MATHEMATICAL ECONOMICS

JOSÉ PEDRO GAIVÃO

Contents

1. Fundamental concepts 2
1.1. Set theory 2
1.2. Functions 3
2. Metric Spaces 4
2.1. Open and closed sets 6
2.2. Compact sets 7
2.3. Limits 8
2.4. Continuous functions 9
3. Fixed point theorems 10
3.1. Banach fixed point 10
3.2. Brouwer fixed point 13
4. Hyperplane separation theorem 16
5. Correspondences 17
5.1. Continuity of correspondences 18
5.2. Kakutani fixed point 20
6. Optimization 22
6.1. Some terminology 23
6.2. Unconstrained optimization 23
6.3. Convex and concave functions 27
6.4. Equality constraints 28
6.5. Inequality constraints 32
6.6. Mixed constrains 34
7. Scalar Differential equations 37
7.1. Initial value problem 38
7.2. First order linear ODEs 38
7.3. Separation of variables 40
7.4. Qualitative theory of scalar ODEs 43
8. Planar Differential Equations 45
8.1. Change of variables 46
8.2. Eigenvalues and eigenvectors 47
8.3. Matrix P 47
8.4. Jordan normal forms 50
8.5. Solution of IVP in Jordan normal form 51
8.6. Solution of IVP 52

Date: December 7, 2022.

8.7. 2nd order scalar linear ODEs 54
8.8. Solution of non-homogeneous IVP 55
8.9. Phase portraits 56
8.10. Transformation of the phase portrait 58
8.11. Trace-determinant plane 60
9. Calculus of Variations 62
9.1. Basic problem of calculus of variations 62
9.2. Necessary conditions 63
9.3. Sufficient condition 65
9.4. Variable terminal conditions 66
10. Optimal Control 68
10.1. Basic optimal control problem 68
10.2. Variable terminal conditions 71
10.3. Scrap value formulation 73
10.4. Infinite horizon 74

1. Fundamental concepts

1.1. **Set theory.** Sets are denoted by capital letters A, B, etc. Examples of sets are the natural numbers $\mathbb{N}$, the integers $\mathbb{Z}$, the real numbers $\mathbb{R}$ or the complex numbers $\mathbb{C}$. Another example is the set of students in Mathematical Economics. A set which is very common is the **empty set** (the set containing no elements) which we denote by $\emptyset$. Objects or **elements** of the sets are denoted by lowercase letters a, b, etc. Of course, there are exceptions to these rules of thumb. If an element a belongs to a set A we write

$$a \in A$$

Otherwise

$$a \notin A$$

We say that A is a **subset** of B if $a \in A$, then $a \in B$. In that case we write

$$A \subset B$$

Of course, if $A \subset B$ and $B \subset A$ then $A = B$.

A set can be **specified** by enumerating its elements

$$A = \{a, b, c, d\}$$

or by sharing a common property, i.e.,

$$B = \{x \colon x \text{ satisfies property P}\}.$$

This reads, the set of elements x such that x satisfies property P. For instance,

$$E = \{x \colon x \text{ is an even integer}\}$$

Alternatively,

$$E = \{x \in \mathbb{Z} \colon x \text{ is even }\}$$

1.1.1. *Constructing new sets from given sets.* The **union** of two sets A and B is the set formed by collecting all elements of A together with the elements of B, i.e.,

$$A \cup B = \{x \colon x \in A \text{ or } x \in B\}$$

Recall that "or" is mathematics is different from "or" in english. The **intersection** of A and B is the set of elements that are both in A and in B, i.e.,

$$A \cap B = \{x \colon x \in A \text{ and } x \in B\}$$

Of course $\cup$ and $\cap$ are **commutative** binary operations, i.e., $A \cup B = B \cup A$ and $A \cap B = B \cap A$. Moreover,$A \cup \emptyset = A$ and $A \cap \emptyset = \emptyset$. When $A \cap B = \emptyset$ we say that A and B are **disjoint**. We also denote by $A \setminus B$ the set of elements that are in A but not in B, i.e.

$$A \setminus B = \{x \in A \colon x \notin B\}$$

Exercise 1 (DeMorgan's laws)**.** Consider the sets A, B and C. Show that

(1) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
(2) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
(3) $A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$

Given a sequence of sets A_1, A_2, and so on, the union of all A_n, $n \in \mathbb{N}$ is

$$\bigcup_{n\in\mathbb{N}} A_n = \{x \colon x \in A_n \text{ for at least one } n \in \mathbb{N}\}$$

Similarly,

$$\bigcap_{n\in\mathbb{N}} A_n = \{x \colon x \in A_n \text{ for every } n \in \mathbb{N}\}$$

The **Cartesian product** of two sets A and B is denoted by $A \times B$ and is defined to be the set of ordered pairs (a, b) where $a \in A$ and $b \in B$[1], i.e.,

$$A \times B = \{(a, b) \colon a \in A \text{ and } b \in B\}$$

As an example,

$$\mathbb{R} \times \mathbb{R} = \{(x, y) \colon x \in \mathbb{R} \text{ and } y \in \mathbb{R}\}$$

Of course, this definition extends to the Cartesian product of any finite number of sets. In particular, the Cartesian product of n copies of a set A is denoted by A^n. For instance, $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$.

Exercise 2. Sketch in a paper the following sets:

(1) $A = \{x \in \mathbb{R} \colon x^2 > 1\}$
(2) $B = \{x \in \mathbb{R} \colon x^3 > 1 \text{ and } x^4 < 16\}$
(3) $C = \{x \colon \mathbb{Z} \colon x^2 < 2 \text{ and } x \text{ is even}\}$
(4) $B \cup C$
(5) $A \times B$
(6) $D = \{(x, y) \in \mathbb{R}^2 \colon x^2 + y^2 \leq 4\}$
(7) $E = \{(x, y) \in \mathbb{R}^2 \colon x^2 \leq y \text{ and } 1 \geq y + x^2\}$

1.2. **Functions.** A **function** f is a rule of assignment together with two sets A and B such that to each element of A associates a unique element in B. The set A is commonly known as **domain** and B the **range**. This notion is usually written as

$$f \colon A \to B$$

It is also common to say that f is a mapping from A to B. Given an element $a \in A$, its corresponding element in B is denoted by $f(a)$, also called the **value** (or image) of f at a. The set of all values is called the **image set**,

$$\text{image}(f) = \{f(a) \colon a \in A\}$$

It is common to define a function by specifying the domain and values that can take. For instance, let $f \colon \mathbb{R} \to \mathbb{R}$ be the function such that

$$f(x) = x^2 + 1$$

[1] An ordered pair (a, b) can be represented in set theory by $\{\{a\}, \{a, b\}\}$.

Another example of a function is the **identity function** $\mathrm{id}\colon A \to A$ where $\mathrm{id}(a) = a$ for every $a \in A$. The **graph** of a function $f : A \to B$ is the subset of the Cartesian product $A \times B$ defined by

$$\operatorname{graph}(f) = \{(a, b) \in A \times B \colon b = f(a)\} = \{(a, f(a) \colon a \in A\}$$

Given two functions $f\colon A \to B$ and $g\colon C \to A$ we define the **composition function** $f \circ g : C \to B$ as $f \circ g(x) = f(g(x))$. Notice that the domain of f has to be equal to the range of g for the composition to make sense.

A function $f\colon A \to B$ is said to be **injective** if $a \neq b$ implies that $f(a) \neq f(b)$. Or equivalently, if $f(a) = f(b)$ then, $a = b$. It is **surjective** if for every $b \in B$ there is $a \in A$ such that $b = f(a)$. If f is both injective and surjective, then it is called **bijective** (or a **one-to-one correspondence**). If f is bijective, then there is a function $f^{-1}\colon B \to A$ called the **inverse** of f. The inverse is defined by the property that $f \circ f^{-1} = f^{-1} \circ f = \mathrm{id}$. Two sets A and B have the same **cardinality** if there is a bijective function mapping A to B. A set A is **finite** if there is a bijection mapping A to $\{1, \ldots, n\}$ for some $n \in \mathbb{N}$. In this case n is the cardinality of A. A set is called **infinite** if it is not finite. We say that A is **countable** if there is a bijection mapping A to $\mathbb{N}$. Otherwise, it is called **uncountable**. For instance, the set of all integers $\mathbb{Z}$ and the set of all rationals $\mathbb{Q}$ is countable. The set of all real numbers $\mathbb{R}$ is uncountable.

Exercise 3. Sketch the graph of the following real valued functions $f : D \to \mathbb{R}$, determine if are injective/surjective and compute its inverse if exists:

(1) $f(x) = x^3 + 1$ with $x \in \mathbb{R}$
(2) $f(x) = x^2 - x$ with $x \in \mathbb{R}$
(3) $f(x) = \sqrt{x + 1}$ with $x > 0$.
(4) $f(x) = \frac{x-1}{x+1}$ with $x > -1$
(5) $f(x) = 2e^{-x}$ with $x \in \mathbb{R}$
(6) $f(x) = \log(x^2 + 1)$ with $x > 0$

Exercise 4.

(1) Find a bijective function mapping $\mathbb{N}$ to $\mathbb{Z}$.
(2) Find a bijective function mapping $\mathbb{R}$ to $]-1, 1[$.

2. Metric Spaces

Let X be a set. A **metric**[2] in X is a function $d : X \times X \to \mathbb{R}$ such that for every $x, y, z \in X$ it satisfies

(1) $d(x, y) \geq 0$
(2) $d(x, y) = 0$ iff[3] $x = y$

[2] Also called distance

[3] if and only if

(3) $d(x, y) = d(y, x)$
(4) $d(x, z) \leq d(x, y) + d(y, z)$

The pair (X, d) is called a **metric space**. The 4th property is known as the **triangle inequality**. It is common to say that X endowed with the metric d is a metric space. The elements of X are called **points**. As an example, let $X = \mathbb{R}^n$ with $n \in \mathbb{N}$ and d be the usual **Euclidean distance**

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$

It is simple to see that d as defined above satisfies the 4 axioms of a metric. Hence, $(\mathbb{R}^n, d)$ is a metric space, also known as **Euclidean space**. As another example, let $X = \mathbb{Z}^2$ with the **Manhattan distance**

$$d_1(x, y) = |x_1 - y_1| + |x_2 - y_2|.$$

It is easy to see that d_1 as defined above is also a metric. Thus $(\mathbb{Z}^2, d_1)$ is a metric space.

Finally, to give a more abstract example, let A be a set and consider the collection X of all bounded[4] functions $f\colon A \to \mathbb{R}$. We endow X with the metric[5]

$$\rho(f, g) = \sup_{a \in A} |f(a) - g(a)|$$

Then (X, ρ) is a metric space. The points of X are functions and ρ measures the distance of any two functions in X.

Informally, a metric space is a set of points for which we can measure distances.

In order to reduce the level of abstraction, throughout the rest of this section we will stick with the Euclidean space $(\mathbb{R}^n, d)$, although the results we will prove are also valid in abstract metric spaces.

Exercise 5.

(1) Show that the following functions $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ satisfy the four axioms of a metric:
 (a) $d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$
 (b) $d(x, y) = |x_1 - y_1| + |x_2 - y_2|$
 (c) $d(x, y) = \max\{|x_1 - y_1|, |x_2 - y_2|\}$
(2) For each of the metrics above sketch the set

$$\{x \in \mathbb{R}^2 \colon d(x, 0) = 1\}.$$

[4] f is bounded if there is $M \geq 0$ such that $|f(a)| \leq M$ for every $a \in A$

[5] The supremum of a set $A \subset \mathbb{R}$, which we denote by $\sup A$, is the least element in $\mathbb{R}$ that is greater than or equal to all elements of A. As an example, $\sup]-4, 5[= 5$. For some sets the supremum might not exist, e.g., $[-1, +\infty[$ has no supremum. In this case it is common to formally write $\sup\, [-1, +\infty[= +\infty$. If the supremum of a set A belongs to A, then we say that A has a maximum and write $\max A$. Similar considerations hold for the infimum and minimum.

2.1. **Open and closed sets.** The **open ball** of radius $r > 0$ centred at $a \in \mathbb{R}^n$ is the set

$$B(a, r) = \{x \in \mathbb{R}^n \colon d(x, a) < r\}$$

Let $A \subset \mathbb{R}^n$ be a subset. We say that A is **open** if for every $a \in A$ there is $r > 0$ such that $B(a, r) \subset A$. A point $a \in A$ is called **interior to** A is there is $r > 0$ such that $B(a, r) \subset A$. The set of interior point to A is denoted by $\operatorname{int}(A)$. Clearly, A is open iff all its points are interior to A, i.e., $\operatorname{int}(A) = A$. Obviously, $\mathbb{R}^n$ is open, and by convention the empty set $\emptyset$ is open. The collection of all open sets is called a **topology**.

We say that A is **closed** if its complement $\mathbb{R}^n \setminus A$ is open.

As an example, in $\mathbb{R}$, the intervals

$$]-1, 4[\quad \text{and} \quad]-\infty, 0[$$

are open[6] and the sets

$$\{0\}, \quad [-1, 4] \quad \text{and} \quad [1, +\infty[$$

are closed. Notice that $\mathbb{R} =]-\infty, +\infty[$ is both open and closed. Of course, there are sets which are neither open nor closed, e.g., the interval $[1, 3[$.

For more examples of open and closed sets, consider in $\mathbb{R}^2$, the open ball

$$B(0, 1) = \{(x, y) \in \mathbb{R}^2 \colon x^2 + y^2 < 1\}$$

and the closed square $[0, 1]^2 = [0, 1] \times [0, 1]$. There are plenty of other examples as the following result shows.

Lemma 2.1. *The following holds:*

1. *Every open ball is an open set.*
2. *An arbitrary union of open sets is open.*
3. *The intersection of a finite number of open sets is open.*

Proof. Exercise. □

Notice that a countable intersection of open sets might not be open as the following exercise shows.

Exercise 6. Consider the open intervals $A_n =]-1/n, 1/n[$ with $n \in \mathbb{N}$. Show that

$$\bigcap_{n \in \mathbb{N}} A_n = \{0\}.$$

A **neighbourhood** of a point $a \in \mathbb{R}^n$ is any set V which contains an open ball $B(a, r)$ for some $r > 0$. Of course, any open ball $B(a, r)$ is a neighbourhood of a.

[6] An equivalent notation for writing open intervals is $(-1, 4)$.

Given a subset $A \subset \mathbb{R}^n$, a point $x \in \mathbb{R}^n$ is called an **accumulation point**[7] of A if any neighbourhood of x contains at least one point in A different from x itself, i.e.,

$$\forall\, r > 0, \quad (B(x,r) \setminus \{x\}) \cap A \neq \emptyset.$$

As an example, 0 is an accumulation point of $]0,1]$. Also, it is an accumulation point of the set $\{1/n \colon n \in \mathbb{N}\}$.

The set of all accumulation points of A is denoted by A', also known as **derivative** of A.

The union $A \cup A'$, which we denote by $\overline{A}$, is called the **closure** of the set A. The following result gives an equivalent characterization of closed set. It basically says that a closed set contains all its accumulation points.

Proposition 2.2. *$\overline{A} = A$ if and only if A is closed.*

Proof. Suppose that A is closed. In order to show that $\overline{A} = A$ we have to show that $A' \subset A$. So let $x \in A'$. If $x \notin A$, because $\mathbb{R}^n \setminus A$ is open, there is $r > 0$ such that $B(x,r) \subset \mathbb{R}^n \setminus A$. But then x cannot be an accumulation point of A. This shows that x has to belong to A.

Now suppose that $\overline{A} = A$. Then A has to be closed because for any $a \in \mathbb{R}^n \setminus A$ there is $r > 0$ such that $B(a,r) \subset \mathbb{R}^n \setminus A$. Otherwise, a would belong to A' which cannot be. □

2.2. **Compact sets.** Given a subset $A \subset \mathbb{R}^n$ we define its **diameter**

$$\operatorname{diam}(A) = \sup\{d(x,y) \colon x, y \in A\}.$$

For instance, the open ball $B(a,r)$ has diameter $2r$. However, $\operatorname{diam}(\mathbb{R}) = \infty$.

We say that A is **bounded** if it has finite diameter.

A bounded and closed set is called **compact**. For instance, any **closed ball**

$$\overline{B}(a,r) = \{x \in \mathbb{R}^n \colon d(x,a) \leq r\}, \quad a \in \mathbb{R}^n,\ r > 0$$

is compact. In $\mathbb{R}$, a closed ball is a closed interval $[a,b]$, $a < b$.

The following theorem gives an equivalent characterization of compactness. An **open cover** of a set $A \subset \mathbb{R}^n$ is a collection of open sets whose union contains A. A **finite subcover** is a finite collection of sets that form a cover.

Theorem 2.3 (Heine - Borel). *A is compact if and only if for every open cover of A we can extract a finite subcover.*

Proof. Check bibliography. □

Exercise 7. Sketch the following sets and decide which are open, closed, bounded and compact.

(1) $A = \{(x,y) \in \mathbb{R}^2 \colon x \geq 0\}$

[7] Also commonly called limit point.

(2) $B = \{(x, y) \in \mathbb{R}^2 \colon 1 \le x^2 + y^2 \le 4\}$
(3) $A \cap B$
(4) $C = \{(x, y) \in \mathbb{R}^2 \colon x^2 + 2x > y\}$
(5) $C \setminus A$
(6) $D = \{(x, y, z) \in \mathbb{R}^3 \colon z > x^2 + y^2\}$
(7) $E = \{(x, y, z) \in \mathbb{R}^3 \colon -1 \le z \le 1\}$
(8) $F = \{(x, y, z) \in \mathbb{R}^3 \colon z^2 \ge x^2 + y^2 \quad \text{and} \quad z \le 0\}$
(9) $D \cap E$
(10) $(\overline{D} \cup F) \cap E$

Exercise 8. Show that:

(1) Any finite union of compact sets is compact.
(2) $A \subset \mathbb{R}^n$ is bounded if and only if $\overline{A}$ is compact.

Exercise 9. In this exercise you will study the Cantor[8] set C. Let $A_0 = [0, 1]$ and define A_1 by cutting A_0 in three equal parts and then removing the middle part from A_0. i.e., $A_1 = [0, 1/3] \cup [2/3, 1]$. The set A_1 is a union of two disjoint intervals. Now for each interval of A_1 we proceed as before. Cut in three equals parts and remove from each its middle part. Call this new set A_2. Continue this process indefinitely to obtain a sequence of sets A_n, $n \ge 0$. The Cantor set is the intersection of all these sets, i.e.,

$$C = \bigcap_{n \ge 0} A_n$$

(1) Obtain the analytic expression for A_2.
(2) Show that each set A_n is a union of 2^n disjoint closed intervals.
(3) Is A_n closed? Why?
(4) Prove that C is compact.

2.3. **Limits.** A **sequence** $\{x_n\}_{n \in \mathbb{N}}$ in a set A is an ordered list of points in A indexed by $\mathbb{N}$. Given a sequence $\{x_n\}_{n \in \mathbb{N}}$ in $\mathbb{R}^n$, a point $x \in \mathbb{R}^n$ is called the **limit** of the sequence $\{x_n\}_{n \in \mathbb{N}}$ if for every neighbourhood V of x there is $N \in \mathbb{N}$ such that x_n belongs to V for every $n \ge N$. A sequence that has a limit is called **convergent**. We also denote its limit by $\lim_{n \to \infty} x_n$. Is is easy to see that the limit of a converging sequence is unique.

Exercise 10. Show that the limit of a converging sequence is unique, i.e., a converging sequence cannot more than one limit.

Using the limits of sequences we obtain another characterization of closed sets.

Proposition 2.4. *A set $A \subset \mathbb{R}^n$ is closed if and only if it contains the limits of converging sequences in A.*

Proof. Exercise. □

[8] Check wikipedia.

A sequence $\{x_n\}_{n\in\mathbb{N}}$ in $\mathbb{R}^n$ is called a **Cauchy sequence** if for any $r > 0$ there is $N \in \mathbb{N}$ such that

$$d(x_n, x_m) < r, \quad \forall\, n, m \geq N.$$

It is easy to see that any convergent sequence is Cauchy. The other direction is the content of the following theorem.

Theorem 2.5. *A sequence in $\mathbb{R}^n$ is convergent if and only if it is a Cauchy sequence.*

The above theorem is not true for arbitrary metric spaces, i.e., there are metric spaces for which some Cauchy sequence do not have a limit. For instance, the metric space $(]0,1[, \ell)$ with the metric $\ell(x,y) = |y-x|$ has a Cauchy sequence $x_n = 1/n$, $n \in \mathbb{N}$ that has no limit in the space $]0,1[$.

The metric spaces for which the theorem holds are called **complete**. In this sense, $\mathbb{R}^n$ is complete, i.e., it contains the limits of all Cauchy sequences. Complete metric spaces have no "holes"!

2.4. **Continuous functions.** Consider a subset $A \subset \mathbb{R}^n$ and a function $f\colon A \to \mathbb{R}^m$. We say that f **is continuous at** $x \in A$ if for every convergent sequence $\{x_n\}_{n\in\mathbb{N}}$ with limit $x \in A$, the value sequence $\{f(x_n)\}_{n\in\mathbb{N}}$ is convergent with limit $f(x)$, i.e.,

$$x = \lim_{n\to\infty} x_n \quad \Longrightarrow \quad f(x) = \lim_{n\to\infty} f(x_n).$$

A **continuous function** is any function that is continuous at each point of its domain. Informally speaking, a continuous function maps convergent sequences into convergent sequences. Well-known examples of continuous functions $f : \mathbb{R} \to \mathbb{R}$ are the polynomial functions and trigonometric functions. Another example of a continuous function is $f :]0, +\infty[\to \mathbb{R}$ with $f(x) = 1/x$. This function is continuous, even though it becomes unbounded for points close to 0 (which does not belong to the domain).

The following is a classical result. Given a set $A \subset \mathbb{R}^n$ we denote by $f(A)$ its **image set**, i.e., the set of all values of f. Formally,

$$f(A) = \{f(a)\colon a \in A\}.$$

Proposition 2.6. *If A is compact and $f : A \to \mathbb{R}^n$ is continuous, then the image set $f(A)$ is also compact.*

Proof. Check bibliography. □

As a corollary of this proposition we obtain the Weierstrass theorem. We say that a function $f : A \to \mathbb{R}$ has a **maximum** if there is $a \in A$ such that $f(x) \leq f(a)$ for every $x \in A$. In a similar way we define the minimum.

Theorem 2.7 (Weierstrass)**.** *Every continuous function $f : A \to \mathbb{R}$ defined on a compact domain $A \subset \mathbb{R}^n$ has a maximum and a minimum.*

Proof. By the previous proposition, $f(A) \subset \mathbb{R}$ is compact. Since $f(A)$ is compact in $\mathbb{R}$ is has a maximum value $M \in f(A)$ and minimum value $m \in f(A)$. Take $x^* \in f^{-1}(M)$ and $x_* \in f^{-1}(m)$. Then

$$\forall\, x \in A \quad f(x_*) \leq f(x) \leq f(x^*).$$

□

An an application of this theorem, consider the problem of finding the maximum of a continuous function $f : A \to \mathbb{R}$ over a compact domain A. The domain may be given by inequalities like in Exercise 7. The Weierstrass theorem says that the maximization problem has a solution. However, it does not give a way to compute it explicitly. Later in the course we will learn a method that allows to solve certain maximization problems explicitly.

Exercise 11. Show, using the definition of continuity, that given two continuous functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, the composition function $f \circ g$ is also continuous.

Exercise 12. For each of the following functions $f \colon D \to \mathbb{R}$ determine its continuity points and, using the Weierstrass theorem, decide if the function has a maximum or minimum.

1. $f(x) = x^2$ where $D = \{x \in \mathbb{R} \colon |x| \leq 1\}$
2. $f(x) = x^3 - x^2 + x - 1$ where $D = [-2, -1] \cup [1, 2]$
3. $f(x) = x \cos^2(1/x)$ where $D = \left\{ \frac{(-1)^n}{2\pi n} \colon n \in \mathbb{N} \right\} \cup \{0\}$
4. $f(x, y) = xy$ where $D = [-1, 1]^2$
5. $f(x, y) = x \log(y)$ where $D =]0, 1]^2$
6. $f(x, y) = e^{-x^2 - y^2}$ where $D = \mathbb{R}^2$

3. Fixed point theorems

Let $A \subset \mathbb{R}^n$ and $f : A \to \mathbb{R}^n$ be a function. We say that $x \in A$ is a **fixed point** of f if

$$f(x) = x.$$

Finding fixed points of functions is of great importance for Economics. We will give sufficient conditions for the existence of fixed points and then give an application to General Equilibrium Theory.

3.1. **Banach fixed point.**

A function $f : A \to \mathbb{R}^n$ is called a **Lipschitz contraction** if there is $\lambda \in]0, 1[$ such that

$$d(f(x), f(y)) \leq \lambda\, d(x, y), \quad \forall\, x, y \in A.$$

Informally speaking, a Lipschitz contraction contracts distances between points.

Theorem 3.1 (Banach)**.** *If $f : A \to \mathbb{R}^n$ is a Lipschitz contraction, $f(A) \subset A$ and A is closed, then f has a unique fixed point $\bar{x} \in A$.*

Proof. Let $x_0 \in A$. Consider the following sequence of points in A,

$$x_{n+1} = f(x_n), \quad n = 1, 2, \dots$$

Since f is a Lipschitz contraction

$$d(x_{n+1}, x_n) = d(f(x_n), f(x_{n-1})) \leq \lambda\, d(x_n, x_{n-1}).$$

Iterating the previous inequality we get

$$d(x_{n+1}, x_n) \leq \lambda^{n-1}\, d(x_2, x_1), \quad n \in \mathbb{N}$$

This shows that $\lim_{n\to\infty} d(x_{n+1}, x_n) = 0$. In fact, with a little effort we can show that $\{x_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence (try to prove it!). Thus, $\{x_n\}_{n\in\mathbb{N}}$ is a convergent sequence (by Theorem 2.5). This means that $\{x_n\}_{n\in\mathbb{N}}$ has a limit point $\bar{x} \in \mathbb{R}^n$. However, A is closed, so it contains all its accumulation points. Therefore, $\bar{x} \in A$. This shows the existence part of the theorem. To show uniqueness, let $\bar{x}_1$ and $\bar{x}_2$ be two fixed points inside A. Then

$$d(\bar{x}_1, \bar{x}_2) = d(f(\bar{x}_1), f(\bar{x}_2)) \leq \lambda\, d(\bar{x}_1, \bar{x}_2)$$

which can only be true if $\bar{x}_1 = \bar{x}_2$. □

Example 3.2. As an example, let $f : \overline{B}(0,1) \to \mathbb{R}^n$ defined by $f(x) = \lambda\, x$ where $\lambda \in]0,1[$ and $\overline{B}(0,1)$ is the closed ball of radius 1 centred at the origin. In other words, f takes any point in the closed ball and contracts it towards the origin by a factor λ. Clearly, f is a Lipschitz contraction because $d(f(x), f(y)) = d(\lambda x, \lambda y) = \lambda d(x,y)$. Since $f(\overline{B}(0,1)) \subset \overline{B}(0,1)$ and $\overline{B}(0,1)$ is closed, the Banach fixed point theorem says that f has a unique fixed point inside $\overline{B}(0,1)$. In fact, that fixed point is the origin because, $f(0) = 0$.

Example 3.3. Consider the function $f : [-1,1] \to [-1,1]$ defined by $f(x) = \sin(\lambda\, x)$ where $\lambda \in]0,1[$. To see that f is a Lipschitz contraction we compute

$$\begin{aligned} |f(y) - f(x)| &= \left| \int_x^y f'(z)\, dz \right| \\ &= \left| \int_x^y \lambda \cos(\lambda\, x)\, dz \right| \\ &\leq \int_x^y \lambda\, dz \\ &\leq \lambda |y - x| \end{aligned}$$

So, by the Banach fixed point theorem, we conclude that 0 is the only fixed point of f in the interval $[-1,1]$.

Example 3.4 (Zeros of functions)**.** The Banach fixed point theorem can be used to show the existence of zeros of functions $g : \mathbb{R} \to \mathbb{R}$. Indeed, let

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

If f is a Lipschitz contraction on a closed set A and $f(A) \subset A$, then it has a unique fixed point $\bar{x}$ in A, i.e., $\bar{x} = \bar{x} + g(\bar{x})$. Then $\bar{x}$ is the unique zero of g in the set A.

Example 3.5 (Price Adjustment equation)**.** Consider a market with single commodity. Let $D(p)$ denote the demand function at price p and $S(p)$ the supply function at price p. A simple model for price adjustment is the following equation

$$p_{t+1} - p_t = k(D(p_t) - S(p_t))$$

where $k > 0$ is some coefficient. We see that an increase of the price is followed by an excess demand and vice-versa. The question is whether the price converges to an equilibrium? An equilibrium price p^* has zero excess demand, i.e., has to satisfy the fixed point equation

$$p^* = f(p^*) \quad \text{where} \quad f(p) = p + k(D(p) - S(p)).$$

To give an answer to the problem, we simplify the demand and supply function and consider:

$$D(p) = a - dp \quad \text{and} \quad S(p) = -b + sp.$$

where a, b, d, s are positive numbers. Then, the excess demand is

$$D(p) - S(p) = a + b - p(d + s)$$

So, the price at equilibrium is given by

$$p^* = \frac{a + b}{d + s}.$$

For $d + s < 2/k$, the function f is a Lipschitz contraction. Hence, by the Banach fixed point theorem, p_t converges as $t \to +\infty$ to the price at equilibrium p^*.

Exercise 13. Sketch the graph of f in Example 3.3 and interpret the fixed point geometrically (also draw the bisectrix, i.e., $y = x$).

Exercise 14. Show that $f :]0, 1/4[\to]0, 1/4[$ defined by $f(x) = x^2$ is a Lipschitz contraction. Can you conclude that f has a unique fixed point?

Exercise 15. Determine whether the following functions $f : D \to \mathbb{R}^n$ are Lipschitz contractions. For each function determine its fixed points.

(1) $f(x) = \frac{1}{4}x(1 - x^2)$ with $D = [-1, 1]$
(2) $f(x) = \arctan(x/2)$ with $D = \mathbb{R}$
(3) $f(x) = \frac{1}{4}\sin(x^3)$ with $D = [-1, 1]$
(4) $f(x) = \sqrt{1 + x}$ with $D = [0, +\infty[$
(5) $f(x, y) = \left(\frac{x}{2} + 5, \frac{y}{3} - 1\right)$ with $D = \mathbb{R}^2$

Exercise 16. Use the Banach fixed point theorem to show that the sequence $x_{n+1} = \sqrt{1+x_n}$, $n \in \mathbb{N}$ and $x_0 = 0$ converges to the golden ratio

$$\frac{1+\sqrt{5}}{2} = \sqrt{1+\sqrt{1+\sqrt{1+\cdots}}}.$$

3.2. **Brouwer fixed point.** A set $K \subset \mathbb{R}^n$ is **convex** if any two points in K can be joining by a straight segment contained in K, i.e.,

$$\forall\, x, y \in K, \quad \forall\, \lambda \in [0,1], \quad \lambda x + (1-\lambda)y \in K.$$

For instance, any open (or closed) ball is convex. The following theorem is due to Luitzen Brouwer whose proof we omit.

Theorem 3.6. *If $K \subset \mathbb{R}^n$ is compact and convex, and $f : K \to K$ is continuous, then f has a fixed point in K.*

The one-dimensional version of the theorem is easy to prove and understand. It says that any continuous function $f : [a,b] \to [a,b]$ has a fixed point. In other words, the graph of f has to cross the diagonal of the square $[a,b] \times [a,b]$. Do a picture to convince yourself.

A corollary of Brouwer fixed point theorem is the following.

Corollary 3.7. *Any continuous function $f : \overline{B}(a,r) \to \overline{B}(a,r)$ has a fixed point.*

Notice that the convexity assumption in the theorem is necessary. Indeed, consider the function $f : D \to D$ defined on the annulus $D = \{(x,y) \in \mathbb{R}^2 \colon 1 \leq x^2 + y^2 \leq 4\}$ given by $f(x,y) = (-y,x)$. Geometrically, f rotates anti-clockwise any point (x,y) by 90 degrees. Clearly, f is continuous, D is compact, but f has no fixed point. Indeed, $f(x,y) = (x,y)$ if and only if $x = y = 0$. However, $(0,0)$ is not in D. This is not in contradiction with the Brouwer fixed point because D is not convex, so we cannot apply the theorem to f.

Example 3.8 (Application in a Pure Exchange Economy)**.** Consider an economy with m consumers and n commodities. No production is possible and consumers engage in exchanging commodities to maximize a utility function. We also assume that consumers have a fixed budget. So each consumer i initially has an amount $w_{i,j}$ of commodity j. Let $p = (p_1, \ldots, p_n)$ denote the price vector of the commodities, i.e., p_j is the unit price of commodity j. If $x_{i,j}(p)$ denotes the final consumer i demand of commodity j, then the following budget constraint holds for each consumer i,

$$\sum_{j=1}^{n} p_j w_{i,j} = \sum_{j=1}^{n} p_j x_{i,j}(p)$$

Summing both sides over i we get the **Walras's law**,

$$\sum_{j=1}^{n} p_j g_j(p) = 0$$

where w_j is the initial aggregate amount of commodity j,

$$w_j = \sum_{i=1}^{m} w_{i,j}$$

and

$$g_j(p) = \sum_{i=1}^{m} x_{i,j}(p) - w_j$$

is the aggregate excess demand. The following natural question arises: is it possible to find a price vector so that the aggregate demand for each commodity does not exceed its initial aggregate amount? Prices with this property are called Walras equilibrium prices.

It is clear that only relative prices matter in this problem. Thus we suppose that

$$p \in \Delta^{n-1} = \{(x_1, \dots, x_n) \in \mathbb{R}^n_+ \colon x_1 + \cdots + x_n = 1\}.$$

Here $\mathbb{R}^n_+$ denotes the set of vectors $x \in \mathbb{R}^n$ with non-negative entries, i.e., $x_j \geq 0$ for every $j = 1, \dots, n$. The set Δ^{n-1}, called the simplex of dimension $n-1$, represents the set of relative price vectors.

Now, formalizing the question above, we want to find $p^* \in \Delta^{n-1}$ such that

$$g_j(p) \leq 0 \quad \text{for every} \quad j = 1, \dots, n$$

In order to apply the Brouwer theorem, we define a function $G : \Delta^{n-1} \to \Delta^{n-1}$. First, it is easy to see that Δ^{n-1} is compact and convex. Now we define the function

$$G(p) = \frac{1}{d(p)} \left(p_1 + \max\{0, g_1(p)\}, \dots, p_n + \max\{0, g_n(p)\}\right)$$

where

$$d(p) = 1 + \sum_{j=1}^{n} \max\{0, g_j(p)\}$$

Note that $d(p) \geq 1$. Thus G is continuous under the reasonable assumption that the demand functions $x_{i,j}(p)$ are also continuous. Applying the Brouwer fixed point theorem to get $p^* \in \Delta^{n-1}$ satisfying

$$p^* = G(p^*).$$

In coordinates, the previous fixed point equation is

$$d(p^*)(p_1^*, \dots, p_n^*) = (p_1^* + \max\{0, g_1(p^*)\}, \dots, p_n^* + \max\{0, g_n(p^*)\})$$

This means that

$$(d(p^*) - 1)p_j^* = \max\{0, g_j(p^*)\}, \quad j = 1, \dots, n$$

By Walras's law, there must exist a positive price $p_k^* > 0$ which has $g_k(p^*) \leq 0$. Indeed, if $g_j(p^*) > 0$ for every j whose price p_j^* is positive, then the sum $\sum_j p_j^* g_j(p^*)$ would also be positive contradicting Walras's law. Therefore, $\max\{0, g_k(p^*)\} = 0$, which implies that $d(p^*) = 1$. Thus, $\max\{0, g_j(p^*)\} = 0$ for every j which implies that

$$g_1(p^*) \leq 0, \ldots, g_n(p^*) \leq 0$$

as we wanted to show.

Exercise 17. Decide which of the following sets are convex:

(1) $A = \{(x,y) \in \mathbb{R}^2 \colon y \geq x^2\}$
(2) $B = \{(x,y) \in \mathbb{R}^2 \colon y < x^2\}$
(3) $C = \{(x,y) \in \mathbb{R}^2 \colon \frac{x^2}{4} + y^2 \leq 1\}$
(4) $A \cap C$
(5) $C \setminus B$
(6) $D = \{(x,y,z) \in \mathbb{R}^3 \colon x+y+z = 1,\ x,y,z \geq 0\}$

Exercise 18. Consider the function $f(x) = \frac{1}{2}(x+1)$ defined on the interval $]0,1[$. Show that $f(]0,1[) \subset]0,1[$. Does f has a fixed point in $]0,1[$? Can you apply the Brouwer fixed point theorem to f ?

Exercise 19. Consider the following matrix

$$A = \begin{pmatrix} 0 & 1/2 & 1 \\ 1 & 0 & 0 \\ 0 & 1/2 & 0 \end{pmatrix}$$

and define the function $f(v) = Av$ where $v \in \Delta^2 = \{(x,y,z) \in \mathbb{R}^3 \colon x+y+z = 1,\ x,y,z \geq 0\}$.

(1) Show that $f(\Delta^2) \subset \Delta^2$.
(2) Can you apply the Brouwer fixed point theorem to f? Can you show that f has a fixed point?
(3) If yes, compute the fixed points of f explicitly.

Exercise 20. Find a continuous function $f \colon [0,1[\to [0,1[$ with no fixed points. Can you apply the Brouwer fixed point theorem to f?

Exercise 21. Find a continuous function $f \colon [0,1] \to [0,1]$ with an infinite number of fixed points.

Exercise 22. Consider the function $f \colon [0,1] \to [0,1]$ defined by

$$f(x) = \begin{cases} 2x, & x \leq 1/2 \\ 2-2x, & x > 1/2 \end{cases}$$

(1) Can you apply the Brouwer fixed point to f?
(2) Compute the fixed points of f and of $f \circ f$ (hint: draw the graph of the functions).

(3) How many fixed points has $f^n = f\circ\cdots\circ f$? Here, f^n denotes the composition of f with itself n times ($f^2 = f\circ f$, $f^3 = f\circ f\circ f$, etc.)

Exercise 23. Consider a pure exchange economy with 2 commodities and 2 consumers. Denote by p_1 and p_2 the relative price of commodity 1 and 2, respectively. The 1st consumer has an initial amount of 1 unit of commodity 1 and 2 units of commodity 2. The 2nd consumer has an initial amount of 2 units of commodity 1 and 1 unit of commodity 2. Consumers engage in exchanging with the following demand functions

$$x_{1,1}(p) = \frac{\alpha(p_1+2p_2)}{p_1}, \quad x_{1,2}(p) = \frac{(1-\alpha)(p_1+2p_2)}{p_2}$$

for the 1st consumer and

$$x_{2,1}(p) = \frac{\alpha(2p_1+p_2)}{p_1}, \quad x_{2,2}(p) = \frac{(1-\alpha)(2p_1+p_2)}{p_2}$$

for the 2nd consumer, where $\alpha \in [0,1]$. Recall that $x_{i,j}(p)$ is the demand of commodity j of consumer i. Check that the Walras's law is satisfied. Determine an equilibrium price for this economy.

4. Hyperplane separation theorem

Given $x, y \in \mathbb{R}^n$ define the inner product

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

Given any $p\in\mathbb{R}^n$ and $c\in\mathbb{R}$ we denote by $H(p,c)$ the **hyperplane**,

$$H(p,c) = \{x\in\mathbb{R}^n\colon p\cdot x = c\}.$$

Notice that the same hyperplane can be represented with infinitely many pairs (p,c). In fact, $H(p,c) = H(\alpha p, \alpha c)$ for any $\alpha\in\mathbb{R}\setminus\{0\}$. Thus, the direction of p is uniquely determined but not its magnitude. Geometrically, $H(p,c)$ can be seen as the set of vectors with base point $h\,p$ that are orthogonal[9] to p. Here, $h = c/(p\cdot p)$ which can be seen as the height of $H(p,c)$. The vector p is orthogonal to the hyperplane. If $c=0$, then $H(p,0)$ is simply the set of vectors in $\mathbb{R}^n$ that are orthogonal to p. As an example,

$$H((1,1),1) = \{(x,y)\in\mathbb{R}^2\colon x+y=1\}.$$

Notice that, hyperplanes in $\mathbb{R}^2$ are lines and hyperplanes in $\mathbb{R}^3$ are planes. Any hyperplane separates $\mathbb{R}^n$ in two regions: the **upper half-space**

$$H^+(p,c) = \{x\in\mathbb{R}^n\colon p\cdot x\geq c\}$$

[9] Also called perpendicular. Recall that two vectors x and y in $\mathbb{R}^n$ are orthogonal iff $x\cdot y = 0$.

and the **lower half-space**

$$H^-(p,c) = \{x \in \mathbb{R}^n \colon p \cdot x \leq c\}$$

Given two subsets A and B of $\mathbb{R}^n$ we say that A and B are **separated by a hyperplane** if there is $p \in \mathbb{R}^n$ and $c \in \mathbb{R}$ such that $A \subset H^-(p,c)$ and $B \subset H^+(p,c)$. The hyperplane separation theorem gives a sufficient condition for any two sets to be separated.

Theorem 4.1 (Separation theorem)**.** *If A and B are disjoint and convex, then A and B can be separated by a hyperplane.*

Without the convexity assumption there is no guarantee that A and B can be separated. For instance, the following sets cannot be separated by a line,

$$\begin{aligned} A &= \{(x,y) \in \mathbb{R}^2 \colon x^2 + y^2 \leq 1\} \\ B &= \{(x,y) \in \mathbb{R}^2 \colon 2 \leq x^2 + y^2 \leq 3\} \end{aligned}$$

Notice that B is not convex, so the theorem does not apply.

Exercise 24. Sketch the following hyperplanes and half-spaces:

(1) $H((1,1),-1)$
(2) $H^+((2,-1),1)$
(3) $H((1,0,-1),1)$
(4) $H^-((0,1,0),2)$

Exercise 25.

(1) Let A be the set of points in $\mathbb{R}^n$ whose first coordinate is equal to $a \in \mathbb{R}$. Find $p \in \mathbb{R}^n$ and $c \in \mathbb{R}$ such that $A = H(p,c)$.
(2) Find a hyperplane $H(p,c)$ that separates $\Delta^2 = \{(x,y,z) \in \mathbb{R}^3_+ \colon x+y+z = 1\}$ and $A = \{(x,y,z) \in \mathbb{R}^3 \colon x = -2\}$.
(3) Find the hyperplane $H(p,c)$ that contains the points $(1,0,0)$, $(1,1,0)$ and $(0,1,1)$.

Exercise 26. Explain, using the hyperplane separation theorem, if it is possible to prove the existence of a hyperplane separating A and B:

(1) $A = \{(x,y) \in \mathbb{R}^2 \colon x^2 + y^2 < 1\}$ and $B = \{(x,y) \in \mathbb{R}^2 \colon x + y = 2\}$
(2) $A = \{(x,y) \in \mathbb{R}^2 \colon x^2 + y^2 < 3\}$ and $B = \{(x,y) \in \mathbb{R}^2_+ \colon x + y = 1\}$
(3) $A = C \cap D$ with $C = \{(x,y) \in \mathbb{R}^2 \colon x^2 + y^2 = 2\}$ and $D = [-6,6] \times [-5,5]$, and $B = [-1,1] \times [-1,1]$

5. Correspondences

To motivate the introduction of correspondences we begin with two examples.

Example 5.1. Given a vector of prices $p = (p_1, \dots, p_n) \in \mathbb{R}^n_+$ and an income $I > 0$ the budget set is

$$\mathcal{B}(p, I) = \{x \in \mathbb{R}^n_+ \colon \sum_{i=1}^n p_i x_i \leq I\}.$$

So, to each pair (p, I) it *corresponds* a budget set $\mathcal{B}(p, I) \subset \mathbb{R}^n$.

Example 5.2. Consider a firm producing a single commodity at a cost

$$c(q) = \begin{cases} 0, & q = 0 \\ a + bq + cq^2, & q > 0 \end{cases}$$

where $a, b, c > 0$. Suppose that the commodity has output price $p > 0$ with $p > b$. Then the profit is

$$\pi(q) = pq - c(q) = \begin{cases} 0, & q = 0 \\ -a + (p - b)q - cq^2, & q > 0 \end{cases}$$

The quantity q^* that maximizes the profit is given by $\pi'(q^*) = 0$, i.e., $q^* = (p-b)/(2c)$. The corresponding profit is $\pi(q^*) = (p-b)^2/(2c) - a$. Therefore, $\pi(q^*) \geq 0$ if and only if $p \geq b + \sqrt{2ac}$. Thus, the profit maximizing choice of output is

$$q(p) = \begin{cases} 0, & p \leq b + \sqrt{2ac} \\ (p - b)/(2c), & p \geq b + \sqrt{2ac} \end{cases}$$

So, for the price $p = b + \sqrt{2ac}$ there corresponds a set of quantities $\{0, \sqrt{a/(2c)}\}$, allowing the producer to earn zero profit.

These examples can be described using correspondences. Given two sets A and B, a **correspondence** F is a rule that to each $a \in A$ associates a subset $F(a) \subset B$. It is common to write $F \colon A \rightrightarrows B$. The set A is called the **domain** of F. The graph of a correspondence F is

$$\operatorname{graph}(F) = \{(x, y) \in A \times B \colon x \in A,\, y \in F(x)\}$$

Example 5.3. As an example, consider the graph of the correspondences $\mathbb{R} \rightrightarrows \mathbb{R}$,

$$F(x) = \begin{cases} [1, 3], & x < 1 \\ \{2\}, & x \geq 1 \end{cases}, \quad G(x) = \begin{cases} [x - 7, x - 5], & x < 0 \\ ]6, 9[, & x = 0 \\ ]x + 7, x + 8], & x > 0 \end{cases}$$

5.1. **Continuity of correspondences.** We say that a correspondence $F \colon A \subset \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ has a **closed graph at** x if for any convergent sequence $\{(x_n, y_n)\} \subset \operatorname{graph}(F)$ with limit $(x, y) \in A \times \mathbb{R}^m$ we have that $y \in F(x)$, i.e., $(x, y) \in \operatorname{graph}(F)$. We say that F has the **closed graph property** if it has closed graph at each $x \in A$.

Proposition 5.4. *If F has the closed graph at x, then $F(x)$ is a closed set.*

Proof. Exercise. □

Proposition 5.5. *If the graph of F is closed, then F has the closed graph property.*

Proof. Exercise. □

The correspondence F in Example 5.3 does not have the closed graph at $x = 1$.

Example 5.6. The budget correspondence in Example 5.1 that to each $p \in \mathbb{R}^n_+$ corresponds $\mathcal{B}(p, I)$ with I fixed, has the closed graph property.

We say that a correspondence $F \colon A \subset \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is **upper hemicontinuous (u.h.c) at** $x \in A$ if for every open set U containing $F(x)$ there is an open set V containing x such that $F(x) \subset U$ for every $x \in A \cap V$. We say that F is **upper hemicontinuous** if it is upper hemicontinuous at each $x \in A$.

The following result shows that upper hemicontinuity is the same as continuity when the correspondence is single-valued. That is, upper hemicontinuity is an extension of continuity of functions to correspondences.

Proposition 5.7. *A function $f : A \subset \mathbb{R}^n \to \mathbb{R}^m$ is continuous at $x \in A$ if and only if the correspondence $F \colon A \subset \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ defined by $F(x) = \{f(x)\}$ is u.h.c at $x \in A$.*

The following result shows that if the set $F(x)$ is closed but F does not have the closed graph at x, then F cannot be upper hemicontinuous at x.

Proposition 5.8. *If F is u.h.c at $x \in A$ and $F(x)$ is closed, then F has the closed graph at x*

Proof. Let $(x_n, y_n) \in \text{graph}(F)$ converge to (x, y) by $y \notin F(x)$. There is a closed ball $\overline{B}(y, r)$ not intersecting $F(x)$. The complement set $U = \mathbb{R}^m \setminus \overline{B}(x, r)$ is open. By u.h.c, there is an open set V containing x such that $F(x) \notin \overline{B}(x, r)$ for every $x \in V$. But for large n, $x_n \in V$, thus $y_n \notin \overline{B}(y, r)$ contradicting the convergence. □

The following result gives a sufficient condition for checking that a correspondence is upper hemicontinuous at a point. We say that F **is bounded near** x if there is a neighbourhood V of x and a closed ball $B \subset \mathbb{R}^m$ such that $F(y) \subset B$ for every $y \in V$.

Proposition 5.9. *If F has the closed graph at x and F is bounded near x, then F is u.h.c at x.*

Proof. See the bibliography. □

With this proposition we finally arrive at a simple criterion to decide if a given correspondence is upper hemicontinuous.

Theorem 5.10. *If the graph of F is compact, then F is upper hemicontinuous.*

Proof. To prove the theorem we use Proposition 5.9. Since the graph of F is closed (because it is compact by assumption), we know by Proposition 5.5 that F has the closed graph property. Moreover, F is bounded near any point $x \in A$. Otherwise, the graph of F would not be compact. Hence, the hypothesis of Proposition 5.9 are met for any point $x \in A$. So, F is upper hemicontinuous. □

Exercise 27. Decide if the following correspondences $F : [0,1] \rightrightarrows [0,1]$ are upper hemicontinuous and/or have the closed graph property:

(1)
$$F(x) = \begin{cases} [1/4, 3/4], & x < 1/2 \\ [1/2, 3/4], & x \geq 1/2 \end{cases}$$

(2)
$$F(x) = \begin{cases} [x^2, (x+1)/2], & x < 1 \\ \{0, 1\}, & x = 1 \end{cases}$$

(3)
$$F(x) = \begin{cases}]x, 1-x[, & x < 1/2 \\ \{1/4, 3/4\}, & x \geq 1/2 \end{cases}$$

(4)
$$F(x) =]x/2, (x+1)/2[$$

5.2. **Kakutani fixed point.** Now we are able to state the main result of this section, i.e., the existence of a fixed point for correspondences. We say that $x \in A$ is a fixed point for F if $x \in F(x)$.

Theorem 5.11 (Kakutani fixed point theorem)**.** *Let $K \subset \mathbb{R}^n$ be a compact and convex set. If*

(1) *$F : K \rightrightarrows K$ is upper hemicontinuous*
(2) *$F(x)$ is convex and non-empty for every $x \in K$*

then F has a fixed point in K.

Proof. Check the bibliography. □

Exercise 28. Explain if the following correspondences satisfy the hypothesis of the Kakutani fixed point theorem. Compute the fixed points if they exist.

(1) $F\colon [0,2] \rightrightarrows [0,2]$ defined by
$$F(x) = \begin{cases} \{1\}, & 0 \leq x < 1 \\ [1, 2], & x = 1 \\ \{2\}, & 1 < x \leq 2 \end{cases}$$

(2) $F\colon [0,20] \rightrightarrows [0,20]$ defined by

$$F(x) = \begin{cases} \{10-x\}, & 0 \le x < 7 \\ [3,20], & x = 7 \\ \{x-4\}, & 7 < x \le 20 \end{cases}$$

(3) $F\colon [-6,20] \rightrightarrows [-6,20]$ defined by

$$F(x) = \begin{cases} \{x+1\}, & -6 \le x < 7 \\ [-6,10], & x = 7 \\ \{(x+5)/2\}, & 7 < x \le 20 \end{cases}$$

(4) $F\colon [-6,20] \rightrightarrows [-6,20]$ defined by

$$F(x) = \begin{cases} \{x+1\}, & -6 \le x < 7 \\ [-6,6] \cup [8,10], & x = 7 \\ \{(x+5)/2\}, & 7 < x \le 20 \end{cases}$$

6. Optimization

Optimization problems arise rather naturally in Economics. Consider for instance, the problem of maximizing utility subject to budget constrains, i.e.,

$$\begin{aligned}&\text{maximize } u(x)\\&\text{subject to } x \in \mathcal{B}(p, I)\end{aligned}$$

where $u : \mathbb{R}^n_+ \to \mathbb{R}$ is a utility function, $I > 0$ the income, $p \in \mathbb{R}^n_+$ the vector price and $\mathcal{B}(p, I)$ the budget set

$$\mathcal{B}(p, I) = \{x \in \mathbb{R}^n_+ \colon p \cdot x \le I\}.$$

One can also consider the problem of minimizing the spending $p \cdot u$ but not going below a certain utility level ℓ, i.e.,

$$\begin{aligned}&\text{minimize } p \cdot x\\&\text{subject to } u(x) \ge \ell\end{aligned}$$

In general, given a function $x \mapsto f(x) \in \mathbb{R}$ and a set $D \subset \mathbb{R}^n$, the **optimization problem** is

$$\begin{aligned}&\text{maximize (minimize) } f(x)\\&\text{subject to } x \in D\end{aligned}$$

The function f is usually called the **objective function** and the set D the **constraint set**.

Recall, by the Weierstrass theorem, if the objective function f is continuous in D and the constraint set D is compact, then f attains a maximum and a minimum in D, i.e., there is $x_*, x^* \in D$ such that

$$f(x_*) \le f(x) \le f(x^*), \quad \forall\, x \in D$$

This means that the optimization problem has a solution when f is continuous and D compact. However, the Weierstrass theorem, does not provide an explicit solution, only its existence. For many practical applications one needs a method to explicitly compute a solution.

Depending on the function f and the set D there are many optimization problems. For instance, the **linear optimization problem** (or linear programming) has

$$f(x) = c_1 x_1 + \cdots + c_n x_n, \quad D = \{x \in \mathbb{R}^n \colon Ax \le b\}$$

where A is an $m \times n$ matrix and $b \in \mathbb{R}^m$. Given two vectors $u, v \in \mathbb{R}^m$ we write $u \le v$ meaning that $u_i \le v_i$ for each component $i = 1, \dots, m$.

Other optimization problems are called **integer programming** where the variables are constrained to integer values, and **convex optimization** where both utility function and the constraint set are convex.

6.1. **Some terminology.** Given $D \subset \mathbb{R}^n$ and a function $f\colon D \to \mathbb{R}$ we say that $x^* \in D$ is a **maximizer of** f **on** D if $f(x) \leq f(x^*)$ for every $x \in D$. The value $f(x^*)$ is called the **maximum of** f **on** D. It is common to write $\max_D f$ for the maximum of f on D. A **local maximizer of** f **on** D is a point $x^* \in D$ such that $f(x) \leq f(x^*)$ for every $x \in D \cap B(x^*, r)$ and some $r > 0$. In this case, the value $f(x^*)$ is called a **local maximum of** f **on** D. Clearly, any maximizer is a local maximizer. The converse is not true. To distinguish maximizers from local maximizers we often call global maximizer (global maximum) to a maximizer (maximum). With the correct modifications we can define minimizer (minimum) and local minimizer (local minimum). Maximizers and minimizers are called **optimal**[10] **points**. If they are local, then **local optimal points**. Their values are called **(local) optimal values**.

Example 6.1. The function $f(x) = 4x^2 - 4x + 2$ on $D = \mathbb{R}$ has a minimum $\min_D f = 1$ with minimizer $x = 1/2$.

Example 6.2. The function $f(x, y) = 1 - x^2 - y^2$ on $D = \mathbb{R}^2$ has a maximum $\max_D f = 1$ with maximizer $x = 0$.

Example 6.3. The function $f(x) = \frac{x^3}{3} - x$ on $D = [-3, 3]$ has a optimal points by Weirestrass theorem. In fact, $\max_D f = 6$ and $\min_D f = -6$ with maximizer $x = 3$ and minimizer $x = -3$. The function also has a local maximizer at $x = -1$ and a local minimizer at $x = 1$.

In order to systematically compute the local optimal points we first study a simpler optimization problem called **unconstrained optimization**, i.e., the problem of maximizing or minimizing a given function inside the interior of its domain.

6.2. **Unconstrained optimization.** Given a set $D \subset \mathbb{R}^n$ and a function $f : D \to \mathbb{R}$ the derivative of f at x is row vector

$$Df(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) & \cdots & \frac{\partial f}{\partial x_n}(x) \end{bmatrix}.$$

We say that f is of class C^1 if its partial derivatives $\frac{\partial f}{\partial x_i}$ are continuous in D. The second derivative of f at x is the matrix

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

We say that f is of class C^2 if its partial derivatives $\frac{\partial^2 f}{\partial x_i \partial x_j}$ are continuous in D. In this case, $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$ and so $D^2 f(x)$ is a symmetric matrix, also called, the **Hessian matrix**.

[10] or extreme

Recall that $\operatorname{int}(D)$ is the set of all interior points to D, i.e., $x \in D$ is interior if there is $r > 0$ such that $B(x,r) \subset D$. We say that $x^* \in \operatorname{int}(D)$ is a **critical**[11] **point** of f if

$$Df(x^*) = 0$$

It is easy to see that local optimal points are critical points.

Proposition 6.4. *If $x^* \in int(D)$ is a local optimal point of f, then x^* is a critical point.*

Proof. Check the bibliography. □

Clearly, not all critical points are local optimal points. For instance, $f(x) = x^3$ with $D = \mathbb{R}$. We have $f'(x) = 3x^2$, thus $x^* = 0$ is the single critical point. However, $x^* = 0$ is not a local optimal point, since the function f increases for $x > 0$ and decreases for $x < 0$.

6.2.1. *Second-order condition.* To decide if a given critical point is a local optimal point we need a criterion, called the second-order condition. The second-order condition is a condition on the Hessian matrix at the critical point. To explain the condition we introduce the following terminology. First, recall that a symmetric square matrix A has all its eigenvalues real. We say that A is **positive semi-definite**, and we write $A \geq 0$, if all eigenvalues of A are non-negative, i.e., ≥ 0. If all eigenvalues are positive (> 0), then A is **positive definite**, and we write $A > 0$. An alternative way to check if $A > 0$ is the **Sylvester's criterion**. This criterion says that A is positive definite if and only if all its leading principal minors are positive. A leading principal minor of A is the determinant of an upper-left i-by-i corner of A. We denote the i-th leading principal minor by Δ_i. Any square matrix A of dimension n has exactly n leading principal minors since $i = 1, \dots, n$. For instance, the matrix

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

has the following leading principal minors

$$\Delta_1 = \begin{vmatrix} 2 \end{vmatrix} = 2, \quad \Delta_2 = \begin{vmatrix} 2 & 0 \\ 0 & 3 \end{vmatrix} = 6, \quad \Delta_3 = \begin{vmatrix} 2 & 0 & 1 \\ 0 & 3 & -1 \\ 1 & -1 & 1 \end{vmatrix} = 1$$

Since all leading principal minors are positive, we conclude that $A > 0$. Finally, we say that A is **negative (semi-)definite** and write $A < 0$ ($A \leq 0$), if $-A$ is positive (semi-)definite. Notice that $A < 0$ if and only if the i-th leading principal minor has sign $(-1)^i$. Summarizing,

- $A > 0$ iff $\Delta_i > 0$ for every $i = 1, \dots, n$.
- $A < 0$ iff $(-1)^i \Delta_i > 0$ for every $i = 1, \dots, n$.

[11] or stationary

A point $x^* \in D$ is called a **strict local maximizer of f on D** if there is $r > 0$ such that $f(x) < f(x^*)$ for every $x \in D \cap B(x^*, r)$ and $x \neq x^*$. Clearly, strict local maximizers are local maximizers but the converse may not be true. Strict local minimizers are defined analogously. We have the following criterion based on the second derivative of f.

Theorem 6.5 (Second-order condition)**.** *Let f be of class C^2 and $x^* \in int(D)$ be a critical point. The following holds:*

- *If $D^2 f(x^*) < 0$, then x^* is a strict local maximizer.*
- *Conversely, if x^* is a local maximizer, then $D^2 f(x^*) \leq 0$.*

Proof. Check the bibliography. □

This theorem has a version for strict local minimizer where the second-order condition reads $D^2 f(x^*) > 0$. In addition, $D^2 f(x^*) \leq 0$ whenever x^* is a local minimizer. As a consequence of the theorem, if neither $D^2 f(x^*) \leq 0$ nor $D^2 f(x^*) \geq 0$, then the critical point x^* is not a local optimal point. In this case, we say that x^* is a **saddle point**. Therefore, critical points are either local optimal points or saddle points.

According to what has been said we have the following condition to check if a critical point is a saddle point.

Proposition 6.6. *If $x^* \in int(D)$ is a critical point,* $\det(D^2 f(x^*)) \neq 0$ *and neither $D^2 f(x^*) < 0$ nor $D^2 f(x^*) > 0$, then x^* is a saddle point.*

Example 6.7. Let $f(x, y) = x^2 - y^2$ with $D = \mathbb{R}^2$. The point $(x^*, y^*) = (0, 0)$ is a critical point of f. Indeed, since $\frac{\partial f}{\partial x}(x, y) = 2x$ and $\frac{\partial f}{\partial y}(x, y) = -2y$ we have $\frac{\partial f}{\partial x}(0, 0) = \frac{\partial f}{\partial y}(0, 0) = 0$. On the other hand, the Hessian matrix of f at the critical point is

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

Therefore, $D^2 f(0, 0)$ is neither positive definite nor negative definite (it has positive and negative eigenvalues). So, $(0, 0)$ is a saddle point.

Example 6.8. Consider the function

$$f(x, y, z) = \frac{x^3}{3} + \frac{y^3}{3} + \frac{z^3}{3} + xy + xz + yz$$

with $D = \mathbb{R}^3$. The derivative of f is

$$Df(x, y, z) = \begin{bmatrix} x^2 + y + z & y^2 + x + z & z^2 + x + y \end{bmatrix}$$

To find the critical points we have to solve the system

$$\begin{cases} x^2 + y + z = 0 \\ y^2 + x + z = 0 \\ z^2 + x + y = 0 \end{cases}$$

Subtracting the 2nd - 1st and 3rd - 1st equations we get an equivalent system

$$\begin{cases} x^2+y+z=0 \\ (y-x)(x+y-1)=0 \\ (z-x)(x+z-1)=0 \end{cases}$$

Now we see that both $x+y$ and $x+z$ cannot be equal to 1, otherwise using the second and third equations of the 1st system we get $y^2+1=0$ and $z^2+1=0$ which has no real solutions. Thus, only $x=y=z$ is possible. In this case, we get $x^2+2x=0$, so $x=0$ or $x=-2$. We conclude that f has critical points $(0,0,0)$ and $(-2,-2,-2)$.

Computing the Hessian matrix we get,

$$D^2f(x,y,z)=\begin{bmatrix} 2x & 1 & 1 \\ 1 & 2y & 1 \\ 1 & 1 & 2z \end{bmatrix}$$

At the critical point $(-2,-2,-2)$ we have

$$D^2f(-2,-2,-2)=\begin{bmatrix} -4 & 1 & 1 \\ 1 & -4 & 1 \\ 1 & 1 & -4 \end{bmatrix}$$

and the leading principal minors are

$$\begin{vmatrix} -4 \end{vmatrix}=-4, \quad \begin{vmatrix} -4 & 1 \\ 1 & -4 \end{vmatrix}=15, \quad \begin{vmatrix} -4 & 1 & 1 \\ 1 & -4 & 1 \\ 1 & 1 & -4 \end{vmatrix}=-50$$

This means that $-D^2f(-2,-2,-2)>0$, that is, $D^2f(-2,-2,-2)<0$. So the critical point $(-2,-2,-2)$ is a local maximizer.

Doing the same for the critical point $(0,0,0)$ we conclude that the leading principal minors of the Hessian matrix are 0, -1 and 2. In this case, neither $D^2f(0,0,0)>0$ nor $D^2f(0,0,0)<0$, so because $\det(D^2f(0,0,0))\neq 0$, we conclude that $(0,0,0)$ is a saddle point.

Exercise 29. Classify the critical points of the following functions,

(1) $f(x,y,z)=x^2+2y^2+3z^2+2xy+2xz$ on $\mathbb{R}^3$
(2) $f(x,y,z,w)=20y+48z+6w+8xy-4x^2-12z^2-w^2-4y^3$ on $\mathbb{R}^4$
(3) $f(x,y,z)=z\log(x^2+y^2+z^2)$ on $\mathbb{R}^3\setminus\{(0,0,0)\}$

Exercise 30. Consider the following function defined on $\mathbb{R}^2$,

$$f(x,y)=(1+y)^3x^2+y^2$$

Show that f has a unique critical point (x^*,y^*) which is a local minimizer. Is (x^*,y^*) a global minimizer ?

6.3. **Convex and concave functions.** The function $f\colon D \to \mathbb{R}$ is **convex on** D if for every $x, y \in D$ and every $\lambda \in]0,1[$,

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

If the previous inequality holds for $\geq$ then f is said to be **concave on** D. If the inequality is strict $<$ (resp. $>$) whenever $x \neq y$, then f is **strictly** convex (resp. concave). Clearly, f is (strictly) convex iff $-f$ is (strictly) concave.

Example 6.9. The function $f(x) = 1 - x^2$ is strictly concave on $\mathbb{R}$ and $g(x) = x^2$ is strictly convex on $\mathbb{R}$.

Example 6.10. Any linear function $f(x) = a_1x_1 + \cdots + a_nx_n$ defined on $D = \mathbb{R}^n$ is both convex and concave on $\mathbb{R}$.

The following theorem provides a simple characterization of convex functions in terms of the second derivative.

Theorem 6.11. *Characterization of convex (concave) functions Let $D \subset \mathbb{R}^n$ be open and convex and $f : D \to \mathbb{R}$ be of class C^2. The following holds:*

- *f is convex iff $D^2f(x) \geq 0$ for every $x \in D$,*
- *f is concave iff $D^2f(x) \leq 0$ for every $x \in D$.*

Furthermore,

- *if $D^2f(x) > 0$ for every $x \in D$, then f is strictly convex,*
- *if $D^2f(x) < 0$ for every $x \in D$, then f is strictly concave.*

Proof. Check bibliography. □

Example 6.12. The function $f(x,y,z) = \log(xyz)$ is strictly concave on $\mathbb{R}^3_+$. Indeed,

$$D^2f(x,y,z) = \begin{bmatrix} -\frac{1}{x^2} & 0 & 0 \\ 0 & -\frac{1}{y^2} & 0 \\ 0 & 0 & -\frac{1}{z^2} \end{bmatrix}$$

Clearly, $D^2f(x,y,z) < 0$ for every $(x,y,z) \in \mathbb{R}^3_+$. Thus, f is strictly concave.

Proposition 6.13. *Let $D \subset \mathbb{R}^n$ be convex and $f\colon D \to \mathbb{R}$ convex (concave). Then*

- *Local minimizers (maximizers) are global minimizers (maximizers),*
- *if f is strictly convex (concave) and it has a minimizer (maximizer), then it is unique.*

Theorem 6.14 (Necessary and sufficient condition)**.** *Let $D \subset \mathbb{R}^n$ be open and convex, and $f : D \to \mathbb{R}$ be differentiable and convex (concave) function. Then $x^* \in D$ is a local minimizer if and only if x^* is a critical point.*

Proof. One direction has been proved. To prove the other direction, suppose that $x^* \in D$ is a critical point and f is convex. The concave case is analogous. Then

$$f(x) - f(x^*) \geq Df(x^*)(x - x^*)$$

So $f(x) \geq f(x^*)$ for every x near x^*, so x^* is a minimizer. □

Exercise 31. Determine if the following functions are (strictly) convex/concave:

(1) $f(x, y) = 2x - y - x^2 + 2xy - y^2$ on $\mathbb{R}^2$.
(2) $f(x, y) = x^a y^b$ on $\mathbb{R}^2_+$ and $a + b \leq 1$ with $a, b \geq 0$.

Exercise 32. Find the largest domain $D \subset \mathbb{R}^2$ on which the following function is concave,

$$f(x, y) = x^2 - y^2 - xy - x^3.$$

6.4. **Equality constraints.** Given C^2 functions f and g_i are defined in some open subset $U \subset \mathbb{R}^n$, the optimization problem with equality constraints is

$$\begin{aligned} \text{maximize (minimize) } & f(x) \\ \text{subject to } & g_1(x) = 0 \\ & g_2(x) = 0 \\ & \vdots \\ & g_m(x) = 0 \end{aligned}$$

This problem has m equality constraints and we assume that $m < n$. We denote by $g = (g_1, \dots, g_n)$ the vector function and we define the constrain set

$$D = \{x \in \mathbb{R}^n \colon g(x) = 0\}$$

So the problem is to find the local optimal points of f on D.

The standard procedure to solve this problem is to define the **Lagrangian function**

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

where $\lambda \in \mathbb{R}^m$ are the **Lagrange multipliers**. The following theorem is due to Lagrange.

Theorem 6.15 (Necessary condition)**.** *If $x^* \in D$ is a local optimal point of f on D and* $\operatorname{rank} Dg(x^*) = m$*, then there is $\lambda^* \in \mathbb{R}^m$ such that (x^*, λ^*) is a critical point of the Lagrangian, i.e.,*

$$\begin{cases} Df(x^*) + \sum_{i=1}^m \lambda_i^* Dg_i(x^*) = 0 \\ g(x^*) = 0 \end{cases}$$

Proof. Check the bibliography. □

The condition $\operatorname{rank} Dg(x^*) = m$ is called **constraint qualification**.

The $\operatorname{rank} A$ of an $m \times n$ matrix A is the maximum number of linearly independent columns (or rows) of A. We say that A has **full rank** whenever $\operatorname{rank} A = m$, i.e., the rows of A are linearly independent, or equivalently, A has m columns whose determinant is non-zero.

Example 6.16. Consider the function $f(x, y) = xy$ subject to the constrain $x^2 + y^2 = 2$. Let $g(x, y) = x^2 + y^2 - 2$ and $D = \{(x, y) \in \mathbb{R}^2 : g(x, y) = 0\}$. By the previous theorem, the critical points of the Lagrangian are candidates for local optimal points of f on D. The critical points satisfy the equations,

$$\begin{cases} \frac{\partial f}{\partial x} + \lambda \frac{\partial g}{\partial x} = 0 \\ \frac{\partial f}{\partial y} + \lambda \frac{\partial g}{\partial y} = 0 \\ g(x,y) = 0 \end{cases} \Leftrightarrow \begin{cases} y + 2\lambda x = 0 \\ x + 2\lambda y = 0 \\ x^2 + y^2 = 2 \end{cases}$$

First, we deal with the top two equations. We have $y = -2\lambda x$, thus $x + 2\lambda y = x(1 - 4\lambda^2) = 0$. This implies that either $x = 0$ or $\lambda = \pm 1/2$. If $x = 0$, then $y = 0$, which is impossible because $x^2 + y^2 = 2$. Therefore, only $\lambda = \pm 1/2$ is possible. To determine x we use the 3rd equation. Hence,

$$2 = x^2 + y^2 = x^2 + 4\lambda^2 x^2 = 2x^2$$

So $x = \pm 1$. Summarizing, we have 4 critical points (x^*, y^*, λ^*),

$$(1, -1, 1/2), \quad (1, 1, -1/2), \quad (-1, 1, 1/2), \quad (-1, -1, -1/2)$$

Since

$$Dg(x, y) = \begin{bmatrix} 2x & 2y \end{bmatrix}$$

we have

$$\operatorname{rank} Dg(\pm 1, \pm 1) = 1$$

So, by Theorem 6.15, all critical points of the Lagrangian are candidates for local optimal points of f on D.

Under convexity assumptions we have the following sufficient condition.

Theorem 6.17 (Sufficient condition under convexity)**.** *If there are $\lambda^* \in \mathbb{R}^m$ and $x^* \in D$ such that (x^*, λ^*) is a critical point of the Lagrangian and $L(\lambda^*, x)$ is a convex (concave) function of the variable x, then x^* is a minimizer (maximizer) of f on D.*

Example 6.18. Consider $f(x, y, z) = x + 2z$ on the constraint set D defined by the constraints $x + y + z = 1$ and $x^2 + y^2 + z = 7/4$. Both the objective function and the constraints are convex. In order to determine the local optimal points of f on D we have to compute the critical points of the associated Lagrangian,

$$L(x, y, z, \lambda_1, \lambda_2) = x + 2z + \lambda_1(x + y + z - 1) + \lambda_2(x^2 + y^2 + z - 7/4)$$

The critical points are solutions of the system

$$\begin{cases} 1+\lambda_1+2\lambda_2 x=0 \\ \lambda_1+2\lambda_2 y=0 \\ 2+\lambda_1+\lambda_2=0 \\ x+y+z=1 \\ x^2+y^2+z=7/4 \end{cases}$$

From the 3rd equation we get $\lambda_1=-2-\lambda_2$. Substituting into the 1st and 2nd equations we obtain

$$x=\frac{1+\lambda_2}{2\lambda_2}, \quad y=\frac{2+\lambda_2}{2\lambda_2}$$

From the 4th equation we get,

$$z=1-x-y=-\frac{3}{2\lambda_2}$$

Thus, by the 5th equation we get

$$\frac{(1+\lambda_2)^2}{4\lambda_2^2}+\frac{(2+\lambda_2)^2}{4\lambda_2^2}-\frac{6\lambda_2}{4\lambda_2^2}=\frac{7}{4}$$

which simplifies after cancellations, $\lambda_2^2=1$. So $\lambda_2=\pm 1$. This means that L has two critical points

$$(x,y,z,\lambda_1,\lambda_2)=(1,3/2,-3/2,-3,1)$$

and

$$(x,y,z,\lambda_1,\lambda_2)=(0,-1/2,3/2,-1,-1)$$

Notice that

$$L(x,y,z,-3,1)=-\frac{11}{4}-2x-3y+x^2+y^2$$

is convex and

$$L(x,y,z,-1,-1)=\frac{11}{4}-y-x^2-y^2$$

is concave. Therefore, $(1,3/2,-3/2)$ is a minimizer and $(0,-1/2,3/2)$ a maximizer of f on D.

Exercise 33. Use Theorem 6.17 to solve the following optimization problems:

(1)

$$\text{maximize } 2x+y$$
$$\text{subject to } x^2+y^2=1$$

(2)

$$\text{minimize } x^2y^2$$
$$\text{subject to } (1/x)^2+(1/y)^2=1$$

(3)

$$\begin{aligned} \text{maximize } & x+4y+z \\ \text{subject to } & x+2y+3z=0 \\ & x^2+y^2+z^2=42 \end{aligned}$$

(4)

$$\begin{aligned} \text{minimize } & x+4z \\ \text{subject to } & x-y+z=2 \\ & x^2+y^2=1 \end{aligned}$$

If the Lagrangian is not convex or concave there is another criterion to determine if the critical points of L give **local** optimal points.

6.4.1. *Second-order condition.* Define the following determinants

$$B_r(x,\lambda) = \begin{vmatrix} 0 & \cdots & 0 & \frac{\partial g_1(x)}{\partial x_1} & \cdots & \frac{\partial g_1(x)}{\partial x_r} \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & \frac{\partial g_m(x)}{\partial x_1} & \cdots & \frac{\partial g_m(x)}{\partial x_r} \\ \frac{\partial g_1(x)}{\partial x_1} & \cdots & \frac{\partial g_m(x)}{\partial x_1} & \frac{\partial^2 L(x,\lambda)}{\partial x_1^2} & \cdots & \frac{\partial^2 L(x,\lambda)}{\partial x_1 \partial x_r} \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1(x)}{\partial x_r} & \cdots & \frac{\partial g_m(x)}{\partial x_r} & \frac{\partial^2 L(x,\lambda)}{\partial x_r \partial x_1} & \cdots & \frac{\partial^2 L(x,\lambda)}{\partial x_r^2} \end{vmatrix}, \quad r=m+1,\ldots,n$$

Then we have the following sufficient criterion for a critical point of L to be a local optimal point.

Theorem 6.19. *Let (x^*,λ^*) be a critical point of the Lagrangian satisfying the constraint qualification condition, i.e., $Dg(x^*)$ has full rank. Then*

1. *If $(-1)^m B_r(x^*,\lambda^*)>0$ for every $r=m+1,\ldots,n$, then x^* is a local minimizer of f on D.*
2. *If $(-1)^r B_r(x^*,\lambda^*)>0$ for every $r=m+1,\ldots,n$, then x^* is a local maximizer of f on D.*

Example 6.20. Consider the optimization problem

$$\begin{aligned} \text{maximize (minimize) } & x^2+y^2 \\ \text{subject to } & 4x^2+2y^2=4 \end{aligned}$$

The critical points of the Lagrangian $L(x,y,\lambda)=x^2+y^2+\lambda(4x^2+2y^2-4)$ satisfy

$$\begin{cases} 2x+8\lambda x=0 \\ 2y+4\lambda y=0 \\ 4x^2+2y^2=4 \end{cases}$$

Taking the 1st and 2nd equations we get $2x(1+4\lambda)=0$ and $2y(1+2\lambda)=0$. Thus either x or y have to be zero. If $x=0$, then $\lambda=-1/2$ and

$y = \pm\sqrt{2}$ by the 3rd equation. If $y = 0$, then $\lambda = -1/4$ and $x = \pm 1$ again by the 3rd equation. So we have 4 critical points,

$$(0, \pm\sqrt{2}, -1/2), \quad (\pm 1, 0, -1/4)$$

Since

$$Dg(x, y) = \begin{bmatrix} 8x & 4y \end{bmatrix}$$

we conclude that both $Dg(0, \pm\sqrt{2})$ and $Dg(\pm 1, 0)$ have full rank, thus all critical points satisfy the constraint qualification condition. Now,

$$B_2(x, y, \lambda) = \begin{vmatrix} 0 & 8x & 4y \\ 8x & 2 + 8\lambda & 0 \\ 4y & 0 & 2 + 4\lambda \end{vmatrix}$$

Hence,

$$B_2(0, \pm\sqrt{2}, -1/2) = 64 \quad \text{and} \quad B_2(\pm 1, 0, -1/4) = -64$$

So $(0, \pm\sqrt{2})$ are local maximizers and $(\pm 1, 0)$ are local minimizers.

Exercise 34. Use Theorem 6.19 to find the local optimal points of f on D where:

(1)

$$f(x, y) = \log(xy)$$
$$D = \{(x, y) \in \mathbb{R}^2 \colon (1/x)^2 + (1/y)^2 = 1\}$$

(2)

$$f(x, y) = x + y$$
$$D = \{(x, y) \in \mathbb{R}^2 \colon xy = 16\}$$

(3)

$$f(x, y, z) = x^2 - z^2$$
$$D = \{(x, y, z) \in \mathbb{R}^3 \colon 2x + z = a,\ x - y = b\}$$

6.5. **Inequality constraints.** Given C^2 functions f and h_i are defined in some open subset $U \subset \mathbb{R}^n$, the optimization problem with inequality constraints is

$$\begin{aligned} \text{maximize (minimize) } & f(x) \\ \text{subject to } & h_1(x) \geq 0 \\ & h_2(x) \geq 0 \\ & \vdots \\ & h_\ell(x) \geq 0 \end{aligned}$$

This problem has ℓ inequality constraints. We denote by $h = (h_1, \ldots, h_\ell)$ the vector function and we define the constrain set

$$D = \{x \in \mathbb{R}^n \colon h(x) \geq 0\}$$

So the problem is to find the local optimal points of f on D.

Given $x \in D$ we say that h_i is an **active constraint for** x if $h_i(x) = 0$. If all vectors $Dh_i(x)$, corresponding to those h_i which are active for x, are linearly independent, then we say that x satisfies the **constraint qualification**. We are now ready to state the necessary conditions for the existence of a local optimal point.

Theorem 6.21 (Kuhn-Tucker necessary conditions)**.** *Suppose that $x^* \in D$ is a local optimal point of f on D and x^* satisfies the constraint qualification. Then there is $\lambda^* = (\lambda_1^*, \dots, \lambda_\ell^*) \in \mathbb{R}^\ell$ such that*

(1) $Df(x^*) + \sum_{i=1}^{\ell} \lambda_i^* Dh_i(x^*) = 0$

(2) $\lambda_i^* h_i(x^*) = 0$ *for every* $i = 1, \dots, \ell$

The conditions (1) and (2) are called the Kuhn-Tucker conditions. They become sufficient under convexity assumptions.

Theorem 6.22 (Sufficient condition)**.** *Suppose that (x^*, λ^*) satisfies*

$$\begin{cases} Df(x^*) + \sum_{i=1}^{\ell} \lambda_i^* Dh_i(x^*) = 0 \\ \lambda_i^* h_i(x^*) = 0, \quad \forall\, i = 1, \dots, \ell \\ h_i(x^*) \geq 0, \quad \forall\, i = 1, \dots, \ell \end{cases}$$

If $\lambda^* \geq 0$ ($\lambda^* \leq 0$) and $L(x, \lambda^*) = f(x) + \sum_i \lambda_i^* h_i(x)$ is a concave (convex) function of the variable x, then x^* is a maximizer (minimizer) of f on D.

Example 6.23. Consider the problem

$$\text{maximize } x^2 - y$$

$$\text{subject to } x^2 + y^2 \leq 1$$

The Kuhn-Tucker conditions are

$$\begin{cases} 2x - 2\lambda x = 0 \\ -1 - 2\lambda y = 0 \\ \lambda(1 - x^2 - y^2) = 0 \\ x^2 + y^2 \leq 1 \end{cases}$$

By the 2nd and 3rd equations we conclude that $\lambda \neq 0$. So $x^2 + y^2 = 1$. The 1st equation gives $x(1 - \lambda) = 0$. Thus, either $x = 0$ or $\lambda = 1$. If $x = 0$, then $y = \pm 1$ and $\lambda = \mp 1/2$. However, if $\lambda = 1$, then $y = -1/2$ and $x = \pm\sqrt{3}/2$. Therefore, we have 4 solutions (x^*, y^*, λ^*)

$$(0, 1, -1/2) \quad (0, -1, 1/2) \quad (\sqrt{3}/2, -1/2, 1) \quad (-\sqrt{3}/2, -1/2, 1)$$

Since

$$\begin{aligned} L(x, y, 1) &= f(x, y) + h(x, y) \\ &= x^2 - y + (1 - x^2 - y^2) \\ &= 1 - y - y^2 \end{aligned}$$

we conclude that $L(x, y, 1)$ is a concave function of (x, y). So $(\pm\sqrt{3}/2, -1/2)$ are maximizers of f on D and solve the desired optimization problem.

Exercise 35. Use Theorem 6.22 to solve the following optimization problems:

(1)

$$\begin{aligned} \text{maximize } & x^2 + 2y \\ \text{subject to } & x^2 + y^2 \leq 5 \\ & y \geq 0 \end{aligned}$$

(2)

$$\begin{aligned} \text{maximize } & \frac{1}{2}x - y \\ \text{subject to } & x + e^{-x} + z^2 \leq y \\ & x \geq 0 \end{aligned}$$

(3)

$$\begin{aligned} \text{minimize } & 2x^2 + 3y^2 \\ \text{subject to } & x + 2y \leq 11 \\ & x \geq 0 \\ & y \geq 0 \end{aligned}$$

Exercise 36. A firm has L units of labour available and produces 3 goods whose values per unit of output are a, b and c, respectively. Producing x, y and z units of the goods requires αx^2, βy^2 and γz^2 units of labour, respectively. Here $a, b, c, \alpha, \beta, \gamma > 0$. Determine the number of units of the goods that maximize the value of the output that can be produced with no more than L units of labour.

6.6. **Mixed constrains.** Let f, $g_1, \dots, g_m$ and $h_1, \dots, h_\ell$ be C^2 functions defined in some open set $U \subset \mathbb{R}^n$. We consider the following optimization problem with mixed constrains

$$\begin{aligned} \text{maximize (minimize) } & f(x) \\ \text{subject to } & g_1(x) = 0 \\ & \vdots \\ & g_m(x) = 0 \\ & h_1(x) \geq 0 \\ & \vdots \\ & h_\ell(x) \geq 0 \end{aligned}$$

There are m equality constraints and ℓ inequality constraints. As before, let D denote the set of points x which satisfy the equality and inequality constraints. Then we say h_i is an active constraint for $x \in D$ if $h_i(x) = 0$. Moreover, $x \in D$ satisfies the constraint qualification if all vectors $Dg_1(x), \dots, Dg_m(x)$ and $Dh_i(x)$, corresponding to those h_i which are active for x, are linearly independent.

Theorem 6.24 (Necessary condition)**.** *Suppose that $x^* \in D$ is a local optimal point of f on D and x^* satisfies the constraint qualification. Then there are $\lambda^* = (\lambda_1^*, \dots, \lambda_m^*) \in \mathbb{R}^m$ and $\mu^* = (\mu_1^*, \dots, \mu_\ell^*) \in \mathbb{R}^\ell$ such that*

(1) $Df(x^*) + \sum_{i=1}^m \lambda_i^* Dg_i(x^*) + \sum_{i=1}^\ell \mu_i^* Dh_i(x^*) = 0$
(2) $\mu_i^* h_i(x^*) = 0$ *for every* $i = 1, \dots, \ell$

This theorem generalizes the theorem of Kuhn-Tucker as it includes equality and inequality constraints. The procedure to find the candidates for local optimal points is the same as in the previous optimization problems. First, one finds all solutions (x^*, λ^*, μ^*) of the system

$$\begin{cases} Df(x) + \sum_{i=1}^m \lambda_i Dg_i(x) + \sum_{i=1}^\ell \mu_i Dh_i(x) = 0 \\ \mu_i h_i(x) = 0, \quad i = 1, \dots, \ell \\ g_i(x) = 0, \quad i = 1, \dots, m \\ h_i(x) \geq 0, \quad i = 1, \dots, \ell \end{cases}$$

Then we may apply the sufficient conditions of Theorem 6.17 and Theorem 6.22. See the following example.

Example 6.25. Consider the problem of finding the rectangle with maximum area and perimeter equal to 4. We can formalize the problem as follows

$$\begin{aligned} &\text{maximize } xy \\ &\text{subject to } x + y = 2 \\ &\qquad\qquad x \geq 0 \\ &\qquad\qquad y \geq 0 \end{aligned}$$

Of course, only solutions with positive x and y are meaningful. We have to solve the system

$$\begin{cases} y - \lambda + \mu_1 = 0 \\ x - \lambda + \mu_2 = 0 \\ \mu_1 x = 0 \\ \lambda_2 y = 0 \\ x + y = 2 \\ x \geq 0 \\ y \geq 0 \end{cases}$$

We have 4 cases, corresponding to μ_1 and μ_2 being or not equal to zero:

- Let $\mu_1 = \mu_2 = 0$. Then $\lambda = x = y$ and $\lambda = 1$.
- Let $\mu_1 = 0$ and $\mu_2 \neq 0$. Then $y = 0$ which implies that $x = 2$, $\lambda = 0$ and $\mu_2 = -2$.
- Let $\mu_1 \neq 0$ and $\mu_2 = 0$. Then $x = 0$ which implies that $y = 2$, $\lambda = 0$ and $\mu_1 = -2$.
- Let $\mu_1 \neq 0$ and $\mu_2 \neq 0$. Then $x = y = 0$. But $x + y = 2$ which is impossible.

So we have 3 solutions. However, only the solution $(x, y, \lambda, \mu_1, \mu_2) = (1, 1, 1, 0, 0)$ is meaningful. So, the square with sides equal 1 is the solution to the problem.

7. Scalar Differential equations

The most familiar differential equation is

$$\frac{dx}{dt} = ax$$

where $a \in \mathbb{R}$ is a constant and $x = x(t)$ is an unknown real-valued function of the real variable $t \in \mathbb{R}$. It is common to say that t represents **time**. In the left-hand-side of the equation we have $\frac{dx}{dt}$ which means the derivative of the function $x(t)$ with respect to t. We shall also use the equivalent notations x' and $\dot{x}$ to represent the derivative. The differential equation tell us that the derivative $x'(t)$ equals $ax(t)$ for every t. In order to solve the equation we have to find such function $x(t)$. A solution to the equation is

$$x(t) = ce^{at} \tag{7.1}$$

where $c \in \mathbb{R}$ is any constant. In fact,

$$x'(t) = ace^{at} = ax(t)$$

so it verifies the differential equation. Moreover, all solutions of the differential equation are of type (7.1). In fact, if $u(t)$ is any other solution of the differential equation, i.e., $\frac{du}{dt} = au$, then

$$\begin{aligned}\frac{d}{dt}(u(t)e^{-at}) &= u'(t)e^{-at} + u(t)(-ae^{-at}) \\ &= au(t)e^{-at} - au(t)e^{-at} \\ &= 0\end{aligned}$$

So $u(t)e^{-at}$ has to be equal to a constant c for all t, since it has zero derivative. This means that $u(t) = ce^{at}$, which shows that the differential equation has essentially a unique solution given by (7.1). The solution (7.1) is called a **general solution**.

Now we develop the theory of scalar ordinary differential equations (ODE). Let $F(t, x_0, \ldots, x_{n-1})$ be a function taking real values. A **scalar ODE** is an equation of the form

$$x^{(n)} = F(t, x, x', x'', \ldots, x^{(n-1)})$$

where $x^{(i)} = \frac{d^i x}{dt^i}$ denotes the i-th derivative of x. By convention $x^{(0)} = x$. The differential equation is called scalar because $x(t)$ takes real values (1 dimensional). Any function $x(t)$ that solves the equation is called a **solution**.

Now we introduce some terminology in order to classify the scalar ODEs. Whenever F does not depend on t, the scalar ODE is called **autonomous**. The highest order of the derivative in the equation is called the **order** of the ODE. For instance,

$$x' = ax, \quad x' + e^t = 2x, \quad x' = x(1-x) \tag{7.2}$$

are first order scalar ODEs and

$$x'' + x' + x = 1, \quad x'' + e^t x^2 = 0 \tag{7.3}$$

are second order scalar ODES. The 2nd equations in (7.2) and (7.3) are non-autonomous. All the others are autonomous.

A scalar ODE is called **linear** if it has the form

$$x^{(n)} + a_1(t)x^{(n-1)} + \cdots + a_n(t)x = b(t) \tag{7.4}$$

where $a_i(t)$ and $b(t)$ are real-valued functions of t. For instance, in examples (7.2) the 1st and 2nd equations are linear but the 3rd is non-linear. In examples (7.3), the 1st equation is linear but the 2nd non-linear.

When $b(t)$ is equal to zero in (7.4) we say that the linear scalar ODE is **homogeneous**.

Exercise 37. Classify the following scalar ODEs:

(1) $x' = -3x + 4 + e^t$
(2) $x'' + 4tx' - 3(1 - t^2)x = 0$
(3) $x' + 3tx = e^x$

7.1. **Initial value problem.** In many cases (and also applications) we will be interested in finding solutions to scalar first order ODEs,

$$x' = f(t, x)$$

which have a prescribed value x_0 at some instant of time t_0. The problem of finding such solution is called the **initial value problem**[12],

$$x' = f(t, x) \quad \text{with} \quad x(t_0) = x_0.$$

Example 7.1. In the simple ODE $x' = ax$ the general solution is $x(t) = ce^{at}$. Now consider the initial value problem

$$x' = ax, \quad x(0) = x_0$$

The initial condition $x(0) = x_0$ will set the value of the constant c. Indeed, taking the general solution we see that $x(0) = ce^0 = c$. So the only solution that satisfies the initial condition is $x(t) = x_0 e^{at}$. This is called a **particular solution**.

7.2. **First order linear ODEs.** Any first order linear ODE is of the form

$$x' + a(t)x = b(t) \tag{7.5}$$

where $a(t)$ and $b(t)$ are given functions. In the following we present a method to solve the differential equation based on the **integrating factor**,

$$\alpha(t) = e^{\int a(t)\, dt}$$

where $\int a(t)\, dt$ denotes a primitive of $a(t)$, i.e., any function $u(t)$ such that $u'(t) = \alpha(t)$ for every t.

[12] IVP for short

Multiplying the integrating factor on both sides of the equation (7.5) we get

$$\alpha(t)x'(t) + a(t)\alpha(t)x(t) = \alpha(t)b(t)$$

Now notice that $\alpha'(t) = a(t)\alpha(t)$. Then

$$\alpha(t)x'(t) + \alpha'(t)x(t) = \alpha(t)b(t)$$

But $(\alpha(t)x(t))' = \alpha(t)x'(t) + \alpha'(t)x(t)$. Thus $(\alpha(t)x(t))' = \alpha(t)b(t)$ which implies that

$$x(t) = \frac{1}{\alpha(t)}\left(\int \alpha(t)b(t)\,dt + c\right) \tag{7.6}$$

where $c \in \mathbb{R}$ is any constant.

Example 7.2. Consider the autonomous first order linear ODE

$$x' + ax = b$$

where a and b are constants. Clearly, if $a = 0$, then the solution is

$$x(t) = bt + c$$

where c is a constant. Now suppose that $a \neq 0$. Then the integrating factor is

$$\alpha(t) = e^{\int a\,dt} = e^{at}$$

Applying formula (7.6) we get

$$\begin{aligned} x(t) &= e^{-at}\left(\int e^{at} b\,dt + c\right) \\ &= e^{-at}\left(\frac{b}{a}e^{at} + c\right) \\ &= \frac{b}{a} + e^{-at}c \end{aligned}$$

So, the general solution is

$$x(t) = \begin{cases} bt + c & a = 0 \\ \frac{b}{a} + e^{-at}c & a \neq 0 \end{cases}$$

If we consider the IVP

$$x' + ax = b, \quad x(0) = x_0$$

then, the initial condition $x(0) = x_0$ sets the value of the constant $c = x_0$ if $a = 0$, and $c = x_0 - b/a$ otherwise. So, the solution is

$$x(t) = \begin{cases} bt + x_0 & a = 0 \\ \frac{b}{a} + e^{-at}\left(x_0 - \frac{b}{a}\right) & a \neq 0 \end{cases}$$

Exercise 38. Find the general solution of

(1) $x' + 2x = 8$
(2) $x' + 3x = e^t$
(3) $x' + 2tx = e^{-t^2}$

(4) $x' + 2tx = 4t$

Exercise 39. For each 1st order linear ODE of Exercise 38 solve the IVP with the initial condition

(1) $x(0) = 0$
(2) $x(0) = -1$
(3) $x(0) = 1$
(4) $x(0) = -2$

Exercise 40. Consider the following model of economic growth in a developing country,

$$X(t) = \sigma K(t), \quad K'(t) = \alpha X(t) + H(t)$$

where $X(t)$ is the total domestic product per year, $K(t)$ the capital stock, $H(t)$ the net inflow of foreign investment per year, all measured at time instant t. Assume that $H(t) = H_0 e^{\mu t}$.

(1) Derive a differential equation for $K(t)$ and find the solution given that $K(0) = K_0$.
(2) If the size of the population $N(t) = N_0 e^{\rho t}$, compute $x(t) = X(t)/N(t)$ which is the domestic product per capita.
(3) Assuming that $\mu = \rho$ and $\rho > \alpha\sigma$ compute $\lim_{t\to+\infty} x(t)$.

7.3. **Separation of variables.** In the following we shall present a method to solve scalar ODEs whose variables x and t can be **separated**. A scalar ODE has separated variables if it can be written in the form

$$x' = g(t)f(x) \tag{7.7}$$

Clearly, any autonomous scalar ODE

$$x' = f(x)$$

has separated variables. Other examples of scalar ODEs with separated variables are

$$x' = -2tx^2, \quad x'x = e^{x+t}\sqrt{1+t^2}$$

However, not every scalar ODE has separated values. For instance,

$$x' = t^2 + x, \quad x' = tx + 1$$

have no separated variables.

Let us now consider the IVP

$$x' = g(t)f(x), \quad x(t_0) = x_0.$$

If $f(x_0) = 0$, then $x(t) = x_0$ solves the IVP. Indeed,

$$x'(t) = (x_0)' = 0 = g(t)f(x_0) = g(t)f(x).$$

So suppose that $f(x_0) \neq 0$. Then $f(x) \neq 0$ for every x close to x_0. Thus, we may write the equation (7.7) as

$$\frac{x'(t)}{f(x(t))} = g(t)$$

for every t close to t_0. Integrating both sides of the equation we get

$$\int_{t_0}^{t} \frac{x'(s)}{f(x(s))}\, ds = \int_{t_0}^{t} g(s)\, ds$$

Making the variable substitution $u = x(s)$ we get $du = x'(s)\, ds$ and

$$\int_{x_0}^{x(t)} \frac{1}{f(u)}\, du = \int_{t_0}^{t} g(s)\, ds$$

The previous equation gives a way to compute the solution of the IVP. Define the functions

$$F(z) = \int_{x_0}^{z} \frac{1}{f(u)}\, du \quad \text{and} \quad G(t) = \int_{t_0}^{t} g(s)\, ds$$

We have the following method to compute $x(t)$:

(1) Find a primitive of $1/f(u)$ and used it to calculate the function $F(z)$
(2) Find a primitive of $g(s)$ and used it to calculate the function $G(t)$
(3) Solve the equation $F(x(t)) = G(t)$ to find $x(t)$.

Example 7.3. Consider the logistic equation

$$x' = \mu x \left(1 - \frac{x}{K}\right), \quad x(0) = x_0$$

where $\mu, K > 0$ and $0 < x_0 < K$. The ODE is autonomous, so it has separated variables. In this case $g(t) = \mu$ and $f(x) = x(1 - x/K)$. To solve this IVP we apply the method above.

(1) First we compute a primitive of $1/f(u) = \frac{1}{u(1-u/K)}$. Writing[13],

$$\frac{1}{u(1-u/K)} = \frac{1}{u} + \frac{1}{K-u}$$

[13] In general, the following fraction decomposition is very useful

$$\frac{1}{(x-a)(x-b)} = \frac{1}{a-b}\frac{1}{x-a} + \frac{1}{b-a}\frac{1}{x-b}, \quad a \neq b$$

we see that

$$\begin{aligned}F(z) &= \int_{x_0}^{z} \frac{1}{u(1-u/K)}\,du \\ &= \int_{x_0}^{z} \frac{1}{u}\,du + \int_{x_0}^{z} \frac{1}{K-u}\,du \\ &= \log u\Big|_{x_0}^{z} + (-\log(K-u))\Big|_{x_0}^{z} \\ &= \log z - \log x_0 - \log(K-z) + \log(K-x_0) \\ &= \log\left(\frac{(K-x_0)z}{x_0(K-z)}\right)\end{aligned}$$

(2) A primitive of $g(s) = \mu$ is μs, hence

$$G(t) = \int_0^t g(s)\,ds = \mu s\Big|_0^t = \mu t$$

(3) Now we solve $F(x(t)) = G(t)$, that is

$$\log\left(\frac{(K-x_0)x(t)}{x_0(K-x(t))}\right) = \mu t$$

Taking exponential we get

$$\frac{(K-x_0)x(t)}{x_0(K-x(t))} = e^{\mu t}$$

So

$$x(t) = \frac{K}{1 + \frac{K-x_0}{x_0}e^{-\mu t}}$$

Example 7.4. Consider the IVP

$$x' = -2tx^2, \quad x(1) = -1$$

The equation has separated variables with $g(t) = -2t$ and $f(x) = x^2$. Applying the method we get

(1)

$$F(z) = \int_{-1}^{z} \frac{1}{u^2}\,du = -\frac{1}{u}\Big|_{-1}^{z} = -\left(1 + \frac{1}{z}\right)$$

(2)

$$G(t) = \int_1^t (-2s)\,ds = -s^2\Big|_1^t = -t^2 + 1$$

(3) Solving

$$-\left(1 + \frac{1}{x(t)}\right) = -t^2 + 1$$

we get

$$x(t) = \frac{1}{t^2 - 2}$$

Exercise 41. Find the solutions of the following IVP

(1) $tx' = (1-t)x$ with $x(1) = 1/e$
(2) $x' = t/x$ with $x(\sqrt{2}) = 1$
(3) $x' = (x-1)(x+1)$ with $x(0) = 0$

7.4. **Qualitative theory of scalar ODEs.** Given a function f taking real values we are interested in the qualitative behaviour of solution to autonomous scalar ODEs

$$x' = f(x), \quad x(0) = x_0 \tag{7.8}$$

Thanks to the following theorem we known that the IVP above has a solution.

Theorem 7.5. *If f is of class C^1, then* (7.8) *has a solution, i.e., there is $\varepsilon > 0$ and a function $x : [-\varepsilon, \varepsilon] \to \mathbb{R}$ such that $x'(t) = f(x(t))$ and $x(0) = x_0$. Moreover, the solution is unique.*

The uniqueness in the theorem means that if $x(t)$ and $y(t)$ are two solutions of the IVP (satisfying the same initial condition), then $x(t) = y(t)$ for every t where both solutions are defined. So different initial conditions give different solutions of $x' = f(x)$. The **maximal interval of existence** of $x(t)$, which we denote by I_{x_0}, is the largest time interval where the solution $x(t)$ is defined. To stress the dependence of $x(t)$ on the initial condition $x(0) = x_0$ we write $x(t; x_0)$ instead.

Example 7.6. Consider the ODE $x' = -x$. The solution is $x(t; x_0) = e^{-t}x_0$ which can be computed as in Example 7.2. The maximal interval of existence is $I_{x_0} = \mathbb{R} =]-\infty, +\infty[$.

Example 7.7. Consider the ODE $x' = x^2$. Suppose that $x_0 > 0$. The solution is $x(t; x_0) = x_0/(1 - x_0\, t)$ which can be computed using the method of separation of variables. The maximal interval of existence is $I_{x_0} = \mathbb{R} =]-\infty, 1/x_0[$.

Exercise 42. Determine the maximal interval of existence for the solutions of the following ODEs:

(1) $x' = e^{-x}$
(2) $x' = \frac{1}{2x}$

From the uniqueness of solutions we derive the following properties of the solutions:

(1) $x(t; x_0)$ is strictly[14] increasing in x_0
(2) $x(t; x_0)$ is monotone (increasing or decreasing) in t

Among all solutions there is one which plays a special role.

Definition 7.1. We say that $x^* \in \mathbb{R}$ is an **equilibrium point** of f if $f(x^*) = 0$.

[14] if $x_0 < y_0$ then $x(t; x_0) < x(t; y_0)$

Clearly, if x^* is an equilibrium point of f then $x(t; x^*) = x^*$. Indeed, $x'(t) = (x^*)' = 0 = f(x^*) = f(x(t))$. The solution is called **equilibrium or stationary** solution.

The importance of equilibrium points is that they attract all bounded solutions.

Proposition 7.8. *If $x(t; x_0)$ is bounded as a function of t, then the maximal interval of existence I_{x_0} where the solution is defined is unbounded and $x(t; x_0)$ converges to an equilibrium point either as $t \to +\infty$ or as $t \to -\infty$.*

Equilibrium points can be classified according to the local behaviour of nearby solutions.

Definition 7.2. We say that an equilibrium point x_0 is **stable** if for every x_0 close to x^*, the solution $x(t; x_0)$ stays close to x^* for every $t \geq 0$. If in addition, $\lim_{t\to+\infty} x(t; x_0) = x^*$ for every x_0 close to x^*, then we call x^* **asymptotically stable**. An equilibrium which is not stable is called **unstable**.

Example 7.9. Let $f(x) = -x$. The solutions of $x' = f(x)$ are

$$x(t; x_0) = e^{-t} x_0$$

Clearly, f has a single equilibrium $x^* = 0$. The corresponding equilibrium solution is $x(t; 0) = 0$. It is clear that $\lim_{t\to+\infty} x(t; x_0) = 0$ for every $x_0 \in \mathbb{R}$. Thus $x^* = 0$ is an asymptotically stable equilibrium point.

The following criterion is useful to determine if a given equilibrium point is asymptotically stable.

Theorem 7.10. *Suppose that f is of class C^1 and x^* is an equilibrium point of f. Then*

(1) *x^* is asymptotically stable if $f'(x^*) < 0$*
(2) *x^* is unstable if $f'(x^*) > 0$.*

Example 7.11. Consider the equation

$$x' = x(1 - x^2)$$

Clearly, there are 3 equilibrium points: $0, \pm 1$. Since $f'(x) = 1 - 3x^2$ we get $f'(0) = 1$ and $f'(\pm 1) = -2$. Thus 0 is unstable and ± 1 are asymptotically stable.

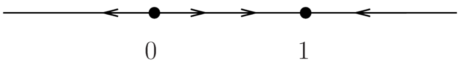
Example 7.12. For the equation $x' = x^2$ we have a single equilibrium at 0. Since $f'(0) = 0$ we cannot apply the criterion in Theorem 7.10. But a closer analysis shows that if $x(t)$ is close to 0 and $x(t) < 0$ then $x'(t) = (x^(t))^2 > 0$. So the solution tends to increase toward 0, i.e., $\lim_{t\to+\infty} x(t) = 0$. Otherwise, if $x(t) > 0$ and close to 0 then $x'(t) = (x(t))^2 > 0$, so $\lim_{t\to+\infty} x(t) = +\infty$. Hence, 0 is unstable.

A **phase portrait** is a geometric representation of the solutions of the ODE. In scalar ODEs, the phase portrait is 1-dimensional. On the x-axis a set of initial conditions is represented by a different curve (with arrows), or point (in the case of equilibrium points). Phase portraits are very useful in studying the qualitative behaviour of solutions. They reveal crucial information such as stable/unstable equilibrium points and the limits of solutions as $t \to \pm\infty$.

Example 7.13. Consider the ODE

$$x' = x(1-x)$$

There are two equilibrium points $x = 0$ and $x = 1$. The first is unstable and the second is asymptotically stable. The phase portrait is



Exercise 43. Determine the phase portrait of the follows ODEs and classify the equilibrium points.

(1) $x' = ax$, with $a \neq 0$
(2) $x' = x - x^3$
(3) $x' = b + x$ with $b \in \mathbb{R}$
(4) $x' = (x+1)(x+2)$
(5) $x' = -x + x^3 + \lambda$ with $\lambda \in \mathbb{R}$
(6) $x' = 1 - \sin x$

8. Planar Differential Equations

Consider a 2nd order linear ODE with constant coefficients

$$x'' + ax' + bx = 0$$

If $y = x'$, then $y' = x'' = -bx - ax' = -bx - ay$. Thus we get a system of two differential equations

$$\begin{aligned} x' &= y \\ y' &= -bx - ay \end{aligned}$$

We can write these equations using matrix notation. Let,

$$X(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} 0 & 1 \\ -b & -a \end{pmatrix}.$$

Then, the above system of scalar ODEs is equivalent to a **planar linear ODE**

$$X' = AX$$

Our goal in this section is to solve this type of differential equations. As in the scalar case, for planar ODEs is common to add an initial condition $X(0) = X_0$ obtaining an IVP,

$$X' = AX, \quad X(0) = X_0 = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \tag{8.1}$$

where

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

For certain matrices, such as A being diagonal ($b = c = 0$), the above IVP can be easily solved. Indeed, let $A = \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}$. Then $X' = AX$ is

$$\begin{aligned} x' &= ax \\ y' &= dy \end{aligned}$$

So the solution is $x(t) = e^{at}x_0$ and $y(t) = e^{bt}y_0$ because both equations are detached. Even if A contains a non-zero element in the top right conner, we can first find a solution for the y-equation and then solve the x-equation.

Exercise 44. Find the solution to the IVP assuming that

$$A = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$$

For a general matrix A the equations of the corresponding planar ODE are not detached, so the above simple procedure cannot be applied. However, if A can be transformed into a simpler form, then we may have a chance to solve our problem. Next, we will show how to transform A into a form that we can solve the IVP.

8.1. **Change of variables.** Consider the change of variables according to the relation

$$X(t) = PY(t) \tag{8.2}$$

where P is a 2-by-2 matrix (not depending on t) that has an inverse, i.e., $\det P \neq 0$[15] We can easily derive the planar ODE for the new variable Y. Since $Y(t) = P^{-1}X(t)$ we get

$$Y'(t) = P^{-1}X'(t) = P^{-1}AX(t) = P^{-1}APY(t)$$

[15] A matrix $P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$ has an inverse P^{-1} if and only if $\det(P) = p_{11}p_{22} - p_{12}p_{21} \neq 0$. The inverse is characterized by $PP^{-1} = P^{-1}P = I$ where I denotes the 2-by-2 identity matrix, i.e., $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Here is a simple formula for computing the inverse

$$P^{-1} = \frac{1}{\det(P)} \begin{pmatrix} p_{22} & -p_{12} \\ -p_{21} & p_{11} \end{pmatrix}$$

Let

$$J = P^{-1}AP$$

Then the IVP (8.1) is equivalent to

$$Y' = JY, \quad Y(0) = Y_0$$

where $Y_0 = P^{-1}X_0$.

Therefore, if J is a matrix as in Exercise 44, then we can solve the IVP for the variable Y and then use the variable relation (8.2) to obtain the solution for the original IVP (8.1). Our problem now is to find a matrix P which gives a matrix J as simple as possible.

8.2. **Eigenvalues and eigenvectors.** Given a **non-zero** vector $v \in \mathbb{R}^2$ we say that v is an **eigenvector** of A if

$$Av = \lambda v$$

for some scalar λ. The constant λ is called an **eigenvalue** of A. The pair (λ, v) is called an **eigenpair**. Clearly, an eigenpair (λ, v) satisfies $(A - \lambda I)v = 0$. Since $v \neq 0$, this means that the matrix $A - \lambda I$ is not invertible. Therefore, the eigenvalues of A are characterized by the following equation

$$\det(A - \lambda I) = \lambda^2 - \operatorname{tr}(A)\lambda + \det(A) = 0$$

where $\operatorname{tr}(A)$ is the **trace** of A, i.e., $\operatorname{tr}(A) = a + d$, the sum of the diagonal of A. This equation is quadratic in λ and can be solved using the quadratic formula

$$\lambda = \frac{\operatorname{tr}(A)}{2} \pm \sqrt{\left(\frac{\operatorname{tr}(A)}{2}\right)^2 - \det(A)}$$

Since there is a choice in the sign, every 2-by-2 matrix A has two eigenvalues λ_1 and λ_2 given by the formula above. Depending on the values of $\operatorname{tr}(A)$ and $\det(A)$ we distinguish 3 cases:

- **(I):** *Real and distinct eigenvalues*: $\lambda_1 \neq \lambda_2$ and $\lambda_1, \lambda_2 \in \mathbb{R}$. This happens then $(\operatorname{tr}(A)/2)^2 > \det(A)$.
- **(II):** *Equal eigenvalues*: $\lambda_1 = \lambda_2$. This happens when $(\operatorname{tr}(A)/2)^2 = \det(A)$.
- **(III):** *Complex conjugate eigenvalues*: $\lambda_1 = \alpha + i\beta$ and $\lambda_2 = \alpha - i\beta$ where $\alpha, \beta \in \mathbb{R}$. This happens when $(\operatorname{tr}(A)/2)^2 < \det(A)$.

In the following we show how to compute the eigenvectors of A for each of the above cases.

8.3. **Matrix P.** Depending on the type of eigenvalues (according to the previous cases) we have the following algorithm to compute the eigenvectors of A.

(I): *Real and distinct eigenvalues*: In this case we solve the following equations to find eigenvectors v_1 and v_2,

$$Av_1 = \lambda_1 v_1 \quad \text{and} \quad Av_2 = \lambda_2 v_2.$$

Each equation is solvable but has infinitely many solutions, i.e., if v_1 solves the 1st equation, then αv_1 with $\alpha \in \mathbb{R}$ is also a solution of the same equation. It is common to choose non-zero solutions v_1 and v_2 that have the simplest possible expression. Then we define the matrix P as having in the 1st column v_1 and v_2 in the 2nd column, i.e.,

$$P = (v_1|v_2)$$

Because v_1 and v_2 solve the equations above, we have

$$AP = PJ \quad \text{where} \quad J = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

(II): *Equal eigenvalues*: Let λ denote the single eigenvalue of A. We have two cases:

(1) If A is diagonal, i.e., $A = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$, then

$$v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad v_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

are eigenvectors of A. So we set $P = (v_1|v_2)$, which equals the identity matrix. Clearly, $AP = PJ$ where $J = A$.

(2) If A is not diagonal, then we find a 1st eigenvector v_1 by solving

$$Av_1 = \lambda v_1$$

We find a 2nd eigenvector v_2 by solving

$$(A - \lambda I)v_2 = v_1$$

Then we define $P = (v_1|v_2)$. A simple computation shows that

$$AP = PJ \quad \text{where} \quad J = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$$

(III): *Complex conjugate eigenvalues*: Let $\lambda_1 = \alpha + i\beta$. We solve the equation

$$Av = (\alpha + i\beta)v$$

As before, this equation has infinitely many solutions. However, because λ_1 is complex, v will also be complex, i.e., we can separate v into real and imaginary parts $v = v_1 + i\, v_2$ where $v_1, v_2 \in \mathbb{R}^2$. Then we set $P = (v_1|v_2)$. As before, a simple computation shows that

$$AP = PJ \quad \text{where} \quad J = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}$$

Example 8.1. We consider 3 examples:

(I): Suppose that

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

Since $\operatorname{tr}(A) = 0$ and $\det(A) = -4$, the eigenvalues are

$$\lambda = \frac{\operatorname{tr}(A)}{2} \pm \sqrt{\left(\frac{\operatorname{tr}(A)}{2}\right)^2 - \det(A)} = \pm\sqrt{4} = \pm 2$$

So $\lambda_1 = 2$ and $\lambda_2 = -2$. First we find an eigenvector v_1. The equation $Av_1 = 2v_1$ is equivalent to the system

$$\begin{cases} x + 3y = 2x \\ x - y = 2y \end{cases}$$

Since both equations are equivalent, we take the 1st and deduce that $x = 3y$. So $(x, y) = (3y, y) = y(3, 1)$. This gives $v_1 = (3, 1)$. Similarly, we find v_2 by solving the equation $Av_2 = -2v_2$. An eigenvector is $v_2 = (-1, 1)$. So

$$P = \begin{pmatrix} 3 & -1 \\ 1 & 1 \end{pmatrix}$$

(II): Suppose that

$$A = \begin{pmatrix} 3 & 0 \\ -2 & 3 \end{pmatrix}$$

Since $\operatorname{tr}(A) = 6$ and $\det(A) = 9$, the eigenvalues are

$$\lambda = \frac{\operatorname{tr}(A)}{2} \pm \sqrt{\left(\frac{\operatorname{tr}(A)}{2}\right)^2 - \det(A)} = 3$$

Because A is not a diagonal matrix we have to find v_1 such that $Av_1 = 3v_1$. So we solve the system

$$\begin{cases} 3x = 3x \\ -2x + 3y = 3y \end{cases}$$

We conclude that $x = 0$ and $y \in \mathbb{R}$. Thus $(x, y) = (0, y) = y(0, 1)$. This gives $v_1 = (0, 1)$. To find v_2 we solve the equation $(A - 3I)v_2 = v_1$, which is equivalent to solving the system

$$\begin{cases} 0 = 0 \\ -2x = 1 \end{cases}$$

This gives $(x, y) = (-1/2, y) = (-1/2, 0) + y(0, 1)$. So $v_2 = (-1/2, 0)$ since $(0, 1)$ is already an eigenvector of A. Thus,

$$P = \begin{pmatrix} 0 & -1/2 \\ 1 & 0 \end{pmatrix}$$

(III): Suppose that

$$A = \begin{pmatrix} 2 & -1 \\ 5 & 0 \end{pmatrix}$$

Since $\mathrm{tr}(A) = 2$ and $\det(A) = 5$, the eigenvalues are

$$\lambda = \frac{\mathrm{tr}(A)}{2} \pm \sqrt{\left(\frac{\mathrm{tr}(A)}{2}\right)^2 - \det(A)} = 1 \pm 2i$$

To find v_1 and v_2 we solve the system

$$\begin{cases} 2x - y = (1+2i)x \\ 5x = (1+2i)y \end{cases}$$

The equations are equivalent. We pick one, say the 2nd equation. It gives $x = \frac{1+2i}{5}y$. Separating into real and imaginary parts we get

$$(x, y) = \left(\frac{1+2i}{5}y, y\right) = y\left(\frac{1}{5}, 1\right) + y\left(\frac{2}{5}, 0\right)i$$

So, $v_1 = (1/5, 1)$ and $v_2 = (2/5, 0)$. Thus

$$P = \begin{pmatrix} 1/5 & 2/5 \\ 1 & 0 \end{pmatrix}$$

8.4. **Jordan normal forms.** In any of the above 3 cases, the matrix $P = (v_1|v_2)$ is always invertible, i.e., $\det(P) \neq 0$. We may write

$$J = P^{-1}AP$$

where J is a matrix belonging to one of the following types:

$$(i)\ \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \qquad \text{(ii)}\ \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \qquad \text{(iii)}\ \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}$$

where $\lambda_1, \lambda_2, \lambda, \alpha, \beta \in \mathbb{R}$. These 3 types of matrices are called **Jordan normal forms**. To summarize the discussion done so far we formulate the following theorem.

Theorem 8.2 (Jordan Normal Form). *Given any 2-by-2 matrix A, there is an invertible matrix P (consisting of eigenvectors of A) such that $J = P^{-1}AP$ is a Jordan normal form.*

The matrix P is computed using the algorithm described in the previous subsection. Notice that J and A have the same eigenvalues. Indeed, this follows from the fact

$$\begin{aligned} \det(J - \lambda I) &= \det(P^{-1}AP - \lambda I) \\ &= \det(P^{-1}(A - \lambda I)P) \\ &= \det(P^{-1})\det(A - \lambda I)\det(P) \\ &= \det(A - \lambda I) \end{aligned}$$

Exercise 45. Find the matrix P and determine the Jordan normal form for the following matrices

(1) $A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$

(2) $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$

(3) $A = \begin{pmatrix} 1 & 1 \\ -1 & 3 \end{pmatrix}$

(4) $A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$

(5) $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$

(6) $A = \begin{pmatrix} 0 & 4 \\ -5 & 4 \end{pmatrix}$

8.5. **Solution of IVP in Jordan normal form.** Consider the IVP

$$Y' = JY, \quad Y(0) = Y_0$$

where J is a Jordan normal form. Write $Y(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix}$ and $Y_0 = \begin{pmatrix} y_{10} \\ y_{20} \end{pmatrix}$ in coordinates. The general solution of this IVP is the following:

(i): Suppose that $J = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$. Then the IVP is equivalent to

$$\begin{cases} y_1' = \lambda_1 y_1, & y_1(0) = y_{10} \\ y_2' = \lambda_2 y_2, & y_2(0) = y_{20} \end{cases}$$

Thus $y_1(t) = e^{\lambda_1 t} y_{10}$ and $y_2(t) = e^{\lambda_2 t} y_{20}$ are the solutions to each scalar ODE. So the solution to the IVP is

$$Y(t) = \begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix} Y_0$$

(ii): Suppose that $J = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$. Then the IVP is equivalent to

$$\begin{cases} y_1' = \lambda y_1 + y_2, & y_1(0) = y_{10} \\ y_2' = \lambda y_2, & y_2(0) = y_{20} \end{cases}$$

The solution to the 2nd ODE is $y_2(t) = e^{\lambda t} y_{20}$. Substituting into the 1st ODE we get the following differential equation for y_1

$$y_1' = \lambda y_1 + e^{\lambda t} y_{20}$$

The solution can be found using (7.6),

$$y_1(t) = e^{\lambda t} y_{10} + t e^{\lambda t} y_{20}$$

Writing in matrix notation we have the solution to the IVP,

$$Y(t) = e^{\lambda t}\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} Y_0$$

(iii): Suppose that $J = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}$. Then

$$\begin{cases} y_1' = \alpha y_1 + \beta y_2, & y_1(0) = y_{10} \\ y_2' = -\beta y_1 + \alpha y_2, & y_2(0) = y_{20} \end{cases}$$

Let $z(t) = y_1(t) - i\, y_2(t)$. Then

$$\begin{aligned} z' &= y_1' - iy_2' \\ &= \alpha y_1 + \beta y_2 - i(-\beta y_1 + \alpha y_2) \\ &= \alpha(y_1 - iy_2) + \beta(y_2 + iy_1) \\ &= \alpha(y_1 - iy_2) + i\beta(y_1 - iy_2) \\ &= (\alpha + i\beta)(y_1 - iy_2) \\ &= (\alpha + i\beta)z \end{aligned}$$

The IVP

$$z' = (\alpha + i\beta)z, \quad z(0) = y_{10} - iy_{20}$$

has solution $z(t) = e^{(\alpha+i\beta)t}(y_{10} - i\, y_{20})$. Using Euler's formula[16] we get

$$\begin{aligned} z(t) &= e^{(\alpha+i\beta)t}(y_{10} - i\, y_{20}) \\ &= e^{\alpha t}(\cos(\beta t) + i\sin(\beta t))(y_{10} - i\, y_{20}) \\ &= e^{\alpha t}\left[\cos(\beta t)y_{10} + \sin(\beta t)y_{20} + i(\sin(\beta t)y_{10} - \cos(\beta t)y_{20})\right] \end{aligned}$$

where we have used the fact $i^2 = -1$. Because $z(t) = y_1(t) - i\, y_2(t)$ we conclude that

$$\begin{cases} y_1(t) = e^{\alpha t}\left(\cos(\beta t)y_{10} + \sin(\beta t)y_{20}\right) \\ y_2(t) = e^{\alpha t}\left(-\sin(\beta t)y_{10} + \cos(\beta t)y_{20}\right) \end{cases}$$

Writing in matrix notation we get

$$Y(t) = e^{\alpha t}\begin{pmatrix} \cos(\beta t) & \sin(\beta t) \\ -\sin(\beta t) & \cos(\beta t) \end{pmatrix} Y_0$$

8.6. **Solution of IVP.** Consider a general IVP

$$X' = AX, \quad X(0) = X_0 \tag{8.3}$$

where A is any given 2-by-2 matrix. By Theorem 8.2 there is an invertible matrix P such that $J = P^{-1}AP$ is in Jordan normal form.

[16] $e^{i\theta} = \cos(\theta) + i\sin(\theta)$

Changing variables $X = PY$, as explained in subsection 8.1, we transform the IVP (8.1) into

$$Y' = JY, \quad Y(0) = Y_0$$

where $Y_0 = P^{-1}X_0$. If $Y(t)$ is the solution to the IVP in Jordan normal form, then $X(t) = PY(t)$ is the solution to the IVP (8.3). The following theorem summarizes our discussion.

Theorem 8.3. *Let λ_1 and λ_2 denote the eigenvalues of A. Denote by P the matrix of eigenvectors of A as in Theorem 8.2. The following holds:*

(i): *If A is diagonal or $\lambda_1 \neq \lambda_2$ and real, then* (8.3) *has solution*

$$X(t) = P \begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix} P^{-1}X_0$$

(ii): *If $\lambda = \lambda_1 = \lambda_2$ and A is not diagonal, then* (8.3) *has solution*

$$X(t) = e^{\lambda t} P \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} P^{-1}X_0$$

(iii): *If $\lambda_1 = \alpha + i\beta$ and $\lambda_2 = \alpha - i\beta$, then* (8.3) *has solution*

$$X(t) = e^{\alpha t} P \begin{pmatrix} \cos(\beta t) & \sin(\beta t) \\ -\sin(\beta t) & \cos(\beta t) \end{pmatrix} P^{-1}X_0$$

Remark 8.4. The solution $X(t)$ of the IVP (8.3) can be written in a more compact way

$$X(t) = e^{At}X_0$$

where e^{At} is the exponential[17] matrix

$$e^{At} = \sum_{n=0}^{\infty} \frac{(At)^n}{n!}.$$

We can relate the exponential matrix e^{At} with e^{Jt} in the following way. Since $A = PJP^{-1}$ we conclude that

$$A^n = (PJP^{-1})(PJP^{-1}) \cdots (PJP^{-1}) = PJ^nP^{-1}$$

which gives

$$e^{At} = \sum_{n=0}^{\infty} \frac{(At)^n}{n!} = \sum_{n=0}^{\infty} \frac{P(Jt)^nP^{-1}}{n!} = P \sum_{n=0}^{\infty} \frac{(Jt)^n}{n!} P^{-1} = Pe^{Jt}P^{-1}.$$

Thus

$$X(t) = Pe^{Jt}P^{-1}X_0$$

[17] Notice the analogy with the series representation of the exponential function

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

where e^{Jt} is one of the three matrices in the theorem above, i.e.,

$$e^{Jt} = \begin{cases} \begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix}, & \text{if } A \text{ is of type (I)} \\ e^{\lambda t} \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, & \text{if } A \text{ is of type (II)} \\ e^{\alpha t} \begin{pmatrix} \cos(\beta t) & \sin(\beta t) \\ -\sin(\beta t) & \cos(\beta t) \end{pmatrix}, & \text{if } A \text{ is of type (III)} \end{cases} \tag{8.4}$$

Exercise 46. Find the solution of $X' = AX$ with $X(0) = X_0$ where

(1) $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

(2) $A = \begin{pmatrix} 1 & 1 \\ -1 & 0 \end{pmatrix}$

(3) $A = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$

(4) $A = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$

(5) $A = \begin{pmatrix} 1 & 1 \\ -1 & 3 \end{pmatrix}$

8.7. **2nd order scalar linear ODEs.** As we have discussed in the beginning of Section 8, any 2nd order linear homogeneous ODE with constant coefficients

$$x'' + ax' + bx = 0 \tag{8.5}$$

can be transformed into a 1st order planar linear ODE

$$X' = AX, \quad A = \begin{pmatrix} 0 & 1 \\ -b & -a \end{pmatrix}, \quad X(t) = \begin{pmatrix} x(t) \\ x'(t) \end{pmatrix}.$$

By Theorem 8.3, this system has the solution

$$X(t) = e^{At}X_0 = Pe^{Jt}P^{-1}X_0$$

where X_0 is a vector of initial conditions. The exponential matrix e^{Jt} is of the three types depending on the eigenvalues of A which can be obtained by solving the **characteristic equation**,

$$\lambda^2 + a\lambda + b = 0.$$

Notice the similarity between the characteristic equation and (8.5). Therefore, the solution $x(t)$ of (8.5) is obtained by a linear combination,

$$x(t) = c_1 w_1(t) + c_2 w_2(t), \quad c_1, c_1 \in \mathbb{R}$$

where $w_1(t)$ and $w_2(t)$ are the functions appearing in (8.4), i.e.,

(i): $w_1(t) = e^{\lambda_1 t}$ and $w_2(t) = e^{\lambda_2 t}$

(ii): $w_1(t) = e^{\lambda t}$ and $w_2(t) = te^{\lambda t}$

(iii): $w_1(t) = e^{\alpha t}\cos(\beta t)$ and $w_2(t) = e^{\alpha t}\sin(\beta t)$

Example 8.5. Suppose we want to solve the IPV

$$x'' - x' - 2x = 0, \quad x(0) = 1, \quad x'(0) = 0.$$

The characteristic equation is $\lambda^2 - \lambda - 2 = 0$, from which we obtain the eigenvalues $\lambda_1 = 2$ and $\lambda_2 = -1$. Thus, the solution of the IVP is of the form

$$x(t) = c_1 e^{2t} + c_2 e^{-t}.$$

Taking into account the initial conditions we determine the values of c_1 and c_2,

$$\begin{cases} x(0) = 1 \\ x'(0)) = 0 \end{cases} \Leftrightarrow \begin{cases} c_1 + c_2 = 1 \\ 2c_1 - c_2 = 0 \end{cases} \Leftrightarrow \begin{cases} c_1 + c_2 = 1 \\ 2c_1 - c_2 = 0 \end{cases} \Leftrightarrow \begin{cases} c_1 = \frac{1}{3} \\ c_2 = \frac{2}{3} \end{cases}$$

Hence,

$$x(t) = \frac{1}{3} e^{2t} + \frac{2}{3} e^{-t}$$

Exercise 47. Solve the IVPs:

(1) $x'' = 4x$, with $x(0) = 2$ and $x'(0) = -1$.
(2) $x'' + x = 0$ with $x(0) = 0$ and $x'(0) = 1$.
(3) $x'' - 2x' + x = 0$ with $x(0) = 1$ and $x'(0) = 1$.

8.8. **Solution of non-homogeneous IVP.** Consider the non-homogeneous IVP,

$$X'(t) = AX(t) + b(t), \quad X(0) = X_0$$

where $b(t)$ is a 2×1 vector-valued function depending on time. This problem has a unique solution $X(t)$. To derive the exact formula for the solution, we write $X(t) = e^{At} Z(t)$ and deduce a differential equation for $Z(t)$. Notice that

$$X'(t) = Ae^{At} Z(t) + e^{At} Z'(t) = AX(t) + e^{At} Z'(t).$$

But $X'(t) = AX(t) + b(t)$, so we get $b(t) = e^{At} Z'(t)$, which implies that $Z'(t) = e^{-At} b(t)$ since $(e^{At})^{-1} = e^{-At}$. By the fundamental theorem of calculus we get

$$Z(t) = X_0 + \int_0^t e^{-As} b(s)\, ds.$$

Concluding, the non-homogeneous IVP has the solution

$$X(t) = e^{At} \left(X_0 + \int_0^t e^{-As} b(s)\, ds \right).$$

Exercise 48. Solve the following IVPs

(1) $\begin{cases} x' = y + e^{-2t}, \\ y' = x + 1, \end{cases} \quad x(0) = 1, \, y(0) = 2$
(2) $x'' + x' - 6x = 2$ with $x(0) = -1$ and $x'(0) = 1$.

8.9. **Phase portraits.** As introduced in scalar ODEs, the phase portrait of a planar ODE is a geometric representation of the solutions. In the planar case, the phase portrait is 2-dimensional and two axis are required. On the (x, y)-plane, a set of initial conditions is represented by a different curve (with arrows), or point (in the case of equilibrium points). As pointed out before, phase portraits are very useful in studying the qualitative behaviour of solutions.

Among the solutions of $X' = AX$, the equilibria are the most simple. We say that $X^* \in \mathbb{R}^2$ is an **equilibrium point** if $AX^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. Clearly, the origin of the plane $X^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ is equilibrium point. The proof of following proposition is left as an exercise.

Proposition 8.6. *The origin is the unique equilibrium point if and only if* $\det(A) \neq 0$*. Moreover,*

1. *if the real part of the eigenvalues is negative, then the origin is asymptotically stable.*
2. *if the real part of the eigenvalues is positive, then the origin is unstable.*

Example 8.7. As an example, suppose that

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$

Then every point $X^* = \begin{pmatrix} 0 \\ y \end{pmatrix}$ with $y \in \mathbb{R}$ is an equilibrium point. So in this case there are infinitely many equilibrium points. Notice that $\det(A) = 0$.

In the following we sketch several phase portraits for each IVP in Jordan normal form,

$$X' = JX, \quad X(0) = X_0 = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \tag{8.6}$$

where J is of type (i)-(iii).

(i): Consider the Jordan normal form

$$J = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

The solution to the IVP is

$$x(t) = e^{\lambda_1 t} x_0$$
$$y(t) = e^{\lambda_2 t} y_0$$

Depending on the signs of the eigenvalues we have the following phase portraits depicted in Figures 1 and 2.

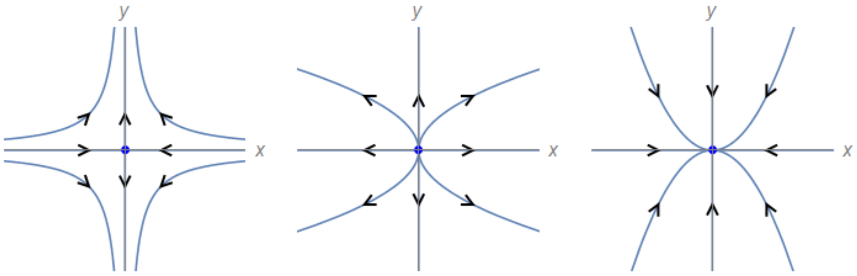


FIGURE 1. The left phase portrait has $\lambda_1 < 0 < \lambda_2$ and is called a **saddle**. The middle phase portrait has $0 < \lambda_2 < \lambda_1$ and is called a **source**. The right phase portrait has $\lambda_2 < \lambda_1 < 0$ and is called a **sink**.

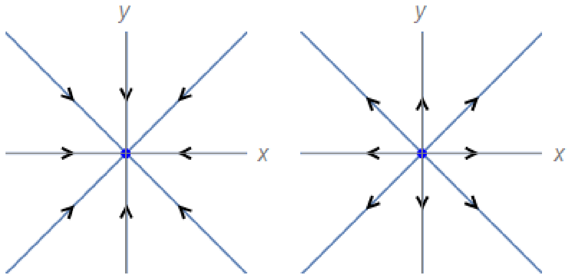


FIGURE 2. The left phase portrait has $\lambda_1 = \lambda_2 < 0$ and the right phase portrait has $\lambda_1 = \lambda_2 > 0$. The 1st is a sink and the 2nd a source.

(ii): Consider the Jordan normal form

$$J = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$$

The solution to the IVP is

$$\begin{aligned} x(t) &= e^{\lambda t} x_0 + t e^{\lambda t} y_0 \\ y(t) &= e^{\lambda t} y_0 \end{aligned}$$

Depending on the sign of λ we have the following phase portraits depicted in Figure 3.

(iii): Consider the Jordan normal form

$$J = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}$$

The solution to the IVP is

$$\begin{aligned} x(t) &= e^{\alpha t} \cos(\beta t) x_0 + e^{\alpha t} \sin(\beta t) y_0 \\ y(t) &= -e^{\alpha t} \sin(\beta t) x_0 + e^{\alpha t} \cos(\beta t) y_0 \end{aligned}$$

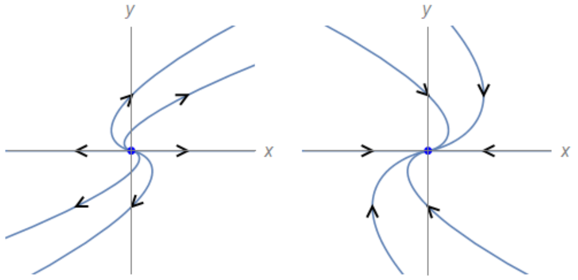


FIGURE 3. The left phase portrait has $\lambda > 0$ and right has $\lambda < 0$. The 1st is called an **unstable node** and the 2nd a **stable node**.

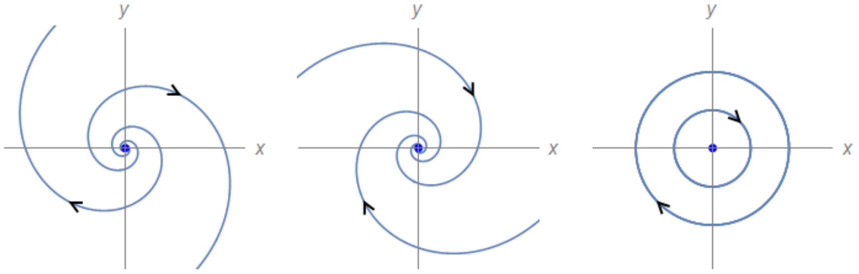


FIGURE 4. In all phase portraits $\beta > 0$. The left phase portrait has $\alpha > 0$ and is an **unstable focus**, the middle phase portrait has $\alpha < 0$ and is a **stable focus**, and finally the right phase portrait has $\alpha = 0$ and is a **center**. Notice that the center is stable but no asymptotically stable. For $\beta < 0$ the rotation around the origin becomes anticlockwise.

Depending on the sign of α and β we have the following phase portraits depicted in Figure 4.

8.10. **Transformation of the phase portrait.** If the matrix A is not in a Jordan normal form, then the phase portrait of the associated IVP can be obtained using the matrix P. Let us show how to sketch the phase portrait of a general IVP through an example. Consider the IVP

$$X' = AX, \quad X(0) = X_0$$

where

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

The matrix A is not in Jordan normal form. To find the normal form we find the eigenvalues of A,

$$\lambda_1 = 1, \quad \lambda_2 = -1$$

The Jordan normal form of A is

$$J = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Then we find the associated eigenvectors and construct the matrix P

$$P = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

In the Y-plane, the IVP is $Y' = JY, \quad Y(0) = Y_0$ and its phase portrait is depicted in Figure 5. To determine the phase portrait in the X-plane we see how P transforms the axis of the Y-plane. Using $X = PY$ we find that

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} = P \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 \\ -1 \end{pmatrix} = P \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Thus, the x-axis in the Y-plane is transformed into an axis spanned by the vector $(1,1)$ and the y-axis in the Y-plane is transformed into an axis spanned by the vector $(1,-1)$. In fact, it is not a coincidence that these are the eigenvectors of A. Then, the phase portrait is sketched accordingly, see Figure 5.

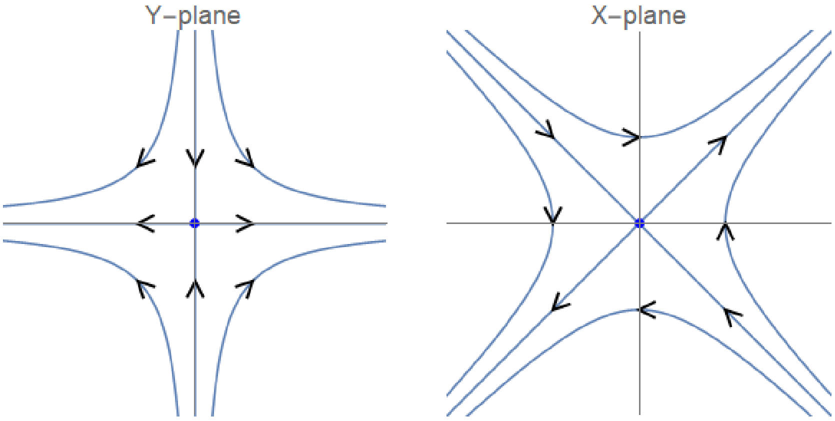


FIGURE 5. Phase portrait of the IVP in the Y and X variables, respectively.

Exercise 49. For each of the following planar ODEs $X' = AX$

$$(i)\ A = \begin{pmatrix} -8 & -5 \\ 10 & 7 \end{pmatrix} \qquad (ii)\ A = \begin{pmatrix} 1/2 & -1/2 \\ 0 & 1 \end{pmatrix}$$

$$(iii)\ A = \begin{pmatrix} -1 & 1 \\ -1 & -3 \end{pmatrix} \qquad (iv)\ A = \begin{pmatrix} 4 & 1 \\ -4 & 0 \end{pmatrix}$$

$$(v)\ A = \begin{pmatrix} 5 & 4 \\ -10 & -7 \end{pmatrix} \qquad (vi)\ A = \begin{pmatrix} -1 & -2 \\ 1 & 1 \end{pmatrix}$$

(1) Find the Jordan normal form of A
(2) Compute the associated matrix P
(3) Compute solution of the associated IVP
(4) Sketch the phase portrait

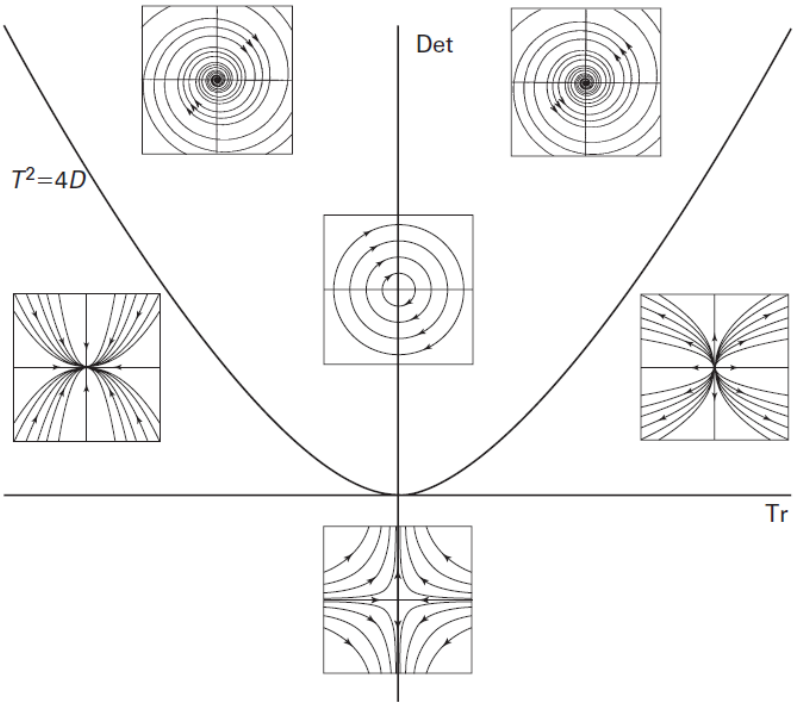


FIGURE 6. Trace-determinant plane.

8.11. **Trace-determinant plane.** We have seen that for

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

the eigenvalues are given by

$$\lambda = \frac{\operatorname{tr}(A)}{2} \pm \sqrt{\left(\frac{\operatorname{tr}(A)}{2}\right)^2 - \det(A)}$$

So knowing $\operatorname{tr}(A)$ and $\det(A)$ we can tell immediately which type of Jordan normal form and which phase portrait (up to a transformation by P) are associated with A. The parabola $\det(A) = \left(\frac{\operatorname{tr}(A)}{2}\right)^2$ in the $(\operatorname{tr}, \det)$-plane separates two types of Jordan normal forms:

- $\det(A) < \left(\frac{\operatorname{tr}(A)}{2}\right)^2$ gives real and distinct eigenvalues (type (i) Jordan normal form)

- $\det(A) > \left(\frac{\mathrm{tr}(A)}{2}\right)^2$ gives complex eigenvalues (type (iii) Jordan normal form)

If $\det(A) = \left(\frac{\mathrm{tr}(A)}{2}\right)^2$, then the eigenvalues are repeated and we may be in a presence of a type (ii) Jordan normal form or a type (i) with equal eigenvalues.

For each of the above cases we have the following phase portraits. Let us consider the case $\det(A) < \left(\frac{\mathrm{tr}(A)}{2}\right)^2$. The others are analysed similarly. This case breaks down in 3 sub-cases. In fact,

- if $\det(A) < 0$, then the eigenvalues have opposite sign, so we have a saddle.
- if $\det(A) > 0$ and $\mathrm{tr}(A) > 0$, then the eigenvalues are both positive, so we have a source.
- if $\det(A) > 0$ and $\mathrm{tr}(A) < 0$, then both eigenvalues are negative, so we have a sink.

All this analysis is summarized in Figure 6 which is called the **trace-determinant plane**. This plane helps to find the type of phase portrait and to decide about the stability of the equilibrium just by looking at the trace and determinant of A.

Exercise 50. Find the solution of the following 2nd order scalar ODEs,

(1) $x'' + bx = 0$ with $b > 0$ (harmonic oscillator)
(2) $x''+ax'+bx = 0$ with $a, b > 0$ (harmonic oscillator with friction)

For each case, discuss the phase portrait in the (x, x')-plane.

9. Calculus of Variations

To motivate the problems of calculus of variations we consider the Ramsey problem: *how much of its income should a nation save?* Consider an economy where $K(t)$ denotes the capital at time t, $C(t)$ the consumption, and $Y(t)$ the net national product. To simplify, we suppose that $Y(t)$ is a function of the capital $C(t)$ alone, i.e., $Y(t) = f(K(t))$ where f is strictly increasing and concave. Ramsey introduces an equation which relates the total output to the sum of consumption and investment

$$f(K(t)) = C(t) + \dot{K}(t).$$

If one wishes to have a high consumption, then investment $\dot{K}(t)$ will be small, yielding lower capital in the future, whence reducing the possibility of future consumption. Associated to this economy there is an utility function $U(C)$ which is strictly increasing and concave. So the problem is to find the optimal consumption $C(t)$ on a time interval $[0, T]$ which maximizes,

$$\int_0^T U(C(t))e^{-rt}\,dt$$

where $r \geq 0$ is a rate of a discounted factor. Usually, one imposes some initial and terminal conditions

$$K(0) = K_0 \quad \text{and} \quad K(T) = K_T.$$

The problem becomes

$$\max \int_0^T U(f(K(t)) - \dot{K}(t))e^{-rt}\,dt \quad \text{subject to} \quad K(0) = K_0,\ K(T) = K_T.$$

9.1. **Basic problem of calculus of variations.** Let $F \in C^2$ be a function of three variables $(t, x, \dot{x})$. The basic problem of calculus of variations is to find a function $x(t)$ of class C^1 which solves the problem

$$\max \int_{t_0}^{t_1} F(t, x(t), \dot{x}(t))\,dt \quad \text{subject to} \quad x(t_0) = x_0,\ x(t_1) = x_1.$$

Instead of maximize, one can defined the similar problem with minimize. To clarify, let us introduce two concepts: the set $\mathcal{A}$ and the functional $\mathcal{J}$. Firstly, $\mathcal{A}$ is the set of **admissible functions** defined by those functions which are of class C^1 and satisfy the initial and terminal condition, i.e.,

$$\mathcal{A} = \left\{x : [t_0, t_1] \to \mathbb{R} \colon x(t) \text{ is of class } C^1 \text{ and } x(t_0) = x_0,\ x(t_1) = x_1\right\}.$$

The elements of $\mathcal{A}$, which are functions, are called admissible functions.

Secondly, $\mathcal{J} : \mathcal{A} \to \mathbb{R}$ is the functional that to each admissible function $x(t)$ it associates the real number

$$\mathcal{J}(x) = \int_{t_0}^{t_1} F(t, x(t), \dot{x}(t))\,dt.$$

Hence, the basic problem of calculus of variations is an optimization problem in a function space $\mathcal{A}$ (not a subset of $\mathbb{R}^n$ as in classical optimization) whose goal is to find the optimal points (either maximizers or minimizers) of $\mathcal{J}$ on $\mathcal{A}$. Recall that $x^* \in \mathcal{A}$ is a maximizer (resp. minimizer) of $\mathcal{J}$ on $\mathcal{A}$ if $\mathcal{J}(x^*) \geq \mathcal{J}(x)$ (resp. $\mathcal{J}(x^*) \leq \mathcal{J}(x)$) for every $x \in \mathcal{A}$.

9.2. **Necessary conditions.** In the following we derive necessary conditions for $x^* \in \mathcal{A}$ to be an optimal point of $\mathcal{J}$ on $\mathcal{A}$. So, suppose that x^* is a maximizer[18], i.e., $\mathcal{J}(x^*) \geq \mathcal{J}(x)$ for every $x \in \mathcal{A}$.

Let $\mu(t)$ be a function of class C^1 with the boundary conditions $\mu(t_0) = \mu(t_1) = 0$. So, $x^*(t) + \alpha\mu(t)$ is an admissible function for every $\alpha \in \mathbb{R}$ (prove it!). Thus, $\mathcal{J}(x^*) \geq \mathcal{J}(x^*+\alpha\mu)$ for every $\alpha \in \mathbb{R}$. Defining $f(\alpha) := \mathcal{J}(x^* + \alpha\mu)$ we have that $\alpha = 0$ is maximizer of $f(\alpha)$. This implies that $\alpha = 0$ is a critical point of $f(\alpha)$, i.e., $f'(0) = 0$. Now we compute $f'(\alpha)$. First notice that

$$f(\alpha) = \int_{t_0}^{t_1} F(t, x^*(t) + \alpha\mu(t), \dot{x}^*(t) + \alpha\dot{\mu}(t))\, dt.$$

Taking the derivative we get,

$$\begin{aligned} f'(\alpha) &= \int_{t_0}^{t_1} \frac{\partial F}{\partial x}(t, x^*(t) + \alpha\mu(t), \dot{x}^*(t) + \alpha\dot{\mu}(t))\mu(t)\, dt \\ &\quad + \int_{t_0}^{t_1} \frac{\partial F}{\partial \dot{x}}(t, x^*(t) + \alpha\mu(t), \dot{x}^*(t) + \alpha\dot{\mu}(t))\dot{\mu}(t)\, dt. \end{aligned}$$

Integrating by parts the second integral in the previous equation we obtain,

$$\begin{aligned} \int_{t_0}^{t_1} \frac{\partial F}{\partial \dot{x}}(t, y(t), \dot{y}(t))\dot{\mu}(t)\, dt &= \left[\frac{\partial F}{\partial \dot{x}}(t, y(t), \dot{y}(t))\mu(t)\right]_{t_0}^{t_1} - \int_{t_0}^{t_1} \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{x}}(t, y(t), \dot{y}(t))\right)\mu(t) \\ &= -\int_{t_0}^{t_1} \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{x}}(t, y(t), \dot{y}(t))\right)\mu(t) \end{aligned}$$

where $y(t) = x^*(t) + \alpha\mu(t)$. Thus

$$f'(\alpha) = \int_{t_0}^{t_1} \left(\frac{\partial F}{\partial x}(t, y(t), \dot{y}(t)) - \frac{d}{dt}\frac{\partial F}{\partial \dot{x}}(t, y(t), \dot{y}(t))\right)\mu(t)\, dt.$$

Since $f'(0) = 0$, we get that

$$\int_{t_0}^{t_1} \left(\frac{\partial F}{\partial x}(t, x^*(t), \dot{x}^*(t)) - \frac{d}{dt}\frac{\partial F}{\partial \dot{x}}(t, x^*(t), \dot{x}^*(t))\right)\mu(t)\, dt = 0,$$

[18] The minimizer case is can be treated in a similar way.

which is valid for every C^1 function $\mu(t)$ vanishing at the boundary points. Therefore[19],

$$\frac{\partial F}{\partial x}(t, x^*(t), \dot{x}^*(t)) - \frac{d}{dt}\frac{\partial F}{\partial \dot{x}}(t, x^*(t), \dot{x}^*(t)) = 0.$$

We summarize this discussion in the following theorem.

Theorem 9.1 (Euler-Lagrange equation)**.** *If $x^* \in \mathcal{A}$ is an optimal point of $\mathcal{J}$ on $\mathcal{A}$, then it is a solution of the Euler-Lagrange equation,*

$$\frac{\partial F}{\partial x}(t, x(t), \dot{x}(t)) - \frac{d}{dt}\frac{\partial F}{\partial \dot{x}}(t, x(t), \dot{x}(t)) = 0.$$

Example 9.2. Suppose we want to steer the state $y(t)$ of an economy over the period $[0, T]$ to a desired level $\ell \in \mathbb{R}$ using a control (policy) $u(t) = c\dot{y}(t)$ where $c > 0$ is some constant. Because $u(t)$ is costly, our goal is to find $u(t)$ which minimizes

$$\int_0^T ((y(t) - \ell)^2 + (u(t))^2 \, dt$$

subject to the boundary conditions $y(0) = y_0$ and $y(T) = \ell$. Ideally, we would like $y(t)$ to approach ℓ as fast as possible and $u(t)$ to be very small. These requirements are both taken into account when we minimize the integral above. To solve this problem using calculus of variations we simply note that we aim to

$$\min \int_0^T ((x(t))^2 + c^2(\dot{x}(t))^2 \, dt \quad \text{subject to} \quad x(0) = y_0 - \ell,\, x(T) = 0$$

where $x(t) = y(t) - \ell$ and $\dot{x}(t) = \dot{y}(t) = \frac{1}{c}u(t)$. In this formulation we have $t_0 = 0$, $t_1 = T$, $x_0 = y_0 - \ell$, $x_1 = 0$ and

$$F(t, x, \dot{x}) = x^2 + c^2\dot{x}^2.$$

To determine the associated Euler-Lagrange equation we compute the derivatives

$$\frac{\partial F}{\partial x} = 2x \quad \text{and} \quad \frac{\partial F}{\partial \dot{x}} = 2c^2\dot{x}.$$

Taking the time derivative we get,

$$\frac{d}{dt}\frac{\partial F}{\partial \dot{x}} = 2c^2\ddot{x}.$$

Thus, the Euler-Lagrange equation is

$$x(t) - c^2\ddot{x}(t) = 0.$$

This equation is a 2nd order scalar linear ODE. Its characteristic equation is $1 - c^2\lambda^2 = 0$, which gives $\lambda_1 = 1/c$ and $\lambda_2 = -1/c$. We conclude

[19] Here we are using a lemma which says that if g is continuous and $\int_{t_0}^{t_1} g(t)\mu(t)\, dt = 0$ for every μ of class C^1 with $\mu(t_0) = \mu(t_1) = 0$, then $g(t) = 0$ for every $t \in [t_0, t_1]$.

that the associated planar system is of type (I). Therefore, the solution $x(t)$ has the form

$$x(t) = a_1 e^{t/c} + a_2 e^{-t/c}.$$

The coefficients a_1 and a_2 can be determined using the boundary conditions. Indeed,

$$\begin{cases} x(0) = y_0 - \ell \\ x(T)) = 0 \end{cases} \Leftrightarrow \begin{cases} a_1 + a_2 = y_0 - \ell \\ a_1 e^{T/c} + c_2 e^{-T/c} = 0 \end{cases} \Leftrightarrow \begin{cases} a_1 = -\frac{(y_0-\ell)e^{-T/c}}{e^{T/c}-e^{-T/c}} \\ a_2 = \frac{(y_0-\ell)e^{T/c}}{e^{T/c}-e^{-T/c}} \end{cases}$$

Thus,

$$x(t) = -\frac{(y_0-\ell)e^{-T/c}}{e^{T/c}-e^{-T/c}} e^{t/c} + \frac{(y_0-\ell)e^{T/c}}{e^{T/c}-e^{-T/c}} e^{-t/c}. \tag{9.1}$$

So, the Euler-Lagrange equation gives a candidate for the solution of our problem

$$u(t) = c\dot{x} = -\frac{(y_0-\ell)e^{-T/c}}{e^{T/c}-e^{-T/c}} e^{t/c} - \frac{(y_0-\ell)e^{T/c}}{e^{T/c}-e^{-T/c}} e^{-t/c}.$$

This is a candidate, because being a solution to the Euler-Lagrange equation is just a necessary condition, may not be sufficient. However, under convexity assumptions, the solution found using the Euler-Lagrange turns out to be the optimal point, i.e., the solution to our problem.

9.3. **Sufficient condition.**

Theorem 9.3. *Suppose that F is a convex (concave) function of the variables $(x, \dot{x})$. If $x^*(t)$ is admissible, i.e., $x^* \in \mathcal{A}$, and is a solution of the Euler-Lagrange equation, then x^* is a minimizer (maximizer) of $\mathcal{J}$ on $\mathcal{A}$.*

Proof. Check the bibliography. □

Example 9.4. In Example 9.2, the function $F(t, x, \dot{x}) = x^2 + c^2\dot{x}^2$ is convex. Hence the solution (9.1) of the Euler-Lagrange equation solves the problem.

Exercise 51. Consider the Ramsey problem where $f(K) = K$, $U(C) = 2\sqrt{C}$ and $r = 2$. Determine the optimal consumption function assuming that $K(0) = 1$ and $K(T) = 2$.

Exercise 52. Solve the following variational problems:

(1) $\max \int_0^1 (4xt - \dot{x}^2)\,dt$, $x(0) = 2$, $x(1) = 2/3$
(2) $\min \int_0^1 (t\dot{x} + \dot{x}^2)\,dt$, $x(0) = 1$, $x(1) = 0$
(3) $\min \int_0^1 (x^2 + 2tx\dot{x} + \dot{x}^2)\,dt$, $x(0) = 1$, $x(1) = 2$.

Exercise 53. Consider the planar curve $\gamma(x) = (x, f(x))$ with $f \in C^1$ that connects the points $A = (x_0, y_0)$ and $B = (x_1, y_1)$. The length of the curve γ is given by

$$L(\gamma) = \int_{x_0}^{x_1} \sqrt{1 + (f'(x))^2}\, dx$$

Determine the curve γ which minimizes the length between A and B. Formulate the problem as a variational problem and solve it using the calculus of variations.

9.4. **Variable terminal conditions.** In most economic applications the initial condition $x(t_0) = x_0$ is fixed because it represents the initial state of an economic system. However, the terminal condition might be unknown or simply less restrictive. In that case, we consider the variational problem of minimizing or maximizing

$$\int_{t_0}^{t_1} F(t, x(t), \dot{x}(t))\, dt \quad \text{subject to} \quad x(t_0) = x_0, \begin{cases} x(t_1) \text{ is free}, & (A) \\ x(t_1) \geq x_1, & (B) \end{cases}$$

Here we consider two different terminal conditions. The condition (A) means that there is no terminal condition, i.e., $x(t_1)$ can take any real value. The condition (B) imposes a lower bound on the terminal value $x(t_1)$.

Theorem 9.5. *Suppose that $x^*(t)$ is a C^1 function that solves the variational problem and satisfies the boundary conditions with either (A) or (B) terminal condition. Then $x^*(t)$ is a solution of the Euler-Lagrange equation*

$$\frac{\partial F}{\partial x}(t, x(t), \dot{x}(t)) - \frac{d}{dt}\frac{\partial F}{\partial \dot{x}}(t, x(t), \dot{x}(t)) = 0,$$

and satisfies the following ***transversality condition****:*

(A):

$$\frac{\partial F}{\partial \dot{x}}(t_1, x^*(t_1), \dot{x}^*(t_1)) = 0$$

(B):

$$\begin{aligned} &\frac{\partial F}{\partial \dot{x}}(t_1, x^*(t_1), \dot{x}^*(t_1)) \leq 0 \\ \textit{but} \quad &\frac{\partial F}{\partial \dot{x}}(t_1, x^*(t_1), \dot{x}^*(t_1)) = 0 \quad \textit{whenever } x^*(t_1) > x_1 \end{aligned}$$

Moreover, if $F(t, x, \dot{x})$ is a concave (convex) function of $(x, \dot{x})$ and an admissible $x(t)$ solves the Euler-Lagrange equation with the appropriate transversality condition, then $x(t)$ is the solution of the maximization (minimization) variational problem.

Proof. Check the bibliography. □

Example 9.6. Consider the following problem

$$\max \int_0^1 (1 - x^2 - \dot{x}^2)dt, \quad x(0) = 1, \quad x(1) \text{ free}$$

Notice that $F(t, x, \dot{x}) = 1 - x^2 - \dot{x}^2$. Thus, $\frac{\partial F}{\partial x} = -2x$ and $\frac{\partial F}{\partial \dot{x}} = -2\dot{x}$. Since $\frac{d}{dt}\frac{\partial F}{\partial \dot{x}} = -2\ddot{x}$, the Euler-Lagrange equation is

$$\ddot{x} = x$$

This 2nd order linear ODE has characteristic equation $\lambda^2 = 1$. Thus, has eigenvalues ±1. So, has general solution

$$x(t) = Ae^t + Be^{-t}.$$

To determine the constants A and B we use the initial condition and the transversality condition. Using the initial condition $x(0) = 1$ we get $1 = A + B$, whence $B = 1 - A$. The transversality condition is

$$\begin{aligned} 0 &= \frac{\partial F}{\partial \dot{x}}(1, x(1), \dot{x}(1)) \\ &= -2\dot{x}(1) \\ &= -2(Ae^1 - (1 - A)e^{-1}) \end{aligned}$$

Hence, $0 = Ae^1 - (1 - A)e^{-1}$ which implies that $A = \frac{1}{e^2+1}$. So we found a candidate solution to the variational problem

$$x(t) = \frac{1}{e^2 + 1}e^t + \left(1 - \frac{1}{e^2 + 1}\right) e^{-t}$$

Since F is a concave function of the variables $(x, \dot{x})$, we conclude that the candidate solution is indeed the solution.

Example 9.7. Consider the same problem as the previous example but with a different terminal condition

$$\max \int_0^1 (1 - x^2 - \dot{x}^2)dt, \quad x(0) = 1, \quad x(1) \geq 2$$

As before, we have the solution to the Euler-Lagrange equation

$$x(t) = Ae^t + (1 - A)e^{-t}$$

where we have used the initial condition to determine B. Now, the constant A has to be determined using the transversality condition,

$$\frac{\partial F}{\partial \dot{x}}(1, x(1), \dot{x}(1)) \leq 0$$

but equal to zero whenever $x(1) > 2$. First we consider the inequality above and get that $-2(Ae^1 - (1 - A)e^{-1}) \leq 0$ which gives $A \geq \frac{1}{e^2+1}$. Now, using the terminal condition $x(1) \geq 2$ we get $Ae^1 + (1 - A)e^{-1} \geq 2$ which gives $A \geq \frac{2e-1}{e^2-1}$. So A has to be greater or equal than the maximum of $\frac{1}{e^2+1}$ and $\frac{2e-1}{e^2-1}$. Since the $\frac{2e-1}{e^2-1} > \frac{1}{e^2+1}$, we conclude that $A \geq \frac{2e-1}{e^2-1}$. However, if $A > \frac{2e-1}{e^2-1}$, then $x(1)$ overshoots in the sense that

$x(1) > 2$ but $\frac{\partial F}{\partial \dot{x}}(1, x(1), \dot{x}(1)) < 0$. Therefore, the only possibility is $A = \frac{2e-1}{e^2-1}$ giving the solution to the variation problem

$$x(t) = \frac{2e-1}{e^2-1}e^t + (1 - \frac{2e-1}{e^2-1})e^{-t}$$

Exercise 54. Solve the variational problems:

(1)

$$\min \int_0^1 (t\dot{x} + \dot{x}^2)\, dt, \quad x(0) = 1, \quad x(1) \geq 1$$

(2)

$$\max \int_0^1 (10 - \dot{x}^2 - 2x\dot{x} - 5x^2)e^{-t}\, dt \quad x(0) = 1, \quad x(1) \text{ free}$$

Exercise 55. Let $A(t)$ denote a pensioner's wealth at time t and w be the pension income (constant) per unit time. The pensioner consumption is given by

$$C(t) = rA(t) + w - \dot{A}(t),$$

where $0 < r < 1$. Now the pensioner wants to maximize

$$\int_0^T U(C(t))e^{-rt}\, dt$$

knowing that $A(0) = A_0$. The pensioner's utility function U is given by $U(C) = 1 - e^{-C}$. Determine the optimal consumption $C(t)$ so that at the end of the period the pensioner retains at least $2A_0$, i.e., $A(T) \geq 2A_0$.

10. Optimal Control

Optimal control is an extension of the calculus of variations. Consider the **state** $x(t)$ of a system that evolves with time according to the following ODE,

$$\dot{x}(t) = g(t, x(t), u(t)), \quad x(t_0) = x_0$$

where $u(t)$ is a function called the **control**. We assume that for a given control, the ODE has a unique solution in a time interval $[t_0, t_1]$. Of course, different control functions originate different solutions.

10.1. **Basic optimal control problem.**

We aim to find the control $u(t)$ which maximizes (or minimizes) the **objective function**

$$\int_{t_0}^{t_1} f(t, x(t), u(t))\, dt$$

where the state of the system evolves according to the ODE above. More precisely, the **basic optimal control problem** is

$$\max_{u(t)\in U} \int_{t_0}^{t_1} f(t, x(t), u(t))\, dt$$
$$\text{subject to } \dot{x}(t) = g(t, x(t), u(t)), \quad x(t_0) = x_0, \quad x(t_1) \text{ free.}$$

In the simplest case the set U of control values, i.e., the values that the control $u(t)$ can take, is assumed to be $\mathbb{R}$. Thus no restriction on the control values. However, in several applications it is important to restrict the $u(t)$ to a small subset U of $\mathbb{R}$.

A similar formulation holds for the minimization problem. However, since we can write a minimization problem as a maximization problem with negative objective function, we shall restrict the following discussion to the maximization problem. See Example 10.6 for a minimization problem. Throughout we assume that f and g are sufficiently regular (C^2 is enough).

Since this is an optimization problem with constraints, as we did in static optimization, we define an auxiliary function

$$H(t, x, u, p) = f(t, x, u) + pg(t, x, u)$$

called the **Hamiltonian function**. The variable p is called the **costate**.

The following theorem gives necessary conditions that an optimal pair $(x^*(t), u^*(t))$ has to satisfy.

Theorem 10.1 (Pontryagin maximum principle)**.** *Suppose that $(x^*(t), u^*(t))$ is an optimal pair, i.e., solves the basic optimal control problem. There exists a function $p(t)$ defined in $[t_0, t_1]$ such that*

(1) *$u^*(t)$ is a maximizer of $u \in U \mapsto H(t, x^*(t), u, p(t))$ for every $t \in [t_0, t_1]$, i.e.,*

$$H(t, x^*(t), u, p(t)) \leq H(t, x^*(t), u^*(t), p(t)), \quad \forall\, u \in U,\ t \in [t_0, t_1].$$

(2) *$(x^*(t), u^*(t), p(t))$ satisfy the equations*

$$\begin{cases} \dot{x}^*(t) = g(t, x^*(t), u^*(t)), & x^*(t_0) = x_0 \\ \dot{p}(t) = -\frac{\partial H}{\partial x}(t, x^*(t), u^*(t), p(t)) \end{cases}$$

(3) *the transversality condition is satisfied*

$$p(t_1) = 0$$

Proof. Check the bibliography. □

Some remarks about the necessary conditions in the previous theorem.

Remark 10.2.

(1) the 1st condition expresses that for every $t \in [t_0, t_1]$ the point $u^*(t)$ is a maximizer of $H(t, x^*(t), u, p(t))$ seen as a function of the single variable $u \in \mathbb{R}$. Hence, in the unconstrained control case $U = \mathbb{R}$, we have that $u^*(t)$ is a critical point.

(2) the differential equations in the 2nd condition are called **Hamiltonian equations**. In fact, they can be written in a more compact form

$$\dot{x} = \frac{\partial H}{\partial p} \quad \text{and} \quad \dot{p} = -\frac{\partial H}{\partial x}$$

(3) the 3rd condition is a transversality condition because $x(t_1)$ is free.

Under convexity assumptions the conditions in the maximum principle theorem are in fact sufficient.

Theorem 10.3. *Suppose that H is concave as a function of the variables (x, u). If $(x(t), u(t), p(t))$ satisfy the conditions (1)-(3) in the maximum principle theorem, then $(x(t), u(t))$ solves the basic optimal control problem.*

Proof. Check the bibliography. □

Example 10.4. Consider the following problem

$$\max_{u(t)\in\mathbb{R}} \int_0^1 (1 - tx(t) - u(t)^2)\, dt, \quad \dot{x} = u(t), \quad x(0) = 1, \quad x(1) \text{ free}$$

First, we write the Hamiltonian

$$H(t, x, u, p) = 1 - tx - u^2 + pu$$

Next, to solve the control problem, we proceed as follows:

(1) Find the maximizer of H wrt to u. Taking the derivative we get $\frac{\partial H}{\partial u} = -2u + p$. Equating to zero we find $u(t) = p(t)/2$.

(2) The Hamiltonian equations are

$$\begin{cases} \dot{x} = u \\ \dot{p} = -\frac{\partial H}{\partial x} = t \end{cases}$$

The 2nd equation can be immediately integrated and we get the general solution $p(t) = \frac{t^2}{2} + A$ where A is a constant to be determined. Substituting $u(t) = p(t)/2$ in the 1st equation we get $\dot{x} = p(t)/2 = \frac{t^2}{4} + \frac{A}{2}$. Integrating we get,

$$x(t) = \frac{t^3}{12} + \frac{At}{2} + B$$

where B is another constant to be determined.

(3) In the last step we use the initial condition and the transversality condition to determined the constants. Using the transversality condition $p(1) = 0$ we get $\frac{1}{2} + A = 0$, i.e., $A = -\frac{1}{2}$. Using the initial condition $x(0) = 1$ we get $B = 1$.

(4) Since H is a concave function of (x, u) we conclude that

$$x(t) = \frac{t^3}{12} - \frac{t}{4} + 1, \quad u(t) = \frac{t^2}{4} - \frac{1}{4}$$

solves the problem.

Exercise 56. Solve the following optimal control problems

(1)

$$\max_{u(t)\in\mathbb{R}} \int_0^2 (e^t x(t) - u(t)^2)\, dt, \quad \dot{x} = -u(t), \quad x(0) = 0, \quad x(2) \text{ free}$$

(2)

$$\max_{u(t)\in\mathbb{R}} \int_0^1 (1 - u(t)^2)\, dt, \quad \dot{x} = x(t) + u(t), \quad x(0) = 1, \quad x(1) \text{ free}$$

(3)

$$\min_{u(t)\in\mathbb{R}} \int_0^1 (x(t) + u(t)^2)\, dt, \quad \dot{x} = -u(t), \quad x(0) = 0, \quad x(1) \text{ free}$$

10.2. **Variable terminal conditions.** Allowing other terminal conditions (as in calculus of variations) we write the **standard optimal control problem**

$$\max_{u(t)\in U} \int_{t_0}^{t_1} f(t, x(t), u(t))\, dt$$

$$\text{subject to } \dot{x}(t) = g(t, x(t), u(t)), \quad x(t_0) = x_0, \quad \begin{cases} x(t_1) \text{ free}, & (A) \\ x(t_1) \geq x_1, & (B) \\ x(t_1) = x_1, & (C) \end{cases}.$$

Theorem 10.5. *Suppose that H is concave as a function of the variables (x, u). If $(x(t), u(t), p(t))$ satisfy the conditions (1)-(2) in the maximum principle theorem and the corresponding transversality condition*

(A): *$p(t_1) = 0$,*

(B): *$p(t_1) \geq 0$ (with $p(t_1) = 0$ if $x(t_1) > x_1$,*

(C): *$x(t_1) = x_1$ and no condition on $p(t_1)$,*

then $(x(t), u(t))$ solves the basic optimal control problem.

Proof. Check the bibliography. □

Example 10.6. Consider the problem

$$\min_{u(t)\in\mathbb{R}} \int_0^1 \frac{x(t)^2}{2} + \frac{u(t)^2}{2}\, dt, \quad \dot{x}(t) = u(t), \quad x(0) = 2, \quad x(1) \geq 4$$

First, we turn the minimization problem into a maximization problem,

$$\max \int_0^1 -\frac{x(t)^2}{2} - \frac{u(t)^2}{2}\,dt, \quad \dot{x}(t) = u(t), \quad x(0) = 2, \quad x(1) \geq 4$$

Now, we write the Hamiltonian

$$H(t, x, u, p) = -\frac{x^2}{2} - \frac{u^2}{2} + pu$$

Next, to solve the control problem, we proceed as follows:

(1) Find the maximize of H wrt to u. Taking the derivative we get $\frac{\partial H}{\partial u} = -u + p$. Equating to zero we find $u(t) = p(t)$.

(2) The Hamiltonian equations are

$$\begin{cases} \dot{x} = u \\ \dot{p} = x \end{cases} \Leftrightarrow \begin{cases} \dot{x} = p \\ \dot{p} = x \end{cases} \Leftrightarrow \begin{cases} \ddot{x} = \dot{p} \\ \dot{p} = x \end{cases} \Leftrightarrow \begin{cases} \ddot{x} = x \\ \dot{p} = x \end{cases}$$

The characteristic equation of the 1st ODE is $\lambda^2 = 1$. Thus it has eigenvalues ± 1, from which we conclude that the general solution is $x(t) = Ae^t + Be^{-t}$. Using the equation $p = \dot{x}$ we get $p(t) = Ae^t - Be^{-t}$.

(3) Using the initial condition $x(0) = 2$ we get $A + B = 2$, i.e., $B = 2 - A$. And the terminal condition $x(1) \geq 4$ implies that $x(1) = Ae + (2 - A)e^{-1} \geq 4$, thus $A \geq \frac{4-2e^{-1}}{e-e^{-1}}$. The transversality condition gives $p(1) = Ae - (2 - A)e^{-1} \geq 0$ which implies $A \geq \frac{2}{e+e^{-1}}$. If $A > \frac{4-2e^{-1}}{e-e^{-1}}$ (i.e. $x(1) > 4$) then $p(1) > 0$ which cannot be. Thus $A = \frac{4-2e^{-1}}{e-e^{-1}}$.

(4) Since H is a concave function of (x, u) we conclude that

$$x(t) = \frac{4 - 2e^{-1}}{e - e^{-1}}e^t + (2 - \frac{4 - 2e^{-1}}{e - e^{-1}})e^{-t}, \quad u(t) = \dot{x}(t).$$

solves the problem.

Exercise 57. Solve the following optimal control problems

(1)

$$\max_{u(t)\in\mathbb{R}} \int_0^1 (1 - x(t)^2 - u(t)^2)\,dt, \quad \dot{x} = u(t), \quad x(0) = 0, \quad x(1) \geq 1$$

(2)

$$\max_{u(t)\in[-1,1]} \int_0^1 x(t)\,dt, \quad \dot{x} = x(t) + u(t), \quad x(0) = 0, \quad x(1) \text{ free}$$

10.3. **Scrap value formulation.** We consider a more general class of optimal control problems that appear in the economics literature

$$\max_{u(t)\in U}\left\{\int_{t_0}^{t_1} f(t,x(t),u(t))e^{-rt}\,dt + S(x(t_1))e^{-rt_1}\right\}$$

$$\text{subject to } \dot{x}(t)=g(t,x(t),u(t)),\quad x(t_0)=x_0,\quad \begin{cases} x(t_1) \text{ free}, & (A)\\ x(t_1)\geq x_1, & (B)\\ x(t_1)=x_1, & (C)\end{cases}.$$

Here $r>0$ is an interest rate (or discount factor) and S is a function of x called the **scrap value function**. Consider the Hamiltonian function

$$H(t,x,u,p)=f(t,x,u)+pg(t,x,u).$$

Theorem 10.7. *Suppose that H is concave as a function of the variables (x,u). If $(x^*(t),u^*(t),p(t))$ satisfy the following conditions*

(1) *$u^*(t)$ is a maximizer of $u\in U\mapsto H(t,x^*(t),u,p(t))$ for every $t\in[t_0,t_1]$, i.e.,*

$$H(t,x^*(t),u,p(t))\leq H(t,x^*(t),u^*(t),p(t)),\quad \forall\, u\in U,\ t\in[t_0,t_1].$$

(2)

$$\begin{cases}\dot{x}^*(t)=g(t,x^*(t),u^*(t)), & x^*(t_0)=x_0\\ \dot{p}(t)-rp(t)=-\frac{\partial H}{\partial x}(t,x^*(t),u^*(t),p(t))\end{cases}$$

(3) **(A):** *$p(t_1)=\frac{\partial S}{\partial x}(x^*(t_1))$,*
(B): *$p(t_1)\geq\frac{\partial S}{\partial x}(x^*(t_1))$ (with $=$ if $x(t_1)>x_1$,*
(C): *$x(t_1)=x_1$ and no condition on $p(t_1)$,*

then $(x^*(t),u^*(t))$ solves the general optimal control problem.

Proof. Check the bibliography. □

Example 10.8. Consider the problem

$$\max_{u(t)\in\mathbb{R}}\int_0^T (x(t)-u(t)^2)e^{-t}\,dt+2x(T)e^{-T},\quad \dot{x}=-x+u,\quad x(0)=1,\quad x(T)\text{ free}$$

Write the Hamiltonian

$$H(t,x,u,p)=x-u^2+p(u-x)$$

Notice that $S(x)=2x$. Next, to solve the control problem, we proceed as follows:

(1) Find the maximizer of H wrt to u. Taking the derivative we get $\frac{\partial H}{\partial u}=-2u+p$. Equating to zero we find $u(t)=p(t)/2$.
(2) The Hamiltonian equations are

$$\begin{cases}\dot{x}=-x+u\\ \dot{p}-p=-1+p\end{cases}\Leftrightarrow\begin{cases}\dot{x}=-x+p/2\\ \dot{p}=-1+2p\end{cases}$$

The general solution of the 2nd ODE is

$$p(t) = 1/2 + Ae^{2t}$$

Substituting into the 1st ODE we get the general solution

$$x(t) = \frac{1}{4} + \frac{A}{6}e^{2t} + Be^{-t}.$$

(3) Using the transversality condition $p(T) = \frac{\partial S}{\partial x}(x(T)) = 2$ we get $1/2 + Ae^{2T} = 2$, thus $A = \frac{3}{2}e^{-2T}$. Using the initial condition $x(0) = 1$ we get $\frac{1}{4} + \frac{A}{6} + B = 1$, i.e., $B = \frac{3}{4} - \frac{A}{6}$.

(4) Since H is a concave function of (x, u) we conclude $(x(t), u(t))$ solve the problem.

Exercise 58. Solve the following optimal control problems

(1)

$$\max_{u(t)\in\mathbb{R}} \left\{ \int_0^1 -\frac{1}{2}u(t)^2\, dt + \sqrt{x(1)} \right\}, \quad \dot{x} = x + u, \quad x(0) = 0, \quad x(1) \text{ free}$$

(2)

$$\max_{u(t)\in\mathbb{R}} \left\{ \int_0^T -e^{-t}(x(t) - u(t))^2\, dt - e^{-T}x(T)^2 \right\}, \quad \dot{x} = u - x + 1, \quad x(0) = 0, \quad x(T) \text{ free}$$

10.4. **Infinite horizon.** Many optimal control problems that appear in the economics literature have an infinite time horizon,

$$\max_{u(t)\in U} \left\{ \int_{t_0}^{+\infty} f(t, x(t), u(t))e^{-rt}\, dt \right\}$$
$$\text{subject to } \dot{x}(t) = g(t, x(t), u(t)), \quad x(t_0) = x_0$$

Some problems impose an extra asymptotic boundary condition

$$\lim_{t\to+\infty} x(t) \geq x_1.$$

Consider the Hamiltonian function

$$H(t, x, u, p) = f(t, x, u) + pg(t, x, u).$$

Theorem 10.9. *Suppose that H is concave as a function of the variables (x, u). If $(x^*(t), u^*(t), p(t))$ satisfy the following conditions*

(1) *$u^*(t)$ is a maximizer of $u \in U \mapsto H(t, x^*(t), u, p(t))$ for every $t \in [t_0, +\infty]$, i.e.,*

$$H(t, x^*(t), u, p(t)) \leq H(t, x^*(t), u^*(t), p(t)), \quad \forall\, u \in U,\ t \in [t_0, +\infty].$$

(2)

$$\begin{cases} \dot{x}^*(t) = g(t, x^*(t), u^*(t)), \quad x^*(t_0) = x_0 \\ \dot{p}(t) - rp(t) = -\frac{\partial H}{\partial x}(t, x^*(t), u^*(t), p(t)) \end{cases}$$

(3)

$$\lim_{t\to+\infty} p^*(t)e^{-rt}(x_1 - x^*(t)) \geq 0$$

(4)

$$\lim_{t\to+\infty} p^*(t)e^{-rt} < \infty \quad \textit{and} \quad p^*(t) \geq 0, \forall t \geq 0$$

then $(x^*(t), u^*(t))$ *solves the general optimal control problem.*

Proof. Check the bibliography. □

Example 10.10. Consider the problem

$$\max_{u(t)\in\mathbb{R}} \int_0^{+\infty} -u(t)^2 e^{-t}\, dt, \quad \dot{x} = ue^{-t}, \quad x(0) = 0, \quad \lim_{t\to+\infty} x(t) \geq 1$$

Write the Hamiltonian

$$H(t, x, u, p) = -u^2 + pue^{-t}$$

Next, to solve the control problem, we proceed as follows:

(1) Find the maximizer of H wrt to u. Taking the derivative we get $\frac{\partial H}{\partial u} = -2u + pe^{-t}$. Equating to zero we find $u(t) = p(t)e^{-t}/2$.

(2) The Hamiltonian equations are

$$\begin{cases} \dot{x} = ue^{-t} \\ \dot{p} - p = 0 \end{cases} \Leftrightarrow \begin{cases} \dot{x} = p(t)e^{-2t}/2 \\ \dot{p} = p \end{cases}$$

The general solution of the 2nd ODE is

$$p(t) = Ae^t$$

Substituting into the 1st ODE we get the general solution

$$x(t) = B - \frac{A}{2}e^{-t}$$

which taking into account the initial condition $x(0) = 0$ we get that $B = \frac{A}{2}$, thus

$$x(t) = \frac{A}{2}\left(1 - e^{-t}\right)$$

(3) To determine the value of A we use the condition $\lim_{t\to+\infty} x(t) \geq 1$ which gives $\frac{A}{2} \geq 1$, hence $A \geq 2$.

(4) We want

$$\lim_{t\to+\infty} p(t)e^{-t}(1 - x(t)) = A\left(1 - \frac{A}{2}\right) \geq 0.$$

Thus, $A \leq 2$, which means that the only possible value for A is $A = 2$. Also notice that $\lim_{t\to+\infty} p(t)e^{-t} = 2$ and $p(t) = 2e^t \geq 0$ for all t.

(5) Since H is a concave function of (x, u) we conclude $(x(t), u(t)) = (1 - e^{-t}, 1)$ solve the problem.

Exercise 59. Solve the following optimal control problems

(1)

$$\max_{u(t)\in\mathbb{R}} \left\{ \int_0^{+\infty} 2\sqrt{x(t)-u(t)}e^{-2t}\,dt \right\}, \quad \dot{x}=u, \quad x(0)=1, \quad \lim_{t\to+\infty} x(t) \geq 0.$$

(2)

$$\max_{u(t)\in\mathbb{R}} \left\{ \int_0^{+\infty} \log(u(t))e^{-t/5}\,dt \right\}, \quad \dot{x}=x/10-u, \quad x(0)=10, \quad \lim_{t\to+\infty} x(t) \geq 0.$$