Applied Analysis

I.M. Sigal with M. Merkli

Introduction

Analysis studies properties of maps. Sometimes those maps are defined on finite dimensional spaces. In this case, they are called functions. Often maps are defined on infinite dimensional spaces, which are spaces of functions. The part of analysis which concentrates on properties of functions is called *classical analysis*, the part dealing with maps is often refered to as *modern analysis.* There is an obvious overlap and interdependence between these two parts.

In this course, we deal with modern analysis. Properties of functions are studied as much as they are needed for understanding maps. More specifically, our emphasis is on applications of modern analysis and the material is selected accordingly. As a result, the parts of analysis whose main role is to support the internal structure is reviewed only briefly. Fortunately, the latter material is well covered in the literature and there are several excellent textbooks on the subject which we often refer to (see [F], [McO], [RS]).

Toronto, 2000-2001

Contents

Chapter IV. Variational Calculus **75**

Chapter V. Dynamical systems **115**

Chapter VI. Stochastic Analysis **143**

Bibliography **171**

Chapter I. Measures and Integrals

1. Motivation

Our objective in this chapter is to generalize the notion of length and volume to complicated subsets of a general set X. Our generalization results in a concept of measure, denoted μ. This concept is used later in the theories of integration and probability.

Example: $X = \mathbb{R}^n$. We want to define a function μ defined on subsets of $\mathbb{R}^n$ and with values in $[0, \infty]$, such that

(i) if $E_1, E_2, \dots$ is a finite or infinite collection of disjoint subsets of $\mathbb{R}^n$, then $\mu(E_1 \cup E_2 \cup \cdots) = \mu(E_1) + \mu(E_2) + \cdots$,

(ii) if E is congruent to F (i.e. if E can be transformed into F by rigid motions: translations, rotations and reflections), then $\mu(E) = \mu(F)$,

(iii) $\mu(Q) = 1$, where $Q = \{x = (x_1, \dots, x_n) \in \mathbb{R}^n : 0 \leq x_j \leq 1, \text{ for } j = 1, \dots, n\}$ is the unit cube.

One can show that a function satisfying (i)–(iii) *cannot be defined on all subsets* of $\mathbb{R}^n$! See e.g. [F].

Let $n = 1$. It is easy to measure intervals in $\mathbb{R}$. We define $\mu(I) = b - a$ if I is one of the following: $(a, b), [a, b], (a, b], [a, b)$, i.e. an open, closed or semiclosed interval. However, a union of disjoint intervals is in general not an interval. So we define its measure by using Property (i) above. We can now go on and form unions of unions of intervals and so forth. If we continue this procedure, what kind of subsets of $\mathbb{R}$ do we end up with? We answer this question in the next section. Meantime, we remark that the collection of subsets of X on which a measure μ is defined is called the *domain of* μ. Thus a measure μ on X is a function on some collection of subsets of X, called the domain of μ, with values in $[0, \infty]$.

2. σ–algebras

Let X be a nonempty set and let $\mathcal{P}(X)$ denote its power set, i.e. the set of all subsets of X.

Definition. *An* algebra $\mathcal{A}$ *of subsets of* X *is a nonempty collection of subsets of* X*, i.e.* $\mathcal{A} \subset \mathcal{P}(X)$*, which is closed under finite unions and under complements, i.e. if* $\{E_j\}_1^n \subset \mathcal{A}$*, then* $\cup_i^n E_j \in \mathcal{A}$*, and if* $E \in \mathcal{A}$*, then* $E^c := X \backslash E \in \mathcal{A}$*.*

Observe that an algebra is also closed under finite intersections. This follows from the relation $\cap E_j = (\cup E_j^c)^c$. Moreover, the whole space X and the empty set ϕ are always contained in an algebra. This follows from $E \cup E^c = X$ and $X^c = \phi$.

Definition. *A* σ–algebra *is an algebra which is closed under countable unions, i.e. if* $\{E_j\}_1^\infty \subset \mathcal{A}$*, then* $\cup_1^\infty E_j \in \mathcal{A}$*.*

Examples. Both $\mathcal{P}(X)$ and $\{\phi, X\}$ are σ–algebras.

Often, σ–algebras interesting to us are difficult to describe. Instead, we show how to produce them, starting from easily imaginable sets such as intervals and boxes (by a box in $\mathbb{R}^n$, we understand a set of the form $\{x \in \mathbb{R}^n : x_i \in I_i, \forall i = 1, \dots, n\}$ for some intervals I_i, i.e. a box is a product of n intervals: $I_1 \times \cdots \times I_n$). Let A be a label set, and let $\mathcal{A}_\alpha$, $\alpha \in A$, be σ–algebras.

Exercise. Show: if $\mathcal{A}_\alpha$ are σ–algebras, then $\cap_\alpha \mathcal{A}_\alpha$ is a σ–algebra.

Hence for any $\mathcal{E} \in \mathcal{P}(X)$, there is a unique smallest σ–algebra containing $\mathcal{E}$. We call this σ–algebra $\mathcal{M}(\mathcal{E})$:

$$\mathcal{M}(\mathcal{E}) := \bigcap_{\substack{\mathcal{A}:\sigma\text{–algebra} \\ \mathcal{E} \subset \mathcal{A}}} \mathcal{A}.$$

We say that $\mathcal{M}(\mathcal{E})$ is *generated* by $\mathcal{E}$.

Example. Denote by $B_{\mathbb{R}^n} = \mathcal{M}(\mathcal{E}_0)$, where $\mathcal{E}_0$ is the set of all open subsets of $\mathbb{R}^n$. $B_{\mathbb{R}^n}$ is called the *Borel-algebra* of $\mathbb{R}^n$. Elements of $B_{\mathbb{R}^n}$ are called *Borel sets.* The same σ–algebra is obtained if instead of open sets, we start with open boxes, or with closed sets or boxes, or semiopen boxes, etc. We formulate this in the case $n = 1$:

Proposition. *$B_{\mathbb{R}}$ is generated by each of the following:*
(a) the open intervals: $\mathcal{E}_1 = \{(a, b) : a < b\}$
(b) the closed intervals: $\mathcal{E}_2 = \{[a, b] : a < b\}$
(c) the half-open intervals: $\mathcal{E}_3 = \{(a, b] : a < b\}$ or $\mathcal{E}_4 = \{[a, b) : a < b\}$.

Proof of (a) and (b). Since $\mathcal{E}_1 \subset \mathcal{E}_0$, then $\mathcal{M}(\mathcal{E}_1) \subset \mathcal{M}(\mathcal{E}_0)$.

Exercise. Show this.

On the other hand, every open subset E of $\mathbb{R}$ is a countable union of open intervals I_j: $E = \cup_1^\infty I_j \in \mathcal{M}(\mathcal{E}_1)$. Hence $\mathcal{M}(\mathcal{E}_0) \subset \mathcal{M}(\mathcal{E}_1)$, so $\mathcal{M}(\mathcal{E}_0) = \mathcal{M}(\mathcal{E}_1)$.

Now we show (b) by proving $\mathcal{M}(\mathcal{E}_1) = \mathcal{M}(\mathcal{E}_2)$. Take any open interval $(a, b) \in \mathcal{E}_1$. We have $(a, b) = \cup_{n=1}^\infty [a + 1/n, b - 1/n] \in \mathcal{M}(\mathcal{E}_2)$. Therefore $(a, b) \in \mathcal{M}(\mathcal{E}_2)$, so $\mathcal{M}(\mathcal{E}_1) \subset \mathcal{M}(\mathcal{E}_2)$. To prove the converse inclusion, take any closed interval $[a, b] \in \mathcal{E}_2 \Rightarrow [a, b] = \cap_{n=1}^\infty (a - 1/n, b + 1/n) \in \mathcal{M}(\mathcal{E}_1) \Rightarrow [a, b] \in \mathcal{M}(\mathcal{E}_1) \Rightarrow \mathcal{M}(\mathcal{E}_2) \subset \mathcal{M}(\mathcal{E}_1) \Rightarrow \mathcal{M}(\mathcal{E}_1) = \mathcal{M}(\mathcal{E}_2)$. ■

3. Measures

Let X be a nonempty set, $\mathcal{M}$ a σ–algebra of subsets from X (we say X is equipped with $\mathcal{M}$).

Definition. *A measure μ is a function on $\mathcal{M}$ with values in* $[0, \infty]$*, s.t. the following hold:*

(a) $\mu(\phi) = 0$,

(b) if $\{E_j\}_1^\infty$ are disjoint subsets from $\mathcal{M}$, then $\mu(\cup_1^\infty E_j) = \sum_1^\infty \mu(E_j)$.

Terminology. • Property (b) is called *countable additivity.*
• A function $\mu : \mathcal{M} \to [0, \infty]$ which satisfies (a) and the condition

(b') if $\{E_j\}_1^n$ are disjoint subsets from $\mathcal{M}$, then $\mu(\cup_1^n E_j) = \sum_1^n \mu(E_j)$,

is called a *finitely additive measure.* Of course, any measure is finitely additive.

- A *measurable space* is a pair $(X, \mathcal{M})$, where X is a nonempty set and $\mathcal{M}$ is a σ–algebra of subsets of X $(\mathcal{M} \subset \mathcal{P}(X))$.
- A *measure space* is a triple $(X, \mathcal{M}, \mu)$, where $(X, \mathcal{M})$ is a measurable space, and μ is a measure on $\mathcal{M}$.
- μ is a *finite measure* if $\mu(X) < \infty$, μ is *σ-finite* if $X = \cup_1^\infty E_j$, where $E_j \subset \mathcal{M}$, and $\mu(E_j) < \infty$.
- Elements of $\mathcal{M}$ are called *measurable sets.*
- $E \in \mathcal{M}$ such that $\mu(E) = 0$ is called a *null set.*
- If some property holds $\forall x \in X \backslash E$, where $\mu(E) = 0$, then we say this property holds *almost everywhere*, written a.e.

Shorthand. Whenever $\mathcal{M}$ is understood, we say μ is defined on X, e.g. by "μ is defined on $\mathbb{R}^n$", we mean μ is defined on $B_{\mathbb{R}^n}$. If $\mathcal{M}$ and μ are understood, we call X a measure space.

Theorem. *Let $(X, \mathcal{M}, \mu)$ be a measure space. Then*

(a) (monotonicity) $E \subset F \to \mu(E) \leq \mu(F)$,

(b) (subadditivity) $\mu(\cup_1^\infty E_j) \leq \sum_1^\infty \mu(E_j)$,

(c) (continuity from below) $E_j \uparrow E \Rightarrow \lim_{j\to\infty} \mu(E_j) = \mu(E)$,

(d) (continuity from above) $E_j \downarrow E \Rightarrow \lim_{j\to\infty} \mu(E_j) = \mu(E)$.

We used the notation $E_j \uparrow E \Leftrightarrow E_1 \subset E_2 \subset \cdots$, and $\cup_1^\infty E_j = E$. Analogously $E_j \downarrow E \Leftrightarrow E_1 \supset E_2 \supset \cdots$, and $\cap_1^\infty E_j = E$.

Proof of (a). If $E \subset F$, then the sets E and $F \backslash E$ are disjoint and $E \cup (F \backslash E) = F$. So by Property (b) in the definition of a measure, and since $\mu(G) \geq 0$ $\forall G \in \mathcal{M}$, we get $\mu(F) = \mu(E) + \mu(F \backslash E) \geq \mu(E)$. ■

Exercise. Prove (b).

4. Borel measures on $\mathbb{R}$

In this section, we discuss a construction of measures on $\mathbb{R}^n$. To fix the ideas, we consider only the case $n = 1$.

Definition. *A function $F : \mathbb{R} \to \mathbb{R}$ is called* right continuous at a *iff* $\lim_{x \downarrow a} F(x) = F(a)$. *If F is right continuous at every point in an open set, then we say that F is right continuous in that set.*

Similarly, we define *left continuous* functions.

Consider the Borel σ–algebra, $B_{\mathbb{R}}$, on $\mathbb{R}$. We showed in Section 1.2 that it is generated by either one of the three: open, closed or semiopen intervals. Let $F : \mathbb{R} \to \mathbb{R}$ be an increasing right continuous function on $\mathbb{R}$. We want to define a measure μ on $B_{\mathbb{R}}$ sucht that

$$\begin{aligned} \mu((a,b]) &= F(b) - F(a), && (4.1) \\ \mu(\cup_1^n (a_j, b_j]) &= \sum_1^n \left(F(b_j) - F(a_j)\right), && (4.2) \end{aligned}$$

where $a_1 < b_1 < a_2 < b_2 < \cdots$. The first question we ask is whether there is a measure μ on $B_{\mathbb{R}}$ that satisfies (4.1) and (4.2). One can show that such a measure does exist. This follows from a general result on the completion of premeasures. We do not explain this result here, but refer the reader to [F], Theorem 1.14 and Proposition 1.15. The next question is uniqueness of such a measure. The answer is given in the following

Theorem. *Let $F : \mathbb{R} \to \mathbb{R}$ be an increasing right continuous function. Then there is a unique measure $\mu = \mu_F$ on $B_{\mathbb{R}}$ such that $\mu_F((a,b]) = F(b) - F(a)$, for any $a, b \in \mathbb{R}$. If G is another such function, then $\mu_F = \mu_G$ iff $F - G$ is constant. Conversely, if μ is a measure on $B_{\mathbb{R}}$ which is finite on all bounded Borel sets, then $\mu = \mu_F$, where F is the increasing right continuous function defined by $F(x) = \mu((0,x])$ if $x > 0$, $F(0) = 0$, and $F(x) = -\mu((x,0])$ if $x < 0$.*

Exercise. Check (4.1) for $F(x)$ defined as in the above theorem. For a proof of the theorem, see [F], Theorem 1.16.

Lebesgue measure. Now we define the simplest and most important measure on $B_{\mathbb{R}}$, the Lebesgue measure m:

$$m := \mu_F \text{ for } \ F(x) = x. \qquad (4.3)$$

Important properties of m are collected in the following

Theorem. *If $E \in B_{\mathbb{R}}$, and $r, s \in \mathbb{R}$, then $E + s \in B_{\mathbb{R}}$, $rE \in \mathbb{R}$, and $m(E + s) = m(E)$, $m(rE) = |r|m(E)$.*

We used the notation $E + s := \{x + s : x \in E\}$ and $rE := \{rx : x \in E\}$. The theorem says that the Lebesgue measure is translation, dilation and reflection invariant.

Proof. The collection $\mathcal{E}$ of all open intervals is invariant under translation, dilation and reflection. Since $B_{\mathbb{R}}$ is generated by $\mathcal{E}$, then $B_{\mathbb{R}}$ must also be invariant under those operations. To prove the second part of the theorem, we define $m_s(E) = m(E+s)$ and $m_r(E) = m(rE)$.

Exercise. Show that (i) m_s and m_r are measures, (ii) on $\mathcal{E}$, m_s coincides with m and m_r with $|r|m$.

By the unique extension theorem (see [F], Theorem 1.14), m_s agrees with m, and m_r agrees with $|r|m$ on the whole $B_{\mathbb{R}}$. ■

Exercise. Show that $m(E) = 0$ if $E = \{x\}$ (singleton), and if E is a countable set ($E = \{x_j\}_1^\infty$).

There are however null sets (with respect to Lebesgue measure) having the cardinality of the continuum! (E is said to have the cardinality of the continuum iff there is a bijection $f : E \to \mathbb{R}$.) A standard example of such a set is the *Cantor set C*. The Cantor set is obtained using the following procedure. Remove from the set $[0, 1]$ the open middle third $(1/3, 2/3)$, then remove from each of the two remaining intervals their open middle thirds $(1/9, 2/9)$ and $(7/9, 8/9)$, respectively, and so forth. One can show that

(a) $m(C) = 0$,
(b) C has the cardinality of the continuum,
(c) C has no isolated points,
(d) for any $x, y \in C$ with $x < y$ there is a $z \notin C$ s.t. $x < z < y$,
thus C is totally disconnected (and nowhere dense).

In a similar way, we can construct Borel and Lebesgue measures on $B_{\mathbb{R}^n}$ for $n > 1$. Another way of defining the Lebesgue measure m on $B_{\mathbb{R}^n}$ is given by $m = \prod_{i}^{n} m_j$, where m_j is the Lebesgue measure on $\mathbb{R}$.
We have set up the basics of measure theory, and will return to this

topic in the last chapter, when we consider probability theory. In the next section, we use measure theory to construct a theory of integration.

5. Integration: Measurable functions

First, we introduce a set of functions which can be in principle integrated with respect to measures defined on some σ–algebra $\mathcal{M}$.

Let X and Y be two nonempty sets. Consider a function (also called *map* or *mapping*) $f : X \to Y$. Such a function induces a map $f^{-1} : \mathcal{P}(Y) \to \mathcal{P}(X)$, defined for $E \in \mathcal{P}(Y)$ by

$$f^{-1}(E) := \{x \in X : f(x) \in E\}.$$

Exercise. Show f^{-1} commutes with unions, intersections and complements, i.e. $f^{-1}(E \cup F) = f^{-1}(E) \cup f^{-1}(F)$, $f^{-1}(E \cap F) = f^{-1}(E) \cap f^{-1}(F)$, and $f^{-1}(E^c) = (f^{-1}(E))^c$.

The properties above imply that if $\mathcal{N}$ is a σ–algebra on Y, then $f^{-1}(\mathcal{N})$ is a σ–algebra on X, with the obvious notation $f^{-1}(\mathcal{N}) := \{f^{-1}(E) : E \in \mathcal{N}\}$.

Definition. *Given two measurable spaces $(X, \mathcal{M})$ and $(Y, \mathcal{N})$, we say a map $f : X \to Y$ is $(\mathcal{M}, \mathcal{N})$*–measurable *(or simply* measurable*, if $\mathcal{M}$ and $\mathcal{N}$ are understood), iff $f^{-1}(E) \in \mathcal{M}$ for every $E \in \mathcal{N}$ (i.e. $f^{-1}(\mathcal{N}) \subset \mathcal{M}$).*

Exercise. Show that the composition of measurable functions is measurable, i.e. if $f : X \to Y$ is $(\mathcal{M}, \mathcal{N})$–measurable, and $g : Y \to Z$ is $(\mathcal{N}, \mathcal{O})$–measurable, then $g \circ f : X \to Z$ is $(\mathcal{M}, \mathcal{O})$–measurable.

Proposition. *If $\mathcal{N}$ is generated by $\mathcal{E}$, then f is $(\mathcal{M}, \mathcal{N})$–measurable iff $f^{-1}(E) \in \mathcal{M}$, $\forall E \in \mathcal{E}$.*

Proof. ($\Rightarrow$) If f is $(\mathcal{M}, \mathcal{N})$–measurable, then $f^{-1}(E) \in \mathcal{M}$, $\forall E \in \mathcal{N}$, hence $f^{-1}(E) \in \mathcal{M}$ $\forall E \in \mathcal{E}$, since $\mathcal{E} \subset \mathcal{M}$.

($\Leftarrow$) Let $\mathcal{N}_0 := \{E \in \mathcal{P}(Y) : f^{-1}(E) \in \mathcal{M}\}$. Then $\mathcal{N}_0$ is a σ–algebra of subsets in Y, and $\mathcal{E} \subset \mathcal{N}_0$. Now since by definition, $\mathcal{N}$ is the *smallest* σ–algebra containing $\mathcal{E}$, then $\mathcal{N} \subseteq \mathcal{N}_0$. Therefore we have $\forall E \in \mathcal{N}$: $f^{-1}(E) \in \mathcal{M}$, i.e. f is $(\mathcal{M}, \mathcal{N})$–measurable. ■

Corollary. *If X and Y are metric spaces (e.g. $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$), then every continuous function $f : X \to Y$ is (B_X, B_Y)–measurable.*

Proof. f is continuous iff $f^{-1}(U)$ is open in X for every U open in Y. Hence the statement follows from the proposition above and the facts that B_X and B_Y are generated by open sets. ■

If $(Y, \mathcal{N}) = (\mathbb{R}^m, B_{\mathbb{R}^m})$, then f is called *$\mathcal{M}$–measurable* (or again, just *measurable*). If also $(X, \mathcal{M}) = (\mathbb{R}^n, B_{\mathbb{R}^n})$, then f is called *Borel-measurable.*

Theorem. *If $f, g : X \to \mathbb{R}$ are $\mathcal{M}$–measurable, then so are $f + g$, fg, $\max(f, g)$, and $\min(f, g)$.*

Proof. Define new functions $F : X \to \mathbb{R}^2$ and $\varphi : \mathbb{R}^2 \to \mathbb{R}$ by $F(x) = (f(x), g(x))$, and $\varphi(r, s) = r + s$. Then $f + g = \varphi \circ F$. Since φ is continuous, it is measurable. We now show that F is measurable. We know that the Borel σ–algebra $B_{\mathbb{R}^2}$ is generated by open rectangles $R = (a, b) \times (c, d)$. By the proposition above, it is enough to show that $F^{-1}(R) \in \mathcal{M}$, for any such open rectangle. Now $F^{-1}(R) = f^{-1}((a, b)) \cap g^{-1}((c, d))$, but since $f^{-1}((a, b))$, $g^{-1}((c, d)) \in \mathcal{M}$, we have that $F^{-1}(R) \in \mathcal{M}$. Therefore, F is measurable, and hence so is $\varphi \circ F = f + g$, since the composition of two measurable functions is measurable.

To prove that fg is measurable, proceed in the same way, taking now $\varphi(r, s) = rs$.

To prove that $h := \max(f, g)$ is measurable, notice that $h^{-1}((a, \infty]) = f^{-1}((a, \infty]) \cup g^{-1}((a, \infty])$. Since f and g are both measurable, then $f^{-1}((a, \infty])$, $g^{-1}((a, \infty]) \in \mathcal{M}$, and therefore $h^{-1}((a, \infty]) \in \mathcal{M}$. But $\{(a, \infty] : a \in \mathbb{R}\}$ generates $B_{\mathbb{R}}$, so we conclude that h is measurable.

To prove that $\min(f, g)$ is measurable, repeat the last argument, replacing $(a, \infty]$ by $[-\infty, a)$. ■

Theorem. *If $\{f_j\}$ is a sequence of $\overline{\mathbb{R}}$–valued measurable functions on $(X, \mathcal{M})$, then the functions $\sup_j f_j(x)$, $\inf_j f_j(x)$, $\limsup_{j\to\infty} f_j(x)$ and $\liminf_{j\to\infty} f_j(x)$ are all measurable. Moreover, if $\lim_{j\to\infty} f_j(x)$ exists, then it is also measurable.*

We used the following definitions:

$$\begin{aligned}\limsup_{j\to\infty} f_j &= \inf_{k\geq 1}\left(\sup_{j\geq k} f_j\right) \quad \left(= \lim_{k\to\infty}(\sup_{i\geq k} f_j)\right), \\ \liminf_{j\to\infty} f_j &= \sup_{k\geq 1}\left(\inf_{j\geq k} f_j\right) \quad \left(= \lim_{k\to\infty}(\inf_{i\geq k} f_j)\right).\end{aligned}$$

The proof of this theorem is given in [F], Proposition 2.7 and Corollary 2.8.

The *characteristic function* or *indicator function* χ_E of a set $E \subset X$ is defined by

$$\chi_E(x) = \begin{cases} 1 & \text{if } x \in E \\ 0 & \text{if } x \in E^c \end{cases} \tag{5.1}$$

A *simple function* is a function which is a finite linear combination of characteristic functions of sets in $\mathcal{M}$ (i.e. measurable sets). Note that if f is a simple function, then the range of f consists of a finitely many numbers: $f = \sum_1^n a_j \chi_{E_j} \Rightarrow \text{Ran} f = \{a_j\}_1^n$.

Exercise. Prove that the converse is also true: $\text{Ran} f = \{a_j\}_1^n \Rightarrow f$ is simple.

In what follows, we assume that in the representation $f = \sum_1^n a_j \chi_{E_j}$ all the a_j's are distinct (otherwise we combine the corresponding terms into one term). We call such a representation the *standard representation* of the simple function in question.

Clearly, if f and g are simple functions, then so are $f+g$, af and fg. It is possible to show that any measurable function can be approximated by simple functions in the following sense:

Theorem. *For any measurable function f, there is a increasing sequence $\{\varphi_j\}$ of simple functions, $\varphi_1 \leq \varphi_2 \leq \cdots \leq f$, such that $\varphi_j \to f$ pointwise, and uniformly on any bounded set.*

Proof. In order to be more explicit, we assume that the range of f lies in the interval $I = [0,1]$. For any $n = 0, 1, 2, \ldots$, we partition I into subintervals, $I_j^{(n)} = [j\, 2^{-n}, (j+1)2^{-n})$, $j = 0, 1, \ldots, 2^n - 1$. Define

$\varphi_n = \sum_{j=0}^{2^n-1} j\, 2^{-n} \chi_{E_j^{(n)}}$, where we defined the sets $E_j^{(n)} = f^{-1}(I_j^{(n)})$. By construction, we have $\varphi_n \leq \varphi_{n+1} \leq f$ for any n. Furthermore, for any $x \in X$, $y = f(x) \in [0,1]$. Let $y_n = j_n\, 2^{-n}$ be the diadic integer with $0 \leq j_n \leq 2^n - 1$ s.t. $y \in I_{j_n}^{(n)}$. Then $y_n \to y$ as $n \to \infty$. On the other hand, since $x \in f^{-1}(I_{j_n}^{(n)})$, we have $\varphi_n(x) = y$. Therefore, $\varphi_n(x) \to f(x)$. ■

Consequently, simple functions will be building blocks of the theory of integration.

6. Integration of nonnegative functions

Consider a measure space $(X, \mathcal{M}, \mu)$ with characteristic functions χ_E on it (see definition (5.1)). We define the integral first for characteristic functions, and then we extend the notion of integral to more complicated functions.

Exercise. Show that $E \in \mathcal{M} \Leftrightarrow \chi_E$ is measurable.

The integral of χ_E with respect to the measure μ is defined as

$$\int \chi_E d\mu := \mu(E). \tag{6.1}$$

Notice that we allow $\mu(E) = \infty$. If μ is a probability measure, then the integral (6.1) is the probability that the event E occurs. Definition (6.1) is readily extended by linearity to simple functions: if φ is a simple function with standard representation $\varphi = \sum_1^n a_j \chi_{E_j}$, then we define the integral of φ (over X) with respect to μ as

$$\int \varphi d\mu := \sum_1^n a_j \mu(E_j). \tag{6.2}$$

Note again that we allow $\int \varphi d\mu = \infty$. If the measure μ is understood, then we simply write $\int \varphi$ for the left hand side of (6.2). We want now to extend the notion of integral to *all* nonnegative measurable functions on X (i.e. measurable functions from X to $[0, \infty)$). Let us denote the set of all nonnegative measurable functions by L^+. Remember that any

$f \in L^+$ can be approximated by a monotonically increasing sequence of simple functions (see the last theorem of Section 5). For $f \in L^+$, it is thus natural to define

$$\int f d\mu := \sup \left\{ \int \varphi d\mu : 0 \leq \varphi \leq f,\ \varphi \text{ is simple} \right\}. \tag{6.3}$$

This definition needs some justification. Namely, we have to show that if f is simple, definition (6.3) coincides with definition (6.2). This can be done using the following

Proposition. *If φ, f are simple functions, and $\varphi \leq f$, then $\int \varphi \leq \int f$. (The integrals are understood in the sense of (6.2))*

Proof. Represent φ and f in their standard form: $\varphi = \sum_j a_j \chi_{E_j}$ and $f = \sum_k b_k \chi_{F_k}$. Since $\varphi \leq f$, then

$$a_j \leq b_k \text{ if } E_j \cap F_k \neq \phi. \tag{6.4}$$

Now $\int \varphi = \sum_j a_j \mu(E_j) = \sum_{k,j} a_j \mu(E_j \cap F_k)$, where we used $E_j = E_j \cap X = E_j \cap (\cup_k F_k) = \cup_k (E_j \cap F_k)$, and the fact that $E_j \cap F_k$ is disjoint from $E_j \cap F_{k'}$, if $k \neq k'$ (that's why we put φ and f in their standard form!). We get thus from (6.4)

$$\begin{aligned} \int \varphi &\leq \sum_{k,j} b_k \mu(E_j \cap F_k) \\ &\leq \sum_k b_k \mu(\cup_j (E_j \cap F_k)) \\ &\leq \sum_k b_k \mu(F_k) \\ &= \int f. \blacksquare \end{aligned}$$

Corollary. *If f is simple, then (6.2) and (6.3) are equivalent.*

Proof. Due to the proposition above, the supremum on the r.h.s. of (6.3) is reached for $\varphi = f$. ■

The next result shows how we can approximate integrals of measurable functions by integrals of simple functions. For a proof, consult [F], Theorem 2.14.

Theorem (monotone convergence, MCT). *If* $\{f_n\} \subset L^+$, $f_n \leq f_{n+1}$ $\forall n$, $\lim_{n\to\infty} f_n = f$ *(pointwise in* $\overline{\mathbb{R}}_+$*), then* $\int f = \lim_{n\to\infty} \int f_n$.

Remark. Since any $f \in L^+$ can be approximated by a monotonically nondecreasing sequence of simple functions $\{\varphi_n\} \subset L^+$, we get from the MCT: $\int f = \lim_n \int \varphi_n$. Notice also that the monotonicity condition cannot be removed as shows the following example: let $f_n = \chi_{(n,n+1)}$ (the characteristic function of the interval $(n, n+1)$. Then $f_n \to 0$ (pointwise), but on the other hand, $\int f_n = 1$, $\forall n$!

Some basic properties of the integral are given by:

Theorem. *Let* $f, g \in L^+$. *Then*

(a) $\int cf = c \int f$, $\forall c \in \mathbb{R}$,

(b) $\int (f + g) = \int f + \int g$,

(c) if $g \leq f$, *then* $\int g \leq \int f$,

(d) the map $A \mapsto \int_A f$ *is a measure on* $\mathcal{M}$.

Exercise. Prove (a)–(d) for simple functions, then use this to prove (a)–(c) for functions in L^+ (hint: use the MCT for the second part).

Using these basic properites of the integral, we are now ready to prove some more refined results.

Theorem. *If* $\{f_n\}$ *is a finite or infinite sequence in* L^+ *and* $f = \sum_1^n f_n$, *then* $\int f = \sum_1^n \int f_n$.

Proof. By induction, the assertion holds clearly for any finite sum: $\int \sum_1^N f_n = \sum_1^N \int f_n$. If the sequence is infinite, then we take the limit $N \to \infty$. Notice that $f^{(N)} := \sum_1^N f_n$ is an increasing sequence, hence $f^{(N)} \to \sum_1^\infty f_n$ (pointwise in $\overline{\mathbb{R}}_+$). Therefore, by the MCT, $\lim_{N\to\infty} \int f^{(N)} = \int \lim_{N\to\infty} f^{(N)}$. ■

Theorem. *If $f \in L^+$, then $\int f = 0 \Leftrightarrow f = 0$ a.e.*

Exercise. Prove this theorem for $f \in L^+$, f simple. (A full proof is given in [F], Proposition 2.16)

Corollary. *If $\{f_n\} \subset L^+$, $f \in L^+$, $f_n \leq f_{n+1}$ $\forall n$, $\lim f_n = f$ a.e., then $\int f = \lim \int f_n$.*

Proof. Let E be s.t. $\mu(E^c) = 0$ and s.t. f_n increases to f for all $x \in E$. So we have $f = f\chi_E$ a.e. and $f_n = f_n\chi_E$ a.e., and $f_n\chi_E \uparrow f\chi_E$ (pointwise), hence by the MCT $\int f = \int f\chi_E = \int \lim f_n\chi_E = \lim \int f_n\chi_E = \lim \int f_n$. ■

Fatou's Lemma. *If $\{f_n\}$ is any sequence in L^+, then $\int \liminf f_n \leq \liminf \int f_n$.*

The proof is given in [F], 2.18.

Corollary. *If $\{f_n\} \subset L^+$, $f \in L^+$, and $f_n \to f$ a.e., then $\int f \leq \liminf \int f_n$.*

For a proof, see [F], 2.19.

7. Integration of complex functions

In this section, we extend the definition of the integral of nonnegative functions to the integral of real functions and then of complex functions. Let as before $(X, \mathcal{M}, \mu)$ be a fixed measure space. If f is a real measurable function on X, then we can write $f = f_+ - f_-$, where $f_+(x) := \max(f(x), 0)$, and $f_-(x) := \max(-f(x), 0)$. Both f_+ and f_-

are nonnegative and measurable, so $\int f_\pm$ is defined. We now define

$$\int f := \int f_+ - \int f_-. \tag{7.1}$$

If f is a complex valued measurable function on X, then Ref and Imf are real valued measurable functions on X, and we define

$$\int f := \int \mathrm{Re} f + i \int \mathrm{Im} f. \tag{7.2}$$

We say that f *is integrable* (on $E \in \mathcal{M}$) iff $\int |f| < \infty$ ($\int_E |f| < \infty$).

Theorem. *The set of real (complex) integrable functions on X forms a real (complex) vector space.*

Proof. To show that this set is a vector space, we have to show that if f and g are integrable, then so are cf ($c \in \mathbb{R}$ or $\mathbb{C}$) and $f+g$. But this is clear from the inequalities $|cf| \leq |c||f|$ and $|f+g| \leq |f|+|g|$. Remember that if $g \leq f$, then $\int g \leq \int f$. ■

We denote the space mentioned in the theorem by either of the following symbols: $L^1(X,\mu)$, $L^1(X)$, $L^1(\mu)$ or simply L^1, depending on what we want to emphasize.

Exercise. Prove that if $f \in L^1$, then $|\int f| \leq \int |f|$.
Now we formulate a basic convergence theorem:

Theorem (dominated convergence, DCT). *Let $\{f_n\} \subset L^1$ be a sequence of functions s.t. $f_n \to f$ a.e., and s.t. $|f_n| \leq g$ $\forall n$, for some nonnegative $g \in L^1$. Then $f \in L^1$ and $\int f_n \to \int f$.*

Proof. Since $|f_n(x)| \leq g(x)$ a.e., then $|f(x)| \leq g(x)$ a.e., but $g \in L^1$, so $f \in L^1$.

Let us assume f_n (and hence f) are real. In the complex case, just do the argument that follows for the real and imaginary part separately. We have $g \pm f_n \geq 0$ and apply Fatou's lemma: $\liminf \int (g \pm f_n) \geq \int \liminf (g \pm f_n) = \int (g \pm f)$. Therefore $\liminf \int \pm f_n \geq \int \pm f$. The plus sign yields $\liminf \int f_n \geq \int f$, and the minus sign yields $\liminf \int (-f_n) = -\limsup \int f_n \geq -\int f$. The combination of these two estimates gives $\int f \le \liminf \int f_n \le \limsup \int f_n \le \int f$, but this means $\liminf \int f_n = \limsup \int f_n = \lim \int f_n = \int f$. ■

Using the DCT, one can show (see e.g. [F], 2.25-2.27):

(a) If $\{f_n\} \subset L^1$ and $\sum_1^\infty \int |f_n| < \infty$, then $\sum_1^\infty f_n$ converges a.e. to an L^1 function, and $\sum_1^\infty \int f_n = \int \sum_1^\infty f_n$.

(b) If $f \in L^1$, then there is a sequence of simple functions $\{\varphi_n\} \subset L^1$ s.t. $\int \varphi_n \to \int f$.

(c) Suppose $f(x,t)$ is in $L^1(d\mu(x))$, $\forall t \in [\alpha, \beta]$, $|f(x,t)| \le g(x)$, uniformly in t, and $g \in L^1$. If also $\lim_{t\to t_0} f(x,t) = f(x,t_0)$ $\forall x$, then $\lim_{t\to t_0} \int f(x,t)d\mu(x) = \int f(x,t_0)d\mu(x)$.

(d) If $f(x,t)$ is in $L^1(d\mu(x))$, $\forall t \in [\alpha, \beta]$, $|\frac{\partial f}{\partial t}(x,t)| \le g(x)$, uniformly in t for some $g \in L^1$, then $\frac{\partial}{\partial t} \int f(x,t)d\mu(x) = \int \frac{\partial f}{\partial t}(x,t)d\mu(x)$.

The integral with respect to the Lebesgue measure is called the *Lebesgue integral.* One can show that if f is Riemann integrable on a finite interval $[a,b]$, then f is Lebesgue integrable on $[a,b]$, and both integrals coincide. The Lebesgue integral is denoted by either $\int f(x)dx$, $\int f dx$ or $\int f$, depending on what we want to display.

8. The Lebesgue integral on $\mathbb{R}^n$

Two key properties of the Lebesgue integral are

(a) translation invariance: $\int f(x+h)dx = \int f(x)dx, \quad \forall h \in \mathbb{R}^n$,

(b) rotation invariance: $\int f(\mathcal{R}x)dx = \int f(x)dx, \quad \forall \mathcal{R} \in \mathcal{O}(n)$.

Here, $\mathcal{O}(n)$ denotes the group of rotations of vectors in $\mathbb{R}^n$. E.g. for $n = 2$, $\mathcal{O}(2)$ consists of rotations $\mathcal{R}_\theta$ by an angle θ:

$$\mathcal{R}_\theta \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \tag{8.1}$$

Rotations are real $n \times n$ matrices $\mathcal{R}$ which satisfy $\mathcal{R}^T\mathcal{R} = \mathcal{R}\mathcal{R}^T = I$, where $\mathcal{R}^T$ is the transpose of $\mathcal{R}$, and I is the $n \times n$ unity matrix. Since

$\det \mathcal{R} = \det \mathcal{R}^T$, we have $\det \mathcal{R} = \pm 1$. We identify matrices with maps in the usual way as $(\mathcal{R}x)_k = \sum_l \mathcal{R}_{kl} x_l$.

Exercise. Show $\mathcal{R}^T\mathcal{R} = \mathcal{R}\mathcal{R}^T = I$ for $\mathcal{R}$ given in (8.1).

Sketch of proof of (a). We prove it first for characteristic functions χ_E. From $\chi_E(x+h) = \chi_{E-h}(x)$ (where $E - h = \{x - h : x \in E\}$), we get

$$\int \chi_E(x+h)dx = \int \chi_{E-h}(x)dx = m(E-h) = m(E) = \int \chi_E(x)dx,$$

thus (a) holds for characteristic functions, and hence for simple functions φ:

$$\int \varphi(x+h)dx = \int \varphi(x)dx.$$

Approximating any nonnegative measurable function by simple functions and using the DCT, we prove (a) for nonnegative measurable functions, then for real and complex valued functions.

Exercise. Fill in the details of this proof.

The proof of (b) follows from the formula:

$$\int f(x)dx = |\det T| \int f(Tx)dx,$$

where T is an $n \times n$ matrix with $\det T \neq 0$. This formula as well as its generalization given below are known from multivariable calculus. ■

The formula above is a special case of the following result:

$$\int_{G(\Omega)} f(x)dx = \int_{\Omega} f(G(x))|\det \mathcal{D}_x G|dx, \tag{8.2}$$

where $G : \Omega \to \mathbb{R}^n$ is a C^1–diffeomorphism and $\mathcal{D}_x G$ is the linear map of $\mathbb{R}^n$ given by the matrix

$$\mathcal{D}_x G = \left(\frac{\partial G_k}{\partial x_l}\right),$$

where $G(x) = (G_1(x), \ldots, G_n(x))$. For a proof of (8.2), consult e.g. [F], Theorem 2.43 and 2.47. Let us finally recall the definition of a C^1–diffeomorphism. A function $G : \Omega \to \mathbb{R}^n$ is called a C^1–diffeomorphism iff

(i) G is injective (meaning $G(x_1) = G(x_2) \Rightarrow x_1 = x_2$),

(ii) $\mathcal{D}_x G$ is invertible for all $x \in \Omega$.

9. L^p–spaces

A great part of this course is devoted to the study of properties of maps of functions. Such maps are defined on spaces of functions. In this section, we introduce the simplest and most commonly used spaces of functions. Consider a fixed measure space $(X, \mathcal{M}, \mu)$. We define the L^p–space for $1 \leq p < \infty$:

$$L^p(X,\mu) := \{f : X \to \mathbb{C} | f \text{ is measurable, and} \int |f|^p d\mu < \infty\}. \quad (9.1)$$

In other words, $f \in L^p(X,\mu) \Leftrightarrow |f|^p \in L^1(X,\mu)$. We also define the L^∞– space:

$$L^\infty(X,\mu) := \{f : X \to \mathbb{C} | f \text{ is measurable, and } \operatorname{ess\,sup}|f| < \infty\}. \quad (9.2)$$

Here, recall that $\operatorname{ess\,sup}|f| := \inf\{\sup |g| : g = f \text{ a.e.}\}$. We often use the following abbreviations for $L^p(X,\mu)$: $L^p(X)$, $L^p(\mu)$, or simply L^p, depending on which part of the measure space $(X, \mathcal{M}, \mu)$ we care to display. L^p, $1 \leq p \leq \infty$ is a vector space. For $p = \infty$, this is obvious, and for $1 \leq p < \infty$, it easily follows from the inequality

$$|f+g|^p \leq 2^p \left(|f|^p + |g|^p\right). \quad (9.3)$$

The latter inequality is obtained as follows: $|f+g|^p \leq (2\max(|f|,|g|))^p \leq 2^p(|f|^p + |g|^p)$.

We define for every $f \in L^p$:

$$||f||_p := \begin{cases} \left(\int |f|^p\right)^{1/p} & \text{if } 1 \leq p < \infty, \\ \operatorname{ess\,sup}|f| & \text{if } p = \infty. \end{cases} \quad (9.4)$$

Clearly, $||f||_p = 0 \Leftrightarrow f = 0$ a.e., and $||cf||_p = |c|\,||f||_p$, $\forall c \in \mathbb{C}$. We have also the *triangle inequality* $||f+g||_p \leq ||f||_p + ||g||_p$, which we prove later.

From these properties, it follows that the map $f \mapsto ||f||_p$ is a *norm*. A vector space equipped with a norm is called a *normed vector space*. Hence L^p is a normed vector space, for every $1 \leq p \leq \infty$. Having defined a norm, we can define the notion of (norm–)convergence as follows. Let $\{f_n\} \subset L^p$ be a sequence. We say that f_n converges to f $(\in L^p)$, iff $||f_n - f||_p \to 0$. We write $f_n \to f$.

A normed vector space B is called *complete* iff every Cauchy sequence converges, i.e. if $\{f_n\} \subset B$ is a Cauchy sequence (meaning $||f_n - f_m|| \to 0$, as $m, n \to \infty$, where $|| \cdot ||$ denotes the norm of B) then $\{f_n\}$ converges (i.e. there is a $f \in B$ such that $||f_n - f|| \to 0$, as $n \to \infty$). Remark that the converse is always true: any convergent sequence is necessarily a Cauchy sequence. Completeness is a very important property of a normed vector space, e.g. when one solves equations. Indeed, often one solves equations by successive approximations, and one wants to know that such approximate solutions converge (to the actual solution). A complete normed vector space is called a *Banach space*.

Theorem. *For $1 \leq p \leq \infty$, L^p is a Banach space. Furthermore, simple functions are dense in L^p (i.e. for every $f \in L^p$, there is a sequence $\{\varphi_n\}$ of simple functions such that $||f - \varphi_n||_p \to 0$, as $n \to \infty$).*

The proof is given in [F], Theorem 6.6, Proposition 6.7 and Theorem 6.8. The second statement of the theorem can be derived from the construction in Section 5. We give it here as an

Exercise. Let φ_n be the simple function constructed in the proof at the end of Section 5, so that $\varphi_n \uparrow f$ p.w. Show that $\varphi_n \to f$ in L_p, provided $f \in L^p$. Hint: consider first the case $f \geq 0$ and use that $|f - \varphi_n|^p \to 0$ p.w., $|f - \varphi_n|^p \leq 2^p(|f| + |\varphi_n|^p) \leq 2^{p+1}|f|^p$ and the Dominated Convergence Theorem.

There are two basic inequalities:

1. *Hölder's inequality*: let $1 \leq p \leq \infty$, $p^{-1} + q^{-1} = r^{-1} \leq 1$, and $f \in L^p$, $g \in L^q$ then $fg \in L^r$ and $||fg||_r \leq ||f||_p ||g||_q$,
2. *Minkowski's inequality*: let $1 \leq p < \infty$ and $f, g \in L^p$, then $||f + g||_p \leq ||f||_p + ||g||_p$.

We prove Hölder's inequality. A proof of Minkowski's inequality can be found in [F], 6.5. Observe first that it suffices to prove the Hölder inequality for the case $r = 1$ (this follows easily from $||fg||_r = (\int |fg|^r)^{1/r} = (\int |f|^r|g|^r)^{1/r}$, and e.g. calling $f^r = f_1$ and $g^r = g_1$). Next notice that the result is trivial if $||f||_p = 0$ or $||g||_p = 0$ (for then $f = 0$ a.e. or $g = 0$ a.e.), or if $||f||_p = \infty$ or $||g||_p = \infty$. If neither of these cases hold, then we can define

$$a = \left|\frac{f(x)}{||f||_p}\right|^p, \quad b = \left|\frac{g(x)}{||g||_q}\right|^q \quad \text{and} \quad \lambda = \frac{1}{p}.$$

Below, we will show that for any $a, b > 0$, and $0 < \lambda < 1$:

$$a^\lambda b^{1-\lambda} \leq \lambda a + (1-\lambda)b. \tag{9.5}$$

Applying this to our case, we get

$$\frac{|f(x)g(x)|}{||f||_p||g||_q} \leq \frac{|f(x)|^p}{p\int |f|^p} + \frac{|g(x)|^q}{q\int |g|^q}.$$

Integrating this inequality, and using $p^{-1} + q^{-1} = 1$, we arrive at Hölder's inequality. We finish the proof by showing (9.5). We can assume $b \neq 0$, so we can divide (9.5) by b on both sides. (9.5) is then equivalent to the inequality (where $t = a/b$)

$$t^\lambda - \lambda t - 1 - \lambda \leq 0.$$

Since $0 < \lambda < 1$, the maximum of the function on the l.h.s. is reached at $t = 1$, and is equal to 0. ■

Remark. Inequality (9.5) is a special case of the very useful *Jensen's inequality*: let φ be a convex function on $[a, b]$, and p_k positive numbers satisfying $\sum_1^n p_k = 1$. We can think about p_k as probabilities. Then $\varphi(\sum_1^n p_k t_k) \leq \sum_1^n p_k \varphi(t_k)$, for all $t_k \in [a, b]$. This inequality indeed implies (9.5) for $\varphi(t) = e^t$.

Exercise. Let $g \in L^\infty$. Show that the operator $T : L^p \to L^p$ defined by $f \mapsto Tf = gf$ (multiplication operator) satisfies $||Tf||_p \leq ||g||_\infty ||f||_p$. In other words, $T : L^p \to L^p$ is bounded.

Exercise. Prove Jensen's inequality.

The L^2 space is special: it has an additional structure - the inner product

$$\langle f, g\rangle = \int \overline{f} g$$

which allows, in addition to sizes and distances, to measure angles, e.g.

$$f \perp g \Leftrightarrow \langle f, g\rangle = 0.$$

Recall that the map $(f, g) \to \langle f, g\rangle \in \mathbb{C}$ is called an *inner product* iff

- $\langle f, g\rangle$ is linear in the second argument, i.e. $\langle f, \alpha g + \beta h\rangle = \alpha \langle f, g\rangle + \beta \langle f, h\rangle$, for any $\alpha, \beta \in \mathbb{C}$,
- $\langle f, g\rangle = \overline{\langle g, f\rangle}$ (this together with the above implies that $\langle f, g\rangle$ is anti-linear in the first argument: $\langle \alpha f + \beta h, g\rangle = \overline{\alpha} \langle f, g\rangle + \overline{\beta} \langle h, g\rangle$),
- $\langle f, f\rangle \geq 0$ with equality iff $f = 0$.

We remark that sometimes, the scalar product is taken to be linear in the first argument and anti-linear in the second one.

An inner product defines a norm according to

$$\|f\| = \sqrt{\langle f, f\rangle}. \tag{9.6}$$

Exercise. Check that (9.6) defines a norm. (Hint: to prove the triangle inequality, use the Schwartz inequality $|\langle f, g\rangle| \leq \|f\| \, \|g\|$ which, in turn, follows from the obvious relation

$$0 \leq \|u - v\|^2 = \langle u \pm v, u \pm v\rangle = \|u\|^2 + \|v\|^2 \pm \langle u, v\rangle \pm \langle v, u\rangle$$

applied to $u = \pm f/\|f\|$ and $v = g/\|g\|$ and to $u = \pm if/\|f\|$ and $v = g/\|g\|$. Observe that for L^2, the Schwartz inequality is a special case of the Hölder inequality: take $p = q = 2$ and $r = 1$.)

Thus a space with an inner product, or an *inner product space*, is also a normed space. A complete inner product space is called a *Hilbert space.* By definition, a Hilbert space is also a Banach space.

Chapter II. Transforms and Distributions

10. Convolution

We consider L^p spaces for the Lebesgue measure m $(dm = dx)$: $L^p(\mathbb{R}^n) = L^p(\mathbb{R}^n, dx)$. We use the following notaion: for a multi-index $\alpha = (\alpha_1, \dots, \alpha_n)$, where the entries $\alpha_1, \dots, \alpha_n$ are nonnegative integers. Let us define

$$|\alpha| = \sum_1^n \alpha_j, \quad \alpha! = \prod_i^n \alpha_j!, \quad \partial^\alpha = \prod_1^n \partial_{x_j}^{\alpha_j}, \quad x^\alpha = \prod_1^n x_j^{\alpha_j}.$$

For two (Lebesgue–)measurable functions f and g, we define their *convolution* $f * g$ as the measurable function given by:

$$(f * g)(x) := \int f(x-y)g(y)dy.$$

By Hölder's inequality, if $f \in L^p$ and $g \in L^q$ with $1/p + 1/q = 1$, then the integral on the r.h.s. is well defined.

Proposition. *The convolution has the following properites:*

(a) $f * g = g * f$*,*

(b) $(f * g) * h = f * (g * h)$*,*

(c) if f and g are $|\alpha|$–times differentiable, then $\partial^\alpha(f * g) = (\partial^\alpha f) * g = f * (\partial^\alpha g)$.

Exercise. Prove this proposition.
The basic inequality for the convolution is

Young's inequality. *Let $1 \leq p, q, r \leq \infty$ and $p^{-1} + q^{-1} = 1 + r^{-1}$. If $f \in L^p$ and $g \in L^q$, then $f * g \in L^r$, and $||f * g||_r \leq ||f||_p ||g||_q$.*

Proof. For $r = \infty$, the proof follows from Hölder's inequality. For $r = 1$, it follows from a change of variables (notice $r = 1 \Rightarrow p = q = 1$):

$$\begin{aligned} ||f * g||_1 &= \left\| \int f(\cdot - y) g(y) dy \right\|_1 \\ &\leq \int ||f(\cdot - y)||_1 |g(y)| dy \\ &= \int ||f||_1 |g(y)| dy \\ &= ||f||_1 ||g||_1. \end{aligned}$$

We used the fact that the Lebesgue measure is translation invariant. The proof for $1 < r < \infty$ follows by the Riesz–Thorin interpolation theorem [F], §6.27. ■

11. Approximation of identity

Let $\varphi \in L^1(\mathbb{R}^n)$ such that $\int \varphi = 1$. Define

$$\varphi_t(x) = t^{-n} \varphi(x/t). \tag{11.1}$$

Then we have $\int \varphi_t = \int \varphi = 1 \ \forall t$.

Theorem. *Let φ as in (11.1), and $f \in L^p$. Then $\varphi_t * f \to f$ as $t \to 0$. The convergence is in L^p if $1 \leq p < \infty$, and uniformly on compact sets if $p = \infty$.*

A proof is given in [F], Theorem 8.14.

To demonstrate the ideas involved in proving this theorem, we prove that for any function $f \in C^1$ which, together with its first derivatives, is bounded, we have

$$\varphi_t * f \longrightarrow f, \quad \text{as } t \to 0,$$

where the convergence is in the supremum (i.e. L^∞) norm. Indeed, since $\int \varphi_t = 1$, we obtain

$$(\varphi_t * f)(x) - f(x) = \int \varphi_t(x-y)\,(f(y) - f(x))\,d^n y.$$

Splitting the integral on the r.h.s. into integrals over the domains

$$\{y \in \mathbb{R}^n : \, |x-y| \leq \sqrt{t}\} \quad \text{and} \quad \{y \in \mathbb{R}^n : \, |x-y| \geq \sqrt{t}\},$$

and estimating the integrand in the first integral by

$$\sup_{|x-y| \leq \sqrt{t}} |f(y') - f(x)| \varphi_\epsilon(y) \leq \sup_{y'} |\nabla f(y')| \sqrt{t}\, \varphi_\epsilon(y),$$

and in the second integral by

$$\sup_{|x-y| \leq \sqrt{t}} |f(y') - f(x)| \varphi_\epsilon(y) \leq \sup_{y'} |f(y')| \varphi_\epsilon(y),$$

we arrive at

$$\begin{aligned} |(\varphi_t * f)(x) - f(x)| \;\; &\leq \;\; \sup_{y'} |\nabla f(y')| \sqrt{t} \int_{|y-x| \leq \sqrt{t}} \varphi_t(x-y) d^n y \\ &\quad + 2 \sup_{y'} |f(y')| \int_{|y-x| \geq \sqrt{t}} \varphi_t(x-y) d^n y. \end{aligned}$$

Now changing the variables of integration as $z = x - y$, we find for the integrals on the r.h.s.

$$\begin{aligned} \int_{|z| \leq \sqrt{t}} \varphi_t(z) d^n z \;\; &\leq \;\; \int \varphi_t = 1 \\ \int_{|z| \geq \sqrt{t}} \varphi_t(z) d^n z \;\; &= \;\; \int_{|z'| \geq t^{-1/2}} \varphi_t(z') d^n z', \end{aligned}$$

where in the second integral, we have changed the variables a second time as $z' = z/t$, and have used that $\varphi_t(z) = t^{-n}\varphi(z/t)$. This gives

$$|(\varphi_t * f)(x) - f(x)| \leq \sup_y |\nabla f(y)|\sqrt{t} + 2\sup_y |f(y)| \int_{|z|\geq t^{-1/2}} \varphi(z)d^n z,$$

and the r.h.s. converges to zero as $t \to 0$. ■

Observe that if φ is smooth, then so is $\varphi_t * f$, whatever rough f is (this follows from *c)* of the last proposition). Thus in this case, $\varphi_t * f$ gives a smooth approximation of f. The operator $f \mapsto \varphi_t * f$ is called an *approximation of identity.*

12. Fourier transform

In this section, we describe one of the most powerful tools in analysis – the Fourier transform. This transform allows us to analyze a fine structure of functions and to solve differential equations. The Fourier transform takes functions of time to functions of frequencies, functions of coordinates to functions of momenta, and vice versa.

Initially, we define the Fourier transform on the Schwartz space $\mathcal{S}(\mathbb{R}^n) = \mathcal{S}$:

$$\mathcal{S} = \{f \in C^\infty(\mathbb{R}^n) : <x>^N |\partial^\alpha f(x)| \text{ is bounded } \forall N \text{ and } \forall \alpha\}, \quad (12.1)$$

where $<x>= (1 + |x|^2)^{1/2}$. On $\mathcal{S}$, we define the Fourier transform $\mathcal{F} : f \mapsto \hat{f}$ by

$$\hat{f}(k) = (2\pi)^{-n/2} \int f(x)e^{-ik\cdot x}dx. \quad (12.2)$$

The next theorem gives the important example of the Fourier transform - the Fourier transform of a Gaussian:

Theorem. *Let A be a $n \times n$ matrix s.t.* $\mathrm{Re}A := (A + A^*)/2$ *is positive definite (i.e. $x \cdot \mathrm{Re}A\, x > 0$ if $x \neq 0$). Then we have*

$$\mathcal{F} : e^{-x\cdot Ax} \mapsto (2\pi)^{-n/2}(\det A)^{-1/2}e^{-k\cdot A^{-1}k} \quad (12.3)$$

Proof. We prove the theorem only for positive definite matrices. If A is positive definite (i.e. if $x \cdot Ax > 0$ for $x \neq 0$), then there is an orthogonal matrix U (i.e. U is real and $UU^T = U^TU = \mathrm{id}$) s.t. $\overline{A} := U^TAU$ is diagonal, say $\overline{A} = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$. Letting $x = Uy$ and noticing that $x \cdot Ax = y \cdot U^TAUy$, and that $\det U = 1$, we get

$$\int e^{-x\cdot Ax}e^{-ik\cdot x}dx = \int e^{-y\cdot \bar{A}y}e^{ik'\cdot y}dy = \prod_1^n \int e^{-\lambda_j y_j^2}e^{ik_j' y_j}dy_j,$$

where $k' = U^Tk$, and we have used $k \cdot Uy = U^Tk \cdot x$.

Exercise. Show that for $n = 1$, $\mathcal{F} : e^{-\lambda x^2} \mapsto (2\pi)^{-1/2}e^{-k^2/\lambda}$.

The last two relations imply the desired statement. ■

The function $e^{-x\cdot Ax}$ is called a Gaussian. It is one of the most common functions in applications. There is another important function whose Fourier transform can be explicitely computed:

$$\mathcal{F} : |x|^{-\alpha} \mapsto \begin{cases} C_{n,\alpha}|k|^{-n+\alpha} & \text{if } \alpha \neq n, \\ C_{n,n}\ln|k| & \text{if } \alpha = n. \end{cases} \tag{12.4}$$

The coefficients are given for $\alpha = 2$ by

$$C_{n,2} = \begin{cases} ((2-n)\sigma_{n-1})^{-1} & \text{for } n \neq 2, \\ \sigma_{n-1} = (2\pi)^{-1} & \text{for } n = 2, \end{cases} \tag{12.5}$$

where σ_n is the volume of the n–dimensional unit spere $S^n = \{x \in \mathbb{R}^{n+1} : |x| = 1\}$. One can easily deduce formula (12.4) modulo the constants (12.5). Indeed, since $|x|^{-\alpha}$ is rotationally invariant, then so is its Fourier transform. Also, since $|x|^{-\alpha}$ is homogeneous of degree $-\alpha$, then its Fourier transform is homogeneous of degree $-n+\alpha$. Hence (12.4) follows. Though it is easy to compute the Fourier transform of $|x|^{-\alpha}$, it is not easy to justify it. Indeed, the function $|x|^{-\alpha}$ is rather singular and definitely does not belong to $\mathcal{S}(\mathbb{R}^n)$.

Exercise. For $n = 1$, compute the Fourier transform of the characteristic function $\chi_{(-a,a)}(x)$, using definition (12.2).

Define also

$$\check{f}(x) = (2\pi)^{-n/2}\int f(k)e^{ix\cdot k}dk. \tag{12.6}$$

Some key properties of the Fourier transform are collected in the following

Theorem. *Assume $f, g \in \mathcal{S}(\mathbb{R}^n)$. Then we have:*

(a) $(-i\partial)^\alpha f \mapsto k^\alpha \hat{f}$, and $x^\alpha f \mapsto (-i\partial)^\alpha \hat{f}$,

(b) $\int \hat{f} g = \int f \hat{g}$,

(c) $\overline{\hat{f}} = \check{\overline{f}}$,

(d) $\int \overline{\hat{f}} \hat{g} = \int f g$,

(e) $fg \mapsto (2\pi)^{-n/2} \hat{f} * \hat{g}$, and $f * g \mapsto (2\pi)^{n/2} \hat{f} \hat{g}$,

(f) $(\hat{f})\check{} = f = (\check{f})\hat{}$.

Properties (a) - (e) hold (with signs changed in (a)) also when $\hat{}$ is replaced by $\check{}$.

Proof. We give a formal proof. Integrating by parts, we compute

$$\begin{aligned} -i(\partial_{x_j} f)\hat{}(k) &= (2\pi)^{-n/2} \int (-i)\partial_{x_j} f(x) e^{-ik\cdot x} dx \\ &= (2\pi)^{-n/2} \int f(x) i\partial_{x_j} e^{-ik\cdot x} dx \\ &= k_j \hat{f}(k). \end{aligned}$$

Exercise. Prove the remaining relations in (a), and prove properties (b) and (c) (formally, without justification of the interchange of the order of integration etc.).

Statement (d) is called the *Plancherel Theorem.* The *adjoint* $\mathcal{F}^*$ of the Fourier transform is defined by $\langle \mathcal{F}^* u, v \rangle = \langle u, \mathcal{F} v \rangle$ for all $u, v \in \mathcal{S}(\mathbb{R}^n)$, where $\langle \cdot, \cdot \rangle$ is the standard inner product in $L^2(\mathbb{R}^n)$. Then (b) and the relation $\hat{\overline{f}} = \overline{\hat{f}}$ show that $\langle f, \mathcal{F} g \rangle = \int \overline{f} \hat{g} = \int \hat{\overline{f}} g = \int \overline{\check{f}} g$, so $\check{f} = \mathcal{F}^* f$. This together with (f) implies that $\mathcal{F}\mathcal{F}^* = \mathrm{id} = \mathcal{F}^*\mathcal{F}$ on $\mathcal{S}$, which is a restatement of the Plancherel theorem.

Now we derive property (e) from properties (b) and (f). Indeed, set $e_k(x) := e^{-ikx}$. Using that $(e_kg)\check{}(k') = \hat{g}(k-k')$, we obtain

$$\begin{aligned}(fg)\check{}(k) &= (2\pi)^{-n/2}\int e_kg\ (\hat{f})\check{} = (2\pi)^{-n/2}\int (e_kg)\check{}\hat{f} \\ &= (2\pi)^{-n/2}\int \hat{g}(k-k')\hat{f}(k')dk' = (2\pi)^{-n/2}(\hat{f}*\hat{g})(k),\end{aligned}$$

where we used (f) in the first equality and (b) (with $\hat{}$ replaced by $\check{}$) in the second one.

The proof of (f) is more subtle. We use an approximation of unity $\varphi_t(x) = t^{-n}\varphi(x/t)$ and compute $\varphi_t * (\hat{f})\check{}$. Let us define $\varphi^x(y) := \varphi(x-y)$. Using property (b), we find

$$\varphi_t * (\hat{f})\check{} = \int \varphi_t^x \cdot (\hat{f})\check{}dy = \int (\varphi_t^x)\check{}\hat{f}dy = \int ((\varphi_t^x)\check{})\hat{}f dy.$$

Exercise. Show that

$$((\varphi_t^x)\check{})\hat{} = ((\hat{\varphi}_t)\check{})^x = t^{-n}(\hat{\varphi})\check{}\left(\frac{x-y}{t}\right).$$

Thus we have

$$\varphi_t * (\hat{f})\check{} = ((\hat{\varphi})\check{})_t * f \tag{12.7}$$

We can choose φ such that $(\hat{\varphi})\check{} \in L^1$, and $\int (\hat{\varphi})\check{}(x)dx = 1$. Indeed, take e.g. $\varphi(x) = (4\pi)^{-n/2}e^{-|x|^2}$ and use the fact that $((e^{-|x|^2})\hat{})\check{} = e^{-|x|^2}$. With this in mind, we take the limit $t \to 0$ in (12.7) and use the properties of the approximation of identity to get

$$\varphi_t * (\hat{f})\check{} \to (\hat{f})\check{} \quad \text{and} \quad ((\hat{\varphi})\check{})_t * f \to f \quad \text{as } t \to 0$$

to obtain $(\hat{f})\check{} = f$. Similaly one shows that $(\check{f})\hat{} = f$. ■

Corollary. *$\mathcal{F}$ extends to a unitary operator on L^2, i.e. to a bounded operator satisfying $\mathcal{F}^* = \mathcal{F}^{-1}$.*

The Hausdorff-Young inequality. *Let $1 \le p \le 2$ and $p^{-1}+q^{-1} = 1$. Then $||\hat{f}||_q \le ||f||_p$. Consequently $\mathcal{F}$ extends to a bounded operator* *from L^p to L^q.*

Proof. Clearly, $||\hat{f}||_\infty \leq ||f||_1$. Moreover, we have shown that $||\hat{f}||_2 = ||f||_2$. For $1 < p < 2$, the result follows from the Riesz-Thorin theorem, [F], section 8.4. We omit the proof of the second statement. ■

Theorem (Riemann-Lebesgue Lemma). *Suppose that g is s.t. $\hat{g} \in L^1$. Then*

i) g is bounded and continuous,

ii) g decays at infinity: $lim_{|x|\to\infty} g(x) = 0$.

Proof. i) The boundedness is easily seen: $\forall x$,

$$|g(x)| = \left| \int e^{ikx} \hat{g}(k) \right| \leq \int |\hat{g}(k)| = \|\hat{g}\|_{L^1}.$$

Next, we show continuity. Since $\hat{g} \in L^1$, then

$$\lim_{h\to 0} (g(x+h) - g(x)) = \lim_{h\to 0} \int e^{ik\cdot x} \left(e^{ik\cdot h} - 1 \right) \hat{g}(k) dk = 0,$$

by the dominated convergence theorem ($|e^{ik\cdot h} - 1||\hat{g}| \leq 2|\hat{g}|$). This shows that g is continuous. Next, let us show *ii)*. Since the Schwartz space $\mathcal{S}$ is dense in L^1, there is a sequence $\varphi_j \in \mathcal{S}$ such that $||\varphi_j - \hat{g}||_1 \to 0$ as $j \to 0$. Thus

$$||\check{\varphi}_j - g||_\infty \leq \int |\varphi_j(k) - \hat{g}(k)| = ||\varphi_j - \hat{g}||_1 \to 0,$$

which shows that $\check{\varphi}_j \to g$ uniformly on $\mathbb{R}^n$. But $\check{\varphi}_j \in \mathcal{S}$, so $\check{\varphi}_j \to 0$ as $|x| \to \infty$, and therefore $g \to 0$ as $|x| \to \infty$. ■

13. Application of the Fourier transform to partial differential equations

Our goal in this section is to apply the Fourier transform in order to solve elementary but very basic partial differential equations (PDE's).

The Poisson equation on $\mathbb{R}^n$:

$$-\Delta u = f, \tag{13.1}$$

where $u : \mathbb{R}^n \to \mathbb{R}$ is an unknown function, $f : \mathbb{R}^n \to \mathbb{R}$ is a given function, and Δ is the Laplace operator (the Laplacian):

$$\Delta u = \sum_{j=1}^{n} \frac{\partial^2 u}{\partial x_j^2}.$$

The Poisson equation first appeared in the problem of determining the electric potential $u(x)$, created by a given charge distribution $\rho(x) = f(x)/(4\pi)$. Since then, it came up in various fields of mathematics, physics, engineering, chemistry, biology and economics.

In order to solve the Poisson equation, we apply the Fourier transform to both sides of (13.1) to obtain:

$$|k|^2 \hat{u}(k) = \hat{f}(k).$$

This equation can be easily solved: $\hat{u} = \hat{f}/|k|^2$. We can now apply the inverse Fourier transform to the last equality to get

$$u = \check{g} * f, \quad \text{where} \quad g(k) = |k|^{-2}. \tag{13.2}$$

But the inverse Fourier transform of $g(k) = |k|^{-2}$ is known:

$$\check{g}(x) = \begin{cases} [(2-n)\sigma_{n-1}]^{-1}|x|^{-n+2} & \text{if } n \neq 2 \\ [2\pi]^{-1} \ln |x| & \text{if } n = 2, \end{cases}$$

where σ_n is the volume of the unit–sphere $S_n = \{x \in \mathbb{R}^{n+1} : |x| = 1\}$ in dimension n.

Explicitely, (13.2) can be written as

$$u(x) = [(2-n)\sigma_{n-1}]^{-1} \int \frac{f(y)}{|x-y|^{n-2}} dy,$$

for $n \neq 2$, and similarly for $n = 2$. In particular, for $n = 3$, we have the celebrated Newton formula

$$u(x) = -\frac{1}{4\pi} \int \frac{f(y)}{|x-y|} dy.$$

Of course, the functions appearing in the above derivation are not necessarily from the Schwartz space $\mathcal{S}$ and therefore these manipulations must be justified. We leave this as an exercise, while proceeding in a similar fashion with other equations.

The heat equation on $\mathbb{R}^n$:

$$\frac{\partial u}{\partial t} = \Delta u \quad \text{and} \quad u|_{t=0} = u_0, \tag{13.3}$$

where $u : \mathbb{R}^n_x \times \mathbb{R}^+_t \to \mathbb{R}$ is an unknown function, and $u_0 : \mathbb{R}^n \to \mathbb{R}$ is a given initial condition. Problem (13.3) is called an *initial value problem.* It first appeared in the theory of heat diffusion. In that case, $u_0(x)$ is a given distribution of temperature in a body at time $t = 0$, and $u(x,t)$ is the unknown temperature–distribution at time t. Presently, this equation appears in various fields of science, including mathematical modeling of stock markets.

As before, we apply the Fourier transform to (13.3) and solve the resulting equation

$$\frac{\partial \hat{u}}{\partial t} = -|k|^2 \hat{u} \quad \text{and} \quad \hat{u}|_{t=0} = \hat{u}_0$$

to get $\hat{u} = e^{-|k|^2 t}\hat{u}_0$. Applying the inverse Fourier transform, and using that $(e^{-|k|^2 t})\check{} = (4\pi t)^{-n/2} e^{-|x|^2/(4t)}$, we obtain

$$u = (4\pi t)^{-n/2} e^{-|x|^2/(4t)} * u_0. \tag{13.4}$$

Remark. Define $\varphi(x) = (2\pi)^{-n/2} e^{-|x|^2/2}$ and $\varphi_s(x) = s^{-n}\varphi(x/s)$. Then

$$u = \varphi_{\sqrt{2t}} * u_0.$$

In particular, $u \to u_0$ as $t \to 0$, as it should be (c.f. the theorem after (11.1)).

Formula (13.4) shows that the heat diffuses over the smaple with velocity $\sim \sqrt{t}$.

The Schrödinger equation on $\mathbb{R}^n$:

$$i\frac{\partial \psi}{\partial t} = -\Delta \psi \quad \text{and} \quad \psi|_{t=0} = \psi_0. \tag{13.5}$$

This is also an initial value problem for the unknown function $\psi : \mathbb{R}^n_x \times \mathbb{R}^+_t \to \mathbb{C}$. Equation (13.5) describes the motion of a free quantum particle. Proceeding as with the heat equation, we obtain

$$\psi = (4\pi i t)^{-n/2} e^{i|x|^2/(4t)} * \psi_0. \tag{13.6}$$

Observe that this formula can be obtained from (13.4) by performing the substitution $t \to t/i$.

Exercise. Derive equation (13.6) using the Fourier transform.

The wave equation on $\mathbb{R}^n$:

$$\frac{\partial^2 u}{\partial t^2} = \Delta u \quad \text{and} \quad u|_{t=0} = u_0 \quad \text{and} \quad \partial_t u|_{t=0} = u_1. \tag{13.7}$$

This is a second order equation in time and consequently, it has two initial conditions u_0 and u_1. The wave equation (13.7) describes various wave phenomena: propagation of light and sound, oscillations of strings, etc. Proceeding as with the heat equation, we find

$$u = \partial_t W_t * u_0 + W_t * u_1, \tag{13.8}$$

where $W_t(x)$ is the inverse Fourier transform of the function $\sin(|k|t)/|k|$. The latter can be computed explicitely for $n = 1, 2, 3$:

$$W_t(x) = \begin{cases} \frac{1}{2}\chi_{\rho^2 \geq 0} & \text{for } n = 1, \\ (2\pi)^{-1}\rho^{-1}\chi_{\rho^2 \geq 0} & \text{for } n = 2, \\ (2\pi)^{-1}\delta(\rho^2) & \text{for } n = 3, \end{cases}$$

where $\rho^2 = t^2 - |x|^2$, and $\chi_{\rho^2 \geq 0}$ stands for the characteristic function of the set $\{(x,t) \in \mathbb{R}^{3+1} : \rho^2 \geq 0\}$, i.e.

$$\chi_{\rho^2 \geq 0} = \begin{cases} 1 & \text{if } \rho^2 \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

and $\delta(x)$ is the δ–function, a generalized function, or distribution, which we study in the next section.

Thus the dependence of W on x and t comes through the combination $\rho^2 = t^2 - |x|^2$, which is the Minkowski–distance in space–time, playing a crucial role in relativity. Observe that $\chi_{\rho\geq 0} = \chi_{|x|\leq t}$ and $\delta(\rho^2) = (2t)^{-1}\delta(t-|x|)$.

Exercise. Prove (13.8), and find $W_t(x)$ for $n=1$.

We examine closer the special case when $n=3$ and $u_0=0$. Then we get

$$\begin{aligned} u = W_t * u_1 &\equiv \frac{1}{4\pi t}\int \delta(|x-y|-t)u_1(y)dy \\ &= \frac{1}{4\pi t}\int_{S(x,t)} u_1(y)dS(y) \\ &= \frac{t}{4\pi}\int_{S(0,1)} u_1(x+tz)dS(z), \end{aligned}$$

where $S(x,t) = \{y \in \mathbb{R}^3 : |y-x| = t\}$ is a sphere of radius t centered at x. We see that only the initial condition evaluated on the sphere $S(x,t)$ matters in order to determine the solution at time t and at position x. This is called the *Huygens' principle.*

14. Dirac's δ–function

In the 1920's, in connection with his work on quantum mechanics, the British physicist P.A.M. Dirac introduced the following peculiar "function":

$$\delta : \mathbb{R} \to \mathbb{R}, \quad \delta(x) = \begin{cases} \infty & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}, \quad \text{and} \quad \int_{-\infty}^{\infty} \delta(x) = 1.$$

Though this function has no mathematical meaning, one can imagine how it has to look like: think about an infinitely high spike concentrated at $x=0$. About 25 years after its introduction, the French mathematician L. Schwartz gave a rigorous definition of Dirac's δ–function as a linear functional:

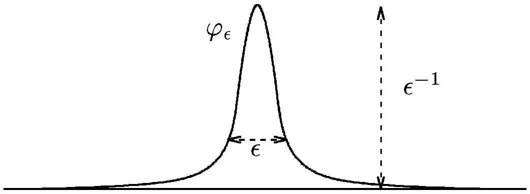
$$\delta : f \mapsto f(0),$$

defined on some appropriate space of functions, say on $C_0^\infty(\mathbb{R})$, the space of infinitely differentiable functions on $\mathbb{R}$ that have compact support. Though this definition makes perfect mathematical sense, it is not very intuitive.

We consider now a picture which is a compromise between Dirac's and Schwartz's views. Let $\varphi \in C_0^\infty(\mathbb{R})$, $\text{supp}\varphi \subset [-1,1]$, $\varphi \geq 0$, $\varphi(0) > 0$ and $\int_{-\infty}^{\infty} \varphi(x)dx = 1$. Define the approximation of identity $\varphi_\epsilon(x) := \epsilon^{-1}\varphi(x/\epsilon)$. We claim that in some appropriate sense we have

$$\varphi_\epsilon \to \delta \quad \text{as } \epsilon \to 0.$$

Intuitively, it is clear that as $\epsilon \to 0$, $\varphi_\epsilon(x)$ "resembles" Dirac's original δ–function closer and closer.



Now let us connect φ_ϵ to Schwartz's definition: for any $f \in C_0^\infty$ we have

$$\int \varphi_\epsilon(x)f(x)dx = \int_{-\epsilon}^{\epsilon} \varphi_\epsilon(x)f(x)dx \approx f(0) \int \varphi_\epsilon(x)dx = f(0),$$

since $\text{supp}\varphi_\epsilon \subset [-\epsilon, \epsilon]$. In fact, we have for C_0^∞-functions, see section 11, that $\varphi_\epsilon * f \to f$ pointwise, as $\epsilon \to 0$, i.e.

$$\int_{-\infty}^{\infty} \varphi_\epsilon(x-y)f(y)dy \to f(x).$$

This shows that for any x, the functional $\delta_{x,\epsilon}(f) := \int \varphi_\epsilon(x-y)f(y)dy$ converges to the functional $\delta_x(f) := f(x)$ in the sense that $\delta_{x,\epsilon}(f) \to \delta_x(f)$, as $\epsilon \downarrow 0$, $\forall f \in C_0^\infty$. This is the Schwartzian viewpoint. On the other hand, formally "$\varphi_\epsilon(x) \to \delta(x)$", or "$\varphi_\epsilon(x-y) \to \delta(x-y)$". The last two relations suggest that we can write, formally again:

$$\delta(f) = \int_{-\infty}^{\infty} \delta(x)f(x)dx,$$

where the left hand side is Schwartz's functional, and the integral in the right hand side is thought of as a convenient heuristic expression. Similarly, we can define the δ–function at a point x_0:

$$\delta_{x_0}(f) = \int_{-\infty}^{\infty} \delta(x - x_0) f(x) dx = f(x_0).$$

In the n–dimensional case, we define for $x = (x_1, \dots, x_n)$:

$$\delta(x) = \prod_1^n \delta(x_j),$$

Using the definition of the δ–function, we obtain $\mathcal{F} : \delta(x - x_0) \to (2\pi)^{-n/2} e^{-ik\cdot x_0}$, hence $\mathcal{F}^{-1} : (2\pi)^{-n/2} e^{-ik\cdot x_0} \to \delta(x - x_0)$, and, by taking the complex conjugate (remember that $\mathcal{F}(\overline{f}) = \overline{\mathcal{F}^*(f)} = \overline{\mathcal{F}^{-1}(f)}$), we arrive at

$$\mathcal{F} : \ (2\pi)^{-n/2} e^{-ik_0\cdot x} \mapsto \delta(k - k_0). \tag{14.1}$$

Exercise. Assuming (14.1) is true, prove formally that $(\hat{f})\check{} = f = (\check{f})\hat{}$, and $(fg)\hat{} = \hat{f} * \hat{g}$.

15. Distributions

In this section, we briefly discuss the theory fo distributions (also called *generalized functions*), generalizing the theory of Dirac's δ–function.

Distributions are defined as continuous linear functionals over certain spaces of "nice" functions. A functional F is a map of some space of functions into $\mathbb{R}$, e.g.

1. $F(\varphi) = \int_0^1 F\varphi dx$ for some fixed $F \in L^1([0,1])$, and where $\varphi \in C([0,1])$,

2. $F(\varphi) = \varphi'(x_0)$ for any $\varphi \in C^1([0,1])$,

3. $F(\varphi) = \frac{1}{2}\int_\Omega |\nabla\varphi|^2$, for a fixed bounded domain $\Omega \subset \mathbb{R}^n$, and for any $\varphi \in C(\Omega)$,

4. $F(\varphi) = \int_\Omega |\varphi|^p$, for a fixed domain $\Omega \subset \mathbb{R}^n$, and for any $\varphi \in L^p(\Omega)$.

A functional F is called *linear* iff $F(a\varphi + b\psi) = aF(\varphi) + bF(\psi)$, for any $a, b \in \mathbb{R}$ (or $\mathbb{C}$ if we are dealing with complex spaces), and any φ, ψ from the space on which F is defined. The functionals in examples 1. and 2. are linear, and the functionals in examples 3. and 4. are not. A functional F is called *continuous* iff $F(\varphi_n) \to F(\varphi)$ whenever $\varphi_n \to \varphi$. Of course, we assume that the space on which F is defined has a notion of convergence. Remember that in a normed space, $\varphi_n \to \varphi$ iff $||\varphi_n - \varphi|| \to 0$. We give now two examples of very important spaces which do have a notion of convergence, but which cannot be equipped with a norm:

a) $C_0^\infty(\Omega)$, the space of C^∞ (i.e. smooth) functions with bounded support. We say $f_n \to f$ iff $||f_n - f||_{K,r} \to 0$ for any compact set $K \subset \Omega$ and any nonnegative integer r, where

$$||f||_{K,r} := \sup_{|\alpha| \le r} \sup_{x \in K} |\partial^\alpha f(x)|.$$

If Ω is bounded, then it suffices to take $K = \overline{\Omega}$.

b) *The Schwartz space* $\mathcal{S}(\mathbb{R}^n)$, defined as

$$\mathcal{S}(\mathbb{R}^n) = \{f \in C^\infty(\mathbb{R}^n) : \sup_x (1 + |x|)^N |\partial^\alpha f(x)| < \infty \text{ for any } N \text{ and } \alpha\}.$$

We say $f_n \to f$ iff $||f_n - f||_{r,m} \to 0$ for all r, m, and where

$$||f||_{r,m} := \sup_{|\alpha| \le r} \sup_{x \in \mathbb{R}^n} (1 + |x|)^m |\partial^\alpha f(x)|.$$

Continuous linear functionals on $\mathcal{S}(\mathbb{R}^n)$ are called *tempered distributions* and the space of tempered distributions is denoted by $\mathcal{S}'(\mathbb{R}^n)$.

Continuous linear functionals over $C_0^\infty(\Omega)$ are simply called distributions, their space is denoted by $\mathcal{D}'(\Omega)$. Clearly, $\mathcal{S}'(\Omega) \subset \mathcal{D}'(\Omega)$.

Distributions are modeled on example 1. above, and we write symbolically

$$F(\varphi) = \int F(x)\varphi(x)dx. \tag{15.1}$$

We say that a family F_ϵ of distributions converges to a distribtuion F, as $\epsilon \to 0$, iff $F_\epsilon(\varphi) \to F(\varphi)$, $\forall \varphi$.

Exercise. Show that $\delta_x \in \mathcal{S}'(\mathbb{R})$, and that $\delta_{x,\epsilon} \to \delta_x$ as $\epsilon \downarrow 0$, $\forall x$. For distributions, we can define many of the notions we have for usual functions if we let us be guided by (15.1). We call supp$F$ (the *support of* F) the smallest closed set $K \subset \mathbb{R}^n$ s.t. $F(\varphi) = 0$ for all φ whose support is disjoint from K. We also define *partial derivatives of distributions* as follows:

$$\partial_{x_j} F(\varphi) := -F(\partial_{x_j}\varphi).$$

Note that if $F(\varphi) = \int F\varphi$, where $F \in C^1$, then $\partial_{x_j}F(\varphi) = \int \partial_{x_j}F\varphi$. Similarly, we define for any multi–index α

$$\partial^\alpha f(\varphi) := (-1)^{|\alpha|}F(\partial^\alpha \varphi).$$

Example. On $\mathbb{R}$, δ' is defined by $\delta'(\varphi) = -\delta(\varphi') = -\varphi'(0)$. We see that any distribution is infinitely many times differentiable!

We define the *convolution of a distribution* F with $\varphi \in C_0^\infty$ by:

$$F * \varphi := F(\varphi^x),$$

where $\varphi^x(y) = \varphi(x-y)$. We also define the *Fourier transform of a distribution* by:

$$(\mathcal{F}F)(\varphi) := F(\mathcal{F}\varphi).$$

This definition needs some justification. One can show that if $\varphi \in \mathcal{S}$, then $\hat{\varphi} \in \mathcal{S}$, so that $F(\hat{\varphi})$ is well defined for a tempered distribution F.

If F is a distribution with compact support, then $\hat{F}$ can be defined as the function $\hat{F}(k) = F(e_k)$, where $e_k(x) = (2\pi)^{-n/2}e^{-ik\cdot x}$.

Exercises. **1)** Show that $\delta(x) = \frac{d}{dx}\chi_{x\geq 0}$, where $\chi_{x\geq 0}$ is the *Heaviside function*: $\chi_{x\geq 0} = 1$ if $x \geq 0$, and $\chi_{x\geq 0} = 0$ for $x < 0$. **2)** Show that the distribution $\delta(x-ct)$ solves the equation $c^{-2}\partial_t^2 f = \partial_x^2 f$. **3)** Find $\hat{\delta}(k)$ and $(\partial_x^m \delta)\hat{}\,(k)$. **4)** Let f be a continuous function on $\mathbb{R}^n \backslash \{0\}$, and define the distribution $PV(f)$ by

$$PV(f)(\varphi) = \lim_{\epsilon \downarrow 0} \int_{|x| \geq \epsilon} f(x)\varphi(x).$$

This distribution is called the *principal value* of the integral of f. Show that $PV(f)$ is a distribution, and find $\hat{F}$ for $F = PV\left((\pi x)^{-1}\right)$ in dimension $n = 1$. **5)** Let $n \geq 3$ and $F_\epsilon(x) := [(2-n)\sigma_{n-1}]^{-1}(|x|^2+\epsilon^2)^{-(n-2)/2}$, where σ_n is the volume of the n–sphere $S^n = \{x \in \mathbb{R}^{n+1} : |x| = 1\}$. Show that $\Delta F_\epsilon(x) \to \delta(x)$ as $\epsilon \to 0$. Show that $u = F_0 * f$ satisfies the differential equation $\Delta u = f$. **6)** Let $n \geq 3$ and let $F_\epsilon(x)$ be as in 5). Show that $\Delta F_\epsilon(x) \to \delta(x)$, as $\epsilon \to 0$. Here again, σ_n is the volume of the n–sphere $S^n = \{x \in \mathbb{R}^{n+1} : |x| = 1\}$.

16. Sobolev spaces

In many respects, the L^p–spaces are easier to work with than the C^p–spaces. One reason is that the L^p–spaces are defined in terms of integrals which are easy to estimate. For instance, we know that the Fourier transform of an L^p–function with $1 \leq p \leq 2$ is an L^q–function, with $q^{-1} = 1 - p^{-1}$. On the other hand, we cannot say much about the Fourier transform of a continuous or bounded continuous function on $\mathbb{R}^n$.

Now we want to introduce an additional structure on L^p–spaces which measures smoothness, similarly to the smoothness properties of functions in C^k. We do so only for $p = 2$, i.e. for the space $L^2(\mathbb{R}^n)$. This is the simplest space among the L^p–spaces. It has an inner product:

$$\langle f, g \rangle := \int \overline{f} g$$

and therefore it is a *Hilbert space* (i.e. an inner product space which is complete with respect to the norm $||f|| := \sqrt{\langle f, f \rangle}$ induced by the inner product). Another advantage of the L^2–space is that the Fourier transform leaves it invariant (i.e. $f \in L^2 \Rightarrow \hat{f} \in L^2$).

We now define for s integer, $s \geq 0$, the new spaces

$$H_s(\mathbb{R}^n) = \{f \in L^2(\mathbb{R}^n) : \partial^\alpha f \in L^2(\mathbb{R}^n)\ \forall\alpha \text{ s.t. } |\alpha| \leq s\}. \qquad (16.1)$$

This definition is very similar to the definition of the $C^s(\mathbb{R}^n)$–spaces: in fact, by replacing $L^2(\mathbb{R}^n)$ in (16.1) by $C(\mathbb{R}^n)$, one obtains the definition of $C^s(\mathbb{R}^n)$. But there is one crucial difference: in the $C^s(\mathbb{R}^n)$–case, the functions f are assumed to be s times continuously differentiable, but in the H_s–case, they are not. Namely, the derivatives $\partial^\alpha f$ in the above definition are understood *in the distributional sense*:

$$\partial^\alpha f(\varphi) = (-1)^{|\alpha|} f(\partial^\alpha \varphi),$$

where $f(\varphi) = \int f\varphi$, and $\varphi \in \mathcal{S}(\mathbb{R}^n)$.

There is another way of defining the spaces $H_s(\mathbb{R}^n)$:

$$H_s(\mathbb{R}^n) = \{f \in L^2(\mathbb{R}^n) : <\!k\!>^s \hat{f}(k) \in L^2(\mathbb{R}^n)\}, \qquad (16.2)$$

where $<\!k\!>= (1 + |k|^2)^{1/2}$. Definition (16.2) has the advantage that it makes sense for an arbitrary $s \in \mathbb{R}$. Besides, it does not require extra explanations. Of course we have to show that definitions (16.1) and (16.2) are equivalent for positive integers s.

Let first f belong to the r.h.s. of (16.2). Then, since $|(\partial^\alpha f)\hat{}(k)| = |k^\alpha \hat{f}(k)| \leq <\!k\!>^{|\alpha|} |\hat{f}(k)|$, we have that $(\partial^\alpha f)\hat{} \in L^2$, and therefore by the Plancherel theorem, $\partial^\alpha f \in L^2$, as long as $|\alpha| \leq s$.

Now let f belong to the r.h.s. of (16.1). So we have in particular that $f \in L^2$, $\partial^s_{x_j} f \in L^2\ \forall j$, and therefore by the Plancherel theorem $\hat{f} \in L^2$, $k_j^s \hat{f} \in L^2$, $\forall j$. The latter implies that $(1 + \sum_1^n |k_j|^s)\hat{f} \in L^2$. Since for some $0 < C_1 < C_2 < \infty$:

$$C_1 <\!k\!>^s \leq 1 + \sum_1^n |k_j|^s \leq C_2 <\!k\!>^s, \qquad (16.3)$$

we then get $<\!k\!>^s \hat{f} \in L^2$, i.e. f belongs to the r.h.s. of (16.2). ■

Exercise. Show formula (16.3).

The space $H_s(\mathbb{R}^n)$ (defined in (16.2)) is called *Sobolev space of order* s. It is a Hilbert space with the inner product and norm

$$\langle f, g\rangle_s := \int \bar{\hat{f}} \hat{g} <\!k\!>^{2s} \quad \text{and} \quad ||f||_{(s)} := \left(\int |\hat{f}|^2 <\!k\!>^{2s}\right)^{1/2}.$$

The spaces H_s, $s \in \mathbb{R}$, have the following properties:

(i) $H_0 = L^2$,

(ii) $H_s \subset H_t$ if $s > t$,

(iii) ∂^α is a bounded map from H_s into $H_{s-|\alpha|}$.

Property (i) is just Plancherel's theorem. Property (ii) follows from $<\!k\!>^s \geq <\!k\!>^t$ if $s \geq t$. To prove (iii), notice that (iii) $\Leftrightarrow ||\partial^\alpha f||_{(s-|\alpha|)} \leq C||f||_{(s)}$ and that the latter estimate is true because $||\partial^\alpha f||^2_{(s-|\alpha|)} = \int |(\partial^\alpha f)\hat{}|^2 <\!k\!>^{2(s-|\alpha|)} = \int |k^\alpha \hat{f}|^2 <\!k\!>^{2(s-|\alpha|)}$, and the last integral is bounded from above by $\int |\hat{f}|^2 <\!k\!>^{2s} = ||\hat{f} <\!k\!>^s ||_2^2 = ||\hat{f}||^2_{(s)}$. ■

The next theorem connects Sobolev spaces and C^s–spaces.

The Sobolev embedding theorem. *If $s > r+n/2$, then $H_s(\mathbb{R}^n) \subset C^r_d(\mathbb{R}^n)$, where $C^r_d(\mathbb{R}^n)$ denotes the elements in $C^r(\mathbb{R}^n)$ that decay at infinity (i.e. functions $f \in C^r(\mathbb{R}^m)$ s.t. $\lim_{|x|\to\infty} f(x) = 0$). Moreover, this inclusion is continuous in the sense that there is a constant C (depending only on s, k, n) such that $||\partial^\alpha f||_\infty \leq C||f||_{(s)}$, uniformly in $f \in H_s(\mathbb{R}^n)$ and every α, $|\alpha| \leq k$.*

Proof. Suppose that $s > n/2 + r$ and $|\alpha| = r$. Then using that $\langle k\rangle^r = \langle k\rangle^s \langle k\rangle^{-s+k}$,

$$||(\partial^\alpha f)\hat{}||_1 \leq || <\!k\!>^r \hat{f}||_1 \leq || <\!k\!>^s \hat{f}||_2 \; ||<\!k\!>^{-s+r}||_2 \leq C||f||_{(s)}.$$

In the last step, we used that $s > n/2+r$, and therefore $C = \|\langle k\rangle^{-s+r}\| < \infty$. We have thus shown that $(\partial^\alpha f)\hat{} \in L^1$. But this implies $\partial^\alpha f \in C_d$ since $\check{L}_1 \subset C_d$ by the Riemann-Lebesgue lemma, see Section 12. ■

Exercise. Define the operator $\Lambda_s : f \mapsto (<\!k\!>^s \hat{f})\check{}$, i.e. $(\Lambda_s f)\hat{} =< k>^s \hat{f}$. Show that $\Lambda_s : H_t \to H_{t-s}$ is an *isometry* (i.e. $||\Lambda_t f||_{(s-t)} = ||f||_{(s)}$).

17. Linear operators

Linear operators or simply *operators* are linear maps from one vector space X into another vector space Y. We denote linear operators usually by capital roman letters, $A, B, \ldots$. For instance

$$A : X \to Y \quad \text{or} \quad A : u \mapsto Au.$$

To define an operator A, we have to give two things: the domain of definition, $\mathcal{D}(A)$ (a subset of X), and a rule that prescribes to each element of $u \in \mathcal{D}(A)$ an element of Y (the image of u). We require this rule to be linear, i.e. $\forall u, v \in \mathcal{D}(A)$, and $\alpha, \beta \in \mathbb{C}$:

$$A(\alpha u + \beta v) = \alpha Au + \beta Av. \tag{17.1}$$

To fix ideas here and in what follows, we consider vector spaces over the complex numbers $\mathbb{C}$, i.e. complex vector spaces. All the material of this section, except for spectral theory, remains unchanged if we substitute $\mathbb{R}$ for $\mathbb{C}$.

The linearity property (17.1) implies that the domain of A can always be taken to be a vector subspace of X. Indeed, if we take $u, v \in \mathcal{D}(A)$, then Au and Av are well defined, and we can add $\alpha u + \beta v$ (for any $\alpha, \beta \in \mathbb{C}$) to the domain $\mathcal{D}(A)$ by defining $A(\alpha u + \beta v) := \alpha Au + \beta Av$. We therefore will always assume that $\mathcal{D}(A)$ is a vector space. This also implies that the range (or image) of A,

$$\operatorname{Ran}(A) := \{Au : u \in \mathcal{D}(A)\} \equiv A\mathcal{D}(A),$$

is a vector space as well. We may assume that $\mathcal{D}(A)$ is *dense* in X, i.e. for any $u \in X$, there is a sequence $\{u_n\} \subset \mathcal{D}(A)$ s.t. $u_n \to u$ as $n \to \infty$. Indeed, if $\mathcal{D}(A)$ is not dense to begin with, we consider instead of the space X simply the space $X' := \overline{\mathcal{D}(A)}$, the closure of $\mathcal{D}(A)$.

Examples. **1)** The identity operator $\mathbb{1} : L^p \to L^p$ has domain $\mathcal{D}(\mathbb{1}) = L^p$.

2) The multiplication operator $M_f : L^p \to L^p$, $u \mapsto fu$ for a fixed $f \in L^\infty$ has domain $\mathcal{D}(M_f) = L^p$.

3) The differentiation operator $\frac{\partial}{\partial x_j} : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ has domain $\mathcal{D}(\frac{\partial}{\partial x_j}) = \{u \in L^2 : \frac{\partial}{\partial x_j} u \in L^2\}$.

4) The Laplacian $\Delta := \sum_1^n \frac{\partial^2}{\partial x_j^2} : L^2(\Omega) \to L^2(\Omega)$ has the domain $\mathcal{D}(\Delta) = H_2(\Omega)$.

5) The Fourier transform $\mathcal{F} : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ has the domain

$L^2(\mathbb{R}^n)$.

6) Integral operators are operators of the form

$$(Ku)(x) = \int K(x,y)u(y)dy,$$

for some function $K(x,y)$ (called the *kernel* or *integral kernel*). The domain and range of the integral operator K depend on the properties of the kernel $K(x,y)$.

In all the previous examples, the operators can be represented as integral operators, but with distributional kernels, e.g. $K(x,y) = f(x)\delta(x-y)$ for M_f, and $K(x,y) = -\delta'(x_j - y_j)\prod_{i\neq j}\delta(x_i - y_i)$ for $\frac{\partial}{\partial x_j}$.

We say the operator $A : X \to Y$ is bounded iff there is a constant C (independent of u) such that

$$||Au|| \leq C||u||, \tag{17.2}$$

for all $u \in \mathcal{D}(A)$. The smallest constant C satisfying (17.2) is called the *norm* of A, and it is denoted by $||A||$. We have

$$||A|| = \sup_{u\neq 0}\frac{||Au||}{||u||} = \sup_{u:\,||u||=1}||Au||, \tag{17.3}$$

and so $||Au|| \leq ||A||\,||u||$. If $A : X \to Y$ is bounded and defined on a dense subset of X and Y is complete, then one can extend A by continuity to the whole space X.

In the examples above, we see that the multiplication operator M_f is bounded with $||M_f|| = ||f||_\infty$.

Exercise. Show that the differentiation operator in example 2) is not bounded by finding a sequence f_n of functions from $\mathcal{D}(\frac{\partial}{\partial x_j})$ such that $||f_n|| \leq 1$, $\forall n$, and $||\frac{\partial}{\partial x_j}f_n||_2 \to \infty$, as $n \to \infty$.

The identity operator in example 3) is clearly bounded, and $||\mathbb{1}|| = 1$. The integral operator K in 4) with a kernel satisfying $K(x,y) \in L^2(\mathbb{R}^n \times \mathbb{R}^n)$ is bounded as an operator from $L^2(\mathbb{R}^n)$ to $L^2(\mathbb{R}^n)$.

We say A is *invertible* iff A has a bounded inverse, i.e. iff there is a bounded operator $A^{-1} : Y \to X$ such that $A^{-1}A = \mathbb{1}_X$ and $AA^{-1} = \mathbb{1}_Y$, where $\mathbb{1}_X$ and $\mathbb{1}_Y$ are the identity operators in X and Y respectively.

Exercises. Show that: **1)** A is invertible iff for every $f \in Y$, the equation $Au = f$ has a unique solution $u(= A^{-1}f) \in X$, i.e. iff A is one–to–one ($Au = 0 \Rightarrow u = 0$) and onto ($\mathrm{Ran}A = Y$). **2)** if A is just one-to-one (i.e. not necessarily onto), then A is invertible as an operator from X to $\mathrm{Ran}\,A \subset Y$, i.e. the equation $Au = f$ has a unique solution $u = A^{-1}f$ for any $f \in \mathrm{Ran}\,A$.

Example: invertibility of the Laplacian on a bounded domain in $\mathbb{R}^n$. Let $\Omega \subset \mathbb{R}^n$ be a bounded domain. Consider first $\Delta : H_k(\Omega) \to H_{k-2}(\Omega)$. Recall that Δ is a bounded operator between these spaces. However, on $H_k(\Omega)$, Δ has an eigenvalue 0 with a constant eigenfunction u_0. Hence Δ is not invertible on $H_k(\Omega)$.

If we restrict ourselves to a smaller space than $H_k(\Omega)$, in particular a space that does not contain constant functions, then Δ has a chance to be invertible on that smaller space. We therefore introduce

$$H_k^{(0)}(\Omega) := \{u \in H_k(\Omega) : u|_{\partial\Omega} = 0\}. \tag{17.4}$$

Notice that if $k > n/2$, then by Sobolev's embedding theorem, functions in $H^{(0)}(\Omega)$ are continuous, so the condition $u|_{\partial\Omega} = 0$ makes sense. Otherwise, we define $H_k^{(0)}(\Omega)$ by completing the space $C_0^{(k)}(\Omega)$ of k-times differentiable functions on Ω, vanishing on $\partial\Omega$ in the norm $\|\cdot\|_{(k)} := \|\cdot\|_{H_k(\Omega)}$, i.e. adding to $C_0^{(k)}(\Omega)$ "limits" of all Cauchy sequences. In order not to worry about that, we assume here $k > n/2$.

Showing that $\Delta : H_k^{(0)}(\Omega) \to H_{k-2}(\Omega)$ is invertible is equivalent to showing that the Dirichlet problem

$$\begin{aligned} \Delta u = f \quad & \text{in } \Omega, \\ u = 0 \quad & \text{on } \Omega \end{aligned} \tag{17.5}$$

has a unique solution in $H_k^{(0)}(\Omega)$, for any $f \in H_{k-2}(\Omega)$.

If $\Omega = B_R(0)$ is a ball of radius R, then for any $f \in H_{k-2}(\Omega)$, there is a solution $u \in H_k(\Omega)$ given by the Poisson integral ($n \geq 3$):

$$u(x) = \int G(x,y) f(y)\, d^n y,$$

where the Green's function G is given by

$$G(x,y) = \frac{1}{(2-n)\sigma_{n-1}} \left(|x-y|^{2-n} - \left(\frac{R}{|y|}\right)^{n-2} \left| x - \frac{Ry}{|y|^2} \right|^{2-n} \right)$$

and where σ_{n-1} is the volume of the $(n-1)$–sphere.

Exercise. Verify the statement above.
For a general bounded domain $\Omega \subset \mathbb{R}^n$, we prove the existence of solutions to problem (17.5) later, using variational methods.

Here we show only uniqueness of solutions of (17.5) for any bounded Ω, i.e. that the operator $\Delta : H_k^{(0)}(\Omega) \to H_{k-2}(\Omega)$ is one-to-one. According to the last exercise above, this implies that the boundary value problem (17.5) has a unique solution for any $f \in \Delta(H_k^{(0)}(\Omega))$. The result of uniqueness is based on the

Poincaré inequality. *Let Ω have a diameter $d < \infty$ in some direction (i.e. it is possible to place Ω between two parallel hyperplanes at a distance d from each other). Then for any $u \in H_1^{(0)}(\Omega)$, we have*

$$\int_\Omega |u|^2 \leq (2d)^2 \int_\Omega |\nabla u|^2. \tag{17.6}$$

Proof. We can assume that this hyperplanes are $\{x_1 = 0\}$ and $\{x_1 = d\}$. Assume u is real and estimate

$$||u||_2^2 = \int_\Omega 1 \cdot |u|^2 = -\int_\Omega x_1 \frac{\partial}{\partial x_1}|u|^2 = -2\mathrm{Re}\int_\Omega x_1 \overline{u} \frac{\partial u}{\partial x_1} \leq 2d \int_\Omega |u| \left| \frac{\partial u}{\partial x_1} \right|.$$

Applying now the Schwarz inequality to the integral on the r.h.s., we obtain

$$||u||_2^2 \leq 2d||u||_2 \left|\left| \frac{\partial u}{\partial x_1} \right|\right|_2 \leq 2d||u||_2\,||\nabla u||_2,$$

where $||\nabla u||_2^2 = \int_\Omega |\nabla u|^2 = \int_\Omega \sum_1^n |\frac{\partial u}{\partial x_n}|^2$. The latter inequality implies $||u||_2 \leq 2d||\nabla u||_2$. ■

If $u \in H_2^{(0)}(\Omega)$, then

$$\int_\Omega |\nabla u|^2 = \int_\Omega \overline{u}\,(-\Delta u) = \int_\Omega \left|\sqrt{-\Delta}\,u\right|^2,$$

where we have used in the last step that $-\Delta$ is positive and self-adjoint and defined $\sqrt{-\Delta}$ by $-\Delta = \sqrt{-\Delta}\,\sqrt{-\Delta}$ (or by Fourier transform). We can now rewrite the Poincaré inequality as

$$||u||_2 \leq 2d||\sqrt{-\Delta}\,u||_2,$$

which implies for $u \in H_2^{(0)}(\Omega)$:

$$||\Delta u||_2 = ||\sqrt{-\Delta}\,\sqrt{-\Delta}\,u||_2 \geq (2d)^{-2}||u||_2.$$

Exercise. Show that for all $u \in H_k^{(0)}(\Omega)$, Ω bounded: $||\Delta u||_2 \geq (2d)^{-2}||u||_2$ implies

$$||\Delta u||_{(k)} \geq C||u||_{(k)}, \tag{17.7}$$

for some $C > 0$, independent of u, and any k.

Formula (17.7) shows that $\Delta u = 0 \Rightarrow u = 0$ for $u \in H_k^{(0)}(\Omega)$. Therefore we can define the inverse Δ^{-1} on the range $\Delta(H_k^{(0)}(\Omega))$: if $v \in \Delta(H_k^{(0)}(\Omega))$, then $v = \Delta u$ for some unique $u \in H_k^{(0)}(\Omega)$ and we define $\Delta^{-1}v = u$. Now from (17.7):

$$||v||_{(k)} \geq C||\Delta^{-1}v||_{(k)}.$$

Hence $\Delta^{-1} : \Delta(H_k^{(0)}(\Omega)) \to H_k^{(0)}(\Omega)$ is bounded. Remark that we have not shown that $\Delta^{-1} : H_{k-2}(\Omega) \to H_k^{(0)}(\Omega)$ is bounded (or equivalently that $\Delta(H_k^{(0)}(\Omega)) = H_{k-2}(\Omega)$, or equivalently that the Dirichlet problem (17.5) has a unique solution for any $f \in H_{k-2}(\Omega)$). We will show this in a later section, using variational methods. This finishes our example.

The *spectrum* of A acting on a space X (i.e. $Y = X$), $\sigma(A)$, is the set in $\mathbb{C}$ defined by

$$\sigma(A) := \{z \in \mathbb{C} : A - z\mathbb{1} \quad \text{is not invertible}\}.$$

For notational convenience, the operator "multiplication by $z \in \mathbb{C}$" will be simply written as z instead of $z\mathbb{1}$. The spectrum of an operator is always a closed set in $\mathbb{C}$. Clearly, eigenvalues of A belong to $\sigma(A)$ (in fact, if λ is an eigenvalue, then $Au_\lambda = \lambda u_\lambda$ for some nonzero $u_\lambda \in X$, so $(A-\lambda)u_\lambda = 0$, and $A - \lambda$ is not invertible). In general, the spectrum can also contain continuous pieces and it can take very peculiar forms.

The complement of the spectrum is called the *resolvent set* $\rho(A)$:

$$\rho(A) := \mathbb{C}\backslash\sigma(A).$$

For $z \in \rho(A)$, $A - z$ has a bounded inverse $(A-z)^{-1}$, called the *resolvent.* The resolvent is analytic in $z \in \rho(A)$.

Exercise. The spectrum of the multiplication operator introduced in example 1) above is $\sigma(M_f) = \overline{\text{Ran} f}$, the differentiation operator 2) has spectrum $\sigma(\frac{\partial}{\partial x_j}) = \mathbb{R}$, and the identity operator 3) has purely discrete spectrum $\sigma(\mathbb{1}) = \{1\}$.

The study of the spectra of operators is called *spectral analysis.* It is considerably simplified if X is a Hilbert space. From now on in the remainder of this chapter we will consider only operators $A : X \to X$, where X is a Hilbert space.

With an operator A we can associate its *adjoint* A^* defined (roughly) by the relation $\langle A^*u, v\rangle = \langle u, Av\rangle$, for all $v \in \mathcal{D}(A)$, and for all u's for which this relation makes sense (those u's form the domain of the operator A^*, $\mathcal{D}(A^*)$).

Exercises. Show that **1)** $\|w\| = \sup_{\|v\|=1} |\langle w, v\rangle|$, and therefore $\|A\| = \sup_{\|u\|,\|v\|=1} |\langle Au, v\rangle|$; **2)** if A is bounded, then so is $\|A^*\|$, and $\|A\| = \|A^*\|$ (Hint: use part 1)).

One can show that $((A-z)^{-1})^* = (A^*-\overline{z})^{-1}$, and therefore $\sigma(A^*) = \overline{\sigma(A)}$.

Exercise. Prove the latter statement for A bounded.

An important class of operators on a Hilbert space is the class of *self–adjoint* operators. By definition, an operator A is called self–adjoint iff $A^* = A$. In particular, every self–adjoint operator is *symmetric*, i.e. $\langle Au, v\rangle = \langle u, Av\rangle$, for all $u, v \in \mathcal{D}(A) = \mathcal{D}(A^*)$. Notice that the converse is not true. However, every symmetric *bounded* operator is self–adjoint.

From the property $\sigma(A^*) = \overline{\sigma(A)}$, we see immediately that if A is self–adjoint, then $\sigma(A) \subset \mathbb{R}$.

If we consider the examples of the operators 1)–4) above, we have the following: M_f is symmetric iff f is a real function; $(\frac{\partial}{\partial x_j})^* = -\frac{\partial}{\partial x_j}$, so the differentiation operator is not symmetric, but $-i\frac{\partial}{\partial x_j}$ is symmetric; the identity operator is obviously symmetric; the integral operator is symmetric if $K(x, y) = \overline{K(y, x)}$.

We will use the following basic facts from the geometry of Hilbert spaces. Let V and W be subspaces of the Hilbert space X. We say *V and W are orthogonal to each other* (written $V \perp W$) iff

$$\langle v, w\rangle = 0, \ \forall v \in V, \forall w \in W.$$

If V and W are orthogonal, we define the *orthogonal sum* (or *direct sum*) by

$$V \oplus W := \{v + w : v \in V \text{ and } w \in W\}.$$

If $V \subset X$ is a subspace, then its *orthogonal complement in X* is defined by

$$V^{\perp} := \{x \in X : \langle v, x\rangle = 0, \ \forall v \in V\}.$$

Exercise. Show that $V^{\perp}$ is a closed subspace of X, even if V is not closed. Recall that a subspace $V \subset X$ is called closed iff the limit of any convergent sequence in V lies in V. More precisely, $V \subset X$ is closed iff $\{v_n\} \subset V$ and $v_n \to v \in X$ implies $v \in V$.

One of the key properties of Hilbert spaces is given in the following

Theorem. *If $V \subset X$ is a closed subspace, then $X = V \oplus V^{\perp}$.*

Proof. We give a complete proof in the case when $\dim V = n < \infty$. Pick an orthonormal basis $\{e_1, \dots, e_n\}$ for V. For an arbitrary $f \in X$, define $f_1 := \sum_{i=1}^n \langle f, e_i\rangle e_i$, and define $f_2 := f - f_1$. We clearly have $f_1 \in V$, moreover, $f_2 \in V^{\perp}$: Indeed, for any $v \in V$, we have $v = \sum_{j=1}^n \langle v, e_j\rangle e_j$, hence

$$\begin{aligned}
\langle v, f_2\rangle &= \langle v, f\rangle - \langle v, f_1\rangle \\
&= \sum_{j=1}^n \langle v, e_j\rangle\langle f, e_j\rangle - \sum_{i=1}^n \langle f, e_i\rangle\langle v, e_i\rangle \\
&= 0.
\end{aligned}$$

We have thus decomposed an arbitrary $f \in X$ as $f = f_1 + f_2$, with $f_1 \in V$ and $f_2 \in V^\perp$. We finish the proof by showing that this decomposition is unique. To do so, assume that there are $f_1' \in V$ and $f_2' \in V^\perp$ such that $f = f_1' + f_2'$. We show $f_1' = f_2$, $f_2' = f_2$. Indeed, $0 = f - f = f_1 - f_1' + f_2 - f_2'$, so $f_1 - f_1' = -(f_2 - f_2')$. But on the other hand, $V \ni f_1 - f_1' \perp f_2 - f_2' \in V^\perp$, thus $f_1 - f_1' = 0 = f_2 - f_2'$. The proof is complete if $\dim V < \infty$. In the general case ($\dim V = \infty$), the same proof is valid, but now define f_1 as being the element in V that minimizes the distance to f (it has to be shown that such an element exists and is unique). ■

Closed subspaces of a Hilbert space can be identified with (i.e. are in one–to–one correspondence to) *projection operators*. A bounded operator P on X is called a projection operator (or simply a projection) iff it satisfies

$$P^2 = P.$$

This relation implies $||P|| \leq ||P||^2$, i.e. $||P|| \geq 1$. We have

$$v \in \text{Ran}P \Rightarrow Pv = v, \quad \text{and} \quad v \in (\text{Ran}P)^\perp \Rightarrow P^*v = 0. \qquad (17.8)$$

Indeed, if $v \in \text{Ran}P$, then there is a $u \in X$ s.t. $v = Pu$, so $Pv = P^2u = Pu = v$; the second statement is left as an

Exercise. Prove that $P^*v = 0$ if $v \perp \text{Ran}P$.

The above mentioned correspondence between projection operators and closed subspaces is given by the following fact. Let $V = \text{Ran}P$. Then V is a closed subspace of X. To show that V is closed, let $\{v_n\} \subset V$, and $v_n \to v \in X$, and show that $v \in V$. Since P is a projection, we have $v_n = Pv_n$, so $||v - Pv|| = ||v - v_n - P(v - v_n)|| \leq ||v - v_n|| + ||P||\,||v - v_n|| \to 0$, as $n \to \infty$. Therefore $v = Pv$, so $v \in V$, and V is closed.

A projection P is called an *orthogonal projection* iff it is selfadjoint, i.e. iff $P = P^*$. Let P be an orthogonal projection, then

$$v \perp \text{Ran}P \Rightarrow Pv = 0. \qquad (17.9)$$

Exercise. Let P be an orthogonal projection. Using (17.8) show that $||P|| \leq 1$, and therefore $||P|| = 1$.

Conversely, given a closed subspace V, define a projection operator P by

$$Pu = v, \quad \text{where} \quad u = v + v^{\perp} \in V \oplus V^{\perp}. \tag{17.10}$$

Exercise. Show that P defined in (17.10) is a projection with $\mathrm{Ran} P = V$.

Finally, we observe that for any operator A on a Hilbert space X, we can write

$$X = \mathrm{Null} A \oplus \overline{\mathrm{Ran} A^*} \tag{17.11}$$

Here, the null space is defined as $\mathrm{Null}\, A = \{u \in X : Au = 0\}$.

Exercise. Show that for a bounded operator A, $\mathrm{Null} A$ is a closed set, and show (17.11).

The space of bounded linear operators $\mathcal{L}(X, Y)$. We assume that X and Y are normed vector spaces over $\mathbb{C}$, and consider the set of all bounded linear operators from X into Y, i.e. each such operator is defined on the entire space X, and its range lies in Y. This set of operators is denoted by $\mathcal{L}(X, Y)$.

For $A, B \in \mathcal{L}(X, Y)$, we define a new operator, called $A + B$, by setting $(A + B)u := Au + Bu$, for all $u \in X$. Also, for $\lambda \in \mathbb{C}$ and $A \in \mathcal{L}(X, Y)$, we define a new operator λA as $(\lambda A)u = \lambda Au$, for all $u \in X$. If in addition to these two operations on operators, we equip the set $\mathcal{L}(X, Y)$ with the norm introduced in (17.3), then $\mathcal{L}(X, Y)$ is a normed vector space.

Exercise. Show that $\mathcal{L}(X, Y)$ is a vector space.
An important question is: when is $\mathcal{L}(X, Y)$ a Banach space? The answer is given in the following theorem, which is not difficult to prove (see e.g. [F], Proposition 5.3):

Theorem. *If Y is a Banach space, then $\mathcal{L}(X, Y)$ is a Banach space.*

The dual space. In the special case when $Y = \mathbb{C}$, the space $\mathcal{L}(X, Y)$ is called *the dual space of X* (or simply the *dual*, or *adjoint space* or *conjugate space* of X), and it is denoted as X'. Hence the elements of $X' := \mathcal{L}(X, \mathbb{C})$ are linear maps from X to $\mathbb{C}$, and they are called *linear functionals*. Remark also that since $\mathbb{C}$ is complete, then the last theorem shows that X' is always a Banach space, whether X is complete or not.

The operator norm induces a norm on X': if $l \in X'$, then

$$||l|| = \sup_{||x||=1} |l(x)|.$$

If X is a space of functions, then X' can be identified with either a space of functions or a space of distributions or a space of measures. Here are some examples of dual spaces:

1) $(L^p)' = L^q$, where $1/p + 1/q = 1$, if $1 \leq p < \infty$ (space of functions),

2) $(L^\infty)'$ is a space of measures which is much larger than L^1,

3) $(H_s)' = H_{-s}$ (space of distributions if $s > 0$).

Note that $(L^p)' \supset L^q$, for $1 \leq p < \infty$ follows from the Hölder inequality. In fact, given $f \in L^q$, define $l_f(u) := \int fu$. Since $|l_f(u)| \leq ||f||_q ||u||_p$, we see that l_f is a bounded linear functional on L^p. It can be shown that in fact any bounded linear functional on L^p can be represented by l_f for some $f \in L^q$.

Chapter III. Equations

Our goal in this chapter is to learn basic tools in solving various equations. Mostly we are interested in differential and integral equations, but the methods developed apply to other types of equations as well.

18. Calculus of maps

We will study equations of the form

$$F(u) = 0, \tag{18.1}$$

where u is an unknown function, and F is a map which takes a function u into another function.

For instance, look at the nonlinear Poisson equation:

$$-\Delta u + g(u) = f,$$

where f is a given function, then the map F is defined by

$$F(u) = -\Delta u + g(u) - f.$$

To solve equation (18.1), we have to choose a space to which the function u belongs; say we assume that u beongs to some Banach space X, and that F maps X into another Banach space Y.

Our goal is to develop a calculus of maps which will allow us to solve equations of the form (18.1).

Let us consider first several examples of maps $F : X \to Y$

1) $F(u) = \Delta u$,

2) $F(u) = f \circ u$ for a given function f,

3) $F(u) = \operatorname{div}(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}})$.

Depending on the problem at hand, we choose different spaces for the examples above. For instance we can choose

1) $X = H_2(\Omega)$ and $Y = L^2(\Omega)$,

2) $X = C^k(\Omega)$ and $Y = C^k(\Omega)$, if the function f is $C^k(\Omega)$,

3) $X = C^k(\Omega)$ and $Y = C^{k-2}(\Omega)$.

Note that the map in 1) is linear, while the map in 2) is linear if f is linear, and nonlinear otherwise.

If $X = Y$, equation (18.1) appears often in the form

$$F(u) = u. \tag{18.2}$$

A solution to (18.2) is called a *fixed point of the map F*.

19. The contraction mapping principle

Let X be a Banach space. Denote by $d(u, v) = ||u - v||$ the distance between the vectors u and v. Remark that actually all we need for the next theorem is that X is a complete metric space (i.e. it does not have to have a norm). A map $F : X \to X$ is called a *strict contraction* iff there is a number $\alpha \in (0, 1)$ s.t.

$$d(F(u), F(v)) \leq \alpha\, d(u, v), \quad \forall u, v \in X.$$

Theorem (the contraction mapping principle). *If F is a strict contraction, then F has a unique fixed point.*

Proof. The proof uses the method of successive approximations. We want to solve the equation $u = F(u)$. Pick some $u_0 \in X$ and define $u_1 = F(u_0), \dots, u_n = F(u_{n-1})$.

We claim that $\{u_n\}$ is a Cauchy sequence in X. In fact, let $n \geq m$, then $d(u_n, u_m) \leq \alpha^m d(u_{n-m}, u_0)$. Next, by the triangle inequality (i.e. $d(v,u) \leq d(v,w) + d(w,u)$, $\forall w \in X$), we get

$$\begin{aligned} d(u_k, u_0) &\leq d(u_k, u_{k-1}) + d(u_{k-1}, u_{k-2}) + \cdots + d(u_1, u_0) \\ &\leq \left(\alpha^{k-1} + \alpha^{k-2} + \ldots + 1\right) d(u_1, u_0) \\ &\leq \frac{1}{1-\alpha} d(u_1, u_0). \end{aligned}$$

The last two inequalities imply

$$d(u_n, u_m) \leq \frac{\alpha^m}{1-\alpha} d(u_1, u_0) \to 0 \quad \text{as } m, n \to \infty.$$

Thus indeed, $\{u_n\}$ is a Cauchy sequence in X. Now since X is complete, there is a $u \in X$ s.t. $u_n \to u$ so $d(F(u_n), F(u)) \leq \alpha d(u_n, u) \to 0$. Then the diagram

$$\begin{tikzcd} u_{n+1} \arrow[d] & = & F(u_n) \arrow[d] \\ u & & F(u) \end{tikzcd}$$

shows that $u = F(u)$. This demonstrates existence of a fixed point, and we finish the proof by showing its uniqueness. Suppose that $F(u) = u$, and $F(v) = v$. Then we have $d(F(v), F(u)) = d(v,u) \leq \alpha d(v,u)$, hence $d(v,u) = 0$ since $\alpha \in (0,1)$, and so $v = u$. ■

Application. Let Y be a Banach space, and $\Phi : Y \times I \to Y$, where $I = [0, T] \subset \mathbb{R}$ is an interval. We consider the differential equation (on the Banach space Y)

$$\partial_t u_t = \Phi(u_t, t), \tag{19.1}$$

with the initial condition $u_t|_{t=0} = u_0$.

Theorem. *Let Φ be continuous in $t \in I$, and Lipshitz continuous in $y \in Y$ (i.e. $||\Phi(u,t) - \Phi(v,t)||_Y \leq C||u - v||_Y$ for some $C < \infty$, and $0 \leq t \leq T$). Then for T sufficiently small, the differential equation (19.1) has a unique solution which is C^1 in t.*

Proof. We rewrite (19.1) as an integral equation

$$u_t = u_0 + \int_0^t \Phi(u_s, s)ds. \tag{19.2}$$

Define the map $F : X \equiv C(I, Y) \to C(I, Y)$ by

$$F(u.)_t = u_0 + \int_0^t \Phi(u_s, s)ds,$$

where $u.$ denotes the map $t \mapsto u_t$, and u_0 is the initial condition considered as a constant map $t \mapsto u_0$. Now equation (19.2) can be rewritten as $u. = F(u.)$, which is a fixed point equation for F. Note that the norm in the Banach space X is given by $||u.||_X = \sup_{t \in I} ||u_t||_Y$. We now show that for T sufficiently small, F is a strict contraction.

$$\begin{aligned} ||F(u.) - F(v.)||_X &:= \sup_{t \in I} ||F(u.)_t - F(v.)_t||_Y \\ &= \sup_{t \in I} || \int_0^t (\Phi(u_s, s) - \Phi(v_s, s))\, ds||_Y \\ &\leq TC \sup_{s \in T} ||u_s - v_s||_Y \\ &= TC||u. - v.||_X. \end{aligned}$$

Thus, for $T < 1/C$, F is a strict contraction, and hence the equation $u. = F(u.)$ has a unique solution in $C(I, X)$. The r.h.s. of (19.2) is differentiable in t, so u_t is differentiable in t, and it satisfies (19.1). ∎

20. The Gâteaux and Fréchet derivatives

The goal of this section is to develop a differential calculus of maps $F : X \to Y$, where X and Y are Banach spaces.

The map F is called *Gâteaux differentiable at* $u \in X$ iff there exists a bounded linear map $\mathcal{D}F(u) \in \mathcal{L}(X, Y)$ s.t. for any $\xi \in X$:

$$\left.\frac{\partial}{\partial \lambda}\right|_{\lambda=0} F(u + \lambda\xi) = \mathcal{D}F(u)\xi. \tag{20.1}$$

The Gâteaux derivative is sometimes called the *gradient map* or the *variational derivative.*

The map F is called *continuously differentiable at* $\overline{u} \in X$ iff it is Gâteaux differentiable for u in a neighbourhood of $\overline{u}$, and moreover, $u \mapsto \mathcal{D}F(u)$ is a continuous map from X to $\mathcal{L}(X, Y)$ at the point $\overline{u}$; i.e. if $u_n \to \overline{u}$ in X, then $\mathcal{D}F(u_n) \to \mathcal{D}F(\overline{u})$ in $\mathcal{L}(X, Y)$. This continuity condition is expressed equivalently by

$$\sup_{||u-\overline{u}||<\epsilon} ||\mathcal{D}F(u) - \mathcal{D}F(\overline{u})|| \to 0, \quad \text{as } \epsilon \to 0. \tag{20.2}$$

The map F is called *continuously differentiable, or* C^1 (written $F \in C^1$) iff it is continuously differentiable for all $u \in X$.

Examples.

1) If $F(u) = Lu$, where L is a linear map, then $\mathcal{D}F(u) = L$ (independently of u). Indeed, $\mathcal{D}F(u)\xi = \frac{\partial}{\partial\lambda}L(u + \lambda\xi)|_{\lambda=0} = \frac{\partial}{\partial\lambda}(Lu + \lambda L\xi)|_{\lambda=0} = L\xi$. Thus if L is bounded, then F is C^1.

2) If $F(u) = f \circ u$ (composition map), for a fixed C^1–function $f : \mathbb{R} \to \mathbb{R}$, and $u : \mathbb{R}^n \to \mathbb{R}$, then $\mathcal{D}F(u)$ is the multiplication operator by $f'(u)$. Indeed, $\mathcal{D}F(u)\xi = \frac{\partial}{\partial\lambda}F(u + \lambda\xi)|_{\lambda=0} = \frac{\partial}{\partial\lambda}f(u(x) + \lambda\xi(x))|_{\lambda=0} = f'(u)\xi$. So if $f'(u)$ is a bounded function, say for some $u \in L^p(\mathbb{R}^n)$, then $F : L^p(\mathbb{R}^n) \to L^p(\mathbb{R}^n)$ is differentiable at u.

Exercises. **1)** Compute $\mathcal{D}F(u)$ for $F : \mathbb{R}^n \to \mathbb{R}^m$, and for $F(u) = \text{div}(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}})$. **2)** Let K be a convex subset of a Banach space X (i.e. if $u, v \in K$, then $su + (1-s)v \in K$, for all $s \in [0, 1]$). Show that if $F : K \to K$ satisfies $||\mathcal{D}F(\psi)|| \leq \alpha$, $\forall\psi \in K$, then F is Lipshitz: $||F(\psi) - F(\varphi)|| \leq \alpha||\psi - \varphi||$, $\forall\psi, \varphi \in K$.

In applications, one often considers composition operators $F(u) = f \circ u$, where f is a fixed function and u belongs to the space of differentiable functions. The statements below are useful in this context.

Exercise. Let $F(u) = f \circ u$, and let Ω be a bounded domain in $\mathbb{R}^n$ with a smooth boundary. Show that if $f \in C^{k+1}(\mathbb{R})$, then $F : C^k(\overline{\Omega}) \to C^k(\overline{\Omega})$, and F is C^1 with $\mathcal{D}F(u)\xi = f'(u)\xi$.

An important result in this direction is the following

Theorem. *Let $F(u) = f \circ u$ and let $\Omega \subset \mathbb{R}^n$ be a bounded domain with smooth boundary. If $f \in C^{k+1}(\mathbb{R})$ with $k > n/2$, then $F : H_k(\Omega) \to H_k(\Omega)$, and F is C^1.*

For a proof, see [McO], page 221.

Discussion, the Fréchet derivative. Though the Gâteaux derivative is straightforward to compute, for theoretical considerations, one needs often a stronger notion of derivative: the *Fréchet derivative.* Before we define the Fréchet derivative, let us remark that equation (20.1) is equivalent to

$$F(u + \lambda\xi) - F(u) = \lambda \mathcal{D}F(u)\xi + o(\lambda), \tag{20.3}$$

were $o(\lambda)$ is a vector in Y satisfying $\lim_{\lambda\to 0} \|o(\lambda)\|/\lambda = 0$. Notice that in general, $o(\lambda)$ depends on ξ.

The map F is called *Fréchet differentiable at $u \in X$* iff there exists a bounded linear map $\mathcal{D}F(u) \in \mathcal{L}(X, Y)$ s.t.

$$F(u + \xi) - F(u) = \mathcal{D}F(u)\xi + o(\|\xi\|) \tag{20.4}$$

as $\|\xi\| \to 0$. The symbol $o(\|\xi\|)$ stands for a map $R : X \to Y$ s.t.

$$\frac{\|R\|}{\|\xi\|} \to 0, \quad \text{as } \|\xi\| \to 0.$$

The operator $\mathcal{D}F(u)$ satisfying (20.4) is called the Fréchet derivative of F at the point u.

From definition (20.4) and equation (20.3), it is clear that if F is Fréchet differentiable at u, with Fréchet derivative $\mathcal{D}F(u)$, then F is Gâteaux differentiable at u with Gâteaux derivative given by the same operator $\mathcal{D}F(u)$. The converse is also true if $F \in C^1$:

Theorem. *If F is continuously differentiable at $u \in X$, with Gâteaux derivative $\mathcal{D}F(u)$, then F is Fréchet differentiable at u, and the Fréchet derivative is given by the same operator $\mathcal{D}F(u)$.*

Proof. Define the function $g : [0,1] \to Y$ by

$$g(t) = F(u + t\xi),$$

for $u, \xi \in X$ fixed. According to the definition of the Gâteaux derivative (20.1), we have

$$g'(t) = \lim_{\tau \to 0} \frac{F(u + (t+\tau)\xi) - F(u + t\xi)}{\tau} = \mathcal{D}F(u + t\xi)\xi.$$

Now using the Mean Value Theorem for g, we obtain

$$\begin{aligned} & \|F(u+\xi) - F(u) - \mathcal{D}F(u)\xi\| \\ = \; & \|g(1) - g(0) - g'(0)\| \\ \leq \; & \sup_{0<t<1} \|g'(t) - g'(0)\| \\ \leq \; & \sup_{0<t<1} \|\mathcal{D}F(u+t\xi) - \mathcal{D}F(u)\| \, \|\xi\| \\ = \; & o(\|\xi\|). \end{aligned}$$

In the last step, we used the continuity (20.2). ■

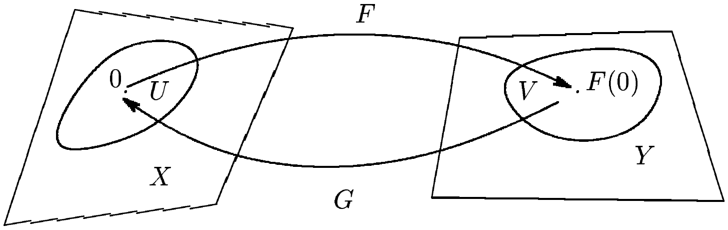
For a detailed discussion of Fréchet and Gâteaux derivatives, we refer to [ZeiI].

In everything that follows, by the derivative $\mathcal{D}F(u)$ we understand the Gâteaux derivative. We point out that in most of our applications, we deal with C^1 maps, in which case the Fréchet and Gâteaux derivatives coincide, according to the last theorem.

21. The inverse function theorem

Let X and Y be two Banach spaces. Recall that a map $G : Y \to X$ is called the *inverse* of the map $F : X \to Y$ iff $G \circ F = \mathbb{1}_X$ and $F \circ G = \mathbb{1}_Y$. Here, $\mathbb{1}_Z$ dentotes the identity on the space Z. We write $G = F^{-1}$. Recall that a *linear* map A is called invertible iff it has a bounded inverse. The following theorem is a generalization of the corresponding theorem in multivariable calculus:

The inverse function theorem. *Let U be an open neighbourhood of $0 \in X$, and let $F : U \to Y$ be a C^1 map s.t. $\mathcal{D}F(0) : X \to Y$ has a bounded inverse (i.e. $\mathcal{D}F(0) : X \to Y$ is bijective). Then there is a neighbourhood V of $F(0)$ in Y and a unique map $G : V \to X$ s.t. $F(G(y)) = y$, for all $y \in V$.*



Proof. Finding the inverse function G is equivalent to solving the equation

$$F(u) = v \tag{21.1}$$

for u, given $v \in V$. By the definition of the Fréchet derivative, we have

$$F(u) = F(0) + \mathcal{D}F(0)u + R(u), \tag{21.2}$$

where the remainder satisfies $R(u) = o(||u||)$. Due to equation (21.2), (21.1) is equivalent to

$$u = \mathcal{D}F(0)^{-1}\left[v - F(0) - R(u)\right]. \tag{21.3}$$

For any $y \in Y$, define the map

$$H_v(u) := \mathcal{D}F(0)^{-1}[v - F(0) - R(u)].$$

Then solving (21.3) is equivalent to solving $H_v(u) = u$ for u, i.e. we need to find a fixed point of H_v. Denote by $B_X(u, r)$ the open ball of radius r centered at u in X.

Claim. *$\exists \epsilon > 0$ and $\delta > 0$, s.t. (i) $H_v : B_X(0, \epsilon) \to B_X(0, \epsilon)$, provided $v \in B_Y(F(0), \delta)$, (ii) $||\mathcal{D}H_v(u)|| \leq 1/2$.*

Given (i) and (ii), we see that H_v for $v \in B_Y(F(0),\delta)$ is a contraction, therefore for all $v \in B_Y(F(0),\delta)$, H_v has a unique fixed point in $B_X(0,\epsilon)$. Call this fixed point $u = u(v)$. It solves $u = H_v(u)$, so $F(u) = v$.

It remains to prove the claim. For some ϵ, δ, let $||u|| \leq \epsilon$ and $||v - F(0)|| \leq \delta$. Then

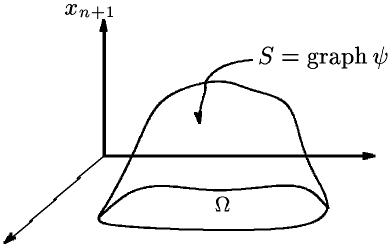
$$\begin{aligned} ||H_v(u)|| &\leq ||\mathcal{D}F(0)^{-1}R(u)|| + ||\mathcal{D}F(0)^{-1}(v - F(0))|| \\ &\leq ||\mathcal{D}F(0)^{-1}||(o(\epsilon) + \delta). \end{aligned}$$

Now find ϵ_1 s.t. $o(\epsilon) \leq \frac{\epsilon}{2||\mathcal{D}F(0)^{-1}||}$ for all $\epsilon \leq \epsilon_1$. Then $||H_v(u)|| \leq \epsilon$ for all $\epsilon \leq \epsilon_1$, provided $v \in B_Y(F(0),\delta)$, $\delta = \frac{\epsilon}{2||\mathcal{D}F(0)^{-1}||}$. Thus (i) follows.

Using that $R(u) + F(0) = F(u) - \mathcal{D}F(0)u$, we find $\mathcal{D}H_v(u) = \mathcal{D}F(0)^{-1}[\mathcal{D}F(0) - \mathcal{D}F(u)]$. Then by continuity of $\mathcal{D}F(u)$ in u, we obtain that there is an ϵ_2 s.t. $||\mathcal{D}H_v(u)|| \leq ||\mathcal{D}F(0)^{-1}||\,||\mathcal{D}F(0) - \mathcal{D}F(u)|| \leq 1/2$, if $||u|| \leq \epsilon_2$.

Now take $\epsilon \leq \min(\epsilon_1, \epsilon_2)$, and $\delta = \frac{\epsilon}{2||\mathcal{D}F(0)^{-1}||}$. Then also (ii) holds.■

Application of the inverse function theorem: existence of surfaces with prescribed mean curvature. Assume S is a hypersurface in $\mathbb{R}^{n+1}$, given as a graph of a function $\psi : \Omega \subset \mathbb{R}^n \to \mathbb{R}$, $S = \text{graph}\psi$. We assume Ω is bounded.



The *mean curvature of S* is given by

$$\operatorname{div}\left(\frac{\nabla\psi}{\sqrt{1+|\nabla\psi|^2}}\right), \tag{21.4}$$

for x in the interior of Ω.

Our question is: given a function $h(x)$, is there a surface $S = \text{graph}\psi$

which has mean curvature $h(x)$, i.e. is there a solution ψ to the equation $h =$(21.4)?

In order to find an answer, we define a map

$$F(\psi) = \operatorname{div}\left(\frac{\nabla\psi}{\sqrt{1+|\nabla\psi|^2}}\right),$$

and we want to solve $F(\psi) = h$. To do so, we want to use the inverse function theorem, and we need to define spaces X and Y s.t.

1) $F : U \to Y$, where U is a neighbourhood of $0 \in X$,

2) F is C^1,

3) $\mathcal{D}F(0)$ has a bounded inverse.

In a first step, let X, Y be the Sobolev spaces $X = H_k(\Omega)$, $Y = H_{k-2}(\Omega)$. In order to show 1), we define for $p \in \mathbb{R}^n$ the smooth function:

$$G(p) = \frac{p}{\sqrt{1+|p|^2}},$$

then $F(\psi) = \operatorname{div} G \circ \nabla\psi$. We know that $\nabla : H_k(\Omega) \to H_{k-1}(\Omega)$.

In the theorem of section 20, we saw that if $k - 1 > n/2$, then since G is smooth, the composition with G leaves $H_{k-1}(\Omega)$ invariant: $G\circ : H_{k-1}(\Omega) \to H_{k-1}(\Omega)$.

Finally, div : $H_{k-1}(\Omega) \to H_{k-2}(\Omega)$, and therefore the composition of these three maps satisfies:

$$F = \operatorname{div} \circ G \circ \nabla : \; H_k(\Omega) \to H_{k-2}(\Omega),$$

which shows 1).

In order to check 2), i.e. $F \in C^1$, we remember from a previous exercise that $\mathcal{D}F(\psi)\xi = \operatorname{div}(\frac{\nabla\xi}{\sqrt{1+|\nabla\psi|^2}})$.

Exercise. Show $\mathcal{D}F(\psi) : H_k(\Omega) \to H_{k-2}(\Omega)$ is bounded, and it is continuous in ψ (i.e. $||\mathcal{D}F(\psi_n) - \mathcal{D}F(\psi)|| \to 0$, as $||\psi_n - \psi|| \to 0$).

Finally, we have to verify that 3) is satisfied, i.e. that $\mathcal{D}F(0)$ has a bounded inverse. Now $\mathcal{D}F(0) = \Delta$, and we have discussed the existence of Δ^{-1} in Section 17, where we saw that for the inverse Laplacian to exitst, we must exclude the possibility of constant eigenfunctions of Δ, hence we need to take $X = H_k^{(0)}(\Omega)$.

Again, we have not shown yet boundedness of $\Delta^{-1} : H_{k-2}(\Omega) \to H_k^{(0)}(\Omega)$, this we will do in a later section, using variational calculus. Modulo the proof of this fact, we have thus shown that the conditions of the inverse function theorem are satisfied, and therefore, for any sufficiently small $h \in H_{k-2}(\Omega)$, the equation $F(\psi) = h$ has a unique solution $\psi \in H_k^{(0)}(\Omega)$. In other words, there exists a surface $S = \text{graph}\psi$ with prescribed small mean curvature h.

22. The implicit function theorem

Consider three Banach spaces X, Y and Z, and a map $F : X \times Y \to Z$. We wish to solve $F(x, y) = 0$ for y, i.e. we want to define y as a function of x by the equation $F(x, y) = 0$. We introduce the *partial Fréchet derivatives*, denoted by $\mathcal{D}_y F(x, y)$, etc.

The implicit function theorem. *Let U and V be neighbourhoods of $0 \in X$ and $0 \in Y$ respectively. Let $F : U \times V \to Z$ be a C^1–map s.t. $F(0,0) = 0$, and suppose $\mathcal{D}_y F(0,0)$ has a bounded inverse. Then there is a neighbourhood W of $0 \in X$ and a map $G : W \to Y$ such that $F(x, G(x)) = 0$, $\forall x \in W$.*

Proof. The proof is similar to that of the inverse function theorem, so we omit some details. We want to solve $F(x, y) = 0$ for y near $(x, y) = (0, 0)$. Expand F in y around 0: $F(x, y) = F(x, 0) + \mathcal{D}_y F(x, 0)y + R(x, y)$, with $R(x, y) = o(||y||)$. So our task is to solve the following equation for y:

$$F(x,0) + \mathcal{D}_y F(x,0)y + R(x,y) = 0,$$

or

$$y = -\mathcal{D}_y F(x,0)^{-1} \left(F(x,0) + R(x,y) \right). \tag{22.1}$$

If we neglect the remainder term $R(x, y)$, then equation (22.1) yields for each given x the corresponding $y = G(x)$. In reality, the remainder is not zero, but small in y, and we can use once again the fixed point argument to show existence of a solution to (22.1). To do so, introduce the map

$$H_x(y) := -\mathcal{D}_y F(x,0)^{-1} \left(F(x,0) + R(x,y) \right), \tag{22.2}$$

then equation (22.1) is equivalent to the fixed point equation $H_x(y) = y$. We now show that the map H_x has a fixed point $y = y(x)$. We have

$$\begin{aligned} \mathcal{D}_y H_x(y) &= -\mathcal{D}_y F(x,0)^{-1} \mathcal{D}_y R(x,y) \\ &= -\mathcal{D}_y F(x,0)^{-1} \left[\mathcal{D}_y F(x,y) - \mathcal{D}_y F(x,0) \right]. \end{aligned}$$

Now since F is a C^1–map, we have that for all (fixed) x, $\mathcal{D}_y F(x,y) - \mathcal{D}_y F(x,0) \to 0$ as $y \to 0$, and hence we get $||\mathcal{D}_y H_x(y)|| \leq 1/2$ if $||y|| \leq \epsilon(x)$ (notice that $||\mathcal{D}_y F(x,0)^{-1}|| \leq C$, for some constant $C > 0$, provided x is in some bounded domain W). Thus H_x is a contraction, and $\forall x \in W \;\; \exists y = y(x) \in B_{\epsilon(x)}$ solving (22.2). ■

23. Theory of bifurcation

Consider a C^1–map $F : \mathbb{R} \times Y \to Z$, s.t. $F(0,0) = 0$. Here, Y and Z are Banach spaces.

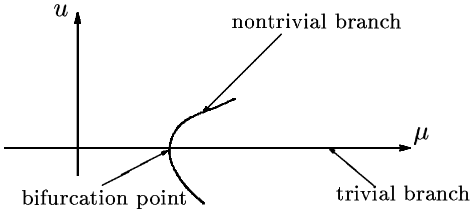
Our problem is to find a function $u = u(\mu)$, implicitely defined by the equation

$$F(\mu, u) = 0. \tag{23.1}$$

By the implicit function theorem, we know that if $\mathcal{D}_u F(0,0)$ has a bounded inverse, then equation (23.1) has a unique solution in a neighbourhood of $(0,0)$. Here, we look at the situation when $\mathcal{D}_u F(0,0)$ does not have a bounded inverse (e.g. if $\mathcal{D}_u F(0,0)$ has a zero eigenvalue). In this situation, the implicit function theorem is not applicable.

More specifically, we look at the following problem. Assume F satisfies $F(\mu, 0) = 0$ for all μ, i.e. $(\mu, 0)$ satisfies (23.1) for all μ. The branch of solutions $\{(\mu, 0) : \mu \in \mathbb{R}\}$ is called the *trivial branch.* The corresponding solutions are called *trivial solutions.* Our task is to find nontrivial sulutions to (23.1) in a vicinity of the trivial branch.

The "curve" $(\mu, u(\mu))$, $\mu \in [-\epsilon, \epsilon]$ is called a *branch of solutions* if $F(\mu, u(\mu)) = 0$. A point $(\mu_0, 0)$ at which a branch of nontrivial solutions appears is called a *bifurcation point.*



From what has been said above, it is clear that we have the following

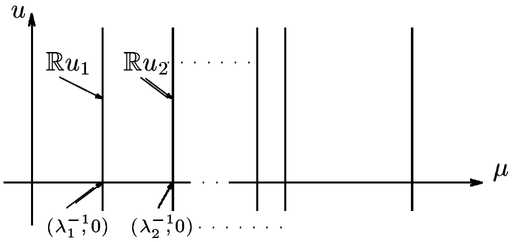
Proposition. *If $(\mu_0, 0)$ is a bifurcation point, then $\mathcal{D}_u F(\mu_0, 0)$ does not have a bounded inverse.*

An important example is given by the case when $F : \mathbb{R} \times Y \to Y$ is linear in u:

$$F(\mu, u) = \mu L u - u,$$

where L is a linear operator on Y. Then $(\mu, 0)$ is the trivial branch of solutions. Let us find the bifurcation points. The candidates for bifurcation points are the points where $\mathcal{D}_u F(\mu, 0)$ is not invertible. We have $\mathcal{D}F(\mu, 0) = \mu L - \mathbb{1}$. Assume here for simplicity that L has purely point spectrum (i.e. only eigenvalues, no continuous spectrum), then $\mu L - \mathbb{1}$ is not invertible iff 0 is an eigenvalue of $\mu L - \mathbb{1}$ (indeed, recall $0 \notin \sigma(A)$ iff A is invertible). Now $(\mu L - \mathbb{1})u_0 = 0 \Leftrightarrow L u_0 = \frac{1}{\mu} u_0$, i.e. $1/\mu$ is an eigenvalue of L. If $1/\mu$ is an eigenvalue of L, then we call μ a *characteristic value* of L. So if $\sigma(L) = \{\lambda_n\}_1^\infty$, with corresponding eigenfunctions u_n, then $(\lambda_n^{-1}, 0)$ are candidates for bifurcation points.

To show that the points $(\lambda_n^{-1}, 0)$ are indeed bifurcation points, notice that $F(\lambda_n^{-1}, u) = 0$ has nontrivial solutions in a neighbourhood of $u = 0$: indeed, $F(\lambda_n^{-1}, u) = \lambda_n^{-1} L u - u = 0$ is the eigenequation for u_n, and it has solutions $u = 0$ and $u = a u_n$, where $a \in \mathbb{R}$ (or $a \in \mathbb{C}$).



Example. Let $Y = Z = L^2([0, 2\pi])$, $L = -\Delta$ with Dirichlet boundary conditions: $u(0) = u(2\pi) = 0$. Recall that the domain of L satisfies $\mathcal{D}(L) = H_2([0, 2\pi]) \subset C([0, 2\pi])$ by the Sobolev embedding theorem, and therefore the boundary conditions $u(0) = u(2\pi) = 0$ make sense.

To find the eigenvalues of L, we need to solve the characteristic equation $-\Delta u = \lambda u$, i.e. $u'' = -\lambda u$. The solutions satisfying the Dirichlet boundary conditions are $u_n = a \sin(\frac{n}{2}x)$, where $a \in \mathbb{R}$, and the eigenvalues are given by $\lambda_n = (\frac{n}{2})^2$, $n = 1, 2, 3, \ldots$. Thus besides the trivial branch $(\mu, 0)$, $\mu \in \mathbb{R}$, the equation $\mu(-\Delta u) - u = 0$ has the branches of solutions $((\frac{2}{n})^2, \mathbb{R}u_n)$, for $n = 1, 2, 3, \ldots$.

Remember that in the the above examples, F is *linear* in u, and as a result, the bifurcating branches are straight lines. In general, if F is nonlinear, we expect the bifurcating branches to be bent, as in the following example.

Example. Let $Y = Z = \mathbb{R}$, and $F(\mu, u) = \mu u - u^3$. Clearly we have $F(\mu, 0) = 0 \ \forall \mu \in \mathbb{R}$, so $(\mu, 0)$ is the trivial branch. We calculate the derivative $\mathcal{D}_u F(\mu, u) = \mu - 3u^2$, so $\mathcal{D}_u F(\mu, 0) = \mu = 0$ has the solution $\mu_0 = 0$, hence $(0, 0)$ is a candidate for a bifurcation point. On the other hand, we can solve the equation $F(\mu, u) = 0$ explicitely, obtaining the solutions $(\mu, 0)$ and $u = \pm\sqrt{\mu}$. This shows that $(0, 0)$ is indeed a bifurcation point (the bifurcation here is called a pitchfork bifurcation because of the shape of its bifurcating branch).

In later sections, we will learn how to find out the qualitative behaviour of the bifurcating branch without actually solving for it. But before, let us find a sufficient condition for a bifurcation to happen at a point $(\mu_0, 0)$ (notice that the last proposition only gave a necessary condition).

24. Sufficient condition for bifurcations: the Krasnoselski theorem

In the last section, we have seen that $(\mu_0, 0)$ is a possible bifurcation point only if $\mathcal{D}_u F(\mu, 0)$ is not invertible. This is however not a sufficient condition, as is demonstrated in the following example.

Example. For $F : \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}^2$ given by $F(\mu, u_1, u_2) = (u_1, u_2) - \mu(u_1 - u_2^3, u_2 + u_1^3)$, we find

$$\mathcal{D}_u F(\mu, u)(\xi_1, \xi_2) = (\xi_1, \xi_2) - \mu(\xi_1 - 3u_2^2\xi_2, \xi_2 + 3u_1^2\xi_1),$$

and therefore $\mathcal{D}_u F(\mu, 0) = (1 - \mu)\mathbb{1}$. Thus $\mathcal{D}_u F(1, 0)$ is not invertible. However, $(1, 0)$ is not a bifurcation point! Indeed, look at the two components of the equation $F(\mu, u) = 0$. Multiplying the first one by $-u_2$, the second one by u_1, we obtain

$$\begin{aligned} -(1-\mu)u_1u_2 - \mu u_2^4 &= 0 \\ (1-\mu)u_1u_2 - \mu u_1^4 &= 0. \end{aligned}$$

Adding the above two equations yields $-\mu(u_1^4 + u_2^4) = 0$, so $u_1 = u_2 = 0$ (for $\mu \neq 0$), which shows that $F(\mu, u_1, u_2) = 0$ has only the trivial solution $(u_1, u_2) = (0, 0)$, $\forall \mu \in \mathbb{R}$ (if $\mu = 0$, then this follows directly from the definition of F). $(1, 0)$ is therefore not a bifurcation point.

We now want to give a sufficient condition for a bifurcation to take place in a slightly specialized case. Namely, let $F : \mathbb{R} \times Y \to Z$ (Y, Z Banach spaces), and let $(\mu_0, 0)$ be a candidate for a bifurcation point, i.e. we assume $\mathcal{D}_u F(\mu, u)$ is not invertible at $(\mu_0, 0)$. Our simplifying assumptions are

i) $Y = Z$ is a Hilbert space (so in particular, $F : \mathbb{R} \times Y \to Y$),

ii) the spectrum of $\mathcal{D}_u F(\mu_0, 0)$ is discrete in the vicinity of $z = 0 \in \mathbb{C}$.

Remark that if $\mathcal{D}_u F(\mu_0, 0)$ is not invertible, then we must have $0 \in \sigma(\mathcal{D}_u F(\mu_0, 0))$. Condition ii) tells us something about "how $\mathcal{D}_u F(\mu_0, 0)$ is not invertible".

In what follows, we use the notation $F_{u\mu} = \partial_\mu \mathcal{D}_u F$, and the following definitions:

(a) the *multiplicity* of an eigenvalue λ of L is $\dim \operatorname{Null}(L-\lambda)$, i.e. the number of linearly independent eigenvectors with eigenvalue λ,

(b) the *algebraic multiplicity* of λ is $\dim \operatorname{span}\{\bigcup_{r\geq 1} \operatorname{Null}(L-\lambda)^r\}$, i.e. the number of linearly independent eigenvectors and root vectors with the eigenvalue λ.

If L is self–adjoint, then the multiplicity and the algebraic multiplicity coincide.

Theorem (Krasnoselski). *Assume that*
(i) 0 *is an eigenvalue of* $\mathcal{D}_u F(\mu_0, 0)$ *of odd algebraic multiplicity, and*
(ii) $\exists\, v_0 \in \operatorname{Null} \mathcal{D}_u F(\mu_0, 0)$ *s.t.* $\langle v_0, F_{u\mu}(\mu_0, 0)v_0\rangle \neq 0$.
Then $(\mu_0, 0)$ *is a bifurcation point.*

Examples. **1)** As in the last example of the previous section, let $F(\mu, u) = \mu u - u^3$. We have $\mathcal{D}_u F(\mu, u) = \mu - 3u^2$, so 0 is an eigenvalue of $\mathcal{D}_u F(0,0)$ of multiplicity 1. Next, $\partial_\mu \mathcal{D}_u F(0,0) = 1$, so the condition (ii) is satisfied as well. Therefore $(0,0)$ is a bifurcation point.

2) Let $\mathcal{D}_u F(\mu, 0) = \mu L - \mathbb{1}$, where L a linear operator. Then $F_{u\mu}(\mu, 0) = L$, and $\langle u_0, F_{u\mu}(\mu, 0)u_0\rangle = \langle u_0, Lu_0\rangle = \mu_0^{-1}||u_0||^2 \neq 0$ $\forall u_0 \in \operatorname{Null}(\mu_0 L - \mathbb{1})$. This yields the following

Corollary. *Let* $\mathcal{D}_u F(\mu, 0) = \mu L - \mathbb{1}$, *where* L *is a linear operator. If* μ_0 *is a characteristic value of* L *of odd algebraic multiplicity, then* $(\mu_0, 0)$ *is a bifurcation point.*

To illustrate the corollary, let us consider the nonlinear eigenvalue problem

$$Lu + f(u) = \lambda u, \quad \text{with} \quad f(u) = o(||u||). \tag{24.1}$$

Then the corollary implies that if λ_0 is an eigenvalue of L of odd algebraic multiplicity, then equation (24.1) has a nontrivial branch of solutions near the bifurcation point $(\lambda_0^{-1}, 0)$.

Before giving the actual proof of the Krasnoselski theorem, let us discuss its idea. Our goal is to solve the equation

$$F(\mu, u) = 0, \tag{24.2}$$

for (μ, u) near the point $(\mu_0, 0)$, i.e. for u and $\mu - \mu_0$ small. We expand $F(\mu, u)$ in u around the point $u = 0$, using that $F(\mu, 0) = 0\,\forall\mu$:

$$F(\mu, u) = \mathcal{D}_u F(\mu, 0)u + R(\mu, u), \tag{24.3}$$

where $R(\mu, u) = o(||u||)$. Substitute this into equation (24.2) to obtain

$$\mathcal{D}_u F(\mu, 0)u = -R(\mu, u). \tag{24.4}$$

Now we want to solve this equation for u. If $\mathcal{D}_u F(\mu, 0)$ were invertible in a neighbourhood of $\mu = \mu_0$, then we would get

$$u = -\mathcal{D}_u F(\mu, 0)^{-1} R(\mu, u),$$

which implies that

$$||u|| \leq ||\mathcal{D}_u F(\mu, 0)^{-1}|| \cdot o\left(||u||\right),$$

so $u = 0$ is the only solution to (24.2). Notice that this is of course exactly the idea of the proof of the implicit function theorem.

Here however, the key point is that $\mathcal{D}_u F(\mu_0, 0)$ is not invertible, more precisely, it has a zero eigenvalue. This implies that for μ close to μ_0, $\mathcal{D}_u F(\mu, 0)^{-1}$ has also a zero eigenvalue or at least an eigenvalue close to zero, so that even if $\mathcal{D}_u F(\mu, 0)^{-1}$ for $\mu \neq \mu_0$ exists as a bounded operator, it blows up (becomes unbounded) as $\mu \to \mu_0$. This fact allows for a nontrivial solution to (24.2) to pop up for $\mu \neq \mu_0$.

To solve equation (24.2), we observe that though $\mathcal{D}_u F(\mu_0, 0)$ is not invertible on the entire space, it is invertible as an operator from $(\mathrm{Null}\,\mathcal{D}_u F(\mu_0, 0))^{\perp}$ to $(\mathrm{Null}\,\mathcal{D}_u F(\mu_0, 0)^*)^{\perp}$, the latter space being the orhtogonal complement of the zero eigenvectors of the adjoint operator $\mathcal{D}_u F(\mu, 0)^*$. So we can solve equation (24.4) on the subspace $(\mathrm{Null}\,\mathcal{D}_u F(\mu_0, 0))^{\perp}$, and afterwards deal with the remaining part of the whole space, namely

$$\mathrm{Null}\,\mathcal{D}_u F(\mu_0, 0),$$

which is *finite dimensional.* This procedure reduces the infinite dimensional problem to a finite dimensional one. The sort of behaviour of an equation we describe here, namely when an equation can be solved on the whole space except on a finite dimensional subspace, where all the action takes place (the solution outside is the trivial one!), and where the solution has to be examined separately on the finite dimensional subspace, is quite reoccurring in applications, and the ideas explained below lie in the foundations of many mathematical methods.

Proof of the Krasnoselski theorem. Let $L(\mu) := \mathcal{D}_u F(\mu, 0)$, and denote by P the projection onto the subspace $\mathrm{Null}\, L(\mu)$, and let $\overline{P} := \mathbb{1} - P$. Then P^* is the projection onto the space $\mathrm{Null}\, L(\mu)^*$. We project the equation $F(\mu, u) = 0$ onto the subspaces $\mathrm{Null}\, L(\mu)^*$ and $(\mathrm{Null}\, L(\mu)^*)^{\perp}$:

$$\begin{aligned} P^* F(\mu, u) &= 0, \qquad (24.5)\\ \overline{P}^* F(\mu, u) &= 0, \qquad (24.6)\end{aligned}$$

and we decompose $u \in X$ along the two subspaces $\mathrm{Null}\, L(\mu)$ and $(\mathrm{Null}\, L(\mu))^{\perp}$: $u = v + w$, where $v \in \mathrm{Ran} P$ and $w \in \mathrm{Ran} \overline{P}$. We have thus two equations, (24.5) and (24.6), for two variables v and w. Observe that since $\dim \mathrm{Ran} P < \infty$, v is a finite–dimensional variable.

To solve equations (24.5) and (24.6), we proceed as follows. First, we solve (24.6) for $w = w(\mu, v)$, and substitute this solution into (24.5) to obtain the equation

$$PF(\mu, v + w(\mu, v)) = 0. \qquad (24.7)$$

In a second step, we solve equation (24.7). This equation is called the *bifurcation equation* or *branching equation.* It describes the bifurcating branches, and usually, it is a system of $n = \dim \mathrm{Null}\, L(\mu)$ algebraic equations for $n+1$ variables μ and v. We consider (24.7) as an equation for v as a function of μ. However, in general, this equation has several solutions v for one given μ. To parametrize these solutions, we proceed as follows. Let $v = (v_1, \dots, v_n)$, and pick one of these variables as a parameter, say v_n, then solve (24.7) for $(\mu, v_1, \dots, v_{n-1})$ as a function of v_n, say $(\mu, v_1, \dots, v_{n-1}) = \gamma(v_n)$.

We now carry out the above mentioned first step: we solve equation (24.6), i.e. we show equation (24.6) has a unique solution w. Define

$$F_1(\mu, v, w) := \overline{P}^* F(\mu, v + w) : \mathbb{R} \times PY \times \overline{P}Y \to \overline{P}^* Y.$$

In order to identify this situation with a standard implicit function theorem, we denote $X := \mathbb{R} \times PY$. Now observe that

(α) F_1 is C^1,

(β) $F_1(\mu, 0, 0) = 0$ for any μ,

(γ) $\mathcal{D}_w F_1(\mu_0, 0, 0)$ is invertible.

Indeed, (α) follows from the condition that F is C^1, (β) results from the relation $F_1(\mu, 0, 0) = \overline{P}^* F(\mu, 0) = 0$, and ($\gamma$) is due to the relation $\mathcal{D}_w F_1(\mu_0, 0, 0) = \overline{P}^* \mathcal{D}_u F(\mu_0, 0) \overline{P}$ plus the fact that the r.h.s. is invertible as an operator from $\overline{P}Y$ to $\overline{P}^* Y$.

The implicit function theorem shows thus that for any (μ, v) sufficiently close to $(\mu_0, 0)$, equation (24.6) has a unique solution, which we denote $w = w(\mu, v)$. This completes the first step of the proof.

Before proceeding to the second step, we prove the following important property of the solution $w(\mu, v)$:

$$w = o(||v||). \tag{24.8}$$

In order to show this, we expand the map $F(\mu, u)$ around $u = 0$:

$$F(\mu, u) = L(\mu)u + R(\mu, u), \tag{24.9}$$

with $R(\mu, u) = o(||u||)$, and where we used the fact that $F(\mu, 0) = 0$ for any μ. This expansion together with equation (24.6) and the fact that $L(\mu)P = 0$ implies

$$\overline{P}^* L(\mu) \overline{P} w + \overline{P}^* R(\mu, u) = 0,$$

and since the operator $\overline{L}(\mu) := \overline{P}^* L(\mu) \overline{P} : \overline{P}Y \to \overline{P}^* Y$ is invertible, we derive

$$w = -\overline{L}(\mu)^{-1} \overline{P}^* R(\mu, u) = o(||u||),$$

which shows (24.8), since $u = v + w$.

We now show the second step, i.e. we solve equation (24.7). For this, we use conditions (i) and (ii) of the theorem. Assume for simplicity that $\dim \mathrm{Null}\, L = 1$, i.e. that the eigenvalue 0 of L is simple, and we show that equation (24.7) has a unique solution for μ as a function of $v \in \mathbb{R}$. Let v_0 and v_0^* be the normalized zero eigenvectors of the operator L and L^* respectively. Then $Pu = \langle v_0^*, u \rangle v_0$. Since $v = sv_0$ for some $s \in \mathbb{R}$, equation (24.7) is equivalent to

$$f(s,\mu) = 0 \quad \text{where} \quad f(s,\mu) = \frac{1}{s}\langle v_0^*, F(\mu, sv_0 + w(\mu, sv_0))\rangle. \qquad (24.10)$$

Using expansion (24.9), we rewrite $f(s,\mu)$ as

$$f(s,\mu) = \langle v_0^*, \mathcal{D}_u F(\mu, 0)(v_0 + w_1)\rangle + \langle v_0^*, s^{-1}R(\mu, su_1)\rangle, \qquad (24.11)$$

where $u_1 := s^{-1}u$ and $w_1 := s^{-1}w$. Since $||w_1||$ and $s^{-1}||\partial_\mu R(\mu, su_1)|| \to 0$ as $s \to 0$, we have that

$$\frac{\partial f}{\partial \mu}(0, \mu_0) = \langle v_0^*, F_{u\mu}(\mu_0, 0)v_0\rangle \neq 0, \qquad (24.12)$$

by condition (ii) of the Krasnoselski theorem. Therefore equation (24.10) has a unique solution $\mu = \mu(s)$, for μ as a function of s, if s is in a neighbourhood of $s = 0$. This completes the second step.

We have shown that in the case when 0 is a simple eigenvalue of L, the solution of the original problem has the form

$$\begin{cases} u & = v + w(\mu, v), \\ \mu & = \mu(v). \end{cases} \qquad (24.13)$$

The second equation defines v as a function of μ and it has several solutions. ■

25. Type of bifurcations and stability

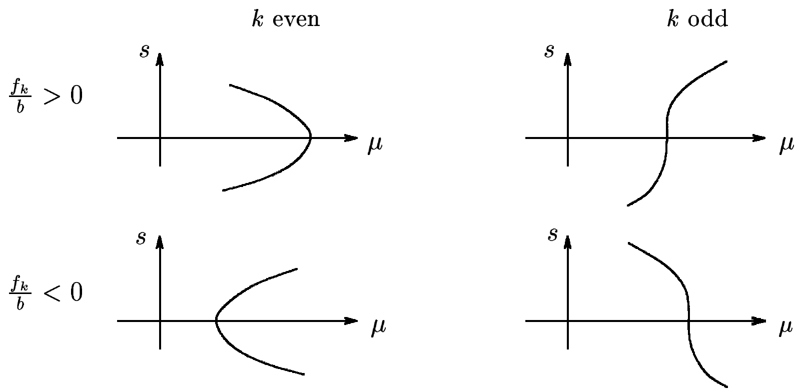
Type of bifurcations. We want to investigate the shape of a bifurcating branch of nontrivial solutions. We will deal with the special case when 0 is a simple eigenvalue of $L := \mathcal{D}_u F(\mu_0, 0)$. We solve equation (24.10) in the leading order in s (for small s). The equivalence relation $A \doteq B$ will stand for an equality in the leading order in s. Let $b = \frac{\partial f}{\partial \mu}(0, \mu_0) \neq 0$ (see equation (24.12)). Then the Taylor expansion theorem yields

$$\mu - \mu_0 \doteq -\frac{f(s, \mu_0)}{\frac{\partial f}{\partial \mu}(s, \mu_0)} \doteq -b^{-1} f(s, \mu_0). \tag{25.1}$$

Since v_0^* is the eigenfunction of L^* with eigenvalue 0, equation (24.11) implies

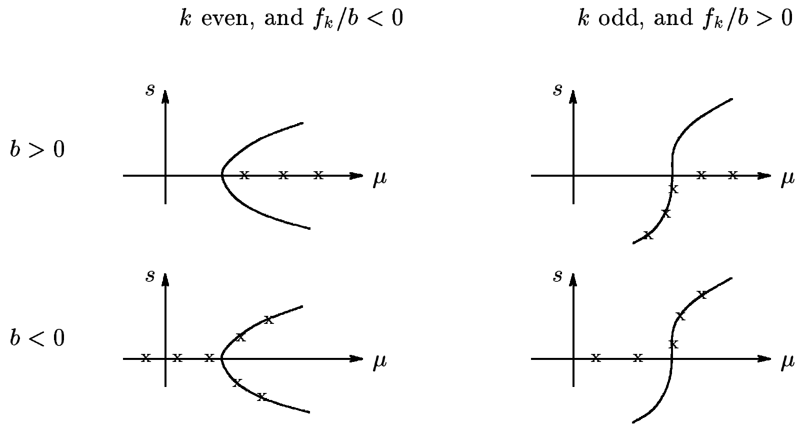
$$f(s, \mu_0) = \langle v_0^*, s^{-1} R(\mu, s u_1) \rangle,$$

hence $f(0, \mu_0) = 0$. Therefore, there is a $k \geq 1$ s.t. $f(s, \mu_0) \doteq f_k s^k$, for some f_k, provided $f(s, \mu_0)$ is C^{k+1} in s (this means in particular that we have to assume that F is in C^{k+2}). Then we can rewrite (25.1) as $\mu - \mu_0 \doteq -\frac{f_k}{b} s^k$, and consequently, we get the following qualitative pictures:



Stability. One of the important consequences of bifurcations is the change of stability at the bifurcation point. We return to this question later, here let us just illustrate this with a picture and an example.

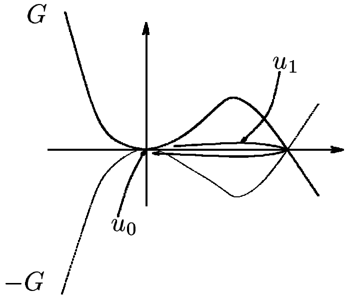
In the picture below, the crossed branches are stable and those which are not crossed are unstable.



Example. Let L be a very large fixed number, and take the family of spaces $[-\alpha/2, \alpha/2] \times [-L/2, L/2]$, where $\alpha \in \mathbb{R}$ is a parameter. On these spaces, consider the equation

$$-\Delta u + G'(u) = 0, \tag{25.2}$$

with the potential G of the form



and with periodic boundary conditions:

$$\begin{aligned} u(-\alpha/2, y) &= u(\alpha/2, y) \quad \forall y, \\ u(x, -L/2) &= u(x, L/2) \quad \forall x. \end{aligned} \tag{25.3}$$

The equation (25.2) with boundary conditions (25.3) has the following branches of solutions: (α, u_0), $\forall \alpha$, and (α, u_1), $\forall \alpha$, where u_0 and u_1 are solutions to the equation

$$-\frac{\partial^2 u}{\partial y^2} + G'(u) = 0 \quad \text{on} \quad [-L/2, L/2],$$

with periodic boundary conditions. The last equation is just Newton's equation if we interpret y as being the time-variable, and $-G$ as a potential (whose derivative is a force). u_0 is a minimizer, and u_1 is a saddle point of the functional

$$\mathcal{E}(u) = \int_{-L/2}^{L/2} \left[\frac{1}{2} \left(\frac{\partial u}{\partial y} \right)^2 + G(u) \right] dy.$$

Exercises. **1)** Check whether any solutions bifurcate from (α, u_0), **2)** find the bifurcation points from the branch (α, u_1), **3)** find the bifurcation points for $-\Delta u + \lambda u + u^3 = 0$ on $[-L, L]^n$ with Dirichlet boundary conditions (i.e. $u = 0$ on the boundary), and where $u \in H_2([-L, L]^n)$.

Chapter IV. Variational Calculus

The variational calculus deals with finding extrema or more generally, critical points, of real functions of an infinite number of variables – the *functionals*. It is one of the key tools in analysis. The Variational Calculus originates in the problem of minimization of energy functionals and finding stationary (i.e. critical) points of action functionals in physics. Presently, it is used in practically every field on science, and in particular to solve nonlinear and linear differential equations.

26. Functionals

Functionals are maps which have $\mathbb{R}$ as the target space. More precisely, let X be a Banach space, and $M \subset X$ a not necessarily open subset of X. Then a *functional* is a map $F : M \to \mathbb{R}$. Usually, X is a functional space. If X has a basis, then functionals on X can be represented as functions of an infinite number of coordinates along the basis. If X is a finite–dimensional space (which we are not concerned with here), then a functional on X is just a usual function of several variables. In the following list of examples of functionals, Ω is a domain in $\mathbb{R}^n$:

1) $\mathcal{E}(u) = \int_\Omega G(u(x))d^nx$, where $G : \mathbb{R} \to \mathbb{R}$, and $u : \Omega \to \mathbb{R}$;

2) $\mathcal{E}(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 d^nx$, where $u : \Omega \to \mathbb{R}$;

3) $\mathcal{E}(u) = \int_\Omega (\frac{1}{2}|\nabla u|^2 + G(u))d^nx$, where G and u as above;

4) $S(\varphi) = \int_0^T [\frac{1}{2}|\frac{\partial\varphi}{\partial t}|^2 - V(\varphi)]dt$, where $\varphi : [0, T] \to \mathbb{R}^m$, and $V : \mathbb{R}^m \to \mathbb{R}$;

5) $S(\varphi) = \int_0^T \int_\Omega [\frac{1}{2}|\frac{\partial\varphi}{\partial t}|^2 - \frac{1}{2}|\nabla\varphi|^2 - G(\varphi)]d^n x dt$, where $\varphi : \Omega \times [0, T] \to \mathbb{R}$, and G as above;

6) $\mathcal{E}(u) = \frac{1}{2}\langle u, Au\rangle$, where $u \in V$, an inner product space, and $A : V \to V$ is a linear operator;

7) $S(\psi) = \frac{1}{2}\int_0^T \int_\Omega (-\mathrm{Im}(\psi\dot{\bar{\psi}}) + |\nabla\psi|^2 + G(|\psi|^2))d^n x dt$, where $\psi : \Omega \times [0, T] \to \mathbb{C}$, and G as above;

8) $S(f) = \int f \log f d^n p$, where $f : \mathbb{R}^n \to \mathbb{R}^+$.

In these examples, we have used different letters for various functionals to indicate their physical origin: the functionals in examples 1)–3) and 6) originate in expressions for energy, the functionals in examples 4), 5) and 7) come from expressions for an action, and 8) is an expression for entropy.

We also have to specify the spaces on which the functionals are defined. Unlike for linear operators, we do not distinguish between the spaces on which the functionals are defined and their domains of definition. The (not necessarily linear or vector) spaces are chosen according to the specific functional and the problem at hand. Of course, we always try to choose the simplest possible space for a given problem.

For instance consider example 1). Assume that the domain Ω is bounded, and that the function G satisfies the estimate

$$|G(u)| \leq C|u|^p + C, \tag{26.1}$$

for some constant $C > 0$. It is then natural to define $\mathcal{E}$ on the space $L^p(\Omega)$.

In example 2), we define $\mathcal{E}$ on the Sobolev–space $H_1(\Omega)$.

In example 3), if Ω is bounded, and G satisfies (26.1), then we define $\mathcal{E}$ on $H_1(\Omega) \cap L^p(\Omega)$.

For the other examples, the spaces are chosen similarly.

Let us now return to the important example 2). There is another, even more popular space on which we can define this functional. Assume we want to vary u among functions in $H_1(\Omega)$ with fixed values on the boundary $\partial\Omega$. So let $g : \partial\Omega \to \mathbb{R}$ be smooth and put

$$H_s^{(g)}(\Omega) = \{u \in H_s(\Omega) : u = g \quad \text{on} \quad \partial\Omega\}.$$

Since the boundary $\partial\Omega$ has n–dimensional Lebesgue-measure zero, we have to be careful about the meaning of "$u = g$ on $\partial\Omega$". There is an equivalent way of defining $H_s^{(g)}$. Let $\tilde{g} \in H_s(\Omega)$ be a smooth real-valued function s.t. $\tilde{g} = g$ on $\partial\Omega$. Then we can define

$$H_s^{(g)}(\Omega) := \{u \in H_s(\Omega) : u - \tilde{g} \in H_s^{(0)}(\Omega)\}.$$

Clearly, this definition is independent of the choice of the extension $\tilde{g}$. This space can also be written as

$$H_s^{(g)}(\Omega) = \tilde{g} + H_s^{(0)}(\Omega).$$

If $g \neq 0$, then $H_s^{(g)}$ is not a vector space, but it is an "affine" space, and for most purposes it is as convenient to study as the vector space H_s itself.

The Gâteaux derivative for functionals. Let us now examine in more detail the notion of Gâteaux derivative in the case of functionals. Recall that the Gâteaux derivative of a map $F : X \to Y$ (where X, Y are Banach spaces) at a point $u \in X$ is the linear operator $\mathcal{D}F(u) : X \to Y$ defined by $\mathcal{D}F(u)\xi = \frac{\partial}{\partial\lambda}F(u_\lambda)|_{\lambda=0}$, where $u_\lambda := u + \lambda\xi$, for $\xi \in X$.

Consider now a functional $F : M \to \mathbb{R}$. If M is an open subset of a Banach space X, then the Gâteaux derivative $\mathcal{D}F(u)$, $u \in M$, is a linear functional on X. Recall that F is C^1 iff $\mathcal{D}F(u)$ is a bounded linear functional, i.e. $\mathcal{D}F(u) \in X'$ and that if X is a Hilbert space, then X' can be identified with X and therefore we can consider $\mathcal{D}F(u) \in X$. This identification is displayed as

$$\mathcal{D}F(u)\xi = \langle \mathcal{D}F(u), \xi \rangle,$$

on the l.h.s. of which $\mathcal{D}F(u)$ appears as an element in X', while on the r.h.s. it appears as an element of X. At first sight, this might seem confusing, but after getting used to it one appreciates the convenience of this ambiguous notation.

Let us now consider a simple example displaying the ambiguity mentioned above (more examples are to follow). Let G be a real differentiable function on $\mathbb{R}$ satisfying the estimate

$$|G(u)| + |G'(u)| \leq C|u|^2,$$

where C is independent of $u \in \mathbb{R}$. Then the functional

$$u \mapsto \int_\Omega G(u(x))d^nx$$

is defined on $L^2(\Omega)$, where $\Omega \subset \mathbb{R}^n$. Its Gâteaux derivative is

$$\left(\mathcal{D}\int_\Omega G \circ u \; d^nx\right)\xi = \int_\Omega G'(u(x))\xi(x)d^nx.$$

Therefore we can either identify the Gâteaux derivative in this case with either the linear functional standing on the l.h.s., or with the $L^2(\Omega)$–function $G'(u(x))$.

Exercise. Let Ω be a bounded domain in $\mathbb{R}^n$. Show that $\frac{1}{2}\int_\Omega |\psi|^2$ is C^1 on $L^2(\Omega)$ and on $H_1^{(0)}(\Omega)$.

Now, if M is not an open subset of a Banach space X, then the situation is more subtle. We cannot in general take a piece of a straight line $u_\lambda = u + \lambda\xi$ in the definition of $\mathcal{D}F(u)$, but rather we have to take "curves" $\lambda \mapsto u_\lambda$, s.t. $u_0 = u$ and $\frac{du_\lambda}{d\lambda}|_{\lambda=0} = \xi$ for a given ξ. Then we define

$$\mathcal{D}F(u)\xi := \frac{d}{d\lambda}F(u_\lambda)|_{\lambda=0}.$$

The "initial velocities" ξ might not span the entire space X, but only a subspace of X. In general, we define the *tangent space* to M at u, T_uM, to be the set of all the $\xi \in X$ s.t. there is an $\epsilon > 0$, and a path $[-\epsilon, \epsilon] \ni \lambda \mapsto u_\lambda \in M$ for which $u_0 = u$, $\frac{du_\lambda}{d\lambda}|_{\lambda=0} = \xi$. Then we have by definition $\mathcal{D}F(u) : T_uM \to \mathbb{R}$, i.e. $\mathcal{D}F(u) \in (T_uM)' =: T_u^*M$.

Exercise. Show that $T_uX = X$, $T_uH_s^{(g)}(\Omega) = H_s^{(0)}(\Omega)$, and if $M = \{u \in X : J(u) = 0\}$, where J is C^1 and X is a Hilbert space, then for $u \in M$, $T_uM = \mathcal{D}J(u)^{\perp}$.

We go now back to the Gâteaux derivative of a map $F : M \to \mathbb{R}$, where $M \subset X$, and X is a Banach space. If F is C^1, then $\forall u \in M$, $\mathcal{D}F(u) \in X'$. Let us now consider the map

$$u \mapsto \mathcal{D}F(u). \tag{26.2}$$

If this map (from M to X') is C^1, then we say that F is C^2. The Gâteaux derivative of the map (26.2) at u is the second derivative of F at u, denoted by $\mathcal{D}^2F(u)$. Continuing in this way, we define the notion of a C^k functional.

The operator $\mathcal{D}^2F(u) : X \to X'$ is called the *Hessian* of F at u. If X is a Hilbert space, then $\mathcal{D}^2F(u)$ is symmetric:

$$\langle \xi, \mathcal{D}^2F(u)\eta\rangle = \langle \mathcal{D}^2F(u)\xi, \eta\rangle, \quad \forall \xi, \eta \in X.$$

In fact, it is self–adjoint in practically all cases of interest. We give some examples of Hessians:

1) the finite–dimensional case: $f : \mathbb{R}^N \to \mathbb{R}$. Then $\mathcal{D}f(u) = \nabla u$ (the gradient), and $\mathcal{D}^2f(u) = \text{Hess}f(u) = (\frac{\partial^2 f}{\partial u_i \partial u_j}(u))$.

2) If $\mathcal{E}(u) = \frac{1}{2}\int |\nabla u|^2$ on $H_2^{(0)}(\Omega)$, then $\mathcal{D}\mathcal{E}(u)\xi := \frac{\partial}{\partial \lambda}\mathcal{E}(u+\lambda\xi)|_{t=0} = \int_\Omega \nabla u \cdot \nabla \xi = -\int_\Omega \Delta u \xi$. Clearly, the linear functional $\mathcal{D}\mathcal{E}(u)\xi = -\int_\Omega \Delta u\xi$ can be identified with the "function" $\mathcal{D}\mathcal{E}(u) = -\Delta u \in L^2(\Omega)$. Similarly, we can compute $\mathcal{D}^2\mathcal{E}(u) = -\Delta$.

Combining examples 1) and 2), we compute the first and second Gâteaux derivatives to get

3) $\mathcal{D}\mathcal{E}(u) = -\Delta u + G'(u) \Rightarrow \mathcal{D}^2\mathcal{E}(u) = -\Delta + G''(u)$;

4) $\mathcal{D}S(\varphi) = -m\frac{\partial^2 \varphi}{\partial t^2} - \nabla V(\varphi) \Rightarrow \mathcal{D}^2S(\varphi) = -m\frac{\partial^2}{\partial t^2} - \text{Hess}V(\varphi)$.

Here, $\text{Hess}V(\varphi) = (\frac{\partial^2 V(\varphi)}{\partial \varphi_i \partial \varphi_j})$ is the standard Hessian of $V : \mathbb{R}^m \to \mathbb{R}$.

In both examples 3) and 4), the second Gâteaux derivatives (or Hessians) are Schrödinger operators. In the second case, the Schrödinger operator is an operator valued $m \times m$ matrix acting on m–vector valued functions.

Exercise. Compute the first and second Gâteaux derivatives in examples 5)–8) at the beginning of this section. (Hint: in example 7), first write the functional $S(\psi)$ in terms of the real vector–functions $\vec{\psi}(x,t) = (\psi_1(x,t), \psi_2(x,t))$ instead of the complex form $\psi(x,t) = \psi_1(x,t) + i\psi_2(x,t)$)

Critical points. Given a C^1–functional $F : M \to \mathbb{R}$, we say that $u_0 \in M$ is a *critical point* (CP) of F iff $\mathcal{D}F(u_0) = 0$ (on $T_{u_0}M$).

Exercise. Find the equations for the critical points in examples 1)–7) given at the beginning of this section.

The equation $\mathcal{D}F(u_0) = 0$ for critical points of F is sometimes called the *Euler* or *Euler–Lagrange equation.*

Theorem. *If u_0 is a minimizer of F and $u_0 \notin \partial M$, then u_0 is a critical point of F.*

Proof. Let u_0 be a minimizer of F, and let ξ be an arbitrary vector from $T_{u_0}M$, and λ sufficiently close to 0 so that there is u_λ s.t. $u_{\lambda=0} = u_0$ and $\frac{du_\lambda}{d\lambda}|_{\lambda=0} = \xi$. Then the function $f(\lambda) := F(u_\lambda)$ has a minimum at $\lambda = 0$, and therefore $\lambda = 0$ is a critical point of this function, $f'(0) = 0$. This is equivalent to $\frac{\partial}{\partial\lambda}F(u_\lambda)|_{\lambda=0} = 0$, which by the definition of the Gâteaux derivative implies that $\mathcal{D}F(u_0)\xi = 0$ (or $\langle \mathcal{D}F(u_0), \xi \rangle = 0$). This holds for every $\xi \in T_{u_0}M$, and we conclude that $\mathcal{D}F(u_0) = 0$. ■

What about the converse statement: is every critical point of F a minimizer (or a maximizer)? As in the calculus of functions of several variables, the answer is negative. Some of the critical points are neither minimizers nor maximizers. They are called *saddle points.* The question then is, how can we classify critical points of F depending on whether they are extrema (minimizers or maximizers) or saddle points? In principle, this can be done as in the case of calculus of functions of several variables: we use the second derivative criterion. A partial result in this direction is given in the following

Theorem. *Let u_0 be a critical point of a functional F. Then*
(i) if u_0 is a minimizer, then $\mathcal{D}^2F(u_0) \geq 0$,
(ii) u_0 is a minimizer if $\mathcal{D}^2F(u_0) > 0$.

Soap films. Let S be a hypersurface in $\mathbb{R}^{n+1}$ (i.e. a n–dimensional surface). Assume S is the graph of a function f:

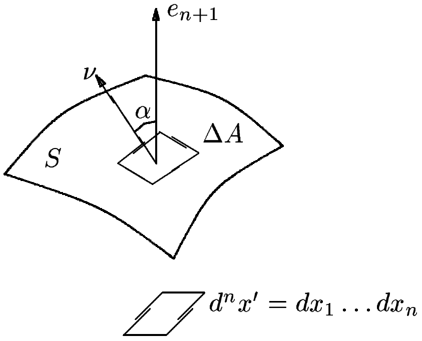
$$x_{n+1} = f(x'), \quad \text{where} \quad x' = (x_1, \dots, x_n),$$

defined on a domain $\Omega \subset \mathbb{R}^n$. Consider the area functional $A(f)$ that measures the area of S.

Lemma. *$A(f)$ can be written as*

$$A(f) = \int_\Omega \sqrt{1 + |\nabla f|^2}\ .$$

Proof. The following picture shows that $\Delta A = \frac{d^n x'}{\cos\alpha}$, where α is the angle between the x_{n+1}-axis (the unit vector e_{n+1}) and the normal ν to S at a given point.



Let $\varphi(x) = x_{n+1} - f(x')$. Then

$$\nu(x) = \frac{\nabla\varphi(x)}{|\nabla\varphi(x)|},$$

and therefore

$$\cos\alpha = \frac{\nabla\varphi(x) \cdot e_{n+1}}{|\nabla\varphi(x)|} = \frac{1}{\sqrt{1 + |\nabla f|^2}}. \blacksquare$$

We define A on $C^2(\Omega)$. Critical points of the functional $A(f)$ are called *minimal surfaces.*

Exercise. Show that the Euler–Lagrange equation for $A(f)$ is

$$\operatorname{div}\left(\frac{\nabla f}{\sqrt{1+|\nabla f|^2}}\right)=0. \tag{26.3}$$

Thus (26.3) yields an equation for a minimal surface. One can show (see below) that

$$h(x):=\operatorname{div}\left(\frac{\nabla f}{\sqrt{1+|\nabla f|^2}}\right)$$

is the mean curvature of S at x, hence we obtained the following

Theorem. *Let S be a smooth hypersurface in $\mathbb{R}^n$ with mean curvature $h(x)$. Then $h=0 \Leftrightarrow S$ is a minimal surface (in a neighbourhood of every $x\in S$; we have $S=graphf$, where f is a critical point of A).*

We give now the definition of different notions of curvatures at a point $x_0\in S$. Pick coordinate systems s.t. $\nabla f(x_0')=0$, where $x_0=(x_0',x_0^{n+1})$. Then we define

- the *principal curvatures at* x_0 as the eigenvalues of $\mathrm{Hess}f(x_0')$,
- the *Gaussian curvature at* x_0 as $\det\mathrm{Hess}f(x_0')$,
- the *mean curvature at* x_0 as $h(x_0)=\operatorname{div}(\frac{\nabla f}{\sqrt{1+|\nabla f|^2}})$.

Lemma. *Let $S=graphf$ for some f s.t. $f(x')\neq 0$. Then the mean curvature at x is given by $h(x)=\operatorname{div}(\frac{\nabla f}{\sqrt{1+|\nabla f|^2}})$.*

This lemma and (26.3) imply the theorem.

Proof. Consider first an arbitrary coordinate system and a function $f:\Omega\to\mathbb{R}$, s.t. $S=\mathrm{graph}f$. We denote as before $x=(x',x_{n+1})\in\mathbb{R}^{n+1}$, $x'=(x_1,\dots,x_n)\in\Omega\subset\mathbb{R}^n$. As we have shown, the unit normal vector to S at x, $\nu(x)$, can be expressed as

$$\nu(x)=\frac{(-\nabla f(x'),1)}{\sqrt{1+|\nabla f(x')|^2}}. \tag{26.4}$$

Now for a given point $x_0 \in S$, let $x = (\overline{x}', \overline{x}_{n+1}) \in \mathbb{R}^{n+1}$ be a special coordinate system s.t. there is a domain $\overline{\Omega} \subset \mathbb{R}^n$ and a function $\overline{f} : \overline{\Omega} \to \mathbb{R}$ s.t. $S = \text{graph}\overline{f}$ and $\nabla \overline{f}(\overline{x}_0') = 0$. Then we can express the normal vector $\nu(x)$ in terms of this function as

$$\nu(x) = \frac{(-\nabla \overline{f}(\overline{x}'), 1)}{\sqrt{1 + |\nabla \overline{f}(\overline{x}')|^2}}.$$

Now compute

$$\operatorname{div} \nu(x) = -\frac{\Delta \overline{f}(\overline{x}')}{(1 + |\nabla \overline{f}(\overline{x}')|^2)^{1/2}} + \frac{|\nabla \overline{f}(\overline{x}')|^2}{(1 + |\nabla \overline{f}(\overline{x}')|^2)^{3/2}},$$

and therefore we get $\operatorname{div} \nu(x_0) = -\Delta \overline{f}(\overline{x}_0')$. By the definition of the mean curvature at the point x_0, $\Delta \overline{f}(\overline{x}_0') = -h(x_0)$, and therefore $\operatorname{div} \nu(x_0) = -h(x_0)$, which together with (26.4) implies the lemma. ■

Instead of defining the surface S as a graph of a function $f : \Omega \to \mathbb{R}$, we can define it as the image of the function $u : \Omega \to \mathbb{R}^{n+1}$ given by $u(x') = (x', f(x'))$, i.e. u is a *parametrization* of Ω. Then we have $|\nabla u| = \sqrt{1 + |\nabla f|^2}$, and so $A(S) = \int_\Omega |\nabla u|$.

Now, instead of minimizing the area integral $\int_\Omega |\nabla u|$, we would like to minimize the energy integral

$$\mathcal{E}(S) = \mathcal{E}(u) = \frac{1}{2} \int_\Omega |\nabla u|^2. \tag{26.5}$$

How are the minimizers of $A(u)$ and $\mathcal{E}(u)$ related? It turns out that they describe the same surface S.

One can further generalize functional (26.5) as

$$\mathcal{E}(u) = \frac{1}{2} \int_\Omega \sum_{i,j} g_{ij}(u) \nabla u^i \cdot \nabla u^j,$$

where $u = (u^1, \ldots, u^m) : \Omega \to \mathbb{R}^m$, and the matrix $g(u) = (g_{ij}(u))$ is positive definite for all u. Moreover, we assume $g(u)$ satisfies $g(u) \geq \delta \mathbb{1}$, with $\delta > 0$. The matrix $g(u)$ yields a Riemannian metric on the target space $\mathbb{R}^m$ and $\sum_{i,j} g_{ij}(u) \nabla u^i \nabla u^j$ is the square of the length of u in this metric. The Euclidean metric used above is given by $g_{ij}(u) = \frac{1}{n}\delta_{i,j}$ if $(i,j) \neq (n+1, n+1)$, and $g_{n+1,n+1} = 1$.

Let $\Omega \subset \mathbb{R}^m$. Then $u : \Omega \to \mathbb{R}^m$ represents a parametrization of the surface $S = \operatorname{range} u \subset \mathbb{R}^m$. The boundary of this surface is $g(\partial\Omega) \equiv \operatorname{range} g$ on $\partial\Omega$. Here, $\mathcal{E}(u)$ defines the "energy" of the surface. The area of the surface is given by

$$A(S) = A(u) = \int_\Omega \sqrt{\sum_{i,j} g_{ij}(u) \nabla u^i \cdot \nabla u^j}.$$

27. Constraints and Lagrange multipliers

Consider a functional $\mathcal{E}$ on a Banach space X. We want to minimize $\mathcal{E}$ not on the entire space X, but rather on a subset M of X defined as

$$M = \{u \in X : J(u) = 0\},$$

where J is another functional on X. The key result here goes back to Lagrange and it is called the method of *Lagrange multipliers*:

Theorem. *Let the functionals $\mathcal{E}$ and J be C^1, and let u_0 be a critical point of $\mathcal{E}$ on M. Then there is a $\lambda \in \mathbb{R}$ (the Lagrange multiplier) s.t. $\mathcal{D}\mathcal{E}(u_0) - \lambda \mathcal{D}J(u_0) = 0$.*

Proof. The fact that u_0 is a critical point of $\mathcal{E}$ in M means that

$$\mathcal{D}\mathcal{E}(u_0)\xi (= \langle \mathcal{D}\mathcal{E}(u_0), \xi \rangle) = 0, \tag{27.1}$$

$\forall \xi \in T_{u_0}M$. Recall the definition of the tangent space

$$T_{u_0}M = \{\xi \in X : \exists \text{ path } u_s \subset M \text{ s.t. } u_{s=0} = u_0, \frac{du_s}{ds}|_{s=0} = \xi\}.$$

Taking u_s as in this definition and differentiating the constraint $J(u_s) = 0$ at $s = 0$, we obtain

$$0 = \mathcal{D}J(u_0)\dot{u}_0 \quad (= \langle \mathcal{D}J(u_0), \dot{u}_0 \rangle),$$

where $\dot{u}_0 = \frac{du_s}{ds}|_{s=0}$. Hence we have $T_{u_0}M \subseteq \mathcal{D}J(u_0)^{\perp}$. In fact, one can show that $T_{u_0}M = \mathcal{D}J(u_0)^{\perp}$, and therefore we can write (27.1) as

$$\langle \mathcal{D}\mathcal{E}(u_0), \xi\rangle = 0, \quad \forall \xi \in \mathcal{D}J(u_0)^{\perp},$$

which is equivalent to saying that there is a $\lambda \in \mathbb{R}$ s.t.

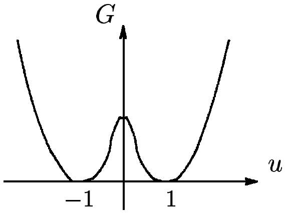
$$\mathcal{D}\mathcal{E}(u_0) - \lambda \mathcal{D}J(u_0) = 0.$$

(i.e. $\mathcal{D}\mathcal{E}(u_0) \perp (\mathcal{D}J(u_0))^{\perp}$ implies $\mathcal{D}\mathcal{E}(u_0)||\mathcal{D}J(u_0)$). ■

Examples. 1) Consider the *Ginzburg–Landau functional*

$$\mathcal{E}(u) = \int \left(\frac{1}{2}|\nabla u|^2 + G(u)\right) d^n x,$$

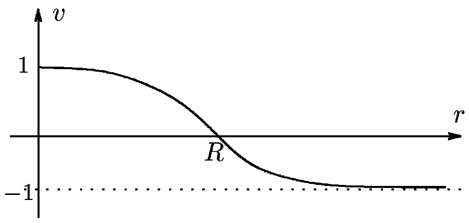
with an even double–well potential G of the following form:



on spherically symmetric functions $u(x) = v(|x|)$. Then we have $\mathcal{E}(u) = \sigma_{n-1}e(v)$, where σ_n is the volume of the unit n–sphere and

$$e(v) = \int_0^{\infty} \left(\frac{1}{2}|v'|^2 + G(v)\right) r^{n-1} dr.$$

Since G has zeroes at ±1, for $e(v)$ to be finite, we must require that $v \to +1$ or -1 as $r \to \infty$. Consider e on functions of the form



i.e. $v(0) = 1$, $v(\infty) = -1$ and $v(R) = 0$. We set $X = H_1^{(g)}(\mathbb{R}^+, r^{n-1}dr)$, where g is a fixed smooth function of r satisfying $g(0) = 0$ and $g(\infty) = -1$, and consider $e(v)$ on X with the constraints

$$v(R) = 0 \quad \text{or equivalently} \quad \int v\delta_R = 0, \tag{27.2}$$

where δ_R is the Dirac distribution concentrated at R. Then a minimizer of this problem satisfies the equation

$$-v'' + G'(v) = \lambda\delta_R,$$

for some λ determined by condition (27.2).

Consider the minimization problem above in a very large ball $B_L := \{x \in \mathbb{R}^n : |x| \leq L\}$, $L >> 1$. Then there is a minimizer $v_R^{(L)}$. One can show that $v_R^{(L)} \to v_R$, where v_R is a minimizer in $\mathbb{R}^n$.

2) Let Ω be a domain in $\mathbb{R}^n$. Consider the Dirichlet functional

$$\mathcal{D}(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 d^n x$$

on the set $H_1^{(g)}(\Omega)$ (or on the set $M = \{u \in H^{(g)}(\Omega) : J(u) = 1\}$, where $J(u) = \frac{1}{p}\int_\Omega |u|^p d^n x$). Though the set M is not a vector space, it is obtained from the vector space $H_1^{(0)}(\Omega)$ through a shift. The equation for the critical point on $H_1^{(g)}(\Omega)$ is

$$\begin{aligned} \Delta u &= 0 \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega. \end{aligned}$$

To find the equation of the critical points on the space M, we use the theorem above to obtain

$$\mathcal{D}\mathcal{E}(u) - \lambda\mathcal{D}J(u) = 0,$$

for some λ determined by the side condition $J(u) = 1$. Since $\mathcal{D}\mathcal{E}(u) = -\Delta u$, and $\mathcal{D}J(u) = |u|^{p-2}u$, we obtain

$$\Delta u + \lambda|u|^{p-2}u = 0,$$

with the constraint $\frac{1}{p}\int_\Omega |u|^p = 1$.

Exercise. Find the equation for critical points of the functional $\frac{1}{2}\int_\Omega |\nabla u|^2$ on the space $M = \{u \in H_1^{(0)} : \int_\Omega |u|^2 = 1\}$.

28. Theory of interfaces

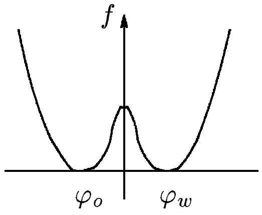
The difference in free energy of two phases is given by the functional:

$$F(\varphi) = \int_\Omega \left(\frac{1}{2}|\nabla\varphi|^2 + \lambda f(\varphi)\right) d^d x, \tag{28.1}$$

where Ω is a large domain, say a ball of radius L filled with two liquids, e.g. oil and water. Here, $\lambda \in \mathbb{R}$ is a parameter. The function $\varphi : \Omega \to \mathbb{R}$ is the difference of the local densities of oil and water:

$$\varphi(x) := n_o(x)\rho_o - n_w(x)\rho_w,$$

where ρ_o and ρ_w are the densities of oil and water, and $n_o(x)$ and $n_w(x)$ are concentrations of oil and water at the point x, respectively (i.e. $n_o(x) + n_w(x) = 1\,\forall x$). We take a potential f of the form



where $\varphi_w = \rho_w$ (i.e. $n_o = 0$, homogeneous water phase), and $\varphi_o = \rho_o$ (i.e. $n_w = 0$, homogeneous oil phase). Hence $F(\varphi) \geq 0$, and F has two absolute minimizers, φ_w and φ_o, so that $F(\varphi_o) = F(\varphi_w) = 0$.

The question is: does $F(\varphi)$ have local minimizers (or saddle points) describing two coexisting phases separated by an interphase? If yes, what does this interface look like?

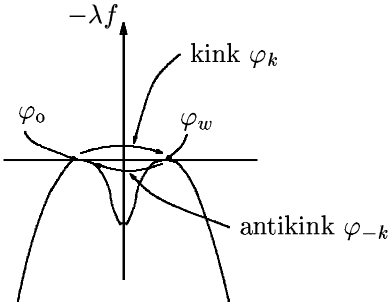
An important tool in our study is the equation for critical points of

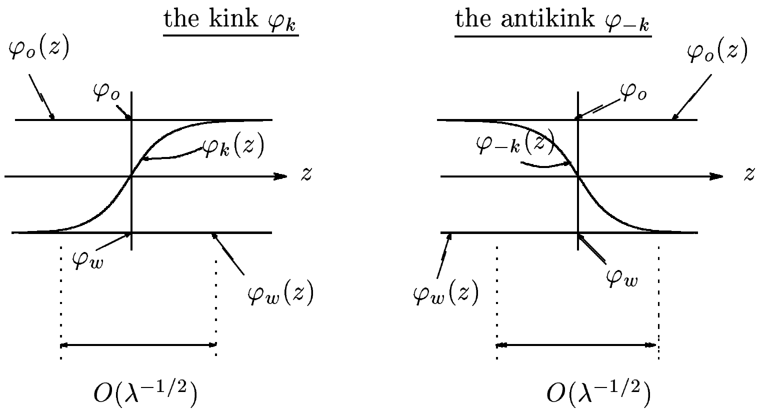
$F(\varphi)$:

$$-\Delta\varphi + \lambda f'(\varphi) = 0. \tag{28.2}$$

In the search of solutions to (28.2), we investigate several possibilities: planar interface, lamellar phase, spherical drops and cylinder solutions.

Planar interface. Let $\Omega = \mathbb{R}^d$. Assume φ depends only on one coordinate, say $z = x_d$. then equation (28.2) becomes simply $\varphi'' = \lambda f'(\varphi)$. If we think about z as a time variable, then this is just Newton's equation with the potential $-\lambda f(\varphi)$. Apart from the constant solutions φ_o and φ_w, there are two other solutions for $-\infty < z < \infty$ with smallest $F(\varphi)$, called the kink φ_k and antikink φ_{-k}. They are depicted in the following pictures:





Note that for large “times” z, any function φ of finite energy $F(\varphi) < \infty$, must tend to a zero of the potential f, otherwise the energy (28.1) becomes infinite.

One can show that the solutions $\varphi_o, \varphi_w, \varphi_k$ and φ_{-k} minimize the energy $F(\varphi)$ under the boundary conditions $\varphi(z) \to \varphi_o$ as $z \to \pm\infty$, $\varphi(z) \to \varphi_w$ as $z \to \pm\infty$, $\varphi(z) \to \varphi_w$ as $z \to -\infty$ and $\varphi(z) \to \varphi_o$ as $z \to +\infty$, or $\varphi(z) \to \varphi_w$ as $z \to -\infty$ and $\varphi(z) \to \varphi_o$ as $z \to +\infty$.

The solution $\varphi_{k/-k}$ describes a separation of oil and water with the interface $z = 0$. The centre of the kink is arbitrary, since $F(\varphi)$ is translational invariant. Therefore, $\varphi_k(a - z)$ and $\varphi_{-k}(z - a)$, for arbitrary a, are also kink and antikink solutions.

Specific examples. **a)** Let $f(\varphi) = \frac{1}{2}(\varphi^2 - 1)^2$, then one can explicitely calculate the kink: $\varphi_k(z) = \tanh(\lambda z)$. **b)** Let

$$f(\varphi) = \begin{cases} \omega_o(\varphi - \varphi_o)^2, & \varphi \geq 0, \\ \omega_w(\varphi - \varphi_w)^2, & \varphi \leq 0, \end{cases} \quad \text{and} \quad \omega_o\varphi_o^2 = \omega_w\varphi_w^2. \tag{28.3}$$

Then equation (28.2) is piecewise linear, and we get

$$\varphi_k(z) = \begin{cases} \varphi_w(1 - e^{-\sqrt{\omega_w}z}), & z \leq 0, \\ \varphi_o(1 - e^{\sqrt{\omega_o}z}), & z \geq 0. \end{cases} \tag{28.4}$$

Lamellar phase. Let again $\Omega = \mathbb{R}^d$. In this situation, layers of oil and water coexist in a periodic array. To get a solution to (27.2), glue together a kink at z_1 and an antikink at z_2. There is no exact solution of this form: the kink and the antikink interact at any distance. They repel each other and as a result, they move away from each other. This means that $F(\varphi)$ is monotonically decreasing as $R \to \infty$. Here, φ is a function consisting of a kink and an antikink glued together at a distance R. However, one can construct a periodic solution corresponding to an array of kinks and antikinks.

Spherical drops and cylinder solutions. In both cases, we look for minimizers of $F(\varphi)$ of the form

$$\varphi(x) = \varphi_R(|x|), \tag{28.5}$$

subject to the boundary condition

$$\varphi_R(0) = \varphi_o, \; \varphi_R(\infty) = \varphi_w \tag{28.6}$$

and the side condition

$$\varphi_R(R) = 0. \tag{28.7}$$

For the spherical drop, we have $x \in \mathbb{R}^3$, and for the cylinder we have $x \in \mathbb{R}^2$.

One can show that such minimizers exist, for every R. Moreover, the theory of Lagrange multipliers implies that they satisfy the equation

$$-\Delta_r \varphi_R + \lambda f'(\varphi_R) = \nu \delta_R, \tag{28.8}$$

where $\nu = \nu(R)$ is the Lagrange multiplier, δ_R is the Dirac distribution concentrated at R, $\delta_R(x) = \delta(|x| - R)$, and

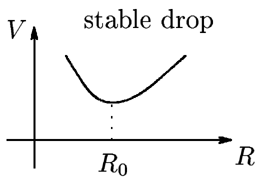
$$\Delta_r = \frac{\partial^2}{\partial r^2} + \frac{d-1}{r}\frac{\partial}{\partial r}$$

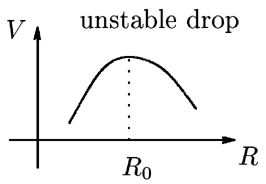
is the radial part of the Laplacian. Here, $d = 2$ for the cylindrical case, and $d = 3$ for the drop. We plug this minimizer into $F(\varphi)$ to define the potential $V(R) := F(\varphi_R)$.

Theorem. *If R_0 is a critical point of $V(R)$, then φ_{R_0} is a critical point of $F(\varphi)$: $\mathcal{D}F(\varphi_{R_0}) = 0$. For such an R_0, φ_{R_0} satisfies the original equation (28.2), i.e.*

$$\varphi''_{R_0} + \frac{d-1}{r}\varphi'_{R_0} = \lambda f'(\varphi_{R_0}). \tag{28.9}$$

We have the following possibilities:





If we think about r as a time variable, then equation (28.9) describes a mechanical particle in the potential $-\lambda f(\varphi)$, subject to friction.

The equilibrium radius can be found also through the equation $\nu(R_0) = 0$, where $\nu = \nu(R)$ is given in (28.8). If we take again for $f(\varphi)$ the quadratic potential (28.3), then φ_R can be found explicitely:

$$\varphi_R(r) = \begin{cases} \varphi_o(1 - \frac{R}{\sinh(\sqrt{\omega_o}R)}\frac{\sinh(\sqrt{\omega_o}r)}{r}), & 0 \leq r \leq R \\ \varphi_w(1 - \frac{R}{e^{-\sqrt{\omega_w}R}}\frac{e^{-\sqrt{\omega_w}r}}{r}), & r \geq R. \end{cases} \tag{28.10}$$

This φ_R satisfies equation (28.8) with $\lambda = 1/2$ and

$$\nu = \nu(R) = \varphi_0\left(\sqrt{\omega_o}\coth\sqrt{\omega_o}R - R^{-1}\right) - \varphi_w\left(\sqrt{\omega_w} + R^{-1}\right).$$

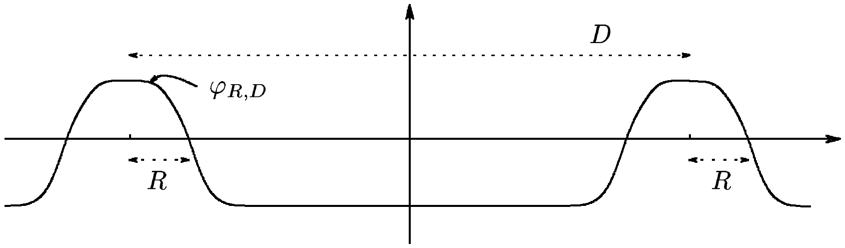
The condition on the continuity of $\varphi_R'(r)$ at $r = R$ is $\nu_R(R) = 0$, and the solution of this equation gives the equilibrium radius of the drop (or cylinder).

Exercises. 1) Show (28.10). Hint: for $0 \leq r \leq R$, assume $\varphi \geq 0$, and solve $-\Delta_r\varphi + f'(\varphi) = 0$ with boundary conditions $\varphi(0) = \varphi_o$ and $\varphi(R) = 0$. Then for $R \leq r < \infty$, assume $\varphi(R) \leq 0$ and solve $-\Delta_r\varphi + f'(\varphi) = 0$ with boundary conditions $\varphi(R) = 0$ and $\varphi(\infty) = \varphi_w$. **2)** Compute the left and right derivatives of φ_R at R. Notice that in general the first derivative has a jump at R, so taking the second derivative gives a Dirac distribution, hence φ_R satisfies

$$-\Delta_r\varphi_R + f'(\varphi_R) = \omega(R)\delta_R,$$

where $\omega(R) = \varphi_R'(R_-) - \varphi_R'(R_+)$. Then $\omega(R_0) = 0 \Leftrightarrow \varphi_{R_0}'(R_{0-}) = \varphi_{R_0}'(R_{0+})$, and φ_{R_0} solves $-\Delta\varphi + f'(\varphi) = 0$ on $0 \leq r < \infty$.

Interaction between spherical drops or cylinders. We glue two drops at a large distance D together to get the function $\varphi_{R,D}$:



Here, R is the stable equilibrium radius. The interaction between drops is given by

$$V_{\mathrm{drop},R}(D) = F(\varphi_{R,D}) - 2F(\varphi_R),$$

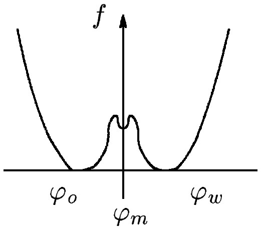
i.e. by the energy of the solution close to the two-drop solution for fixed drops minus the energy of two noninteracting drops.

A more systematic way of defining $V_{\mathrm{drop},R}(D)$ is by setting it equal to $\inf\{F(\varphi) : \varphi$ describes two R–drops fixed at a distance D from each other$\} - 2F(\varphi_R)$.

More refined models. In order to capture better effects of curvature, or to accomodate several homogeneous phases, or both, one makes the following refinements of (28.1):

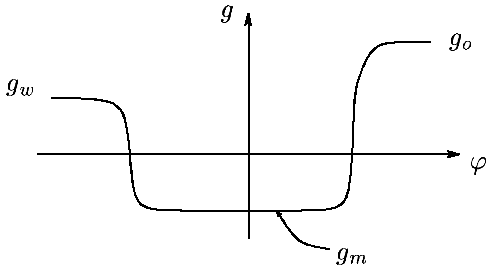
(i) $|\nabla\varphi|^2 \to c|\Delta\varphi|^2 + g(\varphi)|\nabla\varphi|^2$, where $g(\varphi)$ is allowed to assume negative values on some bounded interval,

(ii) $f(\varphi)$ is assumed to have several minima:



Here, the third minimizer in the middle corresponds to the microemulsion state of the oil and water mixture.

(iii) A combination of (i) and (ii) is the *Gompper and Schick model.* In this model, the function g is taken of the form



An example of values of the parameters are $c = 1$, $g_w = g_o = 4.6$, and $g_m = -4.5$.

For the planar interface, the Gompper–Schick model has the "conservation law" (in x):

$$2c\left[\overline{\varphi}'\overline{\varphi}'' - \frac{1}{2}(\overline{\varphi}'')^2\right] - g(\overline{\varphi})(\overline{\varphi}')^2 + \lambda f(\overline{\varphi}) = 0.$$

Here, $\overline{\varphi}$ is a minimizer of the Gompper-Schick model representing a planar interface (i.e. depending only on one variable).

To compute $V(R) := F(\varphi_R)$ in the spherical and cylindrical case, Gompper et al. do an expansion in $1/R$:

$$\varphi_R(r) = \overline{\varphi}(R-r) + \frac{1}{R}\varphi_1(r-R) + \frac{1}{R^2}\varphi_2(r-R) + \cdots .$$

A computation of $V(R)$ to the order $\mathcal{O}(1/R)$ yields

$$V_{\text{sphere}}(R) = \sigma R^2 + \lambda R + \mu,$$

where, for $p_s(z) = 2g(\overline{\varphi})(\overline{\varphi}')^2 + 4c(\overline{\varphi}'')^2$, we have

$$\begin{aligned}\sigma &= \int_{-\infty}^{\infty} p_s(z)dz,\\ \lambda &= 2\int_{-\infty}^{\infty} p_s(z)zdz,\\ \mu &= 2\int_{-\infty}^{\infty} p_s(z)z^2dz.\end{aligned}$$

The equilibrium radius (given by $V'_{\text{sphere}}(R_{\text{equil,sphere}}) = 0$) is thus

$$R_{\text{equil,sphere}} = -\frac{\lambda}{2\sigma}.$$

Exercise. Compute $V_{\text{sphere}}(R)$ for the the piecewise parabolic model (28.3).

Energy of fluctuations. Our goal is to compute the energy of fluctuations near a planar interface. We go back to the free energy functional

$$F(\varphi) = \int \left[\frac{1}{2}|\nabla\varphi|^2 + \lambda f(\varphi)\right].$$

The planar interface is described by the kink solutions $\overline{\varphi} = \varphi_k$. For $f(\varphi) = \frac{1}{2}(\varphi^2 - 1)^2$, $\varphi_k(z) = \tanh(\lambda z)$ as already discussed.

Consider now a fluctuation of the interface around its equilibrium position at $z = 0$. In other words, we look for functions φ whose energy is very close to $\inf F(\varphi)$ and whose zero level set $S := \{x : \varphi(x) = 0\}$ is close to the equilibrium interface $\{z = 0\}$. One expects that up to very tiny perturbations, such functions are of the form $\varphi(x) = \varphi_k(u(x))$, where $u(x) = 0$ on S. For S very close to the plane $\{z = 0\}$, we can write it as a graph of some function $h : \mathbb{R}^{d-1} \to \mathbb{R}$, $S = \operatorname{graph} h = \{(x', h(x')) : x' \in \mathbb{R}^{d-1}\}$. In this case, the simplest $u(x)$ we can take is $u(x) = z - h(x')$. To take into account the second order term, we choose

$$u(x) = \frac{z - h(x')}{\sqrt{1 + |\nabla h(x')|^2}} = z - h(x') - \frac{1}{2}(z - h(x'))|\nabla h(x')|^2 + \cdots.$$

Plug $\varphi(x) = \varphi_k(u(x))$ into the expression for $F(\varphi)$, and compute

$$F(\varphi) = \sigma A(h)\left(1 + \mathcal{O}(\nabla^2 h)\right),$$

where

$$A(h) = \int \sqrt{1 + |\nabla h|^2}\, dx'$$

is the area of the surface graph h, and

$$\sigma := \int |\varphi_k'|^2 dz = \int \sqrt{2} f(\varphi) dz.$$

Thus the energy of the fluctuation φ_{fluct} is proportional to the surface area of the interface S.

29. Minimization: direct methods

The problem we address is the following: given a functional $\mathcal{E}$ on a space M, find a function $u_0 \in M$ (if such exists) that minimizes $\mathcal{E}$:

$$\mathcal{E}(u_0) = \inf_{u \in M} \mathcal{E}(u).$$

Such a function u_0 is called a *minimizer for* $\mathcal{E}$. Thus to begin with, we want to assume that $\mathcal{E}$ is *bounded below*, i.e.

$$E_0 := \inf_{u \in M} \mathcal{E}(u) > -\infty.$$

Let us first analyze a simple but typical finite dimensional situation: $M = \mathbb{R}^N$. How would we minimize a functional on M? We do this in three steps (that will be suitable to be generalized to the infinite dimensional case):

- *Step 1.* Since $\mathcal{E}$ is bounded on M from below by say $E_0 > -\infty$, we can pick a sequence $\{u_n\} \subset M$ s.t. $\mathcal{E}(u_n) \to E_0$, as $n \to \infty$. Such a sequence is called a *minimizing sequence.*

- *Step 2.* We hope that either such a sequence converges or at least contains a convergent subsequence, which for notational convenience we denote again by $\{u_n\}$. The limit of such a subsequence clearly is a candidate for a minimizer if the latter exists. How do we show that $\{u_n\}$ has a convergent subsequence? Assume that

$$\mathcal{E}(u) \to \infty \quad \text{as} \quad ||u|| \to \infty. \tag{29.1}$$

Clearly we can take the sequence $\{u_n\}$ s.t. every element u_n satisfies $\mathcal{E}(u_n) \leq E_0 + 1$ (just throw out those u_n for which $\mathcal{E}(u_n) > E_0 + 1$). Due to (29.1), we have that $||u_n|| \leq C$, $\forall n$ and for some $C > 0$. Hence by the Bolzano–Weierstrass theorem, $\{u_n\}$ has a convergent subsequence.

- *Step 3.* Let $u_0 := \lim_{n\to\infty} u_n$. If $\mathcal{E}$ is *continuous*, then

$$\lim_{n\to\infty} \mathcal{E}(u_n) = \mathcal{E}(u_0).$$

Since on the other hand we have $\lim_{n\to\infty} \mathcal{E}(u_n) = E_0$, we conclude that $\mathcal{E}(u_0) = E_0$, i.e. u_0 is a minimizer of $\mathcal{E}$.

Let us look closer at the last two steps. We assume $M \subset X$, where X is a Hilbert space. A functional $\mathcal{E}$ on M is called *coercive* iff $\mathcal{E}(u) \to \infty$ whenever $||u||_M \to \infty$. A set K s.t. every infinite sequence of elements of K contains a convergent subsequence is called *compact.*

The Bolzano–Weierstrass theorem states that a closed ball in $\mathbb{R}^N$ is compact. This property does not hold in general in the infinite dimensional case. For instance, closed balls in $L^2(\Omega)$ are not compact. As a concrete example, take for instance $\Omega = \mathbb{R}^n$, and $L^2(\Omega) \ni u_n(x) := u(x-n)$, for some $u \in L^2(\Omega)$, $||u||_2 = 1$. Clearly, this sequence does not have a convergent subsequence.

We have however the following weaker result (which follows from Alaoglu's theorem): If X is a (reflexive) Banach space, then every uniformly bounded sequence $\{u_n\}$ in X (i.e. $||u_n|| \leq C$) has a weakly convergent subsequence $\{u_{n_k}\}$, i.e. $\exists u_0$ s.t. $u_{n_k} \xrightarrow{w} u_0$ in X (weak convergence is denoted by $\xrightarrow{w}$). Notice that the finite dimensional spaces, weak convergence is equivalent to strong convergence, and the above statement reduces to the Bolzano-Weierstrass theorem.

Next, the continuity is a property hard to come by for functionals on infinite dimensional spaces. But there is a weaker property which often holds: $\mathcal{E}$ is called *weakly lower semicontinuous (w.l.s.c.)* iff $u_n \xrightarrow{w} u_0$ in M implies $\liminf_{n\to\infty} \mathcal{E}(u_n) \geq \mathcal{E}(u_0)$. We now check that in fact w.l.s.c. suffices to carry out step 3 above. If $u_n \xrightarrow{w} u_0$, then $\liminf_{n\to\infty} \mathcal{E}(u_n) \geq \mathcal{E}(u_0)$. On the other hand, by the definition of $\{u_n\}$, we have $\liminf_{n\to\infty} \mathcal{E}(u_n) = E_0 = \inf_{u\in M} \mathcal{E}(u)$. Therefore $\mathcal{E}(u_0) = E_0$, and hence u_0 is a minimizer of $\mathcal{E}$.

As an example, take $\mathcal{E} = \int_\Omega (\frac{1}{2}|\nabla u|^2 + G(x,u))$.

Proposition. *Let $G(x,u) \geq c|u|^2$, for some $c \geq 0$.* **(a)** *If $c > 0$,* *then $\mathcal{E}$ is coercive on $H_1(\Omega)$; if $c = 0$ and Ω is bounded in one direction, then $\mathcal{E}$ is coercive on $H_1^{(0)}(\Omega)$.* **(b)** *If $c = 0$, i.e. $G(x,u) \geq 0$, then $\mathcal{E}$ is w.l.s.c..*

Proof. **(a)** By the assumption on G, we have $\mathcal{E}(u) \geq \int_\Omega (\frac{1}{2}|\nabla u|^2 + c|u|^2)$. If $c > 0$, then $\mathcal{E}(u) \geq \delta ||u||_{(1)}^2$, where $\delta = \min(1/2, c) > 0$, and therefore $\mathcal{E}$ is coercive on $H_1(\Omega)$, i.e. $\mathcal{E}(u) \to \infty$ whenever $\|u\|_{H_1(\Omega)} \to \infty$.

If $c = 0$, and Ω as specified, then by the Poincaré inequality (c.f. (17.6)), we have $\int_\Omega |u|^2 \leq D^2 \int_\Omega |\nabla u|^2$, for any $u \in H_1^{(0)}(\Omega)$, where D is the smallest diameter of Ω. So we get $\mathcal{E}(u) \geq \frac{1}{4}\min(1, D^{-2})||u||_{(1)}^2$ for every $u \in H_1^{(0)}(\Omega)$. Therefore $\mathcal{E}$ is coercive on $H_1^{(0)}(\Omega)$.

(b) We have

$$\begin{aligned}
\frac{1}{2}\int_\Omega |\nabla u_n|^2 &= \frac{1}{2}\int_\Omega |\nabla(u + u_n - u)|^2 \\
&= \frac{1}{2}\int_\Omega |\nabla u|^2 + \mathrm{Re}\int_\Omega \nabla\overline{u}\cdot\nabla(u_n - u) + \frac{1}{2}\int_\Omega |\nabla(u_n - u)|^2 \\
&\geq \frac{1}{2}\int_\Omega |\nabla u|^2 + \mathrm{Re}\int_\Omega \nabla u\cdot\nabla(u_n - u).
\end{aligned}$$

If $u_n \xrightarrow{w} u$ in $H_1(\Omega)$, then $\int_\Omega \nabla u \cdot \nabla(u_n - u) \to 0$, and therefore

$$\liminf_{n\to\infty} \frac{1}{2}\int_\Omega |\nabla u_n|^2 \geq \frac{1}{2}\int_\Omega |\nabla u|^2.$$

Now we need to show that the potential energy part in $\mathcal{E}$ is also w.l.s.c. Using properties of Sobolev spaces, one can show that (up to a subsequence), if $u_n \xrightarrow{w} u$ in $H_1(\Omega)$, then $u_n \to u$ *a.e.*. Therefore, by Fatou's lemma, we get $\liminf_{n\to\infty}\int_\Omega G(x, u_n) \geq \int_\Omega G(x,u)$, and $\mathcal{E}$ is w.l.s.c. on $H_1(\Omega)$. ■

Exercises. 1) Let $V(x) \geq 0$. Show that if $u_n \xrightarrow{w} u$ in $L^2(\Omega)$, then $\liminf_{n\to\infty}\int V(x)|u_n|^2 \geq \int V(x)|u|^2$.

2) Let Ω be a bounded domain in $\mathbb{R}^n$, and let $f \in L^2(\Omega)$. Show that the functional $\frac{1}{2}\int_\Omega |\nabla u|^2 d^n x - \int_\Omega fu$ is coercive and w.l.s.c. on $H_1^{(g)}(\Omega)$.

3) Let Ω be a bounded domain in $\mathbb{R}^n$, and let $g(u) = (g_{ij}(u))$ be a family of $m \times m$ positive definite matrices satisfying $g(u) \geq \delta \mathbb{1}$, for some $\delta > 0$. Show that the functional $\mathcal{E}(u) = \frac{1}{2}\int_\Omega \sum_{i,j} g_{ij}(u) \nabla u^i \cdot \nabla u^j d^n x$ is coercive and w.l.s.c. on $H_1^{(g)}(\Omega, \mathbb{R}^m)$. Here, $u = (u_1, \dots, u_m) : \Omega \to \mathbb{R}^m$.

30. A key result about existence of minimizers

In this section, we prove a key result about the existence of minimizers. Let X be a reflexive Banach space (for simplicity, think of X as a Hilbert space), and let M be a subset of X. We say that M is *weakly closed in X* iff $u_n \xrightarrow{w} u_0$ in X and $u_n \in M$, imply $u_0 \in M$. We consider a functional $\mathcal{E} : M \to \mathbb{R}$. We have the following

Key Theorem. *Assume M is weakly closed in X and $\mathcal{E}$ is a coercive and w.l.s.c. functional on X. Then $\mathcal{E}$ is bounded below and attains its minimum in M (i.e. there is a minimizer of $\mathcal{E}$ on M).*

Proof. Let $E_0 := \inf_{u \in M} \mathcal{E}(u)$, and let u_n be a minimizing sequence for $\mathcal{E}$, i.e.

$$\lim_{n \to \infty} \mathcal{E}(u_n) = E_0. \tag{30.1}$$

Clearly, we can assume that $\mathcal{E}(u_n) \leq E_0 + 1$ (we get rid of those u_n's in the minimizing sequence for which $\mathcal{E}(u_n) > E_0 + 1$). Then by the coercivity of $\mathcal{E}$, there is a constant K s.t. $||u_n|| \leq K$, $\forall n$. Hence, by Alaoglu's theorem, $\{u_n\}$ contains a weakly convergent subsequence, which we denote again by $\{u_n\}$, $u_n \xrightarrow{w} u_0 \in X$. The element u_0 is a candidate for a minimizer. Since M is weakly closed, $u_0 \in M$. Finally, w.l.s.c. of $\mathcal{E}$ gives $\liminf_{n \to \infty} \mathcal{E}(u_n) \geq \mathcal{E}(u_0)$. This together with equation (30.1) implies that $E_0 \geq \mathcal{E}(u_0)$. On the other hand, $E_0 = \inf_{u \in M} \mathcal{E}(u) \leq \mathcal{E}(u_0)$, and therefore $\mathcal{E}(u_0) = E_0$. This shows that u_0 is a minimizer and that $\inf_{u \in M} \mathcal{E}(u) > -\infty$. ■

This is a simple but powerful result. It says that in order to show that $\mathcal{E}$ has a minimizer, we have to check three conditions:

(α) M is weakly closed,

(β) $\mathcal{E}$ is w.l.s.c.,

(γ) $\mathcal{E}$ is coercive.

We analyzed already conditions (β) and (γ) in the previous section. Now we give examples of weakly closed sets, besides of the obvious ones, $M = X$ or $M = g + X$ (for a fixed g).

(a) Let $I = [-1, 1]$, and let $M \subset H_1(I)$ be given by $M = \{u \in H_1(I) : u(0) = 0\}$.

Proposition. it M is weakly closed in $H_1(I)$.

Proof. By the Rellich–Kondrashov theorem, for Ω bounded, if $u_n \xrightarrow{w} u_0$ in $H_1(\Omega)$, then there is a subsequence $\{u_{n'}\}$ s.t. $u_{n'} \to u_0$ in $C(\Omega)$. If $u_n \in M$, then $u_n(0) = 0$, and therefore $u_0(0) = \lim_{n' \to 0} = 0$. ■

(b) Let Ω be a bounded, smooth domain in $\mathbb{R}^n$. Consider $M \subset H_1^{(0)}(\Omega)$ given by $M = \{u \in H_1^{(0)}(\Omega) : \int_\Omega |u|^p d^n x = 1\}$.

Proposition. *If* $n \geq 3$ *and* $p < \frac{2n}{n-2}$*, then* M *is weakly closed in* $H_1^{(0)}(\Omega)$.

Proof. By the Rellich-Kondrashov theorem, $H_1^{(0)}(\Omega)$ is compactly embedded into $L^p(\Omega)$, for $p < \frac{2n}{n-2}$, and Ω bounded. This means that any weakly convergent sequence $u_n \xrightarrow{w} u_0$ in $H_1^{(0)}(\Omega)$ contains a subsequence $\{u_{n'}\}$ s.t. $u_{n'} \to u_0$ in $L^p(\Omega)$. Hence $||u_0||_p = \lim_{n\to\infty} ||u_n||_p = 1$, and therefore $u_0 \in M$. ■

The key theorem implies that the following functionals have minimizers on the specified sets:

1. $\frac{1}{2}\int_\Omega |\nabla u|^2$ on $H_1^{(g)}(\Omega)$, provided Ω is bounded in one direction,

2. $\frac{1}{2}\int_\Omega \sum_{i,j} g_{ij}(u)\nabla u^i \cdot \nabla u^j$ on $H_1^{(g)}(\Omega, \mathbb{R}^m)$, provided Ω is bounded in one direction, and $g(u) = (g_{ij}(u)) \geq \delta \mathbb{1}$, for some $\delta > 0$,

3. $\int_\Omega (\frac{1}{2}|\nabla u|^2 + G(x,u))$ on $H_1^{(g)}(\Omega)$, if $G(x,u) \geq 0$ and Ω is bounded in one direction.

We have the following special cases for example 3.:

i) $G(x,u) = g(u) \geq 0$. If G has a minimum at u_0, then $u_0(x) \equiv u_0$ is a minimizer: $\mathcal{E}(u_0) = 0$,

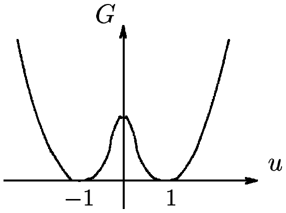
ii) $G(x,u) = V(x)|u|^2$ for some $V(x) \geq 0$, i.e. G is quadratic (remember that in this case the equation for the critical points of $\mathcal{E}$ is *linear*!). We can write $\mathcal{E}(u)$ on $H_2^{(0)}(\Omega)$ as $\mathcal{E}(u) = \int_\Omega \overline{u}(-\frac{1}{2}\Delta + V(x))u$, i.e. $\mathcal{E}$ is the quadratic form of the Schrödinger operator $-\frac{1}{2}\Delta + V(x)$.

We consider now some specific functionals of interest.

α) The Ginzburg–Landau energy functional

$$\mathcal{E}(u) = \int_\Omega \left(\frac{1}{2}|\nabla u|^2 + \lambda G(u)\right) d^n x,$$

where $u : \Omega \to \mathbb{R}$, $\lambda > 0$, and G is of the form of a double-well potential:



We formalize this picture by requiring that $G(u) \geq 0$, and G has strict (in the transverse direction) minima at $|u| = 1$, and $G(u) \to \infty$ as $|u| \to \infty$. The typical and most important example is $G = \frac{1}{4}(|u|^2 - 1)^2$.

We want to consider the case $\Omega = \mathbb{R}^n$, but first we assume that $\Omega = [-L, L]^n$ with L very large (i.e. we take Ω to be a very large box, almost $\mathbb{R}^n$), and we want to find minimizers and saddle points of $\mathcal{E}$ whose properties are independent of the boundary effects, i.e. independent of L. This is a typical physicist's approach. Dealing with a very large box Ω instead of $\mathbb{R}^n$ saves us many technical complications.

To begin with, we require that $u \to \text{Null}G$, as $|x| \to \infty$, otherwise, $\mathcal{E}(u)$ would grow proportionally to $\text{Vol}\Omega$, which we want to avoid.

Since $G \geq 0$, $\mathcal{E}$ is w.l.s.c. and coercive on $H_1^{(g)}(\Omega)$, where $g : \partial\Omega \to \text{Null}G$. The function g is our boundary condition. If $n = 1$, or $u(x)$ depends on one coordinate only, say x_1, then we have four distinct boundary conditions (BC): $u(x_1) \to \pm 1$ as $x_1 \to L$ and $u(x_1) \to \pm 1$, as $x_1 \to -L$. Consequently, we are led to consider $\mathcal{E}$ on the following four spaces:

$$M_{\pm,\pm} = \{u \in H_1(\Omega) : \text{ one of the above BC holds}\}.$$

Clearly, $\mathcal{E}$ attains its strict minimum on $M_{+,+}$ and $M_{-,-}$ at $u_+(x_1) \equiv 1$ and $u_-(x_1) \equiv -1$, respectively. In the phase separation model described by the Ginzburg–Landau functional, these minimizers describe homogeneous phases. Next, a minimizer on $M_{+,-}$ is obtained from a minimizer on $M_{-,+}$ by reflection $u(x_1) \to u(-x_1)$. Thus it suffices to consider only minimizers on $M_{-,+}$. Observe that by the reflection symmetry of $\mathcal{E}$, we look for odd minimizers, i.e. we pass from $M_{-,+}$ to

$$M_{-,+}^{\text{odd}} := \{u \in M_{-,+} : u(-x_1) = -u(x_1)\}.$$

As shown above, $\mathcal{E}$ is w.l.s.c. and coercive on $M_{-,+}^{\text{odd}} \subset H_1(\Omega)$. Moreover, one can show that $M_{-,+}^{\text{odd}}$ is weakly closed in $H_1(\Omega)$. (Recall that we think about $M_{-,+}$ as $M_{-,+} = \tilde{g} + H_1^{(0)}(\Omega)$, where $\tilde{g}$ is a smooth (and odd) extension of the function g from the boundary $\partial\Omega$ to the interior of Ω. Then since $H_1^{(0)}(\Omega)$ is weakly closed, then so is $M_{-,+}$.)

Now by the key theorem above, $\mathcal{E}$ has a minimizer $u_k^{(L)}$ on $M_{-,+}^{\text{odd}}$.

This is the kink discussed in the previous section. One can show that as $L \to \infty$, $u_k^{(L)} \to u_k$, where u_k is a minimizer of $\mathcal{E}$ on $M_{-,+}^{\text{odd}}$ for $L = \infty$, i.e. if $\Omega = \mathbb{R}$.

Of course, by shifting $u_k(x)$ to $u_k(x-h)$, in the case $L = \infty$, we obtain a one–parameter family of minimizers, the kinks centered at different points of $\mathbb{R}$.

As was discussed in the previous section, the kink solutions describe planar interfaces. To find solutions corresponding to spherical drops and cylinders, we take for Ω a ball B of radius L centered at the origin. We minimize the energy functional $\mathcal{E}$, as was described above, on the set of spherically symmetric functions of the form

$$M' := \{u(x) = \varphi(|x|) : |x| \leq L, \varphi \in M_0 \equiv M_{R=0}\},$$

where

$$M_R := \{\varphi \in H^{(g)}(B) : \varphi(R) = 0\},$$

where $g(0) = 1$ and $g(L) = -1$. On this set, the functional $\mathcal{E}$ takes the form

$$\mathcal{E}(u) = \sigma_{n-1} e(\varphi),$$

where σ_n is the volume of the n–dimensional unit sphere, and

$$e(\varphi) = \int_0^L \left(\frac{1}{2}(\varphi')^2 + G(\varphi) \right) r^{n-1} dr.$$

The functional e defined on M_R is w.l.s.c. and coercive. Moreover, as before, M_R is weakly closed. Hence e has a minimizer φ_R on M_R. This minimizer describes a spherical drop, if $n = 3$, or a cylinder if $n = 2$, of fixed radius$R$. Recall that in order to find a true (stable or metastable) sphere or cylinder, we have to find critical points of the energy $V(R) := e(\varphi_R)$.

Ground state of the nonlinear Schrödinger equation. Now we show how to use the Key Theorem in order to prove existence of solutions of differential equations.

Let Ω be a smooth bounded domain in $\mathbb{R}^n$. For $\lambda \in \mathbb{R}$ and $p > 2$, we consider the problem

$$\begin{aligned} -\Delta u - |u|^{p-2}u &= -\lambda u \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{30.2}$$

We want to prove existence of solutions of this boundary value problem. Denote by λ_1 the lowest eigenvalue of $-\Delta$ on Ω with Dirichlet boundary conditions. We have the following

Theorem. *Let $n \geq 3$ and $p \leq \frac{2n}{n-2}$. Then for any $\lambda > -\lambda_1$, there is a solution to the problem (30.2).*

Discussion. Differential equation (30.2) is the Euler–Lagrange equation for the functional

$$F(u) = \int_\Omega \left(\frac{1}{2}|\nabla u|^2 - \frac{1}{p}|u|^p + \frac{\lambda}{2}|u|^2 \right) d^n x.$$

However, for $p > 2$, this functional is not bounded from below. Indeed, take $u_\mu = \mu u$ with a fixed function u, and some $\mu > 0$. Then

$$F(u_\mu) = \mu^2 \frac{1}{2} \int_\Omega \left(|\nabla u|^2 + \lambda |u|^2 \right) d^n x - \frac{\mu^p}{p} \int_\Omega |u|^p \to -\infty, \quad \text{as } \mu \to \infty.$$

Consequently, this functional does not have a minimizer. Taking $u^{(\mu)} := \mu^\alpha u(\mu x)$ with a fixed function u and some μ, we find

$$F(u^{(\mu)}) = \mu^{2+2\alpha-n} \frac{1}{2} \int_\Omega |\nabla u|^2 - \mu^{\alpha p - n} \frac{1}{p} \int_\Omega |u|^p + \mu^{2\alpha - n} \frac{1}{2} \int_\Omega |u|^2.$$

Take now α so that $2 + 2\alpha - n > 0$, and $2 + 2\alpha - n > \alpha p - n$, i.e. $\frac{2}{p-2} > \alpha > \frac{n}{2} - 1$. Since $\frac{2}{p-2} > \frac{n-2}{2}$ (because $\frac{p}{2} < \frac{2}{n-2} + 1 = \frac{n}{n-2}$), this is possible. Then we get $F(u^{(\mu)}) \to \infty$ as $\mu \to \infty$, which shows that F is not bounded from above, and consequently, it has no maximizer either.

To get out of this dilemma, we consider the constraint problem: minimize the functional

$$\mathcal{E}(u) = \frac{1}{2}\int_\Omega \left(|\nabla u|^2 + \lambda |u|^2\right) d^n x, \tag{30.3}$$

subject to the constraint $J(u) = 1$, where

$$J(u) := \frac{1}{p}\int_\Omega |u|^p d^n x. \tag{30.4}$$

If such a problem has a minimizer v_0, then by the Lagrange multiplier theorem, v_0 satisfies the equations

$$-\Delta v_0 + \lambda v_0 - \mu |v_0|^{p-2} v_0 = 0, \tag{30.5}$$

and

$$\frac{1}{p}\int_\Omega |v_0|^p d^n x = 1,$$

for some $\mu \in \mathbb{R}$. But now we have the undesirable coefficient μ. To get rid of this coefficient, we first show that $\mu > 0$. Indeed, multiplying (30.5) by $\overline{v}_0$, integrating the result over Ω, and then integrating by parts, we obtain

$$\mu \int_\Omega |v_0|^p d^n x = \int_\Omega \left(|\nabla v_0|^2 + \lambda |v_0|^2\right) d^n x,$$

so indeed $\mu > 0$. Now we rescale v_0 as

$$u_0(x) := \mu^{\frac{1}{p-2}} v_0(x),$$

then clearly u_0 satisfies (30.2).

Proof of the theorem. We show that the functional $\mathcal{E}$ defined in (30.3) has a minimizer in the set

$$M = \{u \in H_1^{(0)}(\Omega) : J(u) = 1\}.$$

To this end, we have to show that

(α) M is weakly closed,

(β) $\mathcal{E}$ is w.l.s.c.,

(γ) $\mathcal{E}$ is coercive.

Properties (α)–(γ) were proved before as a part of our exercises. So by the key theorem, $\mathcal{E}$ has a minimizer v_0 in M, which by the argument given in the discussion above leads to a (weak) solution of problem (30.3). For a definition of the weak solution, see below. A weak solution can be upgraded to a classical solution by the elliptic regularity argument also described below. ■

Discussion. There is one subtle issue here which we brushed under the rug: the argument above shows that $u_0 \in H_1^{(0)}(\Omega) \cap L^p(\Omega)$. Hence all we know is that $\Delta u_0 \in H_{-1}(\Omega)$ and therefore, we have to specify what we mean by saying that u_0 satisfies (30.2).

Note that if u_0 is a smooth function satisfying (30.2), then multiplying (30.2) by $v \in C_0^\infty(\Omega)$ and integrating by parts, we obtain

$$-\int_\Omega u_0 \, \Delta v - \int_\Omega f(u_0) v = 0, \tag{30.6}$$

where $f(u) = |u|^{p-2}u - \lambda u$. A function $u_0 \in H_1^{(0)}(\Omega)$ is called a *weak solution* to (30.2) if it satisfies (30.6) for any $v \in C_0^\infty(\Omega)$.

Exercises. Prove existence of (weak) solutions of the following boundary value problems (below, Ω is a bounded domain in $\mathbb{R}^n$):

(a) the Dirichlet problem:

$$\begin{aligned} \Delta u &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega, \end{aligned}$$

for every $f \in L^2(\Omega)$ and any smooth $g : \partial\Omega \to \mathbb{R}$;

(b) the nonlinear eigenvalue problem:

$$\begin{aligned} \Delta u + a(x)|u|^{p-1}u &= \mu u \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where $a(x)$ is a smooth and positive function on Ω, $n \geq 3$, $2 < p < \frac{2n}{n-2}$ and $\mu \geq 0$;

(c) the nonlinear Dirichlet problem:

$$\begin{aligned} \nabla\left(|\nabla u|^2 \nabla u\right) &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

for any $f \in L^{4/3}(\Omega)$. (Hint: reduce this to a minimization problem on the Sobolev space

$$H_4^{(0)}(\Omega) = \{u \in L^4(\Omega) : \int_\Omega |\nabla u|^4 < \infty \text{ and } u|_{\partial\Omega} = 0\}).$$

How to gain smoothness: elliptic regularity. Assume we show that the following equation has a (weak) solution in $H_1(\Omega)$:

$$\Delta u = a(x) u^4 \quad \text{in } \Omega,$$

and $u = 0$ on $\partial\Omega$ (Dirichlet boundary conditions). Here, a is smooth and $\Omega \subset \mathbb{R}^3$ is bounded. This is not so good since $\Delta u \in H_{-1}(\Omega)$. But it turns out that in fact u is smooth!

We can show this heuristically in the following way. By the Sobolev embedding theorem (i.e. $H_1(\Omega) \subset L^\alpha(\Omega)$ with $\alpha < \frac{2n}{n-2} = 6$), we have that $u \in L^\alpha(\Omega)$ with $\alpha < 6$. Hence $u^4 \in L^\beta(\Omega)$ with $\beta < 3/2$, so $\Delta u \in L^\beta(\Omega)$ since a is smooth.

Now assume u has a singularity (at 0, say), i.e. $u \sim |x|^{-\sigma}$ around 0, for some $\sigma > 0$. Then $\Delta u \sim |x|^{-\sigma-2}$ around 0, and $\Delta u \in L^\beta(\Omega)$ implies that $|x|^{-\beta(\sigma+2)}$ is locally integrable (in $\mathbb{R}^3$) at 0, so $-2 + \beta(\sigma+2) < 1$ or $\sigma < 3/\beta - 2$. This holds for all $\beta = 3/2 - \epsilon$, $\forall \epsilon > 0$ small. We thus get $\sigma < 3/\beta - 2 \leq 4\epsilon$. This shows that the singularity is very weak: for any $\gamma < \infty$, we can choose ϵ s.t. $|x|^{-\sigma\gamma}$ is locally integrable (take e.g. $\epsilon < \frac{1}{2\gamma}$). This means that $u \in L^\gamma(\Omega)$, for any $\gamma < \infty$ and therefore we have $\Delta u \in L^\gamma(\Omega)$ $\forall \gamma < \infty$, so $u \in C^{2-\epsilon}(\Omega)$ $\forall \epsilon > 0$ (this is again a Sobolev embedding theorem). We can repeat this process to show that actually u is smooth.

31. Saddle points and the Mountain Pass Lemma

Often, we are interested in saddle points of functionals, rather than their extrema. This happens especially when the functionals of interest are not bounded below or above, and therefore have no minima or maxima. Such functionals appear in various applications and a typical example is

$$\mathcal{E}(u) = \int_\Omega \left(\frac{1}{2} |\nabla \psi|^2 - \lambda |\psi|^p \right) d^n x,$$

where Ω is a domain in $\mathbb{R}^n$, $n \geq 3$ and $2 < p < \frac{2n}{n-2}$. We have already shown that such a functional is bounded neither from below nor from above.

Such functionals may appear directly in a problem, or they may occur through equations: for instance assume we want to solve the equation

$$\begin{aligned} \Delta u &= -a(x)|u|^{p-2}u \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{31.1}$$

Here again, Ω is a domain in $\mathbb{R}^n$, $n \geq 3$ and $p > 2$. It turns out that this equation is the Euler–Lagrange equation for the functional

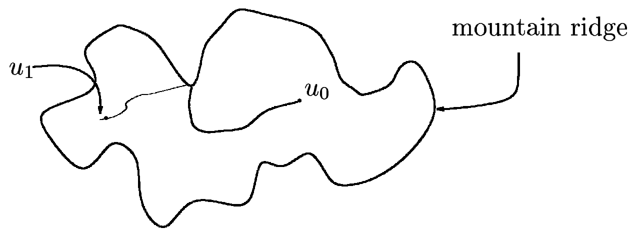
$$\mathcal{E}(u) = \int_\Omega \left(\frac{1}{2} |\nabla u|^2 - \frac{1}{p} a(x)|u|^p \right) d^n x, \tag{31.2}$$

which is a generalization of the one considered above.

Exercise. Show that (31.1) is the Euler–Lagrange equation of (31.2).

One can try to solve the boundary value problem above either directly, or using the variational calculus. The latter is often easier. In this section, we sketch a key technique for finding critical points of functionals. This technique is in particular applicable to functionals of the form (31.2).

The main result is called the Mountain Pass Lemma (MPL). It is due to Ambrosetti and Rabinowitz. Its main idea is the same as that for finding a pass across a mountain ridge as shown in this figure:

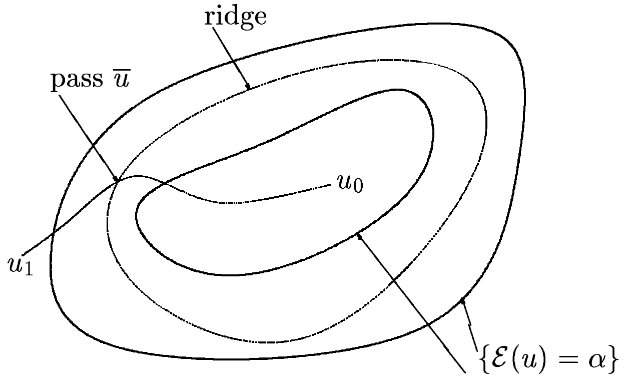


Now we cast this picture into mathematical terms. Consider a functional $\mathcal{E}$ on a (reflexive) Banach space X. Denote by $B(u_0, \rho_0)$ the ball of radius ρ centered at u_0: $B(u_0, \rho_0) = \{u \in X : ||u - u_0|| < \rho\}$. We assume that $\exists u_0, u_1, \alpha$ and $\rho > 0$, s.t.

(i) $\mathcal{E}(u_0) < \alpha$, and $\mathcal{E}|_{\partial B(u_0, \rho)} \geq \alpha$,

(ii) $u_1 \in X \backslash B(u_0, \rho)$, and $\mathcal{E}(u_1) \leq \mathcal{E}(u_0)$.

An important feature of these conditions is that the set $\{u \in X : \mathcal{E}(u) \leq \mathcal{E}(u_0)\}$ is disconnected:



To find a pass in a mountain ridge, we proceed as follows. Consider all paths from u_0 to u_1:

$$\Gamma = \{\gamma : I = [0,1] \to X| \ \ \gamma(0) = u_0, \gamma(1) = u_1, \gamma \text{ continuous}\}$$

Denote the height of the ridge at γ by

$$F(\gamma) := \sup_{t \in I} \mathcal{E}(\gamma(t)).$$

To find the lowest point in the ridge, we minimize the hight of the ridge at γ over all paths in Γ:

$$h = \inf_{\gamma\in\Gamma} \sup_{t\in I} \mathcal{E}(\gamma(t)). \tag{31.3}$$

The number h represents the height of the mountain pass, and it satisfies $h \geq \alpha$.

If the inf sup is attained at some points $\overline{\gamma}$ and $\overline{t} = \overline{t}(\overline{\gamma})$, then h is the value of $\mathcal{E}$ at $\overline{u} = \overline{\gamma}(\overline{t})$, $h = \mathcal{E}(\overline{u})$. The point $\overline{u}$ is the highest point on the mountain pass, and we expect that the slope there is horizontal: $\mathcal{D}\mathcal{E}(\overline{u}) = 0$. To see that this is indeed what happens, we need an unpleasant technical assumption: the *Palais–Smale (PS) condition.* A sequence $\{u_n\}$ is called a *Palais–Smale (or simply PS) sequence* iff

(PS1) $|\mathcal{E}(u_n)| \leq C$, uniformly in n,

(PS2) $||\mathcal{D}\mathcal{E}(u_n)||_{X'} \to 0$, as $n \to \infty$.

The Palais–Smale Condition is the following:

(PS) Any PS sequence contains a convergent subsequence.

Theorem (the Mountain Pass Lemma). *Let $\mathcal{E}$ be C^1 in X, and suppose conditions (i), (ii) and (PS) are satsified. Then $\mathcal{E}$ has a critical point u_0 with $\mathcal{E}(u_0) = h$.*

Idea of the proof. (For details, see E. Zeidler, Applied Functional Analysis, Springer, Theorem 25, p.88) Since for a given γ, $t \mapsto \mathcal{E}(\gamma(t))$ is a continuous real function, on the closed bounded interval $I = [0, 1]$, it reaches its maximum, say at t_γ. Thus $F(\gamma) = \mathcal{E}(\gamma(t_\gamma))$. Now let γ_n be a minimizing sequence for $F(\gamma)$:

$$F(\gamma_n) \to h = \inf_{\gamma} F(\gamma).$$

The minimizing sequence can be picked so that $F(\gamma_n) \leq h + 1$. The key claim is that the sequence $u_n := \gamma_n(t_{\gamma_n})$ is a PS sequence. The first defining property of a PS sequence, (PS1), is immediate:

$$\mathcal{E}(u_n) = \mathcal{E}(\gamma_n(t_{\gamma_n})) = F(\gamma_n) \leq h + 1.$$

The second defining property, (PS2), namely $\mathcal{DE}(u_n) \to 0$, as $n \to \infty$, is intuitively clear, but very tedious to prove. The condition says just that the slope at points of the ridge leading to the pass decreases, i.e. the path becomes more and more horizontal as we approach the top of the pass.

Given that $\{u_n\}$ is a PS sequence, the rest is easy. By the PS–condition, $\{u_n\}$ contains a convergent subsequence, which we denote again by $\{u_n\}$. Let $u_n \to \overline{u}$. Since $u \mapsto \mathcal{DE}(u)$ is continuous, then $\mathcal{DE}(\overline{u}) = \lim \mathcal{DE}(u_n) = 0$, so $\overline{u}$ is a critical point, and $\mathcal{E}(\overline{u}) = \lim \mathcal{E}(u_n) = h$. ■

Example. Let us go back to the functional (31.2). Assume Ω is a bounded domain in $\mathbb{R}^n$, $n \geq 3$, and $2 < p < \frac{2n}{n-2}$ and a continuous. We check first that (31.2) is defined on $H_1^{(0)}(\Omega)$:

$$\left| \int_\Omega a|u|^p \right| \leq \sup_\Omega |a| \int_\Omega |u|^p. \tag{31.4}$$

In the section on Sobolev spaces, we have proven that the following inequality holds, for any $u \in H_1^{(0)}(\Omega)$:

$$||u||_p \leq C_p ||u||_{(1)}, \quad 2 \leq p \leq \frac{2n}{n-2}. \tag{31.5}$$

Recall that $\|u\|_{(k)}$ is the norm in the Sobolev space H_k. Thus, if $u \in H_1^{(0)}(\Omega)$, then $\int_\Omega |\nabla u|^2 < \infty$, and $|\int_\Omega a|u|^p| < \infty$, so $|\mathcal{E}(u)| < \infty$. Therefore, $\mathcal{E}$ is defined on $H_1^{(0)}(\Omega)$.

Theorem. *The functional (31.2) has a critical point in the space $H_1^{(0)}(\Omega)$. Consequently, the boundary value problem (31.1) has a (weak) solution.*

Proof. We check that the conditions of the MPL are satisfied. We choose $u_0 = 0$, so $\mathcal{E}(u_0) = 0$.

(i) We claim that

$$\mathcal{E}(u) \geq \frac{1}{2} c\rho^2, \tag{31.6}$$

provided $||u||_{(1)} = \rho$, where $\rho = \frac{c}{2\sup_\Omega |a|}$, and where c is the (largest) constant entering into the Poincaré inequality (c.f. (17.6)):

$$\int_\Omega |\nabla u|^2 \geq c||u||^2_{(1)},$$

i.e. c is the lowest eigenvalue of $-\Delta$ on Ω.

Indeed, equations (31.4) and (31.5) imply that

$$\left|\int_\Omega a|u|^p\right| \leq \sup_\Omega |a| \, ||u||^p_{(1)}.$$

This together with the Poincaré inequality implies

$$\mathcal{E}(u) \geq ||u||^2_{(1)} \left(c - \sup_\Omega |a| \, ||u||^p_{(1)} \right),$$

which in turn implies (31.6).

Thus condition (i) of the MPL is satisfied with $\alpha = c\rho^2/2$, and $\rho = \frac{c}{2\sup_\Omega |a|}$.

(ii) We claim that condition (ii) of the MPL is satisfied as well: take $u_1 = \lambda u$, with $u \in H_1^{(0)}(\Omega)$ arbitrary but fixed, and λ sufficiently large. Then we get

$$\mathcal{E}(\lambda u) = \lambda^2 \frac{1}{2} \int_\Omega |\nabla u|^2 - \lambda^p \frac{1}{p} \int_\Omega a|u|^p \to -\infty,$$

as $\lambda \to \infty$ since $p > 2$. So for λ sufficiently large, and $u_1 = \lambda u$, we get $\mathcal{E}(u_1) < 0$.

(PS) This condition is the most difficult to check. Let $\{u_n\}$ be a PS sequence, i.e.

$$|\mathcal{E}(u_n)| \leq C \tag{31.7}$$

$$||\mathcal{D}\mathcal{E}(u_n)||_{X'} \to 0 \tag{31.8}$$

We want to show that $\{u_n\}$ contains a convergent subsequence. Equation (31.7) implies that

$$\frac{1}{2}||\nabla u_n||^2 \leq C + \frac{1}{p}\left|\int_\Omega a|u_n|^p\right|, \tag{31.9}$$

and (31.8) implies that

$$\forall\epsilon\ \exists n_0 : ||\mathcal{DE}(u_n)||_{X'} \leq \epsilon, \quad \text{for } n \geq n_0. \tag{31.10}$$

Equation (31.10) and the Poincaré inequality yield

$$\left|\int_\Omega \nabla u_n \cdot \nabla v - \int_\Omega a|u_n|^{p-2}u_n v\right| \leq \epsilon||v||_{(1)} \leq \epsilon(1+\sqrt{c})||\nabla v||,$$

for any $v \in H_1^{(0)}(\Omega)$, provided n is sufficiently large. Take $\epsilon = (1+\sqrt{c})^{-1}$, and $v = u_n$. Then the last estimate gives

$$\left|\int_\Omega a|u_n|^p\right| \leq ||\nabla u_n|| + ||\nabla u_n||^2.$$

Combining this with (31.9) gives

$$\frac{1}{2}||\nabla u_n||^2 \leq C + \frac{1}{p}||\nabla u_n|| + \frac{1}{p}||\nabla u_n||^2.$$

Since $p > 2$, this implies $||\nabla u_n|| \leq C_1$. Thus $\{u_n\}$ is bounded in $H_1^{(0)}(\Omega)$ uniformly in n (Poincaré inequality), so there is a subsequence $\{u_{n_k}\}$ converging weakly in $H_1^{(0)}(\Omega)$.

Now observe that $\mathcal{DE}(u_n) = -\Delta u_n + a|u_n|^{p-2}u_n$. We solve this equation for u_n as

$$u_n = (-\Delta)^{-1}\mathcal{DE}(u_n) - (-\Delta)^{-1}a|u_n|^{p-2}.$$

Since $\mathcal{DE}(u_n) \to 0$ in $H_{-1}(\Omega)$, we have that $(-\Delta)^{-1}\mathcal{DE}(u_n) \to 0$ in $H_1(\Omega)$.

Next, we claim that the operator $u \mapsto \Delta^{-1}G(u)$, where $G(u) := a|u|^{p-2}u$, is compact in $H_1(\Omega)$, i.e. it maps weakly convergent sequences into strongly convergent ones. Indeed, $u_n \xrightarrow{w} u$ in $H_1(\Omega)$ implies that (up to a subsequence) $u_n \to u$ in $L^\alpha(\Omega)$, for any $\alpha < \frac{2n}{n-2}$ (see (31.5)), and in particular for $\alpha = p$. Since a is uniformly bounded, the latter fact implies that $a|u_n|^{p-2}u_n \to a|u|^{p-2}u$ in $L^{\frac{p}{p-1}}(\Omega)$. Now since $\frac{p}{p-1} > \frac{2n}{n-2}$, we have by the Sobolev embedding theorem that

$$(-\Delta)^{-1}a|u_n|^{p-2}u_n \to (-\Delta)^{-1}|u|^{p-2}u$$

in $H_1(\Omega)$. Hence u_n converges in $H_1(\Omega)$, and consequently, the (PS) condition is fulfilled as well.

We verified all the conditions of the MPL for the functional (31.2) on $H_1^{(0)}(\Omega)$, hence by the MPL this functional has a critical point. ■

Chapter V. Dynamical systems

A dynamical system is an evolution equation of the form

$$\frac{du_t}{dt} = F(u_t), \tag{32.1}$$

where $t \in [0, \infty)$ is the "time"–variable, and $t \mapsto u_t$ is a differentiable path in a Banach space X, i.e. a differentiable vector function from $[0, \infty)$ to X. It is assumed that F maps the space X into itself. X is called the *phase space* or *state space* and F is called a *vector field* on X.

Examples. 1) A one dimensional dynamical system is given by $\dot{x} = F(x)$, $x : \mathbb{R} \to \mathbb{R}$. Here, $x = x_t$ and $\dot{x} = \frac{dx}{dt}$.
2) A two dimensional dynamical system coming from Newton's equation $\ddot{x} = -V'(x)$, where $V : \mathbb{R} \to \mathbb{R}$ is the potential corresponding to the "force" $-V'$ (the negative derivative of V) is given by

$$\frac{d}{dt}\begin{pmatrix} x \\ \dot{x} \end{pmatrix} = \begin{pmatrix} \dot{x} \\ -V'(x) \end{pmatrix}.$$

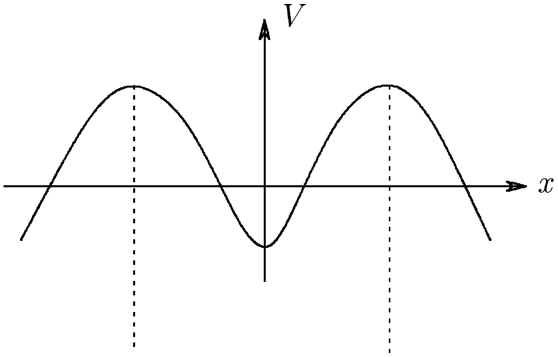
3) The heat equation:

$$\frac{du_t}{dt} = \Delta u_t + g(u_t).$$

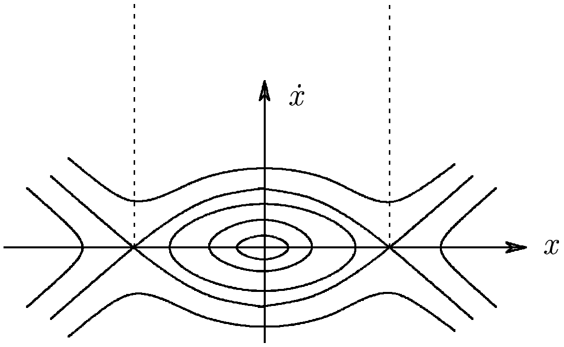
Here, $F(u) = \Delta u + g(u)$.

Other examples are the Schrödinger equation, the wave equation, the Euler equation, the Navier–Stokes equation, the KdV equation, etc.

In the theory of dynamical systems one is interested in the *long time* behaviour of solutions u_t for various *initial conditions* u_0 (i.e. u_0 is considered to be given). One usually seeks qualitative pictures describing possible scenarios of the evolution of the system rather than quantitative ones. For example, assume the potential V in example 2) is of the following form (two hills and a valley in between)



Then trajectories of this system in phase space $\mathbb{R}^2$ (space of points $(x, \dot{x})$) look like



The subject of dynamical systems is vast. We consider one of the central questions which is natural from the viewpoint of this course, namely the *long–time dynamics near equilibria.* However, before proceeding to this topic, let us introduce some basic notions.

32. The flow

If the vector field F is linear, $F(u) = Au$, then the flow is another name for the evolution operator, or exponent $\exp(At)$. Moreover, the notion of flow covers also the nonlinear case. Assume equation (32.1) has a unique solution, u_t, say on the interval $t \in [0, T]$, for some $T > 0$, and for any $u_0 \in X$. This defines a family of maps $\phi_t : X \to X$ by

$$\phi_t(u_0) = u_t,$$

i.e. the map ϕ_t shifts the function on which it acts along the solution of equation (32.1). The family $\{\phi_t\}_{t\geq 0}$ is called the *flow* (generated by F). Flows have the following properties:

(a) $\phi_0 = \mathbb{1}$,

(b) $\phi_t \circ \phi_s = \phi_{t+s}$,

(c) $F(\overline{u}) = 0 \Longleftrightarrow \phi_t(\overline{u}) = \overline{u}$.

A point $\overline{u}$ is called an *equilibrium point* of (32.1) iff $\overline{u}$ does not depend on time, i.e. iff $F(\overline{u}) = 0$.

Property (c) says that the fixed points of the flow are precisely the zeroes of the vector field F, i.e. equilibrium points of (32.1).

Exercise. Prove (a)–(c).

Let $\phi : X \to X$ be a map. We define the *discrete time flow* $\phi^n(u)$ for $n = 0, 1, \ldots$ by $\phi^0 := \mathbb{1}$, and

$$\phi^n := \underbrace{\phi \circ \ldots \circ \phi}_{n \text{ times}}, \; n > 0.$$

For example if ϕ_t is a flow, then $\phi_n = (\phi_1)^n$ is a discrete flow.

Example: linear flow. Let $F(u) = Au$, where A is a bounded linear operator (or a self–adjoint operator). Then the flow is given by

$$\phi_t(u) = e^{At}u.$$

This linear flow exists for all times. This formula shows that ϕ_t generalizes the notion of the exponential mapping or evolution operator (in fact, one can think about a flow as an exponential of, in general, nonlinear maps).

The next theorem shows that the linearization of a flow at an equilibrium point is a linearized flow.

Theorem. *Let F be a C^1–map, and let $\overline{u}$ be an equilibrium point (i.e. $F(\overline{u}) = 0$). Then ϕ_t is C^1, and*

$$\mathcal{D}\phi_t(\overline{u}) = e^{t\mathcal{D}F(\overline{u})}. \tag{32.2}$$

Exercise. Prove this theorem. (Hint: assuming ϕ is C^1, take the Fréchet derivative of the initial value problem $\frac{\partial}{\partial t}\phi_t(u) = F(\phi_t(u))$, and $\phi_0(u) = u$ at $\overline{u}$. The solution of the resulting equation leads to (32.2). Then use this to argue that ϕ_t is in fact C^1).

Our task is to study the behaviour of sulutions to equation (32.1) near an equilibrium point $\overline{u}$. In particular, we want to answer the following questions:

- Do they stay in a neighbourhood of $\overline{u}$ (*Lyapunov stability*)?
- Do they converge to $\overline{u}$ as $t \to +\infty$ (*asymptotic stability*)?
- Do they move away from $\overline{u}$ as t progresses (*instability*)?

To answer these questions, i.e. to understand the stability character of the equilibrium $\overline{u}$, we look for solutions of the form

$$u_t = \overline{u} + \xi_t,$$

where ξ_t is "small" relative to $\overline{u}$. The path $t \mapsto \xi_t$ is called the *fluctuation* of u_t around $\overline{u}$. It will be shown below that in the leading order in the size of ξ_t (i.e. $||\xi_t||$), ξ_t satisfies the linearized equation (c.f. (34.3))

$$\frac{d\xi_t}{dt} = \mathcal{D}F(\overline{u})\xi_t. \tag{32.3}$$

Thus our first task is to understand the behaviour of the linearized dynamics (32.3) near its equilibrium $\overline{\xi} = 0$ ($\Leftrightarrow u_t = \overline{u}$).

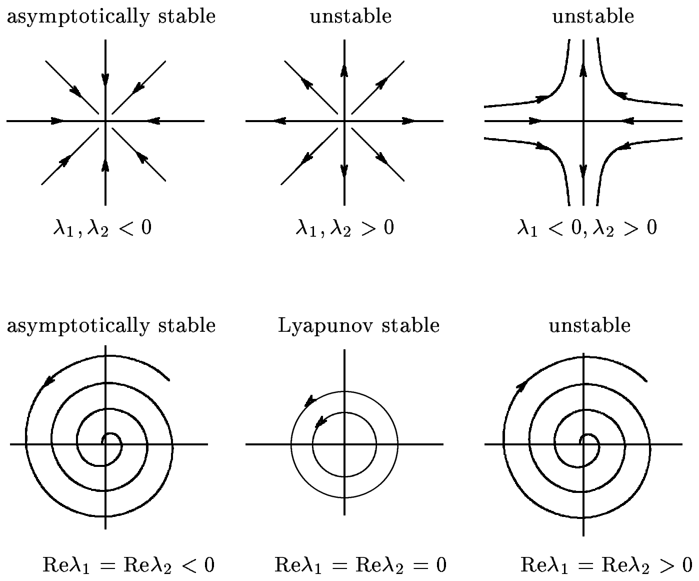
33. Dynamics near equilibrium: linear case

Let A be a (bounded) linear operator on a Banach space X. We consider the evolution equation

$$\frac{du_t}{dt} = Au_t. \tag{33.1}$$

This equation has an equilibrium solution at $\overline{u} = 0$. We want to understand the behaviour of solutions of this equation near this equilibrium point, i.e. for u_t small. This behaviour can be rather complicated, as can be seen from the following simple

Example. Let $X = \mathbb{R}^2$ and let A be a real 2×2 matrix with eigenvalues λ_1, λ_2. According to the nature of the spectrum of A, we have the following qualitative pictures for the flow $\phi_t(u) = e^{At}u$:



asymptotically stable — $\lambda_1, \lambda_2 < 0$

unstable — $\lambda_1, \lambda_2 > 0$

unstable — $\lambda_1 < 0, \lambda_2 > 0$

asymptotically stable — $\mathrm{Re}\lambda_1 = \mathrm{Re}\lambda_2 < 0$

Lyapunov stable — $\mathrm{Re}\lambda_1 = \mathrm{Re}\lambda_2 = 0$

unstable — $\mathrm{Re}\lambda_1 = \mathrm{Re}\lambda_2 > 0$

Exercises. **1)** Justify the picture above, **2)** analyze the stability properties of the static solution of the equation $\ddot{x} + c\dot{x} + kx = 0$, where $c \geq 0, k > 0$ (damped oscillator) for $c > 0$ and $c = 0$.

This example shows that we should try to connect the behaviour of

u_t near $\overline{u} = 0$ with the location of the spectrum of A with respect to the imaginary axis. For instance, if $\sigma(A) \subset \mathbb{C}$ lies in the open left half plane, then we expect that 0 is a stable equilibrium.

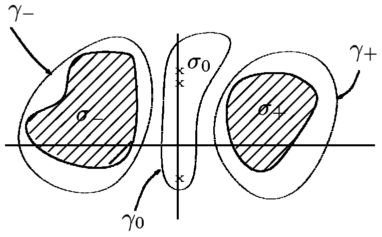
We decompose thus the spectrum of A as

$$\sigma(A) = \sigma_- \cup \sigma_0 \cup \sigma_+,$$

where

$$\sigma_\pm := \sigma(A) \cap \mathbb{C}_\pm, \quad \text{and} \quad \sigma_0 := \sigma(A) \cap i\mathbb{R},$$

with $\mathbb{C}_\pm = \{z \in \mathbb{C} : \mathrm{Re} z \gtrless 0\}$. In other words, σ_- is the part of the spectrum of A lying in the left half plane and so on. We assume that the components σ_-, σ_0 and σ_+ are disjoint sets:



Now we define spectral subspaces associated with these components as $V_n = \mathrm{Ran} P_n$, $n = \pm, 0$, where P_n are bounded operators defined as

$$P_n = \frac{1}{2\pi i} \oint_{\gamma_n} (A - z)^{-1} dz,$$

here, γ_n is a contour encircling the component σ_n but not intersecting or containing the other components (see the figure above). Our main result is

Theorem. *The subspaces V_n, $n = \pm, 0$ have the following properties:*

(i) they span uniquely the entire space X in the sense that any vector $u \in X$ can be uniquely written as a sum

$$u = u_- + u_0 + u_+, \quad \textit{where } u_n \in V_n,\ n = \pm, 0, \tag{33.2}$$

i.e. the space X can be decomposed as $X = V_- \oplus V_0 \oplus V_+$;

(ii) they are invariant under A, i.e. A maps V_n into V_n, $n = \pm, 0$;

(iii) they satisfy

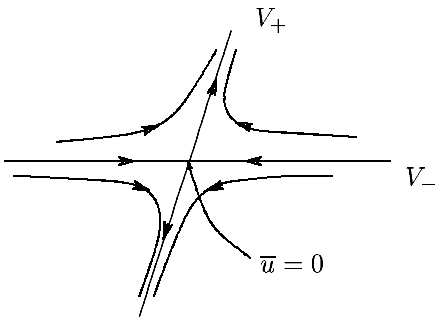
$$\forall u \in V_\pm \text{ we have } ||e^{At}u|| \to 0 \text{ as } t \to \mp\infty,$$
$$\forall u \in V_\pm, u \neq 0 \text{ we have } ||e^{At}u|| \to \infty \text{ as } t \to \pm\infty,$$
$$\forall u \in V_0,\ ||e^{At}u|| = ||u||.$$

Corollary. *Let $V_0 = \{0\}$ (i.e. σ_0 is empty: $\sigma_0 = \phi$, which is called the hyperbolic case). For any initial condition u, the solution of (33.1) satisfies*

$$||e^{At}u - e^{At}u_\pm|| \to 0 \quad \text{as } t \to \pm\infty,$$

where $u_\pm \in V_\pm$ and $u_+ + u_- = u$. In particular, if $\sigma_+ = \sigma_0 = \phi$, then the equilibrium $\overline{u} = 0$ is stable; if $\sigma_+ \neq \phi$, then it is unstable.

V_+, V_0 and V_- are called *unstable, central* and *stable* subspaces. Sometimes, they are also denoted as V_u, V_c and V_s, respectively.



An equilibrium point is called *hyperbolic* iff $V_c = \{0\}$. In the finite dimensional case, i.e. for $X = \mathbb{R}^n$, and A a $n \times n$ matrix, some $n \geq 1$, one can show that

> V_u =span{(root) eigenvectors of A corresponding to eigenvalues with real part > 0},

> V_c =span{(root) eigenvectors of A corresponding to eigenvalues with real part $= 0\}$,
>
> V_s =span{(root) eigenvectors of A corresponding to eigenvalues with real part $< 0\}$.

Recall that ξ is an eigenvector of A with eigenvalue λ iff ξ and λ satisfy the equation $(A - \lambda)\xi = 0$. ξ is called a *root eigenvector* of A with eigenvalue λ iff ξ and λ satisfy the equation $(A-\lambda)^n\xi = 0$ and moreover, $(A - \lambda)^{n-1}\xi \neq 0$, for some integer $n \geq 2$.

Exercises. 1) In the finite dimensional case, where A is a symmetric $n \times n$ matrix, prove the theorem as well as the characterization for V_s, V_0 and V_u given above. (Hint: use that (i) one can choose an orthonormal basis consisting of eigenvectors of A and (ii) $A\xi = \lambda\xi$ implies $e^{At}\xi = e^{\lambda t}\xi$). **2)** Let ξ be a root vector of A of order 2 for an eigenvalue λ, i.e. $(A - \lambda)^2\xi = 0$ but $(A - \lambda)\xi \neq 0$. Show that

$$e^{At}\xi = e^{\lambda t}\left(\mathbb{1} + t(A - \lambda)\right)\xi.$$

This formula shows that $||e^{At}\xi|| \to \infty$ if $\text{Re}\lambda > 0$.

We derive the theorem from properties of the operators P_n given in the following

Proposition. *Under the assumptions on $\sigma(A)$ made above, the operators P_n, $n = \pm, 0$, have the following properties:*

- *(a)* $[P_n, A] = 0$;
- *(b)* $P_n^2 = P_n$ *and* $P_n P_m = 0$ *if* $n \neq m$;
- *(c) the spectrum of A restricted to* $\text{Ran}P_n$ *is* σ_n: $\sigma(A \restriction V_n) = \sigma_n$;
- *(d)* $\forall\epsilon > 0$, $||e^{At}P_\pm|| \leq C_\epsilon e^{(\mu_\pm + \epsilon)t}$, $t \lessgtr 0$ *and* $||e^{At}P_\pm|| \geq c_\epsilon e^{(\mu_\pm + \epsilon)t}$, $t \gtrless 0$, *where* $\mu_- = \sup\{\text{Re}\lambda : \lambda \in \sigma_-\}$ *and* $\mu_+ = \inf\{\text{Re}\lambda : \lambda \in \sigma_+\}$; *moreover,* $\|e^{At}P_0\| = \|P_0\|$;
- *(e)* $P_+ + P_- + P_0 = \mathbb{1}$.

Discussion. Property (a) implies that $\text{Ran}P_n$ is invariant under A: if $u \in \text{Ran}P_n$, then $\exists v : u = P_n v$ and therefore $Au = AP_n u = P_n Av \in$

$\text{Ran}P_n$. Property (e) implies that any $u \in X$ can be decomposed as in (33.2): $u = \mathbb{1}u = P_-u + P_0u + P_+u$. Property (b) shows that this decomposition is unique (the first part of this property shows that the P_n's are projections). Property (d) implies the last statement of the theorem. We will not use property (c).

Sketch of the proof of the proposition. We prove some of the statements, the others are proven similarly.
(a) This statement follows from the representation

$$P_n = \frac{1}{2\pi i} \oint_{\gamma_n} (A-z)^{-1} dz$$

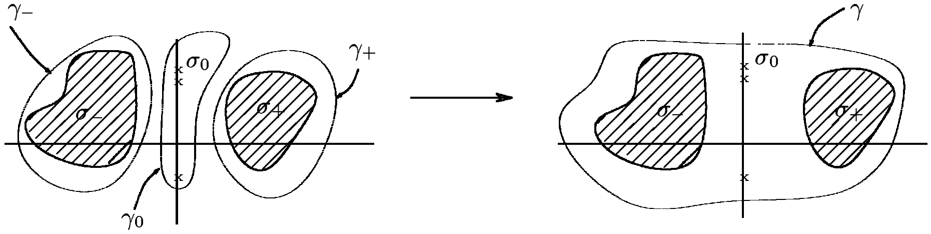
and the fact that $(A-z)^{-1}$ commutes with A.
(e) Deforming the contour of integration $\overline{\gamma} := \gamma_- \cup \gamma_0 \cup \gamma_+$ in the domain of analyticity of $(A-z)^{-1}$, we get from

$$P_- + P_0 + P_+ = \frac{1}{2\pi i} \oint_{\overline{\gamma}} (A-z)^{-1} dz$$

the following equation:

$$P_- + P_0 + P_+ = \frac{1}{2\pi i} \oint_{\gamma} (A-z)^{-1} dz, \tag{33.3}$$

where γ is a contour around the entire spectrum $\sigma(A)$:



We now show that the r.h.s. of (33.3) is equal to $\mathbb{1}$, so the result (e) follows. Choose $\gamma = \{|z| = \rho\}$, where $\rho > ||A||$. Now $(A-z)^{-1} = -z^{-1}(\mathbb{1} - A/z)^{-1}$ and $||A/z|| \leq ||A||/|z| < 1$ if $z \in \gamma$, so we can expand

$(\mathbb{1} - A/z)^{-1}$ into an absolutely converging geometrical series, and

$$(A - z)^{-1} = -z^{-1} \sum_{k=0}^{\infty} \left(\frac{A}{z}\right)^{k}.$$

We obtain therefore

$$\begin{aligned}
\frac{1}{2\pi i} \oint_{\gamma} (A - z)^{-1} dz &= -\sum_{k=0}^{\infty} A^{k} \frac{1}{2\pi i} \oint_{\gamma} z^{-k-1} dz \\
&= \sum_{k=0}^{\infty} A^{k} \begin{cases} 1 & k = 0 \\ 0 & k > 0 \end{cases} \\
&= \mathbb{1}.
\end{aligned}$$

(d) We prove only the part of (d) involving the upper bound on $e^{At}P_{-}$, the other cases are shown in the same way. We use the Cauchy formula to get:

$$e^{At}P_{-} = \frac{1}{2\pi i} \oint_{\gamma_{-}} \frac{e^{At}}{A - z} dz = \frac{1}{2\pi i} \oint_{\gamma_{-}} \frac{e^{zt}}{A - z} dz. \tag{33.4}$$

This equation holds since

$$\begin{aligned}
\frac{e^{At} - e^{zt}}{A - z} &= e^{zt} \frac{e^{(A-z)t} - 1}{A - z} \\
&= e^{zt} (A - z)^{-1} \sum_{k=1}^{\infty} \frac{t^{k}}{k!} (A - z)^{k} \\
&= e^{zt} \sum_{k=1}^{\infty} \frac{t^{k}}{k!} (A - z)^{k-1}
\end{aligned}$$

is analytic in z, so by Cauchy's theorem, we have

$$\oint_{\gamma_{-}} \frac{e^{At} - e^{zt}}{A - z} dz = 0,$$

which proves (33.4).

We take now γ_{-} s.t. the distance from γ_{-} to σ_{-} in $\mathbb{C}$ is bigger than or equal to ϵ, which implies that

$$||(A - z)^{-1}|| \leq \epsilon^{-1}, \ \forall z \in \gamma_{-}.$$

This gives us now immediately

$$\begin{aligned} ||e^{At}P_-|| &\leq \frac{1}{2\pi}\oint_{\gamma_-} e^{\mathrm{Re}zt}||(A-z)^{-1}||\,d|z| \\ &\leq \frac{1}{2\pi}e^{(\mu_-+\epsilon)t}\oint_{\gamma_-} ||(A-z)^{-1}||\,d|z| \\ &\leq C_\epsilon e^{(\mu_-+\epsilon)t}.\blacksquare \end{aligned}$$

Remark. In concrete cases, we can extend the analysis above to unbounded operators A. Using our previous results, we can do the stability analysis for A being a multiple of the Laplacian Δ, which is an unbounded operator. There is a general class of operators on Hilbert spaces, called *selfadjoint* operators, for multiples of which the analysis above can not only be done, but can actually be made more precise. For instance, we can show for such operators that if $\sigma(A) \subset \overline{\mathbb{C}_-}$, then $\overline{u}$ is Lyapunov stable.

34. Nonlinear case: discussion

We return to our nonlinear dynamical system

$$\frac{du_t}{dt} = F(u_t). \tag{34.1}$$

Assume there is an equilibrium state $\overline{u}$, i.e. $F(\overline{u}) = 0$. We want to understand the behaviour of solutions to (34.1) with initial conditions near this equilibrium $\overline{u}$. It is natural to represent such solutions in the form

$$u_t = \overline{u} + \xi_t,$$

where ξ_t is small compared to $\overline{u}$, at least for small values of t.

Plugging the latter formula for u_t into equation (34.1), we derive an equation for ξ_t:

$$\frac{d\xi_t}{dt} = F(\overline{u} + \xi_t).$$

To present this equation in a convenient form, we expand the r.h.s. around $\overline{u}$:

$$F(\overline{u}+\xi_t) = \mathcal{D}F(\overline{u})\xi_t + R(\xi_t),$$

where $R(\xi_t) = o(||\xi_t||)$ and where we have taken into account that $F(\overline{u}) = 0$. Thus the equation for ξ_t becomes

$$\frac{d\xi_t}{dt} = \mathcal{D}F(\overline{u})\xi_t + R(\xi_t). \tag{34.2}$$

This is our basic equation for ξ_t.

The term $R(\xi_t)$ is purely nonlinear in ξ_t and is small compared to $\mathcal{D}F(\overline{u})\xi_t$ for small $||\xi_t||$ (except in the case when $\mathcal{D}F(\overline{u})$ has a zero eigenvalue). Thus it is natural in a first step to omit this term. Consequently, we obtain the linear equation

$$\frac{d\xi_t}{dt} = \mathcal{D}F(\overline{u})\xi_t. \tag{34.3}$$

This equation is called the *linearization* of equation (34.1) around the static solution (equilibrium) $\overline{u}$. It is the key equation in the stability analysis. The analysis of the spectrum of (34.3) determines to a large degree the stability properties of the solution ξ_t. It is called the *linear stability theory.*

Recall that the key element in the stability analysis of (34.3) is the study of the linearized flow

$$\phi_t^{(0)} = e^{At} \quad \text{where } A = \mathcal{D}F(\overline{u}).$$

To transfer the results of the linear stability analysis to the nonlinear equation (34.1), we linearize the nonlinear flow ϕ_t at $\overline{u}$: write $u = \overline{u}+\xi$, then, using $\phi_t(\overline{u}) = \overline{u}$ and (32.2), we get

$$\phi_t(\overline{u}+\xi) - \overline{u} = \phi_t^{(0)}(\xi) + g(\xi),$$

where $g(\xi) = o(||\xi||)$.

For the discrete flow $\phi_n(u)$ define

$$\phi(\xi) := \phi_1(\overline{u}+\xi) - \overline{u} \quad \text{and } \Lambda := \phi_1^{(0)} = e^A.$$

Then we have

$$\phi(\xi) = \Lambda\xi + g(\xi), \tag{34.4}$$

where $g(\xi) = o(||\xi||)$.

Duhamel principle and integral equation for the flow. The Duhamel principle is formulated as follows. Let A be a bounded operator. Then the solution of the inhomogeneous initial value problem (with given initial condition ξ_0 and inhomogeneity f_t)

$$\frac{d\xi_t}{dt} = A\xi_t + f_t \quad \text{and} \quad \xi_t|_{t=0} = \xi_0 \tag{34.5}$$

can be expressed in terms of the *linear* flow $\phi_t^{(0)} = e^{At}$ (also called the linear propagator) as follows:

$$\xi_t = \phi_t^{(0)}(\xi_0) + \int_0^t \phi_{t-s}^{(0)}(f_s)ds. \tag{34.6}$$

On the other hand, if ξ_t is given by (34.6) and is differentiable, then it satisfies the initial value problem (34.5).

Applying the Duhamel principle to equation (34.2), we obtain

$$\xi_t = \phi_t^{(0)}(\xi_0) + \int_0^t \phi_{t-s}^{(0)}\left(R(\xi_s)\right) ds.$$

Expressing ξ_t in terms of the flow ϕ_t for (34.2), $\xi_t = \phi_t(\xi_0)$, and replacing ξ_0 by ξ, we find

$$\phi_t(\xi) = \phi_t^{(0)}(\xi) + G_t(\xi),$$

where G_t is the nonlinear part of the map ϕ_t:

$$G_t(\xi) = \int_0^t \phi_{t-s}^{(0)}\left(R(\phi_s(\xi))\right) ds.$$

Taking $t = 1$ in the previous equation, we obtain an equation corresponding to (34.4).

35. Nonlinear stability

The following theorem shows that stability properties of an equilibrium point can be determined by the stability properties of the corresponding linearized system.

Theorem. *Consider the dynamical system* $\dfrac{du_t}{dt} = F(u_t)$*, with an equilibrium point* $\overline{u}$*:* $F(\overline{u}) = 0$*. Assume* $F : X \to X$ *is* C^1*. Then,*

(i) if $\sigma(\mathcal{D}F(\overline{u})) \subset \mathbb{C}_-$*, then* $\overline{u}$ *is asymptotically stable,*

(ii) if $\sigma(\mathcal{D}F(\overline{u})) \cap \mathbb{C}_+ \neq \phi$*, then* $\overline{u}$ *is not asymptotically stable.*

Here, $\mathbb{C}_\pm$ denote the (open) right and left complex half–planes, respectively. Asymptotic stability of $\overline{u}$ means that $\forall \epsilon > 0\ \exists \delta > 0$ s.t. if the initial condition u_0 satisfies $||u_0 - \overline{u}|| < \delta$, then $||\phi_t(u_0) - \overline{u}|| < \epsilon$, $\forall t > 0$ and moreover, $\lim_{t\to\infty} ||\phi_t(u_0) - \overline{u}|| = 0$.

Proof. Without loss of generality, we assume $\overline{u} = 0$. We show first (i). In a first step, we transform the initial value problem (34.2) into an integral equation, using the Duhamel principle introduced in the previous paragraph. Setting $A = \mathcal{D}F(\overline{u})$, we obtain

$$\xi_t = e^{At}\xi_0 + \int_0^t e^{A(t-s)} R(\xi_s) ds. \tag{35.1}$$

In a second step, we estimate (35.1). First notice that by the definition of $R(\xi)$, $\forall \epsilon\ \exists \alpha$ s.t.

$$||R(\xi)|| \leq \epsilon ||\xi||, \quad \text{if } ||\xi|| \leq \alpha. \tag{35.2}$$

Take initially $||\xi_0|| < \alpha$, and set

$$T := \sup\{t : ||\xi_t|| \leq \alpha\}.$$

Equations (35.1) and (35.2) and the inequality $||e^{At}\xi|| \leq Ce^{\mu t}||\xi||$ (with $\mu = \frac{1}{2}\sup \mathrm{Re}\sigma(A) < 0$), proven before (c.f. (d) in the last proposition above), imply

$$||\xi_t|| \leq Ce^{\mu t}||\xi_0|| + C\int_0^t e^{\mu(t-s)}\epsilon||\xi_s|| ds,$$

provided $t \leq T$. Let

$$a := \sup_{0 \leq t \leq T} e^{-\frac{\mu}{2}t} ||\xi_t||.$$

Then the latter inequality implies

$$a \leq C||\xi_0|| + C\epsilon \sup_{0 \leq t \leq T} e^{-\frac{\mu}{2}t} \int_0^t e^{\mu(t-s)} e^{\frac{\mu}{2}s} \left(e^{-\frac{\mu}{2}s} ||\xi_s|| \right) ds,$$

so, since $\mu < 0$, this yields

$$a \leq C||\xi_0|| - 2C\epsilon a/\mu.$$

Now choose $\epsilon = (-4C/\mu)^{-1}$, then $a \leq C||\xi_0|| + \frac{1}{2}a$, thus $a \leq 2C||\xi_0||$. If we take the initial condition s.t. $||\xi_0|| \leq \frac{\alpha}{4C}$ with C as above, then $a \leq \alpha/2$ and therefore $T = \infty$ (in fact, if there were a $T < \infty$ s.t. $||\xi_T|| = \alpha$, then we would have $e^{-\mu T/2}\alpha = e^{-\mu T/2}||\xi_T|| \leq a \leq \alpha/2$, which is not possible since $e^{-\mu T/2} \geq 1$). Thus we have

$$||\xi_t|| \leq \alpha e^{\frac{\mu}{2}t}/2, \quad \text{provided } \ ||\xi_0|| \leq \tfrac{\alpha}{4C}.$$

This shows in particular that $||\xi_t|| \to 0$ exponentially fast, provided the initial condition is chosen sufficiently close to the equilibrium point $\overline{u} = 0$.

Now we show (ii), where we assume for simplicity that $\sigma(A) \cap \mathbb{C}_+$ contains an eigenvalue λ with corresponding eigenvector $\xi^{(\lambda)}$. We can take $||\xi^{(\lambda)}||$ as small as we wish. Consider then equation (35.1) with initial condition $\xi_0 = \xi^{(\lambda)}$. We have

$$\xi_t = e^{\lambda t}\xi_0 + \int_0^t e^{A(t-s)} R(\xi_s) ds, \tag{35.3}$$

and therefore

$$\begin{aligned} ||\xi_t|| &\geq e^{\text{Re}\lambda t}||\xi_0|| - \int_0^t e^{\text{Re}\lambda(t-s)}||R(\xi_s)||ds \\ &\geq e^{\text{Re}\lambda t}||\xi_0|| + \frac{1 - e^{\text{Re}\lambda t}}{\text{Re}\lambda} \sup_{0 \leq s \leq t} ||R(\xi_s)|| \\ &\geq e^{\text{Re}\lambda t}\left(\|\xi_0\| - \frac{1}{\text{Re}\lambda} \sup_{0 \leq s \leq t} \|R(\xi_s)\| \right) + \frac{1}{\text{Re}\lambda} \sup_{0 \leq s \leq t} \|R(\xi_s)\|. \end{aligned}$$

Recall that Re$\lambda > 0$ and $R(\xi) = o(\|\xi\|)$. If $||\xi_t||$ is very small, then the negative term is very small, so that first term, which grows, dominates. So $||\xi_t||$ grows until the second term in (35.3) balances the first term, if this ever occurs. Thus if ξ_0 is in a very small ball around $\overline{u} = 0$, then ξ_t leaves this ball after a while. ■

36. The stable manifold theorem

Consider a map ϕ on a Banach space X with a fixed point (equilibrium) $\overline{u}$, i.e. $\phi(\overline{u}) = \overline{u}$. We want to understand the behaviour of the orbits $\phi^n(u)$, starting at points u close to $\overline{u}$. Our first step is to establish the existence of a *stable manifold* $M_s = M_s(\overline{u})$ of $\overline{u}$, i.e. an invariant manifold for ϕ s.t. $\phi^n(u) \to \overline{u}$ for all $u \in M_s$. Often in infinite dimensional problems, the manifold M_s is of a finite codimension, which allows us to reduce the original problem to a finite dimensional one (or in some problems, by fixing a finite number of parameters, one can place u onto M_s). This will be explained more carefully later.

We begin with stating the assumptions suggested by our previous analysis:

(A) $\exists$ subspaces $V_s = V_s(\overline{u})$ and $V_{cu} = V_{cu}(\overline{u})$ s.t. $V_s \cap V_{cu} = \{0\}$ and $V_s + V_{cu} = X$,

(B) V_s and V_{cu} are invariant under $\Lambda := \mathcal{D}\phi(\overline{u})$ and $||\Lambda_s|| = p$, $||\Lambda_{cu}^{-1}|| = q^{-1}$, where $\Lambda_\# = \Lambda \restriction V_\#$, $\# = s, cu$ and $0 < p < q$, $p < 1$.

Under conditions (A) and (B), we have

Theorem. *There is a neighbourhood U of $\overline{u}$ and there is a local manifold $M_s = M_s(\overline{u}) \subset U$ s.t.*

(a) M_s is invariant under ϕ,

(b) V_s is tangent to M_s at $\overline{u}$,

(c) $\phi^n(u) \to \overline{u}$ as $n \to \infty$, $\forall u \in M_s$.

Remarks. 1) (a) and (b) hold even if the condition $p < 1$ is removed. 2) Later on, we will give a more complete picture of the flow.

Let us give the proof of the theorem in several steps.

1) Notation. $V_{s,r} = \{s \in V_s : ||s|| \leq r\}$ is the ball in V_s of radius r centered at 0. P_s and P_{cu} denote the projection operators onto the subspaces V_s and V_{cu}: $P_s^2 = P_s$, $\mathrm{Ran} P_s = V_s$ and similarly for P_{cu}, defined by the relation $P_s + P_{cu} = \mathbb{1}$, i.e. $P_s u = x$ and $P_{cu} u = y$, where $x \in V_s$ and $y \in V_{cu}$ satisfy $x + y = u$. Since V_s and V_{cu} are invariant under Λ, the projections P_s and P_{cu} commute with Λ: $[\Lambda, P_s] = 0$, $[\Lambda, P_{cu}] = 0$.

Exercise. Show the operators P_s and P_{cu} defined above are projections.

Redefining the flow ϕ, we can shift $\overline{u}$ to 0, so we assume to begin with that $\overline{u} = 0$.

2) Blow up. Instead of looking at ϕ on a small neighbourhood of $\overline{u} = 0$, we blow it up as $\phi_\epsilon(u) := \frac{1}{\epsilon}\phi(\epsilon u)$, then ϕ_ϵ is defined on a fixed neighbourhood of $\overline{u} = 0$, say the ball, X_2, of radius 2. In the proof below, we require ϵ to be sufficiently small. The neighbourhood U mentioned in the theorem is obtained by applying the inverse transform to X_2, i.e. $U = \epsilon X_2 = X_{2\epsilon}$ (or $U = \overline{u} + X_{2\epsilon}$). We drop from now on the subscript ϵ. Using that $\phi(\overline{u}) = \overline{u} = 0$, we expand the flow in ϵ:

$$\phi(u) = \mathcal{D}\phi(0)u + R(u),$$

where $R(u) = \frac{1}{\epsilon} o(\epsilon ||u||) = o_\epsilon(1) \to 0$ as $\epsilon \to 0$.

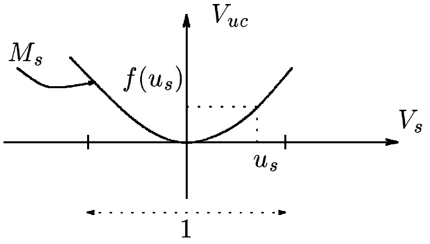
3) Lipschitz continuity. Recall that a map $f : X \to Y$ is called *Lipschitz continuous* iff

$$||f(x) - f(y)||_Y \leq L||x - y||_X \tag{36.1}$$

We set $\mathrm{Lip}(f) := \inf\{L : (36.1) \text{ holds}\}$.

Exercise. Show that $\mathrm{Lip}(g \circ h) \leq \mathrm{Lip}(g)\mathrm{Lip}(h)$. (Hint: use the chain rule, as in derivatives).

4) Ansatz for M_s. We look for M_s as a graph of a map $f : V_{s,1} \to V_{cu}$, i.e. $M = \mathrm{graph} f = \{u_s + f(u_s) : u_s \in V_{s,1}\}$.



5) Main steps of the proof.

(α) Using that M_s is invariant under ϕ, we derive an equation for f:

$$H(f) = f. \tag{36.2}$$

(β) We find a space B on which H is a contraction. Then (36.2) has a unique solution, so $M_s = \text{graph} f$!

(γ) We show that $\forall u \in M_s$, $\phi^n(u) \to \overline{u} = 0$.

Proof. (α) We have that $u \in M_s \Leftrightarrow \exists x \in V_{s,1}$ s.t. $u = x + f(x)$. Since $P_s u = x$, we have

$$u \in M_s \Longleftrightarrow u = P_s u + f(P_s u) =: P_f(u).$$

Now M_s is invariant under the map $\phi \Leftrightarrow \{u \in M_s \Rightarrow \phi(u) \in M_s\} \Leftrightarrow \{u = P_f(u) \Rightarrow \phi(u) = P_f(\phi(u))\}$. A sufficient condition for the last implication to hold is

$$\phi \circ P_f = P_f \circ \phi \circ P_f, \tag{36.3}$$

where both sides are defined as functions on $V_{s,1}$. Equation (36.3) is our equation for f. We transform it to a convenient form. First, project it onto the space V_{cu}, i.e. apply the projector P_{cu} to (36.3). Since we have

$$P_{cu} P_f = P_{cu} (P_s + f \circ P_s) = f \circ P_s, \tag{36.4}$$

we obtain

$$P_{cu} \phi \circ P_f = f \circ P_s \phi \circ P_f. \tag{36.5}$$

Now expand $\phi \circ P_f$ on the l.h.s. as

$$\phi(P_f(x)) = \phi(x + f(x)) = \mathcal{D}\phi(0)(x + f(x)) + R(P_f(x)).$$

Since P_{cu} commutes with $\mathcal{D}\phi(0)$ and since $P_{cu}(x + f(x)) = f(x)$ (this is the same as (36.4)), we have

$$P_{cu}\phi(P_f(x)) = \Lambda_{cu} f(x) + P_{cu} R(P_f(x)),$$

where, recall, $\Lambda_{cu} = \Lambda \restriction V_{cu}$, with $\Lambda := \mathcal{D}\phi(0)$. Substituting this into (36.5) and inverting Λ_{cu} (recall that $||\Lambda_{cu}^{-1}|| = q^{-1} < \infty$ by condition (B)), we obtain

$$f = H(f), \quad \text{where} \quad H(f) := \Lambda_{cu}^{-1} \left[f \circ P_s \phi \circ P_f - P_{cu} R \circ P_f \right]. \tag{36.6}$$

This is our desired equation.

(β) Now we choose a Banach space on which we define the map $H(f)$. This is a crucial step. For the moment we will not worry about smoothness and take

$$B := \{f : V_{s,1} \rightarrow V_{cu} | f \text{ is Lipschitz }, f(0) = 0, \text{ and } \mathrm{Lip}(f) \leq 1\}.$$

We take an unusual norm on B:

$$|||f||| := \sup_{x \neq 0, x \in V_{s,1}} \left(\frac{||f(x)||}{||x||} \right). \tag{36.7}$$

However, one can prove

Lemma. *B is complete in the norm (36.7).*

Observe that B is not a vector space and therefore in not a Banach space. It is a "nonlinear space". A proof of this lemma is simple but slightly tedious. We omit it here and refer for it to [MMcC].

Now we turn to the technical statement:

Proposition. *Let ϵ be sufficiently small. Then H maps B into itself and is a contraction.*

Proof. First, we show that the map H is well defined on B. To this end, we have to show that $v := v_f := P_s\phi \circ P_f$ maps $V_{s,1}$ into itself. Expanding ϕ around 0, we find

$$P_s\phi(P_f(x)) = \Lambda_s x + P_s R(P_f(x)), \tag{36.8}$$

where we recall $\Lambda_s = \mathcal{D}\phi(0) \restriction V_s$ and where we have used that P_s commutes with Λ, and $P_s P_f = P_s$. This expansion implies

$$||P_s\phi(P_f(x))|| \leq (p + o_\epsilon(1))||x||, \tag{36.9}$$

where $p := ||\Lambda_s||$. Hence for $p < 1$ and ϵ sufficiently small (so that $p + o_\epsilon(1) \leq 1$), $P_s\phi \circ P_f$ maps $V_{s,1}$ into itself.

The relation $H(f)(0) = 0$ follows from the definition of $H(f)$ and the relations $f(0) = 0, \phi(0) = 0$ and $R(0) = 0$. In order to see that $H(f)$ is Lipschitz, it is convenient to write it as

$$H(f)(x) = \mathcal{H}(x, f(x), f(v(x))), \tag{36.10}$$

where $v = P_s\phi \circ P_f$ and

$$\mathcal{H}(x, y, z) = \Lambda_{cu}^{-1}\left[z - P_{cu}R(x + y)\right].$$

Now since R is Lipschitz, then so is $\mathcal{H}$ in all its variables. Furthermore, since ϕ is Lipschitz, then so is v. Hence $H(f)(x)$ is Lipschitz (in x) as a composition of several Lipschitz functions.

Now we estimate $\text{Lip}(H(f))$. Denoting by $\text{Lip}_x(\mathcal{H})$ the Lipschitz constant of $\mathcal{H}$ in the variable x and similarly for the other variables, we obtain by the chain rule

$$\text{Lip}(H(f)) \leq \text{Lip}_x(\mathcal{H}) + \text{Lip}_y(\mathcal{H})\text{Lip}(f) + \text{Lip}_z(\mathcal{H})\text{Lip}(f)\text{Lip}(v). \tag{36.11}$$

From equation (36.8), we obtain:

$$\text{Lip} v \leq p + o_\epsilon(1).$$

Furthermore, due to the explicit formula for $\mathcal{H}$ and condition (B), we have $\text{Lip}_x(\mathcal{H}) = o_\epsilon(1)$, $\text{Lip}_y(\mathcal{H}) = o_\epsilon(1)$ and $\text{Lip}_z(\mathcal{H}) \leq q^{-1}$, where

$q := ||\Lambda_{cu}^{-1}||^{-1}$.

Combining the last two relations with equation (36.11), we obtain

$$\mathrm{Lip}(H(f)) \leq \left(\frac{p}{q} + o_\epsilon(1)\right) \mathrm{Lip}(f) + o_\epsilon(1).$$

Thus, since $p/q < 1$, we can arrange that $\mathrm{Lip}(H(f)) \leq \mathrm{Lip}(f) \leq 1$.

Finally, we show that $H : f \mapsto H(f)$ is a contraction. In what follows, we omit the argument x and write

$$H(f) = \mathcal{H}(f, f(v))$$

instead of (36.10). Write

$$H(f_1) - H(f_2) = \mathcal{H}(f_1, f_1(v_1)) - \mathcal{H}(f_2, f_2(v_2)),$$

where $v_i = v_{f_i}$. We telescope the r.h.s. as

$$\begin{aligned} H(f_1) - H(f_2) \quad = \quad & \mathcal{H}(f_1, f_1(v_1)) - \mathcal{H}(f_2, f_1(v_1)) + \mathcal{H}(f_2, f_1(v_1)) \\ & -\mathcal{H}(f_2, f_2(v_1)) + \mathcal{H}(f_2, f_2(v_1)) - \mathcal{H}(f_2, f_2(v_2)). \end{aligned}$$

We now estimate separately each of the differences on the r.h.s.:

$$\begin{aligned} |||\mathcal{H}(f_1, f_1(v_1)) - \mathcal{H}(f_2, f_1(v_1))||| \quad & \leq \quad \mathrm{Lip}_y(\mathcal{H})|||f_1 - f_2|||, \\ |||\mathcal{H}(f_2, f_1(v_1)) - \mathcal{H}(f_2, f_2(v_1))||| \quad & \leq \quad \mathrm{Lip}_z(\mathcal{H})|||f_1 - f_2||| \cdot |||v_1|||, \\ |||\mathcal{H}(f_2, f_2(v_1)) - \mathcal{H}(f_2, f_2(v_2))||| \quad & \leq \quad \mathrm{Lip}_z(\mathcal{H})|||f_2(v_1) - f_2(v_2)||| \\ & \leq \quad \mathrm{Lip}_z(\mathcal{H})\mathrm{Lip}(f)|||v_1 - v_2|||. \end{aligned}$$

We use the following facts:

$$\begin{aligned} \mathrm{Lip}(f) \quad & \leq \quad 1, \\ \mathrm{Lip}_y(\mathcal{H}) \quad & = \quad o_\epsilon(1), \\ |||v_1||| \quad & \leq \quad p + o_\epsilon(1), \\ \mathrm{Lip}_z(\mathcal{H}) \quad & \leq \quad q^{-1}, \\ |||v_1 - v_2||| \quad & = \quad o_\epsilon(1)|||f_1 - f_2||| \end{aligned}$$

to conclude that

$$|||H(f_1) - H(f_2)||| \leq \left(\frac{p}{q} + \left(1 + \frac{1}{q}\right) o_\epsilon(1)\right) |||f_1 - f_2|||.$$

Since $p/q < 1$, we can choose ϵ so small that $H(f)$ is a contraction. ■

Corollary. *There is a unique solution $f \in B$ to (36.6).*

One can show further that $f \in C^k$ and $\mathcal{D}f(0) = 0$. This means that V_s is the tangent plane to $M_s = \text{graph}\, f$ at 0.

(γ) Show that $\forall u \in M_s$, $\phi^n(u) \to 0$.
Let $u \in M_s \leftrightarrow u = x + f(x)$, $x = P_s u$. Expand $\phi(u) = \mathcal{D}\phi(0)u + R(u)$. Commuting P_s through $\mathcal{D}\phi(0)$, we obtain for $x^{(1)} := P_s\phi(u)$:

$$x^{(1)} = P_s\phi(x + f(x)) = \mathcal{D}\phi(0)x + P_s R(x + f(x)).$$

Hence the inequality $||f(x)|| \le |||f|||\,||x||$ implies $||x^{(1)}|| \le \delta ||x||$, where $\delta := p + o_\epsilon(1)(1 + |||f|||)$. Since $p < 1$ by assumption, take ϵ so small that $\delta < 1$. Let $u^{(n)} := \phi^n(u)$. Then for $x^{(n)} = P_s u^{(n)}$, we obtain

$$||x^{(n)}|| \le \delta^n ||x|| \to 0.$$

Now since $u^{(n)} \in M_s$, we can write $u^{(n)} = x^{(n)} + f(x^{(n)})$, and so

$$||u^{(n)}|| \le ||x^{(n)}|| + |||f|||\,||x^{(n)}||,$$

which shows that $||u^{(n)}|| \to 0$. ■

We proved the stable manifold theorem modulo showing that f is as smooth as ϕ and that $\mathcal{D}\phi(0) = 0$. For a proof of the latter statements, see [MMcC].

Now we refine the theorem above. Assume

(C) V_{cu} is the sum of two subspaces $V_{cu} = V_c + V_u$, $V_c \cap V_u = \{0\}$, both invariant under ϕ, and

$$||\Lambda_{sc}|| \cdot ||\Lambda_u^{-1}|| < 1 \tag{36.12}$$

here $\Lambda_{sc} = \Lambda \restriction V_{sc}$, $V_{sc} = V_s + V_c$.

Theorem. *Under assumptions (A)–(C), there is a neighbourhood U of $\overline{u}$ and there are local manifolds $M_{sc}, M_c \subset U$ s.t.*

(a) V_{sc} *and* V_c *are tangent to* M_{sc} *and* M_c *at* $\overline{u}$*, respectively,*

(b) $\phi^n(u) \to M_c$ *as* $n \to \infty$*,* $\forall u \in M_{sc}$*.*

$M_s = M_s(\overline{u}), M_c = M_c(\overline{u}), M_{sc} = M_{sc}(\overline{u})$ are called respectively the (local) stable, central, stable-central manifolds for the fixed point $\overline{u}$.

Proof. The construction of M_{sc} is similar to that of M_s; we use condition (36.12) instead of the conditions in (B) and have to deal with an extra technical problem, namely that $v := P_{sc}\phi \circ P_{f,sc}$ does not map $V_{sc,1}$ into itself. We do not address this problem here (see [L] for details).

The construction of M_c as well as the proof of convergence contains some essential differences. We sketch this construction, leaving out some details that are similar to those above. We start in the same way as with M_s: look for M_c as $M_c = \text{graph} f$, for some $f : V_{c,1} \to V_{su}$, where $V_{su} = V_s + V_u$. Let P_{su} be the projector onto V_{su} (defined as P_s, etc., above) and $\Lambda_{su} = \Lambda \upharpoonright V_{su}$. Define $P_f := P_c + f \circ P_c$ and derive the following equality for f:

$$P_{su}\phi \circ P_f = f \circ P_c\phi \circ P_f. \tag{36.13}$$

Now comes the difference: unlike Λ_u, Λ_{su} is *not* invertible, so we cannot extract f from the l.h.s. We need to proceed differently. Break up (36.13) into two equations:

$$\begin{aligned} P_u\phi \circ P_f &= P_u f \circ P_c\phi \circ P_f, & (36.14)\\ P_s\phi \circ P_f &= P_s f \circ P_c\phi \circ P_f. & (36.15) \end{aligned}$$

The first equation is similar to the one for M_s and can be solved for $P_u f$:

$$P_u f = \Lambda_u^{-1}\left[P_u f \circ P_c\phi \circ P_f - P_u R \circ P_f\right]. \tag{36.16}$$

We cannot do the same with equation (36.15), since Λ_s is not invertible! So instead of the l.h.s., we resolve the r.h.s.: let

$$g_f := P_c\phi \circ P_f : V_{c,1} \to V_{c,1}$$

and show by the implicit function theorem that g_f is invertible. Then (36.15) can be rewritten as

$$P_s f = P_s\phi \circ P_f \circ g_f^{-1}. \tag{36.17}$$

Adding equation (36.16) to (36.17) and using that $P_u f + P_s f = f$, we arrive at a fixed point equation for f: $f = H(f)$.

The proof of the convergence $\phi^n(u) \to M_c$ $\forall u \in M_{sc}$ is trickier. Let $u^{(n)} = \phi^n(u)$, $w^{(n)} = u^{(n)} - P_f(u^{(n)})$ and $u^{(0)} = u$. We want to show that $w^{(n)} \to 0$. Then, since $P_f(u^{(n)}) \in M_c$, we would have

$$\text{dist}(u^{(n)}, M_c) \leq ||u^{(n)} - P_f(u^{(n)})|| \to 0,$$

which is what we want. As before, it suffices to prove a one step contraction. To simplify notation, we write $u' = \phi(u)$, $w := u - P_f(u)$ and

$$w' := u' - P_f(u'). \tag{36.18}$$

We want to show

$$w' = \Lambda w + R(u),\ \ R(u) = O(||u||^2) \text{ and } w_u = O(w_s). \tag{36.19}$$

Then $||w_s'|| \leq \delta ||w_s||$, $\delta = p + o_\epsilon(1) < 1$, $||w_u|| \leq C||w_s||$. So $||w_s^{(n)}|| \leq \delta^n ||w_s||$, $||w_u^{(n)}|| \leq C||w_s^{(n)}|| \leq C\delta^n||w_s||$, and therefore

$$||w^{(n)}|| \leq C_1 \delta^n ||w_s|| \to 0.$$

Proof of (36.19). Equation (36.18) implies

$$w' = u' - P_c u' - f(P_c u') = P_{su} u' - f(P_c u'). \tag{36.20}$$

On the other hand, equation (36.13) gives

$$P_{su}\phi(P_f u) = f(P_c \phi(P_f u)).$$

Subtracting this from (36.20), we obtain

$$w' = P_{su}\left(\phi(u) - \phi(P_f u)\right) - \left(f(P_c\phi(u)) - f(P_c\phi(P_f u))\right). \tag{36.21}$$

Next, using that $P_c w = 0$ and $u = P_f u + w$, we expand

$$\phi(u) - \phi(P_f u) = \Lambda w + R_1(u), \tag{36.22}$$

where $R_1(u) = R(u) - R(P_f(u))$ and

$$f(P_c \phi(u)) - f(P_c \phi(P_f u)) = \mathcal{D} f(\hat{u}) P_c R_1(u), \tag{36.23}$$

where $\hat{u}$ is some point on the interval joining u and $P_f u$. Equations (36.21)–(36.23) give $w' = \Lambda w + R(u)$ with $R(u) = O(||u||^2)$. Thus the first part in (36.19) is proved.

In what follows, we use the notation $u_c = P_c u$ etc. We now show that $w_u = O(w_s)$. The fact that $M_c \subset M_{sc}$ and the definitions

$$M_c = \{u | u_{su} = f_c(u_c)\} = \{u | u_s = P_s f_c(u_c), u_u = P_u f_c(u_c)\},$$

where f_c (denoted simply as f above) defines the central manifold ($M_c = \text{graph} f_c$) and

$$M_{sc} = \{u | u_u = f_{sc}(u_{sc})\},$$

where $M_{sc} = \text{graph} f_{sc}$ imply the following relation between f_c and f_{sc}:

$$P_u f_c(u_c) = f_{sc}\left(P_s f_c(u_c) + u_c\right), \qquad \forall u_c \in V_c.$$

On the other hand, from the definition of w we have

$$w_s = u_s - P_s f_c(u_c) \quad \text{and} \quad w_u = u_u - P_u f_c(u_c).$$

The last three relations (and the fact that we take $u \in M_{sc}$, so that $u_u = f_{sc}(u_{sc})$) give:

$$w_u = f_{sc}(u_{sc}) - P_u f_c(u_c) = f_{sc}(u_{sc}) - f_{sc}(P_s f_c(u_c) + u_c).$$

Applying the mean value theorem and using that

$$u_{sc} - P_s f_c(u_c) - u_c = u_s - P_s f_c(u_c) = w_s,$$

we finally obtain $w_u = \mathcal{D} f_{sc}(\hat{u}) w_s$, where $\hat{u}$ is a point in the segment between u_{sc} and $P_s f_c(u_c) + u_c$. This shows the last part in (36.19). ■

Examples. 1) Consider the nonlinear diffusion equation

$$\frac{du_t}{dt} = \Delta u_t + g(u_t).$$

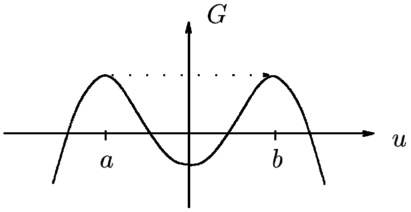
This is a dynamical system with the vector field

$$F(u) := \Delta u + g(u).$$

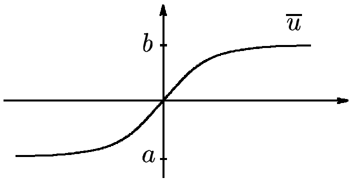
Equilibrium solutions satisfy the equation

$$\Delta u + g(u) = 0. \tag{36.24}$$

Let G be the antiderivative of g, i.e. $G' = g$. We consider functions g s.t. their antiderivatives are of the form



Then equation (36.24) has a solution $\overline{u}$ depending on one coordinate only, say x_1, which correspond to a Newtonian particle in the potential G moving from the top of the left hill to the top of the right hill:



Indeed, (36.24) can be rewritten as Newton's equation

$$u'' = -G'(u),$$

in which x_1 is thought of as the time variable. The solution $\overline{u}$ is called the kink. Note also that $\overline{u}$ is a separatrix for $u'' = -g(u)$, considered as a dynamical system: $u' = v$ and $v' = -g(u)$:

The Fréchet derivative of F at $\overline{u}$ is

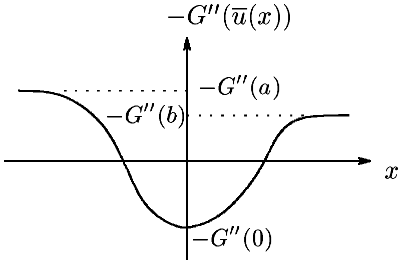
$$\mathcal{D}F(\overline{u}) = \Delta + g'(\overline{u}). \tag{36.25}$$

Our task is to find the spectrum of $T := \mathcal{D}F(\overline{u})$ (recall that $\mathcal{D}\phi_t(\overline{u}) = e^{t\mathcal{D}F(\overline{u})}$, so for $\phi(u) = \phi_1(u)$, we have $\Lambda := \mathcal{D}\phi(\overline{u}) = e^T$).

The operator (36.25) can be rewritten as

$$\Delta + g'(\overline{u}) = -(-\Delta - G''(\overline{u})) =: -H,$$

where H is a Schrödinger operator with the potential $-G''(\overline{u}(x))$ that looks like



It is well known that a Schrödinger operator with such a potential has continuous spectrum in $[\min\{-G''(a), -G''(b)\}, \infty)$, and possibly eigenvalues in the interval $[-G''(0), \min\{-G''(a), -G''(b)\})$.

It is not difficult to see that 0 is an eigenvalue of H. Indeed, since equation (36.24) is invariant under coordinate translation, we know that

$$\mathcal{D}F(\overline{u})\xi = 0, \quad \text{where } \xi = \frac{d\overline{u}}{dx}.$$

Hence ξ is an eigenvector of T with eigenvalue zero. From the shape of $\overline{u}$, it is also clear that $\xi(x) > 0$, which allows us to conclude that 0 is the lowest eigenvalue and it is simple.

Chapter VI. Stochastic Analysis

37. Basic concepts of probability theory

Probability theory deals with the description of events when only partial information is available. On first sight the mathematical framework of probability theory defines it as a part of real analysis. The following table identifies main concepts of probability theory with the corresponding concepts of analysis.

probabilistic notion	analytic notion
probability space $(\Omega, \mathcal{B}, P)$	measure space $(X, \mathcal{M}, \mu)$
event	measurable set
random variable X	measurable function f
expectation of X, $E(X)$	integral of f, $\int f d\mu$
random variable with finite p–th moment	L^p–function
convergence in probability	convergence in measure
almost sure(ly), a.s.	almost every(where), a.e.
characteristic function	Fourier transform

However, probability theory has its own distinctive viewpoint which makes it quite different in philosophy and methods from other parts of analysis. To immerse ourselves into this viewpoint, we adopt entirely the probabilist's terminology and notation. Thus our point of departure is the probability space $(\Omega, \mathcal{B}, P)$, where Ω is a given set, called the *sample space*, $\mathcal{B}$ is a σ–algebra on Ω (elements of $\mathcal{B}$ are called *events*) and P is a *probability measure* on $(\Omega, \mathcal{B})$, i.e. a finite measure, normalized as $P(\Omega) = 1$. The number $P(E)$ is called the *probability of the event* $E \in \mathcal{B}$ (i.e. the probability that the event occurs).

Furthermore, a *random variable* X is a measurable real function $X : \Omega \to \mathbb{R}$. An important class of events is given by

$$\{X > \alpha\},$$

where $\alpha \in \mathbb{R}$ and where we used the typical probabilistic notation for the measurable sets

$$\{X > \alpha\} := \{\omega \in \Omega | X(\omega) > \alpha\}.$$

Example: n tosses of a coin. Here, $\omega = (\omega_1, \dots, \omega_n) \in \Omega$, where $\omega_j \in \{\mathrm{H}, \mathrm{T}\}$ (heads or tails).

Exercise. What is $\mathcal{B}$ in this example ?

To continue with the probabilistic terminology, the *expectation of* X (or the expected value or mean value of X) is defined by

$$E(X) := \int X dP.$$

The *variance* and *standard deviation* are defined respectively as

$$\sigma^2(X) := E\left((X - E(X))^2\right)$$

and

$$\sigma(X) := \sqrt{\sigma^2(X)}.$$

The next important notion is that of the *distribution of* X, P_X:

$$P_X(B) := P\left(X^{-1}(B)\right), \quad \forall B \in B_{\mathbb{R}}, \tag{37.1}$$

where we recall that $B_{\mathbb{R}}$ is the Borel algebra of $\mathbb{R}$, i.e. the algebra of subsets of $\mathbb{R}$ generated by open (or semiopen or closed) intervals. The distribution P_X is a measure on $(\mathbb{R}, B_{\mathbb{R}})$ and $P_X(\mathbb{R}) = 1$.

If $P_X = P_Y$ we say that the random variables X and Y are identically distributed.

Exercises. Let $f : \mathbb{R} \to \mathbb{R}$ be Borel measurable. Show that

(i) $(f \circ X)^{-1} = X^{-1} \circ f^{-1}$,

(ii) $P_X(f^{-1}(B)) = P_Y(B)$,

where $Y = f \circ X$ is a new random variable.

Example. *Gaussian random variables.*
A random variable with *Gaussian* (or *normal*) probability distribution

$$d\nu_\mu^{\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma} dx \tag{37.2}$$

is called a *Gaussian random variable with mean μ and variance σ^2* and denoted by $N_\mu^{\sigma^2}$. In other words, for a Gaussian random variable X, we have $P_X = \nu_\mu^{\sigma^2}$ in the sense that $P_X(B) = \int_B d\nu_\mu^{\sigma^2}(x)$ $\forall B \in B_{\mathbb{R}}$ (Borel σ-algebra of $\mathbb{R}$).

Let us discuss the notion of probability distribution in more detail. There is an easy way to produce new measures. Let $(\Omega', \mathcal{B}')$ be another measurable space and $\varphi : \Omega \to \Omega'$ be a $(\mathcal{B}, \mathcal{B}')$–measurable map, i.e. $\varphi^{-1}(E') \in \mathcal{B}$ whenever $E' \in \mathcal{B}'$. Then φ induces a probability measure P_φ on $(\Omega', \mathcal{B}')$:

$$P_\varphi(E') := P\left(\varphi^{-1}(E')\right) \quad \forall E' \in \mathcal{B}'.$$

Exercise. Show that P_φ is a probability measure on $\mathcal{B}'$ (Hint: Use that φ^{-1} commutes with unions and intersections).

The following result connects integrals with respect to P and P_φ:

Proposition. *If $f : \Omega' \to \mathbb{R}$ is a measurable function, then*

$$\int_{\Omega'} f dP_\varphi = \int_\Omega f \circ \varphi dP.$$

Proof. Check first that it is true for characteristic functions ("indicator functions"):

$$\begin{aligned} \int_{\Omega'} \chi_{E'} dP_\varphi &:= P_\varphi(E') \\ &:= P(\varphi^{-1}(E')) \\ &:= \int_\Omega \chi_{\varphi^{-1}(E')} dP \\ &= \int_\Omega \chi_{E'} \circ \varphi dP, \end{aligned}$$

where we have used that $\chi_{\varphi^{-1}(E')} = \chi_{E'} \circ \varphi$. Now extend this equality first to simple functions and then to measurable functions. ■

If in the last proposition we take $\Omega' = \mathbb{R}$ and $\mathcal{B}' = B_{\mathbb{R}}$ (the Borel σ–algebra), $f = id$ on $\mathbb{R}$ and $\varphi = X$, a random variable, then we obtain

$$E(X) = \int x dP_X(x)$$

and

$$\sigma^2(X) = \int (x - E(X))^2 dP_X(x).$$

Now we consider several random variables $X_1, \dots, X_n$. We can form the measurable vector function $X = (X_1, \dots, X_n) : \Omega \to \mathbb{R}^n$, which can also be considered as a vector random variable. Then the above construction (see (37.1)) defines the measure $P_{(X_1,\dots,X_n)}$ on $(\mathbb{R}^n, B_{\mathbb{R}^n})$, which is called the *joint distribution of* $X_1, \dots, X_n$.

Exercises. 1) Show that

$$(X_1, \dots X_n)^{-1}(E_1 \times \dots \times E_n) = \cap_1^n X_j^{-1}(E_j).$$

2) Using 1) show

$$P_{(X_1,\dots,X_n)}(E_1 \times \dots \times E_n) = P\left(\cap_1^n X_j^{-1}(E_j)\right).$$

Taking $f(x_1, \dots, x_n) = \sum_1^n x_j$, we derive from the proposition above that for instance

$$E\left(\sum_1^n X_j\right) = \int \sum_1^n x_j \; dP_{(X_1,\dots,X_n)}(x_1, \dots, x_n).$$

Here, we have chosen $\Omega' = \mathbb{R}^n$, $\varphi = (X_1, \dots, X_n) : \Omega = \mathbb{R} \to \mathbb{R}^n = \Omega'$.

Example. *Family of Gaussian random variables.*
A family $\{X_\alpha\}$ of random variables labelled by $\alpha \in I$ (I some index set) is called Gaussian iff for any n and any $\alpha_1, \dots, \alpha_n \in I$, $P_{(X_{\alpha_1},\dots,X_{\alpha_n})}$ is Gaussian:

$$dP_{(X_{\alpha_1},\dots,X_{\alpha_n})}(x_1, \dots, x_n) = Ne^{-\langle x-\mu,(2A)^{-1}(x-\mu)\rangle},$$

where $N = (\det(2\pi A))^{-n/2}$ is the normalization constant, A is a positive $n \times n$ matrix and $x = (x_1, \dots, x_n)$, $\mu = (\mu_1, \dots, \mu_n)$. $\langle \cdot, \cdot \rangle$ denotes the standard scalar product in $\mathbb{R}^n$.

38. Independence

This is the first distinctive concept of probability theory. First we address the question: what is the probability of an event F given that an event E has taken place?

The mathematical expression for this probability is

$$P_E(F) := \frac{P(E \cap F)}{P(E)}, \tag{38.1}$$

i.e. the probability of the joint event $E \cap F$ normalized by the probability of E (since we know that the event E occured). The quantity (38.1) is called the *conditional probability* (on E). We say that the events E and F are *independent* iff the conditional probability is independent of E, i.e. iff

$$P_E(F) = P(F),$$

which is equivalent to

$$P(E \cap F) = P(E)P(F).$$

Similarly, we say that a collection $\{E_\alpha\}$, $\alpha \in I$ is independent iff

$$P(\cap_1^n E_{\alpha_j}) = \prod_1^n P(E_{\alpha_j})$$

for any distinct $\alpha_1, \dots, \alpha_n \in I$ and any $n \in \mathbb{N}$. Now we say that the *random variables* $\{X_\alpha\}$ *are independent* iff the events

$$\{X_\alpha \in B_\alpha\} = \{X_\alpha^{-1}(B_\alpha)\}$$

are independent for all Borel sets $B_\alpha \subset \mathbb{R}$.

We now show that $X_1, \dots, X_n$ are independent iff

$$P_{(X_1,\dots,X_n)} = \prod_1^n P_{X_j}.$$

Indeed, by the definition of $P_{(X_1,\dots,X_n)}$, we have

$$P_{(X_1,\dots,X_n)}(E_1 \times \dots \times E_n) = P(\cap_1^n X_j^{-1}(E_j)).$$

Thus if $X_1, \dots , X_n$ are independent, then so are $X_1^{-1}(E_1), \dots , X_n^{-1}(E_n)$ and therefore

$$P(\cap_1^n X_j^{-1}(E_j)) = \prod_1^n P(X_j^{-1}(E_j)) = \prod_1^n P_{X_j}(E_j).$$

This proves the implication ⇒ in the above equivalence.

Exercise. Prove the ⇐ direction (take for simplicity $n = 2$).

(Solution:

$$\begin{aligned} P_{(X_1,X_2)}(B_1 \times B_2) &= P_{X_1}(B_1)P_{X_2}(B_2) \\ \Leftrightarrow P(X_1^{-1}(B_1) \cap X_2^{-1}(B_2)) &= P(X_1^{-1}(B_1))P(X_2^{-1}(B_2)), \end{aligned}$$

which is equivalent to saying that $X_1^{-1}(B_1)$ and $X_2^{-1}(B_2)$ are independent which is again equivalent to the fact that X_1 and X_2 are independent.)

We present some properties of independent random variables. Let $\{X_i\}_1^n$ be independent random variables. Then

(i) $\{f_i(X_i)\}_1^n$ are independent random variables for any Borel measurable functions f_j,

(ii) $P_{X_1+\dots+X_n} = P_{X_1} * \dots * P_{X_n}$,

(iii) $E(\prod_1^n X_j) = \prod_1^n E(X_j)$, provided $X_j \in L^1$ $\forall j$,

(iv) $\sigma^2(\sum_1^n X_j) = \sum_1^n \sigma^2(X_j)$, provided $X_j \in L^2$ $\forall j$.

Remark. Property (i) can be formulated in words as follows: functions of independent random variables are independent random variables. It can be considerably generalized.

We prove statements (i) and (iii) and leave (ii) and (iv) as exercises (see [F], paragraphs 9.4 and 9.6).

Proof of (i). Let $Y_i = f_i(X_i)$. If B_i are Borel subsets, then by an exercise above we know that

$$\begin{aligned} &(Y_1, \dots , Y_n)^{-1}(B_1 \times \dots \times B_n) \\ &= \cap_1^n Y_i^{-1}(B_i) \\ &= \cap_1^n X_i^{-1}(f_i^{-1}(B_i)) \\ &= (X_1, \dots , X_n)^{-1}(f_1^{-1}(B_1) \times \dots \times f_n^{-1}(B_n)). \end{aligned}$$

Hence

$$P_{(Y_1,\dots,Y_n)}(B_1\times\dots\times B_n)=P_{(X_1,\dots,X_n)}(f_1^{-1}(B_1)\times\dots\times f_n^{-1}(B_n)).$$

By the independence of X_j, the r.h.s. equals $\prod_1^n P_{X_j}(f_j^{-1}(B_j))$. Now an exercise above shows that the latter quantity equals $\prod_1^n P_{Y_j}(B_j)$ and therefore

$$P_{(Y_1,\dots,Y_n)}(B_1\times\dots\times B_n)=\prod_1^n P_{Y_j}(B_j),$$

i.e. $Y_1,\dots,Y_n$ are independent.

Proof of (iii). First, let $f(t_1,\dots,t_n)=\prod_1^n |t_j|$. Then

$$f(X_1,\dots,X_n)=\prod_1^n |X_j|$$

and therefore

$$E(\prod_1^n |X_j|)=\int f dP_{(X_1,\dots,X_n)}.$$

By independence, $dP_{(X_1,\dots,X_n)}=\prod_1^n dP_{X_j}$ so by Fubini's theorem

$$\begin{aligned}E(\prod_1^n |X_j|)&=\prod_1^n\int |x_j|dP_{X_j}(x_j)\\&=\prod_1^n E(|X_j|).\end{aligned}$$

This proves that $\prod_1^n X_j\in L^1$. Removing the absolute values in the above argument shows (iii). ■

39. The law of large numbers

The law of large numbers says essentially that the average of many independent trials in the game of chance is approximately equal to the expectation for each trial.

We give a weak form of the law of large numbers. For other forms of this law, see e.g. [F], paragraphs 9.12 and 9.13. Recall that we say that $X_n \to X$ in probability iff $\forall \epsilon > 0$: $P(|X_n - X| \geq \epsilon) \to 0$ as $n \to \infty$ (i.e. X_n converges to X in measure). The key to the proof of the weak law of large numbers is

Chebyshev's inequality. *Let $X \in L^p$, $0 < p < \infty$. Then for any $\alpha > 0$:*

$$P(|X| > \alpha) \leq \left(\frac{||X||_p}{\alpha}\right)^p .$$

Proof. Let $E_\alpha = \{|X| > \alpha\}$. Then

$$\begin{aligned} ||X||_p^p &= \int |X|^p dP \\ &\geq \int_{E_\alpha} |X|^p dP \\ &\geq \alpha^p \int_{E_\alpha} dP \\ &= \alpha^p P(|X| > \alpha). \blacksquare \end{aligned}$$

Theorem (The weak law of large numbers). *Let $\{X_j\}_1^\infty$ be a sequence of independent L^2 random variables with means μ_i and variances σ_i^2. If $\lim_{n\to\infty} n^{-2} \sum_1^n \sigma_i^2 = 0$, then as $n \to \infty$:*

$$\frac{1}{n}\sum_1^n X_j - \frac{1}{n}\sum_1^n \mu_i \to 0 \quad \textit{in probability.}$$

Proof. The random variable $S_n := n^{-1} \sum_1^n (X_j - \mu_j)$ has mean zero and variance $n^{-2} \sum_1^n \sigma_j^2$ (see (iv) in the last section). Therefore

$$||S_n||_{L^2}^2 = \frac{1}{n^2} \sum_1^n \sigma_j^2$$

and so by Chebyshev's inequality we have

$$P(|S_n| > \epsilon) \leq \frac{1}{n^2\epsilon^2}\sum_1^n \sigma_j^2.$$

The r.h.s. vanishes as $n \to \infty$ by our assumption in the theorem. ■

40. The central limit theorem

Let $\{X_j\}_1^\infty$ be i.i.d.r.v. (independent identically distributed random variables) with mean zero and variance σ^2. By the law of large numbers,

$$\frac{1}{n}\sum_1^n X_j \to 0,$$

but by property (iv) above,

$$\sigma^2\left(\frac{1}{\sqrt{n}}\sum_1^n X_j\right) = \sigma^2.$$

The question is: does $\frac{1}{\sqrt{n}}\sum_1^n X_j$ converge to some random variable as $n \to \infty$? If yes, then what is this limit random variable? An answer to this question is given in the

Central limit theorem. *Let $\{X_j\}_1^\infty$ be a sequence of independent identically distributed L^2 random variables with mean μ and variance σ^2. Then as $n \to \infty$, we have*

$$\frac{1}{\sigma\sqrt{n}}\sum_1^n (X_j - \mu) \to N_0^1 \quad \textit{in distribution},$$

where we recall that N_0^1 is the Gaussian random variable of mean 0 *and variance* 1. *In other words,*

$$P\left(\frac{1}{\sigma\sqrt{n}}\sum_1^n (X_j - \mu) \leq s\right) \longrightarrow \int_{-\infty}^s \nu_0^1(x)dx. \qquad (40.1)$$

Proof. For a random variable X, we introduce the characteristic function defined by

$$\varphi_X(t) := E(e^{itX}).$$

This characteristic function can be expressed as

$$\varphi_X(t) = \int e^{itx} dP_X(x),$$

i.e. it is the Fourier transform of the probability distribution dP_X.

Exercise. Show that $\varphi_{N_0^1}(t) = e^{-t/2}$.

If $X \in L^2$ and $E(X) = 0$, then expanding the exponent we have

$$\varphi_X(t) = 1 - \frac{t^2}{2}\sigma^2(X) + o(t^2).$$

Replacing X_j by $\frac{1}{\sigma}(X_j - \mu)$ we may assume that $\mu = 0$ and $\sigma = 1$. Now let $S_n = \frac{1}{\sigma\sqrt{n}} \sum_1^n X_j$. Then using the joint probability distribution, we obtain

$$\varphi_{S_n}(t) = \int e^{\frac{it}{\sigma\sqrt{n}} \sum_1^n x_j} dP_{(X_1,\dots,X_n)}(x_1,\dots,x_n).$$

Using the independence (i.e. $dP_{(X_1,\dots X_n)} = \prod_1^n dP_{X_j}$) and Fubini's theorem (equality of multiple and iterated integrals), we obtain

$$\varphi_{S_n}(t) = \prod_1^n \varphi_{X_j}\left(\frac{t}{\sigma\sqrt{n}}\right).$$

Since the X_j are identically distributed with mean 0 and variance 1 (in particular, all characteristic functions $\varphi_{X_j}(t)$ are equal), this gives

$$\begin{aligned} \varphi_{S_n}(t) &= \left[\varphi_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n \\ &= \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n, \end{aligned}$$

which converges to $e^{-t^2/2}$ as $n \to \infty$. Now since $e^{-t^2/2} = \varphi_{N_0^1}(t)$ (see the last exercise), we have

$$\varphi_{S_n}(t) \longrightarrow \varphi_{N_0^1}(t)$$

as $n \to \infty$. From this, one can derive (40.1). ■

41. Construction of sample spaces

In experiments we are usually given joint probability distributions of random variables and we want to reconstruct their sample space Ω or the space of elementary events.

First, we have to decide which probability measures, say on $\mathbb{R}^n$ (or on a more general space), constitute probability distributions. So consider a family $\{X_\alpha\}_{\alpha\in I}$ of random variables on $(\Omega, \mathcal{B}, P)$, indexed by a set I. If the set I is finite (say it has n elements), then the answer to our question is easy: every probability measure $\tilde{P}$ on $\mathbb{R}^n$ is a joint probability density for some random variables $X_1, \ldots, X_n$. Indeed, take $\Omega = \mathbb{R}^n$, $\mathcal{B} = B_{\mathbb{R}^n}$ and $P = \tilde{P}$, while for X_j, we take the coordinate functions $X_j : \mathbb{R}^n \ni (x_1, \ldots, x_n) \mapsto x_j$. Then $(X_1, \ldots, X_n)$ is the identity map on $\mathbb{R}^n$ and therefore

$$\begin{aligned} P_{(X_1,\ldots,X_n)}(B_1 \times \ldots \times B_n) &:= P((X_1, \ldots, X_n)^{-1}(B_1 \times \ldots \times B_n)) \\ &= P(B_1 \times \ldots \times B_n) \\ &= \tilde{P}(B_1 \times \ldots \times B_n) \end{aligned}$$

as required.

However, if I is infinite, then the problem is more subtle. Let us examine the joint probability distribution for a family $\{X_\alpha\}_{\alpha\in I}$ of random variables on $(\Omega, \mathcal{B}, P)$. Pick any number n and a finite collection $(\alpha_1, \ldots, \alpha_n)$ and let $P_{\alpha_1,\ldots,\alpha_n}$ be the joint probability distribution of $X_{\alpha_1}, \ldots, X_{\alpha_n}$. Then the $P_{\alpha_1,\ldots,\alpha_n}$ have the following properties:

(A) $dP_{\alpha_{\pi(1)},\ldots,\alpha_{\pi(n)}}(x_{\pi(1)}, \ldots, x_{\pi(n)}) = dP_{\alpha_1,\ldots,\alpha_n}(x_1, \ldots, x_n)$ for any permutation π of n indices,

(B) $P_{\alpha_1,\ldots,\alpha_k}(E) = P_{\alpha_1,\ldots,\alpha_n}(E \times \mathbb{R}^{n-k})\ \forall k \leq n,\ \forall E \in B_{\mathbb{R}^k}$.

Sets of the form $E \times \mathbb{R}^{n-k}$ are called cylindrical: the first k coordinates $x_1, \ldots, x_k$ are constrained to be in E while the other coordinates are arbitrary.

Proof. We take for simplicity $n = 2$. Then we have

(A) $P_{\alpha_2,\alpha_1}(E_2 \times E_1) = P(X_{\alpha_2}^{-1}(E_2) \cap X_{\alpha_1}^{-1}(E_1)) = P_{\alpha_1,\alpha_2}(E_1 \times E_2)$,

(B) $P_{\alpha_1,\alpha_2}(E \times \mathbb{R}) = P(X_{\alpha_1}^{-1}(E) \cap X_{\alpha_2}^{-1}(\mathbb{R}))$, but $X_{\alpha_2}^{-1}(\mathbb{R}) = \Omega$, so $P_{\alpha_1,\alpha_2}(E \times \mathbb{R}) = P_{\alpha_1}(E)$. ■

A remarkable fact is that the conditions (A) and (B) are also sufficient to characterize joint probability distributions. In other words, given probability measures $P_{\alpha_1,\dots,\alpha_n}$ on $\mathbb{R}^n$ for any $\alpha_1, \dots, \alpha_n \in I$ satisfying (A) and (B), there is a probability space $(\Omega, \mathcal{B}, P)$ and a family $\{X_\alpha\}_{\alpha \in I}$ of random variables indexed by I s.t. $P_{\alpha_1,\dots,\alpha_n}$ are joint probability distributions for $\{X_\alpha\}_{\alpha\in I}$, i.e.

$$P_{\alpha_1,\dots,\alpha_n} = P_{(X_{\alpha_1},\dots,X_{\alpha_n})} \quad \forall \alpha_1, \dots, \alpha_n \in I.$$

In fact, the space Ω and the random variables X_α, $\alpha \in I$ are explicitely constructed as follows:

$$\Omega = (\mathbb{R}^*)^I,$$

where $\mathbb{R}^* = \mathbb{R} \cup \{\infty\}$ is a one point compactification of $\mathbb{R}$ and

$$(\mathbb{R}^*)^I := \{f : I \to \mathbb{R}^*\}.$$

The random variables are defined as the coordinate functions, i.e. $X_\alpha : \Omega \to \mathbb{R}^*$ is given by

$$X_\alpha(f) = f(\alpha).$$

This construction is due to A. N. Kolmogorov. For a proof, see e.g. [F], paragraph 9.18.

Exercises. 1) Show that

(a) if E_1, E_2 are random events, then so are E_1^c, E_2^c,

(b) the events $\{E_\alpha\}$ are independent iff the random variables $\{X_{E_\alpha}\}$ are independent,

(c) if the events $E_1, E_2, \dots$ are s.t. $\sum_1^\infty P(E_k) < \infty$, then

$$P(\cap_{m=1}^\infty(\cup_{k=m}^\infty E_k)) = 0,$$

i.e. the probability that ω belongs to infinitely many E_k's is zero (the Borel-Cantelli lemma). Hint: see [F] paragraph 9.10.

2) Consider for $a > 0$ a probability distribution of the form

$$\lambda_a = e^{-a} \sum_0^\infty \frac{a^k}{k!} \delta_k,$$

where $\delta_k(x)$ is the δ–function concentrated at k (λ_a is called the *Poisson distribution* with parameter a). Show that

(a) the mean and variance of λ_a are both equal to a,

(b) $\lambda_a * \lambda_b = \lambda_{a+b}$.

3) A fair coin is tossed $10,000$ times. Let X be the number of times it comes up heads. Use the central limit theorem to estimate:

1. the probability that $4950 \leq X \leq 5050$,
2. the number k s.t. $|X - 5000| \leq k$ with probability 0.98. Hint: consult a table of values of $(2\pi)^{-k/2} \int_0^x e^{-t^2/2} dt$.

42. The Wiener process

The *Wiener process* or *Brownian motion* is a basic stochastic process playing the same role in stochastic analysis as the Laplace operator in differential equations. This process was discovered by N. Wiener whose goal was to give a mathematical justification to a physical description by Einstein and Smolochowski of a phenomenon first observed by the Scottish botanist Brown. The phenomenon consists of a seemingly chaotic motion of small particles (pollen) suspended in a fluid such as water or air.

Formally, an abstract Wiener process B_t, $0 \leq t < \infty$ is a *stochastic process* (i.e. a family of random variables indexed by $t \in [0, \infty)$) satisfying the following conditions:

(A) $B_0 = 0$ (almost surely),

(B) the random variables $B_{t_1} - B_{t_0}, \ldots, B_{t_n} - B_{t_{n-1}}$ are independent for any $0 \leq t_0 < t_1 < \ldots < t_n$,

(C) the random variables $B_t - B_s$ are Gaussian with mean zero and variance $C(t-s)$ for some $C > 0$ and for any $0 < s < t$.

Let us explain the meaning of these conditions. Let B_t describe the coordinate of a particle in dimension one at time t. Then condition (A) states that the particle starts at the origin at $t = 0$. It is just a normalization condition.

Condition (B) says that the displacement of the particle at any given moment in time is independent of its previous displacements. This reflects the irregularity of the particle motion (the displacement of the particle at any given moment depends only on how it was hit by a molecule of the fluid it is suspended in at this very moment).

Condition (C) can be understood in the following way. Write $B_t - B_s$ as a sum of independent random variables

$$B_t - B_s = \sum_1^n \left(B_{t_j} - B_{t_{j-1}}\right),$$

where $t_0 = s$ and $t_n = t$. Here, n can be taken arbitrarily large. So on the basis of the central limit theorem, we expect that the r.h.s. is very close to a Gaussian random variable.

The main questions we ask are

(1) Does a Wiener process exist?

(2) Is it unique?

(3) What is its concrete realization?

We now show that axioms (A)–(C) lead to the fulfilment of the conditions of the Kolmogorov theorem (see last section) and therefore a Wiener process exists and, under additional regularity conditions on probability measures not elaborated here, the process is unique. Moreover, the same theorem leads to a concrete realization of the Wiener process.

In order not to carry around the constant C in condition (C), we set for simplicity $C = 1$.

Thus we assume for a moment that we already know that a Wiener process exists and compute its probability distributions.

Exercises. **1)** Let X be a random variable and $T : \mathbb{R} \to \mathbb{R}$ a measurable function. Let $Y = T(X)$. Show that

(a) $P_Y = P_X \circ T^{-1}$,

(b) $dP_Y(x) = |T'|^{-1} dP_X(T^{-1}(x))$.

2) Generalize these statements to vector valued random variables, i.e. for $X = (X_1, \dots, X_n)$ and $T : \mathbb{R}^n \to \mathbb{R}^n$ (hint: use the result $\int f dP_X = E(f(X))$ to show that for any measurable function $f : \mathbb{R} \to \mathbb{R}$,

$$\int f dP_X = \int f \circ T dP_Y = \int f |dP_Y \circ T^{-1}|).$$

Pick arbitrary $0 = t_0 < t_1 < t_2 < \dots < t_n$ and denote $B_j = B_{t_j}$ and $Y_j = B_{t_j} - B_{t_{j-1}} = B_j - B_{j-1}$, $j = 1, \dots, n$. Let us find an expression for $P_{(B_1,\dots,B_n)}$ based on assumptions (A)–(C). First, we observe that

$$(B_1, \dots, B_n) = T(Y_1, \dots, Y_n),$$

where

$$T(y_1, \dots, y_n) = (y_1, y_1 + y_2, \dots, y_1 + \dots + y_n).$$

Now the last exercise gives

$$P_{(B_1,\dots,B_n)} = P_{(Y_1,\dots,Y_n)} \circ T^{-1}.$$

On the other hand, since $Y_1, \dots, Y_n$ are independent Gaussian random variables with mean zero and variances $t_1 - t_0, t_2 - t_1, \dots, t_n - t_{n-1}$, we have

$$dP_{(Y_1,\dots,Y_n)}(y_1, \dots, y_n) = \prod_1^n dP_{Y_j}(y_j) = \prod_1^n \frac{e^{-\frac{y_j^2}{2(t_j - t_{j-1})}}}{\sqrt{2\pi(t_j - t_{j-1})}} dy_j$$

and therefore, since

$$T^{-1}(x_1, \dots, x_n) = (x_1, x_2 - x_1, \dots, x_n - x_{n-1}),$$

we have

$$dP_{(Y_1,\dots,Y_n)}(T^{-1}(x_1,\dots,x_n)) = \prod_1^n d\nu_{x_{j-1}}^{t_j-t_{j-1}}(x_j),$$

where we recall that $d\nu_\mu^{\sigma^2}(x)$ is the normal distribution with mean μ and variance σ^2 (see (37.2)). Remark that $\det T^{-1} = 1$, so exercise 2) above shows that

$$dP_{(B_1,\dots,B_n)}(x_1,\dots,x_n) = \prod_1^n d\nu_{x_{j-1}}^{t_j-t_{j-1}}(x_j).$$

Consequently,

$$dP_{(B_{t_1},\dots,B_{t_n})}(x_1,\dots,x_n) = \prod_1^n d\nu_{x_{j-1}}^{t_j-t_{j-1}}(x_j), \tag{42.1}$$

provided $0 = t_0 < t_1 < \dots < t_n$. For general $t_1,\dots,t_n$, we first time–order $t_1,\dots,t_n$, say, $0 = t_0 < t_{m_1} < \dots < t_{m_n}$, and then we set $dP_{(B_{t_1},\dots,B_{t_n})}(x_1,\dots,x_n) = dP_{B_{t_{m_1}},\dots,B_{t_{m_n}}}(x_{m_1},\dots,x_{m_n})$.

Exercise. Show that the conditions of Kolmogorov's reconstruction theorem are satisfied for the measures

$$dP_{t_1,\dots,t_n}(x_1,\dots,x_n) = \prod_1^n d\nu_{x_{m_{j-1}}}^{t_{m_j}-t_{m_{j-1}}}(x_{m_j}), \tag{42.2}$$

where $m_0 = 0$ and $m_1,\dots,m_n$ are such that $0 = t_{m_0} < t_{m_1} < \dots < t_{m_n}$, for any $t_1,\dots,t_n \in \mathbb{R}^+$.

Now Kolmogorov's reconstruction theorem tells us that the sample space for our process is

$$\Omega = (\mathbb{R}^*)^{[0,\infty)} \equiv \{\omega : [0,\infty) \to \mathbb{R}^*\}$$

and the process $B_t : \Omega \to \mathbb{R}$ is realized as the evaluation functional

$$B_t(\omega) = \omega(t).$$

Unfortunately, the space Ω is too large, it is the space of all functions from $[0,\infty)$ into the compactified line. The hope is that functions from

Ω which are too wild have probability zero, i.e. that our probability measure P is supported, in fact, on a much smaller set, say the set of continuous functions, $\Omega_c = C([0,\infty),\mathbb{R})$. This is indeed true:

Theorem. *Let P be the unique Radon measure on $\Omega = (\mathbb{R}^*)^{[0,\infty)}$ whose finite dimensional probability distributions are given by (42.2). Then Ω_c is a Borel subset of Ω and $P(\Omega_c) = 1$.*

For the definition of Radon measures and a proof of this theorem, see [F].

The next question we address is how smooth in fact the sample paths are. In other words, what is the smallest subspace $\Omega^{(0)}$ of Ω s.t. $P(\Omega^{(0)}) = 1$? One can show that the sample paths of the Wiener process are almost surely nowhere differentiable. So we have to look for Hölder continuous subspaces: let

$$\Omega_\alpha := \{\omega \in \Omega_c : \ |\omega(t') - \omega(t)| \leq C|t'-t|^\alpha \text{ for some } C < \infty\}.$$

One can show that $P(\Omega_\alpha) = 0$ for $\alpha \geq 1/2$ and $P(\Omega_\alpha) = 1$ for $\alpha < 1/2$. We can sharpen the second part of this result as

Theorem. $P(\Omega_{\frac{1}{2},\log}) = 1$*, where we defined*

$$\begin{aligned}\Omega_{\frac{1}{2},\log} \ := \ & \{\omega \in \Omega_c : \ |\omega(t') - \omega(t)| \leq C\sqrt{|t'-t| \ \log|t'-t|^{-1}}, \\ & \textit{for some } C < \infty\}.\end{aligned}$$

In fact, we can say a little bit more: For $C > 0$ and for a.a. ω, there is a $\delta = \delta(\omega) \in (0,1)$ s.t.

$$|B_{t'}(\omega) - B_t(\omega)| \leq C\sqrt{|t'-t| \ \log|t'-t|^{-1}},$$

provided $|t'-t| \leq \delta$. The constant C in the last statement is uniform in ω. See [F] for a proof of this.

Exercises. Show that **(i)** B_t is a Gaussian process,

(ii) $E(e^{i\sum_1^k u_j B_{t_j}}) = e^{-\langle u, Cu\rangle/2}$, where $u = (u_1, \dots, u_k)$ and

$$C = \begin{pmatrix} t_1 & t_1 & \dots & t_1 \\ t_1 & t_2 & \dots & t_2 \\ .. & .. & \dots & .. \\ t_1 & t_2 & \dots & t_k \end{pmatrix}.$$

Note that this implies that

$$E(e^{iuB_t}) = e^{-u^2t/2}. \tag{42.3}$$

(iii) Use the power series for the exponential in (42.3) to show that

$$E(B_t^{2k}) = \frac{(2k)!}{2^k k!} t^k \quad \forall k \in \mathbb{N}.$$

The *n-dimensional Wiener process* is given by $\vec{B}_t = (B_t^1, \dots, B_t^n)$, where B_t^j are independent, one dimensional Wiener processes, i.e.

(a) B_t^j is a one dimensional Wiener process for all j,

(b) $B_{t_k}^j - B_{t_{k-1}}^j$, $1 \le j \le n$, $1 \le k \le m$, are independent random variables for any $0 \le t_1 < \dots < t_m$.

Exercise. Show that for $t > s$, the random variable $\vec{B}_t - \vec{B}_s$ has the probability distribution

$$P(t,x) := (2\pi(t-s))^{-n/2} e^{-\sum_1^n \frac{x_j^2}{2(t-s)}} dx_1 \dots dx_n.$$

Observe that $P(t, x-y)$ is the integral kernel of the operator $e^{t\Delta}$ solving the diffusion equation $\frac{\partial}{\partial t}U = \Delta U$ and $U(0) = id$.

43. Stochastic integrals

In this section, we describe the theory of stochastic integration. One of the main goals of this theory is to provide tools for solving stochastic differential equations. We begin with the simple case of the Wiener integral.

The Wiener integral. Let B_t be the Wiener process on the probability space $(\Omega, \mathcal{B}, P)$. Let $f : J = [0,1] \to \mathbb{C}$. We want to define the integral

$$\int_J f(t) dB_t.$$

The value of this integral is not a number, but a random variable, since the "measure dB_t" is random. To define this integral, we proceed as in the usual theory of integration. For $\Delta = (s,t]$, let $B_\Delta = B_t - B_s$ and for a simple function $f = \sum c_k \chi_{\Delta_k}$ we define

$$I(f) = \sum c_k B_{\Delta_k}.$$

As before, $I(f)$ is independent of the representation of f. For a general measurable function f, we define the integral by approximating f first by simple functions $f^{(n)}$ (e.g. $f^{(n)} = \sum_1^n f(t_k^*) \chi_{\Delta_k}$, $t_k^* \in \Delta_k$) and then we take the limit of the integrals:

$$I(f) = \lim_{n \to \infty} I(f^{(n)}). \tag{43.1}$$

The question is whether and in what sense the r.h.s. converges. Fix $\omega \in \Omega$. Then $t \mapsto B_t(\omega)$ is a usual function. Recall that a function $t \mapsto F_t$ defines a measure iff F is of bounded variation, i.e. iff

$$\operatorname{Var}(F) := \sup \left\{ \sum_1^n |F_{t_j} - F_{t_{j-1}}| : 0 = t_0 < \cdots < t_n = 1 \right\} < \infty.$$

However, the function $t \mapsto B_t(\omega)$ is not of bounded variation with probability one (i.e. for a.e. ω).

Hence we have to understand the convergence in (43.1) differently. To this end, consider the L^2–space $L^2(\Omega, P)$. Observe that the inner product in this space can be written as $\langle X, Y \rangle = E(X\overline{Y})$, where we now consider complex valued random variables. Let L_s^2 denote the subspace of simple functions in $L^2(J)$. We claim that the linear operator $I : f \mapsto I(f)$ is a bounded map from L_s^2 into $L^2(\Omega, P)$. In fact, $I : L_s^2 \to L^2(\Omega, P)$ is an isometry:

$$E(I(f)\overline{I(g)}) = \int_J f\overline{g}. \tag{43.2}$$

Ineed, since $B_\Delta \in L^2$, then so are $\sum c_i B_{\Delta_i}$. Relation (43.2) follows by linearity from the relation

$$E(B_\Delta B_{\Delta'}) = \begin{cases} 0 & \text{if } \Delta \cap \Delta' = \emptyset \\ |\Delta| & \text{if } \Delta = \Delta', \end{cases}$$

which is due to the definition of the Wiener process B_t. ■

Exercise. Show that $E(B_\Delta B_{\Delta'}) = |\Delta \cap \Delta'|$.

Since the subspace of simple functions on J is dense in $L^2(J)$, we can extend this operator to the entire space $L^2(J)$. In other words, the limit in (43.1) must be understood in the sense of L^2–random variables.

The integral (43.1) in called the *Wiener integral* and is written as

$$\int_J f dB.$$

The map $I : f \mapsto \int f dB$ is an isometry from $L^2(J)$ to $L^2(\Omega, P)$.

Example: The Ornstein–Uhlenbeck process. This process is given by

$$V_t = V_0 e^{-\alpha t} + \int_0^t e^{-\alpha(t-s)} dB_s. \tag{43.3}$$

It serves as a model for the velocity of a Brownian particle, which does not exist in the classical sense. The r.v. V_0 can be considered as independent of B_t, $t \geq 0$.

Differentiating (43.3) formally, we obtain the stochastic Langevin equation:

$$dV = -\alpha V dt + dB, \tag{43.4}$$

which expresses the fact that the acceleration is equal to the friction plus a random kick (fluctuation). We understand (43.4) as a symbolic expression for (43.3).

The Ito integral. Now we consider the case when the integrand (which we write now as f_t) is itself a random variable. Hence we want to define the random variable

$$\int f_t dB_t$$

(in shorthand: $\int f dB$). This is a tricky business as the following example shows. Take $f_t = B_t$ and consider two "Riemannian" approximations for the integral $\int B_t dB_t$ by replacing first B_t by the simple functions $B_t^{(1)} = \sum_1^n B_{t_j} \chi_{\Delta_j}(t)$ and then by the simple functions $B_t^{(2)} = \sum_1^n B_{t_{j+1}} \chi_{\Delta_j}(t)$, where $\Delta_j = [t_j, t_{j+1})$. This gives us two approximations for $\int B_t dB_t$, namely $I_n^{(1)} = \sum_1^n B_{t_j} B_{\Delta_j}$ and $I^{(2)} = \sum_1^n B_{t_{j+1}} B_{\Delta_j}$. We compute the difference of the expectations of these integrals:

$$E(I_n^{(2)}) - E(I_n^{(1)}) = \sum_1^n E(B_{\Delta_j}^2) = \sum_1^n |\Delta_j| = |J|.$$

Therefore, $I_n^{(1)}$ and $I_n^{(2)}$ can not converge to the same random variable.

Compare the last relation with a similar relation for a function F_t of bounded variation where we have

$$\sum_1^n |F_{\Delta_j}|^2 \leq \max |F_{\Delta_j}| \mathrm{Var}(F) \longrightarrow 0$$

as $n \to \infty$. Roughly, the difference here is that for a differentiable function $F_\Delta = O(|\Delta|)$, while for the Wiener process, $B_\Delta = O(\sqrt{|\Delta|})$.

To define the stochastic integral $\int f dB$ for stochastic processes f_t, we begin with *simple stochastic processes*

$$f_t = \sum_1^n X_j \chi_{\Delta_j}(t),$$

where X_j are L^2–random variables and $\Delta_j = [t_j, t_{j+1})$. We define the integral of such functions as

$$I(f) := \sum_1^n X_j B_{\Delta_j}.$$

Notice that this corresponds to the choice $I_n^{(1)}$ in the above example, i.e. we choose a "Riemannian" approximation and take the value of the function $X(t)$ to be integrated at the left endpoint t_j of each interval Δ_j. The integral constructed in this way is called the *Itô integral.* Other integrals can be defined, e.g. taking the midpoint of the function in the Riemann sum yields a stochastic integral called the *Stratonovich integral.* We shall deal exclusively with the Itô integral.

In a second step, we approximate a stochastic process f_t by simple processes $f_t^{(n)}$ (say, $f_t^{(n)} = \sum_1^n f_{t_j}\chi_{\Delta_j}$) and define

$$I(f) = \lim_{n\to\infty} I(f_t^{(n)}),$$

provided $I(f_t^{(n)})$ converges in some sense. The point, as in the usual theory of integration, is to find a class of stochastic processes f_t s.t. the above procedure makes sense.

One way to do so is as follows. Let $L_s^2 \subset L^2(J, L^2(\Omega, P))$ be the subspace of simple processes. Of this subspace L_s^2, we pick again a subspace $\mathcal{M}$ s.t. the map

$$I : \mathcal{M} \longrightarrow L^2(\Omega, P)$$

is uniformly bounded (in fact, an isometry). Then we extend I to the closure $\overline{\mathcal{M}}$ of $\mathcal{M}$ in $L^2(J, L^2(\Omega, P))$ (which will happen not to be a proper subspace of the latter!).

Let us define $\mathcal{M}$ right away: $\mathcal{M}$ is a set of simple processes which are *non-anticipating.* A simple process $f_t = \sum X_j\chi_{\Delta_j}(t)$ is called (weakly) non–anticipating (w.r. to B_t) iff the r.v.'s X_j are independent of the r.v.'s B_{Δ_k} with $k \geq j$.

Proposition. *$I : \mathcal{M} \to L^2(\Omega, P)$ is an isometry, i.e.*

$$E\left(I(f)\overline{I(g)}\right) = \int E(f_t\,\overline{g_t})dt. \tag{43.5}$$

Proof. First, observe that $f_t = \sum X_j\chi_{\Delta_j}(t)$ is L^2 iff the X_j's are L^2. Since X_j and B_{Δ_j} are L^2 and independent, $X_jB_{\Delta_j}$ are L^2 r.v.'s. Since X_j^2 and $B_{\Delta_j}^2$ are independent r.v.'s, we have, using $E(B_{\Delta_j}^2) = Var(B_{\Delta_j}) = |\Delta_j|$

$$E\left((X_jB_{\Delta_j})^2\right) = E(X_j^2)E(B_{\Delta_j}^2) = E(X_j^2)|\Delta_j|.$$

Next, since B_{Δ_k} is independent of B_{Δ_j}, X_j and X_k for $j < k$, we have by a theorem about functions of independent r.v.'s that B_{Δ_k} is independent of $X_j B_{\Delta_j} X_k$ and therefore, for $j < k$,

$$E(X_j B_{\Delta_j} X_k B_{\Delta_k}) = E(X_j B_{\Delta_j} X_k) E(B_{\Delta_k}) = 0.$$

The last two relations imply

$$E\left((\sum X_j B_{\Delta_j})^2 \right) = \sum E(X_j^2)|\Delta_j|.$$

Exercise. Show that

$$\int E(f_t^2) dt = \sum E(X_j^2)|\Delta_j|.$$

The last two relations yield (43.5) for $g = f$. The case of arbitrary g is treated similarly. ■

The remaining question is what $\overline{\mathcal{M}}$ is, in other words, what kind of process in $L^2(J, L^2(\Omega, P))$ can be approximated by simple non–anticipating processes.

Given an L^2–process f_t, we approximate it by the simple processes

$$f_t^{(n)} = \sum_0^{n-1} f_{t_j} \chi_{\Delta_j},$$

where $\Delta_j = [t_j, t_{j+i})$ with $t_0 = 0$ and $t_n = 1$. So a non–anticipating process should have a property that for any such approximation, i.e. for any $t_1 < \cdots < t_n$, f_{t_j} are independent of B_{Δ_k} with $k \geq j$. Since $\{t_l\}$ are arbitrary, this says that f_t is independent of B_Δ as long as $\inf \Delta \geq t$. The latter condition is satisfied if $\forall B \in \mathcal{B}_\mathbb{R}$, $f_t^{-1}(B) \in \mathcal{B}^t$, where $\mathcal{B}^t$ is the σ–algebra generated by $B_s^{-1}(C)$, $\forall C \in \mathcal{B}_\mathbb{R}$, $\forall s \leq t$, in other words, if f_t is measurable w.r. to $\mathcal{B}^t$. Such a process is called *non–anticipating* (w.r. (or adapted) to B^t).

Definition. $\mathcal{V}([a,b])$ is the class of non-anticipating processes $f_t(\omega)$ (with $t \in [a,b]$) that are $\mathcal{B}_\mathbb{R} \times \mathcal{B}$–measurable.

Examples. a) $f_t = B_{t/2}$ is non–anticipating while $f_t = B_{2t}$ is not; **b)** $f = \sum X_j \chi_{\Delta_j}$ is non–anticipating iff X_j are $\mathcal{B}^{t_j}$–measurable.

A family $\{\mathcal{B}^t\}$ of σ–algebras s.t. $\mathcal{B}^s \subseteq \mathcal{B}^t$ if $s \leq t$ is called a *filter.*

Exercises. Show that the family $\mathcal{B}^t$ defined above is **(a)** a filter and **(b)** *right continuous* in the sense that $\cap_{t \geq s} \mathcal{B}^t = \mathcal{B}^s$.

For a non–anticipating process f_t we define the *Ito integral* as

$$\int f dB := \lim_{n \to \infty} \sum_{0}^{n-1} f_{t_j} B_{\Delta_j},$$

where $\Delta_j = [t_j, t_{j+1})$, $0 = t_0 < t_1 < \cdots < t_n = 1$ and the convergence is understood in the sense of the norm $\sqrt{E(|X|^2)}$.

Observe that we cannot replace the r.h.s. by $\sum_0^{n-1} f_{t_j^*} B_{\Delta_j}$, where $t_j^* \in \Delta_j$ and $t_j^* \neq t_j$. Indeed, the simple function $\sum_0^{n-1} f_{t_j^*} B_{\Delta_j}$ is not in general non–anticipating.

Exercises. Show that **(a)** $\lim_{n \to \infty} \sum_0^{n-1} [B_{t_{j+1}} - B_{t_j}]^2 = t$ with probability one, **(b)** if $X_t = \int_{t_0}^t f dB$, then $X_{t+h} - X_t = f_t(B_{t+h} - B_t) + o(h^{1/2})$, where $h^{-1} E(|o(h^{1/2})|^2) \to 0$ as $h \to 0$, **(c)** $\int_0^t B dB = \frac{1}{2} B_t^2 - \frac{1}{2} t$ (hint: use the identities $B_{t_j}(B_{t_{j+1}} - B_{t_j}) = \frac{1}{2}(B_{t_{j+1}}^2 - B_{t_j}^2) - \frac{1}{2}(B_{t_{j+1}} - B_{t_j})^2$), **(d)** $\int_0^t s dB_s = t B_t - \int_0^t B_s ds$ (integration by parts formula, hint: use that $\sum_j \Delta(s_j B_j) = \sum_j s_j \Delta B_j + \sum B_{j+1} \Delta s$), **(e)** $\int_0^t B_s^2 dB_s = \frac{1}{3} B_t^3 - \int_0^t B_s ds$.

Properties of the Ito integral. Let $f, g \in \mathcal{V}([0, s])$, α, β constants and let $0 < r < u < s$. Then

(i) $\int_r^s f dB = \int_r^u f dB + \int_u^s f dB$ a.s.,

(ii) $\int_r^s (\alpha f + \beta g) dB = \alpha \int_r^s f dB + \beta \int_r^s g dB$ a.s.,

(iii) $E(\int_r^s f dB) = 0$,

(iv) $\int_r^s f dB$ is $\mathcal{B}^s$–measurable.

Exercise. Prove (i)-(iv). Hint: prove these properties first for simple processes and then argue by continuity.

44. Ito Processes and the Ito Formula

Our goal is to develop key rules, such as integration by parts, which allow us to evaluate stochastic integrals. To this end, we introduce an appropriate class of stochastic processes, the Ito processes, and prove the Ito formula, which corresponds to the chain rule of standard differential calculus.

As before, we assume we have a Brownian motion B_t on a probability space $(\Omega, \mathcal{B}, P)$.

Definition. An *Ito-process* (or stochastic integral) is a process X_t on $(\Omega, \mathcal{B}, P)$ of the form

$$X_t = X_0 + \int_0^t u_s ds + \int_0^t v_s dB_s, \tag{44.1}$$

where u_s and v_s are stochastic processes s.t. $v \in \mathcal{V}$ satisfies the estimate

$$P\left(\int_0^t v_s^2 ds < \infty \ \forall t \geq 0\right) = 1$$

and u_s is non-anticipating and satisfies the estimate

$$P\left(\int_0^t |u_s| ds < \infty \ \forall t \geq 0\right) = 1.$$

Differentiating (44.1) formally with respect to t, we get

$$dX_t = u_t dt + v_t dB_t. \tag{44.2}$$

We use (44.2) as a shorthand for (44.1). Thus for us, (44.2) is just a rule of memorizing (44.1). For example,

$$d\left(\frac{1}{2}B_t^2\right) = \frac{1}{2}dt + B_t dB_t$$

stands for the integration formula

$$\int_0^t B_s dB_s = \frac{1}{2}B_t^2 - \frac{1}{2}t,$$

proven in one of the exercises above.

The key result of this section is the following chain rule formula for computing stochastic integrals of composite functions:

Theorem: Ito's formula. *Let X_t be an Ito process given by $dX_t = u_t dt + v_t dB_t$. Let $g \in C^2(\overline{\mathbb{R}^+} \times \mathbb{R})$. Then $Y_t = g(t, X_t)$ is again an Ito process and it is given by*

$$dY_t = \frac{\partial g}{\partial t}(t, X_t)dt + \frac{\partial g}{\partial x}(t, X_t)dX_t + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(t, X_t)(dX_t)^2,$$

where $(dX_t)^2 = dX_t \cdot dX_t$ is computed according to the rules $dt \cdot dt = dt \cdot dB_t = dB_t \cdot dt = 0$ and $dB_t \cdot dB_t = dt$.

Before proving this theorem, we consider some examples and applications.

Examples. Find solutions to some exercises above using Ito's formula (with $X_t = B_t$ below):

$g(t,x)$	$dg(t,B_t)$
$x^2/2$	$B_t dB_t + dt/2$
$x^3/3$	$B_t^2 dB_t + B_t dt$
tx	$B_t dt + t dB_t$

Exercise. Check that these consequences of Ito's formula coincide with the results of the direct computation of Ito's integrals performed in the exercises above..

One can generalize the last example by letting $g(t,x) = f_t\dot{x}$ to obtain the *Leibnitz* rule:

$$d(f_t B_t) = B_t df_t + f_t dB_t.$$

Rewriting this formula in the integral form we obtain the *integration by parts formula*:

$$\begin{aligned} f_t B_t &= \int_0^t B_s df_s + \int_0^t f_s dB_s, \quad \text{or} \\ \int_0^t f_s dB_s &= f_t B_t - \int_0^t B_s df_s. \end{aligned}$$

Observe that here, f_t is a deterministic function.

Now we explain the main idea behind the proof of Ito's formula. We motivate it by analyzing the proof to the usual chain rule:

$$\begin{aligned} \frac{d}{dt} g(t, x_t) &= \frac{\partial g}{\partial t}(t, x_t) + \frac{\partial g}{\partial x}(t, x_t) \frac{dx_t}{dt}, \quad \text{or} \\ dg(t, x_t) &= \frac{\partial g}{\partial t}(t, x_t) dt + \frac{\partial g}{\partial x}(t, x_t) dx_t. \end{aligned} \tag{44.3}$$

We can prove this formula as follows: expand $\Delta g := g(t+\Delta t, x_{t+\Delta t}) - g(t, x_t)$ in Δt by Taylor's theorem:

$$\Delta g = \frac{\partial g}{\partial t} \Delta t + \frac{\partial g}{\partial x} \Delta x_t + o(\Delta t),$$

where $\Delta x_t := x_{t+\Delta t} - x_t$. Ignoring the higher order terms and replacing Δt by dt and Δg by $dg(t, x_t)$ gives us (44.3).

We can still shortcut this derivation as follows: expand $df(t, x_t) := g(t + dt, x_{t+dt}) - g(t, x_t)$ in dt and set $o(dt) = 0$.

Now we do the same with $Y_t = g(t, X_t)$ while recalling that $dB_t = B_{t+dt} - B_t = O(\sqrt{dt})$ and therefore $o((dX_t)^2) = 0$! Expanding $dY_t := g(t + dt, X_{t+dt}) - g(t, X_t)$ in dt, we obtain

$$\begin{aligned} dY_t &= \frac{\partial g}{\partial t}(t, X_t) dt + \frac{\partial g}{\partial x}(t, X_t) dX_t + \frac{1}{2} \frac{\partial^2 g}{\partial t^2}(t, X_t)(dt)^2 \\ &\quad + \frac{\partial^2 g}{\partial t \partial x}(t, X_t)\, dt\, dX_t + \frac{1}{2} \frac{\partial^2 g}{\partial x^2}(t, X_t)(dX_t)^2 \\ &\quad + o((dt)^2) + o(dt\, dX_t) + o((dX_t)^2), \end{aligned}$$

and therefore (recall that we agreed that $o(dt) = 0$)

$$dY_t = \frac{\partial g}{\partial t} dt + \frac{\partial g}{\partial x} dX_t + \frac{1}{2} \frac{\partial^2 g}{\partial x^2} (dX_t)^2$$

whith $(dX_t)^2 = (u_t dt + v_t dB_t)^2 = v_t^2 (dB_t)^2$.

Finally, we want to show that $(dB_t)^2 = dt$ in the sense that for any $f_s \in \mathcal{V}([0, t])$,

$$\sum f_{s_j} (B_{\Delta_j})^2 \longrightarrow \int_0^t f_s ds \quad \text{in } L^2(\Omega, P).$$

Indeed, compute

$$E\left(\left[\sum f_{s_j}(B_{\Delta_j})^2-\sum f_{s_j}|\Delta_j|\right]^2\right)$$
$$=\sum_{i,j}E\left(f_{s_i}f_{s_j}((B_{\Delta_i})^2-|\Delta_i|)((B_{\Delta_j})^2-|\Delta_j|)\right).$$

If $i<j$, then $f_{s_i}f_{s_j}((B_{\Delta_j})^2-|\Delta_j|)$ are independent of B_{Δ_j} and similarly for $i>j$. Hence all off-diagonal terms on the r.h.s. vanish. Since $f_{s_i}^2$ are independent of $((B_{\Delta_i})^2-|\Delta_i|)^2$, the r.h.s. becomes

$$\sum_i E\left(f_{s_i}^2((B_{\Delta_i})^2-|\Delta_i|)^2\right)=\sum_i E(f_{s_i}^2)E\left(((B_{\Delta_i})^2-|\Delta_i|)^2\right) \quad (44.4)$$

Exercise. Show that $E((B_\Delta)^4)=3|\Delta|^2$ (hint: use the fact that B_Δ is a Gaussian random variable with mean zero and variance $|\Delta|$).

The exercise implies that the expectation in the r.h.s. of (44.4) is equal to $E((B_{\Delta_i})^4)-2|\Delta_i|^2E((B_{\Delta_i}^2))+|\Delta_i|^4=O(|\Delta_i|^2)$ as $|\Delta_i|\to 0$ (i.e. $n\to\infty$), so the r.h.s. of (44.4) is of the order

$$\sum_i E(f_{s_i}^2)|\Delta_i|^2\to 0 \text{ as } n\to\infty.\blacksquare$$

Exercises. 1) Use Ito's formula to write the process $X_t=2+t+e^{B_t}$ in the standard form ($dX_t=u_tdt+v_tdB_t$). **2)** Let X_t,Y_t be Ito precesses. Show that $d(X_tY_t)=X_tdY_t+dX_tdY_t$. Deduce from this the following integration by parts formula;

$$\int_0^t X_sdY_s=X_tY_t-X_0Y_0-\int_0^t Y_sdX_s-\int_0^t dX_sdY_s.$$

3) For $X_t=e^{ct+\alpha B_t}$ show that $dX_t=(c+\alpha^2/2)X_tdt+\alpha X_tdB_t$. **4)** Find f_s s.t. $\sin(B_t)=E(\sin B_t)+\int_0^t f_sdB_s$.

Bibliography

[DFN] Dubrovin, B.A., Novikov, C.P., Fomenko, A.T.: *Modern Geometry - methods and applications.* 2$^{\text{nd}}$ edition, Springer Verlag (1992)

[E] Evans, L.C.: *Partial Differential Equations.* Graduate studies in mathematics Vol. 19, American Mathematical Society (1998)

[F] Folland, G.B.: *Real Analysis: modern techniques and their applications.* Wiley, New York (1984)

[HS] Hislop, P.D., Sigal, I.M.: *Introduction to spectral theory.* Applied Mathematical Sciences Vol. 133, Springer Verlag (1996)

[L] Landford, O.E. III: *Lectures on Dynamical Systems.* Lecture notes, Mathematics Department, ETH Zürich (1997)

[LL] Lieb, E.H., Loss, M.: *Analysis.* Graduate studies in mathematics Vol. 14, American Mathematical Society (1997)

[MMcC] Marsden, J.E., McCracken, M.: *The Hopf bifurcation and its applications.* Applied Mathematical Sciences, Vol. 19, Springer Verlag (1976)

[McO] McOwen, R.C.: *Partial Differential Equations: methods and applications.* Prentice Hall (1996)

[Ø] Øksendal, B.K.: *Stochastic differential equations: an introduction with applications.* 4$^{\text{th}}$ edition, Springer Verlag (1995)

[R] Rozanov, Y.A.: *Introduction to Random Processes.* Springer Verlag (1987)

[RS] Reed, M., Simon, B.: *Functional Analysis, Self-Adjointness.* Methods of Modern Mathematical Physics, Vol. I, Academic Press (1975)

[Z] Zeidler, E.: *Nonlinear Functional Analysis and its Applications.* Vol. I, Springer Verlag (1986)