfied, that is, given a feasible point $\boldsymbol{x}$, if J is the set $\{j \mid g_j(\boldsymbol{x}) = 0\}$, then the set of vectors $\{\nabla g^j(\boldsymbol{x}) \mid j \in J\}$ is linearly independent.[10]

If $\boldsymbol{x}^*$ is a solution to our problem, then there must exist $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_m^*)$ such that

- First Order Conditions for $\boldsymbol{x}$ hold: $\frac{\partial L}{\partial x_i}(\boldsymbol{x}^*) = 0$ for each $i = 1, \ldots, n$,
- Complementary Slackness condition holds for each $j = 1, \ldots, m$:

$$\underbrace{g^j(\boldsymbol{x}^*) > 0 \text{ and } \lambda_j = 0}_{\text{th } j\text{th constraint not binding}} \quad \text{or} \quad \underbrace{g^j(\boldsymbol{x}^*) = 0 \text{ and } \lambda_j \geq 0}_{\text{the } j\text{th constraint binding}}$$

It is worth noting that the envelope theorem continues to hold. That is, suppose the problem evolves (in a differentiable way) with an exogenous parameter α. Denoting by v the value function

$$\frac{dv}{d\alpha} = \left.\frac{\partial L}{\partial \alpha}\right|_{\boldsymbol{x}=\boldsymbol{x}^*,\ \boldsymbol{\lambda}=\boldsymbol{\lambda}^*}$$

An example: different borrowing and saving rates

Suppose a consumer has income $y_1 = 1$ in period 1, and $y_2 = 1$ in period 2. She can save at rate $r_s = 1$ and borrow at rate $r_b = 2$, meaning if she saves a pound in period 1, she has an extra pound in period 2, whereas if she borrows a pound in period 1, she needs to pay back 2 pounds in period 2.

Denoting by c_i her consumption in period i, her preferences over consumption streams (c_1, c_2) are described by a utility function:

$$U(c_1, c_2) = \ln(c_1) + \delta \ln(c_2),$$

where $0 \leq \delta \leq 1$.

First note that both c_1 and c_2 have to be positive for U to be well-defined.

Her consumption possibility set is depicted in Figure 7.

The problem is

$$\max_{c_1, c_2} U(c_1, c_2) \quad \text{s.t.} \quad \begin{cases} c_1 \leq 1 + \frac{1-c_2}{2} \\ c_2 \leq 1 + 1 - c_1 \\ c_1 > 0 \\ c_2 > 0 \end{cases}$$

Since $c_1 > 0$ and $c_2 > 0$ are strict inequalities, these constraints cannot bind, and therefore will not feature in the Lagrangian. So the Lagrangian for the above problem is

$$L = U(c_1, c_2) + \lambda_b(3 - 2c_1 - c_2) + \lambda_s(2 - c_1 - c_2)$$

The FOC for c_1 and c_2:

$$L_{c_1} = 0 \quad = \quad \frac{1}{c_1} - \lambda_s - 2\lambda_b \tag{2}$$

$$L_{c_2} = 0 \quad = \quad \delta\frac{1}{c_2} - \lambda_s - \lambda_b \tag{3}$$

[10] It is worth noting that NDCQ is a mild condition, and in most applications it won't be hard to verify, without having to resort to tedious machinery, that it holds.

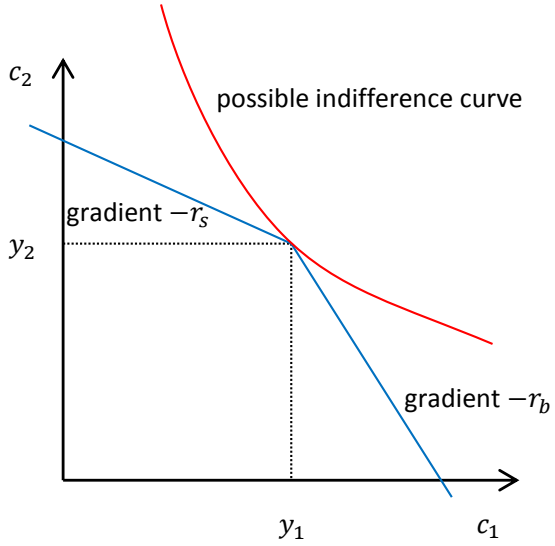


Figure 7: Different borrowing and saving rates

And complementary slackness:

- Saving constraint binds, and $\lambda_s \geq 0$, or
- Borrowing constraint binds, and $\lambda_b \geq 0$, or
- Both constraints bind, and $\lambda_s \geq 0$ and $\lambda_b \geq 0$.

Case 1: Money is being saved
Saving constraint binds ($c_1 + c_2 = 2$, $\lambda_s \geq 0$), borrowing does not ($c_1 < y_1$, $\lambda_b = 0$).

Solve the equalities: putting $\lambda_b = 0$ into 2 and 3 gives $\lambda_s = \frac{\delta}{c_2} = \frac{1}{c_1} (\geq 0)$, i.e. $r_s = \frac{c_2}{c_1}$

Solving this and the binding saving constraint gives $(c_1, c_2) = \left(\frac{2}{(1+\delta)}, \frac{2}{1+\delta}\delta\right)$.

$\lambda_s \geq 0$ is satisfied, and we also need $c_1 < 1$ (non-binding borrowing constraint), but this is not possible since $\delta \leq 1$.

Case 2: Money is being borrowed
Borrowing constraint binds ($2c_1 + c_2 = 3$, $\lambda_b \geq 0$), saving does not ($c_1 > y_1$, $\lambda_s = 0$)

Similarly to above, there is solution $(c_1, c_2) = \left(\frac{3}{2(1+\delta)}, \frac{3}{1+\delta}\delta\right)$ which has $c_1 > y_1$ when $\delta < \frac{1}{2}$. ($\lambda_b = \frac{1+\delta}{3} \geq 0$)

Case 3: Neither borrower nor lender
$(c_1, c_2) = (1, 1)$ (both constraints bind) and $\lambda_b, \lambda_s \geq 0$.

Plug these into FOC for c_1 and c_2 to obtain

$$\begin{aligned} 1 - \lambda_s - 2\lambda_b &= 0 \\ \delta - \lambda_s - \lambda_b &= 0 \end{aligned}$$

which yield $\lambda_s = 2\delta - 1$ and $\lambda_b = 1 - \delta$.

The requirement for $\lambda_s, \lambda_b \geq 0$ is $\frac{1}{2} \leq \delta$.

Thus we have the solution:

$$(c_1, c_2) = \begin{cases} \left(\frac{3}{2(1+\delta)}, \delta\frac{3}{(1+\delta)}\right) & \text{if} \quad 0 \leq \delta < \frac{1}{2} \\ (1,1) & \text{if} \quad \frac{1}{2} \leq \delta \leq 1 \end{cases}$$

12.6 A special case: the Kuhn-Tucker Lagrangian

Suppose some of the inequality constraints are non-negativity constraints. Like we discussed at the very beginning of this chapter, the simpler nature of the non-negativity constraints can allow us to also simplify the Lagrangian function we will employ.

Take, for example, a common economic problem where we are given utility function $u : \mathbb{R}^n \to \mathbb{R}$ and we are asked to solve

$$\max f(\boldsymbol{x}) \quad \text{subject to} \quad g(\boldsymbol{x}) \leq c \quad \text{and} \quad x_i \geq 0 \quad \text{for all} \quad i = 1, \ldots, n. \qquad (\star)$$

We can, of course, solve this using the standard Lagrangian method. Denoting by λ the multiplier for the first constraint, and by μ_i the multiplier for the constraint $-x_i \leq 0$, we have a total of $n+1$ multipliers, and the associated Lagrangian function

$$L(\boldsymbol{x}, \lambda, \boldsymbol{\mu}) \quad = \quad f(\boldsymbol{x}) + \lambda(c - g(\boldsymbol{x})) + \sum_{i=1}^{n} \mu_i x_i$$

Note that for the multipliers μ_i, we know that at the optimal solution (and all local maxima), we will necessarily have for each $i = 1, \ldots, n$:

- $\mu_i \geq 0$, and
- $\mu_i = 0$ or $x_i = 0$.

This simple nature of the multipliers associated with the nonnegativity constraints allows us to work instead with the so-called **Kuhn-Tucker Lagrangian** which does not feature the nonnegativity constraints at all:

$$\mathcal{L}(\boldsymbol{x}, \lambda) \quad = \quad f(\boldsymbol{x}) + \lambda(c - g(\boldsymbol{x}))$$

The first order conditions for $\mathcal{L}$ are different from those of L, because

$$\mathcal{L}_{x_i} \quad = \quad L_{x_i} - \mu_i \quad \leq \quad 0$$

We will first treat the problem via the standard Lagrangian approach. Then using the FOC for the standard Lagrangian, we will derive the equivalent FOC for the KT-Lagrangian.

Remember that

The standard Lagrangian FOC for the problem $(\star)$ are:

(a) For each $i = 1, \ldots, n$:

$$x_i^* > 0 \quad \text{and} \quad \mu_i^* = 0 \qquad \text{or} \qquad x_i^* = 0 \ \text{ and } \ \mu_i^* \geq 0$$

(b)

$$g(\boldsymbol{x}^*) < c \ \text{ and } \ \lambda^* = 0 \qquad \text{or} \qquad g(\boldsymbol{x}^*) = c \ \text{ and } \ \lambda^* \geq 0$$

Now note that

$$x_i^* > 0 \quad \text{and} \quad \mu_i^* = 0 \implies \frac{\partial \mathcal{L}}{\partial x_i}(\boldsymbol{x}^*) = 0$$

$$x_i^* = 0 \ \text{ and } \ \mu_i^* \geq 0 \implies \frac{\partial \mathcal{L}}{\partial x_i}(\boldsymbol{x}^*) \leq 0$$

Therefore

So the KT-Lagrangian FOC for the problem $(\star)$ are:

(a) For each $i = 1, \ldots, n$:

$$x_i^* > 0 \quad \text{and} \quad \tfrac{\partial \mathcal{L}}{\partial x_i}(\boldsymbol{x}^*) = 0 \qquad \text{or} \qquad x_i^* = 0 \ \text{ and } \ \tfrac{\partial \mathcal{L}}{\partial x_i}(\boldsymbol{x}^*) \leq 0$$

(b)

$$g(\boldsymbol{x}^*) < c \ \text{ and } \ \lambda^* = 0 \qquad \text{or} \qquad g(\boldsymbol{x}^*) = c \ \text{ and } \ \lambda^* \geq 0$$

Finally, the envelope theorem continues to hold for the KT-Lagrangian, because $\mathcal{L} = L$ at the optimal solution (*can you see why?*), and we know from before that the envelope theorem holds for the standard Lagrangian L:

$$\frac{dv}{d\alpha} \quad = \quad \left.\frac{\partial L}{\partial \alpha}\right|_{\boldsymbol{x}=\boldsymbol{x}^*,\ \lambda=\lambda^*,\ \boldsymbol{\mu}=\boldsymbol{\mu}^*} \quad = \quad \left.\frac{\partial \mathcal{L}}{\partial \alpha}\right|_{\boldsymbol{x}=\boldsymbol{x}^*,\ \lambda=\lambda^*}$$

An example using the KT-Lagrangian

Consider the choice problem of a consumer with preferences $u(x_1, x_2) = (x_1+1)x_2$ and monetary budget w. He problem can be summarised as

$$\max_{x_1, x_2} (x_1 + 1)x_2 \qquad \text{s.t.} \quad \begin{cases} p_1 x_1 + p_2 x_2 \leq w, \\ x_1 \geq 0, \\ x_2 \geq 0 \end{cases}$$

The associated KT-Lagrangian is

$$\mathcal{L} \quad = \quad (x_1 + 1)x_2 + \lambda(w - p_1 x_1 - p_2 x_2)$$

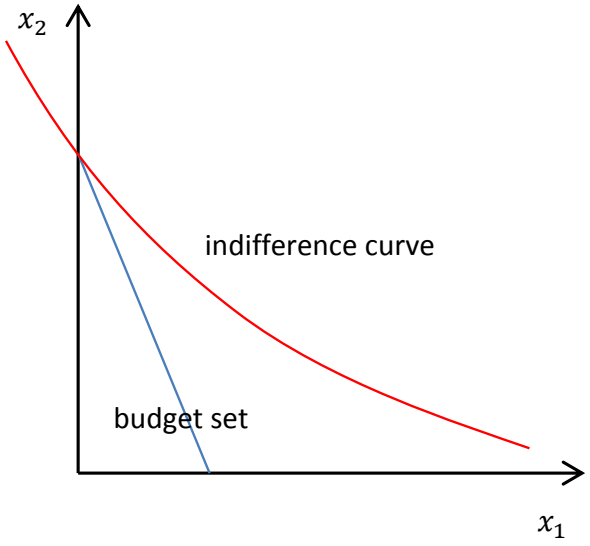


Figure 8: A solution with $x_1 \geq 0$ binding.

and the related first order conditions will be as follows:

Clearly the budget constraint will be met with equality, so $\mathcal{L}_\lambda = w - p_1 x_1 - p_2 x_2 = 0$ and $\lambda^* \geq 0$.

Suppose the constraint $x_1 \geq 0$ binds, i.e., $x_1^* = 0$. Then $x_2^* = \frac{w}{p_2} > 0$. We have to check: $\mathcal{L}_{x_1} = x_2 - \lambda p_1 \leq 0$ and $L_{x_2} = x_1 + 1 - \lambda p_2 = 1 - \lambda p_2 = 0$ (with $\lambda \geq 0$). The last equation gives $\lambda^* = \frac{1}{p_2}$, so we have $x_2^* - \lambda p_1 = \frac{w}{p_2} - \frac{1}{p_2} p_1$. This is ≤ 0 when $p_1 \geq w$.

This gives $(0, \frac{w}{p_2})$ as a solution when $p_1 \geq w$.

Suppose the constraint $x_2 \geq 0$ binds, i.e., $x_2^* = 0$. But this leads to a utility of 0, and therefore cannot be a solution.

Suppose neither nonnegativity constraint binds, i.e., Case $x_1^*, x_2^* > 0$. Then we require $\mathcal{L}_{x_1} = \mathcal{L}_{x_2} = 0$. So $x_2 - \lambda p_1 = x_1 + 1 - \lambda p_2 = 0$. This gives $\lambda = \frac{x_2}{p_1} = \frac{x_1+1}{p_2}$. So the budget constraint gives $p_1 x_1 + p_1 (x_1 + 1) = w$, so $2x_1 p_1 = w - p_1$. So $x_1 = \frac{w-p_1}{2p_1}$ and $x_2 = \frac{1}{p_2}\left[w - \frac{w-p_1}{2}\right] = \frac{w+p_1}{2p_2}$. The condition for $x_1, x_2 > 0$ is $w > p_1$.

This gives $\left(\frac{w-p_1}{2p_1}, \frac{w+p_1}{2p_2}\right)$ as a solution when $w > p_1$.

So we have solution: $(x_1^*, x_2^*) = \left(0, \frac{w}{p_2}\right)$ when $p_1 \geq w$, and $\left(\frac{w-p_1}{2p_1}, \frac{w+p_1}{2p_2}\right)$ when $w > p_1$.

Another example using the KT-Lagrangian

A firm has total revenue $R = 10Q - Q^2 + A/2$ where Q is its output and A is its advertising expenditure. Its total costs are $C = Q^2/2 + 5Q + 1 + A$. The managers of the firm wish to choose Q and A to maximise total revenue subject to a minimum profit constraint, $\pi = R - C \geq \pi_0$, and $A \geq 0$.

1. Comment on the role of advertising. Why will the $\pi \geq \pi_0$ constraint bind?

2. Find the Q and A satisfying the Lagrangian FONCs for optimality when $\pi_0 = 3$. Have these necessary conditions found the true optimum?

3. Find $\frac{\partial R^*}{\partial \pi_0}$ at $\pi_0 = 3$.

Solution.

1. The profit constraint is $\pi = R - C \geq \pi_0$, i.e.:

$$\begin{aligned} 10Q - Q^2 + \frac{A}{2} - \left(\frac{Q^2}{2} + 5Q + 1 + A\right) &\geq \pi_0 \\ \text{or} \quad 5Q - \frac{3Q^2}{2} - \frac{A}{2} - 1 &\geq \pi_0 \end{aligned}$$

   Note that this constraint must bind in the optimal solution, because if it didn't, we could increase A which would then increase R.

2. The KT-Lagrangian does not feature the nonnegativity constraints $A \geq 0$ and $Q \geq 0$:

$$\mathcal{L}(Q, A, \lambda) = 10Q - Q^2 + \frac{A}{2} + \lambda\left(-\pi_0 - \frac{3Q^2}{2} + 5Q - \frac{A}{2} - 1\right)$$

   KT first order conditions for a maximum are:

   - $\mathcal{L}_Q = 10 - 2Q - 3\lambda Q + 5\lambda = 0$
   - Either $A^* = 0$ and $\mathcal{L}_A = \frac{1}{2} - \frac{\lambda}{2} \leq 0$, or $A^* > 0$ and $\frac{1}{2} - \frac{\lambda^*}{2} = 0$
     That is: either $A^* = 0$ and $\lambda^* \geq 1$, or $A^* > 0$ and $\lambda^* = 1$
   - $\mathcal{L}_\lambda = 0$, i.e., $5Q - \frac{3Q^2}{2} - \frac{A}{2} - 1 = \pi_0$ at $(A, Q) = (A^*, Q^*)$.

   If $A^* > 0$ is part of a solution, then $\lambda^* = 1$, so $\mathcal{L}_Q = 15 - 5Q = 0$, so $Q^* = 3$. So $\pi_0 = 3 = 5Q - \frac{3Q^2}{2} - \frac{A}{2} - 1 = \frac{1-A}{2}$, so $A = -5$. Hence there **cannot** be a solution with $A^* > 0$.

   If there is solution with $A^* = 0$, then $\pi_0 = 3 = 5Q - \frac{3Q^2}{2} - 1$, so $4 = 5Q - \frac{3Q^2}{2}$,which solves to give $Q = 2$ or $Q = \frac{4}{3}$.

   Then $L_Q = 0$ gives $\lambda = \frac{10-2Q}{3Q-5}$, which for $Q = 2$ gives $\lambda = 6$, and for $Q = \frac{4}{3}$ gives $\lambda = -\frac{22}{3}$.

   Only the first has $\lambda \geq 0$ as required, so must have $Q = 2$.

   Then $\mathcal{L}_A = \frac{1}{2} - \frac{\lambda}{2} = -\frac{5}{2} \leq 0$ as required.

   So the FOC is solved uniquely by $A^* = 0$, $Q^* = 2$ and $\lambda^* = 6$.

   This solution $(A, Q, \lambda) = (0, 2, 6)$ of the FOC must be the global maximum since the objective function is concave (verify via the Hessian) and the constraint set is a convex set (verify by drawing the graph of $5Q - \frac{3Q^2}{2} - \frac{A}{2} - 1 = \pi_0$ in the Q-A coordinate plane).

3. Finally, the envelope theorem yields:

$$\frac{\partial R}{\partial \pi_0} = \left.\frac{\partial \mathcal{L}}{\partial \pi_0}\right|_{(A^*, Q^*, \lambda^*)} = -\lambda^* = -6$$

$\diamond$