Fundamentals of Dynamical Systems
(Tópicos de Sistemas Dinâmicos)
Licenciatura em Matemática

Salvatore Cosentino

Departamento de Matemática - Universidade do Minho

Campus de Gualtar - 4710 Braga - PORTUGAL

gab: CG - Edifício 6 - 3.48, tel: 253 604086

e-mail: scosentino@math.uminho.pt

url http://w3.math.uminho.pt/~scosentino

2019

Abstract

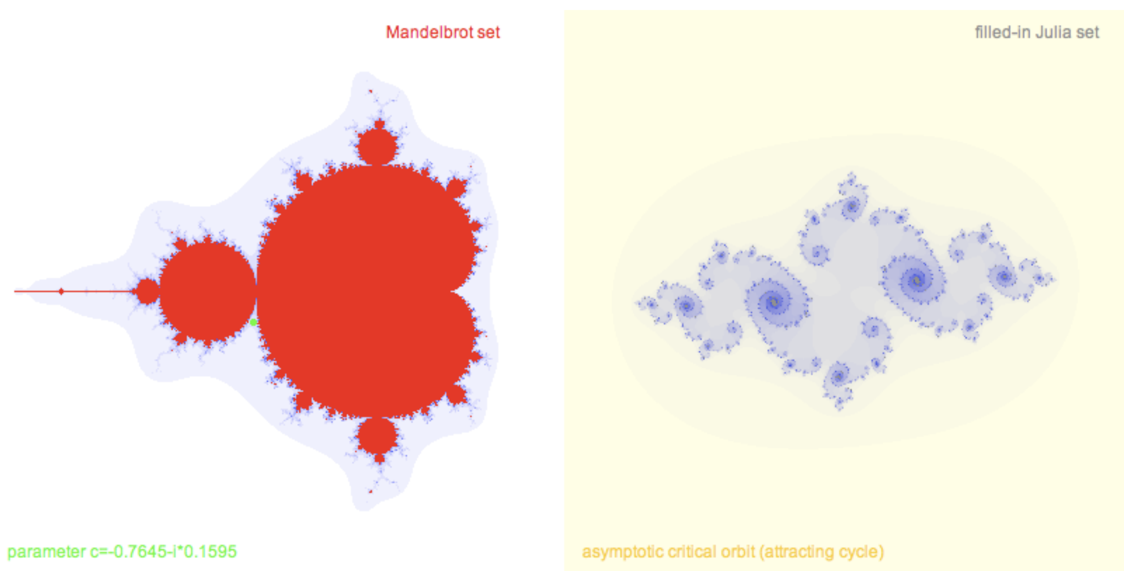
This is an english version of the notes written for my lectures on "Tópicos de Sistemas Dinâmicos" for the "Licenciatura em Matemática" of the University of Minho, during the last decade (available at my page http://w3.math.uminho.pt/~scosentino/salteaching.html). Emphasis is on examples, and on the interplay between different areas of mathematics. Some very important parts of the modern theory of dynamical systems, as hyperbolic theory, hamiltonian systems, or the qualitative theory of differential equations, are almost completely missing. Other interesting results or directions are only sketched. Main references and sources are [KH95, HK03], others are suggested along the text.

e.g. means *exempli gratia*, that is, "for example", and is used to introduce important or interesting examples.

ex: means "exercise", to be solved at home or in the classroom.

□ indicates the end of a proof.

Pictures were made with *Grapher* on my MacBook, or taken from Wikipedia, or produced with my own Java codes, like the one below.



Contents

1 Iterations 4
- 1.1 Exponential growth/decay . . . 4
- 1.2 Babylonians-Heron method to compute square roots . . . 9
- 1.3 From Newton method to Julia and Fatou sets . . . 10
- 1.4 Finite difference equations . . . 13
- 1.5 Interval maps and cobweb plot . . . 15

2 Topological dynamical systems, basic definitions 17
- 2.1 Transformations . . . 17
- 2.2 Trajectories and orbits . . . 18
- 2.3 Periodic orbits and basin of attraction . . . 18
- 2.4 Observables . . . 20
- 2.5 Invariant sets . . . 21
- 2.6 Conjugations . . . 22

3 Differential equations and flows 23
- 3.1 Flows . . . 23
- 3.2 Structure of physical models . . . 24
- 3.3 Integration of one-dimensional systems . . . 26
- 3.4 Existence and uniqueness theorems . . . 30
- 3.5 Oscillations and cycles . . . 34

4 Linear systems 42
- 4.1 Exponential of a linear operator . . . 42
- 4.2 Linear fows . . . 44
- 4.3 Linear systems in the plane . . . 46
- 4.4 Jordan normal form . . . 49
- 4.5 Hyperbolic linear flows . . . 53

5 Numbers and dynamics 55
- 5.1 Decimal expansion and multiplication by ten . . . 55
- 5.2 Bernoulli shifts . . . 57
- 5.3 Rotations of the torus . . . 59
- 5.4 Dyadic adding machine . . . 60
- 5.5 Continued fractions and Gauss map . . . 61
- 5.6 Exponential sums . . . 65

6 Simple orbits and perturbations 67
- 6.1 Topological fixed point theorems . . . 67
- 6.2 Dynamics of contractions . . . 67
- 6.3 Linear maps . . . 70
- 6.4 Order of the line and trajectories . . . 73
- 6.5 Local analysis: attracting and repelling fixed points . . . 74
- 6.6 Transversality and bifurcations . . . 78

7 Statistical description of orbits 81
- 7.1 Probability measures . . . 81
- 7.2 Transformations and invariant measures . . . 83
- 7.3 Invariant measures and time averages . . . 86
- 7.4 Examples of invariant measures . . . 88

8 Recurrences **91**
8.1 Limit sets and recurrent points 91
8.2 Dirichlet theorem on Diophantine approximation 92
8.3 Poincaré recurrence theorem 93
8.4 Transitivity and minimality 95
8.5 Kronecker theorem on irrational rotations 97
8.6 Circle homeomorfisms 99

9 Chaos and mixing **103**
9.1 Sensitive dependence and chaos 103
9.2 Topological mixing 104
9.3 Symbolic dynamics 106
9.4 Cantor sets 107
9.5 Expanding maps 109
9.6 Hyperbolic automorphisms of the torus 111
9.7 Topological entropy 113

10 Ergodicity and convergence of time means **117**
10.1 Ergodicity 117
10.2 Examples of ergodic maps 118
10.3 Normal numbers 120
10.4 Distribution of digits in continued fractions 121
10.5 Unique ergodicity and equidistribution 122

1 Iterations

1.1 Exponential growth/decay

e.g. Fibonacci numbers. Consider the following problem, posed by Leonardo Pisano (alias Fibonacci) in his *Liber Abaci*, 1202:

> *Quot paria cuniculorum in uno anno ex uno pario germinentur.*
> *Quidam posuit unum par cuniculorum in quodam loco, qui erat undique pariete circundatus, ut sciret, quot ex eo paria germinarentur in uno anno: cum natura eorum sit per singulum mensem aliud par germinare; et in secundo mense ab eorum nativitate germinant.*

Let f_n be the number of pairs of rabbits at the n-th month. The offspring one month later, $f_{n+1} - f_n$, is equal to the number of "adult" pairs present in the n-th month, which is f_{n-1}. Therefore, the f_n's satisfy the recursive law

$$f_{n+1} = f_n + f_{n-1}\,, \tag{1.1}$$

which prescribes the successive values of f_n given some initial values f_0 and f_1. The sequence grows quite fast, as you can see: if we take, with Fibonacci, the initial values $f_0 = f_1 = 1$, we get

$$1,\ 1,\ 2\ ,\ 3,\ 5,\ 8,\ 13,\ 21,\ 34,\ 55,\ 89,\ 144,\ 233,\ 377,\ 610,\ 987,\ 1597,\ 2584\ ,\ \dots$$

and the numbers soon become astronomically large. For example, after 10 years we get

$$f_{120} \simeq 8.67 \times 10^{24}\,,$$

larger than the Avogadro number!

In order to see this, you could write a simple code like (in `Java` or `c++`)

```
int Fib(int n)
  {
    if (n==0) return 1;
    else if (n==1) return 1;
    else return Fib(n-1) + Fib(n-2);
  }
```

An applet which computes the sequence is in my page `http://w3.math.uminho.pt/~scosentino/salbestiario.html`.

Also useful would be a formula, or at least an asymptotic formula, for the f_n's, and I'll show you one later. For example, an asymptotic formula would solve a problem like

ex: Estimate the smallest time n such that $f_n > 10^{80}$.

e.g. Duplication of bacteria. Experiments show that a population of bacteria, during a certain initial period at least, double each characteristic time $\tau > 0$. Thus, an initial population of N_0 cells gives origin to $N_1 = 2N_0$ after a time τ, to $N_2 = 4N_0$ cells after a time 2τ, ..., and to

$$N_n = 2^n N_0$$

cells after time $n\tau$. For example, a unique cell gives origin to 1024 cells after a time $t = n\tau$ such that $2^n = 1024$, i.e. $n\tau = (\log_2 1024) \cdot \tau = 10 \cdot \tau$.

Sequences as time series. A (real or complex valued) *sequence* is a collection $(x_n)_{n\in\mathbb{N}}$ of numbers $x_n \in \mathbb{R}$ or $\mathbb{C}$, indexed (hence ordered) by an non-negative integer $n \in \mathbb{N} := \{0, 1, 2, 3, \dots\}$. We may think of the index n as "time", an therefore at the n-th term x_n as the value of some "observable" x at time n (as the number of pairs of rabbits or of bacteria). Physicists call them "time series". Clearly, we may as well define sequences with values in an arbitrary set X, for example in the Euclidean space $\mathbb{R}^d$.

Subsequences are obtained forgetting to observe x at certain times, i.e. are sequences $(y_i)_{i\in\mathbb{Z}}$ defined by $y_i := x_{n_i}$, where $i \mapsto n_i$ is an increasing function of $\mathbb{N}$ into itself.

Sequences may be defined as functions are. Indeed, a sequence with values in the set X is nothing but a function $f : \mathbb{N} \to X$, disguised by the notation $x_n := f(n)$. A second possibility is some recursive law prescribing the value of x_n given the (past) values of $x_0, x_1, \ldots, x_{n-1}$. A third possibility, is using some property that the successive terms must have.

Engineers also use to look at sequences as "discrete-time signals" $x[n] = x(n\tau)$, possibly obtained from an analogic signal $x(t)$, defined for times t in some interval of the real line, sampling its values at integer multiples of some "sampling time" τ.

e.g. **Arithmetic progression.** An *arithmetic progression* $x_n = a + nb$, which may also be defined using the recursion $x_{n+1} = x_n + b$, with initial term $x_0 = a$.

e.g. **The primes sequence.** The sequence $2, 3, 5, 7, 11, 13, 17, 19, 23, \ldots$, whose generic term is the n-th prime number p_n. It is not clear what the recursive law could be.[1]

Limits. We say that the real or complex sequence (x_n) *converges* to some *limit* $a \in \mathbb{R}$ or $\mathbb{C}$, and we write $\lim_{n\to\infty} x_n = a$ or simply $x_n \to a$ (as $n \to \infty$), if for any "precision" $\varepsilon > 0$ there exists a time $\overline{n}$ such that $|x_n - a| < \varepsilon$ for all times $n \geq \overline{n}$. This means that the values x_n are within an arbitrarily small neighborhood of a as long as the time n is sufficiently large.

The basic fact about limits in the real line $\mathbb{R}$ is that monotone (non-decreasing or non-increasing, i.e. satisfying $x_{n+1} \geq x_n$ or $x_{n+1} \leq x_n$, for any n, respectively) bounded (i.e. such that $|x_n| \leq M$ for some $M > 0$ and all n) sequences of real numbers do admit limit. For example, the limit of a bounded increasing sequence is simply the supremum of the set of values.

We also use the notation $x_n \to \pm\infty$ to say that given an arbitrarily large $K > 0$ we can find a time $\overline{n}$ such that $\pm x_n > K$ for all times $n \geq \overline{n}$.

Of course, there exist sequences which do not admit limits in either senses. These are, for example, oscillating sequences, as $x_n = (-1)^n$. We'll encounter sequences with much more wild behavior.

Fundamental sequences. A sequence (x_n) is said *fundamental*, or *Cauchy sequence*, if for any precision $\varepsilon > 0$ there exists a time $\overline{n}$ such that

$$|x_n - x_m| < \varepsilon$$

for all times $n, m > \overline{n}$. Fundamental sequences are clearly bounded. It is obvious that a convergent sequence is fundamental (a triangular argument, since both x_n and x_m are $\varepsilon/2$-near to the limit for sufficiently large n and m). A similar triangular argument shows that a fundamental sequence with a convergent subsequence is itself convergent. Less obvious is that any fundamental sequence in $\mathbb{R}$ is convergent. Indeed, let $X_n := \{x_k \text{ with } k \geq n\}$. It is clear that the X_n are bounded, and therefore by the supremum axiom there exist the numbers $a_n := \inf X_n$. But the sequence (a_n) is bounded and not decreasing, and therefore there exists $a = \lim_{n\to\infty} a_n$ (indeed, $a = \sup\{a_n \text{ with } n \in \mathbb{N}\}$). It is then easy to construct subsequences of (x_n) which converge to a, and this implies that (x_n) itself is convergent to a.

Thus, we may know that a sequence is convergent without knowing its limit! In general, convergence of all fundamental sequences is taken as a definition of (sequential) completeness of a metric space.

Geometric progression. The most important sequence is the *geometric progression*, defined by the recursion

$$x_{n+1} = \lambda x_n ,$$

and an initial term $x_0 = a$ (which we may assume $\neq 0$ to avoid trivialities). Thus, the sequence is

$$x_0 = a \qquad x_1 = a\lambda \qquad x_2 = a\lambda^2 \qquad \ldots \qquad x_n = a\lambda^n \qquad \ldots$$

[1] This is not the place to talk about it, but if you find it intriguing, you may take a look at the wonderful book by Marcus du Sautoy, *The music of primes*, Harper-Collins, 2003 [*A música dos números primos*, Zahar, 2008].

The parameter λ (which may be real or complex) is called *ratio*, since it is the ratio x_{n+1}/x_n between successive terms of the sequence. The geometric sequence clearly converges to zero when $|\lambda| < 1$. It is constant, hence trivially convergent, when $\lambda = 1$, while oscillates between $\pm a$ when $\lambda = -1$ (hence does not converge if $a \neq 0$). We may also observe that $|\lambda^n| \to \infty$ when $|\lambda| > 1$.

ex: Show that the term x_n of a geometric progression is equal to the geometric mean $\sqrt{x_{n+1}x_{n-1}}$ of its neighbors.

Computing limits. First, observe that $x_n \to a$ is equivalent to $x_n - a \to 0$. Therefore, we only need to understand how to "prove" that some sequence converges to zero. One possibility is to "compare" the sequence (x_n) under investigation with a sequence with known behavior, as for example the geometric progression. Indeed, if $|x_n| \leq y_n$ for all n sufficiently large, then $y_n \to 0$ implies $x_n \to 0$ too.

Subsequences and sequential compactness. A *subsequence* of a sequence (x_n) is a sequence (x_{n_i}) obtained selecting only the values x_{n_i} of the original sequence, where $i \mapsto n_i$ is an increasing map $\mathbb{N}_0 \to \mathbb{N}_0$.

Sometimes we are only interested in a rough estimate of the growth of a sequence (x_n). The "limsup" is the limit $\limsup_{n\to\infty} x_n := \lim_{n\to\infty} a_n \in \mathbb{R} \cup \{\infty\}$ of the non-increasing sequence $a_n := \sup\{x_n, x_{n+1}, x_{n+2}, \dots\}$. The "liminf" is the limit $\liminf_{n\to\infty} x_n := \lim_{n\to\infty} b_n \in \mathbb{R}\cup\{-\infty\}$ of the non-decreasing sequence $b_n := \inf\{x_n, x_{n+1}, x_{n+2}, \dots\}$.

The basic fact (that closed and bounded sets of the real line are *sequentially compact*) is that any bounded sequence admits a convergent subsequence.

e.g. **Half-life.** Radioactive decay may be characterized by a "half-life", the time τ after which approximatly half of the initial nuclei decay, between a sufficiently large sample. If q_n denotes the number of nuclei at time $n\tau$, with $n = 0, 1, 2, \dots$, then

$$q_{n+1} = \tfrac{1}{2}\, q_n \,.$$

Thus, the number of nuclei at time $n\tau$ is $q_n = q_0 2^{-n}$, while the product of the decayment is $q_0 - q_n = q_0(1 - 2^{-n})$. Observe that $q_n \to 0$ when $n \to \infty$.

If solar radiation produces radioactive nuclei at a constant rate $\alpha > 0$ (i.e. α nuclei each time interval τ), then the number of radioactive nuclei at time $n\tau$ satisfies the recursion

$$q_{n+1} = \tfrac{1}{2} q_n + \alpha \,. \tag{1.2}$$

Equilibrium is possible when q_0 is equal to $\overline{q} := 2\alpha$, since then $q_1 = \alpha+\alpha = q_0$, $q_2 = \alpha+\alpha = q_1 = q_0$, and so on, $q_n = \overline{q}$ for all $n \in \mathbb{N}$.

What happens if $q_0 \neq \overline{q}$? The recursion says that

$$\begin{aligned}
q_1 &= \tfrac{1}{2} q_0 + \alpha \\
q_2 &= \tfrac{1}{4} q_0 + \tfrac{1}{2}\alpha + \alpha \\
q_3 &= \tfrac{1}{8} q_0 + \tfrac{1}{4}\alpha + \tfrac{1}{2}\alpha + \alpha \\
&\vdots \\
q_n &= \tfrac{1}{2^n} q_0 + \left(\tfrac{1}{2^{n-1}} + \cdots + \tfrac{1}{8} + \tfrac{1}{4} + \tfrac{1}{2} + 1\right)\alpha
\end{aligned}$$

The first term $q_0/2^{n+1} \to 0$ when $n \to \infty$, which means that "future" is independent on the initial condition q_0. The second term converges to 2α when $n \to \infty$, being the sum of a geometric series (if you forgot about it, see below).

A simpler formula for q_n may be obtained using the substitutio $x_n := q_n - \overline{q}$, where $\overline{q} = 2\alpha$ is the equiibrium solution. We get

$$\begin{aligned}
x_{n+1} &= q_{n+1} - 2\alpha \\
&= \tfrac{1}{2} q_n + \alpha - 2\alpha \qquad \text{(using (1.2))} \\
&= \tfrac{1}{2}\, x_n \,,
\end{aligned}$$

So, the difference between q_n and $\overline{q}$ is a geometric progression with ratio $1/2$. Thus $x_n = x_0 2^{-n}$, and therefore

$$q_n = 2\alpha + (q_0 - 2\alpha) \cdot 2^{-n} .$$

It is interesting to observe that $x_n \to 0$, and therefore $q_n \to \overline{q}$, when $n \to \infty$. So, the amount of radioactive nuclei converges to the stationary value independently on its initial value.

ex: After how much time does the radioactive substance decrease to $\frac{1}{32}$-th of its initial value?

ex: Half-life of ^{14}C is estimated to be $\tau \simeq 5730$ years. Show how to date a fossil, assuming that we know the proportion of ^{14}C in a living being. [2]

e.g. **Exponential growth.** Exponential growth of populations in a illimited environment is modeled by the recursion

$$p_{n+1} = \lambda p_n ,$$

where p_n represent the population at time n (measured in units of some fixed time interval $\tau > 0$), given an initial population p_0. The meaning of the parameter λ is the following: at every time interval τ, the increase $p_{n+1} - p_n$ of the population is equal the "offspring" αp_n, where $\alpha > 0$ is some "fertility" coefficient, minus the "deaths" βp_n, where $\beta > 0$ is some "mortality" coefficient. Thus, $\lambda = \alpha - \beta$. An applet with the simulations is in exponentialgrowth.

ex: Discuss the behaviour of solutions p_n for different values of λ.

ex: To a population growing exponentially is added or retired a certain amount β each time interval τ. Te model is therefore

$$p_{n+1} = \lambda p_n + \beta ,$$

where β is a positive or negative parameter. Find the stationary solution, and then the solution with arbitrary initial condition p_0 (consider the substitution $x_n = p_n - \overline{p}$, where $\overline{p}$ is the stationary solution).

ex: For which values of λ and β do the solution p_n converge to the stationary solution when $n \to \infty$?

e.g. **Growth of Fibonacci numbers.** How fast do Fibonacci numbers grow? Define the quotients $q_n := f_{n+1}/f_n$ between neighbor Fibonacci numbers. From (1.1) one deduce the recursive equation

$$q_{n+1} = 1 + 1/q_n \tag{1.3}$$

for the q_n's. We compute:

$$1\,, \quad 2\,, \quad 3/2 = 1.5\,, \quad 5/3 \simeq 1.66666\,, \quad 8/5 = 1.6\,, \quad 13/8 = 1.625\,, \quad 21/13 \simeq 1.61538\,, \quad \dots$$

You may observe the sequence in the following applet. It turns out that the sequence (q_n) converge (try to prove it!), namely, $q_n \to \phi$ as $n \to \infty$. Taking the limits in the recursive equation (1.3) we see that $\phi = 1 + 1/\phi$, and therefore ϕ is the positive root of the quadratic polynomial $x^2 - x - 1$,

$$\phi = \frac{1+\sqrt{5}}{2} \simeq 1.6180339887498948482\dots$$

Hence, for large values of n we may approximate Fibonacci law as

$$f_{n+1} \approx \phi f_n ,$$

an exponential growth with rate ϕ. In particular, we expect $f_n \sim \phi^n$.

The limit ϕ is a famous irrational, the Greeks' "ratio/proportion". As described by Euclid[3]:

[2] J.R. Arnold and W.F. Libby, Age determinations by Radiocarbon Content: Checks with Samples of Known Ages, *Sciences* **110** (1949), 1127-1151.

[3] Euclid, *Elements*, Book VI, Definition 3.

> *"A straight line is said to have been cut in extreme and mean ratio when, as the whole line is to the greater segment, so is the greater to the less."*

If a is the greater part and b the less of a line of lenght $a+b$, Euclid's requirement is

$$\frac{a+b}{a} = \frac{a}{b}$$

There follows that the ratio $\phi = a/b$ satisfies $1 + 1/\phi = \phi$. This division of an interval is used in Book IV of the Elements to construct a regular pentagon. Observe that, as follows from the quadratic equation, ϕ^{-1} is equal to $\phi - 1$.

ex: Show that ϕ is irrational using its geometric definition (see Euclid's *Elements*, or [HW59] section 4.6.)

e.g. **The invention of chess.** Legend says that Sissa invented chess, and offered the game to the king of Persia. Asked for a reward, he suggested that he wanted one grain of rice on the first square of the chessboard, two grains on the second, four grains on the third, and so on. The king didn't take it seriously, but a computation shows that the reward amounts to

$$1 + 2 + 4 + 8 + \cdots + 2^{63} \simeq 1.84 \times 10^{19}$$

grains of rice. Now, if 1 Kg of rice contains something like 30000 grains, the above number amounts to roughly 6.13×10^{11} tons of rise (which you may want to compare with People's Republic of China's production in 2017, which has been, according to FAO, about 2.14×10^{8} tons!).

Series. A *series* is a formal infinite sum $\sum_{n=0}^{\infty} x_n$, or $\sum_{n\geq 0} x_n$, where the $x_n \in \mathbb{R}$ are elements of some given real (or complex) sequence. If the sequence (s_n) of *partial sums*, defined as $s_n := \sum_{k=0}^{n} x_k$ (which are honest numbers) converges to some limit, say $\lim_{n\to\infty} s_n = s$, then we say the series is *convergent* (or *summable*), and that its *sum* is $\sum_{n\geq 0} x_n := s$.

A series $\sum_n x_n$ is *absolutely convergent* is the series $\sum_n |x_n|$, formed with the absolute values of its terms, is convergent. Of course, absolute convergence is stronger than mere convergence. Indeed, convergent but not absolutely convergent series are quite interesting and strange objects[4] (see, for example, the last book by Hardy [Har49]).

e.g. **Harmonic series.** The *harmonic series*

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \dots$$

diverges. Indeed, its generic term $1/n$, for $n \geq 1$, is bigger than the integral $\int_n^{n+1} dx/x$, hence the partial sums $\sum_{k=1}^{n} 1/k$ are greater than the logarithm $\log(n+1)$.

Geometric series. The identity $(1 + \lambda + \lambda^2 + \lambda^3 + ... + \lambda^n)(\lambda - 1) = \lambda^{n+1} - 1$ shows that, if $\lambda \neq 1$, the sum of the first $n+1$ terms of the geometric progression (with $a = 1$) is

$$1 + \lambda + \lambda^2 + \lambda^3 + ... + \lambda^n = \frac{\lambda^{n+1} - 1}{\lambda - 1}$$

In particular, when $|\lambda| < 1$, the *geometric series* $\sum_{n=0}^{\infty} \lambda^n$ is absolutely convergent, and its sum is

$$1 + \lambda + \lambda^2 + \lambda^3 + ... + \lambda^n + ... = \frac{1}{1-\lambda}\,.$$

e.g. **Dichotomy paradox.** Using the above formula for the sum of the geometric series, you may try to convince Zeno that

$$1/2 + 1/4 + 1/8 + 1/16 + 1/32 + ... = 1\,.$$

[4] According to Abel (1828), "divergent series are the invention of the devil, and it is shameful to base on them any demonstration whatsoever."

e.g. Decimal expansions. Also, you may convince yourself that $0.99999\dots$, which by definition is the sum of the series

$$\frac{9}{10}+\frac{9}{100}+\frac{9}{1000}+\frac{9}{10000}+\dots$$

is actually equal to 1. Moreover, you may learn how to recognize rational numbers as $0.33333\dots$ or $1.285714285714\dots$ from their periodic expansion. Indeed, a real number is rational if and only if its base 10 (or any other base $d \geq 2$) expansion is eventually periodic.

ex: Say if the following series are convergent, and, if so, compute their sum.

$$1+1/2+1/4+1/8+1/16+\dots \qquad 1+10+100+1000+\dots \qquad 1+1/10+1/100+1/1000+\dots$$

$$\sum_{n=0}^{\infty}(4/5)^n \qquad 9/10+9/100+9/1000+\dots \qquad 0.3333\dots$$

Convergence tests. Deciding convergence or divergence of a series is not easy. The only tool at our disposal is comparison with known series, and essentially the only known non-trivial series is the geometric one. Comparison means the obvious observation that $0 \leq x_n \leq y_n$ for any n sufficiently large implies the following two conclusions: $\sum_n y_n < \infty \Rightarrow \sum_n x_n < \infty$, and $\sum_n x_n = \infty \Rightarrow \sum_n y_n = \infty$.

Now, if $|x_n| \leq C\lambda^n$ for some constant $C > 0$ and any n sufficiently large, then the partial sums of the series $\sum_n x_n$ are bounded by a constant times the partial sums of the geometric series $\sum_n \lambda^n$, therefore the series $\sum_n x_n$ is absolutely convergent whenever $|\lambda| < 1$. This happens when $\limsup_{n\to\infty} |x_n|^{1/n} < 1$ (*root test*) or when $\limsup_{n\to\infty} |x_{n+1}/x_n| < 1$ (*ratio test*).

1.2 Babylonians-Heron method to compute square roots

Consider the problem to find the side ℓ of a square given the value $a > 0$ of its area, i.e. to find the number which we call $\ell = \sqrt{a}$.

Babylonian-Heron algorithm. A clever method, descibed by Heron [5], but probably already used by Babylonians [6] [7], is as follows. We start with a rectangle with basis x_0 and height y_0, "simple" numbers such that $x_0 y_0 = a$ (for example, if the area is an integer like $a = 2$, we may start with $x_0 = 3/2$ and $y_0 = 4/3$). We choose a second rectangle is such a way that its sides are nearer than the sides of the first rectangle. An obvius way to do it is to take as new basis the arithmetic mean $x_1 = (x_0 + y_0)/2$, which forces to take $y_1 = a/x_1$ as second height. And so on, if we are not satisfied yet. The recursion for that basis reads

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right).$$

Observe that if both a and the initial conjecture x_0 are rationals (the only numbers known to Babylonians), then all the x_n's are also rationals.

The algorithm converges, and quite fast. We could, as the Babylonians, put an initial guess $x_0 = 3/2$ for $\sqrt{2}$ (quite reasonable, since $1^2 < 2 < 2^2$), and find

$$x_1 = \frac{17}{12} \simeq 1.41666666666 \qquad x_2 = \frac{577}{408} \simeq 1.41421568627 \qquad x_3 = \frac{665857}{470832} \simeq 1.41421356237$$

As you see, the sequence stabilizes quite fast.

[5] *"Since 720 has not its side rational, we can obtain its side within a very small difference as follows. Since the next succeeding square number is 729, which has 27 for its side, divide 720 by 27. This gives 26 2/3. Add 27 to this, making 53 2/3, and take half this or 26 5/6. The side of 720 will therefore be very nearly 26 5/6. In fact, if we multiply 26 5/6 by itself, the product is 720 1/36, so the difference in the square is 1/36. If we desire to make the difference smaller still than 1/36, we shall take 720 1/36 instead of 729 (or rather we should take 26 5/6 instead of 27), and by proceeding in the same way we shall find the resulting difference much less than 1/36."* Heron of Alexandria, *Metrica*, Book I.

[6] Carl B. Boyer, *A history of mathematics*, John Wiley & Sons, 1968.

[7] O. Neugebauer, *The exact sciences in antiquity*, Dover, 1969.

Error estimate. As a first attempt to explain this miracle, we could start looking at the recursive equations for the bases and the heights of the rectangles:

$$x_{n+1} = \frac{x_n + y_n}{2} \qquad 1/y_{n+1} = \frac{1/x_n + 1/y_n}{2}$$

(so, the next height is the "harmonic mean" of the base and height). We see that the x_n's and the y_n's form decreasing and increasing sequences, respectively (disregarding the first guess, of course), namely

$$y_2 \leq y_3 \leq \cdots \leq y_n \leq \cdots \leq x_n \leq \cdots \leq x_3 \leq x_2 \,,$$

The real root is somewhere between, namely $y_n \leq \sqrt{a} \leq x_n$. Hence, we have an explicit control of the error. A computation shows that the lenghts of those intervals, the differences $\varepsilon_n = x_n - y_n$ satisfy the recursion

$$\varepsilon_{n+1} < \frac{1}{2} \cdot \varepsilon_n$$

So, and initial "error" $\varepsilon_1 \leq 1$ (an easy achievement, since we easily recognize squares of integers) reduces to at least $\varepsilon_n \leq 2^{-n}$ after n iterations. The true error is actually much smaller. Indeed, in our example we may compute

$$\varepsilon_1 = \frac{17}{12} - 2\frac{12}{17} = \frac{1}{204} \simeq .005 \qquad \text{and} \qquad \varepsilon_2 = \frac{577}{408} - 2\frac{408}{577} = \frac{1}{235416} \simeq 0.000004$$

So that the first improved guess x_1 has already one correct decimals, and the second, x_2 has already four correct decimals!

Irrationals. What Babylonians didn't suspect is that if you start with a rational guess for $\sqrt{2}$, you get an infinite sequence of rational approximations, but the process never stops. This is due to

Theorem 1.1 (Pythagoras)**.** *The square root of* 2 *is not rational.*

ex: A formula by Heron says that the area of a triangle with sides of lenghts a, b and c, and semi-perimeter $s = (a + b + c)/2$, is given by

$$\mathrm{A} = \sqrt{s(s-a)(s-b)(s-c)}$$

Estimate the area of a triangle with sides 7, 8 and 9.

ex: Estimate $\sqrt{13}$ with an error < 0.01 or 0.001.

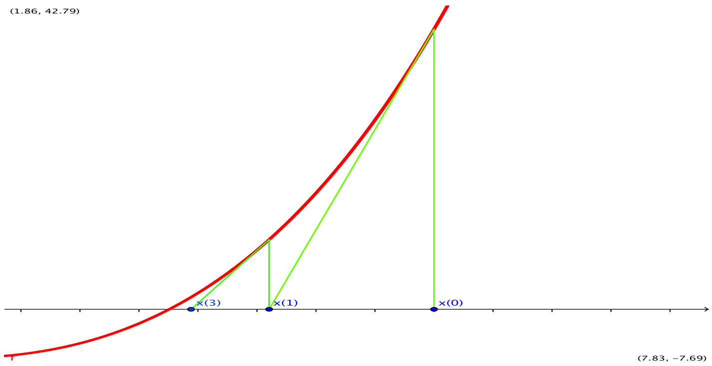
ex: Estimate how many iterations are necessary to obtain the first n correct decimals of $\sqrt{2}$ using Babylonians' method.

ex: Prove Pythagora's theorem 1.1 above (take a look at [HW59]).

1.3 From Newton method to Julia and Fatou sets

Finding $\sqrt{a}$ means solving the polynomial equation $z^2 - a = 0$. What about finding roots of a generic polynomial $p(x) \in \mathbb{R}[x]$?

Newton-Raphson iterative scheme. "Newton method" is a method proposed by Joseph Raphson around 1690 to approximate roots of a polynomial $p(x)$ (Newton used it to solve $x^3 - 2x - 5 = 0$). It consists in starting with an initial conjecture x_0 near to some root, and then improve it using the linear approximation $p(x_0) + p'(x_0)(x - x_0)$.



Search for a root of $x^3 - 2x - 5$ using Newton iterations.

This idea leads to the recursion

$$x_{n+1} = x_n - \frac{p(x_n)}{p'(x_n)}\,.$$

It is clear that if the sequence converges, i.e. $x_n \to x_\infty$, and if $p'(x_\infty) \neq 0$, then the limit x_∞ is a root.

ex: Use Newton method to solve Newton's problem, i.e. find the roots of $x^3 - 2x - 5$.

ex: Show that Newton method to solve $x^2 - a = 0$ corresponds to babylonian-Heron iterative scheme.

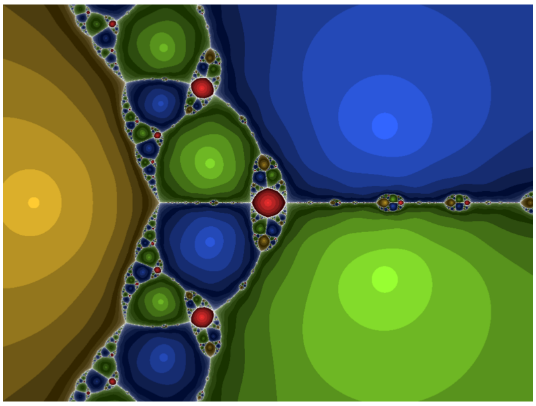
ex: Use Newton method to approximate the Greeks' *ratio*, the positive root of $x^2 - x - 1$. Then, compare with the babylonian-Heron method (i.e., estimate $\sqrt{5}$, then sum 1 and divide by 2).

ex: Write and implement Newton method to find n-th roots, i.e. to solve $x^n - a = 0$.

Newton's fractals. In 1879 Cayley observed that the above method could be also used to approximate complex roots of complex polynomials $p(z) \in \mathbb{C}[z]$. It amounts to iterate the rational function

$$f(z) = z - \frac{p(z)}{p'(z)}$$

The problem is therefore to understand when, i.e for which initial values z_0, the sequence z_n converges to one of the roots. The "basins of attraction" of the different roots draw beautiful and unexpected patterns in the complex plane.



Basins of attraction of the roots of $2z^3 - 2z + 2$ in $\mathbb{C}$
(from http://en.wikipedia.org/wiki/Newton_fractal).

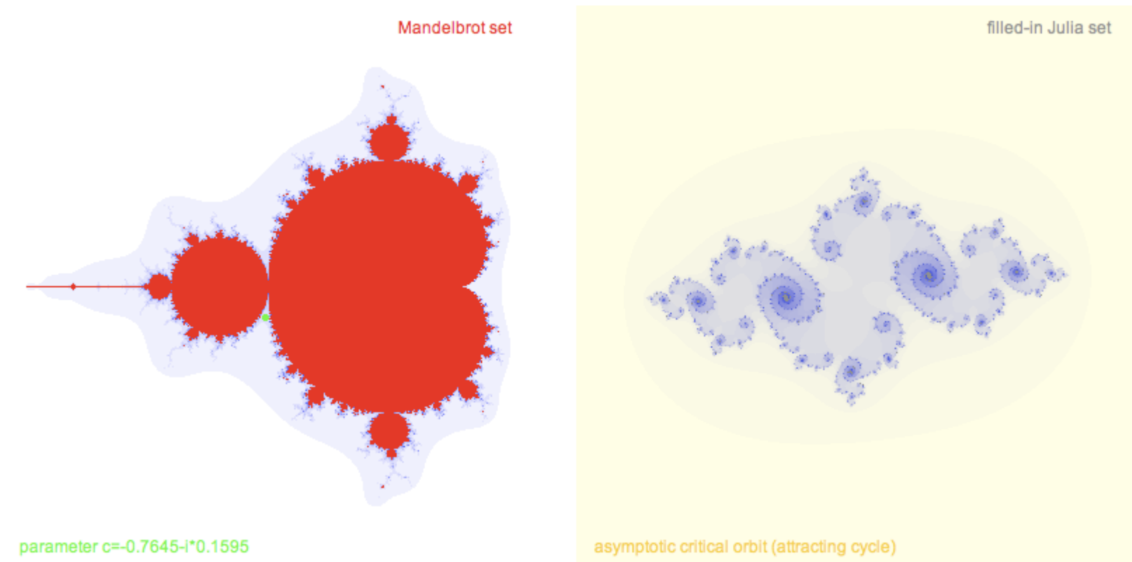
Iteration of rational functions in the Riemann sphere. The natural generalization is to take a rational function $f(z) \in \mathbb{C}(z)$, which is an endomorphism of the Riemann sphere $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, i.e. try to understand its rajectories, i.e. the iteration $z_{n+1} = f(z_n)$.

Most studied is iteration of the family of quadratic polynomials

$$f(z) = z^2 + c$$

depending on a parameter $c \in \mathbb{C}$. Its beauty was foreseen by Gaston Julia[8] and Pierre Fatou[9] at the beginning of the XX century, revealed with the help of the first modern computers by Benoît Madelbrot, and then studied by a variety of great mathematicians (like Adrian Douady, Dennis Sullivan, John Milnor, Misha Lyubich, Jean-Christophe Yoccoz, Curtis McMullen, ...) starting from the 80's of the last century.

Below, you may find a picture of what Julia and Fatou could only dream about.



Mandelbrot set (left) and Julia set (right) of the polynomial $z^2 + c$ with $c \simeq -0.7645 - i \cdot 0.1595$.
(from http://w3.math.uminho.pt/~scosentino/bestiario/julia.html)

The red hearts on the left form the *Mandelbrot set*, the set of those values of the parameter c such the orbit of critical point $z_0 = 0$ is bounded. The almost invisible grey points on the right form the *filled-in Julia set*, the set of initial values z_0 with bounded orbits (once fixed a value of

[8] G. Julia, Mémoire sur l'iteration des fonctions rationnelles, *Journal de Mathématiques Pures et Appliquées*, **8** (1918), 47-245.

[9] P. Fatou, Sur les substitutions rationnelles, *Comptes Rendus de l'Académie des Sciences de Paris*, **164** (1917) 806-808, and **165** (1917), 992-995.

c). Blue colors, which help to to see the Julia set, are chosen depending on the speed with which other trajectories diverge to ∞.

Much more beautiful pictures, and then movies and so on, may be found in this page by Jos Leys: http://www.josleys.com

1.4 Finite difference equations

Fibonacci model is the prototype of

Recursive linear equations. A *recursive linear equation* (or "finite difference linear equation") is a law

$$a_p x_{n+p} + a_{p-1} x_{n+p-1} + \cdots + a_1 x_{n+1} + a_0 x_n = f_n \tag{1.4}$$

which defines a sequence (x_n) given a set of "initial conditions" $x_0, x_1, \ldots, x_{p-1}$ and the known sequence (external force) f_n . Above, $a_0 \neq 0, a_1, \ldots, a_{p-1}, a_p \neq 0$ are real or complex parameters. It is a discrete version of a linear ordinary differential equation of degree p with constant coefficients. When $f_n = 0$ for all n, we get a *homogeneous recursive equation*

$$a_p x_{n+p} + a_{p-1} x_{n+p-1} + \cdots + a_1 x_{n+1} + a_0 x_n = 0\,. \tag{1.5}$$

The set of solutions of the homogeneous equation (1.5) is a vector space $\mathcal{H}$ of dimension p, and the set of solutions of (1.4) is an affine space modeled on $\mathcal{H}$, i.e. has the form $(z_n) + \mathcal{H}$, where (z_n) is any (particular) solution of (1.4).

Eigenfunctions. The general recipe is: "linear homogeneous equations have exponential solutions". The conjecture $x_n = z^n$ solves the recursive equation (1.5) if z is a root of the *characteristic polynomial*

$$P(z) = a_p z^p + a_{p-1} z^{p-1} + \cdots + a_1 z + a_0$$

In particular, if P has p distinct roots (in $\mathbb{C}$), say $z_1, z_2, \ldots, z_p$, then the general solution of the homogeneous equation is a linear combination

$$x_n = c_1 z_1^n + c_2 z_2^n + \cdots + c_p z_p^n$$

where the $c_1, c_2, \ldots, c_p$ are constants which depend on the initial conditions $x_0, x_1, \ldots x_{p-1}$.

ex: Find an explicit formula for the Fibonacci numbers f_n's (which is known as *Binet's formula*).

ex: Discuss what happens when there are non-simple roots.

Generating functions. Given a sequence (x_n), defined anyway, we may consider the (formal) power series

$$F(z) := \sum_{n \geq 0} x_n z^n$$

If the series has a non-zero radius of convergence (since the radius of convergence R is given by Hadamard formula $1/R = \limsup_{n\to\infty} \sqrt[n]{|x_n|}$, this happens when the x_n's grow at most exponentially, i.e. when $|x_n| \leq C\lambda^n$ for some $C > 0$ and $\lambda > 0$), it defines an analytic function $F(z)$ in some neighborhood of the origin. Then, the original sequence may be recovered computing derivatives, since

$$x_n = \frac{F^{(n)}(0)}{n!}\,.$$

For this reason, $F(z)$ is called *generating function* of the sequence (x_n).

You may find interesting the following characterization of rational functions.

Theorem 1.2. *A power series $\sum_{n\geq 0} x_n z^n$ represents a rational function $F(z) = \frac{P(z)}{Q(z)} \in \mathbb{C}(z)$ iff the coefficients x_n satisfy a recursive linear homogeneous equation.*

e.g. Generating function of the Fibonacci numbers. If f_n denotes the n-th Fibonacci number, starting from $f_0 = f_1 = 1$, then the power series $\sum_{n\geq 0} f_n z^n$ represents the rational function

$$F(z) = \frac{1}{1 - z - z^2}$$

in a neighborhood of the origin. Observe that it has a pole with smallest absolute value at $1/\phi$, and deduce that $\limsup_{n\to\infty} |f_n|^{1/n} = \phi$ (so that $f_n \sim \phi^n$, as we already knew).

ex: Give examples of sequences which do not satisfy any (finite) recursion.

ex: Consider the recursive equaiton

$$x_{n+2} = 2x_{n+1} + x_n \, .$$

Find the geral solution. Find the solution with $x_0 = 0$ and $x_1 = 1$, and compute explicitely the first few terms of the sequence. Show that the quotients $q_n := x_{n+1}/x_n$ converge to $1 + \sqrt{2}$ when $n \to \infty$, and therefore

$$\frac{x_{n+1} - x_n}{x_n} \to \sqrt{2}$$

Obtain rational approximations of $\sqrt{2}$.

Recursive systems. A linear homogeneous recursive system is a law

$$x_{n+1} = A x_n$$

for some vector valued sequence $x_n \in \mathbb{R}^k$, given a square matrix $A \in \mathrm{Mat}_{k\times k}(\mathbb{R})$. The solution is

$$x_n = A^n x_0 \, ,$$

where $x_0 \in \mathbb{R}^k$ is the initial condition. The computation of powers A^n of a square matrix A is simplified if we can diagonalize it. For example, if the matrix has k distinct and real eigenvalues, then in the basis formed by the eigenvectors it is a diagonal matrix, say $A = \mathrm{diag}(\lambda_1, \dots, \lambda_k)$, and its n-th power is simply the diagonal matrix $A^n = \mathrm{diag}(\lambda_1^n, \dots, \lambda_k^n)$.

A finite difference equation of order p like

$$a_p y_{n+p} + a_{p-1} y_{n+p-1} + \cdots + a_1 y_{n+1} + a_0 y_n = 0$$

is equivalent to a recursive linear homogeneous system $x_{n+1} = A x_n$ for the vector values sequence $x_n := (y_n, y_{n-1}, \dots, y_{n-p-1})$.

ex: Write and solve the system which corresponds to Fibonacci problem.

e.g. Arithmetic-geometric mean. Given two positive numbers x and y, define recursively

$$a_{n+1} = \frac{1}{2}(a_n + g_n) \qquad g_{n+1} = \sqrt{a_n \, g_n} \, ,$$

starting with $a_0 = (x+y)/2$ and $g_0 = \sqrt{x\, y}$, the arithmetic and the geometric mean of x and y, respectively. The arithmetic-geometric mean inequality (the fact that $(x+y)^2 \geq 0$) says that $g_n \leq a_n$, and therefore

$$g_{n+1} = \sqrt{a_n \, g_n} \geq \sqrt{g_n \, g_n} = g_n$$

Since both sequences a_n and g_n are between the minimum and the maximum of x and y, this implies that g_n converges, to some (positive) limit p. The sequence a_n also converges, and to the same limit, since

$$a_n = g_{n+1}^2 / g_n \to p$$

The common limit is called *arithmetic-geometric mean* of x and y, say $p =: \mathrm{AGM}(x, y)$. What is not trivial is a formula for the limit, and this is due to Gauss: it says that

$$\mathrm{AGM}(x, y) = \frac{\pi}{4} \frac{x + y}{K\left(\frac{x-y}{x+y}\right)}$$

where

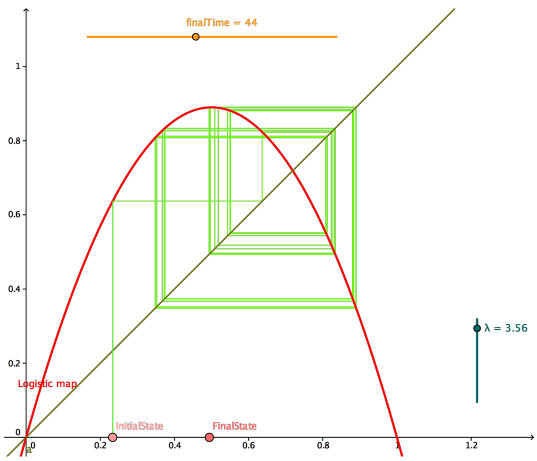
$$K(k) := \int_0^{\pi/2} \frac{d\theta}{\sqrt{1 - k^2 \sin^2\theta}}$$

is the "complete elliptic integral of the first kind".

1.5 Interval maps and cobweb plot

Graphical analysis. Consider a transformation $f : I \to \mathbb{R}$ defined in an interval I of the real line. Iteration is possible when $f(I) \subset I$. One can follows trajectories using a "cobweb plot": drawing vertical and horizontal lines connecting the points

$$(x, f(x)) \mapsto (f(x), f(x)) \mapsto (f(x), f^2(x)) \mapsto (f^2(x), f^2(x)) \mapsto (f^2(x), f^3(x)) \mapsto \ldots$$



Cobweb plot of the quadratic map $f(x) = \lambda x(1 - x)$ when $\lambda = 3.56$.

e.g. **Affine interval maps.** As we have already seen, affine maps behave quite predictably. Indeed, the trajectories of an affine map like

$$f(x) = \lambda x + \alpha$$

with $\lambda \neq 1$, are sent, by the change of variable $y = x - \overline{x}$, where $\overline{x} = \alpha/(1 - \lambda)$ is the stationary solution, into the trajectories of $g(y) = \lambda y$, and the latter are geometric sequencies. If $\lambda = 1$, trajectories are simply arithmetic series.

e.g. **The quadratic family.** As soon as the interval map is not affine, trajectories are not easily understood. The simplest interval maps which are not affine are quadratic polynomials. Consider the *quadratic family*, the collection of interval maps $f_\lambda : \mathbb{R} \to \mathbb{R}$ defined by

$$f_\lambda(x) := \lambda x (1 - x) \tag{1.6}$$

depending on a (real) parameter λ. For $0 \leq \lambda \leq 4$, formula (1.6) defines a transformation of the unit interval, that we denote by the same symbol $f_\lambda : [0, 1] \to [0, 1]$. For small λ, the trajectories are previsible. As λ approaches 4, they become quite wild.

ex: Try to understand the dynamic of the following maps, defined in convenient intervals (some are easy, other are hard, if not impossible).

$$f(x) = \pm x^3 \qquad f(x) = x^{1/3} \qquad f(x) = x^3 \pm x$$

$$f(x) = x^2 + 1/4 \qquad f(x) = |1 - x| \qquad f(x) = x^2 - 2 \qquad f(x) = \sin x \qquad f(x) = \cos x$$

$$f(x) = x(1 - x) \qquad f(x) = 2x(1 - x) \qquad f(x) = 3x(1 - x) \qquad f(x) = 4x(1 - x)$$

2 Topological dynamical systems, basic definitions

2.1 Transformations

Transformations. In these notes, we'll be mainly interested in discrete time dynamical systems, i.e. actions of $\mathbb{N}_0$ or $\mathbb{Z}$ on some space X, generated by a transformation/map

$$f : X \to X \,.$$

Apart from some special cases, X will be a topological space (or even a metric space). The transformation will be continuous, or at least piecewise continuous. In such cases we speak of *topological dynamical system*.

The "(forward) iterates" of a transformation f are the transformations $f^n : X \to X$, with $n \in \mathbb{N}_0$, defined inductively according to

$$f^0 = \mathrm{id} \qquad \text{and} \qquad f^{n+1} = f \circ f^n \qquad \text{if } n \geq 0$$

(warning! with this notation $f^2(x)$ is not the square of $f(x)$, but $f(f(x)) \ldots$).

In general, if $n \in \mathbb{N}$ and $A \subset X$, then $f^{-n}(A)$ denotes the set

$$f^{-n}(A) = \{x \in X \text{ t.q. } f^n(x) \in A\} \,.$$

If f is invertible (e.g. is an homeomorphism), we can also define the backward iterates, and therefore the transformations $f^n : X \to X$ for all $n \in \mathbb{Z}$.

We have therefore an action $\Phi : \mathbb{N}_0 \times X \to X$, or $\Phi : \mathbb{Z} \times X \to X$ if f is invertible, defined by $\Phi_n(x) = f^n(x)$.

Phase/states space. In the following, (X, d) will be a metric space equipped with its natural topology τ, locally compact (any point admits a compact neighborhood) and separable (admits a countable dense subset, and therefore, being a metric space, a countable basis for the topology). For example, regions of $\mathbb{R}^n$, intervals of the line, the circle $\mathbb{R}/\mathbb{Z}$, the torus $\mathbb{R}^n/\mathbb{Z}^n$, the complex plane $\mathbb{C}$, the Riemann sphere $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, Cantor sets, and Cartesian products of finite spaces. Also, in order to avoid trivialities, we'll always assume tacitly that X is not a finite set.

Translations in homogeneous spaces. The simplest, tautological, way to build actions is algebraic. Let G be a topological group (a group equipped with a Hausdorff topology such that the group operations $(g, g') \mapsto gg'$ and $g \mapsto g^{-1}$ are continuous). Given a closed subgroup $\Gamma \subset G$, one can consider the homogeneous space $X = G/\Gamma = \{g\Gamma\,;\, g \in G\}$, equipped with the quotient topology (the finest topology in G/Γ such that the projection $\pi : G \to G/\Gamma$ is continuous). If Γ is not too large or wild, for example if Γ is discrete, X is a sufficiently large and interesting space.

Every subgroup $S \subset G$ acts on the homogeneous space $X = G/\Gamma$, the action $S \times G/\Gamma \to G/\Gamma$ being $(s, g\Gamma) \mapsto sg\Gamma$. The space of orbits is the quotient $S\backslash G/\Gamma$.

In particular, a cyclic subgroup $S = \{s^n\}_{n\in\mathbb{Z}}$ generates an action $\Phi : \mathbb{Z} \times X \to X$ defined by $\Phi_n(g\Gamma) = s^n g\Gamma$, which consists in iterating the left translations $g\Gamma \mapsto sg\Gamma$ of a generator.

e.g. **Translations of the torus.** The n-dimensional *torus* is the quotient space $\mathbb{T}^n := \mathbb{R}^n/\mathbb{Z}^n$ of the additive Abelian group $\mathbb{R}^n$ modulo the discrete subgroup $\mathbb{Z}^n$ of integer vectors. Observe that the torus is itself an Abelian group. The one-dimensional case $\mathbb{T}^1 = \mathbb{R}/\mathbb{Z}$ is also called *circle*, because it is isomorphic to the unit circle of the complex plane, the multiplicative Abelian group $\mathbb{S}^1 = \{z \in \mathbb{C}\,,\, |z| = 1\}$, under the eponential map $x \mapsto e^{2\pi i x}$.

Any $\alpha \in \mathbb{R}^n$ defines a translation $T_\alpha : \mathbb{R}^n \to \mathbb{R}^n$, according to $T_\alpha(x) = x + \alpha$. As explained above, the translation defines a *rotation* $R_\alpha : \mathbb{T}^n \to \mathbb{T}^n$, according to $R_\alpha(x + \mathbb{Z}^n) := x + \alpha + \mathbb{Z}^n$.

Generic properties. We will often want to talk about "most trajectories", or "almost all trajectories".

Being X a topological space, one could consider (probability or infinite) measures on the Borel σ-algebra of X. Given such a measure μ, one says that a properties is satisfied fo μ-almost all points if the subset $N \subset X$ of those points which do not have the property has measure $\mu(N) = 0$.

The topological couterpart of the dichotomy "zero-one probability" is possible when X is a *Baire space*, i.e. a Hausdorff (any two distinct points have dosjoint neighborhoods) topological space where a countable intersection of dense open sets is dense. Baire theorem says that examples of Baire spaces are complete metric spaces. A subset $R \subset X$ is said *residual* if it contains a countable intersection of dense open sets. A subset $M \subset X$ is said *meager* if it is a countable union of "nowhere dense" subsets (subsets such that the closure has empty inteiror), i.e. if its complementar $X \backslash M$ is residual. A property is said *generic* if the subset $P \subset X$ of those points with this property is residual.

2.2 Trajectories and orbits

Trajectories. Given a transformation $f : X \to X$, we are mainly interested in the asympthotic behavior of the "history" of a point $x \in X$, the sequence of points

$$x \mapsto f(x) \mapsto f^2(x) \mapsto f^3(x) \mapsto ...$$

obtained recursively applying f to the point x. If X is the space state of a physical system, and if the system is (prepared) in the state x at time $t = 0$, then it will be in the state $f(x)$ at time 1, in the state $f^2(x) = f(f(x))$ at time 2, and so on.

The *trajectory* of $x \in X$ is the sequence $(x_n)_{n\in\mathbb{N}_0}$, the function that, given the "initial condition" $x_0 = x$, produces the states $x_n = f^n(x)$ of the system at each time $n \geq 0$. Thus, the trajectory of x is the solution of the recurrence

$$x_{n+1} = f(x_n)$$

with initial condition $x_0 = x$.

Orbits. The *(forward) orbit* of $x \in X$ is the image of its trajectory, i.e. the set

$$\mathcal{O}^+_f(x) := \{f^n(x)\}_{n\in\mathbb{N}_0}$$

(we put the supscript "+" to remind that we are only allowed to go forward in time, since in general f will not be invertible). It is the "future" of a point.

A point x may have more than one pre-image, and therefore its "past" is not unique. The *big orbit* of a point $x \in X$ is the set

$$\mathcal{GO}_f(x) := \{x' \in X : \exists\, n, m \geq 0 : f^n(x') = f^m(x)\}$$

i.e the set of points which have eventually the same future of x.

If f is invertible, we may also define the *complete orbit* of a point x as

$$\mathcal{O}_f(x) := \{f^n(x)\}_{n\in\mathbb{Z}}$$

(which actually coincide with the big-orbit), the past and future of a point.

Observe that "being in the same big-orbit" is an equivalence relation, and therefore X is a disjoint union of equialence classes, i.e. orbits. It must be said that the quotient space, the space of orbits X/f, may be messy if trajectories are not regular (and this is when things get interesting!). For example, if there exists a dense orbit, then the quotient topology in X/f is the trivial topology. Thus, the space of orbits, as a topological space, does not contain much informations on the dynamics of the system.

2.3 Periodic orbits and basin of attraction

Fixed points. The simplest orbits are (composed of) *fixed points* of f, those states $p \in X$ such that

$$f(p) = p\,.$$

Geometrically, fixed points are the intersections of the graph of f with the "diagonal" $\Delta \subset X \times X$. If X is a linear space, fixed points are roots of the equation $f(x) - x = 0$.

Periodic orbits. A point $p \in X$ is said *periodic* with *period* $n \geq 1$ if $f^n(p) = p$ and n is the smallest of those times $k \geq 1$ such that $f^k(p) = p$. Thus, the orbit of the periodic point p is a *cycle*, a finite set

$$\left\{p,\ f(p), f^2(p), ..., f^{n-1}(p)\right\}$$

of points which are permuted by the transformation f.

A point x may have a finite orbit without being periodic: this happens when there exists a time $k \geq 1$ such that $f^k(x)$ is a periodic point. Such points are called *pre-periodic*.

$\mathrm{Fix}(f^n)$ denotes the set of those fixed points of the transformation f^n, that is the set of those periodic points of f whose period divides n.

$$\mathrm{Per}_f = \cup_{n \geq 1} \mathrm{Fix}(f^n)$$

denotes the set of periodic points of the map f. Observe that any of the sets $\mathrm{Fix}(f^n)$ is closed, because f^n is continuous, but their union Per_f may not be closed.

Convergent trajectories. If a trajectory is convergent, then its limit is a fixed point of f. Indeed, if $f^n(x) \to p$, the continuity of f implies that

$$f(p) = f\left(\lim_{n\to\infty} f^n(x)\right) = \lim_{n\to\infty} f^{n+1}(x) = p\,.$$

Basin of attraction. Let p be a fixed point of $f : X \to X$. The *basin of attraction*, or *stable set*, of p is the set of those points $x \in X$ whose trajectories converge to p, i.e.

$$W^s(p) := \left\{x \in X \text{ t.q. } \lim_{n\to\infty} f^n(x) = p\right\}$$

Uniqueness of limits of convergent sequences in a metric space implies that stable sets of different fixed points are disjoint.

e.g. **Endomorphisms of linear spaces.** The simplest dynamical systems are endomorphisms of a linear space. For example, endomorphisms of $\mathbb{R}^n$ are defined, in the canonical basis, by matrices $A \in \mathrm{Mat}_{n\times n}(\mathbb{R})$, according to $f(x) = Ax$ (vectors are column vectors, and the product is the usual product between matrices). The origin is a fixed point, by linearity. Other fixed points are the eigenvectors with eigenvalue $\lambda = 1$, non-trivial solutions of the homogeneous equation $Ax = x$. Periodic points with period n are eigenvectors with eigenvalue λ such that $\lambda^n = 1$.

e.g. **Collatz/Kakutani/Syracuse problem.** Consider the *Collatz map* $f : \mathbb{N} \to \mathbb{N}$, defined as

$$f(n) = \begin{cases} n/2 & \text{if } n \text{ is even} \\ 3n+1 & \text{if } n \text{ is odd} \end{cases}$$

It is clear that $4 \mapsto 2 \mapsto 1$ is a cycle. *Collatz conjecture* ("...an extraordinarily difficult problem, completely out of reach of present day mathematics", according to Lagarias) affirms that this is the only cycle and that any initial condition will eventually fall in this cycle.

ex: (Dynamics on finite state spaces) Study the dynamics, i.e. the structure of orbits, of an arbitrary transformation of a finite set $X = \{1, 2, \ldots, n\}$. Observe that the study of the dynamics of invertible transformations consists essentially in the study of the symmetric group S_n, permutations of X.

ex: (Affine maps) Draw some orbits of the transformations of the complex plane $f : \mathbb{C} \to \mathbb{C}$ defined by

$$f(z) = z + \alpha \qquad \text{or} \qquad f(z) = \lambda z$$

for different values of the parameters $\alpha, \lambda \in \mathbb{C}$. Explain for which values of those parameters there exists periodic orbits.

ex: Find, when possible, periodic orbits (of small period) of the transformations of the interval

$$f(x) = \pm x^3 \qquad f(x) = x^{1/3} \qquad f(x) = x^3 \pm x$$

$$f(x) = x^2 + 1/4 \qquad f(x) = |1 - x| \qquad f(x) = x^2 - 2 \qquad f(x) = \sin x \qquad f(x) = \cos x$$

$$f(x) = x(1 - x) \qquad f(x) = 2x(1 - x) \qquad f(x) = 3x(1 - x) \qquad f(x) = 4x(1 - x)$$

ex: Find the basins of attraction of the fixed points of $f(x) = x^2$ and $f(x) = x^3$, considered as transformations of the real line $\mathbb{R}$.

ex: Let $f : \mathbb{R} \to \mathbb{R}$ be the linear homogeneous transformation of the line defined by $f(x) = \lambda x$. Study the basin of attraction of $p = 0$ dependning on the "multiplier" λ. Do the same for $f(z) = \lambda z$ defined in the complex line $\mathbb{C}$.

ex: Find the basin of attraction of the origin for the linear maps $f : \mathbb{R}^2 \to \mathbb{R}^2$ defined, in the canonical basis, by the following 2×2 real matrices:

$$\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \qquad \begin{pmatrix} 2 & 0 \\ 0 & 1/3 \end{pmatrix} \qquad \begin{pmatrix} 1/2 & 0 \\ 0 & 1/3 \end{pmatrix} \qquad \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \qquad \frac{1}{5}\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \qquad 2\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

e.g. Squaring complex numbers. Consider the transformation $f : \mathbb{C} \to \mathbb{C}$ of the complex plane defined by "squaring", i.e.

$$f(z) = z^2 \,.$$

The basin of attraction of the fixed point 0 is the unit disk $\mathbb{D} = \{|z| < 1\}$. Indeed, if $|z| = \lambda < 1$, then $|f^n(z)| = \lambda^{2^n} \to 0$ as $n \to \infty$. We may also extend f to an endomorphism of the Riemann sphere $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, and then, by the same reasoning, we see that the basin of attraction of ∞ is the exterior of the disk, the set $\mathbb{D}^- = \{|z| > 1\}$. Meanwhile, it is not obvious to describe the basin of attraction of the fixed point $p = 1$. It is clear that its basin belongs to the unit circle $\mathbb{S} = \{|z| = 1\}$, and that it contains ± 1, its square roots $\pm i$, the square roots of these points, and so on ... a countable and dense subset of the unit circle.

2.4 Observables

Observables. *Observáveis* are functions $\varphi : X \to \mathbb{R}$ or $\mathbb{C}$. If the system is initially in the state x, and therefore is observed the value $\varphi(x)$ of the observable φ, after a time n observation of φ will give the value $\varphi(f^n(x))$.

Invariant functions. Particularly interesting are observable which do not change with time, that physicists call *first integrals*. The function/observable $\varphi : X \to \mathbb{R}$ is *invariant* sif

$$\varphi \circ f = \varphi$$

oi.e. if it is constant in each orbit. Oserve thaif φ is invariant, $I \subset \mathbb{R}$ and $A = \varphi^{-1}(I)$, then $f^{-1}(A) = A$. The existence of an invariant function contains the following information: if we know that $\varphi(x) = a$, then future and past of X belong to the level set $\Sigma_a = \{x \in X \text{ t.q. } \varphi(x) = a\}$, i.e. $\mathcal{GO}_f(x) \subset \Sigma_a$. Invariant functions, therefore, reduce the allowed phase space of trajectories.

Lyapunov functions. Also useful are monotone observable, which increase or decrease along trajectories, known in physics as *Lyapunov functions*. For example, if we know that $\varphi \circ f \leq \varphi$, and $\varphi(x) = a$, then the future of x doeas not leave the sub-level set $\Sigma_{\leq a} = \{x \in X \text{ s.t. } \varphi(x) \leq a\}$, and the past of x comes from x $\Sigma_{\geq a} = \{x \in X \text{ s.t. } \varphi(x) \geq a\}$.

e.g. Energy. The energy $E(q,p) = p^2/2 + q^2/2$, which is a constant of the motion for the harmonic oscillator $\ddot{q} = -q$ (here $p = \dot{q}$), is a Lyapunov function for the dumped oscillator $\ddot{q} = -\alpha\dot{q} - q$, since its time derivative is $\frac{d}{dt}E = -\alpha p^2 \leq 0$.

ex: Show that, if $\varphi : X \to \mathbb{R}$ is invariant, $I \subset \mathbb{R}$ and $A = \varphi^{-1}(I)$, then $f^{-1}(A) = A$.

ex: Show that the characteristic function of a set $A \subset X$ is invariant iff $f^{-1}(A) = A$.

Time means. The *time mean* (or *Birkhoff mean*) of the observable φ up to time $n \geq 0$ is the observable $\overline{\varphi}_n$ defined by

$$\overline{\varphi}_n(x) := \frac{1}{n+1}\sum_{k=0}^{n} \varphi\left(f^k(x)\right)$$

i.e. the value of $\overline{\varphi}_n$ at the point x is the arithmetic mean of the values of φ along the "n-orbit of x", the set $\left\{x, f(x), f^2(x), ..., f^n(x)\right\}$. If the limit

$$\overline{\varphi}(x) = \lim_{n\to\infty} \overline{\varphi}_n(x)$$

exists, it has the meaning of "asympthotic mean value" of φ along the orbit of x. Also observe that $\overline{\varphi}(x) = (\overline{\varphi} \circ f)(x)$ at points where the limit exists.

If, in particular, 1_A denotes the characteristic function of a subset $A \subset X$, then the limit

$$\overline{1_A}(x) = \lim_{n\to\infty} \frac{1}{n+1} \text{card}\left\{0 \leq k \leq n \text{ s.t. } f^k(x) \in A\right\}$$

if it exists, represents the "asympthotic fraction of time that the trajectory of x spend inside A", i.e. the asymptotic "frequency" with which the trajectory of x visit the subset A.

2.5 Invariant sets

Invariant sets. The characteristic function of a subset $A \subset X$ is invariant iff $f^{-1}(A) = A$. This motivates the folowing definition: a subset $A \subset X$ is *invariant* if

$$f^{-1}(A) = A$$

This condition implies that $f(A) \subset A$, and therefore a point inside an invariant set has all its history, past and future, inside the invariant set.

Observe that $\mathcal{GO}_f(x)$ is the smaller invariant set which contains x, and therefore a subset is invariant iff it a union of big orbits. If f is invertible, $\mathcal{O}_f(x)$ is the smaller invariant set which contains x, so that a subset is invariant iff it is a union o complete orbits, i.e. if $A = \cup_{x\in A}\mathcal{O}_f(x)$.

We also say that a subset $A \subset X$ is $+$*invariant* (positively invariant) if $f(A) \subset A$, and $-$*invariant* (negatively invariant) if $f^{-1}(A) \subset A$. In particular, if A is +invariantt, it is possible to define the restriction of f to A, i.e. dynamical system $f|_A : A \to A$.

ex: Discover the possible implications between the conditions

$$f^{-1}(A) = A\,, \qquad f(A) \subset A\,, \qquad f^{-1}(A) \subset A\,,$$
$$f(A) = A\,, \qquad \text{e} \qquad f^{-1}(A) = A = f(A)$$

for a generic transformation, or a transfomation which is injective, surjective, or one-to-one.

ex: Consider a set C equal to $\mathcal{GO}_f(x)$, $\mathcal{O}_f(x)$ or $\mathcal{O}_f^+(x)$ for some $x \in X$, and determine the invariance properties of C, $\overline{C}$, ∂C and C'.

ex: Let $A \subset X$. Show that $\bigcup_{n\geq 0} f^n(A)$ is +invariant, indeed the smallest +invariant set which contains A.

If f is invertible, show that $\bigcup_{n\in\mathbb{Z}} f^n(A)$ is invariant, indeed the smallest invariant set which contains A.

ex: Let $\varphi : X \to \mathbb{R}$ be an observable, and $A \subset X$ be the set of those points $x \in X$ such that the limit $\overline{\varphi}(x) = \lim_{n\to\infty} \overline{\varphi}_n(x)$ exists. Shows that A is invariant, and that the observable $\overline{\varphi} : A \to \mathbb{R}$ is also invariant w.r.t. the restriction $f|_A : A \to A$.

2.6 Conjugations

Conjugations. The topological dynamical systems $f : X \to X$ and $g : Y \to Y$ are *(topologically) conjugated* if there exists a homeomorphism $h : X \to Y$, called *conjugation*, such that

$$h \circ f = g \circ h$$

This means that arrows in the following diagram commute:

$$\begin{array}{ccc} X & \xrightarrow{f} & X \\ \downarrow h & & \downarrow h \\ Y & \xrightarrow{g} & Y \end{array}$$

This condition may be also written $f = h^{-1} \circ g \circ h$, and is clearly an equivalence relation. By induction, we see that $f^n = h^{-1} \circ g^n \circ h$ for all times $n \geq 0$. In particular, a conjugation sends orbits of f into orbits of g, and vice-versa. The idea is that two conjugated transformations are indistinguishable from the topological point of view (we are just changing the names of the points).

A weaker notion is the following. A continuous and onto function $h : X \to Y$ is a *semi-conjugation* beween the dynamical systems $f : X \to X$ and $g : Y \to Y$ if $h \circ f = g \circ h$. In this case, g is called *factor* of f. The h-image of an orbit of f is an orbit of g, but each orbit of g may have more than one pre-image. Informally, the dynamics of f is richer than the dynamics of g. Meanwhile, when the set where h fails to be bijective is small, the two dynamics are still nearby.

3 Differential equations and flows

3.1 Flows

The main way in which dynamical systems enter in physics is through differential equations.

Flows of vector fields. Let X be a differentiable manifold (as, for example, an open region of $\mathbb{R}^n$), and let v be a vector field on X. If we assume that the autonomous differential equation

$$\dot{x} = v(x)$$

with any given initial condition $x(0) = x$, has solutions $t \mapsto x(t)$ which exist for any time $t \in \mathbb{R}$ (as is the case when v is smooth and X is compact), then the *flow* of the vector field v is the action $\Phi : \mathbb{R} \times X \to X$ given by $\Phi_t(x) = x(t)$. Indeed, it is clear that Φ_0 is the identity map, and that

$$\Phi_t \circ \Phi_s = \Phi_{t+s}$$

for any $t, s \in \mathbb{R}$. Therefore, $\Phi_{-t} = (\Phi_t)^{-1}$.

Conversely, given a one-parameter group of diffeomorphisms Φ_t, one define the phase velocity according to

$$v(x) := \frac{d}{dt}\,\Phi_t(x)\bigg|_{t=0} = \lim_{t\to 0} \frac{\Phi_t(x) - x}{t}$$

The group property then implies that the curve $t \mapsto x(t) = \Phi_t(x)$ satisfies

$$\dot{x}(t) = \lim_{s\to 0} \frac{\Phi_{t+s}(x) - \Phi_t(x)}{t} = \lim_{s\to 0} \frac{\Phi_s(\Phi_t(x)) - \Phi_t(x)}{t} = v(x(t))$$

and therefore is a solution of the autonomous differential equation $\dot{x} = v(x)$ with initial condition $x(0) = x$.

Also interesting are *semi-flows* Φ_t, which are defined only for non-negative times $t \geq 0$.

A flow or semi-flow is called *continuous time dynamical system*, and indeed our basic definitions in the previous chapter are adaptations of physicists' ideas and terminolgy about flows of vector fields. The map $t \mapsto \Phi_t(x)$ is called *trajectory* of the (initial) point x, and its image $\mathcal{O}^+(x) = \{\Phi_t(x) \,:\, t \in \mathbb{R}_+\}$ is called *(forward) orbit* of x. If it happens, as usual in classical mechanics, that flows are defined for all times $t \in \mathbb{R}$, then the set $\mathcal{O}(x) = \{\Phi_t(x) \,:\, t \in \mathbb{R}\}$ is called *orbit* of x.

From flows to maps, discretization. Given a flow Φ_t on X, one could specialize to discrete time looking at the system at multiples integers $n\tau$ of a given time-unit $\tau > 0$, and this amounts to iterate the transformation $f = \Phi_\tau$.

More interesting is the following construction.

Poincaré maps. Let Φ_t be the flow of the autonomous differential equation $\dot{x} = v(x)$ on a manifold X, and let $Y \subset X$ be a submanifold of codimension one which is transversal to the flow (i.e. the tangent space T_xY does not contain the vectors $v(x)$ for any $x \in Y$).

If $x_0 \in Y$ is a periodic point, say $\Phi_\tau(x_0) = x_0$ for some period $\tau > 0$, then nearby points $x \in Y$ also return to Y after some time near to τ. Thus, one could define, in a sufficiently small neighborhood $U \subset Y$ of x_0, a *first return/Poincaré map* $f : U \to Y$, sending a point $x \in U$ into $\Phi_{\tau(x)}(x)$ if $\tau(x)$ is the smallest positive time $t > 0$ such that $\Phi_t(x) \in Y$. This construction is even possible around a point which is not periodic, provided its orbit returns to Y sufficiently near.

Moreover, it may be also happens that the flow allows a *global (Poincaré) section*, a codimension one submanifold $Y \subset X$ transversal to the vector field v such that the orbit of any point $y \in Y$ eventually returns to Y after a minimal time

$$\tau(y) := \inf\{t > 0\,\text{s.t.}\;\; \Phi_t(y) \in Y\} < \infty\,,$$

called *first return time*. This allows to consider a globally defined *first return/Poincaré map* $f : Y \to Y$, according to

$$f(y) := \Phi_{\tau(y)}(y)\,.$$

e.g. Linear flows on the two-torus and rotations of the circle. A constant vector field $v = (\alpha, \beta)$ generates a linear flow

$$g_t : (x, y) \mapsto (x + \alpha t, y + \beta t)$$

on the plane, which is invariant under translations by integer vectors (being constant), and therefore it defines a flow on the two dimensional torus $\mathbb{T}^2 := \mathbb{R}^2/\mathbb{Z}^2$. The circle $\mathbb{R}/\mathbb{Z} \simeq C := \{(x + \mathbb{Z}, 0 + \mathbb{Z})\} \subset \mathbb{T}^2$ is transversal to the vector field if $\beta \neq 0$. The orbit of a point $(x, 0) \in C$ goes back to the section after a time $\tau = 1/\beta$ to the point $g_\tau(x, y) = (x + \alpha/\beta, 0)$. Thus, the first return map is $f : x + \mathbb{Z} \mapsto x + \alpha/\beta + \mathbb{Z}$, a rotation of the circle by an "angle" α/β.

Suspension flows. Poincaré construction of a first return map out of a flow admits an inverse. Given a map $f : X \to X$, one can define the *mapping torus* X_f as the Cartesian product $X \times [0, 1]$, with coordinates (x, t) with $x \in X$ and $t \in [0, 1]$, modulo the equivalence relation $(x, 1) \sim (f(x), 0)$. The flow of the vertical vector field $\partial/\partial t$ on X_f (which is a smooth manifold if X is) is called *suspension* of f. It is clear that it admits a global Poincaré section $X \times \{0\} \simeq X$, and its first return map is precisely f.

More generally, given a map $f : X \to X$ and a *roof function* $\tau : X \to \mathbb{R}_+$ bounded away from 0, one can consider the space $X_{f,\tau} = Y/\sim$ obtained as

$$Y = \{(x, t) \, : \, x \in X \, , \, 0 \leq t \leq \tau(x)\}$$

modulo the equivalence relation $(x, \tau(x)) \sim (f(x), 0)$. The flow of the vertical vector field $\partial/\partial t$ on X_f is called *suspension* of f with *height* τ. Again, it admits a global Poincaré section $X \times \{0\} \simeq X$, and its first return map is f.

3.2 Structure of physical models

Classical mechanics is the natural source of interesting dynamical systems.

Newtonian mechanics. According to greeks, the "velocity" $\dot{q} = \frac{d}{dt}q$ of a planet, where $q \in \mathbb{R}^3$ is its position in our Euclidean space and t is time, was determined by gods or whatever forced planets to move around circles. Then came Galileo, and showed that gods could at most determine the "acceleration" $\ddot{q} = \frac{d^2}{dt^2}q$, since the laws of physics should be written in the same way by an observer in any reference system at uniform rectilinear motion with respect to the fixed stars. Finally came Newton, who decided that what gods determined was to be called "force", and discovered that the trajectories of planets, fulfilling Kepler's experimental three laws[10] , were solutions of his famous (second order differential) equation

$$m\ddot{q} = F$$

where m is the mass of the planet, and where the attractive force F between the planet and the Sun is proportional to the product of their masses and inverse proportional to the square of their distance.

Later, somebody noticed that most observed forces were "conservative", could be written as $F = -\nabla V$, for some real valued function $V(q)$ called "potential energy". There follows that Newton equations can be written as $m\ddot{q} = -\nabla V$, and that the "total energy"

$$E = \frac{1}{2}m\|\dot{q}\|^2 + V(q)$$

[10] In *Astronomia nova*, 1609, and *Harmonices mundi*, 1619, Johannes Kepler published his three laws of planetary motions:
i) planets moves in ellipses with focus at the Sun,
ii) the radius vector describes equal areas in equal times,
iii) the squares of the periods are to each other as the cubes of the mean distance from the Sun.
It was with the purpose to derive Kepler laws from a second order differential equation $m\ddot{q} = F$ that Isaac Newton realized that the force of gravitational attraction between the Sun and a planet (hence between any two bodies!) should be proportional to m/ρ^2 (*Philosophiae naturalis principia mathematica*, 1687).

is constant along trajectories. The function $\frac{1}{2}m\|\dot{q}\|^2$ is called "kinetic energy" of the system.

An alternative (and indeed useful) formulation of Newtonian mechanics is the one developed by Lagrange. He defined (what we now call) the "Lagrangian" of the system as the difference between the kinetic energy and the potential energy

$$L(q,\dot{q}) = \frac{1}{2}m\|\dot{q}\|^2 - V(q)$$

and observed that Newton equations are equivalent to the (Euler)-Lagrange equations

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) = \frac{\partial L}{\partial q}$$

The product $p = m\dot{q} = \partial L/\partial \dot{q}$ is called "(linear) momentum", and, since p/m is the gradient of the kinetic energy $K(p) = \|p\|^2/2m$, Hamilton could write Newton's second order differential equations as the system of first order differential equations

$$\dot{q} = \frac{\partial H}{\partial p} \qquad \dot{p} = -\frac{\partial H}{\partial q}$$

where $H(q,p) = K(p) + V(q)$ is the total energy as function of q and p, nowdays called "Hamiltonian". It is a simple check that the energy is a constant of the motion, since

$$\frac{d}{dt}H = \frac{\partial H}{\partial q}\cdot\dot{q} + \frac{\partial H}{\partial p}\cdot\dot{p} = \frac{\partial H}{\partial q}\cdot\frac{\partial H}{\partial p} - \frac{\partial H}{\partial p}\cdot\frac{\partial H}{\partial q} = 0$$

Hamiltonian flows. The modern abstract formulation of classical mechanics is as follows. Let (X,ω) be a symplectic manifold, i.e. a differentiable manifold X of even dimension $2n$, equipped with a smooth closed differential two-form ω such that $\omega^n \neq 0$. Darboux theorem says that locally one can choose "canonical" coordinates $(q_1,...,q_n,p_1,..,p_n)$ such that $\omega = \sum_{k=1}^n dp_k \wedge dq_k$. The standard example is the cotangent bundle $T^*\mathbb{R}^n$ of the Euclidean vector space $\mathbb{R}^n$, whose coordinates are positions q_k and momenta p_k.

Let $H : X \to \mathbb{R}$ be a smooth function, called "Hamiltonian" and thought as the "energy" of the system. Typically, it has the form "kinetic energy+potential energy", where the kinetic energy is a positive definite quadratic form in the momenta p, and the potential energy is a function V depending on the positions q and possibly on the momenta p. The Hamiltonian vector field v is defined by the identity $dH = i_v\omega$, and the *Hamiltonian flow* is the flow of v. In canonical coordinates, the equations of motion read

$$\dot{q_k} = \frac{\partial H}{\partial p_k} \qquad \dot{p_k} = -\frac{\partial H}{\partial q_k}$$

It happens that the Hamiltonian flow Φ_t preserves the energy, namely $H(\Phi_t(x)) = H(x)$ for any $x \in X$ and any time $t \in \mathbb{R}$, as follows form the fact that $\pounds_v H = 0$.

Also, according to Liouville theorem, it preserves the volume form ω^n, defined in canonical coordinates by the volume element $dq_1 \wedge \dots dq_n \wedge dp_1 \wedge \dots \wedge dp_n$. In particular, if the phase space if compact, it preserves a probability measure.

e.g. **Free motion.** Free motion in an inertial frame is described by the Lagrangian $L = \frac{1}{2}m\|\dot{q}\|^2$. The equations of motion are

$$\ddot{q} = 0\,.$$

Solutions are straight lines $q(t) = c + vt$, for same initial position $q(0) = c$ and velocity $\dot{q}(0) = v$.

e.g. **Free fall.** Free fall near the Earth's surface is modeled by the Lagrangian $L = \frac{1}{2}m\|\dot{q}\|^2 - mgz$, where $g \simeq 9.8\ \mathrm{m/s}^2$ is the gravitational acceleration and z is the height of the particle (assumed much smaller than the Earth's diameter), the third coordinates of $q = (x,y,z)$. The equation of motion for the height is

$$\ddot{z} = g\,.$$

Solution are parabolae $z(t) = c + vt - \frac{1}{2}gt^2$, for some initial height $c = z(0)$ and some initial velocity $v = \dot{z}(0)$.

Geodesic flows. The simplest mechanical system, the free motion of a particle, belongs to the class of geodesic flows. Let (M, g) be a Riemannian manifold, g beeing the Riemannian metric. Let SM be the unit tangent bundle of M. If M is geodesically complete, to every unit vector $v \in SM$ there corresponds a unique geodesic line (i.e. a local isometry) $c : \mathbb{R} \to M$ such that $\dot{c}(0) = v$. The *geodesic flow* is the action $\Phi : \mathbb{R} \times SM \to SM$, defined as $\Phi_t(v) = \dot{c}(t)$.

Particularly interesting are geodesic flows over homogeneous spaces. Apart from the rather trivial exemple of flat spaces, a source of interesting dynamical properties is the geodesic flow on a manifold with constant negative curvature. The proptotype is as follows. The group $G = PSL(2, \mathbb{R})$ can be seen as the orientation preserving isometry group of the Poincaré half-plane $\mathbb{H}$, equipped with the hyperbolic metric of sectional curvature -1. Its action is transitive. Since the stabilizer of a point in the half-plane is isomorphic to the group of rotations $SO(2)$, we can identify $S\mathbb{D}$ with G. Now, let Γ be a discrete cocompact subgroup of G with no torsion. The quotient space $\Sigma = \mathbb{D}/\Gamma$ is a compact Riemann surface, which comes equipped with a Riemannian metric of sectional curvature -1, and its unit tangent bundle is diffeomorphic to G/Γ. The geodesic flow on $S\Sigma$ is then the algebraic flow $\Phi : \mathbb{R} \times G/\Gamma \to G/\Gamma$ defined as $\Phi_t(g\Gamma) = e_t g\Gamma$, where

$$e_t = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix}$$

3.3 Integration of one-dimensional systems

Some techniques to integrate ordinary differential equations (ODEs) like $\dot{x} = v(x, t)$ when the phase space is one or two-dimensional.

Integrating simple ODEs. The simplest case occurs when the velocity field v does not depend on the phase space variable x, hence

$$\dot{x} = v(t)\,,$$

where $v(t)$ is some given (piecewise) continuous function of time. This just says that x must be a primitive of v, and the fundamental theorem of calculus (i.e. Leibniz and/or Newton's discovery) tells us how to compute such a primitive:

$$x(t) = x_0 + \int_{t_0}^{t} v(s)ds\,.$$

Here you may observe that this class of ODEs have "symmetries". The line field does not depend on x, hence slopes of solutions are the same along horizontal lines (t = constant) in the extended phase space $X \times \mathbb{R}$. There follows that any translate $\varphi(t) + c$ of a solution $\varphi(t)$ is still a solution.

Autonomous first order ODEs and their flows. A first order ODE of the form

$$\dot{x} = v(x)\,,$$

where the velocity field v does not depend on time, is called *autonomous.* Most fundamental equations of physics (those describing closed systems, without external forces) can be written as autonomous first order ODEs, and this corresponds to time-invariance of physical laws.

Here you may notice symmetries again. The line field v of an autonomous equation is constant along vertical lines (x = constant) of the extended phase space $X \times \mathbb{R}$. Hence any translate $\varphi(t+s)$ of a solution $\varphi(t)$ is still a solution. This is the manifestation of time-invariance of a law codified by an autonomous ODE. This also implies that there is no loss of generality in restricting to an initial time $t_0 = 0$.

Equilibrium solutions. First, we observe that an autonomous equation may admit constant solutions. Indeed, if x_0 is a *singular point* of the vector field v, i.e. a point where $v(x_0) = 0$, then the constant function

$$x(t) = x_0 \qquad \forall \quad t \in \mathbb{R}$$

obviously solves the equation. Such solutions, which do not change with time, are called *equilibrium*, or *stationary*, solutions.

Solutions near non-singular points. The trick used to "guess" other solutions, when the phase space is one-dimensional, i.e. $X \subset \mathbb{R}$, is a first instance of the method of "separation of variables". Fix a *non-singular point* of the velocity field, i.e. a point x_0 where $v(x_0) \neq 0$. We want to solve the Cauchy problem with initial condition $x(t_0) = x_0$. First, rewrite the equation $dx/dt = v(x)$ formally as "$dx/v(x) = dt$" (multiply by dt and divide by $v(x)$, so that all x's are on the left and all t's are on the right). Instead of trying to make sense to this last expression (which is possible, of course, and here you can appreciate the beauty of Leibniz' notation dx/dt for derivatives!), observe that it is suggesting that $\int dx/v(x) = \int dt$. Now assume that the velocity field v is continuous and let $J = (x_-, x_+)$ be the maximal interval containing x_0 where v is different from zero. Integrating, from x_0 to $x \in J$ on the left and from t_0 to t on the right, we obtain a differentiable function $x \mapsto t(x)$ defined as

$$t(x) = t_0 + \int_{x_0}^{x} \frac{dy}{v(y)}$$

for any $x \in J$. Now, observe that the derivative dt/dx is equal to $1/v$. Since, by continuity, $1/v$ does not change its sign in J, our $t(x)$ is a strictly monotone continuously differentiable function. We can invoke the inverse function theorem and conclude that the function $t(x)$ is invertible. This prove that the above relation defines actually a continuously differentiable function $t \mapsto x(t)$ in some interval $I = t(J)$ of times around t_0. Finally, you may want to check that the function $t \mapsto x(t)$ solves the Cauchy problem: just compute the derivative (using the inverse function theorem),

$$\begin{aligned} \dot{x}(t) &= 1/\left(\frac{dt}{dx}(x(t))\right) \\ &= v(x)\,, \end{aligned}$$

and check the initial condition. Observe that the function $t(x) - t_0$ has then the interpretation of the "time needed to go from x_0 to x".

At the end of the story, if you are lucky enough and know how to invert the function $t(x)$, you'll get an explicit solution as

$$x(t) = F^{-1}\left(t - t_0 + F(x_0)\right)\,,$$

where F is any primitive of $1/v$. Close inspection of the above reasoning shows that the local solution you've found is indeed the unique one. Namely, we have the following

Theorem 3.1. *Let $v(x)$ be a continuous velocity field and let x_0 be a non-singular point of v. Then there exist one and only one solution of the Cauchy problem $\dot{x} = v(x)$ with initial condition $x(t_0) = x_0$ in some sufficiently small interval I around t_0. Moreover, the solution $x(t)$ is the inverse function of*

$$t(x) = t_0 + \int_{x_0}^{x} \frac{dy}{v(y)}\,,$$

defined in some small interval J around x_0.

Proof. Here we give the pedantic proof. Let J be as above. Define a function $H : \mathbb{R} \times J \to \mathbb{R}$ as

$$H(t, x) = t - t_0 - \int_{x_0}^{x} \frac{dy}{v(y)}\,.$$

If $t \mapsto \varphi(t)$ is a solution of the Cauchy problem, then computation shows that $\frac{d}{dt} H(t, \varphi(t)) = 0$ for any time t. There follows that H is constant along the solutions of the Cauchy problem. Since $H(t_0, x_0) = 0$, we conclude that the graph of any solution belongs to the level set $\Sigma = \{(t, x) \in \mathbb{R} \times J \text{ s.t. } H(t, x) = 0\}$. Now observe that H is continuously differentiable and that its differential $dH = dt + dx/v(x)$ is never zero. Actually, both partial derivatives $\partial H/\partial t$ and $\partial H/\partial x$ are always different from zero. Hence we can apply the implicit function theorem and conclude that the level set Σ is, in some neighborhood $I \times J$ of (t_0, x_0), the graph of a unique differentiable function $x \mapsto t(x)$, as well as the graph of a unique differentiable function $t \mapsto x(t)$, the inverse of t, which as we have already seen solves the Cauchy problem. $\square$

On the failure of uniqueness near singular points. The interval $I = t(J)$ where the solution is defined need not be the entire real line: solutions may reach the boundary of J, i.e. one of the singular points $x_\pm$ of the velocity field, in finite time. Since singular points are themselves equilibrium solutions, this imply that solutions of the Cauchy problem at singular points may not be unique, under such mild conditions (continuity) for the velocity field. Later we'll see Picard's theorem, which prescribes stronger regularity conditions on the velocity field v under which the Cauchy problem admits unique solutions for any initial condition in the extended phase space.

e.g. **Counter-example.** Both curves $x(t) = 0$ and $x(t) = t^3$ solve the equation

$$\dot{x} = 3x^{2/3}$$

with initial condition $x(0) = 0$. The problem here is that the velocity field $v(x) = 3x^{2/3}$, although continuous, is not differentiable and not even Lipschitz at the origin. You may notice that the solution starting, for example, at $x_0 = 1$ reaches (or better comes from) the singular point $x_- = 0$ in finite time, since

$$t(x_-) - t(x_0) = \int_1^0 \frac{1}{3} y^{-2/3} dy = -1 \,.$$

One-dimensional Newtonian motion in a time independent force field. The one-dimensional motion of a particle of mass m subject to a force $F(x)$ that does not depend on time is described by the Newton equation

$$m\ddot{x} = -U'(x) \,,$$

where the potential $U(x) = -\int F(x)dx$ is some primitive of the force. The total energy

$$E\,(x, \dot{x}) = \frac{1}{2} m\dot{x}^2 + U(x)$$

(which of course is defined up to an arbitrary additive constant) of the system is a constant of the motion, i.e. is constant along solutions of the Newton equation. In particular, once a value E of the energy is given (depending on the initial conditions), the motion takes place in the region where $U(x) \leq E$, since the kinetic energy $\frac{1}{2} m\dot{x}^2$ is non-negative. Conservation of energy allows to reduce the problem to the first order ODE

$$\dot{x}^2 = \frac{2}{m} \left(E - U(x)\right) \,,$$

which has the unpleasant feature to be quadratic in the velocity $\dot{x}$. Meanwhile, if we are interested in a one-way trajectory going from some x_0 to x, say with $x > x_0$, we may solve for $\dot{x}$ and find the first order autonomous ODE

$$\dot{x} = \sqrt{\frac{2}{m} \left(E - U(x)\right)} \,.$$

There follows that the time needed to go from x_0 to x is

$$t(x) = \int_{x_0}^{x} \frac{dy}{\sqrt{\frac{2}{m}\left(E - U(y)\right)}} \,.$$

The inverse function of the above $t(x)$ will give the trajectory $x(t)$ with initial position $x(0) = x_0$ and initial positive velocity $\dot{x}(0) = \sqrt{\frac{2}{m}\left(E - U(x_0)\right)}$, at least for sufficiently small times t.

The exponential. The exponential function, according to Walter Rudin "the most important function in mathematics" ([Ru87], 1st line of page 1), is the unique solution of the autonomous differential equation

$$\dot{x} = x$$

with initial condition $x(0) = 1$. If we try a power series like $a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \ldots$, the differential equation gives the recursion $n\, a_n = a_{n-1}$ for the coefficients, while the initial condition yelds $a_0 = 1$. Thus, the solution is $x(t) = 1 + t + t^2/2 + t^3/6 + \ldots$.

Actually, it is convenient to complexify time, i.e. take $z = t + i\theta \in \mathbb{C}$ with $t, \theta \in \mathbb{R}$, and define the *exponential* as the power series

$$\exp(z) := 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24} + \cdots = \sum_{n\geq 0} \frac{z^n}{n!}$$

Since $\limsup_{n\to\infty}(1/n!)^{1/n} = 0$, the radius of convergence is ∞, hence the power series defines an entire function, i.e. a holomorphic function $\exp : \mathbb{C} \to \mathbb{C}$. Deriving each term of the series, we easily verify that indeed $\exp' = \exp$. The initial condition $\exp(0) = 1$ is obvious. From absolute convergence of the series and algebraic manipulation we also get the group property

$$\exp(z + w) = \exp(z) \cdot \exp(w)$$

for any $z, w \in \mathbb{C}$, saying that exp is a homomorphism of the additive group $\mathbb{C}$ into the multiplicative group $\mathbb{C}^\times = \mathbb{C}\backslash\{0\}$. In particular, $\exp(-z) = 1/\exp(z)$, so that the exponential $\exp(z)$ is never 0. This also justifies our notation $\exp(z) = e^z$, where

$$e := \exp(1) = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots \simeq 2.71828182845904523536028747135266249775 72\ldots$$

(another famous irrational, actually a transcendental number!). For real time $z = t$, we recover the familiar model of "exponential growth" $t \mapsto e^t$, a strictly increasing function from the additive group $\mathbb{R}$ onto the multiplicative group $\mathbb{R}_+ =]0, \infty[$, growing faster than any power t^n as $t \to \infty$. For pure imaginary times, say $z = i\theta$ with $\theta \in \mathbb{R}$, we get the *Euler's formula*

$$e^{i\theta} = \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \ldots\right) + i\left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \ldots\right) = \cos(\theta) + i\sin(\theta)$$

(and of course you may take the last identity as the "definition" of the trigonometric functions!). So, $\theta \mapsto e^{i\theta}$ defines a periodic function with period 2π, sending the real line $\mathbb{R}$ onto the unit circle $\mathbb{S} = \{z \in \mathbb{C}\,\text{s.t.}\ |z| = 1\}$. There follows from the group property that

$$\exp(t + i\theta) = e^t\left(\cos(\theta) + i\sin(\theta)\right) .$$

Finally, the exponential exp is a periodic entire function with period $i2\pi$ which only omits the value 0, a holomorphic bijection of the cylinder $\mathbb{C}/i2\pi\mathbb{Z}$ onto $\mathbb{C}\backslash\{0\}$.

e.g. **Interest rates and the exponential.** Let x be the annual interest payed for a deposit (so that an interest of 0.2% mean $x = 0.02$). If the interest is payed once each year, an initial deposit of a euros increases to

$$a + xa = a \cdot (1 + x)$$

after one year. If, however, the interest is "computed" every six months, the same initial deposit produces

$$a + \frac{x}{2}a + \left(a + \frac{x}{2}a\right)\frac{x}{2} = a \cdot \left(1 + \frac{x}{2}\right)^2$$

after one year. By induction, we see that if the interest is computed every $12/n$ months, after one year we get a final capital of

$$a \cdot \left(1 + \frac{x}{n}\right)^n$$

The limit of the gain factor as $n \to \infty$,

$$E(x) = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n$$

is another definition of the exponential function. If the argument lives in the Riemann sphere, you may think that $\exp(z) = (1 - z/\infty)^\infty$ has a zero of order ∞ at the point $p = \infty \in \overline{\mathbb{C}}$.

e.g. Population dynamics. The exponential models the dynamics of a population in a unlimited environment. The *Malthusian/exponential model* [11] is

$$\dot{N} = \lambda N$$

where $N(t)$ is the population at time t, and $\lambda > 0$ is some growth constant (the difference $\alpha - \beta$ between the natality rate and the mortality rate). The solution is $N(t) = N(0)e^{\lambda t}$. If we retire specimen at fixed rate $\alpha > 0$

$$\dot{N} = \lambda N - \alpha$$

we have a non-trivial stationary solution $\overline{N} = \alpha/\lambda$, and the difference $x(t) = N(t) - \overline{N}$ is still exponential.

This behaviour has to be compared with the *super-exponential model*

$$\dot{N} = \lambda N^2 .$$

which undergoes a catastrophe (infinite population) in finite time! Indeed, the solution with $N(0) = N_0 > 0$ is $N(t) = N_0/(1 - \lambda t/N_0)$.

A more realistic model of population dynamics in a finite environment is the *logistic equation*[12]

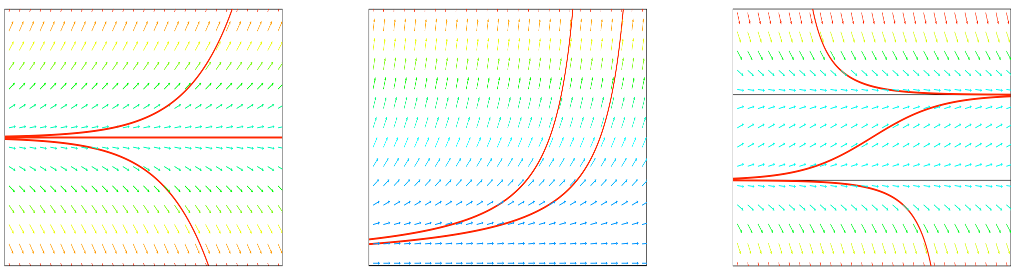
$$\dot{N} = \lambda N(1 - N/M)$$

where $\lambda > 0$ and the constant $M > 0$ is a maximal population. Observe that $\dot{N} \simeq \lambda N$ if $N \ll M$, and that $\dot{N} \to 0$ when $N \to M$. The relative population $x(t) = N(t)/M$ satisfies the "adimensional" logistic equation

$$\dot{x} = \lambda x(1 - x) .$$

Here we see two equilibria: the trivial equilibrium $x(t) = 0$ and the maximum allowed polpulation $x(t) = 1$. The generic solution with initial condition $0 < x(0) < 1$ is

$$x(t) = \frac{1}{1 + \left(\frac{1}{x_0} - 1\right) e^{-\lambda t}} \quad ,$$



Exponential growth, super-exponential growth and logistic model.

3.4 Existence and uniqueness theorems

Solutions of a differential equation. Here we consider a generic first order ODE of the form

$$\dot{x} = v(x, t)$$

where the velocity field v is a (continuous) function defined in some extended phase space $X \times \mathbb{R}$. The phase space X may be some interval of the real line, an open subset of some Euclidean $\mathbb{R}^n$, or a differentiable manifold.

The problem we address is the existence and uniqueness of solutions of the initial value (or Cauchy) problem. A *local solution* passing through the point $(x_0, t_0) \in X \times \mathbb{R}$ is a solution $t \mapsto \varphi(t)$,

[11] T.R. Malthus, *An Essay on the Principle of Population*, London, 1798.

[12] Pierre François Verhulst, Notice sur la loi que la population pursuit dans son accroissement, *Correspondance mathématique et physique* **10** (1838), 113-121.

defined in some neighborhood I of t_0, such that $\varphi(t_0) = x_0$. Eventually, we'll be interested also in the possibility of extending such local solutions to larger intervals of times.

The basic existence theorem is [13]

Theorem 3.2 (Peano)**.** *Let $v(x,t)$ be a continuous velocity field in some domain A of the extended phase space $\mathbb{R}^2$. Then for any point $(x_0, t_0) \in A$ passes at least one integral curve of the differential equation $\dot{x} = v(x,t)$.*

Proof. (Idea) Natural guesses for the solutions are Euler lines starting through (x_0, t_0). If we restrict to a sufficiently small neighborhood of (x_0, t_0), we can assume that the velocity field is bounded, say $|v(x,t)| \leq K$, and that all such Euler lines lies in the "papillon" made of two triangles touching at (x_0, t_0) with slopes $\pm K$. Construct a family of Euler lines, graphs of $\varphi_n(t)$, such that the maximal step ε_n of the n-th line goes to 0 as $n \to \infty$. One easily sees that the family (φ_n) is bounded and equicontinuous. By the Ascoli-Arzelá theorem it admits a (uniformly) convergent subsequence. Finally, we claim that the sublimit $\varphi_{n_i} \to \varphi$ solves the differential equation. □

Both existence and uniqueness may fail. The Hamilton-Jacobi equation

$$(\dot{x})^2 - xt + 1 = 0$$

cannot have solutions satisfying the initial condition $x(0) = 0$, for otherwise we would have a negative "kinetic energy" $(\dot{x})^2 = -1$ at that point!

Some regularity of the functions involved in a differential equation is also needed to ensure the uniqueness of solutions. For example, both curves $t \mapsto 0$ and $t \mapsto t^3$ solve the equation

$$\dot{x} = 3x^{2/3}$$

with initial condition $x(0) = 0$. The problem here is that the velocity field $v(t,x) = 3x^{2/3}$, although continuous, is not differentiable and not even Lipschitz at the origin.

Uniqueness of solutions. A velocity field $v(t,x)$, defined in a domain $I \times D$ of the extended phase space $\mathbb{R} \times \mathbb{R}^n$, is *locally Lipschitz* w.r.t. to the variable x if for any $(t_0, x_0) \in I \times D$ there is a neighborhood $J \times U \ni (t_0, x_0)$ and a constant $L \geq 0$ such that

$$\|v(t,x) - v(t,y)\| \leq L \cdot \|x - y\| \qquad \forall\ (t,x), (t,y) \in J \times U$$

If $v(t,x)$ has continuous derivative w.r.t. x, i.e. if the Jacobian

$$D_x v(t,x) = \left(\frac{\partial v_i}{\partial x_j}(t,x) \right)$$

exists and is continuous, then $v(t,x)$ is locally Lipschitz in any compact convex domain $I \times K \subset \mathbb{R} \times \mathbb{R}^n$. The basic uniqueness theorem is the following classical result by Lindelöf [14] and Picard.

Theorem 3.3 (Picard-Lindelöf)**.** *Let $v(t,x)$ be a continuous velocity field defined in some domain D of the extended phase space $\mathbb{R} \times X$. If v is locally Lipschitz (for example continuously differentiable) w.r.t. the second variable x, then there exist one and only one local solution of $\dot{x} = v(t,x)$ passing through any point $(t_0, x_0) \in D$.*

Geometrically, the uniqueness theorem says that through any point (t_0, x_0) of the domain D there pass one and only one solution. Hence solutions, considered as curves in the extended phase space, cannot intersect each other.

In a domain where Picard's theorem applies, if two local solutions agree in a common interval of times then they are indeed restrictions of a unique solution defined in the union of the respective domains. There follows that solutions are always extendible to a maximum domain. Such solutions are called *maximal solutions*.

[13] G. Peano, Sull'integrabilità delle equazioni differenziali del primo ordine, *Atti Accad. Sci. Torino* **21** (1886), 677-685. G. Peano, Demonstration de l'intégrabilité des équations différentielles ordinaires, *Mathematische Annalen* **37** (1890) 182-228.

[14] M. E. Lindelöf, Sur l'application de la méthode des approximations successives aux équations différentielles ordinaires du premier ordre, *Comptes rendus hebdomadaires des séances de l'Académie des sciences* **114** (1894), 454-457. Digitized version online via http://gallica.bnf.fr/ark:/12148/bpt6k3074

Strategy of the proof of the Picard's theorem. The first observation is that a function $\varphi(t)$ is a solution of the Cauchy problem for $\dot{x} = v(t,x)$ with initial condition $\varphi(t_0) = x_0$ iff

$$\varphi(t) = x_0 + \int_{t_0}^{t} v\left(s, \varphi(s)\right) ds$$

Now, we notice that the above identity is equivalent to the statement that φ is a fixed point of the so called *Picard's map* $\phi \mapsto \mathcal{P}\phi$, sending a function $t \mapsto \phi(t)$ into the function

$$\left(\mathcal{P}\phi\right)(t) = x_0 + \int_{t_0}^{t} v\left(s, \phi(s)\right) ds$$

At this point, one must chose cleverly the domain of the Picard's map, which is the space of functions where we think a solution should be. It will be a certain space $\mathcal{C}$ of continuous functions, defined in an appropriate neighborhood I of t_0, equipped with a norm that makes it a complete metric space (hence a Banach space). The Lipschitz condition, together with continuity, satisfied by the velocity field will imply that if the interval I is sufficiently small then the Picard's map $\mathcal{P} : \mathcal{C} \to \mathcal{C}$ is a contraction. The contraction principle (theorem 6.4) finally guarantees the existence and uniqueness of the fixed point of $\mathcal{P}$ in $\mathcal{C}$.

Picard's iterations. The contraction principle actually says that the fixed point, i.e. the solution we are looking for, is the limit of any sequence ϕ, $\mathcal{P}\phi$, ..., $\mathcal{P}^n\phi$, ... of iterates of the Picard map starting with any initial guess $\phi \in \mathcal{C}$. In other words, the existence part of the theorem is "constructive", it gives us a procedure to find out the solution, or at least a sequence of functions which approximate the solution.

e.g. **Simple ODEs.** Consider the simple ODE $\dot{x} = v(t)$ with initial condition $x(t_0) = x_0$. Picard's recipe, starting from the initial guess $\phi(t) = x_0$ gives, already at the first step,

$$\left(\mathcal{P}\phi\right)(t) = x_0 + \int_{t_0}^{t} v(s) ds$$

which is the solution we know.

e.g. **The exponential.** Suppose you want to solve $\dot{x} = x$ with initial condition $x(0) = 1$. You start with the guess $\phi(t) = 1$, and then compute

$$\left(\mathcal{P}\phi\right)(t) = 1 + t \qquad \left(\mathcal{P}^2\phi\right)(t) = 1 + t + \frac{1}{2}t^2 \qquad \dots \qquad \left(\mathcal{P}^n\phi\right)(t) = 1 + t + \frac{1}{2}t^2 + \dots + \frac{1}{n!}t^n$$

Hence the sequence converges (uniformly on bounded intervals) to the Taylor series of the exponential function

$$\left(\mathcal{P}^n\phi\right)(t) \to 1 + t + \frac{1}{2}t^2 + \dots + \frac{1}{n!}t^n + \dots = e^t \,,$$

which is the solution we already knew.

Details of the proof of the Picard's theorem. Choose a sufficiently small rectangular neighborhood

$$I \times B = [t_0 - \varepsilon, t_0 + \varepsilon] \times \overline{B}_\delta\left(x_0\right)$$

around (t_0, x_0), where $B = \overline{B}_\delta\left(x_0\right)$ denotes the closed ball with center x_0 and radius δ in X. There follows from continuity of v that there exists K such that

$$|v(t,x)| \leq K$$

for any $(t,x) \in I \times B$. There follows from the local Lipschitz condition for v that there exists M such that

$$|v(t,x) - v(t,y)| \leq M|x - y|$$

for any $t \in I$ and any $x, y \in B$. Now restrict, if needed, the (radius of the) interval I in such a way to get both the inequalities $K\varepsilon \leq \delta$ and $M\varepsilon < 1$. Let $\mathcal{C}$ be the space of continuous functions $t \mapsto \phi(t)$ sending I into B. Equipped with the sup norm

$$\|\phi - \varphi\| = \sup_{t \in I} |\phi(t) - \varphi(t)|$$

this is a complete space. One verifies that the Picard's map sends $\mathcal{C}$ into $\mathcal{C}$, since

$$|\left(\mathcal{P}\phi\right)(t) - x_0| \leq \int_{t_0}^{t} |v\left(s, \phi(s)\right)| ds \leq K\varepsilon \leq \delta.$$

Finally, given two functions $\phi, \varphi \in \mathcal{C}$, one sees that

$$|\left(\mathcal{P}\phi\right)(t) - \left(\mathcal{P}\varphi\right)(t)| \leq \int_{t_0}^{t} |v\left(s, \phi(s)\right) - v\left(s, \varphi(s)\right)| ds \leq M\varepsilon \sup_{t \in I} |\phi(t) - \varphi(t)|$$

hence $\|\mathcal{P}\phi - \mathcal{P}\varphi\| < M\varepsilon \|\phi - \varphi\|$. Since $M\varepsilon < 1$, this proves that the Picard's map is a contraction and the fixed point theorem allows to conclude. □

We may not be able to solve them! Last but not least, we must keep in mind that we are not able to solve all equations. Actually, although we may prove the existence and the uniqueness for large classes of equations, we are simply not able to explicitly integrate the really interesting differential equations...

Ultimately we must recur to numerical methods to find approximate solutions and to qualitative analysis

Dependence on initial data and parameters Consider a family of ODEs

$$\dot{x} = v(t, x, \lambda)$$

where λ is a real parameter. We want to understand how solutions depend on the parameter λ. A basic instrument is the[15]

Theorem 3.4 (Grönwall's lemma). *Let $\phi(t)$ and $\psi(t)$ be two non-negative real valued functions defined in interval $[a, b]$ such that*

$$\phi(t) \leq K + \int_{a}^{t} \psi(s)\phi(s) ds$$

for any $a \leq t \leq b$ and some constant $K \geq 0$. Then

$$\phi(t) \leq K e^{\int_a^t \psi(s) ds}.$$

Proof. First, assume $K > 0$. Define

$$\Phi(t) = K + \int_{a}^{t} \psi(s)\phi(s) ds$$

and observe that $\Phi(a) = K > 0$, hence $\Phi(t) > 0$ for all $a \leq t \leq b$. The logarithmic derivative is

$$\frac{d}{dt} \log \Phi(t) = \frac{\psi(t)\phi(t)}{\Phi(t)} \leq \psi(t)$$

where we used the hypothesis $\phi(t) \leq \Phi(t)$. Integrating the inequality we get, for $a \leq t \leq b$,

$$\log \Phi(t) \leq \Phi(a) + \int_{a}^{t} \psi(s) ds.$$

[15] T. H. Gronwall, Note on the derivative with respect to a parameter of the solutions of a system of differential equations, *Ann. of Math* **20** (1919), 292-296.

Exponentiation gives the result, since

$$\phi(t) \leq \Phi(t) \leq K \cdot e^{\int_a^t \psi(s)ds}$$

The case $K = 0$ follows taking the limit of the above inequalities for a sequence of $K_n > 0$ decreasing to zero. □

Continuous dependence on initial conditions. If $x(t)$ and $y(t)$ are two solutions of the same differential equation

$$\dot{x} = v(t, x)$$

then

$$x(t) - y(t) = x(0) - y(0) + \int_{t_0}^{t} (v(s, x(s)) - v(s, y(s)))\, ds$$

If $L(s)$ denotes the Lipschitz constant of $v(s, \cdot)$, we get

$$\|x(t) - y(t)\| \leq \|x(0) - y(0)\| + \int_{t_0}^{t} L(s)\|x(s) - y(s)\| ds$$

The Gronwall's lemma 3.4 gives the estimate

$$\|x(t) - y(t)\| \leq e^{\int_{t_0}^t L(s)ds} \|x(0) - y(0)\|$$

Observe that the above control also gives an alternative proof of uniqueness of solutions given a Lipschitz condition on the vector field.

Theorem 3.5 (smooth dependence on parameters)**.** *Let $v(t, x, \lambda)$ be a family of vector fields defined on some domain of the extended phase space $D \subset \mathbb{R} \times X$ depending on a parameter $\lambda \in \Lambda \subset \mathbb{R}$. If v is of class C^k with $k \geq 1$, then in some neighborhood of any $(t_0, x_0, \lambda_0) \in D \times \Lambda$ the local solutions of*

$$\dot{x} = v(t, x, \lambda)$$

with initial condition $x(t_0) = x_0$ are differentiable (indeed C^k) functions of (t, x, λ).

A proof may be found in [BN05].

Warning. Continuous dependence does not exclude sensitive dependence on both initial conditions and parameters, even in the linear case! For example, the distance between solutions of $\dot{x} = \mu x$ with different $x(0)$ and/or μ may diverge for large time ...

3.5 Oscillations and cycles

The first remarkable natural phenomena are, of course, periodic motions.

Harmonic oscillator. The *harmonic oscillator* is the (phenomenon modeled by the) Newton equation

$$\ddot{q} = -\omega^2 q\,.$$

This is a quite universal equation, since it describes small oscillations around a "generic" stable equilibrium of any one-dimensional Newtonian system[16] (indeed, take a Newton equation $m\ddot{x} =$

[16] "The harmonic oscillator, which we are about to study, has close analogs in many other fields; although we start with a mechanical example of a weight on a spring, or a pendulum with a small swing, or certain other mechanical devices, we are really studying a certain *differential equation*. This equation appears again and again in physics and other sciences, and in fact is a part of so many phenomena that its close study is well worth our while. Some of the phenomena involving this equation are the oscillations of a mass on a spring; the oscillations of charge flowing back and forth in an electrical circuit; the vibrations of a tuning fork which is generating sound waves; the analogous vibrations of the electrons in an atom, which generate light waves; the equations for the operation of a servosystem, such as a thermostat trying to adjust a temperature; complicated interactions in chemical reactions; the growth of a colony of bacteria in interaction with the food supply and the poison the bacteria produce; foxes eating rabbits eating grass, and so on; ..."
Richard P. Feynman [Fe63]

$-dU'(x)$ of a particle in a potential field U. An equilibrium position is a zero of the force, i.e. a point x_0 where $U'(x_0) = 0$. It is "stable" if x_0 is a local minimum of the potential, so that the Taylor expansion of the potential around x_0 in powers of $q = x - x_0$ starts with $U(x) = \alpha + \frac{1}{2}\beta q^2 + \dots$, for some positive second derivative $U''(x_0) = \beta$. If we are only interested in small displacements of x around x_0, we can safely disregard high order terms and approximate the Newton equation as $m\ddot{q} \simeq -\beta q$, which is an harmonic oscillator with resonant frequency $\omega = \sqrt{\beta/m}$).

Call $p = \dot{q}$ the momentum. The Newton equation $\ddot{q} = -\omega^2 q$ is equivalent to Hamilton's first order equations

$$\begin{aligned} \dot{q} &= p \\ \dot{p} &= -\omega^2 q \,. \end{aligned}$$

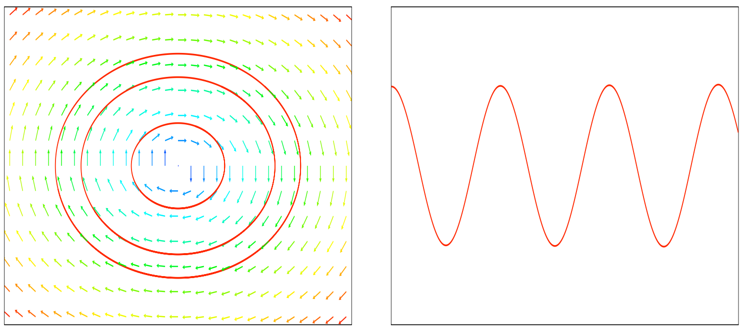
If we define the complex variable $z = \omega q + i\dot{q}$, Newton equation then takes the form of a first order linear equation in the complex line, namely $\dot{z} = -i\omega z$, whose solution is $z\,(t) = e^{-i\omega t} z\,(0)$.

In terms of the original (physical) variables, the solutions read

$$q\,(t) = q_0 \cos(\omega t) + \frac{v_0}{\omega} \sin(\omega t) = A \sin(\omega t + \phi)$$

where the amplitude A and the initial phase ϕ depend on the initial conditions $q(0) = q_0$ and $\dot{q}(0) = v_0$. So, all trajectories are periodic with common period $2\pi/\omega$, and orbits are ellipses in the q-$\dot{q}$ plane, determined by the conserved energy

$$E = \frac{1}{2}\left(\dot{q}^2 + \omega^2 q^2\right) = \omega^2 A^2 \,.$$



Harmonic oscillator, phase curves and time series.

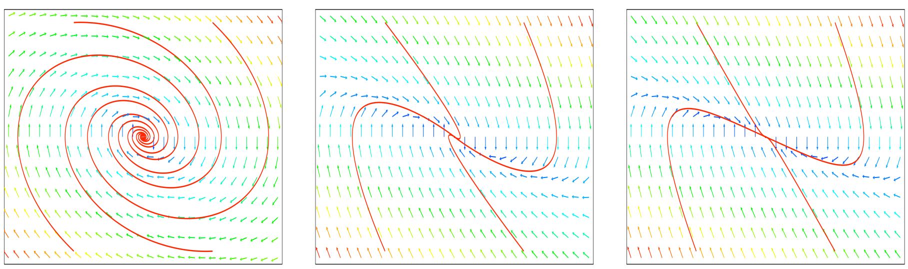
Dumped oscillations. Adding friction to an harmonic oscillator we get

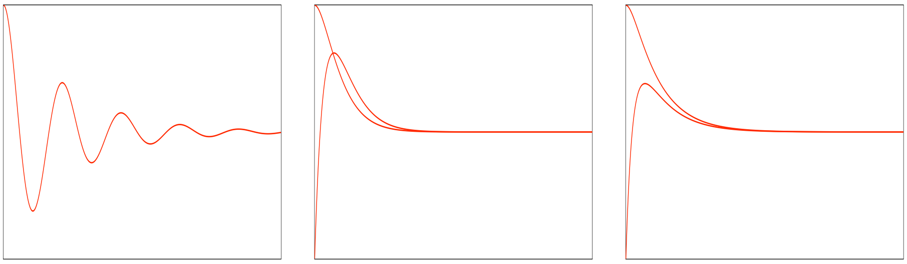
$$\ddot{q} = -2\alpha\dot{q} - \omega^2 q \,,$$

where $\alpha > 0$ is some friction coefficient. The guess $q(t) = e^{-\alpha t} y(t)$ gives $\ddot{y} = \delta y$ where the "discriminant" is $\delta = \omega^2 - \alpha$. Find the general solution, draw pictures and discuss the cases $\alpha^2 < \omega^2$ (under-critical damping), $\alpha^2 = \omega^2$ (critical damping), and $\alpha^2 > \omega^2$ (overcritical damping). Show that the energy

$$E(q, \dot{q}) = \frac{1}{2}\dot{q}^2 + \frac{1}{2}\omega^2 q^2$$

decreases with time outside equilibrium points.





Underdamped, critical and overdamped oscillations (phase portrait and time series).

Mathematical pendulum. The Newton equation

$$I\ddot{\theta} = -mg\ell \sin\theta$$

models the motion of an idealized pendulum (meaning a point mass attached to a wire of negligible weight, under a constant gravitational force) with mass m and length ℓ, where $I = m\ell^2$ is the moment of inertia, g is the gravitational acceleration (near the Earth's surface), and θ is the angle of the wire with the origin $\theta = 0$ located at the stable equilibrium point. The energy

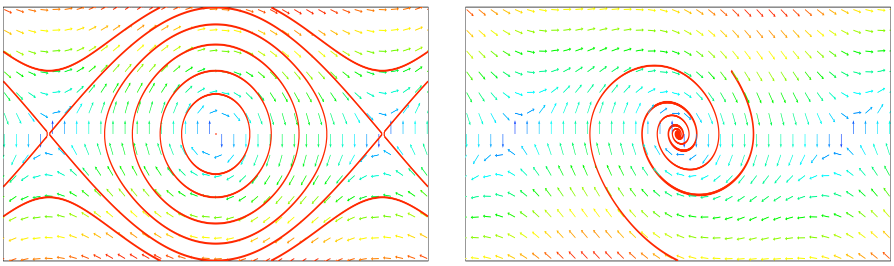
$$E = \frac{1}{2}\dot{\theta}^2 - mg\ell\cos\theta$$

is a constant of the motion. We can define the resonant frequency $\omega = \sqrt{mg\ell/I} = \sqrt{g/\ell}$ and write the equation as

$$\ddot{\theta} = -\omega^2 \sin\theta$$

Observe that in the limit of small oscillations we could replace $\sin\theta \simeq \theta$ and we are back to the harmonic oscillator $\ddot{\theta} = -\omega^2\theta$. To simplify thinks, let's take $\omega = 1$. Solving the energy for $\dot{\theta}^2$ the we see that the motion with energy E is given implicitly by the "elliptic integral"

$$t = \int \frac{d\theta}{\sqrt{2(E - \cos(\theta))}}$$



Phase portrait of a mathematical pendulum (without and with friction)

Jacobi elliptic functions. What does a mathematician/physicist do when he/she face an integral and doesn't see how to solve in terms of known functions? He/she gives a name to it.

Define $k = \sqrt{\frac{E+1}{2}}$ and then $x = \frac{1}{k}\sin(\theta/2)$. The conservation of energy reads

$$\dot{x} = \sqrt{(1 - x^2)(1 - k^2x^2)}$$

There follows that time is given by the so called *Jacobi's elliptic integral of the first kind*

$$t = \int \frac{dx}{\sqrt{(1 - x^2)(1 - k^2x^2)}}$$

The solution, actually the inverse function $x = \mathrm{sn}(t, k)$ as a function of t and the parameter k, is "named" *Jacobi elliptic function*.

This is the beginning of a long and interesting story. You may want to know that sn, as well its relatives, is a quotient of products of *Jacobi's theta functions*, hence, we are at the intersection between complex analysis, algebraic geometry, number theory, ...

Kepler problem. Kepler problem deals with the motion of two point-like bodies (planets and/or stars) under mutual gravitational interaction. Let $m_1, m_2 > 0$ be their masses, and $q_1, q_2 \in \mathbb{R}^3$ their positions, respectively. Gravitational interaction is described by the conservative force $-\nabla V$ with potential energy

$$V(q_1, q_2) = G\frac{m_1 m_2}{|q_1 - q_2|}$$

where G is the gravitational constant. This force verifies the "third law of dynamics", hence the total linear and angular momentum

$$P = m_1\dot{q}_1 + m_2\dot{q}_2 \quad \text{and} \quad M = m_1 q_1 \wedge \dot{q}_1 + m_2 q_2 \wedge \dot{q}_2$$

are conserved. This implies that the center of mass moves at uniform rectilinear speed and that the motion of the two bodies takes place in a plane orthogonal to the angular momentum M. If we choose a Galileian reference system where $P = 0$ and M is parallel to the z-axis (in particular M is supposed different from the zero vector, a case which leads to a collision ...) , the full system is described by the single vector $q_2 - q_1$ in the x-y plane, which we write in polar coordinates as $\rho e^{i2\pi\theta}$. It turns out that the two-body problem is equivalent to the motion of a single point mass $m = \frac{m_1 m_2}{m_1+m_2}$ moving on a plane under the influence of a potential energy $V(\rho) = -G\frac{m}{\rho}$, the (conserved) energy beeing

$$E = \frac{1}{2}m\left(\dot{\rho}^2 + \rho^2\dot{\theta}^2\right) + V(\rho)$$

Observe that if one of the bodies is much bigger than the other (like the Sun and the Earth), say $m_1 \gg m_2$, then the center of mass nearly coincides with the position q_1 of the bigger body, while the reduced mass m is essentially the mass m_2 of the smaller one (hence it looks like the Earth moving around the Sun, as Galileo had suggested).

Central forces. Consider the Newton equation

$$m\ddot{r} = F(|r|)\,\hat{r}$$

describing the motion of a particle (planet) of mass m in a central force field F. Conservation of angular momentum implies that the motion is planar, hence we may take $r \in \mathbb{R}^2$. In polar coordinates $r = \rho e^{i\theta}$, the equations reed

$$\begin{aligned} \ddot{\rho} - \rho\dot{\theta}^2 &= F(\rho)/m \\ \rho\ddot{\theta} + 2\dot{\rho}\dot{\theta} &= 0\,. \end{aligned}$$

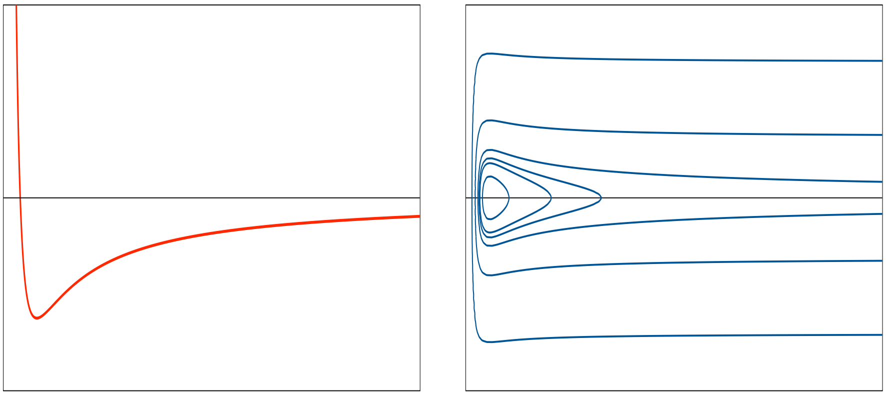
The second equation says that the "areal velocity" $\ell = \rho^2\dot{\theta}$ is a constant of the motion (Kepler's second law).

Taking Newton's gravitational force $F(\rho) = -\frac{GmM}{\rho^2}$ (where M is the mass of the Sun and G is the gravitational constant), the first equation may be written as

$$m\ddot{\rho} = -\frac{\partial}{\partial\rho}V_\ell(\rho)\,,$$

where we defined the "effective potential energy" as $V_\ell(\rho) = \frac{1}{2}m\frac{\ell^2}{\rho^2} - G\frac{mM}{\rho}$. The conserved energy is

$$E = \frac{1}{2}m\dot{\rho}^2 + \frac{1}{2}m\frac{\ell^2}{\rho^2} - G\frac{mM}{\rho}\,.$$



Kepler's effective potential and some energy level sets.

Now we set $\rho = 1/x$ and look for a differential equation for x as a function of θ. Computation shows that $dx/d\theta = -\dot{\rho}/\ell$, and, using conservation of ℓ, that $d^2x/d\theta^2 = -\rho^2\ddot{\rho}/\ell^2$. There follows that the first Newton equation reads

$$\frac{d^2x}{d\theta^2} + x = -\frac{1}{\ell^2 x^2 m} F\left(1/x\right) .$$

we get

$$\frac{d^2x}{d\theta^2} + x = -\frac{GM}{\ell^2} .$$

The general solution of this second order linear differential equation is

$$x\left(\theta\right) = \frac{GM}{\ell^2}\left(1 + e\cos\left(\theta - \theta_0\right)\right) ,$$

for some constants e and θ_0. Back to the original radial variable we get the solution

$$\rho\left(\theta\right) = \frac{\ell^2/GM}{1 + e\cos\left(\theta - \theta_0\right)} ,$$

Hence, orbits are conic sections with eccentricity e and focus at the origin: an ellipse for $0 \leq e < 1$ (corresponding to negative energy, hence to planets, and this is Kepler's first law), a parabola for $e = 1$ (corresponding to zero energy), an hyperbola for $e > 1$ (corresponding to positive energy). .

Phenomenological models. A number of phenomenological models (i.e. models which are not fundamental laws of nature), like the ones below, are also a source of interesting dynamical behaviour.

e.g. **Lotka-Volterra system.** The *Lotka-Volterra system* is the first-order non-linear differential equation

$$\begin{aligned} \dot{x} &= ax - bxy \\ \dot{y} &= -cy + dxy \end{aligned}$$

It has been proposed by Vito Volterra[17] to model competition between x preys and y predators, and by Alfred J. Lotka[18] to model the cyclic behavior of certain chemical reactions, like the abstract sceme

$$A + X \to 2X \qquad X + Y \to 2Y \qquad Y \to B$$

[17]Vito Volterra, Variazioni e fluttuazioni del numero d'individui in specie di animali conviventi, *Mem. Acad. Lincei* **2** (1926), 31-113. Vito Volterra, *Leçons sur la Théorie Mathématique de la Lutte pour la Vie*, Paris 1931.

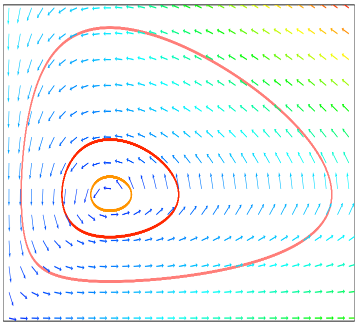
[18]Alfred J. Lotka, *J. Amer. Chem. Soc* **27** (1920), 1595. Alfred J. Lotka, *Elements of physical biology*, Williams & Wilkins Co. 1925.

Preys increase exponentially at rate a and are killed at rate proportional to the probability of beeing captured by a predator, while predators decrease exponentially at rate c and increase at rate proportional to the probability of capturing preys.

The function

$$H(x,y) = dx + by - c\log x - a\log y$$

is a constant of the motion, i.e. $\frac{d}{dt}H(x(t),y(t)) = 0$. There follows that orbits are contained (and actually are) in the level curves of H.



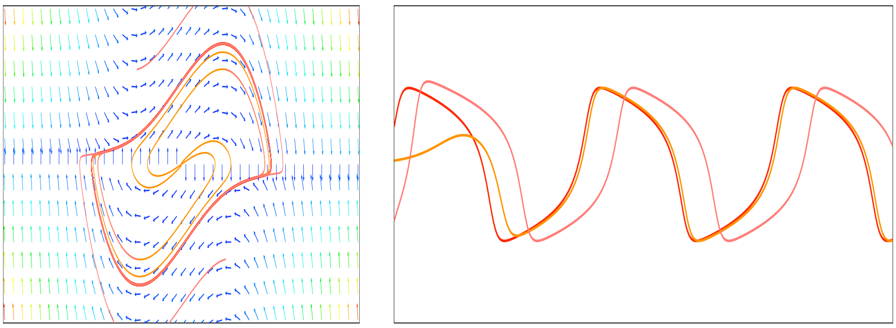
Phase portrait of the Lotka-Volterra system.

ex: Discuss the possible dynamics depending on the values of the parameters.

e.g. Van der Pols oscillator. The *van der Pol oscillator*[19] is the second-order non-linear differential equation

$$\ddot{q} - \mu(1-q^2)\dot{q} + q = 0$$

which models current in a circuit with a non-liner element.



Phase portrait and time series of the Van der Pols oscillator.

e.g. Brusselator. The *Brusselator* is an auto-cathalytic model proposed by Ilya Prigogine and collaborator [20] which models the abstract reaction

$$A \to X \qquad B + X \to Y + C \qquad 2X + Y \to 3X \qquad X \to D$$

and reads

$$\begin{aligned} \dot{x} &= \alpha - (\beta+1)x + x^2 y \\ \dot{y} &= \beta x - x^2 y \end{aligned}$$

[19] B. van der Pol, A theory of the amplitude of free and forced triode vibrations, *Radio Review* **1** (1920), 701-710 and 754-762. B. van der Pol and J. van der Mark, Frequency demultiplication, *Nature* **120** (1927), 363-364.

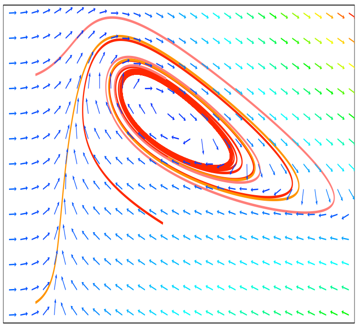
[20] I. Prigogine and R. Lefever, Symmetry breaking instabilities in dissipative systems, *J. Chem. Phys.* **48** (1968), 1655-1700. P. Glansdorff and I. Prigogine, *Thermodynamic theory of structure, stability and fluctuations*, Wiley, New York 1971. G. Nicolis and I. Prigogine, *Self-organization in non-equilibrium chemical systems*, Wiley, New York 1977.

ex: Observe what happens to the concentrations X e Y, namely x and y, when the concentrations $[A] \sim \alpha$ and $[B] \sim \beta$ are kept constant.

ex: Simulate the system

$$\begin{aligned}\dot{x} &= \alpha - (b+1)x + x^2 y\\ \dot{y} &= bx - x^2 y\\ \dot{b} &= -bx + \delta\end{aligned}$$

for the concentrations of X, Y and B, obtained when the concentration $[A] \sim \alpha$ is mantained constant and B in injected with constant velocity $v \sim \delta$.

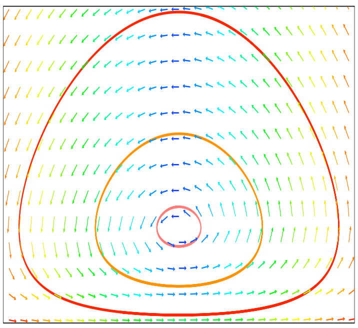


Phase portrait of the Brussellator.

e.g. **Goodwin oscillator.** A system modeling the interaction protein-mRNA was poposed by Goodwin[21]

$$\begin{aligned}\dot{M} &= \tfrac{1}{1+P} - \alpha\\ \dot{P} &= M - \beta\end{aligned}$$

where M nd P denote the relative concentrations of mRNA and protein, respectively.



Phase portrait of the Goodwin oscillator.

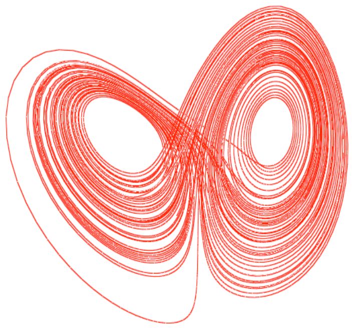
e.g. **Lorenz attractor.** Finally, we mention the *Lorenz system* [22]

$$\begin{aligned}\dot{x} &= \sigma(y-x)\\ \dot{y} &= x(\rho - z) - y\\ \dot{z} &= xy - \beta z\end{aligned}$$

For values of the parameters like $\sigma \simeq 10$, $\rho \simeq 28$ and $\beta \simeq 8/3$, one observe trajectories which diverge from one another, and yet oscillate all along the figure-eight above.

[21]B.C. Goodwin, *Temporal organization in cells*, Academic Press, London/New York 1963. B.C. Goodwin, Oscillatory behaviour in enzymatic control processes, *Adv. Enzyme Regul.* **3** (1965), 425-438.

[22]E.N. Lorenz, Deterministic nonperiodic flow, *J. Atmspheric Science* **20** (1963), 130-141.



Some orbits of the Lorenz attractor.

This strange phenomenon motivated an important part of the modern theory of dynamical systems.

4 Linear systems

The simplest higher-dimensional systems are described by linear differential equations. They provide models for the local behaviour of more general systems.

4.1 Exponential of a linear operator

Linearity & exponentials. The exponential $x(t) = e^{\lambda t}$ is the unique solution of the differential equation $\dot{x} = \lambda x$ with initial condition $x(0) = 1$. Moreover, it satisfies the functional equation $x(t+s) = x(t)\,x(s)$, which says that $\exp : \mathbb{R} \to \mathbb{R}^\times$ defines a homomorphism from the additive group $\mathbb{R}$ into the multiplicative group $\mathbb{C}^\times$. If we try to solve a system of linear homogeneous differential equations like

$$\dot{x} = Ax\,,$$

with $x \in \mathbb{R}^n$ and $A \in \mathrm{Mat}_{n\times n}(\mathbb{R})$, we are tempted to look for a solution as

$$x(t) = e^{tA}x(0)\,.$$

In the following, we recall how to give a meaning to such an expression, and prove that it solves the problem. The functional equation will say that e^{tA} is a one-parameter subgroup of the general linear group $\mathrm{GL}_n(\mathbb{R})$. The practical computation of the exponential of a matrix will make use of diagonalization, commutativity, and related considerations. More important, some qualitative aspects of solutions will derive simply from considerations on the spectrum of A, the eigenvalues of its complexification.

Exponential of a linear operator. The *exponential* of the square matrix $A = (a_{ij}) \in \mathrm{Mat}_{n\times n}(\mathbb{C})$ is the square matrix e^A, or $\exp(A)$, defined by the power series

$$\boxed{\begin{aligned} e^A &:= \sum_{k=0}^{\infty} \frac{1}{k!}A^k \\ &= I + A + \frac{1}{2}A^2 + \frac{1}{6}A^3 + \dots \end{aligned}} \tag{4.1}$$

This definition makes sense because each entry of r.h.s. above is the sum of an absolutely convergent series. To see this, observe that the operator norm $\|A\| := \sup_{v\in\mathbb{C}^n,\,\|v\|=1}\|Av\|$ is multiplicative, i.e. satisfies $\|AB\| \le \|A\|\,\|B\|$. This implies the bound

$$\|A^k/k!\| \le \|A\|^k/k!$$

There follows, since all norms in a finite dimensional vector space are equivalent, that the absolute value of each entry of the series (4.1) is bounded by a constant times the convergent series $\sum_{k=0}^{\infty}\|A\|^k/k! = e^{\|A\|}$. Bytheway, this also implies the bound

$$\left\|e^A\right\| \le e^{\|A\|}\,.$$

It is clear that if the matrix A is real, then also its exponential e^A is real.

If A and B are similar matrices, i.e. $A = U^{-1}BU$ for some $U \in \mathrm{GL}_n(\mathbb{C})$, then also their exponentials are similar, since powers of A are $A^n = U^{-1}B^nU$ for all $n \ge 0$, and therefore one easily justifies the following computation

$$\begin{aligned} e^A &= I + U^{-1}BU + \tfrac{1}{2}U^{-1}B^2U + \dots \\ &= U^{-1}\left(I + B + \tfrac{1}{2}B^2 + \dots\right)U \\ &= U^{-1}\,e^B\,U\,. \end{aligned} \tag{4.2}$$

Therefore, if L is a linear operator defined in a finite-dimensional vector space isomorphic to $\mathbb{C}^n$ or $\mathbb{R}^n$, represented in some fixed basis by the matrix A, then the formula (4.1) defines a linear operator

$$e^L = I + L + \frac{1}{2}L^2 + \frac{1}{6}L^3 + \dots$$

According to formula (4.2), this definition does not depend on the chosen basis.

Exponential of diagonalizable matrices. If A is a diagonal matrix with eigenvalues λ_k's, i.e.

$$\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n) := \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}$$

(missing entries are zero) then a straightforward computation shows that its exponential is also diagonal, and indeed

$$e^\Lambda = \mathrm{diag}\left(e^{\lambda_1}, \dots, e^{\lambda_n}\right) = \begin{pmatrix} e^{\lambda_1} & & & \\ & e^{\lambda_2} & & \\ & & \ddots & \\ & & & e^{\lambda_n} \end{pmatrix}.$$

In particular, if A is diagonalizable, i.e. $A = U^{-1}\Lambda U$ with Λ diagonal and $U \in \mathrm{GL}_n(\mathbb{C})$, then its exponential is similar to the diagonal matrix e^Λ, namely $e^A = U^{-1}e^\Lambda U$. Thus, exponentials of diagonalizable matrices are easy to compute, provided we know the change of coordinates U that diagonalizes the matrix.

An important consequence is a relation between the exponential and the principal invariants of a square matrix, the determinant and the trace. It says that

$$\boxed{\det\left(e^A\right) = e^{\mathrm{tr}A}} \tag{4.3}$$

This formula is obvious if A is diagonalizable, and follows by continuity in the general case, because the set of diagonalizable matrices is dense in the space $\mathrm{Mat}_{n\times n}(\mathbb{C})$ of complex square matrices (a generic degree n complex polynomial has n distinct roots).

ex: Show that if v is an eigenvector of the linear operator L with eigenvalue λ, then v is also an eigenvector of e^L, with eigenvalue e^λ.

One-parameter groups of matrices. Given a matrix $A \in \mathrm{Mat}_{n\times n}(\mathbb{C})$, we may consider the family of matrices

$$G(t) := e^{tA},$$

parametrized by a "time" $t \in \mathbb{R}$. It is clear that $G(0) = I$. The series of functions $t \mapsto (e^{tA})_{ij}$ which define the entries of e^{tA} converge uniformly in any bounded interval of the real line, as well the series of their derivatives. In particular, the time derivatives may be computed term-wise. The result is that

$$\frac{d}{dt}G(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!}A^{k+1} = A\,G(t) = G(t)\,A \tag{4.4}$$

In particular, A commutes with $G(t)$.

The derivative of $F(t) := e^{tA}e^{-tA}$ is equal, by the Leibniz rule applied to every entry of the product, to $F'(t) = A\,F(t) - F(t)\,A = 0$, because A commutes with $G(t)$. By the mean value theorem, $F(t) = F(0) = I$. Therefore, $G(t) = e^{tA}$ is invertible, and its inverse is $(e^{tA})^{-1} = e^{-tA}$. Thus, the exponential sends $\exp : \mathrm{Mat}_{n\times n}(\mathbb{C}) \to \mathrm{GL}_n(\mathbb{C})$.

Theorem 4.1. *Let $A \in \mathrm{Mat}_{n\times n}(\mathbb{C})$. The unique solution of the linear differential equation*

$$\dot X = AX \qquad \text{or} \qquad \dot X = XA\,,$$

with initial condition $X(0) = X_0 \in \mathrm{GL}(n, \mathbb{C})$, is

$$X(t) = e^{tA}\,X_0 \qquad \text{or} \qquad X(t) = X_0\,e^{tA}\,,$$

respectively.

Proof. It is clear, by the above computation, that $e^{tA}X_0$ or X_0e^{tA} are solutions of the two problems. In order to prove uniqueness, we may observe that if $X(t)$ is a solution, then the matrix $X(t)\,e^{-tA}$ (or $e^{-tA}X(t)$ in the second case) does not depend on time, since its derivative is zero, and therefore is constant and equal to its initial value X_0. □

Observe that the two differential equations in the above theorem are not the same, the product between matrices does not commute, in general. Indeed, if A and B do not commute, the three exponentials e^{A+B} and e^Ae^B and e^Be^A may all be different from each other. What is true is the following.

Theorem 4.2. *If A and B commute, i.e. if $AB = BA$, then*

$$e^{A+B} = e^Ae^B = e^Be^A\,.$$

Proof. If A commutes with B, the all its powers A^k also commute with all the powers B^j, and therefore with the exponentials e^{tB} and e^{tA}, and viceversa. There follows that the derivative of

$$H(t) = e^{t(A+B)} - e^{tA}e^{tB}$$

is, using formulas 4.4,

$$H'(t) = (A+B)\,e^{t(A+B)} - A\,e^{tA}e^{tB} - e^{tA}e^{tB}\,B = (A+B)\,H(t)$$

By the uniqueness theorem 4.1, $H(t) = e^{t(A+B)}H(0)$. But $H(0) = 0$, therefore $H(t) = 0$ for all times t, and in particular for $t = 1$. □

In particular, since all multiples tA of A commute, the family of the $G(t) = e^{tA}$, with $t \in \mathbb{R}$, is a *one-parameter subgroup* of the general linear group $\mathrm{GL}_n(\mathbb{C})$, i.e. satisfies

$$e^{0A} = I \qquad \text{and} \qquad e^{tA}e^{sA} = e^{(t+s)A}\,.$$

In other words, the correspondence $t \mapsto e^{tA}$ is an homomorphism of the additive group $\mathbb{R}$ into $\mathrm{GL}_n(\mathbb{C})$. Its image is a curve in the linear group, which passes through the identity for $t = 0$, and solves the differential equation $\dot{G} = AG$. The matrix A is called *generator* of the subgroup $\{G(t)\}_{t\in\mathbb{R}}$, and may be obtained as the derivative

$$A = \dot{G}(0) = \lim_{t\to 0} \frac{G(t) - I}{t}\,.$$

Thus, A is the velocity of the curve $G(t)$ at time $t = 0$.

4.2 Linear fows

Linear systems. A *homogeneous linear system* with constant coefficients is an autonomous differential equation

$$\dot{x} = L(x) \tag{4.5}$$

for $x(t) \in \mathbb{R}^n$, defined by a linear vector field $L \in \mathrm{End}(\mathbb{R}^n)$. The origin is an equilibrium solution, since $L(0) = 0$ by linearity. Fixed a basis of $\mathbb{R}^n$, e.g. the canonical basis, the system may be written in matrix notation as

$$\dot{x} = Ax\,,$$

where $x(t) = (x_1(t), x_2(t), \ldots, x_n(t))^\top \in \mathbb{R}^n$ is a column vector, $A = (a_{ij}) \in \mathrm{Mat}_{n\times n}(\mathbb{R})$ is the matrix which represents the linear vector field L in the chosen basis, and Ax denotes the usual product between matrices. By the proof of theorem 4.1, the solution with initial condition $x(0) = x_0 \in \mathbb{R}^n$ is given by

$$x(t) = e^{tA}x_0\,.$$

The flow of the linear vector field L is the one-parameter group of linear maps $\Phi_t = e^{tL}$, given, in the chosen basis, by $\Phi_t(x) := e^{tA}x$.

Thus, if we want to understand solutions of a linear system, we must compute the exponential of the linear vector field, in some convenient basis.

Diagonalizable linear systems. Assume that A is diagonalizable, and has n real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ (not necessarily distinct) with linearly independent eigenvectors $v_1, v_2, \ldots, v_n$, respectively, so that $Av_k = \lambda_k v_k$ and the v_k's form a basis of $\mathbb{R}^n$. Then the solution of (4.5) with initial conditions $x(0) = \sum_k a_k v_k$is a superposition

$$x(t) = \sum_{k=1}^{n} e^{t\lambda_k} a_k v_k$$

The qualitative asymptotic behavoiur of solutions is therefore decided by the signs of the eigenvalues.

For example, if all the eigenvalues are negative, i.e. $\lambda_k < 0$ for all $k = 1, \ldots, n$, then all solutions decay, exponentially fast to the origin, i.e.

$$\|x(t)\| \leq e^{-\alpha t} \|x(0)\|$$

for some $\alpha = \min_k |\lambda_k| > 0$. The origin is then an "asymptotically stable" equilibrium, or a "sink".

If, on the other side, all the eigenvalues are positive, i.e. $\lambda_k > 0$ for all $k = 1, \ldots, n$, then all solutions different from the equilibrium solution diverge exponentially fast, i.e.

$$\|x(t)\| \geq e^{\beta t} \|x(0)\|$$

for some $\beta = \min_k \lambda_k > 0$. The origin is an "asymptotically unstable equilibrium", or a "source".

More interesting is the mixed situation of a saddle, with some stable directions and some unstable directions. The case with some zero eigenvalue, i.e. some indifferent directions, is clearly non generic, although physically interesting (the harmonic oscillator is such a case!).

On the other side, generic real matrices are not diagonalizable. To understand their exponentials, we must complexify and use the Jordan normal form.

Complexification. The *complexification* of the real vector space $\mathbb{R}^n$ is the complex vector space $\mathbb{C}^n := \mathbb{R} \oplus i\mathbb{R}$, i.e. the set of vectors $z = x \oplus iy \approx x + iy$, with $x, y \in \mathbb{R}^n$, equipped with the natural sum and multiplication by complex scalars.

The complexification of the linear map $x \mapsto L(x)$ defined, in the canonical basis of $\mathbb{R}^n$, by a matrix $A \in \mathrm{Mat}_{n\times n}(\mathbb{R})$ according to $x \mapsto Ax$, is the linear operator $z \mapsto L^{\mathbb{C}}(z)$ defined by the same matrix, i.e. according to $z = x + iy \mapsto Az = Ax + iAy$.

The *spectrum* of the linear operator L (in a finite dimensional linear space) is the set $\sigma(L) \subset \mathbb{C}$ of the eigenvalues of its complexification $L^{\mathbb{C}}$, i.e. complex roots of the characteristic polynomial $P_A(t) := \det(t - A)$. By Gauss' fundamental theorem of arithmetic, the characteristic polynomial factorizes as a product

$$P_A(t) = \prod_{\lambda \in \sigma(L)} (t - \lambda)^{m_\lambda} .$$

The integer exponent m_λ is called *(algebraic) multiplicity* of the eigenvalue λ. It is clearl that $\sum_{\lambda\in\sigma(A)} m_\lambda = n$.

Complexification in dimension two. The relevant example, for our purposes, is the following. Let $x \mapsto L(x)$ be the linear operator defined, in the canonical basis $e_1 = (1, 0)$ and $e_2 = (0, 1)$ of $\mathbb{R}^2$, by the real matrix

$$A = \begin{pmatrix} \alpha & \omega \\ -\omega & \alpha \end{pmatrix}$$

Thus, $L(e_1) = \alpha e_1 - \omega e_2$ and $L(e_2) = \omega e_1 + \alpha e_2$. Then the complexified operator $z \mapsto L^{\mathbb{C}}(z)$ is defined, in the basis $v_+ = e_1 + ie_2$ and $v_- = e_1 - ie_2$ of $\mathbb{C}^2$, by the diagonal matrix

$$\Lambda = \begin{pmatrix} \lambda & 0 \\ 0 & \overline{\lambda} \end{pmatrix}$$

where $\lambda = \alpha + i\omega$. Indeed, a computation shows that $A(e_1 + ie_2) = (\alpha + i\omega)(e_1 + ie_2)$ and $A(e_1 - ie_2) = (\alpha - i\omega)(e_1 - ie_2)$.

Vice-versa, let $L^{\mathbb{C}}$ be the complexification of a real operator defined by the real two-by-two matrix A, and let v_+ be an eigenvector of $L^{\mathbb{C}}$ with eigenvalue $\lambda = \alpha + i\omega$, so that, in the canonical basis, $Av_+ = \lambda v_+$. Then $v_- := \overline{v_+}$ is an eigenvector of $L^{\mathbb{C}}$ with eigenvalue $\overline{\lambda} = \alpha - i\omega$. Indeed, since the entries of A are real, the roots of the characteristic polynomial comes in pairs of conjugated complex numbers, and one check that $A\,v_- = A\,\overline{v_+} = \overline{Av_+} = \overline{\lambda}\,\overline{v_+} = \overline{\lambda}\,v_-$. There follows that, in the real basis $e_+ = (v_+ + v_-)/2$ and $e_- = (v_+ - v_-)2i$, which is therefore a basis of the real vector space $\mathbb{R}^2 \subset \mathbb{C}^2$, the real operator L is represented by the matrix A as above.

So, a diagonalizable complexified real linear operator in the plane with a couple of complex conjugate eigenvaules $\lambda_\pm = \alpha \pm i\omega$ "corresponds" to a two-by-two real matrix which is the sum of a multiple of the identity αI and an anti-symmetric matrix Ω as below

$$\alpha I + \Omega := \alpha \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix} .$$

Since any matrix commute with any multiples of the identity, by theorem 4.2 we may compute separately the exponentials of $t\rho I$ and $t\Omega$, and then multiply the results. The flow of the diagonal part is simply $e^{t\alpha I} = e^{\alpha t} I$. A computation (using the power series of the trigonometric functions $\sin t$ and $\cos(t)$) shows that the flow defined by the antisymmetric matrix Ω above is

$$e^{t\Omega} = \begin{pmatrix} \cos(\omega t) & \sin(\omega t) \\ -\sin(\omega t) & \cos(\omega t) \end{pmatrix}$$

i.e. it is a clockwise rotation $R_{t\omega}$ by an angle $t\omega$. Multiplying, we finally get

$$e^{tA} = e^{t(\alpha I + \Omega)} = e^{\alpha t} \begin{pmatrix} \cos(\omega t) & \sin(\omega t) \\ -\sin(\omega t) & \cos(\omega t) \end{pmatrix}$$

So, the flow of A is a rotation with angular frequency ω (or frequency $\nu = \omega/2\pi$) together with stretching/contraction with exponential rate α. Orbits are logarithmic spirals entering or coming from the origin, depending on the sign of α.

The case $\alpha = 0$ corresponds to pure rotations (this is the case of the harmonic oscillator $\ddot{x} = -\omega^2 x$).

4.3 Linear systems in the plane

Linear systems in the plane. We have now all the tools to understand the general linear system of differential equations

$$\begin{aligned} \dot{x} &= ax + by \\ \dot{y} &= cx + dy \end{aligned}$$

in the plane $\mathbb{R}^2$, defined by a real 2×2 matrix

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

Let λ_+ and λ_- be the eigenvalues of the complexification of A, i.e. the complex roots (possibly equal) of the characteristic polynomial $\det(tI - A)$. The product $\lambda_+\lambda_-$ of the eigenvalues is $q = \det(A) = ad - bc$, and the sum $\lambda_+ + \lambda_-$ of the eigenvalues is $p = \operatorname{tr}(A) = a + d$. Eigenvalues are therefore

$$\lambda_\pm = \frac{p \pm \sqrt{\Delta}}{2} ,$$

where the "discriminant" is $\Delta = p^2 - 4q$.

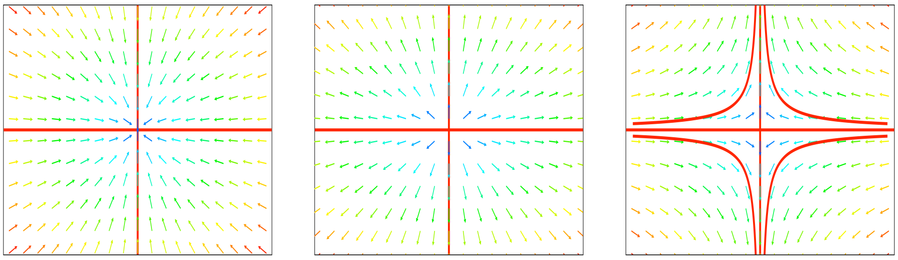
Real eigenvalues. If the matrix A is diagonalizable over the reals, i.e. admits two linearly independent eigenvectors with real eigenvalues $\rho_\pm \in \mathbb{R}$ (possibly equal), then the system is linearly equivalent to

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} \rho_+ & 0 \\ 0 & \rho_- \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Solutions are

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} e^{\rho_+ t} & 0 \\ 0 & e^{\rho_- t} \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$$

The origin is called *stable node* if $\rho_\pm < 0$, *unstable node* if $\rho_\pm > 0$, or *saddle* if $\rho_- < 0 < \rho_+$.



Stable node, unstable node and saddle.

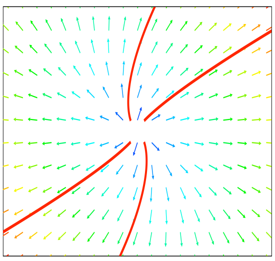
If the matrix A admits just one eigenvector, with eigenvalue $\rho \in \mathbb{R}$, then one can show that the system is linearly equivalent to

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} \rho & 1 \\ 0 & \rho \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Solutions are

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = e^{\rho t} \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$$

The origin is called *degenerate node*, stable or unstable, depending on the sign of ρ.



Degenerate unstable node.

ex: Show that if a two-by-tro real matrix A has only one eigenvector v, with eigenvalue λ, then one can find a second linearly independent vector w such that $Aw = \lambda w + v$. In this new basis v and w, the linear map $x \mapsto Ax$ is induced by the matrix

$$\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}.$$

Complex eigenvalues. If the matrix A has no real eigenvalue, then its complexification admits two complex conjugate eigenvalues λ and $\overline{\lambda}$. If the eigenvaues are purely imaginary, say $\lambda_\pm = \pm i\omega$, with $\omega > 0$, then, by the previous discussion, the system is linearly equivalent to

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

This is an harmonic oscillator $\ddot{x} = -\omega^2 x$ with angular frequency ω. Solutions are

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} \cos(\omega t) & \sin(\omega t) \\ -\sin(\omega t) & \cos(\omega t) \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$$

Orbits are ellipsis, and the origin is called *(indiffeent) focus*. Trajectories which start near the origin stay near the origin for all times, still not being asymptotic to the origin.

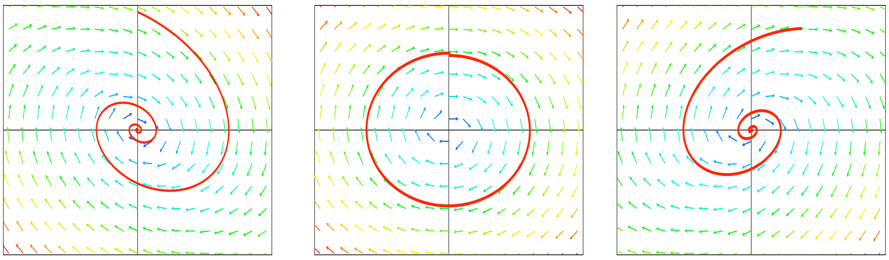
The generic case is a complexified matrix with complex eigenvalues $\lambda_\pm = \rho \pm i\omega$, with real part $\rho \neq 0$. The system is linearly equivalent to

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} \rho & \omega \\ -\omega & \rho \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Solutions are

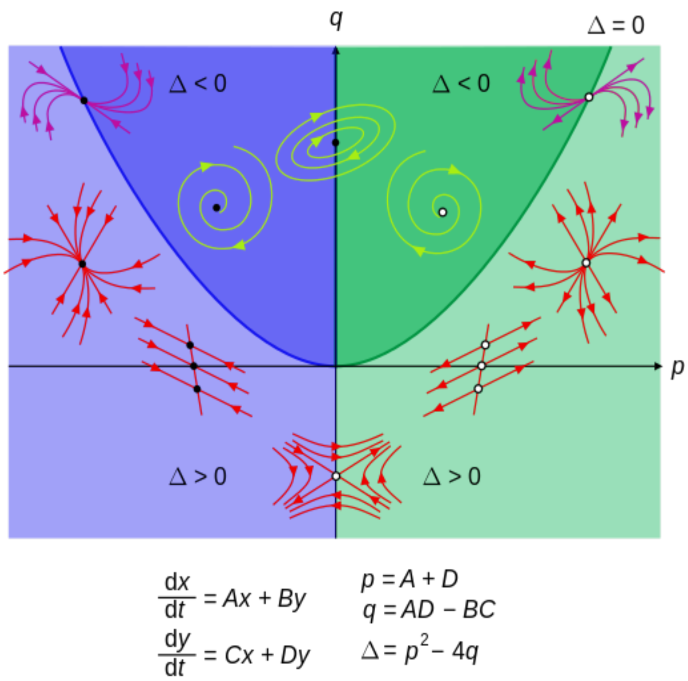
$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = e^{\rho t} \begin{pmatrix} \cos(\omega t) & \sin(\omega t) \\ -\sin(\omega t) & \cos(\omega t) \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$$

orbits are logarithmic spirals that comes or enter into the origin, depending on the sign of ρ. The origin is called *unstable focus* if $\rho > 0$, or *stable focus* if $\rho < 0$.



Stable, indifferent and unstable focus.

Global picture. It is clear that the stability or unstability of nodes or foci is preserved under small perturbations of the parameters (the entries of the matrix A). Here is a famous picture of the different phase portraits, depending on the trace and determinant of the matrix.



By Maschen, from Wikimedia Commons.

ex: Discuss the degenerate cases when one of the eigenvalues is zero (so that the matrix A is not invertible).

ex: Consider the "inverted oscillator"

$$\dot{q} = p$$
$$\dot{p} = q$$

Find the nature of the equilibrium, and determine the generic solution.

ex: Sketch the phase portrait (i.e. some orbits near the equilibrium in the phase space) of the following linear systems.

$$\begin{cases} \dot{x} = x - y \\ \dot{y} = x + y \end{cases} \qquad \begin{cases} \dot{x} = 2x + y \\ \dot{y} = x + y \end{cases} \qquad \begin{cases} \dot{x} = 4x \\ \dot{y} = 2x - y \end{cases}$$

$$\begin{cases} \dot{x} = 6x + 5y \\ \dot{y} = x + 2y \end{cases} \qquad \begin{cases} \dot{x} = -x + 2y \\ \dot{y} = 3y \end{cases} \qquad \begin{cases} \dot{x} = -7x + y \\ \dot{y} = -4x - 3y \end{cases}$$

$$\begin{cases} \dot{x} = y \\ \dot{y} = -4x \end{cases} \qquad \begin{cases} \dot{x} = -x + 5y \\ \dot{y} = -5x - y \end{cases} \qquad \begin{cases} \dot{x} = x + 5y \\ \dot{y} = -5x + y \end{cases}$$

ex: The current $I(t)$ in a LRC circuit is a solution of the homogeneous differential equation

$$L\ddot{I} + R\dot{I} + \frac{1}{C}I = 0$$

Write the corresponding linar system for $x(t) = I(t)$ and $y(t) = \dot{I(t)}$, and sketch the possible phase portraits, depending on the relative values of the positive parameters L, R and C.

4.4 Jordan normal form

In the higher-dimensional case, the useful normal form to understand exponentials is the Jordan normal form.

Generalized eigenspaces. Let $L : \mathbb{C}^n \to \mathbb{C}^n$ linear operator defined in a complex linear space $\mathbb{C}^n$. Given a scalar λ, let L_λ denotes the operator $L - \lambda$. If the kernel of L_λ is not trivial, then λ is an eigenvalue of L, and $V_\lambda = \text{kernel}(L_\lambda)$ is its associated proper space, made of eigenvectors v such that $Lv = \lambda v$.

A non-zero vector $v \in \mathbb{C}^n$ is said *generalized eigenvector* if it is in the kernel of some power of L_λ, i.e. if there exists $\lambda \in \mathbb{C}$ and some minimal integer $m \geq 1$ such that $L_\lambda^m v = 0$. The non-zero integer m is called *period* of v, and the vector v itself is also called L_λ-*cyclic* (meaning that the orbit of v by the map L_λ is formed by m distinct non-zero vector).

If the period is $p = 1$, then v in an eigenvector of L. In general, the m vectors

$$v_1 = L_\lambda^{m-1} v \qquad v_2 = L_\lambda^{m-2} v \qquad \dots \qquad v_m = v \tag{4.6}$$

are all generalized eigenvectors, since , $L_\lambda^k v_k = L_\lambda^k l_\lambda^{m-k} v = L_\lambda^m v = 0$, and the first one, v_1, is an eigenvector of L with eigenvalue λ.

Theorem 4.3. *If v is a L_λ-cyclic vector of period m, then the m vectors (4.6) are linearly independent and generate a L-invariant subspace of generalized eigenvectors.*

Proof. If $a_1 v_1 + a_2 v_2 + \cdots + a_m v_m = 0$ for some non-zero vector $(a_1, a_2, \dots, a_m)$, then $p(L_\lambda)v = 0$ where $p(t)$ is the non-zero polynomial

$$p(t) = a_1 t^{m-1} + a_2 t^{m-2} + \cdots + a_{m-1} t + a_m$$

But also $q(L_\lambda)v = 0$, where $q(t) = t^m$, because v is L_λ-cyclic with period m. If $h(t)$ denotes the maximum common divisor between the polynomials $p(t)$ and $q(t)$, then there exist polynomials $f(t)$ and $g(t)$ such that $h(t) = f(t)\,p(t) + g(t)\,q(t)$. Therefore, also $h(L_\lambda)v = 0$. But $h(t)$ is a power of t (since it divides t^m) of degree $\deg(h) = k \leq m - 1$ (since it divides $p(t)$). Thus, $h(t) = t^k$, and therefore $L_\lambda^k v = 0$. This contradicts the fact that m is the period of v. Thus, the vectors are linearly independent.

The v_k's are all generalized eigenvectors, because $L_\lambda^k v_k = L_\lambda^k l_\lambda^{m-k} v = L_\lambda^n v = 0$. Finally, the subspace generated by the v_k's is L-invariant, because

$$Lv_k = L(L_\lambda^{m-k} v) = L_\lambda^{m-k+1} v + \lambda\, L_\lambda^{m-k} v = v_{k-1} + \lambda v_k$$

where, clearly, we set $v_0 = (L - \lambda I)^n v = 0$. □

The kernels $\mathrm{kernel}(L_\lambda^k)$ are called *generalized eigenspaces* of order k, and one easily sees that $\mathrm{kernel}(L_\lambda) \subset \mathrm{kernel}(L_\lambda^2) \subset \cdots \subset \mathrm{kernel}(L_\lambda^n)$. Moreover, if λ is en eigenvalue of L, then the space of generalized eigenvectors associated to the eigenvalue λ is equal to $\mathrm{kernel}(L_\lambda^n)$ (which, of course, may coincide with $\mathrm{kernel}(L_\lambda^k)$ for some smaller $k \leq n$).

Jordan blocks. If it happens that the vectors (4.6) span the whole space $\mathbb{C}^n$, i.e. if $m = n$, then the entire space is said *cyclic*. The computation in the proof above shows that

$$Lv_1 = \lambda v_1 \qquad \text{and} \qquad Lv_k = \lambda v_k + v_{k-1} \qquad \text{for } 2 \leq k \leq n\,.$$

Therefore, the matrix which represents the linear operator L in this basis $v_1, v_2, \ldots, v_n$ is

$$J_\lambda = \begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{pmatrix} \tag{4.7}$$

In particular, v_1 is the unique eigenvector, with eigenvalue λ, the proper space $V_\lambda = \mathrm{kernel}(L_\lambda)$ being the line $\mathbb{C}v_1$. Thus, the geometric multiplicity of λ is equal to 1. The matrix (4.7) is called *Jordan block* of dimension n, and the basis (4.6) is called *Jordan basis*, or *Jordan chain* of lenght n. The vector v_n is called *generator*, or *lead vector* of the Jordan chain.

Observe that a Jordan block of dimesion n has the form

$$J_\lambda = \lambda I + N$$

where N is the nilpotent (upper triangular) matrix

$$N = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{pmatrix} \tag{4.8}$$

which satisfies $Ne_k = e_{k-1}$, if the e_k's denote the column vectors of the canonical basis of $\mathbb{C}^n$, and $N^n = 0$.

The characteristic polynomial of a Jordan block J of lenght n is $P_J(z) = (z-\lambda)^n$, and therefore the algebraic multiplicity of the eigenvalue λ is n. The minimal polynomial (the monic polynomial of minimal degree such that $f(J) = 0$) is also $M_J(z) = (z - \lambda)^n$ (to be compared with the minimal polynomial of the diagonalizable matrix $\Lambda = \lambda I$ of order n, which is only $M_\Lambda(z) = (z - \lambda)$).

e.g. **Derivative and quasi-polynomials.** The paradigmatic example is the derivative operator ∂, defined by $(\partial f)(t) := f'(t)$ on complex-values functions $f(t)$ of a real variable t. Its eigenfunctions are the exponentials $e^{\lambda t}$, since $\partial(e^{\lambda t}) = \lambda\, e^{\lambda t}$. On the other side, it is nilpotent on the space of polynomials $p(t)$ of fixed degree, say $\deg(p) < n$, where $\partial^n = 0$. Exponentials and polynomials combine to form the spaces $Q_{\lambda,n} \approx \mathbb{C}^n$ of quasi-polynomials $f(t) = p(t)e^{\lambda t}$, where λ is a fixed complex exponent and the $p(t)$'s are polynomials of $\deg(p) < n$. These are cyclic spaces for the derivative, since $(\partial - \lambda)^n = 0$, and a generating vector is $t^{n-1}e^{\lambda t}$. The eigenvector of ∂ is, of course, $f(t) = e^{\lambda t}$, and has eigenvalue λ. A Jordan basis is

$$e^{\lambda t} \qquad t\, e^{\lambda t} \qquad \tfrac{1}{2}\, t^2\, e^{\lambda t} \qquad \ldots \qquad \tfrac{1}{(n-1)!}\, t^{n-1}\, e^{\lambda t}$$

In this basis, the operator ∂ is represented by the matrix (4.7).

Flow of a Jordan block. The exponential of t times a Jordan block J_λ of dimension n, which defines the flow of the linear differential equation

$$\dot{v} = J_\lambda \, v$$

defined in a cyclic space, is easily computed. Indeed, since N commute with λI, we may compute separately the two exponentials and then multiply. But since N is nilpotent, namely $N^n = 0$, the series defining the exponential terminates, and indeed

$$e^{tN} = I + tN + \frac{t}{2} N^2 + \cdots + \frac{t^{n-1}}{(n-1)!} N^{n-1} \, .$$

So, the exponential of tJ_λ is simply

$$e^{tJ_\lambda} = e^{t\lambda} \left(I + tN + \frac{t}{2} N^2 + \cdots + \frac{t^{n-1}}{(n-1)!} N^{n-1} \right) \, .$$

It is clear, since polynomials corrections are negligible compared with exponential growth or decay, that the asymptotic behaviour of solutions of the linear system $\dot{v} = J_\lambda v$ only depends on the sign of the real part of λ, provided it is not zero.

Theorem 4.4. *If $\Re(\lambda) < 0$, then for all $0 < \alpha < |\Re(\lambda)|$ there exists a constant C such that*

$$\|e^{tJ_\lambda} v\| \leq C \, e^{-\alpha t} \, \|v\| \qquad \text{for} \;\; t \geq 0 \, .$$

Proof. If we write a generic vector as a superposition $v = \sum_k a_k v_k$ of the vectors v_k's of the Jordan basis, we see that

$$e^{tJ_\lambda} v = e^{\lambda t} \left(\sum_i a_i \, p_{ik}(t) \right) v_k$$

where the $p_{ik}(t)$'s are certain polynomials of degree $< n$, which only depend on the dimension n of the Jordan block. Assume that $\Re(\lambda) = -\rho < 0$. Take any $0 < \alpha < \rho$, and set $\varepsilon = \rho - \alpha > 0$. We may define a norm on the cyclic space according to $\|v\|_\lambda := \max_k |a_k|$. Then, if $M = \max_{i,k} M_{ik}$ denotes the maximal value of the $M_{ik} = \sup_{t \geq 0} |e^{-t\varepsilon} p_{ik}(t)|$, we clearly have

$$\|e^{tJ_\lambda} v\|_\lambda \leq M \, e^{-\alpha t} \, \|v\|_\lambda$$

for all $t \geq 0$. Since all norms in a finite dimensional vector space are equivalent, this finally implies that claimed inequality, for some other constant C, holds for the standard or any other norm in the cyclic space. □

Thus, if $\Re(\lambda) < 0$, all vectors are exponentially contracted by the flow of J_λ, and decay to zero exponentially fast as $t \to \infty$.

Reversing the arrow of time, one shows that if $\Re(\lambda) = \rho > 0$ and $\rho > \beta > 0$, then there exists a constant C such that

$$\|e^{-tJ_\lambda} v\| \leq C \, e^{-\beta t} \, \|v\| \qquad \forall \, t \geq 0$$

Thus, if $\Re(\lambda) > 0$, all vectors are exponentially stretched by the flow of J_λ, and decay to zero exponentially fast as $t \to -\infty$.

Jordan normal form. It happens that any linear operator in a finite dimensional complex vector space is a direct sum of Jordan blocks.

Theorem 4.5 (Jordan normal form)**.** *Let L be a linear operator in a finite-dimensional complex vector space $\mathbb{C}^n$. The total space splits as a direct sum $\mathbb{C}^n = E_{\lambda_1} \oplus E_{\lambda_2} \oplus \cdots \oplus E_{\lambda_d}$ of cyclic L-invariant subspaces.*

Therefore, if we chose a Jordan basis in any invariant cyclic subspace E_{λ_k}, the matrix that represents the linear operator L in the resulting basis is block diagonal as

$$J = \begin{pmatrix} J_{\lambda_1} & 0 & \dots & 0 \\ 0 & J_{\lambda_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & J_{\lambda_d} \end{pmatrix} \tag{4.9}$$

where each $J_{\lambda_k} = \lambda_k I + N_k$ is a Jordan block as (4.7). The λ_k's are the eigenvalues of L, the roots of the characteristic polynomial $P_A(z) = \det(zI - A)$, where A is the matrix that represents L in the canonical basis. Indeed, the characteristic polynomial factorizes as a product $P_A(z) = \prod_{\lambda\in\sigma(A)} (z-\lambda)^{m_\lambda}$, where m_λ is the *(algebraic) muliplicity* of the eigenvalue λ, which is equal to the sum of the dimensions of the Jordan blocks with $\lambda_k = \lambda$, i.e. to the dimension of the generalized eigenspace kernel(L_λ^n). The *geometric multiplicity* of the eigenvalue λ is the dimension of the proper space kernel(L_λ), which is equal to the cardinality of those Jordan blocks with $\lambda_k = \lambda$. The minimal polynomial of A is a product $M_A(z) = \prod_{\lambda\in\sigma(A)}(z-\lambda)^{\mu_\lambda}$, where μ_λ is the dimension of the largest Jordan block with $\lambda_k = \lambda$.

If A is the matrix that represents the linear operator L in the canonical basis (or in any other basis), then there exists an invertible matrix $G \in \mathrm{GL}_n(\mathbb{C})$ (whose columns are the vectors of the Jordan bases) such that $G^{-1} A G = J$. The *canonical form* J is unique modulo permutations of the blocks. In particular, the matrix A may be represented as a sum

$$A = \Lambda + N$$

of a *semi-simple*, i.e. diagonalizable, matrix $\Lambda = G(\lambda_1 I \oplus \lambda_2 I \oplus \dots) G^{-1}$ and a nilpotent matrix $N = G(N_1 \oplus N_2 \oplus \dots) G^{-1}$ which commute, i.e. such that $\Lambda N = N\Lambda$.

Clear proofs of the Jordan normal form theorem 4.5 can be found in the classical [HS74], or in any good reference on linear algebra, as for example [La87, Ax97].

Normal form of real operators. We now consider a linear operator $L : \mathbb{R}^n \to \mathbb{R}^n$, defined, in the canonical basis, by a matrix $A \in \mathrm{Mat}_{n\times n}(\mathbb{R})$. We may think at A as a complex matrix, representing the complexified operator $L^{\mathbb{C}} : \mathbb{C}^n \to \mathbb{C}^n$, and as such conjugated to a block diagonal matrix as (4.9) above. Eigenvalues are real, or come in couples of complex conjugated pairs $\lambda_\pm = \alpha \pm i\omega$, since the characteristic polynomial has real coefficients.

Theorem 4.6 (Jordan normal form for real operators)**.** *Let L be a linear operator on the real vector space $\mathbb{R}^n$. The total space splits as a direct sum of invariant subspaces E_λ or $E_{\lambda,\overline{\lambda}}$, namely*

$$\mathbb{R}^n = \left(\bigoplus_{\lambda\in\mathbb{R}} E_\lambda\right) \oplus \left(\bigoplus_{\lambda\in\mathbb{C}\backslash\mathbb{R}} E_{\lambda,\overline{\lambda}}\right),$$

where the operator is represented by a matrix of the form (4.7), for some real eigenvalue λ, or by a matrix of the form

$$J_{\lambda,\overline{\lambda}} = \begin{pmatrix} R_{\lambda,\overline{\lambda}} & I & & & \\ & R_{\lambda,\overline{\lambda}} & I & & \\ & & \ddots & \ddots & \\ & & & R_{\lambda,\overline{\lambda}} & I \\ & & & & R_{\lambda,\overline{\lambda}} \end{pmatrix} \tag{4.10}$$

with

$$R_{\lambda,\overline{\lambda}} = \alpha I + \Omega = \begin{pmatrix} \alpha & \omega \\ -\omega & \alpha \end{pmatrix} \qquad \text{and} \qquad I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

for some couple of complex conjugated eigenvalues $\lambda = \alpha + i\omega$ and $\overline{\lambda} = \alpha - i\omega$, respectively.

Proof. According to the Jordan normal form theorem 4.5, there exists a basis of $\mathbb{C}^n$ such that the complexified operator $L^{\mathbb{C}}$ is represened by a block diagonal Jordan matrix.

Consider a Jordan block with real eigenvalue λ. If $z = x + iy$ is a L_λ-cyclic vector, then either x or y is a real L_λ-cyclic vector. This real cyclic vector generates, therefore, a real Jordan chain of the same real dimension as the complex dimension of original block.

We now consider a Jordan block with complex eigenvalue $\lambda = \alpha + i\omega$. If λ is not real, then the complexified operator also admts an eigenvalue $\overline{\lambda} = \alpha - i\omega$, and a corresponding Jordan block of equal dimension. Indeed, if v is L_λ-cyclic, then $\overline{v}$ is $L_{\overline{\lambda}}$-cyclic. Proceeding as in the two dimensional case, one easily sees that this couple of complex Jordan blocks give origin to a real Jordan block of the form (4.10). □

The invariant subspaces E_λ or $E_{\lambda,\overline{\lambda}}$ of theorem 4.6 are also referred to as *root spaces*, using a terminology borrowed from the theory of Lie algebras.

4.5 Hyperbolic linear flows

Stable and unstable spaces. Given a linear vector field L on $\mathbb{R}^n$, defined in the canonical basis by a real matrix A, we are interested in its flow $\Phi_t = e^{tL}$, which solves the linear system

$$\dot{x} = Ax\,.$$

We already saw that the asymptotic behavior of the flow e^{tL} in each root space depends on the sign of the real part of the corresponding eigenvalue.

One can write the total space as a direct sum of three invariant subspaces

$$\mathbb{R}^n = E^- \oplus E^0 \oplus E^+$$

where the *stable space* E^- is the direct sum of those root spaces with $\Re(\lambda) < 0$, the *unstable space* E^+ is the direct sum of those root spaces with $\Re(\lambda) > 0$, and finally the *neutral space* E^0 is the direct sum of those root spaces with $\Re(\lambda) = 0$.

Sinks and sources. The linear system, or better its equilibrium point 0, is calld a *sink* if all the eigenvalues have negative real part, i.e. $\Re(\lambda) < 0$, so that that $\mathbb{R}^n = E^-$. It is called a *source* if all the eigenvalues have positive real part, so that $\mathbb{R}^n = E^+$. It is clear that reversing the arrow of time transfoms a sink to a source, and vice-versa, since $(e^{tL})^{-1} = e^{-tL}$.

Theorem 4.7. *The linear system $\dot{x} = L(x)$ is a sink iff it satisfies one of the following equivalent conditions:*

i) all the eigenvalues of L have negative real part,

ii) all solutions decay $e^{tL}v \to 0$ when $t \to \infty$,

iii) there exist a positive $\alpha > 0$ and a constant C such that for all $v \in \mathbb{R}^n$

$$\left\|e^{tL}v\right\| \leq C\,e^{-\alpha t}\,\|v\| \qquad \text{for times } t \geq 0\,. \tag{4.11}$$

Proof. It is obvious that iii) ⇒ ii). It is also clear that ii) ⇒ i), because if some eigenvalue has $\Re(\lambda) \geq 0$, then one easily find, in the corresponding Jordan chain, a solution which does not decay to zero. Finally, to see that i) ⇒ iii), we note that this holds in each Jordan block according to theorem 4.4. But if we have norms in each subspace of a direct sum decomposition (as for example the restrictions of the Euclidean norm), we can define a norm on the total by space taking their maximum (or their sum, or the square root of the sum of theirs squares). With respect to this norm, we then have the inequality (4.11) for some $\alpha > 0$ strictly smaller than all the $|\Re(\lambda)|$'s and some maximal constant C. Again, by the equivalence of all norms, the same inequaity holds w.r.t. to the any norm in $\mathbb{R}^n$, for some possibly different constant C. □

Thus, all trajectories of a sink decay exponentially fast to the origin. Conversely, all trajectories of a source are exponentially stretched, i.e. satisfy an inequality like

$$\left\|e^{tL}v\right\| \geq C\, e^{\beta t}\, \|v\|$$

for some $\beta > 0$ and all $t \geq 0$, and therefore diverge exponentially fast as $t \to \infty$, provided the initial condition is not the equilibrium, i.e. $v \neq 0$. If the linear field has non-real eigenvalues, trajectories may decay or diverge along logarithmic spirals.

Hyperbolic linear flows. A linear vector field L is called *hyperbolic* if the spectrum of its complexification is disjoint from the imaginary axis, i.e. if all the eigenvalues λ, real or complex, have non-zero real part $\Re(\lambda) \neq 0$. The total space of a hyperbolic vector field therefore splits as a direct sum

$$\mathbb{R}^n = E^- \oplus E^+$$

of a stable and an unstable invariant subspace.

Of course, sinks and sources are hyperbolic, but the most interesting case is when both the stable and the unstable subspaces are not-empty. Reasoning as in the proof of theorem 4.7, one shows that

Theorem 4.8. *Let L be a hyperbolic linear field. The phase space is a direct sum of two invariant subspaces $\mathbb{R}^n = E^- \oplus E^+$, the stable and the unsable subspaces, and there exist positive constants $\alpha, \beta > 0$ and a constant C such that*

$$\|e^{tL}v\| \leq C\, e^{-\alpha t}\, \|v\| \qquad \text{if } v \in E^- \text{ and } t \geq 0$$

and

$$\|e^{-tL}v\| \leq C\, e^{-\beta t}\, \|v\| \qquad \text{if } v \in E^+ \text{ and } t \geq 0$$

Thus, the flow of a hyperbolic linear vector field contracts vectors in the stable space and stretch vectors in the unstable space. Indeed, the stable and the unstable subspaces $E^{\pm}$ may be characterized/defined as the sets of those vectors satisfying $e^{\pm tL}v \to 0$ for $t \to \infty$, respectively. If both spaces are not empty, generic trajectories, not starting in $E^- \cup E^+$, diverge for $t \to \pm\infty$.

It turns out that the hyperbolic vector fields are precisely the structurally stable linear vector fields. This is the starting point of a large area of the modern theory of dynamical systems, called *hyperbolic theory*. Classical references are [HS74, PM78].

5 Numbers and dynamics

Another important source of interesting dynamics is, quite surprisingly, elementary number theory.

5.1 Decimal expansion and multiplication by ten

Decimal expansion. When children we learn to represent numbers as decimals, like

$$3.14159265358979323846264338327950288419716939937510\ldots$$

Of course, there is nothing special with the number 10, it is but the number of fingers in our hands. Any other integer $d \geq 2$ would work. Representing a non-negative (for simplicity) real number $x \in \mathbb{R}_+$ in base 10 means writing x as the sum of a convergent series

$$\begin{aligned} x &= \text{“}X_m \ldots X_2 X_1 X_0 . x_1 x_2 x_3 \ldots\text{”} \\ &:= X_m \cdot 10^m + \cdots + X_2 \cdot 10^2 + X_1 \cdot 10 + X_0 + \frac{x_1}{10} + \frac{x_2}{10^2} + \frac{x_3}{10^3} + \ldots \\ &= \sum_{n=0}^{m} X_n \cdot 10^n + \sum_{n=1}^{\infty} x_n \cdot 10^{-n} \end{aligned}$$

where $X_n, x_n \in \{0, 1, 2, \ldots, 9\}$ and $m \geq 0$ (the series above is absolutely convergent because it is bounded by 9 times the geometric series $\sum_{n=1}^{\infty}(1/10)^n$).

The finite sum

$$[x] := \sum_{n=0}^{m} X_n \cdot 10^n \in \mathbb{Z}$$

is the *integral part* of x, the largest of those integers n such that $n \leq x$. The possibly infinite sum

$$\{x\} := 0.x_1 x_2 x_3 \cdots = \sum_{n=1}^{\infty} x_n \cdot 10^{-n} \in [0, 1)$$

is the *fractional part* of x, the difference $\{x\} = x - [x]$. Consequently, $[x] + \{x\} = x$.

Some representations terminate, i.e. have $x_n = 0$ starting from some $n \geq N$, and some others are *recurring* (or *eventually periodic*), i.e. of the form

$$[x] + 0.x_1 x_2 \ldots x_k \overline{a_1 a_2 \ldots a_n} := [x] + 0.x_1 x_2 \ldots x_k a_1 a_2 \ldots a_n a_1 a_2 \ldots a_n a_1 a_2 \ldots a_n \ldots$$

for some finite recurring word $a_1 a_2 \ldots a_n$ (and of course a terminating decimal is a recurring one with recurring word $\overline{0}$).

The representation is unique, hence defines a bijection between $\mathbb{R}$ and the space of infinite words $X_m \ldots X_2 x_1 X_0 . x_1 x_2 x_3 \ldots$ as above, if we do not admit recurrent 9's, i.e. if we substitute $\ldots x_{k-1}\overline{9}$ with $\ldots (x_{k-1}+1)\overline{0}$ (where we assume $x_{k-1} \neq 9$).

Division algorithm. The iterative scheme to obtain the decimal representation of a rational number is the “division algorithm” that we also learn when children. Consider a positive rational number $x = p/q$ with $p, q \in \mathbb{N}$:

$$\frac{p}{q} = x_0 . x_1 x_2 x_3 \ldots .$$

The integer $[x] = x_0$ is “the number of times q is contained in p”, i.e. the unique integer such that

$$p = x_0 \cdot q + r_0$$

for some rest r_0 which is an integer $0 \leq r_0 < q$. Hence, $p/q = x_0 + r_0/q$ and $0 \leq r_0/q < 1$. The “geometric” meaning of x_1 is that the point r_0/q lies between $0.x_1$ and $0.x_1 + 0.1$. Multiplying by 10 and then by q this means that

$$x_1 \cdot q \leq 10 \cdot r_0 < x_1 \cdot q + q$$

or, equivalently, that x_1 is the unique integer between 0 and 9 such that

$$10 \cdot r_0 = x_1 \cdot q + r_1$$

where, again, the rest r_1 is a non-negative integer $0 \leq r_1 < q$. And so on. Hence, the digits of the decimal expansion of p/q are iteratively determined by

$$10 \cdot r_{n-1} = x_n \cdot q + r_n \qquad \text{where} \qquad 0 \leq r_n < q$$

Since the possibilities for the rests are finite, they necessarily recurr. On the other side, a simple computation shows that a recurring decimal is a (series converging to a) rational number.

Theorem 5.1. *The rational numbers are precisely those real numbers whose representation in base* 10 *(or any other base $d \geq 2$) is (eventually) repeating/recurring.*

Meanwhile, there exist irrational numbers. For example,

$$0.101001000100001\cdots = \frac{1}{10} + \frac{1}{10^3} + \frac{1}{10^6} + \frac{1}{10^{10}} + \frac{1}{10^{15}} + \dots$$

is irrational, since it is not recurring.

Indeed, almost all numbers are irrational, in a precise probabilistic sense, since rationals are countable.

The weight of the rationals. Consider the unit interval $I = [0,1]$, and and imagine to cut out all its rational points. What is left is a set, $I \backslash \mathbb{Q}$, whose lenght is equal to the lenght of the interval! Indeed, the rationals are countable, for example those inside I may be ordered according to

$$0 \qquad 1 \qquad 1/2 \qquad 1/3 \qquad 2/3 \qquad 1/4 \qquad 3/4 \qquad 1/5 \qquad 2/5 \qquad 3/5 \qquad 4/5 \qquad \dots$$

say $I \cap \mathbb{Q} = \{r_1, r_2, r_3, \dots\}$. Given an (arbitrarily small) $\varepsilon > 0$, we may even cut out a whole interval $J_n = (r_n - \ell_n/2, r_n + \ell_n/2)$ of finite diameter $\ell_n = \varepsilon/2^n$ around each r_n. The measure of what is left of the unit interval is

$$\text{lenght}\,(I \backslash (\cup_n J_n)) \geq 1 - \sum_n \varepsilon/2^n = 1 - \varepsilon$$

In other words, the rationals inside the unit interval have neighborhoods of arbitrarily small lenght! Mathematicians say that

Theorem 5.2. *Rationals form a set of Lebesgue measure zero inside the real line.*

Therefore, almost all numbers are irrationals. In other words, if we "choose" a random number in the interval $[0,1]$, with respect to the uniform distribution giving probability $|b-a|$ to any interval $[a,b] \subset [0,1]$, it will be irrational "with probability one".

ex: Show that the decimal representation of a (reduced) rational p/q terminates iff the denominator is of the form $q = 2^\alpha 5^\beta$ for some non-negative integers α and β.

ex: Write $1/3$ in base 2, and $2/3$ in base 3 and 7.

ex: Show that the decimal (or any other base) representation of a rational number is repeating (observe that the possibilities for the rests r_n are finite). Then show the converse: a repeating decimal represents a rational number (compute the sum of the series).

ex: Give examples of non-repeating decimal expansions (see [HW59], section 9.4).

ex: Prove that *Euler's number*

$$e := \sum_{n=0}^{\infty} \frac{1}{n!}$$

is irrational (Fourier's idea: assume that $e = p/q$ for some positive integers p and q, and deduce that $x = q!\,(e - \sum_{n=0}^{q} 1/n!)$ is then an integer. Estimate the series $x = \sum_{n=q+1}^{\infty} q!/n!$ and prove that $0 < x < 1$).

Multiplication by an integer. The representation of real number in base 10 is strictly related to the dynamics of a particular transformation acting on the circle. Let $d \geq 2$ be an integer.

The transformation $F : \mathbb{R} \to \mathbb{R}$ sending each number x to its multiple $d{\cdot}x$ has a trivial dynamics, since all trajectories diverge, apart from the fixed point 0. Things get interesting if we do not allow trajectories to escape, i.e. if we force them to a bounded domain. One way to do it is considering the quotient $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, and define the transformation $f : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ as

$$x + \mathbb{Z} \mapsto d \cdot x + \mathbb{Z}$$

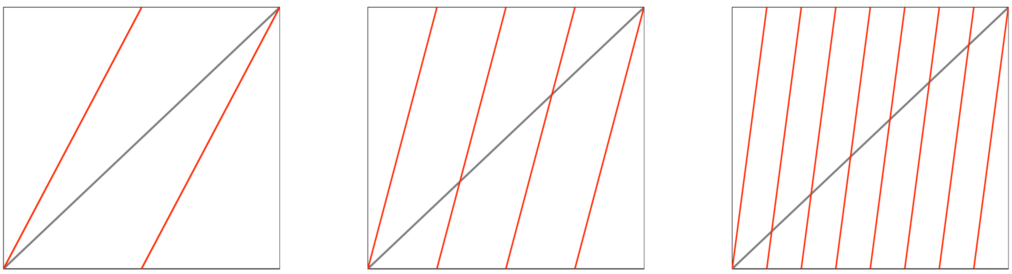
If $\pi : \mathbb{R} \to \mathbb{R}/\mathbb{Z}$ denotes the projection of the real line over the circle, then $\pi \circ F = f \circ \pi$, i.e., F is a lift of f.

Alternatively, we could have defined a transformation of the unit interval $[0, 1]$ sending $x \mapsto \{d \cdot x\} = d \cdot x - [d \cdot x]$, thus avoiding the identification $0 \sim 1$.

If $x = 0.x_1x_2x_3\ldots$ is the representation of $x \in \mathbb{R}/\mathbb{Z} \simeq [0, 1)$ in base d, then f sends

$$0.x_1x_2x_3\ldots \mapsto 0.x_2x_3x_4\ldots$$

The simplest case is that of the *doubling map*, $f(x + \mathbb{Z}) = 2x + \mathbb{Z}$.



Graph of the doubling map, and its first two iterates.

ex: Find the cardinality of the inverse image by f of a generic point in $\mathbb{R}/\mathbb{Z}$.

ex: Find periodic and pre-periodic points of f, show that they are dense in the circle.

ex: Show that the identification $h : \mathbb{R}/\mathbb{Z} \to \mathbb{S}^1$, given by

$$x + \mathbb{Z} \mapsto e^{2\pi i x}\,,$$

is a topologcal conjugation between the doubling map and the restriction of the squaring map $z \mapsto z^2$ to the unit circle $\mathbb{S} \subset \mathbb{C}$. State the corresponding result for the multiplication by an arbitrary integer $d \geq 2$.

5.2 Bernoulli shifts

The abstract version of multiplication by 10 on the circle is the shift on the space of Bernoulli trials, a map which is basic in probability.

Infinite words. Let $\mathcal{A} = \{1, 2, ..., z\}$ be an "alphabet" made of $z \geq 2$ words, i.e. a finite set equipped with the discrete topology, and $\Sigma^+ := \mathcal{A}^{\mathbb{N}}$ be the topological product of infinite copies of $\mathcal{A}$. Points of Σ^+ are actually sequences $(x_n)_{n\in\mathbb{N}}$ with values in $\mathcal{A}$, but are more conveniently denoted as

$$x = x_1 x_2 x_3 \dots x_n \dots$$

with $x_n \in \mathcal{A}$, and interpreted as "infinite words" in the letters of the alphabet $\mathcal{A}$. In probability theory, the x_k's represents the outcomes of a sequence of trials of some experience with z possible outcomes (a dice with z faces).

The *product topology* is the weakest topology on $\mathcal{A}^{\mathbb{N}}$ such that all the projections $\pi_n : \Sigma^+ \to \mathcal{A}$, sending $x \mapsto x_n$, are continuous. A basis for this topology is the family $\mathcal{C}$ of centered cylinders. A *centered cylinder* is a subset

$$C_\alpha := \left\{x \in \Sigma^+ \text{ s.t. } x_1 = \alpha_1, x_2 = \alpha_2, \dots, x_k = \alpha_k\right\}$$

where $\alpha = \alpha_1\alpha_2 \dots \alpha_k \in \mathcal{A}^k$ is a finite word of lenght $k \in \mathbb{N}$. More colloquially, C_α is the set of those infinite words x starting with the finite word α, i.e. of the form $x = \alpha*$, with an obvious meaning of the symbol "$*$ ". Thus, a basis $\mathcal{C}$ of the product topology is the countable family of C_α, when α ranges in the set $\bigcup_k \mathcal{A}^k$ of all finite words in the letters of $\mathcal{A}$. By definition, an open set of the topological product Σ^+ is a union $A = \cup_\alpha C_\alpha$ of centered cylinders.

Observe that the family of centered cylinders is a basis of a topology because it is covering, since obviously $\Sigma^+ = C_1 \cup C_2 \cup ... \cup C_z$, and because the intersection of two cylinders is the empty set or one of the two cylinders. Indeed, two cylinders C_α and C_β have non-empty intersection iff one of the two words, say $\alpha = \alpha_1\alpha_2 \dots \alpha_k$, is the initial string of the other word, in the sense that $\beta = \alpha_1\alpha_2 \dots \alpha_k\beta_{k+1} \dots \beta_{k+i}$, and in this case $C_\alpha \cap C_\beta = C_\beta$. The idea is that the longer is the word α the smaller is the cylinder C_α.

Ultrametrics. The product topology is metrizable. This means that there exist metrics on Σ^+ which induce the product topology. One possibility is the metric

$$d_\lambda(x, y) = \sum_{n=1}^{\infty} \lambda^{-n} \cdot |x_n - y_n|$$

for some $\lambda > 1$ (for example $\lambda = z$). Another possibility, simpler to deal with, is to define $\mathrm{ord}(x, y) := \min\{n \in \mathbb{N} : x_n \neq y_n\}$, the smallest place where the two words x and y differ, and then a distance as

$$d_\infty(x, y) = z^{-\mathrm{ord}(x,y)}$$

if $x \neq y$, and zero otherwise. It is clear that centered cylinders C_α are both closed and open balls for this metric, as strange as it may seem. It turns out that this is indeed an *ultrametric*, triangular inequality being a consequence of the stronger ultrametric inequality

$$d_\infty(x, y) \leq \max\{d_\infty(x, z), d_\infty(z, y)\}$$

Between the paradoxical properties of ultrametric spaces, one verifyies that any point of a ball is its center. The space Σ^+ is the abstract example of a Cantor set: a compact, perfect and totally disconnected metric space.

ex: Show that d_∞ is an ultrametric.

ex: Show that any point of a ball in a ultrametric space is its center, and that balls are both open and closed.

Bernoulli shift. The *Bernoulli shift* is the transformation $\sigma : \Sigma^+ \to \Sigma^+$ which "forgets the first letter" of the infinite word, sending

$$x_1x_2x_3 \dots \mapsto \sigma(x_1x_2x_3) := x_2x_3x_4 \dots$$

It is continuous, because the inverse image of any centered cylinder is a union of centered cylinders, hence an open set. It is not invertible, and indeed the inverse image of any point is made of z different points (the choices for the first letter of the infinite word).

In probability, letters of the alphabet $\mathcal{A}$ repesents the possible outcomes of an experience, as tossing a coin or drowing a dice. An infinite word $x_1x_2x_3\dots x_n\dots$ therefore represent the sucessive results of a countable set of experiences, ordered, for example, by time n. Iteration of σ means forgetting the oucomes of the first experiences.

ex: Describe periodic and pre-preriodic points of σ. Show that they are dense in Σ^+.

ex: Consider the alphabet $\mathcal{A} = \{0, 1, 2, \dots, 9\}$. Define a map $h : \mathcal{A}^{\mathbb{N}} \to [0, 1]$ as

$$x_1x_2x_3\cdots \mapsto 0.x_1x_2x_3\dots .$$

Show that h is a semi-conjugation between the shift σ andt the multiplication by 10 on the circle.

5.3 Rotations of the torus

Rotations of the circle. The *circle*, or *one-dimensional torus*, is the quotient $\mathbb{T} := \mathbb{R}/\mathbb{Z}$ of the commutative group $\mathbb{R}$ modulo its subgroup $\mathbb{Z}$, equipped with the quotient topology. We denote by $\pi : \mathbb{R} \to \mathbb{T}$ the projection $x \mapsto \pi(x) := x + \mathbb{Z}$. The euclidean metric on the real line induces a metric on the circle, defined by

$$\begin{aligned} d(x+\mathbb{Z}, y+\mathbb{Z}) &= \min_{x'\in x+\mathbb{Z}, y'\in y+\mathbb{Z}} |x'-y'| \\ &= \min_{n\in\mathbb{Z}} |x-y+n| \end{aligned}$$

Thus, the distances between the classes of x and y is the minimal distance between the subsets $x+\mathbb{Z}$ and $y+\mathbb{Z}$ of the real line. Observe that the diameter of the circle, i.e. the maximal distance between two points, is $1/2$.

Rotations of the circle are the transformations $R_\alpha : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ defined by

$$x + \mathbb{Z} \mapsto x + \alpha + \mathbb{Z}$$

where $\alpha \in \mathbb{R}$. Observe that $\mathbb{R}/\mathbb{Z}$ is a commutative group, and the R_α are its translations. Also, rotations are the isometries of the circle which preserve the orientation. If we identify the circle with the unit circle $\mathbb{S}^1 := \{z \in \mathbb{C} \text{ t.q. } |z| = 1\} \subset \mathbb{C}$ in the complex plane, by means of the homeomorphism $x + \mathbb{Z} \mapsto e^{2\pi i x}$, rotations are the transformations $z \mapsto e^{i2\pi\alpha}z$.

It is interesting to observe that trajectories of a circle rotations are the sucessive points where a billard ball hits the boundary circle if thrown inside a circular billard.

Theorem 5.3. *A rotation R_α has periodic points iff α is rational.*

Proof. If α is rational, and equal to the reduced fraction p/q, then all points are periodic with period q, since $x + q\alpha + \mathbb{Z} = x + \mathbb{Z}$ for any x. On the other side, if α is irrational, there exists no natural $n \geq 1$ such that $x + \mathbb{Z} = x + n\alpha + \mathbb{Z}$, independently on x. □

Indeed, what we showed is that all orbits of a rational rotation are periodic, hence finite. On the other side, all orbits of an irrational rotation are infinite, and this will be the interesting case.

ex: Observe that any point $x+\mathbb{Z}$ of the circle $\mathbb{R}/\mathbb{Z}$ has a unique representative $\{x\}$ in the interval $[0, 1)$, the fractional part of x, and show that the distance between two points of the circle is given by the explicit formula

$$d(x + \mathbb{Z}, y + \mathbb{Z}) = \min\{|\{x\} - \{y\}| \,, 1 - |\{x\} - \{y\}|\}$$

Rotations of a torus. The n-dimensional *torus* is the quotient $\mathbb{T}^n := \mathbb{R}^n/\mathbb{Z}^n$, equipped with the quotient metric. *Rotations* are the homeomorphisms $R_\alpha : \mathbb{R}^n/\mathbb{Z}^n \to \mathbb{R}^n/\mathbb{Z}^n$ defined by

$$x + \mathbb{Z}^n \mapsto x + \alpha + \mathbb{Z}^n$$

where now $\alpha \in \mathbb{R}^n$.

ex: Try to understand possible orbits of a torus rotation.

5.4 Dyadic adding machine

p-adic number fields and integers. The field $\mathbb{R}$ of real numbers may be considered (actually constructed) as the completion of the rational number field $\mathbb{Q}$ with respect to the Euclidean norm $|x|_\infty := \max\{\pm x\}$. This means that real numbers are equivalence classes of fundamental sequences of rationals, two fundamental sequences (x_n) and (y_n) being in the the same class, i.e. representing the same real '$x := \lim_{n\to\infty} x_n$", if $|x_n - y_n|_\infty \to 0$.

It happens that there exist other "norms" (positive and homogeneous functionals $x \mapsto |x| \in \mathbb{Q}$ satisfying the triangular inequality) on $\mathbb{Q}$ which respect the multiplicative structure, i.e. such that $|xy| = |x|\,|y|$. Such norms are called *valuations*.

Let $p = 2, 3, 5, 7, \ldots$ be a rational prime, also called *place* in this context. The *order* of a non-zero rational $x \in \mathbb{Q}\backslash\{0\}$ at the place p is the unique integer $\mathrm{ord}_p(x) = n$ such that $x = p^n a/b$ for some $a, b \in \mathbb{Z}$ which are not divided by p. The *p-adic valuation/place* is the absolute value on $\mathbb{Q}$ defined as

$$|x|_p := p^{-\mathrm{ord}(x)} \qquad \text{if } x \neq 0$$

and $|0|_p = 0$ otherwise. Clearly $\mathrm{ord}_p(xy) = \mathrm{ord}_p(x) + \mathrm{ord}_p(y)$ (just like the degree of polynomials), and this gives homogeneity of $\|\cdot\|_p$. Triangular inequality follows from the observation that

$$\mathrm{ord}_p(x + y) \geq \min\{\mathrm{ord}_p(x), \mathrm{ord}_p(y)\}$$

One can show that those, together with the euclidean norm, are the only valuations on $\mathbb{Q}$, modulo trivial equivalences. The *p-adic (topological) number field* $\mathbb{Q}_p$ is the completion of $\mathbb{Q}$ with respect to $|\cdot|_p$ (uniqueness is trivial, and existence may be proved as usual considering equivalence classes of fundamental sequences, the only annoying issue being keeping track of the field operations). The p-adic valuation, naturally extended to $\mathbb{Q}_p$, is "non-Archimedean" (i.e. does not satisfy the "Archimedean property" that for all $\varepsilon > 0$ and all N there exists an integer n such that $n\varepsilon > N$) since triangular inequality is enhanced by the stronger "ultra-metric" inequality

$$|x + y|_p \leq \max\{|x|_p, |y|_p\}\,.$$

This causes many paradoxical properties. For example, closed balls are open as well (hence called "clopen"), and any point of a ball is its center. The ultrametric inequality also implies that

$$|b_1 + b_2 + \ldots b_n|_p \leq \max_{1\leq k\leq n} |b_k|_p\,.$$

Consequently, a series $\sum_{n=0}^{\infty} b_n$ converges (for the p-adic metric, of course!) iff the norm of its terms $|b_n|_p \to 0$ as $n \to \infty$ (there is no room for divergent series like the harmonic series in the p-adic world!).

The ring of *p-adic integers* $\mathbb{Z}_p$ is the closure of $\mathbb{Z}$ in $\mathbb{Q}_p$. One can describe the p-adic integers as the inductive limit $\mathbb{Z}_p = \lim\limits_{\leftarrow} (\mathbb{Z}/p^n\mathbb{Z})$, and represent a p-adic integer as a series

$$z = \ldots z_n \ldots z_2 z_1 z_0 := \sum_{n=0}^{\infty} z_n p^n \tag{5.1}$$

with $z_n \in \mathcal{A}_p := \{0, 1, 2, \ldots, p-1\} \approx \mathbb{Z}/p\mathbb{Z}$, which converges in $\mathbb{Q}_p$ because the norm of the generic term $z_n p^n$ is bounded by $|z_n p^n|_p = p^{-n} \to 0$ as $n \to \infty$. Thus, as a topological space (not as a ring!), $\mathbb{Z}_p$ is isomorphic to the topological poduct $\Sigma^+ = \mathcal{A}^{\mathbb{N}}$, the space of infinite words (written backwards!) $(\ldots z_n \ldots z_3 z_2 z_1)$ in the letters of the alphabeth $\mathcal{A}_p$. Observe that $\mathbb{Z}_p = \{x \in \mathbb{Q}_p \text{ s.t. } |x|_p \leq 1\}$, i.e. the ring of p-adic integers is the clopen ball of radius one around 0 in $\mathbb{Q}_p$.

Any p-adic number $x \in \mathbb{Q}_p$ can be represented uniquely as $x = z + r$ where $z = [x]_p = \sum_{n\geq 0} x_n p^n \in \mathbb{Z}_p$ is the "p-adic integer part" and $r = \{x\}_p = \sum_{n=1}^{N} x_n p^{-n} \in \mathbb{Z}[1/p]$ is the "p-adic fractional part". In symbols,

$$x = \dots x_n \dots x_2 x_1 x_0 x_{-1} \dots x_{-N} = \sum_{n=-N}^{\infty} x_n p^n$$

The quotient $\mathbb{Q}_p/\mathbb{Z}_p = \mathbb{Z}[1/p]/\mathbb{Z}$ is a discrete (additive) group where the norm $|\cdot|_p$ takes values p^n with $n \in \mathbb{N}$. Multiplication by p is a uniform contraction $x \mapsto px$ of $\mathbb{Q}_p$ with Lipschitz constant p^{-1}, and its inverse $x \mapsto xp^{-1}$ uniformly expands distances by a factor p. Thus, $\mathbb{Z}_p$ is the disjoint union $\cup_{a_i=0}^{p-1}(a_i + p\mathbb{Z}_p)$ of p clopen balls of radius p^{-1} (and so on, iterating the contraction). Also, one can represent the field of p-adic numbers as a union $\mathbb{Q}_p = \bigcup_{n\in\mathbb{N}} p^{-n}\mathbb{Z}_p$.

ex: Compute the following sums in $\mathbb{Z}_2$.

$$\dots 011 + \dots 001 \qquad \dots 0101 + \dots 1010 \qquad \dots 111 + \dots 001$$

Find the additive opposite of $1 = \dots 001$ in $\mathbb{Z}_2$.

Adding machine. Consider the ring $\mathbb{Z}_2$ of dyadic integers, thought as an additive topological group. The *dyadic adding machine* (or *"Kakutani-von Neumann odometer"*) is the translation $f : \mathbb{Z}_2 \to \mathbb{Z}_2$, defined by

$$z \mapsto z + 1$$

Observe that f changes the first (starting from the right) digit of $x = \dots x_2 x_1 x_0$, and consequently its n-th iterate changes the n-th digit of x. Thus, we have one more example of a translation in a compact topological group without periodic points.

ex: Show that $z \mapsto z + 1$ is a homeomorphism of $\mathbb{Z}_2$, and find its inverse.

5.5 Continued fractions and Gauss map

Continued fractions. Any real number $x \in \mathbb{R}$ can be represented (uniquely for irrationals, and with only a minor ambiguity for rationals) as a *continued fraction*

$$x \sim [a_0; a_1, a_2, a_3, \dots] := a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{\ddots}}}}$$

with $a_0 \in \mathbb{Z}$ and "partial quotients" $a_n \in \mathbb{N}$ if $n \geq 1$. This means that x is equal to the limit of the *convergents*, the finite continued fractions defined as

$$p_n/q_n = [a_0; a_1, a_2, \dots, a_n] := a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{a_n}}}}$$

as $n \to \infty$. Observe that finite continued fractions are rationals, and obey the recursion

$$[a_0; a_1, a_2, \dots, a_n, a_{n+1}] = [a_0; a_1, a_2, \dots, a_n + 1/a_{n+1}] \tag{5.2}$$

Continued fractions constitute the fundamental tool to investigate rational approximations to real numbers, because they provide base-free, hence intrinsic, rational approximations. Thus, while Earthlings with ten fingers write $\pi = 3.1415\dots$ and Martians with three fingers write $\pi = 10.0102\dots$, they all agree to write $\pi = [3; 7, 15, 1, 292, \dots]$). Moreover, they provide the best rational approximations, in a certain precise sense [HW59, Kh35].

Construction and Gauss map. The continued fraction converging to a given number $x \in \mathbb{R}$ is given essentially by Euclid's algorithm to find the m.c.d. of two integers. One starts with $a_0 = \lfloor x \rfloor \in \mathbb{Z}$ (here the "floor" function $\lfloor x \rfloor$ returns the smallest integer n such that $n \leq x < n+1$), and write $x = a_0 + x_0$ for some $x_0 = \{x\} \in [0,1)$. Then define the *Gauss map* $G : (0,1] \to [0,1]$ as

$$G(x) := 1/x - \lfloor 1/x \rfloor \,, \tag{5.3}$$

(thus, $G(x)$ is the fractional part of the inverse of x) and inductively define the partial quotients $a_n \in \mathbb{N}$ and the "rests" $x_n \in [0,1)$ as

$$a_{n+1} = \lfloor 1/x_n \rfloor \qquad x_{n+1} = G(x_n)\,,$$

provided all the $x_0, x_1, \ldots, x_n \neq 0$. Then,

$$x = a_0 + x_0 = a_0 + \cfrac{1}{a_1 + x_1} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + x_2}} = \cdots = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{a_n + x_n}}}}$$

If some $x_n = 0$, the iteration stops and x is equal to a finite continued fraction as above. Conversely, if $x = p/q$ is rational, all the x_n's are positive rationals, and have strictly decreasing denominators (for if $x_n = a/b$, then $1/x_{n+1} = x_n - a_n = (a - a_n b)/b = c/b$, and $c < b$ because $x_n - a_n < 1$). So, there must be some first x_n which is an integer, and the algorithm stops. Thus, finite continued fractions correspond to rationals (and are unique if we demand the the last non-zero partial quotient be $a_n > 1$).

Convergence of the convergents. We must therefore understand the case of infinite continued fractions, which, as we already know, correspond to irrationals. The key observation is that the convergents $p_n/q_n = [a_0; a_1, a_2, \ldots, a_n]$ of a continued fraction $[a_0, a_1, a_2, a_3, \ldots]$ are determined by the partial quotients a_n's according to the following recursive equation.

Theorem 5.4. *The convergents $p_n/q_n = [a_0; a_1, a_2, \ldots, a_n]$ are obtained from the coefficients a_k's by the recursions*

$$\begin{aligned} p_n &= a_n p_{n-1} + p_{n-2} \\ q_n &= a_n q_{n-1} + q_{n-2} \end{aligned} \tag{5.4}$$

given the initial conditions $p_0 = a_0$, $q_0 = 1$, and $p_{-1} = 1$, $q_{-1} = 0$ (or $p_{-2} = 0$ and $q_{-2} = 1$) .

Proof. The proof is by induction. The first two values are easily verified. Assume the results holds until n, and compute

$$\begin{aligned} \frac{p_{n+1}}{q_{n+1}} &= [a_0; a_1, a_2, \ldots, a_n, a_{n+1}] \\ &= [a_0; a_1, a_2, \ldots, a_n + 1/a_{n+1}] \\ &= \frac{(a_n + 1/a_{n+1})p_{n-1} + p_{n-2}}{(a_n + 1/a_{n+1})q_{n-1} + q_{n-2}} \\ &= \frac{a_{n+1}(a_n p_{n-1} + p_{n-2}) + p_{n-1}}{a_{n+1}(a_n q_{n-1} + q_{n-2}) + q_{n-1}} \\ &= \frac{a_{n+1}p_n + p_{n-1}}{a_{n+1}q_n + q_{n-1}} \end{aligned}$$

□

It is important to write the recursion (5.4) in matrix notation as

$$\begin{pmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{pmatrix} = \begin{pmatrix} p_{n-1} & p_{n-2} \\ q_{n-1} & q_{n-2} \end{pmatrix} \begin{pmatrix} a_n & 1 \\ 1 & 0 \end{pmatrix} ,$$

whose solution, taking care of the initial conditions, is the backward product

$$\begin{pmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{pmatrix} = \begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix} \dots \begin{pmatrix} a_n & 1 \\ 1 & 0 \end{pmatrix} \tag{5.5}$$

of $n+1$ integer matrices with determinant -1. In particular,

$$p_n q_{n-1} - p_{n-1} q_n = (-1)^{n+1} ,$$

which says that the matrix with columns p_n, q_n and p_{n-1}, q_{n-1} is unimodular, i.e. belongs to the group $\mathrm{GL}_2(\mathbb{Z})$ of (invertible) integer matrices with determninant ± 1 (two by two matrices whose rows and columns are relatively prime integers, a group which contains much arithmetical information!). This shows that the fractions p_n/q_n obtained using the recurrence in theorem 5.4 are reduced.

There also follows that

$$\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n} = \frac{(-1)^n}{q_{n+1} q_n}$$

Also, one can easily show that convergents with n even form an increasing sequence, and convergents with n odd form a decreasing sequence. Another consequence of the recursion (5.4) is that the denominators q_n of an infinite continued fraction (i.e. such that $a_n \geq 1$ for all $n \geq 1$) satisfy

$$q_{n+2} \geq q_{n+1} + q_n \geq 2 q_n ,$$

and therefore grow exponentially fast:

$$q_{n+1} \geq 2^{n/2} .$$

(indeed, they grow at least like the Fibonacci sequence starting with $f_0 = 1$ and $f_1 = 1$, hence like $q_n \geq c\phi^n$, where $\phi = (1+\sqrt{2})/2$ is the "ratio" and c is some positive constant). This implies that

$$\left| \frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n} \right| = \frac{1}{q_{n+1} q_n} \leq \frac{2}{2^n}$$

and therefore the sequence of the convergents p_n/q_n is fundamental. Its limit $\lim_{n\to\infty} p_n/q_n = x$ is called "value" of the continued fraction, and denoted by

$$x = [a_0; a_1, a_2, a_3, \dots] := \lim_{n\to\infty} [a_0; a_1, a_2, \dots, a_n] .$$

It is also possible to find a lower bound to the difference between an irrational x and its convergents, and the two-sided estimate reads as follows:

$$\frac{1}{q_n(q_{n+1} + q_n)} < \left| x - \frac{p_n}{q_n} \right| < \frac{1}{q_{n+1} q_n}$$

ex: Use the quadratic equation $\phi^2 - \phi - 1 = 0$ to show that the "ratio" ϕ has the simplest continued fraction, namely

$$\frac{1+\sqrt{5}}{2} = [1; 1, 1, 1, 1, 1, \dots]$$

(observe that $\phi^{-1} = \phi - 1$ is a root of $x^2 + x - 1 = 0$, hence $x = 1/(1+1/x)$, and so on). Its convergents are 1 , 2 , $3/2$, $5/3$, $8/5$, $13/8$, $21/13$, $34/21$, $\dots$, ratios between successive Fibonacci numbers. It is also the (irrational) number with worse rational approximations, namely $|\phi - p/q| > (1/\sqrt{5})/q^2$ for any rational p/q.

ex: Also, the most famous irrational has a simple continued fraction. Show that

$$\sqrt{2} = [1; 2, 2, 2, 2, 2, \dots]$$

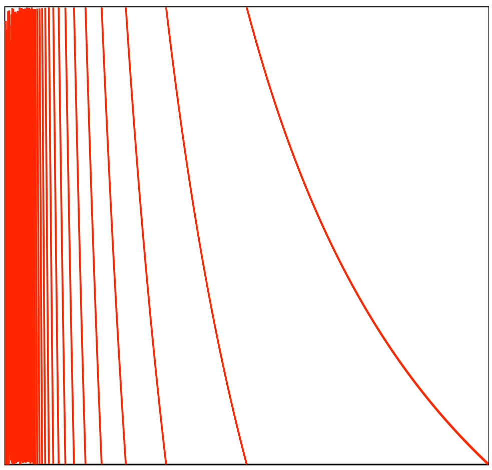
(observe that $1 + \sqrt{2}$ is the positive root of $x^2 - 2x - 1$. Hence $x = 2 + 1/x$, and so on). Its convergents are 1 , $3/2$, $7/5$, $17/12$, $41/29$, $99/70$, $239/169$, $577/408$, ...

Continued fractions and Bernoulli shift. The continued fraction development, the map

$$x \mapsto [a_0; a_1, a_2, \dots]$$

realizes a conjugation between the restriction of the Gauss map (5.3) to the irrationals, the transformation $G : (0,1]\backslash\mathbb{Q} \to (0,1]\backslash\mathbb{Q}$, and the shift $\sigma : \mathbb{N}^{\mathbb{N}} \to \mathbb{N}^{\mathbb{N}}$ over an alphabeth $\mathcal{A} = \mathbb{N}$ of infinite letters. Indeed,

$$G : [0; a_1, a_2, a_3, \dots] \mapsto [0; a_2, a_3, a_4, \dots] .$$



Graph of the Gauss map.

ex: Find the (largest two or three) fixed points of the Gauss map, and compute their values.

Periodic continued fractions and quadratic irrationals. *Quadratic irrationals* (or *quadratic surds*) are irrational roots of quadratic polynomials with integer coefficients

$$f(x) = ax^2 + bx + c$$

(with $a, b, c \in \mathbb{Z}$), i.e. numbers like

$$x = \frac{\alpha + \sqrt{\beta}}{\delta}$$

where $\alpha, \beta, \delta \in \mathbb{Z}$, $\delta \neq 0$ and $\beta > 0$ which is not a square.

Theorem 5.5 (Lagrange). *The continued fraction of an irrational number $x \in \mathbb{R}\backslash\mathbb{Q}$ is periodic iff x is a quadratic irrational.*

See [Kh35, HW59].

5.6 Exponential sums

Arithmetic progressions . The dynamics of an *arithmetic progression*

$$a \qquad a+\alpha \qquad a+2\alpha \qquad a+3\alpha \qquad \dots \qquad a+n\alpha \qquad \dots,$$

obtained from the initial condition $x_0 = a$ using the recursion $x_{n+1} = x_n + \alpha$, is quite trivial. All trajectories $x_n = a + n\alpha$ diverge, provided $\alpha \neq 0$.

Something interesting happens if we compute time averages of the basic character of the real line, the observable $e : \mathbb{R} \to \mathbb{S} \subset \mathbb{C}$ given by

$$e(x) := e^{2\pi i x}.$$

Apart from a constant factor $e^{2\pi i a}$ and the normalization $1/N$, the Birkhoff averages of an arithmetic progression are

$$S_N(\alpha) = \sum_{n=0}^{N-1} e^{2\pi i \alpha n}.$$

ex: Show that the sum of the first n terms of an arithmetic progression $x_k = a + k\alpha$ is

$$\sum_{k=0}^{n-1} x_k = \frac{n}{2}(x_0 + x_{n-1}) = na + \frac{n(n-1)}{2}\alpha$$

Exponential sums. Sums as

$$E(N) = \sum_{n=1}^{N} e^{2\pi i x_k}$$

are called *exponential sums*, and contain "spectral information" about the distribution of the sequence of numbers (x_n) modulo 1. Triangular inequality gives the trivial bound $|E(N)| \leq N$, i.e. $E(N) = \mathcal{O}(N)$. If the different exponentials $e^{2\pi i x_n}$ were "uncorrelated", as successive positions of a random walk in the plane, we should expect $E(N) = \mathcal{O}(\sqrt{N})$. This, of course, does not happen with "deterministic" generic sequences. The best we can hope is some bound as $E(N) = o(N)$ (which, in our case, would mean that the Birkhoff averages $\overline{\varphi}_n \to 0$).

ex: Observe that, for integer $q \geq 1$, the complex number $z = e^{2\pi i/q}$ is a non-trivial q-th root of unity. Hence,

$$1 + z + z^2 + \dots + z^{q-1} = 0.$$

Deduce that if $\alpha = p/q \in \mathbb{Q}$ with $p \in \mathbb{Z}$, then

$$\sum_{n=0}^{q-1} e^{2\pi i (p/q) n} = 1$$

so that the exponential sum $S_n(p/q)$ is periodic, and in particular is $\mathcal{O}(1)$.

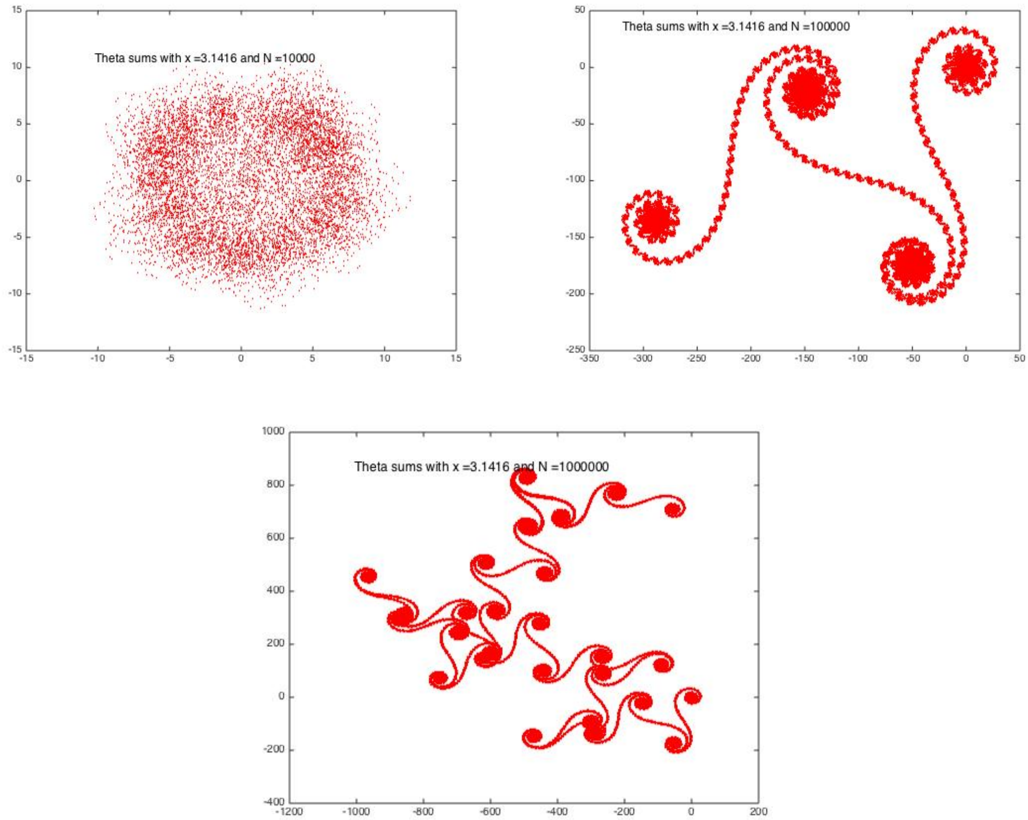
Gauss sums. Much more interesting are exponential sums defined by a "quadratic progression" $x_n = \alpha n^2$. These are

$$G_N(\alpha) = \sum_{n=0}^{N-1} e^{2\pi i \alpha n^2}.$$

When $\alpha = p/q$ is a rational, they are called *(quadratic) Gauss sums*, and they are extremely interesting objects in number theory, as well as in the Fourier analysis on finite fields. These sums are also obviously related to the *Jacobi theta function*, defined for complex $z \in \mathbb{C}$ and $\tau \in \mathbb{H} := \{x + iy \in \mathbb{C} \, : \, y > 0\}$ (the Poincaré upper half-space, a model for the hyperbolic plane) by the series

$$\theta(z, \tau) := \sum_{n=0}^{\infty} e^{\pi i \tau n^2 + 2\pi i z}$$

If you plot the sums for a large number of values of N, given an irrational α or a rational with large denominator, you see "curlicues" as in the following pictures:



Theta sums with $\alpha \simeq \pi$, and $N = 10000$, 100000 and 1000000.

Observe the axis: the sums are of the order of $\sqrt{N}$, as typical trajectories of a random walk!

ex: You may also explore what happens with other exponents, such as $\sqrt{n}$, and get interesting patterns or phenomena.

6 Simple orbits and perturbations

6.1 Topological fixed point theorems

To find periodic points of a transformation $f : X \to X$, namely fixed points of its iterates f^n, may be difficult. For example, when $f(x)$ is a polynomial of degree $d > 1$, its iterates are polynomials of exponentially growing degree.

Fixed point theorems in intervals. In real dimension one, connected and convex sets coincide, and are called intervals. This "miracle" is responsable for two very simple criteria to prove the existence of fixed points of continuous interval transformations. They say that if a compact interval is squeezed or enlarged, at least one of its points remains fixed.

Theorem 6.1 (fixed point theorem for intervals). *Let $f : I \to \mathbb{R}$ be a continuous transformation defined in an interval $I \subset \mathbb{R}$.*
i) If $J \subset I$ is a compact interval such that $f(J) \subset J$, then f has a fixed point in J.
ii) If $J \subset I$ is a compact interval such that $J \subset f(J)$, then f has a fixed point in J.

The proof is an elementary application of Bolzano theorem to the continuous function $f(x)-x$. A more abstract proof, which can be generalized to higher dimension (with the help of some non-trivial algebraic topology), is as follows. Suppose that$f$ has no fixed points in J. Then the function

$$g(x) = \frac{f(x) - x}{|f(x) - x|}$$

(possible since the denominator does not vanish) would define a continuous map of an interval (J itself in case i) or some sub-interval of J in case ii)), which is a connected space, onto the disconnected space $\{-1, 1\}$.

ex: Prove theorem 6.1.

ex: Find examples of continuous functions $f : I \to I$ and non-compact intervals J such that $f(J) \subset J$ or $J \subset f(J)$ which do not contain fixed points of f.

Other topological fixed point theorems. In higher, but finite, dimension, part i) of theorem 6.1 generalizes as

Theorem 6.2 (Brouwer). *A continuous map of the closed unit disk $D \subset \mathbb{R}^n$ into itself has a fixed point.*

The idea is that if a continous map $f : D \to D$ had no fixed point, the same formula as above (associating to each point x of the disk the intersection between the ray passing through x and $f(x)$ and the boundary sphere) would define a continuous map $g : D \to \partial D$ from the disk into the unit sphere. That such a map cannot exist is quite clear intuitively, but needs some non-trivial algebraic topology to rigorously prove it.

In infinite dimension, one has the

Theorem 6.3 (Shauder-Tychonov). *A continuous transformation $f : K \to K$ of a compact and convex subset $K \subset X$ of a Banach space X (or of a locally convex topological vector space) has a fixed point.*

6.2 Dynamics of contractions

Contractions. Let (X, d) be a metric space. A map $f : X \to X$ is called *contraction* (or λ-*contraction* if one wants to keep track of the constant λ) is it is Lipschitz and has Lipschitz constant $\lambda < 1$, i.e. if there exists a $0 \le \lambda < 1$ such that for all $x, x' \in X$

$$d(f(x), f(x')) \le \lambda \cdot d(x, x') \tag{6.1}$$

Clearly, a constant transformation, sending any $x \in X$ into $f(x) = p$, is trivially a contraction. A linear homogeneous transformation $f(z) = \lambda z$ of the complex plane $\mathbb{C}$ or of the real line $\mathbb{R}$ is a contracton provided $|\lambda| < 1$. Observe that a contraction, as any Lipschitz map, is continuous (take $\delta = \varepsilon/\lambda$ in the ε-δ definition).

e.g. **Smooth contractions.** By the mean value theorem, a continuously differentiable transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ (or of a convex subset $X \subset \mathbb{R}^n$) is a contraction provided there exists a positive $\lambda < 1$ such that $|f'(x)| \leq \lambda$ for any $x \in \mathbb{R}^n$.

ex: Show that a contraction of a compact metric space X cannot be invertible, provided the space contains more than one point (compare the diameters X and $f(X)$)

ex: Give non-trivial (i.e. non constant) examples of contractions of

$$[0,1] \qquad [0,1]\times[0,1] \qquad B_r(x) = \{y \in \mathbb{R}^n \text{ t.q. } d(x,y) < r\} \qquad \mathbb{S}^1 = \{z \in \mathbb{C} \text{ t.q. } |z| = 1\}$$

Contraction principle. The dynamics of a contraction is described by the following fundamental theorem, which we state with all details.

Theorem 6.4 (Contraction principle/Banach fixed point theorem)**.** *All trajectories of a contraction $f : X \to X$ of a metric space (X, d) are fundamental sequences, and the distance between any two trajectories tends to zero exponentially fast. If X is complete, then f admits one and only one fixed point p. The trajectory of any initial point $x_0 \in X$ converges exponentially fast to the fixed point, i.e.*

$$d(f^n(x), p) \leq C\,\lambda^n\,,$$

where $C > 0$ is a positive constant and $0 \leq \lambda < 1$ is the Lipschitz constant of f.

Proof. Let $f : X \to X$ be a λ-contraction. Let $x_0 \in X$ be any initial point, and let (x_n) be its trajectory, defined by the recursion $x_{n+1} = f(x_n)$. Iterating (6.1), one sees that

$$d(x_{k+1}, x_k) \leq d(x_1, x_0) \cdot \lambda^k$$

Using k times the triangular inequality and then the convergence of the geometric series of ratio $\lambda < 1$, we get

$$\begin{aligned} d(x_{n+k}, x_n) &\leq \sum_{j=0}^{k-1} d(x_{n+j+1}, x_{n+j}) \leq d(x_1, x_0) \cdot \sum_{j=0}^{k-1} \lambda^{n+j} \\ &\leq d(x_1, x_0) \cdot \lambda^n \cdot \sum_{j=0}^{\infty} \lambda^j \leq \frac{\lambda^n}{1-\lambda} \cdot d(x_1, x_0). \end{aligned}$$

This implies that (x_n) is fundamental, since we can make $\lambda^n \cdot d(x_1, x_0)/(1-\lambda)$ smaller that any $\varepsilon > 0$ choosing a sufficiently large $n = n(\varepsilon)$. Continuity of f implies that the limit $p = \lim_{n\to\infty} x_n$, which exists if X is complete, is a fixed point of f. Uniqueness is clear, for if p and p' were two different fixed points, then by (6.1) their distance $\delta = d(p, p') > 0$ would be $\leq \lambda\delta$, which is impossible if $\lambda < 1$. Again by (6.1) and finite induction, the distance between any two trajectories $x_n = f^n(x_0)$ and $x'_n = f^n(x'_0)$ decay as $d(x_n, x'_n) \leq \lambda^n \cdot d(x_0, y_0)$. In particular, the distance between an arbitrary trajectory and the fixed point p is bounded by $d(x_n, p) \leq \lambda^n \cdot d(x_0, p)$, proving our last assertion with $C = d(x_0, p)$. $\square$

ex: Show that a transformation $f : X \to X$ of a complete metric space (X, d) such that

$$d(f(x), f(x')) < d(x, x')$$

for all distinct $x, x' \in X$ may fail to have fixed points (think at a decreasing sequence forming a divergent series).

ex: Let $a > 0$ and $x_0 > 0$. Show that the sequence (x_n) defined by

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right)$$

converges to $\sqrt{a}$. This is Babylonians-Heron method to approximate square roots (the sequence is a trajectory of the transformation $f(x) = (x + a/x)/2$, which is a contraction once restricted to the closed interval $[\sqrt{a}, \infty) = f(\mathbb{R}_+) \dots)$

Stability of contractions. A contraction $f : X \to X$ of a complete metric space X may be thought as a "machine" that compute the fixed point $p = \lim_{n\to\infty} f^n(x)$ starting with any initial guess $x \in X$.

We pose the question whether contraction are stable, in some sense to be specified. We want to decide whether a small perturbation of a contraction f, say $g : X \to X$, produces a fixed point p' near to p. The point is to decide what "small" means. If we only require something like $d_\infty(f, g) := \sup_{x\in X} d(f(x), g(y)) < \delta$, the transformation g needs not be a contraction, no matter haw small δ is chosen (to see this, try to visualize a δ-neighborhood of the graph of a contraction of an interval, and fit there the graph of a transformation g with arbitrarily large derivative). It is clear that we also needs some control on the derivatives. One possibility is to assume that X has a linear and differentiable structure, e.g. is a subset of some Euclidean $X \subset \mathbb{R}^n$, and look for f and g smooth. The condition

$$\|f - g\|_{\mathcal{C}^1} := \sup_{x\in X} \|f(x) - g(x)\| + \sup_{x\in X} \|f'(x) - g'(x)\| < \delta$$

clearly implies that, if f is a λ-contratraction and $\delta < 1 - \lambda$, then also g is a contraction and has Lipschitz constant $\leq \lambda + \delta$. Simpler, however, is to formulate a stability result inside the class of contractions.

Theorem 6.5. *Let $f : X \to X$ be a λ-contraction of the complete metric space (X, d), and let $p \in X$ be its fixed point. For every $\varepsilon > 0$ there exists some $0 < \delta < 1 - \lambda$ such that if $g : X \to X$ is a $(\lambda + \delta)$-contraction at distance $d_\infty(f, g) < \delta$ from f, and if p' is the fixed point of g, then*

$$d(p, p') < \varepsilon$$

Proof. If p' is the fixed point of g, we know that $g^n(p) \to p'$ when $n \to \infty$. By triangle inequality we see that

$$\begin{aligned} d(p, p') &\leq \sum_{n=0}^{\infty} d\left(g^{n+1}(p), g^n(p)\right) \leq d(g(p), p) \cdot \sum_{n=0}^{\infty} (\lambda + \delta)^n \\ &\leq \delta \cdot \sum_{n=0}^{\infty} (\lambda + \delta)^n = \frac{\delta}{1 - (\lambda + \delta)} \end{aligned}$$

and this quantity is $< \varepsilon$ provided that δ is sufficiently small. $\square$

Equivalence classes of linear contractions. Contraction of the real line also provide a simple example of how to use the dynamics to built a conjugation between two transformations.

Let $f : x \mapsto \alpha x$ and $g : x \mapsto \beta x$ be two linear contractions of $\mathbb{R}$, with $0 < \alpha, \beta < 1$. The origin is the common fixed point. The set $A = [-1, -\alpha) \cup (\alpha, 1]$ is a "fundamental domain" for the action of f on the punctured real line $\mathbb{R}^\times := \mathbb{R}\backslash \{0\}$, in the sense that for any $x \in \mathbb{R}\backslash \{0\}$ there exists a unique time $n(x) \in \mathbb{Z}$ such that $f^{n(x)}(x) \in A$. Similarly, a fundamental domain for the action of g on $\mathbb{R}^\times$ is $B = [-1, -\beta) \cup (\beta, 1]$. Let $H : \overline{A} \to \overline{B}$ be any homeomorphism such that $H(-1) = -1$, $H(-\alpha) = -\beta$, $H(\alpha) = \beta$ and $H(1) = 1$ (for example, an affine homeomorphism). It is easy to check that the recipe

$$h(x) = \begin{cases} 0 & \text{se } x = 0 \\ g^{-n(x)}(H(f^{n(x)}(x))) & \text{if } x \neq 0 \end{cases}$$

defines a homeomorphism $h : \mathbb{R} \to \mathbb{R}$. Since $n(x) = n(f(x)) + 1$ (why?), we see that

$$\begin{aligned} (h \circ f)(x) &= g^{-n(f(x))}\left(H\left(f^{n(f(x))} f(x)\right)\right) = g^{-n(x)+1}\left(H\left(f^{n(x)-1}(f(x))\right)\right) \\ &= g\left(g^{-n(x)}\left(H\left(f^{n(x)}(x)\right)\right)\right) = (g \circ h)(x) \end{aligned}$$

and therefore h is a topological conjugation between f and g.

The case when $-1 < \alpha, \beta < 0$ is analogous. On the other side, it is not difficult to see that the contractions $x \mapsto \alpha x$ and $x \mapsto -\alpha x$, with $\alpha \neq 0$, having opposed orientations, cannot be conjugated. The result is that non-trivial linear contractions of the real line fit into two classes of topological conjugated transformations,

It is important to observe that a conugation h between $f : x \mapsto \alpha x$ and $g : x \mapsto \beta x$ cannot be diffeentiable, unless $\alpha = \beta$. Indeed, if $f = h^{-1} \circ g \circ h$ and if h is differentiable, then the chain rule implies that $f'(0) = g'(0)$, hence that $\alpha = \beta$.

ex: Show that the linear conractions $x \mapsto \alpha x$ and $x \mapsto -\alpha x$, with $\alpha \neq 0$, cannot be conjugated (observe that a conjugation is a homemorphism of the line, in particular monotone).

6.3 Linear maps

Linear systems are the only dynamical systems we can explicitely solve. They serve as models of the local behavior of generic systems near a fixed point.

Linear maps. A linear transformation of the Euclidean vector space $\mathbb{R}^n$ (which we may think equipped with the standard Eucidean structure) is an endomorphism $f : \mathbb{R}^n \to \mathbb{R}^n$, defined, in the canonial basis, by a square matrix $A = (a_{ij}) \in \mathrm{Mat}_{n\times n}(\mathbb{R})$, according to

$$f(x) = Ax$$

Here, we think at $x = (x_1, x_2, \dots, x_n)^\top$ as a column vector, and therefore at Ax as the usual raws-by-columns product between matrices. Thus, the map is defined, in coordinates, by $x_i \mapsto \sum_j a_{ij} x_j$.

A linear change of coordinates $h : x \mapsto y = Ux$, with $U \in GL_n(\mathbb{R})$ an invertible square matrix, defines a linear (and therefore topological) conjugation between f and the linear transformation $g : y \mapsto By$, defined by the matrix $B = UAU^{-1}$. Thus, we are free to change coordinates.

Observe that the origin is a fixed point of f, i.e. $f(0) = 0$, and it is the unique fixed point iff the matrix $A - I$ is invertible, i.e. if 1 is not an eigenvalue of A.

If x is an eigenvector of A with eigenvalue λ, i.e. a non-trivial solution of the homogeneous equation $Ax = \lambda x$, then iterations are simple. Indeed, $f^n(x) = \lambda^n x$, and therefore the asymptotic behaviour of the trajectory of the eigenvector x depends on the absolute value of its eigenvalue. Trajectories converge to the origin when $|\lambda| < 1$, and diverge when $|\lambda| > 1$. In the exceptional cases with $\lambda = \pm 1$, we have a fixed point or a periodic point with period 2.

In order to understand the possible global pictures, we start with the smallest non-trivial case.

Linear maps in the plane. The simplest non-trivial case is that of a linear map endomorphism of the plane $\mathbb{R}^2$, defined, in the canonial basis, by a two-by-two real square matrix A. The qualitiative behavior of trajectories depends on the eigenvalues of A, and on their geometric multiplicity. Remember that the eigenvalues $\lambda_\pm$ are the roots of the characterstic polynomial

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

which is a degree two polynomial with real coefficients, and therefore they are a couple of real numbers $\lambda_\pm = \alpha \pm \beta$, possibly overlapping, or a couple of complex conjugate numbers $\lambda_\pm = \alpha \pm i\omega$.

If the matrix A is diagonalizable (as a real matrix in a real vector space), i.e. admits two eigenvalues $\lambda_\pm$ (possibly equal) and two linearly independent eigenvectors $v_\pm$, then the system is linearly conjugated to a diagonal system

$$f(x,y) = (\lambda_+ x, \lambda_- y)\,.$$

If both eigenvalues have absolute values $|\lambda_\pm| < 1$, then f is a contraction and the orbit of any point converges (exponentially fast) to $f^n(x,y) \to 0$. Trajectories move along curves

$$y = Cx^\alpha\,,$$

for some constant C and some exponent $\alpha = \log|\lambda_-|/\log|\lambda_+| > 0$. The basin of attraction of the origin is the whole plane, and the origin, which is an attractng fixed point, is called a *stable node*, or *sink*.

If both eigenvalues have absolute values $|\lambda_\pm| > 1$, then the inverse f^{-1} is a contraction, all backward trajectories converge to $f^{-n}(x,y) \to 0$ as $n \to \infty$, and all forward tarjectories of points different from the origin diverge, i.e. $|f^n(x,y)| \to \infty$ as $n \to \infty$. The basin of attraction of the origin is $\{0\}$ itself, and the origin is called a *unstable node*, or *source*.

If one of eigenvalue has $|\lambda_-| < 1$ and the other $|\lambda_+| > 1$, what happens is the following: trajectories starting at the "stable line" $E^- = \mathbb{R}v_- \approx \{(0,y)\}$, the eigenspace of the eigenvalue λ_-, converge to the origin, while trajectories starting at the "unstable line" $E^+ = \mathbb{R}v_+ \approx \{(x,0)\}$, the eigenspace of the eigenvalue λ_+, diverge. A generic trajectory, starting at a point which does not belong to $E^- \cup E^+$, i.e. (x,y) with both $x \neq 0 \neq y$, also diverge (since the y coordinate decays but the x coordinate explodes), moving along curves

$$y = Cx^\beta\,,$$

for some constant C and some exponent $\beta = \log|\lambda_-|/\log|\lambda_+| < 0$. The origin is then called a *saddle*, and the linear map *hyperbolic*.

The next case is when A has only one eigenvalue λ, with geometric multiplicity one (i.e. admits just a one-dimensional family of eigenvectors). It can be shown that the system is linearly conjugated with

$$f(x,y) = (\lambda x + y, \lambda y)\,.$$

(any eigenvector is proportional to $(1,0)$). One easily check that iterations of this map are

$$f^n(x,y) = \lambda^{n-1}(\lambda x + ny, \lambda y)\,.$$

Therefore, if $|\lambda| < 1$ all trajectories converge to the origin, which is then called a *degeneate stable node*. If $|\lambda| > 1$, then all trajectories of points different from the origin diverge. The origin is then called *degenerate source*. It is clear, however, that this is not a generic situation. A small (generic) perturbation of the matrix leads to one of the preceeding or the following case.

Finally, it may happen that the characteristic polynomial has no real roots, but a couple of complex conjugated roots $\lambda_\pm = \rho e^{\pm i\theta}$, for some $\rho = |\lambda_\pm| > 0$ and $\theta \notin \pi\mathbb{Z}$. This means that the complexification of A, the linear operator $A^{\mathbb{C}} : \mathbb{C}^2 \to \mathbb{C}^2$ defined in the canonical basis by the same matrix as A, admits two linearly independent eigenvectors $v_\pm$, corresponding to the two complex eigenvalues $\lambda_\pm$. Moreover, since $A = \overline{A}$, we may take $v_- = \overline{v_+}$. But then, in the basis of $\mathbb{R}^2 \subset \mathbb{C}^2$ defined by $e_1 = (v_+ + v_-)/2$ and $e_2 = (v_+ - v_-)/2i$, the map $f : \mathbb{R}^2 \to \mathbb{R}^2$ is induced/defined by the matrix

$$B = \rho \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}\,.$$

This means that f is a (counter-clockwise) rotation by an angle θ followed by a homothety/scaling with ratio ρ. Iterations of B are simply

$$B^n = \rho^n \begin{pmatrix} \cos(n\theta) & -\sin(n\theta) \\ \sin(n\theta) & \cos(n\theta) \end{pmatrix}$$

To understand trajectories, it is easier to identify the plane with the complex line $\mathbb{R}^2 \approx \mathbb{C}$, and use polar coordinates $(x, y) \approx x + iy = re^{i\varphi}$. Then $f^n(re^{i\varphi}) = r\rho^n e^{i(\varphi+n\theta)}$, and therefore trajectories move along logarithmic spirals

$$r = Ce^{\gamma\varphi},$$

for some constant C and some exponent $\gamma = (\log \rho)/\theta$, which may be positive or negative, or along circles $r = C$, if it happens that $\rho = 1$. In particular, if $|\lambda_\pm| = \rho < 1$, then all trajectories converge to the origin $f^n(x, y) \to 0$ as $n \to \infty$, which is then called a *stable focus*. If $|\lambda_\pm| = \rho > 1$, then the trajectories of all points different from the origin diverge, and the origin is then called an *unstable focus*.

ex: Describe what happens in the exceptional situation when $f(x, y) = (x + y, y)$, i.e. the only eigenvalue is 1 and it has geometric multiplicity one.

General linear maps, Jordan normal form. Let $f(x) = Ax$ be a linear system defined by a real $n \times n$ matrix A, and consider its complexification, i.e. the linear operator $A : \mathbb{C}^n \to \mathbb{C}^n$, acting on $\mathbb{C}^n = \mathbb{R}^n \oplus i\mathbb{R}^n$ according to $A(x + iy) := Ax + iAy$. According to the Jordan normal form theorem, the complexified linear space is a direct sum $\mathbb{C}^n = \bigoplus E_\lambda$ of *generalized eigenspaces*, or *root spaces*, E_λ, which are invariant under A and where the action of A is

$$\lambda I + N$$

where λ is the eigenvalue, and N is a nilpotent operator. More precisely, the matrix which represents the restriction of A on $E_\lambda \subset \mathbb{C}^n$ is a Jordan block

$$J = \begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{pmatrix} \tag{6.2}$$

(empty entries are all zero). Moreover, generalized eigenspaces with non-real eigenvalues come in pairs of generalized eigenspaces E_λ and $E_{\overline{\lambda}}$, whose vectors are related by a complex conjugation. As in the two-dimensional situation, one can then contruct an invariant subspace $E_{\lambda,\overline{\lambda}} \subset \mathbb{R}^n$ where the acion of A is given by the Jordan block

$$\begin{pmatrix} \rho R_\theta & I & & & \\ & \rho R_\theta & I & & \\ & & \ddots & \ddots & \\ & & & \rho R_\theta & I \\ & & & & \rho R_\theta \end{pmatrix} \tag{6.3}$$

where

$$\rho R_\theta = \rho \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} \quad \text{and} \quad I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and $\lambda = \alpha + i\beta = \rho e^{i\theta}$.

The conclusion if that the phase space $\mathbb{R}^n$ of the real linear system $f(x) = Ax$ splits as a direct sum of invariant subspaces, E_λ or $E_{\lambda,\overline{\lambda}}$, where A acts as (6.2) or as (6.3), respectively.

It is clear that the asymptotic behavior of iterations of A on each root space depends on the absolute value $|\lambda|$ of the corresponding eigenvalue. Indeed, one can write the total space as a direct sum of three invariant subspaces

$$\mathbb{R}^n = E^- \oplus E^0 \oplus E^+$$

where the *stable space* E^- is the direct sum of those root spaces with $|\lambda| < 1$, the *unstable space* E^+ is the direct sum of those root spaces with $|\lambda| > 1$, and finally the *neutral space* E^0 is the direct sum of those root spaces with $|\lambda| = 1$.

One can then show that the restriction of f to E^- is eventually contracting (i.e. some power is contracting), and therefore $f^n(x) \to 0$ as $n \to \infty$ if $x \in E^-$. This happens because the exponential contraction dominates the nilpotent part of each Jordan block for sufficiely large n, and therefore $\|A^n v\| \leq C\mu^n \|v\|$ for some constants C and $\mu < 1$ and every $v \in E^-$. Similarly, one shows that the inverse of the restriction of f to E^+ is eventually contracting.

A linear map is called *hyperbolic* if all the eigenvalues have $|\lambda| \neq 1$, i.e. if the spectrum of the complexification is disjoint from the unit circle of the complex plane. This means that the phase space splits as a direct sum

$$\mathbb{R}^n = E^- \oplus E^+$$

of a stable and an unstable subspace. If an hyperbolic map has eigenvaules with absolute values both $|\lambda| < 1$ and $|\lambda| > 1$, then the origin is called a *saddle*.

6.4 Order of the line and trajectories

The order of the real line cause restrictions of possible trajectories of monotone transformations.

Increasing maps of the interval. Let $f : I \to I$ be a continuous increasing map if the interval $I \subset \mathbb{R}$. Any trajectory $(x_n)_{n\in\mathbb{N}_0}$ is monotone, increasing or decreasing (depending whether $f(x) > x$ or $f(x) < x$, respectively). The monotone trajectory may converge, i.e. $x_n \to p$ to some fixed point, if bounded, or may diverge $x_n \to \pm\infty$, is unbounded. In particular, if the interval I is compact, the second possibility is excluded, and all trajectories converge to some fixed point. In this case, there exists a not-empty compact set $F \subset I$ made of fixed points, and any x in each connected component A_k of $I \backslash F = \bigcup_k A_k$ has a trajectory contained in A_k which converge some point $x_\infty \in \partial A_k$.

ex: Show that a homeomorphism $f : I \to I$ of an interval $I \subset \mathbb{R}$ cannot have periodic points of period larger than 2. When does it have periodic points of period 2 ?

ex: Let $I \subset \mathbb{R}$ be compact interval and $f : I \to I$ a continuous and increasing function. Show that any trajectory converges to a fixed point. Discuss the dynamics of f

ex: Discuss the dynamics of a continuous and decreasing map $f : I \to I$ of a compact interval o $I \subset \mathbb{R}$ (observe that if f is decreasing then f^2 is increasing).

ex: (difficult!) Let $I \subset \mathbb{R}$ be an interval and $f : I \to I$ and $g : I \to I$ be two homeomprphisms of I without fixed points. Show that hey are topologically conjugated.

Sharkovskii order. The order of the real line also implies restrictions on the possible periods of a map. A striking result by Alexander N. Sharkovskii[23] says that there exists an order $\prec$ on the naturals, which looks like

$$\begin{aligned} 1 &\prec 2 \prec 2^2 \prec 2^3 \prec \cdots \prec 2^m \prec \cdots \prec 2^k \cdot (2n-1) \prec \dots \\ \dots &\prec 2^k \cdot 3 \prec \cdots \prec 2 \cdot 3 \prec \dots \prec 2n-1 \prec \cdots \prec 9 \prec 7 \prec 5 \prec 3 \end{aligned}$$

such that if a continuous transformation $f : \mathbb{R} \to \mathbb{R}$ has an orbit of period k and if $j \prec k$ then it also has an orbit of period j. In particular, the existence of an orbit of period 3 implies the existence of orbits of all periods!

ex: Try to figure a transformation of the real line with an orbit of period 3.

[23] A.N. Sharkovskii, Co-existence of cycles of a continuous mapping of the line into itself, *Ukrainian Math. J.* **16** (1964), 61-71.

6.5 Local analysis: attracting and repelling fixed points

Differentiability of a transformation and the contraction principle helps to understand the trajectories of points which are near to the fixed or periodic points.

Attracting and repelling fixed points. Let $f : X \to X$ be a transformation of class $\mathcal{C}^1$ defined in some open subset $X \subset \mathbb{R}^n$, and let $p \in X$ be a fixed point of f.

We say that the fixed point p is *attracting* if its basin of attraction $W^s(p)$ is a neighborhood of p, i.e. if p its admits a neighborhood B such that $f^n(x) \to p$ for all $x \in B$. The following criterium is a simple consequence of the contraction principle.

Theorem 6.6. *If $|f'(p)| < 1$, then p is attractiong.*

Proof. By continuity of f', there exist $\lambda < 1$ and a ball $B = B_\varepsilon(p)$ around p such that $|f'(x)| < \lambda$ for all $x \in B$. By the mean value theorem, $f\left(\overline{B}\right) \subset \overline{B}$, since if $d(x,p) \leq \varepsilon$ then

$$d(f(x), p) \leq \lambda \cdot d(x,p) < \varepsilon$$

Moreover, the mean value theorem also implies that $d(f(x), f(x')) \leq \lambda \cdot d(x,x')$ if $x, x' \in \overline{B}$. Thus, $f|_{\overline{B}} : \overline{B} \to \overline{B}$ is a contraction, and the contraction principle says that trajectories of all points $x \in \overline{B}$ converge exponentially to p. □

We say the fixed point p is *repelling* if it admits a neighborhood B such that the trajectory of any $x \in B$, different from p, leaves B in finite time, i.e. $f^n(x) \notin B$ for some $n \geq 1$. The following criterium use the orde of the real line.

Theorem 6.7. *Let $f : X \to X$ be a transformation of class $\mathcal{C}^1$ defined in some open interval $X \subset \mathbb{R}$. If $|f'(p)| > 1$, then p is repelling.*

Proof. By continuity of f', there exist $\lambda > 1$ and an interval $B = [p-\varepsilon, p+\varepsilon]$ around p sach that $|f'(x)| > \lambda$ for all $x \in B$. Also observe that f is strictly increasing or decreasing, depending on the sign of $f'(p)$, and therefore sends bijectively intervals onto intervals. take a point $x \in B$ different from p, and suppose that $f^k(x) \in B$ for all times $0 \leq k \leq n$. The chain rule implies that the derivative of f^n at points c between p and x grow exponentially, since

$$\left|(f^n)'(c)\right| = \left|f'\left(f^{n-1}(c)\right)\right| \cdot \left|f'\left(f^{n-2}(c)\right)\right| \dots |f'(c)| > \lambda^n$$

The mean value theorem implies that n cannot be arbitrarily large, since

$$d(p, f^n(x)) \geq \lambda^n \cdot d(p,x) \qquad \text{and} \qquad d(p, f^n(x)) \leq \varepsilon$$

are not compatible for large n. Thus, there exists a time $n \geq 1$ duch that $f^n(x) \notin B$. □

It must be said that this result is local, it does not say anything about the basin of attraction of p. Also, the condition $|f'(p)| > 1$ is not sufficient to estabilish a similar result in higher dimension (there may be directions where f dilates distances, and directions where it contracts distances ...)

ex: Show by examples that the basin of attraction of a repelling fixed point p can be larger than $\{p\}$.

ex: Find a good definition of *attractiong peiodic orbit* (observe that the derivative of f^n is constant along a periodic orbit of period n, then consider iterations of f^n ...)

ex: Consider the family of quadratic maps

$$x \mapsto \lambda x^2$$

depending on the parameter λ. Find the basin of attraction of the fixed point $p = 0$, and describe the speed of convergence of convergent trajectories.

ex: If p is a fixed point of $f : \mathbb{R} \to \mathbb{R}$ such that $f'(p) = 1$, then everything can happen! The basin of attraction of p can be a neighborhood of p, or just $\{x\}$, or may contain an half-neighborhood like $[p, p \pm \varepsilon)$ ...

Consider the examples

$$x \mapsto x \pm x^3 \qquad \text{e} \qquad x \mapsto x \pm x^2$$

and find others.

e.g. The quadratic family. The *quadratic family* is the family of transformations of the unit interval $f_\lambda : [0, 1] \to [0, 1]$, defined according to

$$f_\lambda(x) = \lambda x(1 - x)$$

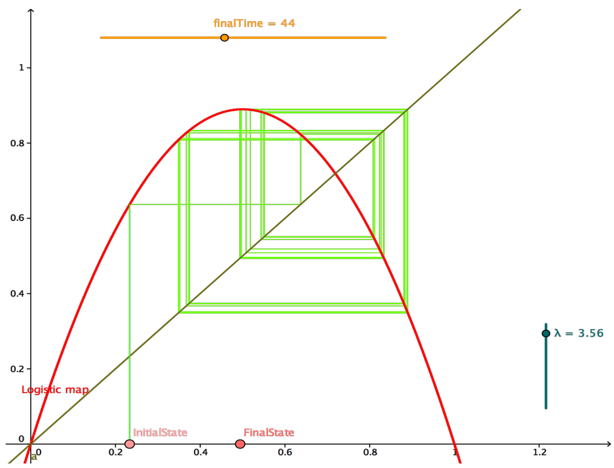
Here the parameter λ takes values in the interval $[0, 4]$. It is also called *logistic* (from the French "logement"), since it is a model of population growth in a limited environment, x being the relative population, the quotient of the actual population over the maximal allowed population.

Fixed points are 0, which is attracting for $0 \leq \lambda < 1$, and $p_\lambda = \frac{\lambda - 1}{\lambda}$, which appears when $\lambda > 1$ (remeber that our phase space is only the unit interval and not the whole real line) and is attracting when $1 < \lambda < 3$.

If $\lambda \in [0, 1)$ then all trajectories converge to 0. Indeed, trajectories are bounded and decreasing sequences, and 0 is the unique fixed point.

If $\lambda \in (1, 3]$ then all trajectories converge to p_λ. This is not so obvious.

What is really interesting is to observe what happens for increasing values of $\lambda > 3$. You may take a look at my applet in http://w3.math.uminho.pt/~scosentino/salbestiario.html.



Cobweb plot of the logistic map, for $\lambda \simeq 3.56$.

e.g. Convergence for Newton method Let $F \in \mathbb{R}[x]$ be a polynomial with real coefficients. Newton method to find the roots of F, i.e. to solve the equation $F(x) = 0$, consists in choosing a first approximation x_0, and then iterate

$$x_{n+1} = x_n - \frac{F(x_n)}{F'(x_n)} .$$

This means that we try to refine our bet x_n using the linear approximation (first order Taylor)

$$F(x) \simeq F(x_n) + F'(x_n) \cdot (x - x_n)$$

Clearly, we may iterate provided the derivative stays away from zero in a neighborhood of the root we want to approximate.

It is clear that if the sequence (x_n) converges to some p, and if $F'(p) \neq 0$, then the limit $p = \lim_{n\to\infty} x_n$ is a root of the polynomial F. Conversely, if p is a root of F, and if $F'(p) \neq 0$ (so that it is also different from zero in a neighborhood of p), then p is a fixed point of the map

$$x \mapsto f(x) := x - \frac{F(x)}{F'(x)}$$

The derivative of f at p is

$$f'(p) = 1 - \frac{(F'(p))^2 - F(p)\, F''(p)}{(F'(p))^2} = 0$$

Therefore, p is an attracting fixed point of f: the trajectory of any initial guess x_0 sufficiently close to p converges to p.

Indeed, since the derivative is $f'(p) = 0$, any root of F is a super-attracting fixed point of f, and the convergence is much better than exponential.

Theorem 6.8. *Let p be a non-critical root of the polynomial $F \in \mathbb{R}[x]$, i.e. a root where $F'(p) \neq 0$. Then Newton's iterations starting from any x_0 sufficiently near the root p converge to this root, and the convergence is "quadratic", i.e. the error $\varepsilon_n = |x_n - p|$ decreases as*

$$\varepsilon_{n+1} \leq K \cdot \varepsilon_n^2$$

for some $K > 0$.

Proof. We may assume, without loss of generality, that the root we are looking for is the origin, so that $F(0) = 0$. Now, suppose we are at x_n after n iterations. Taylor's formula with Lagrange estimate of the error around x_n says that

$$F(x) = F(x_n) + F'(x_n)\cdot(x - x_n) + \frac{1}{2}F''(y)\cdot(x - x_n)^2$$

for some y between x and x_n. Taking $x = 0$ (the root!) and dividing by $F'(x_n)$ we get

$$0 = F(0) = F(x_n) - F'(x_n) - x_n + \frac{1}{2}F''(y)\cdot x_n^2$$

and therefore

$$x_n - \frac{F(x_n)}{F'(x_n)} = \frac{1}{2}\frac{F''(y)}{F'(x_n)}\, x_n^2$$

But the l.h.s. is x_{n+1}, so that

$$x_{n+1} = \frac{1}{2}\frac{F''(y)}{F'(x_n)}\, x_n^2$$

Since $F'(0) \neq 0$ (and polynomials have continuous derivatives), there is an interval $I =]-\varepsilon, \varepsilon[$ around the root 0 where $M = \sup_{x\in I} |F''(x)| < \infty$ and $\delta = \inf_{x\in I} |F'(x)| > 0$. Let $K = M/2\delta$. There follows that the distance $\varepsilon_n = |x_n - 0|$ between the n-th iterate and the root satisfies the iterative bound

$$|\varepsilon_{n+1}| \leq K \cdot |\varepsilon_n|^2$$

□

ex: Check that Newton's method applied to the quadratic polynomial $z^2 - a$, with $a > 0$, corresponds to Heron's algorithm.

ex: Estimate $\sqrt{17}$.

ex: Write down Newton's repice to solve $z^n - a = 0$, with $a > 0$ and $n \geq 2$.

ex: Use Newton's method to estimate the roots of

$$z^2 + 1 + z \qquad z^3 - z - 1 \qquad z^5 + z + 1 \qquad z^3 - 2z - 5$$

Linearization in the complex plane. Let $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ be a rational function defined in the Riemann sphere $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. Any fixed point p has its basin of attraction B_p. Looking for fast methods to compute the iterates, in 1871 E. Schrœder had the idea to look for local conformal conjugations of f with simpler rational functions, like affine functions $g : z \mapsto \lambda z$. The method amount to solve the functional equation

$$h \circ f\,|_{B_p} = g \circ h\,,$$

where $h : B_p \to B$ is an holomorphic function. E. Schrœder, G. Kœnig and J.H. Poincaré solved the problem with $|\lambda| \neq 1$, and then Carl S. Siegel solved the case $|\lambda| = 1$ around 1940.

Theorem 6.9 (Kœnigs)**.** *Let z_0 be a fixed point of f with multiplier $f'(z_0) = \lambda$ such that $|\lambda| \neq 0, 1$. Then there exists a conformal map ϕ, unique up to a non-zero factor, from a neighborhood of z_0 onto a neighborhood of 0 such that $\phi \circ f = \lambda \cdot \phi$.*

Proof. We assume that z_0 is attracting, i.e. $|\lambda| < 1$, since the repelling case follows considering the local inverse of f. Also, after conjugation, we can assume that $z_0 = 0$, hence the map has the form

$$f(z) = \lambda z + a_2 z^2 + \dots .$$

Now define $\phi_n(z) = f^n(z)/\lambda^n$. There exists a $\delta > 0$ and a constant $c < |\lambda| < 1$ such that, for $|z| < \delta$,

$$|\phi_{n+1}(z) - \phi_n(z)| \;\leq\; k \cdot (c/|\lambda|)^n$$

for some $k > 0$. Hence the sequence of holomorphic functions ϕ_n converges uniformly in a small ball around 0. The functional equation $\phi \circ f = \lambda \cdot \phi$ follows immediatly from its definition.

Comparing coefficients it is easy to see that any conjugation of $z \mapsto \lambda z$ to itself is a constant multiple of the identity, as long as $|\lambda| \neq 0, 1$. Uniqueness follows. □

Theorem 6.10 (Böttcher)**.** *Let z_0 be a superattracting fixed point of f, where*

$$f(z) \;=\; z_0 + a_p (z - z_0)^p + \dots$$

with $a_p \neq 0$ and $p \geq 2$. Then there exists a conformal map ϕ, unique up to multiplication by a $(p-1)$-root of unity, from a neighborhood of z_0 onto a neighborhood of 0 such that $\phi \circ f = \phi^p$.

Proof. (scketch) We can assume that $z_0 = 0$ and that $a_p = 1$. As in Koenigs proof, we look for the conjugation as a limit af the functions

$$\phi_n(z) = f^n(z)^{p^{-n}} .$$

It can be shown that the ϕ_n converge uniformly in some sufficiently small ball around 0, and the functional equation follows from the definition. Uniqueness, up to a $(p-1)$-root of unity, can be checked comparing power series. □

6.6 Transversality and bifurcations

Transversality. Let $f : I \to \mathbb{R}$ be a transformation o fclass $\mathcal{C}^1$ defined in some interval $I \subset \mathbb{R}$, and let p be a fixed point of f. If $f'(p) \neq 1$, then this fixed point is "isolated", i.e. it is the unique fixed point of f in some neighborhood of p. Indeed, a fixed point is a solution of

$$F(x) = f(x) - x = 0$$

Now, if $f'(p) \neq 1$ then $F'(p) \neq 0$. The inverse function theorem then says that F is invertible in a neighborhood B of p, and this implies that p is the unique zero of F in B, so that p is the unique fixed point of f in B.

Fixed points satisfying the condition $f'(p) \neq 1$ are called *transversal*, because the tangent to the graph of f at p is transversal to the (tangent to the) graph of the identity function.

Persistence. The condition $f'(p) \neq 1$ is an open condition, and this suggests that it may stable under small perturbations of f.

Theorem 6.11. *Let $f : I \to \mathbb{R}$ be a transformation of class $\mathcal{C}^1$, and p be a transversal fixed point of f. Then all transformations $g : I \to \mathbb{R}$ sufficiently $\mathcal{C}^1$-near to f have one, and only one, fixed point in some neigborhood of p, which is also transversal.*

Proof. Let $g = f - h$ be a perturbation of f, with $\|h\|_{\mathcal{C}^1} = \|h\|_\infty + \|h'\|_\infty < \delta$. A fixed point of g is a solution of $g(x) - x = 0$, i.e. of

$$F(x) = h(x)$$

if we define $F(x) = f(x) - x$. We know that F in some neighborhood B' of p, hence a fixed point of g inside B is a solution of $x = \left(F^{-1} \circ h\right)(x)$, which means a fixed point of $F^{-1} \circ h$. The strategy, now, it to show that $F^{-1} \circ h$ is a contraction in some neighborhood of p. If the closed neighborhhod $B = \overline{B_r(p)}$ is sufficiently small, then the inverse of F has bounded derivative, say $\left|\left(F^{-1}\right)'(x)\right| < M$ in $F(B)$. If δ is sufficiently small, then the derivative $\left|\left(F^{-1} \circ h\right)'(x)\right| < M \cdot \delta$ is uniformly $\leq \lambda := M\delta < 1$ in B, and therefore $F^{-1} \circ h$ has good chances to be a contraction. We must show that the image $\left(F^{-1} \circ h\right)(B)$ is contained in B. Now, given $x \in B$, triangular inequality, the mean value theorem and the chain rule, imply o

$$\begin{aligned} d\left(\left(F^{-1} \circ h\right)(x), p\right) &\leq d\left(F^{-1}(h(x)), F^{-1}h(p)\right) + d\left(F^{-1}(h(p)), p\right) \\ &\leq d\left(F^{-1}(h(x)), F^{-1}h(p)\right) + d\left(F^{-1}(h(p)), F^{-1}(0)\right) \\ &\leq M \cdot \delta \cdot r + M \cdot \delta \end{aligned}$$

(whre we used the fact tha p is a fixed point of f) and this quantity is $< r$ whenever δ is sufficiently small. The contraction principle then says that a fixed point $p' \in B$ of g exists and is unique. The derivative of g at this point is δ-near the derivative of f in p, and therefore this fixed point is also transversal. □

ex: Let $f : \mathbb{R} \to \mathbb{R}$ be a transformation of class $\mathcal{C}^1$, and let p a periodic point of period n such that $(f^n)'(p) \neq 1$. Show that all transformations sufficiently $\mathcal{C}^1$-near to f have a periodic point of period n near p. (consider the iterate f^n and apply the above theorem)

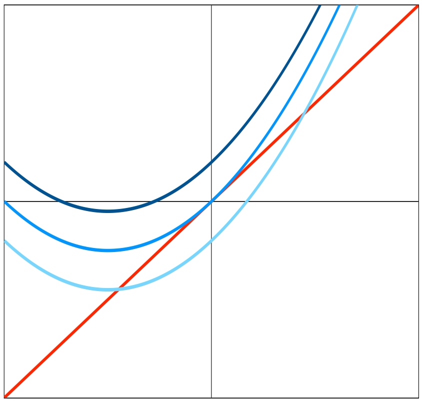
ex: Let $f : V \to \mathbb{R}^n$ be a transformation of class $\mathcal{C}^1$ defined in some open set $V \subset \mathbb{R}^n$, and let p be a fixed point of f. Transversality of p now means that the derivative (Jacobian) operator $f'(p)$ does not have 1 has an eigenvalue. State and prove the analogous of theorem 6.11in this case.

Bifurcations. Non-transversal fixed points need not be persistent, and may disappear or change thei nature under generic perturbations. This phenomenon is called *bifurcation*. The idea of bifurcation theory is to treat families f_λ of transformations, defined in some neigborhood of a fixed or periodic point of $f = f_0$, and describe possible changes in the dynamics when the parameter λ varies.

Consider, for example, the family

$$f_\lambda(x) = x + x^2 - \lambda$$

defined in the real line. The origin is a non-transversal fixed point of f_0, where $f_0'(0) = 1$. If $\lambda \neq 0$ is small, then f_λ has two fixed points $\pm\sqrt{\lambda}$, one repelling and the other attracting, if $\lambda > 0$, or none if $\lambda < 0$.



Graphs of $f(x) = x + x^2 - \lambda$, for $\lambda = -0.2$, 0 and 0.2 (different kinds of blue), compared with the diagonal (red).

The family

$$f_\lambda(x) = x + x^3 + \lambda x$$

shows a different behavior. The problem is to decide which phenomena are "generic", and possibly "stable", in some sense to be specified.

If we admit the existence of a sufficent number of derivatives, an arbitrary family of transformations with a non-transversal fixed point at the origin when $\lambda = 0$ is

$$\begin{aligned} f_\lambda(x) &= a_\lambda + b_\lambda x + c_\lambda x^2 + \ldots \\ &= \left(a'\lambda + a''\lambda^2 + \ldots\right) + \left(1 + b'\lambda + b''\lambda^2 + \ldots\right)x + \left(c + c'\lambda + c''\lambda^2 + \ldots\right)x^2 + \ldots \end{aligned}$$

The generic case is when $c \neq 0$ (i.e. $f_0 = x + cx^2 + \ldots$) and a generic perturbation has $a' \neq 0$ (i.e. the constant term of f_λ is different from zero when $\lambda \neq 0$). It is then not diffult to convince onself that this family behaves qualitatively like the simpler family $f_\lambda(x) = x + x^2 - \lambda$ above. A small perturbation of f_0 may destroy the fixed point, in one direction, or create two new fixed points, in the other direction.

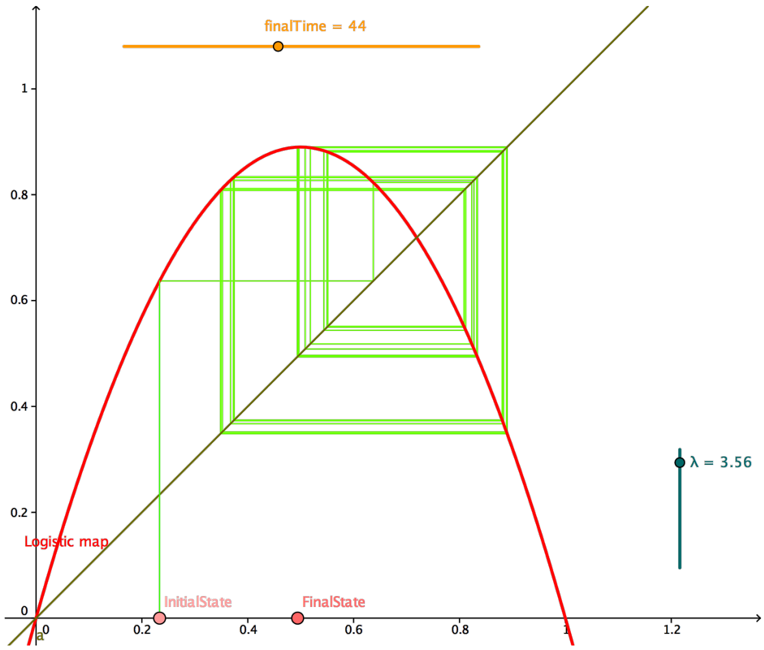
ex: State and prove the above result (observe that looking for roots of $f_\lambda(x) = x$, as function of λ, amount to to define look for functions $\lambda \mapsto x(\lambda)$ which satisfy $G(\lambda, x) = f_\lambda(x) - x = 0$, and this problem is treated by the implicit function theorem).

Period-doubling and Feigenbaum universality. Also interesting is the case of a family interval transformations f_λ such that f_0 has a fixed point at the origin with multiplies $f_0'(0) = -1$. Such a fixed point is transversal, hence persistent. Meanwhile, $(-1)^2 = 1$, and therefore the derivative of f_0^2 at 0 is $\left(f_0^2\right)'(0) = 1$. This says that the origin is not tranversal as a fixed point of the second iterate f^2. A small peturbation may produce periodic points of period 2 in a neighborhood of the persistend fixed point 0.

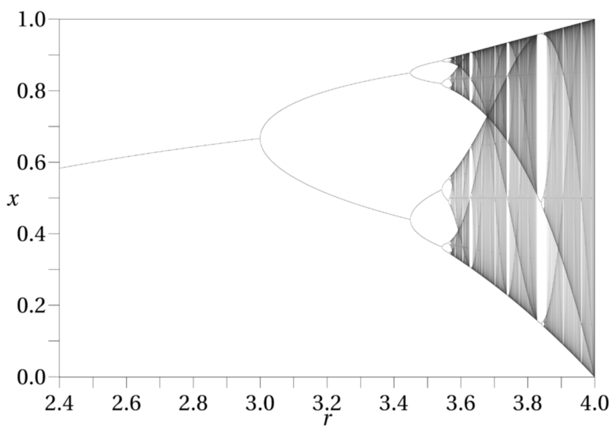
This kind of bifurcation is called "period-doubling". An example is that of the family

$$f_\lambda(x) = -x + x^2 + \lambda x$$

Another example occurrs in the quadratic family $f_\lambda(x) = \lambda x(1 - x)$, when we pass the value $\lambda = 3$ of the parameter. You may check this with my applet http://w3.math.uminho.pt/~scosentino/bestiario/logistic.html.



Doing simulations with a computer, Mitchell J. Feigenbaum discovered, in the '70 of the last century, that certai families of transformations produce a "cascade" of period-doublings, i.e. there is a sequence of values $\lambda_1 < \lambda_2 < ... < \lambda_n < \lambda_{n+1}...$ of the parameter such that, when passing through λ_{n+1} orbits of period 2^{n+1} are created in a neighborhood of orbits of period 2^n, created by the previous value λ_n. This phenomenon is easily observed, and indeed seems to be "universal": it happens for almost all families, provided we find the region where it takes place. The following picture is obtained if we plot the parameter λ, within some interval, versus a typical orbit of f_λ, say $\{x_{100}, x_{101}, \ldots, x_{200}\}$ starting from a random initial point x_0. Here is what you get.



Bifurcation diagram for the logistic family.
(from https://en.wikipedia.org/wiki/Period-doubling_bifurcation)

Even more misterious is that, as already observed by Feigenbaum, the limit $\lambda_\infty = \lim_{n\to\infty} \lambda_n$ seems to exist, it is achieved exponencially, i.e. $|\lambda_\infty - \lambda_n| \simeq$const$\times\delta^{-n}$ where

$$\delta = \lim_{n\to\infty} \frac{\lambda_n - \lambda_{n-1}}{\lambda_{n+1} - \lambda_n} \simeq 4.669201609102990671853\ldots$$

seem to be independent from the family! This mistery was explained later by Lanford, Epstein, Sullivan, ...

7 Statistical description of orbits

Together with the topological point of view, a source of informations about dynamical systems is their statistical description. The idea is to measure the relative size of those points whose orbits have certain definite properties. This is done looking for invariant probability measures, and the main result is the Birkhoff-Khinchin ergodic theorem. To state and prove the Birkhoff-Khinchin ergodic theorem, we need to recall many standard facts and results of integration theory. You can find most of them in the classical [Ru87] or [Ha74].

7.1 Probability measures

Probability spaces. A *measurable space* is a pair $(X, \mathcal{E})$, a non-empty set X together with a σ-algebra of subsets $\mathcal{E}$. Recall that a (Boolean) algebra is a nonempty family $\mathcal{A}$ of subsets of X which contains X, which contains the complement of any of its elements, and which is closed under finite unions and intersections. A σ-algebra is an algebra which is also closed under countable unions and intersections. Given any family $\mathcal{C}$ of subsets of X, there exists a minimal σ-algebra $\sigma(\mathcal{C})$ which contains all the elements of $\mathcal{C}$, which is called the σ-algebra generated by $\mathcal{C}$.

If (X, τ) is a topological space, the *Borel σ-algebra* is $\sigma(\tau)$, the smallest σ-algebra which contains all open sets.

A *measure* on the measurable space $(X, \mathcal{E})$ is a σ-additive function $\mu : \mathcal{E} \to [0, \infty]$ such that $\mu(\emptyset) = 0$. Here σ-additivity means that, if (S_n) is a countable family of pairwise disjoint elements of $\mathcal{E}$, then

$$\mu(\cup_n S_n) = \sum_n \mu(S_n)$$

The triple $(\Omega, \mathcal{E}, \mu)$ is said a *measure space*, or *probability space* if it happens that $\mu(X) = 1$. Given a probability space, measurable sets $A \in \mathcal{E}$ are commonly called "events", and the number $\mu(A)$ is called "probability of the event A". Basic properties of probability measures are the following: probability measures are monotone, i.e. $\mu(S) \leq \mu(T)$ if $S \subset T$, and σ-subadditive, i.e. if (S_n) is a countable family of elements of $\mathcal{E}$ then

$$\mu(\cup_n S_n) \leq \sum_n \mu(S_n)$$

Probability measures are continuous from below and from above, in the following sense: if $S_n \uparrow S$ then $\mu(S_n) \uparrow \mu(S)$, and if $S_n \downarrow S$ then $\mu(S_n) \downarrow \mu(S)$. Both continuity properties are equivalent, and indeed a simple argument shows that they are equivalent to continuity from above at $\emptyset$: if $S_n \downarrow \emptyset$ then $\mu(S_n) \downarrow 0$. Moreover, continuity is equivalent to σ-aditivity if the set function μ is only assumed (finitely) additive.

A subset $E \subset X$ *has zero measure* if it is contained in a measurable set $S \in \mathcal{E}$ with $\mu(S) = 0$. If any set with zero measure belongs to $\mathcal{E}$, then the measure space $(X, \mathcal{E}, \mu)$ is said *complete*. Any measure space can be canonically completed, extending the measure to the σ-algebra $\overline{\mathcal{E}}$ made of $\mathcal{E}$ and of subsets of zero measure. A property (like continuity of a function, or convergence of a sequence of functions) holds *μ-a.e.* ("almost everywhere" with respect to the measure μ) if the set of points of X where it does not hold has zero measure.

Construction of probability measures. Measures are never "explicitely" given as functions on a σ-algebra. A set function $\mu : \mathcal{P}(X) \to [0, \infty]$ is an *exterior measure* if it is monotone, σ-subadditive, and if $\mu(\emptyset) = 0$. It happens that, given an exterior measure μ, the family of *μ-measurable sets*, defined as

$$\mathcal{E} = \{E \subset X \text{ such that } \mu(S) = \mu(S \cap E) + \mu(S \cap E^c) \text{ for any } S \subset X\}$$

is a σ-algebra, and that μ is a complete measure if restricted on $\mathcal{E}$ (the proof is quite long and delicate, but the only idea it uses is the following: in order to check that $E \in \mathcal{E}$ it is indeed sufficient, by virtue of monotonicity and subadditivity of μ, to check that $\mu(S) \geq \mu(S \cap E) + \mu(S \cap E^c)$ for any $S \subset X$). A strategy to construct interesting measures on uncountable spaces is: start with an exterior measure (it is very easy to produce exterior measures, for example by means of variational principles) and then check that the σ-algebra of measurable sets is sufficiently big for our purpose.

The idea of Carathéodory is the following. A *probability measure* on an algebra $\mathcal{A}$ of subsets of X is an additive function $m : \mathcal{A} \to [0,1]$ such that $m(\emptyset) = 0$, $m(X) = 1$, and such that $A_n \downarrow \emptyset$ implies $m(A_n) \downarrow 0$. Given a probability measure m on a algebra $\mathcal{A}$, the recipe

$$\mu(S) = \inf \left\{ \sum m(A_n) \text{ with } S \subset \cup_n A_n \text{ e } A_n \in \mathcal{A} \right\}$$

defines an exterior measure on $\mathcal{P}(X)$, hence the above construction produces a measure μ on the σ-algebra of μ-measurable sets, which contains $\mathcal{A}$ and so contains $\sigma(\mathcal{A})$. One then checks that $\mu(A) = m(A)$ for any $A \in \mathcal{A}$, so that μ is an "extension" of the measure m. Carathéodory's extension theorem is then stated in the following form:

Theorem 7.1 (Carathéodory's extension theorem)**.** *Given a probability measure m on a algebra $\mathcal{A}$ of subsets of X, there exists a unique measure μ on $\sigma(\mathcal{A})$ which extends m.*

The following corollary of Carathéodory's theorem is also useful, for example when trying to prove that some event has a definite probability.

Theorem 7.2 (Approximation theorem)**.** *Let $(X, \mathcal{E}, \mu)$ be a probability space, and let $\mathcal{A}$ be an algebra of subsets of Xsuch that $\sigma(\mathcal{A}) = \mathcal{E}$. Then, for any $A \in \mathcal{E}$ and any $\varepsilon > 0$, we can find a $A_\varepsilon \in \mathcal{A}$ such that*

$$\mu(A \Delta A_\varepsilon) < \varepsilon.$$

Indeed, one easily sees that the family $\mathcal{C} = \{A \in \mathcal{E} \text{ s.t. } \forall \varepsilon > 0 \ \exists A_\varepsilon \in \mathcal{A} \text{ s.t. } \mu(A \Delta A_\varepsilon) < \varepsilon\}$ is a σ-algebra. Since $\mathcal{A}$ is obviously contained in $\mathcal{C}$, this implies that $\mathcal{E} = \sigma(\mathcal{A}) \subset \mathcal{C} \subset \mathcal{E}$.

Lebesgue measure. The collection $\mathcal{I}$ of intervals of the real line is a *semi-algebra*, i.e. the intersection of two elements of $\mathcal{I}$ is in $\mathcal{I}$ and the complement of an element of $\mathcal{I}$ is a union of elements of $\mathcal{I}$. The function $m : \mathcal{I} \to [0, \infty]$, defined as $m([a,b]) = |b-a|$ if a e b are finite, and ∞ if the interval is unbounded, is monotone and gives value zero to the empty set. Postulating additivity, the function m extends to a measure on the algebra $\mathcal{A}$ made of disjoint unions of elements of $\mathcal{I}$ (this is not trivial!, the proof uses the Heine-Borel theorem about compact subsets of the real line). The function $\mu : \mathcal{P}(\mathbb{R}) \to [0, \infty]$, defined as

$$\mu(E) = \inf \left\{ \sum m(C_n) \text{ with } E \subset \cup_n C_n \text{ e } C_n \in \mathcal{A} \right\}$$

is then an exterior measure on the real line. The σ-algebra $\mathcal{L}$ of μ-measurable sets, called *Lebesgue σ-algebra*, contains the Borel sets, because it contains the intervals. The restriction $\ell = \mu|_{\mathcal{L}}$, as well as $\mu\big|_{\mathcal{B}(\mathbb{R})}$, is called *Lebesgue measure.*

Observe that Lebesgue measure on the real line is not a probability measure, having infinite mass. Nevertheless, one can easily define probability measures on bounded intervals taking normalized restrictions of Lebesgue measure. For example, take $X = [0,1]$, and $\mathcal{E} = \mathcal{B}(X) = \{X \cap B \text{ with } B \in \mathcal{B}(\mathbb{R})\}$, the Borel subsets of the interval. The restriction of ℓ to $\mathcal{E}$ is a probability measure, called Lebesgue measure on the unit interval.

The very same construction works in $\mathbb{R}^n$, starting with the semi-algebra of "rectangles" measured by the "euclidean volume", and produces a measure ℓ on $\mathcal{B}(\mathbb{R}^n)$, also called Lebesgue measure. Lebesgue measure is the unique measure over the Borel sets of the euclidean space which is invariant under traslations, i.e. $\ell(\lambda + B) = \ell(B)$ for any $\lambda \in \mathbb{R}^n$ and any Borel set B, and which is normalized to give measure one to the unit square, i.e. $\ell([0,1]^n) = 1$.

The axiom of choice allows one to "give examples" of subsets wich are not Lebesgue-measurable (for example, the set made of one point for each orbit of an irrational rotation of the circle).

The following result is useful (see [Mat95] for a proof). Below, $B_\varepsilon(x) = \{y \in \mathbb{R}^n \text{ s.t. } \|x - y\|_2 < \varepsilon\}$ denotes the open ball of radius $\varepsilon > 0$ and center $x \in \mathbb{R}^n$ w.r.t. the Euclidean distance $\|x-y\|_2^2 = \sum_{i=1}^n |x_i - y_i|^2$.

Theorem 7.3 (Lebesgue density theorem)**.** *Let $A \subset \mathbb{R}^n$ be a Lebesgue-measurable set. For ℓ-almos any $x \in A$ the density*

$$\lim_{\varepsilon \to 0} \frac{\ell(A \cap B_\varepsilon(x))}{\ell(B_\varepsilon(x))} = 1$$

Kolmogorov extension. Let X be a finite space, equipped with the discrete topology, and let Σ^+ be the topological product $X^{\mathbb{N}} = \{x : \mathbb{N} \to X\}$, its point indentified with sequences $x = (x_1, x_2 \ldots, x_n, \ldots)$ with $x_n \in X$. Let $\mathcal{C}$ be the collection of *cylinders* of X, the subsets of the form

$$C_B = \left\{x \in \Sigma^+ \text{ s.t. } (x_1, x_2 \ldots, x_n) \in B\right\}$$

with B an open subset of X^n. Cylinders form a basis of the product topology of Σ^+, which makes Σ^+ a compact metrizable space. In particular, the Borel σ-álgebra of Σ^+ is $\mathcal{B} = \sigma(\mathcal{C})$. Let $\mu_1, \mu_2, \mu_3, \ldots, \mu_n, \ldots$ be probability measures defined on the Borel sets of $X, X^2, \ldots, X^n, \ldots$, respectively. The sequence (μ_n) is said *consistent* if

$$\mu_{n+1}(B \times X) = \mu_n(B)$$

for any n and any Borel subset $B \subset X^n$. The (most elementary version of) Kolmogorov extension theorem says that

Theorem 7.4 (Kolmogorov extension theorem)**.** *Given a consistent family of probability measures as above, there exists a unique probability measure μ, defined on the Borel σ-algebra of Σ^+, such that*

$$\mu(C_B) = \mu_n(B)$$

for any cylinder C_B.

Proof. The proof consists in the following two steps. First, observe that cylinders form an algebra, and use consistency of the μ_n's to verify that the formula above does define a function $\mu : \mathcal{C} \to [0, 1]$ on cylinders (i.e. it does not depend on the different ways the same cylinder may be presented) which is additive and properly normalized. Then, use compactness of X to check that μ is continuous at $\emptyset$, in order to apply Carathéodory theorem. Indeed, let (A_n) be a sequence of cylinders such that $A_n \downarrow \emptyset$, and assume by contradiction that $\mu(A_n) > \delta > 0$ for any n. This implies that $A_n \neq \emptyset$ for any n, but, since the A_n are compact, then the Cantor intersection theorem says that $\cap_n A_n \neq \emptyset$, contrary to the hypothesis. □

Kolmogorov theorem is the key tool in probability theory, since it allows one to construct measures which describe an infinite sequence of trials starting with some rule which gives information about the n-th trial given the knowledge of the first $n-1$. It actually works with much more general spaces and in a more general setting. Also, one can easily adapt the construction to $\prod_{n\in\mathbb{N}} X_n$, the topological product of a countable family of finite spaces. In some precise sense, this is a universal model of a dynamical system.

e.g. **Bernoulli trials.** If $X = \{0, 1\}$, then $\Sigma^+ = X^{\mathbb{N}}$ is the state space of infinite Bernoulli trials with two possible outcomes: success and failure. Let $\mu_1 : \mathcal{P}(X) \to [0, 1]$ be a any probability measure, defined by $\mu_1(\{1\}) = p$. Kolmogorov construction can be applied postulating the independence of different trials, i.e. declaring that the family formed by the cylinders $\{x_n = 1\}$ is an independent family, and giving measure p to each $\{x_n = 1\}$. The resulting probability space $(\Sigma^+, \mathcal{B}, \mu)$ describes the infinite independent Bernoulli trials. Of course, the very same construction can be made when X is a finite space with any finite numer z of elements.

7.2 Transformations and invariant measures

Measurable transformations. A transformation $f : X \to X$ of the measurable space $(X, \mathcal{E})$ is said *measurable* if $f^{-1}(A) \in \mathcal{E}$ for any $A \in \mathcal{E}$. A measurable transformation f is said an *endomorphism* of the measurable space, or an *automorphism* if it is invertible and its inverse is measurable too.

Observe that an endomorphism f of a measurable space $(X, \mathcal{E})$ acts naturally on the space of measures on $\mathcal{E}$ by "push forward": if μ is a measure, then $f_*\mu$, defined by $(f_*\mu)(A) = \mu\left(f^{-1}(A)\right)$ for any $A \in \mathcal{E}$, is also a measure.

Let f be an endomorphism of the measurable space $(X, \mathcal{E})$. A probability measure μ on $\mathcal{E}$ is *invariant* (w.r.t. the transformation f) if $f_* \mu = \mu$, namely if

$$\mu\left(f^{-1}(A)\right) = \mu(A)$$

for any $A \in \mathcal{E}$. If this happens, we also say that f is an *endomrphism* (resp. an *automorphism*) of the probability space $(X, \mathcal{E}, \mu)$. The meaning of this definition is that "mean values" of integrable observables $\varphi : X \to \mathbb{R}$ with respect to invariant probability measures do not change with time, in the sense that $\int_X \varphi d\mu = \int_X (\varphi \circ f) d\mu$.

Given an endomorphism f of the probability space $(X, \mathcal{E}, \mu)$, one says that an event $A \in \mathcal{E}$ is *invariant mod* 0 if $\mu\left(A \bigtriangleup f^{-1}(A)\right) = 0$. The set of invariant mod 0 events form a sub-σ-algebra of $\mathcal{E}$, denoted by $\mathcal{E}_f$.

How to prove that a measure is invariant. The very definition of invariance does not help too much if we want to prove that a certain measure μ on the σ-algebra $\mathcal{E}$ is invariant w.r.t. the measurable transformation $f : X \to X$. The trick is the following. Suppose that we can prove that $\mu\left(f^{-1}(C)\right) = \mu(C)$ for any $C \in \mathcal{C}$, where $\mathcal{C}$ is some subset of $\mathcal{E}$. Caratheodory theorem implies that $f_* \mu$ and μ are the same measure on the σ-algebra $\sigma(\mathcal{C})$ generated by $\mathcal{C}$. On the other side, the family of those $A \in \mathcal{E}$ such that $\mu\left(f^{-1}(A)\right) = \mu(A)$ is easily seen to be a σ-algebra. Hence, if it happens that $\sigma(\mathcal{C}) = \mathcal{E}$, then μ is actually invariant. In other words, in order to prove that μ is invariant it is sufficient to check that $\mu\left(f^{-1}(C)\right) = \mu(C)$ for any C belonging to a family of subsets of X which generate the σ-algebra $\mathcal{E}$.

Observables as random variables. When dealing with a endomorphism $f : X \to X$ of the probability space $(X, \mathcal{E}, \mu)$, one should consider *measurable* observables $\varphi : X \to \mathbb{R}$ (or $\mathbb{C}$), those functions such that $\varphi^{-1}(A) \in \mathcal{E}$ for any Borel set $A \subset \mathbb{R}$. In the context of probability theory they are called "random variables", and the sequence of observables $\varphi \circ f^n$ may be interpreted as a "random process". If φ is integrable, the Lebesgue integral $\int_X \varphi d\mu$ is interpreted as the "mean value" of φ. Of course, invariance of a measurable observable must be intended modulo sets of zero measure. Then, one can consider the Banach spaces $L^p(\mu)$ of (equivalence classes of) observables equipped with the L^p-norm

$$\|\varphi\|_p = \left(\int |\varphi|^p \, d\mu\right)^{1/p}$$

and use the full power of integration theory to get informations about the dynamical system. In particular, $L^2(\mu)$ is a Hilbert space if equipped with the inner product

$$\langle \varphi, \psi \rangle = \int_X \varphi \overline{\psi} d\mu$$

Conditional mean. Recall that, given a measurable space $(X, \mathcal{E})$, a measure ν is said *absolutely continuous* w.r.t. the measure μ if $\nu(A) = 0$ whenever $\mu(A) = 0$. The following technical result (which may be proved using Hilbert space techniques) is particularly useful:

Theorem 7.5 (Radon-Nikodym). *Let $(X, \mathcal{E}, \mu)$ be a probability space, and let ν be a finite measure over $\mathcal{E}$ which is absolutely continuous with respect to μ. Then there exists a nonnegative integrable random variable ρ (called the Radon Nikodym derivative of ν w.r.t. μ and denoted by $d\nu/d\mu$) such that*

$$\nu(A) = \int_A \rho d\mu$$

for any $A \in \mathcal{E}$.

A particularly important tool, taken from the theory of probability, is the conditional mean. Let $(X, \mathcal{E}, \mu)$ be a probability space, and let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{E}$. Given an integrable random variable φ, there exists a unique random variable $\varphi_{\mathcal{F}}$, called the *conditional mean* of φ w.r.t. $\mathcal{F}$, which is $\mathcal{F}$-measurable (i.e. the inverse image of any Borel set belongs to $\mathcal{F}$) and such that

$$\int_A \varphi_{\mathcal{F}} d\mu = \int_A \varphi d\mu$$

for any $A \in \mathcal{F}$. Indeed, if $\varphi \geq 0$, then one can define $\varphi_{\mathcal{F}}$ as equal to the Radon-Nikodym derivative of the measure $A \mapsto \int_A \varphi d\mu$, defined on $\mathcal{F}$, with respect to the restriction $\mu|_{\mathcal{F}}$. The general case is treated by linearity, writing φ as a difference of two non-negative random variables. Uniqueness is intended μ-a.e., i.e. modulo sets of zero probability. The conditional mean is monotone, namely if $\varphi \geq 0$ then $\varphi_{\mathcal{F}} \geq 0$, and preserves the mean value, since $\int_X \varphi_{\mathcal{F}} d\mu = \int_X \varphi d\mu$. It can be considered as a "projection" of φ onto the space of $\mathcal{F}$-measurable random variable, preserving the mean value. In particular, if $\mathcal{N}$ is the trivial σ-algebra made of events of measure 0 or 1, then $\varphi_{\mathcal{N}}$ is constant a.e. and equal to $\int_X \varphi d\mu$.

Topological dynamical systems and Borel measures. If we are interested in the dynamics of a continuous transformation $f : X \to X$ of a topological space X, it is natural to consider the Borel σ-algebra $\mathcal{B}$, the smallest σ-algebra of subsets of X which contain all open sets. The map f is then an endomorphism of $(X, \mathcal{B})$. Probability measures on $\mathcal{B}$ are said *Borel probability measures*. If, moreover, X is a compact metric space, one can consider the space $\mathcal{C}^0(X, \mathbb{R})$ of bounded continuous real valued functions of X (observe that, since X is compact, any continuous function is automatically bounded), equipped with the sup norm

$$\|\varphi - \psi\|_\infty = \sup_{x \in X} |\varphi(x) - \psi(x)|$$

These observables are clearly integrable w.r.t. to any Borel probability measure μ, and the mean value map

$$\varphi \mapsto \int_X \varphi d\mu$$

is a bounded, positive definite (in the sense that $\int_X \varphi d\mu \geq 0$ for any $\varphi \geq 0$) linear functional on $\mathcal{C}^0(X, \mathbb{R})$. The basic fact about Borel measures is the converse of that, namely

Theorem 7.6 (Riesz-Markov representation theorem)**.** *Let X be a compact metric space. Given any bounded and positive definite linear functional L on $\mathcal{C}^0(X, \mathbb{R})$ such that $L(1) = 1$, there exists a unique Borel probability measures μ such that*

$$L(\varphi) = \int_X \varphi d\mu$$

for any $\varphi \in \mathcal{C}^0(X, \mathbb{R})$

The space of invariant probability measures. The space Prob of probability measures on a measurable space $(X, \mathcal{E})$ has a natural convex structure: convex combinations of probability measures are also probability measures. An arbitrary measurable transformation $f : X \to X$ of a measurable space may not admit any invariant probability measure. On the other side, if μ_0 and μ_1 are invariant probability measures, so are their convex combinations $\mu_t = (1-t)\mu_0 + t\mu_1$ for any $t \in [0, 1]$. This means that the set Prob_f of invariant probabilty measures on $\mathcal{E}$ is a convex set: if it contains two points, it contains the whole segment between them.

Now, let (X, d) be a compact metric space and let $\mathcal{B}$ its Borel σ-algebra. The space Prob of probability measures on $\mathcal{B}$ can be equipped with a natural topology, called the *weak* topology*, which says essentially that two measures are near if they give nearby mean values to some well behaved observables. Formally, one says that a sequence of measures (μ_n) converge weakly* to a measure μ, which we denote simply as $\mu_n \to \mu$, if

$$\int_X \varphi d\mu_n \to \int_X \varphi d\mu$$

for any (bounded) continuous function $\varphi : X \to \mathbb{R}$. The space $\mathcal{C}^0(X, \mathbb{R})$ of bounded continuous real valued functions on X, equipped with the sup norm, is a separable Banach space. In particular, it admits a countable set of points $\{\varphi_n\}_{n \in \mathbb{N}}$ which is dense in its unit sphere. Given that, one defines, for any couple of Borel probability measures μ and ν, a distance

$$d(\mu, \nu) = \sum_{n=1}^{\infty} 2^{-n} \cdot \left| \int_X \varphi_n d\mu - \int_X \varphi_n d\nu \right|$$

It turns out that d is indeed a metric, and that it induces the weak* topology on Prob. The important fact (somewhere called "Helly's theorem"), which follows from the Ascoli-Arzela theorem together with the above Riesz-Markov representation theorem, is that Prob, equipped with the weak* topology, is a compact space: any sequence (μ_n) of Borel probability measures admits a weakly* convergent subsequence $\mu_{n_i} \to \mu$.

Now, we are in position to prove the existence of invariant probability measures for certain well behaved dynamical systems.

Theorem 7.7 (Krylov-Bogolyubov). *A continuous transformation $f : X \to X$ of a metrizable compact space X admits at least one Borel invariant probability measure.*

Proof. Take any Borel probability measure μ_0 on X, and inductively define a family of probability measures μ_n by $\mu_{n+1} = f_* \mu_n$. Consider the family of Cesaro means

$$\overline{\mu}_n = \frac{1}{n+1} \sum_{k=0}^{n} \mu_k$$

Since the space of Borel probability measures on a compact metrizable space is compact w.r.t. weak* convergence, there exist a weakly* convergent subsequence $\overline{\mu}_{n_i} \to \mu$. One then easily sees that

$$\begin{aligned}
\int_X (\varphi \circ f)\, d\mu &= \lim_{i\to\infty} \frac{1}{n_i+1} \sum_{k=0}^{n_i} \int_X (\varphi \circ f)\, d\mu_k \\
&= \lim_{i\to\infty} \frac{1}{n_i+1} \sum_{k=0}^{n_i} \int_X \varphi d\mu_{k+1} \\
&= \lim_{i\to\infty} \frac{1}{n_i+1} \sum_{k=0}^{n_i} \int_X \varphi d\mu_k + \frac{1}{n_i+1} \left(\int_X \varphi d\mu_{n_i+1} - \int_X \varphi d\mu_0 \right) \\
&= \int_X \varphi d\mu
\end{aligned}$$

for any bounded continuous observable φ, hence that μ is an invariant measure. $\square$

7.3 Invariant measures and time averages

The relevance of invariant measures when studying the dynamics of continuous transformations is due to the following crucial observations.

Invariant measures and time averages. Assume that, for a given point $x \in X$, the time averages

$$\overline{\varphi}(x) = \lim_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} \varphi\left(f^k(x)\right)$$

do exist for any bounded continuos observable φ. One easily shows that the functional $\mathcal{C}_b^0(X, \mathbb{R}) \to \mathbb{R}$ defined by $\varphi \mapsto \overline{\varphi}(x)$ is linear, bounded and positive definite. There follows from the Riesz-Markov representation theorem that there exists a unique Borel probability measure μ_x on X such that

$$\overline{\varphi}(x) = \int_X \varphi\, d\mu_x$$

for any $\varphi \in \mathcal{C}_b^0(X, \mathbb{R})$. The invariance property $\overline{\varphi}(x) = (\overline{\varphi} \circ f)(x)$ for time averages then implies that $\int_X (\varphi \circ f)\, d\mu_x = \int_X \varphi d\mu_x$ for any φ, hence that μ_x is an invariant probability measure. In the language of physicists, this says that "time averages" along the orbit of x are equal to "space averages" with respect to the measure μ_x.

One is thus lead to consider the following questions. Do there exist points x for which time averages exists? Given an invariant measure μ, do there exist, and how many, points x such that $\mu = \mu_x$?

e.g. Periodic orbits. Here is a trivial but important example. Let p be a periodic point with period n. The time average $\overline{\varphi}(p)$ of any observable φ exists, and is equal to the arithmetic mean of φ along the orbit, namely

$$\overline{\varphi}(p) = \frac{1}{n}\sum_{k=0}^{n-1} \varphi(f^n(p))$$

If μ_p denotes the normalized sum $\frac{1}{n}\sum_{k=0}^{n-1}\delta_{f^n(p)}$ of Dirac masses placed on the orbit of p, this amount to say that $\overline{\varphi}(p) = \int_X \varphi d\mu_p$.

Let p be a fixed point, and $\varphi : X \to \mathbb{R}$ be an observable which is continuous at p. If $x \in W^s(p)$, then the time average $\overline{\varphi}(x)$ exists and is equal to $\varphi(p)$, i.e. time averages of points in the basin of attraction of p are described by the Dirac measure $\mu_p = \delta_p$.

The Birkhoff-Khinchin ergodic theorem. Ergodic theorems are the milestones of ergodic theory, and deal with various type of convergence of the time means $\overline{\varphi}_n$ for certain classes of observables φ. In particular, the Birkhoff-Khinchin ergodic theorem must be thougth as the generalization of the Kolmogorov strong law of large numbers, as it says that time means of certain well-behaved observables exist almost everywhere. The Birkhoff-Khinchin ergodic theorem was actually preceeded by the von Neumann's "statistic" ergodic theorem, which says that

Theorem 7.8 (von Neumann "statistic" ergodic theorem)**.** *Let U be a unitary operator on a Hilbert space H, let $H_U = \{v \in H \text{ s.t. } Uv = v\}$ denote the closed subspace of those vectors which are fixed by U, and $P_U : H \to H_U$ denote the orthogonal projection onto H_U. Then, for any vector $v \in H$ we have*

$$\lim_{n\to\infty}\left\|\frac{1}{n+1}\sum_{k=0}^{n} U^k v - P_U v\right\|_H = 0$$

If $f : X \to X$ is an endomorphism of the probability space $(X, \mathcal{E}, \mu)$, one can consider the "shift" operator $U : L^2(\mu) \to L^2(\mu)$ given by $(U\varphi)(x) = \varphi(f(x))$. It is clearly unitary, its fixed point set is the space of invariant L^2-observable. The von Neumann theorem then asserts convergence of time means $\overline{\varphi}_n \to \overline{\varphi}$ in $L^2(\mu)$. Here, we prove the

Theorem 7.9 (Birkhoff-Khinchin "individual" ergodic theorem)**.** *Let $f : X \to X$ be an endomorphism of the probability space $(X, \mathcal{E}, \mu)$, and let $\varphi \in L^1(\mu)$ be an integrable observable. Then the limit*

$$\overline{\varphi}(x) = \lim_{n\to\infty}\frac{1}{n+1}\sum_{k=0}^{n}\varphi\left(f^k(x)\right)$$

exists for μ-almost any $x \in X$. Moreover, the observable $\overline{\varphi}$ is in $L^1(\mu)$, is invariant, and satisfies

$$\int \overline{\varphi} d\mu = \int \varphi d\mu$$

Proof. (by A. Garsia, as explained in [KH95]) Let $\mathcal{E}_f$ be the invariant σ-algebra. For any $\psi \in L^1$, set $\psi_n = \max_{k\le n}\sum_{k=0}^{n-1}\varphi\circ f^k$ and observe that $E_\psi = \{x \in X \text{ s.t. } \psi_n(x) \to \infty\} \in \mathcal{E}_f$. One easily sees that the sequence $\psi_{n+1} - \psi_n \circ f$ is decreasing, and converges to ψ at the points of E_ψ. The monotone convergence theorem and the invariance of μ imply that

$$0 \le \int_{E_\psi}(\psi_{n+1} - \psi_n)\, d\mu = \int_{E_\psi}(\psi_{n+1} - \psi_n \circ f)\, d\mu \to \int_{E_\psi}\psi d\mu = \int_{E_\psi}\psi_{\mathcal{E}_f} d\mu\,|_{\mathcal{E}_f}$$

In particular, if $\psi_{\mathcal{E}_f} < -\varepsilon < 0$ then $\mu(E_\psi) = 0$. On the other side,

$$\limsup \frac{1}{n}\sum_{k=0}^{n-1}\psi\circ f^k(x) \le \limsup \frac{1}{n}\psi_n \le 0$$

on $X \backslash E_{\psi}$. Applying twice these observations to the observables $\varphi - \varphi_{\mathcal{E}_f} - \varepsilon$ and $-\varphi + \varphi_{\mathcal{E}_f} - \varepsilon$, with $\varepsilon > 0$, we find

$$\limsup \frac{1}{n} \sum_{k=0}^{n-1} \varphi \circ f^k(x) - \varphi_{\mathcal{E}_f} - \varepsilon \leq 0 \qquad \liminf \frac{1}{n} \sum_{k=0}^{n-1} \varphi \circ f^k(x) - \varphi_{\mathcal{E}_f} + \varepsilon \geq 0$$

μ-almost everywhere. Since ε was arbitrary, the limit $\overline{\varphi}(x)$ exists and is equal to $\varphi_{\mathcal{E}_f}(x)$ for μ-almost every x. The rest of the theorem then follows easily from the properties of the conditional mean. □

7.4 Examples of invariant measures

Haar measures. Any locally compact topological group G admits a *Haar measure*, a measure μ on its Borel sets which is left-invariant, i.e. satisfies $L_g \mu = \mu$ for any $g \in G$. Moreover, the Haar measure is unique up to a constant factor. It is an exercise that μ is a finite measure, hence can be renormalized to give a probability measure, iff G is compact. There follows that translations on compact topological groups admits invariant probability measures.

On the other side, for some groups G, called *unimodular*, the Haar measure μ is both left and right invariant. If $\Gamma \subset G$ is a lattice, i.e. a subgroup such that $\mu(G/\Gamma) < \infty$, then the normalized Haar measure on the homogeneous space G/Γ is an invariant probability measure for any left translation $g\Gamma \mapsto sg\Gamma$.

Rotations of the circle. Lebesgue probability measure ℓ on the circle is invariant for the rotations $R_\alpha : x + \mathbb{Z} \mapsto x + \alpha + \mathbb{Z}$, with $\alpha \in \mathbb{R}$. Indeed, rotations of the circle are isometries, and the Lebesgue measure $\ell(I)$ of an interval is its "lenght".

Coverings of the circle. Lebesgue probability measure ℓ on the circle is invariant for the maps $f : x + \mathbb{Z} \mapsto \lambda \cdot x + \mathbb{Z}$, with $\lambda \in \mathbb{Z} \backslash \{0\}$. This comes from the fact that the inverse image of a sufficiently small interval I with lenght $\ell(I)$ is the disjoint union of $|\lambda|$ intervals with lenght $\ell(I) / |\lambda|$.

Bernoulli shifts. Consider the Bernoulli shift $\sigma : \Sigma^+ \to \Sigma^+$ over the alphabet $X = \{1, 2, ..., z\}$. Let p be a 'probability on X", i.e. a finite set of nonegative numbers $p_1, p_2, ..., p_z$ such that $p_1 + p_2 + ... + p_z = 1$. Given a centered cylinder C_α, we define $\mu(C_\alpha)$ as equal to the product $p_{\alpha_1} p_{\alpha_2} ... p_{\alpha_n}$. This function μ extends in a unique way as a finitely additive function on the algebra $\mathcal{A}$ generated by the centered cylinders, the algebra which contains all finite unions of centered cylinders as well as the empty set and Σ^+. One then show that μ is σ-additive on $\mathcal{A}$ (for example, showing that if a decreasing sequence $A_1 \supset A_2 \supset ...$ has empty intersection then $\mu(A_n) \to 0$). Since centered cylinders generates the topology of Σ^+, Carathéodory theorem implies that there exists a unique extension, which we still call μ, of this measure on the Borel σ-algebra of Σ^+. This measure is called the *Bernoulli measure* defined by p.

As for the "physical" meaning of this measure, you may imagine that X represents the possible outcomes when tossing a coin with z sides, and p_k is the probability of obtaining the k-th side. Then points in Σ^+ represent the outcomes of an infinite sequence of tossings, and the very definition of μ says that each trial is described by the probability p, and each trial is "independent" from any finite collection of different trials.

It is not surprising that μ is indeed an invariant probability measure. This comes from the fact that the inverse image $\sigma^{-1}(A)$ of any $A \in \mathcal{A}$ is the disjoint union of z elements $B_1, B_2, ..., B_z$ of the algebra (obtained from A chosing the first letter in z different ways) with measures $\mu(B_k) = p_k \cdot \mu(A)$, so that

$$\mu\left(\sigma^{-1}(A)\right) = \sum_{k=1}^{z} p_k \cdot \mu(A) = \mu(A)$$

Absolutely continuous invariant measures for maps and flows. Let U be a domain in some euclidean $\mathbb{R}^n$, and let ℓ denote the Lebesgue measure on U, given locally as

$$d\ell = dx = dx_1 dx_2 ... dx_n \, .$$

Thus, the volume of an open set $A \subset U$ is $\ell(A) = \int_A dx$. A local diffeomorphism $f : U \to U$ of class $\mathcal{C}^1$ preserves the measure vol iff

$$\sum_{x \in f^{-1}\{x'\}} \frac{1}{|\det f'(x)|} = 1$$

for any point $x' \in U$, as one can check using the change of coordinates formula. Also interesting is to see wheather f preserves an absolutely continuous measure $d\mu = \rho d\ell$, definied by $\mu(A) = \int_A \rho \, dx$. This happens iff the "density" ρ satisfies the equation

$$\sum_{x \in f^{-1}\{x'\}} \frac{\rho(x)}{|\det f'(x)|} = \rho(x')$$

for any point $x' \in U$.

Now, let Φ_t be the flow of a vector field $v = \sum_{k=1}^n v_k \frac{\partial}{\partial x_k}$ on U. The above obviously applies, considering the Jacobian of the diffeomorphisms Φ_t. Since

$$\det \Phi_t' = \int_0^t \mathrm{div} v \circ \Phi_s \, ds$$

we get the result that Lebesgue measure ℓ is invariant under the flow of v iff

$$\mathrm{div} v = \sum_{k=1}^n \frac{\partial v_k}{\partial x_k} = 0$$

In general, the absolutely continuous measure $d\mu = \rho d\ell$ is invariant under the flow of v iff its density satisfies $\mathrm{div}\,(\rho v) = 0$.

Hamiltonian flows. Consider a symplectic manifold (X, ω). Liouville measure $d\mu = \omega^n$ is invariant under the Hamiltonian flow of any Hamiltonian function H. If X has finite volume, it can be normalized to give an invariant probability measure.

Geodesic flows. Consider a geodesic flow on the unit tangent bundle $\pi : SM \to M$ of the Riemannian manifold (M, g). Let $d\mu = \sqrt{g}\, dx$ denote the Riemannian volume form on M, and let $d\sigma_m$ denotes the Lebesgue probability measure on the sphere $S_m M = \pi^{-1}\{m\}$. The Liouville measure ℓ, defined locally as $d\mu(m) \times d\sigma_m$, is invariant under the geodesic flow.

Gauss map. Any irrational real number $x \in (0, 1]$ has a unique continued fraction representation of the form

$$x = [0; a_1, a_2, a_3, \dots] = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{\ddots}}}}$$

where the a_n are nonnegative integers. The equality sign and the "infinite fraction" above mean that the sequence of finite continued fractions

$$p_n/q_n = [0; a_1, a_2, \dots, a_n] = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{a_n}}}}$$

which are called “convergents”, do converge to x as $n \to \infty$. The sequence of partial quotients a_n is inductively constructed as follows. First, observe that if $a_1 = [1/x]$ and $x_1 = 1/x - a_1$ we may write

$$x = \frac{1}{a_1 + x_1}$$

with $x_1 \in [0,1]$. Then, since $x_1 \neq 0$, for otherwise x would be rational, we may define $a_2 = [1/x_1]$ and $x_2 = 1/x_1 - a_2$ to get

$$x = \cfrac{1}{a_1 + \cfrac{1}{a_2 + x_2}}$$

Inductively, we see that

$$x = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{... + \cfrac{1}{a_n + x_n}}}}$$

where $x_n = 1/x_{n-1} - a_n$ and $a_n = [1/x_{n-1}]$. This amounts to say that the sequence (x_n) is the trajectory of x under the *Gauss map* $G :]0,1] \to]0,1]$, defined as

$$x \mapsto 1/x - [1/x]$$

Observe that G is not defined at the origin, hence to iterate G we need to avoid all the preimages of 0, which are the rationals. This is not a problem if we want to study the statistical properties of G with respect to Lebesgue measure, since rationals form a subset of zero measure. The Gauss map admits an absolutely continuous invariant measure $\mu = \rho dx$, defined as

$$\mu(A) = \frac{1}{\log 2} \cdot \int_A \frac{1}{1+x} dx$$

for any Borel subset $A \subset]0,1]$. The denominator $\log 2$ is there to normalize the measure, so we just have to check the invariance criterium for the density $\rho(x) = 1/(1+x)$. Since any $x' \in]0,1]$ has one preimage $x_k = 1/(x'+k)$ in each interval $]1/(k+1), 1/k]$, we compute

$$\begin{aligned} \sum_{x \in G^{-1}\{x'\}} \frac{\rho(x)}{|G'(x)|} &= \sum_{k \geq 1} \frac{x_k^2}{1+x_k} \\ &= \sum_{k \geq 1} \left(\frac{1}{x'+k} - \frac{1}{x'+k+1} \right) \\ &= \frac{1}{1+x'} = \rho(x') \end{aligned}$$

and we are done.

8 Recurrences

8.1 Limit sets and recurrent points

Omega and alpha limit sets. Let $f : X \to X$ be a continuous transformation of a topological aspace X. The simplest thing that an infinite (i.e. not periodic) orbit can do is to be (the image of a) convergent (trajectory). In this case, as we already know, the limit must be a fixed point of the transformation.

Trajectories, even when not convergent, may have at least convergent subsequences. The *ω-limit set* of a point $x \in X$ is the set

$$\omega_f(x) := \bigcap_{n=0}^{\infty} \overline{\bigcup_{k \geq n} \{f^k(x)\}}$$

that is, the set of those points $x' \in X$ such that there exists a sequence of times $n_i \to \infty$ (i.e. an increasing map $i \mapsto n_i$) such that $f^{n_i}(x) \to x'$ when $i \to \infty$. Observe that, if the orbit of x is not finite (i.e. if x is not periodic), then the ω-limit set of x is the derived set of its forward orbit, i.e. $\omega_f(x) = \mathcal{O}_f^+(x)'$. It is clear that $\omega_f(x)$ is a closed (possibly empty) and +invariant subset of X.

$\mathrm{Lim}_f = \cup_{x \in X} \omega_f\,(x)$ denotes the set of ω-limit points of all the $x \in X$. If x is periodic, then $\omega_f\,(x)$ coincides with its orbit. There follows that

$$\mathrm{Per}_f \subset \mathrm{Lim}_f\,.$$

If f is invertible, we may also define the *α-limit set* of $x \in X$ as $\alpha_f(x) := \omega_{f^{-1}}(x)$, i.e. the set of points $x' \in X$ such that there exists a sequence of times $n_i \to \infty$ such that $f^{-n_i}(x) \to x'$ when $i \to \infty$. In this case, both $\omega_f(x)$ and $\alpha_f(x)$ are closed and invariant subsets of X. $\mathrm{Lim}_{f^{-1}} = \cup_{x \in X} \alpha_f\,(x)$ denotes the set of all α-limit points of an invertible transformation f.

Limit sets in compact spaces. Both the ω and the α-limit sets of a generic point can be empty. For example, all the limit points for the translation $f(x) = x + 1$ of the real line are empty.

This may happens, of course, only if the phase space X is not compact. Indeed, if X is compact, then the trajectory of any point admits convergent subsequences (by sequencial compactness, which holds for compact metric spaces), and therefore $\omega_f\,(x) \neq \emptyset$ for all $x \in X$. For the same reason, if f is a homeomorphism of a compact metic space, $\alpha_f\,(x) \neq \emptyset$ for all points $x \in X$. In particular, the sets $\mathrm{Lim}_{f^{\pm 1}}$ are not empty.

ex: Show that $\omega_f(x)$ is closed and +invariant. Show that if f is a homeomorfism, then $\omega_f(x)$ and $\alpha_f(x)$ are closed and invariant.

ex: Give examples such that $\omega_f(x)$ and $\alpha_f(x)$ are empty.

ex: Show that $\mathrm{Per}_f \subset \mathrm{Lim}_f$.

Recurrent points. Let $f : X \to X$ be a topological dynamical system. The point $x \in X$ is *recurrent* if $x \in \omega_f(x)$. It is clear that this is equivalent to asking that given any neighborhood B of x there exists a time $n \geq 1$ such that $f^n\,(x) \in B$. Indeed, chosing smaller neighborhoods (if $f^n(x) \neq x$, so that x is not already periodic with period n), this also implies that the trajectory of x passes infinitely often in a any such neighborood, i.e. $f^n(x) \in B$ for infinitely many times $n \geq 1$.

Rec_f denotes the set of recurrent points for f. A periodic point is obviously recurrent, therefore

$$\mathrm{Per}_f \subset \mathrm{Rec}_f\,.$$

If f is a homeomorphism, it also makes sense to consider the set $\mathrm{Rec}_{f^{-1}}$, the set of those points $x \in X$ such that $x \in \alpha_f(x)$.

ex: Define a partial order in X as follows: $x \prec x'$ if for any neighborhood U of x and V of x' there exists a time $n \geq 1$ such that $f^n\,(U) \cap V \neq \emptyset$. Show that x is recurrent iff $x \prec x$.

ex: Show that $\mathrm{Per}_f \subset \mathrm{Rec}_f$.

ex: Give examples which show that both Rec_f and $\mathrm{Rec}_{f^{-1}}$ may be empty.

Non-wandering set. The point x is *wandering*[24] if it admits a neighborood which is disjoint from all its iterates, i.e. if there exists an open set U containing x such that $U \cap f^n(U) = \emptyset$ for all times $n \geq 1$. Conversely, the point x is not wandering is for any neighborhood U of x there exists a time $n \geq 1$ such that $f^n(U) \cap U \neq \emptyset$.

The *non-wandering set* NW_f is the set of those points x which are not wandering. This is the where the interesting dynamics takes place, since other points are "forgotten" as time passes.

The non-wandering set NW_f is closed (the set of wandering points is open by definition, since any point in a sufficiently small open neighborhood of a wandering point is itself wandering) and +invariant. It contains the ω-limit points of all points in X, as well as the recurring points. Thus, the inclusions are

$$\mathrm{Per}_f \subset \mathrm{Lim}_f \subset \mathrm{NW}_f \qquad \text{e} \qquad \mathrm{Per}_f \subset \mathrm{Rec}_f \subset \mathrm{NW}_f$$

If f is an homeomorphism, NW_f, which is equal to $\mathrm{NW}_{f^{-1}}$, is also invariant, and contains the ω- and α-limits of all points of X.

If X is compact, then $\mathrm{NW}_f \neq \emptyset$, since any point $x \in X$ have $\omega_f(x) \neq \emptyset$ and $\mathrm{Lim}_f \subset \mathrm{NW}_f$.

ex: Show that the non-wandering set of a homeomorphism is closed, invariant, and contains the ω and α-limit sets.

ex: Show that if f is a homeomorphism, then $\mathrm{NW}_f = \mathrm{NW}_{f^{-1}}$.

ex: Give examples that show that NW_f may be empty.

ex: Show that $\mathrm{Per}_f \subset \mathrm{Rec}_f \subset \mathrm{NW}_f \subset \mathrm{Rec}^{\varepsilon}_f$. Give exemple that show that these inclusions may be strict.

ex: Show that $\mathrm{Per}_f \subset \mathrm{Rec}_f \subset \mathrm{NW}_f$ and therefore $\overline{\mathrm{Per}_f} \subset \overline{\mathrm{Rec}_f} \subset \mathrm{NW}_f$. More difficult is to find example showing that these inclusion may be stricd.

ex: Find the non-wandering sets of linear maps of the plane.

8.2 Dirichlet theorem on Diophantine approximation

Rotations of the circle and Dirichlet theorem on Diophantine approximation. Consider a rotation $R_\alpha : x + \mathbb{Z} \mapsto x + \alpha + \mathbb{Z}$ of the circle $\mathbb{T} = \mathbb{R}/\mathbb{Z}$. If α is rational, all points are trivially recurrent, being periodic. When α is irrational, recurrence of a point $x + \mathbb{Z}$ means that for any $\varepsilon > 0$ there exist an infinity of times $q \in \mathbb{N}$ such that that

$$d(x + \mathbb{Z}, x + q\alpha + \mathbb{Z}) < \varepsilon$$

or, equivalently, that for any $\varepsilon > 0$ there exist an infinity of rationals p/q such that

$$|q\alpha - p| < \varepsilon \qquad \text{i.e.} \qquad \left|\alpha - \frac{p}{q}\right| < \frac{\varepsilon}{q}$$

Indeed, much more is true, and is a consequence of the following classical result by Dirichlet on Diophantine approximation (see [HW59]).

[24] The greek word for "wandering" was $\pi\lambda\alpha\nu\eta\tau\eta\varsigma$, i.e. *planet*.

Theorem 8.1 (Dirichlet, 1842). *For any number θ and any positive integer $Q \in \mathbb{N}$ there exist $p \in \mathbb{Z}$ and $q \in \mathbb{Z}\backslash\{0\}$ such that*

$$|q\theta - p| < 1/Q \qquad \text{and} \qquad |q| \leq Q\,. \tag{8.1}$$

and, a fortiori,

$$|\theta - p/q| < 1/q^2\,. \tag{8.2}$$

Proof. Divide the unit interval $[0,1)$ into the Q subintervals

$$[0,1/Q)\,, \quad [1/Q,2/Q)\,, \quad [2/Q,3/Q)\,, \quad \dots\,, \quad [(Q-1)/Q,1)$$

of equal length $1/Q$, and consider the $Q+1$ points[25]

$$\{0\}\,, \quad \{\theta\}\,, \quad \{2\theta\}\,, \quad \dots\,, \quad \{Q\theta\}$$

inside $[0,1)$. By the box principle (which Dirichlet stated to prove this theorem!), at least two of those points, say $\{k\theta\}$ and $\{k'\theta\}$ with $k > k'$, belong to the same subinterval. Therefore, there exist integers a, a' such that $|k\theta - a - (k'\theta - a')| < 1/Q$. The theorem follows taking $q = k - k'$ and $p = a - a'$, and observing that $q \leq Q$. □

For rational θ, there are only finitely many integers q and p satisfying the above inequalities (8.1). Indeed, if $\theta = a/b$ and $p/q \neq a/b$ (we may assume that both are reduced fractions), then

$$|q\theta - p| = \frac{|qa - pb|}{|b|} \geq \frac{1}{|b|}$$

(because the numerator is the absolute value of a non-zero integer) and therefore no fraction different from a/b may satisfy the inequalities (8.1) if Q is larger that $|b|$.

On the other hand, if θ is irrational and p_1/q_1, p_2/q_2, $\dots$, p_n/q_n are any finite number of fractions satisfying (8.2), we may consider an integer Q larger than the inverse of

$$\varepsilon = \min_{1\leq k\leq n} |q_k\theta - p_k| > 0$$

and produce, by theorem 8.1, one more fraction p/q satisfying (8.2). Thus,

Theorem 8.2 (Dirichlet, 1842). *For any irrational number θ there exist infinitely many reduced fractions p/q such that*

$$|\theta - p/q| < 1/q^2\,.$$

In particular, any point $x + \mathbb{Z}$ is recurrent for an irrational rotation of the circle.

8.3 Poincaré recurrence theorem

If f satisfies a condition (natural in physics) like "preserving a probability measure", then there are a lot of recurrent points, actually almost any point is recurrent. If, moreover, the probability measure is diffuse, i.e. any non-empty open set has positive measure, then the set of recurrent points is also dense. These results, discovered by Henri Poincaré around 1890, motivated the modern theory of dynamical systems. They show how weak informations on the transformation f (or the flow of a differential equation) may yeld significative qualitative information about "almost all" orbits of the system. Here follow the precise statements, together with all the necessary technical details. If you don't know the meaning of some words, like "measurable" or "Borel set", don't worry, just try to understand what's going on. Poincaré himself didn't know, yet!

You may look at this wonderful lecture on Poincaré recurrence theorem by Etienne Ghys in YouTube: https://www.youtube.com/watch?v=21fHNMccrY8#t=1741

[25] As usual, $\{x\}$ denotes the "fractional part" of x, so that any real number may be written as a sum $x = [x] + \{x\}$ for some unique $[x] \in \mathbb{Z}$ and $\{x\} \in [0,1)$.

Theorem 8.3 (Poincaré recurrence theorem)**.** *Let $f : X \to X$ be an endomorphism of a probability space $(X, \mathcal{E}, \mu)$, and let $A \in \mathcal{E}$. Then the set*

$$A^{rec} = \{x \in A \text{ t.q. } f^n(x) \in_{i.o.} A\}$$

of those points of A whose orbit passes through A infinitely often has total probability, namely $\mu(A^{rec}) = \mu(A)$

Proof. For $k \geq 1$, let

$$B_k = \{x \in A \text{ s.t. } f^n(x) \notin A \; \forall \, n \geq k\}$$

be the set of those points of A which never return in A after $n \geq k$ iterates. Observe that $B_k = A \cap (\cap_{n \geq k} f^{-n}(X \backslash A))$ and that $A^{\text{rec}} = A \backslash (\cup_{k \geq 1} B_k)$. In particular, this shows that A^{rec} is measurable. It is clear that $f^{-nk}(B_k) \cap B_k = \emptyset$ for any $n \geq 1$, since a point in the intersection would be a point $x \in B_k \subset A$ such that $f^{kn}(x) \in A$, and $kn \geq k$, contradicting the definition of B_k. For the same reason, $f^{-nk}(B_k) \cap f^{-mk}(B_k) = \emptyset$ for any $n > m \geq 0$. Therefore, the sets $f^{-nk}(B_k)$, for $n \in \mathbb{N}$, are pairwise disjoint. They also have all the same measure $\mu\left(f^{-nk}(B_k)\right) = \mu(B_k)$, because μ is invariant. This implies that $\mu(B_k) = 0$, because

$$\sum_{n \geq 1} \mu(B_k) = \sum_{n \geq 1} \mu\left(f^{-nk}(B_k)\right) = \mu\left(\cup_{n \geq 1} f^{-nk}(B_k)\right) \leq \mu(X) = 1\,.$$

There follows that $\mu(A^{\text{rec}}) = \mu(A)$. □

Now, let $f : X \to X$ be a continuous transformation of a metrizable topological space X, and let μ be an invariant Borel probability measure. Assume that (the topology of) X admits a countable basis $(U_i)_{i \in \mathbb{N}}$. We can apply the above theorem 8.3 to every open set U_i, and this easily implies that the set of recurrent points has full measure, i.e.

$$\mu(\text{Rec}_f) = 1\,.$$

In particular, since any set of full measure is dense in the support of a Borel measure, we get the following general result.

Theorem 8.4 (topologic Poincaré recurrence theorem)**.** *Let $f : X \to X$ be a continuous transformation of a separable metrizable topological space X. The support of any invariant Borel probability measure μ is contained in the closure of the set of recurrent points, namely*

$$\text{supp}(\mu) \subset \overline{\text{Rec}_f}\,.$$

In particular, if f admits an invariant measure μ which is diffuse (i.e. gives positive measure to any nonempty open set) then the set of recurrent points is dense in X, namely

$$\overline{\text{Rec}_f} = X\,.$$

Observe that if f is a homeomorphism, then the same applies to $\text{Rec}_{f^{-1}}$, and the support of any invariant Borel probability measure is contained in the closure of $\text{Rec}_f \cap \text{Rec}_{f^{-1}}$.

If you don't like the above proof, here is another, perhaps more elementary, of the last statement.

Proof. (of the last statement of theorem 8.4) Assume that the continuous map $f : X \to X$ preserves a diffuse Borel probability measure μ. For each $n \geq 1$, let

$$R_n := \left\{x \in X \text{ s.t. } \exists \, k \geq 1 \text{ s.t. } d(f^k(x), x) < 1/n\right\}$$

be the set of "$1/n$-recurrent" points. It is plain that $\text{Rec}_f = \cap_{n=1}^{\infty} R_n$. The sets R_n are clearly open. To show that Rec_f is dense we must therefore show that each R_n is dense, since then the Baire theorem implies that also their countable intersection is dense. So, take any nonempty ball $B = B_r(p)$ with diameter $2r < 1/n$. Its inverse images $f^{-1}(B)$, $f^{-2}(B)$, $f^{-3}(B)$, … have all the same measure by invariance, which is positive, i.e. $\mu(B) > 0$ (because the measure μ is diffuse). Since $\mu(X) = 1$, the $f^{-n}(B)$, for $n = 0, 1, 2, 3, \ldots$, cannot be pairwise disjoint. There follows that there exist $k > 0$ and $n \geq 0$ such that $f^{-(n+k)}(B) \cap f^{-n}(B) \neq 0$, and this implies that B contains a $1/n$-recurrent point (for a point x in the intersection has both images $f^n(x)$ and $f^{n+k}(x) = f^k(f^n(x))$ in B, hence at distance $< 1/n$). Since B was arbitrary, this proves that each R_n is dense, and Baire theorem implies that Rec_f is dense too. □

8.4 Transitivity and minimality

Transitive transformations. Let X be a complete and separable metric space. A continuous transformation $f : X \to X$ is *(topologically) +transitive* if it satisfies one of the following equivalent conditions:

i) for any two not-empty open sets $U, V \subset X$ there exist a time $n \geq 0$ such that $f^n(U) \cap V \neq \emptyset$

ii) there exists a point $x \in X$ such that $\omega_f(x) = X$

iii) there exists a residual subset $R \subset X$ of points x such that $\omega_f(x) = X$

Proof. (of the equivalence) The implications iii) ⇒ ii) ⇒ i) are obvious, since, if $\omega_f(x) = X$, then the trajectory of x visits infinitely often all non-empty open subsets of X. To show that i) ⇒ iii), the fist observation is that condition i) amounts to say that, for all not-empty open set V, its orbit $\bigcup_{n\geq 0} f^{-n}(V)$ is dense, and, moreover, its orbits $\bigcup_{n\geq k} f^{-n}(V) = \bigcup_{n\geq 0} f^{-n}\left(f^{-k}(V)\right)$ are also dense for all $k \geq 0$. Let $(U_i)_{i\in\mathbb{N}}$ be a countable basis for the toplogy f X. The family of $\bigcup_{n\geq k} f^{-n}(U_i)$, with $k \geq 0$ eand $i \geq 1$, is a family of open and dense subsets of X. Its countable intersection $R = \bigcap_{i\in\mathbb{N}} \bigcap_{k\geq 0} \bigcup_{n\geq k} f^{-n}(U_i)$ is a residual set, and a point $x \in R$ has a trajectory which visits infinitely often all the open sets U_i, i.e. $\omega_f(x) = X$. □

Also clear is that i) implies that X does not have isolated points (unless it has finite cardinality, trivial case in which X is a single orbit). This, in turns, implies that $\mathcal{O}_f^+(x)' = X$ if $x \in R$.

ex: Prove the implications iii) ⇒ ii) ⇒ i) above.

ex: If $f : X \to X$ is +transitive, then it snon-wandering set is $\mathrm{NW}_f = X$, since NW_f contains the ω-limit sets of points $x \in X$.

ex: If $f : X \to X$ is +transitive, then Rec_f is a residual set (observe that if $\omega_f(x) = X$ then $x \in \omega_f(x)$).

Transitive hoemomorphisms. There exists a weaker notion, only meaningful for invertible transformations. A homeomorphism $f : X \to X$ is *(topologically) transitive* if it satisfies one of the following three conditions:

i) for any two not-empty open sets $U, V \subset X$ there exists a time $n \in \mathbb{Z}$ such that $f^n(U) \cap V \neq \emptyset$

ii) there exists a point $x \in X$ with dense orbit, i.e. such that $\overline{\mathcal{O}_f(x)} = X$

iii) there exists a residual set of points $x \in X$ with dense orbits, i.e. such that $\overline{\mathcal{O}_f(x)} = X$

Proof. (of the equivalence) The implications iii) ⇒ ii) ⇒ i) are obvious, since if the complete orbit $\mathcal{O}_f(x)$ of x is dense, it visits at least once each not-empty open subset of X. To show that i) ⇒ iii), we first observe that condition i) is equivalent to say that the orbit $\bigcup_{n\in\mathbb{Z}} f^n(V)$ of any not-empty open set V is dense in X. Let $(U_i)_{i\in\mathbb{N}}$ be a countable basis for the topology of X . The family $U_i^{\pm} = \bigcup_{n\in\mathbb{Z}} f^n(U_i)$ is therefore a family of dense and open sets. Its countable intersection $R = \bigcap_{i\in\mathbb{N}} U_i^{\pm}$ is a residual set, and the complete orbit of any point $x \in R$ visits at first once any of the open sets U_i. Therefore, if $x \in R$ then $\overline{\mathcal{O}_f(x)} = X$. □

Observe that $f : X \to X$ is a transitive homeomorphism iff f^{-1} is a transitive homeomorphism. A transitive homeomorphsms need not be also +transitive. Indeed, it may have no recurrent points and an empty non-wandering set, provided X is not compact!

It is interesting to observe that transitivity is a kind of "dynamical connectedness", in the following precise sense.

Theorem 8.5. *A homeomorphism $f : X \to X$ is transitive iff X does not contain the disjoit union of two open invariant not-empty sets.*

Proof. The implication $\Rightarrow$ is obvious. To show the reverse implication $\Leftarrow$, observe that, if $U, V \subset X$ ore two not-empty open setsthen $U^{\pm} = \bigcup_{n\in\mathbb{Z}} f^n(U)$ and $V^{\pm} = \bigcup_{n\in\mathbb{Z}} f^n(V)$ are open, not-empty invariant sets. If $U^{\pm} \cap V^{\pm} \neq \emptyset$, there exist times $n, m \in \mathbb{Z}$ such that $f^n(U) \cap f^m(V) \neq \emptyset$, which implies $f^{n-m}(U) \cap V \neq \emptyset$. □

A consequence is the following useful criterium to decide when a homeomorphism cannot be transitive.

Theorem 8.6. *If $f : X \to X$ is a transitive homeomorphism, then all continuous invariant observable $\varphi : X \to \mathbb{R}$ is constant.*

Proof. Indeed, if φ is not constant, then it takes at least two values, say $a < b$. But the, if $c = (a + b)/2$, both $U = \{\varphi < c\}$ and $V = \{\varphi > c\}$ are invariant open disjoint and not-empty sets. □

ex: Give examples of homeomorphisms $f : X \to X$ which are transitive but not +transitive.

ex: Show that a homeomorphism $f : X \to X$ is +transitive iff is transitive and its non-wandering set is the whole X (the implication $\Leftarrow$ is obvious).

ex: It may happen that a transformation $f : X \to X$ is +transitive but some iterate f^n, with $n > 1$, is not. A trivial example is a permutation of a finite set, since some iterate is the identity transformation. In general, if X is compact, what happens is the following. There exist some finite covering $X = X_1 \cup X_2 \cup ... \cup X_k$, where k divides n and the 's a compact sets with nohere dense intersections $X_i \cap X_j$ if $i \neq j$, suah that $f(X_i) = X_{i+1 \bmod k}$ and the restrictions $f^n|_{X_i}$ are +transitive. The idea is to chose a point $x \in X$ such that $\omega_f(x) = X$, and then define $X_i = \omega_{f^n}\left(f^i(x)\right) \ldots$

Minimal sets. Let $f : X \to X$ be a continuous transformation. A closed not-empty $K \subset X$ is *minimal* if it is +invariant and if it does not contain proper closed +invariant subsets.

The orbit of a periodic point is an example of a minimal set.

If K is minimal, then the orbit of any $x \in K$ is dense in K, for otherwise its closure $\overline{\mathcal{O}_f^+(x)}$ would be a proper +invariant closed subset of K. This implies that $x \in \omega_f(x)$, and therefore that all points of a minimal set are recurrent. If Min_f denotes the union of all minimal subsets of X, the inclusions are

$$\mathrm{Per}_f \subset \mathrm{Min}_f \subset \mathrm{Rec}_f$$

Of course, an arbitrary transformation $f : X \to X$ may not admit any minimal subsets. This is the case of a non-trivial traslation $x \mapsto x + a$ of the real line.

If X is compact, we may consider the family $\mathcal{C}$ of those subsets $C \subset X$ which are closed, not-empty and +invariant, equipped with the natural partial order given by inclusion "$\subset$". The family is not empty, since it contains X itself. By Zorn lemma, $\mathcal{C}$ contains a minimal element K, which is clearly a minimal set. More generally, we proved the following result.

Theorem 8.7. *If a continuous transformation $f : X \to X$ admits a compact $C \subset X$ such that $f(C) \subset C$, then it admits at least a minimal subset $K \subset C$.*

A consequence is that a continuous transformation $f : X \to X$ defined in a compact space X admits at least one recurrent point (which may be unique), since $\mathrm{Min}_f \subset \mathrm{Rec}_f$.

Minimal transformations. A continuous transformation $f : X \to X$ is *minimal* if it satisfies one of the following equivalent conditions:

i) all orbits $\mathcal{O}_f^+(x)$ are dense in X

ii) X does not contain a proper close and +invariant subset, and therefore is a minimal set.

Clearly, a minimal transformation if +transitive. If X is a discrete space, minimality implies that X is made of a single orbit, which may be finite. If X is not dicrete, a minimal transformation cannot have periodic points.

Minimal homeomorfisms. A homeomorphism $f : X \to X$ is *minimal* if it satisfies one of the following equivalent conditions:

i) all complete orbits $\mathcal{O}_f(x)$ are dense in X

ii) X does not contain a proper close and invariant subset.

Minimal homeomorphisms are transitive. The discussion we made above about minimal sets may be repeated in this context. In particular, a homeomorphism $f : X \to X$ defined in a compact space X admits at least a minimal set $K \subset X$, which in this case is a closed not-empty invariant set which does not contain any proper closed invariant subsets.

ex: Give examples of transformations $f : X \to X$ such that $\mathrm{Min}_f = \emptyset$.

ex: Prove the implications i) $\Leftrightarrow$ ii) above in the definition of "minimal transformation".

ex: Prove the implications i) $\Leftrightarrow$ ii) above in the definition of "minimal homeomorphism".

8.5 Kronecker theorem on irrational rotations

Irrational rotations of the circle. A non-homogeneous version of Dirichlet's theorem 8.1 was discovered by Kronecker. In its original formulation[26], it says that, given an irrational α, for any integer $Q > 0$ and any $y \in \mathbb{R}$ there exist integers p and $q > Q$ such that

$$|q\alpha - p - y| < 3/q \tag{8.3}$$

Let $R_\alpha(x + \mathbb{Z}) = x + \alpha + \mathbb{Z}$ denotes the rotation of the circle $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ by the irrational angle $\alpha \notin \mathbb{Q}$. If we don't mind about the exact bound $3/q$ for the error, it says that for all $x + \mathbb{Z}$ and $x' + \mathbb{Z}$ in $\mathbb{T}$ (the y above is $x' - x$) and any precision $\varepsilon > 0$ there exists a time $q > 0$ such that $\mathrm{d}(R_\alpha^q(x + \mathbb{Z}), x' + \mathbb{Z}) < \varepsilon$. In our language,

Theorem 8.8 (Kronecker, 1884). *An irrational rotation of the circle is minimal (i.e. all its orbit are dense in the circle).*

Different proofs are presented in [HW59] (XXIII, Theorems 438, 439 and 440). Here we give just two.

Proof. Let $F \subset \mathbb{T}$ be the closure of an orbit of an irrational rotation of the circle. If F is not the whole circle, then its complementar $I = \mathbb{T}\backslash K$, which is a not-empty open set, is a countable union of open intervals (arcs of the circle). Let J be (one of) the intervals of I of maximal lenght (why does it exist?), say $|J| > 0$. We claim that its images $f^n(J)$, with $n \in \mathbb{Z}$, are pairwise disjoint. Indeed, two such intervals $f^n(J)$ and $f^m(J)$, with $n \neq m$, cannot coincide, for otherwise the boundary points would be periodic points of the rotation (which is irrational), and also cannot have not-empty intersection, for otherwise their union would be an interval of I of bigger lenght. Since the rotation preserves the lenghts, all $f^n(J)$ have the same positive lenght $|J|$, and this is impossible because the circle has finite (unit) lenght. □

A less abstract proof (with a worse constant than in Kronecker's statement), along the ideas of Dirichlet theorem, is as follows.

[26] L. Kronecker, Die Periodensysteme von Funktionen Reeller Variablen, *Berliner Sitzungsberichte* (1884), 1071-1080.

Proof. Given $\varepsilon > 0$, chose a positive integer Q such that $1/Q < \varepsilon$. Divide the circle into Q intervals of lenght $1/Q$. By the box principle, at least two of the $Q+1$ points

$$x + \mathbb{Z}, \quad R_\alpha(x + \mathbb{Z}), \quad \dots, \quad R_\alpha^q(x + \mathbb{Z})$$

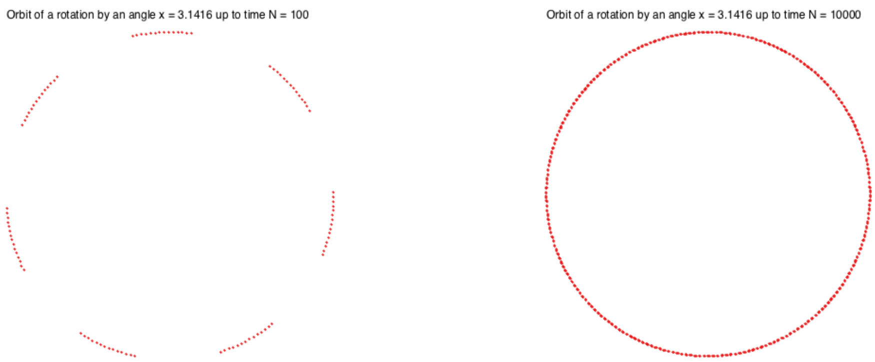
belong to the same interval, say $R_\alpha^i(x+\mathbb{Z})$ and $R_\alpha^j(x+\mathbb{Z})$ with $0 \leq i < j \leq Q$. Since rotations are isomeries,

$$\mathrm{d}(R_\alpha^i(x+\mathbb{Z}), R^j(x+\mathbb{Z})) = \mathrm{d}(R_\alpha^k(x+\mathbb{Z}), x+\mathbb{Z}) < 1/Q < \varepsilon\,,$$

with $1 \leq k = j - i \leq Q$. Thus, the rotation R_α^k displaces points a (positive) distance $< \varepsilon$. It is clear then that the images $R_\alpha^{nk}(x+\mathbb{Z})$, with $n \geq 0$, pass infinitely often in a ε-neighborhood of each point $x' + \mathbb{Z}$ of the circle. Thus, there exist integers $q = nk > Q$ and p such that $|q\alpha - p - y| < 1/Q < \varepsilon$, where $y = x' - x$. □

As explained in [HW59], Kronecker theorem has a nice physical interpretation. It implies that orbits in a square billard are either periodic, if the angle of incidence of the first hit to the boundary is a rational multiple of π, or dense in the square, otherwise. This is just the starting point of modern theory of billards, a major area in dynamical systems.

The theorem has also an "algebraic" side. Observe that the orbit of $0 + \mathbb{Z}$, the identity of the abelian group $\mathbb{R}/\mathbb{Z}$, is the cyclic subgroup generated by $\alpha + \mathbb{Z}$. Therefore Kronecker theorem says that *the closed and proper subgroups of* $\mathbb{R}/\mathbb{Z}$ *are the finite subgroups.*



Orbits of a rotation by an angle $\alpha \simeq \pi$ up to time 100 and 10000.

e.g. **Example of a non-measurable set.** If you believe the axiom of choice, you may consider a set B made of one (exactly one!) point for any orbit of an irrational rotation R_α of the circle. The images $B_n = R_\alpha^n(B)$, for $n \in \mathbb{Z}$, are pairwise disjoint and cover the circle. If B, hence all its images, were Lebesgue-measurable, then

$$\sum_{n\in\mathbb{Z}} |B| = \sum_{n\in\mathbb{Z}} |B_n| = |\cup_{n\in\mathbb{Z}} B_n| = |\mathbb{R}/\mathbb{Z}| = 1$$

since rotations preserve Lebesgue measure, so that $|B_n| = |B|$. But there exists no size $b = |B| \geq 0$ such that $\sum_{n\in\mathbb{Z}} b = 1$.

ex: Also instructive is to see why rational rotations are not transitive, using theorem 8.6, since this extends to the higher dimensional torus. If $\alpha = p/q$ is rational, then the function $\varphi(x+\mathbb{Z}) = \sin(2\pi q x)$ is well defined in the circle $\mathbb{R}/\mathbb{Z}$, continuous, non-constant, and clearly invariant under the rotation R_α.

Rotations of the torus. Kronecker's theorem is actually much more general [Kr84]. We say that the frequencies/numbers $\omega_1, \omega_2, \dots, \omega_k$ are *linearly independent over the rationals* if the only rational solution of the equation

$$n_1\omega_1 + n_2\omega_2 + \cdots + n_k\omega_k = 0$$

is the trivial solution $n_1 = n_2 = \cdots = n_k = 0$. An important example: the logarithms $\omega_k = \log p_k$ of different primes p_k are linearly independent, as follows from the uniqueness of prime decomposition.

Theorem 8.9 (Kronecker, 1884)**.** *Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_n) \in \mathbb{R}^n$. If $\theta_1, \theta_2, \dots, \theta_n, 1$ are linearly independent over the rationals, then any orbit*

$$\mathbf{x} + \mathbb{Z}\boldsymbol{\theta} + \mathbb{Z}^n$$

is dense in the torus $\mathbb{R}^n/\mathbb{Z}^n$.

This means that for any $\mathbf{x} = (x_1, x_2, \dots, x_n) \in [0,1)^n \approx \mathbb{T}^n := \mathbb{R}^n/\mathbb{Z}^n$ and any precision $\varepsilon > 0$ we can find integers $p_1, p_2, \dots, p_n$ and $q \in \mathbb{Z}$ such that $|q\theta_k - p_k - x_k| < \varepsilon$ for all $k = 1, 2, \dots, n$. Chapter XXIII of [HW59] contains some different proofs.

8.6 Circle homeomorfisms

While studying vector fields on the 2-dimensional torus $\mathbb{R}^2/\mathbb{Z}^2$, Henri Poincaré [27] was led to the necessity to claffify possible dynamics of circle homeomorphisms. A model is that of rotations, where the dichotomy between closed or dense orbits reflects the rationality of the "rotation angle". He discovered an invariant which plays a similar role for generic homeomorphisms.

Homeomorphisms of the circle. A homeomorphism $F : \mathbb{R} \to \mathbb{R}$ such that $F(x+1) = F(x)+1$ defines a orientation preserving homeomorphism $f : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$, according to $f(x+\mathbb{Z}) := F(x)+\mathbb{Z}$. Conversely, a bijection $f : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ is an *orientation preserving homeomorfism*of the circle if there exists a homeomorpfism $F : \mathbb{R} \to \mathbb{R}$ such that $F(x+1) = F(x)+1$ (observe that this condition implies that F is strictly increasing) and $f(x+\mathbb{Z}) = F(x)+\mathbb{Z}$. Such F is called *lift* of f. Clearly, the lift is not unique, but any two lifts F and G of f differs by an integer constant, i.e. $F(x) = G(x)+n$ for some $n \in \mathbb{Z}$.

For example, a lift of the rotation R_α is $F(x) = x+\alpha$. It is clear that $F_\lambda(x) = x+\alpha+\lambda\sin(2\pi x)$ induces a homeomorphism of the circle $f_\lambda(x+\mathbb{Z}) = x+\alpha+\lambda\sin(\pi x)+\mathbb{Z}$, provided that λ is sufficiently small, that may be considered a small variation of the rigid rotation $f_0 = R_\alpha$.

Rotation number. Let $f : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ be a orientation preserving homeomorphism if the circle, and let $F : \mathbb{R} \to \mathbb{R}$ be one of its lifts.

The *rotation number* of f is the "angle"

$$\rho(f) := \tau(F) + \mathbb{Z} \in \mathbb{R}/\mathbb{Z} \tag{8.4}$$

where $\tau(F)$ is the *translation number* of F, defined by

$$\tau(F) := \lim_{n\to\infty} \frac{F^n(x) - x}{n} \tag{8.5}$$

where x is an arbitrary point of the line. This makes sense once we prove that the limit exists and does not depend on the initial point x, and that its class in the circle doees not depend on the particular lift.

For example, the translation number of the translation $F(x) = x+\alpha$ is α, and therefore the rotation number of the rotation R_α is $\alpha+\mathbb{Z}$.

The main ingredient of the existence proof is the following fact, of independent importance, about subaditive sequences.

Theorem 8.10 (subadditive sequence lemma)**.** *Let $(a_n)_{n\in\mathbb{N}}$ be a quasi-subadditive real sequence, i.e. a sequence such that*

$$a_{n+m} \leq a_n + a_m + c$$

for any $n, m \in \mathbb{N}$ and some $c \geq 0$. Then there exists the limit

$$\lim_{n\to\infty} \frac{a_n}{n} \in \mathbb{R} \cup \{-\infty\}.$$

[27] H. Poincaré, Sur les courbes définies par les équations différentialles, *J. Math. Pures App.* Série IV **1** (1885), 167-244.

Proof. It is clear that existence of the limit $\lim_{n\to\infty} a_n/n$ is equivalent to existence of the limit $\lim_{n\to\infty} b_n/n$ for the sequence $b_n = a_n + c$. The sequence (b_n) is now subadditive, i.e.

$$b_{n+m} \leq b_n + b_m \, .$$

Subadditivity implies $b_n \leq nb_1$. Thus the sequence (b_n/n) is bounded above, hence there eixsts $\lambda = \liminf_{n\to\infty} b_n/n < \infty$. Given $\varepsilon > 0$, there exists a natural m such that $b_m/m < \lambda + \varepsilon$. A generic positive integer as $n = km + r$, with k a non-negative integer and $0 \leq r < m$. Let $iB = \max_{1\leq i<m} b_i$. Subadditivity also implies

$$\begin{aligned} b_n/n &\leq (b_{km} + b_r)/n \leq (kb_m + b_r)/n \\ &\leq b_m/m + b_r/n \leq \lambda + \varepsilon + B/n \end{aligned}$$

By the arbitrarity of ε, the inequality above implies that $\limsup_{n\to\infty} b_n/n \leq \lambda$. Thus, the limit $\lim_{n\to\infty} b_n/n$ exists and is equal to λ. □

Theorem 8.11. *The limit $\tau(F)$ in (8.5) exists.*

Proof. The lift F and its iterates F^n are increasing homeomorphisms of the real line satisfying $F^n(x+1) = F^n(x)+1$ for all $x \in \mathbb{R}$. In particular, $F^n(x) - x$ are periodic functions of period one. This implies that

$$\max_{x,x'} |(F^n(x) - x) - (F^n(x') - x')| \leq 1$$

since, by periodicity, we may compute the maximum inside the unit interval $[0, 1]$, and we know that F^n is increasing and that the image $F^n([0,1])$ is an interval of unit lenght. Let $a_n = F^n(x) - x$. The above inequality implies that the sequence (a_n) is quasi-subadditive, i.e.

$$a_{n+m} \leq a_n + a_m + c$$

for all $n, m \geq 0$ and some constant c. Indeed,

$$\begin{aligned} F^{n+m}(x) - x &= F^n(F^m(x)) - F^m(x) + F^m(x) - x \\ &= F^n(x) - x - F^n(x) + x + F^n(F^m(x)) - F^m(x) + F^m(x) - x \\ &\leq F^n(x) - x + F^m(x) - x + 1 \end{aligned}$$

so that we may chose $c = 1$. The theorem follows from theorem 8.10. □

Theorem 8.12. *The limit $\tau(F)$ in (8.5) does not depend on x.*

Proof. We already saw that $|(F^n(x) - x) - (F^n(x') - x')| \leq 1$. Therefore,

$$\left| \frac{F^n(x) - x}{n} - \frac{F^n(x') + x}{n} \right| \leq 1/n$$

for all x, x' and all n. This implies that $\tau(F)$ is independent on the chosen point x. □

Theorem 8.13. *The class $\rho(f) = \tau(F) + \mathbb{Z}$ does not depend on the lift F of f.*

Proof. Two different lifts , say F and G, differ by an integer, i.e. $G(x) = F(x)+k$ for some $k \in \mathbb{Z}$. This implies that $G^n(x) - x = F^n(x) - x + nk$ and therefore that $\tau(F) = \tau(G) + k$. □

Finally,

Theorem 8.14. *The rotation number $\rho(f)$ is invariant under topological conjugations.*

Proof. Let $h : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ be a connjugation between the homeomorphisms f and g. If H is a lift of H and F is a lift of F, then $H \circ F \circ H^{-1}$ is a lift of g. But the difference $\left(H \circ F \circ H^{-1}\right)^n (x) - F^n(x)$ is bounded, independently of x and n. Indeed, we may observe that

$$\left|\left(H \circ F \circ H^{-1}\right)^n\right| = \left|H \circ F^n \circ H^{-1}\right|, \qquad |H(x) - x| \qquad \text{and} \qquad \left|H^{-1}(x) - x\right|$$

are bounded by a constant which does not depend on x, and use triangular inequality. This implies that $\tau(F) = \tau\left(H \circ F \circ H^{-1}\right)$, and therefore that $\rho(f) = \rho(g)$. □

Of course, the rotation number of a rotation $R_\alpha(x + \mathbb{Z}) = x + \alpha + \mathbb{Z}$ is α itself.

ex: Show that $\rho(f^q) = q \cdot \rho(f) + \mathbb{Z}$ (observe that, if F is a lift of f, then F^n is a lift of f^n).

Poincaré classification theorem The rotation number contains the following information about the dynamics of an homeomorphism.

Theorem 8.15 (Poincaré)**.** *The rotation number $\rho(f)$ is rational iff the homeomorphism f admits periodic points.*

Proof. ($\Leftarrow$) If $F^q(x) = x + p$ with integers $q \geq 1$ and p, then $F^{nq}(x) - x = np$ for all n, and therefore $\tau(F) = p/q$.

($\Rightarrow$) Observing that $\rho(f^q) = q \cdot \rho(f) \bmod \mathbb{Z}$, it is sufficient to prove that $\rho(f) = 0$ implies that f has a fixed point. Now, if f does not have fixed points and F is a lift of f, then the function $F(x) - x$ has values in $\mathbb{R}\backslash\mathbb{Z}$. But the image of the real line by a continuous function is an interval. Therefore, there exists a lift F such that $F(x) - x$ takes values in the open unit interval $(0, 1)$. Since $F - \mathrm{id}$ is periodic with period one, its maximum and minimum are both different from 0 and 1. Thus, there exists $\varepsilon > 0$ such that $\varepsilon < F(x) < 1 - \varepsilon$ for all $x \in [0, 1]$. IIterating, this implies $n\varepsilon < F^n(0) < n(1 - \varepsilon)$ and therefore that $\tau(F)$ is not integer. □

Indeed, one can also prove tha if $\rho(f)$ is rational then all periodic points share the same period. Thus, in order to understand the structure of orbits of a homeomorphism with rational rotation mumber is sufficient to study the case of zero rotation number, i.e. homeomorphisms with a fixed point. If $C = \mathrm{Fix}(f)$ is the set of fixed points (which may be any compact subset of the circle), then f induces hoemomorphism in any connected component I of the open set $\mathbb{R}/\mathbb{Z}\backslash C$. Images $f^n(x)$ of points $x \in I \subset (\mathbb{R}/\mathbb{Z})\backslash C$ converge to points in $\partial I \subset C$ when $n \to \pm\infty$.

The dynamics of homeomorhisms with irrational rotation number is described by the following result.

Theorem 8.16 (Poincaré)**.** *Let $f : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ be a orientation preserving homeomorphism with irrational rotation number. Then*

i) either f is minimal, i.e. the orbit of all points are dense in the circle,

ii) or there exist en invariant compact subset $K \subset \mathbb{R}/\mathbb{Z}$, perfect and with empty interior (i.e. a Cantor set) such that the ω-limitset of all points of the circle is equal to K.

Proof. By Zorn lemma, the family of not-empty compact invariant sets of the circle, equipped with the natural partial order given by inclusion, admits a minimal set K. By minimality, the orbit of any point of K is dense in K. The boundary ∂K and the derived set K' are also compact and invariant, so they must be empty or coincide with K itself. Since f has no fixed oints, K cannot be finite. By the Bolzano-Weierstrass theorem $K' \neq \emptyset$, hence $K' = K$, i.e. K is perfect.

Now, if $\partial K = \emptyset$, then $K = \mathbb{R}/\mathbb{Z}$ and therefore f is minimal. If, otherwise, $\partial K = K$, then K has empty interior. Let $x \in (\mathbb{R}/\mathbb{Z}) \setminus K$ let I be the connected compontnet of $(\mathbb{R}/\mathbb{Z}) \setminus K$ which contains X. The images $f^n(I)$ are pairwise disjoint (beacause F has no fixed points), and therefore $\operatorname{diam}(f^n(I)) \to 0$ when $n \to \infty$. If $x' \in \partial I \subset K$, then $\omega_f(x') = K$, and the previous observation implies that also $\omega_f(x) = K$, because $d(f^n(x), f^n(x')) \leq \operatorname{diam}(f^n(I)) \to 0$ when $n \to \infty$. In particular, this shows that the minimal set K is unique. □

More interesting is the following result, also due to Poincaré.

Theorem 8.17 (Poincaré classification theorem)**.** *Let $f : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ be a orientation preserving homeomorphism with irrational rotation number ρ.*

i) If f is minimal, then it is topologically conjugated to the rotation R_ρ.

ii) If f is not minimal, then the rotation R_ρ os a factor of f.

Indeed, if f is minimal we may construct a conjugation H between one orbit of f and one orbit of R_ρ, and then define the full conjugation h by continuity, using the fact that orbits are dense. This is possible because orbits of f have the "same order" han orbits of a rotation. If f is not minimal, it is possible o construct a semiconjugation $h : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ such that $\mathbb{R}/\mathbb{Z}$ is the image $h(K)$ of the minimal set of f. Somehow, the semiconjugation "forgets" $(\mathbb{R}/\mathbb{Z}) \setminus K$, the wandering set of f.

This was a starting point of a beautiful story, starting with Denjoy in the 30's of the last century, and due to mathematicians like Michaël Hermann, Adrien Douady, Jean-Christophe Yoccoz, . . . See [MS93].

9 Chaos and mixing

9.1 Sensitive dependence and chaos

Regular points and loss of memory. Iterations of a continuous transformation $f : X \to X$ of a metric space divide in a natural manner the phase space into two classes of points, depending whether orbits are stable or unstable under small perturbations.

The point $x \in X$ is *regular* if the family $\{f^n\}_{n\geq 0}$ is equicontinuous at x, i.e. if for all $\varepsilon > 0$ there is a neighborhood B of x such that for all $x' \in B$ and all times $n \geq 0$

$$d\left(f^n\left(x\right), f^n\left(x'\right)\right) < \varepsilon$$

So, one orbit for each ε-ball contain informations on orbits of all regular points. In particular, if X is compact and all point is regular, we only need a finite number of orbits to describe the dynamics up to an error ε.

The point $x \in X$ is *not regular* if there exists $\delta > 0$ such that in any neighborhood B of x there are points existem $x' \in B$ such that

$$d\left(f^n\left(x\right), f^n\left(x'\right)\right) > \delta$$

for some time $n \geq 1$. This means tha f has "sensitive dependence on initial conditions" near x. In some sense, trajectories of points nearby x "lose their memory" of x. If the set of non-regular points is compact, the δ above may be chose uniformly for all of the points. This suggest the following definition.

Sensitive dependence on initial conditions. The continuous transformation $f : X \to X$ has *sensitive dependence on initial conditions* if all points are uniformly not-regular, i.e. if there exists $\delta > 0$ such that for all $x \in X$ and all neighborhoods B of x, there exist $x' \in B$ and a time $n \geq 1$ such that

$$d\left(f^n\left(x'\right), f^n\left(x\right)\right) > \delta$$

In other words, no matter how small our sensibility ε is, in a ε-neighborhood of any point x there is another point x' such that the futures of x and x' is uncorrelated, being at macroscopic (relative to ε) distance δ after some time n. Thus, a small change in the initial conditions may produce a large change at later times, a phenomenon popularized as "butterfly effect" by Edward Lorenz.

Of course this phenomenon is unexpected when the phase space is compact, for otherwise there is plenty of space for orbits to diverge from each other. Observe also that sensitive dependence is not compatible with preserving distances, hence isometries (like rotations of a torus) cannot have such a property. Thus, in order to display this kind of chaotic behaviour, a map must somehow "stretch" and "fold", as our examples below will show.

Chaos. The combination of sensitive dependence on initial condition and a dense set of periodic points is usually referred as *chaos* [28].

e.g. **Julia and Fatou sets.** The above dichotmy is particularluy meaningful for endomorphisms of the Riemann sphere $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, rational transformations $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$

$$z \mapsto f(z) = \frac{p(z)}{q(z)}$$

where q and q are polynomial. A point $z \in \overline{\mathbb{C}}$ is *regular* if it admits a neighborhood U such that $\{f^n|_U\}_{n\geq 1}$ is a "normal family" (i.e. every sequence of elements of the family admits a locally uniformly convergent subsequence).

[28] The Greek word $\chi\alpha o\varsigma$, which we may translate as "abysm", contains the same root $\chi\alpha$- (and probably comes from) of the verbs $\chi\alpha\iota\nu\varepsilon\iota\nu$ and $\chi\alpha\sigma\chi\varepsilon\iota\nu$, whch mean "open-itself", "open the mouth" or "yawn" (cfr. $\chi\alpha\sigma\mu\alpha$, i.e. "chasm"). It has been used in some greek cosmogonies to mean "the desordered mixture of elements preceeding the formation of the $\chi o\sigma\mu o\sigma$, the ordered universe".

The set $F \subset \overline{\mathbb{C}}$ of regular points, which is an open subset of the Riemann sphere, is called *Fatou set*. The complementar set, the compact $J = \overline{\mathbb{C}} \backslash F$, is called *Julia set*. The Julia set is where the interesting, i.e. desordered, dynamics takes place.

For example, if $f(z) = z^n$, then the Julia set is the unit circle $\mathbb{S}$. For perturbations, like for example $f(z) = z^2 + c$, the Julia set becomes a very irregular curve, typically of Hausdorff dimension > 1, or a "dust" like a Cantor set.

The investigation on the dynamics of rational mapst starded at the beginning of the last century (1918-19) with Gaston Julia e Pierre Fatou. The contemporary theory, due essentially to sophisticated ideas in complex analysis, is one of the greatest successes of the modern theory of dynamical systems. A great introduction is in a famous lectures notes by John Milnor [Mi91].

9.2 Topological mixing

Mixing. A continuous map $f : X \to X$ is *topologically mixing* if for any two not-empty open sets $U, V \subset X$there exists a time $n \geq 0$ such that for al times $k \geq n$

$$f^k(U) \cap V \neq \emptyset$$

This definition captures the idea tha the future $f^k(U)$, with $k \geq n$, of any open set is "asympthotically independent" on its present, since it intersets stably all other not-empty open set V.

It is clear that a topologically mixing map is also + transitive. In particular, if f is topologically mixing then $\mathrm{NW}_f = X$ and $\omega_f(x) = X$ is a generic property.

Theorem 9.1. *A topological mixing map of a non trivial metric space has sensitive dependence on the initial condition.*

Proof. Indeed, let U and V be two disjoint open sets at distance at least $2\delta > 0$ (which exist if X contains at leat two distinct points). Given $x \in X$, the orbit of any neighborhood B os x intersects any not-empty open set starting from some time $n \geq 0$. This easily implies, by the triangular inequality, that there exists a point $x' \in B$ such that $d(f^n(x'), f^n(x)) > \delta$. $\square$

In particular, an isometry (as a torus rotation) cannot be topologically mixing.

ex: Does there exist a minimal homeomorphism (hence topological transitive) which is not topologically mixing?

ex: Does there exist a topologically transitive map which is nor minimal neither topologically mixig?

ex: (difficult) A continuous map $f : X \to X$ is *weakly mixing* if the product map $f \times f : X \times X \to X \times X$, defined by

$$(x, x') \mapsto (f(x), f(x'))$$

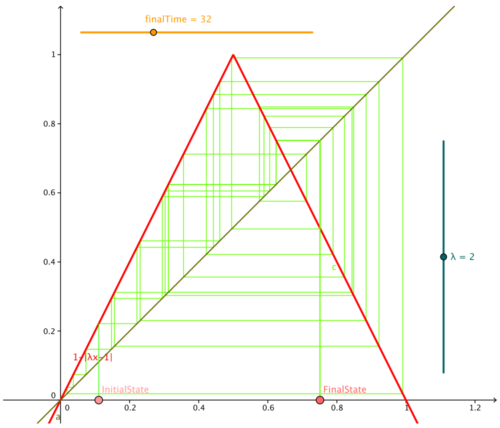
is topologically mixing. Show that a weakly mixing map of a non-trivial space X (i.e. which contains at least two points) has sensitive dependence on initial conditions. Show that all iterates f^n of a weakly mixing map of a compact space are +transitive. Show that

$$\text{mixing} \;\Rightarrow\; \text{weak mixing} \;\Rightarrow\; +\text{transitive}$$

and give examples which show that the reverse implications are false.

e.g. Tent map. One of the paradigms of chaotic maps is the *tent map* $T : [0,1] \to [0,1]$, defined by

$$T(x) = \begin{cases} 2x & \text{if } x < 1/2 \\ 2 - 2x & \text{if } x \geq 1/2 \end{cases}$$



Cobweb diagram of the tent map.

Iterations of the tent map are simple, since T is piecewise affine, and compositions of affine maps are affine maps (they for a group). Indeed, it is clear (and not difficult to prove by induction) that in any dydadic interval as $I_{k,n} = \left[\frac{k}{2^n}, \frac{k+1}{2^n}\right]$, with $k = 0, 1, 2, ..., 2^n - 1$, the iterate T^n is given by

$$T^n(x) = \begin{cases} 2^n x + k & \text{if } k \text{ is even} \\ -2^n x + k + 1 & \text{if } k \text{ is odd} \end{cases}$$

In particular, T^n is a strictly increasing or decreasing bijection of $I_{k,n}$ onto $[0,1]$. The fixed point theorem then implies that T^n has exactly one fixed point in each of these intervals $I_{k,n}$ (which is repelling since the derivative of T^n is $2^n > 1$, and only coincides with one of the boundary points when $k = 0$), and therefore $|\mathrm{Fix}(T^n)| = 2^n$.

Moreover, since any nnot-empty open interval $U \subset [0,1]$ contains one of the dyadic intervals $I_{k,n}$, if n is sufficiently large, this says that periodic points are dense in the interval.

Finally, the tent map is topologically mixing. Indeed, since any not-empty open set $U \subset [0,1]$ contains one of the $I_{k,n}$, its image under T^n is $T^n(U) = [0,1]$, and, a fortiori, $T^k(U) = [0,1]$ for all times $k \geq n$ because T is onto. This impies that $T^k(U) \cap V \neq \emptyset$ for all times $k \geq n$ and any other not-empty open set $V \subset [0,1]$. Thus, there exists a residual set of points x such that $\omega_f(x) = [0,1]$, i.e. with essentially unpredictably trajectory.

ex: Show that $h : x \mapsto \sin^2(\pi x/2)$ is a topological conjugation between the tent map T and the transformation $f_4 : [0,1] \to [0,1]$ of the quadratic family, defined by $f_4(x) = 4x(1-x)$. This show that f_4 has the same properties than T, e.g. it is topologically mixing and has a dense set of periodic points. Moreover, and this is surprising, this also provides explicit formulae for the trajectories of f_4. Indeed, if the initial condition is $x_0 = \sin^2(\pi\theta)$, then $x_n = f_4^n(x_0)$ is given by $x_n = \sin^2(2^n\pi\theta)$.

ex: Discuss also the dynamics of the map $S : [0,1] \to [0,1]$ defined by

$$S(x) = \begin{cases} 2x & \text{if } x < 1/2 \\ 2x - 1 & \text{if } x \geq 1/2 \end{cases}$$

Observe that S is not continuous, but it is not much different from the tent map.

9.3 Symbolic dynamics

Bernoulli shifts. The abstract archetypal mixing map is the *Bernoulli shift* $\sigma : \Sigma^+ \to \Sigma^+$ over a finite alphabet $\mathcal{A} \approx \{1, 2, ..., z\}$ made of $z > 1$ letters.

Any open not-empty set $U \subset \Sigma^+$ contains some centered cylinder $C_\alpha \subset U$, and, if $|\alpha|$ denotes the lenght of the finite word α, it is clear that $\sigma^n(C_\alpha) = \Sigma^+$ for all times $n \geq |\alpha|$. Therefore, $\sigma^n(U)$ intersect any other not-empty open set for such large times. Thus, σ is topoligically mixing. Being mixing, it is +transitive, and therefore a generic point has a dense orbit.

Indeed, in this case it is quite simple to exhibit points with dense orbits. Since the set of finite words in the alphabet is countable, we may enumerate finite words as $\alpha_1, \alpha_2, \alpha_3, \ldots$, and then observe that the trajectory of the point $x = \alpha_1\alpha_2\alpha_3\ldots$ passes through all centered cylinders, hence through all not-empty open sets.

Less obvious is to give example of points x such that $\omega_\sigma(x) = \Sigma^+$, which is also a generic property. An example is

$$x = \alpha_1\alpha_1\alpha_2\alpha_1\alpha_2\alpha_3\alpha_1\alpha_2\alpha_3\alpha_4\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5\ldots$$

whose trajectory visits all centered cylinders infinitely often

As we already saw, the Bernoulli shift also have dense periodic points, and fixed points of σ^n have cardinality $|\mathrm{Fix}(\sigma^n)| = z^n$. In particular, it is chaotic.

Besides, there are points whose orbit is dense in a proper subset of Σ^+. For example, the restriction of the shift on $(\mathcal{A}\backslash\{1\})^{\mathbb{N}} \subset \Sigma^+$, formed by infinite words which do no use the letter "1" (or any other letter), is topologically mixing (we may just repeat the discussion above). Therefore, a generic point $x \in (\mathcal{A}\backslash\{1\})^{\mathbb{N}}$ has an orbit which is dense inside $(\mathcal{A}\backslash\{1\})^{\mathbf{N}}$.

ex: Give examples of points $x \in \Sigma^+$ such that $\omega_\sigma(x) = X$.

ex: Give example of not pre-periodic points $x \in \Sigma^+$ tais such that the closure of the orbit $\mathcal{O}^+_\sigma(x)$ is a proper subset of Σ^+.

ex: Let $\Sigma = \mathcal{A}^{\mathbb{Z}}$ be the space of bi-infinite words $x = \ldots x_{-2}x_{-1}x_0x_1x_2\ldots x_n$ in the letters of the finite alphabet $\mathcal{A} = \{1, 2, ..., z\}$, equipped with the product topology. Verify that the (full) shift $\sigma : \Sigma \to \Sigma$, defined by $(\sigma(x))_k = x_{k+1}$, is a homeomorphism. Find the cardinality of $\mathrm{Fix}(\sigma^n)$, and prove that periodic points are dense. Show that $\sigma : \Sigma \to \Sigma$ is topologically mixing.

ex: The *Backer's map* is the transformation $f : [0,1] \times [0,1] \to [0,1] \times [0,1]$ defined by

$$(x, y) \mapsto \begin{cases} (2x, y/2) & \text{if } 0 \leq x \leq 1/2 \\ (2x-1, (y+1)/2) & \text{if } 1/2 < x \leq 1 \end{cases}$$

Discuss its dynamics. Consider the Bernoulli shift . $\sigma : \Sigma \to \Sigma$ on $\Sigma = \{0,1\}^{\mathbb{Z}}$. Show that the map $h : \Sigma \to [0,1] \times [0,1]$, defined by

$$\ldots x_{-2}x_{-1}x_0x_1x_2\cdots \mapsto \left(\sum_{n=0}^{\infty} \frac{x_{-n}}{2^n}, \sum_{n=1}^{\infty} \frac{x_n}{2^n}\right)$$

is a semi-conjugation between σ and f.

Topological Markov chains. The rescrtiction of the shift σ to an invariant subset of Σ^+ or Σ is called *symbolic dynamical system.*

The simplest way to produce invariant subsets uses trasition matrices (an idea which comes from the theory of Markov chains in probability). Let $A = (a_{ij})$ be a "transition matrix", i.e. a $z \times z$ matrix with entries 0 or 1. Let

$$\Sigma^+_A = \left\{x \in \Sigma^+ \text{ such that } a_{x_n x_{n+1}} = 1 \ \forall\ n \geq 0\right\} .$$

It is clear that $\sigma\left(\Sigma^+_A\right) \subset \Sigma^+_A$. The restriction

$$\sigma_A := \sigma\Big|_{\Sigma^+_A} : \Sigma^+_A \to \Sigma^+_A$$

is called *topological Markov chain*, or *subshift of finite type*. The idea is that the letters of the alphabet represents the possible states of a system, and transition from the state $x_n = i$ at time n to the state $x_{n+1} = j$ at time $n+1$ is possible iff $a_{ij} = 1$.

A finite word $\alpha = \alpha_1\alpha_2\dots\alpha_n$ is *admissible* if $a_{\alpha_k\alpha_{k+1}} = 1$ for all $k = 1, 2, ..., n-1$, i.e. if it is a piece of a admissible word of Σ_A^+. It is useful to introduce a *Markov graph* $\mathcal{G}_A$, whose vertices are the letters/states of the alphabet $\mathcal{A} \approx \{1, 2, \dots, z\}$, with oriented edges from i to j whenever $a_{ij} = 1$. Admissible words are therefore admissible paths, i.e. sequences of consecutive letters joined by the edges of the graph.

A little reflection shows that the cardinality of admissible words of lenght $n+1$ which start with the letter i and end with the letter j, i.e. the number of different paths of lenght $n+1$ in $\mathcal{G}_A$ joining i to j, is equal to the entry $(A^n)_{ij}$ of the n-th power of the transition matrix. Therefore,

$$|\mathrm{Fix}\,(\sigma_A^n)| = \mathrm{Tr}(A^n)\,.$$

This number may be estimated using the Perron-Frobenius theorem, and it grows like $\sim \lambda^n$, if λ is the largest eigenvalue of A.

The topological Markov chain $\sigma_A : \Sigma_A^+ \to \Sigma_A^+$ is *irreducible* if for any two states i and j there exist a time $n \geq 1$ such that the ij entry of the n-th power of A is not zero, i.e. $(A^n)_{ij} \neq 0$.

The topological Markov chain $\sigma_A : \Sigma_A^+ \to \Sigma_A^+$ is *transitive* if there exists a time $n \geq 1$ such that all the entries of A^n are striclty positive. It is clear that this implies that each row and each column of A has at least a non-zero entry (for otherwise $A^n = AA^{n-1} = A^{n-1}A$ would have a zero entry), and this in turn implies that all the entries of A^k are strictly positive whenever $k \geq n$. Clearly, a transitive Markov chain is irreducible, but the converse is false.

The relevant results are

Theorem 9.2. *An irreducible topological Markov chain is topologically transitive.*

Theorem 9.3. *A transitive topological Markov chain is topologically mixing and has dense periodic orbits.*

ex: Prove the above theorems, following the lines of the case of the Bernoulli shift.

Coding. Symbolic dynamical systems are abstract models for dynamical systems. One of the central idea in dynamical systems is indeed to "code" an actual map $f : X \to X$ with a symbolic system. A possible strategy is to divide tha phase space X into pieces $B_1, B_2, ..., B_z$, hoping that the history of any point $x \in X$, defined by the infinite word $x_0x_1x_2\dots x_n\dots$ with $x_n = i$ iff $f^n(x) \in B_i$, determines uniquely the starting point x. We define the transition matrix $A = (a_{ij})$ such that $a_{ij} = 1$ if $B_j \subset f(B_i)$ and $a_{ij} = 0$ otherwise. If f is sufficiently chaotic, to any infinite word $x_0x_1x_2\dots x_n\dots \in \Sigma_A^+$ there corresponds a unique $x = \bigcap_{n=0}^\infty f^{-n}(B_{x_n})$. The hope is to be able to show that f is conjugated, or at least semi-conjugated (since there will be ambiguities at the boundaries of the B_i's), to σ_A.

9.4 Cantor sets

Orbit closures of sufficiently chaotic maps have often complicated structures. If disconnected, they are typically Cantor sets, i.e. perfect and totally disconnected compact sets.

Middle-third Cantor set. The archetype if the *middle-third Cantor set*

$$K := \left\{ \sum_{n=1}^{\infty} \frac{x_n}{3^n} \text{ with } x_n \in \{0, 2\} \right\} \subset [0, 1]\ ,$$

the set of those numbers in the unit interval such that their base 3 representation does not use the letter "1".

Another popular definition is $K = [0,1] \setminus \bigcup_{k=1}^{\infty} I_k$, where the open intervals I_k are defined inductively as follows: $I_1 = (1/3, 2/3)$ is the central middle-third of the unit interval, $I_2 = (1/9, 2/9)$ and $I_3 = (7/9, 8/9)$ are the central middle-third intervals of the two components of $[0,1] \setminus I_1$, and so on.

One more definition is $K = \bigcap_{k \geq 0} K_n$, where

$$K_n = \left\{ \sum_{k=1}^{\infty} \frac{x_n}{3^n} \text{ with } x_1, x_2, ..., x_n \in \{0,2\} \text{ and } x_k \in \{0,1,2\} \text{ se } k > n \right\}$$

denotes the compact set of those numbers in the unit interval such that their base 3 representation does not use the letter "1" at the first n places. Observe that the K_n's form a deacreasing family, i.e. $\dots \subset K_{n+1} \subset K_n \subset \dots \subset K_0 = [0,1]$, and that each K_n is a disjoint union of 2^n closed intervals of lenght 3^{-n}.

In particular, K is compact, being a countable intersection of compact sets.

K does not contain isolated points, and therefore $K' = K$, i.e. it is "perfect". Indeed, if $x = 0.x_1x_2x_3\dots$ is the base 3 representation of $x \in K$, we may change just the n-th digit (from 0 to 2 or vice-versa), and contruct a sequence of distinct points of K converging to x.

K is "totally disconnected", i.e. the connected component of each $x \in K$ is $\{x\}$ itself. Indeed, any two distict point are at a distance larger that 3^{-n} for some sufficiently large n, and therefore cannot be contained in the same connected component of K_n.

The strange properties of the Cantor sets become less misterious if one observe that it is homeomorphic to the topological product $\{0,2\}^{\mathbb{N}}$, the space of the Bernoulli shift over an alphabet of two letters. The homeomorphism is simply

$$\{0,2\}^{\mathbb{N}} \ni x_1x_2\dots x_n\dots \mapsto \sum_{n=1}^{\infty} \frac{x_n}{3^n} \in K$$

The function $\{0,2\}^{\mathbb{N}} \to \{0,2\}^{\mathbb{N}} \times \{0,2\}^{\mathbb{N}}$, defined by

$$x_1x_2x_3x_4\dots\dots \mapsto (x_1x_3\dots\,,\, x_2x_4\dots)$$

induces a homeomorphism of K onto $K \times K$. By induction, we see that K is homeomorphic to any finite power K^n. Indeed, one can prove that K is also homeomorphic to the countable Cartesian product $K^{\mathbb{N}}$ (provided one understand the product topology on this space).

Observe that $\{0,2\}^{\mathbb{N}}$ is trivially homeomorphic to $\{0,1\}^{\mathbb{N}}$, and that the binary representation is a continuous map of $\{0,1\}^{\mathbb{N}}$ onto the unit interval $[0,1]$, given explicitely by $x_1x_2x_2 \mapsto \sum_{k \geq 0} x_k/2^k$. Thus, there exists a continuous map of K onto the unit interval $[0,1]$. By the Schröder-Bernstein theorem, K has the cardinality of the interval.

Another much appreciated property of the Cantor set is its "self-similarity", a property which makes of K the prototype of a "fractal set". It is clear, indeed, that any of the closed intervals which form K_n contains an affine copy of K itself (we must only make an homothety of ratio 3^n and an appropriate translation).

Finally, the "lenght" (i.e. the Lebesgue measure) of K is

$$|K| = \lim_{n \to \infty} |K_n| = \lim_{n \to \infty} 2^n \cdot 3^{-n} = 0\,.$$

The Cantor set is very "small", while containing the same cardinality of points as the whole interval!

e.g. **Cantor sets from the quadratic family.** Consider the quadratic family $f_\lambda : \mathbb{R} \to \mathbb{R}$ (this time defined in the whole real line), defined by $x \mapsto \lambda x\,(1-x)$, where $\lambda > 0$. The trajectory of any point outside the unit interval $I = [0,1]$ diverges. We may therefore define the set

$$\Lambda = \bigcap_{n \geq 0} f_\lambda^{-n}\,(I)$$

of those points with bounded orbits. If $\lambda > 4$, a picture shows that $f_\lambda^{-1}\,([0,1])$ is the disjoint union of two closed not-empty intervals I_0 and I_1 contained in $[0,1]$. If λ is sufficiently large, it is also clear that $|f'_\lambda(x)|$ is uniformly larger than one at the points of $I_0 \cap I_1$. By induction, one can show that this implies that $f_\lambda^{-(n+1)}(I)$ is a disjoint union of 2^{n+1} compact intervals striclty contained, in pairs, in the 2^n compact intervals which form $f_\lambda^{-n}(I)$. There follows that Λ is a Cantor set, and that the restriction $f_\lambda|_\Lambda : \Lambda \to \Lambda$ is topologically conjugated to the Bernoulli shift $\sigma : \Sigma^+ \to \Sigma^+$ in the alphabet $\{0, 1\}$.

9.5 Expanding maps

The obvious way to force sensitive dependence on initial conditions is "stretching", for example dilating distances at least along some directions, and "folding", for example, taking quotients.

Expanding maps. A continuous transformation $f : X \to X$ of a metric space is *expanding* if there exist $\lambda > 1$ and $\varepsilon > 0$ such that for all distinct $x, x' \in X$ at distance $d(x, x') < \varepsilon$

$$d(f(x), f(x')) > \lambda \cdot d(x, x')$$

This last, which looks like an opposite of a contraction, is a local condition, since otherwise compact spaces would not admit expanding maps with large ε. On the other side, it is precisely in compact phase spaces X that stretching induces chaotic orbits, since there is not enough space to escape, and divergent trajectories are forced to come back, eventually.

The mere existence of expanding maps also implies strong topological restrictions on the possible phase spaces. If X is a manifold, then its universal cover must be $\mathbb{R}^n$, and even then, its fundamental group cannot be arbitrary. For example, between all the orientable compact surfaces, only the torus $\mathbb{R}^2/\mathbb{Z}^2$ admits expanding transformations!

Expansive maps. A weaker notion is also useful. A continuous transformation $f : X \to X$ of a metric space is *(positively) expansive* if there exists $\delta > 0$ such that for all distinct $x, x' \in X$ there exists a time $n \geq 0$ such that

$$d(f^n(x), f(x')) > \delta$$

Equivalently, if there exists a $\delta > 0$ such that if the orbit of two points $x, x' \in X$ stays at distance $d(f^n(x), f^n(x')) < \delta$ for all times $n \geq 0$ then the points coincide, i.e. $x = x'$. Thus, expansive maps have sensitive dependence on initial conditions.

It is clear that an expanding map is expansive.

ex: Give examples of expanding transformations of $\mathbb{R}$, of $\mathbb{R}/\mathbb{Z}$ and of $\mathbb{R}^2/\mathbb{Z}^2$.

ex: Shows that there is no expansive map $f : I \to I$ defined in a compact interval $I \subset \mathbb{R}$ (observe that such map would be locally injective, hence strictly increasing or decreasing ...)

ex: Can an expanding map of a compact space be an homeomorphism? The answer is yes if the space is finite, and an example is not so difficult. On the other side, one can show (but it is not easy!) t that an infinite compact space does not admits expanding homeomorphisms.

e.g. Decimal expansion. The most famous expanding map is of course "multiplication by 10", the circle map $f : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ defined by

$$f(x + \mathbb{Z}) = 10 \cdot x + \mathbb{Z}$$

If $x = 0.x_1x_2x_3\dots$, with $x_n \in \{0, 1, 2, ..., 9\}$ is the representation of $x \in [0, 1)$ in base 10, $x \in [0, 1)$, then

$$f(0.x_1x_2x_3 \dots + \mathbb{Z}) = 0.x_2x_3x_4 \dots + \mathbb{Z}$$

Periodic and pre-periodic points, which correspond to rationals, are dense.

If the circle $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ is equipped with the standard metric, it is clear that if $d(x, x') < 1/20$ then $d(f(x), f(x')) = 10 \cdot d(x, x')$. Therefore, f is expanding.

Sensitive dependence on initial conditions can also easly recognized. Indeed, if $d(x,x') < 1/2\cdot 10^{-n}$, then $d(f^n(x), f^n(x')) = 10^n \cdot d(x,x')$. Therefore, for all $\varepsilon > 0$ and all $x \in \mathbb{R}/\mathbb{Z}$, there exist another point $x' \in \mathbb{R}/\mathbb{Z}$ and a time $n \geq 0$ such that

$$d(x,x') < \varepsilon \qquad \text{e} \qquad d(f^n(x), f^n(x')) > 1/4\,.$$

It is also clear that for any not-empty interval $I \subset \mathbb{R}/\mathbb{Z}$, there exists a time $n \geq 0$ such that $f^k(I) = \mathbb{R}/\mathbb{Z}$ for all times $k \geq n$ (it is sufficient to observe that I contains some $J = [k/10^2, (k+1)/10^n]$ for n sufficiently large, and that $f^n(J) = \mathbb{R}(\mathbb{Z})$. This, f is topologically mixing.

Le $\alpha = \alpha_1\alpha_2\ldots\alpha_n$ be a finite word in the letters of the alphabet $\{0,1,2,...,9\}$. There esists a residual set of points $x \in \mathbb{R}/\mathbb{Z} \approx [0,1)$ such that their base 10 representation contains the word α infinitely often (in the sense that, if $x = 0.x_1x_2x_3\ldots$, there exist an infinity of times $k \geq 0$ such that $x_{k+1}x_{k+2}\ldots x_{k+n} = \alpha_1\alpha_2\ldots\alpha_n$). Moreover, since finite words are countable, there exists a residual set of points $x \in \mathbb{R}/\mathbb{Z} \approx [0,1)$ such that their base 10 representation contains all finite words in the alphabet $\{0,1,2,...,9\}$ infinitely often. This means that a “generic” infinite book contains all the possible finte books infinitely often! More is true, as showed by Émile Borel (see the paragraph on normal numbers below).

ex: Give example of points in the residual sets described above.

Linear expanding maps of the circle. There is, of course, nothing special with the number 10 used above, the number of fingers in our hands. A *standard expanding map* of the interval is a map $f : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$, defined by

$$x + \mathbb{Z} \mapsto z\cdot x + \mathbb{Z}$$

where $z \in \mathbb{Z}$ an integer such that $|z| > 1$. It is expanding, topologically mixing and has sensitive dependence on initial conditions, and have dense and countable set of periodic points. Proofs are just a rewriting of the above proofs that we gave for $z = 10$.

The map f is a factor of the Bernoulli shift over an alphabet made of $|z|$ letters, and the set where the semi-conjugation fails to be one-to-one is small (it is made of rationals).

Besides periodic orbits and dense infinite orbits, such maps also admit more complicated orbit closures. For example, the expanding map with $z = 3$ clearly preserves the Cantor set K, thought as a subset of the circle (i.e. with the points 0 and 1 identified), i.e. $f(K) \subset K$, and the restriction $f|_K : K \to K$ is clearly topologically mixing. Thus, there exist orbits of $f : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ which are dense in K.

Non-linear expanding maps of the circle. We now consider a generic, not necessarily linear, expanding map $g : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ of class $\mathcal{C}^1$, i.e. such that any of its lifts $G : \mathbb{R} \to \mathbb{R}$ is continuously differentiable. Being F' periodic with period 1, there exists $\lambda > 1$ such that $|G'(x)| > \lambda$ for all $x \in \mathbb{R}$, and G' does not change sign. In particular, the degree of g has absolute value strictly bigger than one, since

$$|\deg(g)| = |G(1) - G(0)| = \left|\int_0^1 G'(x)\,dx\right| = \int_0^1 |G'(x)|\,dx > \int_0^1 \lambda\,dx > 1.$$

Theorem 9.4. *Any expanding map $g : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ of class $\mathcal{C}^1$ and degree $\deg(f) = z$ is topologically conjugated to the standard expanding map $f : x + \mathbb{Z} \mapsto z\,x + \mathbb{Z}$.*

Proof. For simplicity, we assume that G is increasing, i.e. that $z > 1$. The idea is to first define a conjugation between the pre-images of a fixed point, and then extend it to the whole circle using the facts that such pre-images are dense.

Let $x_k^i = i/\lambda^k$, with $i = 0,1,...,\lambda^k - 1$. Then $f(x_k^i) = x_{k-1}^{i'}$, where i' is the unique integer between 0 and $z^{k-1} - 1$ such that $i = i' \bmod z^{k-1}$. Let p be the fixed point of G, a lift of g. Since G is strictly increasing and $G(p+1) = p + z$, there exist $p = y_1^0 < y_1^1 < ... < y_1^{z-1} < p+1$ such that $g(y_1^i) = p + i$. Inductively (in k) we define the points y_k^i, with $i = 0,1,...,z^k - 1$ such that

$$y_{k-1}^i = y_k^{zi} < y_k^{zi+1} < ...y_k^{zi+z-1} < y_k^{zi+z} = y_{k-1}^{i+1}$$

and $G(y_k^i) = y_{k-1}^{i'}$, where i' is the unique integer between 0 and $z^{k-1}-1$ such that $i = i' \mod z^{k-1}$. For any interval $I_k^i = \pi\left(\left[y_k^i, y_k^{i+1}\right]\right)$ we have $g^k(I_k^i) = \mathbb{R}/\mathbb{Z}$ (remeber that $\pi : \mathbb{R} \to \mathbb{R}/\mathbb{Z}$ is the projection of the line onto the circle). Since g is expanding, i.e. there exists $\lambda > 1$ such that $|G'(x)| > \lambda$ for all x, any of these intervals has lenght $\left|I_k^i\right| < \lambda^{-k}$, and therefore the family of points $\left\{y_k^i\right\}_{k\in\mathbb{N},\ i=0,1,\dots,z^k-1}$ is dense in $[p, p+1]$. The function

$$H : \left\{y_k^i\right\}_{k\in\mathbb{N},\ i=0,1,\dots,z^k-1} \to \left\{x_k^i\right\}_{k\in\mathbb{N},\ i=0,1,\dots,z^k-1}$$

defined by $H(y_k^i) = x_k^i$ is strictly monotone. The density of the points $\left\{y_k^i\right\}$ and $\left\{x_k^i\right\}$ allows to extend H as a homeomorphism $H : [p, p+1] \to [0,1]$, which in turn defines a homeomorphism $h : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$. Finally, one easily see that $f \circ h = h \circ g$. $\square$

In particular, given an expanding map of the circle $g : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ of class $\mathcal{C}^1$, all maps sufficiently near to g in the $\mathcal{C}^1$ topology is topologically conjugated to f. This follows from the fact that expansiveness is an open condition, and that the degree is locally constant. Therefore,

Theorem 9.5. *Continuously differentiable expanding maps of the circle are $\mathcal{C}^1$-structurally stable.*

9.6 Hyperbolic automorphisms of the torus

Expansiveness is not necessary to produce chaos. It was Anosov, following the work by Hadamard and Hopf on the geodesic flow on surfaces with negative curvature, who discovered a large class of chaotic transformations (and flows), where chaos is due to some non-trivial way of streching and folding.

Automorphisms of the torus. Let $\mathbb{T}^n := \mathbb{R}^n/Z^n$ be the n-dimensional torus. A linear map $L : \mathbb{R}^n \to \mathbb{R}^n$ defined, in the canonical basis, by a $n\times n$ matrix with integer entries $A \in \mathrm{Mat}_{n\times n}(\mathbb{Z})$, induces an endomorphism of the torus $f : \mathbb{T}^n \to \mathbb{T}^n$ according to

$$f_A(x + \mathbb{Z}^n) := Ax + \mathbb{Z}^n\,.$$

This is clear, since a matrix with integer entries sends the lattice $\mathbb{Z}^n$ into itself, i.e. $A(\mathbb{Z}^n) \subset \mathbb{Z}^n$. If if happens that $\det A = \pm 1$, then A is invertible and its inverse A^{-1} also has integer entries. This implies that f_A is invertible too, i.e. is an *automorphism* of the torus.

Modular group. The existence of non-trivial automorphisms of the torus is due to arithmetical reasons. For example, orientation preserving automorphisms of the 2-dimensional torus are induced by 2×2 integer matrices with determinant one, which form the *modular group* $\mathrm{SL}_2(\mathbb{Z})$. It is made of matrices

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

such that a, b, c, d are integers satisfying $ad - bc = 1$. But this means that lines and columns of A are made of pairs of relatively prime integers! Simple non-trivial (different from the identity) examples are

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \qquad \begin{pmatrix} 3 & 2 \\ 1 & 1 \end{pmatrix} \qquad \dots$$

And much more can be produced using the group structure. Indeed, $\mathrm{SL}_2(\mathbb{Z})$ is a “large group”, and one of the most interesting group in mathematics, since it contains informations about primes, and also is related to the hyperbolic geometry of the Poincaré upper half-space $\mathbb{H}$. Indeed, isometries of $\mathbb{H}$ are induced by fractional linear transformations $z \mapsto (az+b)/(cz+d)$ with $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{R})$, and the quotient $\mathbb{H}/\mathrm{PSL}_2(\mathbb{Z})$ is an interesting hyperbolic surface, called “modular orbifold”.

In general, orientation preserving automorphisms of the n-dimensional torus are induced by matrices $A \in \mathrm{SL}_n(\mathbb{Z})$. The homogeneous space $\mathrm{SL}_n(\mathbb{R})/\mathrm{SL}_n(\mathbb{Z})$ is the space of lattices $\Gamma \subset \mathbb{R}^n$ with unit co-volume (the volume of a fundamental region). Indeed, any $G \in \mathrm{SL}_n(\mathbb{R})$ sends the standard lattice $\mathbb{Z}^n$, which has a fundamental region $[0,1]^n$ of volume one, into a lattice $G(\mathbb{Z}^n)$ with a fundamental region of volume one (because $\det G = 1$), and the stabilizer of the standard lattice is precisely $\mathrm{SL}_n(\mathbb{Z})$.

ex: Shows that for any pair of relatively prime integers p and q there exists a matrix $A \in \mathrm{SL}_2(\mathbb{Z})$ having p and q in the first column (or row).

Hyperbolic automorphisms of the torus. Let $f_A : \mathbb{T}^n \to \mathbb{T}^n$ be the automorphism of the torus induced by a matrix $A \in \mathrm{SL}_n(\mathbb{Z})$. If some power A^q of A has eigenvalue 1, and $v \in \mathbb{R}^n$ is a corresponding eigenvector, then the entire line $\mathbb{R}v + \mathbb{Z}^n \subset \mathbb{T}^n$ is made of periodic points (of period which divides q) of the automorphism f_A. This line may be dense in the torus or in some subtorus, depending on the rationality properties of the coordinates of v.

A square matrix A, and the corresponding endomorphism of the torus, is called *hyperbolic* if it does not have eigenvalues with absolute value one, i.e. if its spectrum is disjoint from the unit circle of the complex plane. If $\det A = 1$, this also implies that (the complexification of) A has eigenvalues with both $|\lambda| > 1$ and $|\lambda| < 1$, since their product must be one. Thus, A dilates distances in some diections and contracts distances in some other directions.

Theorem 9.6. *Let $f_A : \mathbb{T}^n \to \mathbb{T}^n$ be a hyperbolic automorphism of the torus. The set of periodic points of f_A is the set $\mathbb{Q}^n/\mathbb{Z}^n$ of points with rational coordinates. In particular,* Per_{f_A} *is dense in the torus.*

Proof. If $x+\mathbb{Z}^n$ is a periodic point of period $q \geq 1$, then $A^q x = x+k$, or, equivalently, $(A^q - I)x = k$, for some $k \in \mathbb{Z}^n$. If the eigenvalues of A are not roots of one, then $A^q - I$ is invertible, and it is clear that the entries of its inverse are rationals. There follows that $x = (A^q - I)^{-1}k$ has rational coordinates. Thus, periodic points are rational.

On the other side, for any fixed natural $q \geq 1$, we may consider the finite set $Q_q \subset \mathbb{T}^n$ of those points of the torus with coordinates that are integer multiples of $1/q$ (it has cardinality q^n). It is clear that $f_A(Q_n) = Q_n$, because A multiplies coordinates by integers, and f_A is invertible. But if we have a permutation of a finite set, any point is periodic. Since the denominator q was arbitrary, this prove that all rational points are periodic. □

ex: Show that $|\mathrm{Per}_n(f_A)| = |\det(A^n - \mathrm{id})|$.

Arnold's cat map. The classical example of a hyperbolic automorphism of the torus $\mathbb{T}^2$ is induced by the unimodular matrix

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

and reads

$$f_A((x,y) + \mathbb{Z}^2) = (2x+y, x+y) + \mathbb{Z}^2$$

It is known as *Arnold's cat map.*

The eigenvalues of A are

$$\lambda_\pm = \frac{3 \pm \sqrt{5}}{2}.$$

Moreover, since A is symmetric, one can find eigenvectors $v_\pm$ which form an orthogonal basis, and they are vectors with irrational slopes. Thus, $\mathbb{R}^2$ is the orthogonal direct sum $\mathbb{R}^2 = E^+ \oplus E-$ of the eigenspaces. The linear map $x \mapsto Ax$ dilates vectors of $E^+\backslash\{0\}$ by a factor $\lambda_+ > 1$, and contracts vectors of $E^-\backslash\{0\}$ by a factor $\lambda_- < 1$. Observe that f_A preserves areas, since $\det A = \lambda_+\lambda_- = 1$.

Theorem 9.7. *The Arnold cat map is topologically mixing, hence chaotic.*

Proof. (sketch) The projections of the lines $x + E^\pm \subset \mathbb{R}^2$ into $\mathbb{T}^2$ contain orbits of a minimal translation of the torus, because the slopes $\lambda_\pm$ are irrational, and therefore they are dense in the torus. Let $R \subset \mathbb{T}^2$ be a small square with sides of lenght ℓ parallel to the $E^\pm$'s. The image $f^n(R)$ is a "rectangle" with sides $\ell \cdot \lambda_+^n$ and $\ell \cdot \lambda_-^n$, still parallel to the lines $E^\pm$, respectively. As n grows, the complementar set $\mathbb{T}^2\backslash f^n(R)$ does not contan balls of radius greater than some δ, where $\delta \to 0$ as $n \to \infty$. Thus, $f^n(R)$ intersects stably any not-empty open subset of the torus. □

9.7 Topological entropy

Coverings, nets and separated sets. The following notions of size of a metric space (X, d) are due to Kolmogorov's school [29] [30].

An *ε-covering* of (X, d) is a covering of $X \subset \bigcup_\alpha C_\alpha$ by subsets of diameters $\mathrm{diam}(C_\alpha) < 2\varepsilon$. Call $C_\varepsilon(X, d)$ the minimal cardinality of an ε-covering of X.

An *ε-net* for (X, d) is a collection $N \subset X$ of points such that any point of X is at a distance not exceeding ε from some point of A, i.e. $X \subset \bigcup_{p \in N} B_\varepsilon(p)$. Call $N_\varepsilon(X, d)$ the minimal cardinality of an ε-net for X. If X is a centered space (any subset of diameter $2r$ is contained in a ball of radius r centered in some point of X) then $N_\varepsilon(X, d) = C_\varepsilon(X, d)$.

A subset $S \subset X$ is said *ε-separated* (or *ε-distinguishable*) if its points are a distance greater than ε from each other, i.e. if $d(p, p') > \varepsilon$ for all $p, p' \in S$ such that $p \neq p'$. The collection of disjoint balls $B_{\varepsilon/2}(p)$, where p ranges in a ε-separated set S, is also called *ε-packing*. Call $S_\varepsilon(X, d)$ the maximal cardinality of a set of ε-separated points inside X.

These three definitions make sense if the above extremal cardinalities are finite for every $\varepsilon > 0$, and it is not difficult to see that this happens simultaneously. The class of metric spaces with this property is called the class of *totally bounded sets* and the main examples are compact spaces.

The (base 2, for example) logarithms of these quantities have interpretations related to the probabilistic theory of transmission of signals, and are called

$$\begin{array}{rl} \log C_\varepsilon(X,d) & \textit{absolute } \varepsilon\textit{-entropy of } (X,d) \\ \log N_\varepsilon(X,d) & \varepsilon\textit{-entropy of } (X,d) \\ \log S_\varepsilon(X,d) & \varepsilon\textit{-capacity of } (X,d) \end{array}$$

ex: Show that an ε-net defines an ε-covering, and any ε-covering determines a 2ε-net, so that

$$C_\varepsilon(X, d) \leq N_\varepsilon(X, d) \leq C_{2\varepsilon}(X, d) \tag{9.1}$$

ex: Show that a maximal ε-separated set is a ε-net, and that any ε-ball centered at a point of a minimal ε-net cannot contain more than one point of a 2ε-separated set, so that

$$S_{2\varepsilon}(X, d) \leq N_\varepsilon(X, d) \leq S_\varepsilon(X, d) \tag{9.2}$$

Box-counting dimensions. The *upper* and *lower box counting dimension* (also known as *Minkowski dimensions* or *metric dimensions*) of the metric space (X, d) are defined as

$$\overline{\dim}_b(X, d) := \limsup_{\varepsilon \searrow 0} -\frac{\log N_\varepsilon(X, d)}{\log \varepsilon}$$

$$\underline{\dim}_b(X, d) := \liminf_{\varepsilon \searrow 0} -\frac{\log N_\varepsilon(X, d)}{\log \varepsilon}$$

We get the same values if we substitute $S_\varepsilon(X, d)$ or $C_\varepsilon(X, d)$ to $N_\varepsilon(X)$ in the above formulas (use (9.1) and (9.2), and compare the counting functions at the values ε and 2ε).

For reasonable self-similar metric spaces the two limits coincide, and their common value $\dim_b(X)$ is simply called *box counting dimension*, and denoted by $\dim_b(X)$.

An important observation if that box dimensions do not change under scalings of the metric, i.e. if we measure distances as $d'(x, y) = \lambda\, d(x, y)$ instead of $d(x, y)$, for some fixed $\lambda > 0$.

ex: Show that the box-counting dimension of the n-dimensional cube $[0, 1]^n$ is what you expect, namely $\dim_b([0, 1]^n) = n$.

[29] A. N. Kolmogorov, On certain asymptotic characteristics of completely bounded metric spaces, *Dokl. Akad. Nauk SSSR* **108**, 3 (1956), 385-389.

[30] A.N. Kolmogorov and V.M. Tihomirov, ε-entropy and ε-capacity of sets in functional spaces, *Uspekhi Mat. Nauk* **14** (1959), 3-86. [Translated in *Amer. Math. Soc. Transl.*, series 2, **17** (1961), 277-364.]

ex: Show that the box counting dimension of the middle-third Cantor set is $\dim_b(K) = \log 2/\log 3$.

ex: Compute the box dimension of the space $\Sigma^+ = \mathcal{A}^{\mathbb{N}}$ of infinite words in an alphabet of z letters, equipped with the ultrametric

$$d_\lambda(x, y) = \lambda^{-\min\{k\geq 1 \text{ s.t. } x_k \neq y_k\}} .$$

Observe that a centered cylinder C_α, defined by a finite word $\alpha = \alpha_1\alpha_2\dots\alpha_n$ of $|\alpha| = n$ letters, is a closed ball $\overline{B_r(x)}$ of radius/diameter $r = \lambda^{-(n+1)}$ centered at any one of its points $x \in C_\alpha$, and that the distance between any two different centered cylinders C_α and C_β, defined by finite words of the same lenght $|\alpha| = |\beta| = n$, is $d(C_\alpha, C_\beta) \geq \lambda^{-n}$.

ex: Consider the unit interval $I = [0, 1]$ equipped with the Euclidean metric d, and define new metrics

$$d_\alpha(x, y) := d(x, y)^\alpha ,$$

for $\alpha \leq 1$. Verify that these are indeed metrics, and compute the box-counting dimension of the metric spaces (I, d_α).

Topological entropy. Bowen and Dinaburg adapted Kolmogorov's ideas to define an invariant of topological dynamical systems, which measures the asymptotic exponential rate of divergence of orbits. Let $f : X \to X$ be a continuous transformation of a metrizable topological space X. If d is a metric on X which induzes its toplogy, we may define a family of "dynamical metrics", depending on time $n \geq 0$, according to

$$d_n\left(x, y\right) := \max_{0\leq k\leq n} d\left(f^k\left(x\right), f^k\left(y\right)\right) \tag{9.3}$$

That is, $d_n\left(x, y\right)$ is the 'maximal distance between the n-trajectories of x and y".

If we fix a precision $\varepsilon > 0$, then $N_\varepsilon\left(X, d_n\right)$ is the "minimal number of n-orbits necessary to describe all the n-orbits with an error $\leq \varepsilon$", and $S_\varepsilon\left(X, d_n\right)$ is the "maximal number of n-orbits which an instrument with sensibility ε can distinguish". If X is compact, then these numbers are finite, and are monotone non-decresing as $n \nearrow \infty$ and $\varepsilon \searrow 0$.

The *topological entropy* of the continuous transformation $f : X \to X$ of the compact metric space X is finally defined as the exponential growth rate of $N_\varepsilon\left(X, d_n\right)$, namely, the iterated limit

$$h_{\text{top}}\left(f\right) := \lim_{\varepsilon\searrow 0}\ \limsup_{n\to\infty} \frac{\log N_\varepsilon(X, d_n)}{n} \tag{9.4}$$

By inequalities (9.1) and (9.2), we may also use $S_\varepsilon(X, d_n)$ or even the quantity $C_\varepsilon(X, d_n)$, the cardinality of a minimal ε-cover of (X, d_n). This is useful because one easily shows that

$$C_\varepsilon(X, d_{n+m}) \leq C_\varepsilon(X, d_n) \cdot C_\varepsilon(X, d_m) \tag{9.5}$$

Therefore, for any fixed $\varepsilon > 0$, the sequence $c_n = \log C_\varepsilon(X, d_n)$ is subadditive, i.e. satisfies $c_{n+m} \leq c_n + c_m$. By theorem 8.10, there exists the limit $\lim_{n\to\infty} c_n/n$. Thus, again by inequalities (9.1) and (9.2), the limsup in the definition (9.8) of topological entropy may be substituted by a limit, i.e.

$$h_{\text{top}}\left(f\right) = \lim_{\varepsilon\searrow 0}\ \lim_{n\to\infty} \frac{\log C_\varepsilon(X, d_n)}{n} .$$

Moreover, the different characterizations are useful to get upper bounds (from nets) and lower bounds (from separated sets) for the entropy, and therefore, in some simple cases where the two are equal, the exact value.

The notation suggests that the iterated limit which defines the topological entropy does not depend on the actual metric d, but only on the topology it induces on X. This is the case, at least for compact X's.

Theorem 9.8. *The topological entropy does not depend on the metric used to define the topology of the compact space X.*

Proof. Let d and d' two equivalent metrics generating the same topology of X. Since X is compact, the identity transformation is a uniformly continuous homeomorphisms between (X,d) and (X,d'). Thus, for any $\varepsilon > 0$ there exists $\delta > 0$ such that if $d'(x,y) < \delta$ then $d(x,y) < \varepsilon$. This clearly implies that $C_\varepsilon(X,d_n) \leq C_\delta(X,d'_n)$. Since the inverse homeomorphism is also uniformly continuous, the reverse inequality also holds. □

Theorem 9.9. *If $f : X \to X$ is a factor of $g : Y \to Y$, then*

$$h_{\text{top}}(f) \leq h_{\text{top}}(g)$$

In particular, topologically conjugated dynamical systems share the same topological entropy.

Proof. This is more or less the same argumet as before. Let d and d' be the metrics of X and Y, respectively. The semi-conjugation $h : Y \to X$ is uniformly continuous, because Y (and X) is compact. Therefore, for any $\varepsilon > 0$ there exists a $\delta > 0$ such that the h-image of a δ-ball in Y is contained in a ε-ball of X, i.e. $h(B_\delta(y)) \subset B_\varepsilon(h(y))$. This clearly implies that $C_\varepsilon(X,d_n) \leq C_\delta(Y,d'_n)$. Taking limits this shows the first result. The second follows changing the roles of f and g. □

For non-compact phase spaces X, one still can define the entropy taking the supremum over compact subsets $K \subset X$. Then one shows invariance under equicontinuous homeomorphisms (see, for example, [Wa82]).

It turns out that the second limit $\varepsilon \to 0$, in the definition of the topological entropy, is also unnecessary, provided the map is expansive and we take $\varepsilon > 0$ sufficiently small. We can explicitely see this phenomenon in the simple examples below.

ex: Let $f : X \to X$ be a topological dynamical system, and d_n, for $n \geq 1$, be the dynamical metrics defined in (9.3). Observe that is $A \subset X$ is a set of d_n-diameter $< \varepsilon$ and $B \subset X$ is a set of d_m-diameter $< \varepsilon$, then $A \bigcap f^{-n}(B)$ is a set of d_{n+m}-diameter $< \varepsilon$. Deduce inequality (9.5).

ex: Show that $h_{\text{top}}(f^n) = n\, h_{\text{top}}(f)$.

e.g. **Entropy of isometries.** Contraction, isometries, or, in general, Lipschitz maps $f : X \to X$ with Lipschitz constant ≤ 1, have zero entropy $h_{\text{top}}(f) = 0$. This is obvious since the dynamical metrics d_n do not depend on time n.

e.g. **Entropy of linear expanding endomorphisms of the circle.** Consider the expanding endomorphism $f : x + \mathbb{Z} \mapsto zx + \mathbb{Z}$ of the circle $\mathbb{T}$, with degree z such that $z \geq 2$ (a similar computation works also for orientation reversing expanding maps, having negative z). One explicitly computes $N_\varepsilon(\mathbb{T},d_n) \leq z^{n+m}$ and $S_\varepsilon(\mathbb{T},d_n) \geq z^{n+m}$, for $\varepsilon \approx z^{-m}$. More precisely, consider the set

$$A_{m+m} = \left\{ p_k = \frac{k}{z^{n+m}} + \mathbb{Z}\,, \text{ with } k = 0,1,2,\dots,z^{n+m}-1 \right\} ,$$

made of "dyadic" points of the circle with denominator z^{n+m}, of cardinality z^{n+m}. For m sufficiently large, any two successive points of A_{n+m} are at a distance $d_n(p_{k+1},p_k) = z^m$. Therefore, A_{n+m} is a ε-net for the metric d_n and $z^{-m-1} < \varepsilon \leq z^{-m}$, as well as a ε-separated set for the metric d_n and $z^{-m} < \varepsilon \leq z{-}m+1$. These estimates imply that

$$\lim_{\varepsilon\to 0} \limsup_{n\to\infty} \frac{\log N_\varepsilon(\mathbb{T},d_n)}{n} \leq \log z \qquad \text{and} \qquad \lim_{\varepsilon\to 0} \limsup_{n\to\infty} \frac{\log S_\varepsilon(\mathbb{T},d_n)}{n} \geq \log z \,.$$

Thus, the topological entropy of f is

$$h_{\text{top}}(f) = \log |z| \,.$$

e.g. Entropy of Bernoulli shifts and topological Markov chains. Consider the Bernoulli shift $\sigma : \Sigma^+ \to \Sigma^+$ over an alphabet $\mathcal{A}$ of $z \geq 2$ letters. Using the ultrametric $d(x,y) = z^{-\min\{k\geq 1\, x_k \neq y_k\}}$ (for which cylinders are clopen balls, centered at any of their points), one explicitely computes $N_\varepsilon(\Sigma^+, d_n) \approx z^{n+m}$ for $\varepsilon = z^{-m}$, since a ε-net for the metric d_n on Σ^+ is given by one point for each cylinder C_α with $|\alpha| = n+m$. Therefore,

$$h_{\mathrm{top}}(\sigma) = \log z\,.$$

Now, consider the topological Markov chain $\sigma_A : \Sigma_A^+ \to \Sigma_A^+$, induced by a $z \times z$ transition matrix A. As above, we must count the number of admissible sylinders C_α, i.e. defined by admissible words α, with $|\alpha| = n+m$, and this number is equal to $N_\varepsilon(\Sigma_A^+, d_n) = \sum_{i,j} (A^{n+m-1})_{ij}$, the number of Markov paths of lenght $n+m$, starting with i and ending with j. The above sum is clearly bounded from above and from below by

$$c\,\|A^{n+m-1}\| \leq \sum_{i,j} (A^{n+m-1})_{ij} \leq C\,\|A^{n+m-1}\|\,,$$

for some positive constants c and C. The right inequality is obvious. The left inequality comes from the fact that $\sum_{i,j}(A^{n+m-1})_{ij} = \|A^{n+m-1}\|_1$, since all the entries of A are non-negative, and the fact that all norms in a finite dimensional Euclidean space are equivalent. Therefore,

$$h_{\mathrm{top}}(\sigma_A) = \lim_{n\to\infty} \frac{1}{n} \log \|A^n\| = r(A)\,,$$

where the *spectral radius* $r(A)$ is the maximal absolute value of the eigenvalues of A.

10 Ergodicity and convergence of time means

10.1 Ergodicity

Ergodic maps. Let $f : X \to X$ be an endomorphism of the measurable space $(X, \mathcal{E})$. The invariant probability measure μ is said *ergodic* if any of the following equivalent conditions is satisfied:

i) for any observable $\varphi \in L^1(\mu)$, the time average

$$\overline{\varphi}(x) = \lim_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} \varphi\left(f^k(x)\right)$$

exists and is equal to the mean value $\int_X \varphi d\mu$ for μ-almost any $x \in X$,

ii) any invariant event $A \in \mathcal{E}$ has probability $\mu(A) = 0$ or 1, namely the invariant σ-algebra $\mathcal{E}_f$ is equal to the trivial σ-algebra $\mathcal{N}$ generated by events of zero measure,

iii) any invariant (measurable) observable φ is constant μ-a.e.

If this happens, one also says that f is an *ergodic endomorphism* of the probability space $(X, \mathcal{E}, \mu)$.

Condition i) is the physical meaning of ergodicity, as it says that "time averages are almost everywhere constant and equal to space averages". In particular, taking φ equal to the characteristic function of any event A, almost any trajectory spend in A a fraction of time asymptotically proportional to $\mu(A)$, as dreamed by Boltzmann in his "ergodic hypothesis".

Conditions ii) or iii) are what one usually check in order to prove ergodicity of a probability measure.

Proof. (of the equivalence) To see that i) $\Rightarrow$ ii), let A be an invariant event, and φ its characteristic function. Invariance of A implies that φ is invariant, hence that $\overline{\varphi} = \varphi$. There follows fom i) that $\mu(A) = \int_X \varphi d\mu = \varphi(x)$ for some $x \in X$, hence that $\mu(A) = 0$ or 1, the only values of characteristic functions.

Conditions ii) and iii) are clearly equivalent, since any invariant event defines an invariant function (its characteristic function), and conversly level sets of invariant functions are invariant events.

Finally, in order to show that iii) $\Rightarrow$ i), let $\varphi \in L^1(\mu)$ be an integrable observable. According to the Birkhoff-Khinchin ergodic theorem 7.9, the time average $\overline{\varphi}(x)$ exists for μ-almost any $x \in X$ and $\int_X \overline{\varphi} d\mu = \int_X \varphi d\mu$. Since $\overline{\varphi}$ is invariant mod 0, by iii) it is constant with probability one. This implies that $\overline{\varphi}(x) = \int_X \varphi d\mu$ for μ-almost any $x \in X$. □

Warning. Ergodic dynamical systems exist, and some are listed below. On the other side, to show that a physically interesting system is ergodic turns out to be extremely difficult, and very few examples are known. The most famous are some "billards", systems of hard spheres inside a billard table interacting via elastic collisions, studied by Yakov Sinai in the sixties...

Ergodic measures as extremal measures. We already saw that the space Prob_f of invariant probability measure is a convex and closed subset of the compact space Prob. Here, we observe that ergodic measures are the "indecomposable" elements of this set.

Theorem 10.1. *Ergodic invariant measures are the extremals of* Prob_f. *Namely, an invariant measure μ is ergodic iff it cannot be written as a convex combination*

$$\mu = t\mu_1 + (1-t)\,\mu_0$$

where $t \in (0,1)$ of two distinct invariant measures μ_0 and μ_1

Proof. First, observe that if ν is an invariant measure which is absolutely continuous w.r.t. the ergodic measure μ, then $\nu = \mu$. Indeed, one easily verifies that the Radon-Nykodim derivative $\rho = d\nu/d\mu$ is an invariant function, and ergodicity of μ implies that it is constant and equal to one

μ-a.e. Now, let μ be an ergodic measure, and assume that $\mu = t\mu_1 + (1-t)\,\mu_0$ for some $t \in (0,1)$. Since both μ_0 and μ_1 are absolutely continuous w.r.t. μ, they coincide with μ, hence, are not different. To prove the converse, assume that the invariant measure μ is not ergodic, hence there exists an invariant event C such that $0 < \mu(C) < 1$. Let μ_0 and μ_1 be the "conditional probability measures" defined as $\mu_1(A) = \mu(A \cap C)/\mu(C)$ and $\mu_0(A) = \mu(A \cap C^c)/\mu(C^c)$. Clearly they are different, both are invariant, and $\mu = \mu(C)\,\mu_1 + (1-\mu(C))\,\mu_0$. □

Ergodic decomposition. In the first lines of the above proof, we actually showed that any two ergodic invariant measure μ and ν are either equal or "mutually singular", namely, if $\mu \neq \nu$ then there exists a measurable set A such that $\mu(A) = \nu(A^c) = 1$ and $\mu(A^c) = \nu(A) = 0$. This suggests that maybe any invariant measure could be "disintegrated" along a partition whose atoms are the support of all the different ergodic measure, in other word that μ is a "convex combination", namely an integral, of the ergodic measures. This is true, sometimes, but both its statement and proof are quite technical: we just quote the result.

Theorem 10.2 (Ergodic decomposition)**.** *Let $f : X \to X$ be a continuous transformation of the compact metrizable space X There exists a partition $\mathcal{P} = \{P_e\}_{e\in E}$ of X (modulo sets of zero measure) into invariant measurable sets indexed by a Lebesgue space E, and a measurable map $E \ni e \mapsto \mu_e \in Prob_f$ with values in the space of ergodic Borel probability measures and with the property that $\mu_e(P_e) = 1$ for any $P_e \in \mathcal{P}$, such that any invariant Borel probability measure μ can be written as an integral*

$$\mu = \int_E \mu_e \, d\overline{\mu}(e)$$

where $\overline{\mu}$ is some probability measure on E.

Observe that the above theorem contains the statement that any continuous transformation of a compact space admits at least one ergodic Borel probability measure.

10.2 Examples of ergodic maps

Bernoulli shift. Let $\sigma : \Sigma^+ \to \Sigma^+$ be the Bernoulli shift over the alphabet $\mathcal{A} = \{1, 2, ..., z\}$, let $p = \{p_1, p_2, \ldots, p_z\}$ be any probability on $\mathcal{A}$, and μ the Bernoulli invariant measure defined by p.

Theorem 10.3. *The Bernoulli invariant measure μ is ergodic w.r.t. σ^+.*

Proof. First observe that, given two centered cylinders C_α and C_β, the definition of μ implies that there exists a time $n \geq 1$ such

$$\mu\left(C_\alpha \cap \sigma^{-k}(C_\beta)\right) = \mu(C_\alpha) \cdot \mu\left(\sigma^{-k}(C_\beta)\right) = \mu(C_\alpha) \cdot \mu(C_\beta)$$

whenever $k \geq n$. Indeed, one can take $n = |\alpha| + 1$, and the above reflect the "independence" of the different trials encoded in the construction of the Bernoulli measure. By aditiviity, the same holds true for any couple of elements of $\mathcal{A}$, the algebra made of finite unions of centered cylinders. Now, assume that $A \in \mathcal{B}$ is invariant. Since any Borel set $A \in \mathcal{B}$ can be aproximated in measure by an elements of $\mathcal{A}$, given any $\varepsilon > 0$ one can find an $A_\varepsilon \in \mathcal{A}$ such that $\mu(A \Delta A_\varepsilon) < \varepsilon$. Using the above result, we can find an $n \geq 1$ such that

$$\mu\left(A_\varepsilon \cap \sigma^{-n}(A_\varepsilon)\right) = \mu(A_\varepsilon) \cdot \mu\left(\sigma^{-n}(A_\varepsilon)\right) = \mu(A_\varepsilon)^2$$

where the last equality comes from invariance of μ. Then, observe that the symmetric difference between $A \cap \sigma^{-n}(A)$ and $A_\varepsilon \cap \sigma^{-n}(A_\varepsilon)$ is contained in $(A \Delta A_\varepsilon) \cup \sigma^{-n}(A \Delta A_\varepsilon)$. This gives

$$\begin{aligned} \left|\mu\left(A \cap \sigma^{-n}(A)\right) - \mu\left(A_\varepsilon \cap \sigma^{-n}(A_\varepsilon)\right)\right| &\leq \mu(A \Delta A_\varepsilon) + \mu\left(\sigma^{-n}(A \Delta A_\varepsilon)\right) \\ &\leq 2 \cdot \mu(A \Delta A_\varepsilon) < 2\varepsilon \end{aligned}$$

which, together with
$$\left|\mu\left(A\right)^2-\mu\left(A_\varepsilon\right)^2\right|\leq 2\cdot\mu\left(A\Delta A_\varepsilon\right)<2\varepsilon$$
gives
$$\left|\mu\left(A\right)-\mu\left(A\right)^2\right|<4\varepsilon$$
Since $\varepsilon>0$ was arbitrary, we just showed that the measure of any invariant Borel set A satisfies $\mu\left(A\right)=\mu\left(A\right)^2$, hence it is either 0 or 1. □

Observe that this proof is very similar to the argument in the Kolmogorov zero-one law for tail events in the theory of stochastic processes.

Now, let φ_k be the the characteristic function of $\{x\in\Sigma^+ \text{ s.t. } x_1=k\}$. The observables $\varphi_k\circ\sigma^n$ form a sequence of independent and identically distributed random variables with mean p_k. One can interprete the event $\{\varphi_k\circ\sigma^n=1\}=\{x\in\Sigma^+ \text{ s.t. } x_n=k\}$ as "sucess in the n-th trial", where the probability of sucess in each trial is p_k. The Birkhoff-Khinchin ergodic theorem, together with the ergodicity of μ, gives the result that
$$\mu\left\{x\in\Sigma^+ \text{ s.t. } \frac{1}{n+1}\left(\varphi_k+\varphi_k\circ\sigma^1+\varphi_k\circ\sigma^2+...+\varphi_k\circ\sigma^n\right)(x)\to p_k\right\}=1$$
which is the Kolmogorov strong law of large numbers.

Expanding endomorphisms of the circle. Let $f:x+\mathbb{Z}\mapsto z\cdot x+\mathbb{Z}$ be an expanding endomorphism of the circle $\mathbb{T}=\mathbb{R}/\mathbb{Z}$ of degree z such that $|z|>1$. Lebesgue probability measure ℓ si clearly invariant under f. We claim that

Theorem 10.4. *Lebesgue probability measure ℓ is an ergodic measure for f.*

Proof. To prove ergodicity, let A be an invariant Borel set, and assume that $\ell\left(A\right)<1$. We must show that the complement $B=\mathbb{T}\backslash A$, that has positive measure, has indeed probability one. The argument goes as follows: if $\ell\left(B\right)>0$, then, according to Lebesgue density theorem, B contains nearly all the mass of some nonempty interval. Namely, given any $\varepsilon>0$, we can find an open interval I_n with lenght $\ell\left(I_n\right)=|z|^{-n}$ and centered at a density point of B such that
$$\ell\left(B\cap I_n\right)>\left(1-\varepsilon\right)\cdot\ell\left(I_n\right)$$
Now observe that the restriction $f^n\left|_{I_n}\right.$ is an injective map sending I_n onto the circle minus one point, in particular, $\ell\left(f^n\left(I_n\right)\right)=1$. Since f uniformly dilatates lenghts by a factor $|z|$, there follows that
$$\frac{\ell\left(f^n\left(B\cap I_n\right)\right)}{\ell\left(f^n\left(I_n\right)\right)}=\frac{\ell\left(B\cap I_n\right)}{\ell\left(I_n\right)}$$
Since, moreover, A is invariant, its complement B is +invariant, and this implies that the left-hand side above is equal to $\ell\left(B\right)$. There follows that
$$\ell\left(B\right)=\frac{\ell\left(B\cap I_n\right)}{\ell\left(I_n\right)}>\left(1-\varepsilon\right)$$
and, since ε was arbitrary, that $\ell\left(B\right)=1$. □

Hyperbolic automorphisms of a torus. Let $f_A:\mathbb{T}^2\to\mathbb{T}^2$ be an hiperbolic automorphism of the torus, induced by the unimodular matrix $A\in\mathrm{SL}_2(\mathbb{Z})$. Lebesgue measure on the torus is an invariant probability measure, since $\det(A)=1$.

Theorem 10.5. *Lebesgue measure is an ergodic invariant measure for f_A.*

Proof. Let φ be an invariant square integrable function in $\mathbb{T}^2$, and consider its Fourier series

$$\varphi(x) \sim \sum_{k\in\mathbb{Z}^2} \widehat{\varphi}(k) e^{2\pi i k\cdot x}$$

One compute

$$(\varphi \circ f_A)(x) = \sum_{k\in\mathbb{Z}^2} \widehat{\varphi}(k) e^{2\pi i k\cdot Ax} .$$

Since φ is invariant, the Fourier coefficients must verify $\widehat{\varphi}(k) = \widehat{\varphi}(kA)$, and consequently

$$\widehat{\varphi}(k) = \widehat{\varphi}(kA^n) \tag{10.1}$$

for all $n \in \mathbb{N}$ and $k \in \mathbb{Z}^2$. Fix a wave number $k \in \mathbb{Z}^2$ different from 0. If the sequence of integer vectors kA^n were bounded, then some of the $v = kA^n$ would be periodic, say $vA^m = v$ (since bounded integer vectors are finite). But then A would have an eigenvalue λ which is a root of one, i.e. $\lambda^m = 1$. Since A is hyperbolic, this canot happen. There follows that $\|kA^n\| \to \infty$ for all $k \neq 0$. By the Riemann-Lebsesgue lemma, $\widehat{\varphi}(k) \to 0$ as $\|k\| \to \infty$, and therefore, by invariance (10.1), all the non-zero Fourier coefficients vanishes, i.e. $\widehat{\varphi}(k) = 0$ for all $k \neq 0$, since

$$\widehat{\varphi}(k) = \widehat{\varphi}(kA^n) \to 0$$

as $n \to \infty$. There follows that φ is a constant function. □

10.3 Normal numbers

Normal numbers. In particular, Lebesgue measure ℓ is ergodic w.r.t. multiplication by 10 in the unit circle, the map $f(x + \mathbb{Z}) = 10 \cdot x + \mathbb{Z}$. Identify the circle with the interval $[0, 1[$, and let $x = 0, x_1x_2x_3...$ be the base 10 expression of a point of the circle, which is unique outside a subset of Lebesgue measure zero. For $k = 0, 1, 2, ..., 9$, let φ_k be the characteristic function of the interval $[k/10, (k + 1)/10)$, i.e. the observable which is equal to $\varphi_k(x) = 1$ if $x_1 = k$ and $\varphi_k(x) = 0$ otherwise. The time mean of φ_k is

$$\frac{1}{n+1}\sum_{j=0}^{n} \varphi_k\left(f^j(x)\right) = \frac{1}{n+1} \cdot \mathrm{card}\left\{1 \le j \le n+1 \text{ s.t. } x_j = k\right\}$$

that is the number of k's within the first $n + 1$ digits of the decimal expansion of x. The limit as $n \to \infty$, if it exists, is the "asymptotic frequency" of k's contained in the expansion of x. Ergodicity of μ implies that there exists a set $A_k \subset [0, 1[$ of Lebesgue measure one where the limit $\overline{\varphi_k}(x)$ exists and is equal to $\int \varphi_k d\ell = 1/10$. Since the intersection $A_0 \cap A_1 \cap ... \cap A_9$ has still probability one, the result is that Lebesgue almost any number $x \in [0, 1[$ contains in its decimal expansion any of the letters $0, 1, 2, ..., 9$ with asymptotic frequency $1/10$.

Actually, one could repeat the same argument considering any finite word $b = b_1b_2...b_n$ in the alphabeth $\{0, 1, 2, ..., 9\}$, and show that there is a set $A_b \subset [0, 1[$ of probability one such that the base 10 expansion of any $x \in A_b$ contains the word b with asymptotic frequency 10^{-n}. A real number x whose base 10 expansion contains any finite word with the right asymptotic frequency is called 10-*normal* (meaning "normal in base 10"). Since finite words in the alphabeth $\{0, 1, 2, ..., 9\}$ are countable, and a countable union of zero measure sets still has zero measure, we just showed that Lebesgue almost any real number is normal in base 10. Indeed, as first observed by Émile Borel[31],

Theorem 10.6 (Borel)**.** *Lebesgue almost any real number is normal in every base $d \geq 2$.*

It is not so easy to give examples of normal numbers, actually of series whose sum is a normal number. Much more difficult is to show that a "given" number, such as π, $\sqrt{2}$ or e ..., is normal. Here we quote Mark Kac:[32]

[31] E. Borel, Les probabilités dénombrables et leurs applications arithmétiques, *Rendiconti del Circolo Matematico di Palermo* **27** (1909), 247-271.

[32] Mark Kac, *Statistical independence in probability, analysis, and number theory*, Carus Math. Monographs, **12**, New York 1959 (pag. 18).

"As is often the case, it is much easier to prove that an overhelming majority of objects possess a certain property that to *exhibit* even one such object. The present case is no exception. It is quite difficult to exhibit a 'normal' number! The simplest example is the number (written in decimal notation) $x = 0.1234567891011\dots$ where after the decimal point we write the positive integers in succession. The proof that this number is normal is by no means trivial."

10.4 Distribution of digits in continued fractions

Continued fractions and Gauss map. Numbers in the unit interval $(0,1]$ are uniquely represented, i.e. "coded", by continued fractions. If we disregard rationals, which form a set of zero Lebesgue measure, we are left with infinite continued fractions $[0; a_1, a_2, a_3, \dots]$, i.e. one-sided infinite sequences $(a_1, a_2, a_3, \dots) \in \mathbb{N}^{\mathbb{N}}$. Recall that the *Gauss map* $G : (0,1] \to [0,1]$ is defined as

$$G(x) := 1/x - \lfloor 1/x \rfloor \qquad \text{if } x \neq 0$$

(but we may also define $G(0) = 0$). Observe that for any rational $r \in \mathbb{Q}$ there exists a time n such that $G^n(r) = 0$. The infinite sequence of the continued fraction expansion of $x \sim [0; a_1, a_2, a_3, \dots] \in [0,1]\backslash\mathbb{Q}$ is a coding of the orbit of x. Indeed,

$$G([0; a_1, a_2, a_3, \dots]) = [0; a_2, a_3, a_4, \dots]$$

This means that $a_n = \lfloor 1/G^{n-1}(x) \rfloor$, or, equivalently, $a_n = k$ if $G^n(x) \in \left[\frac{1}{k+1}, \frac{1}{k}\right)$. In the language of dynamical systems, the Gauss map (restricted to the full measure set of irrationals) is conjugated to the one-sided shift $\sigma : \mathbb{N}^{\mathbb{N}} \to \mathbb{N}^{\mathbb{N}}$, the conjugation being the continued fraction representation $x \sim [0; a_1, a_2, a_3, \dots]$. In particular, the equivalence relation coming from the action of $\mathrm{PSL}_2(\mathbb{Z})$ corresponds to "being in the same great orbit" of the Gauss map.

Ergodicity and distribution of digits. It is essentially due to Gauss himself the crucial observation that the absolutely continuous measure μ with density

$$d\mu(x) = \frac{1}{\log 2}\,\frac{1}{1+x}\,dx \tag{10.2}$$

is an invariant probability measure for G, meaning that $\mu(G^{-1}(B)) = \mu(B)$ for all Borel subsets $B \in (0,1]$. It is sufficent to check invariance for intervals. The measure of an interval $[a,b]$ is

$$\frac{1}{\log 2}\int_a^b \frac{dx}{1+x} = \frac{1}{\log 2}\log\frac{1+b}{1+a}\,.$$

The preimage $G^{-1}([a,b])$ is a union of intervals $\left[\frac{1}{b+n}, \frac{1}{a+n}\right]$ with $n = 1,2,3,\dots$. Therefore, its measure is

$$\begin{aligned}
\frac{1}{\log 2}\sum_{n=1}^{\infty}\int_{\frac{1}{b+n}}^{\frac{1}{a+n}} \frac{dx}{1+x} &= \frac{1}{\log 2}\sum_{n=1}^{\infty}\log\frac{1+\frac{1}{a+n}}{1+\frac{1}{b+n}} \\
&= \frac{1}{\log 2}\sum_{n=1}^{\infty}\log(1+a+n) - \log(a+n) + \log(b+n) - \log(1+b+n) \\
&= \frac{1}{\log 2}\log\frac{1+b}{1+a}
\end{aligned}$$

Indeed, more is true [33]

Theorem 10.7 (Knopp, 1926)**.** *The Gauss measure μ is ergodic for the Gauss map.*

[33] K. Knopp, Mengentheoretische Behandlung einiger Probleme der diophantischen Ap- proximationen und der transfiniten Wahrscheinlichkeiten, *Math. Ann.* **95** (1926), 409-426.

There follows from the Birkhoff-Khinchin ergodic theorem 7.9 that time-averages of integrable observables φ converge μ-a.e. and are equal to the μ- averages, i.e.

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \varphi(G^n(x)) = \int_0^1 \varphi(x)\, d\mu(x) \qquad \mu \text{ - a.e.}$$

In particular, we may compute the distribution of digits in the continued fraction representation of a number in the unit interval, simply taking for φ the indicator function of some digit $d \in \mathbb{N}$ in $\mathbb{N}^{\mathbb{N}} \approx (0,1]$. The result is the *Gauss-Kuzmin distribution* (conjectured by Gauss and proved by Kuzmin[34], see also [Ar78]):

Theorem 10.8 (Gauss-Kuzmin, 1928)**.** *For almost every real number x, the asymptotic frequency of the digit d in the continued fraction representation $x \sim [a_0; a_1, a_2, a_3, \dots]$ is*

$$p_d = \frac{1}{\log 2} \log\left(1 + \frac{1}{d(d+2)}\right)$$

The ergodicity of the Gauss map w.r.t. the Gauss measure imply many other "surprising" results, for clever choices of the observable φ. For example, if we choose $\varphi(x) = \log(a_1)$ then the Birkhoff averages are the geometric means of the first n partial quotients. There follows that for almost all numbers $x \sim [a_0; a_1, a_2, a_3, \dots]$ the limit $\lim_{n\to\infty} \sqrt[n]{a_1 a_2 a_3 \dots a_n}$ exists and is a constant, equal to

$$\prod_{n=1}^{\infty} \left(1 + \frac{1}{n(n+2)}\right)^{\log_2 n} \simeq 2.6854\dots,$$

a number now called *Khinchin constant* [Kh35]. A similar result is: the n-th root of the denominators q_n of the convergents of almost all numbers converge to

$$\lim_{n\to\infty} \sqrt[n]{q_n} = e^{\pi^2/(12\log 2)} \simeq 3.2758\dots,$$

a number called *Khinchin-Lévy constant* [35] [36]. On the other hand, the arithmetic mean of the partial quotients is unbounded for almost all numbers.

10.5 Unique ergodicity and equidistribution

Unique ergodicity. A homeomorphism $f : X \to X$ of a compact metric space (X, d) is *uniquely ergodic* if it admits one, and only one, invariant Borel probability measure μ. The above discussion implies that this unique invariant measure is ergodic.

This notion is the probabilistic counterpart of minimality, and indeed both minimality and unique ergodicity are often observed simultaneously (this means that, although equivalence of the two is false, it is not easy to think at a couterexample!). Observe that we defined unique ergodicity in the context of continuous transformations. The reason is that this notion is interesting due to the following[37]

Theorem 10.9 (Oxtoby)**.** *Let $f : X \to X$ be a homeomorphism of a compact metric space X. The following statements are equivalent:*

i) f is a uniquely ergodic,

ii) there exists an invariant Borel probability measure μ such that, for any continuous observable φ, the time averages $\overline{\varphi}(x)$ exist and are equal to $\int_X \varphi d\mu$ for any initial condition $x \in X$.

iii) there exists an invariant Borel probability measure μ such that, for any continuous observable φ, the convergence

$$\frac{1}{n+1}\sum_{k=0}^{n} \varphi\left(f^k(x)\right) \to \int_X \varphi d\mu$$

as $n \to \infty$ holds and is uniform in $x \in X$.

[34] R.O. Kuzmin, Ob odnoi zadache Gaussa, *Doklady akad. nauk,* ser. A (1928), 375-380.

[35] A.Y. Khinchin, Einige Sätze über Kettenbrüche, mit Anwendungen auf die Theorie der Diophantischen Approximationen, *Math. Ann.* **92** (1924), 115-125.

[36] P. Lévy, Sur les lois de probabilité dont dependent les quotients complets et incomplets d'une fraction continue, *Bull. Soc. Math.* **57** (1929), 178-194.

[37] J.C. Oxtoby, Ergodic sets, *Bull. Amer. Math. Soc.* **58** (1952), 116-136.

Weyl equidistribution theorem. The classical example of equidistribution was discovered by Hermann Weyl[38], and refines Dirichlet and Kronecker theorems, 8.1 and 8.8, on irrational rotations of the circle.

Theorem 10.10 (Weyl, 1916)**.** *An irrational rotation of the circle is uniquely ergodic.*

Proof. Let $R_\alpha : x + \mathbb{Z} \mapsto x + \alpha + \mathbb{Z}$ with $\alpha \notin \mathbb{Q}$. We must check that time means of continuous observables φ converge uniformly to the average $\int_0^1 \varphi \, dx$. According to Weierstrass theorem, trigonometric polinomials are dense in the space of continuous functions of the circle. Trigonometric polynomials are finite superpositions of the characters $e_k(x + \mathbb{Z}) := e^{i2\pi kx}$, with $k \in \mathbb{Z}$ (the characters of the abelian group $\mathbb{R}/\mathbb{Z}$). Therefore, by a simple triangular argument, it suffices to check that uniform convergence of Birkhoff sums holds for any of the e_k. A computation gives, for $k \neq 0$,

$$\left| \frac{1}{n+1} \sum_{j=0}^{n} e_k\left(R_\alpha^j(x + \mathbb{Z})\right) \right| = \left| \frac{1}{n+1} \sum_{j=0}^{n} e^{i2\pi kj\alpha} \right| \leq \frac{2}{n+1} \cdot \frac{1}{|1 - e^{i2\pi k\alpha}|} \to 0$$

as $n \to \infty$, uniformly in x. On the other side, it is obvious that time averages of the constant character e_0 are constant and equal to 1. $\square$

The theorem owes its name to the fact that

$$\frac{1}{n+1} \sum_{j=0}^{n} \varphi(x + j\alpha) \to \int_0^1 \varphi dx$$

uniformly for any continuous function φ on the circle, and this is interpreted as saying that the sequence of points $\{x, x + \alpha, x + 2\alpha, x + 3\alpha, ...\}$ is "equidistributed" w.r.t. Lebesgue measure.

We also observe that the convergence of time means also holds for Riemann integrable functions, since any such function ψ can be approximated by a couple of continuous functions $\varphi_- \leq \psi \leq \varphi_+$ such that the mean $\int (\varphi_+ - \varphi_-)dx$ is arbitrarily small.

On the other side, mean values of Lebesgue masurable functions need not converge. For example, the time mean of the characteristic function of the orbit of a point of the circle converge to one, while its mean value is clearly zero, since the orbit is countable.

Weyl's theorem extends to higher-dimensional tori. Here we state a version for flows.

Linear flows on tori. Consider the torus $X = \mathbb{R}^n/\mathbb{Z}^n$ of dimension $n \geq 2$, and the linear flow $\phi_t : x + \mathbb{Z}^n \mapsto x + t\alpha + \mathbb{Z}^n$ defined by the differential equation

$$\dot{x} = \alpha$$

where $\alpha \in \mathbb{R}^n$. The "frequency vector" $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$ is said *non-resonant* if the scalar product $\langle k, \alpha \rangle = \sum_{j=1}^n \alpha_j k_j \neq 0$ for any $k \in \mathbb{Z}^n \backslash \{0\}$. As above, one can approximate any continuous function on the torus with trigonometric polynomials. One then checks that

$$\frac{1}{T} \int_0^T e^{i2\pi \langle k, x + t\alpha \rangle} dt = \frac{e^{i2\pi \langle k, x \rangle}}{i2\pi \langle k, \alpha \rangle} \frac{e^{i2\pi \langle k, \alpha \rangle T} - 1}{T} \to 0$$

as $T \to \infty$, for any $k \in \mathbb{Z}^n \backslash \{0\}$, while the time mean of the observable 1 is constant and equal to one. There follows that

Theorem 10.11. *A non-resonant linear flow on the torus is uniquely ergodic w.r.t. to Lebesgue measure.*

e.g. **Digits of powers of two.** Look at the sucessive powers of two, written in base ten:

2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768, 65536,
131072, 262144, 524288, 1048576, 20197152, 4194304, 8388608, 16777216, . . .

[38] H. Weyl, Über die Gleichverteilung von Zahlen mod. Eins, *Math. Ann.* **77** (1916), 313-352.

The last digit recurrs every four iterations. This happens also to the last two digits, although with a much larger period. The reason is quite dull, since there are a finite number of possibilities and the digits on the left do not interfere. More interesting is to observe the first digit. Although the initial time serie

$$2\,,4\,,8,\ 1,\ 3,\ 6,\ 1,\ 2,\ 5,\ 1,\ 2,\ 4,\ 8,\ 1,\ 3,\ 6,\ 1,\ 2,\ 5,\ 1,\ 2,\ 4,\ 8,\ 1,\ \dots$$

looks periodic, this is just an accident of the first few iterations. Moreover, any of the letters $k = 1, 2, \dots, 9$ will eventually appear, and with a definite asymptotic frequency. To see this, observe that the first digit of 2^n is equal to k iff

$$k \cdot 10^m \leq 2^n < (k+1) \cdot 10^m$$

for some integer $m \geq 0$, i.e. iff

$$\log_{10} k + m \leq n \log_{10} 2 < \log(k+1) + m$$

If we denote $\alpha = \log_{10} 2$, which is irrational, then the above inequality means that the image $R_\alpha^n(0 + \mathbb{Z})$ of the origin under the n-th iterate of the irrational rotation R_α belongs to the interval $I_k = [\log_{10} k, \log_{10}(k+1))$ of the unit circle $\mathbb{R}/\mathbb{Z}$. The lenght of this interval is

$$|I_k| = \log_{10}\left(1 + \frac{1}{k}\right) .$$

By Weyl theorem 10.10, if $C_k(N)$ counts the number of times that the letters k appears as the first digit of 2^n for $1 \leq n \leq N$, then

$$\lim_{N\to\infty} \frac{C_k(N)}{N} \to \log_{10}\left(1 + \frac{1}{k}\right)$$

as $n \to \infty$. Thus, for example, the letter 1 appears about 30% of times, while the letter 7 appears only less than 6% of times.

References

[AA67] V.I. Arnold and A. Avez, *Problèmes ergodiques de la mécanique classique*, Gauthier-Villars, 1967.

[Ap69] T.M. Apostol, *Calculus*, John Wiley & Sons, New York 1969.

[Ar78] V.I. Arnold, *Metodi geometrici della teoria delle equazioni differenziali ordinarie*, Editori Riuniti - MIR, Roma 1978.

[Ar79] V.I. Arnold, *Metodi matematici della meccanica classica*, Edizioni MIR - Editori Riuniti, Roma 1979.

[Ar85] V.I. Arnold, *Equações diferenciais ordinárias*, MIR 1985.

[AS64] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, 1964.

[Ax97] S. Axler, *Linear Algebra Done Right*, second edition, Springer, 1997.

[Bi65] P. Billingsley, *Ergodic Theory and Information*, Wiley, 1965.

[BV12] L. Barreira e C. Valls, *Sistemas dinâmicos, uma introdução*, IST Press, 2012.

[BN05] P. Buttà e P. Negrini, *Note del corso di Sistemi Dinamici*, Università di Roma "La Sapienza", 2005.

[BS03] M. Brin and G. Stuck, *Introduction to Dynamical Systems*, Cambridge University Press, 2003.

[CG93] L. Carleson and T.W. Gamelin, *Complex dynamics*, UTX, Springer-Verlag, 1993.

[Chaos] P. Cvitanović, R. Artuso, P. Dahlqvist, R. Mainieri, G. Tanner, G. Vattay, N. Whelan and A. Wirzba, *Chaos: Classical and Quantum*, `http://ChaosBook.org` (Niels Bohr Institute, Copenhagen 2008).

[CR48] R. Courant and H. Robbins, *What is mathematics?*, Oxford University Press, 1948.

[De89] R.L. Devaney, *An introduction to chaotic dynamical systems*, Addison-Wesley, 1989.

[De92] R.L. Devaney, *A first course in chaotic dynamical systems*, Addison-Wesley, 1992.

[EW10] M. Einsiedler and T. Ward, *Ergodic Theory with a view towards Number Theory*, GTM **259**, Springer, 2010.

[Fa85] K. J. Falconer, *The geometry of fractal sets*, Cambridge University Press, 1985.

[Fe63] R.P. Feynman, R.B. Leighton and M. Sands, *The Feynman lectures on physics*, Addison-Wesley, Reading, 1963.

[Gh07] E. Ghys, Résonances et petits diviseurs, in *L'héritage scientifique de Kolmogorov*, Berlin 2007.

[GL] E. Ghys and J. Leys, Lorenz and Modular Flows: A Visual Introduction

[Ha74] P. Halmos, *Measure theory*, Springer-Verlag. New York 1974.

[Har49] G.H. Hardy, *Divergent series*, Oxford University Press, 1949.

[HK03] B. Hasselblatt and A. Katok, *A first course in dynamics: with a panorama of recent developments*, Cambridge University Press, 2003.

[HS74] M.W. Hirsch and S. Smale, *Differential equations, dynamical systems and linear algebra*, Academic Press (Pure and Applied Mathematics. A series of Monographs and Textbooks), San Diego 1974.

[HSD04] M.W. Hirsch, S. Smale and R.L. Devaney, *Differential Equations, Dynamical Systems, and an Introduction to Chaos*, 2nd ed., Elsevier Academic Press, 2004.

[HW59] G.H. Hardy and E.M. Wright, *An Introduction to the Theory of Numbers*, fourth edition, Oxford University Press, 1959.

[Kh35] A.Ya. Khinchin, *Continued Fractions*, 1935 [translation by University of Chicago Press, 1954].

[KH95] A. Katok and B. Hasselblat, *Introduction to the modern theory of dynamical systems*, Encyclopedia of mathematics and its applications, Cambridge University Press, 1995.

[Kr84] L. Kronecker, Die Periodensysteme von Funktionen Reeller Variablen, *Berliner Sitzungsberichte* (1884), 1071-1080.

[La87] S. Lang, *Linear Algebra*, Third Edition, UTM Springer, 1987.

[Ma75] B. Mandelbrot, *Les object fractals: forme, hasard, et dimension*, Flammarion, Paris 1975.

[Mat95] P. Mattila *Geometry of Sets and Measures in Euclidean Spaces: Fractals and rectifiability*, Cambridge University Press, 1995.

[Mi91] J. Milnor, *Dynamics in one complex variable*, IMS preprint, 1991.

[MS93] W. de Melo and W. van Strien, *One-Dimensional Dynamics*, Springer-Verlag, 1993.

[MSW02] D. Mumford, C. Series and D. Wright, *Indra's Pearls: The Vision of Felix Klein*, Cambridge University Press, 2002.

[PM78] J. Palis jr. e W. de Melo, *Introdução aos sistemas dinâmicos*, Projeto Euclides, IMPA, 1978.

[Ro99] J.C. Robinson, *Dynamical Systems, Stability, Symbolic Dynamic and Chaos*, CRC Press, Cambridge 1999.

[Ro04] J.C. Robinson, *An introduction to ordinary differential equations*, Cambridge University Press, 2004.

[Ru87] W. Rudin, *Real and complex analysis*, McGraw-Hill, 1987.

[Sm67] S. Smale, Differentiable dynamical systems, *Bull. of the AMS* **73** (1967), 747-817.

[SS03] E.M. Stein and R. Shakarchi, *Fourier Analysis. An Introduction*, Princeton University Press, 2003.

[St94] S.H. Strogatz, *Nonlinear Dynamics and Chaos*, Addison-Wesley, 1994.

[Wa82] P. Walters, *An Introduction to ergodic theory*, Graduate Texts in Math. **79**, Springer-Verlag, 1982