DYNAMICAL SYSTEMS FROM ODE'S TO ERGODIC THEORY

Carlangelo Liverani
Dipartimento di Matematica
Università di Roma "Tor Vergata"

Rome, February 11, 2026

Contents

List of Figures

Chapter 1

The origins: Differential equations

As this book is about Dynamical Systems, let's start by defining the object of study. The concept of Dynamical System is a very general one and it appears in many branches of mathematics from discrete mathematics, number theory, probability, geometry and analysis and has wide applications in physics, chemistry, biology, economy and social sciences.

Probably the most general formulation of such a concept is the action of a monoid over an algebra. Given a monoid $\mathbb{G}$ and an algebra $\mathcal{A}$, the (left)-action of $\mathbb{G}$ on $\mathcal{A}$ is simply a map $f : \mathbb{G} \times \mathcal{A} \to \mathcal{A}$ such that[1]

1. $f(gh, a) = f(g, f(h, a))$ for each $g, h \in \mathbb{G}$ and $a \in \mathcal{A}$;
2. $f(e, a) = a$ for every $a \in \mathcal{A}$, where e is the identity element of $\mathbb{G}$;
3. $f(g, a + b) = f(g, a) + f(g, b)$ for each $g \in \mathbb{G}$ and $a, b \in \mathcal{A}$;
4. $f(g, ab) = f(g, a)f(g, b)$ for each $g \in \mathbb{G}$ and $a, b \in \mathcal{A}$;

In our discussion we will be mainly motivated by physics. In fact, we will consider the cases in which $\mathbb{G} \in \{\mathbb{N}, \mathbb{Z}, \mathbb{R}_+, \mathbb{R}\}$[2] is interpreted as *time* and

[1] In an alternative, one can consider the action on a vector space, if one wants to include, e.g, stochastic processes.

[2] Although even in physics other possibilities are very relevant, e.g. in the case of Statistical Mechanics it is natural to consider the action of the space translations, i.e. the groups $\{\mathbb{Z}^d, \mathbb{R}^d\}$ for some $d \in \mathbb{N}$, $d > 1$.

$\mathcal{A}$, interpreted as the *observables* of the system,[3] is a commutative algebra consisting of functions over some set X. In addition, we will restrict ourselves to situations where the action over the algebra is induced by an action over the set X (this is a map $f : \mathbb{G} \times X \to X$ that satisfies condition 1, 2 above).[4] Indeed, given an action f of $\mathbb{G}$ on X and an algebra $\mathcal{A}$ of functions on X such that, for all $a \in \mathcal{A}$ and $g \in \mathbb{G}$, $b(\cdot) := a(f(g, \cdot)) \in \mathcal{A}$, it is natural to define $\tilde{f}(g, a)(x) := a(f(g, x))$ for all $g \in \mathbb{G}$, $a \in \mathcal{A}$ and $x \in X$. It is then easy to verify that $\tilde{f}$ satisfies conditions 1–4 above.

We will call *discrete time Dynamical System* the ones in which $\mathbb{G} \in \{\mathbb{N}, \mathbb{Z}\}$ and *continuous time Dynamical Systems* the ones in which $\mathbb{G} \in \{\mathbb{R}_+, \mathbb{R}\}$. Note that, in the first case, $f(n, x) = f(n-1+1, x) = f(1, f(n-1, x))$, hence defining $T : X \to X$ as $T(x) = f(1, x)$, holds $f(n, x) = T^n(x)$.[5] Thus in such a case we can (and will) specify the Dynamical System by writing only (X, T). In the case of continuos Dynamical Systems we will write $\phi_t(x) := f(t, x)$ and call ϕ_t a flow (if the group is $\mathbb{R}$) or a semi-flow (if the group is $\mathbb{R}_+$) and will specify the Dynamical System by writing (X, ϕ_t). In fact, in this notes we will be interested only in Dynamical Systems with more structure i.e. *topological*, *measurable* or *smooth* Dynamical Systems. By topological Dynamical Systems we mean a triplet $(X, \mathcal{T}, T)$, where $\mathcal{T}$ is a topology and T is continuos (if $B \in \mathcal{T}$, then $T^{-1}B \in \mathcal{T}$). By smooth we consider the case in which X has a differentiable structure and T is r-times differentiable for some $r \in \mathbb{N}$. Finally, a measurable Dynamical Systems is a quadruple (X, Σ, T, μ) where Σ is a σ-algebra, T is measurable (if $B \in \Sigma$, then $T^{-1}B \in \Sigma$) and μ is an invariant measure (for all $B \in \Sigma$, $\mu(T^{-1}B) = \mu(B)$).[6]

So far for general definitions that, to be honest, are not very inspiring. Indeed, what characterizes the modern Dynamical Systems is not so much the setting but rather the type of questions that are asked, first and foremost:

- **Which behaviors are visible in nature?** (stability and bifurcation theory).
- **What happens for very long times?** (statistics and asymptotic theory)

The rest of this book will deal in various ways with such questions.

The original motivation for the above setting and for these questions comes from the study of the motion which, after Newton, typically appears as so-

[3] Again other possibilities are relevant, e.g. the case of Quantum Mechanics (in the so called Heisenberg picture) where the algebra of the observable is non commutative and consists of the bounded operators over some Hilbert space.

[4] Again relevant cases are not included, for example all Markov Process where the evolution is given by the action of some semigroup.

[5] Obviously $T^2(x) = T \circ T(x) = T(T(x))$, $T^3(x) = T \circ T \circ T(x) = T(T(T(x)))$ and so on.

[6] The definitions for continuos Dynamical Systems are the same with $\{\phi_t\}$ taking the place of T.

lution of an *ordinary differential equation* (ODE). It is then natural to start with a brief reminder of basic ODE theory.[7]

In section 1.1 I will recall the theorem of existence and uniqueness of the solutions of an ODE. In addition, I will state the Gronwall inequality, a very useful inequality for estimating the growth rate of the solution of an ODE. Finally, a theorem yielding the smooth dependence of the solutions of an ODE from an external parameter or from the initial conditions is provided.

In section 1.2 is given a very brief account of linear equations with constant coefficients (by discussing the exponential of a matrix) and of Floquet theory. That is the study of the solutions of a linear equation with coefficients varying periodically in time. The basic result being that the asymptotic properties of the solutions can be understood by looking at the solutions after one period.

Finally, section 1.3 discusses the possibility of qualitative understanding the behavior of the solutions of ODE that cannot be solved explicitly (essentially all the ODEs). The arguments are very naive and are intended only to convince the reader that a) something can be done; b) a more sophisticated theory needs to be developed in order to have a satisfactory picture.

1.1 Few basic facts about ODE: a reminder

Our starting point is the initial Cauchy problem for ODE. That is, given a separable Banach space $\mathcal{B}$,[8] $V \in \mathcal{C}^0_{\rm loc}(\mathcal{B}\times\mathbb{R},\mathcal{B})$,[9] and $x_0 \in \mathcal{B}$, find an open interval $0 \ni I \subset \mathbb{R}$ and $x \in \mathcal{C}^1(I,\mathcal{B})$ such that

$$\begin{aligned} \dot{x}(t) &= V(x(t),t) \\ x(0) &= x_0. \end{aligned} \tag{1.1.1}$$

Remark 1.1.1 *I will be mainly interested in the case $\mathcal{B}=\mathbb{R}^d$, for some $d\in\mathbb{N}$. Thus, the reader uncomfortable with Banach spaces can safely substitute*

[7] In fact, also the solutions of a partial differential equation (PDE) may give rise to a Dynamical System, yet the corresponding theory is typically harder to investigate.

[8] A Banach spaces is a complete normed vector spaces. This means that a Banach space is a vector space V, over $\mathbb{R}$ or $\mathbb{C}$, equipped with a norm $\|\cdot\|$ such that every Cauchy sequence in V has a limit in V. By *separable* we mean that there exists a countable dense set. Check [RS80, Kat66] for more details or [DS88] for a lot more details.

[9] Given two Banach spaces $\mathcal{B}_1,\mathcal{B}_2$, an open set $U\subset\mathcal{B}_1$, and $q\in\mathbb{N}$ by $\mathcal{C}^q(U,\mathcal{B}_2)$ we mean the continuous functions from U to $\mathcal{B}_2$ that are q time (Fréchet) differentiable and the q-th differentials are continuous (see Problem 1.18 for a very quick discussion of differentiation in Banach spaces). Such a vector space can be equipped with the norm $\|\cdot\|_{\mathcal{C}^q}$ given by the sup of all its derivatives till the order q included. If we then consider the subset for which such a norm is finte, then we have again a vector space which is, in fact, a Banach space. We will call such a Banach space $\mathcal{C}^q(U,\mathcal{B}_2,\|\cdot\|_{\mathcal{C}^q})$, yet, when no confusion can arise, we will abuse of notation and call it simply $\mathcal{C}^q(U,\mathcal{B}_2)$. By $\mathcal{C}^q_{\rm loc}(U,\mathcal{B}_2)$ we mean the vector space of the functions $f:U\to\mathcal{B}_2$ such that, for each $u\in U$ and $R>0$ such that $\overline{B(u,R)}=\{v\in\mathcal{B}_1\ :\|v-u\|\le R\}\subset U$, $f\in\mathcal{C}^q(B(u,R),\mathcal{B}_2,\|\cdot\|_{\mathcal{C}^q})$. Note that, in general, $\mathcal{C}^q_{\rm loc}$ is not a Banach space (in fact, it is a Fréchet space).

$\mathbb{R}^d$ to $\mathcal{B}$ in all the subsequent arguments. Yet, it is interesting that the theory can be developed for general Banach spaces at no extra cost. For simplicity, in the following we will always assume that all the Banach spaces are separable even if not explicitly mentioned. In essence, this is just a fancy way of saying that much of the following depends only on the Banach structure of $\mathbb{R}^d$, that is on the fact that $\mathbb{R}^d$ is a complete vector space with a norm (e.g. the euclidean norm) and, for example, nowhere is used the fact that $\mathbb{R}^d$ has a finite basis.

I will also briefly consider ODE on (finite dimensional) manifolds. Not much extra theory is needed in order to do this, since ODE on manifolds can always be reduced to the case $\mathbb{R}^d$ case, see section 1.1.5.

The first problem that comes to mind is

Question 1 *Does the Chauchy problem (1.1.1) always admit a solution? If there exists a solution is it unique?*

To address such an issue it is convenient to consider the equation[10]

$$x(t) = x_0 + \int_0^t V(x(s), s)ds \tag{1.1.2}$$

Problem 1.1 *Show that for each finite open interval $0 \in I \subset \mathbb{R}$, if $x \in \mathcal{C}^1(I, \mathcal{B})$ is a solution of (1.1.1), then it is a solution of (1.1.2). Show that if $x \in \mathcal{C}^0(I, \mathcal{B})$ is a solution of (1.1.2) then $x \in \mathcal{C}^1(I, \mathcal{B})$ and solves (1.1.1).*

1.1.1 Existence and uniqueness

The issue of existence and uniqueness of the solutions of (1.1.1) can be solved by applying the clasical Banach fixed point Theorem (see A.1.1), provided we make a stronger assumption on V.

Theorem 1.1.2 (Existence and Uniqueness theorem for ODE) *For each $V \in \mathcal{C}^1_{\text{loc}}(\mathcal{B} \times \mathbb{R}, \mathcal{B})$ and $x_0 \in \mathcal{B}$ there exists $\delta \in \mathbb{R}_+$ such that there exists a unique solution of (1.1.1) in $\mathcal{C}^1((-\delta, \delta), \mathcal{B})$.*[11]

Proof. Let $\delta \in (0, 1)$. The reader can verify that the vector space $\mathcal{C}^0([-\delta, \delta], \mathcal{B})$, equipped with the norm $\|u\|_\infty := \sup_{t\in[-\delta,\delta]} \|u(t)\|_{\mathcal{B}}$ is a Banach space.[12] By definition there exist $\delta_0, R_0 \geq 0$ such that, for all $\delta \leq \delta_0$ and

[10] The most convenient meaning of the integral of a function with values in a Banach space is the *Bochner sense*, which reduces to the usual Lebesgue integral in the case $\mathcal{B} = \mathbb{R}^d$, see [Yos95] for definition and properties. Yet, for our purposes the equivalent of the Riemannian integral suffices and it is defined in the obvious manner. See Problem 1.20 for details.

[11] We equip $\mathcal{B} \times \mathbb{R}$ with the norm $\|(x, t)\| \leq \sup\{\|x\|_{\mathcal{B}}, |t|\}$, where $\|\cdot\|_{\mathcal{B}}$ is the norm of $\mathcal{B}$.

[12] The uniform limit of continuous functions is a continuos function.

$R \leq R_0$, $V \in \mathcal{C}^1(D_R, \mathcal{B})$, where $D_R = \{y \in \mathcal{C}^0([-\delta, \delta], \mathcal{B}) \ : \ \|y - x_0\|_\infty \leq R\}$. We can then define the operator $K : D_R \to \mathcal{C}^0([-\delta, \delta], \mathcal{B})$ by[13]

$$K(u)(t) := x_0 + \int_0^t V(u(s), s)ds.$$

Let $M_\delta = \sup_{|t|\leq\delta} \sup_{u\in D_R} \{\|V(u,t)\| + \|\partial_u V(u,t)\|\}$, note that M_δ is a decreasing function of δ. Then, for each $u \in D_R$ and $|t| \leq \delta$, (recall Problem 1.22)

$$\|K(u(t)) - x_0\| \leq \delta M_\delta \leq R$$

provided we chose $\delta M_\delta \leq R$. Thus K maps D_R into D_R. In addition, for each $u, v \in D_R$,

$$\|K(u) - K(v)\|_\infty \leq \delta M_\delta \|u - v\|_\infty \leq \frac{1}{2}\|u - v\|_\infty,$$

provided we chose $2\delta M_\delta \leq 1$. We can then apply Theorem A.1.1 and obtain a unique solution of the equation $Ku = u$ in D_R. This shows the existence and uniqueness of the solution of (1.1.2). The Theorem follows then by remembering Problem 1.1. □

Remark 1.1.3 *Note that in the proof of Theorem A.1.1 one can chose the same δ for an open set of initial condition.*

Remark 1.1.4 *The hypotheses of the above Theorem can be easily weakened to the case of V locally Lipschitz in x and continuous in t, yet only continuity does not suffice for uniqueness as shown by the example*

$$\dot{x} = \sqrt{x}$$
$$x(0) = 0.$$

which has the infinitely many solutions $x_a(t) = 0$ for $t \leq a$ and $x_a(t) = \frac{1}{4}(t-a)^2$ for $t \geq a$, $a \in \mathbb{R}$.[14]

Remark 1.1.5 *The restriction to an interval of size δ in Theorem A.1.1 cannot be avoided as shown by the example*

$$\dot{x} = x^2$$
$$x(0) = 1.$$

Its solution $x(t) = (1-t)^{-1}$ is not continuous, nor bounded, for $t = 1$.

[13] The meaning of $\mathcal{C}^0(K, \mathcal{B}_2)$ where K is a closed set of $\mathcal{B}_1$ is the usual one.

[14] If $\mathcal{B}$ is finite dimensional, then $V \in \mathcal{C}^0$ suffices for the existence of a solution. This follows by a direct application of Schauder fixed point Theorem to (1.1.2). For informations on such a fixed point theorem and fixed point theorems in general see [Zei86].

We have seen a mechanism whereby the solution cannot be defined for all times, the next Lemma shows that, for $\mathcal{C}^1$ vector fields, the above is the *only* mechanism.[15]

Lemma 1.1.6 *In the hypotheses of Theorem 1.1.2, if $x \in \mathcal{C}^1_{\text{loc}}((-\underline{\delta}, \delta), \mathcal{B})$ is a solution of (1.1.1) for some $\underline{\delta}, \delta > 0$, and if there exists $M > 0$ such that $\sup_{t\in[0,\delta)} \|x(t)\| \leq M$, then there exists $\bar{\delta} > \delta$ and $\bar{x} \in \mathcal{C}^1((-\underline{\delta}, \bar{\delta}), \mathcal{B})$ that solves (1.1.1) (i.e. the solution can be extended for longer times).*

PROOF. Let $\{t_n\}$ be any sequence that converges to δ, then

$$\|x(t_n) - x(t_m)\| \leq \int_{t_n}^{t_m} \|V(x(s), s)\| ds \leq |t_n - t_m| \sup_{\|z\|\leq M} \sup_{s\in[0,\delta)} \|V(z, s)\|.$$

Thus $\{x(t_n)\}$ is a Cauchy sequence and admits a limit $x_* \in \mathcal{B}$ such that

$$x_* = \lim_{n\to\infty} x(t_n) = \lim_{t\to\delta} x(t) = x_0 + \int_0^\delta V(x(s), s) ds.$$

We can then consider the equation

$$y(t) = x_* + \int_0^t V(y(s), s + \delta) ds.$$

By Theorem 1.1.2 there exists δ_1 and $y \in \mathcal{C}^1((-\delta_1, \delta_1), \mathcal{B})$ which satisfy the above equation. Let then $\bar{\delta} = \delta + \delta_1$ and define

$$\bar{x}(t) := \begin{cases} x(t) & \text{fot all } t \in (-\underline{\delta}, \delta) \\ y(t - \delta) & \text{fot all } t \in [\delta, \bar{\delta}). \end{cases}$$

Clearly $\bar{x} \in \mathcal{C}^0((-\underline{\delta}, \bar{\delta}), \mathcal{B})$ and, for $t \in [\delta, \bar{\delta})$ holds true

$$\begin{aligned} \bar{x}(t) &= x_* + \int_\delta^t V(\bar{x}(s), s) ds = x_0 + \int_0^\delta V(\bar{x}(s), s) ds + \int_\delta^t V(\bar{x}(s), s) ds \\ &= x_0 + \int_0^t V(\bar{x}(s), s) ds. \end{aligned}$$

Thus, again remembering Problem 1.1, the Lemma follows. □

Remark 1.1.7 *Applying repeatedly Lemma 1.1.6 it follows that there exists a maximal open interval $J \subset \mathbb{R}$ such that the Cauchy problem (1.1.1) has a unique solution belonging to $\mathcal{C}^1_{\text{loc}}(J, \mathcal{B})$.*

[15] I state the result for positive times, for negative times it is the same.

1.1.2 Growald inequality

We have seen that the escape (growth) to infinity is the only obstruction to enlarging the domain of the solution.[16] The question remains: how large the maximal interval J in Remark 1.1.7 can be?

To understand better how the solution of an ODE can grow, we need a technical but extremely useful Lemma.

Lemma 1.1.8 (Integral Gronwall inequality) *Let $L, T \in \mathbb{R}_+$ and $\xi, f \in \mathcal{C}^0([0,T],\mathbb{R})$. If, for all $t \in [0,T]$,*

$$\xi(t) \leq L \int_0^t \xi(s)\, ds + f(t),$$

then

$$\xi(t) \leq f(t) + L \int_0^t e^{L(t-s)} f(s)\, ds.$$

PROOF. Let us first consider the case in which $f \equiv 0$. In this case the Lemma asserts $\xi(t) \leq 0$. Indeed, since ξ is a continuos function there exists $t_* \in [0, (2L)^{-1}] \cap [0,T] =: I_1$ such that $\xi(t_*) = \sup_{t\in I_1} \xi(t)$. But then,

$$\xi(t_*) \leq L \int_0^{t_*} \xi(s)\, ds \leq \xi(t_*) L t_* \leq \frac{1}{2}\xi(t_*)$$

which implies $\xi(t_*) \leq 0$ and hence $\xi(t) \leq 0$ for each $t \in I_1$. If $I_1 = [0,T]$, then we are done, otherwise letting $t_1 := (2L)^{-1}$ we have

$$\xi(t) \leq L \int_{t_1}^t \xi(s)\, ds$$

and we can make the same argument as before in the interval $[t_1, 2t_1]$. Iterating we have $\xi(t) \leq 0$ for all $t \in [0,T]$.

To treat the general case we reduce it to the previous one. Let

$$\zeta(t) := \xi(t) - f(t) - L \int_0^t e^{L(t-s)} f(s)\, ds.$$

Then

$$\begin{aligned}
\zeta(t) &\leq L \int_0^t \xi(s)\, ds - \int_0^t L e^{L(t-s)} f(s)\, ds \\
&= L \int_0^t \zeta(s)\, ds + L \int_0^t \left\{ f(s) ds + L \int_0^s e^{L(s-\tau)} f(\tau) d\tau \right\} \\
&\quad - \int_0^t L e^{L(t-s)} f(s)\, ds.
\end{aligned}$$

[16] Of course, this is the case only for regular vector fields. For other possibilities, think of the case of collisions among planets.

Next, notice that

$$\int_0^t ds L \int_0^s e^{L(s-\tau)} f(\tau)\, d\tau = L \int_0^t d\tau f(\tau) \int_\tau^t ds e^{L(s-\tau)}$$
$$= \int_0^t f(s)\{e^{L(t-s)} - 1\} ds.$$

Thus,

$$\zeta(t) \le L \int_0^t \zeta(s)\, ds.$$

We have then reduced the problem to the previous case which implies that it must be $\zeta(t) \le 0$ from which the Lemma follows. □

Let us see the usefulness of the above Lemma in a concrete example. Let $L(\mathcal{B}, \mathcal{B})$ be the Banach space of the linear bounded operators from $\mathcal{B}$ to $\mathcal{B}$.[17]

Lemma 1.1.9 *For each $A \in \mathcal{C}^1_{\text{loc}}(\mathbb{R}, L(\mathcal{B}, \mathcal{B}))$, consider the Cauchy problem*

$$\dot{x}(t) = A(t)x(t)$$
$$x(0) = x_0.$$

If $\|A(t)\| \le L$ for all $t \in \mathbb{R}$, then $\|x(t)\| \le e^{Lt}\|x_0\|$ for all $t \in \mathbb{R}$. In particular, the solution is defined on all $\mathbb{R}$.

Proof. If we write the equation in the equivalent integral form we have

$$\|x(t)\| \le \|x_0\| + \int_0^t \|A(s)x(s)\|\, ds \le \|x_0\| + L \int_0^t \|x(s)\|\, ds.$$

Let $\xi(t) := \|x(t)\|$, apply Lemma 1.1.8 for any $T \in \mathbb{R}_+$, the Lemma follows. □

Problem 1.2 *Explain why Lemma 1.1.9 does not apply to the following setting: $\mathcal{B} = \mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$ and*

$$\dot{x}(t, z) = \alpha(z, t)\partial_z x(t, z),$$

for some $\alpha \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$, $\alpha(z, T + t) = \alpha(z, t)$, $T > 0$. Compare with Problem 1.24.

[17] The norm of $L \in L(\mathcal{B}, \mathcal{B})$ is given by $\|L\| := \sup_{\substack{v \in \mathcal{B} \\ \|v\|=1}} \|Lv\|$. If $\mathcal{B} = \mathbb{R}^d$, then $L(\mathcal{B}, \mathcal{B})$ is just the vector space of the $d \times d$ matrices.

1.1.3 Flows

In this section we analyze the case in which the vector field is time independent and grows at most linearly.

Lemma 1.1.10 *Given $V \in \mathcal{C}^1_{\mathrm{loc}}(\mathcal{B}, \mathcal{B})$, if there exists $L, M \geq 0$ such that $\|V(x)\| \leq L\|x\| + M$, then the solution of (1.1.1) exists for all times and for all initial conditions.*

PROOF. We argue by contradiction. Choose any initial condition $x_0 \in \mathcal{B}$ and let $I(x_0) = (-\delta_-(x_0), \delta_+(x_0))$ be the maximal interval on which the solution is defined. If $\delta_+(x_0) < \infty$, then for each $t \leq \delta_+(x_0)$

$$\|x(t)\| \leq \|x_0\| + L \int_0^t \|x(s)\| ds + Mt.$$

Thus Gronawall inequality implies

$$\|x(t)\| \leq e^{Lt} \left\{ \|x_0\| + ML^{-1} \right\}$$

for $t \in [0, \delta_+(x_0))$. Then, by Lemma 1.1.6, the solution can be extended, contrary to the assumption that $(-\delta_-(x_0), \delta_+(x_0))$ was the maximal interval. A similar argument holds for negative t. □

For each $x_0 \in \mathcal{B}$ and $t \in \mathbb{R}$ let $x(t, x_0)$ be the solution of (1.1.1) at time t.

Lemma 1.1.11 *For each V as in Lemma 1.1.10, setting $\phi_t(x_0) := x(t, x_0)$, $\phi_{-t} = \phi_t^{-1}$ for $t \geq 0$, we have that $(\mathcal{B}, \phi_t)$, $t \in \mathbb{R}$, is a Dynamical System.*

PROOF. All we need to prove is that ϕ_t is an action of $\mathbb{R}$ on $\mathcal{B}$. First of all note that ϕ_t is indeed invertible. If not then there would be $x, x' \in \mathcal{B}$ such that $\phi_t(x) = \phi_t(x')$. But then the uniqueness of the solutions of the ODE implies $x = x'$. Moreover it is easy to check that $\phi_{-t}(x_0) = x(-t, x_0)$. Finally, $\phi_t(\phi_s(x)) = \phi_{t+s}(x)$. □

Remark 1.1.12 *We have thus proved that a large class of vector fields gives rise to* flows.

1.1.4 Dependence on a parameter

Having established the existence and uniqueness of the solution, the next natural questions present itself.

Question 2 *How do the solutions depend on the initial condition? How do the solutions depend on a change of the vector field?*

To discuss such issues it is convenient to analyze first the second question. More precisely, given $V \in \mathcal{C}^2_{\rm loc}(\mathcal{B}\times\mathbb{R}\times\mathbb{R}^d, \mathcal{B})$ we consider the Chauchy problem

$$\begin{aligned} \dot{x}(t) &= V(x(t), t, \lambda) \\ x(0) &= x_0. \end{aligned} \tag{1.1.3}$$

Clearly the solution will depend on the parameter λ. The question is then: calling $x(t,\lambda)$ the solution of (1.1.3), for a given $t \in \mathbb{R}$ what can we say about the function $x(t,\cdot)$?

For simplicity let us consider the case $V \in \mathcal{C}^2(\mathcal{B} \times \mathbb{R} \times \mathcal{B}_1, \mathcal{B})$, the more general case $V \in \mathcal{C}^2_{\rm loc}(\mathcal{B} \times \mathbb{R} \times \mathcal{B}_1, \mathcal{B})$ is similar and is left to the reader.

Theorem 1.1.13 (Smooth dependence on a parameter) *Given two Banach spaces $\mathcal{B}, \mathcal{B}_1$, let $V \in \mathcal{C}^2(\mathcal{B} \times \mathbb{R} \times \mathcal{B}_1, \mathcal{B})$. Let $X(t, x_0, \lambda)$ be the unique solution of (1.1.3), then $X(t, x_0, \cdot) \in \mathcal{C}^1_{\rm loc}(\mathcal{B}_1, \mathcal{B})$.*

PROOF. For each $x_0 \in \mathcal{B}$ consider the ODE for $\xi \in \mathcal{C}^1_{\rm loc}(\mathbb{R} \times \mathcal{B}_1, L(\mathcal{B}_1, \mathcal{B}))$

$$\begin{aligned} \dot{\xi}(t,\lambda) &= \partial_x V(X(t, x_0, \lambda), t, \lambda) \cdot \xi(t,\lambda) + \partial_\lambda V(X(t, x_0, \lambda), t, \lambda) \\ \xi(0,\lambda) &= 0. \end{aligned} \tag{1.1.4}$$

We claim that $\xi(t) = \partial_\lambda X(t, x_0, \lambda)$.[18] To verify the claim it suffices to prove that there exists $C > 0$ such that, for $h \in \mathcal{B}_1$ small enough, if $\zeta(t, h, \lambda) := X(t, x_0, \lambda+h) - X(t, x_0, \lambda) - \xi(t)h$, then $\|\zeta(t,h)\| \le C\|h\|^2$. By Taylor formula we have[19]

$$\begin{aligned} \dot{\zeta}(t,h) &= V(X(t, x_0, \lambda + h), t, \lambda + h) - V(X(t, x_0, \lambda), t, \lambda) \\ &\quad - \partial_x V(X(t, x_0, \lambda), t) \cdot \xi(t)h - \partial_\lambda V(X(t, x_0, \lambda), t, \lambda)h \\ &= \partial_x V(X(t, x_0, \lambda), t) \cdot \zeta(t,h) + R(t) \end{aligned} \tag{1.1.5}$$

where, in the last line, we have used

$$\begin{aligned} &V(X(t, x_0, \lambda + h), t, \lambda) - V(X(t, x_0, \lambda), t, \lambda) \\ &= \partial_x V(X(t, x_0, \lambda), t, \lambda) \cdot (X(t, x_0, \lambda + h), t, \lambda) - X(t, x_0, \lambda)) \\ &\quad + \mathcal{O}(\|X(t, x_0, \lambda + h), t, \lambda) - X(t, x_0, \lambda)\|^2), \end{aligned}$$

and

$$\begin{aligned} \|R(t)\| &\le C\left(\|X(t, x_0, \lambda + h) - X(t, x_0, \lambda)\|^2 + \|h\|^2\right) \\ &\le 2C(\|\zeta(t,h)\|^2 + (1 + \|\xi(t)\|^2)\|h\|^2). \end{aligned}$$

[18] If $\mathcal{B} = \mathbb{R}^d$ e $\mathcal{B}_1 = \mathbb{R}^m$ then ξ is just a $d \times m$ matrix.

[19] Note that we cannot Taylor expand $X(t, x_0, \lambda + h)$ with respect to h, since we do not know yet that X is differentiable with respect to λ.

with $C = \|V\|_{\mathcal{C}^2}$. Note that $\zeta(0) = 0$. We can then conclude by using Lemma 1.1.8. Indeed such a Lemma applied to (1.1.4) implies $\|\xi(t)\| \leq e^{C_1 t}$, for some $C_1 > 0$. Next, let $T > 0$ be the maximal time such that $\|\zeta(t,h)\| \leq 1/2$ and $e^{2C_1 T} \leq 2$. Then, for $t \leq T$, (1.1.5) yields

$$\|\zeta(t,h)\| \leq \int_0^t 2C\|\zeta(s)\|ds + 3\|h\|^2$$

and Lemma 1.1.8, again, implies the announced estimate. □

Problem 1.3 *Prove the analogous of Theorem 1.1.13 when* $V \in \mathcal{C}^1_{\mathrm{loc}}$.

The above theorem allow to easily prove the following fundamental result on the smooth dependence on parameters of an ODE.

Theorem 1.1.14 (Smooth dependence on initial conditions) *Let* $V \in \mathcal{C}^r(\mathcal{B}\times\mathbb{R},\mathcal{B})$, $r \geq 1$. *For* $x_0 \in \mathcal{B}$ *let* $X(t,x_0)$ *be the unique solution of* (1.1.1). *Then, for each* $t \in \mathbb{R}$, $X(t,\cdot) \in \mathcal{C}^r_{\mathrm{loc}}(\mathcal{B},\mathcal{B})$. *Moreover,* $\xi = \partial_{x_0} X$ *solves*

$$\begin{aligned} \dot{\xi}(t) &= \partial_x V(X(t,x_0),t)\cdot\xi(t) \\ \xi(0) &= \mathbb{1}. \end{aligned} \tag{1.1.6}$$

Proof. Set $z = x - x_0$ and consider the resulting equation

$$\begin{aligned} \dot{z} &= V(z+x_0,t) =: \bar{V}(z,t,x_0) \\ z(0) &= 0. \end{aligned}$$

One can then consider x_0 as an external parameter, applying Theorem 1.1.13 yields the result for $r = 1$. On the other hand, (1.1.6) is itself a differential equation depending on a parameter with a $\mathcal{C}^1$ vector field and a $\mathcal{C}^1$ dependence on the parameter x_0 , provided $r \geq 2$. So we can apply Theorem 1.1.13 again, and so on for r times, which proves the theorem. □

1.1.5 ODE on Manifolds–few words

Let us remind that a *topological manifold* is a second countable Hausdorff space which is locally homeomorphic to Euclidean space. A *chart* over a topological manifold M is a pair (U,ϕ) such that $U \subset M$ is an open set and $\phi : U \to \mathbb{R}^n$, for some $n \in \mathbb{N}$, is an homeomorphism between U and the open set $\Phi(U)$. An *atlas* on a topological manifold is a countable collection of charts $\{(U_\alpha,\phi_\alpha)\}$. We say that an atlas is $\mathcal{C}^k$ if $\phi_\alpha \circ \phi_\beta^{-1}$ is $\mathcal{C}^k$ when is defined. We say that two $\mathcal{C}^k$ atlas are equivalent if their union is a $\mathcal{C}^k$ atlas. A $\mathcal{C}^k$ manifold is a topological manifold equipped with an equivalence class of $\mathcal{C}^k$ atlas (often called a *differentiable structure*).

Although most often we will be concerned with manifolds embedded in some $\mathbb{R}^d$, also other possibilities will be relevant. Let us consider two examples.

Problem 1.4 *Show that $\mathbb{R}^d$ is a $\mathcal{C}^\infty$ manifold.*[20]

Problem 1.5 *Let $f \in \mathcal{C}^k(\mathbb{R}^d, \mathbb{R})$, and consider $M = \{(x, y) \in \mathbb{R}^d \times \mathbb{R} \,:\, y = f(x)\}$. Consider the atlas consisting of the chart (M, ϕ) where $\phi(x, y) = x$. This is a $\mathcal{C}^\infty$ manifold.*

Problem 1.6 *Check that $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$ is a $\mathcal{C}^\infty$ manifold.*

Given two differentiable manifolds ($\mathcal{C}^k$ manifolds with $k \geq 1$) M_1, M_2 and a map $f : M_1 \to M_2$ we say that $f \in \mathcal{C}^r(M_1, M_2)$, $r \leq k$, if for each atlas $\{(U_\alpha, \phi_\alpha)\}$ of M_1 and atlas $\{(V_\beta, \psi_\beta)\}$ of M_2, holds true $\psi_\beta \circ f \circ \phi_\alpha^{-1} \in \mathcal{C}^r$ on their domains of definition.

Given a differentiable manifold M and $x \in M$, we say that two curves $\gamma_1, \gamma_2 \in \mathcal{C}^1((-1, 1), M)$, such that $\gamma_1(0) = \gamma_1(0) = x$, are equivalent at x if for each chart (U, ϕ) such that $x \in U$ holds true $(\phi \circ \gamma_1)'(0) = (\phi \circ \gamma_2)'(0)$. A *tangent vector* at x is an equivalence class of curves.

Problem 1.7 *Show that if M is localy homeorphic to $\mathbb{R}^d$, then the set of tangent vectors at any $x \in M$ form canonically a d dimensional vector space.*[21]

We will use $\mathcal{T}_x M$ to designate the *tangent space* at x, that is the set of the tangent vectors at x. The tangent bundle is the disjoint union of the tangent spaces, i.e. $\mathcal{T}M = \cup_{x \in M} \{x\} \times \mathcal{T}_x M$. Finally, a *vector field* is a section of the tangent bundle, i.e. $\tilde{V} : M \to \mathcal{T}M$ such that $\tilde{V}(x) = (x, V(x))$, $V(x) \in \mathcal{T}_x M$. Form now on, with a slight abuse of notation, we will identify $\tilde{V}$ with V. Also, given $f \in \mathcal{C}^1(M_1, M_2)$, since the image of a $\mathcal{C}^1$ curve is a $\mathcal{C}^1$ curve, ve have naturally defined a map $f_* : \mathcal{T}M_1 \to \mathcal{T}M_2$.

Problem 1.8 *If $f \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}^n)$ discuss the relation between f_* and the derivative Df.*

We have finally the language to define O.D.E. on manifolds, in fact the Cauchy problem is exactly given again by (1.1.1), only now V is a, possibly time dependent, $\mathcal{C}^1$ vector field.

Problem 1.9 *Suppose that x_0 belongs to some chart (U, ϕ), show that the solution of*

$$\begin{aligned} \dot{x} &= V(x, t) \\ x(0) &= x_0 \end{aligned}$$

for a sufficiently small time can be obtained by the solution of an appropriate O.D.E. in $\phi(U)$.

[20] Note that, contrary to $\mathcal{C}^k$, $\mathcal{C}^\infty$ is not a Banach space (there is no good norm). It is possible to give to it the structure of a Fréchet space [RS80], but we will refrain from such subtleties. We just consider $\mathcal{C}^\infty = \cap_{n \in \mathbb{N}} \mathcal{C}^n$ as a vector space.

[21] If (U, ϕ) is a chart containing x, and γ_1, γ_2 two curves, think of the curves $\gamma_\lambda(t) = \gamma_1(\lambda t)$ and $\phi^{-1}(\phi(\gamma_1(t)) + \phi(\gamma_2(t)) - \phi(x))$.

Problem 1.10 *Given a finite atlas $\{(U_\alpha, \phi_\alpha)\}$, show that there exists a smooth partition of unity subordinated to the atlas, that is a collections $\{\varphi_\alpha\} \in \mathcal{C}^\infty(M, \mathbb{R})$ such that $\sum_\alpha \varphi_\alpha = 1$ and $\operatorname{supp} \varphi_\alpha \subset U_\alpha$.*

Problem 1.11 *Given a smooth vector field V consider*

$$\begin{aligned} \dot{x} &= V(x) \\ x(0) &= x_0 \end{aligned} \tag{1.1.7}$$

with $x_0 \in U_\alpha$ for some element of an atlas $\{(U_\alpha \phi_\alpha)\}$. Let $z_\alpha(t)$ be the solution of

$$\begin{aligned} \dot{z}_\alpha &= (\phi_\alpha)_* V(z_\alpha) \\ z_\alpha(0) &= \phi_\alpha(x_0) \end{aligned}$$

and suppose that $\phi_\alpha^{-1}(z(1)) \in U_\beta$. Consider then the solution of

$$\begin{aligned} \dot{z}_\beta &= (\phi_\beta)_* V(z_\beta) \\ z_\beta(1) &= \phi_\beta(\phi_\alpha^{-1}(z_\alpha(1))). \end{aligned}$$

Show that there exists $t_1 > 1$ such that

$$\begin{aligned} x(t) &= \phi_\alpha^{-1}(z_\alpha(t)) \quad \text{for } t \in [0, 1] \\ x(t) &= \phi_\beta^{-1}(z_\beta(t)) \quad \text{for } t \in (1, t_1) \end{aligned}$$

is a solution of (1.1.7) in the time interval $[0, t_1)$.

Remark 1.1.15 *We have seen that the theory of ODE on manifolds can be reduced locally to the case of $\mathbb{R}^d$. Yet, the reader should be aware that the global properties of the solutions can be very different. We will comment at length on this issue later on.*

1.2 Linear ODE and Floquet theory

Let us briefly discuss the simplest possible differential equation: the affine ones. For simplicity, we restrict ourselves to the case $\mathcal{B} = \mathbb{R}^d$ for some $d \in \mathbb{N}$.

1.2.1 Linear equations

Consider

$$\begin{aligned} \dot{x} &= Ax \\ x(0) &= x_0. \end{aligned} \tag{1.2.8}$$

Problem 1.12 *Show, by induction, that for each $n \in \mathbb{N}$ the solution of (1.2.8) satisfies*

$$x(t) = \sum_{k=0}^{n} \frac{1}{k!} A^k t^k x_0 + \int_0^t dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{n-1}} dt_n A^{n+1} x(t_n).$$

Taking the limit for $n \to \infty$ in the above expression one readily obtains $x(t) = \sum_{n=0}^{\infty} \frac{1}{n!} A^n t^n x_0$. That this is a solution can be verified directly inserting this formula in (1.2.8) (and noticing that the series and the series obtained by deviating term by term are uniformly convergent). By the standard analytic functional calculus for matrices (and operators, see Appendix C) we can thus write

$$x(t) = e^{At} x_0. \tag{1.2.9}$$

The above discussion provides a general solution for all equations of the type (1.2.8).

In reality life it is not that simple: if one has a concrete matrix A and wants to compute e^{At}, this may be quite unpleasant. A general strategy, although not necessarily the simplest one, is to perform a linear change of variables $x = Uz$. Then $\dot{z} = U^{-1}AUz$, and U is chosen so that $\Lambda = U^{-1}AU$ is in Jordan normal form. Then

$$x(t) = Uz(t) = Ue^{\Lambda t} z_0 = Ue^{\Lambda t} U^{-1} x_0.$$

It suffices then to know how to take exponentials of Jordan blocks, and this can be computed by using the defining series.

Problem 1.13 *Compute $e^{\Lambda t}$ for*

$$\Lambda = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} \; ; \quad \Lambda = \begin{pmatrix} a & 1 \\ 0 & a \end{pmatrix} \; ; \quad \Lambda = \begin{pmatrix} a & 1 & 0 \\ 0 & a & 1 \\ 0 & 0 & a \end{pmatrix}.$$

Another, equivalent, point of view is to look for solutions of the type $x(t) = e^{at}v$, substituting in the first of (1.2.8) one obtains $av = Av$. Thus, as we know already, each eigenvalue of A provides a solution of (1.2.8) (ignoring the initial condition). If there exists real eigenvectors $\{v_i\}_{i=1}^d$ which span all $\mathbb{R}^d$ then one can write the general solution, depending on d parameters α_i, as $x(t) = \sum_{i=1}^d \alpha_i v_i e^{a_i t}$, where a_i is he eigenvalue associated to the eigenvector v_i. One can then satisfy the initial condition by solving $x_0 = \sum_{i=1}^d \alpha_i v_i$. The same can be done is the eigenvectors are complex, by working in $\mathbb{C}^d$ instead then $\mathbb{R}^d$. If Jordan blocks are present one can look for solutions of the form $x(t) = \sum_{k=0}^p \frac{1}{(p-k)!} t^k e^{at} v_k$, compare this formula with your solution of Problem 1.13.

Remark 1.2.1 *Note that if the matrix A does not have eigenvalues with zero real part, then (by spectral decomposition) one can write $\mathbb{R}^d = V_- \oplus V_+$, where $AV_\pm = V_\pm$ and A restricted to V_- has eigenvalues with negative real part while on V_+ has eigenvalues with positive real part. Hence if $x_0 \in V_-$ it will hold $\lim_{n\to\infty} x(t) = 0$, and if $x_0 \in V_+$ it will hold $\lim_{n\to\infty} \|x(t)\| = \infty$. If $x_0 \notin V_-$ we can write it as $x_0 = x_- + x_+$, where $x_\pm \in V_\pm$. Hence $\lim_{n\to\infty} \|x(t)\| = \infty$ and the trajectory will escape to infinity while getting exponentially close to the subspace V_+. This is our first* long time *result.*

A slightly more complex situation is given by

$$\begin{aligned} \dot{x} &= Ax + b(t) \\ x(0) &= x_0, \end{aligned} \tag{1.2.10}$$

where $b \in \mathcal{C}^0(\mathbb{R}, \mathbb{R}^d)$. The solution of (1.2.10) is given by[22]

$$x(t) = e^{At}x_0 + \int_0^t e^{A(t-s)}b(s)ds. \tag{1.2.11}$$

1.2.2 Floquet theory

Let us consider the simplest case of a linear time dependent equation: there exists a continuous function $A \in \mathcal{C}^0_{\text{loc}}(\mathbb{R}, L(\mathbb{R}^d, \mathbb{R}^d))$ and $T \in \mathbb{R}_+$ such that, for all $t \in \mathbb{R}$, $A(t+T) = A(t)$. More precisely, let $\Phi(x_0, s, t)$ be the solution of the Cauchy problem[23]

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) \\ x(s) &= x_0. \end{aligned} \tag{1.2.12}$$

Problem 1.14 *Verify the following facts for each $x_0, y_0 \in \mathcal{B}$ and for each $a, b, t, s, \tau \in \mathbb{R}$*

- $\Phi(ax_0 + by_0, s, t) = a\Phi(x_0, s, t) + b\Phi(y_0, s, t)$,
- $\Phi(x_0, s, t) = \Phi(\Phi(x_0, s, \tau), \tau, t)$,
- $\Phi(x_0, s + T, t + T) = \Phi(x_0, s, t)$.

By the first property of Problem 1.14 there exists $K \in \mathcal{C}^1_{\text{loc}}(\mathbb{R}^2, L(\mathbb{R}^d, \mathbb{R}^d))$ such that $\Phi(x_0, s, t) = K(s, t)x_0$, the second property implies that $K(\tau, t)K(s, \tau) = K(s, t)$, the third that $K(s + T, t + T) = K(s, t)$. The next step is the first occurrence in this book of a very simply but very powerful idea to analyze

[22] Look for a solution of the form $x(t) = e^{At}z(t)$ and find the differential equation for z.
[23] The solution is well defined for all times by Lemma 1.1.10.

dynamical systems: a Poincaré section. Essentially the idea consist in looking at the system only at specially selected moments in time. In this case it is convenient to look at $t \in \{nT\}_{n\in\mathbb{Z}}$. That is, we want to investigate $\Phi(x_0, 0, nT) =: F(x_0, n)$.

Lemma 1.2.2 *The couple $(\mathbb{R}^d, F)$ is a discrete Dynamical System.*

PROOF. We have to show that F is an action of $\mathbb{Z}$ on $\mathbb{R}^d$. Let $f(x_0) := F(x_0, 1)$.

$$\begin{aligned} F(x_0, n) &= \Phi(x_0, 0, nT) = \Phi(\Phi(x_0, 0, (n-1)T), (n-1)T, nT)) \\ &= \Phi(\Phi(x_0, 0, (n-1)T), 0, T)) = f(\Phi(x_0, 0, (n-1)T)) = f^n(x_0). \end{aligned}$$

In addition, note that the uniqueness of the solutions of the ODE implies that if $f(x_0) = 0$, then $x_0 = 0$. Now, by construction, $f(x_0) = K(0, T)x_0$, thus $K(0, T)$ is an invertible matrix. Hence $F(x_0, -n) = f^{-n}(x_0)$ for all $n \in \mathbb{N}$. □

By using the functional calculus (see Problem C.19) one can define $B := T^{-1} \ln K(0, T)$, so $e^{BT} = K(0, T)$. Let us now consider $P(t) := K(0, t)e^{-Bt}$.[24]

$$\begin{aligned} P(t+T) &= K(0, t+T)e^{-B(t+T)} = K(T, t+T)K(0, T)K(0, T)^{-1}e^{-Bt} \\ &= K(0, t)e^{-Bt} = P(t). \end{aligned}$$

We have just proven the following result.

Theorem 1.2.3 (Floquet theorem) *The solutions of the equation (1.2.12) can be written as $x(t) = P(t)e^{Bt}K(s, 0)x_0$ where $P(t+T) = P(t)$ is periodic.*

Note that the matrix B can be complex valued. This can be avoided at a little extra cost.

Problem 1.15 *Prove that the solutions of the equation (1.2.12) can be written as $x(t) = P(t)e^{Bt}x_0$ where B is real and $P(t+2T) = P(t)$ is periodic of period $2T$.*

Note that Theorem 1.2.3 implies that the long time behavior is completely contained in the eigenvalues of the matrix B often called *floquet exponents*.

Problem 1.16 *Find the solutions of*

$$\dot{x} = a(t)Ax$$

where $a \in \mathcal{C}^0(\mathbb{R}, \mathbb{R})$ is periodic of period T and A is a fixed matrix.

Problem 1.17 *Given a fixed matrix A and a function at matrix values $B(t)$ of period T, consider the equation $\dot{x} = (A + \varepsilon B(t))x$, $\varepsilon \in \mathbb{R}$. Show that, for ε small enough, calling ν_i the Floquet exponents and setting $\lambda_i = e^{\nu_i}$ (often called* Floquet multiplier*), the λ_i are ε-close to the eigenvalues of A.*

[24] Note that the kernel of $K(0, T)$ must be $\{0\}$, otherwise it would violate the uniqueness of the solutions of the differential equation. Hence, $0 \notin \sigma(K(0, T))$.

1.3 Qualitative study of ODE

The previous discussion has shed some light on the behavior of linear ODE, unfortunately the interesting ODE are typically non linear. Although some nonlinear ODE can be solved explicitly (see any ODE book for examples) typically this is not possible, hence the need of a qualitative theory. As for the qualitative study of functions this can be done quite naively in one dimension, while higher dimensions requires some non trivial theory. Let us see such a naive qualitative theory for ODE via few examples.

1.3.1 The one dimensional case

This situation is very similar to the study of the graph of a function of one variable. Indeed to draw the graph one studies the first derivative and here the first derivative is specified by the equation. Let us consider a couple of simple examples. Consider

$$\begin{aligned}\dot{x} &= e^{-x^2} + x - 2 = V(x)\\ x_0 &= 0.\end{aligned}$$

One cannot integrate the function $V(x)^{-1}$ (which would yield an explicit solution of the ODE), yet from the equation follows that there exists a close to 2 such that $\dot{x}$ is negative if $x \leq a$ and positive otherwise. This implies that the solution starts to be decreasing and keeps decreasing forever.

Next, consider

$$\begin{aligned}\dot{x} &= 1 - 2tx\\ x_0 &= a.\end{aligned}$$

Such an equation cannot be solved by separation of variables, yet the above arguments still apply. In particular, for $t \geq 0$, we have $\dot{x}(t) < 0$ iff $x(t) > \frac{1}{2t}$. On the other hand if $x(t) > \frac{1}{2t}$ it will be so forever. In fact, consider $g(t) = x(t) - \frac{1}{2t}$, then $g'(t) = \dot{x}(t) + \frac{1}{2t^2}$. So if $g(t_*) = 0$, then $g'(t_*) > 0$ hence for $t < t_*$ one has $g(t) < 0$. Thus the solution will increase until it will intersect the curve $\frac{1}{2t}$ and then it will start decreasing but always staying above such a curve. Accordingly, for $t \geq t_*$ we can write $x(t) = \frac{1+\alpha(t)}{2t}$ with $\alpha \geq 0$. Then $\dot{x}(t) = -\alpha(t)$, that is

$$\frac{1}{2t} \leq x(t) = \frac{1}{2t_*} - \int_{t_*}^{t} \alpha(s)ds \tag{1.3.13}$$

moreover $-\frac{1+\alpha(t)}{2t^2} + \frac{\dot{\alpha}(t)}{2t} = -\alpha(t)$

$$\dot{\alpha}(t) = -(2t - \frac{1}{t})\alpha(t) + \frac{1}{t}$$

which means that either $\alpha(t) \leq \frac{1}{2t^2-1}$ or it is decreasing. But if it is decreasing it must decrease to zero otherwise (1.3.13) would be false for large t. Accordingly it must be $\lim_{t\to\infty} \alpha(t) = 0$.

1.3.2 Autonomous equations in two dimensions

In this case the basic idea is to consider one component as a function of the other and in this way reduce to the previous case. Let us see some examples.

Van Der Pol equation

Consider the equation

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= (1-3x^2)y - x. \end{aligned} \tag{1.3.14}$$

Clearly $(0,0)$ is the unique zero of the vector field. If we linearise (1.3.14) around zero we have

$$\frac{d}{dt}(x,y) = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

The matrix has eigenvalues $\lambda_\pm = \frac{1\pm\sqrt{3}i}{2}$ hence the fixed point is repelling and the solutions spiral away from it.

The next question is if a similar motion takes place also far away from the origin. To this end we want to forget the time dependence and concentrate only on the shape of the trajectories. Thus we can represent trajectories on the xy plane. Indeed, apart from the point $(0,0)$, either $\dot{x}$ or $\dot{y}$ are different from zero. In the first case one can locally invert $x(t)$ and write $y(x) = y(t(x))$. When this is possible one obtains

$$\frac{dy}{dx} = 1 - 3x^2 - \frac{x}{y},$$

which can be studied as in the previous examples. With a bit of work one can see that the trajectory spirals around zero, but exactly how?

To better understand the behaviour of the solution we introduce a "Lyapunov" like function.

$$L(x,y) = 2(x - x^3 - y)^2 + (x-y)^2 + 3x^2.$$

If $(x(t), y(t))$is a solution of (1.3.14), then a direct computation yields

$$\frac{d}{dt}L(x(t),y(t)) = x^2\left[6 - x^2 - 3(x-y)^2 - 3y^2\right].$$

Accordingly, L is decreasing outside an ellipse. Since $2ab \le a^2 + b^2$,[25]

$$\begin{aligned} L(x,y) &= 3(x-y)^2 - 4(x-y)x^3 + 2x^6 + 3x^2 \ge (x-y)^2 + 3x^2 \\ &= 4x^2 - 2xy + y^2 \ge 2x^2 + \frac{1}{2}y^2. \end{aligned}$$

Hence, the level sets $K_\alpha = \{(x,y) \in \mathbb{R}^2 \ : \ L(x,y) \le \alpha\}$ are contained in the ellipses $\{(x,y) \in \mathbb{R}^2 \ : \ 2x^2 + \frac{1}{2}y^2 \le \alpha\}$ and hence are compact.

Thus, far away from the origin the trajectory spirals inwardly. It follows, by the continuity with respect to the initial data, that there exists an $a_* \ge 0$ such that the corresponding solution is a periodic orbit.

Lotka-Volterra equation

$$\begin{aligned} \dot{x} &= ax - Ax^2 - \lambda xy \\ \dot{y} &= -dy + \lambda xy. \end{aligned}$$

This equation is meant to describe the evolution of two populations one feeding on the other (predator-prey). It also has periodic solutions, try to prove it using qualitative methods.

Second order in one dimension

Consider the equation

$$\begin{aligned} \ddot{x} &= -\gamma\dot{x} + \frac{x^2}{1+x^4} \\ x(0) &= 0; \quad \dot{x}(0) = v. \end{aligned}$$

Setting $(z,w) = (x,\dot{x})$, we can write it as

$$\begin{aligned} \dot{z} &= w \\ \dot{w} &= -\gamma w + \frac{z^2}{1+z^4} \end{aligned}$$

which is the type discussed above.

Clearly if we consider still higher dimensional cases the above naive approach cannot help us very much, hence the need of a more sophisticated theory.

[25] It follows from $(a-b)^2 \ge 0$.

Problems

1.18. Given two Banach spaces $\mathcal{B}_1, \mathcal{B}_2$ and a function$f : \mathcal{B}_1 \to \mathcal{B}_2$ we can define the partial derivative at $x \in \mathcal{B}_1$ in the direction $v \in \mathcal{B}_1$ (Gâteaux derivative) by

$$\partial_v f(x) = \lim_{h \to 0} h^{-1} \left[f(x + hv) - f(x) \right],$$

if the limit exists. On the other hand we say that f is Fréchet differentiable at x if there exists $A \in L(\mathcal{B}_1, \mathcal{B}_2)$ (the space of the continous linear operators from $\mathcal{B}_1$ to $\mathcal{B}_2$) such that

$$\lim_{h \to 0} \frac{\|f(x + h) - f(x) - Ah\|}{\|h\|} = 0,$$

and A is called the Fréchet differential at of f at x (often written $Df(x)$). Show that if f is Fréchet differentiable at zero, then it is continuous and Gâteaux differentiable.

1.19. Let $f \in \mathcal{C}^0(\mathcal{B}_0, \mathcal{B}_1)$ and $g \in \mathcal{C}^0(\mathcal{B}_1, \mathcal{B}_2)$ such that f is Fréchet differentiable at $x \in \mathcal{B}_0$ and g is Fréchet differentaible at $f(x) \in \mathcal{B}_1$. Show that $g \circ f \in \mathcal{C}^0(\mathcal{B}_0, \mathcal{B}_2)$ is Fréchet differentaible at x and that $D(g \circ f)(x) = Dg(f(x)) \cdot Df(x) \in L(\mathcal{B}_0, \mathcal{B}_2)$. Of course, this is nothing else than a glorified version of the *chain rule*.

1.20. Given a compact interval $I \subset \mathbb{R}$, a Banach space $\mathcal{B}$, and a continuous function $f \in \mathcal{C}^0(I, \mathcal{B})$, shows that one can define the equivalent of the Riemannian integral.

1.21. Prove the fundamental theorem of calculus in this setting. That is, for $f \in \mathcal{C}^1(\mathcal{B}_1, \mathcal{B}_2)$ let $Df(x) \in L(\mathcal{B}_1, \mathcal{B}_2)$ be the Fréchet differential at $x \in \mathcal{B}_1$, then for each $x, y \in \mathcal{B}_1$

$$f(y) = f(x) + \int_0^1 Df(x + t(y - x)) \cdot (x - y) dt.$$

1.22. Show that, for all $f \in \mathcal{C}^0([a, b], \mathcal{B})$,

$$\left\| \int_a^b f(t) dt \right\| \leq \int_a^b \|f(t)\| dt.$$

1.23. Study the solutions of the following equations for all possible initial conditions and $p \in \mathbb{N}$

$$\dot{x} = |x|^p$$
$$\dot{x} = x(\ln |x|)^p$$

1.24. Let $K \in \mathcal{C}^1(\mathbb{R} \times [0,1])$. Show that the equation

$$\partial_t u(t,s) = \int_0^1 K(t+s,\tau) u(t,\tau)^2 d\tau$$
$$u(0,s) = s^2.$$

has a unique continuos solution for t small enough.

1.25. Under the same hypotheses of Problem 1.17 show that if $\int_0^T B(s)ds = 0$ and the eigenvalues of A have all multiplicity one, then the Floquet multiplier differ from the eigenvalues of e^{AT} only of order ε^2.

1.26. Study the equation

$$(1+x)y\dot{y} + (x+y^2) = 0.$$

1.27. Study the equation (Bernoulli)

$$\dot{y} + p(x)y = q(x)y^n.$$

1.28. Study the equation

$$\ddot{x} = -\gamma\dot{x} - x^3.$$

Hints to solving the Problems

In this section, and in the parallel sections in later chapters, I provide hints for solving some of the Problems.

It is a very good idea to try very hard to solve the problems *before* looking at the hints: it is impossible to appreciate the solution if one has no feeling for the difficulties in the problem. The only way I know to get such a feeling is to *seriously* try to solve it.

Also, keep in mind that I suggest one way to proceed, often other ways are possible and maybe better.

1.1 The proof is the same as the standard proof for the case $\mathcal{B} = \mathbb{R}^d$. However, to see this, you have to do Problems 1.18 and 1.20 to understand exactly what the derivative and integral mean in this more general case.

1.12 For $n = 0$ it is just (1.1.2). To verify it for any n it suffices to show that

$$\int_0^t dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{n-1}} dt_n 1 = \frac{t^n}{(n+1)!}.$$

This follows since the domain of integration is $D = \{x \in [0,t]^{n+1} : t_{n+1} \le t_n \le \cdots \le t\}$. On the other hand, for each permutation σ of the

set $\{1,\dots,n+1\}$ the sets $D_\sigma = \{x \in [0,t]^{n+1} \,:\, t_{\sigma_{n+1}} \le t_{\sigma_n} \le \cdots \le t\}$ have the same measure, all the D_σ are disjoint and the union of all of them gives $[0,t]^{n+1}$.

1.15 First notice that if a matrix has no eigenvalues on the negative axis (including the zero) then the contour γ in C.3.3 can be taken symmetric around the real axis and, by using C.3.3 with the standard definition of ln with a cut on the negative real axis, this defines $\ln K(0,T)$ with real entries (since the formula for his complex conjugate is the same). In general, the spectrum of $K(0,T)$ can be split in $\sigma(K(0,T) = \alpha \cup \beta$ where $\alpha \cap [\mathbb{R}_- \cup \{0\}] = \emptyset$ and $\beta \subset \mathbb{R}_-$. Let P_α and $P_\beta = \mathbb{1} - P_\alpha$ be the associated spectral projectors, then we have the decomposition $K(0,T) = C + D$ where $C = P_\alpha K(0,T) P_\alpha$, $D = P_\beta K(0,T) P_\beta$. Consequently, $CD = DC = 0$ and $\sigma(C) = \alpha \cup \{0\}$ and $\sigma(D) = \beta \cup \{0\}$. Since we want to define a logarithm, we do not want zero in the spectrum, so we define $\tilde C = C + P_\beta$ and $\tilde D = D + P_\alpha$. The reader can check that $\sigma(\tilde C) = \alpha \cup \{1\}$ and $\sigma(\tilde D) = \beta \cup \{1\}$. Note that $\tilde D^2 = D^2 + P_\alpha$, hence $\sigma(\tilde D^2) \subset \mathbb{R}_+$. Hence $B = \frac{1}{T} \ln \tilde C + \frac{1}{2T} \ln \tilde D^2$ is real and, cince $[\tilde C, \tilde D] = 0$,

$$\begin{aligned} e^{2BT} &= \left[e^{\ln \tilde C}\right]^2 e^{\ln \tilde D^2} = \tilde C^2 \tilde D^2 \\ &= \left[(C + P_\beta)(D + P_\alpha)\right]^2 = K(0,T)^2 = K(0,2T). \end{aligned}$$

The rest of the argument remains the same.

1.17 Show that the solution satisfies

$$x(t) = e^{At} x_0 + \varepsilon \int_0^t e^{A(t-s)} B(s) x(s) ds.$$

and apply the perturbation theory in Appendix C.

1.20 Let $I = [a,b]$. Since the function is continuos, it is uniformly continuous, hence for $\varepsilon > 0$ there exists $\delta > 0$ such that, for each partition $\xi = \{[x_0,x_1],\dots,[x_{n-1},x_n]\}$, $x_0 = a, x_n = b$, $x_{n+1} - x_n \le \delta$, holds $\sup_{z,y \in [x_{n+1},x_n]} \|f(z) - f(y)\| \le \varepsilon$. Accordingly, for each choice of $z_n, y_n \in [x_{n+1}, x_n]$ we have

$$\left\| \sum_{k=0}^{n-1} f(z_k)(x_{k+1} - x_k) - \sum_{k=0}^{n-1} f(y_k)(x_{k+1} - x_k) \right\| \le \varepsilon.$$

By similar arguments, one can compare the sum defined on one partition with the sum defined on a finer partition. Finally, the sum over different partitions can be compared with the sum over the coarser partition,

which is finer than both. This shows that all sufficiently fine partitions yield the same approximate value, hence one can consider the partitions $\xi_n = \{[a + i\frac{b-a}{n}, a + (i+1)\frac{b-a}{n}]\}_{i=0}^{n-1}$ and define

$$\int_I f(t)dt := \lim_{n\to\infty} \sum_{i=0}^{n-1} f(a + i\frac{b-a}{n})\frac{b-a}{n}.$$

By the above discussion, this is equivalent to the same limit taken along any other partition, the diameter of whose elements tends uniformly to zero.

1.24 Consider the Banach space $\mathcal{B} = \mathcal{C}^0([0,1], \mathbb{R})$. Then $u(t,\cdot) \in \mathcal{B}$ and one can apply Theorem 1.1.2.

1.25 By Problem 1.17 we know that the solution at time T is given by the matrix $D(\varepsilon) := e^{AT}\left[\mathbb{1} + \varepsilon\int_0^T e^{-As}B(s)e^{As}ds\right]$. By the results in Appendix C it follows that, for ε small enough, the eigenvalues of $D(\varepsilon)$ are still simple and analytic on ε. Thus, let $\lambda(\varepsilon)$ one of such eigenvalues and $\Pi(\varepsilon)$ the associated eigenprojector. We have $D(\varepsilon)\Pi(\varepsilon) = \lambda(\varepsilon)\Pi(\varepsilon)$. Differentiating yields $\dot{D}(\varepsilon)\Pi(\varepsilon)+D(\varepsilon)\dot{\Pi}(\varepsilon) = \dot{\lambda}(\varepsilon)\Pi(\varepsilon)+\lambda(\varepsilon)\dot{\Pi}(\varepsilon)$. Multiplying on the right by $\Pi(\varepsilon)$, since $\Pi(\varepsilon)D(\varepsilon) = D(\varepsilon)\Pi(\varepsilon)$, we have

$$\Pi(\varepsilon)\dot{D}(\varepsilon)\Pi(\varepsilon) = \dot{\lambda}(\varepsilon)\Pi(\varepsilon).$$

Since $\Pi(\varepsilon)v = \langle a(\varepsilon), v\rangle b(\varepsilon)$ for some vectors a, b analytic in ε, $\dot{\lambda}(\varepsilon) = \langle a(\varepsilon), \dot{D}(\varepsilon)b(\varepsilon)\rangle$. We can now apply such a general formula to our specific case:

$$\begin{aligned}\langle a(0), \dot{D}(0)b(0)\rangle &= \langle a(0), e^{AT}\int_0^T e^{-As}B(s)e^{As}b(0)ds\rangle \\ &= \langle a(0), e^{AT}\int_0^T e^{-As}B(s)e^{As}b(0)ds\rangle \\ &= \lambda(0)\int_0^T \langle a(0), B(s)b(0)\rangle ds = 0.\end{aligned}$$

Notes

This chapter is super condensed and has no pretension to exhaust the theory of ODE. If one wants to have a better understanding of the field and some ideas of how an ODE can be solved in special cases better consult [HS74, Arn92, CL55].

Chapter 2

Local behavior

By local behavior we mean the study of the motion in a neighborhood of a point. As we have seen in the linear case, the motion can leave the neighborhood in a fixed time but it is also possible that it stays in the neighborhood for an unlimited time. In the latter case we will have the first example of how to tackle one of our stated goals: the study of the motion for long times. We start with a trivial case.

2.1 Flow box theorem

Let us consider the differential equation

$$\dot{x} = V(x) \tag{2.1.1}$$

where $V \in \mathcal{C}^2_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^d)$. By the results of the previous chapter there exist $\delta_-, \delta_+ : \mathbb{R}^d \to \mathbb{R}_+$ and $\phi : \{(z,t) \in \mathbb{R}^d \times \mathbb{R} \ : \ t \in (-\delta_-(z), \delta_+(z))\} =: D \to \mathbb{R}^d$ such that $\phi(z,t)$ is the solution of (2.1.1) with initial condition z. We would like to study the solution in a neighborhood of $x_0 \in \mathbb{R}^d$ such that $V(x_0) \neq 0$.

Theorem 2.1.1 (Flow box Theorem) *In the hypotheses above there exists a neighborhood U of x_0 and a change of variables $\Theta \in \mathcal{C}^1(U, \mathbb{R}^d)$ such that $\Theta(\phi(x,t)) = \Theta(x) + t(0, \dots, 0, 1)$, for each $x \in U$, $(x,t) \in D$.*

Proof. Let $S = \{x \in \mathbb{R}^d \ : \ \langle x - x_0, V(x_0)\rangle = 0\}$ and $\{e_i\}_{i=1}^{d-1} \subset S$ the an orthonormal base.[1] For $r > 0$ small enough let $D_r = \{z \in \mathbb{R}^d \mid |z_i| \leq r\}$.

[1] That is $\langle e_i, e_j \rangle = \delta_{ij}$.

Then define $\Xi : D_r \to U$ by $\Xi(\xi) = \phi(x_0 + \sum_{i=1}^{d-1} \xi_i e_i, \xi_d)$. Note that Ξ is invertible since if $\Xi(\xi) = \Xi(\xi')$, $\xi'_d \leq \xi_d$, it would be

$$\phi(x_0 + \sum_{i=1}^{d-1} \xi_i e_i, \xi_d - \xi'_d) = x_0 + \sum_{i=1}^{d-1} \xi'_i e_i.$$

That is there would be $x \in S$ and $\tau = \xi_d - \xi'_d \in (0, 2r)$ such that $\phi(x, \tau) \in S$. But $\langle V(x_0), \phi(x, 0)\rangle = \langle V(x_0), \phi(x, \tau)\rangle = 0$ by definition and, for $t \in [0, 2r]$,

$$\frac{\langle V(x_0), \phi(x,t)\rangle}{dt} = \langle V(x_0), V(\phi(x,t))\rangle > 0$$

provided that r is chosen small enough. Hence $\xi_d = \xi'_d$ and, consequently, $\xi = \xi'$. We can then define $\Theta = \Xi^{-1}$ and, for each $x = \Xi(\xi)$,

$$\begin{aligned}\Theta(\phi(x,t)) &= \Theta(\phi(\phi(x_0 + \sum_{i=1}^{d-1} \xi_i e_i, \xi_d), t) = \Theta(\phi(x_0 + \sum_{i=1}^{d-1} \xi_i e_i, \xi_d + t)) \\ &= \Theta(\Xi(\xi + (0, \ldots, 0, t))) = \xi + (0, \ldots, 0, t) \\ &= \Theta(x) + (0, \ldots, 0, t).\end{aligned}$$

□

2.2 Behavior close to a fixed point

Here we consider a more interesting situation: the study of the solutions of (2.1.1) in a neighborhood of x_0 such that $V(x_0) = 0$ and $\det(D_{x_0} V) \neq 0$.

Problem 2.1 *Note that the condition* $\det(D_{x_0} V) \neq 0$ *can always be achieved by a small* $\mathcal{C}^1$ *change of the vector field. On the contrary, a zero of the vector field cannot be eliminated by small* $\mathcal{C}^1$ *changes of the vector field: prove that if* $V(x_0) = 0$ *and* W *is a vector field* $\mathcal{C}^1$ *close enough to* V*, then there exists a* x_* *close to* x_0 *such that* $W(x_*) = 0$*, and* $D_{x_*} W$ *is close to* $D_{x_0} V$*. In this sense we will say that the above conditions are* generic *(more on this concept later).*

Let us understand the behavior of the equation in the vicinity of x_0. First of all, by a translation, we can assume without loss of generality $x_0 = 0$. Then we can develop V by the Taylor formula to obtain

$$\dot{x} = Ax + R(x) \tag{2.2.2}$$

where $\|R(x)\| \leq C\|x\|^2$ and $\|D_x R\| \leq C\|x\|$, for all $\|x\| \leq 1$.

Problem 2.2 *Show that, by a linear change of variable, one can transform A in its Jordan canonical form. Show then that, by an arbitrary small $\mathcal{C}^1$ change of the vector field one can eliminate all the Jordan blocks and insure that all the eigenvalues have real part different from zero: this is called the* hyperbolic case.

For now, in view of Problem 2.2, we limit ourselves to the hyperbolic case.

We will start by considering the case in which all the eigenvalues of A have real part strictly smaller than zero.

Problem 2.3 *Prove that if A is diagonal with eigenvalues with real part strictly smaller than zero, then there exists $\sigma > 0$ such that, for all $x \in \mathbb{C}^n$,*[2]

$$\Re(\langle x, Ax\rangle) \leq -\sigma\langle x, x\rangle \tag{2.2.3}$$

Prove that, for a general diagonalizable matrix A with all the eigenvalues with real part strictly smaller than zero, there exists a strictly positive matrix B such that, for all $x \in \mathbb{C}^n$,

$$\Re(\langle x, BAx\rangle) \leq -\sigma\langle x, Bx\rangle.$$

That is, we have the same inequality for the scalar product $\langle\cdot,\cdot\rangle_B := \langle\cdot, B\cdot\rangle$.

Problem 2.4 *Prove that, if $\Re(\langle x, Ax\rangle) \leq -\sigma\langle x, x\rangle$, then the solutions of the equation $\dot{x} = Ax$ satisfy $\|x(t)\| \leq e^{-\sigma t}\|x(0)\|$.*

Till the end of this section, we assume that all the eigenvalues of A are strictly negative, hence we assume (2.2.3) (with respect to the appropriate scalar product). In this case, it is well known that the linear part of (2.2.2) has solutions that tend to zero exponentially fast, the question is: does the same holds true for the solutions of the equation (2.2.2)?

To see it, consider $z := \langle x, x\rangle$. Let $z^{\frac{1}{2}} = \|x\| \leq \frac{\sigma}{2C}$, then, recalling Problem 2.3,

$$\begin{aligned}\frac{d}{dt}z &= \langle x, Ax + R(x)\rangle + \langle Ax + R(x), x\rangle \\ &\leq \langle x, (A + A^*)x\rangle + 2C\|x\|^3 = 2\Re(\langle x, (A + A^*)x\rangle) + 2C\|x\|^3 \\ &\leq -2\sigma z + 2Cz^{\frac{3}{2}} \leq -\sigma z\end{aligned}$$

which, seeting $z(t) = e^{-\sigma t}\zeta(t)$, implies $\dot{\zeta} \leq 0$, hence $z(t) \leq e^{-\sigma t}\zeta(0)$ and

$$\|x(t)\| \leq e^{-\frac{\sigma}{2}t}\|x(0)\|. \tag{2.2.4}$$

That is, also the solutions of (2.2.2) tend exponentially fast to zero.

[2] As usual $\langle x, y\rangle := \sum_{i=1}^n \bar{x}_i y_i$ where $\bar{a}$ is the complex conjugate of a. Moreover by A^* we mean the adjoint of A.

Remark 2.2.1 *What we have just seen is that, locally, $F(x) := \langle x, x\rangle$ is a Lyapunov function for (2.2.2). Given a differential equation like (2.1.1), where 0 is a fixed point, a local Lyapunov function on an open set $U \ni 0$ is any $L \in \mathcal{C}^1(U, \mathbb{R})$ such that $L(0) = 0$, $L \geq 0$ and $\langle \nabla_x L, V(x)\rangle < 0$ for all $x \in U \setminus \{0\}$. Then, for each solution $x(t)$ of (2.1.1) holds*

$$\frac{dL(x(t))}{dt} = \langle \nabla_{x(t)} L, V(x(t))\rangle < 0.$$

This readily implies that $\lim_{t\to\infty} x(t) = 0$. (Prove it !).

Yet, the above result is far from being satisfactory: it is possible to tend to zero in many different ways and it would be nice to understand better how this happens.

Let us start with a very simple example: $x \in \mathbb{R}$, $A = -1$, $R(x) = bx^2$. Then the equation reads

$$\dot{x} = -x + bx^2. \tag{2.2.5}$$

If we consider the change of variables

$$z = \Psi(x) = \frac{x}{1 - bx}$$

we have

$$\dot{z} = \frac{-x + bx^2}{1 - bx} + \frac{bx(-x + bx^2)}{(1 - bx)^2} = -\frac{x}{1 - bx} = -z.$$

Thus, in a neighborhood of zero of size smaller than b^{-1} there exists a smooth diffeomorphism that conjugates the solution of (2.2.5) with its linear part.

One can then suspect that this is always the case. This is not so: consider

$$\begin{aligned} \dot{x} &= -2x + cy^2 \\ \dot{y} &= -y \end{aligned} \tag{2.2.6}$$

Let us consider a change of variables

$$\begin{aligned} z &= x + \alpha x^2 + \beta xy + \gamma y^2 + q(x, y) \\ \eta &= y + p(x, y) \end{aligned}$$

where q is of third order and p of second. Substituting in (2.2.6) one can see that it is always possible to choose $p \equiv 0$, while the first of the (2.2.6) yields

$$\dot{z} = -2x + cy^2 - 2x(2\alpha x + \beta y) - y(\beta x + 2\gamma y) + \mathcal{O}(3)$$

where by $\mathcal{O}(3)$ we designate third order terms. If we try to impose the right hand side of the above equation equal to $-2z$ (up to second order) we obtain

$$-2\alpha x^2 - 2\beta xy - 2\gamma y^2 = -4\alpha x^2 - 3\beta xy - (2\gamma + c)y^2$$

that does not admit any solutions if $c \neq 0$.

So there is no hope of finding a C^3 conjugation with the linear part.

What can be salvaged?

2.2.1 Grobman–Hartman

One can look for a less regular change of variables. This may not make sense for the o.d.e. itself but it is meaningful for the associated flows.

Theorem 2.2.2 (Grobman–Hartman) *If ϕ_t is the flow associated to the vector filed V, $V(x_0) = 0$, and ϕ_t^0 is the flow associated to the linearized vector field $D_{x_0}V$, that we assume hyperbolic (see Problem 2.2), then for all $t_* > 0$, there exists a local homeomorphism Ξ such that $\Xi \circ \phi_{t_*} = \phi^0_{t_*} \circ \Xi$.*

Proof. We do the proof in the case $t_* = 1$, the other cases being similar. Thus let us fix some small $r > 0$ and consider a smooth non increasing function $g : \mathbb{R}_+ \to [0, 1]$ such that $g(x) = 1$ for $x \le r$ and $g(x) = 0$ for $x \ge 2r$, with $-g' \le C$. We can then define the functions $\varphi : \mathbb{R}^d \to [0, 1]$ $F_0, F : \mathbb{R}^d \to \mathbb{R}^d$ as $\varphi(x) := g(\|x\|)$ and[3]

$$F_0(x) := e^A x$$
$$F(x) := e^A x + \varphi(x)\left[\phi_1(x) - e^A x\right] =: F_0(x) + \Delta(x),$$

where ϕ_1 is the time one flow associated to (2.2.2). We are considering first the case in which all the eigenvalues of A have strictly negative real part. Clearly, for $\|x\| \le r$ the two functions are simply the time one map of the linear flow and the time one map of (2.2.2), moreover, they are globally Lip. Since we will be interested only in x in the ball of radius r, the modification outside such a ball is totally irrelevant, and it has been done only to facilitate the exposition of the following argument.

Problem 2.5 *Show that, for r small enough, F is a diffeomorphism. Prove that $\|\Delta\|_\infty \le 4Cr^2$.*

The idea is to consider the maps $F_0, F : \mathbb{R}^d \to \mathbb{R}^d$ and to show that they can be conjugated, that is there exists an homeomorphism $\Xi : \mathbb{R}^d \to \mathbb{R}^d$ such that $\Xi \circ F = F_0 \circ \Xi$.

Let us look for a solution in the form $\Xi(x) = x + \Phi(x)$, then we have

$$F_0(x + \Phi(x)) = F(x) + \Phi(F(x))$$

or, setting $\xi = F(x)$,

$$\Phi(\xi) = F_0(F^{-1}(\xi) + \Phi \circ F^{-1}(\xi)) - \xi.$$

We define then the operator $K : \mathcal{C}^0(\mathbb{R}^d) \to \mathcal{C}^0(\mathbb{R}^d)$ defined by

$$K(\Phi)(\xi) := F_0(F^{-1}(\xi) + \Phi \circ F^{-1}(\xi)) - \xi$$

[3]Here and in the following of the proof, we use the norm determined by the scalar product introduced in Problem 2.3.

then our problem boils down to establishing the existence of a fixed point for K. First of all, by Problem 2.5, for each $\|\xi\| \geq 2r + 4Cr^2$ we have $\|x\| \geq 2r$. Hence, recalling Problem 2.4 and equation (2.2.4), it follows

$$\|K(\Phi)(\xi)\| = \|F_0(F_0^{-1}(\xi) + \Phi \circ F_0^{-1}(\xi)) - \xi\| \leq \|F_0(\Phi \circ F_0^{-1}(\xi)\| \leq e^{-\sigma}\|\Phi\|_\infty.$$

On the other hand, if $\|x\| < 2r$, then $\|\xi\| < 2r + 4Cr^2$, and

$$\|K(\Phi)(\xi)\| \leq e^{-\sigma}\left[\|x\| + \|\Phi\|_\infty\right] + 2r + 4Cr^2 \leq 4r + 4Cr^2 + e^{-\sigma}\|\Phi\|_\infty.$$

Thus the set $\{h \in \mathcal{C}^0 \ : \ \|h\|_\infty \leq (4r + 4r^2C)(1 - e^{-\sigma})^{-1}\}$ is invariant for the operator K. Next, given two functions $h, g \in \mathcal{C}^0(\mathbb{R}^d)$, holds

$$\begin{aligned}\sup_{\xi \in \mathbb{R}^d} \|K(h)(\xi) - K(g)(\xi)\| &= \sup_{x \in \mathbb{R}^d} \|F_0(x + h(x)) - F_0(x + g(x))\| \\ &\leq e^{-\sigma}\|h - g\|_\infty.\end{aligned}$$

Thus, the contracting mapping theorem yields the wanted result.

Problem 2.6 *What can be done if all the eigenvalues of A have strictly positive real part?*

We have then, topologically, the behavior of a source, a node, or a stable or unstable focus are the same as the ones of the linear part of the equation. But the generic case is the one in which both eigenvalues with positive and negative real parts are present; do the same conclusions hold for such a more general situation? The answer is yes. To see it consider that in such a case $\mathbb{R}^d$ is naturally split into two spaces $V \oplus W$, invariant for A and such that A restricted to V has only eigenvalues with negative real part, while restricted to W has eigenvalues with positive real part. Then the spaces are invariant for F_0 as well, on one F_0 contracts, on the other expands. Call d_s the dimension of V and d_u the dimension of W. Clearly $d_s + d_u = d$.

Then each $e \in \mathbb{R}^d$ has a unique splitting as $e = v + w$, $v \in V$, $w \in W$. It is then convenient to define the projections $p_1 : \mathbb{R}^d \to V$ and $p_2 : \mathbb{R}^d \to W$ $p_1(e) = v$, $p_2(e) = w$. Moreover, we can split $\mathcal{C}^0(\mathbb{R}^d, \mathbb{R}^d)$ as $\mathbb{V} \oplus \mathbb{W}$ where $\mathbb{V} := \{f \in \mathcal{C}^0(\mathbb{R}^d, \mathbb{R}^d) \ : \ p_2 \circ f = 0\}$ and $\mathbb{W} := \{f \in \mathcal{C}^0(\mathbb{R}^d, \mathbb{R}^d) \ : \ p_1 \circ f = 0\}$. We can then write canonically f as $(f_1, f_2) := (p_1 \circ f, p_2 \circ f)$. Analogously, we can write $(x_1, x_2) = (p_1(e), p_2(e))$.

Accordingly our conjugation equation $F_0 \circ \Xi = \Xi \circ F$, becomes

$$\begin{aligned}B\Xi_1 &= \Xi_1 \circ F \\ D\Xi_2 &= \Xi_2 \circ F\end{aligned}$$

where $F_0((x_1, x_2)) =: (Bx_1, Dx_2)$. We transform the first equation as we did for the contracting case, while on the second we act as you probably did if

you solved Problem 2.6:

$$\Xi_1 = B\Xi_1 \circ F^{-1}$$
$$\Xi_2 = D^{-1}\Xi_2 \circ F.$$

Again, we look for solutions of the form $\Xi_i(x) = x_i + \Phi_i(x)$, where Φ_i are bounded. Substituting such a form for Ξ, one can see that bounded functions are mapped into bounded functions (thanks to Problem 2.5), hence the contracting map argument applies, and the existence of a unique conjugation is established. □

2.3 Dominated Splitting and center manifold

Let $U \subset \mathbb{R}^d$ be an open set containing zero and let us consider a vector field $V \in \mathcal{C}^k(U, \mathbb{R}^d)$, $k \geq 1$, such that $V(0) = 0$ and $A := D_0V$ has a spectrum that *splits* into two disjoint parts. More precisely, assume there exists real numbers $\alpha < \beta$, such that $\sigma(A) = \Sigma_1 \cup \Sigma_2$ where $\mu \in \Sigma_1$ implies $\Re(\mu) \geq \beta$ and $\mu \in \Sigma_2$ implies $\Re(\mu) \leq \alpha$. Let $\mathbb{V}_1, \mathbb{V}_2$ be the eigenspaces associated to Σ_1, Σ_2, respectively.

We say that a manifold W is locally invariant at zero under the flow ϕ_t generated by the vector field V if there exists $\delta > 0$ such that, for all $t \in \mathbb{R}$, there exists $\delta_t \in (0, \delta]$ such that $\phi_t(W \cap B(0, \delta_t)) \subset W$.

Note that, letting $\widetilde{R}(x) := V(x) - Ax$, we can then write the differential equation as

$$\dot{x} = Ax + \widetilde{R}(x). \tag{2.3.7}$$

In the special case $\tilde{R} \equiv 0$, the differential equation is linear and the subspaces $\mathbb{V}_i$ are invariant manifolds for the above differential equation. It is then natural to wonder if there exists invariant manifolds also for the non linear case. Note that the nonlinearity is small only in a neighborhood of zero, it is then natural to look for local invariant manifolds at zero.

We are thus interested in the solutions of (2.3.7) only in a neighborhood of zero. It is then convenient to modify the equation outside the ball $B(0, \delta)$ so that the dynamics is linear outside such a ball. This will allow us to look for a globally invariant manifold for the modified dynamics with the property of bein locally invariant for the original one.

Namely, let $\varphi \in \mathcal{C}^\infty(\mathbb{R}_+, [0, 1])$, be a decreasing function such that $\varphi(t) = 1$ for $t \leq \delta$ and $\varphi(t) = 0$ for $t \geq 2\delta$. We then define $R(x) = \widetilde{R}(x)\varphi(\|x\|)$. Clearly, if we construct an invariant manifold for the differential equation

$$\dot{x} = Ax + R(x),$$

then it is a locally invariant manifold for (2.3.7) as well. By the variation of constant formula we have

$$x(t) = e^{At}x(0) + \int_0^t e^{A(t-s)}R(x(s))ds.$$

To put the problem into a more general context it is convenient to define, for a given τ small enough, the map $F \in \mathcal{C}^k$ such that $F(x(0)) = x(\tau)$.

Problem 2.7 *Prove that*

1. *F is invertible;*
2. *we can choose $\delta > 0$ such that $F(B(0, 3e^{\|A\|\tau}\delta)) \supset B(0, 2\delta)$;*
3. *$F(0) = 0$, $D_0F = e^{A\tau}$ and $D_xF = e^{A\tau}$ for $\|x\| \geq 3e^{\|A\|\tau}\delta$;*
4. *for each $\varepsilon > 0$ we can chose δ such that $\|D_xF - e^{A\tau}\|_\infty \leq \varepsilon$;*
5. *for some $\beta > \beta' > \alpha' > \alpha$, $\|e^{-A\tau}|_{\mathbb{V}_1}\| \leq e^{-\beta'\tau}$ and $\|e^{A\tau}|_{\mathbb{V}_2}\| \leq e^{\alpha'\tau}$.*

Problem 2.8 *Show that a manifold W is locally invariant at zero for (2.3.7) if and only if it is so for F.*

The above shows the relevance of the following theorem

Theorem 2.3.1 *Let $F \in \mathcal{C}^k(\mathbb{R}^d, \mathbb{R}^d)$, $k \geq 1$, be an invertible map from $\mathbb{R}^d$ to itself such that it enjoys the properties of Problem 2.7 and, for a sufficiently small ε, $\|D_xF - D_0F\|_\infty \leq \varepsilon$. Then, there exists a $\mathcal{C}^{k-1}$ locally invariant manifold W. Also, W is dim$(\mathbb{V}_1)$ dimensional and tangent to $\mathbb{V}_1$ at zero.*

Proof. By the hypotheses $\sigma(D_0F)$ splits in two parts $\tilde{\Sigma}_1, \tilde{\Sigma}_2$. Let $\mathbb{V}_1, \mathbb{V}_2$ be the associated eigenspaces. By a change of variable we can assume that $\mathbb{V}_1 = \{(\xi, 0)\}_{\xi \in \mathbb{R}^{d_1}}$ and $\mathbb{V}_2 = \{(0, \eta)\}_{\xi \in \mathbb{R}^{d_2}}$. Also, let $\Pi_1(\xi, \eta) = (\xi, 0)$, $\Pi_2 = \mathbb{1} - \Pi_1$, $\Pi_1 D_0F \Pi_1 = \Lambda$ and $\Pi_2 D_0F \Pi_2 = \Gamma$. In addition,[4] the hypotheses imply that $\|\Lambda^{-1}\| \leq e^{-\beta}$ and $\|\Gamma\| \leq e^{\alpha}$ with $\alpha < \beta$.

The basic idea is to consider manifolds that can be described by a function $G : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ via $W = \{(\xi, G(\xi)\}_{\xi \in \mathbb{R}^{d_1}}$. Obviously we need to limit the set to which G might belong. To this end we define,

$$\Omega = \{G \in \mathcal{C}^k(\mathbb{R}^{d_1}, \mathbb{R}^{d_1}) \ : \ G(0) = 0, \ \|DG\|_\infty \leq 1\}.$$

Let

$$F(\xi, \eta) = (\Lambda\xi + A(\xi, \eta), \Gamma\eta + B(\xi, \eta)).$$

[4] For convenience I am renaming the constants α, β and, possibly, substituting F^n to F in order to offsets the constants coming from the equivalence of the norms in the new coordinates.

If $\|\eta\| \leq \|\xi\|$ and ε is small enough, we have that there exists $\beta' > \alpha$ such that

$$\|\Lambda\xi + A(\xi,\eta)\| \geq e^{\beta'}\|\xi\|.$$

Thus, for each $G \in \Omega$ the map $T_G(\xi) = \Lambda\xi + A(\xi, G(\xi))$ is invertible. Moreover, for $\|\xi\| \geq C\delta$ we have $T_G(\xi) = \Lambda\xi$. We can then describe the evolution of the manifolds of interest:

$$F(\xi, G(\xi)) = (T_G(\xi), S_G \circ T_G^{-1}(T_G(\xi)))$$

where $S_G(\xi) = \Gamma G(\xi) + B(\xi, G(\xi))$. Again note that, for $\|\xi\| \geq C\delta$ we have $S_G(\xi) = \Gamma G(\xi)$. It follows that the image manifold is described by the operator $K : \Omega \to \mathcal{C}^k(\mathbb{R}^d, \mathbb{R}^d)$

$$K(G)(\xi) = S_G \circ T_G^{-1}(\xi).$$

For $G \in \Omega$, $K(G)(0) = 0$. Also

$$D[K(G)] = \left[(\Gamma DG + \partial_\xi A + \partial_\eta ADG)(\Lambda + \partial_\xi B + \partial_\eta BDG)^{-1}\right] \circ T_G^{-1}.$$

Note that, if $DG(0) = 0$, then also $D(K(G))(0) = 0$.

From the above computations it follows that, for ε small enough, there exists $\sigma \in [0,1]$ such that

$$\|D[K(G)]\|_\infty \leq \sigma\|DG\|_\infty + C\varepsilon < \|DG\|_\infty. \tag{2.3.8}$$

Accordingly, $K(\Omega) \subset \Omega$. A direct computation shows that, for $G_1, G_2 \in \Omega$,

$$\begin{aligned}
&\|T_{G_1} - T_{G_2}\|_\infty \leq C_\# \varepsilon \|G_1 - G_2\|_\infty \\
&\|S_{G_1} - S_{G_2}\|_\infty \leq (e^\alpha + C_\#\varepsilon)\|G_1 - G_2\|_\infty.
\end{aligned}$$

On the other hand, for all $\xi \in \mathbb{R}^{d_1}$,

$$\begin{aligned}
\|T_{G_1}^{-1}(\xi) - T_{G_2}^{-1}(\xi)\| &= \|T_{G_2}^{-1} \circ T_{G_2} \circ T_{G_1}^{-1}(\xi) - T_{G_2}^{-1}(\xi)\| \\
&\leq (e^{-\beta} + C_\#\varepsilon)\|T_{G_2} \circ T_{G_1}^{-1}(\xi) - T_{G_1} \circ T_{G_1}^{-1}(\xi)\| \\
&\leq C_\#(e^{-\beta} + C_\#\varepsilon)\varepsilon\|G_1 \circ T_{G_1}^{-1}(\xi) - G_2 \circ T_{G_1}^{-1}(\xi)\|.
\end{aligned}$$

To conclude we introduce the norm[5]

$$|\!|\!|G|\!|\!| = \sup_{\xi \in \mathbb{R}^{d_1}} \|G(\xi)\| \cdot \|\xi\|^{-1}.$$

[5]This norm is necessary only because we do not assume $\alpha < 0$. If we would do so, then the usual sup norm would work perfectly.

Remark that if $G \in \Omega$, then $\|\!|G|\!\| \leq 1$. Next, note that

$$\begin{aligned}\|K(G_1)(\xi) - K(G_2)(\xi)\| &\leq \|S_{G_1} \circ T_{G_1}^{-1}(\xi) - S_{G_1} \circ T_{G_2}^{-1}(\xi)\| \\ &\quad + \|S_{G_1} \circ T_{G_2}^{-1}(\xi) - S_{G_2} \circ T_{G_2}^{-1}(\xi)\| \\ &\leq (e^{\alpha} + C_{\#}\varepsilon)\|T_{G_1}^{-1}(\xi) - T_{G_2}^{-1}(\xi)\| + (e^{\alpha} + C_{\#}\varepsilon)\|G_1 \circ T_{G_2}^{-1}(\xi) - G_2 \circ T_{G_2}^{-1}(\xi)\|.\end{aligned}$$

Accordingly,

$$|\!|\!|K(G_1)(\xi) - K(G_2)|\!|\!| \leq \left[C_{\#}(e^{-\beta} + \varepsilon)\varepsilon e^{-\beta'} + (e^{\alpha} + C_{\#}\varepsilon)e^{-\beta'}\right] |\!|\!|G_1 - G_2|\!|\!|$$

Hence, provided ε is small enough, there exists $\sigma \in (0,1)$, such that for each $G_1, G_2 \in \Omega$

$$|\!|\!|K(G_1) - K(G_2)|\!|\!| \leq \sigma \, |\!|\!|G_1 - G_2|\!|\!| \, .$$

The above implies that K has a unique fixed point $G = \lim_{n\to\infty} K^n(0)$. In addition, G is of the form $G(\xi) = \|\xi\|\hat{G}(\xi)$ with $\hat{G} \in \mathcal{C}^0$.

We leave to the reader the task of checking that the contraction takes place in $\mathcal{C}^{k-1}$ as well. In particular, if $k \geq 2$, it is trivial to check that $DG(0) = 0$. □

From the above, we directly obtain the following very useful result.

Theorem 2.3.2 (Center Manifold Theorem) *Let $F \in \mathcal{C}^k$ be an invertible map from $\mathbb{R}^d$ to itself such that it enjoys the properties (1-4) of Problem 2.7. Moreover assume that the spectrum of the matrix A now splits into three disjoint parts $\Sigma_- \cup \Sigma_0 \cup \Sigma_+$ such that $\mu \in \Sigma_-$ implies $\Re(\mu) \leq \alpha < 0$, $\mu \in \Sigma_0$ implies $\alpha < \Re(\mu) < \beta$ and $\mu \in \Sigma_+$ implies $\Re(\mu) \geq \beta > 0$. Let $\mathbb{V}_0$ be the eigenspace associated to Σ_0 and d_0 be its dimension. Then, there exists a $\mathcal{C}^{k-1}$ d_0-dimensional locally invariant manifold W. In addition, W is tangent to $\mathbb{V}_0$ at zero.*

Proof. Let $\mathbb{V}_+, \mathbb{V}_0, \mathbb{V}_-$ be the eigenspaces associated with the splitting of the spectrum and d_+, d_0, d_- be their dimensions. Simply apply Theorem 2.3.1 to F with the splittings $\Sigma_1 = \Sigma_+ \cup \Sigma_0$, $\Sigma_2 = \Sigma_-$ and to F^{-1} with the splitting $\Sigma_1 = \Sigma_+$, $\Sigma_2 = \Sigma_- \cup \Sigma_0$. In such a way, we obtain two invariant manifolds: W^+ (the weak unstable manifold) and W^- (the weak stable manifold), respectively of dimension $d_+ + d_0$ and $d_- + d_0$. The reader can easily check that the hypotheses of the implicit function theorem apply and prove that $W = W^+ \cap W^-$ is a d_0 dimensional $\mathcal{C}^{k-1}$ locally invariant manifold tangent to $\mathbb{V}_0$ in zero.[6] □

[6] To show that the matrix at zero is invertible, remember (2.3.8) which says that the manifolds are graphs of functions with derivative strictly less than one.

2.4 Hadamard-Perron

Theorem 2.3.2 is quite general but it has a couple of disadvantages: a slightly annoying loss of regularity (from $\mathcal{C}^k$ to $\mathcal{C}^{k-1}$) and, most importantly, it does not provides any information on the dynamics when restricted to the invariant manifold which, in fact, can be pretty much anything. To eliminate such shortcoming it is necessary to consider situations in which there are no eighevalues with zero real part. This gives rise to a sharper results: the Hadamard-Perron theorem. We will discuss it in the simplest possible setting, also we will repeat several arguments to make this section independent on the previous one.

Definition 2.4.1 *Given a smooth map $T : X \to X$, X being a Riemannian manifold, and a fixed point $p \in X$ (i.e. $Tp = p$) we call (local) stable manifold (of size δ) a manifold $W^s(p)$ such that*[7]

$$W^s(p) = \{x \in B_\delta(x) \subset X \mid \lim_{n\to\infty} d(T^n x, p) = 0\}.$$

Analogously, we will call (local) unstable manifold (of size δ) a manifold $W^u(p)$ such that

$$W^u(p) = \{x \in B_\delta(x) \subset X \mid \lim_{n\to\infty} d(T^{-n} x, p) = 0\}.$$

It is quite clear that $TW^s(p) \subset W^s(p)$ and $TW^u(p) \supset W^u(p)$ (Problem 2.10). Less clear is that these sets deserve the name "manifold." Yet, if one thinks of a linear map it is obvious that the stable and unstable manifolds at zero are just segments in the stable and unstable direction, the next Theorem shows that this is a quite general situation.

Theorem 2.4.2 (Hadamard-Perron) *Consider an invertible map $T : U \subset \mathbb{R}^2 \to \mathbb{R}^2$, $T \in \mathcal{C}^1(U, \mathbb{R}^2)$, such that $T0 = 0$ and*

$$D_0 T = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix} \tag{2.4.9}$$

where $0 < \mu < 1 < \lambda$.[8] *That is, the map T is hyperbolic at the fixed point 0. Then there exists unique $\mathcal{C}^1$ stable and unstable manifolds at 0. Moreover, $\mathcal{T}_0 W^{s(u)} = E^{s(u)}$ where $E^{s(u)}$ are the expanding and contracting subspaces of $D_0 T$.*[9]

[7] Sometime we will write $W^s_\delta(p)$ when the size really matters. By $B_\delta(x)$ we will always mean the open ball of radius δ centered at x.

[8] Notice that if $D_0 T$ has eigenvalues $0 < \mu < 1 < \lambda$ then one can always perform a change of variables such that (2.4.9) holds.

[9] By $\mathcal{T}_0 W^{s(u)}$ I mean the tangent space to the manifold (curve) W^u (or W^s) at the point zero.

Remark 2.4.3 *There is an issue not completely addresses in our formulation of Hadamard-Perron theorem: the uniqueness of the manifolds.*[10] *It is not hard to prove that the $W^{s(u)}$ are indeed the only sets satisfying Definition 2.4.1 (see Problem 2.13).*

The proof of Theorem 2.4.2 will be done in two steps: first we will show the existence of the invariant manifolds and then we will prove the regularity.

2.4.1 Invariant manifolds–existence

We will deal explicitly only with the unstable manifold since the stable one can be treated exactly in the same way by considering T^{-1} instead of T.

Proof of existence of the unstable manifold. Since the map is continuously differentiable for each $\varepsilon > 0$ we can choose $\delta > 0$ so that, in a 2δ-neighborhood of zero, we can write

$$T(x) = D_0 T x + R(x) \tag{2.4.10}$$

where $\|R(x)\| \leq \varepsilon\|x\|$, $\|D_x R\| \leq \varepsilon$.

The first step is to decide how to represent manifolds. In the present case, since we deal only with curves, it seems very reasonable to consider the set of curves $\Gamma_{\delta,c}$ passing through zero and "close" to being horizontal, that is the differentiable functions $\gamma : [-\delta, \delta] \to \mathbb{R}^2$ of the form

$$\gamma(t) = \begin{pmatrix} t \\ u(t) \end{pmatrix}$$

and such that $\gamma(0) = 0$; $\|(1,0) - \gamma'\|_\infty \leq c$. It is immediately clear that any smooth curve passing through zero and with tangent vector, at each point, in the cone $\mathcal{C} := \{(a, b) \in \mathbb{R}^2 \mid |\frac{b}{a}| \leq c\}$, can be associated to a unique element of $\Gamma_{\delta,c}$, just consider the part of the curve contained in the strip $\{(x, y) \in \mathbb{R}^2 \mid |x| \leq \delta\}$. Moreover, if $\gamma \in \Gamma_{\delta,c}$ then $\gamma \subset B_{2\delta}(0)$, provided $c \leq 1/2$.

Notice that it suffices to specify the function u in order to identify uniquely an element in $\Gamma_{\delta,c}$. It is then natural to study the evolution of a curve through the change in the associated function.

To this end let us investigate how the image of a curve in $\Gamma_{\delta,c}$ under T looks like.

$$T\gamma(t) = \begin{pmatrix} \lambda t + R_1(t, u(t)) \\ \mu u(t) + R_2(t, u(t)) \end{pmatrix} := \begin{pmatrix} \alpha_u(t) \\ \beta_u(t) \end{pmatrix}.$$

At this point the problem is clearly that the image it is not expressed in the way we have chosen to represent curves, yet this is easily fixed. First of

[10] Namely the doubt may remain that a less regular set satisfying Definition 2.4.1 exists.

all, $\alpha_u(0) = \beta_u(0) = 0$. Second, by choosing $\varepsilon < \lambda$, we have $\alpha_u'(t) > 0$, that is, α_u is invertible. In addition, $\alpha_u([-\delta, \delta]) \supset [-\lambda\delta + \varepsilon\delta, \lambda\delta - \varepsilon\delta] \supset [-\delta, \delta]$, provided $\varepsilon \le \lambda^{-1}$. Hence, α_u^{-1} is a well defined function from $[-\delta, \delta]$ to itself. Finally,

$$|\frac{d}{dt}\beta_u \circ \alpha_u^{-1}(t)| = \left|\frac{\beta_u'(\alpha_u^{-1}(t))}{\alpha_u'(\alpha_u^{-1}(t))}\right| \le \frac{\mu c + \varepsilon}{\lambda - \varepsilon} \le c$$

where, again, we have chosen $\varepsilon \le \frac{c(\lambda-\mu)}{1+c}$.

We can then consider the map $\tilde{T} : \Gamma_{\delta,c} \to \Gamma_{\delta,c}$ defined by

$$\tilde{T}\gamma(t) := \begin{pmatrix} t \\ \beta_u \circ \alpha_u^{-1}(t) \end{pmatrix} \tag{2.4.11}$$

which associates to a curve in $\Gamma_{\delta,c}$ its image under T written in the chosen representation. It is now natural to consider the set of functions $B_{\delta,c} = \{u \in \mathcal{C}^1([-\delta, \delta]) \mid u(0) = 0,\ |u'|_\infty \le c\}$ in the vector space $Lip([-\delta, \delta])$.[11] As we already noticed $B_{\delta,c}$ is in one-one correspondence with $\Gamma_{\delta,c}$, we can thus consider the operator $\hat{T} : Lip([-\delta, \delta]) \to Lip([-\delta, \delta])$ defined by

$$\hat{T}u = \beta_u \circ \alpha_u^{-1} \tag{2.4.12}$$

From the above analysis follows that $\hat{T}(B_{\delta,c}) \subset B_{\delta,c}$ and that $\hat{T}u$ determines uniquely the image curve.

The problem is then reduced to studying the map $\hat{T}$. The easiest, although probably not the most productive, point of view is to show that $\hat{T}$ is a contraction in the sup norm. Note that this creates a little problem since $\mathcal{C}^1$ it is not closed in the sup norm (and not even $Lip([-\delta, \delta])$ is closed). Yet, the set $B^*_{\delta,c} = \{u \in Lip([-\delta, \delta]) \mid u(0) = 0, \sup_{t,s\in[-\delta,\delta]} \frac{|u(s)-u(t)|}{|t-s|} < c\}$ is closed (see Problem 2.11). Thus $\overline{B_{\delta,c}} \subset B^*_{\delta,c}$. This means that, if we can prove that the sup norm is contracting, then the fixed point will belong to $B^*_{\delta,c}$ and we will obtain only a Lipschitz curve. We will need a separate argument to prove that the curve is indeed smooth.

Let us start to verify the contraction property. Notice that

$$\alpha_u^{-1}(t) = \lambda^{-1}t + \lambda^{-1}R_1(\alpha_u^{-1}(t), u(\alpha_u^{-1}(t))),$$

thus, given $u_1, u_2 \in B_{\delta,c}$, by Lagrange Theorem

$$\begin{aligned} |\alpha_{u_1}^{-1}(t) - \alpha_{u_2}^{-1}(t)| &\le \lambda^{-1}|\langle \nabla_\zeta R_1,\ (\alpha_{u_1}^{-1}(t) - \alpha_{u_2}^{-1}(t), u_1(\alpha_{u_1}^{-1}(t)) - u_2(\alpha_{u_2}^{-1}(t)))\rangle| \\ &\le \frac{\varepsilon}{\lambda}\left\{|\alpha_{u_1}^{-1}(t) - \alpha_{u_2}^{-1}(t)| + |u_1(\alpha_{u_2}^{-1}(t)) - u_2(\alpha_{u_2}^{-1}(t))|\right\}. \end{aligned}$$

[11] This are the Lipschitz functions on $[-\delta, \delta]$, that is the functions such that $\sup_{t,s\in[-\delta,\delta]} \frac{|u(s)-u(t)|}{|t-s|} < \infty$.

This implies immediately

$$|\alpha_{u_1}^{-1}(t) - \alpha_{u_2}^{-1}(t)| \leq \frac{\lambda^{-1}\varepsilon}{1-\lambda^{-1}\varepsilon}\|u_1 - u_2\|_\infty. \tag{2.4.13}$$

On the other hand

$$\begin{aligned}|\beta_{u_1}(t) - \beta_{u_2}(t)| &\leq \mu|u_1(t) - u_2(t)| + |\langle \nabla_\zeta R_2,\, (0, u_1(t) - u_2(t))\rangle| \\ &\leq (\mu+\varepsilon)\|u_1 - u_2\|_\infty. \end{aligned}\tag{2.4.14}$$

Moreover,

$$|\beta_u'(t)| \leq \mu + \varepsilon. \tag{2.4.15}$$

Collecting the estimates (2.4.13, 2.4.14, 2.4.15) readily yields

$$\begin{aligned}\|\hat{T}u_1 - \hat{T}u_2\|_\infty &\leq \|\beta_{u_1}\circ\alpha_{u_1}^{-1} - \beta_{u_1}\circ\alpha_{u_2}^{-1}\|_\infty + \|\beta_{u_1}\circ\alpha_{u_2}^{-1} - \beta_{u_2}\circ\alpha_{u_2}^{-1}\|_\infty \\ &\leq \left\{[\mu+\varepsilon]\frac{\lambda^{-1}\varepsilon}{1-\lambda^{-1}\varepsilon} + (\mu+\varepsilon)\right\}\|u_1 - u_2\|_\infty \\ &\leq \sigma\|u_1 - u_2\|_\infty,\end{aligned}$$

for some $\sigma \in (0,1)$, provided ε is chosen small enough.

Clearly, the above inequality immediately implies that there exists a unique element $\gamma_* \in \Gamma_{\gamma,c}$ such that $\tilde{T}\gamma_* = \gamma_*$, this is the *local* unstable manifold of 0. □

2.4.2 Invariant manifolds–regularity

As already mentioned, a separate argument is needed to prove that γ_* is indeed a $\mathcal{C}^1$ curve.

To prove this, one possibility could be to redo the previous fixed point argument trying to prove contraction in $\mathcal{C}^1_{Lip}$ (the $\mathcal{C}^1$ functions with Lipschitz derivative); yet this would require to increase the regularity requirements on T. A more geometrical, more instructive and more inspiring approach is the following.

Proof of the regularity of the unstable manifold. Let $\delta > 0$ such that the arguments of section 2.4.1 apply. We want to define local cone fields in the region $\{\xi = (\xi_x, \xi_y) \in \mathbb{R}^2 \,:\, |\xi_x| < \delta\}$. For each $|u| \leq c\delta$ and $0 < \theta \leq c\delta$ we define the affine cone field $\mathcal{C}_\theta(\xi, u) := \{\xi + (a,b) \in \mathbb{R}^2 \,:\, |b - au| \leq \theta\,|a|\}$.[12] As we need to perform a local argument we must localise the cones. To this end we will intersect them with cylinders of the form $D_h(\xi) = \{\xi + (a,b) \in$

[12] A set $\mathcal{C}$ is a cone iff, for all $y \in \mathcal{C}$ and $\alpha \in \mathbb{R}$, $\alpha y \in \mathcal{C}$. A set $\mathcal{C}$ is an affine cone if there exists z such that $\{y - z \,:\, y \in \mathcal{C}\}$ is a cone.

$\mathbb{R}^2 \,:\, |a| \le h\}$. We define thus a local affine cone field (that in the following we will simply call *cone field*) by

$$\mathcal{C}_{\theta,h}(\xi,u) = \mathcal{C}_\theta(\xi,u) \cap D_h(\xi) = \{\xi + (a,b) \in \mathbb{R}^2 \,:\, |a| \le h;\ |b-au| \le \theta\,|a|\}.$$

By the construction in Section 2.4.1, $D_h(\xi) \cap \gamma_* \subset \mathcal{C}_{c\delta,h}(\xi,0)$ for each $\xi \in \gamma_*$. We will study the evolution of such a cone field on γ_*.

For all $\eta \in \mathcal{C}_{\theta,h}(\xi,u)$, if $(a,b) = \eta - \xi$ and $(\alpha,\beta) = T\eta - T\xi$, it holds

$$(\alpha,\beta) = D_0T(a,b) + \mathcal{O}(\varepsilon|a|) = (\lambda a, \mu b) + \mathcal{O}(\varepsilon|a|).$$

and, at the same time, since T is $\mathcal{C}^1$, $\|(\alpha,\beta) - D_\xi T(a,b)\| \le \varepsilon\theta|a|$ provided $h \le h_\theta$ for some h_θ small enough. Thus, setting $(\alpha',\beta') = D_\xi T(a,ua)$ and $u' = \frac{\beta'}{\alpha'}$, one can compute

$$\begin{aligned}\|(\alpha,\beta) - (\alpha',\beta') - (0,\mu(b-ua))\| &\le \|(D_\xi T - D_0T)(0,b-ua)\| + \theta\varepsilon|a| \\ &\le C\theta\varepsilon|a|.\end{aligned}$$

Hence,

$$\left|\frac{\beta}{\alpha} - u'\right| \le \left|\frac{\beta}{\alpha} - \frac{\beta'}{\alpha}\right| + \left|\frac{\beta'}{\alpha}\right|\left|1 - \frac{\alpha}{\alpha'}\right| \le \frac{\mu\theta}{\lambda - C\varepsilon} + \frac{(\mu + C\varepsilon)C\theta\varepsilon}{(\lambda - C\varepsilon)^2}.$$

Accordingly, if $h \le h_\theta$, then there exists $\sigma \in (0,1)$ such that

$$D_h(T\xi) \cap T\mathcal{C}_{\theta,h}(\xi,u) \subset \mathcal{C}_{\sigma\theta,h}(T\xi,u'). \tag{2.4.16}$$

A similar, but rougher, computation yields

$$D_h(T\xi) \cap T\mathcal{C}_{\theta,h}(\xi,u) \subset \mathcal{C}_{\theta,h}(T\xi,0). \tag{2.4.17}$$

Finally, let $\xi \in \gamma_*$, then, for each $n \in \mathbb{N}$, $T^{-n}\xi \in \gamma_*$ and $\gamma_* \cap D_{h_n}(T^{-n}\xi) \subset \mathcal{C}_{c\delta,h_n}(T^{-n}\xi,0)$. Thus, for all $h_n \le h_{\sigma^n c\delta}$, (2.4.16) implies[13]

$$\begin{aligned}\gamma_* \cap D_{h_n}(\xi) &\subset T^n\mathcal{C}_{c\delta,h_n}(T^{-n}\xi,0) \cap D_{h_n}(\xi) \\ &= T^{n-1}\left(T\mathcal{C}_{c\delta,h_n}(T^{-n}\xi,0) \cap D_{h_n}(T^{n-1}\xi)\right) \cap D_{h_n}(\xi) \\ &\subset T^{n-1}\mathcal{C}_{\sigma c\delta,h_n}(T^{-n+1}\xi, v_{n,1}) \cap D_{h_n}(\xi) \\ &\subset \mathcal{C}_{\sigma^n c\delta,h_n}(\xi, v_n)\end{aligned} \tag{2.4.18}$$

where $(a, av_{n,k}(\xi)) = D_{T^{-n}\xi}T^k(1,0)$, for some $a \in \mathbb{R}_+$, and $v_n(\xi) = v_{n,n}(\xi)$. The last relevant fact is that the limit

$$v_* = \lim_{n\to\infty} v_n \tag{2.4.19}$$

[13]Remember that the map T expands in the first coordinate, hence $TD_h(\xi) \supset D_h(T\xi)$ provided $\xi \in \mathcal{C}_{c\delta,\delta}(0,0)$ and h is small enough.

exists. The proof of this fact is left as an entertainment for the reader (see Problem 2.12). Using (2.4.18), (2.4.19) and remembering that γ_* admits the parametrization $\gamma_*(t) = (t, u_*(t))$ we can compute the derivative. Indeed, let τ so that $(\tau, u_*(\tau)) = \xi \in \gamma_*$, then for each $\varepsilon > 0$ let m so that $\sigma^m c\delta \leq \frac{\varepsilon}{2}$ and $|v_m - v_*| \leq \frac{\varepsilon}{2}$, then for each $h \leq h_m$ holds

$$\begin{aligned}\left|\frac{u_*(\xi+h) - u_*(\xi) - v_* h}{h}\right| &\leq \left|\frac{u_*(\xi+h) - u_*(\xi) - v_m h}{h}\right| + \frac{\varepsilon}{2} \\ &\leq c\sigma^m \delta + \frac{\varepsilon}{2} \leq \varepsilon.\end{aligned}$$

That is, γ_* is differentiable and

$$\gamma_*'(\tau) = (1, v_*). \tag{2.4.20}$$

□

Problem 2.9 *Prove Theorem 2.4.2 in the hypotheses at the beginning of Section 2.3 when $\alpha < 0 < \beta$.*

There is another point of view that can be adopted in the study of stable and unstable manifolds: to "grow" the manifolds. This is done by starting with a very short curve in $\Gamma_{\delta,c}$, e.g. $\gamma_0(t) = (t, 0)$ for $t \in [\lambda^{-n}\delta, \lambda^n\delta]$, and showing that the sequence $\gamma_n := T^n\gamma_0$ converges to a curve in the strip $[-\delta, \delta]$, independent of γ_0. From a mathematical point of view, in the present case, it corresponds to spell out explicitly the proof of the fixed point theorem. Nevertheless, it is a more suggestive point of view and it is more convenient when the hyperbolicity is non uniform. For example consider the map.[14]

$$T\begin{pmatrix} x \\ y \end{pmatrix} := \begin{pmatrix} 2x - \sin x + y \\ x - \sin x + y \end{pmatrix} \tag{2.4.21}$$

then 0 is a fixed point of the map but

$$D_0 T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

is not hyperbolic, yet, due to the higher order terms, there exist stable and unstable manifolds (see Problems 2.15, 2.16, 2.17).

Problems

2.10. Show that, if p is a fixed point, then $TW^s(p) \subset W^s(p)$ and $TW^u(p) \supset W^u(p)$.

[14] Some times this is called *Lewowicz map*

2.11. Prove that the set $B^*_{\delta,c}$ in section 2.4.1 is closed with respect to the sup norm $\|u\|_\infty = \sup_{t\in[-\delta,\delta]} |u(t)|$.

2.12. Prove that the limit in (2.4.20) is well defined and depend continuously on ξ.

2.13. Prove that, in the setting of Theorem 2.4.2, the unstable manifold is unique.

2.14. Show that Theorem 2.4.2 holds assuming only $T \in \mathcal{C}^1(U, U)$.

2.15. Consider the Lewowicz map (2.4.21), show that, given the set of curves $\Gamma_{\delta,c} := \{\gamma : [-\delta, \delta] \to \mathbb{R}^2 \mid \gamma(t) = (t, u(t));\ \gamma(0) = 0;\ |u'(t)| \in [c^{-1}t, ct]\}$, it is possible to construct the map $\tilde{T} : \Gamma_{\delta,c} \to \Gamma_{\delta(1+c^{-1}\delta),\, c}$ in analogy with (2.4.11).

2.16. In the case of the previous problem show that for each $\gamma_i \in \Gamma_{\delta,c}$ holds $d(\tilde{T}\gamma_1, \tilde{T}\gamma_2) \le (1 - c\delta)d(\gamma_1, \gamma_2)$.

2.17. Show that for the Lewowicz map, zero has a unique unstable manifold.

Hints to solving the Problems

2.1. Use the implicit function theorem on the one parameter vector fields $V(\lambda) = V + \lambda(W - V)$.

2.3. If A is diagonal, the claim is trivial. For a general diagonalizable matrix, let U be such that $U^{-1}AU = \Lambda$, diagonal. Set $B = (UU^*)^{-1}$, then

$$\begin{aligned}\Re(\langle x, BAx\rangle) &= \Re(\langle U^{-1}x, U^{-1}AUU^{-1}x\rangle) = \Re(\langle U^{-1}x, \Lambda U^{-1}x\rangle)\\ &\le -\sigma\langle U^{-1}x, U^{-1}x\rangle = -\sigma\langle x, Bx\rangle.\end{aligned}$$

2.5. By the variation of the constants method it follows that

$$\phi_t(x) = e^{At}x + \int_0^t e^{A(t-s)}R(\phi_s(x))ds.$$

Hence

$$\|\Delta(x)\| \le \sup_{\|x\|\le 2r} \|\phi_1(x) - e^A x\| \le 4Cr^2.$$

2.12 By (2.4.17) and arguing as in (2.4.18) it follows

$$\begin{aligned}T^n\mathcal{C}_{c\delta,h_n}(T^{-n}\xi, 0) \cap D_{h_n}(\xi) &\subset T^{n-1}\mathcal{C}_{c\delta,h_n}(T^{-n+1}\xi, 0) \cap D_{h_n}(\xi)\\ &\subset \mathcal{C}_{\sigma^{n-1}c\delta,h_n}(\xi, v_{n-1}(\xi)).\end{aligned}$$

Since, for a small enough, $T^n(T^{-n}\xi + (a,0)) = \xi + aD_{T^{-n}\xi}T^n(1,0) + o(a)$, it follows that $(a, v_n(\xi)a) \in \mathcal{C}_{\sigma^{n-1}c\delta,h_n}(\xi, v_{n-1}(\xi))$. Hence $|v_n(\xi) - v_{n-1}(\xi)| \leq \sigma^{n-1}c\delta$. From this the Problem easily follows.

2.13. This amounts to showing that the set of points that are attracted to zero are exactly the manifolds constructed in Theorem 2.4.2. Use the local hyperbolicity to show that.

2.16. Grow the manifolds, that is, for each $n > 1$ define $\delta_n := \frac{\rho}{n}$. Show that one can choose ρ such that $\delta_{n-1} \geq \delta_n(1 + c^{-1}\delta_n)$. according to Problem 2.15 it follows that $\tilde{T} : \Gamma_{\delta_n,c} \to \Gamma_{\delta_{n-1},c}$. Moreover,

$$d(\tilde{T}^{n-1}\gamma_1, \tilde{T}^{n-1}\gamma_2) \leq \prod_{i=1}^{n}(1 - c\delta_i)d(\gamma_1, \gamma_2).$$

Finally, show that, setting $\gamma_n(t) = (0,t) \in \Gamma_{\delta_n,c}$, the sequence $\tilde{T}^{n-1}\gamma_n$ is a Cauchy sequence that converges in $\mathcal{C}^0$ to a curve in $\Gamma_{1,c}$ invariant under $\tilde{T}$.

Notes

The content of this section is quite standard and rather sketchy, it is intended only to introduce the reader to some basic ideas and techniques. The treatment of the Hadamard-Perron Theorem follows mostly [HK95].

Chapter 3

Bifurcation Theory (the minimum)

Continuing the analysis of the previous section we would like to place it on a more systematics ground: we worried only about hyperbolic fixed points; are more complex situations relevant? To answer to such a question it is first necessary to understand its meaning, that is:
what does it mean to be irrelevant?

3.1 Generic Vector fields

By relevant we mean situations which are *typical*. We would like to summarise the content of Section 2 as follows:

Theorem 3.1.1 *We understand the typical local behavior of the solutions of the differential equations*

$$\dot{x} = V(x) \tag{3.1.1}$$

where $V \in \mathcal{C}^1_{\text{loc}}(\mathbb{R}^d, \mathbb{R}^d)$.

However, to make sense of Theorem 3.1.1 it is necessary to give a technical meaning to the words *behavior, local* and *typical*.

3.1.1 Local behavior

We say that we *understand* the behavior of a vector field in an open set U if it is *equivalent* to a vector field whose associated ODE can be explicitly solved.

Definition 3.1.2 *We say that two vector fields V, W are equivalent in the open set U, if, for each $t > 0$, there exists a homeomorphism $F : U \to U$ such that, calling ϕ_t^V, ϕ_t^W the flows generated by the vector fields, hold $\phi_t^V \circ F = F \circ \phi_t^W$.*

Definition 3.1.3 *We say that we have a* local *understanding of the ODE (3.1.1) in a region K if, for each point $x \in K$, there exists a neighborhood of x in which the equation (3.1.1) is equivalent to an equation with an affine vector field.*[1]

If we could consider only neighborhoods U in which $V(x) \neq 0$ with, at most, the exception of isolated points where the linear part is hyperbolic, then we understand already the local behavior. In fact, either $V(\bar{x}) \neq 0$ and then the flow box Theorem tells us that the field has the same local behavior than a constant vector field; or, if $V(\bar{x}) = 0$, then Grobmann-Hartman Theorem tells us that the field has the same local behavior than its linear part.

Of course, this is not always the case (think of the case $V \equiv 0$), our claim is that the above situation is *typical*.

3.1.2 Typical

Definition 3.1.4 *Given a topological space Ω, we say that a set $A \subset \Omega$ is* generic *if it is open and dense. A set is* typical *if it is the countable intersection of* generic *sets (this is also called a* residual *set).*

Since $\mathcal{C}^1(\mathbb{R}^d, \mathbb{R}^d)$ is a Banach space, its topology is determined by the norm.

Problem 3.1 *Prove that the finite intersection of a generic set is generic. Prove that, in a metric space, a residual set is dense.*

Problem 3.2 *Give an example of a typical set in $[0, 1]$ with zero Lebesgue measure.*

Next, for each $K \subset \mathbb{R}^d$, let us define[2]

$$A_K := \{V \in \mathcal{C}^1_{\text{loc}}(\mathbb{R}^n, \mathbb{R}^n) \ : \ \forall\, x \in K,\ V(x) = 0 \text{ implies } \partial_x V \text{ hyperbolic } \}$$

Remark 3.1.5 *In the following we will prove that, for K compact, A_K is generic, hence $A_{\mathbb{R}^d}$ is typical. Note that the same holds for*

$$\{V \in \mathcal{C}^1_{\text{loc}}(\mathbb{R}^n, \mathbb{R}^n) \ : \ \forall\, x \in K,\ V(x) = 0 \ \textit{implies}\ \det(\partial_x V) \neq 0\}.$$

[1] Note that, if K is compact, then finitely many such neighborhoods will cover K. On the other hand if, for example, $K = \mathbb{R}^d$, then countably many neighborhoods will do the job.

[2] Since our analysis is local, the following can be trivially adapted to the case $\mathcal{C}^1_{\text{loc}}(U, \mathbb{R}^n)$, for some open set U. We avoid it to simplify notation.

Yet, it is convenient to consider small generic sets (see Problems 3.25, 3.26). This allows to obtain a generic understanding with the least effort.

Problem 3.3 *Prove that, for each compact set $K \subset \mathbb{R}^d$, if $V \in A_K$, then V has only finitely many zeroes in K.*

Problem 3.4 *Prove that, for each compact set $K \subset \mathbb{R}^d$, A_K is open.*

To prove that A_K is generic we need to establish the density, this is not entirely obvious and we need a result of independent interest.

Theorem 3.1.6 (Sard–baby version) *Let $F \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}^d)$, and $A = \{x \in \mathbb{R}^d \ : \ \det(D_x F) = 0\}$, then $F(A)$ has zero Lebesgue measure.*

PROOF. Let $Q_\delta(x) := \{z \in \mathbb{R}^d \ : \ |x_i - z_i| \leq \delta \quad \forall i \in \{1, \dots, d\}\}$, clearly it suffices to prove that for each $\bar{x} \in \mathbb{R}^d$ the Lebesgue measure of $F(A \cap Q_1(\bar{x}))$ is zero. Now, for each $n \in \mathbb{N}$ and $k \in \{-n, \dots, 0, \dots, n\}^d =: S_n$, let $x_k := \frac{k}{n}$ and $\Delta_k := Q_{1/2n}(\bar{x} + x_k)$. Clearly $Q_1(\bar{x}) \subset \cup_{k \in S_n} \Delta_k$. We will consider only the Δ_k such that $\Delta_k \cap A \neq \emptyset$. For each such Δ_k let us chose $\xi_k \in \Delta_k \cap A$.

Next, consider the function $\Psi : Q_1(\bar{x})^2 \to \mathbb{R}$ defined by

$$\Psi(x, y) := \begin{cases} \frac{\|F(x) - F(y) - D_x F(x-y)\|}{\|x - y\|} & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}$$

Since $F = \mathcal{C}^1$ we have $\Psi \in \mathcal{C}^0$, hence for each $\varepsilon > 0$ there exists $n_\varepsilon \in \mathbb{N}$ such that

$$\sup_{\|x-y\| \leq n^{-1}} \Psi(x, y) < \varepsilon$$

for each $n > n_\varepsilon$. Since $\xi_k \in A$, there exists $v_k \in \mathbb{R}^d$, $\|v_k\| = 1$, such that $\langle v_k, D_{\xi_k} F w \rangle = 0$ for all $w \in \mathbb{R}^d$. Hence, setting $C = \|DF\|_\infty$ and for n large enough,

$$F(\Delta_k) \subset \{F(\xi_k) + w + t v_k \in \mathbb{R}^d \ : \ \langle w, v_k \rangle = 0; \ \|w\| \leq C n^{-1}; \ |t| \leq \frac{\varepsilon}{n})\}.$$

Thus, calling λ the Lebesgue measure,

$$\lambda(F(\Delta_k)) \leq 4^{d-1} C^{d-1} n^{-d-1} \frac{\varepsilon}{n} = \lambda(\Delta_k) \cdot 4^{d-1} C^{d-1} \varepsilon.$$

Thus

$$\lambda(F(A \cap Q_1(x))) \leq 4^{d-1} C^{d-1} \sum_{k \in S_n} \lambda(\Delta_k) \varepsilon = 4^d C^{d-1} \varepsilon,$$

as announced. □

Problem 3.5 *Use Sard's Theorem to show that, for each compact set $K \subset \mathbb{R}^d$, A_K is dense in $\mathcal{C}^1(\mathbb{R}^d, \mathbb{R}^d)$. Prove that $A_{\mathbb{R}^d}$ is typical.*

3.2 Generic families of vector fields

Our next aim is to consider a situation in which the system has a control parameter. That is, it is described by the equations of the type

$$\dot{x} = V(x, \lambda) \tag{3.2.2}$$

where $x \in \mathbb{R}^d$ and $\lambda \in [-2, 2]$ is the parameter that, in principle, can be varied. Now by *local* understanding in a region K we mean that for each point $(\bar{x}, \bar{\lambda}) \in K \times [-1, 1] =: K^1$ we can find a neighborhood of the form $U \times (\lambda - \varepsilon, \lambda + \varepsilon)$ in which we are able to understand the behavior of the solutions of (3.2.2).

Let us now try to understand the local picture for typical families of vector fields. In analogy with the previous section, for any $K \subset \mathbb{R}^d$, let us consider

$$\bar{A}_K := \{V \in \mathcal{C}^1 \ : \forall\, (x, \lambda) \in K^1,\ V(x, \lambda) = 0 \text{ implies}$$
$$\partial_x V(x, \lambda) \text{ hyperbolic } \}$$

Problem 3.6 *Prove that if $V \in \bar{A}_K$, then for each $(\bar{x}, \bar{\lambda}) \in K^1$ there exists an open set of the form $U \times (-\varepsilon + \bar{\lambda}, \varepsilon + \bar{\lambda}) =: U \times I$ such that either $V(x, \lambda) \neq 0$ or there exists $X \in \mathcal{C}^1(I, K)$ such that $V(X(\lambda), \lambda) = 0$ for each $\lambda \in K$ and there are no other zeroes in $U \times I$. Then, prove that $\bar{A}_K$ is open.*

Clearly the above situations can be treated exactly as we did in the previous section and are therefore locally understandable. Unfortunately, the above does not exhaust all the possibilities.

Lemma 3.2.1 *For each K with non empty interior $\bar{A}_K$ is not generic.*

Proof. Since $\bar{A}_K$ is open, the problem must be the density. To see this let us consider, for example, the case $d = 1$, a compact set K with interior containing zero and the family

$$V(x, \lambda) = \lambda a + \lambda x + b x^2. \tag{3.2.3}$$

Now let us consider any $W \in \mathcal{C}^1(\mathbb{R} \times [-1, 1], \mathbb{R})$ and look at $\tilde{V}(x, \lambda, \mu) := V(x, \lambda) + \mu W(x, \lambda)$. The claim is that for each μ sufficiently small, then $\tilde{V}(x, \lambda, \mu) \notin \bar{A}_K$. In fact, there exists $(x(\mu), \lambda(\mu)) \in K$ such that both $\tilde{V}(x(\mu), \lambda(\mu), \mu) = 0$ and $\partial_x \tilde{V}(x(\mu), \lambda(\mu), \mu) = 0$. To see this we define the function $F : \mathbb{R}^3 \to \mathbb{R}^2$

$$F(x, \lambda, \mu) := \begin{pmatrix} \lambda a + \lambda x + b x^2 + \mu W(x, \lambda) \\ \lambda + 2bx + \mu \partial_x W(x, \lambda) \end{pmatrix} = \begin{pmatrix} \tilde{V} \\ \partial_x \tilde{V} \end{pmatrix}, \tag{3.2.4}$$

clearly we are looking for $(x(\mu), \lambda(\mu))$ such that $F(x(\mu), \lambda(\mu)) = 0$. Since $F(0, 0, 0) = 0$ we can apply the implicit function theorem provided

$$\begin{pmatrix} \partial_x F & \partial_\lambda F \end{pmatrix} \Big|_{x=0;\, \lambda=0;\, \mu=0} = \begin{pmatrix} 0 & a \\ 2b & 1 \end{pmatrix}$$

is invertible, that is if $ab \neq 0$. We have thus seen that the family has an open neighborhood disjoint from $\bar{A}_K$, hence the latter set cannot be dense. □

Thus, to have a generic situation, we need to consider a larger set.

The above example sugegst to ask that $\partial_\lambda V \neq 0$ if $\det(\partial_x V) = 0$. This is a good idea, but it does not suffice to have a nice theory. As we have seen,[3] and we will see later on, it is natural to have some condition on the second derivative. It is then convenient to consider at least $\mathcal{C}^r$ vector fields, $r \geq 2$. Accordingly, from now on the genericity will be according to the $\mathcal{C}^r$ topology. This would not have changed the previous discussion, see Problem 3.32.

The above can be made precise in many ways. Here is a simple, but not totally satisfactory, possibility. For $K \subset \mathbb{R}^d$ let $K^1 := K \times [-1, 1]$.

$$B_K = \left\{ V \in \mathcal{C}^r_{\text{loc}} \ : \ \forall\, (x,\lambda) \in K^1,\ V(x,\lambda) = 0 \Longrightarrow \operatorname{rank} (\partial_x V \ \partial_\lambda V) = d \right\}$$

Let us understand how the vector fields in B_K look like.

Lemma 3.2.2 *If $V \in B_K$ and $V(\bar{x}, \bar{\lambda}) = 0$, then there exists $\varepsilon > 0$ and a neighborhood $U \ni \bar{x}$ such that the set of zeroes of the vector field $V(x, \lambda)$ in $U \times (\bar{\lambda} - \varepsilon, \bar{\lambda} + \varepsilon)$ consists of a smooth curve.*

Proof. First suppose, without loss of generality, that $(\bar{x}, \bar{\lambda}) = (0, 0)$.

If $\det(\partial_x V(0,0)) \neq 0$, then we can argue as in Problem 3.4. The implicit function theorem yields an $\varepsilon > 0$, a neighborhood U of zero and a function $x \in \mathcal{C}^r([-\varepsilon, \varepsilon], \mathbb{R}^d)$ such that $V(x(\lambda), \lambda) = 0$ are the only zeroes of the vector fields $V(\cdot, \lambda)$, $\lambda \in [-\varepsilon, \varepsilon]$, in U.

On the contrary, if $\det(\partial_x V(0,0)) = 0$ then the approach based on a direct application of the implicit function theorem fails. The problem is that the curve of the fixed points it is not a graph over λ so one need to change variables before applying the implicit functions theorem, let us see how.

The null space of $\partial_x V(0,0)$ must have dimension one, otherwise we would have $\operatorname{rank} (\partial_x V(0,0) \ \ \partial_\lambda V(0,0)) < d$, let $v \in \mathbb{R}^d$, $\|v\| = 1$, be the unique vector such that $\partial_x V(0,0) v = 0$. Consider the change of variables $(\lambda, x) = F(\xi, \tau)$ defined by

$$\begin{aligned} x &= \xi - \tau v \\ \lambda &= \langle \xi, v \rangle. \end{aligned} \tag{3.2.5}$$

It is easy to check that F^{-1} is defined by

$$\begin{aligned} \tau &= \lambda - \langle x, v \rangle \\ \xi &= \lambda v + x - \langle x, v \rangle v. \end{aligned}$$

[3] In applying the implicit function theorem to (3.2.4).

Then define the field $\tilde{V} := V \circ F$. Since $F(0,0) = 0$, $\tilde{V}(0,0) = 0$. To apply the implicit function theorem in the new variables, we need $\partial_\xi \tilde{V}$ to be invertible, but $\partial_\xi \tilde{V}(\xi,\tau) = \partial_x V(x,\lambda) + \partial_\lambda V(x,\lambda) \otimes v$.[4] It follows that $\partial_\xi \tilde{V}(0,0)$ must be invertible, otherwise there would exists $z \in \mathbb{R}^d$ such that, for all $\eta \in \mathbb{R}^d$,

$$0 = \langle z, \partial_\xi \tilde{V}(0,0)\eta\rangle = \langle z, \partial_x V(0,0)\eta\rangle + \langle z, \partial_\lambda V(0,0)\rangle\langle v, \eta\rangle.$$

Choosing $\eta = v$ follows $\langle z, \partial_\lambda V(0,0)\rangle = 0$ and hence $\partial_x V(0,0)^T z = 0$. This would mean that all the columns of the rectangular matrix $(\partial_x V(0,0) \;\; \partial_\lambda V(0,0))$ are orthogonal to z, contradicting the definition of B_K.

So we can apply the implicit function theorem and obtain (for ξ, τ in a neighborhood of zero) a $\mathcal{C}^1$ function $\xi(\tau)$ such that $\tilde{V}(\xi(\tau),\tau) = 0$. Then, setting $(x(\tau),\lambda(\tau)) := F(\xi(\tau),\tau)$ we have a $\mathcal{C}^1$ curve in $\mathbb{R}^{d+1}$ such that $V(x(\tau),\lambda(\tau)) = 0$ and no other zero is present in the neighborhood of zero. To study such a curve, we need to compute some derivative. Differentiating $\tilde{V}(\xi(\tau),\tau) = 0$ with respect to τ yields

$$\begin{aligned}\partial_x V(\xi(\tau) - \tau v, &\langle \xi(\tau), v\rangle)(\xi'(\tau) - v) \\ &+ \partial_\lambda V(\xi(\tau) - \tau v, \langle \xi(\tau), v\rangle)\langle \xi'(\tau), v\rangle = 0.\end{aligned} \tag{3.2.6}$$

For $\tau = 0$ yields

$$(\partial_x V(0,0) + \partial_\lambda V(0,0\rangle \otimes v)\, \xi'(0) = 0$$

which implies $\xi'(0) = 0$. Moreover,

$$\frac{d\lambda}{d\tau} = \langle \xi'(\tau), v\rangle, \tag{3.2.7}$$

hence $\frac{d\lambda}{d\tau}(0) = 0$. While

$$x'(\tau) = \frac{dx(\tau)}{d\tau} = \xi'(\tau) - v. \tag{3.2.8}$$

hence $x'(0) = v$. □

Problem 3.7 *Show that $B_{\mathbb{R}^d}$ is typical.*

We thus have a typical set, yet it contains behaviors that we have never analyzed: equilibrium points with a derivative having a one-dimensional kernel and equilibrium points with no kernel but a non-hyperbolic derivative. It would be convenient if we could limit the appearance of such situations to a bare minimum. To do this systematically would require the development of a formalism beyond the present goals. Yet, for the case of one-parameter families, it is still possible to do it naively, provided one is willing to put up with some boring computations.

[4] Given two vectors $v, w \in \mathbb{R}^d$, by $v \otimes w$ is the matrix with elements $(v \otimes w)_{ij} = v_i w_j$.

Definition 3.2.3 *Given $V \in \mathcal{C}^r$ we call a point $(\bar{x}, \bar{\lambda}) \in \mathbb{R}^{d+1}$ such that $V(\bar{x}, \bar{\lambda}) = 0$ and $V(\cdot, \bar{\lambda}) \notin A_{\overline{U}}$, for a neighborhood U of $\bar{x}$, a* bifurcation point.[5] *Let $(\bar{x}, \bar{\lambda})$ be a bifurcation point, we call such point* non degenerate, *if* $\operatorname{rank}\big(DV(\bar{x}, \bar{\lambda})\big) = d - 1$, $\langle w, D^2V(v, v)\rangle \neq 0$ *where v, w are such that $DVv = DV^T w = 0$. We call the the bifurcation point* regular *if it is non degenerate or if* $\det\big(DV(\bar{x}, \bar{\lambda})\big) \neq 0$ *but there are two eigenvalues with zero real part and* $\operatorname{Tr}\big(\Pi_0 \big[\frac{d}{d\lambda} A(\lambda)\big] \Pi_0\big) \neq 0$ *where Π_0 is the eigenprojector on the eigenspace associated to the above eigenvalues and $A(\lambda) = \partial_x V(x(\lambda), \lambda)$, where $x(\lambda)$ is determined by $V(x(\lambda), \lambda) = 0$.*

The idea is then to define the new sets

$$\tilde{B}_K = \{V \in B_K \ : \text{ all the bifurcation points are regular}\}.$$

Let us show that the elements of $\tilde{B}_K$ enjoy a nice characterization.

Lemma 3.2.4 *In $\tilde{B}_K$ the bifurcation points are isolated.*

Proof. Let us start analyzing the case of non-degenerate bifurcation points. Suppose, without loss of generality, that the bifurcation point is at $(0, 0)$. Note that, by continuity, the condition on the second derivative holds in a neighborhood of zero. By Lemma 3.2.2 we know that the zeroes of V lie on a curve $(x(\tau), \lambda(\tau))$, with the derivative with respect to τ given by (3.2.8), (3.2.7). Also, there exists unique normalized vectors w, v such that $\partial_x V(0,0)v = [\partial_x V(0,0)]^T w = 0$. By (3.2.7) it follows that if $\lambda'(\tau) = 0$, then $\langle \xi'(\tau), v\rangle = 0$, and then (3.2.6) implies $\partial_x V(\xi'(\tau) - v) = 0$. That is, if $\lambda'(\tau) = 0$, then $(x(\tau), \lambda(\tau))$ is a bifurcation point. Hence, to show that the difurcation point is isolated, it suffices to prove that $\tau = 0$ is the only point for which $\lambda' = 0$. To this end, we can compute

$$\lambda''(0) = \langle \xi''(0), v\rangle.$$

Differentiating (3.2.6) at zero yields

$$\partial_x^2 V(0,0)(v, v) + (\partial_x V(0,0) + \partial_\lambda V(0,0) \otimes v)\xi''(0) = 0.$$

If we multiply the above by w we have

$$\langle w, \partial_x^2 V(0,0)(v, v)\rangle + \langle w, \partial_\lambda V(0,0)\rangle \lambda''(0) = 0$$

since the zero is a non degenrate bifurcagtion point $\langle w, \partial_\lambda V(0,0)\rangle \neq 0$, and $\langle w, \partial_x^2 V(0,0)(v, v)\rangle \neq 0$. It follows $\lambda''(0) \neq 0$, hence zero is an isolated zero of λ'.

[5] That is the vector field $V(\cdot, \bar{\lambda})$ is not generic.

We are left with the case $\det(\partial_x V(\bar{x},\bar{\lambda})) \neq 0$ but with an eigenvalue which has a zero real part. This means that we have two purely imaginary eigenvalues. Let Π_0 be the eigenprojection associated with such two eigenvalues. By perturbation theory (see Appendix C) it follows that there exists a rank two projector family $\Pi(x,\lambda)$ such that $\Pi\partial_x V = \partial_x V \Pi$ and $\Pi(\bar{x},\bar{\lambda}) = \Pi_0$.

Problem 3.8 *Show that a two-by-two real matrix has purely imaginary eigenvalues iff its trace is zero and the determinant is positive.*

Then we have $\mathrm{Tr}(\Pi_0 \partial_x V(\bar{x},\bar{\lambda})\Pi_0) = 0$. Now, let $x(\lambda)$ be the curve of the zeroes of V, then

$$\frac{d}{d\lambda}\,\mathrm{Tr}(\Pi\partial_x V \Pi)\Big|_{\lambda=0} = \mathrm{Tr}\left(\Pi_0 \left[\frac{d}{d\lambda}\partial_x V\right]\Pi_0\right)$$

since $\Pi^2 = \Pi$ implies $\Pi\left(\frac{d}{d\lambda}\Pi\right)\Pi = 0$.[6] This concludes the argument. □

Problem 3.9 *Show that $\tilde{B}_K$ is still generic.*

Thus, to achieve a typical local understanding of the behavior of one parameter families of vector fields we have to worry only about families with, at most, one regular bifurcation point. Let us suppose, without loss of generality, that the regular bifurcation point is at $(0,0)$, then by Taylor expansion

$$V(x,\lambda) = a(\lambda) + A(\lambda)x + \frac{1}{2}\langle x, B(\lambda), x\rangle + R(x,\lambda), \tag{3.2.9}$$

where B is a vector of $d \times d$ symmetric matrices and $a(0) = 0$, $R(0,\lambda) = \partial_x R(0,\lambda) = \partial_x^2 R(0,\lambda) = 0$.

Due to the previous discussion we need to consider only the following cases

a) $A^T(0)$ has one, and only one, zero eigenvalue w and $\langle w, a'(0)\rangle \neq 0$;

b) $A(0)$ has two purely imaginary conjugated eigenvalues.

3.3 One dimension

In the one dimensional case (b) cannot take place. Then in (3.2.9) we have $a = a'(0) \neq 0$, $A(0) = 0$, $c = B(0) \neq 0$. Then $V(x,\lambda) = 0$ has no solutions if $ac > 0$, while for $ac < 0$ there are the two solutions $x = \pm\sqrt{-\frac{\lambda b}{B}} + \mathcal{O}(\lambda)$. We have therefore the generic picture: either two points collide and kill each other or there is a creation of two zeroes of the vector field.

[6] Indeed, $\mathrm{Tr}(\Pi'\partial_x V\Pi) = \mathrm{Tr}(\Pi\Pi'\partial_x V\Pi) = \mathrm{Tr}((\Pi\Pi'\Pi\partial_x V) = 0$. Analogously, $\mathrm{Tr}(\Pi\partial_x V\Pi') = 0$.

Problem 3.10 *Study the solutions of*

$$\dot{x} = \frac{B}{2}x^2 + g(x)$$

near zero when $g(0) = g'(0) = g''(0) = 0$.

Problem 3.11 *Prove that the two equilibrium points of the vector field* (3.2.9) *are one attractive and the other repulsive.*

The above scenario is called a *saddle-node* bifurcation.

A natural question is if there exists a simpler standard form of the above bifurcation. Indeed, we can try to kill some of the terms in 3.2.9 by a change of variable.

Problem 3.12 *Show that with a change of variables of the type* $x = \alpha\lambda + \rho z$, *one can change the vector field* (3.2.9) *to the from* $\tilde{V}(z,\lambda) = \lambda + bz^2 + \mathcal{O}(\lambda^2) + o(z^2)$.

The above is the *normal form* of the saddle node bifurcation. This type of reduction can be made for each bifurcation and gives rise to the large field of normal form theory which, unfortunately, goes beyond the scopes of the present notes.

3.4 Two dimensions

3.4.1 A zero eigenvalue

In this case the vector field must have the form (possibly after a linear change of variable to put $\partial V_x(0,0)$ in diagonal form)

$$V(x,\lambda) = \begin{pmatrix} 0 & 0 \\ 0 & \nu \end{pmatrix} x + b\lambda + \frac{1}{2}\begin{pmatrix} \langle x, B_1 x\rangle \\ \langle x, B_2 x\rangle \end{pmatrix} + \lambda C x + \mathcal{O}(\lambda^2) + o(\|x\|^2), \quad (3.4.10)$$

with $b_1, B_1 \neq 0$. It is straightforward to prove that the scenario is identical to the one-dimensional case. We leave the details to the reader.

3.4.2 Two imaginary eigenvalues: Hopf bifurcation

In this case, the vector field must have the form (possibly after a translation and a linear change of variable to put $\partial V_x(0,0)$ in chosen form, see Problem 3.33)

$$V(x,\lambda) = Ax + R(x,\lambda), \quad (3.4.11)$$

with $A = \begin{pmatrix} 0 & -\omega_0 \\ \omega_0 & 0 \end{pmatrix}$ for some $\omega_0 > 0$, $R(0,0) = \partial_x R(0,0) = 0$ and $\mathrm{Tr}(\partial_{xx} R(0,0) A^{-1} \partial_\lambda R(0,0) - \partial_{x\lambda} R(0,0)) \neq 0$.

In the above situation, no new fixed point can appear, yet one expects something to happen.

Theorem 3.4.1 (Hopf bifurcation) *As λ goes trhough zero, it appears a* periodic orbit *circling the fixed point.*

Proof. To minimize the computations, we start by performing some changes of variables that reduce the ODE to a simpler one.

Problem 3.13 *Show that, with a change of coordinates of the type $x = \xi + \alpha(\lambda)$, the remainder R in (3.4.11) can be made to satisfy $R(0,\lambda) = 0$, for each λ small enough, $\partial_\xi R(0,0) = 0$ and $\mathrm{Tr}(\partial_{\xi\lambda} R(0,0)) \neq 0$.*

Problem 3.14 *Show that with a further change of variables $x = D(\mu) z$, $\lambda = \mu\rho(\mu)$ one can put (3.4.11) in the form*

$$\dot{z} = [\omega(\mu) J + \mu \mathbb{1}]\, z + R(z,\mu), \text{ where } J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \tag{3.4.12}$$

with $\omega(0) = \omega_0$ and $R(0,\mu) = \partial_z R(0,\mu) = 0$.

Problem 3.15 *Find the solutions of (3.4.12) in the case $R \equiv 0$.*

Given that the solutions of the linear part of (3.4.12) rotate around zero almost in circles, it may occur the idea to treat the problem in polar coordinates. In fact this point of view is quite advantageous and we will adopt it. The reader who wants to appreciate the advantages of this choice is invited to try to do the following analysis in Euclidean coordinates.

The polar coordinates can be written as $x = \rho v(\theta)$, where $\rho \in \mathbb{R}_+$, $\theta \in \mathbb{R}$ and $v(\theta) := (\cos\theta, \sin\theta)$.

Remark 3.4.2 *Note that such a change of coordinates is singular for $\rho = 0$. In addition, it is not globally one-one. Yet, to consider θ in the universal cover of S^1 rather than in S^1 will be very useful in the following.*

If we substitute such coordinates in (3.4.12), we obtain

$$\dot{\rho} v(\theta) + \rho n(\theta) \dot{\theta} = \mu \rho v(\theta) + \omega(\mu) \rho n(\theta) + R(\rho v(\theta), \mu),$$

where $n(\theta) := (-\sin\theta, \cos\theta)$. That is

$$\begin{aligned} \dot{\rho} &= \mu\rho + \langle v(\theta), R(\rho v(\theta),\mu)\rangle =: \mu\rho + a(\theta,\rho,\mu) \\ \dot{\theta} &= \omega(\mu) + \rho^{-1} \langle n(\theta), R(\rho v(\theta),\mu)\rangle =: \omega(\mu) + b(\theta,\rho,\mu), \end{aligned} \tag{3.4.13}$$

where $a(\theta,0,\mu) = \partial_\rho a(\theta,0,\mu) = b(0,\mu) = 0$. In addition, note for later use that, $\partial^2_\rho a(\theta,0,0)$ and $\partial_\rho b(\theta,0,0)$ are homogeneous trigonometric polynomials of degree three, while $\partial^3_\rho a(\theta,0,0)$ and $\partial^2_\rho b(\theta,0,0)$ are of degree four. By Problem 3.35 it follows that we can write $a(\theta,\rho,\mu) = a_0(\theta,\mu)\rho^2 + a_1(\theta,\rho,\mu)\rho^3$ and $b(\theta,\rho,\mu) = b_0(\theta,\mu)\rho + b_1(\theta,\rho,\mu)\rho^2$. Finally, the reader can easily verify that $a \in \mathcal{C}^r$, while $b \in \mathcal{C}^{r-1}$.

Note that the equation (3.4.13) is well defined also for $\rho = 0$ but in such a case, instead of a fixed point, it has the periodic orbit $(\rho(t),\theta(t)) = (0,\omega_0 t)$. Thus in polar coordinates for $\rho = 0$ we have a rotation, this captures the behavior of the system much better than the fixed point in Euclidean coordinates.

Problem 3.16 *Solve* (3.4.13) *in the case* $b = 0$, $a = \rho^2$. *Do it for* $b = 0$, $a = \mu\rho^2 + \rho^3$.

Since for small ρ we have $\dot\theta > 0$, it is convenient to use θ rather than t to parameterize the motion (here is now evident the advantage of using the universal cover of S^1). Calling again ρ the distance from the origin as a function of θ we have

$$\frac{d\rho}{d\theta} = \frac{\mu\rho + a(\theta,\rho)}{\omega + b(\theta,\rho)} =: \frac{\mu}{\omega(\mu)}\rho + \beta(\theta,\mu)\rho^2 + \gamma(\theta,\rho,\mu)\rho^3, \tag{3.4.14}$$

where

$$\begin{aligned}\beta(\theta,\mu) &= \omega(\mu)^{-1}a_0(\theta,\mu) - \mu\omega(\mu)^{-2}b_0(\theta,\mu)\\ \gamma(\theta,0,\mu) &= \mu\omega(\mu)^{-3}b_0^2 + a_0b_0\omega(\mu)^{-2} - \mu b_1\omega(\mu)^{-2} + a_1\omega(\mu)^{-1}.\end{aligned}$$

Note, that $\beta(\theta,0)$ is a trigonometric homogeneous polynomial of third degree while $\gamma(\theta,0,0)$ is the sum of two monomial, one of degree four and one of degree six.

It is now convenient to perform a last change of variables: $\rho = \nu r$, $\mu = \pm\nu^2$, $\nu \geq 0$.[7] Under such changes of variables (3.4.14) becomes

$$\frac{dr}{d\theta} = \pm\frac{\nu^2}{\omega(\pm\nu^2)}r + \beta(\theta,\pm\nu^2)\nu r^2 + \nu^2\gamma(\theta,\nu r,\pm\nu^2)r^3, \tag{3.4.15}$$

Remark 3.4.3 *The reader my wonder what is going on: if the coefficients would not depend on* θ, *then the periodic orbit would be circular and would correspond to a zero in the above vector field. Such a zero would occur for* $r = \mathcal{O}(\nu^{-1}\beta\gamma^{-1})$, *thus it seems that I have just done the wrong scaling. The point is that the above naive analysis is correct only if we consider the average (with respect to* θ*) of the coefficients, but the average of* β *is zero! This is a very simple instance of a general theory called* averaging.

[7] In fact, we have two different changes of variable according to the sign of μ.

Remark 3.4.4 *In the following we will choose the case in which $\mu > 0$, hence the change of variable with the plus is selected. The computations for $\mu < 0$ are completely analogous and are left to the reader.*

Let us call $r(\theta, \xi, \nu)$ the solution of (3.4.15) with initial condition ξ and parameter ν.

Problem 3.17 *Prove that, for each $\theta \in [0, 2\pi]$ the function $r(\theta, \cdot, \cdot)$ are $\mathcal{C}^{r-1}$.*

We are finally ready to prove the existence of a periodic orbit. Clearly, an orbit is periodic if and only if $r(0, \xi, \nu) = r(2\pi, \xi, \nu)$. In other words, if we look at the motion only when it crosses the $\{\theta = 0 \mod 2\pi\}$ line, then we see the orbit always at the same point. We have thus another instance of a *Poincaré section*.

In concrete, if we consider the map $S : \mathbb{R}^2_+ \to \mathbb{R}_+$ defined by $S(\xi, \nu) := r(2\pi, \xi, \nu)$, then the periodic orbits of the flow correspond to the fixed points of the maps $S(\cdot, \nu)$.[8]

Our last task is thus to study such a maps. The right idea is to develop them in power series of ν. Note that $r(\theta, \xi, 0)$ satisfies the Cauchy problem

$$\begin{aligned} &\frac{dr}{d\theta} = 0 \\ &r(0, \xi, 0) = \xi. \end{aligned}$$

Thus $S(\xi, 0) = \xi$. To compute the derivative we must compute $\eta := \partial_\nu r(\theta, \xi, \nu)$. Such a derivative satisfies the equation obtained by differentiating (3.4.15) (see Theorem 1.1.13)

$$\begin{aligned} \frac{d\eta}{d\theta} &= \frac{2\nu}{\omega} r - \frac{2\nu^3\omega'}{\omega^2} r + \frac{\nu^2}{\omega}\eta + \beta r^2 + 2\nu r\eta\beta + 2\nu^2 r^2 \partial_{\nu^2}\beta \\ &\quad + 2\nu\gamma r^3 + 3\nu^2 r^2 \eta\gamma + 2\nu^3 r^3 \partial_{\nu^2}\gamma + \nu^3 r^3 (r + \nu\eta)\partial_{\nu r}\gamma \\ \eta(0, \xi, \nu) &= 0. \end{aligned} \tag{3.4.16}$$

Setting $\nu = 0$ in the above equation yields $\eta(\theta, \xi, 0) = \xi^2 \int_0^\theta \beta(\varphi, 0) d\varphi$. Accordingly, $\partial_\nu S(\xi, 0) = 0$ (see Problem 3.36).

To conclude we need to compute the second derivative at $\nu = 0$. Setting $\zeta(\theta, \xi) = \partial_\nu \eta(\theta, \xi, 0)$ and differentiating (3.4.16), yields

$$\begin{aligned} &\frac{d\zeta}{d\theta} = \frac{2}{\omega_0}\xi + 4\beta\xi\eta(\theta, \xi, 0) + 2\gamma(\theta, 0, 0)\xi^3 \\ &\zeta(0, \xi, 0) = 0. \end{aligned}$$

[8] I mean the non trivial ones, since zero is always a trivial fixed point by construction.

which yields

$$\zeta(\theta,\xi)=\frac{2\theta}{\omega_0}\xi+4\xi\int_0^\theta\beta(\varphi,0)\eta(\varphi,\xi,0)d\varphi+2\xi^3\int_0^\theta\gamma(\varphi,0,0)d\varphi.$$

Next, note that $\frac{d\eta(\varphi,\xi,0)}{d\varphi}=\xi^2\beta(\varphi,0)$, hence

$$\int_0^\theta\beta(\varphi,0)\eta(\varphi,\xi,0)d\varphi=\frac{\eta(\theta,\xi,0)^2}{2\xi^2}=\frac{\xi^2}{2}\left(\int_0^\theta\beta(\varphi,0)d\varphi\right)^2.$$

Thus, setting $\bar\gamma=\int_0^{2\pi}\gamma(\varphi,0,0)d\varphi$, we have[9]

$$S(\xi,\nu)=(1+\frac{2\pi}{\omega_0}\nu^2)\xi+\xi^3\bar\gamma\nu^2+\nu^3\xi\Gamma(\xi,\nu)\tag{3.4.17}$$

To study the solution of $S(\xi,\nu)=\xi$ for $\nu\neq0$ and $\xi\neq0$ it is convenient to introduce the function $F(\xi,\nu)=\nu^{-2}\xi^{-1}(S(\xi,\nu)-\xi)=\frac{2\pi}{\omega_0}+\xi^2\bar\gamma+\nu\Gamma(\xi,\nu)$.

If $\bar\gamma>0$, then $F(\xi,0)$ has no solutions different from zero and the same must hold for small ν.

If $\bar\gamma<0$, then $\xi_0=\sqrt{-\frac{2\pi}{\omega_0\bar\gamma}}$ is the only positive solution of $F(\xi,0)=0$. We can then apply the implicit function theorem since $F(\xi_0,0)=0$ and

$$\partial_\xi F(\xi_0,0)=\frac{2\pi}{\omega_0}+3\xi^2\bar\gamma=-\frac{4\pi}{\omega_0}\neq0.$$

As a conclusion we have a unique $\xi(\nu)=\xi_0+\mathcal{O}(\nu)$ such that $S(\xi(\nu),\nu)=\xi(\nu)$ for $\nu\neq0$.

Problem 3.18 *Compute, in terms of the Tailor coefficients of V, what it means $\bar\gamma=0$ and shows that it is not possible for $V\in\tilde B_{\mathbb{R}^2}$.*

□

3.5 The Hamiltonian case

It is important to note that non-generic situations may appear due to symmetries or other types of constraints. To give an example of such a situation, let us consider a Hamiltonian vector field, that is a vector field of the type $V(x,p)=(\partial_pH,-\partial_xH)$ for some function $H(x,p)$. In this case

$$DV=\begin{pmatrix}\partial_{xp}H & \partial_{pp}H\\ -\partial_{xx}H & -\partial_{xp}H\end{pmatrix}.$$

[9] Since $S(0,\nu)=0$, the coefficient of ν^3 must have the form $\xi\Gamma$.

Note that the trace of DV is always zero. Hence, if $V(x,p)=0$ and $\det DV \neq 0$, either the fixed point is hyperbolic or has two purely imaginary eigenvalues. This means that having two purely imaginary eigenvalues is generic for Hamiltonian vector fields, contrary to the general ones. Analogously, the situation for a one-parameter family, already when $(x,p)\in\mathbb{R}^2$, is more complex. For example, at a generic bifurcation point, the vector field will have two, not one, zero eigenvalues.

In fact, for mechanical systems, the Hamiltonian often has the form $H(x,p)=\frac{1}{2}p^2+U(x)$, for some function U. Hence, $V(x,p)=(p,-\partial_x U)$, which means that the zeroes of the vector field are the critical points of U. Let us discuss Hamiltonian systems in which the Hamiltonian is of the above type.

We start with the so called *one degree of freedom*, i.e. $x,p\in\mathbb{R}$.

Problem 3.19 *Show that if U has a minimum, then the fixed point is a* center*, while if U has a maximum, then the corresponding fixed point is hyperbolic.*

We thus have a new phenomenon: a center that is stable under small perturbations!

Let us consider the case in which a one-parameter family of potentials $U(x,\lambda)$ has a degenerate minimum at zero, i.e. $U(0,\lambda)=0, \partial_x U(0,\lambda)=0, \partial_x^2 U(0,0)=0$. This means that $U(x)=\lambda x^2+a(\lambda,x)x^3$ and

$$V(x,\lambda)=(p, 2\lambda x+a_1(x,\lambda)x^2)$$

Problem 3.20 *Show that in the above family, we have the collision of two fixed points (a center and a saddle) that collide and exchange type.*

This means that the zeroes of the vector fields are $p=0$, $x(\lambda)=0$ and $x(\lambda)\sim -\frac{2\lambda}{a_1(0,0)}$. We then have a new phenomenon: two fixed points that cross and exchange type.[10]

Even more singular situations may happen if more constraints are present. Consider, for example the above situation when, for some reason, the Hamiltonian is constrained to being symmetrical: $H(x,p)=H(-x,p)$. Then it would have the form $U(x)=\lambda x^2+a(\lambda,x)x^4$.

Problem 3.21 *Show that in the above case one has one fixed point that evolves into three fixed points. Moreover, show that if only one fixed point is present, the fixed point is unstable, then of the three fixed points, two are unstable and one is stable. This is called a* peach fork *bifurcation.*

[10] Hence, the set of fixed points no longer forms a smooth curve in the x,λ space.

Next let us consider the case of two degree of freedom, i.e. $x, p \in \mathbb{R}^2$. Limited to the case of a minimum. In such a situation, at the point of minimum, we have

$$\partial_x V(x,p) = \begin{pmatrix} 0 & \mathbb{1} \\ -\partial_x^2 U & 0 \end{pmatrix}. \tag{3.5.18}$$

where $\partial_x^2 U$ is a positive symmetric matrix, let ω_1^2, ω_2^2 be its eigenvalues.

Problem 3.22 *Show that the eigenvalues of $\partial_x V$, at the fixed point, are $\pm\omega_i$.*

Another surprise: a stable situation with four imaginary eigenvalue (an higher dimensional center).

Problem 3.23 *Consider the linear equation (obtained by the matrix (3.5.18) after a change of variables)*

$$\begin{aligned} \dot{x} &= p \\ \dot{p} &= \begin{pmatrix} -\omega_1^2 & 0 \\ 0 & -\omega_2^2 \end{pmatrix} x \end{aligned}$$

Show that $p_i^2 + x_i^2$ are invariant of the motion, i.e. the motion takes place on two-dimensional tori.

Remark 3.5.1 *Contrary to the case of one degree of freedom, in which the conservation of the Hamiltonian implies that the center is stable for the full motion, in higher dimensions it is not clear if the center is stable or not for the full dynamics. Indeed, this is a rather complex matter at present, and it is not yet completely clarified. Part of the answer is the subject of the so called KAM theory.*[11] *We will discuss some aspects of KAM theory in the following.*

Problems

3.24. Compute $\tilde{V} = V \circ F$ where V is given by (3.2.3) and F by (3.2.5), i.e. $F(\xi, \tau) = (\xi - \tau, \xi)$. Show, by direct computation, that $\tilde{V}(\xi, \tau) = 0$ has solution $\xi(\tau) = -\frac{b}{a}\tau^2 + \mathcal{O}(\tau^3)$.

3.25. Prove that the set $\{A \in GL(n, \mathbb{R}) \ : \ \det(A) \neq 0\}$ is generic with respect to the topology induced by the norm.

3.26. Prove that the set $\{A \in GL(n, \mathbb{R}) \ : \ A$ is hyperbolic$\}$ is generic.

3.27. Prove that $\{A \in \mathcal{C}^0([-1,1], GL(n,\mathbb{R})) \ : \ \operatorname{rank}(A(\lambda)) \geq n-1 \ \forall \lambda \in [-1,1]\}$ is generic.

[11] KAM stands for Kolmogorov, Arnold, and Moser.

3.28. Prove that the set $\{A \in GL(n,\mathbb{R}) \ : \ A$ is hyperbolic and has only simple eigenvalues$\}$ is generic (i.e. Jordan blocks are atypical).

3.29. Show that if $A \in GL(2,\mathbb{R})$ and its eigenvalues have zero real part, then $\mathrm{Tr}(A) = 0$.

3.30. If $A \in \mathcal{C}^1([-1,1], GL(n,\mathbb{R}))$ and $\Pi \in \mathcal{C}^1([-1,1], GL(n,\mathbb{R}))$ is an eigenprojector, show that $\frac{d}{d\lambda}\,\mathrm{Tr}(\Pi A) = 2\,\mathrm{Tr}(\Pi \frac{d}{d\lambda} A)$.

3.31. Show that the set $\{A \in \mathcal{C}^1([-1,1], GL(n,\mathbb{R}))$: at most two eigenvalues have zero real part$\}$ is generic.

3.32. Prove that the set

$$A_K := \{V \in \mathcal{C}^r(\mathbb{R}^n, \mathbb{R}^n) \ : V(x) = 0 \text{ implies } \partial_x V \text{ hyperbolic } \forall\, x \in K\}$$

is generic in the $\mathcal{C}^r$ topology.

3.33. Show that any matrix $A \in GL(2,\mathbb{R})$ with two eigenvalue with zero trace and positive determinant is conjugate to a matrix of the form

$$\begin{pmatrix} 0 & -\omega \\ \omega & 0 \end{pmatrix}$$

for some $\omega > 0$.

3.34. Let $f \in \mathcal{C}^r(\mathbb{R}^{d+1})$ and write the elements of $\mathbb{R}^{d+1}$ as $(\xi_1, \ldots, \xi_d, t)$. If $f(\xi, 0) = \partial_t^k f(\xi, 0) = 0$ for all $k \le s < r$, then there exists $g \in \mathcal{C}^{r-s}$ such that $f(\xi, t) = t^s g(\xi, t)$.

3.35. Let $f \in \mathcal{C}^r(\mathbb{R}^{d+1})$ and write the elements of $\mathbb{R}^{d+1}$ as $(\xi_1, \ldots, \xi_d, t)$. Then, for all $s < r$, there exists $g \in \mathcal{C}^{r-s}$ such that $f(\xi, t) = \sum_{k=0}^{s-1} f^k(\xi, 0) t^k + t^s g(\xi, t)$.[12]

3.36. Show that if $p(\theta)$ is a product of an odd number of functions equal either to $\sin\theta$ or $\cos\theta$, then $\int_0^{2\pi} p(\theta) = 0$.

Hints to solving the Problems

3.1 The finite case is easy. The countable case follows from the Baire category theorem.

[12] Essentially this is Taylor formula where one controls the smoothness of the remainder. This issue is relevant in the applications, but often not investigated in standard textbooks.

3.2 For each $\varepsilon \in \mathbb{R}$ let

$$U_\varepsilon = \cup_{q\in\mathbb{N}} \cup_{p\in 0,\dots,q} \left\{ x \in [0,1] \ : \ |x - \frac{p}{q}| < \frac{\varepsilon}{q^3} \right\}.$$

Then the U_ε are generic in $[0,1]$. Yet, their Lebesgue measure is bounded by

$$\sum_{q=1}^{\infty} \sum_{p=0}^{q} \frac{2\varepsilon}{q^3} \leq \sum_{q=1}^{\infty} \frac{4\varepsilon}{q^2} \leq \gamma\varepsilon$$

for some fixed constant γ. Accordingly, $\cap_n U_{1/n}$ is a typical set of zero measure.

3.3 Let $\bar{x} \in K$ such that $V(\bar{x}) = 0$. Then, by assumption $D_{\bar{x}}V$ is invertible, so $V(\bar{x}+\xi) = 0$ can be written as

$$D_{\bar{x}}V^{-1}(D_{\bar{x}}V\xi - V(\bar{x}+\xi)) = \xi.$$

Since $D_{\bar{x}}V\xi - V(\bar{x}+\xi) = o(\|\xi\|)$, it follows that the above equation has the unique solution $\xi = 0$ in a sufficiently small neighborhood of zero. Hence, there exists a neighborhood of $\bar{x}$ in which there are no other zeroes. Next, for each point in K consider a neighborhood as follows: if the V is different from zero at such a point, then consider a neighborhood for which the vector field is different from zero. If the vector field is zero at the point, then consider the above neighborhood in which the point is the only zero. In such a way, we have a covering of K, we can then extract a finite subcover, hence proving the statement.

3.4 Let $V \in A_K$ and $\{x_i\}_{i=1}^M$ be the zeroes of V. Then for each vector field $W \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}^d)$, $\|W\| \leq 1$, consider the family $V(x,\mu) := V(x) + \mu W(x)$. For each $i \in \{1,\dots,M\}$, use the implicit function theorem to show that there exists $\varepsilon_i, \delta_i > 0$ and $X_i \in \mathcal{C}^1([-\varepsilon_i,\varepsilon_i],\mathbb{R}^n) \to \mathbb{R}^d$, $X_i(x_i) = 0$, such that $V(X_i(\mu),\mu) = 0$ and $V(x,\mu) = 0$, $\|x - x_i\| \leq \delta_i$, $|\mu| \leq \varepsilon_i$ implies that $x = X_i(\mu)$. Verify (using perturbation theory) that, for μ small enough $\partial_x V(X(\mu),\mu)$ is hyperbolic. Next, set $\delta = \min \delta_i$ and $\rho := \inf_{|x-x_i|\geq\delta} \|V(x)\|$. Clearly $V(x,\mu) \neq 0$ if $|x - x_i| \geq \delta$ and $\mu < \rho$. Hence a neighborhood of V of size $\min\{\varepsilon_i, \rho\}$ belongs to A_K, hence A_K is open.

3.5 If $Z_K = \{z \in K \ : \ \det(D_x V) = 0\}$, then $V(Z_K)$ is a zero measure set by Sard's Theorem. Let $Z \subset \mathbb{R}^d$ be a zero measure set and, for each $v \in \mathbb{R}^d$, define $Z(v) = \{z \in \mathbb{R}^d \ : \ z - v \in Z\}$. Show that for each $\varepsilon > 0$ there exists $v \in \mathbb{R}^d$, $\|v\| \leq \varepsilon$, such that $0 \notin Z(v)$. Given $V \in \mathcal{C}^1(\mathbb{R}^d,\mathbb{R}^d)$, use this to show that for each $\varepsilon > 0$ there exists $v \in \mathbb{R}^d$, $\|v\| = 1$ such that $V_\varepsilon(x) := V(x) + \varepsilon v$ has the property that $\det(D_x V_\varepsilon) = \det(D_x V) = 0$

implies $V_\varepsilon(x) \neq 0$. An application of the implicit function theorem then shows that the zeroes of V_ε are isolated. Finally, construct $\tilde{V}_\varepsilon$, $\|V_\varepsilon - \tilde{V}_\varepsilon\|_{\mathcal{C}^1} \leq \varepsilon$, such that the zeroes are unchanged but the derivative is hyperbolic, hence $\tilde{V}_\varepsilon \in A_K$. This last step can be performed locally, so it suffices to show how to perform it around one single point. First of all, note that, by continuity, there exists $\alpha > 0$ such that $V_\varepsilon(x) = 0$ implies $\|(D_x V_\varepsilon)^{-1}\| \leq \alpha^{-1}$. Next, let $x_0 \in K$ such that $V_\varepsilon(x_0) = 0$. Then $V_\varepsilon(x) = D_{x_0} V_\varepsilon(x - x_0) + o(x - x_0)$. Thus, there exists $\delta > 0$ such that, for all $\|x - x_0\| \leq \delta$,[13]

$$\|V_\varepsilon(x)\| \geq \frac{\alpha}{2}\|x - x_0\|.$$

Finally, consider the vector field $\tilde{V}_t(x) = V_\varepsilon(x) + t(x - x_0)\varphi(x - x_0)$. Where $\varphi \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ is some fixed function such that the support of φ is contained in the ball of radius δ, $\varphi(0) = 1$, $\nabla\varphi(0) = 0$ and $\|\varphi\|_\infty \leq 1$. Then

$$\|\tilde{V}_t(x)\| \geq (\frac{\alpha}{2} - t)\|x - x_0\|$$

so if $t < \frac{\alpha}{2}$, the filed $\tilde{V}_t(x)$ has the same zeroes than V_ε. Moreover, $D_{x_0}\tilde{V}_t = D_{x_0} V_\varepsilon + t\mathbb{1}$ which is hyperbolic and

$$\|V_\varepsilon - \tilde{V}_t\|_{\mathcal{C}^1} \leq 2t\delta + t\|\varphi\|_{\mathcal{C}^1}$$

which can be made smaller than ε by choosing t sufficiently small.

3.7 It suffices to show that B_K is generic for each compact $K \subset \mathbb{R}^d$. The openness comes from the fact that a small perturbations cannot change the condition on the rank. For the density, consider the set $\Omega := \{(x, \lambda) \in K^1 : \operatorname{rank}(\partial_x V \ \partial_\lambda V) < d\}$. Using the same strategy as in Theorem 3.1.6 show that $V(\Omega)$ has zero Lebesgue measure.[14] This means that, for each $\varepsilon > 0$ there exists $v \in \mathbb{R}^d$, $\|v\| \leq \varepsilon$, such that for each $(x, \lambda) \in K^1$ such that $V(x, \lambda) = -v$ holds $\operatorname{rank}(\partial_x V \ \partial_\lambda V) = d$. We can then consider the vector field $V_\varepsilon = V + v$ and argue as in the first part of Problem 3.5.

3.13 We know from the discussion Lemma 3.2.2 that there exists $x(\lambda)$ such that $V(x(\lambda), \lambda) = 0$, we can then set $\alpha(\lambda) = x(\lambda)$. We get then the wanted equation with the new remainder given by $R(\xi + x(\lambda), \lambda) - R(x(\lambda), \lambda)$. The other properties of R are obtained by direct computation.

[13] Note that, by the uniform continuity of the derivative on K, δ can be chosen independent of the point.

[14] In fact, this is nothing else than another special case of the general Sard Theorem.

3.14 Remember that the change of variable must be performed on the equation $\dot{x} = V(x, \lambda)$, so the vector field changes as $D^{-1}V(Dz)$. In addition, since $\partial_\xi R(0,0) = 0$, Problem 3.35 implies that we can write $\partial_\xi R(0,\lambda) = C(\lambda)\lambda$ for some $\mathcal{C}^{r-1}$ matrix C. Choose $D(\lambda) = D_0(\lambda)(\mathbb{1} + D_1(\lambda))$. Since we do not want to change the form of $\partial_x V$ at first order in λ, we impose $[D_0, A] = 0$. Show that this implies $D_0(\lambda) = \begin{pmatrix} 1 & -a(\lambda) \\ a(\lambda) & 1 \end{pmatrix}$. Show that one can choose a such that

$$D_0^{-1}\partial_x V(0,\lambda)D_0 = A + \lambda H(\lambda)$$

with $H_{11} = H_{22} \geq 0$. Note then that $H_{ii}(0) \neq 0$ since $\operatorname{Tr} H(0) \neq 0$ by hypothesis. Next, choose $D_1 = \begin{pmatrix} 0 & 0 \\ 0 & \lambda b(\lambda) \end{pmatrix}$. Show that b can be chosen so that

$$(\mathbb{1} + D_1)^{-1}D_0^{-1}\partial_x V(0,\lambda)D_0(\mathbb{1} + D_1) = A + \lambda\tilde{H}(\lambda)$$

with $\tilde{H}_{ii} = H_{ii}$ and $\tilde{H}_{12} = -\tilde{H}_{21}$. The problem is then solved by choosing ρ.

3.30 Using a "dot" to mean differentiation holds $\frac{d}{d\lambda}\operatorname{Tr}(\Pi A) = \operatorname{Tr}(\dot{\Pi}A + \Pi\dot{A})$. If B is the portion of the spectrum associated with $\Pi(0)$ and γ a curve surrounding it and no other part of the spectrum, then

$$\dot{\Pi}(0) = \frac{1}{2\pi i}\int_\gamma (z - A(0))^{-1}\dot{A}(0)(z - A(0))^{-1}dz$$

Thus

$$\begin{aligned}\operatorname{Tr}(\dot{\Pi}A) &= \operatorname{Tr}(\Pi\dot{A}) + \frac{1}{2\pi i}\int_\gamma z \operatorname{Tr}\left((z - A(0))^{-1}\dot{A}(0)(z - A(0))^{-1}\right)dz \\ &= \operatorname{Tr}(\Pi\dot{A}) + \frac{1}{2\pi i}\int_\gamma z \operatorname{Tr}\left((z - A(0))^{-1}\dot{A}(0)\right)dz = \operatorname{Tr}(\Pi\dot{A}).\end{aligned}$$

Notes

The present discussion is intended only to give a flavor of the subject and of how it can be systematically developed. For a more complete (and advanced) treatment of bifurcation theory, see [Arn83, CH82]. As a historical curiosity, note that the bifurcation theory can be traced back to antiquity, notably to the *Floating bodies* treatise by Archimed.

Chapter 4

Global Behavior–regular motion

Different local behaviors have been analyzed in the previous chapter. Unfortunately, such analysis is insufficient if one wants to understand the *global* behavior of a Dynamical System. To make precise what we mean by global behavior we need some definitions.

Definition 4.0.1 *Given a Dynamical System (X, ϕ_t), $t \in \mathbb{N}$ or $\mathbb{R}_+$, a set $A \subset X$ is called* invariant *if, for all t, $\emptyset \neq \phi_t^{-1}(A) \subset A$.*

Essentially, the global understanding of a system entails a detailed knowledge of its invariant sets and the dynamics in their neighborhoods. This is, in general, very hard to achieve; essentially, the rest of this book is devoted to the study of some special cases.

Remark 4.0.2 *We start with some simple considerations in the case of continuous Dynamical Systems (this is part of a general theory called Topological Dynamical Systems*[1]*) and then we will address more subtle phenomena that depend on the smoothness of the systems.*

4.1 Long time behavior and invariant sets

We are interested in the long-time behavior of a system, and we look at it locally (i.e., in the neighborhood of a point). Then, three cases are possible:

[1] Recall that a Topological Dynamical Systems is a couple (X, ϕ_t) where X is a topological space and ϕ_t is a continuous action of $\mathbb{R}$ (or $\mathbb{R}_+, \mathbb{N}, \mathbb{Z}$) on X.

either the motion leaves the neighborhood and never returns, or leaves the neighborhood but eventually comes back, or never leaves. Clearly, in the first case, the neighborhood in question has little interest in the study of the long-time behavior. This is made precise by the following.

Definition 4.1.1 *Given a Dynamical System (X, ϕ_t), a point $x \in X$ is called* wandering *if there exists a neighborhood U of x and a $t_0 \geq 1$ such that, for all $t \geq t_0$, $\phi_t(U) \cap U = \emptyset$. A point that is not wandering is called* non-wandering. *The set of non-wandering points is called $NW(\{\phi_t\})$ or simply NW if no confusion arises.*

Problem 4.1 *If $\phi_t \in \mathcal{C}^0$, then the set NW is closed and forward invariant (i.e. $\phi_t(NW) \subset NW$ for each $t \geq 0$). If the ϕ_t are open maps, then NW is also invariant.*

Problem 4.2 *Construct an example of a topological dynamical systems in which the non-wandering set is not invariant.*

Problem 4.3 *Show that if A is invariant, then the sets $\Lambda = \cap_{t=0}^{\infty} \phi_t^{-1} \overline{A}$ and $\Omega = \overline{\cup_{t=0}^{\infty} \phi_t(A)}$ are non-empty, invariant and, more, $\phi_t^{-1}(\Lambda) = \Lambda$ and $\phi_t^{-1}(\Omega) = \Omega$*

The relevance for the long time behavior is emphasized by the following lemma.

Lemma 4.1.2 *If $K \subset X$ is compact and $K \cap NW = \emptyset$, then for all $x \in K$ there exists T such that $\phi_t(x) \notin K$ for all $t \geq T$. In addition, if K is invariant, then T can be chosen independent of x.*

PROOF. If all the points in K are wandering, then for each $x \in K$ there exists a neighborhood $U(x)$ and a time $t(x)$ such that $\phi_t U(x) \cap U(x) = \emptyset$ for all $t \geq t(x)$. Clearly $\{U(x)\}_{x \in K}$ is an open covering of K, hence we can extract a finite subcover. Let $\{U_i\}_{i=1}^N$ be such a subcover, let $\{t_i\}$ be the corresponding associated times. If $x \in K$ then $x \in U_i$ for some $i \in \{1, \ldots, N\}$, and $\phi_t(x) \notin U_i$ for $t \geq t_i$. If $\phi_t(x) \notin K$ for all $t \geq t_i$, then we are done. If there exists $t \geq t_i$ such that $\phi_t(x) \in K$, then $\phi_t(x)$ must belong to another U_j, that will leave forever for $t \geq t_j$. It is then clear that $\phi_t(x)$ cannot remain in K for a time longer than $\sum_i t_i$, nor can the trajectory return for more than N times.

If K is invariant then it follows that if $x \notin K$ then $\phi_t x \notin K$ for all $t \geq 0$. Thus, once a point exits K, it can never come back. The above argument then shows that each point must exist forever in a time at most $\sum_i t_i$. □

Corollary 4.1.3 *If $K \subset X$ is compact and invariant, then either there exists $t \in \mathbb{R}_+$ such that $\phi_t^{-1} K = \emptyset$ or $NW \cap K \neq \emptyset$.*

PROOF. If $NW \cap K = \emptyset$, then Lemma 4.1.2 imply that there exists $t \in \mathbb{R}_+$ such that $\phi_t K \cap K = \emptyset$, hence $\phi_t^{-1} K = \emptyset$. □

To see the connection to long time behavior and invariant sets, we need an extra definition

Definition 4.1.4 *Given a topological Dynamical System (X, ϕ_t), $t \in I \in \{\mathbb{R}, \mathbb{Z}, \mathbb{R}_+, \mathbb{N}\}$, and $x \in X$ we call $\omega(x)$ (*the ω-limit set of x*) the accumulation points of the set $\cup_{t \geq 0}\{\phi_t(x)\}$. If t belongs to $\mathbb{R}$ or $\mathbb{Z}$, then the α-limit set is defined analogously with $t \leq 0$.*

Problem 4.4 *Show that the ω-limit sets are closed sets such that $\phi_t(\omega) = \omega$ (hence if ϕ_t is invertible then the omega limits are invariant).*

Theorem 4.1.5 *For each $x \in X$ we have $\omega(x) \subset NW$. In addition, if X is a proper metric space,[2] then for each $z \in X$ either holds $\lim_{t\to\infty} d(\phi_t(x), z) = \infty$, or $\lim_{t\to\infty} d(\phi_t(x), NW) = 0$.*

PROOF. Let $x \in X$. If $z \in \omega(x)$, then for each neighborhood U of z we have $\{t_n\} \subset \mathbb{R}_+$ such that $\phi_{t_n}(x) \in U$. Thus $\phi_{t_{n+1}-t_n} U \cap U \supset \{\phi_{t_{n+1}}(x)\} \neq \emptyset$. Hence $z \in NW$.

Let us come to the second part of the Theorem. If the two alternatives do not hold, then there exists a compact set (a closed ball) that contains infinitely many points of the orbit of x all at a finite distance from NW. This implies that the orbit has an accumulation point (hence an element of $\omega(x)$) not in NW contradicting the first part of the Theorem. □

In particular the above Theorem shows that all the interesting long time dynamical behavior happens in a neighborhood of the non-wandering set.

Problem 4.5 *Given a discrete topological dynamical system (X, T), let $A = NW(T)$. Since A is forward invariant, one can consider the restriction S of T to A. Find an example in which $NW(S)$ is strictly smaller than A.*

Definition 4.1.6 *Given a Dynamical System (X, ϕ_t), a point $x \in X$ is called* recurrent *if $x \in \omega(x)$. The set of recurrent points is called $R(\{\phi_t\})$, or simply R if no confusions arises.*

Problem 4.6 *Consider a linear system $\dot{x} = Ax$. Show that if A is hyperbolic, then $NW = \{0\}$.*

[2] That is, a distance d is defined and the base for the topology is made of the sets $B_r(x) = \{y \in X \ : \ d(x, y) < r\}$ (this is called a *metric space*). A *proper* metric space is one in which all the closed balls $\{y \in X \ : \ d(x, y) \leq r\}$ are compact.

Problem 4.7 *Consider a saddle-node bifurcation in one dimension. Show that in a small neighborhood of the bifurcation point, when two fixed points x_1, x_2 are present, $NW = \{x_1, x_2\}$. Show that this may not be the case in higher dimensions.*

Problem 4.8 *Consider the ODE $\dot{x} = \begin{pmatrix} 0 & -\omega_0 \\ \omega_0 & 0 \end{pmatrix}$, $\omega_0 > 0$, $2\pi\omega_0 \notin \mathbb{Q}$. Show that $NW = \mathbb{R}^2$, while for each $x \in \mathbb{R}^2$ holds $\omega(x) = \{z \in \mathbb{R}^2 \ : \ \|z\| = \|x\|\}$.*

Problem 4.9 *In the case of the Hopf bifurcation in two dimensions when the fixed point O is repelling, and hence the periodic orbit γ is attracting, show that (in a neighborhood of O for the bifurcation parameter small enough) $NW = \{O\} \cap \gamma$.*

Remark 4.1.7 *We have thus seen examples in which the ω-limit sets can be a point or a periodic orbit,* **do other possibilities exists?**

This question is going to lead us on a long journey.

4.2 Poincaré-Bendixon

The first result is for surfaces.

Theorem 4.2.1 (Poincaré-Bendixon) *Let Σ be a surface on which the Jordan Theorem applies and (Σ, ϕ_t) a flow generated by a C^1 vector field. Assume that $x \in \Sigma$ has a compact omega limit set which contains no fixed points, then $\omega(x)$ is a periodic orbit.*

PROOF. Let x be a point with a compact omega limit set which does not contain fixed points, then let $\xi \in \omega(x)$. Note that $\omega(\xi) \subset \omega(x)$ since if $z \in \omega(\xi)$ then there exists $\{t_n\}$ such that $d(z, \phi_{t_n}(\xi)) \leq n^{-1}$. But then, since the flow is $\mathcal{C}^1$, there are neighborood of U_n of ξ such that $d(z, \phi_{t_n}(\zeta)) \leq 2n^{-1}$ for each $\zeta \in U_n$. Since $\xi \in \omega(x)$, there eists times $\{s_n\}$ such that $\phi_{s_n}(x) \in U_n$, hence $d(z, \phi_{s_n+t_n}(x)) \leq 2n^{-1}$, that is $z \in \omega(x)$.
Our first goal is to show that $\omega(\xi)$ contains a closed orbit.
Let V be the vector field generating the flow, and $\zeta \in \omega(\xi)$. By assumption $V(\zeta) \neq 0$, let n be a vector normal to $V(\zeta)$. Let $\mathbb{S}$ be the line passing through ζ and parallel to n. Let $S \subset \mathbb{S}$ be a segment containing ζ and such that, at each point $z \in S$, $V(z) \neq 0$ and $\langle V(z), V(\zeta)\rangle \geq \frac{1}{2}\|V(z)\|\|V(\zeta)\|$. Since ζ is an accumulation point of $\{\phi_t(\xi)\}_{t\in\mathbb{R}_+}$, the trajectory of ξ intersects S infinitely many times, let $\{\phi_{t_n}(\xi)\}_{n\in\mathbb{N}} \subset S$ be the intersections of $\{\phi_t(\xi)\}_{t\in\mathbb{R}_+}$ with S. We can then consider the close curve γ_n consisting of $\{\phi_t(\xi)\}_{t\in[t_n,t_{n+1}]}$ and the segment $I_n \subset S$ connecting $\phi_{t_n}(\xi)$ with $\phi_{t_{n+1}}(\xi)$. If I_n consists of just one

point, then the trajectory is periodic. Otherwise, by Jordan's Curve Theorem, such a curve divides the plane into two open connected components.
Let us call A_n the component of the side of S toward which point $V(\zeta)$, and B_n the other. Note that a trajectory can cross the boundary of the two components only through I. Therefore, a trajectory can change the connected component only going from B_n to A_n through I. Consequently, trajectories in A_n stay in A_n forever. In addition, the trajectory $\phi_{t_{n+1}+s}(\xi)$ for small $s > 0$ enters in A_n, thus $A_{n+1} \subset A_n$. Moreover, if the trajectory is not periodic, then the inclusion is strict as $\phi_{t_n}(\xi) \notin A_{n+1}$. In addition, since the trajectory accumulates to ζ, it follows that $\zeta \in A_n$ for all n.
Next, we follow the trajectory of $\phi_t(x)$. By assumpton there exists a sequence of times $\{s_n\}$ such that $\phi_{s_n}(x)$ converges to $\phi_{t_1}(\xi)$. Thus $\phi_{s_n+t_2-t_1}(\xi)$ converges to $\phi_{t_2}(\xi)$. But this means that, for t large enough, $\phi_t(x) \in A_2$. Hence, the trajectory will remain in A_2 forever, and since $\phi_{t_1}(\xi) \notin A_2$, it cannot accumulate to it, contrary to the hypothesis.
It follows that it must be $\phi_{t_1}(\xi) = \phi_{t_2}(\xi)$, that ξ belongs to a periodic trajectory of period $T = t_2 - t_1$.[3]
Next, we show that $\omega(x)$ consists of exactly a periodic orbit. If $\phi_{s_n}(x)$ is close enough to $\phi_{t_1}(\xi)$, then it will intersect again S, and, arguing as before, will do so closer to $\phi_{t_1}(\xi)$, hence it will intersect S, infinitely many time converging monotonically to $\phi_{t_1}(\xi)$. Thus, for each ε there exists $\bar{n}$ such that $\phi_t(x)$ will be in an ε-neighborhrood of the periodic trajectory for each $t \geq s_{\bar{n}}$. To conclude let $\xi' \in \omega(x) \setminus \{\xi\}$. Then $\phi_t(x)$ must accumulate to ξ', and the previous argument shows that ξ' must belong to an ε-neighborhood of the periodic trajectory. Since ε is arbitrary, it follows that ξ' belongs to the periodic trajectory, that is, $\omega(x)$ is just one periodic trajectory. □

4.3 Equations on the Torus

The conclusion of our previous chapters is that a generic family of vector fields in $\mathbb{R}^2$ can have a very limited choice of bounded invariant sets: either a fixed point and the associated stable and unstable manifolds, or (by Poincaré-Bendixon) a periodic orbit. Yet, one can have a differential equation on different manifolds, notably the torus $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$.

Problem 4.10 *Consider the vector fields $V(x) = \omega \in \mathbb{R}^2$ on $\mathbb{T}^2$ and show that the orbit of the associated flow can be everywhere dense.*

The above problem shows that on $\mathbb{T}^2$ it is possible to have a new ω-limit set: $\mathbb{T}^2$ itself! Can such a situation take place for an open set or a dense set of vector fields? To understand the situation, it is useful to generalize the setting of Problem 4.10.

[3] For a slighly more detailed argument, see [HS74].

Definition 4.3.1 *A closed non self-intersecting curve $\gamma \in \mathcal{C}^r(S^1, \mathbb{T}^2)$, $r \geq 1$, is called a* global (cross) section *for the flow associated to V if*

a) γ' is always transversal to V.[4]

b) for each $x \in \mathbb{T}^2$ there exists $t \in \mathbb{R}_+$ such that $\phi_t(x) \in \gamma$.

Given a cross section γ we can define the return time $\tau : \gamma \to \mathbb{R}_+$ as the first $t > 0$ such that $\phi_t(x) \in \gamma$ and the Poincaré map $f : \gamma \to \gamma$ as $f(x) = \phi_{\tau(x)}(x)$.

Problem 4.11 *Show that if $\gamma \in \mathcal{C}^r(S^1, \mathbb{T}^2)$ is a global cross section and f is the associate Poincaré map, then $f \in \mathcal{C}^r$ and (γ, f) is a Dynamical Systems that describe the dynamics of the flow when it returns to γ.*

Lemma 4.3.2 (Siegel) *Let $V \in \mathcal{C}^r(\mathbb{T}^2, \mathbb{R}^2)$ be a nowhere zero vector field. If the associated flow has no periodic orbits, then there exists a global section $\gamma \in \mathcal{C}^r$. In addition, if $f : \gamma \to \gamma$ is the Poincaré map associated to the flow, then $f \in \mathcal{C}^r(\gamma, \gamma)$.*

Proof. The (nice) idea is to construct a section close to an orbit. Let ϕ_t be the flow associated with the vector field V. Note that Corolalry 4.1.3 implies $NW \neq \emptyset$. Let $x \in NW$ and consider an open segment, of length less than $1/2$, Σ, $x \in \Sigma$, transversal to the vector field (similar to the construction in the Flow Box Theorem 2.1.1). Since x is non-wandering and due to Theorem 2.1.1, there exists $z \in \Sigma$, $z \neq x$, and $t \in \mathbb{R}$ such that $\phi_t(z) \in \Sigma$, this being the first return to Σ. Since there are no periodic orbits $z \neq \phi_t(z)$.

We will construct a global section close to $\{\phi_s(z)\}_{s=0}^t \cup \Sigma$. Note that the closed curve that one obtains joining z to $\phi_t(z)$ along Σ cannot be homotopic to a point. Otherwise, the curve would have an interior homeomorphic to a disk in $\mathbb{R}^2$ from which the orbits cannot escape either in the future or the past. By the Poincaré-Bendixon theorem, this would imply the existence of a periodic orbit, contrary to the hypothesis. To properly explain the construction it is convenient to introduce a flow box type system of coordinates near such an orbit.

For $s \in [-1/2, 1/2]$ let $\varphi(s) = z + s(x - z)\|x - z\|^{-1} \subset \Sigma$. Clearly $\varphi(0) = z$, $\varphi(\|x - z\|) = x$, and holds $\varphi([-1/2, 1/2]) \supset \Sigma$. Next, for each $y \in \Sigma$ let $s \in [-1/2, 1/2]$ be the unique number such that $y = \varphi(s)$ and $\tau(s) = \inf\{t > 0 \ : \ \phi_t(y) \in \Sigma\}$ be the first return time to the section. By Theorem 2.1.1 and Corollary 1.1.14 there exists $\delta \in (0, 2\|x - z\|)$ such that $\tau \in \mathcal{C}^r([-\delta, \delta], \mathbb{R}_+)$. For $A := \{(s, t) \in \mathbb{R}^2 \ : \ s \in [-\delta, \delta], t \in [0, \tau(s))\}$ let us define the map $\Xi : A \to \mathbb{T}^2$ by $\Xi(s, t) = \phi_t(\varphi(s))$. Note that this map is $\mathcal{C}^r$ and invertible (provided δ is chosen small enough), hence it can be used as a change of coordinates. Note that this are essentially the coordinates used in

[4] That is, the vectors $\{\gamma'(t), V(\gamma(t))\}$ span $\mathbb{R}^2$ for all $t \in S^1$.

the flow box theorem, only now they are used in a long neighborhood of an orbit.

The next step is to understand how the orbit comes back. Indeed, if we use standard flow box coordiantes (s', t') in a neighborhood of Σ, then $(s, t) = (s', t')$ for $t \geq 0$ but for t close to $\tau(s)$ we are again in the neighborhood of Σ corresponding to $t' < 0$. The change of coordinates can then be described by the function θ such that $\phi_{\tau(s)}\varphi(s) = \varphi(\theta(s))$. Then (s, t) corresponds to $(\theta(s), t - \tau(s))$.

Problem 4.12 *Let $\tau_0 = \tau(0)$, then $\langle x - z, \frac{d}{ds}\phi_{\tau_0}(\varphi(s))|_{s=0}\rangle > 0$.*[5]

The above problem means simply that $\theta' > 0$.

To conclude we must analyze two possibilities: either $\phi_\tau z$ is closer to x than z or vice versa. The two cases are treated exactly in the same way so we discuss only the first, that is $\theta(0) > 0$. We can then chose $\varepsilon \in (0, \delta)$ such that $\theta(-\varepsilon) > 0$. Consider a line $(\varepsilon - 2\varepsilon\tau_0^{-1}t, t)$, $t \in [0, \tau_0]$, obviously it is always transversal to the flow. If we look at it in the standard flow box coordinates in a neighborhood of Σ we see that it start as a decreasing curve and, since $\theta' > 0$, it reappears (for $t' < 0$) as a still decreasing curve. It is then easy to see that it can be smoothly deformed, in a neighborhood of Σ, into a closed curve that is always transversal to the flow. We have thus constructed a smooth transversal section it remains to show that it is global.

Problem 4.13 *Consider a piecewise smooth closed curve Γ in $\mathbb{T}^2$. Show that $\mathbb{T}^2 \setminus \Gamma$ is either disconnected (and one connected component is isomorphic to an open set in $\mathbb{R}^2$) or it is isomorphic to a cylinder.*

If the above section would not be global, then there would be trajectories that stay forever in a set (either a piece of $\mathbb{R}^2$ or a cylinder) to which Poincaré-Bendixon applies. But this would imply the presence of a periodic orbit, contrary to the asumption. □

Problem 4.14 *Show that, in the setting of the above theorem, the sign of f' cannot change and that the condition $f' \neq 0$ is generic.*

It is important to notice that, given a topological Dynamical System (M, f) and a function $\tau \in \mathcal{C}^0(M, \mathbb{R}_+ \setminus \{0\})$ (called *roof function*) one can always see them as a Poincaré section and a return time of a flow. The resulting object is called a *suspension* or *standard flow* and is constructed as follows.

Consider the set $\tilde{\Omega} = \{(x, s) \in M \times \mathbb{R}_+ \ : \ s \in [0, \tau(x)]\}$ with the topology induced by $M \times \mathbb{R}_+$ equipped with the product topology.

[5] This is really a consequence of the fact that the torus is orientable, yet it can be proven directly in several ways.

Problem 4.15 *Consider the relation* $(x,s) \sim (y,t)$ *iff* $x = y$ *and* $s = t$ *or* $s = \tau(x)$, $t = 0$ *and* $y = f(x)$ *or* $t = \tau(y)$, $s = 0$ *and* $x = f(y)$. *Prove that it is an equivalence relation.*

One can then consider the space of the equivalence classes $\Omega = \tilde{\Omega}/\sim$ with the induced topology, this is the space on which the flow is defined: let $t \leq \inf \tau$, define

$$\phi_t(x,s) = \begin{cases} (x, s+t) & \text{if } t < \tau(x) - s \\ (f(x), t+s-\tau(x)) & \text{if } t \geq \tau(x) - s \end{cases}$$

and extend ϕ_t by the group property.

Theorem 4.3.3 *Let* $V \in \mathcal{C}^2(\mathbb{T}^2, \mathbb{R}^2)$ *be a nowhere zero generic vector field with no periodic orbits. Then for each point* $y \in \mathbb{T}^2$, $\omega(y) = \mathbb{T}^2$.

Proof. By Lemma 4.3.2 we have a smooth global section γ with a Poincaré map g. Let $h : S^1 \to \gamma$ be a parametrization of γ. If we set $f = h^{-1} \circ g \circ h$, we can consider the return map as $\mathcal{C}^2$ map on the unit circle such that $f' \neq 0$ at each point. Note that a periodic point for the map f corresponds to a periodic orbit for the flow, hence f cannot have periodic orbits. The claim follows then by Lemma 4.5.2 in which it is proven that a smooth circle map with no periodic orbits has dense orbits. □

The final natural question is:
In the hypotheses of Theorem 4.3.3, is it possible to conjugate the flow to a rigid rotation of the torus? if yes, to which one and how smooth is the conjugation?

Motivated by the above question and results, we will now study orientation-preserving circle maps. It turns out to be interesting and helpful to study their properties in relation to their increasing smoothness.

4.4 Circle maps: topology

Here, and in the following, we study a Dynamical System (S^1, f) where f is an orientation-preserving homeomorphism of S^1 (i.e., f is invertible and $f(S^1) = S^1$).

To begin with, we assume continuity only.

First of all, note that one can lift the map f to the universal cover $\mathbb{R}$ of the circle, that is defining $\pi : \mathbb{R} \to S^1$ as $\pi(x) = x \mod 1$, it is possible to find $F \in \mathcal{C}^0(\mathbb{R}, \mathbb{R})$ such that

$$f \circ \pi = \pi \circ F.$$

Problem 4.16 *Construct explicitly such an F. Show that*

$$F(x+1) = F(x) + 1.$$

Problem 4.17 *If there exists $L > 0$ such that $-L \leq a_{m+n} \leq a_n + a_m + L$ for all $n, m \in \mathbb{N}$, then the limit $\lim_{n\to\infty} \frac{a_n}{n}$ exists.*

Lemma 4.4.1 *Let $f : S^1 \to S^1$ be an homeomorphism and $F \in \mathcal{C}^0(\mathbb{R}, \mathbb{R})$ a lift of f. Then the limit*

$$\tau(f) := \lim_{|n|\to\infty} \frac{F^n(x)}{n} \mod 1$$

exists and is independent both from the point and the lift.

Proof. Applying Problem 4.17 to the sequence $F^n(x)$ the existence of the limit follows. The other assertions depend on the already mentioned equality $F(x+1) = F(x) + 1$. □

Lemma 4.4.2 *Show that $\tau(f) \in \mathbb{Q}$ if and only if f has a periodic orbit.*

Proof. If $f^q(x) = x$ and F is a lift then it must be $F^q(x) = x+p$ for some $p \in \mathbb{N}$. This immediately implies $F^{kq}(x) = x + kp$ and hence $\tau(f) = \frac{p}{q} \in \mathbb{Q}$. On the other hand, if $\tau(f) = \frac{p}{q} \in \mathbb{Q}$, we have $\tau(f^q) = p \mod 1 = 0$. It thus suffices to prove that $\tau(f) = 0$ implies f has a fixed point. Let us do a proof by contradiction: we suppose that f has no fixed points. Note that this is the same than saying that $G(\mathbb{R}) \cap \mathbb{Z} = \emptyset$ where $G(x) = F(x) - x$. Since G is continuous this implies $\max G - \min G < 1$. Let $\alpha = \min G$, $\beta = \max G$. Note that, by properly choosing the lift F, one can insure tat $[\alpha, \beta] \subset (0, 1)$. Then

$$F^n(x) = G(F^{n-1}(x)) + F^{n-1}(x) \geq \alpha + F^{n-1}(x) \geq n\alpha$$

hence $\tau(f) \geq \alpha$, analogously $\tau(f) \leq \beta$ which contradicts $\tau(f) = 0$. □

Problem 4.18 *Given $f \in \mathcal{C}^0(S^1, S^1)$, for any interval $I \subset S^1$, if $f(I) \subset I$, then f has a fixed point in I.*

Problem 4.19 *If $\tau(f) \notin \mathbb{Q}$, then for each $n \in \mathbb{N} \setminus \{0\}$ and $x, y \in S^1$, $\{f^k(y)\}_{k\in\mathbb{N}} \cap [x, f^n(x)] \neq \emptyset$.*

Problem 4.20 *If $\tau(f) \notin \mathbb{Q}$, then for each $x \in S^1$ there exist infinitely many $n \in \mathbb{Z}$ such that $\{f^k x\}_{|k|<n} \cap [x, f^n x] = \emptyset$.*

Lemma 4.4.3 *For any homomorpfism $f : S^1 \to S^1$ with $\tau(f) \notin \mathbb{Q}$ and any $x, y \in S^1$ holds $\omega(x) = \omega(y)$.*

PROOF. If $z \in \omega(x)$, then there exists $\{n_j\}$ such that $\lim_{j\to\infty} f^{n_j}(x) = z$. But then Problem 4.19 implies that for each $j \in \mathbb{N}$ there exists $k_j \in \mathbb{N}$ such that $f^{k_j}(y) \in [f^{n_j}(x), f^{n_{j+1}}(x)]$. Clearly $\lim_{j\to\infty} f^{k_j}(y) = z$, thus $z \in \omega(y)$. Reversing the role of x and y the Lemma follows. □

Problem 4.21 *Let f be a homeomorphism of S^1 with irrational rotation number show that for each $\varepsilon > 0$ there exists a homeomorphism f_ε, $\|f - f_\varepsilon\|_\infty \le \varepsilon$, with $\tau(f_\varepsilon) \in \mathbb{Q}$.*

Problem 4.22 *Note that τ is a map from circle homomorphisms to $[0,1]$. Show that it is a continuous map.*

Problem 4.23 *Let f_λ be a one parameter family of homeomorphisms such that $\tau(f_0) < \tau(f_1)$. Suppose that $\tau(f_\lambda)$ is increasing, what can you say on the possible intervals in which it is not strictly increasing?*

4.5 Circle maps: differentiable theory

In this section we investigate the consequences of assuming that the map enjoys some regularity.

Lemma 4.5.1 *Assume $f \in \mathcal{C}^2(S^1, S^1)$ and $\ln f' \in \mathcal{C}^1(S^1, \mathbb{R})$.*[6] *If $\tau(f) \notin \mathbb{Q}$ and $x_0 \notin \omega(x_0)$, then*

$$\sum_{n=0}^{\infty} (f^n)'(x_0) < \infty.$$

PROOF. Let $U(x_0) \ni x_0$ be the largest open interval not intersecting $\omega(x_0)$, call $K(x_0)$ its closure. First of all, the invariance of the ω-limit set implies $\{f^n(\partial K(x_0))\}_{n=1}^{\infty} \subset \omega(x_0)$. This implies that either $f^n K(x_0) \cap U(x_0) = \emptyset$ or $f^n K(x_0) \supset K(x_0)$ but the latter would imply the existence of a fixed point for f^n, which is impossible, hence all the sets $\{f^n U(x_0)\}_{n\in\mathbb{Z}}$ must be disjoint. We can now conclude thanks to a typical distortion estimate: let $K_n(x_0) := f^n(K(x_0))$, then, setting $D := \left|\frac{f''}{f'}\right|_\infty$,

$$\begin{aligned} 1 > \sum_{n\in\mathbb{N}} |K_n(x_0)| &= \sum_{n\in\mathbb{N}} \int_{K(x_0)} (f^n)'(x)dx = \sum_{n\in\mathbb{N}} (f^n)'(x_0) \int_{K(x_0)} \frac{(f^n)'(x)}{(f^n)'(x_0)} dx \\ &\ge \sum_{n\in\mathbb{N}} (f^n)'(x_0) \int_{K(x_0)} e^{-\sum_{k=0}^{n-1} |\ln f'(f^k(x)) - \ln f'(f^k(x_0))|} dx \\ &\ge \sum_{n\in\mathbb{N}} (f^n)'(x_0) \int_{K(x_0)} e^{-\sum_{k=0}^{n-1} D|K_k(x_0)|} dx \ge |K(x_0)| e^{-D} \sum_{n\in\mathbb{N}} (f^n)'(x_0). \end{aligned}$$

[6] These hypotheses can be slightly weakened, see [HK95].

□

Lemma 4.5.2 *If $\tau(f) \notin \mathbb{Q}$, then, for all $x \in S^1$, $\omega(x) = S^1$.*

PROOF. We use the same notation as in Lemma 4.5.1. If the Lemma is false then there exists $x \in S^1$ such that $\omega(x) \neq S^1$. But by Lemma 4.4.3 all the omega limit sets are equal, hence there exists $x_0 \in S^1$ such that $x_0 \notin \omega(x_0)$. Note that if there exists $n \in \mathbb{N}$, $n \neq 0$, such that $f^n(x_0) \in K(x_0)$ then, by the invariance of $\omega(x_0)$, it must be $f^n(x_0) \neq \partial K(x_0) \subset \omega(x_0)$ and then Problem 4.19 implies that there are infinitely many k such that $f^k(x_0) \in [x_0, f^n(x_0)] \subset K(x_0)$, but this is impossible since such an interval does not contain accumulation points of the forward trajectory. Thus, for each $n \in \mathbb{Z}$, $n \neq 0$, $f^n(x_0) \notin K(x_0)$, accordingly there exist $\delta > 0$ such that each interval $[x_0, f^n(x_0)]$ has length at least δ.

Next, choose $L > 0$, by Lemma 4.5.1 there exists $m \in \mathbb{N}$ such that $(f^n)'(x_0) < L^{-1}$, for all $n > m$. We can then apply Problem 4.20 to find an $|n| > m$ such that $\{f^k x\}_{|k|<n} \cap [x_0, f^n(x_0)] = \emptyset$. Suppose $n < 0$ and let $J_- = [x_0, f^n(x_0)]$, then for each $k \in \{1, \dots, -n-1\}$, $f^k J_- = [f^k x_0, f^{n+k} x_0]$, since the extreme of such an interval do not belong to J it follows that $f^k J_- \cap J_- = \emptyset$ (otherwise the first would be contained in the second and there would be a fixed point). Thus, setting $J = [x_0, f^{|n|}(x_0)]$, for all $k \in \{1, \dots, -n-1\}$, holds $f^k J \cap J = \emptyset$. The same result follows, setting $J_- = [x_0, f^{-n}(x_0)]$, for $n > 0$. Finally we conclude with another distortion argument

$$
\begin{aligned}
|f^{-|n|} J| &= \int_J (f^{-|n|})'(x) dx = \frac{1}{(f^{|n|})'(x_0)} \int_J \frac{(f^{|n|})'(f^{-|n|}(f^{|n|}(x_0)))}{(f^{|n|})'(f^{-|n|} x)} dx \\
&\geq \frac{1}{(f^{|n|})'(x_0)} \int_J e^{-\sum_{k=0}^{|n|-1} D|f^k J|} dx \geq L e^{-D} \delta.
\end{aligned}
$$

Then choosing $L > e^D \delta^{-1}$ leads a length of $|f^{-|n|} J|$ larger than one, which contradicts the fact that f is an homeomorphism. □

The above fact can be used to prove the following result (due to Poincaré).

Theorem 4.5.3 *If $\tau(f) = \omega \notin \mathbb{Q}$, then f is $\mathcal{C}^0$-conjugate to $R_\omega(x) = x + \omega \mod 1$.*

PROOF. See [HK95] Theorem 11.2.7. □

4.6 Circle maps: smooth theory

We have seen that the qualitative behavior of smooth circle maps with irrational rotation number is similar to the behavior of the rigid rotation in

Problem 4.10. What it is not clear is if the two dynamics can be smoothly conjugated (i.e. in the spirit of the flow box theorem, but globally). This latter problem turns out to be extremely subtle and to require much finer number theoretical consideration than distinguishing between rational and irrationals.

Since we have seen that more smoothness allows to obtain stronger results, it is natural to start by considering analytic functions.

To make the following easier, we will limit ourselves to the case of a maps close to the identity. That is maps with a covering $F : \mathbb{R} \to \mathbb{R}$ of the form $F(x) = x + \omega + f(x)$, where $f(x+1) = f(x)$ is "small".

4.6.1 Analytic KAM theory

To define the sense in which f is small we assume first that f is an analytic function. That is f is a restriction to the real axes of a function, that abusing notation we will still call f, holomorphic in a strip. Let $D_\alpha = \{z \in \mathbb{C} \ : \ |\Im(z)| \leq \frac{\alpha}{2\pi}\}$ and consider the function space

$$\mathbb{B}_\alpha = \{g \in \mathcal{C}^0(D_\alpha, \mathbb{C}) \ : \ g(z+1) = g(z) \ \forall z \in D_\alpha \, , \ g \text{ holomorphic in } \mathring{D}_\alpha\}.$$

This is a Banach space when equipped with the norm $\|g\|_\alpha = \sup_{z \in D_\alpha} |g(z)|$.

Theorem 4.6.1 *If $\tau(F) = \omega$ and there exist $\alpha_0 \in (0,1)$, $C_0 > 0$ such that if $\|f - \int_{S^1} f\|_{\alpha_0} \leq C_0 \alpha_0^3 10^{-10}$ and $\omega > 0$ satisfies*

$$\left|\omega - \frac{p}{q}\right| \geq \frac{C_0}{q^2}$$

for each $p, q \in \mathbb{N}$, then there exists $h \in \mathbb{B}_{\alpha_0/2}$ such that, setting $H(x) = x + h(x)$, $\|h\|_{\alpha_0/2} \leq 3 C_0^{-\frac{1}{3}} \|f\|_{\alpha_0}^{\frac{1}{3}}$ and, for all $x \in \mathbb{R}$,

$$H^{-1} \circ F \circ H(x) = x + \omega. \tag{4.6.1}$$

A natural question is: do irrational numbers with the above properties exist? The answer is yes (for example, all the quadratic irrationals satisfy such inequalities), but a bit of theory is needed to see it. For a quick introduction to these problems solve the Problems 4.28, 4.29, 4.30, 4.31, 4.32, 4.33, 4.34.

Remark 4.6.2 *Note that we can always reduce to the case $\int f = 0$ by subtracting the average to f and adding it to ω. As an exercise, you can show that given the map $F(x) = x + \omega + \xi + f(x)$, with f zero average and norm small as in Theorem 4.6.1, there exists a ξ for which the map is conjugated to $x + \omega$.*

Remark 4.6.3 *The unaware reader can be horrified by the* 10^{-10} *in the statement of the above theorem. Such a ridiculous number is partly due to my prioritizing readability over optimality, but it is also inherent to the method. It is well known among specialists that obtaining optimal estimates for KAM-type theorems is a very challenging problem. Indeed, it is a field of research currently active.*

PROOF OF THEOREM 4.6.1. First of all remark that, setting $\hat{f}_0 = \int_{S^1} f$, we have

$$\omega + \hat{f}_0 - \|f - \hat{f}_0\|_{\alpha_0} \leq \tau(F) \leq \omega + \hat{f}_0 + \|f - \hat{f}_0\|_{\alpha_0}$$

thus, since $\tau(F) = \omega$,

$$|\hat{f}_0| \leq \|f - \hat{f}_0\|_{\alpha_0}. \tag{4.6.2}$$

Next, note that if H is invertible, then equation (4.6.1) is equivalent to, for each $z \in D_{\alpha_0/2}$,

$$h(z+\omega) - h(z) = f(z + h(z)). \tag{4.6.3}$$

In fact, we are interested to solving the above equation only for real z. In the following to avoid confusion I will use z for a complex variable and x for a real one.

It is natural to introduce the linear operator $L_\omega g(x) = g(x+\omega) - g(x)$. If such an operator were invertible, then we could write

$$h = L_\omega^{-1} f \circ H, \tag{4.6.4}$$

that looks like a fixed point problem and hopefully can be studies with known techniques.

We have thus to study the operator L_ω. The best is to compute it in Fourier series:

$$L_\omega g(x) = \sum_{k\in\mathbb{Z}} e^{2\pi i k x}(e^{2\pi i \omega k} - 1)\hat{g}_k$$

where $g(x) = \sum_{k\in\mathbb{Z}} e^{2\pi i k x}\hat{g}_k$. Thus, provided $\hat{g}_0 = 0$,

$$L_\omega^{-1} g(x) = \sum_{k\in\mathbb{Z}\setminus\{0\}} e^{2\pi i k x} \frac{\hat{g}_k}{e^{2\pi i \omega k} - 1}.$$

Thanks to the fact that $\omega \notin \mathbb{Q}$, the coefficients in the above formula are well defined. Yet, it remains the issue of the convergence of the series. Indeed, the coefficients can be very large since,[7]

$$\left|e^{2\pi i \omega k} - 1\right| \geq 2 \inf_{p\in\mathbb{N}} |\omega k - p| \geq 2C_0 |k|^{-1}.$$

[7]Note that $|e^{ix} - 1| \geq |\sin x| \geq \frac{2x}{\pi}$, provided $x \in [0, \pi/2]$. On the other hand if $x \in [\pi/2, \pi]$, then $|e^{ix} - 1| \geq |1 - \cos x| \geq 1$. Hence we can use the simple, but not very sharp, estimate $|e^{2\pi i x} - 1| \geq \inf_{p\in\mathbb{Z}} 2|x - p|$.

This is the main difficulty of the present problem: the infamous *small divisors*. Clearly, due to the small divisors L_ω^{-1} is not a bounded operator. This makes it very hard to study directly (4.6.4). To bypass this problem we need an idea.

The idea that we will use if due to Kolomogorov and goes as follows: instead of solving (4.6.4) consider the change of variables $H_0(x) = x + h_0(x)$ where $h_0 = L_\omega^{-1}(f - \hat{f}_0)$. Of course such a change of variable it is not the right one since

$$h_0(x+\omega) - h_0(x) = f(x) - \hat{f}_0, \tag{4.6.5}$$

yet one can try to write

$$H_0^{-1} \circ F \circ H_0(x) = x + \omega + f_1(x) \tag{4.6.6}$$

and hope that f_1 is much smaller that f. If this is the case one can iterate the procedure and hope that it converges to a limiting change of variables that is the one we are looking for.

To implement the above idea the first thing we need is to connect the analysis via Fourier series to the analytic properties of the functions.

Consider the norm

$$|g|_\alpha := \sum_{k\in\mathbb{Z}} e^{\alpha|k|}|\hat{g}_k|.$$

Let us call $\mathcal{B}_\alpha$ the Banach space of the periodic functions (of period one) on $\mathbb{R}$ equipped with the above norm.

Note that, for $\beta < \alpha$,[8]

$$\begin{aligned} |L_\omega^{-1} g|_\beta &\le \sum_{k\in\mathbb{Z}} \frac{|k|}{2C_0} e^{\beta|k|}|\hat{g}_k| \le \frac{|g|_\alpha}{2C_0} \sup_{k\in\mathbb{Z}} |k| e^{-(\alpha-\beta)|k|} \\ &\le \frac{|g|_\alpha}{2eC_0(\alpha-\beta)} \end{aligned} \tag{4.6.7}$$

Thus $L_\omega^{-1} : \mathcal{B}_\alpha \to \mathcal{B}_\beta$ is a bounded operator for each $\alpha > \beta$.

[8]Here we use that, for each $n \in \mathbb{N}$ and $\sigma > 0$,

$$\sup_{k\in N} k^n e^{-\sigma k} \le \sup_{x\in\mathbb{R}_+} x^n e^{-\sigma x} = \left(\frac{n}{\sigma}\right)^n e^{-n} \le e^{-1}\sigma^{-n} n!.$$

The last inequality is an application of Stirling formula. If you do not remember it, here is the baby version used above,

$$n! = e^{\sum_{k=1}^n \ln k} \ge e^{\int_1^n \ln x dx} = e^{n\ln n - n + 1} = n^n e^{-n+1}.$$

The point is that there is a connection between the above Banach spaces, namely we can define $\Xi : \mathbb{B}_\beta \to \mathcal{B}_\alpha$, by $\Xi g(x) = g(x)$, for all $x \in \mathbb{R}$.[9] To see the relation between the norms, let us compute the Fourier coefficients

$$[\widehat{\Xi g}]_k = \frac{1}{i}\int_0^1 e^{2\pi i k x} g(x) dx$$

Problem 4.24 *Show that* $|[\Xi g]_k| \le e^{-\alpha|k|}\|g\|_\alpha$.

Hence, for $\alpha > \beta$, $\|\Xi\|_{\mathbb{B}_\alpha \to \mathcal{B}_\beta} \le 2(1 - e^{\beta-\alpha})^{-1}$. Note also that we can easily define the inverse: if $g \in \mathcal{B}_\alpha$, then define

$$\Xi^{-1} g(z) = \sum_{k\in\mathbb{Z}} e^{2\pi i k z}\hat{g}_k$$

Problem 4.25 *Verify that the above is really the inverse of* Ξ.

If $g \in \mathcal{B}_\alpha$, then

$$\|\Xi^{-1} g\|_\alpha \le \sum_{k\in\mathbb{Z}} e^{|k|\alpha}|\hat{g}_k| = |g|_\alpha.$$

Thus $\|\Xi^{-1}\|_{\mathcal{B}_\alpha \to \mathbb{B}_\alpha} \le 1$.

Problem 4.26 *Show that, for each* $\alpha > \beta$, $\alpha - \beta < 2$, *setting* $h_0 = \Xi^{-1} L_\omega^{-1} \Xi(f - \hat{f}_0)$, *holds*

$$\|h_0\|_\beta \le \frac{4\|f - \hat{f}_0\|_\alpha}{C_0(\alpha-\beta)^2}$$
$$\|h_0'\|_\beta \le \frac{64\pi}{C_0(\alpha-\beta)^3}\|f - \hat{f}_0\|_\alpha.$$

The point of the spaces $\mathbb{B}_\alpha$ is that the equation (4.6.6) for f_1 reads

$$f_1(x) = h_0(x) - h_0(x + \omega + f_1(x)) + f(x + h_0(x)). \tag{4.6.8}$$

To study such equation in $\mathcal{B}_\alpha$ is highly non trivial, while $\mathbb{B}_\alpha$ is much better suited to estimate the norms of composition of functions.

To study (4.6.8) in $\mathbb{B}_\alpha$ the first step is to verify that it makes sense. Obviously one can see it as the restriction to the real axes of an equation involving functions defined on the complex plane, yet it is necessary to check that the composition is well defined, that is we have to carefully analyze domains and ranges of the various functions. For later use we carry out all the needed estimates in the following Lemma.

[9] In other words we simply take the restriction of the function to the real axis.

Lemma 4.6.4 *Given functions $f \in \mathbb{B}_\alpha$ and $h \in \mathbb{B}_\beta$, $\alpha > \beta > \alpha/2$ such that, setting $F(z) = z + \omega + f(z)$, we have $\tau(F) = \omega$, $\|f - \hat{f}_0\|_\alpha \leq \frac{\alpha-\beta}{2\pi}$ and h satisfies (4.6.5), it follows that $\|h\|_\beta \leq \frac{\alpha-\beta}{16\pi}$, $\|h'\|_\beta \leq \frac{1}{16}$, $H(z) = z + h(z)$ is invertible, $H^{-1} \in \mathbb{B}_\gamma$, $\gamma \leq 2\beta - \alpha$, and there exists a function $f_1 \in \mathbb{B}_\gamma$ with $\|f_1 - \int_{S^1} f_1\|_\gamma \leq \frac{1}{2}\|f - \int_{S^1} f\|_\alpha$ satisfying*

$$H^{-1} \circ F \circ H(z) = z + \omega + f_1(z) =: F_1(z).$$

PROOF. First of all H is invertible when restricted to the real axis since $H' \geq \frac{1}{2}$. Let $H^{-1}(z) = z + \psi(z)$, clearly

$$\psi(z) = -h(z + \psi(z)).$$

So the inverse is the fixed point of the operator $K(\psi)(z) = -h(z + \psi(z))$ which is well defined on the set $A = \{\psi \in \mathbb{B}_\gamma \ : \ \|\psi\|_\gamma \leq \frac{\alpha-\beta}{2\pi}\}$. It is easy to verify that such a fixed point exists and is unique.

Note that the function f_1 must satisfy equation (4.6.8). To solve (4.6.8) we must look for a fixed point for the operator

$$\mathcal{K}(\varphi)(z) = h(z) - h(z + \omega + \varphi(z)) + f(z + h(z))$$

on the set $A = \{\varphi \in \mathbb{B}_\gamma \ : \ \|\varphi - \hat{f}_0\|_\gamma \leq \frac{1}{4}\|f - \hat{f}_0\|_\alpha\}$. Note that the composition of functions is well defined, hence so is $\mathcal{K}$.

Let us check that $\mathcal{K}(A) \subset A$.

$$\begin{aligned}\mathcal{K}(\varphi)(z) - \hat{f}_0 &= h(z) - h(z+\omega) + h(z+\omega) - h(z+\omega+\varphi(z)) + f(z+h(z)) - \hat{f}_0 \\ &= f(z+h(z)) - f(z) + h(z+\omega) - h(z+\omega+\varphi(z)).\end{aligned}$$

Thus, using the estimate in Problem 4.35 and recalling (4.6.2),

$$\|\mathcal{K}(\varphi) - \hat{f}_0\|_\gamma \leq \|f'\|_\beta \|h\|_\gamma + \|h'\|_\beta \|\varphi\|_\gamma \leq \frac{1}{8}\|f\|_\alpha + \frac{1}{16}\|\varphi - \hat{f}_0\|_\gamma + \frac{1}{16}|\hat{f}_0| \leq \frac{1}{4}\|f\|_\alpha.$$

In addition, if $\varphi, \tilde{\varphi} \in A$, then

$$\|\mathcal{K}(\varphi) - \mathcal{K}(\tilde{\varphi})\|_\gamma \leq \|h'\|_\beta \|\varphi - \tilde{\varphi}\|_\gamma \leq \frac{1}{16}\|\varphi - \tilde{\varphi}\|_\gamma.$$

Thus, by the usual contraction argument, there exists $f_1 \in A$ such that $\mathcal{K}(f_1) = f_1$. On the other hand F_1 is conjugated to F and hence it has rotation number ω. Thus (4.6.3) implies $|\int_{S^1} f_1| \leq \|f_1 - \int_{S^1} f_1\|_\gamma$ and

$$\left\| f_1 - \int_{S^1} f_1 \right\|_\gamma \leq \left\| f_1 - \int_{S^1} f_0 \right\|_\gamma + \left| \int_{S^1} f_0 - f_1 \right| \leq \frac{1}{2}\|f - \hat{f}_0\|_\alpha.$$

□

Since we need to restrict the domain several time it is convenient to do it in a systematic fashion. Let $\rho_k := e^{-k\tau}\alpha$, and apply Lemma 4.6.4 with $\beta = \rho_2$ and $\gamma = \rho_4$. A simple computation shows that the condition on β, γ are satisfied if $e^{-\tau} \geq \frac{2}{3}$. Then, setting $\varepsilon = \|f - \hat{f}_0\|_\alpha$, Lemma 4.6.4 applies provided $\varepsilon \leq \min\{\frac{\tau\alpha}{\pi e}, \frac{C_0\tau^3\alpha^3}{128e^3\pi}\}$.[10] We then choose $\tau_0 = \alpha^{-1}C_0^{-\frac{1}{3}}\varepsilon^{\frac{1}{3}}$. Hence $\min\{\frac{\tau\alpha}{\pi e}, \frac{C_0\tau^3\alpha^3}{128e^3\pi}\} = \frac{C_0\tau^3\alpha^3}{128e^3\pi}$ provided $\varepsilon \leq 10^3e^3\sqrt{C_0}$.

We now implement an iterative procedure by setting: $f_0 = f$,

$$h_n(z+\omega) - h_n(z) = f_n(z)\,, \quad H_n(z) = z + h_n(z)\,, \quad F_n(z) = z + \omega + f_n(z),$$
$$H_n^{-1} \circ F_n \circ H_n(z) = z + \omega + f_{n+1}(z).$$

In addition, we set $\alpha_0 = \alpha$, $\alpha_{n+1} = e^{-4\tau_n}\alpha_n$, $\varepsilon_{n+1} = \frac{\varepsilon_n}{2}$ and $\tau_n = \alpha_n^{-1}C_0^{-\frac{1}{3}}\varepsilon_n^{\frac{1}{3}}$. Note that this choices imply that Lemma 4.6.4 can be applied at each stage of the iteration. Now, if $\alpha_n \geq \frac{1}{2}\alpha_0$, holds $\varepsilon_n = 2^{-n}\varepsilon$, $\tau_n \leq 2\alpha_0^{-1}2^{-n/3}C_0^{-\frac{1}{3}}\varepsilon^{\frac{1}{3}}$. This implies $\alpha_n = \alpha_0 e^{-4\sum_{k=0}^{n-1}\tau_k} \geq e^{-40\alpha_0^{-1}C_0^{-\frac{1}{3}}\varepsilon^{\frac{1}{3}}}\alpha_0$ which is always larger than $\alpha_0/2$ provided $\varepsilon \leq C_0\left[\frac{\alpha_0 \ln 2}{40}\right]^3$. Note that all our condition on ε are satisfied if $\varepsilon \leq \frac{1}{5}C_0\alpha_0^3 10^{-5}$.

We have thus a sequence of changes of variables $H_n(z) = z + h_n(z)$, the next question is if it exists $H(z) = \lim_{n\to\infty} H_0 \circ H_1 \circ \cdots \circ H_n(z)$. It suffices to prove that the sequence is uniformly bounded on $D_{\alpha_0/2}$

$$\begin{aligned}|H_0 \circ H_1 \circ \cdots \circ H_n(z) - z| &\leq \sum_{k=0}^{n} \|h_k\|_{\alpha_k} \leq \sum_{k=0}^{n} \frac{e^2\varepsilon_k}{C_0\tau_k^2\alpha_k^2} \\ &\leq \sum_{k=0}^{\infty} 2^{-k/3}\varepsilon^{\frac{1}{3}}e^2C_0^{-\frac{1}{3}} \leq \varepsilon^{\frac{1}{3}}5e^2C_0^{-\frac{1}{3}}\end{aligned}$$

Similarly it follows that the H_n form a Chauchy sequence, hence they have a limit $H \in \mathbb{B}_{\alpha_0/2}$ with $\|\mathrm{id} - H\|_{\alpha_0/2} \leq \varepsilon^{\frac{1}{3}}5e^2C_0^{-\frac{1}{3}}$. From this it follows also (see Problem 4.35)

$$\|1 - H'\|_{\alpha_0/4} \leq \frac{40\pi e^2\varepsilon^{\frac{1}{3}}}{\alpha_0 C_0^{\frac{1}{3}}} \leq \frac{1}{2}, \tag{4.6.9}$$

provided $\varepsilon \leq 10^{-10}C_0\alpha^3$. Hence H is invertible and this concludes the proof. □

[10] Just use Problem 4.26 and the fact that $1 - e^{-x} = \int_0^x e^{-y}dy \geq e^{-1}x$, for $x \in (0,1)$, to check the hypotheses of the Lemma.

4.6.2 Smooth KAM theory

The final question is: do similar results hold assuming less smoothness? The answer is yes, yet to explore optimal results it is not an easy task. Here we content ourselves with a partial, but enlightening, result.

Theorem 4.6.5 *For each $r > 4$,[11] if $\tau(F) = \omega$, $\|f - \hat{f}_0\|_{\mathcal{C}^r} \leq 10^{-17} C_0 (r-4)^9$ and $\omega > 0$ satisfies*

$$\left| \omega - \frac{p}{q} \right| \geq \frac{C_0}{q^2}$$

for each $p, q \in \mathbb{N}$, then there exists $\mathfrak{h} \in \mathcal{C}^1$ such that, setting $\mathcal{H}(x) = x + \mathfrak{h}(x)$, $\mathcal{H}$ is invertible and

$$\mathcal{H}^{-1} \circ F \circ \mathcal{H}(x) = x + \omega.$$

PROOF. The basic idea is to write $f = \hat{f}_0 + \sum_{m=0}^{\infty} \tilde{f}_m$ where

$$\tilde{f}_m(x) = \sum_{e^{am} \leq |k| < e^{a(m+1)}} \hat{f}_k e^{2\pi i k x}$$

and $a > 1$ is a parameter to be chosen later.[12] Then one can apply Theorem 4.6.1 one $\tilde{f}_m$ at a time. Indeed, let $\alpha_m = b(m+1)e^{-a(m+1)}$, for some $a, b > 0$ to be chosen later, where then

$$\begin{aligned} \|\tilde{f}_m\|_{\alpha_m} &\leq \sum_{e^{am} \leq |k| < e^{a(m+1)}} |\hat{f}_k| e^{\alpha_m |k|} \leq \sum_{e^{am} \leq |k| < e^{a(m+1)}} |f|_{\mathcal{C}^r} (2\pi)^{-r} |k|^{-r} e^{\alpha_m |k|} \\ &\leq \sum_{e^{am} \leq k < e^{a(m+1)}} 2|f|_{\mathcal{C}^r} (2\pi)^{-r} e^{-ram} e^{b(m+1)} \\ &\leq 2|f|_{\mathcal{C}^r} e^{-(ar-a-b)m+a}. \end{aligned}$$

If $|f|_{\mathcal{C}^r}$ is small enough, we can apply Theorem 4.6.1 to $\tilde{f}_0$. Indeed, let $\tilde{F}_0(z) = z + \omega + \xi_0 + \tilde{f}_0(z)$, then $\xi_0 - \|\tilde{f}_0\|_\infty \leq \tau(F_0) - \omega \leq \xi_0 + \|\tilde{f}_0\|_\infty$, so there exists $|\xi_0| \leq \|\tilde{f}_0\|_\infty$ such that $\tau(F_0) = \omega$. Hence, there exist $\tilde{h}_0$ such that, setting $\tilde{H}_0(z) = z + \tilde{h}_0(z)$ and $\tilde{F}_0(z) = z + \xi_0 + \tilde{f}_0(z)$,

$$\tilde{H}_0^{-1} \circ \tilde{F}_0 \circ \tilde{H}_0(z) = z + \omega =: R_\omega(z).$$

The obvious next step is to compute $\mathbb{f}_1$ such that, for each $n \in \mathbb{N}$,

$$\tilde{H}_0^{-1} \circ \left(R_\omega + \sum_{k=0}^{1} \tilde{f}_k \right) \circ \tilde{H}_0(z) = z + \omega + \mathbb{f}_1(z).$$

[11] In fact, by a more sophisticated proof, $r > 3$ suffices [Her83].

[12] This choice (a la Panley Wiener) for the decomposition of f is not optimal, yet it makes the latter computations simpler.

This is possible if $|f|_{\mathcal{C}^r}$ is small enough. We can then try to iterate the above procedure by applying Theorem 4.6.1 to $\mathbb{f}_1$ and so on.

To this end we set up the following iterative scheme: $\mathbb{f}_0 = \tilde{f}_0$, $\mathcal{H}_{-1} = \mathrm{id}$. For $k \in \mathbb{N}_0$ let $F_k(z) = z + \omega + \varsigma_k + \hat{f}_0 + \sum_{j=0}^{k} \tilde{f}_j(z)$, $\tau(F_k) = \omega$, $\int_{S^1} \mathbb{f}_k = 0$

$$\mathbb{F}_k(z) = z + \omega + \xi_k + \mathbb{f}_k(z)\,; \quad \tau(\mathbb{F}_k) = \omega \tag{4.6.10}$$

$$H_k^{-1} \circ \mathbb{F}_k \circ H_k(z) = z + \omega\,; \quad H_k(z) = z + h_k(z) \tag{4.6.11}$$

$$\mathcal{H}_k = \mathcal{H}_{k-1} \circ H_k \tag{4.6.12}$$

$$\mathcal{H}_k^{-1} \circ F_{k+1} \circ \mathcal{H}_k = \mathbb{F}_{k+1}. \tag{4.6.13}$$

Note that, for each $k \in \mathbb{N}_0$,

$$\begin{aligned} \mathcal{H}_k^{-1} \circ F_k \circ \mathcal{H}_k(z) &= H_k^{-1} \circ \mathcal{H}_{k-1}^{-1} \circ F_k \circ \mathcal{H}_{k-1} \circ H_k(z) \\ &= H_k^{-1} \circ \mathbb{F}_k \circ H_k = R_\omega. \end{aligned} \tag{4.6.14}$$

The rest of the proof consists in a rather tedious verification that the induction is well posed and in estimating the norms of the objects involved.

Let us assume by induction that there exists $B > 1$ such that, for each $k \in \mathbb{N}$ and $j < k$, $\|\mathbb{f}_j\|_{\alpha_j/2} \le B\|\tilde{f}_j\|_{\alpha_j}$. In addition, we write $\mathcal{H}_k(z) = z + \mathfrak{h}_k(z)$ and, setting $3\delta := a(r-4) - b$, assume that

$$\|\mathfrak{h}_{k-1}\|_{\alpha_{k-1}/4} \le 10^{-3} \sum_{j=0}^{k-1} e^{-\delta j}\alpha_j =: 10^{-3} A_{k-1}$$

$$\|\mathfrak{h}'_{k-1}\|_{\alpha_k/8} \le \frac{1}{4} - \frac{1}{2k+1}.$$

Note that this is obviously true for $k = 0$. Remark that Theorem 4.6.1 implies that there exists a solution $h_k \in \mathbb{B}_{\alpha_k/4}$ to (4.6.11) provided $\|\mathbb{f}_k\|_{\alpha_k/2} \le C_\star \alpha_k^3$, with $C_\star = C_0 10^{-11}$. Under the above hypotheses,

$$\begin{aligned} \|\mathbb{f}_k\|_{\alpha_k/2} &\le B\|\tilde{f}_k\|_{\alpha_k} \le 2B|f|_{\mathcal{C}^r} e^{-(ar-a-b)k+a} \\ &\le 2B|f|_{\mathcal{C}^r} b^{-3}(k+1)^{-3} e^{-3\delta k + 4a}\alpha_k^3 \le C_\star \delta^6 e^{-3\delta k}\alpha_k^3 \le C_\star \delta^6 e^{-3\delta k}\alpha_k^3 \end{aligned}$$

provided $\delta > 0$ and $|f|_{\mathcal{C}^r} \le \frac{1}{2} C_\star B^{-1} b^3 e^{-4a}\delta^6$. Thus, by Theorem 4.6.1,

$$\|h_k\|_{\alpha_k/4} \le 3C_0^{-\frac{1}{3}} \|\mathbb{f}_k\|_{\alpha_{k/2}}^{\frac{1}{3}} \le 3C_0^{-\frac{1}{3}} C_\star^{\frac{1}{3}} \delta^2 e^{-\delta k}\alpha_k \le 10^{-3}\delta^2 e^{-\delta k}\alpha_k.$$

Moreover (see Problem 4.35)

$$\|h'_k\|_{\alpha_k/8} \le 16\pi \|h_k\|_{\alpha_k/4}\alpha_k^{-1} \le 4 \cdot 10^{-2}\delta^2 e^{-\delta k} < 1/4.$$

By (4.6.12) it follows

$$\mathfrak{h}_k(z) = h_k(z) + \mathfrak{h}_{k-1}(z + h_k(z)),$$

which is well posed in $\mathbb{B}_{\alpha_k/4}$ provided $a \geq 2$, since this implies that $\alpha_k(1+4\cdot 10^{-3}\delta^2 e^{-\delta k}) \leq \alpha_{k-1}$. Moreover $\|\mathfrak{h}_k\|_{\alpha_k/4} \leq 10^{-3}A_k$

$$\|\mathfrak{h}_k'\|_{\alpha_k/8} \leq \left[\frac{1}{4} - \frac{1}{2k+1}\right](1+4\cdot 10^{-2}\delta^2 e^{-\delta k}) + 4\cdot 10^{-2}\delta^2 e^{-\delta k} \leq \frac{1}{4} - \frac{1}{2k+2}.$$

Equation (4.6.13), also recalling (4.6.14), is equivalent to

$$\begin{aligned}
\tilde{\mathfrak{f}}_{k+1}(z) &= \mathfrak{f}_{k+1}(z) + \xi_{k+1} \\
\tilde{\mathfrak{f}}_{k+1}(z) &= \varsigma_{k+1} - \varsigma_k + \mathfrak{h}_k(z+\omega) - \mathfrak{h}_k(z+\omega+\tilde{\mathfrak{f}}_{k+1}(z)) \\
&\quad + \tilde{f}_{k+1}(x+\mathfrak{h}_k(x)).
\end{aligned} \tag{4.6.15}$$

Since $\mathcal{H}_k$ is invertible this implies that $\tilde{\mathfrak{f}}_{k+1}$ is well defined on the real line. This implies that

$$\begin{aligned}
\mathbb{F}_{k+1}(x) =& x+\omega+\varsigma_{k+1}-\varsigma_k+\mathfrak{h}_k(x+\omega)-\mathfrak{h}_k(x+\omega+\tilde{\mathfrak{f}}_{k+1}(x))+\tilde{f}_{k+1}(x+\mathfrak{h}_k(x)) \\
=:& x+\omega+g(x).
\end{aligned}$$

By induction it follows that, for all $q \in \mathbb{N}$,

$$\mathbb{F}_{k+1}^q(z) = z + n\omega + \sum_{j=0}^{n-1} g(\mathbb{F}_{k+1}^j(x)).$$

As usual remark that $\mathbb{F}_{k+1}^q(z) - z$ cannot be an integer since otherwise we would have a periodic point and we would have $\tau(F_{k+1}) = \frac{p}{q} \in \mathbb{Q}$, contrary to the hypothesis. It follows that for each $q \in \mathbb{N}$ there exists $p \in \mathbb{N}$ such that, for all $x \in \mathbb{R}$,

$$p - 1 \leq x + q\omega + \sum_{j=0}^{q-1} g(\mathbb{F}_{k+1}^j(x)) \leq p.$$

Since, $\tau(F) = \omega$ it follows

$$\left|\frac{1}{q}\sum_{j=0}^{q-1} g(\mathbb{F}_{k+1}^j(x))\right| \leq \frac{1}{q}$$

hence, by the arbitrariness of q,

$$|\varsigma_{k+1} - \varsigma_k| \leq \frac{1}{4}\|\tilde{\mathfrak{f}}_{k+1}\|_\infty + \|\tilde{f}_{k+1}\|_\infty$$

Using the above estimate in equation (4.6.15) yields $\|\mathfrak{f}_{k+1}\|_\infty \leq 4\|\tilde{f}_{k+1}\|_\infty$, hence

$$|\varsigma_{k+1} - \varsigma_k| \leq 2\|\tilde{f}_{k+1}\|_\infty. \tag{4.6.16}$$

To obtain an estimate of the $\|\cdot\|_{\alpha_{k+1}}$ norm of $\tilde{\mathfrak{f}}_{k+1}$ from equation (4.6.15) we consider the operator $\mathcal{K} : D \to \mathbb{B}_{\alpha_{k+1}/2}$, where

$$D = \{\varphi \in \mathbb{B}_{\alpha_{k+1}/2} \ : \ \|\varphi\|_{\alpha_{k+1}/2} \leq \frac{1}{2} B \|\tilde{f}_{k+1}\|_{\alpha_{k+1}}\},$$

defined by

$$\mathcal{K}(\varphi) = \varsigma_{k+1} - \varsigma_k + \mathfrak{h}_k(z+\omega) - \mathfrak{h}_k(z+\omega+\varphi(z)) + \tilde{f}_{k+1}(x+\mathfrak{h}_k(x)).$$

The operator is well defined if $e^{-a} \leq \frac{1}{4}$ and $|f|_{\mathcal{C}^r} \leq e^{-2a}\frac{b}{4B}$. Moreover $\mathcal{K}(D) \subset D$ provided $B \geq 8$. By the usual contraction theorem it follows $\|\tilde{\mathfrak{f}}_{k+1}\|_{\alpha_{k+1}/2} \leq \frac{1}{2} B \|\tilde{f}_{k+1}\|_{\alpha_{k+1}}$. Thus $\|\mathfrak{f}_{k+1}\|_{\alpha_{k+1}/2} \leq \|\tilde{\mathfrak{f}}_{k+1} - \int_{S^1} \tilde{\mathfrak{f}}_{k+1}\|_{\alpha_{k+1}/2} \leq B\|\tilde{f}_{k+1}\|_{\alpha_{k+1}}$, whereby concluding the induction.

The last thing we must prove is that the change of coordinate $\mathcal{H}_n$ is convergent. Note that

$$|\mathcal{H}_n'(x)| \leq \prod_{k=0}^{n} \|\tilde{H}_k'\|_{\frac{\alpha_k}{8}} \leq \prod_{k=0}^{n} e^{4\cdot 10^{-2}\delta^2 e^{-\delta k}} \leq e^{4\cdot 10^2 \delta}.$$

It is then easy to see that the $\mathcal{H}_n$ form a Chauchy sequence in $\mathcal{C}^1$. The theorem follows by collecting all the above inequalities and setting $B = 8$, $a = 2$, $b = (r-4)/3$ and recalling the condition $|f|_{\mathcal{C}^r} \leq \frac{1}{2} C_\star B^{-1} b^3 e^{-4a} \delta^6$. □

Problems

4.27. If M is a $\mathcal{C}^r$ manifold, $f \in \mathcal{C}^r(M, M)$ is a diffeomorphism and $\tau \in \mathcal{C}^r(M, (0, \infty))$, show that the associated suspension flow is defined on a $\mathcal{C}^r$ manifold and is $\mathcal{C}^r$.

4.28. Consider the Dynamical System $([0,1], T)$ where

$$T(x) = \frac{1}{x} - \left\lfloor \frac{1}{x} \right\rfloor = \frac{1}{x} \mod 1$$

($\lfloor a \rfloor$ is the integer part of a). This is called the *Gauss map*. Prove that for each $x \in \mathbb{Q} \cap [0,1]$ holds $\lim_{n\to\infty} T^n(x) = 0$.

4.29. Prove that any infinite continuous fraction of the form

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ddots}}}$$

with $a_i \in \mathbb{N}$ defines a real number.

4.30. Prove that, for each $a \in \mathbb{N}$,

$$x = \cfrac{1}{a + \cfrac{1}{a + \cfrac{1}{a + \ddots}}} = \frac{-a + \sqrt{a^2+4}}{2}.$$

4.31. Prove that, for all $s > 2$, for Lebesgue almost all numbers $x \in [0,1]$ there exists $C > 0$ such that[13]

$$\left| x - \frac{p}{q} \right| \geq \frac{C}{q^s}$$

for all $p, q \in \mathbb{N}$.

4.32. Let $f_a(x) = \frac{1}{a+x}$. Given a sequence $[a_0, a_1, \ldots, a_n]$ show that

$$f_{a_0} \circ \cdots \circ f_{a_n}(x) = \cfrac{1}{a_0 + \cfrac{1}{a_1 + \ddots \cfrac{1}{a_n + x}}} = \frac{p_n + p_{n-1}x}{q_n + q_{n-1}x},$$

where $p_{n+1} = a_{n+1}p_n + p_{n-1}$ and $q_{n+1} = a_{n+1}q_n + q_{n-1}$, $p_{-1} = 0$, $q_{-1} = 1$, $p_0 = 1$, $q_0 = a_0$. In addition, show that, for all $n \in \mathbb{N}$, $p_n q_{n-1} - q_n p_{n-1} = (-1)^n$ and decude that p_n, q_n have no common divisor different from one. Finally, verify that

$$f_{a_0} \circ \cdots \circ f_{a_n}(x) - f_{a_0} \circ \cdots \circ f_{a_{n+1}}(x) = \frac{(-1)^{n+1}[x^2 + a_{n+1}x - 1]}{(q_n + q_{n-1}x)(q_{n+1} + q_n x)}.$$

4.33. Let $\omega \in [0,1)$. Show that there exists infinitely many $p, q \in \mathbb{N}$ such that

$$\left| \omega - \frac{p}{q} \right| \leq \frac{1}{q^2}.$$

4.34. Let $\omega \in [0,1)$ have the continuous fraction expansion given by $[a_0, a_1 \ldots]$. Suppose that $\inf_n a_n > 0$ and $\sup_n a_n < \infty$.[14] Show that there exists a constant $c > 0$ such that for all $p, q \in \mathbb{N}$

$$\left| \omega - \frac{p}{q} \right| \geq \frac{c}{q^2}.$$

[13] The composition below is often called *iterated function system*, it can be naturally viewed as a time dependent dynamical system.

[14] Such numbers ω are called of *constant type*.

4.35. For each $\varphi \in \mathbb{B}_\alpha$ and $\beta < \alpha$ show that $\|\varphi'\|_\beta \leq \frac{2\pi\|\varphi\|_\alpha}{\alpha-\beta}$.

4.36. Let us consider an holomorphic function $f : U \subset \mathbb{C} \to \mathbb{C}$ where U is an open set containing zero. Assume that $f(0) = 0, f'(0) = e^{2\pi i\omega}$. Prove that, if ω is Diophantine, then it is possible to find an open set $D \subset U$ on which f is conjugated to the map $f_\omega(z) = e^{2\pi i\omega}z$.

Hints to solving the Problems

4.2 Consider a system $([0,1], T)$ such that T is piecewise linear, it has an unstable fixed point at x_0 and an attracting fixed point at $z \in (0, x_0)$ so that the set $[z, x_0]$ is forward invariant. Finally arrange so that $T(0) = x_0$ and $T(x) \leq x_0$ for x near zero.

4.10 The equation $\dot{x} = \omega = (\omega_1, \omega_2)$ on $\mathbb{T}^2$ has the solution $x(t) = (x_1(t), x_2(t)) = x_0 + \omega t \mod 1$. If one looks at the flow only at the times $\tau_n = n\omega_1^{-1}$, then $x(n\tau) = x_0 + (0, \alpha n) \mod 1$ where $\alpha := \frac{\omega_2}{\omega_1}$. One can then consider the circle map $f : S^1 \to S^1$ defined by $f(z) = z + \alpha \mod 1$. Clearly, if the orbits of such a map are dense in S^1 the original flow will be dense in $\mathbb{T}^2$. The density follows in the case $\alpha \notin \mathbb{Q}$. In fact this implies that f has no periodic orbits. Then $\{f^n(0)\}$ is made of distinct points and contains a converging subsequence (by compactness) hence for each $\varepsilon > 0$ exists $\bar{n} \in \mathbb{N}$ such that $|z - f^{\bar{n}}(z)| \leq \varepsilon$, that is $f^{\bar{n}}$ is a rotation by less than ε. Hence the orbit $\{f^{k\bar{n}}(z)\}$ enters in the ε-neighborhood of each point of S^1.

4.11 By assumption the return time τ is well defined for all $x \in \gamma$. We want thus to solve the equation $\phi_t(\gamma(s)) = \gamma(r)$ with $s, r \in S^1$, $t \in \mathbb{R}$. Thus, if we set $F(s, r, \tau) = \phi_t(\gamma(s)) - \gamma(r)$, we are reduced to solve the equation $F = 0$. To do so we can apply the implicit function theorem B.1.1. Note that, by Theorem 1.1.14 and the assumptions, $F \in \mathcal{C}^r$. Since, by hypothesis, there exists $\bar{r}$ such that $\phi_{\tau(x)}(\gamma(\bar{s})) = \gamma(\bar{r})$, where $\gamma(\bar{s}) = x$, we have $F(\bar{s}, \tau(x), \bar{r}) = 0$. To apply Theorem B.1.1 we need that $\partial_{r,s}F$ be invertible, thi is true since

$$\partial_{r,s}F = \begin{pmatrix} -\gamma'(r) & V(\phi_\tau(\gamma(s))) \end{pmatrix}$$

and the section is transversal to the flow. Hence, the result follows from Problem B.2.

4.13 Suppose that there exists $\varphi(r, s)$, $\varphi \in \mathcal{C}^0([0,1] \times \mathbb{T}^1, \mathbb{R})$, such that $\varphi(1, \cdot)$ is a parametrization of Γ and $\varphi(0, s) = y$ for some fixed $y \in \mathbb{T}^2$ (i.e. Γ is homotopic to y).

4.12 First of all notice that if $\xi(t)$ is the derivative with respect to the initial condition and $\xi(0) = \lambda V(x(0))$, for some λ, then $\xi(t) = \lambda V(x(t))$ for all t. Define then $\omega(x, y) = x_1 y_2 - x_2 y_1$ and verify that $x, y \neq 0$ and $\omega(x, y) = 0$ imply that there exists $\lambda \in \mathbb{R}$ such that $x = \lambda y$.[15] This means that $\omega(\xi(t), V(x(t)))$ cannot change sign. Hence the result.

4.17 Let $\liminf_{n\to\infty} \frac{a_n}{n} = a > -\infty$, then for each $\varepsilon, m > 0$ exists $\bar{n} \in \mathbb{N}$, $\bar{n} > m$, such that $|a_{\bar{n}} - a\bar{n}| \leq \varepsilon\bar{n}$. Let $l \in \mathbb{N}$, $l > \bar{n}$, and write $l = k\bar{n} + r$, $r < \bar{n}$, then

$$\begin{aligned} a - \varepsilon \leq \frac{a_l}{l} &\leq \frac{ka_{\bar{n}} + kL + a_r}{l} \leq \frac{k\bar{n}(a+\varepsilon) + kL + a_r}{l} \\ &= a + \varepsilon + \frac{L}{m} + \frac{a_r}{l}. \end{aligned}$$

From which the claim follows.

4.18 Stetting $I = [a, b]$ note that $g(x) = f(x) - x$ has a zero in I.

4.19 This is the same than saying $\bigcup_{k\in\mathbb{N}} f^{-k}[x, f^n(x)] = S^1$. Argue by contradiction. Consider $f^{-kn}[x, f^n(x)]$, this are contiguous intervals. If they do not cover all S^1, then their length must go to zero. Choose a subsequence $f^{-k_j n}x$ which has a limit, call it z. Then

$$z = \lim_{j\to\infty} f^{-k_j n}(x) = \lim_{j\to\infty} f^{-k_j n}(f^n(x)) = \lim_{j\to\infty} f^n(f^{-k_j n}(x)) = f^n(z).$$

Hence f must have a periodic point contradicting $\tau(f) \notin \mathbb{Q}$.

4.20 Since $\tau(f) \notin \mathbb{Q}$, for each $x \in S^1$, $f^k(x) \neq f^j(x)$ for all $k \neq j \in \mathbb{Z}$. By compactness $\{f^k(x)\}_{k\in\mathbb{N}}$ has accomulation points, let z be one such point. Consider the subseguence $\{f_{n_j}(x)\}_{j\in\mathbb{N}}$, such that $|f^{n_j}(x) - z| \leq |f^k(x) - z|$ for all $k < n_{j+1}$ and $|f^{n_{j+1}}(x) - z| < |f^{n_j}(x) - z|$. Then $f^k(x) \notin [x, f^{n_{j+1}-n_j}(x)]$ for all $k \leq n_{j+1} - n_j$, otherwise there would exists $l \leq n_{j+1}$ such that $f^l(x) \in [f^{n_j}(x), f^{n_{j+1}}(x)]$, but this would imply that $f^l(x)$ is closer to z than $f^{n_j}(x)$ which is not possible by the defintion of n_j.

4.26 For the second inequality use Problem 4.35.

4.28 If $x = \frac{p_0}{q_0}$, $p_0 \leq q_0$, then $q_0 = k_1 p_0 + p_1$, with $p_1 < p_0$, and $T(x) = \frac{p_1}{p_0}$. Let $q_1 = p_0$ and go on noticing that $p_{i+1} < p_i$.[16]

[15] By the way, ω is a symplectic form and its existence implies that the manifold is orientable.

[16] This is nothing else that the *Euclidean algorithm* to find the greatest common divisor of two integers [Euc78, Elements, Book VII, Proposition 1 and 2]. The greatest common

4.29 Note that if you fix the first n $\{a_i\}$, this corresponds to specifying which elements of the partition $\{[\frac{1}{i+1}, \frac{1}{i}]\}$ are visited by the trajectory of $\{T^i x\}$, T being the Gauss map. By the expansivity of the map readily follows that x must belong to an interval of size λ^{-n} for some $\lambda > 1$.

4.30 Note that $T(x) = x$, where T is the Gauss map. Study periodic continuous fractions of period two.

4.31 To see it consider the sets $I_{p,q} := [\frac{p}{q} - Cq^{-s}, \frac{p}{q} - Cq^{-s}]$. If $p \leq q$, then $I_{p,q} \subset [0,1]$. Clearly if $\alpha \notin I_{p,p}$ for all $q \geq p \in \mathbb{N}$, then α satisfies the Diophantine condition. But $\sum_{q \geq p} |I_{p,q}| \leq C \sum_{q=1}^{\infty} q^{-s+1}$ which converges provided $s > 2$ and can be made arbitrarily small by choosing C small. Accordingly, almost all numbers are Diophantine for any $s > 2$.

4.32 By induction.

4.33 The result is trivial for rational numbers. By Problem 4.29, $\omega = \lim_{n\to\infty} f_{a_0} \circ \cdots \circ f_{a_n}(0)$. Moreover, $f_a([0,\infty)) \subset [0, a^{-1}]$. Thus for each $n \in \mathbb{N}$ there exists $x_n \in [0, a_{n+1}^{-1}]$ such that $\omega = f_{a_0} \circ \cdots \circ f_{a_n}(x_n)$. Thus, be the monotonicity of the f_a it follows that either $\omega \in [f_{a_0} \circ \cdots \circ f_{a_n}(0), f_{a_0} \circ \cdots \circ f_{a_{n+1}}(0)]$ or $\omega \in [f_{a_0} \circ \cdots \circ f_{a_{n+1}}(0), f_{a_0} \circ \cdots \circ f_{a_n}(0)]$. One can then use the equalities of Problem 4.32 to conclude all the rationals $f_{a_0} \circ \cdots \circ f_{a_n}(0)$ satisfy

$$|\omega - f_{a_0} \circ \cdots \circ f_{a_n}(0)| \leq \frac{1}{a_{n+1} q_n^2}.$$

You did not like this argument? Here is an interesting alternative. Problem 4.32 implies that

$$f_{a_0} \circ \cdots \circ f_{a_n}(0) = \sum_{k=0}^{n} \frac{(-1)^k}{q_k q_{k-1}}.$$

Since the odd and even partial sum of an alternating series form monotone sequences that converge to the limit from opposite sides, it follows

divisor is clearly the last non-zero p_i. This provides also a remarkable way of writing rational numbers: *continuous fractions*

$$\frac{p_0}{q_0} = \cfrac{1}{k_1 + \cfrac{1}{k_2 + \ddots + \cfrac{1}{k_n}}}.$$

that

$$|\omega - f_{a_0} \circ \cdots \circ f_{a_n}(0)| \leq |f_{a_0} \circ \cdots \circ f_{a_n}(0) - f_{a_0} \circ \cdots \circ f_{a_{n+1}}(0)| \leq \frac{1}{a_{n+1}q_n^2}.$$

4.34 As we have argued at the end of the hint of Problem 4.33, $\omega \in [f_{a_0} \circ \cdots \circ f_{a_n}(0), f_{a_0} \circ \cdots \circ f_{a_{n+1}}(0)] =: I_n$. Note that if $q < q_n$ then

$$\left|\frac{p}{q} - \frac{p_n}{q_n}\right| \geq \frac{1}{q_n q}\ ; \quad \left|\frac{p}{q} - \frac{p_n}{q_n}\right| \geq \frac{1}{q_{n+1}q}.$$

But $|I_n| = \frac{1}{q_n q_{n+1}}$ so it cannot contain any rational number with denominator strictly less than q_n. Accordingly, $\frac{p}{q} \notin I_n$ and thus $|\omega - \frac{p}{q}| \geq \frac{1}{q_{n+1}q} > \frac{1}{q_{n+1}q_n}$. In other words the fraction determined by $[a_0, \ldots, a_n]$ are the best approximation of ω among all the numbers with denominator smaller than q_n. Since,

$$|\omega - f_{a_0} \circ \cdots \circ f_{a_n}(0)| \geq |f_{a_0} \circ \cdots \circ f_{a_n}(0) - f_{a_0} \circ \cdots \circ f_{a_{n+2}}(0)| \geq \frac{1}{(a_{n+1}+2)q_n^2}.$$

the result follows by simple computations.

4.35 Since φ is holomorphic by Rienmann formula we have

$$\varphi'(z) = \frac{1}{2\pi i}\int_\gamma \frac{\varphi(\zeta)}{(z-\zeta)^2}d\zeta$$

where γ is a simple closed curve in D_α surrounding $z \in D_\beta$. For γ we chose the curve $\{z + \frac{\alpha-\beta}{2\pi}e^{i\theta}\}_{\theta\in[0,2\pi]}$. Hence

$$\|\varphi'\|_\beta \leq \frac{1}{2\pi}\int_0^{2\pi} \frac{2\pi|\varphi|_\alpha}{\alpha-\beta}d\theta = \frac{2\pi|\varphi|_\alpha}{\alpha-\beta}.$$

4.36 Mimic Theorem 4.6.1.

Notes

Lemma 4.3.2 is due to Siegel [Sie45], see [NZ99] for a detailed treatment of flows on surfaces. A detailed treatment of circle rotations can be found in [Her83, Her86]. A general treatment of KAM theory for Hamiltonian Systems, with an emphasis on concrete applications, can be found in [CC95].

Chapter 5

Global behavior: more stuff is out there

Every *Dynamical System* studied so far exhibited fairly simple motions, allowing for a detailed understanding of its behavior. Yet, we have not yet addressed the problem of *long time predictions* in systems with more than two dimensions.

Although this is not the proper occasion for a historical excursus, it is worthwhile to stress that the first Dynamical Systems were widely investigated have been the planetary motions. Not surprisingly, the main emphasis in such investigations was accurate prediction of future positions. Nevertheless, exactly from the effort of accurately predicting future motions stemmed the consciousness of the existence of very serious obstructions to such a program. Specifically, in the work of Poincaré [Poi87] appeared for the first time the phenomena of instability with respect to initial conditions, a central concept in the understanding of modern Dynamical Systems. In fact, we will see briefly that such Instability phenomena can already be observed in very simple systems–such as a periodically forced pendulum–that exhibit a so called "homoclinic tangle" [Mos01, PT93].

The realization that many relevant systems are very sensitive with respect to the initial conditions dealt a strong blow to the idea that it is always possible to predict the future behavior of a system,[1] yet the work of many physicists

[1] Without going to the extreme of some authors of the eighteenth century arguing that, given the present state of the universe, a sufficiently powerful mind (maybe God) could predict all the future. Think, more modestly, of an isolated system and imagine to use some numerical scheme to try to solve the equations of motion for an arbitrarily long time with an arbitrary precision.

(and we must mention at least Boltzmann) and mathematicians (in particular, the so called *Russian School* with people like Kolmogorov, Anosov, Sinai, but also some western mathematicians, like Birkhoff, Smale, Ruelle and Bowen, gave important contributions) led to the understanding that, although precise predictions were not possible, it was possible and, at times, even easy to make statistical predictions. The concept of statistical properties of a Dynamical System will be addressed in the following chapters. This chapter is dedicated to making precise, in a simple example, the nature of the above mentioned instability.

5.1 A pendulum–The model and a question

We will study a seemingly trivial example: a forced pendulum. To be more concrete, let us imagine a pendulum of length $l = 1$ meter, mass $m = 1$ kilogram and remember that the gravitational constant (on the Earth's surface) is approximately $g = 9.8$ meters per second squared. The Hamiltonian of the system reads [Gal83]

$$H = \frac{1}{2l^2m}p^2 - mgl\cos\theta, \tag{5.1.1}$$

where θ is the angle, counted counterclockwise, formed by the pendulum with the vertical direction ($\theta = 0$ corresponds to the configuration in which the pendulum assumes the lowest possible position) and $p = l^2m\dot{\theta}$ is the associated momentum. Thus, (θ, p) are the coordinates of the pendulum. The phase space $\mathcal{M}$ where the motion takes place consists of $\mathbb{T}^1 \times \mathbb{R}$.

The equations of motion associated with the Hamiltonian (5.1.1) represent the motion of an ideal pendulum in a vacuum, feeling only the force of gravity. Clearly, this is a highly idealized situation with no counterpart in reality. Every system interacts with the rest of the universe. Thus, the only hope for the idea of *isolated systems* to be fruitful is that the interaction with the exterior does not significantly affect the behavior of the system. Let us try to see what this can mean in reality.

The first issue is clearly friction. Let us imagine that we have set up the pendulum in a reasonable vacuum and reduced the friction at the suspension point so that the loss of energy is negligible on the time scale of a few minutes. Does such a system behave as an isolated pendulum within such a time frame? One problem is that the suspension point is still in contact with the rest of the world. If the pendulum is in a lab not so distant from a street (a rather common situation), then the traffic will induce some vibrations. It is then natural to ask: what happens if the suspension point of the pendulum vibrates?

In fact, nothing much happens for small pendulum oscillations (this is a consequence of Komogorv-Arnold-Moser theory, a highly non trivial fact), but if we start close to the vertical configuration, it is conceivable that a motion that would be oscillatory for the unperturbed pendulum could gather enough energy from the external force as to change its nature and become rotatory, this would create a substantial difference between the unperturbed (ideal) and the perturbed (more realistic) case.

This is exactly the question we want to address:

Question: *Can we really predict the motion for a reasonable time if the initial condition is close to the vertical ?*

We will assume that the frequency of vibration ω is of the order of one hertz[2] and the amplitude of the oscillations is very, very small. Hence, as good mathematicians, we will call such an amplitude ε. In other words, the suspension point moves vertically according to the law $\varepsilon \cos \omega t$.

The Hamiltonian of the vibrating pendulum is then given by (see Problem 5.1)

$$H_\varepsilon(\theta, p, t) = \frac{1}{2l^2 m} p^2 - mgl \cos\theta - \varepsilon m\omega^2 l \cos\omega t \cos\theta. \tag{5.1.2}$$

Accordingly, the equation of motion are (see Problem 5.1)[3]

$$\begin{aligned} \dot\theta &= \frac{\partial H_\varepsilon}{\partial p} = \frac{p}{l^2 m} \\ \dot p &= -\frac{\partial H_\varepsilon}{\partial \theta} = -mgl \sin\theta - \varepsilon m\omega^2 l \cos\omega t \sin\theta. \end{aligned} \tag{5.1.3}$$

It is well known that the function H is an integral of motion for the solutions of (5.1.3) for $\varepsilon = 0$, that is: H computed along the solutions of the associated equations of motion is constant.[4] The physical meaning of H is the energy of the system. Clearly, the energy H_ε is not constant in general since the vibration can add or subtract energy to the pendulum.

5.2 Instability–unperturbed case

Let us first recall a few basic facts about the unperturbed pendulum. The equations of motion are given by the (5.1.3) setting $\varepsilon = 0$. It is obvious that

[2] One hertz corresponds to one oscillation every second, and it can be the order of magnitude for the frequency of a vibration transmitted through the ground (R waves) at a reasonable distance. Thus we are assuming $\omega = 2\pi$.

[3] Here we write the Hamilton equations associated with the Hamiltonian, see [Arn99, Gal83] for the general theory.

[4] See [Arn99, Gal83] for this general fact or do Problem 5.4 for the simple case at hand.

there exist two fixed points: $(0,0)$ which corresponds to the pendulum at rest and is clearly stable , and $(\pi,0)$, which corresponds to the pendulum in the vertical position and is certainly unstable. Our interest here is to analyze the motions that start close to the unstable equilibrium and to make more precise what it is meant by *instability*.

5.2.1 Unstable equilibrium

If we want to have an idea of how the motion looks near a fixed point the natural first step is to study the linearization of the equation of motion near such a point. In our case, using the coordinates $(\theta_0, p) = (\theta - \pi, p)$, they look like

$$\begin{aligned} \dot{\theta}_0 &= \frac{p}{l^2 m} \\ \dot{p} &= mgl\theta_0. \end{aligned} \tag{5.2.4}$$

Let $\omega_p = \sqrt{\frac{g}{l}}$, the general solution of (5.2.4) is

$$(\theta_0(t), p(t)) = (\alpha e^{\omega_p t} + \beta e^{-\omega_p t}, ml^2\omega_p\{\alpha e^{\omega_p t} - \beta e^{-\omega_p t}\}),$$

where α and β are determined by the initial conditions. Note that if the initial condition has the form $\alpha(1,\, ml\sqrt{gl})$ it will evolve as $\alpha e^{\omega_p t}(1,\, ml\sqrt{gl})$. While if the initial condition is of the form $\beta(1,\, -ml\sqrt{gl})$ it will evolve as $\beta e^{-\omega_p t}(1,\, -ml\sqrt{gl})$. In other words the directions $(1,\, ml\sqrt{gl})$ and $(1,\, -ml\sqrt{gl})$ are invariant for the linear dynamics. The first direction is expanded (and because of this is called *unstable direction*) while the second is contracted (*stable direction*).

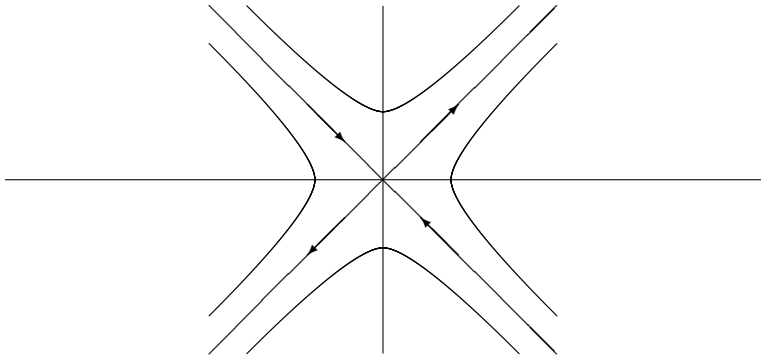

Figure 5.1: Unstable fixed point (phase portrait)

Let us imagine starting the motion from an initial condition of the type $(\pi + \theta_0, 0)$, $\theta_0 \in [-\delta, \delta]$, where $\delta \leq 10^{-4}$ represents the precision with which we are able to set the initial condition (one tenth of a millimeter); what will happen under the linear dynamics?

Our initial condition corresponds to choosing, at time zero, $\alpha = \beta \leq \frac{\delta}{2}$. As time goes on, the coefficient of β becomes exponentially small while the coefficient of α increases exponentially, thus a good approximation of the position of the pendulum after some time is given by

$$\theta_0(t) \approx \alpha e^{\omega_p t}. \tag{5.2.5}$$

Since $\omega_p \approx 3.13$ seconds^{-1}, it follows that after about 2.5 seconds the position of the pendulum can be anywhere up to a distance of about 10 centimeters from the unstable position.

This means that the unstable position is really unstable, and if we tray, as best as we can, to put the pendulum in the unstable equilibrium (always imagining that the friction has been properly reduced) it will typically fall after a few seconds, and it will fall in a direction that we are not able to predict (since it depends on the sign of δ, our unknown mistake). Nevertheless, after the ideal pendulum starts falling in one direction, the subsequent motion is completely predictable, as we will see shortly.

An obvious objection to the above analysis is that I did not show that the linearized equation describes a motion really close to the one of the original equations. The answer to this question is particularly simple in this setting and is addressed in the next subsection.

5.2.2 The unstable trajectories (separatrices)

Given the already noted fact that, for $\varepsilon = 0$, H is a constant of motion, the phase space $\mathcal{M}$ is naturally foliated in the level curves of H, on which the motion must take place. This allows us to obtain a fairly accurate picture of the motions of the unperturbed pendulum. In fact, the level curves are given by the equations

$$\frac{p^2}{2l^2 m} - mgl\cos\theta = E$$

where E is the energy of the motion. It is easy to see that $E = -mgl$ corresponds to the stable fixed point $(\theta, p) = (0, 0)$; $-mgl < E < mgl$ corresponds to oscillations of amplitude $\arccos\left[\frac{E}{mgl}\right]$; $E > mgl$ corresponds to rotatory motions of the pendulum. The last case $E = mgl$ is of particular interest to us: obviously, it corresponds to the unstable fixed point $(\pi, 0)$, yet there are two other solutions that travel on the two curves

$$p = \pm ml\sqrt{2lg(1 + \cos\theta)}.$$

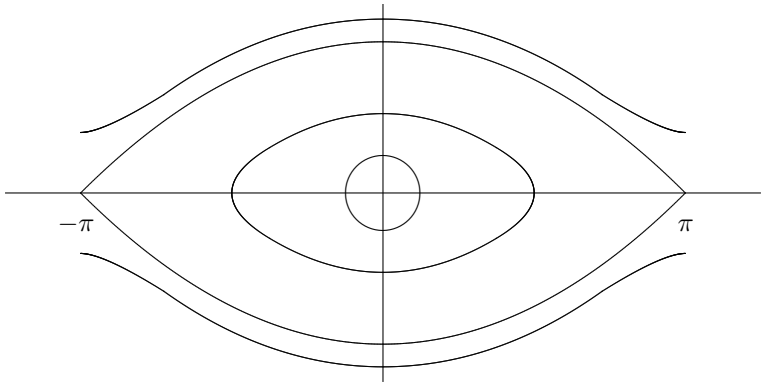

Figure 5.2: Unperturbed pendulum (phase portrait)

These two curves are the ones that separate the oscillatory motions from the rotatory ones and, for this reason, are called *separatrices*. It is very important to understand the motion along such trajectories, luckily the two differential equations

$$\dot{\theta} = \pm\sqrt{2\frac{g}{l}(1+\cos\theta)}. \tag{5.2.6}$$

can be integrated explicitly (see Problem 5.5) yielding, for $\theta(0) = 0$,

$$\theta(t) = 4\arctan e^{\pm\omega_p t} - \pi. \tag{5.2.7}$$

These orbits are asymptotic to the unstable fixed point both at $t \to +\infty$ and at $-\infty$ and, for $|t|$ large, agree with the linear behaviour of section 5.2.1. This situation is somewhat atypical, as we will see briefly.

5.3 The perturbed case

5.3.1 Reduction to a map

The motion of the above system takes place on the cylinder $\mathcal{M} = S^1 \times \mathbb{R}$. By the theorem of existence and uniqueness for the solutions of differential equations follows immediately the possibility to define the maps $\phi_\varepsilon^t : \mathcal{M} \to \mathcal{M}$ associating to the point (θ, p) the point reached by the solution of (5.1.3) at time t, when starting at time 0 from the initial condition (θ, p). In such a way, we define the *flow* ϕ_ε^t associated to the (5.1.3).

Clearly $\phi_\varepsilon^0(\theta, p) = (\theta, p)$, that is, the map corresponding to time zero is the identity. Moreover, if $\varepsilon = 0$ the system is autonomous (the vector field does not depend on the time) hence the flow defines a group: for each $t, s \in \mathbb{R}$

$$\phi_0^{t+s}(\theta, p) = \phi_0^t(\phi_0^s(\theta, p)).$$

This corresponds to the obvious fact that the motion for a time $t + s$ can be obtained first as the motion from time 0 to time s, and then pretending that the time s is the initial time and following the motion for time t.

Of course, the above fact does not hold anymore when $\varepsilon \neq 0$. In this case, the maps ϕ_ε^t depend on our choice of the initial time (if we define them by starting from time 1 instead then time 0, in general we obtain different maps). Nevertheless, due to the fact that the external force is periodic something can be saved of the above nice property.

Let us define the map $T_\varepsilon : \mathcal{M} \to \mathcal{M}$ by

$$T_\varepsilon = \phi_\varepsilon^{\frac{2\pi}{\omega}},$$

then (see Problem 5.3), for each $n \in \mathbb{Z}$,

$$T_\varepsilon^n = \phi_\varepsilon^{\frac{2n\pi}{\omega}}. \tag{5.3.8}$$

The interest of (5.3.8) is that, for many purposes, we can study the map T_ε instead than the more complex object ϕ_ε^t. Morally, it means that if we look at the system *stroboscopically*, that is only at the times $\frac{2\pi}{\omega} n$ with $n \in \mathbb{Z}$, then it behaves like an autonomous (time independent) system.[5] Another interesting fact is that the flow ϕ_ε^t (and hence also the map T_ε) is area preserving (see Problem 5.7).[6]

5.3.2 Perturbed pendulum, $\varepsilon \neq 0$

The situation for the case $\varepsilon \neq 0$ is more complex, and no easy way exists to study these motions.

As a general strategy, to study the behavior of a system (in our case, the map T_ε) it is a good idea to start by investigating simple cases and then move on from there. In our systems, the simplest motion consists of the equilibrium solutions. These are the time independent solutions.[7] Because of the special

[5] Another instance of a very simple case of a very fruitful and general strategy: to look at the system only when some special event happens–in our case, at each time in which the suspension point has its maximum height.

[6] This also is a special instance of a more general fact: the Hamiltonian nature of the system, see [Arn99, Gal83] if you want to know more.

[7] That is, equilibrium solutions for the map T_ε. These are *periodic* solutions for the flows of period $\frac{2\pi}{\omega}$. In fact, $T_\varepsilon x = x$ means $\phi^{\frac{2\pi}{\omega}} x = x$.

type of perturbation chosen, the fixed points of the system for the case $\varepsilon = 0$ remain unchanged when $\varepsilon \neq 0$ (see Problem 5.8 for a brief discussion of a more general case).

Next, we can study the infinitesimal nature of the fixed points. It is natural to expect that the nature of the two fixed points does not change if ε is small, yet to verify this requires some checking. We will discuss explicitly only the fixed point $(\pi, 0)$.

The first step is to make precise the sense in which the case $\varepsilon \neq 0$ is a perturbation of the case $\varepsilon = 0$. This can be achieved by obtaining an explicit estimate of the size of

$$R_\varepsilon = \varepsilon^{-1}(T_0 - T_\varepsilon).$$

Let $z(t) = (z_1(t), z_2(t)) = \phi_0^t(x) - \phi_\varepsilon^t(x)$, then substituting in (5.1.3) and subtracting the general case from the case $\varepsilon = 0$ it yields

$$\begin{aligned} |\dot{z}_1| &\leq \frac{|z_2|}{ml^2} \\ |\dot{z}_2| &\leq mgl|z_1| + \varepsilon m\omega^2 l. \end{aligned}$$

In order to get better estimates, it is convenient to define the new variables $\zeta_1 = z_1$ and $ml^2\omega_p\zeta_2 = z_2$. In these new variables, the preceding equations read

$$\begin{aligned} |\dot{\zeta}_1| &\leq \omega_p|\zeta_2| \\ |\dot{\zeta}_2| &\leq \omega_p|\zeta_1| + \varepsilon\frac{\omega^2}{\omega_p l}. \end{aligned} \tag{5.3.9}$$

Which implies $\|\dot{\zeta}\| \leq \omega_p\|\zeta\| + \varepsilon m\omega^2 l$. Taking into account that, in our situation, $ml^2\omega_p > 1$, it follows (see Problem 5.9)

$$\|R\|_{\mathcal{C}^0} \leq \frac{m\omega^2}{l\omega_p}(e^{2\pi\frac{\omega_p}{\omega}} - 1) \leq 69.$$

Unfortunately, the above norm does not suffice for our future needs. We will see quite soon that it is necessary to estimate also the first derivatives of R, that is the $\mathcal{C}^1$ norm.

To do so, the easiest way is to use the differentiability with respect to the initial conditions of the solutions of our differential equation. Fixing any point $x \in \mathcal{M}$ and calling $\xi^\varepsilon(t) = d_x\phi_\varepsilon^t\xi(0)$ we readily obtain:[8]

[8]The vector $\xi_\varepsilon(t)$ is nothing else than the derivative $\frac{d\phi_\varepsilon^t(x+s\xi(0))}{ds}|_{s=0}$, the following equation is then obtained by exchanging the derivative with respect to t with the derivative with respect to s.

$$\begin{aligned}\dot{\xi}_1^\varepsilon &= \frac{\xi_2^\varepsilon}{l^2 m} \\ \dot{\xi}_2^\varepsilon &= -mgl\cos\theta\,\xi_1^\varepsilon - \varepsilon m\omega^2 l\cos\omega t\cos\theta\,\xi_1^\varepsilon\end{aligned} \tag{5.3.10}$$

One can then estimate the $\mathcal{C}^1$ norm of R by estimating $\|\xi^\varepsilon(\frac{2\pi}{\omega}) - \xi^0(\frac{2\pi}{\omega})\|$, since $\xi^\varepsilon(\frac{2\pi}{\omega}) = D_{(\theta,p)}T_\varepsilon\xi^\varepsilon(0)$. Doing so, one obtains[9]

$$\|R\|_{\mathcal{C}^1} \le \frac{2m\omega^2}{l\omega_p} e^{3\pi\frac{\omega_p}{\omega}} := d_1 \le 690. \tag{5.3.11}$$

5.4 Infinitesimal behavior (linearization)

As a first application of the above considerations, let us study the linearization of T_ε at $x_f = (\pi, 0)$. From (5.3.10) follows (see Problem 5.12)

$$\begin{aligned}D_{x_f}T_0 &= \begin{pmatrix} \cosh\frac{2\pi\omega_p}{\omega} & \frac{\sinh\frac{2\pi\omega_p}{\omega}}{ml^2\omega_p} \\ ml^2\omega_p\sinh\frac{2\pi\omega_p}{\omega} & \cosh\frac{2\pi\omega_p}{\omega}\end{pmatrix} \\ D_{x_f}T_\varepsilon &= D_{x_f}T_0 + \mathcal{O}(d_1\varepsilon)\end{aligned} \tag{5.4.12}$$

The eigenvalues of $D_{x_f}T_\varepsilon$ are then $\lambda_\varepsilon = e^{\frac{2\pi\omega_p}{\omega}} + \mathcal{O}(d_2\varepsilon)$,[10] λ_ε^{-1}, where $d_2 = 2d_1\omega_p ml^2 \simeq 4400$. In addition, calling v_ε, $\langle v_\varepsilon, v_0\rangle = 1$, the eigenvector associate to λ_ε, holds true $\|v_0 - v_\varepsilon\| \le d_3\varepsilon$, $d_3 = 4\lambda_0^{-1}\omega_p^2\omega^2 l^4 d_1 \simeq 1200$.[11]

Clearly, if ε is sufficiently small, then $\lambda_\varepsilon > 1$. This means that the hyperbolic nature of the unstable fixed point remains unchanged under small perturbations (see Problem 5.13 for a case when the perturbation is not so small).[12]

If one does a similar analysis at the fixed point $(0, 0)$ one finds that the eigenvalues have modulus one: that is, the infinitesimal motion is a rotation around the fixed point, exactly as in the $\varepsilon = 0$ case.

Hence, the comments made at the end of subsection 5.2.1 for the unperturbed pendulum hold for the perturbed pendulum as well. Only now the is no

[9] The following bounds are not sharp, working more, one can obtain better estimates, but this would not make much of a difference in the sequel.

[10] In this chapter we will adopt the strict convention that $\mathcal{O}(x)$ means a quantity bounded, in absolute value, by x.

[11] This follows by the fact that the eigenvalues of $D_{x_f}T_0$ are $e^{\pm\frac{2\pi\omega_p}{\omega}} \simeq (23)^{\pm 1}$, a simple perturbation theory of matrices (see Problems 5.10, 5.11) and the already mentioned fact that the map T_ε is area preserving, thus the determinant of its derivative must be one.

[12] As we will see later in detail, hyperbolicity means that there is a direction in which the maps expand (the eigenvector v_ε^u associated to the eigenvalue λ_ε) and a direction in which the map contracts (the eigenvector v_ε^s associated to the eigenvalue λ_ε^{-1})

longer an integral of motion (the energy) that controls globally the behavior of the system.

Imagining that the map is linear (which is clearly false but, as we will see, qualitatively not so wrong) this would mean that the distance between two trajectories can be expanded by almost a factor 23 in a second. Initial conditions that are δ close at time zero will be about 23δ far apart after 1 second. If such a state of affair could persist (and we will see it may) after one minute the two configurations would differ roughly by a factor $10^{80}\delta$, which means that not even knowing the initial condition plus or minus a quark could we predict the final one. This is certainly a rather worrisome perspective, but much more work it is needed to decide if this may indeed be the case.

5.5 Local behavior (Hadamard-Perron Theorem)

The next step is to try to go from the above infinitesimal analysis to a local picture in a small neighborhood of the fixed points.

It is natural to expect that the two fixed points are still stable and unstable respectively, yet this is a far from trivial fact.

The stability of the point $(0,0)$ can be proven by invoking the so called KAM Theorem (this exceeds the scope of the present book and we will not discuss such matters, see [Gal83] for such a discussion).[13]

The study of the local behavior around the point x_f is instead a bit easier and can be performed by applying the Hadamard-Perron Theorem 2.4.2 to conclude that, in a neighborhood of $(\pi,0)$, there exists two curves $x^u_\varepsilon(s) = (\theta^u_\varepsilon(s), p^u_\varepsilon(s))$, $x^s_\varepsilon(s)$ that are invariant with respect to the map T_ε. Namely, there exists $\delta_\varepsilon > 0$ such that $T_\varepsilon x^s_\varepsilon([-\delta_\varepsilon,\delta_\varepsilon]) \subset x^s_\varepsilon([-\delta_\varepsilon,\delta_\varepsilon])$ and $T^{-1}_\varepsilon x^u_\varepsilon([-\delta_\varepsilon,\delta_\varepsilon]) \subset x^u_\varepsilon([-\delta_\varepsilon,\delta_\varepsilon])$; these are called the local stable and unstable manifold of zero, respectively. Essentially δ_ε is determined by the requirement that the non-linear part of T_ε be smaller than the linear part.

Clearly, for $\varepsilon = 0$ $x^s_0 = x^u_0 = x_0$ and it coincides with the homoclinic orbit of the unperturbed pendulum. In addition, by Hadarmd-Perron and the estimates of the previous section, we can choose δ_ε such that

$$\|x^u_\varepsilon - x_0\| \leq 2d_3\varepsilon\|x_0\|. \tag{5.5.13}$$

[13] In some sense this implies that we can indeed predict the motion for an extremely long time if we consider only oscillations close to the configuration $(0,0)$, so in that case the assumption that the pendulum is isolated is legitimate. Yet, this depends on the precision we are interested in and tends to degenerate if the amplitude of the oscillations is rather large. A complete analysis would be a very complicated matter but we will have an idea of the type of problems that can arise by considering extremely large oscillations, close to a full rotation of the pendulum.

and the analogous for the stable manifold. We have obtained a local picture of the behavior of the map T_ε, yet this does not suffice to answer our original question. To do so, we need to follow the motion for at least a full oscillation: this requires global information.

To gain a more global knowledge, we can try to construct a larger invariant set for the map T_ε. A natural way to do so is to iterate: define $W^u = \cup_{n=0}^{\infty} T_\varepsilon^n x^u([-\delta_\varepsilon, \delta_\varepsilon])$. Since $T_\varepsilon x^u([-\delta_\varepsilon, \delta_\varepsilon]) \supset x^u([-\delta_\varepsilon, \delta_\varepsilon])$, it is clear that each time we iterate, we get a longer and longer curve. The set W^u is then clearly a manifold, and it is called the global unstable manifold.[14]

The global manifold, as the name clearly states, is a global object: it carries information on the dynamics for arbitrarily long times. Yet, the procedure by which it has been defined is far from constructive, and the truth is that, besides the sketchy considerations above, at the moment we know very little of it. The next step is to gain a more detailed understanding of a large portion of W^u.

5.6 A more global understanding (Melnikov)

From the above considerations follows that the stable and unstable manifolds $(\theta_\varepsilon^s(s),\, p_\varepsilon^s(s))$, $(\theta_\varepsilon^u(s),\, p_\varepsilon^u(s))$, $|s| \le \delta_\varepsilon$, of T_ε at 0, are ε close to the homoclinic orbit of the unperturbed pendulum, $(\theta_0(t),\, p_0(t))$, $\theta_0(0) = 0$.

Note, however, that while $x_0 = (\theta_0, p_0)$ is invariant under the unperturbed flow, the same does not apply to $(\theta_\varepsilon^{s,u}(s),\, p_\varepsilon^{s,u}(s))$ under ϕ_ε^t. The invariant object is the time-space surface $(\tau, x_\varepsilon^{s,u}(s,\tau)) := (\tau, \phi_\varepsilon^\tau(\theta_\varepsilon^s(s),\, p_\varepsilon^s(s)))$ where $(s,\tau) \in [-\delta_\varepsilon, \delta_\varepsilon] \times [0, \frac{2\pi}{\omega}]$ and and $\tau = t \mod \frac{2\pi}{\omega}$.[15]

We can choose freely the parameterization of our curves in such a surface, and some are more convenient than others. The separatrix of the unperturbed pendulum is most conveniently parametrized by time, hence $\phi^t(\theta_0(s), p_0(s)) = (\theta_0(s+t), p_0(s+t))$. Note that the separatrix can be visualized as a graph of $(\theta, G(\theta))$. Analogously, for ε small enough, the perturbed unstable manifold of T_ε will be the graph of $(\theta, G_\varepsilon^u(\theta))$, for $\theta \in [0, \frac{3}{2}\pi]$. Given $\theta \in [0, \frac{3}{2}\pi]$, let $S_n = 2\pi\omega^{-1}n$. Let $z_n := (\theta_n, G_\varepsilon^u(\theta_n)) = \phi_\varepsilon^{-S_n}(\theta, G_\varepsilon^u(\theta))$, by Hadamrad-Perron we know that $|G_\varepsilon^u(\theta) - G(\theta)| \le C\theta$ for $\theta \in [0, \delta]$, also $|\theta_n| \le Ce^{-an}$ for

[14] Applying the above procedure to the unperturbed problem yields the full separatrix.

[15] A standard way to bring the present non-autonomous setting into the more familiar autonomous one is to introduce the fake variables $(\varphi, \eta) \in S^1 \times \mathbb{R}$ and the new, time independent, Hamiltonian $\bar H_\varepsilon(\theta, p, \varphi, \eta) := H_\varepsilon(\theta, p, \varphi) + \frac{2\pi}{\omega}\eta$. The Hamilton equations yield $\varphi(t) = \frac{2\pi}{\omega}t + \varphi(0)$ and hence the equations for θ, p reduce to (5.1.3). Since $\bar H_\varepsilon$ is now conserved under the motion we can restrict the system to the three dimensional manifold $\bar H_\varepsilon = 0$. In such a manifold, we have the *weak* stable and unstable manifolds (now flow invariant) $(x_\varepsilon^{s,u}(s,\varphi), \varphi, -\frac{2\pi}{\omega}H_\varepsilon((x_\varepsilon^{s,u}(s,\varphi), \varphi))$.

some $C, a > 0$. The basic idea is to compute

$$\begin{aligned} H_0(\theta, G^u_\varepsilon(\theta)) &= H_0(\theta_n, G^u_\varepsilon(\theta_n)) + \int_0^{S_n} \frac{dH_0 \circ \phi^s_\varepsilon(\theta_n, G^u_\varepsilon(\theta_n))}{ds} ds \\ &= H_0(z_n) + \int_0^{S_n} \langle \nabla H_0, J\nabla H_0 + \varepsilon J \nabla H_1 \rangle \circ \phi^{s-S_n}_\varepsilon(z) ds \\ &= H_0(z_n) + \varepsilon \int_{-S_n}^0 \langle \nabla H_0, J\nabla H_1 \rangle \circ \phi^s_\varepsilon(z) ds. \end{aligned}$$

The results of section 5.4 implies that, for some $C > 0, \alpha > 0$,

$$\|\phi^{-s}_\varepsilon(\theta, G^u_\varepsilon(\theta))\| \leq Ce^{-\alpha s},$$

for all $s \geq 0$. In addition, by (5.5.13), setting $(\theta(s), p(s)) = \phi^{-s}_0(\theta, G(\theta))$

$$\|\phi^{-s}_\varepsilon(\theta, G^u_\varepsilon(\theta)) - \phi^{-s}_0(\theta, G(\theta))\| \leq C \min\{\varepsilon e^{\beta s}, e^{-\alpha s}\},$$

for some $\beta > 0$. Thus, taking the limit $n \to \infty$, yields

$$H_0(\theta, G^u_\varepsilon(\theta)) = H_0(0) + \varepsilon \int_{-\infty}^0 \langle \nabla H_0, J\nabla H_1 \rangle \circ \phi^s_0(z) + o(\varepsilon).$$

Consequently,

$$\begin{aligned} G^u_\varepsilon(\theta) - G_0(\theta) &= \frac{G^u_\varepsilon(\theta)^2 - G_0(\theta)^2}{G^u_\varepsilon(\theta) + G_0(\theta)} = \frac{2(H_0(\theta, G^u_\varepsilon(\theta)) - H_0(\theta, G_0(\theta)))}{G^u_\varepsilon(\theta) + G_0(\theta)} \\ &= \frac{2(H_0(\theta, G^u_\varepsilon(\theta)) - H_0(0))}{G^u_\varepsilon(\theta) + G_0(\theta)} \\ &= 2\varepsilon \frac{\int_{-\infty}^0 \langle \nabla H_0, J\nabla H_1 \rangle \circ \phi^s_0(\theta, G(\theta)) ds + \mathcal{O}(\varepsilon)}{G^u_\varepsilon(\theta) + G_0(\theta)} \end{aligned}$$

This allows us to conclude

$$G^u_\varepsilon(\theta) - G_0(\theta) = \varepsilon \frac{\int_{-\infty}^0 \langle \nabla H_0, J\nabla H_1 \rangle \circ \phi^s_0(\theta, G(\theta)) ds}{G_0(\theta)} + o(\varepsilon)$$

Arguing analogously for the stable manifold yields

$$G^u_\varepsilon(\theta) - G^s_\varepsilon(\theta) = \varepsilon \frac{\int_{\mathbb{R}} \langle \nabla H_0, J\nabla H_1 \rangle \circ \phi^s_0(\theta, G(\theta)) ds}{G_0(\theta)} + o(\varepsilon). \tag{5.6.14}$$

The separatrix of the unperturbed pendulum is most conveniently parametrized by time, hence

$$\phi^t(\theta_0(s), p_0(s)) = (\theta_0(s+t), p_0(s+t)) = (\theta_0(s+t), G(\theta_0(s+t)) =: x_0(s+t).$$

Setting $\Delta(s) = \varepsilon^{-1}[G^u_\varepsilon(\theta_0(s)) - G^s_\varepsilon(\theta_0(s))]G_0(\theta_0(s))$, one can compute

$$\Delta(\sigma) = \int_{-\infty}^{\infty} \{H_1,\, H\}_{x_0(t+\sigma)} dt + o\left(d_4 e^{2\omega_p|\sigma|}\right), \tag{5.6.15}$$

where and explcit computation yields $d_4 \simeq 4 \cdot 10^6$, and the curly brackets stand for the so called *Poisson brackets* ($\{f,\, g\}_x = \langle J\nabla_x f,\, \nabla_x g\rangle$).
The integral in (5.6.15) is called *Melnikov integral* and provides an expression, at first order in ε, of the distance between the stable and the unstable manifold. All we are left with is to compute the integrals in (5.6.15). This turns out to be an exercise in complex analysis, and it is left to the reader (see Problem 5.15), the result is:[16]

$$\int_{-\infty}^{\infty} \{H_1(\cdot, t),\, H\}_{x_0(t+\sigma)} dt = 8\pi m l \frac{\omega^4 e^{-\frac{\pi\omega}{2\omega_p}}}{\omega_p^2(e^{\frac{\pi\omega}{\omega_p}} - 1)} \sin\omega\sigma.$$

We have thus gained a very sharp control on the shape of the above manifolds.[17] In particular, $\Delta(\pm 1/4) \simeq \pm 76 + \mathcal{O}(4\cdot 10^7\varepsilon) \neq 0$ provided $\varepsilon \leq 1.5\cdot 10^{-6}$, that is the two manifolds intersect. To understand a bit better such an intersection (we would like to know that in the region $\sigma \in [-1/4, 1/4]$ there is only one *transversal* intersection) it suffices to notice that (5.6.14) provides a control on the angle between x^u_ε and x_0.

This intersections are called *homoclinic* intersection and their very existence is responsible for extremely interesting phenomena as can be readily seen by trying to draw the stable and unstable manifolds (see Figure 5.3 for an approximate first idea); we will discuss this issue in detail shortly.[18]

We have gained much more global information on the map T_ε, yet it does not suffice to answer to our question. The next section is devoted to obtaining

[16] A simple computation yields:

$$\{H_1,\, H\}_{x_0(t+s)} = -\frac{\omega^2}{l} p(t+s)\cos\omega t \sin\theta(t+s).$$

Then, by using (5.2.7) and looking at Problem 5.6, one readily obtains:

$$\{H_1,\, H\}_{x_0(t)} = 4\frac{\omega^2}{l}\frac{\cos\omega(t-s)\sinh\omega_p t}{(\cosh\omega_p t)^3}.$$

Finally, use Problem 5.15.

[17] Note that ε must be exponentially small with respect to ω. In many concrete problems (notably the so-called *Arnold diffusion* it happens that this it is not the case. One can try to solve such an obstacle by computing the next terms of the ε expansion of Δ. In fact, it turns out that it is possible to express Δ as a power series in ε with all the terms exponentially small in ω . Yet this is a quite complex task far beyond our scope.

[18] Note that the intersection corresponds to a homoclinic orbit for the map T_ε (that is, an orbit which approaches the fixed point x_f both in the future and in the past). This is what it is left of the homoclinic orbit of the unperturbed pendulum.

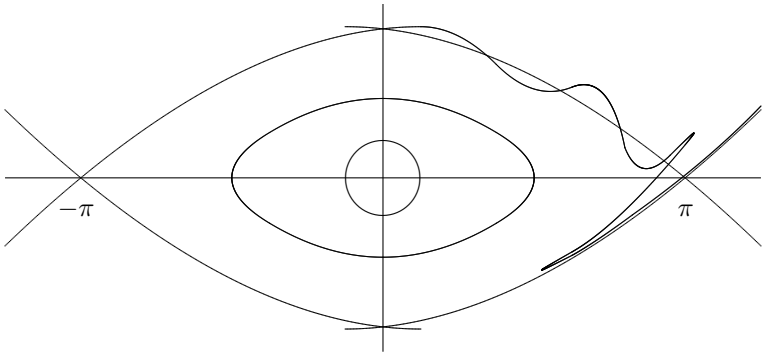

Figure 5.3: Perturbed pendulum

a really global picture. Up to now, we have used mainly analytic tools. Next, geometry will play a much more significant rôle.[19]

5.7 Global behavior (an horseshoe)

We want to explicitly construct trajectories with special properties. A standard way to do so is to start by studying the evolution of appropriate regions and to use judiciously the knowledge so gained. Let us see what this means in practice.

The starting point is to note that we understand the shape of the invariant manifolds, but not very well the dynamics on them, this is our next task. Since points on the unstable manifolds are pulled apart by the dynamics, the estimate must be done with a bit of care. In fact, we will use a way of arguing that is typical when instabilities are present, we will see many other instances of this type of strategy in the sequel.

For each x in the unstable manifold (zero included) let us call $D^u_x T_\varepsilon := D_x T_\varepsilon v^u(x)$, where $v^u(0) = v^u$ and if $x = x^u_\varepsilon(t)$ then $v^u(x) = \|\dot{x}^u_\varepsilon(t)\|^{-1}\dot{x}^u_\varepsilon(t)$, that is the derivative of the map computed along the unstable manifold. A useful idea in the following is the concept of *fundamental domain*. Define $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ by $x^u_\varepsilon(t) = x^u_\varepsilon(\alpha(t))$. Then $[t, \alpha(t)]$ is a fundamental domain and has the property that, setting $t_i := \alpha^i(t)$, the sets $\alpha^i[t_0, t_1]$ intersect only at the boundary.

[19] What comes next is the first example in this book of what is loosely called *a dynamical argument*.

Lemma 5.7.1 (Distortion) *For each x, y in the same fundamental domain of the unstable manifold, $\delta_0 > 0$, and $n \in \mathbb{N}$ such that $\|T_\varepsilon^n x\| \le \delta_0$, holds*[20]

$$e^{-\delta_0 C_2} \le \left| \frac{D_x^u T_\varepsilon^n}{D_y^u T_\varepsilon^n} \right| \le e^{\delta_0 C_2},$$

where $C_2 = \sup_{t \le 0} \left| \frac{\ddot{\alpha}(t)}{\dot{\alpha}(t)} \right|$.

PROOF. The proof is a direct application of the chain rule:

$$\left| \frac{D_x^u T_\varepsilon^n}{D_y^u T_\varepsilon^n} \right| = \prod_{i=1}^n \left| \frac{D^u_{T^i x} T_\varepsilon}{D^u_{T^i y} T_\varepsilon} \right| \le \operatorname{Exp}\left[\sum_{i=1}^n |\log(|D^u_{T^i x} T_\varepsilon) - \log(|D^u_{T^i y} T_\varepsilon|)| \right]$$
$$\le \operatorname{Exp}\left[\sum_{i=1}^n C_2 \|T^i x - T^i y\| \right] = \operatorname{Exp}\left[\sum_{i=1}^n C_2 \|x_\varepsilon^u(t_i) - x_\varepsilon^u(t_{i-1})\| \right] \le e^{C_2 \delta_0}.$$

The other inequality is obtained by exchanging the rôle of x and y. □

Next, we would like to consider the evolution of a small box constructed around the fix point.

Consider the following small parallelogram: $Q_\delta := \{\xi \in \mathbb{R}^2 \mid \xi = a v^u + b v^s$ for some $a, b \in [-\frac{\delta}{2}, \frac{\delta}{2}]\}$, $\delta \ll \delta_0$. Next, consider the first $n \in \mathbb{N}$ such that $T_\varepsilon^n Q_\delta \cap \{\theta = 0\} \neq \emptyset$. Our first task is to understand the shape of $T_\varepsilon^n Q_\delta$ near $\{\theta = 0\}$. Since a fundamental domain in the latter region is of order one, while at the boundary of Q_ε is of order δ, Lemma 5.7.1 implies that the expansion is proportional to $C\delta^{-1}$. By the area preserving of the map, it follows that $T_\varepsilon^n Q_\delta$ must be contained din a $C\delta^2$ neighborhood of the unstable manifold, see Figure 5.4.

By the previous section's considerations on the shape of the invariant manifolds $T^n Q_\delta \cap T^n Q_\delta \neq \emptyset$, moreover they intersect *transversally*.[21]

This is all that is needed to construct a horseshoe (see section ???). In particular, in our case, it means that $T^{2n_0} Q_\delta \cap Q_\delta \neq \emptyset$, in fact the intersection is transversal and consists of three strips almost parallel to the unstable sides. One contains zero, and it is the least interesting for us, the other two cross above and below the unstable manifold, respectively. The width of such a strip is about δ^{-3}. We will discuss in the next chapters all the implications of

[20] This quantity is commonly called *Distortion* because it measures how much the map differs from a linear one (notice that if T is linear then $\frac{D_x T}{D_y T} = 1$). Although apparently an innocent quantity, it is hard to overstate its importance in the study of hyperbolic dynamics.

[21] The meaning of *transversally* is the following: the square Q_δ has two sides parallel to v^u (the unstable direction), which we will call unstable sides, and two sides parallel to v^s (the stable direction), which we will call stable sides. Then the intersection is transversal if it consists of a region with again four sides: two made of the image of the unstable sides and two made of images of the stable sides of Q_δ.

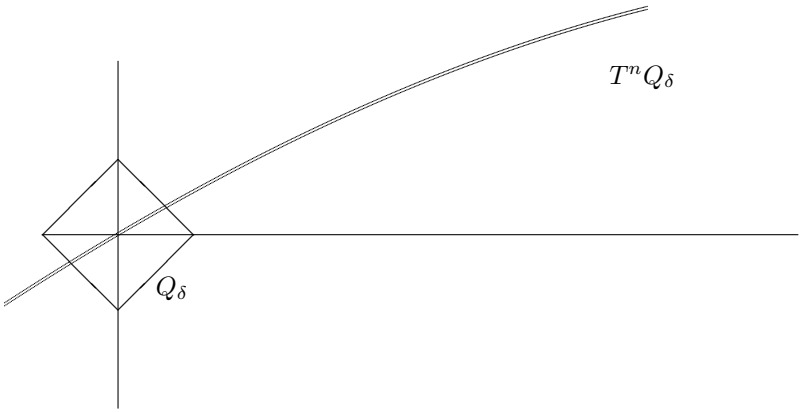

Figure 5.4: The evolution of the small box Q_δ

this situation, here it suffices to notice that if we have two initial conditions in $T^{-2n_0}Q_\delta \cap Q_\delta$ at a distance h, after $2n_0$ iterations the two points will be in Q_δ again but at a distance $h\varepsilon^{-1}$. Since to decide if after that there will be a rotation or an oscillation we need to know the final position with a precision of order δ, we need to know the initial position with a precision $\mathcal{O}(\delta\varepsilon) = \mathcal{O}(\delta^3)$.

Note that in the above construction, we have lost almost all the points, only the ones that come back to Q_δ at time $2n_0$ are under control. Nevertheless, we can consider the set $\Lambda := \cup_{k\in\mathbb{Z}} \bigcap T_\varepsilon^{2kn_0} Q_\delta$. This is clearly a measure zero set, yet it is far from empty (it contains uncountably many points) and it is made of points that at times multiples of $2n_0$ are always in Q_δ. When they arrive in Q_δ they will rotate if they are above the separatrices and oscillate otherwise. Let us call these two subsets of Q_δ R and O. Given a point $\xi \in Q_\delta$, we can associate to it the doubly infinite sequence $\sigma \in \{0,1\}^{\mathbb{Z}}$ by the rule $\sigma_i = 1$ iff $T^{2n_0 i}\xi \in R$. The reader can check that the correspondence is onto.

5.8 Conclusion–an answer

If $\varepsilon = 10^{-6}$ and δ is a millimeter then we need to know the initial condition with a precision of 10^{-9} meters if we want to decide if the point will come back or rotate when it will get almost vertical again (this will happen in about 6 seconds). By the same token if we want to answer the same question, but for the second time, the pendulum gets close to the unstable position, we need to know the initial condition with a precision of the order 10^{-15} meters, and

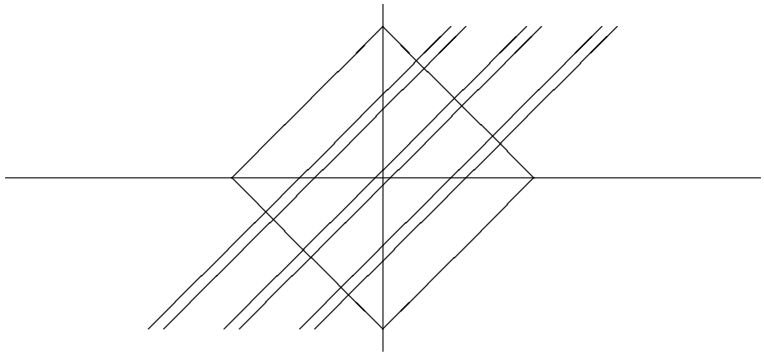

Figure 5.5: Horseshoe construction

this just to predict the motion for about 12 seconds.[22]

We can finally answer to our original question:

Answer: *NO!*

Nevertheless, as we mentioned at the beginning, the above answer it is not the end of the story. In fact, there exist many other very relevant questions that can be answered.[23] The rest of the book deals with a particular type of question: can we meaningfully talk about the *statistical behavior* of a system?

Problems

5.1. Derive the Lagrangian, Hamiltonian and equations of motion for a pendulum attached to a point vibrating with frequency ω and amplitude ε. (Hint: see [LL76, Gal83] on how to do such things. Remember that two Lagrangian that differ by a total time derivative give rise to the same equation of motion and are thus equivalent.)

[22] Remark that it is not just a matter of precision on the initial condition, it is also a matter of how one actually does the prediction. If the method is to integrate numerically the equation of motion, then one has to insure that the precision of the algorithm is of the order of 10^{-15}. This maybe achieved by working in double precision but if one wants to make predictions of the order of one minute, it is quite clear that the numerical problem becomes very quickly intractable.

[23] For example: which type of motions are possible? This is a *qualitative* question. Such types of questions give rise to the qualitative theory of Dynamical Systems [PT93, HK95], an extremely important part of the theory of dynamical systems, although not the focus here.

5.2. Consider the systems of differential equations $\dot{x} = f(x)$, $x \in \mathbb{R}^n$ and f smooth and bounded. Prove that the associated flow form a group. (Hint: use the uniqueness of the solutions of the ordinary differential equation)

5.3. Consider the systems of differential equations $\dot{x} = f(x,t)$, $x \in \mathbb{R}^n$ and f smooth, bounded and periodic in t of period τ. Let ϕ^t be the associated flow. Define $T = \phi^\tau$, prove that $T^n = \phi^{n\tau}$.

5.4. Show that the Hamiltonian is a constant of motion for the pendulum. (Hint: Compute the time derivative)

5.5. Prove (5.2.7). (Hint: Write (5.2.6) in the integral form

$$t = \int_0^t \frac{\dot{\theta}(s)}{\sqrt{\frac{2g}{l}(1+\cos\theta(s))}} ds.$$

Using some trigonometry and changing variable, obtain

$$t = \int_0^{\theta(t)} \frac{1}{2\omega_p \cos\frac{\theta}{2}} d\theta.$$

and compute it.)

5.6. If $\theta(t)$ is the motion obtained in the previous problem, show that

$$\sin\theta(t) = 2\frac{\sinh\omega_p t}{(\cosh\omega_p t)^2};\quad \cos\theta(t) = \frac{2}{(\cosh\omega_p t)^2} - 1;$$
$$\cos^2\frac{\theta(t)+\pi}{4} = \frac{1}{1+e^{2\omega_p t}}.$$

5.7. Consider the systems of differential equations $\dot{x} = f(x,t)$, $x \in \mathbb{R}^n$ and f smooth. Suppose further that $\mathrm{d}ivf = 0$ (that is $\sum_{i=1}^n \frac{\partial f_i}{\partial x_i} = 0$). Show that the associated flow preserves the volume. (Hint: note that this is equivalent to saying that $|\det d\phi^t| = 1$, moreover by the group property and the chain rule for differentiating it suffices to check the property for small t. See that $d\phi^t = \mathbb{1} + Dft + \mathcal{O}(t^2) = e^{Dft+\mathcal{O}(t^2)}$. Finally, remember the formula $\det e^A = e^{\mathrm{T}rA}$.)

5.8. Let $T, T_1 : \mathbb{R}^2 \to \mathbb{R}^2$ be a smooth maps such that $T0 = 0$ and $\det(\mathbb{1} - D_0T) \neq 0$. Consider the map $T_\varepsilon = T + \varepsilon T_1$ and show that, for ε small enough, there exists points $x_\varepsilon \in \mathbb{R}^2$ such that $T_\varepsilon x_\varepsilon = x_\varepsilon$. (Hint: Consider the function $F(x,\varepsilon) = x - T_\varepsilon x$ and apply the Implicit Function Theorem to $F = 0$.)

5.9. Let $x(t) \in \mathbb{R}^n$ be a smooth curve satisfying $\|\dot{x}(t)\| \leq a(t)\|x(t)\| + b(t)$, $x(0) = x_0$, $a, b \in \mathcal{C}^0(\mathbb{R}, \mathbb{R}_+)$, prove that

$$\|x(t) - x_0\| \leq \int_0^t e^{\int_s^t a(\tau)d\tau} \left[a(s)\|x_0\| + b(s)\right] ds.$$

(Hint: Note that $\|x(t)-x_0\| \leq \int_0^t \|\dot{x}(s)\|ds$. Transform then the differential inequality into an integral inequality and apply Gronwall inequality, Lemma 1.1.8.)

5.10. Given two by two matrices A, B such that A has eigenvalues $\lambda \neq \mu$, show that the matrix $A_\varepsilon = A + \varepsilon B$, for ε small enough, has eigenvalues $\lambda_\varepsilon, \mu_\varepsilon$ analytic as functions of ε. Show that the same holds for the eigenvectors. (Hint:[24] consider z in the resolvent of A, that is $(z-A)^{-1}$ exists. Then $(z - A_\varepsilon) = (z - A)(\mathbb{1} - \varepsilon(z - A)^{-1}B)$. Accordingly, if ε is small enough, $(z - A_\varepsilon)^{-1} = \left\{\sum_{n=0}^{\infty} \varepsilon^n \left[(z - A)^{-1}B\right]^n\right\}(z - A)^{-1}$. Finally, if γ, γ' are curves on the complex plane containing λ and μ, respectively, verify that

$$\Pi_\varepsilon := \frac{1}{2\pi i}\int_\gamma (z - A_\varepsilon)^{-1}dz \quad \Pi'_\varepsilon := \frac{1}{2\pi i}\int_{\gamma'} (z - A_\varepsilon)^{-1}dz$$

are commuting projectors and $A_\varepsilon = \lambda_\varepsilon\Pi_\varepsilon + \mu_\varepsilon\Pi'_\varepsilon$. Finally verify that

$$\lambda_\varepsilon\Pi_\varepsilon := \frac{1}{2\pi i}\int_\gamma z(z - A_\varepsilon)^{-1}dz \quad \mu_\varepsilon\Pi'_\varepsilon := \frac{1}{2\pi i}\int_{\gamma'} z(z - A_\varepsilon)^{-1}dz.$$

The statement follows then from the fact that the right-hand side of the above equalities is written as a power series in ε.[25])

5.11. Given two by two matrices A, B such that A has eigenvalues $\lambda \neq \mu$, show that the matrix $A_\varepsilon = A + \varepsilon B$ has eigenvalues $\lambda_\varepsilon, \mu_\varepsilon$ such that $|\lambda_\varepsilon - \lambda| \leq C\varepsilon\|B\|$ and $|\mu_\varepsilon - \mu| \leq C\varepsilon\|B\|$. Compute C. (Hint: By Problem 5.10 we know that $\lambda_\varepsilon, \mu_\varepsilon$ are differentiable function of ε and the same holds for the corresponding eigenvector $v_\varepsilon, \tilde{v}_\varepsilon$. Let us discuss λ_ε since the other eigenvalues can be treated in the same way. One possibility is to use the above formula for $\lambda_\varepsilon\Pi_\varepsilon$ to obtain the wanted estimates.

In alternative, let v, w, $\langle w, v\rangle = 1$ and $\|v\| = 1$, be the eigenvectors of A, with eigenvalue λ and of A^*, with eigenvalue $\bar{\lambda}$, respectively. Hence $\Pi_0 =$

[24] Of course, for matrices one could argue more directly by looking at the characteristic polynomial. Yet the strategy below has the advantage to work even in infinitely many dimensions (that is, for operators over Banach spaces).

[25] This is a very simple case of the very general problem of perturbation of point spectrum, see [Kat66] if you want to know more.

$v\otimes w$ and $\|\Pi_0\| = \|w\|$. Normalize v_ε such that $\langle v_\varepsilon, w\rangle = 1$. Differentiate then the above constraint and the defining equation $(A+\varepsilon B)v_\varepsilon = \lambda_\varepsilon v_\varepsilon$, obtaining (the prime refers to the derivative with respect to ε)

$$\begin{aligned} &Av'_\varepsilon + Bv_\varepsilon + \varepsilon Bv'_\varepsilon = \lambda'_\varepsilon v_\varepsilon + \lambda_\varepsilon v'_\varepsilon \\ &\langle v'_\varepsilon, w\rangle = 0. \end{aligned}$$

Multiplying the first for w yields $\lambda'_\varepsilon = \langle w, Bv_\varepsilon\rangle + \varepsilon\langle w, Bv'_\varepsilon\rangle$. Setting $\tilde{A} := A - \lambda\Pi_0$ we have

$$v'_\varepsilon = (\lambda - \tilde{A})^{-1}\left[Bv_\varepsilon + \varepsilon Bv'_\varepsilon - \lambda'_\varepsilon v_\varepsilon - (\lambda - \lambda_\varepsilon)v'_\varepsilon\right].$$

Next, consider ε_0 such that, for $\varepsilon < \varepsilon_0$ holds

$$\|v'_\varepsilon\| \le 4\|(\lambda - \tilde{A})^{-1}\|\,\|B\|\,\|w\| = 4\|(\lambda - \tilde{A})^{-1}\|\,\|B\|\,\|\Pi_0\| =: C_0, \tag{5.8.16}$$

then $\|v_\varepsilon - v\| \le \varepsilon C_0$ and $|\lambda'_\varepsilon| \le \|B\|\,\|w\|(1 + 2\varepsilon C_0)$. If $4\varepsilon_0 C_0 < 1$, then, indeed, (5.8.16) holds true.)

5.12. Compute $D_0 T$. (Hint: solve (5.3.10) for $\varepsilon = 0$, $\theta = \pi$, $p = 0$ and $t = \frac{2\pi}{\omega}$.)

5.13. Compute $D_0 T_\varepsilon$ and see that, if ω is sufficiently large, the eigenvalues have modulus one (the unstable point becomes stable!). (Hint: setting $\xi := \xi_1$ equation (5.3.10) yields $\ddot{\xi} = \omega_p^2\xi + \varepsilon\frac{\omega^2}{l}\cos\omega t\xi$. It is then convenient to write $\xi := \bar{\xi} + \varepsilon\eta + \varepsilon^2\zeta$ where $\ddot{\bar{\xi}} = \omega_p^2\bar{\xi}$ and $\ddot{\eta} = \omega_p^2\eta + \frac{\omega^2}{l}\cos\omega t\,\bar{\xi}$. One can look for a solution of the latter equation of the form

$$\bar{\eta} = Ae^{\omega_p t}\cos\omega t + Be^{\omega_p t}\sin\omega t + Ce^{-\omega_p t}\cos\omega t + De^{-\omega_p t}\sin\omega t.$$

This allows to compute $D_0 T_\varepsilon(\alpha,\beta) = (\xi_1(\frac{2\pi}{\omega}), \xi_2(\frac{2\pi}{\omega})) + \mathcal{O}(\varepsilon^2)$, where $(\xi_1(0), \xi_2(0)) = (\alpha, \beta)$. Finally, one can verify that, for ε small and ω large enough the eigenvalues of $D_0 T_\varepsilon$ are imaginary, hence the equilibrium is linearly stable.)

5.14. Given an Hamiltonian $H : \mathbb{R}^2 \to \mathbb{R}$, for each solution $x(t)$ of the associated equations of motion show that $\langle \nabla_{x(t)} H, \dot{x}(t)\rangle = 0$.

5.15. Compute the following integrals (5.6.15):

$$\int_{\mathbb{R}} e^{iat}(\cosh t)^{-n}\sinh t\,dt,$$

$a \in \mathbb{R}$ and $n \in \mathbb{N}$, $n > 1$.[26] (Hint: By a change of variable, one can consider only the case $a > 0$. Consider the integral on the complex

[26] The result, for $a > 0$, is:

$$\int_{\mathbb{R}} e^{iat}(\cosh t)^{-n}\sinh t = 2\pi i\sum_{k=0}^{\infty}\frac{\phi_{n,k}^{(n-1)}(i\frac{2k+1}{2}\pi)}{(n-1)!},$$

plane, show that the integral on the half circle $Re^{i\phi}$, $\phi \in [0, \pi]$, goes to zero as $R \to \infty$, then check that the poles of the integrand, on the complex plane, lie on the imaginary axis, finally use the residue theorem to compute the integrals.)

5.16. Do the same analysis carried out for the pendulum with a vibrating suspension point in the case of a pendulum subject to an external force $\varepsilon \cos \omega t$ and in presence of a small friction $-\varepsilon^2 \gamma \dot{\theta}$.

Notes

As already mentioned in the text, the first to realize that the motions arising from differential equations can be very complex was probably Poincaré [Poi87]. At that time, the main problem in celestial mechanics (the famous n-body problem) was to find all the integrals of motion. Dirichlet and Weierstrass worked on this problem, but Poincaré was the first to raise serious doubt on the existence of such integrals (which would have implied regular motions). For more historical remarks, see [Mos01]. In fact, all the content of this chapter is inspired by the more sophisticated, but more qualitative, analysis in [Mos01].

---

where

$$\phi_{n,k}(z) = e^{iza} \sinh z \left(\frac{z - i\frac{2k+1}{2}\pi}{\cosh z} \right)^n .$$

For $n = 3$, the above formula yields

$$\int_{\mathbb{R}} e^{iat} (\cosh t)^{-3} \sinh t = \pi a^2 e^{-\frac{\pi}{2}a} (1 - e^{-\pi a})^{-1}.$$

Chapter 6

Qualitative statistical properties: general facts

From the previous chapter, we learned that long-time predictions may be impossible even for seemingly simple Dynamical Systems. Yet, surprisingly, it is exactly such an unpredictability that makes statistical predictions possible. In this chapter, we explain how to make sense of sentences like: *such and such will happen with probability p.*

For simplicity, we will mainly consider Discrete Dynamical Systems, even though we will briefly comment on flows.

6.1 Dynamical systems

Before diving into the specific situation we aim to discuss, let us make a few general comments on the general concept of dynamical systems. This can be stated in very general terms, to give an example, let us consider the following definition

Definition 6.1.1 *By Dynamical Systems we mean the action of a group G on a set X. More precisely, il $M(X)$ is the group of morphisms of X where the group operation is the composition, then the action of G on X is simply a group homomorphism f from G to X.*

The dynamics is given by $x \to f(g)(x)$ for $x \in X$ and $g \in G$. To substantiate this very abstract definition, let us give some examples.

1. $G = \mathbb{N}$, $X = \mathbb{Z}$, $f(n)(z) = z + n$.
2. $G = \mathbb{N}$, $X = \mathbb{Z}^N$, $f(n)(z)_i = z_{i+\omega_i n}$, for some $\omega \in \mathbb{Z}^N$.

3. $G = \mathbb{Z}$, $X = \mathbb{T}$, $f(n)(x) = x + \omega n \mod 1$, for some $\omega \in \mathbb{R}$.
4. $G = \mathbb{Z}$, $X = \mathcal{C}^0(\mathbb{T})$, $f(n)(g)(x) = g(x + \omega n)$, for some $\omega \in \mathbb{R}$.
5. $G = \mathbb{Z}$, $X = (\mathcal{C}^0)'(\mathbb{T})$, $\int_{\mathbb{T}} h(x)[f(n)(\mu)](dx) = \int h(x + \omega n)\mu(dx)$, for some $\omega \in \mathbb{R}$.
6. $G = \mathbb{N}$, $X = \mathbb{R}^N$, $f(n)(v) = A^n v$, for some $N \times N$ matrix A.
7. $G = \mathbb{R}$, $X = \mathbb{C}^{2N}$, $f(n)(v) = e^{At} v$, for some $N \times N$ matrix A.
8. $G = \mathbb{Z}^N$, $X = \mathbb{Z}^N$, $f(n)(z)_i = z_{i+n_i}$.
9. $G = SL(2, \mathbb{Z})$, $X = SL(2, \mathbb{Z})$, $f(g)(x) = gx$.

Note that 3, 4, 5 are the same dynamical system, seen from different points of view: the first follows the evolution of points, the second of observables, and the last of measures. In the case in which $A_{i,j} \geq 1$ and $\sum_i A_{i,j} = 1$, example 6 describes a Markov chain. In the case in which A is self-adjoint, $A = A^*$, example 7 describes a system of N quantum spins. In 8 we can interpret G as translations, then the dynamical system can be used to describe a classical static mechanical spine model. The last model is the left multiplication by a group element, an operation that is extremely common in many mathematical fields.

Here we will restrict to the case in which $G \in \{\mathbb{N}, \mathbb{Z}, \mathbb{R}_+, \mathbb{R}\}$ and X is often a Riemannian manifold. In the following, we will be interested in the point of view illustrated by example 5: the evolution of measures. Namely, let X be a topological space, $\mathcal{M}_1(x)$ the space of its probability Borel measures, and $f : X \to X$ a measurable map. Then we are interested in $\mu \to f_*\mu$, where $f_*\mu(A) = \mu(f^{-1}(A))$ for all measureble sets $A \subset X$.
Note that $f_*\delta_x = \delta_{f(x)}$, so the dynamical system $(\mathcal{M}_1(x), f_*)$ contains, as a invariant subset, the dynamical system (X, f). It thus seems quite useless to look at the infinite-dimensional system $(\mathcal{M}_1(x), f_*)$ if we are ultimately interested in the finite-dimensional dynamical system (X, f), which is often the case. Nevertheless, $(\mathcal{M}_1(x), f_*)$ has at least two critical advances

1. the map f_* is linear while f is often nonlinear. It is well known that trading infinite dimensions for linearity is often a good bargain.
2. $(\mathcal{M}_1(x), f_*)$ contains other important invariant sets that bring a new perspective on the properties of (X, f).

In the following, we will exemplify point (2) above. Suppose that $\mu_* \in \mathcal{M}_1(X)$ has the property that $f_*\mu_* \ll \mu_*$, that is, f is a regular map with respect to μ_*.

Lemma 6.1.2 *If $f_*\mu_* \ll \mu_*$, then the set $\mathcal{M}_* = \{\mu \in \mathcal{M}_1(X) \,:\, \mu \ll \mu_*\}$ is a forward invariant set (that is $f_*\mathcal{M}_* \subset \mathcal{M}_*$).*

PROOF. Let $\mu \in \mathcal{M}_*$, then we can write $d\mu = hd\mu_*$ for some $h \in L^1(X, \mu_*)$. Thus, for each $\varphi \in L^\infty(X, \mu_*)$,

$$\begin{aligned} \left|\int_X \varphi df_*\mu\right| = \left|\int_X \varphi \circ f d\mu\right| = \left|\int_X \varphi \circ f h d\mu_*\right| \\ \leq \|\varphi\|_{L^\infty(X,\mu_*)} \|h\|_{L^1(X,\mu_*)}. \end{aligned} \tag{6.1.1}$$

This implies that $f_*\mu \ll \mu_*$. If not there would exists a measurable set A such that $f_*\mu(A) > 0$ but $\mu_*(A) = 0$, which would lead to a contradiction choosing $\varphi = \mathbb{1}_A$ in (6.1.1). □

The above lemma implies that for all $h \in L^1(X, \mu_*)$ there exists $h_1 \in L^1(X, \mu_*)$ such that, setting $d\mu = hd\mu_*$ and $df_*\mu = h_1 d\mu_*$. Clearly the relation between h and h_1 is lienar, hence the exists a linear operator $\mathcal{L} : d\mu_0 = L^1(X, \mu_*) \to L^1(X, \mu_*)$ such that $h_1 = \mathcal{L}h$. In additon, equation (6.1.1) implies that $\mathcal{L}$ is an L^1 isometry. We have thus obtained the new dynamical system $(L^1(X, \mu_*), \mathcal{L})$ which describes the evolution of the densities.

Remark 6.1.3 *If μ_* is non atomic, the dynamical system $(L^1(X, \mu_*), \mathcal{L})$ is not trivially reducible to (X, f). We will see that it enlightens interesting and non-trivial properties of the dynamics.*

The operator $\mathcal{L}$ has some general properties that will be useful in the following.

Lemma 6.1.4 *The operator $\mathcal{L}$ is positive (sends positive functions in positive functions) and, for all $\varphi \in L^\infty(X, \mu_*)$ and $h \in L^1(X, \mu_*)$,*

$$\varphi \mathcal{L}h = \mathcal{L}(\varphi \circ T\, h).$$

PROOF. If $\varphi, h \geq 0$, $\varphi \in L^\infty(X, \mu_*)$ and $h \in L^1(X, \mu_*)$, then $\varphi \circ T \geq 0$ and

$$0 \leq \int \varphi \circ T h d\mu = \int_X \varphi \mathcal{L}h d\mu$$

which implies $\mathcal{L}h \geq 0$, μ-a.s. . Indeed, if $A = \{x \in X \,:\, \mathcal{L}h(x) < 0\}$ then $\int_X \mathbb{1}_A \mathcal{L}h d\mu < 0$, unless $\mu(A) = 0$. Next, we have, for all $\psi \in L^\infty(X, \mu_*)$,

$$\int_X \psi\varphi\mathcal{L}h d\mu_* = \int_X \psi \circ T\, (\varphi \circ T h)\, d\mu_* = \int_X \psi \mathcal{L}(\varphi \circ T h)$$

from which the Lemma follows by the arbitrariness of ψ. □

The first object that one typically studies when presented with a new dynamic system are fixed points. To start with, note that if $\mathcal{L}h = h$, then $d\mu_*$ is an invariant measure since

$$\int_X \varphi \circ T h d\mu_* = \int_X \varphi \mathcal{L} h d\mu_* = \int_X \varphi h d\mu_*.$$

However, it may happen that $\mathcal{L}h = h$ has no solution in $L^1(X, \mu_*)$. Consider for example $X = [-1, 1]$, $f(x) = \frac{x}{2}$ and μ_* be the Lebesgue measure. Then if μ is an invariant probability measure, we have, for each $\varphi \in \mathcal{C}^0([-1, 1], \mathbb{R})$,

$$\int_{-1}^{1} \varphi(x)\mu(dx) = \lim_{n\to\infty} \int_{-1}^{n} \varphi(f^n(x))\mu(dx) = \varphi(0).$$

That is, the only invariant measure is δ_0 which does no belog to $L^1([-1, 1], \mathbb{R})$.

6.2 Measurable Dynamical Systems

We will start considering the case in which $(L^1(X, \mu_*), \mathcal{L})$ has a fixed point h, $\mathcal{L}h = h$. This means that the measure $d\mu = h d\mu_*$ is invariant. It is then natural to consider μ as the reference measure. The idea to consider (X, T) together with the measure μ naturally leads to the notion of a *Measurable Dynamical System*.

Definition 6.2.1 *By a Measurable Dynamical System with discrete time, we mean a triplet (X, T, μ) where X is a measurable space,[1] μ is a measure and T is a measurable map from X to itself that preserves the measure (i.e., $\mu(T^{-1}A) = \mu(A)$ for each measurable set $A \subset X$).*

An equivalent characterization of invariant measure is $\mu(f \circ T) = \mu(f)$ for each $f \in L^1(X, \mu)$ since, for each measurable set A, $\mu(\chi_A \circ T) = \mu(\chi_{T^{-1}A}) = \mu(T^{-1}A)$, where χ_A is the characteristic function of the set A.

Remark 6.2.2 *In the following we will always assume $\mu(X) < \infty$ (and quite often $\mu(X) = 1$, i.e. μ is a probability measure). Nevertheless, the reader should be aware that there exists a very rich theory pertaining to the case $\mu(X) = \infty$, see [Aar97].*

Definition 6.2.3 *By Dynamical System with continuous time we mean a triplet (X, ϕ^t, μ) where X is a measurable space, μ is a measure and ϕ^t is a measurable group ($\phi^t(x)$ is a measurable function for each t, $\phi^t(x)$ is a measurable function of t for almost all $x \in X$; $\phi^0 =$identity and $\phi^t \circ \phi^s = \phi^{t+s}$ for each $t, s \in \mathbb{R}$) or semigroup ($t \in \mathbb{R}^+$) from X to itself that preserves the measure (i.e., $\mu((\phi^t)^{-1}A) = \mu(A)$ for each measurable set $A \subset X$).*

[1] By measurable space we simply mean a set X together with a σ-algebra that defines the measurable sets.

The above definitions are very general, this reflects the wideness of the field of Dynamical Systems. In the present book we will be interested in much more specialized situations.

In particular, X will always be a topological compact space. The measures will alway belong to the class $\mathcal{M}^1(X)$ of Borel probability measures on X.[2] For future use, given a topological space X and a map T let us define $\mathcal{M}_T$ as the collection of all Borel measures that are T invariant.[3]

Often X will consist of finite unions of smooth manifolds (eventually with boundaries). Analogously, the dynamics (the map or the flow) will be smooth in the interior of X.

Let us see few examples to get a feeling of how a Dynamical System can look like.

6.2.1 Examples

Rotations

Let $\mathbb{T}$ be $\mathbb{R}$ mod 1. By this we mean $\mathbb{R}$ quotiented with respect to the equivalence relations $x \sim y$ if and only if $x - y \in \mathbb{Z}$. $\mathbb{T}$ can be though as the interval $[0, 1]$ with the points 0 and 1 identified. We put on it the topology induced by the topology of $\mathbb{R}$ via the defined equivalence relation. Such a topology is the usual one on $[0, 1]$, apart from the fact that each open set containing 0 must contain 1 as well. Clearly, from the topological point of view, $\mathbb{T}$ is a circle. We choose the Borel σ-algebra. By μ we choose the Lebesgue measure m, while $T : \mathbb{T} \to \mathbb{T}$ is defined by

$$Tx = x + \omega \mod 1,$$

for some $\omega \in \mathbb{R}$. In essence, T translates, or rotates, each point by the same quantity ω. It is easy to see that the measure μ is invariant (Problem 6.9).

Bernoulli shift

A Dynamical System needs not live on some differentiable manifold, more abstract possibilities are available.

Let $\mathbb{Z}_n = \{1, 2, ..., n\}$, then define the set of two sided (or one sided) sequences $\Sigma_n = \mathbb{Z}_n^{\mathbb{Z}}$ ($\Sigma_n^+ = \mathbb{Z}_n^{\mathbb{Z}_+}$). This means that the elements of Σ_n are sequences $\sigma = \{..., \sigma_{-1}, \sigma_0, \sigma_1, ...\}$ ($\sigma = \{\sigma_0, \sigma_1, ...\}$ in the one sided case) where $\sigma_i \in \mathbb{Z}_n$. To define the measure and the σ-algebra a bit of care is necessary. To start with, consider the *cylinder sets*, that is the sets of the form

$$A_i^j = \{\sigma \in \Sigma_n \mid \sigma_i = j\}.$$

[2] Remember that a Borel measure is a measure defined on the Borel σ-algebra, that is the σ-algebra generated by the open sets.

[3] Obviously, for each $\mu \in \mathcal{M}_T$, (X, T, μ) is a Dynamical System.

Such sets will be our basic objects and can be used to generate the algebra $\mathcal{A}$ of the cylinder sets via unions and complements (or, equivalently, intersections and complements). We can then define a topology on Σ_n (the product topology, if $\{1,\dots,n\}$ is endowed by the discrete topology) by declaring the above algebra made of open sets and a basis for the topology. To define the σ-algebra we could take the minimal σ-algebra containing $\mathcal{A}$, yet this it is not a very constructive definition, neither a particular useful one, it is better to invoke the Carathèodory construction.

Let us start by defining a measure on $\mathbb{Z}_n$, that is n numbers $p_i > 0$ such that $\sum_{i=1}^n p_i = 1$. Then, for each $i \in \mathbb{Z}$ and $j \in \mathbb{Z}_n$,

$$\mu(A_i^j) = p_j.$$

Next, for each collection of sets $\{A_{i_l}^{j_l}\}_{l=1}^s$, with $i_l \neq i_k$ for each $l \neq k$, we define

$$\mu(A_{i_1}^{j_1} \cap A_{i_2}^{j_2} \cap \dots \cap A_{i_s}^{j_s}) = \prod_{l=1}^{s} p_{j_l}.$$

We now know the measure of all finite intersection of the sets A_i^j. Obviously $\mu(A^c) := 1 - \mu(A)$ and the measure of the union of two sets A, B obviously must satisfy $\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$. We have so defined μ on $\mathcal{A}$. It is easy to check that such a μ is σ-additive on $\mathcal{A}$; namely: if $\{A_i\} \subset \mathcal{A}$ are pairwise disjoint sets and $\cup_{i=1}^\infty A_i \in \mathcal{A}$, then $\mu(\cup_{i=1}^\infty A_i) = \sum_{i=1}^\infty \mu(A_i)$. The next step is to define an outer measure[4]

$$\mu^*(A) := \inf_{\substack{B \in \mathcal{A} \\ B \supset A}} \mu(B) \quad \forall A \subset \Sigma_n.$$

Finally, we can define the σ-algebra as the collection of all the sets that satisfy the *Carathèodory's criterion*, namely A is measurable (that is belongs to the σ-algebra) iff

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c) \quad \forall E \subset \Sigma_n.$$

The reader can check that the sets in $\mathcal{A}$ are indeed measurable.

The Carathèodory Theorem then asserts that the measurable sets form a σ-algebra and that on such a σ-algebra μ^* is numerably additive, thus we have our measure μ (simply the restriction of μ^* to the σ-algebra).[5] The σ-algebra so obtained is nothing else than the completion with respect to μ of the minimal σ-algebra containing $\mathcal{A}$ (all the sets with zero outer measure are measurable).

[4] An outer measure has the following properties: i) $\mu^*(\emptyset) = 0$; ii) $\mu^*(A) \leq \mu^*(B)$ if $A \subset B$; iii)$\mu^*(\cup_{i=1}^\infty A_i) \leq \sum_{i=1}^\infty \mu^*(A_i)$. Note that μ^* need not be additive on all sets.

[5] See [LL01] if you want a quick look at the details of the above Theorem or consult [Roy88] if you want a more in depth immersion in measure theory. If you think that the above construction is too cumbersome see Problem 6.19.

The map $T : \Sigma_n \to \Sigma_n$ (usually called *shift*) is defined by

$$(T\sigma)_i = \sigma_{i+1}.$$

We leave to the reader the task to show that the measure is invariant (see Problem 6.17).

To understand what's going on, let us consider the function $f : \Sigma \to \mathbb{Z}_n$ defined by $f(\sigma) = \sigma_0$. If we consider T^t, $t \in \mathbb{N}$, as the time evolution and f as an observation, then $f(T^t\sigma) = \sigma_t$. This can be interpreted as the observation of some phenomenon at various times. If we do not know anything concerning the state of the system, then the probability to see the value j at the time t is simply p_j. If $n = 2$ and $p_1 = p_2 = \frac{1}{2}$, it could very well be that we are observing the successive outcomes of tossing a fair coin where 1 means head and 2 tail (or vice versa); if $n = 6$ it could be the outcome of throwing a dice and so on.

Dilation

Again $X = \mathbb{T}$ and the measure is Lebesgue. T is defined by

$$Tx = 2x \mod 1.$$

This map it is not invertible (similarly to the one sided shift). Note that, in general, $\mu(TA) \neq \mu(A)$ (e.g., $A = [0, \frac{1}{2}]$).

Toral automorphism (Arnold cat)

This is an automorphism of the torus and gets its name by a picture draw by Arnold [AA68]. The space X is the two dimensional torus $\mathbb{T}^2$. The measure is again Lebesgue measure and the map is

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \mod 1 := L\begin{pmatrix} x \\ y \end{pmatrix} \mod 1.$$

Since the entries of L are integers numbers it is clear that T is well defined on the torus; in fact, it is a linear toral automorphism. The invariance of the measure follows from $\det L = 1$.

Hamiltonian Systems

Up to now we have seen only examples with discrete time. Typical examples of Dynamical Systems with continuous time are the solutions of an ODE or a PDE. Let us consider the case of an Hamiltonian system. The simplest case is when $X = \mathbb{R}^{2n}$, the σ-algebra is the Borel one and the measure μ is the Lebesgue

measure m. The dynamics is defined by a smooth function $H : X \to \mathbb{R}$ via the equations

$$\frac{dx}{dt} = J\text{grad}H(x)$$

where $\text{grad}(H)_i = (\nabla H)_i = \frac{\partial H}{\partial x_i}$ and J is the block matrix

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

The fact that m is invariant with respect to the Hamiltonian flow is due to the Liouville Theorem (see [Arn99] or Problem 5.7).

Such a dynamical system has a natural decomposition. Since H is an integral of the motion, for each $h \in \mathbb{R}$ we can consider $X_h = \{x \in X \mid H(x) = h\}$. If $X_h \neq \emptyset$, then it will typically consist of a smooth manifold,[6] let us restrict ourselves to this case. Let σ be the surface measure on X_h, then $\mu_h = \frac{\sigma}{\|\text{grad}H\|}$ is an invariant measure on X_h and (X_h, ϕ_t, μ_h) is a Dynamical System (see Problem 6.11).

Geodesic flow

Along the same lines any geodesic flow on a compact Riemannian manifold naturally defines a dynamical system.

6.3 Return maps and Poincaré sections

Normally in Dynamical Systems there is a lot of emphasis on the discrete case. One reason is that there is a general device that allows to reduce the study of many properties of a continuous time Dynamical System to the study of an appropriate discrete time Dynamical System: Poincaré sections (we have already seen an instance of this in the introduction). Here we want to make few comments on this precious tool that we will largely employ in the study of billiards.

Let us consider a smooth Dynamical System (X, ϕ^t, μ) (that is a Dynamical Systems in continuous time where X is a smooth manifold and ϕ^t is a smooth flow). Then we can define the vector field $V(x) := \frac{d\phi^t(x)}{dt}|_{t=0}$.[7]

Consider a smooth compact submanifold (possibly with boundaries) Σ of codimension one such that $\mathcal{T}_x\Sigma$ (the tangent space of Σ at the point x) is

[6] By the implicit function theorem this is locally the case if $\nabla H \neq 0$.

[7] Very often it is the other way around: the vector field is given first and then the flow–as we saw in the introduction.

transversal to $V(x)$.[8] We can then define the *return time* $\tau_\Sigma : \Sigma \to \mathbb{R}^+ \cup \{\infty\}$ by

$$\tau_\Sigma = \inf\{t \in \mathbb{R}^+ \backslash \{0\} \mid \phi^t(x) \in \Sigma\},$$

where the inf is taken to be ∞ if the set is empty. Next we define the *return map* $T_\Sigma : D(T) \subset \Sigma \to \Sigma$, where $D(T) = \{x \in \Sigma | \tau_\Sigma(x) < \infty\}$, by

$$T_\Sigma(x) = \phi^{\tau_\Sigma(x)}(x).$$

It is easy to check that there exists $c > 0$ such that $\tau_\Sigma \geq c$ (Problem 6.14).

To define the measure, the natural idea is to project the invariant measure along the flow direction: for all measurable sets $A \subset \Sigma$, define[9]

$$\nu_\Sigma(A) := \lim_{\delta \to 0} \frac{1}{\delta} \mu(\phi^{[0,\,\delta]}(A)). \tag{6.3.2}$$

See Problem 6.13 for the existence of the above limit; see Problem 6.14 for the proof that τ_Σ is finite almost everywhere and Problem 6.15 for the proof that $(\Sigma, T_\Sigma, \nu_\Sigma)$ is a dynamical system. The reader is invited to meditate on the relation between this Dynamical System and the original one.

6.4 Suspension flows

A natural question is if it is possible to construct a flow with a given Poincaré section, the answer is that there are infinitely many flows with a given section. Let us construct some of them. Given a dynamical system (Σ, T, ν) consider $\tilde{X} := \Sigma \times R^+$. Define the flow $\phi_t((x, s)) = (x, s + t)$. We then define in $\tilde{X}$ the equivalence relation $(x, t) \sim (y, s)$ iff $s = t + n$ and $y = T^n x$ or $t = s + n$ and $x = T^n y$ for some $n \in \mathbb{N}$. A moment of reflection shows that the set X of equivalence classes is nothing else than the set $\Sigma \times [0, 1]$ with the points $(x, 1)$ and $(Tx, 0)$ identified. Clearly the flow is naturally quotiented over the equivalence classes and yields a quotient flow on X, such a flow is called a *suspension flow*.

A more general construction can by obtained by applying a time change to the above example. Alternatively, one can can choose any smooth function $\tau : \Sigma \to \mathbb{R}^+$, that will be called a *ceiling function* and consider the set $X_\tau = \{(x, t) \in \Sigma \times \mathbb{R}^+ \mid t \in [0, \tau(x)]\}$ with the points $(x, \tau(x))$ and $(Tx, 0)$ identified. A moment of reflection should show that the topology of X_τ does not depend on τ and is then the same than the suspension defined above. The flow is again defined by $\phi_t(x, s) = (x, s + t)$ for $t \leq \tau(x) - s$. Such flows are called *special flows*.

[8] That is $\mathcal{T}_x\Sigma \oplus V(x)$ form the full tangent space at x.

[9] We use the notation: $\phi^I(A) := \cup_{t \in I} \phi^t(A)$ for each $I \subset \mathbb{R}$.

6.5 Invariant measures

A very natural question is: given a space X and a map T does there always exists an invariant measure μ? A non exhaustive, but quite general, answer exists: Krylov-Bogoluvov Theorem.

First of all we need a useful characterization of invariance.

Lemma 6.5.1 *Given a compact metric space X and Borel measurable map T continuous apart from a compact set K,*[10] *a Borel measure μ, such that $\mu(K) = 0$, is invariant if and only if $\mu(f \circ T) = \mu(f)$ for each $f \in \mathcal{C}^0(X)$.*

Proof. To prove that the invariance of the measure implies the invariance for continuous functions is obvious, since each such function can be approximated uniformly by simple functions–that is, a sum of characteristic functions of measurable sets–for which the invariance is immediate.[11] The converse implication is not so obvious.

The first thing to remember is that the Borel measures, on a compact metric space, are regular [RS80]. This means that for each measurable set A the following holds[12]

$$\mu(A) = \inf_{\substack{G \supset A \\ G = \mathring{G}}} \mu(G) = \sup_{\substack{C \subset A \\ C = \overline{C}}} \mu(C). \tag{6.5.3}$$

Next, remember that for each closed set A and open set $G \supset A$, there exists $f \in \mathcal{C}^0(X)$ such that $f(X) \subset [0,1]$, $f|_{G^c} = 0$ and $f|_A = 1$ (this is Urysohn Lemma for Normal spaces [Roy88]). Hence, setting $B_A := \{f \in \mathcal{C}^{(0)}(X) \mid f \geq \chi_A\}$,

$$\mu(A) \leq \inf_{f \in B_A} \mu(f) \leq \inf_{\substack{G \supset A \\ G = \mathring{G}}} \mu(G) = \mu(A). \tag{6.5.4}$$

Accordingly, for each A closed, we have

$$\mu(T^{-1}A) \leq \inf_{f \in B_A} \mu(f \circ T) = \inf_{f \in B_A} \mu(f) = \mu(A).$$

In addition, using again the regularity of the measure, for each A Borel holds[13]

[10] This means that, if $C \subset X$ is closed, then $T^{-1}C \cup K$ is closed as well.

[11] This is essentially the definition of integral.

[12] This is rather clear if one thinks of the Carathéodory construction starting from the open sets.

[13] Note that, by hypothesis, if C is compact and $C \cap K = \emptyset$, then TC is compact.

$$\begin{aligned}\mu(T^{-1}A) &= \sup_{\substack{U\supset K\\ U=\mathring{U}}} \mu(T^{-1}A\backslash U) \leq \sup_{\substack{U\supset K\\ U=\mathring{U}}} \sup_{\substack{C\subset T^{-1}A\backslash U\\ C=\overline{C}}} \mu(T^{-1}(TC))\\ &\leq \sup_{\substack{U\supset K\\ U=\mathring{U}}} \sup_{\substack{C\subset A\\ C=\overline{C}}} \mu(T^{-1}C) \leq \sup_{\substack{C\subset A\\ C=\overline{C}}} \mu(C) = \mu(A).\end{aligned}$$

Applying the same argument to the complement A^c of A it follow that it must be $\mu(T^{-1}A) = \mu(A)$ for each Borel set. □

Proposition 6.5.2 (Krylov–Bogoluvov) *If X is a metric compact space and $T : X \to X$ is continuous, then there exists at least one invariant (Borel) measure.*

Proof. Consider any Borel probability measure ν and define the following sequence of measures $\{\nu_n\}_{n\in\mathbb{N}}$:[14] for each Borel set A

$$\nu_n(A) = \nu(T^{-n}A).$$

The reader can easily see that $\nu_n \in \mathcal{M}^1(X)$, the sets of the probability measures. Indeed, since $T^{-1}X = X$, $\nu_n(X) = 1$ for each $n \in \mathbb{N}$. Next, define

$$\mu_n = \frac{1}{n}\sum_{i=0}^{n-1} \nu_i.$$

Again $\mu_n(X) = 1$, so the sequence $\{\mu_i\}_{i=1}^{\infty}$ is contained in a weakly compact set (the unit ball) and therefore admits a weakly convergent subsequence $\{\mu_{n_i}\}_{i=1}^{\infty}$; let μ be the weak limit.[15] We claim that μ is T invariant. Since μ is a Borel measure it suffices to verify that for each $f \in \mathcal{C}^0(X)$ holds $\mu(f \circ T) = \mu(f)$ (see Lemma 6.5.1). Let f be a continuous function, then by the weak convergence we have[16]

[14] Intuitively, if we chose a point $x \in X$ at random, according to the measure ν and we ask what is the probability that $T^n x \in A$, this is exactly $\nu(T^{-n}A)$. Hence, our procedure to produce the point $T^n x$ is equivalent to picking a point at random according to the evolved measure ν_n.

[15] This depends on the Riesz-Markov Representation Theorem [RS80] that states that $\mathcal{M}(X)$ is exactly the dual of the Banach space $\mathcal{C}^0(X)$. Since the weak convergence of measures in this case correspond exactly to the weak-* topology [RS80], the result follows from the Banach-Alaoglu theorem stating that the unit ball of the dual of a Banach space is compact in the weak-* topology. But see 1.6.22 if you want a more elementary proof.

[16] Note that it is essential that we can check invariance only on continuous functions: if we would have to check it with respect to all bounded measurable functions we would need that μ_n converges in a stronger sense (*strong convergence*) and this may not be true. Note as well that this is the only point where the continuity of T is used: to insure that $f \circ T$ is continuous and hence that $\mu_{n_j}(f \circ T) \to \mu(f \circ T)$.

$$\mu(f\circ T)=\lim_{j\to\infty}\frac{1}{n_j}\sum_{i=0}^{n_j-1}\nu_i(f\circ T)=\lim_{j\to\infty}\frac{1}{n_j}\sum_{i=0}^{n_j-1}\nu(f\circ T^{i+1})$$
$$=\lim_{j\to\infty}\frac{1}{n_j}\left\{\sum_{i=0}^{n_j-1}\nu_i(f)+\nu(f\circ T^{n_j})-\nu(f)\right\}=\mu(f).$$

□

The reason why the above theorem is not completely satisfactory is that it is not constructive and, in particular, does not provide any information on the nature of the invariant measure. On the contrary, in many instances the interest is focused not just on any Borel measure but on special classes of measures, for example measures connected to the Lebesgue measure which, in some sense, can be thought as reasonably physical measures (if such measures exists).

In the following examples we will see two main techniques to study such problems: on the one hand it is possible to try to construct explicitly the measure and study its properties in the given situations (expanding maps, strange attractors, solenoid, horseshoe); on the other hand one can try to *conjugate*[17] the given problem with another, better understood, one (logistic map, circle maps). In view of the second possibility the last example is very important (Markov measures). Such an example gives just a hint to the possibility to construct a multitude of invariant measures for the shift which, as we will see briefly, is a standard system to which many other can be conjugated.

6.5.1 Examples

Contracting maps

Let $X\subset\mathbb{R}^n$ be compact and connected, $T:X\to X$ differentiable with $\|DT\|\leq\lambda^{-1}<1$ and $T0=0\in X$. In this case 0 is the unique fixed point and the delta function at zero is the only invariant measure.[18]

Expanding maps

The simplest possible case is $X=\mathbb{T}$, $T\in\mathcal{C}^2(\mathbb{T})$ with $|DT|\geq\lambda>1$, (see Figure 6.1 for a pictorial example).[19]

[17] See Definition 6.9.2 for a precise definition and Problem 6.41 and 6.42 for some insight.

[18] The reader will hopefully excuse this physicist language, naturally we mean that the invariant measure is defined by $\delta_0(f)=f(0)$. The property that there exists only one invariant measure is called *unique ergodicity*, we will see more of it in the sequel, e.g. see example 6.6.1.

[19] Note that this generalizes Examples 6.2.1.

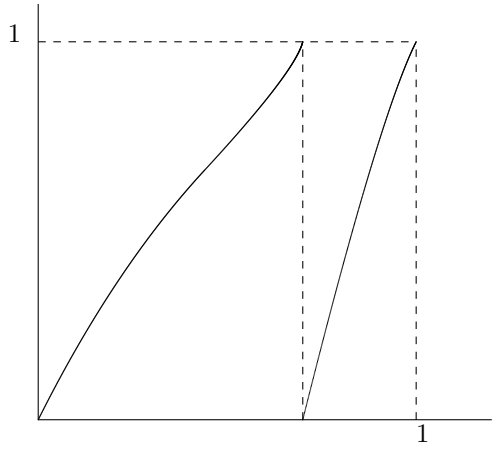

Figure 6.1: Graph of an expanding map on $\mathbb{T}$

We would like to have an invariant measure absolutely continuous with respect to Lebesgue. Any such measure μ has, by definition, the Radon-Nikodym derivative $h = \frac{d\mu}{dm} \in L^1(\mathbb{T}, m)$, [Roy88]. In Proposition 6.5.2 we saw how a measure evolves by defining the operator

$$T_*\mu(f) = \mu(f \circ T) \tag{6.5.5}$$

for each $f \in \mathcal{C}^0$ and $\mu \in \mathcal{M}(X)$ (see also footnote 15 at page 118). If we want to study a smaller class of measures we must first check that T_* leaves such a class invariant. Indeed, if μ is absolutely continuous with respect to Lebesgue then $T_*\mu$ has the same property. Moreover, if $h = \frac{d\mu}{dm}$ and $h_1 = \frac{dT_*\mu}{dm}$ then (Problem 6.20)

$$h_1(x) =: \mathcal{L}h(x) = \sum_{y \in T^{-1}(x)} |D_y T|^{-1} h(y).$$

The operator $\mathcal{L} : L^1(\mathbb{T}, m) \to L^1(\mathbb{T}, m)$ is called *Transfer operator* or Ruelle-Perron-Frobenius operator, and has an extremely important rôle in the study of the statistical properties of the system. Notice that $\|\mathcal{L}h\|_1 \leq \|h\|_1$.[20] The key property of $\mathcal{L}$, in this context, is given by the following inequality (this type of inequality is commonly called of Lasota-York type) (Problem 6.21): if $f' \in L^1$, then

$$\left| \frac{d}{dx} \mathcal{L}h(x) \right| \leq \lambda^{-1} |\mathcal{L}h'(x)| + C|\mathcal{L}h(x)| \tag{6.5.6}$$

[20] Here $\|f\|_1 := \int |h(x)|dx$ is the standard norm in L^1.

where $C = \frac{\|D^2T\|_\infty}{\|DT\|_\infty^2}$.

The above inequality implies $\|(\mathcal{L}h)'\|_1 \leq \lambda^{-1}\|h'\|_1 + C\|h\|_1$. Iterating such a relation yields

$$\|(\mathcal{L}^n h)'\|_1 \leq \lambda^{-n}\|h'\|_1 + \frac{C}{1-\lambda^{-1}}\|h\|_1,$$

for all $n \in \mathbb{N}$. This, in turn, implies that the $\sup_{n\in\mathbb{N}} \|\mathcal{L}^n h\|_\infty < \infty$. Consequently, the sequence $h_n := \frac{1}{n}\sum_{i=0}^{n-1} \mathcal{L}^i h$ is compact in L^1 (this is a consequence of standard embedding theorems[21] [LL01] but see Problem 6.22 for an elementary proof). In analogy with Lemma 6.5.2, we have that there exists $h_* \in L^1$ such that $\mathcal{L}h_* = h_*$. Thus $d\mu := h_* dm$ is an invariant measure of the type we are looking for.

In fact, it is possible to obtain some more information on such measure. Equation (6.5.6) implies that $\mathcal{L}$ is a well defined operator also when restricted to $\mathcal{C}^0$ or $\mathcal{C}^1$. Moreover, for each $h \in \mathcal{C}^0$ and $n \in \mathbb{N}$,

$$\begin{aligned}|\mathcal{L}^n h|_\infty &\leq |\mathcal{L}^n 1|_\infty |h|_\infty \leq |h|_\infty(\|\mathcal{L}^n 1\|_1 + \|(\mathcal{L}^n 1)'\|_1) \leq |h|_\infty \frac{C+1}{1-\lambda^{-1}} \\ &=: C_1|h|_\infty.\end{aligned}$$

Using the above equation and iterating (6.5.6) yields, for each $h \in \mathcal{C}^1$ and $n \in \mathbb{N}$,

$$|(\mathcal{L}^n h)'|_\infty \leq \lambda^{-n} C_1 |h'|_\infty + C_1^2 |h|_\infty.$$

In other words we have a Lasota-Yorke type inequality for $\mathcal{L}$ acting on $\mathcal{C}^0, \mathcal{C}^1$ instead of $L^1, W^{1,1}$. In particular note that one can apply the above inequalities to the average $h_n := \frac{1}{n}\sum_{i=0}^{n-1} \mathcal{L}^i h$, when $h \in \mathcal{C}^1$. Then the compactness follows by Ascoli-Arzelá Theorem and it follows that the invariant density is continuous (in fact, Lipschitz as already argued in the Perron-Frobenius Theorem).

Logistic maps

Consider $X = [0,1]$ and

$$T(x) = 4x(1-x).$$

This map is not an everywhere expanding map ($D_{\frac{1}{2}}T = 0$), yet it can be conjugate with one, [UvN47].

To see this consider the continuous change of variables $\Psi : [0,1] \to [0,1]$ defined by

$$\Psi(x) = \frac{2}{\pi}\arcsin\sqrt{x},$$

[21] Indeed the space $\mathcal{C}^1$ closed with respect to the norm $\|f\| = \|f\|_1 + \|f'\|_1$ is a well known Banach space: the Sobolev space $W^{1,1}$.

thus $\Psi^{-1}(x) = \left(\sin \frac{\pi}{2} x\right)^2$. Accordingly,

$$\begin{aligned} \tilde{T}(x) &:= \Psi \circ T \circ \Psi^{-1}(x) = \Psi(4 \sin^2 \tfrac{\pi}{2} x \cos^2 \tfrac{\pi}{2} x) \\ &= \Psi([\sin \pi x]^2) = \tfrac{2}{\pi} \arcsin[\sin \pi x] \end{aligned}$$

which yields[22]

$$\tilde{T}(x) = \begin{cases} 2x & \text{for } x \in [0, \frac{1}{2}] \\ 2 - 2x & \text{for } x \in [\frac{1}{2}, 1]. \end{cases}$$

The map $\tilde{T}$ is called *tent map* for its characteristic shape, see figure 6.2. What

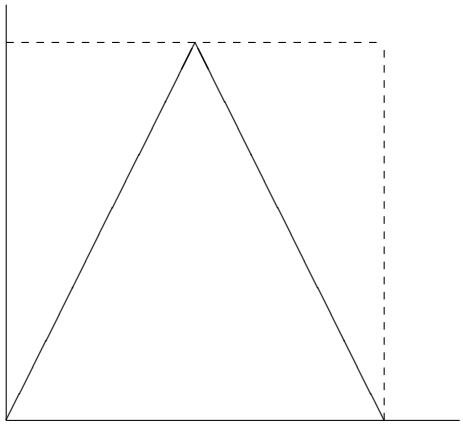

Figure 6.2: Graph of tent map

is more interesting is that the Lebesgue measure is invariant for $\tilde{T}$, as the reader can easily check. This means that, if we define $\mu(f) := m(f \circ \Psi^{-1})$, it holds true

$$\mu(f \circ T) = m(f \circ T \circ \Psi^{-1}) = m(f \circ \Psi^{-1} \circ \tilde{T}) = m(f \circ \Psi^{-1}) = \mu(f).$$

Hence, $([0,1], T, \mu)$ is a Dynamical System. In addition, a trivial computation shows

$$\mu(dx) = \frac{1}{\pi \sqrt{x(1-x)}} dx,$$

thus μ is absolutely continuous with respect to Lebesgue.

[22] Remember that the range of arcsin is $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and $\sin \pi x = \sin \pi(1-x)$.

Circle maps

A circle map is an order preserving continuous map of the circle. A simple way to describe it is to start by considering its lift. Let $\hat{T} : \mathbb{R} \to \mathbb{R}$, such that $\hat{T}(0) \in [0, 1]$, $\hat{T}(x+1) = \hat{T}(x)+1$ ad it is monotone increasing. The circle map is then defined as $T(x) = \hat{T}(x) \mod 1$. Circle maps have a very rich theory that we do not intend to develop here, we confine ourselves to some facts (see [HK95] for a detailed discussion of the properties below). The first fact is that the *rotation number*

$$\rho(T) = \lim_{n\to\infty} \frac{1}{n}\hat{T}^n(x).$$

is well defined and does not depend on x.

We have already seen a concrete example of circle maps: the rotation R_ω by ω. Clearly $\rho(R_\omega) = \omega$. It is fairly easy to see that if $\rho(T) \in \mathbb{Q}$ then the map has a periodic orbit. We are more interested in the case in which the rotation number is irrational. In this case, with the extra assumption that T is twice differentiable (actually a bit less is needed) the Denjoy theorem holds stating that there exists a continuous invertible function h such that $R_{\rho(T)} \circ h = h \circ T$, that is T is *topologically conjugated* to a rigid rotation. Since we know that the Lebesgue measure is invariant for the rotations, we can obtain an invariant measure for T by pushing the Lebesgue measure by h, namely define

$$\mu(f) = m(f \circ h^{-1}).$$

The natural question if the measure μ is absolutely continuous with respect to Lebesgue is rather subtle and depends, once again, on KAM theory. In essence the answer is positive only if T has more regularity and the rotation number is not very well approximated by rational numbers (in some sense it is 'very irrational').

Strange Attractors

We have seen the case in which all the trajectories are attracted by a point. The reader can probably imagine a case in which the attractor is a curve or some other simple set. Yet, it has been a fairly recent discovery that an attractor may have a very complex (strange) structure. The following is probably the simplest example. Let $X = Q = [0, 1]^2$ and

$$T(x,\, y) = \begin{cases} (2x,\, \frac{1}{8}y + \frac{1}{4}) & \text{if } x \in [0,\, 1/2] \\ (2x - 1,\, \frac{1}{8}y + \frac{3}{4}) & \text{if } x \in]1/2,\, 1]. \end{cases}$$

We have a map of the square that stretches in one direction by a factor 2 and contract in the other by a factor 8.

Note that T it is not continuous with respect to the normal topology, so Proposition 6.5.2 cannot be applied directly. This problem can be solved in at

least two ways: one is to *code* the system and we will discuss it later (see Examples 6.9.1), the other is to study more precisely what happens iterating a measure in special cases.

In our situation, since $T^n Q$ consists of a multitude of thinner and thinner strips, it is clear that there can be no invariant measure absolutely continuous with respect to Lebesgue.[23] Yet, it is very natural to ask what happens if we iterate the Lebesgue measure by the operator T_*. It is easy to see that $T_* m$ is still absolutely continuous with respect to Lebesgue. In fact, T_* maps absolutely continuous measures into absolutely continuous measures. Once we note this, it is very tempting to define the transfer operator. An easy computation yields

$$\mathcal{L}h(x) = \chi_{TQ}(x) \sum_{y \in T^{-1}(x)} |\det(D_y T)|^{-1} h(y) = 4\chi_{TQ}(x) h(T^{-1}(x)).$$

Since the map expands in the unstable direction, it is quite natural to investigate, in analogy with the expanding case, the *unstable derivative* D^u, that is the derivative in the x direction, of the iterate of the density.

$$\|D^u \mathcal{L}h\|_1 \leq \frac{1}{2} \|D^u h\|_1 \quad \forall h \in \mathcal{C}^1(Q) \tag{6.5.7}$$

To see the consequences of the above estimate, consider $f \in \mathcal{C}^{(1)}(Q)$ with $f(0,y) = f(1,y) = 0$ for each $y \in [0,1]$, then if μ is a measure obtained by the measure hdm ($h \in \mathcal{C}^1$) with the procedure of Proposition 6.5.2,[24] we have

$$\begin{aligned}\mu(D^u f) &= \lim_{j\to\infty} \frac{1}{n_j} \sum_{i=0}^{n_j-1} (T_*)^i m(h D^u f) = \lim_{j\to\infty} \frac{1}{n_j} \sum_{i=0}^{n_j-1} m(\mathcal{L}^i h D^u f) \\ &= -\lim_{j\to\infty} \frac{1}{n_j} \sum_{i=0}^{n_j-1} m(f D^u \mathcal{L}^i h)\end{aligned}$$

where we have integrated by part. Remembering (6.5.7) we have

$$\mu(D^u f) = 0,$$

for all $f \in \mathcal{C}^{(1)}_{\mathsf{per}}(Q) = \{f \in \mathcal{C}^{(1)}(Q) \mid f(0,y) = f(1,y)\}$. The enlargement of the class of functions is due to the obvious fact that, if $f \in \mathcal{C}^{(1)}_{\mathsf{per}}(Q)$, then $\tilde{f}(x,y) = f(x,y) - f(0,y)$ is zero on the vertical (stable) boundary and $D^u \tilde{f} = D^u f$.

[23] In fact, if μ is an invariant measure, $T_*\mu = \mu$, it follows

$$\mu(\chi_{T^n Q}) = T_*^n \mu(\chi_{T^n Q}) = \mu(\chi_Q) = 1,$$

so μ must be supported on $\Lambda = \cap_{n=0}^{\infty} T^n Q$.

[24] As we noted in the proof of Proposition 6.5.2, the only part that uses the continuity of T is the proof of the invariance. Thus, in general we can construct a measure by the averaging procedure but its invariance is not automatic.

This means that the measure μ, when restricted to the horizontal direction, is μ-a.e. constant (see Problem 6.36). Such a strong result is clearly a consequence of the fact that the map is essentially linear, one can easily imagine a non linear case (think of dilations and expanding maps) and in that case the same argument would lead to conclude that the measure, when restricted to unstable manifolds, is absolutely continuous with respect to the restriction of Lebesgue (these type of measures are commonly called SRB from Sinai, Ruelle and Bowen).

We can now prove that indeed the measure μ is invariant. The discontinuity line of T is $\{x = \frac{1}{2}\}$. Points close to $\{x = \frac{1}{2}\}$ are mapped close to the boundary of Q, so if $f(0, y) = f(1, y) = 0$, then $f \circ T$ is continuous. Hence, the argument of Proposition 6.5.2 proves that $\mu(f \circ T) = \mu(f)$ for all f that vanish at the stable boundary. Yet, the characterization of μ proves that $\mu(\{(x, y) \in Q \mid x \in \{0, 1\}\}) = 0$, thus we can obtain $\mu(f \circ T) = \mu(f)$ for all continuous functions via the Lebesgue dominated convergence theorem and the invariance follows by Lemma 6.5.1.

Horseshoe

This very famous example consists of a map of the square $Q = [0, 1]^2$, the map is obtained by stretching the square in the horizontal direction, bending it in the shape of an horseshoe and then superimposing it to the original square in such a way that the intersection consists of two horizontal strips.[25] Such a description is just topological, to make things clearer let us consider a very special case:

$$T(x,\, y) = \begin{cases} (5x \mod 1,\ \frac{1}{4}y) & \text{if } x \in [1/5,\, 2/5] \\ (5x \mod 1,\ \frac{1}{4}y + \frac{3}{4}) & \text{if } x \in [3/5,\, 4/5]. \end{cases}$$

Note that T is not explicitly defined for $x \in [0,\, 1/5[\cup[\frac{2}{3},\, \frac{3}{5}[\cup]4/5,\, 1]$ since for this values the horseshoe falls outside Q, so its actual shape is irrelevant. Since the map from Q to Q is not defined on the full square, we can have a Dynamical System only with respect to a measure for which the domain of definition of T, and all of its powers, has measure one. We will start by constructing such a measure.

The first step is to notice that the set

$$\Lambda = \cap_{n \in \mathbb{Z}} T^n Q \tag{6.5.8}$$

of the points which trajectories are always in Q is $\neq \emptyset$. Second, note that $\Lambda = T\Lambda = T^{-1}\Lambda$, such an invariant set is called *hyperbolic set* as we will see in ???. We would like to construct an invariant measure on Λ. Since Λ is a compact set and T is continuous on it we know that there exist invariant measures; yet, in

[25] We have already seen something very similar in the introduction.

analogy with the previous examples, we would like to construct one *coming from Lebesgue*.

As already mentioned we must start by constructing a measure on $\Lambda_- = \cap_{n\in\mathbb{N}\cup\{0\}} T^{-n}Q$ since $T^k\Lambda_- \subset \Lambda_-$. To do so it is quite natural to construct a measure by *subtracting* the mass that leaks out of Q. namely, define the operator $\tilde{T} : \mathcal{M}(X) \to \mathcal{M}(X)$ by

$$\tilde{T}\mu(A) := \mu(TA \cap Q).$$

Again we consider the evolution of measures of the type $d\mu = hdm$. For each continuous f with supp$(f) \subset Q$ holds

$$\tilde{T}\mu(f) = \mu(f \circ T^{-1}\chi_Q) = \int_{T^{-1}Q} fh \circ T|\det DT|dm.$$

We can thus define the operator $\mathcal{L}$ that evolves the densities:

$$\mathcal{L}h(x) = \frac{5}{4}\chi_{T^{-1}Q\cap Q}(x)h(Tx).$$

Clearly $\tilde{T}\mu(f) = m(f\mathcal{L}h)$.

Note that $\tilde{T}m(1) = \frac{1}{2}$, thus $\tilde{T}$ does not map probability measures into probability measures; this is clearly due to the mass leaking out of Q. Calling D^s (stable derivative) the derivative in the y direction, follows easily

$$\|D^s\mathcal{L}h\|_1 \leq \frac{1}{4}\|D^s h\|_1$$

for each h differentiable in the stable direction.

On the other hand, if $\|D^s h\|_1 \leq c$ and $\Delta = [0, 1/4] \cup [3/4, 1]$,

$$\begin{aligned}|\tilde{T}\mu(1)| = \int_{Q\cap TQ} h &= \int_\Delta dy \int_0^1 dx h(x, y)\\ &= \int_\Delta dy \int_0^1 dx \int_0^1 d\xi h(x, \xi) + \mathcal{O}(\|D^s h\|_1)\\ &= |\Delta|\|h\|_1 + \mathcal{O}(\|D^s h\|_1) = \frac{1}{2}\mu(1) + \mathcal{O}(\|D^s h\|_1).\end{aligned}$$

It is then natural to define $\hat{\mathcal{L}}h := 2\mathcal{L}h$ and $\hat{T} = 2\tilde{T}$. Thus $\|D^s\hat{\mathcal{L}}h\|_1 \leq \frac{1}{2}\|D^s h\|_1$. This means that $\{\frac{1}{n}\sum_{i=0}^{n-1}\hat{T}^i\mu\}$ are probability measures. Accordingly, there exists an accumulation point μ_* and $\mu_*(D^s f) = 0$ for each f periodic in the y direction. By the same type of arguments used in the previous examples, this means that μ_* is constant in the y direction, it is supported on Λ_- by construction and $\tilde{T}\mu_* = \frac{1}{2}\mu_*$ (*conformal invariance*) : just the measure we where looking for.

We can now conclude the argument by evolving the measure as usual:

$$T_*\mu_*(f) = \mu_*(f \circ T)$$

for all continuous f with the support in Q. Now the standard argument applies. In such a way we have obtained the invariant measure supported on Λ.

Markov Measures

Let us consider the shift (Σ_n^+, T). We would like to construct other invariant measures bedside Bernoulli. As we have seen it suffices to specify the measure on the algebra of the cylinders. Let us define

$$A(m; k_1, \dots, k_l) = \{\sigma \in \Sigma_n^+ \mid \sigma_{i+m} = k_i \; \forall \; i \in \{1, \dots, l\}\};$$

this are a basis for the algebra of the cylinders.

For each $n \times n$ matrix P, $P_{ij} \geq 0$, $\sum_j P_{ij} = 1$ by the Perron-Frobenius theorem (see Secion (A.3.2)) there exists $\{p_i\}$ such that $pP = p$. Let us define

$$\mu(A(m; k_1, \dots, k_l)) = p_{k_1} P_{k_1 k_2} P_{k_2 k_3} \dots P_{k_{l-1} k_l}.$$

The reader can easily verify that μ is invariant over the algebra $\mathcal{A}$ and thus extends to an invariant measure. This is called Markov because it is nothing else than a Markov chain together with its stationary measure.[26]

These last examples (strange attractor, solenoid, horseshoe) show only a very dim glimpse of a much more general and extremely rich theory (the study of SRB measures) while the last (Markov measures) points toward another extremely rich theory: Gibbs (or equilibrium) measures. Although this it is not the focus here, we will see a bit more of this in the future.

One of the main objectives in dynamical systems is the study of the long time behavior (that is the study of the trajectories $T^n x$ for large n). There are two main cases in which it is possible to study, in some detail, such a long time behavior. The case in which the motion is rather regular[27] or close to it (the main examples of this possibility are given by the so called KAM [Arn92] theory and by situations in which the motions is attracted by a simple set); and the case in which the motion is very irregular.[28] This last case may seem surprising since the irregularity of the motion should make its study very difficult. The reason why such systems can be studied is, as usual,

[26] The probabilistic interpretation is that the probability of seeing the state k at time one, given that we saw the state l at time zero, is given by P_{lk}. So the process has a bit of memory: it remembers its state one time step before. Of course it is possible to consider processes that have a longer–possibly infinite–memory. Proceeding in this direction one would define the so called *Gibbs measures*.

[27] Typically, quasi periodic motion, remember the small oscillation in the pendulum.

[28] Remember the example in the introduction.

because we ask the right questions,[29] that is we ask questions not concerning the fine details of the motion but only concerning its statistical or qualitative properties.

The first example of such properties is the study of the invariant sets.

6.6 Ergodicity

Definition 6.6.1 *A measurable set A is invariant for T if $T^{-1}A \subset A$.*

A dynamical system (X, T, μ) is ergodic if each invariant set has measure zero or one.

The definition for continuous dynamical systems being exactly the same.

Note that if A is invariant then $\mu(A \backslash T^{-1}A) = \mu(A) - \mu(T^{-1}A) = 0$, moreover $\Lambda = \cap_{n=0}^{\infty} T^{-n}A \subset A$ is invariant as well. In addition, by definition, $\Lambda = T\Lambda$, which implies $\Lambda = T^{-1}\Lambda$ and $\mu(A \backslash \Lambda) = 0$. This means that, if A is invariant, then it always contains a set Λ invariant in the stronger (maybe more natural) sense that $T\Lambda = T^{-1}\Lambda = \Lambda$. Moreover, Λ is of full measure in A. Our definition of invariance is motivated by its greater flexibility and the fact that, from a measure theoretical point of view, zero measure sets can be discarded.

In essence, if a system is ergodic then most trajectories explore all the available space. In fact, for any A of positive measure, define $A_b = \cup_{n \in \mathbb{N} \cup \{0\}} T^{-n}A$ (this are the points that eventually end up in A), since $A_b \supset A$, $\mu(A_b) > 0$. Since $T^{-1}A_b \subset A_b$, by ergodicity follows $\mu(A_b) = 1$. Thus, the points that never enter in A (that is, the points in A_b^c) have zero measure. Actually, if the system has more structure (topology) more is true (see Problem 6.26).

The reader should be aware that there are many equivalent definitions of ergodicity. In the following, we give a relevant one, but see Problems 6.31, 6.32 and Theorem 6.7.5 for other possibilities.

Lemma 6.6.2 *Show that a Dynamical Systems (X, T, μ) is ergodic if and only if the transfer operator $\mathcal{L}$ acting on $L^1(X, \mu)$ has 1 as a simple eigenvalue.*[30]

PROOF. Since we want to connect the concept of ergodicity to the spectral properties of $\mathcal{L}$, it is natural to consider $L^1(X, \mu)$ as a space of complex functions. Since μ is invariant, we have $\mathcal{L}1 = 1$.
Let us suppose that 1 is a simple eigenvalue and assume that there exists a measurable invariant set A, $\mu(A) \notin \{0, 1\}$. By invariance $f^{-1}(A) \subset A$; that is

[29] Of course, the "right questions" are the ones that can be answered.

[30] That is, there are no other invariant measures absolutely continuous with respect ot μ.

$\mathbb{1}_A \circ T = \mathbb{1}_{f^{-1}(A)} \leq \mathbb{1}_A$. But the invariance of μ implies $\mu(A) = \mu(T^{-1}(A))$, hence $\mathbb{1}_A \circ T = \mathbb{1}_A$, μ-a.s.. Recalling Lemma 6.1.4 we have

$$\mathcal{L}\mathbb{1}_A = \mathcal{L}\mathbb{1}_A \circ T = \mathbb{1}_A \mathcal{L}1 = \mathbb{1}_A$$

which contradicts the assumption that 1 is a simple eigenvalue.
Next, suppose that (X, T, μ) is ergodic. We start to notice that $\mathcal{L}h = h$, with $h = h_1 + ih_2$, with h_i real, then by linearity,

$$h_1 + ih_2 = \mathcal{L}h_1 + i\mathcal{L}h_2$$

which implies $\mathcal{L}h_i = h_i$. We can then restrict ourselves to real invariant functions. Hence, let $h \in L^1(X, \mu)$ be real and $\mathcal{L}h = h$, then by Lemma 6.1.4,

$$\mathcal{L}(|h| \pm h) \geq 0$$

which implies $\mathcal{L}|h| \geq \pm\mathcal{L}h = \pm h$, that is $\mathcal{L}|h| \geq |h|$. Yet,

$$0 \leq \int_X (\mathcal{L}|h| - |h|)d\mu = 0$$

implies $\mathcal{L}|h| = |h|$. Suppose now that $|h| = 1$, then

$$\mathcal{L}1 = 1 = h^2 = h\mathcal{L}h = \mathcal{L}(h \circ T\, h).$$

This allows us to write

$$0 = \int_X \mathcal{L}(1 - h \circ T\, h)d\mu = \int_X (1 - h \circ T\, h)d\mu.$$

But $1 - h \circ T\, h \geq 0$, thus it must be $1 - h \circ T\, h = 0$ or $h = h \circ T$. By ergodicity, this implies h is constant, hence proportional to 1. We are thus left with the possibility of multiple positive eigenvectors. Let $h \geq 0$, $\mathcal{L}h = h$, then for each $a \in \mathbb{R}$ let $\Gamma_a(x) = \max\{a, h(x)\}$, then

$$\mathcal{L}\Gamma_a \geq \mathcal{L}h = h \qquad \mathcal{L}\Gamma_a \geq \mathcal{L}a = a,$$

which implies $\mathcal{L}\Gamma_a \geq \Gamma_a$. But since $\int_X (\mathcal{L}\Gamma_a - \Gamma_a)d\mu = 0$, it must be $\mathcal{L}\Gamma_a = \Gamma_a$. Next, let $A = \{x \in X \ : \ h(x) \geq a\}$ and notice that $x \in A$ iff $\Gamma_a(x) = h(x)$. We have

$$0 = \int_A (\Gamma_a - h)d\mu = \int_X \mathbb{1}_A \mathcal{L}(\Gamma_a - h)d\mu = \int_X \mathbb{1}_A \circ T(\Gamma_a - h)d\mu.$$

Since $\Gamma_a - h \geq 0$ it must be $\mathbb{1}_{T^{-1}(A)}(\Gamma_a - h) = 0$. This is possible if either $\mu(A) = 0$ or $T^{-1}(A) \subset A$. Consequently h is μ=a.s. constant and hence proportional to 1. That is 1 is a simple eigenvalue of $\mathcal{L}$. □

6.6.1 Examples

Rotations

The ergodicity of a rotation depends on ω. If $\omega \in \mathbb{Q}$ then the system is not ergodic. In fact, let $\omega = \frac{p}{q}$ $(p, q \in \mathbb{N})$, then, for each $x \in \mathbb{T}$ $T^q x = x + p \mod 1 = x$, so T^q is just the identity. An alternative way of saying this is to notice that all the points have a periodic trajectory of period q. It is then easy to exhibit an invariant set with measure strictly larger than 0 but strictly less than 1. Consider $[0, \varepsilon]$, then $A = \cup_{i=1}^{q-1} T^{-i}[0, \varepsilon]$ is an invariant set; clearly $\varepsilon \leq \mu(A) \leq q\varepsilon$, so it suffices to choose $\varepsilon < q^{-1}$.

The case $\omega \notin \mathbb{Q}$ is much more interesting. First of all, for each point $x \in \mathbb{T}$ we have that the closure of the set $\{T^n x\}_{i=0}^{\infty}$ is equal to $\mathbb{T}$, which is to say that the orbits are dense.[31] The proof is based on the fact that there cannot be any periodic orbit. To see this suppose that $x \in \mathbb{T}$ has a periodic orbit, that is there exists $q \in \mathbb{N}$ such that $T^q x = x$. As a consequence there must exist $p \in \mathbb{Z}$ such that $x + p = x + q\omega$ or $\omega \in \mathbb{Q}$ contrary to the hypothesis. Hence, the set $\{T^k 0\}_{k=0}^{\infty}$ must contain infinitely many points and, by compactness, must contain a convergent subsequence k_i. Hence, for each $\varepsilon > 0$, there exists $m > n \in \mathbb{N}$:

$$|T^m 0 - T^n 0| < \varepsilon.$$

Since T preserves the distances, calling $q = m - n$, holds

$$|T^q 0| < \varepsilon.$$

Accordingly, the trajectory of $T^{jq}0$ is a translation by a quantity less than ε, therefore it will get closer than ε to each point in $\mathbb{T}$ (i.e., the orbit is dense). Again by the conservation of the distance, since zero has a dense orbit the same will hold for every other point.

Intuitively, the fact that the orbits are dense implies that there cannot be a non trivial invariant set, henceforth the system is ergodic. Yet, the proof it is not trivial since it is based on the existence of Lebesgue density points [Roy88] (see Problem 6.44). It is a fact from general measure theory that each measurable set $A \subset \mathbb{R}$ of positive Lebesgue measure contains, at least, one point $\bar{x}$ such that for each $\varepsilon \in (0, 1)$ there exists $\delta > 0$:

$$\frac{m(A \cap [\bar{x} - \delta, \bar{x} + \delta])}{2\delta} > 1 - \varepsilon.$$

Hence, given an invariant set A of positive measure and $\varepsilon > 0$, first choose δ such that the interval $I := [\bar{x} - \delta, \bar{x} + \delta]$ has the property $m(I \cap A) > (1 - \varepsilon)m(I)$. Second, we know already that there exists $q, M \in \mathbb{N}$ such that $\{T^{-kq} x\}_{k=1}^{M}$

[31] A system with a dense orbit called *Topologically Transitive*.

divides $[0,\,1]$ into intervals of length less that $\frac{\varepsilon}{2}\delta$. Hence, given any point $x \in \mathbb{T}$ choose $k \in \mathbb{N}$ such that $m(T^{-kq}I \cap [x-\delta,\, x+\delta]) > m(I)(1-\varepsilon)$ so,

$$\begin{aligned} m(A \cap [x-\delta,\, x+\delta]) &\geq m(A \cap T^{-kq}I) - m(I)\varepsilon \\ &\geq m(A \cap I) - m(I)\varepsilon \geq (1-2\varepsilon)2\delta. \end{aligned}$$

Thus, A has density everywhere larger than $1-2\varepsilon$, which implies $\mu(A) = 1$ since ε is arbitrary.

The above proof of ergodicity it is not so trivial but it has a definite dynamical flavor (in the sense that it is obtained by studying the evolution of the system). Its structure allows generalizations to contexts whit a less rich algebraic structure. Nevertheless, we must notice that, by taking advantage of the algebraic structure (or rather the group structure) of $\mathbb{T}$, a much simpler and powerful proof is available.

Let $\nu \in \mathcal{M}_T^1$, then define

$$F_n = \int_{\mathbb{T}} e^{2\pi i n x} \nu(dx), \quad n \in \mathbb{N}.$$

A simple computation, using the invariance of ν, yields

$$F_n = e^{2\pi i n \omega} F_n$$

and, if ω is irrational, this implies $F_n = 0$ for all $n \neq 0$, while $F_0 = 1$. Next, consider $f \in \mathcal{C}^{(2)}(\mathbb{T}^1)$ (so that we are sure that the Fourier series converges uniformly, see Problem 6.35), then

$$\nu(f) = \sum_{n=0}^{\infty} \nu(f_n e^{2\pi i n \cdot}) = \sum_{n=0}^{\infty} f_n F_n = f_0 = \int_{\mathbb{T}} f(x)dx.$$

Hence m is the unique invariant measure (unique ergodicity). This is clearly much stronger than ergodicity (see Problem 6.6.2)

Expanding maps

Next, we prove that any smooth invariant map has a unique invariant measure absolutely continuos with respect to Lebesgue and hence it is ergodic with respect to such a measure. Let $h \in L^1$ be the density of an invariant measure and A, of positive measure, an invariant set. For each $\varepsilon > 0$ there exists $f_\varepsilon \in \mathcal{C}^1$ such that $\|f_\varepsilon - \mathbb{1}_A\|_1 \leq \varepsilon$. Calling $f_{\varepsilon,n} = \frac{1}{n}\sum_{i=0}^{n-1} \mathcal{L}^i f_\varepsilon$ and noting that, by invariance, $\varphi_n := \frac{1}{n}\sum_{i=0}^{n-1} \mathcal{L}^i \mathbb{1}_A = \mathbb{1}_A \frac{1}{n}\sum_{i=0}^{n-1} \mathcal{L}^i 1$, we have, by taking subsequeces, that f_n converges in $\mathcal{C}^0$ to some invariant density $\bar{f}_\varepsilon$ while φ_n converges to $\mathbb{1}_A h$, where h is the invariant density to which converges $\frac{1}{n}\sum_{i=0}^{n-1} \mathcal{L}^i 1$ (or rather the chosen

subsequence). On the other hand $\|\bar{f}_\varepsilon - \mathbb{1}_A h\|_1 \leq \varepsilon$. Since the $\bar{f}_\varepsilon$ are all uniformly Lipschitz, hence equicontinuous, (see the end of Example 6.5.1, Expanding maps) by Ascoli-Arzelá we can extract a converging subquence. This means that $\mathbb{1}_A$ is the uniform limit of continuos functions, hence it is continuos hence A is either empty of everything, thus the map is ergodic. The uniqueness of the invariant measure follows by similar arguments.

Baker

This transformation gets its name from the activity of bread making, it bears some resemblance with the horseshoe. The space X is the square $[0,\,1]^2$, μ is again Lebesgue, and T is a transformation obtained by squashing down the square into the rectangle $[0,\,2] \times [0,\,\frac{1}{2}]$ and then cutting the piece $[1,\,2] \times [0,\,\frac{1}{2}]$ and putting it on top of the other one. In formulas

$$T(x,\,y) = \begin{cases} (2x,\,\dfrac{1}{2}y) \mod 1 & \text{if } x \in [0,\,\dfrac{1}{2}) \\ (2x,\,\dfrac{1}{2}(y+1)) \mod 1 & \text{if } x \in [\dfrac{1}{2},\,1]. \end{cases}$$

This transformation is ergodic as well, in fact much more. We will discuss it later.

Translations ($\mathbb{T}^1$)

Let us consider the flow $(\mathbb{T}^1, \phi_t, m)$ where $\phi_t(x) = x + \omega t \mod 1$, for some $\omega \in \mathbb{R} \setminus \{0\}$. This is just a translation on the unit circle. The proof of ergodicity is trivial and it is left to the reader.

We conclude the chapter with a theorem very helpful to establish the ergodicity of a flow.

Theorem 6.6.3 *Consider a flow (X, ϕ_t, μ) and a Poincarè section Σ such that the set $\{x \in X \mid \cup_{t\in\mathbb{R}}\, \phi_t(x) \cap \Sigma = \emptyset\}$ has zero measure. Then the ergodicity of the flow (X, ϕ_t, μ) is equivalent to the ergodicity of the section $(\Sigma, T_\Sigma, \mu_\Sigma)$.*

The proof, being straightforward, is left to the reader.

6.6.2 Examples

Translations ($\mathbb{T}^2$)

Let us consider the flow $(\mathbb{T}^2, \phi_t, m)$ where $\phi_t(x) = x + \omega t \mod 1$, for some $\omega \in \mathbb{R}^2 \setminus \{0\}$. This is a translation on the two dimensional torus. To investigate we will use Theorem 6.6.3. Consider the set $\Sigma := \{(x, y) \in \mathbb{T}^2 \mid x = 0\}$, this

is clearly a Poincaré section, unless $\omega_1 = 0$ (in which case one can choose the section $y = 0$). Obviously Σ is a circle and the Poincaré map is given by

$$T(y) = y + \frac{\omega_2}{\omega_1} \mod 1.$$

The ergodicity of the flow is then reduced to the ergodicity of a circle rotation, thus the flow is ergodic only if ω_1 and ω_2 have an irrational ratio.

The properties of the invariant sets of a dynamical systems have very important reflections on the statistics of the system, in particular on its time averages. Before making this precise (see Theorem 6.7.5) we state few very general and far reaching results.

6.7 Some basic Theorems

In this section, we present some basic theorems and constructions fundamental in Ergodic theory.

6.7.1 Ergodic Theorems

Theorem 6.7.1 *(Birkhoff) Let (X, T, μ) be a dynamical system, then for each $f \in L^1(X, \mu)$*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(T^j x)$$

exists for almost every point $x \in X$. In addition, setting

$$f^+(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(T^j x),$$

holds

$$\int_X f^+ d\mu = \int_X f d\mu.$$

Proof

Since the task at hand is mainly didactic, we will consider explicitly only the case of positive bounded functions, the completion of the proof is left to the reader.

Let $f \in L^\infty(X, d\mu)$, $f \geq 0$, and

$$S_n(x) \equiv \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x).$$

For each $x \in X$, there exists

$$\overline{f}^+(x) = \limsup_{n\to\infty} S_n(x)$$
$$\underline{f}^+(x) = \liminf_{n\to\infty} S_n(x).$$

The first remark is that both $\overline{f}^+$ and $\underline{f}^+$ are invariant functions. In fact,

$$S_n(Tx) = S_n(x) + \frac{1}{n} f(T^n x) - \frac{1}{n} f(x)$$

so, tacking the limit the result follows.[32]

Next, for each $n \in \mathbb{N}$ and k, $j \in \mathbb{Z}$ we define

$$D_{n,l,j} = \left\{ x \in X \;\middle|\; \overline{f}^+(x) \in \left[\frac{l}{n}, \frac{l+1}{n}\right); \; \underline{f}^+(x) \in \left[\frac{j}{n}, \frac{j+1}{n}\right) \right\},$$

by the invariance of the functions follows the invariance of the sets $D_{n,l,j}$. Also, by the boundedness, follows that for each n exists n_0 such as

$$\bigcup_{j,l\in\{-n_0,\dots,n_0\}} D_{n,l,j} = X.$$

The key observation is the following.

Lemma 6.7.2 *For each $n \in \mathbb{N}$ and l, $j \in \mathbb{Z}$, setting $A = D_{n,l,j}$, holds*

$$\frac{l+1}{n}\mu(A) < \int_A f d\mu + \frac{3}{n}\mu(A)$$
$$\frac{j}{n}\mu(A) > \int_A f d\mu - \frac{3}{n}\mu(A)$$

From the Lemma follows

$$0 \le \int_X (\overline{f}^+ - \underline{f}^+) d\mu = \sum_{l,\, j=-n_0}^{n_0} \int_{D_{n,l,j}} (\overline{f}^+ - \underline{f}^+) d\mu$$
$$\le \sum_{l,\, j=-n_0}^{n_0} \left[\frac{l+1}{n} - \frac{j}{n}\right] \mu(D_{n,l,j}) < \frac{6}{n} \sum_{l,\, j=-n_0}^{n_0} \mu(D_{n,l,j}) = \frac{6}{n}.$$

Since n is arbitrary we have

$$\int_X (\overline{f}^+ - \underline{f}^+) d\mu = 0$$

[32] Here we have used the boundedness, this is not necessary. If $f \in L^1(X, d\mu)$ and positive, then $S_n(Tx) \ge S_n(x) - f(x)$, so $\underline{f}^+(Tx) \ge \underline{f}^+(x)$ and it is and easy exercise to check that any such function must be invariant.

which implies $\overline{f}^+ = \underline{f}^+$ almost everywhere (since $\overline{f}^+ \geq \underline{f}^+$ by definition) proving that the limit exists. Analogously, we can prove

$$\int_X (f - f^+) d\mu = 0.$$

Proof of the Lemma 6.7.2 We will prove only the first inequality, the second being proven in exactly the same way.

For each $x \in A$ we will call $k(x)$ the first $m \in \mathbb{N}$ such that

$$S_m(x) > \frac{l-1}{n},$$

by construction $k(x)$ must be finite for each $x \in A$. Hence, setting $X_k = \{x \in A \mid k(x) = k\}$, $\cup_k X_k = A$, and for each $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$\mu\left(\bigcup_{k=1}^{N} X_k\right) \geq \mu(A)(1-\varepsilon).$$

Let us call

$$Y = A \backslash \bigcup_{k=1}^{N} X_k.$$

Then $\mu(Y) \leq \mu(A)\varepsilon$, also set $L = \sup_{x \in A} |f(x)|$. The basic idea is to follow, for each point $x \in A$, the trajectory $\{T^i x\}_{i=0}^{M}$, where $M > N$ will be chosen sufficiently large. If the point would never visit the set Y, we could group the sum $S_M(x)$ in pieces all, in average, larger than $\frac{l-1}{n}$, so the same would hold for $S_M(x)$. The difficulties come from the visits to the set Y.

For each $n \in \{0, ..., M\}$ define

$$\widetilde{f}_n(x) = \begin{cases} f(T^n x) & \text{if } T^n x \notin Y \\ \dfrac{l}{n} & \text{if } T^n x \in Y \end{cases}$$

and

$$\widetilde{S}_M(x) = \frac{1}{M} \sum_{n=0}^{M-1} \widetilde{f}_n(x).$$

By definition $y \in Y$ implies $y \notin X_1$, i.e. $f(y) \leq \frac{l-1}{n}$. Accordingly, $\widetilde{f}(x) \geq f(T^n x)$ for each $x \in A$. Note that for each n we change the function $f \circ T^n$ only at some points belonging to the set Y and $\frac{l}{n}$ can be taken less or equal than L (otherwise $\mu(A) = 0$), consequently

$$\int_A f d\mu = \int_A S_M d\mu \geq \int_A \widetilde{S}_M d\mu - L\mu(Y) \geq \int_A \widetilde{S}_M d\mu - L\mu(A)\varepsilon.$$

We are left with the problem of computing the sum. As already mentioned the strategy consists in dividing the points according to their trajectory with respect to the sets X_n. To be more precise, let $x \in A$, then by definition it must belong to some X_n or to Y. We set $k_1(x)$ equal to j is $x \in X_j$ and $k_1(x) = 1$ if $x \in Y$. Next, $k_2(x)$ will have value j if $T^{k_1(x)}x \in X_j$ or value 1 if $T^{k_1(x)} \in Y$. If $k_1(x) + k_2(x) < M$, then we go on and define similarly $k_3(x)$. In this way, to each $x \in A$ we can associate a number $m(x) \in \{1, ..., M\}$ and indices $\{k_i(x)\}_{i=1}^{m(x)}$, $k_i(x) \in \{1, ..., N\}$, such that $M - N \leq \sum_{i=1}^{m(x)-1} k_i(x) < M$, $\sum_{i=1}^{m(x)} k_i(x) \geq M$. Let us call $K_p(x) = \sum_{j=1}^{p} k_j(x)$. Using such a division of the orbit in segments of length $k_i(x)$ we can easily estimate

$$\begin{aligned}\widetilde{S}_M(x) =& \frac{1}{M}\left\{\sum_{i=1}^{m(x)-1} k_i(x)\left[\frac{1}{k_i(x)}\sum_{j=K_{i-1}(x)}^{K_i(x)-1}\widetilde{f}_j(x)\right] + \sum_{i=K_{m(x)-1}(x)}^{M-1}\widetilde{f}(T^i x)\right\}\\ \geq& \frac{1}{M}\sum_{i=1}^{m(x)-1} k_i(x)\frac{l-1}{n} \geq \frac{M-N}{M}\frac{l-1}{n}.\end{aligned}$$

Putting together the above inequalities we get

$$\begin{aligned}\int_A f d\mu &\geq \left\{\frac{(M-N)(l-1)}{Mn} - L\varepsilon\right\}\mu(A)\\ &\geq \frac{l+1}{n}\mu(A) - \left\{\frac{2}{n} + \frac{N(l-1)}{Mn} + L\varepsilon\right\}\mu(A).\end{aligned}$$

which, by choosing first ε sufficiently small and, after, M sufficiently large, concludes the proof. □

To prove the result for all function in $L^1(X, \mu)$ it is convenient to deal at first only with positive functions (which suffice since any function is the difference of two positive functions) and then use the usual trick to cut off a function (that is, given f define f_L by $f_L(x) = f(x)$ if $f(x) \leq L$, and $f_L(x) = L$ otherwise) and then remove the cut off. The reader can try it as an exercise. □

Birkhoff theorem has some interesting consequences.

Corollary 6.7.3 *For each $f \in L^1(X, \mu)$ the following holds*

1. *$f^+ \in L^1(X, \mu)$;*
2. *$f^+(Tx) = f_+(x)$ almost surely.*

The proof is left to the reader as an easy exercise (see Problem 6.23).

Another interesting fact, that starts to show some connections between averages and invariant sets, emerges by considering a measurable set A and

its characteristic function χ_A. A little thought shows that the ergodic average $\chi_A^+(x)$ is simply the average frequency of visit of the set A by the trajectory $\{T^n x\}$ (Problem 6.32).

Birkhoff theorem implies also convergence in L^1 and L^2 (see also Problem 6.30). Yet, it is interesting to note that convergence in L^2 can be proven in a much more direct way.

Theorem 6.7.4 (Von Neumann) *Let (X, T, μ) be a Dynamical System, then for each $f \in L^2(X, \mu)$ the ergodic average converges in $L^2(X, \mu)$.*

PROOF. We have already seen that it can be useful to lift the dynamics at the level of the algebra of function or at the level of measures. This game assumes different guises according to how one plays it, here is another very interesting version.

Let us define $U : L^2(X, \mu) \to L^2(X, \mu)$ as

$$Uf := f \circ T.$$

Then, by the invariance of the measure, it follows $\|Uf\|_2 = \|f\|_2$, so U is an L^2 contraction (actually, and L^2-isometry). If T is invertible, the same argument applied to the inverse shows that U is indeed unitary, otherwise we must content ourselves with

$$\|U^* f\|_2^2 = \langle UU^* f, f\rangle \leq \|UU^* f\|_2 \|f\|_2 = \|U^* f\|_2 \|f\|_2,$$

that is $\|U^*\|_2 \leq 1$ (also U^* is and L^2 contraction).

Next, consider $V_1 = \{f \in L^2 \mid Uf = f\}$ and $V_2 = \mathrm{Rank}(\mathbb{1} - U)$. First of all, note that if $f \in V_1$, then

$$\|U^* f - f\|_2^2 = \|U^* f\|_2^2 - \langle f, U^* f\rangle - \langle U^* f, f\rangle + \|f\|_2^2 \leq 0.$$

Thus, $f \in V_1^* := \{f \in L^2 \mid U^* f = f\}$. The same argument applied to $f \in V_1^*$ shows that $V_1 = V_1^*$. To continue, consider $f \in V_1$ and $h \in L^2$, then

$$\langle f, h - Uh\rangle = \langle f - U^* f, h\rangle = 0.$$

This implies that $V_1 \subset V_2^\perp$ and $V_2^\perp \subset V_1$, that is $V_2^\perp = V_1$.

Problem 6.1 *Let V be a linear subspace of an Hilbert space H, prove that $H = \overline{V} \oplus V^\perp$.*

Problem 6.2 *Let V be a linear subspace of an Hilbert space H, prove that $(V^\perp)^\perp = \overline{V}$.*

Due to the above problems, since V_1 is a closed space, we have

$$L^2 = V_1 \oplus V_1^{\perp} = V_1 \oplus (V_2^{\perp})^{\perp} = V_1 \oplus \overline{V_2}.$$

Finally, if $g \in V_2$, then there exists $h \in L^2$ such that $g = h - Uh$ and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} U^i g = \lim_{n\to\infty} \frac{1}{n}(h - U^n h) = 0.$$

On the other hand if $f \in V_1$ then $\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} U^i f = f$. The only function on which we do not still have control are the g belonging to the closure of V_2 but not in V_2. In such a case there exists $\{g_k\} \subset V_2$ with $\lim_{k\to\infty} g_k = g$. Thus,

$$\left\| \frac{1}{n} \sum_{i=0}^{n-1} U^i g \right\|_2 \leq \left\| \frac{1}{n} \sum_{i=0}^{n-1} U^i g_k \right\|_2 + \|g - g_k\|_2 \leq \left\| \frac{1}{n} \sum_{i=0}^{n-1} U^i g_k \right\|_2 + \frac{\varepsilon}{2},$$

provided we choose k large enough. Then, by choosing n sufficiently large we obtain

$$\left\| \frac{1}{n} \sum_{i=0}^{n-1} U^i g \right\|_2 \leq \varepsilon.$$

We have just proven that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} U^i = P$$

where P is the orthogonal projection on V_1. □

Let us investigate the relationship between ergodicity and averages a bit further. From an intuitive point of view, a function from X to $\mathbb{R}$ can be thought as an "observable," since to each configuration it associates a value that can represent some relevant property of the configuration (the property that we observe). So, if we observe the system for a long time via the function f, what we see should be well represented by the function f^+. Furthermore, notice that there is a simple relations between invariant functions and invariant sets. More precisely, if a measurable set A is invariant, then its characteristic function χ_A is a measurable invariant function; if f is an invariant function then for each measurable set $I \in \mathbb{R}$ the set $f^{-1}(I)$ is a measurable invariant set (if the implications of the above discussions are not clear to you, see Problem 6.31).

As a byproduct of the previous discussion, it follows that if a system is ergodic then for each function $f \in L^1(X, \mu)$ the function f_+ is almost everywhere constant and equal to $\int_X f$. We have just proven another interesting characterization of the ergodic systems:

Theorem 6.7.5 *A Dynamical System (X, T, μ) is ergodic if and only if for each $f \in L^1(X, \mu)$ the ergodic average f^+ is constant; in fact, $f^+ = \mu(f)$ a.e..*

In other words, if we observe the time average of some observable for a sufficiently long time then we obtain a value close to its space average. The previous observation is very important especially because the space average of a function does not depend on the dynamics. This is exactly what we were mentioning previously: the fact that the dynamics is sufficiently 'complex' allows us to ignore it completely, provided we are interested only in knowing some average behavior. The relevance of ergodic theory for physical systems is largely connected to this fact.

6.7.2 Recurrence Theorems

Next, we discuss another very general result, of a somewhat disturbing nature, is Poincaré return theorem.

Theorem 6.7.6 (Poincaré) *Given a dynamical systems (X, T, μ) and a measurable set A, with $\mu(A) > 0$, there exists infinitely many $n \in \mathbb{N}$ such that*

$$\mu(T^{-n}A \cap A) \neq 0.$$

The proof is rather simple (by contradiction) and the reader can certainly find it out by herself (see Problem 6.24).[33]

A natural question is how long it takes, on average, to come back to a set A. Let $A \subset X$ be a measurable set, and let us define the return time

$$\tau_A(x) = \inf\{n \in \mathbb{N} \ : \ f^n(x) \in A\}. \tag{6.7.9}$$

Problem 6.3 *Check that $\tau : A \to \mathbb{N} \cup \{\infty\}$ is a measurable function.*

Lemma 6.7.7 (Kač) *Given a dynamical systems (X, T, μ) and a measurable set A, with $\mu(A) > 0$,*

$$\int_A \tau_A(x)\mu(dx) = 1 - \mu(Y),$$

where $Y = \{x \in X \ : \ T^n(x) \notin A \, \forall n \in \mathbb{N}\}$.

[33] An unsettling aspect of the theorem is due to the following possibility. Consider a room full of air, the motion of the molecules can be thought to happen accordingly to Newton equations, i.e. it is an Hamiltonian systems, hence a dynamical system to which Poincaré theorem applies. Let A be the set of configurations in which all the air is in the left side of the room. Since we ignore, in general, the past history of the room, it could very well be that at some point in the past the system was in a configuration belonging to A–maybe some silly experiment was performed. So there is a positive probability for the system to return to the same state. Therefore, the disturbing possibility of sudden death by decompression.

PROOF. Consider the set $\tilde{X}_A = \cup_{n=1}^{\infty} T^{-n}(A)$, clearly $T^{-1}(\tilde{X}_A) \subset \tilde{X}_A$. This means that $\tilde{X}_A \setminus T^{-1}(\tilde{X}_A)$ has zero measure. We can then define $X_A = \cap_{n=0}^{\infty} T^{-n}\tilde{X}_A$, clearly $\tilde{X}_A \setminus X_A$ has zero measure and $T(X_A) = X_A = T^{-1}(X_A)$. Also, $B = A \setminus X_A$, must have zero measure, otherwise Poincarè theorem would imply that there exists $m \in \mathbb{N}$ such that $T^{-m}B \cap B \neq \emptyset$, but $T^{-m}B \subset T^m(A) \subset X_A$, which is a contradiction. The same argument shows that τ_A is everywhere finite on $A_* = A \cap X_A$. We can thus resctrict to the dynamical systems $(X_A, T, \bar{\mu})$, where $\bar{\mu} = \mu(X_A)^{-1}\mu$. By construction, τ_A is almost everywhere finite on X_A. Let $E_n = \{x \in A_* \ : \ \tau_A(x) = n\}$ and $R_n = \{x \notin A_* \ : \ \tau_A(x) = n\}$. Note that all these sets are disjoint, hence their measure must tend to zero as $n \to \infty$. Then

$$T^{-1}R_n = E_{n+1} \cup R_{n+1}.$$

Consequently,

$$\mu(R_n) = \mu(E_{n+1}) + \mu(R_{n+1}) = \sum_{k=n+1}^{\infty} \mu(E_k).$$

It follows

$$\begin{aligned} 1 - \mu(Y) = \mu(X_A) &= \sum_{n=1}^{\infty} \mu(E_n) + \mu(R_n) = \sum_{n=1}^{\infty}\sum_{k=n}^{\infty} \mu(E_k) \\ &= \sum_{k=1}^{\infty} k\mu(E_k) = \int_A \tau_A(x)\mu(dx). \end{aligned}$$

□

The above result is somewhat comforting as far as the comment in footnote 33 is concerned. The reader can easily calculate the average time for a catastrophic event to occur and see that it is extremely large.

6.7.3 Kakutani Towers

To conclude the section, we present a construction that is often very useful in ergodic theory: the Kakutani tower associated to a positive measure set A of a measurable dynamical system (X, T, μ).
Let $A_k = \{x \in A \ : \ \tau_A(x) = k\}$, and consider the set $Y = \{(x, j) \in A \times \mathbb{N} \cup \{0\} \ : \ j < \tau_A(x)\}$ and the map $S : Y \to Y$ defined by

$$S((x,j)) = \begin{cases} (x, j+1) & \text{if } j+1 < \tau_A(x) \\ (T^{\tau(x)}(x), 0) & \text{if } j = \tau_A(x) - 1. \end{cases}$$

The index, j stands for the floor in the tower. The σ-algebra on each floor is simply the σ-algebra on X. Also, define the measure on Y. For each measurable set $B \subset Y$,

$$\nu(B) = \sum_{k=0}^{\infty} \mu\left(\{x \in A \ : \ (x,k) \in B\}\right).$$

Consider the map $\pi : Y \to X$ defined by

$$\pi(x,k) = T^k(x).$$

Let τ_A be as in (6.7.9), and $A_k = \{x \in A \ : \ \tau_A(x) = k\}$, $k \in \mathbb{N}$, and $A_0 = A$. Note that, for each measurable set $B \subset X$,

$$\pi^{-1}(B) = \cup_{k \geq 0} \cup_{j<k} \{(x,j) \in Y \ : \ x \in A_k \cap T^{-j}B\},$$

which is the union of measurable sets, hence π is measurable. The main property of Kakutani towers rests in the following Theorem that establishes the relation between the tower and the original dynamical system.

Theorem 6.7.8 *Given the above definitions, we have*

$$T \circ \pi = \pi \circ S,$$

$\pi_* \nu = \mu$, $\nu(Y) = 1$, *and* (A, S, ν) *is a measurable dynamical system. Conversely, if* (A, S, ν) *is a measurable dynamical system for some measure* ν, *then* (Y, T, μ) *is a measurable dynamical system for the measure*

$$\mu(B) = \nu((A \cap B, 0)) + \sum_{j=1}^{\infty} \sum_{n=j+1}^{\infty} \nu\left((A_n \cap T^{-j}B, 0)\right).$$

Proof. Let $(x,l) \in Y$, with $l < \tau_A(x) - 1$, then

$$\pi \circ S(x,l) = \pi(x, l+1) = T^{l+1}(x) = T \circ \pi(x,l).$$

If $l = \tau_A(x) - 1$, then

$$T \circ \pi(x,l) = T^{\tau_A(x)}(x) = \pi(T^{\tau_A(x)}(x), 0) = \pi \circ S(x,l).$$

It remains to compute $\pi_* \nu$.
Let $B \subset X$ be a measurable set, then $(x,0) \in \pi^{-1}(B)$ iff $x \in A \cap B$; while, for $k > 0$, $(x,k) \in \pi^{-1}B$ iff $\pi(x,j) = T^j(x) \in B$ with $j < \tau_A(x)$, that is $x \in A_n = \{x \in A \ : \ \tau_A(x) = n\}$ with $n > j$. Accordingly,

$$\pi^{-1}(B) = (A \cap B, 0) \bigcup \left[\cup_{j=1}^{\infty} (\cup_{n>j} A_n \cap T^{-j}(B), j)\right]. \tag{6.7.10}$$

Let $R_n = \{x \notin A \,:\, \tau_A(x) = n\}$. Since $T^{-1}R_n = R_{n+1} \cup A_{n+1}$ we have

$$\begin{aligned}
\nu(\pi^{-1}(B)) =& \mu(B \cap A) + \sum_{j=1}^{\infty} \sum_{n>j} \mu\left(A_n \cap T^{-j}(B)\right) \\
=& \mu(B \cap A) + \sum_{j=1}^{\infty} \sum_{n>j} \left[\mu\left(T^{-1}R_{n-1} \cap T^{-j}(B)\right) - \mu\left(R_n \cap T^{-j}(B)\right)\right] \\
=& \mu(B \cap A) + \sum_{j=0}^{\infty} \sum_{n>j} \mu\left(R_n \cap T^{-j}(B)\right) \\
& - \sum_{j=1}^{\infty} \sum_{n>j} \mu\left(R_n \cap T^{-j}(B)\right) = \mu(B \cap A) + \sum_{n=1}^{\infty} \mu\left(R_n \cap B\right) \\
=& \mu(B \cap A) + \mu(B \cap A^c) = \mu(B).
\end{aligned}$$

In particular, ν is a probability measure, since $1 = \mu(X) = \nu(\pi^{-1}X) = \nu(Y)$. To conclude the proof, note that, for each measurable set $B \in Y$ we have

$$\begin{aligned}
\nu(S^{-1}(B)) &= \nu(\pi^{-1} \circ T^{-1} \circ \pi(B)) = \mu(T^{-1} \circ \pi(B)) = \mu(\pi(B)) \\
&= \nu(\pi^{-1} \circ \pi(B)) \geq \nu(B).
\end{aligned}$$

Since $\nu(S^{-1}(B)) = 1 - \nu(S^{-1}(B^c)) \leq 1 - \nu(B^c) \geq \nu(B)$. Thus, ν is invariant for S, and the first statement of the Lemma follows.
To prove the last part, for $B \subset A$ let $\bar{\mu}(B) = \nu((B, 0))$. By (6.7.10) we can write, for all $B \subset X$,

$$\mu(B) = \bar{\mu}(A \cap B) + \sum_{j=1}^{\infty} \sum_{n=j+1}^{\infty} \bar{\mu}\left(A_n \cap T^{-j}B\right).$$

Then

$$\begin{aligned}
&\mu(T^{-1}B) = \bar{\mu}(A \cap T^{-1}B) + \sum_{n=2}^{\infty} \bar{\mu}\left(A_n \cap T^{-n}B\right) + \sum_{j=2}^{\infty} \sum_{n=j+1}^{\infty} \bar{\mu}\left(A_n \cap T^{-j}B\right) \\
&= \sum_{n=2}^{\infty} \nu\left((A_n \cap T^{-n}B, n-1)\right) + \sum_{n=1}^{\infty} \bar{\mu}(A_n \cap T^{-1}B) + \sum_{j=2}^{\infty} \sum_{n=j+1}^{\infty} \bar{\mu}\left(A_n \cap T^{-j}B\right) \\
&= \sum_{n=1}^{\infty} \nu\left((A_n \cap T^{-n}B, n-1)\right) + \sum_{j=1}^{\infty} \sum_{n=j+1}^{\infty} \bar{\mu}\left(A_n \cap T^{-j}B\right) \\
&= \nu(S^{-1}(A \cap B, 0)) + \sum_{j=1}^{\infty} \sum_{n=j+1}^{\infty} \bar{\mu}\left(A_n \cap T^{-j}B\right) \\
&= \nu((A \cap B, 0)) + \sum_{j=1}^{\infty} \sum_{n=j+1}^{\infty} \bar{\mu}\left(A_n \cap T^{-j}B\right) = \mu(B).
\end{aligned}$$

□

Kakutami towers are a powerful tool for investigating measurable systems. To get a feeling, solve the following two problems.

Problem 6.4 *Prove Kač theorem using Kakutani towers.*

Problem 6.5 *Given the measurable dynamical system (X, T, μ), prove that, calling T_A the return map to a set A, $\mu(A) > 0$, the tripe (A, T_A, μ) is a measurable dynamical system.*

6.8 Mixing

We have argued the importance of ergodicity, yet from a physical point of view ergodicity may be relevant only if it takes places at a sufficiently fast rate (i.e., if the time average converges to the space average on a physically meaningful time scale). This has prompted the study of stronger statistical properties of which we will give a brief, and by no mean complete, account in the following.

Definition 6.8.1 *A Dynamical System (X, T, μ) is called mixing if for every pairs of measurable sets A, B we have*

$$\lim_{n\to\infty} \mu(T^{-n}(A) \cap B) = \mu(A)\mu(B).$$

Obviously, if a system is mixing, then it is ergodic. In fact, if A is an invariant set for T, then $T^{-n}A \subset A$, so, calling A^c the complement of A, we have

$$\mu(A)\mu(A^c) = \lim_{n\to\infty} \mu(T^{-n}A \cap A^c) = 0,$$

and the measure of A is either one or zero.

An equivalent characterization of mixing is the following:

Proposition 6.8.2 *A Dynamical System (X, T, μ) is mixing if and only if*

$$\lim_{n\to\infty} \int_X f \circ T^n g d\mu = \int_X f d\mu \int_X g d\mu$$

for every $f, g \in L^2(X, \mu)$ or for every $f \in L^\infty(X, \mu)$ and $g \in L^1(X, \mu)$.[34]

The proof is rather straightforward and it is left as an exercise to the reader (see Problem 6.33) together with the proof of the next statement.

[34] The quantity $\int_X f \circ Tg - \int_X f \int_X g$ is called "correlation," and its tending to zero–which takes places always in mixing systems–it is called "decay of correlation."

Proposition 6.8.3 *A Dynamical System (X, T, μ), with X a compact metric space, T continuous and μ Borel, is mixing if and only if for each probability measure λ absolutely continuous with respect to μ*

$$\lim_{n\to\infty} \lambda(f \circ T^n) = \mu(f)$$

for each $f \in \mathcal{C}^0(\mathbb{T}^2)$.

This last characterization is interesting from a mathematical point of view. Define, as usual, the evolution of a measure via the equation

$$(T_*\lambda)(f) \equiv \lambda(f \circ T)$$

for each continuous function f. If for each measure, absolutely continuous with respect to the invariant one, the evolved measure converges weakly to the invariant measure, then the system is mixing (and thus the evolved measures converge strongly). This has also a very important physical meaning: if the initial configuration is known only in probability, the probability distribution is absolutely continuous with respect to the invariant measure, and the system is mixing, then, after some time, the configurations are distributed according to the invariant measure. Again the details of the evolution are not important to describe relevant properties of the system.

6.8.1 Examples

Rotations

We have seen that the translations by an irrational angle are ergodic. They are not mixing. The reader can easily see why.

Bernoulli shift

The key observation is that, given a measurable set A, for each $\varepsilon > 0$ there exists a set $A_\varepsilon \in \mathcal{A}$, thus depending only on a finite subset of indices,[35] with the property[36]

$$\mu(A_\varepsilon \backslash A) \leq \varepsilon.$$

Then, given A, B measurable, and for each $\varepsilon > 0$, let A_ε, B_ε be such an approximation, and I_A, I_B the defining sets of indices, then

$$\left|\mu(T^{-m}A \cap B) - \mu(A)\mu(B)\right| \leq 4\varepsilon + \left|\mu(T^{-m}A_\varepsilon \cap B_\varepsilon) - \mu(A_\varepsilon)\mu(B_\varepsilon)\right|.$$

[35] Remember, this means that there exists a finite set $I \subset \mathbb{Z}$ such that it is possible to decide if $\sigma \in \Sigma_n$ belongs or not to A_ε only by looking at $\{\sigma_i\}_{i\in I}$.

[36] This follows from our construction of the σ-algebra and by the definition of outer measure, see Examples 6.2.1–Bernoulli shift.

If we choose m so large that $(I_A+m)\cap I_B = \emptyset$, then by the definition of Bernoulli measure we have

$$\mu(T^{-m}A_\varepsilon \cap B_\varepsilon) = \mu(T^{-m}A_\varepsilon)\mu(B_\varepsilon) = \mu(A_\varepsilon)\mu(B_\varepsilon),$$

which proves

$$\lim_{m\to\infty} \mu(T^{-m}A \cap B) = \mu(A)\mu(B).$$

Dilation

This system is mixing. In fact, let $f,\, g \in \mathcal{C}^1(\mathbb{T})$, then we can represent them via their Fourier series $f(x) = \sum_{k\in\mathbb{Z}} e^{2\pi i k x} f_k$, $f_{-k} = \overline{f}_k$. It is well known that $\sum_{k\in\mathbb{Z}} |f_k| < \infty$ and $|f_k| \le \frac{c}{|k|}$, for some constant c depending on f. Therefore,

$$f(T^n x) = \sum_{k\in\mathbb{Z}} e^{2\pi i 2^n k x} f_k,$$

which implies that the only Fourier coefficients of $f\circ T^n$ different from zero are the $\{2^n k\}_{k\in\mathbb{Z}}$. Hence,

$$\left|\int_{\mathbb{T}} f\circ T^n g - \int_{\mathbb{T}} f \int_{\mathbb{T}} g\right| = \left|\sum_{k\in\mathbb{Z}} f_k g_{2^n k} - f_0 g_0\right| \le c2^{-n} \sum_{k\in\mathbb{Z}} |f_k|.$$

The previous inequalities imply the exponential decay of correlations for each smooth function. The proof is concluded by a standard approximation argument: given $f,\, g \in L^2(X,\, d\mu)$, for each $\varepsilon > 0$ exists $f_\varepsilon,\, g_\varepsilon \in \mathcal{C}^1(X)$: $\|f - f_\varepsilon\|_2 < \varepsilon$ and $\|g - g_\varepsilon\|_2 < \varepsilon$. Thus,

$$\left|\int_{\mathbb{T}} f\circ T^n g - \int_{\mathbb{T}} f \int_{\mathbb{T}} g\right| \le \left|\int_{\mathbb{T}} f_\varepsilon \circ T^n g_\varepsilon - \int_{\mathbb{T}} f_\varepsilon \int_{\mathbb{T}} g_\varepsilon\right| + 2(\|f\|_2 + \|g\|_2)\varepsilon,$$

which yields the result by choosing first ε small and then n sufficiently large.

6.9 Stronger statistical properties

One very fruitful idea in the realm of measurable dynamical systems is the idea of *entropy*. In some sense the entropy measure the complexity of the motions from a measure theoretical point of view.

To define it one starts by considering a partition of the space into measurable sets $\xi := \{A_1, \dots A_n\}$ and defines[37]

$$H_\mu(\xi) = -\sum_i \mu(A_i)\log\mu(A_i).$$

[37] The case of a countable partition, or even an uncountable partition, can be handled and it is very relevant, but outside the aims of this book, see [Roh67] for a complete treatment of the subject.

Given two partitions $\xi = \{A_i\}, \eta = \{B_j\}$ we define $\xi \vee \eta := \{A_i \cap B_j\}$. Let then be

$$\xi_{-n}^T := \xi \vee T^{-1}(\xi) \vee \cdots \vee T^{-n+1}(\xi).$$

It is then possible to prove that the sequence $H_\mu(\xi_{-n}^T)$ is sub-additive, hence the limit

$$h_\mu(T, \xi) := \lim_{n\to\infty} \frac{1}{n} H_\mu(\xi_{-n}^T)$$

exists.

Definition 6.9.1 *The entropy of T with respect to μ is defined as*

$$h_\mu(T) := \sup\{h_\mu(T, \xi) \mid H(\xi) < \infty\}$$

If a system has positive metric entropy this means that the motion has a high complexity and it is very far from being regular. One of the main property of entropy is that it is a metric invariant, that is if two systems are metrically conjugate (see the following), then they have the same metric entropy.

Even more extreme form statistical behaviors are possible, to present them we need to introduce the idea of equivalent systems. This is done via the concept of conjugation that we have already seen informally in Example 6.5.1 (logistic map, circle map).

Definition 6.9.2 *Two Dynamical Systems (X_1, T_1, μ_1), (X_2, T_2, μ_2) are (measurably) conjugate if there exists a measurable map $\phi : X_1 \to X_2$ almost everywhere invertible*[38] *such that $\mu_1(A) = \mu(\phi(A))$ and $T_2 \circ \phi = \phi \circ T_1$.*

Clearly, the conjugation is an equivalence relation. Its relevance for the present discussion is that conjugate systems have the same ergodic properties (Problem 6.42).[39]

We can now introduce the most extreme form of stochasticity.

Definition 6.9.3 *A dynamical system (X, T, μ) is called Bernoulli if there exists a Bernoulli shift (M, ν, σ) and a measurable isomorphism $\phi : X \to M$ (i.e., a measurable map one one and onto apart from a set of zero measure and with measurable inverse) such that, for each $A \in X$,*

$$\nu(\phi(A)) = \mu(A)$$

and

$$T = \phi^{-1} \circ \sigma \circ \phi.$$

[38] This means that there exists a measurable function $\phi^{-1} : X_2 \to X_1$ such that $\phi \circ \phi^{-1} = \mathrm{id}$ μ_2-a.e. and $\phi^{-1} \circ \phi = \mathrm{id}$ μ_1-a.e..

[39] Of course the reader can easily imagine other forms of conjugacy, e.g. topological or differential conjugation.

That is a system is Bernoulli if it is isomorphic to a Bernoulli shift. Since we have seen that Bernoulli systems are very stochastic (remind that they can be seen as describing a random event like coin tossing) this is certainly a very strong condition on the systems. In particular it is immediate to see that Bernoulli systems are mixing (Problem 6.42).

6.9.1 Examples

Dilation

We will show that such a system is indeed Bernoulli. The map ϕ is obtained by dividing $[0,\,1)$ in $[0,\,\frac{1}{2})$ and $[\frac{1}{2},\,1)$. Then, given $x \in \mathbb{T}$, we define $\phi : \mathbb{T} \to \Sigma_2^+$ by

$$\phi(x)_i = \begin{cases} 1 & \text{if } T^i x \in [0,\,\frac{1}{2}) \\ 2 & \text{if } T^i x \in [\frac{1}{2},\,1) \end{cases}$$

the reader can check that the map is measurable and that it satisfy the required properties. Note that the above shows that the Bernoulli measure with $p_1 = p_2 = \frac{1}{2}$ is nothing else than Lebesgue measure viewed on the numbers written in basis two. This may explain why we had to be so careful in the construction of the Bernoulli measure.

Baker

Let us define ϕ^{-1}; for each $\sigma \in \Sigma_2$

$$x = \sum_{i=0}^{\infty} \frac{\sigma_{-i}}{2^{i+1}},$$
$$y = \sum_{i=1}^{\infty} \frac{\sigma_i}{2^i}.$$

Again the rest is left to the reader.

Forced Pendulum

In the introduction we have seen that there exists a square Q with stable and unstable sides such that, calling T the map introduced by the flow at a proper time, $TQ \cap Q \supset Q_0^u \cup Q_1^u$. Where Q_i^u are rectangles that go from one stable side of Q to the other and, in analogy, $T^{-1}Q \cap Q \supset Q_0^s \cup Q_1^s$.

We can use this fact to code the dynamics similarly to what we have done for the Backer map. Namely, given the set $\Lambda = \bigcap_{n\in\mathbb{Z}} T^n Q$ (this set it is non empty–see Example 6.5.1–Horseshoe) and $\phi : \Lambda \to \Sigma_2$ define by

$$[\phi(x)]_k = \begin{cases} i \in \{0,1\} & \text{if } k \geq 0 \text{ and } T^k x \in Q_i^u \\ i \in \{0,1\} & \text{if } k < 0 \text{ and } T^k x \in Q_i^s. \end{cases}$$

It is easy to verify that ϕ is onto and that it is a.e. invertible. It remains to specify the measure on the Horseshoe, we can just pull back any invariant measure on the shift and we will get an invariant measure on the set Λ.

Let us conclude with a final remark on the physical relevance of the concept just introduced. As we mentioned, if f is an observable, then its ergodic average represents the result of an observation over a very long time (the time scale being determined by the mixing properties of the system). Yet, in reality, it may happen that we look for too short a time or, after studying a certain quantity, we can get a grant to buy the needed apparatus to perform more precise measurements. What would we see in such a case? Clearly, we would not see a constant, even for an ergodic system, and we would interpret the non constant part as fluctuations. In many cases it may happen that this fluctuations have a very special nature: they are Gaussian. In such a case we say that the system satisfies the Central Limit Theorem (CLT). Let us be more precise: define $S_n f := \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} f \circ T^i$.

Definition 6.9.4 *Given a Dynamical System (X, T, μ) and a class of observables $\mathcal{A} \subset L^2(X, \mu)$ we say that the class $\mathcal{A}$ satisfies the CLT if $\forall f \in \mathcal{A}$, $\mu(f) = 0$,*

$$\lim_{n\to\infty} \mu(\{x \mid S_n f \geq t\}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2\sigma^2}} dx,$$

where (the variance) σ is defined by $\sigma^2 = \mu(f) + 2\sum_{i=1}^{\infty} \mu(f \circ T^i f)$.[40]

The relevance of the above theorem is the following: if the system is ergodic and satisfies the CLT, then $\frac{1}{n} \sum_{i=0}^{n-1} f \circ T^i - \mu(f) = \mathcal{O}(\frac{1}{\sqrt{n}})$, we have thus the precise scale on which the fluctuations should appear.

In this book we will be mainly interested in the question of how to establish if a given system is ergodic or not.

Unfortunately, neither ergodicity is a typical property of dynamical systems, nor is regular motion. It is a frustrating fact of life that generically dynamical systems present some kind of mixed behavior. Nevertheless, there are some class of systems that are known to be ergodic and among them the hyperbolic systems are probably the most relevant. We will discuss them in the next chapters.

[40] This definition is a bit stricter than usual because, in general, there may be cases in which the fluctuations are Gaussian but the formula for the variance does not hold as written.

Problems

6.6. Given a measurable Dynamical Systems (X, T, μ) verify that, for each measurable set A, if $T(A)$ is measurable, then $\mu(TA) \geq \mu(A)$.

6.7. Set $\mathcal{M}^1(X) = \{\mu \in \mathcal{M} \mid \mu(X) = 1\}$ and $\mathcal{M}^1_T(X) = \mathcal{M}^1(X) \cap \mathcal{M}_T(X)$. Prove that $\mathcal{M}^1_T(X)$ and $\mathcal{M}^1(X)$ are convex sets in $\mathcal{M}(x)$.

6.8. Call $\mathcal{M}^e(X) \subset \mathcal{M}^1(X)$ the set of ergodic probability measures. Show that $\mathcal{M}^e(X)$ consists of the extremal points of $\mathcal{M}_T(X)$.

6.9. Prove that the Lebesgue measure is invariant for the rotations on $\mathbb{T}$.

6.10. Consider a rotation by $\omega \in \mathbb{Q}$, find invariant measures different from Lebesgue.

6.11. Prove that the measure μ_h defined in Examples 6.2.1 (Hamiltonian systems) is invariant for the Hamiltonian flow.

6.12. Given a Poincaré section prove that there exists $c > 0$ such that $\inf \tau_\Sigma \geq c > 0$.

6.13. Show that ν_Σ, defined in (6.3.2) is well defined.

6.14. Show that the return time τ_Σ is finite ν_Σ-a.e. .

6.15. Show that ν_Σ is T_Σ invariant. Verify that, collecting the results of the last exercises, $(\Sigma, T_\Sigma, \nu_\Sigma)$ is a Dynamical System.

6.16. something about holomorphic dynamics?

6.17. Prove that the Bernoulli measure is invariant with respect to the shift.

6.18. Let Σ_p be the set of periodic configurations of Σ. If μ is the Bernoulli measure prove that $\mu(\Sigma_p) = 0$

6.19. Consider the Bernoulli shift on $\mathbb{Z}$ and define the following equivalence relation: $\sigma \sim \sigma'$ iff there exists $n \in \mathbb{Z}$ such that $T^n\sigma = \sigma'$ (this means that two sequences are equivalent if they belong to the same orbit). Consider now the equivalence classes (the space of orbits) and choose[41] a representative from each class, call the set so obtained K. Show that K cannot be a measurable set.

6.20. Compute the transfer operator for maps of $\mathbb{T}$. Prove that $\|\mathcal{L}h\|_1 \leq \|h\|_1$.

6.21. Prove the Lasota-York inequality (6.5.6).

[41] Attention !!!: here we are using the *Axiom of choice*.

6.22. Prove that for each sequence $\{h_n\} \subset \mathcal{C}^{(1)}(\mathbb{T})$, with the property $\sup_{n\in\mathbb{N}} \|h_n'\|_1 + \|h_n\|_1 < \infty$, it is possible to extract a subsequence converging in L^1.

6.23. Prove Corollary 6.7.3.

6.24. Prove Theorem 6.7.6

6.25. Let $U \subset X$ of positive measure, consider

$$f_U(x) = \lim \frac{1}{n} \sum_{i=0}^{n-1} \chi_U(T^i x).$$

Show that the limit exists and that the set $A_0 := \{x \in U \mid f_U(x) = 0\}$ has zero measure.

6.26. A topological Dynamical System (X, T) is called *Topologically transitive*, if it has a dense orbit. Show that if $(\mathbb{T}^d, T, m)$ is ergodic and T is continuous, then the system is topologically transitive.

6.27. Give an example of a system with a dense orbit which it is not ergodic.

6.28. Give an example of an ergodic system with no dense orbit.

6.29. Give an example of a Dynamical Systems which does not have any invariant probability measure.

6.30. Prove that Birkhoff theorem implies Von Neumann theorem.

6.31. Prove that if (X, T, μ) is ergodic, then all $f \in L^1(X, \mu)$ such that $f \circ T = f$ are a.e. constant. Prove also the converse.

6.32. For each measurable set A, let

$$F_{A,n}(x) = \frac{1}{n} \sum_{i=0}^{n-1} \chi_A(T^i x).$$

be the average number of times x visits A in the time n. Show that there exists $F_A = \lim_{n\to\infty} F_{A,n}$ a.e. and prove that, if the system is ergodic, $F_A = \mu(A)$.

6.33. Prove Proposition 6.8.2 and Proposition 6.8.3.

6.34. Show that the irrational rotations are not mixing.

6.35. Prove that if $f \in \mathcal{C}^2(\mathbb{T})$, then its Fourier series converges uniformly.

6.36. Let ν be a Borel measure on $Q = [0,1]^2$ such that $\nu(\partial_x f) = 0$ for all $f \in \mathcal{C}^1_{\rm per}(Q) = \{f \in \mathcal{C}^1(Q) \mid f(0,y) = f(1,y)\ \forall\ y \in [0,1]\}$. Prove that there exists a Borel measure ν_1 on $[0,1]$ such that $\nu = m \times \nu_1$.

6.37. Prove that is a flow is ergodic (mixing) so is each Poincarè section. Prove that is a map is ergodic so is any suspension on the map. Give an example of a mixing map with a non-mixing suspension (constant ceiling).

6.38. Consider $([0,1], T)$ where

$$T(x) = \frac{1}{x} - \left[\frac{1}{x}\right]$$

($[a]$ is the integer part of a), and

$$\mu(f) = \frac{1}{\ln 2}\int_0^1 f(x)\frac{1}{1+x}dx.$$

Prove that $([0,1], T, \mu)$ is a Dynamical System.[42]

6.39. In view of the two previous exercises explain why it is problematic to study the statistical properties of the Gauss map on a computer.

6.40. Choose a number in $[0,1]$ at random according to Lebesgue distribution. Assuming that the Gauss map is mixing (which it is, see ???) compute the average percentage of numbers larger than n in the associated continuous fraction.

6.41. Let (X_0, T_0, μ_0) be a Dynamical System and $\phi : X_0 \to X_1$ an homeomorphism. Define $T_1 := \phi \circ T_0 \circ \phi^{-1}$ and $\mu_1(f) = \mu_0(f \circ \phi^{-1})$. Prove that (X_1, T_1, μ_1) is a Dynamical System.

6.42. Let (X_0, T_0, μ_0) be measurably conjugate to (X_1, T_1, μ_1), then show that one of the two is ergodic if and only if the other is ergodic. Prove the same for mixing.

6.43. Show that the systems described in Examples ??–strange attractor and horseshoe, are Bernoulli.

6.44. Prove Lebesgue density theorem: for each measurable set A, $m(A) > 0$, there exists $x \in A$ such that for each $\varepsilon > 0$ exists $\delta > 0$ such that $m(A \cap [x-\delta, x+\delta]) > (1-\varepsilon)2\delta$.

[42] The above map is often called *Gauss map* since to him is due the discovery of the above invariant measure.

Hints to solving the Problems

6.1 The first point is to define an orthogonal projection on the closed subspace V. For each $h \in H$, let $\alpha = \inf_{v \in V} \|h - v\|$. Let $\{v_n\}_{n \in \mathbb{N}} \subset V$ be such that $\lim_{n \to \infty} \|h - v_n\| = \alpha$. For each $w \in V$, and $t \in \mathbb{R}$, we have

$$\alpha^2 \leq \|h - v_n + tw\|^2 = \|h - v_n\|^2 - 2\langle h - v_n, w\rangle t + t^2 \|w\|^2.$$

For each $\varepsilon > 0$, there exists $n_\varepsilon \in \mathbb{N}$ such that, for all $n \geq n_\varepsilon$, we have $\|h - v_n\|^2 \leq \alpha^2 + \varepsilon$. Hence, for $n \geq n_\varepsilon$,

$$\|w\|^2 t^2 - 2\langle h - v_n, w\rangle t + \varepsilon \geq 0.$$

The above can happen for all $t \in \mathbb{R}$ only if $|\langle h - v_n, w\rangle| \leq \sqrt{\varepsilon}\|w\|$ for each $w \in V$. Accordingly, for $n, m \geq n_\varepsilon$,

$$\begin{aligned} \|v_n - v_m\|^2 &\leq |\langle v_n - h, v_n - v_m\rangle| + |\langle v_m - h, v_n - v_m\rangle| \\ &\leq 2\sqrt{\varepsilon}\|v_n - v_m\|. \end{aligned}$$

That it, $\{v_n\}$ is Chauchy, let $v \in \overline{V}$ be its limit, then $h - v \in V^\perp$. We can then write $h = v + (h - v)$ which shows that $H = \overline{V} \oplus V^\perp$.

6.2 It follows from Problem 6.1 which implies

$$H = V^\perp \oplus \overline{V} = V^\perp \oplus (V^\perp)^\perp.$$

6.3 If $x \in \tau_A^{-1}(n)$ for some $n \in \mathbb{N}$, than $T^n(x) \in A$, and $T^k(x) \notin A$, that is $Tk(x) \in A^c$, for all $k \in \{1, \ldots, n-1\}$. Thus

$$\tau_A^{-1}(n) \in T^{-n}(A) \cap \left[\bigcap_{i=1}^{n-1} T^{-i} A^c\right],$$

which is measurable once T is measurable.

6.4 Simply note that,

$$\int_A \tau_A(x)\mu(dx) = \sum_{n=1}^{\infty}\sum_{j=0}^{n-1} \mu(A_n) = \sum_{n=1}^{\infty}\sum_{j=0}^{n-1} \nu((A_n, j)) = \nu(Y) = 1.$$

6.5 Let $T_A(x) = T^{\tau_A(x)}(x)$ be the first return map; it suffices to check the invariance of μ. For each measurable set $B \subset A$,

$$S^{-1}(B, 0) = \cup_{n=1}^{\infty}(T^{-n}B \cap A_n, n-1),$$

and

$$\cup_{n=1}^{\infty} T^{-n}B \cap A_n = T_A^{-1}B.$$

Thus,

$$\mu(B) = \nu((B,0)) = \nu(S^{-1}(B,0)) = \sum_{n=0}^{\infty} \nu((T^{-n}B \cap A_n, n-1))$$
$$= \sum_{n=0}^{\infty} \mu(T^{-n}B \cap A_n) = \mu(T_A^{-1}B).$$

6.8 Use Krein-Milman Theorem [DS88].

6.11 Use the properties of H to deduce $\langle \nabla_{\phi^t x} H,\, d_x\phi^t \nabla_x H\rangle = \|\nabla_x H\|^2$, and thus $d_x\phi^t \nabla_x H = \frac{\|\nabla_x H\|^2}{\|\nabla_{\phi^t x} H\|^2} \nabla_{\phi^t x} H + v$ where $\langle \nabla_{\phi^t x} H,\, v\rangle = 0$. Then study the evolution of an arbitrarily small parallelepiped with one side parallel to $\nabla_x H$–or look at the volume form if you are more mathematically incline–remembering the invariance of the volume with respect to the flow.

6.13 Use the invariance of μ and the fact that, by Problem 6.12, if $A \subset \Sigma$ then $\mu(\phi^{[0,\delta]}(A) \cap \phi^{[n\delta,\,(n+1)\delta]}A) = 0$ provided $(n+1)\delta \le c$.

6.14 Let $\delta < c$ and $\Sigma_\delta := \phi^{[0,\delta]}\Sigma$, apply Poincaré return theorem to Σ_δ.

6.17 Check it on the algebra $\mathcal{A}$ first.

6.18 Σ_p is the countable union of zero measure sets.

6.19 Show that $K \cap T^n K \subset \Sigma_p$, then by using Problem 6.18 show that if K is measurable $\sum_{i=-\infty}^{\infty} \mu(T^n K) = 1$ which, by the invariance of μ, is impossible.

6.20 Use the equivalent definition $\int g\mathcal{L}f dm = \int fg \circ T dm$.

6.22 Consider partitions $\mathcal{P}_n$ of $\mathbb{T}$ in intervals of size $\frac{1}{n}$. Define the conditional expectation $\mathbb{E}(h|\mathcal{P}_n)(x) = \frac{1}{m(I(x))}\int_{I(x)} h dm$, where $x \in I(x) \in \mathcal{P}_n$. Prove that $\|\mathbb{E}(h|\mathcal{P}_n) - h\|_1 \le \frac{1}{n}\|h'\|_1$. Notice that the functions $\mathbb{E}(h_n|\mathcal{P}_m)$ have only m distinct values and, by using the standard diagonal trick, construct an subsequence h_{n_j} such that all the $\mathbb{E}(h_{n_j}|\mathcal{P}_m)$ are converging. Prove that h_{n_j} converges in L^1.

6.24 Note that $\mu(T^{-n}A \cap T^{-m}A) \neq 0$ then, supposing without loss of generality $n < m$, $\mu(A \cap T^{-m+n}A) \neq 0$. Then prove the theorem by absurd remembering that $\mu(X) < \infty$.

6.25 The existence follows from Birkhoff theorem, it also follows that A_0 is an invariant set, then

$$0 = \int_{A_0} f_U = \int_{A_0} \chi_U = \mu(A_0).$$

6.26 For each $n \in \mathbb{N}$, $x \in \mathbb{T}^d$ consider $B_{\frac{1}{m}}(x)$–the ball of radius $\frac{1}{m}$ centered at x. By compactness, there are $\{x_i\}$ such that $\cup_i B_{\frac{1}{m}}(x_i) = \mathbb{T}^d$. Let

$$A_{m,i} = \{y \in \mathbb{T}^d \mid T^k y \cap B_{\frac{1}{M}}(X_I) = \emptyset \ \ \forall k \in \mathbb{N}\},$$

clearly $A_{m,i} = \cap_{k\in\mathbb{N}} T^{-k} B_{\frac{1}{m}}(x_i)^c$ has the property $T^{-1}A_{m,i} \supset A_{m,i}$. It follows that $\tilde{A}_{m,i} = \cup_{n\in\mathbb{N}} T^{-n} A_{m,i} \supset A_{m,i}$ is an invariant set and it holds $\mu(\tilde{A}_{m,i} \backslash A_{m,i}) = 0$. Since $A_{m,i}$ it is not of full measure, $\tilde{A}_{m,i}$, and thus $A_{m,i}$, must have zero measure. Hence, $\bar{A}_m = \cap_i A_{m,i}$ has zero measure. This means that $\cup_{m\in\mathbb{N}} \bar{A}_m$ has zero measure. Prove now that, for each $y \in \mathbb{T}^d$, the trajectories that never get closer than $\frac{2}{m}$ to y are contained in $\bar{A}_m$, and thus have measure zero. Hence, almost every point has a dense orbit.)
Extend the result to the case in which X is a compact metric space and μ charges the open sets (that is: if $U \subset X$ is open, then $\mu(U) > 0$.

6.27 A system with two periodic orbits, and the measure supported on them. Along such lines more complex examples can be readily constructed.

6.28 A non transitive system with a measure supported on a periodic orbit.

6.29 $X = \mathbb{R}^d$, $Tx = x + v$, $v \neq 0$.

6.30 Note that the ergodic average is a contraction in L^∞, an isometry in L^2 and that $L^1 \subset L^2$ (since the measure is finite). Use Lebesgue dominate convergence theorem to prove convergence in L^2 for bounded functions. Use Fatou to show that if $f \in L^2$ then $f^+ \in L^2$ and a 3-ε argument to conclude.

6.32 Birkhoff theorem and Theorem 6.7.5.

6.33 Note that for each measurable set A and $\varepsilon > 0$ there exists $f \in \mathcal{C}^0(X)$ such that $\mu(|f - \chi_A|) < \varepsilon$ –by Uryshon Lemma and by the regularity of Borel measures. To prove that $\mu(T^{-n}A \cap B) \to \mu(A)\mu(B)$ choose $d\lambda = \mu(B)^{-1}\chi_B d\mu$ and use the invariance of μ to obtain the uniform estimate $\lambda(|f \circ T^n - \chi_A \circ T^n|) \leq \mu(B)^{-1}\mu(|f - \chi_A|)$.

6.35 Remember that $f_n = \frac{1}{2\pi}\int_{\mathbb{T}} e^{2\pi i n x} f(x) dx$. Thus

$$f_n = \frac{1}{(2\pi i n)^2 2\pi} \int_{\mathbb{T}} e^{2\pi i n x} f^{(2)}(x) dx.$$

6.36 The measure ν_1 is nothing else then the marginal with respect to x, that is: for each continuous function $f : [0,1] \to \mathbb{R}$ define $\tilde{f} : Q \to \mathbb{R}$ by $\tilde{f}(x,y) = f(y)$, then $\nu_1(f) = \nu(\tilde{f})$. To prove the statement use Fourier series. If f is smooth enough $f(x,y) = \sum_{k\in\mathbb{Z}} \hat{f}_k(y)e^{2\pi ikx}$ where the Fourier series for f and $\partial_x f$ converge uniformly. Then notice that $0 = \nu(\partial_x e^{2\pi ik\cdot}) = 2\pi ik\nu(e^{2\pi ik\cdot})$ implies $\nu(f) = \nu(\hat{f}_0) = m \times \nu_1(f)$.

6.38 Write $\mu(f \circ T) = \sum_{i=1}^{\infty} \int_{\frac{1}{i+1}}^{\frac{1}{i}} f \circ T(x)\mu(dx)$, change variable and use the identity $\frac{1}{a^2+a} = \frac{1}{a} - \frac{1}{a+1}$ to obtain a series with alternating signs.

6.39 The computer uses only rational numbers. It is quite amazing that these type of pathologies arises rather rarely in the numerical studies carried out by so many theoretical physicist.

6.40 Define $f(x) = [x^{-1}]$, then the entries of the continuous fraction of x are $\{f \circ T^i\}$. The quantity one must compute is then $m(\lim_{k\to\infty} \frac{i}{k} \sum_{i=0}^{k-1} \chi_{[n,\infty)} \circ f \circ T^i) = \mu([n,\infty))$.

6.44 We have seen in Examples 6.9.1-Dilations that Lebesgue measure is equivalent to Bernoulli measure and that the cylinder correspond to intervals. It then suffices to prove the theorem for the latter. Let $A \subset \Sigma^+$ such that $\mu(A) > 0$, then, for each $\varepsilon > 0$,there exists $A_\varepsilon \in \mathcal{A}$ such that $A_\varepsilon \supset A$ and $\mu(A_\varepsilon) - \mu(A) < \varepsilon\mu(A)$. Since $A_\varepsilon \in \mathcal{A}$, it exists $n_\varepsilon \in \mathbb{N}$ such that it is possible to decide if $\sigma \in A_\varepsilon$ only by looking at $\{\sigma_1, \ldots, \sigma_{n_\varepsilon}\}$. Consider all the cylinders $\mathcal{I}\{A(0; k_1, \ldots, k_{n_\varepsilon})\}$, clearly if $I \in \mathcal{I}$ then $I \cap A_\varepsilon$ is either I or $\emptyset$. Let $\mathcal{I}_+ = \{I \in \mathcal{I} \mid I \cap A_\varepsilon = I\}$ and $\mathcal{I}_+ = \{I \in \mathcal{I} \mid I \cap A_\varepsilon = \emptyset\}$. Now suppose that for each $I \in \mathcal{I}_+$ holds $\mu(I \cap A) \leq (1-\varepsilon)\mu(I)$ then

$$\mu(A) = \sum_{I\in\mathcal{I}_+} \mu(A \cap I) \leq (1-\varepsilon)\mu(A_\varepsilon) < \mu(A),$$

which is absurd. Thus there must exists $I \in \mathcal{I}_+$: $\mu(A \cap I) > (1-\varepsilon)\mu(I)$.

Notes

Give references for SRB and Gibbs, mention entropy, K-systems. diffeo with holes, strange attractors, history of the field

. . . . . .

Chapter 7

Quantitative Statistical Properties, a class of 1-d examples

Given a Dynamical System it is in general very hard to study its ergodic properties, especially if the goal is to have a *quantitative* understanding. To make clear what is meant by a *quantitative understanding* and which type of obstacles may prevent it, I devote this chapter to the study of a simple, but highly non-trivial, class of examples: one dimensional smooth expanding maps.

7.1 The problem

Recall from Examples 6.5.1 that a one dimensional smooth expanding map is a map $T \in \mathcal{C}^2(\mathbb{T}^1, \mathbb{T}^1)$ such that $|DT| \geq \lambda > 1$.

We know already that such maps have a unique absolutely continuous invariant measure (see sections 6.5.1, 6.6.1 Expanding maps).

We would like first to understand other invariant measures in order to have a clearer picture of which measurable Dynamical Systems can be associated with the topological Dynamical System $(\mathbb{T}^1, T)$. This is still at the qualitative level. In addition, we would like to have tools to compute such invariant measures with a given precision, which is a first quantitative issue.

Next, we would like to study statistical properties more in depth. To this end, we will restrict to the case $(\mathbb{T}^1, T, \mu)$, where μ is the measure absolutely continuous with respect to Lebesgue. The type of questions we would like to address is

If we make finite time and precision measurements, what do we observe?

Remember that a measurement is represented by the evaluation of a function. The fact that the measurement has a finite precision corresponds to the fact that the function has some uniform regularity (otherwise, we could identify the point with an arbitrary precision). The fact that the measure is made for a finite time means we can only measure finite-time averages. In other words, we would like to understand the behavior of

$$\sum_{k=0}^{N-1} f \circ T^k$$

for large, but finite, N.

7.2 Invariant measures

Let $\mathcal{M}$ be the set of probability (Borel) measures on $\mathbb{T}^1$. We can then consider the new Dynamical System $(\mathcal{M}, T')$, where $T'\mu(f) = \mu \circ T$ for all $f \in \mathcal{C}^0(\mathbb{T}^1, \mathbb{R})$. The invariant measures are the fixed points of T', let us call them $\mathrm{Fix}(T')$. If $\mu \in \mathrm{Fix}(T')$ then for each $h \in L^\infty(\mathbb{T}^1, \mu)$, $h \geq 0$, $\mu(h) = 1$, we can consider the new probability measure defined by $\mu_h(f) = \mu(hf)$, for all $f \in \mathcal{C}^0(\mathbb{T}^1, \mathbb{R})$. Note that

$$|T'\mu_h(f)| = |\mu(hf \circ T)| \leq |h|_{L^\infty(\mu)}\mu(|f| \circ T) = |h|_{L^\infty(\mu)}\mu(|f|).$$

Hence $T'\mu_h$ is absolutely continuous with respect to μ and $\frac{dT'\mu_h}{d\mu} \in L^\infty(\mu)$. We can then define the operator $\mathcal{L}_\mu : L^\infty(\mathbb{T}^1, \mu) \to L^\infty(\mathbb{T}^1, \mu)$ by $\mathcal{L}_\mu h := \frac{dT'\mu_h}{d\mu}$.

Let $\{I_i\}$ be a partition in interval of $\mathbb{T}^1$ such that $T|_{I_i}$ is invertible, $T(I_i) = \mathbb{T}^1$ and $\cup_i I_i = \mathbb{T}^1$. Call S_i the inverse of the i-th branch of T. Then, setting $\rho_i := \frac{dT'\mu_{\mathbb{1}_{I_i}}}{d\mu}$

$$\begin{aligned} T'\mu_h(f) &= \sum_i \mu(h\mathbb{1}_{I_i} f \circ T) = \sum_i \mu(\mathbb{1}_{I_i}(h \circ S_i f) \circ T) \\ &= \mu\left(\left[\sum_i \rho_i h \circ S_i\right] f\right). \end{aligned}$$

Thus, setting $\rho = \sum_i \rho_i \circ T\mathbb{1}_{I_i}$ we have

$$\frac{dT'\mu_h}{d\mu} = \sum_i (\rho h) \circ S_i =: \mathcal{L}_\rho(h).$$

It follows that $\mathcal{L}_\rho(1) = 1$ and, for each $h \in L^\infty(\mu)$,

$$\mu(\mathcal{L}_\rho(h)) = T'\mu_h(1) = \mu(h).$$

Problem 7.1 *Compute ρ and $\mathcal{L}_\rho$, in the case in which μ is the unique invariant measure absolutely continuous with respect to Lebesgue.*

The relevant fact is that one has the following (partial) converse.

Lemma 7.2.1 *For $\rho \in \mathcal{C}^0$, $\rho \geq 0$, let $\mathcal{L}_\rho(h)(x) := \sum_{y \in T^{-1}x} \rho(y)h(y)$. If there exists $\lambda \in \mathbb{R}$, $h \in \mathcal{C}^0$, $h > 0$, such that $\mathcal{L}_\rho h = \lambda h$, then there exists a measure $\mu \in \mathcal{M}$ such that $\mu(\mathcal{L}_\rho f) = \lambda\mu(f)$ for all $f \in \mathcal{C}^0$ and there exists an invariant measure absolutely continuous with respect to μ.*

PROOF. By continuity there exists $\gamma > 0$ such that $h \geq \gamma > 0$. Thus

$$|\mathcal{L}_\rho^n f| \leq \gamma^{-1}|f|_\infty \mathcal{L}_\rho^n h = \lambda^n \gamma^{-1}|h|_\infty |f|_\infty.$$

Hence, calling m the Lebesgue measure, $\frac{1}{n}\sum_{k=0}^{n-1} \lambda^{-k}(\mathcal{L}'_\rho)^k m$ is a weakly compact sequence. Accordingly the same arguments used in Krylov-Bogoliubov Theorem 6.5.2 imply that there exists a measure μ such that $\lambda^{-1}\mathcal{L}'_\rho\mu = \mu$.

Next, define $\nu(f) := \mu(hf)$. Clearly ν is a measure absolutely continuous with respect to μ, in addition

$$\nu(f \circ T) = \lambda^{-1}(\mathcal{L}'_\rho\mu)(hf \circ T) = \lambda^{-1}\mu(f\mathcal{L}_\rho h) = \mu(fh) = \nu(f).$$

□

7.3 Absolutely continuous invariant measure: revisited

We have already seen that there exists a unique invariant measure with respect to Lebesgue. Here we study this issue by a slightly different technique. Although the main idea is always to study the spectrum of the transfer operator, it is interesting to see how this can be achieved in many different ways, each way having its own advantages and disadvantages. Consider the transfer operator

$$\mathcal{L}h(x) := \sum_{y \in T^{-1}x} |D_y T|^{-1} h(y) \tag{7.3.1}$$

Problem 7.2 *Show that if $d\mu = hdm$, where m is the Lebesgue measure, then $\mu(f \circ T) = m(f\mathcal{L}h)$.*

Problem 7.3 *Show that, for each $n \in \mathbb{N}$,*

$$\mathcal{L}^n h(x) := \sum_{y \in T^{-n}x} |D_y T^n|^{-1} h(y)$$

Notice that, since DT cannot be zero, then its sign is constant. We limit ourselves, for simplicity, to the case $DT \geq \lambda$.

Problem 7.4 *Show that*

$$\begin{aligned}\frac{d}{dx}\mathcal{L}^n h(x) &= \sum_{y\in T^{-1}x} (D_y T)^{-2} h'(y) - D_y^2 T (D_y T)^{-3} h(y) \\ &= \mathcal{L}((DT)^{-1}h') - \mathcal{L}(D^2 T (DT)^{-2} h)\end{aligned}$$

7.3.1 A functional analytic setting

Let us consider first the Sobolev space $W^{1,1}$ and the space L^1.[1] Then, for each $h \in L^1(\mathbb{T}^1, m)$,

$$\int_{\mathbb{T}^1} |\mathcal{L}h| dm \leq \int_{\mathbb{T}^1} 1 \cdot \mathcal{L}|h| dm = \int_{\mathbb{T}^1} 1 \circ T |h| dm = \int_{\mathbb{T}^1} |h| dm \tag{7.3.2}$$

that is $\mathcal{L}$ is a bounded operator on L^1 and its norm is bounded by one.

In addition, remembering Exercise 7.2,

$$\int_{\mathbb{T}^1} |\frac{d}{dx}\mathcal{L}h| dm \leq \lambda^{-1} |h'|_{L^1} + D|h|_{L^1}, \tag{7.3.3}$$

where $D := \sup D^2 T (DT)^{-2}$.

Problem 7.5 *Iterate the* (7.3.2), (7.3.3) *and prove, for all* $n \in \mathbb{N}$,

$$\begin{aligned}|\mathcal{L}^n h|_{L^1} &\leq |h|_{L^1} \\ |\mathcal{L}^n h|_{W^{1,1}} &\leq \lambda^{-n} |h|_{W^{1,1}} + B|h|_{L^1}\end{aligned}$$

where $B = 1 + (1 - \lambda^{-1})^{-1} D$.

Since $W_{1,1}$ controls the L^∞ norm,[2] then we have that there exists $C > 0$ such that $|\mathcal{L}^n 1|_\infty < C$ for each $n \in \mathbb{N}$.

Using such a fact we can obtain similar inequalities in the Hilbert spaces L^2 and $W^{1,2}$. Indeed

$$\|\mathcal{L}^n h\|_{L^2}^2 = \int_{\mathbb{T}^1} h(\mathcal{L}^n h) \circ T^n \leq \|h\|_{L^2} \left[\int_{\mathbb{T}^1} (\mathcal{L}^n h)^2 \circ T^n\right]^{\frac{1}{2}} = \|h\|_{L^2}$$

$$\left[\int_{\mathbb{T}^1} (\mathcal{L}^n h)^2 \mathcal{L}^n 1\right]^{\frac{1}{2}} \leq C^{\frac{1}{2}} \|h\|_{L^2} \|\mathcal{L}^n h\|_{L^2}$$

[1] For an open set $U \subset \mathbb{R}$, the spaces $W^{p,q}(U)$ are the completion of $\mathcal{C}^\infty(U, \mathbb{C})$ with respect to the norms $\left[|f|_{L^q}^q + |f'|_{L^q}^q + \cdots + |f^{(p)}|_{L^q}^q\right]^{\frac{1}{q}}$. Note that they are all Banach spaces by construction but the $W^{p,2}$ are also Hilbert spaces (**Exercise**: write the scalar product).

[2] If $f \in \mathcal{C}^\infty$, then the mean value theorem asserts $\int h = h(\xi)$ for some ξ. Then $h(x) = h(\xi) + \int_\xi^x h'(z) dz$. Thus $|h|_\infty \leq |h|_{L^1} + |h'|_{L^1} = |h|_{W^{1,1}}$. The result extends then to all elements of $W^{1,1}$ by a standard approximation argument.

Which implies $\|\mathcal{L}^n h\|_{L^2} \leq C^{\frac{1}{2}} \|h\|_{L^2}$ for each $n \in \mathbb{N}$. Hence,

$$\|\frac{d}{dx}\mathcal{L}^n h\|_{L^2} \leq \lambda^{-n} C^{\frac{1}{2}} \|h'\|_{L^2} + D_n \|h\|_{L^2}.$$

Iterating as before we have, for all $n \in \mathbb{N}$,

$$\begin{aligned} |\mathcal{L}^n h|_{L^2} &\leq C|h|_{L^2} \\ |\mathcal{L}^n h|_{W^{1,2}} &\leq A\lambda^{-n}|h|_{W^{1,2}} + B|h|_{L^2}, \end{aligned} \tag{7.3.4}$$

for some appropriate constants A, B, C depending only on the map T.

To prove the existence of an invariant measure absolutely continuous with respect to Lebesgue we can try to mimic the Krylov-Bogolubov approach, but to do so we need a compactness result to substitute the weak compactness of the unit ball of the dual of a Banach space. This takes us in a very interesting detour in some fact of functional analysis.

7.3.2 Deeper in Functional analysis

Since we are on a circle it is a good idea to use Fourier series. For each function $h \in \mathcal{C}^\infty(\mathbb{T}, \mathbb{C})$ let h_k be its Fourier coefficients and define

$$(\mathbb{A}_n h)(x) = \sum_{|k| \leq m} h_k e^{2\pi i k x} \tag{7.3.5}$$

Clearly, for all $m > 0$,

$$\begin{aligned} |h - \mathbb{A}_m|^2_{L^2} &= \sum_{|k|>m} |h_k|^2 = \sum_{|k|>m} |h_k|^2 |k|^{-2} |k|^2 \leq m^{-2} \sum_{|k|>m} |(h')_k|^2 \\ &\leq m^{-2} |h'|^2_{L^2} \leq m^{-2} |h|^2_{W^{1,2}}. \end{aligned} \tag{7.3.6}$$

Using the above fact we can prove.

Lemma 7.3.1 *The unit ball of $W^{1,2}$ is (sequentially) compact in L^2.*

PROOF. Consider a sequence $\{h_m\} \subset W^{1,2}$, $|h_m|_{W^{1,2}} \leq 1$. Since $\mathbb{A}_l$ are all finite rank operators, $\{\mathbb{A}_l h_m\}$ for l fixed are contained in a bounded finite dimensional (hence compact) set, thus there exists a converging subsequence for all l while (7.3.6) shows that the sequences for fixed m are all convergent. Using the usual diagonalization trick we can then extract a converging subsequence. □

Consider now $h_n := \frac{1}{n}\sum_{k=0}^{n-1} \mathcal{L}^k 1$. By the above lemma $\{h_n\}$ is relatively compact and thus we can extract a subsequence $\{h_{n_j}\}$ converging in L^2. Let h_* be the limit. Note that $\int h_n = 1$ for all $n \in \mathbb{N}$, thus $h_* \not\equiv 0$ and $\int h_* = 1$.

Problem 7.6 *Show that $\mathcal{L}h_* = h_*$, that is $d\mu := h_* dm$ is an invariant measure absolutely continuous with respect to Lebesgue and with L^2 density.*

Of course, at this point it is natural to ask if μ is the only measure with such a property or there exist others. To answer such a question we need some more facts.

7.3.3 Even deeper in Functional analysis

Since we have to do it, let us do in the following general setting.

Consider two Banach space $(\mathbb{B}, \|\cdot\|)$ and $(\mathbb{B}_0, |\cdot|)$ such that $\mathbb{B} \subset \mathbb{B}_0$ and

i. $|h| \leq \|h\|$ for all $h \in \mathbb{B}$,

ii. if $h \in \mathbb{B}$ and $|h| = 0$, then $h = 0$.

iii. There exists $C > 0$: for each $\varepsilon > 0$ there exists a finite rank operator $\mathbb{A}_\varepsilon \in L(\mathbb{B}, \mathbb{B})$ such that $\|\mathbb{A}_\varepsilon\| \leq C$ and $|h - \mathbb{A}_\varepsilon h| \leq \varepsilon \|h\|$ for all $h \in \mathbb{B}$.[3]

In addition consider a bounded operator $\mathcal{L} : \mathbb{B}_0 \to \mathbb{B}_0$, constants $A, B, C \in \mathbb{R}_+$, and $\lambda > 1$, such that

a. $|\mathcal{L}^n| \leq C$ for all $n \in \mathbb{N}$,

b. $\mathcal{L}(B) \subset B$

c. $\|\mathcal{L}^n h\| \leq A\lambda^{-n}\|h\| + B|h|$ for all $h \in \mathbb{B}$ and $n \in \mathbb{N}$.

In particular $\mathcal{L}$ can be seen as a bounded operator on $\mathbb{B}$.

Theorem 7.3.2 *The spectral radius of the operator $\mathcal{L} \in L(\mathbb{B}, \mathbb{B})$ is bounded by 1 while the essential spectral radius is bounded by λ^{-1}.*[4]

We can now prove our main result.

Proof of Theorem 7.3.2. The first assertion is a trivial consequence of (c), (a) and (i).

[3] In fact, this last property can be weakened to: The unit ball $\{h \in \mathbb{B} : \|h\| \leq 1\}$ is relatively compact in $\mathbb{B}_0$. We use the present stronger condition since, on the one hand, it is true in all the applications we will be interested in and, on the other hand, drastically simplifies the argument. Note also that, if one uses the Fredholm alternative for compact operators rather than finite rank ones (Theorem E.0.1), then one can ask the $\mathbb{A}_\varepsilon$ to be compact instead than finite rank making easier their construction in concrete cases.

[4] The definition of *essential spectrum* varies a bit from book to book. Here we call essential spectrum the complement, in the spectrum, of the isolated eigenvalues with associated finite dimensional eigenspaces (which is also called the Fredholm spectrum).

The second part is much deeper. Let $\mathcal{L}_{n,\varepsilon} := \mathcal{L}^n \mathbb{A}_\varepsilon$, clearly such an operator is finite rank, in addition

$$\|\mathcal{L}^n h - \mathcal{L}_{n,\varepsilon} h\| \leq A\lambda^{-n}\|(\mathbb{1}-\mathbb{A}_\varepsilon)h\| + B|(\mathbb{1}-\mathbb{A}_\varepsilon)h| \leq A(1+C)\lambda^{-n}\|h\| + B\varepsilon\|h\|.$$

By choosing $\varepsilon = \lambda^{-n}$ we have that there exists $C_1 > 0$ such that

$$\|\mathcal{L}^n - \mathcal{L}_{n,\varepsilon}\| \leq C_1 \lambda^{-n}.$$

For each $z \in \mathbb{C}$ we can now write

$$\mathbb{1} - z\mathcal{L} = (\mathbb{1} - z(\mathcal{L} - \mathcal{L}_{n,\varepsilon})) - z\mathcal{L}_{n,\varepsilon}.$$

Since

$$\|z(\mathcal{L} - \mathcal{L}_{n,\varepsilon})\| \leq |z| C_1 \lambda^{-n} < \frac{1}{2},$$

provided that $|z| \leq \frac{1}{2C_1}\lambda^n$. Thus, given any z in the disk $D_n := \{|z| < \frac{1}{2C_1}\lambda^n\}$ the operator $B(z) := \mathbb{1} - z(\mathcal{L} - \mathcal{L}_{n,\varepsilon})$ is invertible.[5] Hence

$$\mathbb{1} - z\mathcal{L} = \left(\mathbb{1} - z\mathcal{L}_{n,\varepsilon} B(z)^{-1}\right) B(z) =: (\mathbb{1} - F(z))B(z).$$

By applying Fredholm analytic alternative (see Theorem E.0.1 for the statement and proof in a special case sufficient for the present purposes) to $F(z)$ we have that the operator is either never invertible or not invertible only in finitely many points in the disk D_n. Since for $|z| < 1$ we have $(\mathbb{1} - z\mathcal{L})^{-1} = \sum_{n=0}^{\infty} z^n \mathcal{L}^n$, the first alternative cannot hold hence the Theorem follows. □

7.3.4 The harvest

We are finally in the position to use all the above result to gain a deep understanding of the properties of the Dynamical Systems under consideration.

Problem 7.7 *Show that Theorem 7.3.2 implies that there exists $\sigma \in (0,1)$, $\{\theta_k\}_{k=1}^p$ and $L > 0$ such that*

$$\mathcal{L} = \sum_{k=1}^{p} e^{i\theta_k} \Pi_{\theta_k} + R$$

where Π_{θ_k} and R are operators on $W^{1,2}$ such that $\Pi_{\theta_k}\Pi_{\theta_j} = \delta_{jk}\Pi_{\theta_k}$ and $R\Pi_{\theta_k} = \Pi_{\theta_k} R = 0$. Moreover $|R^n| \leq L\sigma^n$.(Hint: Read section 6 of the Third Chapter of [Kat66] and recall that the operator is power bounded to exclude Jordan blocks.)

[5] Clearly $B(z)^{-1} = \sum_{n=0}^{\infty} [z(\mathcal{L} - \mathcal{L}_{n,\varepsilon})]^n$.

The above implies that

$$\Pi_\theta := \lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} e^{-i\theta k}\mathcal{L}^k = \begin{cases} \Pi_{\theta_i} & \text{iff } \theta = \theta_j \\ 0 & \text{otherwise.} \end{cases} \tag{7.3.7}$$

Problem 7.8 *Using equations (7.3.4) show that, for each $h \in L^2$*

$$\|\Pi_\theta h\|_{W^{1,2}} \le C\|h\|_{L^2}.$$

(Hint: prove it first for $h \in W^{1,2}$ and then do a density argument).

Next, note that Exercise 7.6 implies that $h_* = \Pi_0 1 \neq 0$, that is one is in the spectrum on $\mathcal{L}$, this means that the spectral radius of $\mathcal{L}$ is one.

Accordingly, if $\Pi_\theta h = h$ we have $h \in W^{1,2} \subset \mathcal{C}^0$ and[6]

$$|h| = |\Pi_\theta h| \le \lim_{j\to\infty} \frac{1}{n_j}\sum_{k=0}^{n_j-1} \mathcal{L}^k|h| = \Pi_0|h| \le |h|_\infty h_*.$$

This means that all the eigenvectors of the peripheral spectrum are of the form $h = gh_*$ with $g \in \mathcal{C}^0$. Thus, if h_i is an $W^{1,2}$ orthonormal a base of the eigenspace associated to an eigenvalue θ, then the eigenprojector must have the form

$$\Pi_\theta h = \sum_i h_i \int \ell_i \cdot h,$$

with $\ell_i \in L^2$ and $\int \ell_i h_j = \delta_{ij}$. Hence $\Pi_\theta \mathcal{L} = e^{i\theta}\Pi_\theta$ implies

$$e^{i\theta}\sum_k h_k \int \ell_k \cdot h = \sum_k h_k \int \ell_k \cdot \mathcal{L}h = \sum_k h_k \int \ell_k \circ T \cdot h.$$

That is $e^{i\theta}\ell_k = \ell_k \circ T$. But then if we set $f_k := \bar{\ell}_k h_* \in L^2$, we have

$$\mathcal{L}f_k = e^{i\theta}\mathcal{L}(\bar{\ell}_k \circ T h_*) = e^{i\theta}\bar{\ell}_k \mathcal{L}h_* = e^{i\theta}\bar{\ell}_k h_* = e^{i\theta} f_k$$

By the above facts, this implies $\Pi_\theta f_k = f_k \in W^{1,2}$, that is $\ell_k \in \mathcal{C}^0$. But then for each $p \in \mathbb{N}$ we can set $h_p := \bar{\ell}_k^p h_*$ obtaining

$$\mathcal{L}h_p = e^{ip\theta}h_p.$$

Since the the peripheral spectrum consists of finitely many eigenvalues it follows that there must exist $p \in \mathbb{N}$ such that $p\,\theta = \theta \mod 2\pi$, that is the

[6] Remember that exercise 7.8 implies that the sequence in (7.3.7) converges in L^2, accordingly there exists a subsequence that converges almost everywhere with respect to Lebesgue.

spectrum on the unit circle must be the union of finitely many cyclic groups. In turn this implies that there exists $\bar{p} \in \mathbb{N}$ such that $\bar{p}\,\theta = 0 \mod 2\pi$, hence $\ell_k^{\bar{p}} = \ell_k^{\bar{p}} \circ T$. But this implies that if we define the sets $A_L := \{x \in \mathbb{T} : |\ell_k^{\bar{p}}| \leq L\}$, $L \in \mathbb{R}$, they are all invariant. So if χ_L is the characteristic function of the set A_L, then $\chi_L \circ T = \chi_L$ and $\mathcal{L}(\chi_L h_*) = \chi_L h_*$. We can thus produce a lot of eigenvalues of $\mathcal{L}$, but we know that such eigenvalues form a finite dimensional space. The only possibility is that only finitely many of the A_L are different. This is like saying that ℓ_k takes only finitely many values. But $\ell_k^{\bar{p}}$ is a continuous function, so it must be constant. Hence ℓ_k can assume only $\bar{p}$ different values, thus, again by continuity, must be constant. Finally this implies $\theta = 0$.

The conclusion is that one is the only eigenvalue on the unit circle and that the associated eigenprojector has rank one. So one is a simple eigenvalue and h_* is the only invariant density for the map.

7.3.5 conclusions

If we have any probability measure ν absolutely continuous with respect to Lebesgue and with density $h \in W^{1,2}$, then setting $d\mu = h_* dm$, for each $\varphi \in W^{1,2}$ we have

$$|\mu(\varphi \circ T^n) - \nu(\varphi \circ T^n)| = \left| \int \varphi \mathcal{L}^n (h - h_*) \right| \leq \|\varphi\|_{1,2} C \sigma^n \|h - h_*\|_{1,2}$$

where σ is the largest eigenvalue of modulus smaller than one (or λ^{-1} is no such eigenvalue exist).

Remark 7.3.3 *The above means that the evolution of the present* chaotic *system, if seen at the level of the absolutely continuous measures, becomes simply a dynamics with an uniformly attracting fixed point, the simplest dynamics of all!*

7.4 General transfer operators

In the previous sections we have been very successful in studying the measure absolutely continuous with respect to Lebesgue. We have seen in §7.2 (crf. Lemma 7.2.1) that to study other invariant measures one has to analyze more general transfer operators. Here we will restrict ourselves to studying

$$\mathcal{L}_\phi h := \mathcal{L}(e^\phi h)$$

where $\mathcal{L}$ is the usual transfer operator. This are called *transfer operators with weight* and ϕ is sometime called the *potential*. We will consider first the case of $\phi : \mathbb{T}^1 \to \mathbb{C}$ and specialize to real potential later on.

For convenience, and also for didactical purposes, we will use the Banach spaces $\mathcal{C}^1$ and $\mathcal{C}^0$. Hence, form now on, we will assume $T \in \mathcal{C}^2(\mathbb{T}^1, \mathbb{T}^1)$ and $\phi \in \mathcal{C}^1(\mathbb{T}^1, \mathbb{C})$.

The first step is to compute the powers of $\mathcal{L}_\phi$ and study how they behave with respect to derivation.

Problem 7.9 *Show that, for each $n \in \mathbb{N}$, holds true*

$$\mathcal{L}_\phi^n h = \mathcal{L}^n \left[e^{\phi_n} h \right],$$

where $\phi_n = \sum_{k=0}^{n-1} \phi \circ T^k$.

Problem 7.10 *Show that for each $n \in \mathbb{N}$ and $h \in \mathcal{C}^1$ holds true*

$$\frac{d}{dx}\mathcal{L}_\phi^n h = \mathcal{L}_\phi^n \left[\frac{h'}{(T^n)'} - \frac{(T^n)''}{[(T^n)']^2} h + \frac{(\phi_n)'}{(T^n)'} h \right]$$

Note that $|\mathcal{L}_\phi^n h|_\infty \leq |h|_\infty \mathcal{L}_{\Re(\phi)}^n 1$. In addition,[7]

$$\begin{aligned} \left| \frac{(T^n)''(y)}{[(T^n)'(y)]^2} \right| &= \left| \frac{\frac{d}{dy}\prod_{k=0}^{n-1} T'(T^k y)}{[(T^n)'(y)]^2} \right| \\ &\leq \sum_{k=0}^{n-1} \left| \frac{T''(T^k y)}{(T^{n-k})'(T^k y)} \right| \leq \sum_{k=0}^{n-1} |T''|_\infty \lambda^{-n+k+1} \leq \frac{|T''|_\infty}{1-\lambda^{-1}}. \end{aligned}$$

Analogously,

$$\left| \frac{(\phi_n)'}{(T^n)'} \right| \leq \frac{|\phi'|_\infty}{1-\lambda^{-1}}.$$

The above inequalities imply

$$\left| \frac{d}{dx}\mathcal{L}_\phi^n h \right| \leq \lambda^{-n} \mathcal{L}_{\Re(\phi)}^n |h'| + B\mathcal{L}_{\Re(\phi)}^n |h|. \tag{7.4.8}$$

Which, taking the sup over x, yields

$$\left| \frac{d}{dx}\mathcal{L}_\phi^n h \right|_\infty \leq \lambda^{-n} |h'|_\infty \mathcal{L}_{\Re(\phi)}^n 1 + B_* |h|_\infty \mathcal{L}_{\Re(\phi)}^n 1,$$

Note that the above inequality implies that the spectral radius is bounded by $\rho = \lim_{n\to\infty} \|\mathcal{L}_{\Re(\phi)}^n 1\|_{\mathcal{C}^0}^{\frac{1}{n}}$ while the essential spectral radius is bounded by $\lambda^{-1}\rho$. The reader should notice that for positive potentials the above bounds are essentially sharp while for non positive, or complex, potential typically there will be cancellations that induce a smaller spectral radius. To control exactly such cancellations is, in general, a very hard problem.

[7] The quantity estimated here is usually called *distortion*. In fact, it measure how much the maps distorts intervals.

7.4.1 Real potential

In this section we will restrict to the case of $\phi \in \mathcal{C}^1(\mathbb{T}^1, \mathbb{R})$, i.e. real potentials.

If we define the cone $\mathcal{C}_a := \{h \in \mathcal{C}^1 \ : \ h > 0 \ |h'(x)| \leq a h(x)\}$, then equation (7.4.8), for $h > 0$, implies that, for each $\sigma \in (0, \lambda^{-1})$, $\mathcal{L}_\phi \mathcal{C}_a \subset \mathcal{C}_{\sigma_a}$ provided $a \geq B(\sigma - \lambda^{-1})^{-1}$.[8] We can then apply the theory of Appendix A to conclude the following.

Lemma 7.4.1 *For each real potential $\phi \in \mathcal{C}^1(\mathbb{T}^1, \mathbb{R})$, the transfer operator $\mathcal{L}_\phi$ has the Perron-Frobenius property, i.e. it has a simple strictly positive maximal eigenvalue ϱ and all the other eigenvalues are strictly smaller in modulus. In particular, the maximal eigenvalue $\varrho(t)$ of $\mathcal{L}_{\tau\phi}$, $\tau \in \mathbb{R}$, is analytic in τ.*[9]

The above, together with Lemma 7.2.1 imply that there exists μ_ϕ, h_ϕ such that $\mathcal{L}'_\phi \mu_\phi = e^\varrho \mu_\phi$, $\mathcal{L}_\phi h_\phi = e^\varrho h_\phi$, $h_\phi \in \mathcal{C}_a$. Moreover, $\nu_\phi(\varphi) = \mu_\phi(\varphi h_\phi)$ is the invariant measure associated to the potentitial ϕ.

7.4.2 Variational principle

Given the above facts it is natural to ask what is the maximal eigenalue of $\mathcal{L}_\phi$. To answer this question, we have to introduce new concepts: *the topological entropy* and the *the topological pressure*.

To this end let us define a dynamical ball

$$B_n(x, \varepsilon) = \{z \in \mathbb{T} \ : \ |T^k(x) - T^k(z)| \leq \varepsilon, k \in \{0, \dots, n\}\}.$$

We call $\mathcal{S}_{\varepsilon,n}$ the set of (ε, n)-covering sets, that is the finite sets of points E such that $\bigcup_{x\in E} B_n(x, \varepsilon) = \mathbb{T}^1$. We call $\mathcal{N}_{\varepsilon,n}$ the set of (ε, n)-separting sets, that is the finite sets of points E such that, for all $x, y \in E, x \neq y$, $B_n(x, \varepsilon) \cap B_n(y, \varepsilon) = \emptyset$. We then set

$$S(T, \phi, \varepsilon, n) = \inf_{E\in\mathcal{S}_{\varepsilon,n}} \sum_{x\in E} e^{\sum_{k=0}^{n-1} \phi\circ T^k(x)}$$

$$N(T, \phi, \varepsilon, n) = \sup_{E\in\mathcal{N}_{\varepsilon,n}} \sum_{x\in E} e^{\sum_{k=0}^{n-1} \phi\circ T^k(x)}$$

We are now ready to introduce the topological pressure

$$\begin{aligned} P_{top}(T, \phi) &= \lim_{\varepsilon\to 0} \liminf_{n\to\infty} \frac{1}{n} \ln S(T, \phi, \varepsilon, n) \\ &= \lim_{\varepsilon\to 0} \limsup_{n\to\infty} \frac{1}{n} \ln N(T, \phi, \varepsilon, n). \end{aligned} \tag{7.4.9}$$

[8] Note that this cone is almost the same than the one in Example 6.6.1, more precisely is its infinitesimal version.

[9] This follows from the fact that the maximal eigenvalue must always be simple and the results in Appendix C.4. This class of potentials is relevant in the so called *thermodynamic formalism*.

Problem 7.11 *Prove that the limits in (7.4.9) are both well defined and equal.*

It follows that, for each $E \in \mathcal{S}_{\varepsilon,n}$

$$1 = \nu_\phi(1) \leq C\mu_\phi(1) \leq C \sum_{x\in E} \mu_\phi(B_n(x,\varepsilon)) = Ce^{-n\varrho} \sum_{x\in E} \mu_\phi(\mathcal{L}_\phi^n \mathbb{1}_{B_n(x,\varepsilon)}).$$

Note that, for $\varepsilon < 1$, the ball $B_n(x,\varepsilon)$ can contain at most a preimmage, under T^n, of a point $z \in \mathbb{T}^1$. Hence, by the usual distortion estimates,

$$\mathcal{L}_\phi^n \mathbb{1}_{B_n(x,\varepsilon)}(z) \leq Ce^{\sum_{k=0}^{n-1} \phi\circ T^k(x)},$$

which implies

$$S(T,\phi,\varepsilon,n) \geq ce^{-n\varrho}.$$

Analogously, if $E \in \mathcal{N}_{\varepsilon,n}$,

$$1 = \nu_\phi(1) \geq c\mu_\phi(1) \geq c \sum_{x\in E} \mu_\phi(B_n(x,\varepsilon)) \geq ce^{-(n+m)\varrho} \sum_{x\in E} \mu_\phi(\mathcal{L}_\phi^{n+m} \mathbb{1}_{B_n(x,\varepsilon)}).$$

Note that, if $\lambda^m \geq \varepsilon^{-1}$, then each $z \in \mathbb{T}^1$ has at least a preimage, under T^{n+m} in $B_n(x,\varepsilon)$, thus

$$N(T,\phi,\varepsilon,n) \leq ce^{-(n+m)\varrho}.$$

From the above facts and the definition (7.4.9) follows that

$$P_{top}(T,\phi) = \varrho.$$

We have thus identified the maximal eigenvalue of $\mathcal{L}_\phi$. To have a more explicit expression we need the following deep local characterization of the entropy.

Theorem 7.4.2 (Theorem [BK83]) *For each invariant measure ν we have that for ν almost all $x \in \mathbb{T}^1$*

$$h_\nu(T) = \lim_{\varepsilon\to 0} \limsup_{n\to\infty} -\frac{1}{n} \ln\left(\nu(B_n(x,\varepsilon))\right)$$

Arguing as before we have

$$ce^{-(n+m)\varrho} e^{\sum_{k=0}^{n+m-1} \phi\circ T^k(x)} \leq \nu_\phi(B_n(x,\varepsilon)) \leq Ce^{-n\varrho} e^{\sum_{k=0}^{n-1} \phi\circ T^k(x)}$$

thus, recalling Birkhoff theorem,

$$h_{\nu_\phi}(T) = \varrho - \lim_{\varepsilon\to 0} \limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \phi \circ T^k(x) = P_{top}(T,\phi) - \int_{\mathbb{T}^1} \phi d\nu_\phi.$$

Finally, arguing as in [HK95, Theorem 20.3.7], we can establish the *varitional principle*

Theorem 7.4.3 *Let $\mathcal{M}(T)$ be the set of invariant probability measures for T, then*

$$P_{top}(T,\phi) = \sup_{\nu\in\mathcal{M}(T)} h_\nu(T) + \int_{\mathbb{T}^1} \phi d\nu.$$

7.5 Limit Theorems

Given $f \in \mathcal{C}^1$, $n \in \mathbb{N}$ and $a \in \mathbb{R}_+$ let

$$A_{a,n}(f) := \left\{ x \in \mathbb{T}^1 \ : \ \left| \frac{1}{n} \sum_{k=0}^{n-1} f \circ T^k(x) - \mu(f) \right| \geq a \right\}. \tag{7.5.10}$$

By the ergodic theorem $\lim_{n\to\infty} \mu(A_{a,n}(f)) = 0$. A natural question is:

Question 3 *How large is $m(A_{a,n})$?*

Note that we can write $\frac{1}{n}\sum_{k=0}^{n-1} f \circ T^k(x) - \mu(f) = \frac{1}{n}\sum_{k=0}^{n-1} \hat{f} \circ T^k(x)$ where $\hat{f} := f - \mu(f)$. So we can reduce the question to the study of zero average function. A more refined question could be.

Question 4 *Does it exists a sequence $\{c_n\}$ such that*

$$\frac{1}{c_n} \sum_{k=0}^{n-1} \hat{f} \circ T^k(x)$$

converges in some sense to a non zero finite object?

7.5.1 Large deviations. Upper bound

Note that it suffices to study the set

$$A^+_{a,n}(f) := \left\{ x \in \mathbb{T}^1 \ : \ \frac{1}{n} \sum_{k=0}^{n-1} f \circ T^k(x) - \mu(f+a) \geq 0 \right\}.$$

since $A_{a,n}(f) = A^+_{a,n}(f) \cap A^+_{a,n}(-f)$. On the other hand, setting $\hat{f} := f - \mu(f)$, for each $\lambda \geq 0$ we have

$$\begin{aligned} m(A^+_{a,n}(f)) &= m(\{x \ : \ e^{\lambda \sum_{k=0}^{n-1} (\hat{f} \circ T^k(x) - a)} \geq 1\}) \leq e^{-n\lambda a} m(e^{\lambda \sum_{k=0}^{n-1} \hat{f} \circ T^k}) \\ &= e^{-n\lambda a} m(e^{\lambda \sum_{k=0}^{n-1} \hat{f} \circ T^k}). \end{aligned}$$

Accordingly,

$$m(A^+_{a,n}(f)) \leq e^{-n\lambda a} m(\mathcal{L}^n_\lambda 1) \tag{7.5.11}$$

where we have defined the operator $\mathcal{L}_\lambda g := \mathcal{L}(e^{\lambda \hat{f}} g)$, $\mathcal{L}$ being the Transfer operator of the map T.

By Lemma 7.4.1 $\mathcal{L}_\lambda$ has a maximal eigenvalue α_λ depending analytically on λ. Hence by the same argument used in Lemma 7.2.1 there exists $c \in \mathbb{R}$ such that

$$m(A^+_{a,n}(f)) \leq e^{-n(\lambda a - \ln \alpha_\lambda) + c}.$$

Since λ has been chosen arbitrarily we have obtained

$$m(A^+_{a,n}(f)) \leq e^{-n\tilde{I}(a)+c} \tag{7.5.12}$$

where $\tilde{I}(a) := \sup_{\lambda\in\mathbb{R}^+}\{\lambda a - \ln \alpha_\lambda\}$. The problem is then reduced to studying the function $I(a)$ which is commonly called *rate function*. Note that $\tilde{I}$ is not necessarily finite. Indeed if $a > \|\hat{f}\|_\infty$, then clearly $m(A^+_{a,n}(f)) = 0$.

To better understand the rate function it is helpful to make a little digression into convex analysis.

Recall that a function $f : \mathbb{R}^d \to \mathbb{R}^d$ is convex if for each $x, y \in \mathbb{R}^d$ and $t \in [0, 1]$ we have $f(ty + (1 - t)x) \leq tf(y) + (1 - t)f(x)$ (if the inequality is everywhere strict, then the function is *stricly convex*.

Problem 7.12 *Show that if $f \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$, then f is convex iff $\frac{\partial^2 f}{\partial x^2}$ is a positive matrix.*[10] *Give a condition for strict convexity.*

Problem 7.13 *If a function $f : D \subset \mathbb{R}^d \to \mathbb{R}$, D convex,*[11] *is convex and bounded, then it is continuous.*

Given a function $f : \mathbb{R}^d \to \mathbb{R}$ let us define its *Legendre transform* as

$$f^*(x) = \sup_{y\in\mathbb{R}^d}\{\langle x, y\rangle - f(y)\} \tag{7.5.13}$$

Remark that f^* can take the value $+\infty$.

Problem 7.14 *Prove that f^* is convex.*

Problem 7.15 *Prove that $f^{**} \leq f$.*

Problem 7.16 *Prove that is $f \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$ is strictly convex, then the function $h(y) := \frac{\partial f}{\partial y}(y)$ is invertible and f^* is strictly convex. Moreover, calling g the inverse function of h, we have*

$$f^*(x) = \langle x, g(x)\rangle - f \circ g(x).$$

Problem 7.17 *Show that if $f \in \mathcal{C}^2$ is strictly convex, then $f^{**} = f$.*

Problem 7.18 *Show that, for each $x, y \in \mathbb{R}^d$, $\langle x, y\rangle \leq f^*(x) + f(y)$, (Young inequality).*

[10] A matrix $A \in GL(\mathbb{R}, d)$ is called *positive* if $A^T = A$ and $\langle v, Av\rangle \geq 0$ for each $v \in \mathbb{R}^d$.
[11] A set D is convex if, for all $x, y \in D$ and $t \in [0, 1]$, olds true $ty + (1 - t)x \in D$.

From the above discussion it follows that the rate function is defined very similarly to the Legendre transform of the logarithm of the maximal eigenvalue, which is commonly called *pressure of* $\hat{f}$. In fact, setting $I(a) = \max_{\lambda\in\mathbb{R}}(\lambda a - \ln \alpha_\lambda)$ we will se that, for $a \geq 0$, $I(a) = \tilde{I}(a)$. Unfortunately, to see that the rate function is exactly a Legendre transform takes some work. Let us start by studying the function α_λ.

Lemma 7.5.1 *There exists continuous functions $C_\lambda > 0$ and $\rho_\lambda \in (0,1)$ such that, for $\lambda \leq 0$, $\mathcal{L}_\lambda = \alpha_\lambda \Pi_\lambda + Q_\lambda$, $\Pi_\lambda Q_\lambda = Q_\lambda \Pi_\lambda = 0$, $\|Q_\lambda^n\|_{\mathcal{C}^1} \leq C_\lambda \rho_\lambda^n \alpha_\lambda^n$. Also $\Pi_\lambda(g) = h_\lambda \ell_\lambda(g)$, $\ell_\lambda(h_\lambda) = 1$, $\ell_\lambda(h'_\lambda) = 0$. In addition, $\mu_\lambda(\cdot) := \ell_\lambda(h_\lambda\, \cdot)$ is an invariant probability measure. Moreover everything is analytic in λ.*

Proof. As we have seen, there exists $h_\lambda \in \mathcal{C}^1$ and a measure ℓ_λ, both analytic in λ, such that the projection on the maximal eigenvalue of $\mathcal{L}_\lambda$ reads $\Pi_\lambda(h) = h_\lambda \ell_\lambda(h)$. Obviously

$$\mathcal{L}_\lambda h_\lambda = \alpha_\lambda h_\lambda, \tag{7.5.14}$$

and $\alpha_0 = 1$, $h_0 = h$ and $\ell_0 = m$. Notice that h_λ and ℓ_λ are not uniquely defined: by $\Pi_\lambda^2 = \Pi_\lambda$ follows $\ell_\lambda(h_\lambda) = 1$ but one normalization can be chosen freely.

Problem 7.19 *Show that the normalization of ℓ_λ, h_λ can be chosen so that $\ell_\lambda(h'_\lambda) = 0$.*

□

Lemma 7.5.2 *The functions α_λ and $\ln \alpha_\lambda$ are convex. Moreover,*

$$\left| \frac{d}{d\lambda} \ln \alpha_\lambda \right| \leq |\hat{f}|_\infty.$$

Proof. Note that

$$\frac{d^2}{d\lambda^2} \ln \alpha_\lambda = \frac{\alpha''_\lambda \alpha_\lambda - (\alpha'_\lambda)^2}{\alpha_\lambda^2}, \tag{7.5.15}$$

thus the convexity of $\ln \alpha_\lambda$ implies the convexity of α_λ.

In view of the above fact we can differentiate (7.5.14) obtaining

$$\mathcal{L}'_\lambda h_\lambda + \mathcal{L}_\lambda h'_\lambda = \alpha'_\lambda h_\lambda + \alpha_\lambda h'_\lambda. \tag{7.5.16}$$

Applying ℓ_λ yields

$$\frac{d\alpha_\lambda}{d\lambda} = \alpha_\lambda \ell_\lambda(\hat{f} h_\lambda)) = \alpha_\lambda \mu_\lambda(\hat{f}). \tag{7.5.17}$$

Thus $\alpha_0' = 0$. Note that, as claimed,

$$\left|\frac{d}{d\lambda}\ln\alpha_\lambda\right| \leq |\mu_\lambda(\hat{f})| \leq |\hat{f}|_\infty.$$

Differentiating again yields

$$\frac{d^2\alpha_\lambda}{d\lambda^2} = \alpha_\lambda\mu_\lambda(\hat{f})^2 + \alpha_\lambda\ell_\lambda'(\hat{f}gh_\lambda) + \alpha_\lambda\ell_\lambda(\hat{f}h_\lambda'). \tag{7.5.18}$$

On the other hand, from (7.5.16) we have

$$(\mathbb{1}\alpha_\lambda - \mathcal{L}_\lambda)h_\lambda' = \mathcal{L}_\lambda(f_\lambda h_\lambda),$$

where $f_\lambda = \hat{f} - \mu_\lambda(\hat{f})$. Since, by construction, $\Pi_\lambda h_\lambda' = \Pi_\lambda(f_\lambda h_\lambda) = 0$, the above equation can be studied in the space $\mathbb{V}_\lambda = (\mathbb{1}-\Pi_\lambda)\mathcal{C}^1$ in which $\mathbb{1}\alpha_\lambda - \mathcal{L}_\lambda$ is invertible.

Setting $\hat{\mathcal{L}}_\lambda := \alpha_\lambda^{-1}\mathcal{L}_\lambda$, we have

$$h_\lambda' = (\mathbb{1} - \hat{\mathcal{L}}_\lambda)^{-1}\hat{\mathcal{L}}_\lambda(f_\lambda h_\lambda). \tag{7.5.19}$$

Doing similar considerations on the equation $\ell_\lambda(\mathcal{L}_\lambda) = \alpha_\lambda\ell_\lambda(g)$, we obtain

$$\begin{aligned}\alpha_\lambda'' &= \alpha_\lambda\mu_\lambda(\hat{f})^2 + \alpha_\lambda\ell_\lambda(f_\lambda(\mathbb{1} - \hat{\mathcal{L}}_\lambda)^{-1}(\mathbb{1} + \hat{\mathcal{L}}_\lambda)(f_\lambda h_\lambda)) \\ &= \alpha_\lambda\mu_\lambda(\hat{f})^2 + \alpha_\lambda\sum_{n=1}^{\infty}\ell_\lambda(f_\lambda\hat{\mathcal{L}}_\lambda^n(\mathbb{1} + \hat{\mathcal{L}}_\lambda)(f_\lambda h_\lambda)) \\ &= \frac{(\alpha_\lambda')^2}{\alpha_\lambda} + \left[\mu_\lambda(f_\lambda^2) + 2\sum_{n=1}^{\infty}\ell_\lambda(f_\lambda\hat{\mathcal{L}}_\lambda^n(f_\lambda h_\lambda))\right]\alpha_\lambda.\end{aligned} \tag{7.5.20}$$

Finally, notice that

$$\ell_\lambda(f_\lambda\hat{\mathcal{L}}_\lambda^n(f_\lambda h_\lambda)) = \ell_\lambda(\hat{\mathcal{L}}_\lambda^n(f_\lambda\circ T^n f_\lambda h_\lambda)) = \mu_\lambda(f_\lambda\circ T^n f_\lambda)$$

and

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\mu_\lambda\left(\left[\sum_{k=0}^{n-1}f_\lambda\circ T^k\right]^2\right) &= \lim_{n\to\infty}\frac{1}{n}\sum_{k,j=0}^{n-1}\mu_\lambda(f_\lambda\circ T^k f_\lambda\circ T^j) \\ &= \mu_\lambda(f_\lambda^2) + \lim_{n\to\infty}\frac{2}{n}\sum_{k=1}^{n-1}(n-k)\mu_\lambda(f_\lambda\circ T^k f_\lambda) \\ &= \mu_\lambda(f_\lambda^2) + 2\sum_{k=1}^{\infty}\mu_\lambda(f_\lambda\circ T^k f_\lambda).\end{aligned} \tag{7.5.21}$$

The above two facts and equations (7.5.15), (7.5.20) yield

$$\frac{d^2}{d\lambda^2} \ln \alpha_\lambda = \lim_{n\to\infty} \frac{1}{n} \mu_\lambda \left(\left[\sum_{k=0}^{n-1} f_\lambda \circ T^k \right]^2 \right) \geq 0. \tag{7.5.22}$$

□

Note that equation (7.5.17) implies $\alpha_0' = 0$, hence $\alpha_\lambda' \geq 0$ for $\lambda \geq 0$. Since the maximum of $\lambda a - \ln \alpha_\lambda$ is taken either at $\alpha_\lambda a = \alpha_\lambda'$ or at infinity (if $a > \sup_{\lambda>0} \frac{\alpha_\lambda'}{\alpha_\lambda}$), it follows that

$$\tilde{I}(a) = \sup_{\lambda\geq 0}(\lambda a - \ln \alpha_\lambda) = \sup_{\lambda}(\lambda a - \ln \alpha_\lambda) = I(a)$$

as announced. In fact, more can be said.

Lemma 7.5.3 *Either the rate function I is strictly convex, or there exists $\beta \in \mathbb{R}, \phi \in \mathcal{C}^0$ such that $f - \beta = \phi - \phi \circ T$.*

Proof. By Problem 7.16 it suffices to prove that $\ln \alpha_\lambda$ is strictly convex. On the other hand equations (7.5.15) and (7.5.22) imply that if the second derivative of $\ln \alpha_\lambda$ is zero for some λ, then

$$\begin{aligned}
\mu_\lambda \left(\left[\sum_{k=0}^{n-1} f_\lambda \circ T^k \right]^2 \right) &= n \left[\mu_\lambda(\hat{f}^2) + 2 \sum_{k=1}^{n-1} \frac{n-k}{n} \mu_\lambda(f_\lambda \circ T^k \, f_\lambda) \right] \\
&= -2n \sum_{k=n}^{\infty} \ell_\lambda(f_\lambda \hat{\mathcal{L}}_\lambda^k (f_\lambda \, h_\lambda)) - 2 \sum_{k=1}^{n-1} k \ell_\lambda(f_\lambda \hat{\mathcal{L}}_\lambda^k (f_\lambda \, h_\lambda)) - \alpha_\lambda \mu_\lambda(\hat{f})^2 \\
&\leq C(\lambda) \left[n\rho_\lambda^n + \sum_{k=0}^{\infty} k \rho_\lambda^k \right]
\end{aligned}$$

Accordingly, the sequence $\sum_{k=0}^{n-1} f_\lambda \circ T^k$ is bounded in $L^2(\mathbb{T}^1, \mu_\lambda)$ and hence weakly compact. Let $\sum_{k=0}^{n_j-1} f_\lambda \circ T^k$ a weakly convergent subsequence,[12] that is there exists $\phi_\lambda \in L^2$ such that for each $\varphi \in L^2$ holds

$$\lim_{j\to\infty} \mu_\lambda(\varphi \sum_{k=0}^{n_j-1} f_\lambda \circ T^k) = \mu_\lambda(\varphi \phi_\lambda).$$

[12] Such a subsequence always exists [LL01].

It follows that, for each $\varphi \in \mathcal{C}^1$,

$$\begin{aligned}\mu_\lambda(\varphi[f_\lambda - \phi_\lambda + \phi_\lambda \circ T]) &= \mu_\lambda(\varphi f_\lambda) + \lim_{j\to\infty} \sum_{k=0}^{n_j-1} \mu_\lambda(\varphi f_\lambda \circ T^{k+1} - \varphi f_\lambda \circ T^k) \\ &= \lim_{j\to\infty} \mu_\lambda(\varphi f_\lambda \circ T^{n_j}) = \lim_{j\to\infty} \ell_\lambda(f_\lambda \hat{\mathcal{L}}_\lambda^{n_j}(\varphi h_\lambda)) \\ &= \mu_\lambda(\varphi)\mu_\lambda(f_\lambda) = 0.\end{aligned}$$

thus, since $\mathcal{C}^1$ is dense in L^2, it follows

$$f_\lambda = \phi_\lambda - \phi_\lambda \circ T\,, \qquad \mu_\lambda - \text{a.s.} \tag{7.5.23}$$

A function with the above property is called a *coboundary*, in this case an L^2 coboundary since we know only that $\phi_\lambda \in L^2(\mathbb{T}, \mu_\lambda)$. In fact, this it is not not enough to conclude the Lemma: we need to show, at least, that $\phi_\lambda \in \mathcal{C}^0$.

First of all notice that, since for each $\beta \in \mathbb{R}$ we have $f_\lambda = \phi_\lambda + \beta - (\phi_\lambda + \beta) \circ T$, we can assume without loss of generality $\mu_\lambda(\phi_\lambda) = 0$. But then

$$\hat{\mathcal{L}}_\lambda(f_\lambda\, h_\lambda) = \hat{\mathcal{L}}_\lambda(\phi_\lambda\, h_\lambda) - \phi_\lambda\, h_\lambda = -(\mathbb{1} - \hat{\mathcal{L}}_\lambda)\phi_\lambda\, h_\lambda.$$

Hence

$$\phi_\lambda = h_\lambda^{-1}(\mathbb{1} - \hat{\mathcal{L}}_\lambda)^{-1}\hat{\mathcal{L}}_\lambda(f_\lambda\, h_\lambda) \in \mathcal{C}^1.$$

□

Remark 7.5.4 *The above result is quite sharp. Indeed, it shows that if I is not strictly convex, then for each invariant measure ν holds $\nu(f) = \beta = \mu(f)$. So it suffices to find two invariant measures for which the average of f differs (for example the average on two periodic orbits) to infer that I is strictly convex.*

Problem 7.20 *Set $\sigma := \alpha''(0)$. Show that, for a small, $I(a) = \frac{a^2}{2\sigma} + \mathcal{O}(a^3)$. Show that if $a > |f|_\infty$, then $I(a) = +\infty$.*

The above discussion allows to conclude

$$m(A_{a,n}^+(f)) \le m(\mathcal{L}_{\lambda_-}^n h) \le Ce^{-\frac{a^2}{2\sigma^2}n + \mathcal{O}(a^3 n)}.$$

Since similar arguments hold for the set $A_{a,n}^+(-f)$, it follows that we have an exponentially small probability to observe a deviation from the average. Moreover, the expected size of a deviation is of order $n^{-\frac{1}{2}}$, to see if this is really the case we a lower bound.

7.5.2 Large deviations. Lower bound

Let $I = (\alpha, \beta)$, fix $c \in (0, \frac{\beta-\alpha}{2})$ and let us consider a $\lambda \in \mathbb{R}$ such that $\mu_\lambda(\hat{f}) \in (\alpha + c, \beta - c) = I_c$. Let $S_n = \sum_{k=0}^{n-1} \hat{f} \circ T^k$, then $\mu_\lambda(S_n) = n\mu_\lambda(\hat{f})$ and, by (7.5.21)

$$\mu_\lambda \left(\left[\sum_{k=0}^{n-1} \hat{f} \circ T^k - n\mu_\lambda(\hat{f}) \right]^2 \right) \leq C_\lambda n,$$

where C_λ depends continuously by λ. Thus, setting $A_{n,I} = \{x \in \mathbb{T}^1 : \frac{1}{n} S_n(x) \in I\}$,

$$\begin{aligned} \mu_\lambda(A^c_{n,I}) &\leq \mu_\lambda \left(\left\{ \left| \sum_{k=0}^{n-1} f_\lambda \circ T^k \right| \geq cn \right\} \right) \\ &\leq c^{-2} n^{-2} \mu_\lambda \left(\left| \sum_{k=0}^{n-1} f_\lambda \circ T^k \right|^2 \right) \leq C_\lambda c^{-2} n^{-1}. \end{aligned}$$

It follows that there exists $n_\lambda \in \mathbb{N}$ such that, for all $n \geq n_\lambda$, $\mu_\lambda(A_{n,I}) \geq \frac{1}{2}$. We can then write

$$\frac{1}{2} \leq \ell_\lambda(A_{n,I} h_\lambda) \leq C_\# e^{-(n+m)\ln \alpha_\lambda} \ell_\lambda \left(\mathcal{L}^{n+m}_\lambda (\mathbb{1}_{A_{n,I}}) \right). \tag{7.5.24}$$

To conclude we must analyse a bit the characteristic function of $A_{n,I}$. First of all, notice that if $|T^k x - T^k y| \leq \varepsilon$ for each $k \leq n$, then $|T^k x - T^k y| \leq \lambda^{-n+k}\varepsilon$ for all $k \leq n$. Accordingly, for each $z \in [x, y]$

$$\begin{aligned} |D_x T^n - D_z T^n| &\leq |D_x T^n| \cdot (e^{\sum_{k=0}^{n-1} |\ln D_{T^k x} T - \ln D_{T^k z} T|} - 1) \\ &\leq |D_x T^n| (e^{C_\# \sum_{k=0}^{n-1} \lambda^{-k}\varepsilon} - 1) \leq C_\# |D_x T^n|. \end{aligned}$$

By a similar estimate follows $|D_x T^n - D_z T^n| \geq C_\# |D_x T^n|$ as well. Moreover,

$$|S_n(x) - S_n(y)| \leq \sum_{k=0}^{n-1} |f|_{\mathcal{C}^1} C_\# \lambda^{-k} \varepsilon \leq C_\# \varepsilon.$$

We can then write $A_{n,I} \supset \cup_l J_l \supset A_{n,I_c}$ where J_l are disjoint intervals such that $|T^n J_l| \leq \varepsilon$. Choosing ε small enough it follow that the oscillation of S_n on each J_l is smaller than c. Moreover

$$\begin{aligned} \|\mathcal{L}^n \mathbb{1}_{J_l}\|_{BV} &= \sup_{|\varphi|_\infty \leq 1} \int_{J_l} \varphi' \circ T^n \leq \sup_{|\varphi|_\infty \leq 1} \int_{J_l} \frac{d}{dx} \left[(DT^n)^{-1} \varphi \circ T^n \right] + B|J_l| \\ &\leq 2 \sup_{x \in J_l} |D_x T^n|^{-1} + B|J_l| \leq C_\# |J_l|. \end{aligned}$$

We can then continue our estimate started in (7.5.24),

$$\begin{aligned}\frac{1}{2} &\leq C_{\#} e^{-(n+m)\ln\alpha_\lambda + n\lambda\beta + mC_{\#}} \sum_l \ell_\lambda \left(\mathcal{L}^{n+m}(\mathbb{1}_{J_l})\right) \\ &= C_{\#} e^{-(n+m)\ln\alpha_\lambda + n\lambda\beta + mC_{\#}} \sum_l \ell_\lambda \left(m(J_l)(1+\mathcal{O}(\rho^m))\right) \\ &\leq C_{\#} e^{-n(\ln\alpha_\lambda - \lambda\beta)} m(A_{n,I}),\end{aligned}$$

where we have chosen m large but fixed. The above computations imply that, for each $L > 0$,

$$m(A_{n,I}) \geq C_L e^{-J_L(I)n}$$

where $J_L(I) = \max_{\{\lambda \leq L \,:\, \mu_\lambda(f)\in I_c\}} \lambda a - \ln\alpha_\lambda$. Note that, if f is not a coboundary and hence $\ln\alpha_\lambda$ is strictly convex, the maximum of $\lambda\beta - \ln\alpha_\lambda$ is attained at some finite value, hence, for L large enough, $J_L(I) = \sup_{\{\lambda\in\mathbb{R}\,:\,\mu_\lambda(f)\in I_c\}} \lambda\beta - \ln\alpha_\lambda$. This implies that

$$m(A^+_{a,n}) \geq C_{\#} e^{-J(a)n}$$

where $J(a) = \sup_{\{\lambda\,:\,\mu_\lambda(f)>a\}} \lambda a - \ln\alpha_\lambda$.

The surprising fact is that the upper and lower bound are essentially the same. To see this a little argument is needed.

7.5.3 Large deviations. Conclusions

In fact, it is possible to give a variational characterization of the rate function in the spirit of general Large deviation theory [Var84, Str84, DZ98].

Lemma 7.5.5 *Let $\mathcal{M}_T$ be the set of invariant probability measures invariant with respect to T. Then*

$$I(a) = -\sup_{\{\nu\in\mathcal{M}_T\,:\,\nu(f)\geq a\}} h_\nu(T) = J(a).$$

Proof. By section 7.4.2 we have that, for each $\nu \in \mathcal{M}_T$, $\ln\alpha_\lambda = \sup_{\nu\in\mathcal{M}_T}\{h_\nu(T) + \lambda\nu(f)\} = h_{\mu_\lambda}(T) + \lambda\mu_\lambda(f)$. Thus for each $\nu \in \mathcal{M}_T$ such that $\nu(f) \geq a$, we can write

$$I(a) \leq \max_{\lambda\geq 0}\{\lambda(a-\nu(f)) - h_\nu(T)\} = -h_\nu(T).$$

On the other and

$$I(a) = \sup_{\lambda\geq 0}\{\lambda(a-\mu_\lambda(f)) - h_{\mu_\lambda}(T)\}.$$

If $a > \sup \mu_\lambda(f)$, then $I(a) = +\infty$, otherwise let λ_* be such that $\mu_{\lambda_*}(f) = a$,[13] then

$$I(a) \geq -h_{\mu_{\lambda_*}}(T) \geq - \sup_{\{\nu \in \mathcal{M}_T \,:\, \nu(f) \geq a\}} h_\nu(T).$$

Finally, since μ_λ and h_{μ_λ} depend smoothly from λ,

$$J(a) = \sup_{\{\lambda \,:\, \mu_\lambda(f) > a\}} \lambda a - \lambda \mu_\lambda(f) - h_{\mu_\lambda}(T) = I(a).$$

□

7.5.4 The Central Limit Theorem

We can now address the second question we have posed. From the above discussion is clear that we must chose $c_n = \sqrt{n}$.

Let $f \in BV$ and set $\hat{f} := f - \mu(f)$, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \hat{f} \circ T^k(x) = 0 \quad m - \text{a.e.}$$

Let us set $\Psi_n := \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} \hat{f} \circ T^k$. We can consider Ψ_n a random variable with distribution $F_n(t) := \mu(\{x \,:\, \Psi_n(x) \leq t\})$. It is well known that, for each continuous function g holds[14]

$$\mu(g(\Psi_n)) = \int_{\mathbb{R}} g(t) dF_n(t)$$

where the integral is a Riemann-Stieltjes integral. It is thus clear that if we can control the distribution F_n, we have a very sharp understanding of the probability to have small deviations (of order $\sqrt{n}$) from the limit. From the work in the previous section it follows that there exists $\delta > 0$ such that, for each $|\lambda| \leq \delta\sqrt{n}$,

$$\begin{aligned} \varphi_n(\lambda) := \mu(e^{i\lambda\Psi_n}) = \mu(\mathcal{L}^n_{i\lambda/\sqrt{n}} h) &= (1 - \frac{\sigma^2\lambda^2}{2n} + \mathcal{O}(\lambda^3 n^{-\frac{3}{2}} + \rho^n)\|f\|_{BV})^n \\ &= e^{-\frac{\sigma^2\lambda^2}{2}}(1 + \mathcal{O}(\lambda^3 n^{-\frac{1}{2}} + n\rho^n)\|f\|_{BV}). \end{aligned} \tag{7.5.25}$$

[13] Actually one must show that the sup is a max.

[14] If $g \in \mathcal{C}^1_0$, then

$$\int_{\mathbb{R}} g dF_n = -\int_{\mathbb{R}} F_n(t) g'(t) dt = -\int_{\mathbb{R}} dt \int_{\mathbb{T}^1} dx \chi_{\{z \,:\, \Psi_n(z) \leq t\}}(x) g'(t).$$

Applying Fubini yields

$$\int_{\mathbb{R}} g dF_n = -\int_{\mathbb{T}^1} dx \int_{\mathbb{R}} dt \chi_{\{z \,:\, \Psi_n(z) \leq t\}}(x) g'(t) = -\int_{\mathbb{T}^1} dx \int_{\Psi_n(x)}^{\infty} g'(t) dt = \int_{\mathbb{T}^1} dx g(\Psi_n(x)).$$

The above quantity is called *characteristic function* of the random variable and determines the distribution (at continuity points) via the formula

$$F_n(b) - F_n(a) = \lim_{\Lambda\to\infty} \frac{1}{2\pi} \int_{-\Lambda}^{\Lambda} \frac{e^{-ia\lambda} - e^{-ib\lambda}}{i\lambda} \varphi_n(\lambda) d\lambda,$$

as can be seen in any basic book of probability theory.[15]

Formula (7.5.25) means in particular that

$$\lim_{n\to\infty} m(e^{\lambda\Psi_n}) = e^{-\frac{\sigma^2\lambda^2}{2}} =: \varphi(\lambda).$$

What can we infer from the above facts? First of all a simple computation shows that

$$g(t) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-it\lambda} \varphi(\lambda) d\lambda = \frac{1}{\sqrt{\pi}\sigma} e^{-\frac{t^2}{2\sigma^2}}$$

a random variable with such a density is called a Gaussian random variable with zero average and variance σ. Accordingly, formula (7.5.25) can be interpreted by saying that there exists a Gaussian random variable G such that

$$\frac{1}{n} \sum_{k=0}^{n-1} \hat{f} \circ T^k \sim \frac{1}{\sqrt{n}} G(1 + \mathcal{O}(n^{-\frac{1}{2}}))$$

in distribution. But what does this means concretely. Actual estimates are made difficult by the fact that the distribution under study not necessarily have a density, thus we are Fourier transforming function that behave quite badly at infinity. To overcome such a problem we can smoothen the quantities involved.

Let $j \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}_+)$ such that $\int_{\mathbb{R}} j(t)dt = 1$, $j(t) = j(-t)$, and $j(t) = 0$ for all $|t| > 1$, for each $\varepsilon > 0$ defined then $j_\varepsilon(t) := \varepsilon^{-1} j(\varepsilon^{-1} t)$ and

$$F_{n,\varepsilon}(t) := \int_{\mathbb{R}} j_\varepsilon(t - s) F_n(s) ds. \tag{7.5.26}$$

A simple computation shows that, for each $a, b \in \mathbb{R}$, holds

$$F_n(b + \varepsilon) - F_n(a - \varepsilon) \geq F_{n,\varepsilon}(b) - F_{n,\varepsilon}(a) \geq F_n(b - \varepsilon) - F_n(a + \varepsilon)$$

that is: if the measurements have a precision worst than 2ε, then $F_{n,\varepsilon}$ is as good as F_n to describe the resulting statistics. On the other hand calling $\varphi_{n,\varepsilon}$

[15] In the case when there exists a density, that is an L^1 function f_n such that $F_n(b) - F_n(a) = \int_a^b f_n(t)dt$, then the formula above becomes simply

$$f_n(t) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-it\lambda} \varphi_n(\lambda) d\lambda,$$

and follows trivially by the inversion of the Fourier transform.

the characteristic function associated to $F_{n,\varepsilon}$, holds $\varphi_{n,\varepsilon}(\lambda) = \varphi_n(\lambda)\hat{j}(\varepsilon\lambda)$, where $\hat{j}$ is the Fourier transform of j. Since now $F_{n,\varepsilon}$ is the law of a smooth random variable it has a density $f_{n,\varepsilon}$ and

$$f_{n,\varepsilon}(t) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\lambda t}\varphi_n(\lambda)\hat{j}(\varepsilon\lambda)d\lambda$$

since j is smooth it follows that there exists $C > 0$ such that $|\hat{j}(\lambda)| \leq C(1+\lambda^2)^{-2}$. We can finally use formula (7.5.25) to obtain a quantitative estimate

$$\begin{aligned} f_{n,\varepsilon}(t) &= \frac{1}{2\pi}\int_{-\varepsilon\sqrt{n}}^{\varepsilon\sqrt{n}} e^{-i\lambda t}\varphi_n(\lambda)\hat{j}(\varepsilon\lambda)d\lambda + \mathcal{O}(\varepsilon^{-5}n^{-\frac{3}{2}}) \\ &= \frac{1}{2\pi}\int_{-\varepsilon\sqrt{n}}^{\varepsilon\sqrt{n}} e^{-i\lambda t}\varphi(\lambda)\hat{j}(\varepsilon\lambda)d\lambda + \mathcal{O}(\varepsilon^{-5}n^{-\frac{3}{2}} + n^{-\frac{1}{2}}) \\ &= g(t) + \mathcal{O}(\varepsilon + \varepsilon^{-5}n^{-\frac{3}{2}} + n^{-\frac{1}{2}}) = g(t) + \mathcal{O}(n^{-\frac{1}{2}}) \end{aligned}$$

provided we choose $n^{-\frac{1}{2}} \geq \varepsilon \geq n^{-5}$. Which, as announced, means that, if the precision of the instrument is compatible with the statistics, the typical fluctuations in measurements are of order $\frac{1}{\sqrt{n}}$ and Gaussian. This is well known by sperimentalists who routinely assume that the result of a measurement is distributed according to a Gaussian.[16]

7.6 Perturbation theory

To answer the questions posed at the beginning we need some perturbation theorems. Few such results are available (e.g., see [Kif88], [BY93] or [Bal00a] for a review), here we will follow mainly the theory developed in [KL99, GL06] adapted to the special cases at hand.

For simplicity let us work directly with the densities and in the case $d = 1$. Then $\mathcal{L}$ is the transfer operator for the densities. We will start by considering an abstract family of operators $\mathcal{L}_\varepsilon$ satisfying the following properties.

Condition 1 *Consider a family of operators $\mathcal{L}_\varepsilon$ with the following properties*

1. *A uniform Lasota-Yorke inequality:*

$$\|\mathcal{L}_\varepsilon^n h\|_{BV} \leq A\lambda^{-n}\|h\|_{BV} + B|h|_{L^1}, \quad |\mathcal{L}_\varepsilon^n h|_{L^1} \leq C|h|_{L^1} \ ;$$

2. $\int \mathcal{L}h(x)dx = \int h(x)dx$;

[16] Note however that our proof holds in a very special case that has little to do with a real experimental setting. To prove the analogous statement for a realistic experiment is a completely different ball game.

3. *For $L : BV \to BV$ define the norm*

$$|||L||| := \sup_{\|h\|_{BV} \leq 1} |Lf|_{L^1},$$

that is the norm of L as an operator from $BV \to L^1$. Then we require that there exists $D > 0$ such that

$$|||\mathcal{L} - \mathcal{L}_\varepsilon||| \leq D\varepsilon.$$

Condition 1-(3) specifies in which sense the family $\mathcal{L}_\varepsilon$ can be considered an approximation of the unperturbed operator $\mathcal{L}$. Notice that the condition is rather weak, in particular the distance between $\mathcal{L}_\varepsilon$ and $\mathcal{L}$ as operators on BV can be always larger than 1. Such a notion of closeness is completely inadequate to apply standard perturbation theory, to get some perturbations results it is then necessary to drastically restrict the type of perturbations allowed, this is done by Conditions 1-(1,2) which state that all the approximating operators enjoys properties very similar to the limiting one.[17]

To state a precise result consider, for each operator L, the set

$$V_{\delta,r}(L) := \{z \in \mathbb{C} \mid |z| \leq r \text{ or } \operatorname{dist}(z, \sigma(L)) \leq \delta\}.$$

Since the complement of $V_{\delta,r}(L)$ belongs to the resolvent of L it follows that

$$H_{\delta,r}(L) := \sup\left\{\|(z - L)^{-1}\|_{BV} \mid z \in \mathbb{C} \setminus V_{\delta,r}(L)\right\} < \infty.$$

By $R(z)$ and $R_\varepsilon(z)$ we will mean respectively $(z - \mathcal{L})^{-1}$ and $(z - \mathcal{L}_\varepsilon)^{-1}$.

Theorem 7.6.1 ([KL99]) *Consider a family of operators $\mathcal{L}_\varepsilon : BV \to BV$ satisfying Conditions 1. Let $H_{\delta,r} := H_{\delta,r}(\mathcal{L})$; $V_{\delta,r} := V_{\delta,r}(\mathcal{L})$, $r > \lambda^{-1}$, $\delta > 0$, then, if $\varepsilon \leq \varepsilon_1(\mathcal{L}, r, \delta)$, $\sigma(\mathcal{L}_\varepsilon) \subset V_{\delta,r}(\mathcal{L})$. In addition, if $\varepsilon \leq \varepsilon_0(\mathcal{L}, r, \delta)$, there exists $a > 0$ such that, for each $z \notin V_{\delta,r}$, holds true*

$$|||R(z) - R_\varepsilon(z)||| \leq C\varepsilon^a.$$

PROOF.[18] To start with we collect some trivial, but very useful algebraic identities.

[17] Actually only Condition 1-(1) is needed in the following. Condition 1-(2) simply implies that the eigenvalue one is common to all the operators. If 1-(2) is not assumed, then the operator $\mathcal{L}_\varepsilon$ will always have one eigenvalue close to one, but the spectral radius could vary slightly, see [LMD03] for such a situation.

[18] This proof is simpler than the one in [KL99], yet it gives worst bounds, although sufficient for the present purposes.

For each operator $L : BV \to BV$ and $n \in \mathbb{Z}$ holds

$$\frac{1}{z}\sum_{i=0}^{n-1}(z^{-1}L)^i(z-L)+(z^{-1}L)^n = \mathbb{1} \tag{7.6.27}$$

$$R(z)(z-\mathcal{L}_\varepsilon)+\frac{1}{z}\sum_{i=0}^{n-1}(z^{-1}\mathcal{L})^i(\mathcal{L}_\varepsilon-\mathcal{L})+R(z)(z^{-1}\mathcal{L})^n(\mathcal{L}_\varepsilon-\mathcal{L}) = \mathbb{1} \tag{7.6.28}$$

$$(z-\mathcal{L}_\varepsilon)\left[G_{n,\varepsilon}+(z^{-1}\mathcal{L}_\varepsilon)^nR(z)\right] = \mathbb{1}-(z^{-1}\mathcal{L}_\varepsilon)^n(\mathcal{L}_\varepsilon-\mathcal{L})R(z) \tag{7.6.29}$$

$$\left[G_{n,\varepsilon}+(z^{-1}\mathcal{L}_\varepsilon)^nR(z)\right](z-\mathcal{L}_\varepsilon) = \mathbb{1}-(z^{-1}\mathcal{L}_\varepsilon)^nR(z)(\mathcal{L}_\varepsilon-\mathcal{L}), \tag{7.6.30}$$

where we have set $G_{n,\varepsilon} := \frac{1}{z}\sum_{i=0}^{n-1}(z^{-1}\mathcal{L}_\varepsilon)^i$.

Let us start applying the above formulae. For each $h \in BV$ and $z \notin V_{r,\delta}$ holds

$$\begin{aligned}\|(z^{-1}\mathcal{L}_\varepsilon)^n(\mathcal{L}_\varepsilon-\mathcal{L})R(z)h\|_{BV} &\le (r\lambda)^{-n}A\|(\mathcal{L}_\varepsilon-\mathcal{L})R(z)h\|_{BV}+\frac{B}{r^n}|(\mathcal{L}_\varepsilon-\mathcal{L})R(z)h|_{L^1}\\ &\le [(r\lambda)^{-n}A2C_1+Br^{-n}D\varepsilon]H_{r,\delta}\|h\|_{BV} < \|h\|_{BV}\end{aligned}$$

Thus $\|(z^{-1}\mathcal{L}_\varepsilon)^n(\mathcal{L}_\varepsilon-\mathcal{L})R(z)\|_{BV} < 1$ and the operator on the right hand side of (7.6.29) can be inverted by the usual Neumann series. Accordingly, $(z-\mathcal{L}_\varepsilon)$ has a well defined right inverse. Analogously,

$$\|(z^{-1}\mathcal{L}_\varepsilon)^nR(z)(\mathcal{L}_\varepsilon-\mathcal{L})h\|_{BV} \le (r\lambda)^{-n}A\|R(z)(\mathcal{L}_\varepsilon-\mathcal{L})h\|_{BV}+Br^{-n}|R(z)(\mathcal{L}_\varepsilon-\mathcal{L})h|_{L^1}.$$

This time to continue we need some informations on the L^1 norm of the resolvent. Let $g \in BV$, then equation (7.6.27) yields

$$\begin{aligned}|R(z)g|_{L^1} &\le \frac{1}{r}\sum_{i=0}^{n-1}|(z^{-1}\mathcal{L})^ig|_{L^1}+\|R(z)(z^{-1}\mathcal{L})^ng\|_{BV}\\ &\le \frac{1}{r^n(1-r)}|g|_{L^1}+H_{\delta,r}A(r\lambda)^{-n}\|g\|_{BV}+H_{\delta,r}Br^{-n}|g|_{L^1}\\ &\le r^{-n}(H_{\delta,r}B+(1-r)^{-1})|g|_{L^1}+H_{\delta,r}A(r\lambda)^{-n}\|g\|_{BV}\end{aligned}$$

Substituting, we have

$$\begin{aligned}&\|(z^{-1}\mathcal{L}_\varepsilon)^nR(z)(\mathcal{L}_\varepsilon-\mathcal{L})h\|_{BV} \le \{(r\lambda)^{-n}AH_{\delta,r}2C_1[1+Br^{-n}]\\ &+Br^{-2n}[H_{\delta,r}B+(1-r)^{-1}]D\varepsilon\}\|h\|_{BV} < 1,\end{aligned}$$

again, provided ε is small enough and choosing n appropriately. Hence the operator on the right hand side of (7.6.30) can be inverted, thereby providing a left inverse for $(z-\mathcal{L}_\varepsilon)$. This implies that z does not belong to the spectrum of $\mathcal{L}_\varepsilon$.

To investigate the second statement note that (7.6.28) implies

$$R(z) - R_\varepsilon(z) = \frac{1}{z}\sum_{i=0}^{n-1}(z^{-1}\mathcal{L})^i(\mathcal{L}_\varepsilon - \mathcal{L})R_\varepsilon(z) - R(z)(z^{-1}\mathcal{L})^n(\mathcal{L}_\varepsilon - \mathcal{L})R_\varepsilon(z).$$

Accordingly, for each $\varphi \in BV$ holds

$$|R(z)\varphi - R_\varepsilon(z)\varphi|_{L^1} \leq \{r^{-n}(1-r)^{-1}\varepsilon + H_{\delta,r}(\lambda r)^{-n}2AC_1 + H_{\delta,r}B\varepsilon\}\|R_\varepsilon(z)\varphi\|_{BV}.$$

□

7.6.1 Deterministic stability

The $\mathcal{L}_\varepsilon$ are Perron-Frobenius (Transfer) operators of maps T_ε which are $\mathcal{C}^1$–close to T, that is $d_{\mathcal{C}^1}(T_\varepsilon, T) = \varepsilon$ and such that $d_{\mathcal{C}^2}(T_\varepsilon, T) \leq M$, for some fixed $M > 0$. In this case the uniform Lasota-Yorke inequality is trivial. On the other hand, for all $\varphi \in \mathcal{C}^1$ holds

$$\int (\mathcal{L}_\varepsilon f - \mathcal{L} f)\varphi = \int f(\varphi \circ T_\varepsilon - \varphi \circ T).$$

Now let $\Phi(x) := (D_x T)^{-1} \int_{Tx}^{T_\varepsilon x} \varphi(z)dz$, since

$$\Phi'(x) = -(D_x T)^{-1} D_x^2 T \Phi(x) + D_x T_\varepsilon (D_x T)^{-1}\varphi(T_\varepsilon x) - \varphi(Tx)$$

follows

$$\int (\mathcal{L}_\varepsilon f - \mathcal{L} f)\varphi = \int f\Phi' + \int f(x)[(D_x T)^{-1} D_x^2 T\Phi(x) + (1 - D_x T_\varepsilon (D_x T)^{-1})\varphi(T_\varepsilon x)].$$

Given that $|\Phi|_\infty \leq \lambda^{-1}\varepsilon|\varphi|_\infty$ and $|1 - D_x T_\varepsilon (D_x T)^{-1}|_\infty \leq \lambda^{-1}\varepsilon$, we have

$$\int (\mathcal{L}_\varepsilon f - \mathcal{L} f)\varphi \leq \|f\|_{BV}\lambda^{-1}|\varphi|_\infty\varepsilon + |f|_{L^1}\lambda^{-1}(B+1)\varepsilon|\varphi|_\infty \leq D\|f\|_{BV}\varepsilon|\varphi|_\infty.$$

By Lebesgue dominate convergence theorem we obtain the above inequality for each $\varphi \in L^\infty$, and taking the sup on such φ yields the wanted inequality.

$$|\mathcal{L}_\varepsilon f - \mathcal{L} f|_{L^1} \leq D\|f\|_{BV}\varepsilon.$$

We have thus seen that all the requirements in Condition 1 are satisfied. See [Kel82] for a more general setting including piecewise smooth maps.

7.6.2 Stochastic stability

Next consider a set of maps $\{T_\omega\}$ depending on a parameter $\omega \in \Omega$. In addition assume that Ω is a probability space and consider a measure P on Ω. Consider the process $x_n = T_{\omega_n} \circ \cdots \circ T_{\omega_1} x_0$ where the ω are i.i.d. random variables distributed accordingly to P and let E_μ be the expectation of such process when x_0 is distributed according to μ. Then, calling $\mathcal{L}_\omega$ the transfer operator associated to T_ω, we have

$$E(f(x_{n+1}) \mid x_n) = \mathcal{L}_P f(x_n) := \int_\Omega \mathcal{L}_\omega f(x_n) P(d\omega).$$

Then if

$$|\mathcal{L}_\omega h|_{BV} \leq \lambda_\omega^{-1} |h|_{BV} + B_\omega |h|_{L^1}$$

integrating yields

$$|\mathcal{L}_P h|_{BV} \leq E(\lambda_\omega^{-1}) |h|_{BV} + E(B_\omega) |h|_{L^1}$$

And the operator $\mathcal{L}_P$ satisfy a Lasota-Yorke inequality provided that $E(\lambda^{-1}) < 1$ and $E(B) < \infty$.

In addition, if for some map T and associated transfer operator $\mathcal{L}$,

$$E(|\mathcal{L}_\omega h - \mathcal{L} h|) \leq \varepsilon |h|_{BV}$$

then we can apply perturbation theory and obtain stochastic stability.

7.6.3 Computability

If we want to compute the invariant measure and the rate of decay of correlations, we can use the operator P_t defined in (7.3.6) and define $\mathcal{L}_{t,m} = P_t \mathcal{L}^m$. By a direct computation it follows

$$|\mathcal{L}_{t,m} h|_{BV} \leq 4^d \sigma^m |h|_{BV} + B |h|_{L^1}.$$

We can then chose the smallest m so that $4^d \sigma^m = \sigma_1 < 1$. Moreover, we also saw that

$$|\mathcal{L}_{t,m} h - \mathcal{L} h| \leq t^{-1} |h|_{BV}.$$

So we are again in the realm of our perturbation theory and we have that the finite dimensional operator $\mathcal{L}_{t,m}$ has spectrum close to the one of the transfer operator. We can then obtain all the info we want by diagonalizing a matrix.

7.6.4 Linear response

Linear response is a theory widely used by physicists. In essence it says the follow: consider a one parameter family of systems T_s and the associated (e.g.) invariant measures μ_s, then, for a given observable f one want to study the response of the system to a small change in s, and, not surprisingly, one expects $\mu_s(f) = \mu_0(f) + s\nu(f) + o(s)$. That is one expects differentiability in s. Yet differentiability is is not ensured by Theorem 7.6.1. Is it possible to ensure conditions under which linear response holds? The answer is yes (for example if holds if the maps are sufficiently smooth and the dependence on the parameter is also smooth in an appropriate sense). To prove it one need a sophistication of Theorem 7.6.1 that can be found in [GL06].

7.6.5 The hyperbolic case

One can wonder is the previous approach can be applied to uniformly hyperbolic systems and partially hyperbolic system. The answer is yes although the work in this direction is still in progress and the price to pay is the need to consider rather unusual functional spaces (space of anysotropic distributions). Just to give a vague idea let us look at a totally trivial example: toral automorphisms.

Then one can consider the norms:

$$\|f\|_{p,q} := \sum_{k\in\mathbb{Z}^{2d}\setminus\{0\}} |f_k| \frac{|k|^p}{1+|\langle v^s, k\rangle|^{p+q}} + |f_0|,$$

where f_k are the Fourier coefficients of f and v^s is the unit vector in the stable direction. Then

$$\begin{aligned} \|[\mathcal{L}f\|_{p,q} &\le C_1\|f\|_{p,q}, \\ \|[\mathcal{L}^n f\|_{p,q} &\le C_3\mu^n\|f\|_{p,q} + B\|f\|_{p-1,q+1}. \end{aligned} \tag{7.6.31}$$

we have thus the Lasota-Yorke inequality. Moreover on can easily check the relative compactness of $\{\|f\|_{p,q} \le 1\}$ with respect to the topology induced by the norm $\|\cdot\|_{p-1,q+1}$, hence our previous theory applies almost verbatim.

To have a more precise idea of what can be done, see [GL06, BT07].

Hints to solving the Problems

7.19 Let ℓ_λ, h_λ be analytic. Let us define $z_\lambda = e^{-\int_0^\lambda \ell_\xi(h'_\xi)d\xi}$, define $\hat{h}_\lambda = z_\lambda h_\lambda$ and $\hat{\ell}_\lambda = z_\lambda^{-1}\ell_\lambda$ and check that they are normalized as required.

Notes

Large deviations are taken from Lai-Sang article and Keller book.

The stochastic stability is reasonably well understood (Cowienson) but what about the smooth dependence from a parameter (linear response)? Counterexamples in $d = 1$ but unknown in higher dimensions. The uniformly hyperbolic case is well understood but not much is know on how to apply the present ideas to the partially hyperbolic case and to the case of systems with discontinuities, although a concentrated effort is taking place to extend the theory in such directions.

Chapter 8

Uniformly hyperbolic systems

The concept of ergodicity is a very important one in dynamical systems, yet it turns out to be surprisingly difficult to establish if a system is or not ergodic and very few examples have been fully analyzed. Nonetheless, in this chapter we will see that a very simple idea introduced by Hopf [Hop39, Hop40] allows to discuss the ergodicity in some special cases. The relevance of Hopf's idea is that, properly generalized, it allows to prove ergodicity in a vast class of systems. Much in the following chapters will deal with such a generalization.

8.1 A Basic Example

To explain the Hopf approach we will study a very simple case: a slight generalization of Arnold's cat, see Examples 6.2.1. Let $T : \mathbb{T}^2 \to \mathbb{T}^2$ (here by $\mathbb{T}^2$ we mean $\mathbb{R}^2 \mod 1$) be defined by

$$T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & a \\ a & 1+a^2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mod 1 \tag{8.1.1}$$

It is obvious that if $a \in \mathbb{Z}$, then T is well defined and it is a linear automorphism of $\mathbb{T}^2$. Moreover, for all $x \in \mathbb{T}^2$

$$D_x T = \begin{pmatrix} 1 & a \\ a & 1+a^2 \end{pmatrix} \equiv L.$$

Since $\det L = 1$, Lebesgue measure is preserved. It is immediate to see that

there exists $\lambda > 1$; v_+, $v_- \in \mathbb{R}^2$:

$$\begin{aligned} Lv_+ &= \lambda v_+ \\ Lv_- &= \lambda^{-1} v_-. \end{aligned}$$

We will call v_+ the unstable eigenvector (direction) and v_- the stable eigenvector (direction). Remark that, since $L^* = L$, $\langle v_+, v_- \rangle = 0$.

The dynamical system just described is a basic model of hyperbolic systems (see next chapter) and will appear in various disguises in this book.

Proposition 8.1.1 *The Arnold cat is ergodic.*

Sections 8.1.1 and 8.2.1 contain two different proofs of the above proposition.

8.1.1 An algebraic proof

A first idea to studying the ergodic properties of this system is to imitate what we have done for the Rotations (Examples 6.6.1) and the Dilations (Examples 6.8.1): use Fourier series. Let us see how such an approach would work.

Let $f,\ g \in C^{(m)}(\mathbb{T}^2)$, then[1]

$$f \circ T^n(x) = \sum_{k \in \mathbb{Z}^2} e^{2\pi i \langle k,\, L^n x \rangle} f_k = \sum_{k \in \mathbb{Z}^2} e^{2\pi i \langle k,\, x \rangle} f_{L^{-n}k},$$

so

$$\begin{aligned} \int_{\mathbb{T}^2} f \circ T^{2n} g = \int_{\mathbb{T}^2} f \circ T^n g \circ T^{-n} &= \sum_{k \in \mathbb{Z}^2} f_{L^{-n}k} g_{L^n k} \\ &= f_0 g_0 + \sum_{k \in \mathbb{Z}^2 \setminus \{0\}} f_{L^{-n}k} g_{L^n k}. \end{aligned}$$

It is well known [RS80] that $f \in C^{(m)}(\mathbb{T}^2)$ implies[2]

$$|f_k| \leq \frac{\|f^{(m)}\|_1}{\|k\|^m} \text{ for } k \neq 0$$

hence

$$\left| \sum_{k \in \mathbb{Z}^2 \setminus \{0\}} f_{L^{-n}k} g_{L^n k} \right| \leq \sum_{k \in \mathbb{Z}^2 \setminus \{0\}} \frac{\|f^{(m)}\|_1 \|g^{(m)}\|_1}{\|L^{-n}k\|^m \|L^n k\|^m}.$$

[1] Note that $e^{2\pi i \langle k, T^n x \rangle} = e^{2\pi i \langle k, L^n x \rangle}$.

[2] Here for $\|f^{(m)}\|_1$ we mean $\sup_{\substack{i+j=m \\ i,j \geq 0}} \frac{1}{(2\pi)^m} \int_{\mathbb{T}^2} |\partial_{x_1}^i \partial_{x_2}^j f| dx_1 dx_2$; and $\|k\| = \sqrt{k_1^2 + k_2^2}$.

For each $k \in \mathbb{Z}^2$ holds $\|k\|^2 = \langle k, v^+\rangle^2 + \langle k, v^-\rangle^2$ hence one of the two terms must be larger than $\|k\|^2/2$.[3] Moreover if $k \neq 0$ $\|L^n k\| \geq 1$ for each $n \in \mathbb{Z}$. Using the above facts it yields

$$\begin{aligned}\left|\sum_{k\in\mathbb{Z}^2\setminus\{0\}} f_{L^{-n}k} g_{L^n k}\right| &\leq \sum_{k\in\mathbb{Z}^2\setminus\{0\}} \frac{\|f^{(m)}\|_1 \|g^{(m)}\|_1 2^{m/2}}{\lambda^{nm}\|k\|^m} \\ &\leq \text{const.}\|f^{(m)}\|_1\|g^{(m)}\|_1\lambda^{-nm},\end{aligned}$$

where the constant does not depend on f or g and we have assumed $m \geq 3$ to insure the convergence of the series.

Accordingly, for each $f,\ g \in C^{(3)}(\mathbb{T}^2)$ we have

$$\left|\int_{\mathbb{T}^2} f\circ T^n g - \int_{\mathbb{T}^2} f \int_{\mathbb{T}^2} g\right| \leq \text{const.}\|f^{(3)}\|_1\|g^{(3)}\|_1\lambda^{-3n/2}.$$

To obtain the final result we need an approximation argument. If $f,\, g \in L^2(\mathbb{T}^2)$ we can choose $f_n,\, g_n \in C^{(3)}(\mathbb{T}^2)$ such that they converge to f and g, respectively, in L^2.

Then, for each $\varepsilon \geq 0$, choose $m \in \mathbb{N}$ such that

$$\|f - f_m\|_2 + \|g - g_m\|_2 \leq \varepsilon.$$

Accordingly,

$$\begin{aligned}\left|\int_{\mathbb{T}^2} f\circ T^n g - \int_{\mathbb{T}^2} f \int_{\mathbb{T}^2} g\right| \leq &\left|\int_{\mathbb{T}^2} f_m\circ T^n g_m - \int_{\mathbb{T}^2} f_m \int_{\mathbb{T}^2} g_m\right| \\ &+ 2\|f - f_m\|_2\|g\|_2 + 2\|f_m\|_2\|g - g_m\|_2 \\ \leq &2(\|g\|_2 + \|f\|_2)\varepsilon + \varepsilon,\end{aligned}$$

where we have chosen n large enough depending on m and ε. We have just proven mixing.

The above result is certainly rather satisfactory: non only it proves the mixing–hence the ergodicity–of the map but gives an explicit estimate on the rate of decay and shows how such a rate depends on the regularity of the functions.[4] Therefore, an eventual critique can not concern the type of result but only the method; indeed the method does have a shortcoming.

The use of Fourier series is strictly related to the group structure of $\mathbb{T}^2$ and the linearity of the map. Clearly, in more general systems, where both

[3] Here we have normalized the eigenvalues so that $\|v^\pm\| = 1$.

[4] In fact, the obtained estimate it is not optimal: using the Diofantine properties of the stable and unstable directions a better estimate can be obtained (see Problem 2.1).

such properties may fail, such a technique has no hope whatsoever of being applied.[5] In some sense, much of the theory of hyperbolic systems may be viewed as an attempt to find an alternative proof of the above facts. Such a proof must be *dynamical*, meaning that it must use properties of the dynamics and as little as possible of the structure of the space.

The best way to gain a real feeling of what is meant by *dynamical* is to see such a type of argument in action.

8.2 An Idea by Hopf

The following argument, due to Hopf [Hop39, Hop40] is exactly such a dynamical proof of ergodicity. Let $f : \mathbb{T}^2 \to \mathbb{R}$ be a continuous function. We want to prove that for almost every $x \in \mathbb{T}^2$ the time averages converge as $n \to +\infty$ to the average value of f, i.e., $\int_{\mathbb{T}^2} f d\mu$. Once this is established one can obtain the same property for all integrable functions by an approximation argument, this proves ergodicity due to the characterization provided by Theorem 6.7.5 (see also Problem 1.6.31). From Birkhoff Ergodic Theorem (BET) we know that the time averages converge almost everywhere to a function $f^+ \in L^1(\mathbb{T}^2, \mu)$ which is invariant on the orbits of T, i.e., $f^+ \circ T = f^+$, and has the same average value as f, i.e., $\int f^+ d\mu = \int f d\mu$. Further, applying BET to f and T^{-1} we obtain that the time averages in the past

$$\frac{f(x) + f(T^{-1}x) + \cdots + f(T^{-n+1}x)}{n}$$

converge almost everywhere as $n \to +\infty$ to $f^- \in L^1(\mathbb{T}^2, \mu)$, $f^- \circ T = f^-$ and $\int f^- d\mu = \int f d\mu$.

The next Lemma is part of the usual magic of the ergodic theory.

Lemma 8.2.1 *The functions f^+ and f^- coincide almost everywhere.*

Proof. Let

$$\mathcal{A}_+ = \{x \in \mathbb{T}^2 \mid f_+(x) > f_-(x)\};$$

by definition $\mathcal{A}_+$ is an invariant set, hence

$$\int_{\mathcal{A}_+} [f_+(x) - f_-(x)]\, d\mu(x) = \int_{\mathcal{A}_+} f(x) d\mu(x) - \int_{\mathcal{A}_+} f(x) d\mu(x) = 0$$

which implies $\mu(\mathcal{A}_+) = 0$ and $f_+ \leq f_-$ μ-almost everywhere. The same argument, this time applied to the set $\mathcal{A}_- = \{x \in \mathbb{T}^2 \mid f_-(x) > f_+(x)\}$, implies the converse inequality. □

[5] In fact, there are very few cases in which this type of idea has produced relevant results, notably the case of geodesic flows on surfaces of constant negative curvature.

8.2.1 A dynamical proof

For $x \in \mathbb{T}^2$ let us denote by $W^u(x)$ ($W^s(x)$) the line in $\mathbb{T}^2$ passing through x and having the direction of the unstable eigenvector (the stable eigenvector), i.e., the eigenvector with eigenvalue λ (λ^{-1}). We call $W^u(x)$ ($W^s(x)$) the unstable (stable) leaf (or manifold) of x. The leaves of x have the following property. If $y \in W^u(x)$ ($y \in W^s(x)$) then the distance (computed along the leaf)

$$d(T^n y,\, T^n x) = \lambda^{-|n|} d(y,\, x) \to 0 \quad \text{as} \quad n \to -\infty(+\infty).$$

Hence for $y, z \in W^{u(s)}(x)$

$$|f(T^n y) - f(T^n z)| \to 0 \quad \text{as} \quad n \to -\infty(+\infty).$$

It follows that for $y, z \in W^{u)}(x)$ either $f^-(y)$ and $f^-(z)$ are both defined and equal or they are both undefined; the same can be said for $f^+(y)$ and $f^+(z)$ if $y, z \in W^s(x)$.

It is interesting to notice that $W^u(x)$ is an infinitely long line in the direction v_+ that fills densely the torus (see Problem 2.6). This implies that the collection (foliation) $\{W^u(x)\}_{x\in\mathbb{T}^2}$ of this global manifolds has a quite complex structure (see Problem 2.7). For this reason it turns out to be much more convenient to deal only with *local manifolds*.

A local manifold of size δ is simply a piece of $W^u(x)$ of size δ centered at x. In the following by $W^u(x)$ and $W^s(x)$ we will always mean local manifolds (lines) of some length. The exact length is, most of the times, irrelevant an often will not be specified (in the following it will be frequently chosen so that the lines do not wrap around the torus more than once).

Up to now we have seen that f^+ is constant along a.e. stable lines while f^- is constant along a.e. unstable line, since they are equal a.e. it seems obvious that they must be equal and constant. Yet, in the last sentence there are a lot of almost everywhere and, being measure theory a rather subtle subject, it is better to spell out the reasoning in full detail.[6]

Let us choose any point $x \in \mathbb{T}^2$ and prove that there is a neighborhood of x in which f^+ is a.e. constant. Since x is arbitrary this implies that f^+ is a.e. constant.[7] Chose a square Q_δ of size $2\delta < \frac{1}{4}$ centered at x with sides parallel to v_+ and v_- respectively. Let $\phi : [-\delta, \delta]^2 \to Q_\delta$ be defined by $\phi(\alpha, \beta) = x + \alpha v_+ + \beta v_-$, where we have chosen $\|v_\pm\| = 1$. It is then convenient to transport the problem in $[-\delta, \delta]^2$ by ϕ: doing so the Lebesgue measure is sent in the Lebesgue measure and that $f^+ \circ \phi$ is a.e. constant in the vertical direction (α constant), while $f^- \circ \phi$ is a.e. constant in the horizontal

[6] We have already seen in Examples 6.6.1–Rotations that these type of arguments must employ measure theory in a non trivial way.

[7] Please, note this apparently naïve idea to look at the problem first locally and then globally, we will see much more of it in the following.

direction. This corresponds simply to a change of variables and from now on we will confuse Q_δ and $[-\delta,\delta]^2$ since this does not create any ambiguity.

There are three full measure sets to consider:
$\widetilde{\mathcal{B}}_+ = \{\xi \in Q_\delta \mid f^+(\xi) \text{ is defined}\}$; $\widetilde{\mathcal{B}}_- = \{\xi \in Q_\delta \mid f^-(\xi) \text{ is defined}\}$ and $G = \{\xi \in \widetilde{\mathcal{B}}_+ \cap \widetilde{\mathcal{B}}_- \mid f^+(\xi) = f^-(\xi)\}$.

Let us call $W^s_\alpha := \{(a,b) \in Q_\delta \mid a = \alpha\}$ the segment in Q_δ parallel to the stable direction passing through the point $(\alpha, 0)$, and $W^u_\beta := \{(a,b) \in Q_\delta \mid b = \beta\}$ the segment in Q_δ parallel to the unstable direction passing through the point $(0,\beta)$. The previous discussion proves that there exist $\mathcal{B}_\pm \in [-\delta,\delta]$ such that $\widetilde{\mathcal{B}}_+ = \cup_{\alpha\in\mathcal{B}_+} W^s_\alpha$ and $\widetilde{\mathcal{B}}_- = \cup_{\beta\in\mathcal{B}_-} W^u_\beta$.

Since m is the product of two one dimensional Lebesgue measures[8] Fubini theorem [Roy88] implies that $\mathcal{B}_\pm$ are measurable sets of full measure. Again by Fubini Theorem, it follows

$$4\delta^2 = m(Q_\delta) = m(\widetilde{\mathcal{B}}_+ \cap G) = \int_{\mathcal{B}_+} d\alpha \int_{-\delta}^{\delta} d\beta \chi_{W^s_\alpha \cap G}(\alpha,\beta).$$

This implies immediately that there exists a set $I \subset \mathcal{B}_+$, of full measure, such that, for each $\alpha \in I$ the set $J_\alpha = \{\beta \in \mathcal{B}_- \mid (\alpha,\beta) \in G\}$ is measurable and has full measure as well; the same holds for $E = \cup_{\alpha\in I} W^s_\alpha$.

Finally, let $z, y \in E$, $z = (a,b)$ and $y = (c,d)$. If $a = c$, then $z, y \in W^s_a$ and $f^+(z) = f^+(y)$. On the other hand, if $a \neq c$ then by choosing $\beta \in J_a \cap J_c$ it follows

$$\begin{aligned} f^+(z) = f^+(W^s_a) = f^+(a,\beta) = f^-(a,\beta) \\ = f^-(W^u_\beta) = f^-(c,\beta) = f^+(c,\beta) = f^+(y). \end{aligned}$$

That is, f^+ is constant on E, hence f^+ (and f^-) is a.e. constant on Q_δ. By the arbitrariness of Q_δ follows that $f^+ = f^- =$constant a.e..

Up to now we have proved that f^+ is a.e. constant only if $f \in \mathcal{C}^{(0)}(\mathbb{T}^2)$, to prove ergodicity we need the same result for each $f \in L^1(\mathbb{T}^2)$. This can be easily obtained by an approximation argument; yet, it is probably more interesting to prove directly that all invariant sets have measure zero or one.

Let us consider a T-invariant measurable subset A. Let

$$f_n \to \chi_A \quad \text{in} \quad L^1(\mathbb{T}^2, \mu)$$

be a sequence of uniformly bounded continuous approximations to the indicator function.[9] We will use the fact that the time average is continuous with

[8] Here, to have an unambiguous notation, we should use m_n for the Lebesgue measure in $\mathbb{R}^n$, then we just said $m_2 = m_1 \times m_1$. For simplicity, I have suppressed all the subscript hoping not to confuse the reader too much.

[9] If the existence of such a sequence $\{f_n\}$ it is not obvious, consider the following: for

respect to the L^1 norm to establish that the time average of χ_A must be constant on $\mathbb{T}^2$. Indeed, if we denote by $\|\cdot\|_1$ the $L^1(\mathbb{T}^2, m)$ norm, then

$$\begin{aligned}\|f_n^+ - \chi_A^+\|_1 &= \left\| \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \left(f_n \circ T^i - \chi_A \circ T^i\right) \right\|_1 \\ &= \lim_{N\to\infty} \frac{1}{N} \left\| \sum_{i=1}^{N} \left(f_n \circ T^i - \chi_A \circ T^i\right) \right\|_1\end{aligned}$$

by the Lebesgue Dominated Convergence Theorem.

Using the invariance of the measure we obtain

$$\|f_n^+ - \chi_A^+\|_1 \leq \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \|f_n - \chi_A\|_1 = \|f_n - \chi_A\|_1.$$

Since the time averages $f_n^+ = m(f_n)$ a.e. in $\mathbb{T}^2$ and $\lim_{n\to\infty} m(f_n) = m(A)$, the Lebesgue dominated convergence theorem implies $\|m(A) - \chi_A^+\|_1 = 0$, that is $\chi_A^+ = m(A)$ a.e.. In addition, the invariance of A forces $\chi_A^+ = \chi_A$ so that either A or A^c has measure zero. In view of the arbitrariness of the invariant set A it follows that T must be ergodic.

8.2.2 What have we done?

The question remains of how and if such an argument can be extended to more general systems. The answer must lie in the possibility to generalize the main ingredients of the previous proof. Such ingredients are essentially two: a) the *existence* of two foliations on which f^+ (f^- respectively) are constant; b) some *regularity* property of such foliations.

In general the foliations will be provided by the stable and unstable manifolds (the existence of which is the content of the next chapter). A careful look at the previous proof should convince the reader that the needed regularity is a property of the type: consider two manifolds W_1^s, W_2^s and define a map $\phi : W_1^s \to W_2^s$ by $\phi(x) = W^u(x) \cap W_2^s$ (this is often called *holonomy map* or *Poincaré transformation*[10], we will use the first name), then ϕ is measurable and *absolutely continuous* that is : if $A \subset W_2^s$ has positive measure so has $\phi^{-1}A$. The absolutely continuity property of stable and unstable foliations will be the topic of chapter 7.

---

each $\varepsilon > 0$, by the regularity of the Lebesgue measure, there exists $C_\varepsilon \subset A \subset G_\varepsilon$ (C_ε closed and G_ε open) such that $m(G_\varepsilon) - m(C_\varepsilon) \leq \varepsilon$. Then Uryshon lemma implies that there exists $f_\varepsilon \in \mathcal{C}^{(0)}(\mathbb{T}^2)$ such that $f_\varepsilon(\mathbb{T}^2) \subset [0,1]$, $f_\varepsilon|_{C_\varepsilon} = 1$ and $f_\varepsilon|_{G_\varepsilon^c} = 0$. Thus $\|f_\varepsilon - \chi_A\|_1 \leq m(G_\varepsilon \backslash C_\varepsilon) \leq \varepsilon$.

[10] Note that if one could define a flow along the unstable direction–and in our case it is possible–then the above map would indeed be a Poincaré map with respect to such a flow.

Of course, the above comments are very imprecise, their only aim is to give an idea of what is coming next. In the mean time, to start building some feeling for the foliations and their properties, see Problems 2.12, 2.13 and 2.14.

8.3 About mixing

We continue our investigations with a discussion of an other dynamical proofs in which we will see the role of hyperbolicity and some basic ideas associated to it at work. The final goal will be to obtain a dynamical proof of the following.

Proposition 8.3.1 *The Arnold cat is mixing.*

We will start by proving Topological Mixing.

Definition 8.3.2 *A smooth Dynamical System is topologically mixing if for each two open sets U and V there exists an integer $n \in \mathbb{N}$ such that*

$$T^{-m}U \cap V \neq \emptyset \quad \forall m \geq n.$$

Note that the all point in the above definition is that it holds for all n large enough (see Problem 2.3).

Remark that it suffices to have the above property for any class of sets that can be used as a basis for the topology. The most convenient choice is given by the so called "rectangles." Such sets are an extremely important tool in hyperbolic theory and we have already met them several times–although I will not insist on them in the present book–here they appear in the simplest possible form.

Definition 8.3.3 *By rectangle we mean a quadrilater (i.e. a region with boundaries consisting of four segments) with sides parallel to the stable or unstable directions.*

Proposition 8.3.4 *The Arnold cat is topologically mixing.*

Proof. Let us consider two rectangles A and B. A first key observation is that, for each $m \in \mathbb{N}$, $T^m A$ and $T^m B$ are rectangles as well. The second key observation is that they have a very special shape: in the stable direction their size has contracted by a factor λ^m while in the unstable direction the size has expanded by the same factor. Hence, provided m is chosen large enough, $T^m A$ and $T^m B$ are very thin in the stable direction and very elongated in the unstable direction. This property of stretching and squeezing, that we are witnessing here, is the cornerstone of almost all arguments in hyperbolic

theory. Of course, similar, but symmetrical, arguments hold for $T^{-m}A$ and $T^{-m}B$. We can then choose $m \in \mathbb{N}$ so large that the length of the unstable sides of $T^m B$ is larger than 2 and, at the same time, the same is true for the stable side of $T^{-m}A$. It is then a trivial geometric observation, best seen on the covering of $\mathbb{T}^2$, that $T^n A \cap T^{-n} B \neq \emptyset$, for each $n \geq m$, thus $T^{-2n}A \cap B \neq \emptyset$, which suffices to prove the topological mixing. □

The reader who starts to appreciate the spirit of the game may be unhappy about the previous proof. The problem is that we have used a bit too heavily the structure of the foliation (straight lines) and of $\mathbb{T}^2$ (the covering).

It is then quite natural to wonder if a more flexible and dynamical proof is available. Here it is.

ANOTHER PROOF OF PROPOSITION 8.3.4. Let us start by a preliminary result.

Given any rectangle A let us call A_c a rectangle of half the size and situated at its center.[11]

Lemma 8.3.5 *If $T^{-n}A_c \cap A_c \neq \emptyset$ for some $n \in \mathbb{N}$ such that $\lambda^n > 4$, then $T^{-mn}A \cap A \neq \emptyset$ for all $m \in \mathbb{N}$.*

PROOF. By construction $T^{-n}A$ intersects A completely from one unstable side to the other (see figure 8.1)

This means that $T^{-2n}A \supset T^{-n}(T^{-n}A \cap A)$, which is a very thin rectangle contained in $T^{-n}A$ and that crosses it from one unstable side to the other. Accordingly $T^{-2n}A$ will intersect A completely (from one unstable side to the other). By induction the result follows. □

Note that the $n \in \mathbb{N}$ required by the above statement always exists (see Problem 2.3).

Next, let $A, B \subset \mathbb{T}^2$ be two rectangles and let $n_B \in \mathbb{N}$ such that Lemma 8.3.5 applies to B. We then consider the Dynamical Systems $(\mathbb{T}^2, T^{n_B}, m)$, this is ergodic as well (see Problem 2.2).[12] Consequently, for each integer $i \in \{1, \dots, n_B - 1\}$ there exists $k_i \in \mathbb{N}$ such that

$$T^{-k_i n_B}(T^{-i}A_c) \cap B_c \neq \emptyset,$$

and the unstable size of A times $\lambda^{-k_i n_B}$ is smaller than one quarter of the unstable size of B (see Problem 2.4). This implies immediately that

$$T^{-k n_B}(T^{-i}A) \cap B \neq \emptyset \quad \forall k \geq k_i. \tag{8.3.2}$$

[11] This may seem a silly construction but it is a rather general trick used to exploit topological mixing and we will see it again under the name of *core* of a rectangle in chapter 8.

[12] This is the crucial property always needed to obtain mixing in hyperbolic systems: ergodicity of all the powers of the map.

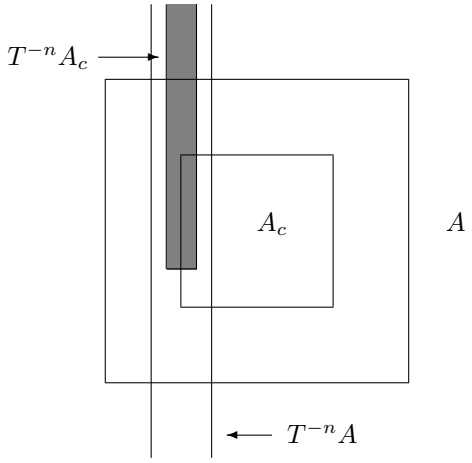

Figure 8.1: Intersection between A and $T^{-n}A$

In fact, $T^{-k_i n_B}(T^{-i}A)$ crosses B from one unstable side to the other and touches B_c, thus (8.3.2) can be proved by the same type of arguments used in Lemma 8.3.5.

Finally, set $k_m := \max\{k_i \mid i \in \{1, \dots, n_B\}\}$. For each $n > k_m n_B$ we can write $n = kn_B + i$ where $0 < i < n_B$, thus

$$T^{-n}A \cap B = T^{kn_B}(T^{-i}A) \cap B \neq \emptyset,$$

by (8.3.2). □

By the same arguments one can prove the following (see Problem 2.5).

Lemma 8.3.6 *Given any stable segment W^s of length δ, and any unstable segment W^u of length $L > \lambda\delta^{-1}$, then it holds $W^s \cap W^u \neq \emptyset$.*

To start discussing the problem of mixing we need to adopt a point of view among the many possible. We will take the one that looks at the measures (see Proposition 6.8.3 and Problem 1.6.33) which, by now, should be rather familiar to the reader. Calling μ_0 a measure absolutely continuous with respect to Lebesgue we would like to study the asymptotic behavior of $\mu_n := {T_*}^n \mu_0$. Thanks to Proposition 6.8.3 we need to study only the weak convergence. The first observation is that such a set of measures is compact hence we can study the set of its limit points Γ (of course with the goal of showing that it consists of only one point).[13] Such a set is simply the set of

[13]Note that such accumulation points are not necessarily invariant measures, this is why we considered accumulation points of averages in section 6.5.

limits of convergent subsequences. Since the measure μ_0 is absolutely continuous with respect to m there exists a function $h \in L^1(\mathbb{T}^2)$, $h \geq 0$, such that

$$\mu_0(f) = m(hf).$$

A lesson that we have learned from the computation in Fourier transform and from the Hopf argument is that the regularity of the functions do matter considerably and that it may be useful to consider, at first, regular functions and then obtain the wanted result by an approximation argument. Accordingly, we will restrict ourself to the case $h \in C^{(1)}(\mathbb{T}^2)$ and establish two fundamental facts.[14]

Lemma 8.3.7 *If $\bar{\mu} \in \Gamma$ then $\bar{\mu}$ is absolutely continuous with respect to Lebesgue. In addition, $\bar{h} = \frac{d\bar{\mu}}{dm} \in L^\infty(\mathbb{T}^2, m)$.*

PROOF. We notice that the sequence μ_n is uniformly absolutely continuous with respect to Lebesgue, that is $\forall f \in \mathcal{C}^{(0)}(\mathbb{T}^2)$ such that $f \geq 0$

$$\mu_n(f) = \int_{\mathbb{T}^2} h \circ T^{-n} f \leq \|h\|_\infty \|f\|_1.$$

This implies $\bar{\mu}(f) \leq \|h\|_\infty m(f)$ and

$$\bar{\mu}(A) = \sup_{\substack{C \subset A \\ C = \overline{C}}} \bar{\mu}(C) = \sup_{\substack{C \subset A \\ C = \overline{C}}} \inf_{\{f \in \mathcal{C}^{(0)} \mid f > \chi_C\}} \bar{\mu}(f) \leq \|h\|_\infty m(A), \tag{8.3.3}$$

where we have used (6.5.3) and (6.5.4). Clearly (8.3.3) implies the absolute continuity. Hence, by the Radon-Nikodym theorem [Roy88], there exists $\bar{h} \in L^1(\mathbb{T}^2, m)$ such that $d\bar{\mu} = \bar{h} dm$.

Next, let $A = \{x \in \mathbb{T}^2 \mid \bar{h}(x) > \|h\|_\infty\}$. If $m(A) \neq 0$, then

$$\|h\|_\infty m(A) < \int_A \bar{h} dm = \bar{\mu}(A) \leq \|h\|_\infty m(A)$$

which is a contradiction, thus $\bar{h} \leq \|h\|_\infty$ a.e.. □

The next argument is very similar to what we have already seen in Examples 6.5.1–Strange Attractors. Let us call D^u the derivative along the unstable direction (if v^+ is the normal vector in the unstable direction then $D^u f := \langle \nabla f, v^+ \rangle$).

Lemma 8.3.8 *There exists $c > 0$: for each $f \in \mathcal{C}^{(1)}(\mathbb{T}^2)$*

$$|\mu_n(D^u f)| \leq \lambda^{-n} c \|f\|_\infty.$$

[14] Actually, this regularity condition on h will be needed only in Lemma 8.3.8.

PROOF.

$$\begin{aligned}\mu_n(D^u f) &= \int_{\mathbb{T}^2} h(D^u f)\circ T^n = \int_{\mathbb{T}^2} h\langle(\nabla f)\circ T^n,\, v^+\rangle \\ &= \int_{\mathbb{T}^2} h\langle L^{-n}\nabla(f\circ T^n),\, v^+\rangle = \lambda^{-n}\sum_{i=1}^{2}\int_{\mathbb{T}^2} h\partial_{x_i}(f\circ T^n)v_i^+ \\ &= -\lambda^{-n}\int_{\mathbb{T}^2} D^u h f\circ T^n,\end{aligned}$$

where the last equality is obtained by integrating by parts with respect to both coordinates. Accordingly,

$$|\mu_n(D^u f)| \le \lambda^{-n}\|\nabla h\|_1\|f\|_\infty.$$

□

From the above results it follows that if $\bar\mu \in \Gamma$ then there exists $\bar h \in L^\infty(\mathbb{T}^2)$ such that, for each $f \in L^1(\mathbb{T}^2, m)$,

$$\bar\mu(f) = \int f\bar h dm$$

and for each $f \in \mathcal{C}^{(1)}(\mathbb{T}^2)$, $\bar\mu(D^u f) = 0$. This two facts together imply that $\bar h$ is constant almost everywhere.

To see this we start by a **local** argument showing that $\bar h$ is constant along the unstable direction. We have already done a similar argument, in Examples 6.5.1–Strange Attractors, by using Fourier series, let us see here a more measure theoretical argument to convince the reader that the global structure of $\mathbb{T}^2$ has nothing to do with the result.

Let us consider an arbitrary rectangle R of size smaller than $1/4$. Consider an arbitrary $f \in \mathcal{C}^{(1)}(\mathbb{T}^2)$ with support contained in $\overset{\circ}{R}$. Then consider coordinates in R parallel to its sides (since this is achieved by rotations and rigid translations it leaves invariant the Lebesgue measure). As before, the unstable sides are horizontal. Let us call x the coordinate along the stable direction and y the one along the unstable direction. In such coordinates $R = [0, a]\times[0, b]$ (we have translates the origin at the bottom left corner of R). Given $f \in \mathcal{C}^{(1)}$, we define

$$\begin{aligned}\tilde f(x,\, y) &= f(x,\, y) - \frac{1}{a}\int_0^a f(\xi,\, y)d\xi, \\ F(x,\, y) &= \int_0^x \tilde f(\xi,\, y)d\xi.\end{aligned}$$

Then $F|_{\partial R} = 0$ so F can be extended to a function on $\mathbb{T}^2$ by setting $F = 0$ outside R. Note, that F is continuous and differentiable everywhere apart

from the boundary ∂R where the derivative can be discontinuous. In the new coordinates D^u becomes simply the derivative with respect to x.

$$\int_{\mathbb{T}^2} \bar{h}f = \int_R \bar{h}f = \int_0^a dx \int_0^b dy \bar{h}\tilde{f} + \frac{1}{a}\int_0^b dy \int_0^a dx \bar{h}(x,\, y) \int_0^a d\xi f(\xi,\, y),$$

and, setting $\tilde{h}(y) = \frac{1}{a}\int_0^a d\xi \bar{h}(\xi,\, y)$, $\bar{f}(y) = \int_0^a d\xi f(\xi,\, y)$,

$$\int_{\mathbb{T}^2} \bar{h}f = \int_0^a dy \int_0^b dx \bar{h}\partial_x F + \int_0^b dy \tilde{h}(y)\bar{f}(y) = \int_{\mathbb{T}^2} \bar{h}D^u F + \int_0^b dy \tilde{h}(y)\bar{f}(y).$$

At this point a small obstacle appears, due to the fact that F is not $\mathcal{C}^{(1)}$. The problem is easily solved by approximating F by $\mathcal{C}^{(1)}$ functions F_ε such that $\|D^u F - D^u F_\varepsilon\|_1 \leq \varepsilon$. Then

$$\left|\int_{\mathbb{T}^2} \bar{h}D^u F\right| = \left|\int_{\mathbb{T}^2} \bar{h}D^u F - \int_{\mathbb{T}^2} \bar{h}D^u F_\varepsilon\right| \leq \|\bar{h}\|_\infty \varepsilon.$$

Hence, $\int_{\mathbb{T}^2} \bar{h}D^u F = 0$ also if the derivative is not continuous, consequently

$$\int_{\mathbb{T}^2} \bar{h}f = \int_{\mathbb{T}^2} \tilde{h}f. \tag{8.3.4}$$

By the arbitrariness of f (8.3.4) implies that $\bar{h} = \tilde{h}$ almost everywhere in $\mathring{R}$. Since R is arbitrary it follows that $\bar{h}$ is constant a.e. along the unstable direction.

A **global** argument is now needed to show that $\bar{h}$ must be constant.[15]

Proof of Proposition 8.3.1–a shortcut. Consider a line $\ell_a = \{x = a\}$. Clearly for each point $p = (a, y) \in \ell_a$ W^u_p intersects again ℓ_a at the point $(a, y + \omega_+ \mod 1)$ where $(1, \omega_+)$ is the unstable direction. Then we can consider the Dynamical Systems $(\ell_a, R_{\omega_+}, m)$, and the function $h_a = \bar{h}(a, y)$. By the previous discussion (and Fubini Theorem) it follows that, for almost every a, the function h_a is an $L^1(\ell_a, m)$ invariant function for the rotation R_{ω_+}; but we know that the irrational rotations are ergodic (see Examples 6.6.1), thus h_a =constant which implies immediately $\bar{h}$ constant. □

The above proof is simple but uses quite heavily the global properties of the foliation and of $\mathbb{T}^2$ to reduce the problem to one already studied (the irrational rotations). Clearly it is not clear how such a trick could work in more general situations. Again we would like a more flexible and dynamical argument.

[15] The fact that the argument is global, i.e. uses some properties of $\mathbb{T}^2$, reflects the fact that it is not as general has the Hopf argument which, instead, is of a completely local nature, as we will see better later.

PROOF OF PROPOSITION 8.3.1–DYNAMICAL. We will use a strategy already employed to prove the ergodicity of irrational rotations based on the existence of density points. Morally, this allows us to consider only rectangles. By topological mixing we can ensure that any two rectangle are crossed by the same unstable line (although it is more convenient to take preimmages of the rectangle and show that they must intersect a given unstable segment), so it is not possible that $\bar{h}$ has values different in the two rectangles. This very naïve argument can be made precise as follows.

If $\bar{h}$ it is not a.e. constant then there exists two sets A and B of positive measure such that $\bar{h}|_A > \bar{h}|_B$ a.e.. Let x_A and x_B be density points, of A and B respectively, and choose two rectangle R_A and R_B of the same size, smaller than $\frac{1}{4}$, and such that

$$\begin{aligned} m(A \cap R_A) &\geq \alpha m(R_A) \\ m(B \cap R_B) &\geq \alpha m(R_B) \end{aligned} \tag{8.3.5}$$

where $\alpha \in [0, 1)$ will be chosen later.

Let us consider $h \circ T^n$, clearly $h \circ T^n|_{T^{-n}A} > h \circ T^n|_{T^{-n}B}$ and the relations (8.3.5) hold for $T^{-n}A$, $T^{-n}R_A$ and $T^{-n}B$, $T^{-n}R_B$.

Next, let $\hat{R}_A \subset R_A$ and $\hat{R}_B \subset R_B$ be two shorter rectangles obtained by the original ones by chopping off a quarter of the length in the stable direction from each side. Let n_0 be so large that the stable length of the rectangles time λ^{n_0} is larger than one. Now chose another rectangle R, of size $\rho \leq \frac{1}{4}$, as you please. By topological mixing it follows that there exists $n > n_0$ such that $T^{-n}\hat{R}_A \cap R \neq \emptyset$ and $T^{-n}\hat{R}_B \cap R \neq \emptyset$. In addition, by the construction of $\hat{R}_A$ and $\hat{R}_B$ and the choice of n_0, it follows that $T^{-n}R_A$ and $T^{-n}R_B$ cross $\tilde{R}$ completely from one unstable side to the other, where $\tilde{R}$ is a rectangle containing R at its center and of double size. Moreover, the same quantitative argument of Lemma 8.3.6 shows that it is possible to choose n such that the stable length of $T^{-n}R_A$, $T^{-n}R_B$ is shorter than $8\lambda^2$.

Let L_A, L_B the two rectangles contained in $T^{-n}R_A \cap \tilde{R}$ and $T^{-n}R_B \cap \tilde{R}$, respectively, that cross $\tilde{R}$ from an unstable side to the other. Chose

$$\alpha = 1 - \frac{m(L_B)}{4m(R_B)} = 1 - \frac{m(L_A)}{4m(R_A)}.$$

The all point is that, on almost all the unstable lines in $\tilde{R}$, $\bar{h} \circ T^n$ is constant, so if one of this unstable lines intersects both $T^{-n}A$ and $T^{-n}B$ we have a contradiction. Thus, it must be

$$m\left(\left[\bigcup_{x \in T^{-n}A} W^u_x \cap L_B\right] \bigcap \left[\bigcup_{x \in T^{-n}B} W^u_x \cap L_B\right]\right) = 0.$$

Fubini theorem implies

$$m\left(\bigcup_{x\in T^{-n}A} W^u_x \cap L_B\right) = m\left(\bigcup_{x\in T^{-n}A} W^u_x \cap L_A\right) \geq m(T^{-n}A \cap L_A),$$

and

$$m\left(\bigcup_{x\in T^{-n}B} W^u_x \cap L_B\right) \geq m(T^{-n}B \cap L_B),$$

This yields:

$$\begin{aligned} m(L_B) &\geq m(T^{-n}A \cap L_A) + m(T^{-n}B \cap L_B) \\ &\geq m(T^{-n}A \cap T^{-n}R_A) - m(T^{-n}R_A \backslash L_A) \\ &\quad + m(T^{-n}B \cap T^{-n}R_B) - m(T^{-n}R_B \backslash L_B) \\ &\geq 2\{\alpha m(T^{-n}R_B) - m(T^{-n}R_B) + m(L_B)\} \\ &\geq \frac{3}{2} m(L_B) \end{aligned}$$

which is a contradiction. This shows that is not possible that the unstable manifolds starting at $T^{-n}A$ systematically avoid $T^{-n}B$.

Hence, $\bar{h}$ is constant, but then $\bar{h} = \int_{\mathbb{T}^2} \bar{h} = \bar{\mu}(1) = \mu_0(1)$. We have just proved that Γ consists of only one measure: the Lebesgue measure. Thus

$$\lim_{n\to\infty} \int_{\mathbb{T}^2} hf \circ T^n dm = \int_{\mathbb{T}^2} h dm \int_{\mathbb{T}^2} f dm,$$

for each g, $f \in C^{(1)}(\mathbb{T}^2)$. The mixing follows by the same approximation argument used in the Fourier series analyses. □

8.4 Shadowing

In this section we explore the topological complexity of the dynamics of our model systems. I have already remarked that when such a strong instability with respect to the initial condition is present it is impossible to follow exactly an orbit of the system. In fact if we compute (e.g. with a computer) the orbit of the initial point $x \in \mathbb{T}^2$, due to round off errors we do not get an orbit but rather a *pseudo-orbit*.

Definition 8.4.1 *Give an systems (X, T), X Riemannian manifold, an infinite sequence $\{x_i\}_{i\in\mathbb{Z}} \subset \mathbb{T}^2$ is called an ε-pseudo orbit if, for all $i \in \mathbb{Z}$,*

$$d(x_{i+1}, Tx_i) \leq \varepsilon.$$

Which means exactly that at each step an error of order ε is allowed.

The following result, beside being very useful, is a partial replay to the argument that it is not possible to follow orbits on a computer. Although the result is quite general, we state, and prove, it in our special context.

Proposition 8.4.2 *For each $\delta > 0$ there exists and $\varepsilon > 0$ such that, if $\{x_i\}$ is a ε-pseudo-orbit for the Arnold cat, then there exists $\xi \in \mathbb{T}^2$ such that*

$$d(x_i, T^i\xi) \leq \delta \quad \forall i \in \mathbb{Z}.$$

That is, there exists an orbit that δ-shadows the pseudo-orbit, moreover such an orbit is unique.

PROOF. As usual we consider rectangular (better yet: square) neighborhood of points. So, let $Q_\varepsilon(x)$ be a square of size ε centered at x with sides parallel to the stable and unstable direction, respectively.

Next, let us consider $TQ_\delta(x_0)$, since $d(Tx_0, x) \leq \varepsilon$, if $\frac{\delta}{2\lambda} + \varepsilon < \frac{\delta}{2}$ and $\frac{\lambda\delta}{2} - \varepsilon > \frac{\delta}{2}$, then $TQ_\delta(x_0)$ crosses $Q_\delta(x_1)$ completely from the stable side to the other stable side. Thus, provided we choose $\delta \geq \frac{2\lambda}{\lambda-1}\varepsilon$, we have the picture of the intersection between rectangle that we have already learned to like.

Of course the same transversal intersection takes place for each $TQ_\delta(x_i)$ and $Q_\delta(x_{i+1})$. This immediately implies that $T^nQ_\delta(x_0)$ crosses $Q_\delta(x_n)$ from one stable side to the other (see figure 8.2)

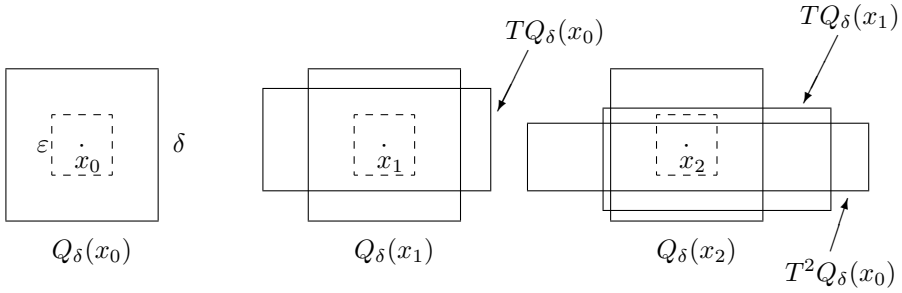

Figure 8.2: Intersection between $T^nQ_\delta(x_0)$ and $Q_\delta(x_n)$

Thus $K_n = T^{-n}(T^nQ_\delta(x_0) \cap Q_\delta(x_n))$ is a sequence of nested $(K_{n+1} \subset K_n)$ vertical rectangles. The unstable side of K_n is of size $\lambda^{-n}\delta$ while the stable side is of size δ.

Clearly, if $\xi \in K_n$, then

$$d(T^i\xi, x_i) < \delta \quad \forall i \in \{0, \ldots, n\}.$$

We can then consider the vertical line $K_\infty = \cap_{n\in\mathbb{N}} K_n$, by construction K_∞ consists of points whose orbit δ shadows $\{x_i\}_{i\in\mathbb{N}}$. By doing the same exact construction in the past we obtain an horizontal line $\tilde{K}_\infty$ of points that δ shadows $\{x_{-i}\}_{i\in\mathbb{N}}$. The theorem is then proven by choosing $\{\xi\} = \tilde{K}_\infty \cap K_\infty$.

the uniqueness should be obvious from the construction. In alternative the reader can prove it by contradiction. □

The above theorem is not so helpful from the measure theoretical point of view, since it could happen that the set of trajectories that shadow pseudo-orbits are of measure zero. (*say more*)

Nevertheless, it is very useful from the topological point of view (see Problem 2.15 for a dim glimpse to such possibilities).

8.5 Markov partitions

In all the above constructions the concept of rectangle has played a key rôle. In this section we present a construction that is the glorification of such a point of view.

Consider the stable and unstable manifolds of zero and prolong them until they meet (of course when they meet we meet an old friend: an homoclinic intersection) few times.

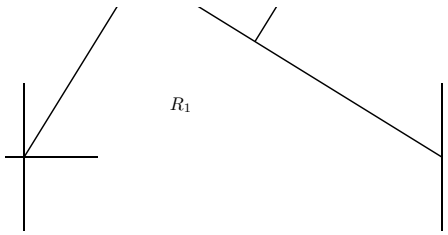

Figure 8.3: Markov partition

Clearly in such a way we have obtained a partition of $\mathbb{T}^2$. Such a partition consists of rectangles with sides that are either stable or unstable manifolds. We call them respectively the stable and the unstable sides of the rectangles. A partition is Markov if the preimage of each unstable side of a rectangle is contained in the unstable side of a rectangle and the image of every stable side is contained in the stable side of a rectangle. The reader can check that it is possible to use the above construction to have a Markov partition with (for example) three rectangles (see Figure 8.3 where the case $a = 1$ is drown).

Problems

8.1 Use the Diofantine properties of the stable and unstable direction to obtain better estimates of the decay of correlations. The Diofantine property refers to the following fact: if we normalize the eigenvectors in such a way that $v_\pm = (1, \omega_\pm)$, then $\omega_\pm$ are irrational numbers that are badly approximated by rationals: there exists $c \geq 0$ such that $|\omega_\pm - \frac{p}{q}| \geq \frac{c}{q^2}$ for each $p, q \in \mathbb{N}$.

8.2 Prove that the dynamical System $(\mathbb{T}^2, T^n, m)$ (where T is the Arnold cat map) is ergodic for each $n \in \mathbb{N}$. (Hint: the same proof as for $n = 1$.)

8.3 Let (X, T, μ) be a Dynamical Systems where X is a compact metric space, T is continuous, and μ charges the open sets (i.e. if $U \subset X$ is open, then $\mu(U) > 0$). Prove that for each $U \subset X$ open, there exist infinitely many $n \in \mathbb{N}$ such that $T^{-n}U \cap U \neq \emptyset$. (Hint: Poincaré Theorem.)

8.4 Let (X, T, μ) be an ergodic Dynamical Systems where X is a compact metric space, T is continuous, and μ charges the open sets. Prove that for each $U, V \subset X$ open, there exist infinitely many $n \in N$ such that $T^{-n}U \cap V \neq \emptyset$. (Hint:For each $k \in \mathbb{N}$, $A = \cup_{n \leq k} T^{-n}U$ is an invariant open set, if it does not intersect V, then $m(A) < 1$, thus, by ergodicity, $m(A) = 0$ which implies $U = \emptyset$.)

8.5 Prove Lemma 8.3.6. (Hint: As in the proof of Topologically mixing consider $T^{-n}W^s$, T^nW^u and chose n so large that $\lambda\delta > 2$ while the length L of W^u must satisfy $\lambda^{-n}L > 2$.)

8.6 Show that for each $x \in \mathbb{T}^2$ the global unstable manifold $W^u(x)$ is dense in $\mathbb{T}^2$. (Hint: *An algebraic proof*–Let us normalize $v_+ = (1, \omega)$, then ω is irrational. Clearly $W^u(x) = \{x + tv_+ \mod 1\}_{t \in \mathbb{R}}$. Consider a point $y = (y_1, y_2)$ and chose $t_0 = y_1 - x_1$, then, for each $n \in \mathbb{Z}$, $x + (t_0 + n)v_1^+ \mod 1 = (y_1, R_\omega^n \xi \mod 1)$ where $\xi = x_2 + (y_1 - x_1)\omega \mod 1$. Now, we know that R_ω has dense orbits (see Examples 6.6.1–Rotations), thus the result.
A dynamical proof–It follows Lemma 8.3.6 plus the fact that $T^{-n}W^u$ is shorter than W^u.)

8.7 Consider the global unstable foliation $\{W^u(x)\}$ and choose an interval of length (in the horizontal direction) one from each fiber.[16] Let K be the set obtained by the union of all such segments. Prove that K it is not measurable. (Hint: Define $R : \mathbb{T}^2 \to \mathbb{T}^2$ by $R(x, y) = (x, R_\omega y)$. Then, remember Problem 1.6.19.)

[16] The Axiom of choice again.

8.8 Let $W^u(x)$, $W^s(x) \subset U \subset \mathbb{R}^2$, $U = \overset{\circ}{U}$ and $\bar{U}$ compact, smooth manifolds ($\mathcal{C}^{(1)}$ curves) such that, the $\{W^{u(s)}(x)\}_{x\in U}$ are pairwise disjoint, $\partial W^{u(s)}(x) \subset \partial U$, if $z \in W^u(x) \cap W^s(y)$, then the angle between $W^u(x)$ and $W^s(y)$ at z is larger than some $\theta > 0$. In addition, assume that, calling $v^{u(s)}(x)$ the unit tangent vector to $W^{u(s)}(x)$ at x, $v^{u(s)} \in \mathcal{C}^{(1)}$. We will call such two foliation " $\mathcal{C}^{(1)}$ uniformly transversal foliations." Show that to each such a foliation it is associated a change of variable (a diffeomorphism $\Psi : U \to U$) and that to each change of variables is associated such a foliation. (Hint: ...)

8.9 Consider two $\mathcal{C}^{(1)}$ uniformly transversal foliations (as in Problem 2.8). Prove that if $f \in L^\infty$ is constant along almost every fiber of the two foliations, then it is constant almost everywhere. (Hint: Do the argument locally and change variables so that the foliations becomes straight.)

8.10 Consider the Bernoulli measures μ_p^B defined on Σ_2^+ (the one sided sequences with two symbols) by choosing $p_0 = p$ and $p_1 = 1 - p$ (see Examples 6.2.1–Bernoulli shift). Show that, if $p \neq p'$ then μ_p^B and $\mu_{p'}^B$ are mutually singular. (Hint: All the dynamical systems $(\Sigma_2^+, \tau, \mu_p^B)$ are ergodic–See Examples ?? and ??.)

8.11 Let μ_p be the measure on $[0,1]$ obtained from μ_p^B by the binary representation of the real numbers let

$$F_p(x) := \mu_p([0,x]).$$

Show that, for each $p \in (0,1)$, $F_p : [0,1] \to [0,1]$ is one one, onto, continuous. In addition, show that there exists $c \in \mathbb{R}^+$ such that, for each $p, q \in [\frac{1}{4}, \frac{3}{4}]$, holds

$$|F_p(x) - F_q(x)| \leq c|p - q|.$$

(Hint: Note that the cylinder correspond to intervals with end points made of binary rationals. It is then immediately clear that all the measures μ_p give positive measures to the open sets. To prove the last inequality prove the representation

$$F_p(x) = \sum_{n=0}^{\infty} \sigma_n \prod_{i=0}^{n} p^{\sigma_i}(1-p)^{1-\sigma_i}$$

where σ is the binary representation of x.)

8.12 Construct $\phi : [0,1] \to [0,1]$, invertible and continuous, such that there exists $A \subset [0,1]$ with $m(A) = 0$ while $m(\phi(A)) = 1$. (Hint: Any of the above F_p will do.)

8.13 Construct a continuous foliation Ψ in $[0,1]^2$made of $\mathcal{C}^\infty$ leaves (that is Ψ is a isomomorphism of $[0,1]^2$ and $\Psi(\cdot,y) \in \mathcal{C}^\infty$). In addition, the foliation must be made of straight lines in $\{(x,y) \in [0,1]^2 \mid x \in [0,\frac{1}{4}] \cup [\frac{3}{4},1]\}$ but is should not be absolutely continuous in the region $\{(x,y) \in [0,1]^2 \mid x \in [\frac{1}{4},\frac{3}{4}]\}$. (Hint: Let $\varphi \in \mathcal{C}^{(\infty)}(\mathbb{R})$, $\varphi(\mathbb{R}) = [0,1]$, $\varphi(x) = 0$ for $x < 0$ and $\varphi(x) = 1$ for $x > \frac{1}{2}$. Then, using ϕ from Problem 2.12, define

$$\Psi(x,y) = \begin{cases} (x,y) & \text{if} \quad x \in [0,\frac{1}{4}] \\ (x,[1-\varphi(x-\frac{1}{4})]y + \varphi(x-\frac{1}{4})\phi(y)) & \text{if} \quad x \in [\frac{1}{4},\frac{3}{4}] \\ (x,\phi(y)) & \text{if} \quad x \in [\frac{3}{4},1]. \end{cases}$$

Clearly the leaves $\Psi(\cdot,y)$ are $\mathcal{C}^{(\infty)}$, yet the foliation it is not absolutely continuous.)

8.14 Find two $\mathcal{C}^{(0)}$ uniformly transversal foliations in $[0,1]^2$, with $\mathcal{C}^{(\infty)}$ leaves, such that the Hopf argument does not apply. (Hint: Call Ψ_p, $p \in [\frac{1}{4},\frac{3}{4}]$ the foliation constructed in the Problem 13 starting from the function F_p defined in the Problem 11. Choose a sequence p_n converging to one quarter, e.g. $p_n = \frac{1}{4} + \frac{1}{4^n}$, then let $x_n = \frac{1}{2} - \frac{1}{2n}$. Finally define the foliation

$$\Psi(x,y) = \begin{cases} \Psi_{p_n}(x_n + (x_{n+1} - x_n)x, y) & \text{for } x \in [x_n, x_{n+1}] \\ (x, F_{\frac{1}{4}}(y)) & \text{for } x \in [\frac{1}{2},1] \end{cases}$$

Further define the function $g : [0,1] \to [0,1]$ to be one on a set of full measure for $\mu_{\frac{1}{4}}$ and of zero measure for μ_{p_n} and zero otherwise. The functions f^+, f^- defined by

$$f^-(x,y) = \begin{cases} 0 & \text{for } x \in [0,\frac{1}{2}) \\ 1 & \text{for } x \in [\frac{1}{2},1] \end{cases}$$

and

$$f^+(x,\Psi(x,y)) = g(x),$$

are then constant on the vertical and the Ψ foliation respectively. Moreover they clearly are equal Lebesgue almost everywhere, nevertheless they are certainly not constant.)

8.15 Show (first without using Markov Partitions and then by using Markov partitions) that the Arnold cat has at least e^{cn} periodic orbits of period n, for some $c > 0$. (Hint: If we have a rectangle R of size ε, then $T^{-n}R \cap R \neq \emptyset$ for some $n \leq c \ln \varepsilon^{-1}$. Then, if $x \in T^{-n}R \cap R$ we consider the pseudo orbit $x_k = T^i x$ where $i = k \mod n$. Then Proposition

8.4.2 implies the existence of a periodic orbit in an ε-neighborhood R_ε of R. On the other hand the boxed $T^{-k}R_\varepsilon$, $k \in \{0, \dots, n\}$ invade a part of $\mathbb{T}^2$ of measure $c\varepsilon^2 \ln \varepsilon^{-1}$. The argument is then concluded taking boxes in the remaining space and continuing until all the available space is exhausted. On the other hand, if one takes in account Markov partions, then the number of periodic orbits is given–appart from the non-invertibility of the coding–by the number of periodic simbolic sequences of period n.)

Notes

Hopf history and ref

Mention Young-Robinson example

. . . . . .

Chapter 9

Non-uniform hyperbolicity an introduction

In this chapter, we discuss what is probably the simplest example of non-uniform hyperbolic behaviour. This is not intended to be a discussion of the general theory; it is just a taste of it. The theory of non-uniformly hyperbolic systems is rather vast, starting with Pesin theory, till the results on the Henon map and their generalizations.

9.1 Pomeau-Manneville map

Let us consider the map

$$f(x) = \begin{cases} f_0(x) := x + 2^\gamma x^{1+\gamma} & \text{if } x \in [0, \frac{1}{2}] \\ f_1(x) := 2x - 1 & \text{if } x \in (\frac{1}{2}, 1], \end{cases} \tag{9.1.1}$$

for some $\gamma \in (0, 1]$. Such a map was introduced as a model for the phenomena of intermittency; indeed, the trajectory has a hyperbolic character away from zero, but in a neighborhood of zero, the motion is very regular.
Let us first study the latter regime: consider the preimages of a point x under the map f_0.

Lemma 9.1.1 *for $x_0 \in [0, 1]$ let $x_n = f_0^{-n}(x_0)$. Then, for each $n \in \mathbb{N}$,*

$$(x_0^{-1} + 2^\gamma \gamma n)^{-\frac{1}{\gamma}} \leq x_n \leq 2^{-1} \gamma^{-\frac{1}{\gamma}} n^{-\frac{1}{\gamma}}$$

PROOF. Let $A = x_0^{-\gamma}$ consider the sequence $a_n = (A + 2^\gamma \gamma n)^{-\gamma^{-1}}$, then

$$\begin{aligned} a_{n-1} &= a_n + \frac{2^\gamma}{(A + 2^\gamma \gamma n)^{\frac{1+\gamma}{\gamma}}} + \frac{2^{\gamma-1}(\gamma^{-1}+1)}{(A + 2^\gamma \gamma n)^{\gamma^{-1}+2}} \xi_n^2 \\ &= a_n + 2^\gamma a_n^{1+\gamma} + 2^{\gamma-1}(\gamma^{-1}+1) a_n^{2+\gamma} \xi_n^2. \end{aligned} \tag{9.1.2}$$

Accodringly, $a_{n-1} \geq f_0(a_n)$. Consequently,

$$a_n \leq f_0^{-1}(a_{n-1}) \leq f_0^{-n}(a_0) = f_0^{-n}(x_0) = x_n.$$

Next, suppose that $x_k \leq c k^{-\frac{1}{\gamma}}$ for all $1 \leq k \leq n$, then

$$x_n = f_0(x_{n+1}) = x_{n+1} + 2^\gamma x_{n+1}^{1+\gamma}.$$

If $x_{n+1} > c(n+1)^{-\frac{1}{\gamma}}$, then

$$\begin{aligned} c n^{-\frac{1}{\gamma}} &\geq x_n > c(n+1)^{-\frac{1}{\gamma}} + 2^\gamma c^{1+\gamma} (n+1)^{-\frac{1+\gamma}{\gamma}} \\ &\geq c n^{-\frac{1}{\gamma}} - \frac{c}{\gamma} n^{-\frac{1}{\gamma}-1} + 2^\gamma c^{1+\gamma} (n+1)^{-\frac{1+\gamma}{\gamma}} \\ &= c n^{-\frac{1}{\gamma}} + \frac{c}{\gamma} n^{-\frac{1}{\gamma}-1} \left[2^\gamma c^\gamma \gamma \left(1 + \frac{1}{n}\right)^{-\frac{1+\gamma}{\gamma}} - 1 \right] \\ &\geq c n^{-\frac{1}{\gamma}} + \frac{c}{\gamma} n^{-\frac{1}{\gamma}-1} \left[2^\gamma c^\gamma \gamma - 1 \right] \end{aligned}$$

this is a contradiction if we choose $c = 2^{-1}\gamma^{-\frac{1}{\gamma}} \geq \frac{1}{2}$, which also ensures $x_1 \leq c$. □

The basic ided is to study the return map $F : [\frac{1}{2}, 1] \to [\frac{1}{2}, 1]$. That is, let $\tau(x) = \inf\{n \in \mathbb{N} \,:\, f^n(x) \in [\frac{1}{2}, 1]\}$, and

$$F(x) = f^{\tau(x)}(x).$$

If we choose $x_0 = 1$, then $x_1 = \frac{1}{2}$. So $f^n([x_{n+1}, x_n]) = [\frac{1}{2}, 1]$. Accordingly, setting $z_n = f_1^{-1}(x_n)$, $\tau(x) = n$ for all $x \in [z_n, z_{n-1}]$, and $F([z_n, z_{n-1}]) = [\frac{1}{2}, 1]$. That is, F is a Markov map with a countably infinite number of branches. To study such a map, we need first to investigate the distortion $D(x) = \frac{F''(x)}{F'(x)^2}$.

Lemma 9.1.2 *There exists $K > 0$ such that for each $x \in [\frac{1}{2}, 1]$*

$$D(x) \leq K.$$

PROOF. By a direct computation, for $x \in [z_n, z_{n-1}]$,

$$\begin{aligned} D_n(x) &= \frac{(f \circ f^{n-1})''(x)}{f'(f^{n-1}(x))^2 (f^{n-1})'(x)^2} \\ &= \frac{f''(f^{n-1})(x)(f^{n-1})'(x)^2 + f'(f^{n-1}(x))(f^{n-1})''(x)}{f'(f^{n-1}(x))^2 (f^{n-1})'(x)^2} \\ &= D_1(f^{n-1}(x)) + \frac{1}{f'(f^{n-1}(x))} D_{n-1}(x) \\ &= \sum_{k=1}^{n} D_1(f^{n-k}(x)) \prod_{j=1}^{k-1} \frac{1}{f'(f^{n-j}(x))} = \sum_{k=1}^{n} \frac{D_1(f^{n-k}(x))}{(f^{k-1})'(f^{n-k}(x))}. \end{aligned}$$

In addition, for $w, y \in [z_n, z_{n-1}]$, let $w_k = f^k(w)$ and $y_k = f^k(y)$, we have

$$\begin{aligned} \frac{(f^n)'(w)}{(f^n)'(y)} &= \operatorname{Exp}\left[\sum_{k=2}^{n-1} (\ln(f_0'(w_k)) - \ln(f_0'(y_k))\right] \\ &\le \operatorname{Exp}\left[\sum_{k=2}^{n-1} \left\| \frac{f_0''}{f_0'} \right\|_\infty |w_k - y_k|\right] \\ &\le \operatorname{Exp}\left[\left\| \frac{f_0''}{f_0'} \right\|_\infty \sum_{k=2}^{n-1} |x_{m-k} - x_{m-k-1}|\right] \le \operatorname{Exp}\left[\frac{1}{2} \left\| \frac{f_0''}{f_0'} \right\|_\infty\right] =: C. \end{aligned}$$

Since $f^{k-1}([x_{n-k-1}, x_{n-k}]) = [\frac{1}{2}, 1]$, by the mean value theorem there must exists $\xi_k \in [x_{n-k-1}, x_{n-k}]$ such that $(f^{k-1})'(\xi_k) = [2|x_{n-k-1} - x_{n-k}|]^{-1}$. Let $\zeta_k \in [z_n, z_{n-1}]$ be such that $f^{n-k}\zeta_k = \xi_k$, then

$$\frac{1}{(f^{k-1})'(f^{n-k}(x))} = 2|x_{n-k-1} - x_{n-k}| \frac{(f^{k-1})'(f^{n-k}(\zeta_k))}{(f^{k-1})'(f^{n-k}(x))} \le 2C|x_{n-k-1} - x_{n-k}|,$$

from which the result readily follows. □

By Lemma 9.1.2, the first return map F has a unique invariant measure absolutely continuous with respect to Lebesgue. Accordingly, so has the Kakutani tower $([\frac{1}{2}, 1], S)$. Let ν be such a measure. Then Theorem 6.7.8 implies that there exists a measure $\mu = \pi_* \nu$ which is absolutely continuous and $([0, 1], f, \mu)$ is a measurable dynamical system. In addition,

$$\int_0^1 g \circ f^n \varphi d\mu = \pi_* \nu(g \circ f^n) = \nu(\varphi \circ \pi g \circ f^n \circ \pi) = \nu(g \circ \pi \circ S^n \varphi \circ \pi).$$

In other words, the decay of correlation for the map f and the map S coincide.

9.2 Young towers

.....

Appendices

Appendix A

Fixed Points Theorems (an idiosincratic selection)

In this appendix, I provide some standard and less standard fixed-point theorems. These constitute a very partial introduction to the subject. The choice of the topics is motivated by the needs of the previous chapters.

A.1 Banach Fixed Point Theorem

Theorem A.1.1 (Fixed point contraction) *Given a Banach space $\mathcal{B}$, a bounded closed set $A \subset \mathcal{B}$ and a map $K : A \to \mathcal{B}$ if*

i) $K(A) \subset A$,

ii) there exists $\sigma \in (0,1)$ *such that* $\|K(v) - K(w)\| \leq \sigma \|v - w\|$ *for each* $v, w \in A$,

then there exists a unique $v_* \in A$ *such that* $Kv_* = v_*$.

PROOF. Since A is bounded $\sup_{x,y \in A} \|x - y\| = L < \infty$, i.e. it has a finite diameter. Let $a_0 \in A$ and consider the sequence of points defined recursively by $a_{n+1} = K(a_n)$ and the sequence of sets $A_0 = A$ and $A_{n+1} = K(A_n) \subset A$. Let $d_n := \sup_{x,y \in A_n} \|x - y\|$ be the diameter of A_n. Then if $x, y \in A_n$, we have

$$\|K(y) - K(x)\| \leq \sigma \|x - y\| \leq \sigma d_n.$$

That is $d_{n+1} \leq \sigma d_n \leq \sigma^n L$. This means that, for each $n, m \in \mathbb{N}$, $a_n, a_0 \in A$ and $a_m, a_{n+m} \in A_m$, hence $\|a_{n+m} - a_m\| \leq \sigma^m L$. That is, $\{a_n\} \subset A$ is a Cauchy sequence and, being $\mathcal{B}$ a Banach space, it must have an accumulation

point $v_* \in \mathcal{B}$. Moreover, since A is closed, it must be $v_* \in A$. Clearly

$$\begin{aligned}\|Kv_* - v_*\| &= \lim_{n\to\infty} \|Kv_* - a_n\| = \lim_{n\to\infty} \|Kv_* - Ka_{n-1}\| \\ &\leq \lim_{n\to\infty} \sigma\|v_* - a_{n-1}\| = 0.\end{aligned}$$

Hence, v_* is a fixed point. Next, suppose there exists $u \in A$ such that $Ku = u$. Then

$$\|u - v_*\| = \|K(u - v_*)\| \leq \sigma\|u - v*\|$$

implies $u = v_*$. □

Corollary A.1.2 *Given a Banach space $\mathcal{B}$ and a map $K : \mathcal{B} \to \mathcal{B}$ with the property that there exists $\sigma \in (0, 1)$ such that $\|K(v) - K(w)\| \leq \sigma\|v - w\|$ for each $v, w \in \mathcal{B}$, then there exists a unique $v_* \in \mathcal{B}$ such that $Kv_* = v_*$.*

PROOF. To prove the theorem, for each $L \in \mathbb{R}_+$ consider the sets $B_L := \{v \in \mathcal{B} : \|v\| \leq L\}$. Then $\|K(v)\| \leq \|K(v) - K(0)\| + \|K(0)\| \leq \sigma\|v\| + \|K(0)\| \leq \sigma L + \|K(0)\|$. Thus, for each $L \geq (1 - \sigma)^{-1}\|K(0)\|$ we have that $K(B_L) \subset B_L$. The existence follows by applying Theorem A.1.1. The uniqueness follows from the same argument used at the end of the proof of Theorem A.1.1. □

A.2 Brouwer's Fixed Point Theorems

The basic problem addressed in this section is to study the existence of fixed points for continuous maps $f : D \to D$, for some domain D. The remarkable feature of the theorems that we are going to present is that they relate the geometrical properties of the domain of a map to the existence of a fixed point. However, one should note that the fixed point may not be unique. In the following, I provide elementary proofs, which will also be constructive. Other proofs based on algebraic topology exist, but are outside the scope of this book.
We present a sequence of results that build on each other, progressively increasing the level of generality.

A.2.1 Maps on a symplex

We start by recalling the definition of a simplex.

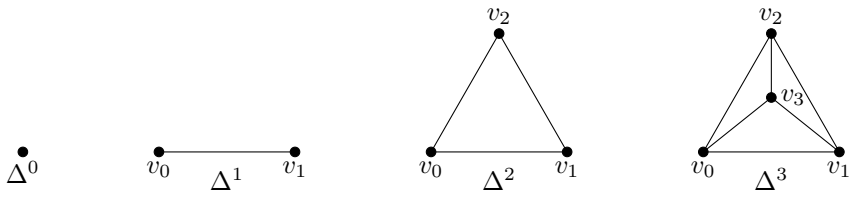

Figure A.1: Low-dimensional examples

Definition A.2.1 (Geometric n-simplex) *Let $v_0, v_1, \ldots, v_n$ be* affinely independent points *in $\mathbb{R}^m$, $m \geq n$.*[1] *The* n-simplex *spanned by these points is*

$$\Delta^n(v_1, \ldots, v_{n+1}) = \Big\{ x \in \mathbb{R}^m : x = \sum_{i=1}^{n+1} \lambda_i v_i,\ \lambda_i \geq 0,\ \sum_{i=1}^{n+1} \lambda_i = 1 \Big\}.$$

The standard n-simplex *in $\mathbb{R}^{n+1}$ is*

$$\Delta^n := \Delta^n(e_1, \ldots, e_{n+1}) = \Big\{ (x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} : x_i \geq 0,\ \sum_{i=1}^{n+1} x_i = 1 \Big\}.$$

Definition A.2.2 (Coloring) *Let Δ^n be the standard n–simplex, and let T be a simplicial subdivision (triangulation) of Δ^n. We call $\mathcal{V}(T)$ the set of vertices of the simplicial decomposition of T. A* s-coloring *of T is a function $\ell : \mathcal{V}(T) \to \{1, \ldots, n+1\}$ such that if v lies on the face of Δ^n opposite e_i (that is, $v_i = 0$), then $\ell(v) \neq i$. A simplex with vertices in $\mathcal{V}(T)$ is* fully colored *if, calling V the set of its vertices, $\ell|_V$ is invertible on its image.*

The basis tool that we will use is the following combinatorial lemma.

Lemma A.2.3 (Sperner's Lemma) *Let Δ^n be the standard n–simplex. Let T be a simplicial subdivision (triangulation) of Δ^n. Any s-colouring of T contains at least one fully colored simplex.*

Proof. The proof is by induction on n.
Let us start with $n = 1$. Here Δ^1 is the interval with endpoints e_0, e_1. The labeling rule forces e_0 to have label 0 and e_1 to have label 1. If all the subdivisions have vertices with the same color, then e_0 and e_1 would have the same color, contrary to the assumption.
Assume the lemma is true for dimension $n-1$. Consider Δ^n. By assumption,

[1] A set of points $v_0, v_1, \ldots, v_n \in \mathbb{R}^m$ is called *affinely independent* if the collection of vectors $v_1 - v_0,\ v_2 - v_0,\ \ldots,\ v_n - v_0$ are linearly independent.

there is at least one fully colored $(n-1)$–simplex $\Delta(v_1, \dots v_n)$, $v_i \in \mathcal{V}(T)$, lying on the boundary $\partial\Delta^n$. Let $\Delta_1 := \Delta(v_1, \dots v_{n+1}) \in T$ be the n-simplex containing $\Delta(v_1, \dots, v_n)$. If $\ell(v_{n+1}) \neq \ell(v_i)$ for all $i \leq n$, then we have a fully colored simplex and we are done. Otherwise, there is a unique j such that $v_{n+1} = v_j$. We then consider the simplex $\Delta(v_1, \dots, v_{j-1}, v_{j+1}, \dots, v_{n+1})$, which is fully colored by construction. Note that each face of an element of T belongs to two elements of T, unless it belongs to $\partial\Delta^n$ in which case it belongs to a unique element of T. So there exists a unique $v_{n+2} \in \mathcal{V}(T)$ such that $v_{n+2} \neq v_j$ and $\Delta_2 := \Delta(v_1, \dots, v_{j-1}, v_{j+1}, \dots, v_{n+1}, v_{n+2}) \in T$. Again, either is fully colored, or we can erase the vertex with the same color as v_{n+2} and obtain another fully covered $n-1$-simplex. In this way, we can construct a sequence of simplices $\{\Delta_k\} \in T$.

Next, we show that $\Delta_k = \Delta_j \implies k = j$. Suppose the contrary, and let k be the smallest integer for which there exists $j < k$ such that $\Delta_k = \Delta_j$. Let $w_l^+ \in \mathcal{V}(T)$ be the last vertex added to obtain Δ_l and $w_l^- \in \mathcal{V}(T)$ the unique vertex in Δ_l such that $\ell(w_l^+) = \ell(w_l^-)$. Consequently, if $V(\Delta)$ are the verteces of Δ, we have $V(\Delta_k) = [V(\Delta_{k-1}) \backslash \{w_{k-1}^-\}] \cup \{w_k^+\}$ and $V(\Delta_{j+1}) = [V(\Delta_j) \backslash \{w_j^-\}] \cup \{w_{j+1}^+\}$. By contruction $\Delta(V(\Delta_k) \setminus \{w_k^+\})$, $\Delta(V(\Delta_k) \setminus \{w_k^-\})$, $\Delta(V(\Delta_j) \setminus \{w_j^-\})$, and $\Delta(V(\Delta_j) \setminus \{w_j^+\})$ are all fully coloured. Since, by hypothesis, $V(\Delta_k) = V(\Delta_j)$, it must be $w_k^\pm \in \{w_j^-, w_j^+\}$, otherwise $\Delta(V(\Delta_k) \setminus \{w_k^\pm\})$ could not be fully colored. So, either $w_k^\pm = w_j^\pm$, or $w_k^\pm = w_j^\mp$. If $w_k^+ = w_j^+$, and $j > 1$, then it must be $\Delta_{k-1} = \Delta_{j-1}$ contradicting the hypothesis that k is the smaller integer for which this happens. If $j = 1$, then note that $w_1^+ \notin \partial\Delta^n$ while $w_k^+ \in \partial\Delta^n$ since otherwise Δ_{k-1} would have a vertex outiside Δ^n. It remains the possibility $w_k^+ = w_j^-$, this implies $\Delta_{k-1} = \Delta_{j+1}$ again contradicting the hypothesis unless $k = j + 2$. But this would imply $\Delta_j = \Delta_{j+2}$ which is impossible, as one can check directly.

The above implies that all the Δ_k are different, but they are only finitely many, so the construction must eventually stop, and the only possibility to stop is when a fully colored simplex appears, whereby concluding the proof. □

We are now ready to prove our first fixed-point Lemma in the simple case where the domain is a simplex.

Theorem A.2.4 (Fixed Point Theorem for simplices) *Every continuous map $f\colon \Delta^n \to \Delta^n$ has a fixed point.*

PROOF. Let $x \in \Delta^n$ such that $f_i(x) \geq x_i$ for each $i \in \{1, \dots, n+1\}$, then

$$0 = 1 - 1 = \sum_{i=1}^{d+1} (f_i(x) - x_i), \tag{A.2.1}$$

which implies $f(x) = x$. It thus suffices to show that such a point exists. We argue by contradiction, assume that for every x there exists some i with

$f_i(x) < x_i$.
For each $k \in \mathbb{N}$, consider a triangulation T_k of Δ^n with simplices of size smaller than 2^{-k}. For each vertex v of T_k, we set $\ell(v) = \arg\max_i\{v_i - f_i(v)\}$. By our assumption, we have $v_{\ell(v)} > f_{\ell(v)}(v)$. If v lies on the face $\{x_j = 0\}$, then clearly $f_j(v) \geq 0 = v_j$, so $\ell(v) \neq j$. Thus, we have defined an s-coloring of T_k. It follows that there exists a simplex $\Delta_k \in T_k$ which is fully colored. Let $x_k \in \Delta_k =: \Delta(v_{k,1}, \dots v_{k,n+1})$, by compactness the sequence $\{x_k\}$ admits a convergent subsequence $\{x_{k_j}\}$. Let $\bar{x} = \lim_{j\to\infty} x_{k_j}$. It follows that $\bar{x} = \lim_{j\to\infty} v_{k_j,l}$, for each $l \in \{1, \dots, n+1\}$. Since the Δ_k are fully colored, for each i and j there exists $l_{j,i}$ such that $f(v_{k_j,l_{i,j}})_i < (v_{k_j,l_{i,j}})_i$. By the continuity of f, it follows

$$\bar{x}_i \leq f(\bar{x})_i$$

for each $i \in \{1, \dots, n+1\}$, hence the contradiciton. The lemma follows. □

A.2.2 Maps on finite-dimensional convex sets

To obtain a more general result, we need to recall a useful characterization of convex sets.

Lemma A.2.5 *Let $K \subset \mathbb{R}^n$ be a non-empty compact convex set with nonempty interior. Then K is homeomorphic to the standard n–simplex Δ^n.*

Proof. Choose $x_0 \in \mathrm{int}(K)$ and $z_0 = (\frac{1}{n+1}, \dots, \frac{1}{n+1}) \in \mathbb{R}^{n+1}$. Let R be a rotation that sends e_{d+1} into the vector $[n+1]^{-\frac{1}{2}}(1, \dots, 1)$. Consider the map $\Phi_0(x) = z_0 + R(x - x_0, 0)$ and let $\tilde{K} = \Phi_0(K)$. By construction, $\tilde{K}$ belongs to the same hyperplane containing Δ^n. For each $z \in \tilde{K}$, the half line $\{z_0 + t(z - z_0) \ : \ t \geq 0\}$ intersects the boundary ∂K at a unique point $a(z)$ and the boundary $\partial\Delta^n$ at a unique point $b(z)$. Define a continuous map $\phi_1 : \tilde{K} \to \Delta^n$ by

$$\phi_1(x) = z_0 + \frac{\|b(z)\|}{\|a(z)\|}(z - z_0).$$

Clearly, $\phi = \phi_1 \circ \phi_0$ is the wanted homeomorphism. □

Theorem A.2.6 (Brouwer Fixed Point Theorem) *For every non-empty compact convex set $K \subset \mathbb{R}^n$ and continuous map $f\colon K \to K$, f has a fixed point.*

Proof. By Lemma A.2.5, there exists a homeomorphism $\phi : K \to \Delta^n$. Define $F = \phi \circ f \circ \phi^{-1} : \Delta^n \to \Delta^n$. Theorem A.2.4 implies that there exist $\bar{x} \in \Delta^n$ such that $F(\bar{n}) = \bar{x}$. Hence, setting $x_* = \phi^{-1}(\bar{x})$ we have $f(x_*) = x_*$. □

A.2.3 Maps on compact convex sets

To conclude this survey, we show how Brouwer's result can be extended to the infinite-dimensional setting by an approximation procedure. Note that this result is less constructive than the previous ones, as it is based on a compactness argument.

Theorem A.2.7 (Schauder Fixed-Point Theorem) *Let $\mathcal{B}$ be a Banach space and $K \subset \mathcal{B}$ a nonempty, compact, convex subset. Let $f : K \to K$ be continuous. Then f has a fixed point.*

PROOF. Since K is compact, for each $\varepsilon > 0$ there exists a finite set $\{x_1, \ldots, x_N\} \subset \mathring{K}$ such that

$$K \subset \bigcup_{i=1}^{N} B_\varepsilon(x_i),$$

where $B_\varepsilon(x_i)$ denotes the open ball of radius ε around x_i. Let

$$K_\varepsilon := \operatorname{conv}\{x_1, \ldots, x_N\} \subset K$$

be the convex hull of the points $\{x_i\}$. Then K_ε is a compact, convex, and finite-dimensional set since it is contained in $\operatorname{span}\{x_1, \ldots, x_N\}$. Next, define

$$\phi_i(x) = \begin{cases} \varepsilon - \|x - x_i\| & \text{for} \|x - x_i\| \le \varepsilon \\ 0 & \text{otherwise.} \end{cases}$$

and

$$P_\varepsilon(x) = \left[\sum_{i=1}^{N} \phi_i(x)\right]^{-1} \sum_{i=1}^{N} \phi_i(x) x_i.$$

Note that $P_\varepsilon(\mathcal{B}) = K_\varepsilon$, P_ε is continuos and, for all $x \in K$

$$\|P_\varepsilon(x) - x\| = \left\| \left[\sum_{i=1}^{N} \phi_i(x)\right]^{-1} \sum_{i=1}^{N} \phi_i(x)(x_i - x) \right\| \le \varepsilon. \tag{A.2.2}$$

We can then define the continuous function

$$f_\varepsilon := P_\varepsilon \circ f : K_\varepsilon \to K_\varepsilon.$$

By Brouwer's fixed-point theorem, there exists

$$x_\varepsilon \in K_\varepsilon \quad \text{such that} \quad f_\varepsilon(x_\varepsilon) = x_\varepsilon.$$

Since K is compact, there exists a convergent subsequence $\{x_{\varepsilon_j}\}$, let x_* be the limit. Consequently, reacalling (A.2.2), we have

$$\|f(x_{\varepsilon_j}) - x_{\varepsilon_j}\| = \|f(x_{\varepsilon_j}) - f_{\varepsilon_j}(x_{\varepsilon_j})\| = \|f(x_{\varepsilon_j}) - P_{\varepsilon_j}(f(x_{\varepsilon_j}))\| \leq \varepsilon_j.$$

Taking the limit $j \to \infty$, by the continuity of f, we have the wanted fixed point $f(x_*) = x_*$. □

A.3 Hilbert metric and Birkhoff theorem

One may wonder if there are cases in which the fixed point provided by the Brower and Shauder theory is unique. In general, the answer is negative, but much more can be said for linear maps. In particular, we will see that the Banach fixed-point theorem can produce unexpected results if used with respect to an appropriate metric. We thus start with a short digression on projective metrics.
Projective metrics are widely used in geometry, not to mention the importance of their generalizations (e.g. Kobayashi metrics) for the study of complex manifolds [IK00]. It is quite surprising that they play a major rôle also in our situation, [Liv95].

Here we limit ourselves to a few words on the Hilbert metric, a quite important tool in hyperbolic geometry.

A.3.1 Projective metrics

Let $C \subset \mathbb{R}^n$ be a strictly convex compact set. For each two point $x, y \in C$ consider the line $\ell = \{\lambda x + (1-\lambda y) \mid \lambda \in \mathbb{R}\}$ passing through x and y. Let $\{u, v\} = \partial C \cap \ell$ and define[2]

$$\Theta(x, y) = \left| \ln \frac{\|x-u\|\|y-v\|}{\|x-v\|\|y-v\|} \right|$$

(the logarithm of the cross ratio). By remembering that the cross ratio is a projective invariant and looking at Figure A.2, it is easy to check that Θ is indeed a metric. Moreover, the distance of an inner point from the boundary is always infinite. One can also check that if the convex set is a disc, then the disc with the Hilbert metric is nothing but the Poincaré disc.

The objects that we will use in our subsequent discussion are not convex sets but rather convex cones, yet their projectivization is a convex set, and one can define the Hilbert metric on it (whereby obtaining a semi-metric for the original cone). It turns out that there exists a more algebraic way of defining such a metric, which is easier to use in our context. Moreover, there exists

[2] Remark that u, v can also be ∞.

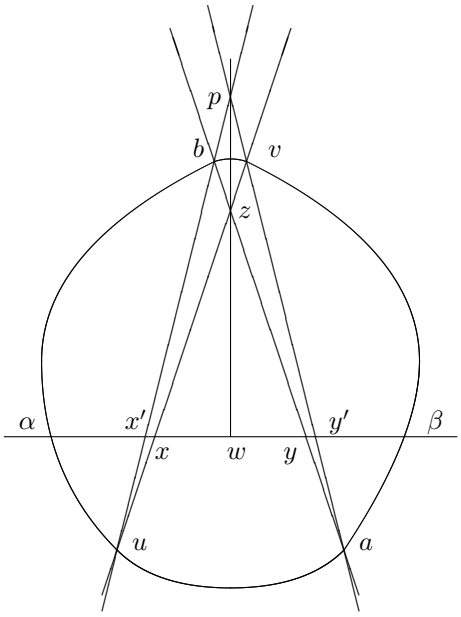

Figure A.2: Hilbert metric

a simple connection between vector spaces with a convex cone and vector lattices (in a vector lattice one can always consider the positive cone). This justifies the next digression in lattice theory.[3]

Consider a topological vector space $\mathbb{V}$ with a partial ordering "$\preceq$," that is a vector lattice.[4] We require the partial order to be "continuous," i.e. given $\{f_n\} \in \mathbb{V} \lim_{n\to\infty} f_n = f$, if $f_n \succeq g$ for each n, then $f \succeq g$. We call such vector lattices "integrally closed." [5]

We define the closed convex cone [6] $\mathcal{C} = \{f \in \mathbb{V} \mid f \neq 0,\ f \succeq 0\}$ (hereafter, the term "closed cone" $\mathcal{C}$ will mean that $\mathcal{C} \cup \{0\}$ is closed), and the equivalence

[3] For more details, see [Bir57], and [Nus88] for an overview of the field.

[4] We are assuming the partial order to be well-behaved with respect to the algebraic structure: for each $f, g \in \mathbb{V}$ $f \succeq g \iff f - g \succeq 0$; for each $f \in \mathbb{V}$, $\lambda \in \mathbb{R}^+ \backslash \{0\}$ $f \succeq 0 \implies \lambda f \succeq 0$; for each $f \in \mathbb{V}$ $f \succeq 0$ and $f \preceq 0$ imply $f = 0$ (antisymmetry of the order relation).

[5] To be precise, in the literature "integrally closed" is used in a weaker sense. First, $\mathbb{V}$ does not need a topology. Second, it suffices that for $\{\alpha_n\} \in \mathbb{R}$, $\alpha_n \to \alpha$; $f, g \in \mathbb{V}$, if $\alpha_n f \succeq g$, then $\alpha f \succeq g$. Here we will ignore these and other subtleties: our task is limited to a brief account of the results relevant to the present context.

[6] Here, by "cone," we mean any set such that, if f belongs to the set, then λf belongs to it as well, for each $\lambda > 0$.

relation "$\sim$": $f \sim g$ iff there exists $\lambda \in \mathbb{R}^+ \backslash \{0\}$ such that $f = \lambda g$. If we call $\widetilde{\mathcal{C}}$ the quotient of $\mathcal{C}$ with respect to $\sim$, then $\widetilde{\mathcal{C}}$ is a closed convex set. Conversely, given a closed convex cone $\mathcal{C} \subset \mathbb{V}$, enjoying the property $\mathcal{C} \cap -\mathcal{C} = \emptyset$, we can define an order relation by

$$f \preceq g \iff g - f \in \mathcal{C} \cup \{0\}.$$

Henceforth, each time that we specify a convex cone, we will assume the corresponding order relation and vice versa. The reader must therefore be advised that "$\preceq$" will mean different things in different contexts.

It is then possible to define a projective metric Θ (Hilbert metric),[7] in $\mathcal{C}$, by the construction:

$$\begin{aligned}
\alpha(f,\, g) &= \sup\{\lambda \in \mathbb{R}^+ \mid \lambda f \preceq g\} \\
\beta(f,\, g) &= \inf\{\mu \in \mathbb{R}^+ \mid g \preceq \mu f\} \\
\Theta(f,\, g) &= \log\left[\frac{\beta(f,\, g)}{\alpha(f,\, g)}\right]
\end{aligned}$$

where we take $\alpha = 0$ and $\beta = \infty$ if the corresponding sets are empty.

The relevance of the above metric in our context is due to the following Theorem by Garrett Birkhoff.

Theorem A.3.1 *Let $\mathbb{V}_1$, and $\mathbb{V}_2$ be two integrally closed vector lattices; $\mathcal{L} : \mathbb{V}_1 \to \mathbb{V}_2$ a linear map such that $\mathcal{L}(\mathcal{C}_1) \subset \mathcal{C}_2$, for two closed convex cones $\mathcal{C}_1 \subset \mathbb{V}_1$ and $\mathcal{C}_2 \subset \mathbb{V}_2$ with $\mathcal{C}_i \cap -\mathcal{C}_i = \emptyset$. Let Θ_i be the Hilbert metric corresponding to the cone $\mathcal{C}_i$. Setting $\Delta = \sup_{f,\, g \in \mathcal{L}(\mathcal{C}_1)} \Theta_2(f,\, g)$ we have*

$$\Theta_2(\mathcal{L}f,\, \mathcal{L}g) \leq \tanh\left(\frac{\Delta}{4}\right) \Theta_1(f,\, g) \qquad \forall f,\, g \in \mathcal{C}_1$$

($\tanh(\infty) \equiv 1$).

Proof. The proof is provided for the reader's convenience.

Let $f,\, g \in \mathcal{C}_1$, on the one hand if $\alpha = 0$ or $\beta = \infty$, then the inequality is obviously satisfied. On the other hand, if $\alpha \neq 0$ and $\beta \neq \infty$, then

$$\Theta_1(f,\, g) = \ln \frac{\beta}{\alpha}$$

where $\alpha f \preceq g$ and $\beta f \succeq g$, since $\mathbb{V}_1$ is integrally closed. Notice that $\alpha \geq 0$, and $\beta \geq 0$ since $f \succeq 0$, $g \succeq 0$. If $\Delta = \infty$, then the result follows from $\alpha \mathcal{L}f \preceq \mathcal{L}g$ and $\beta \mathcal{L}f \succeq \mathcal{L}g$. If $\Delta < \infty$, then, by hypothesis,

$$\Theta_2\left(\mathcal{L}(g - \alpha f),\, \mathcal{L}(\beta f - g)\right) \leq \Delta$$

[7] In fact, we define a semi–metric, since $f \sim g \Rightarrow \Theta(f,\, g) = 0$. The metric that we describe corresponds to the conventional Hilbert metric on $\widetilde{\mathcal{C}}$.

which means that there exist λ, $\mu \geq 0$ such that

$$\lambda \mathcal{L}(g - \alpha f) \preceq \mathcal{L}(\beta f - g)$$
$$\mu \mathcal{L}(g - \alpha f) \succeq \mathcal{L}(\beta f - g)$$

with $\ln \frac{\mu}{\lambda} \leq \Delta$. The previous inequalities imply

$$\frac{\beta + \lambda\alpha}{1 + \lambda} \mathcal{L}f \succeq \mathcal{L}g$$
$$\frac{\mu\alpha + \beta}{1 + \mu} \mathcal{L}f \preceq \mathcal{L}g.$$

Accordingly,

$$\begin{aligned}\Theta_2(\mathcal{L}f, \mathcal{L}g) &\leq \ln \frac{(\beta + \lambda\alpha)(1 + \mu)}{(1 + \lambda)(\mu\alpha + \beta)} = \ln \frac{e^{\Theta_1(f,g)} + \lambda}{e^{\Theta_1(f,g)} + \mu} - \ln \frac{1 + \lambda}{1 + \mu} \\ &= \int_0^{\Theta_1(f,g)} \frac{(\mu - \lambda)e^{\xi}}{(e^{\xi} + \lambda)(e^{\xi} + \mu)} d\xi \leq \Theta_1(f, g) \frac{1 - \frac{\lambda}{\mu}}{\left(1 + \sqrt{\frac{\lambda}{\mu}}\right)^2} \\ &\leq \tanh\left(\frac{\Delta}{4}\right) \Theta_1(f, g).\end{aligned}$$

□

Remark A.3.2 *If $\mathcal{L}(\mathcal{C}_1) \subset \mathcal{C}_2$, then it follows that $\Theta_2(\mathcal{L}f, \mathcal{L}g) \leq \Theta_1(f, g)$. However, a uniform rate of contraction depends on the diameter of the image being finite.*

In particular, if an operator maps a convex cone strictly inside itself (in the sense that the diameter of the image is finite), then it is a contraction in the Hilbert metric. This implies the existence of a "positive" eigenfunction (provided the cone is complete with respect to the Hilbert metric), and, with some additional work, the existence of a gap in the spectrum of $\mathcal{L}$ (see [Bir79] for details). The relevance of this theorem for the study of invariant measures and their ergodic properties is obvious.

It is natural to wonder about the strength of the Hilbert metric compared to other, more usual, metrics. While, in general, the answer depends on the cone, it is nevertheless possible to state an interesting result.

Lemma A.3.3 *Let $\|\cdot\|$ be a norm on the vector lattice $\mathbb{V}$, and suppose that, for each $f, g \in \mathbb{V}$,*

$$-f \preceq g \preceq f \Longrightarrow \|f\| \geq \|g\|.$$

Then, given $f, g \in \mathcal{C} \subset \mathbb{V}$ for which $\|f\| = \|g\|$,

$$\|f - g\| \leq \left(e^{\Theta(f,g)} - 1\right) \|f\|.$$

PROOF. We know that $\Theta(f, g) = \ln \frac{\beta}{\alpha}$, where $\alpha f \preceq g$, $\beta f \succeq g$. This implies that $-g \preceq 0 \preceq \alpha f \preceq g$, i.e. $\|g\| \geq \alpha \|f\|$, or $\alpha \leq 1$. In the same manner, it follows that $\beta \geq 1$. Hence,

$$\begin{aligned} g - f \preceq &(\beta - 1) f \preceq (\beta - \alpha) f \\ g - f \succeq &(\alpha - 1) f \succeq -(\beta - \alpha) f \end{aligned}$$

which implies

$$\|g - f\| \leq (\beta - \alpha)\|f\| \leq \frac{\beta - \alpha}{\alpha}\|f\| = \left(e^{\Theta(f,\, g)} - 1\right)\|f\|.$$

□

Many normed vector lattices satisfy the hypothesis of Lemma 1.3, e.g. Banach lattices.[8]

A.3.2 An application: quantitative Perron-Frobenius

Consider a matrix $L : \mathbb{R}^n \to \mathbb{R}^n$ of all strictly positive elements: $L_{ij} \geq \gamma > 0$. The Perron-Frobenius theorem states that there exists a unique eigenvector v^+ such that $v_i^+ > 0$, in addition, the corresponding eigenvalue λ is simple, maximal and positive. There are quite a few proofs of this theorem; one is based on Birkhoff's theorem. Consider the cone $\mathcal{C}^+ = \{v \in \mathbb{R}^2 \mid v_i \geq 0\}$, then obviously $L\mathcal{C}^+ \subset \mathcal{C}^+$. Moreover an explicit computation (see

Problem A.1 *shows that*

$$\Theta(v, w) = \ln \sup_{ij} \frac{v_i w_j}{v_j w_i}. \tag{A.3.3}$$

Then, setting $M = \max_{ij} L_{ij}$, it follows that

$$\Theta(Lv, Lw) \leq 2 \ln \frac{M}{\gamma} := \Delta < \infty.$$

We then have a contraction in the Hilbert metric, and the result follows from the usual fixed points theorems. Note that, since $\Theta(v, \lambda v) = 0$, for all $\lambda \in \mathbb{R}^+$, the fixed point $v_+ \in \mathbb{R}^n$ is only projective, that is $Lv_+ = \lambda v_+$ for some $\lambda \in \mathbb{R}$; in other words, we have an eigenvalue.

Remark that L^* satisfies the same conditions as L, thus there exists $w^+ \in \mathcal{C}^+$, $\mu \in \mathbb{R}^+$, such that $L^* w^+ = \mu w^+$. Next, define $\rho_1(v) = |\langle w^+, v\rangle|$ and

[8] A Banach lattice $\mathbb{V}$ is a vector lattice equipped with a norm satisfying the property $\| \, |f| \, \| = \|f\|$ for each $f \in \mathbb{V}$, where $|f|$ is the least upper bound of f and $-f$. For this definition to make sense it is necessary to require that $\mathbb{V}$ is "directed," i.e. any two elements have an upper bound.

$\rho_2(v) = \|v\|$. It is easy to check that there are two homogeneous forms of degree one adapted to the cone.

In addition, if $\rho_1(v) = \rho_2(v)$, then $\rho_1(L^n v) = \rho_1(L^n w)$. Hence, by Lemma A.3.3

$$\begin{aligned} \|L^n v - L^n w\| &\leq \left(e^{\Theta(L^n v, L^n w)} - 1\right) \min\{\|L^n v\|, \|L^n w\|\} \\ &\leq K\Lambda^n \min\{\|L^n v\|, \|L^n w\|\}, \end{aligned} \tag{A.3.4}$$

for some constant K depending only on v, w. The estimate A.3.4 means that all the vectors in the cone grow at the same rate. In fact, for all $v \in \text{int}\mathcal{C}$,

$$\|\lambda^{-n} L^n v - \lambda^{-n} L^n w\| \leq K\Lambda^n.$$

Hence, $\lim_{n\to\infty} \lambda^{-n} L^n v = v_+$.

Finally, consider $\mathbb{V}_1 = \{v \in \mathbb{V} \mid \langle w^+, v\rangle = 0\}$. Clearly $L\mathbb{V}_1 \subset \mathbb{V}_1$ and $\mathbb{V}_1 \oplus \text{span}\{v_+\} = \mathbb{V}$. Let $w \in \mathbb{V}_1$, clearly there exists $\alpha \in \mathbb{R}^+$ such that $\alpha v_+ + w \in \mathcal{C}$,[9] thus

$$\|L^n w\| \leq \|L^n(\alpha v_+ + w) - \alpha L^n v_+\| \leq L\Lambda^n \lambda^n.$$

This immediately implies that L restricted to the subspace $\mathbb{V}_1$ has spectral radius less than $\lambda\Lambda$. In other words, λ is the maximal eigenvalue; it is simple, and any other eigenvalue must be smaller than $\lambda\Lambda$. We have thus obtained an estimate of the spectral gap between the first and the second eigenvalue.

Notes

For more details on Hilbert metrics see [Bir79], and [Nus88] for an overview of the field.

[9] this is a special case of the general fact that any vector can be written as the linear combination of two vectors belonging to the cone.

Appendix B

Implicit function theorem (a quantitative version)

In this appendix we recall the implicit function Theorem. We provide an explicit proof because we use in the text a quantitative version of the theorem so it is important to keep track of the various constants.

B.1 The theorem

Let $n, m \in \mathbb{N}$ and $F \in \mathcal{C}^1(\mathbb{R}^{m+n}, \mathbb{R}^m)$ and let $(x_0, \lambda_0) \in \mathbb{R}^n \times \mathbb{R}^m$ such that $F(x_0, \lambda_0) = 0$. For each $\delta > 0$ let $V_\delta = \{(x, \lambda) \in \mathbb{R}^{n+m} \ : \ \|x - x_0\| \leq \delta, \|\lambda - \lambda_0\| \leq \delta\}$.

Theorem B.1.1 *Assume that $\partial_x F(x_0, \lambda_0)$ is invertible and choose $\delta > 0$ such that $\sup_{(x,\lambda)\in V_\delta} \|\mathbb{1} - [\partial_x F(x_0, \lambda_0)]^{-1} \partial_x F(x, \lambda)\| \leq \frac{1}{2}\}$. Let $B_\delta = \sup_{(x,\lambda)\in V_\delta} \|\partial_\lambda F(x, \lambda)\|$ and $M = \|\partial_x F(x_0, \lambda_0)^{-1}\|$. Set $\delta_1 = (2MB_\delta)^{-1}\delta$ and $\Lambda_{\delta_1} := \{\lambda \in \mathbb{R}^m \ : \ \|\lambda - \lambda\| < \delta_1\}$. Then there exists $g \in \mathcal{C}^1(\Lambda_{\delta_1}, \mathbb{R}^m)$ such that all the solutions of the equation $F(x, \lambda) = 0$ in the set $\{(x, \lambda) \in \mathcal{B}_1 \times \mathcal{B}_2 \ : \ \|\lambda - \lambda_0\| < \delta_1, \|x - x_0\| < \delta\}$ are given by $(g(\lambda), \lambda)$. In addition,*

$$\partial_\lambda g(\lambda) = -(\partial_x F(g(\lambda), \lambda))^{-1} \partial_\lambda F(g(\lambda), \lambda).$$

We will do the proof in several steps.

B.1.1 Existence of the solution

Let $A(x, \lambda) = \partial_x F(x, \lambda)$, $M = \|A(x_0, \lambda_0)^{-1}\|$.

We want to solve the equation $F(x, \lambda) = 0$, various approaches are possible. Here we will use a simplification of Newton method, made possible by the

fact that we already know a good approximation of the zero we are looking for. Let λ be such that $\|\lambda-\lambda_0\| < \delta_1 \leq \delta$. Consider $U_\delta = \{x \in \mathbb{R}^n \ : \ \|x-x_0\| \leq \delta\}$ and the function $\Theta_\lambda : U_\delta \to \mathbb{R}^n$ defined by[1]

$$\Theta_\lambda(x) = x - A(x_0, \lambda_0)^{-1} F(x, \lambda). \tag{B.1.1}$$

Problem B.1 *Prove that, for $x \in U(\lambda)$, $F(x, \lambda) = 0$ is equivalent to $x = \Theta_\lambda(x)$.*

Next,

$$\|\Theta_\lambda(x_0) - \Theta_{\lambda_0}(x_0)\| \leq M\|F(x_0, \lambda)\| \leq MB_\delta\delta_1.$$

In addition, $\|\partial_x\Theta_\lambda\| = \|\mathbb{1} - A(x_0, \lambda_0)^{-1}A(x, \lambda)\| \leq \frac{1}{2}$. Thus,

$$\|\Theta_\lambda(x) - x_0\| \leq \frac{1}{2}\|x - x_0\| + \|\Theta_\lambda(x_0) - x_0\| \leq \frac{1}{2}\|x - x_0\| + MB_\delta\delta_1 \leq \delta.$$

The existence of $x \in U_\delta$ such that $\Theta_\lambda(x) = x$ follows then by the standard fixed point Theorem A.1.1. We have so obtained a function $g : \{\lambda \ : \ \|\lambda - \lambda_0\| \leq \delta_1\} = \Lambda_{\delta_1} \to \mathbb{R}^n$ such that $F(g(\lambda), \lambda) = 0$. it remains the question of the regularity.

B.1.2 Lipschitz continuity and Differentiability

Let $\lambda, \lambda' \in \Lambda_{\delta_1}$. By (B.1.1)

$$\|g(\lambda) - g(\lambda')\| \leq \frac{1}{2}\|g(\lambda) - g(\lambda')\| + MB_\delta|\lambda - \lambda'|$$

This yields the Lipschitz continuity of the function g. To obtain the differentiability we note that, by the differentiability of F and the above Lipschitz continuity of g, for $h \in \mathbb{R}^m$ small enough,

$$\|F(g(\lambda + h), \lambda + h) - F(g(\lambda), \lambda) + \partial_x F[g(\lambda + h) - g(\lambda)] + \partial_\lambda F h\| = o(\|h\|).$$

Since $F(g(\lambda + h), \lambda + h) = F(g(\lambda), \lambda) = 0$, we have that

$$\lim_{h\to 0} \|h\|^{-1}\|g(\lambda + h) - g(\lambda) + [\partial_x F]^{-1}\partial_\lambda F h\| = 0$$

which concludes the proof of the Theorem, the continuity of the derivative being obvious be the obtained explicit formula.

[1] The Newton method would consist in finding a fixed point for the function $x - A(x, \lambda)^{-1}F(x, \lambda)$. This gives a much faster convergence and hence is preferable in applications, yet here it would make the estimates a bit more complicated.

B.2 Generalization

First of all note that the above theorem implies the inverse function theorem. Indeed if $f : \mathbb{R}^n \to \mathbb{R}^n$ is a function such that $\partial_x f$ is invertible at some point x_0, then one can consider the function $F(x, y) = f(x) - y$. Applying the implicit function theorem to the equation $F(x, y) = 0$ it follows that $y = f(x)$ are the only solution, hence the function is locally invertible.

The above theorem can be generalized in several ways.

Problem B.2 *Show that if F in Theorem B.1.1 is $\mathcal{C}^r$, then also g is $\mathcal{C}^r$.*

Problem B.3 *Verify that if $\mathcal{B}_1, \mathcal{B}_2$ are two Banach spaces and in Theorem B.1.1 we have $\mathcal{B}_1$ instead of $\mathbb{R}^n$ and $\mathcal{B}_2$ instead of $\mathbb{R}^m$ the Theorem remains true and the proof remains exactly the same.*

As I mentioned the statement of Theorem B.1.1 is suitable for quantitative applications.

Problem B.4 *Suppose that in Theorem B.1.1 we have $F \in \mathcal{C}^2$, then show that we can chose*

$$\delta = [2\|D\partial_x F\|_\infty]^{-1} .$$

Appendix C

Perturbation Theory (a super-fast introduction)

The following is really super condensate (although self-consistent). If you want more details see [RS80, Kat66] in which you probably can find more than you are looking for.

C.1 Bounded operators

In the following we will consider only *separable* Banach spaces, i.e. Banach spaces that have a countable dense set.[1]

Given a Banch space $\mathcal{B}$ we can consider the set $L(\mathcal{B},\mathcal{B})$ of the linear bounded operators from $\mathcal{B}$ to itself. We can then introduce the norm $\|B\| = \sup_{\|v\|\leq 1} \|Bv\|$.

Problem C.1 *Show that $(L(\mathcal{B},\mathcal{B}), \|\cdot\|)$ is a Banach space. That is that $\|\cdot\|$ is really a norm and that the space is complete with respect to such a norm.*

Problem C.2 *Show that the $n \times n$ matrices form a Banach Algebra.*[2]

Problem C.3 *Show that $L(\mathcal{B},\mathcal{B})$ form a Banach algebra.*[3]

[1] Recall that a Banach space is a complete normed vector space (in the following we will consider vector spaces on the field of complex numbers), that is a normed vector space in which all the Cauchy sequences have a limit in the space. Again, if you are uncomfortable with Banach spaces, in the following read $\mathbb{R}^d$ instead of $\mathcal{B}$ and matrices instead of operators, but be aware that we have to develop the theory without the use of the determinant that, in general, is not defined for operators on Banach spaces.

[2] A Banach Algebra $\mathcal{A}$ is a Banach space where the multiplication between elements is defined with the usual properties of an algebra and, in addition, for each $a, b \in \mathcal{A}$ holds $\|ab\| \leq \|a\| \cdot \|b\|$.

[3] The multiplication is given by the composition.

To each $A \in L(\mathcal{B}, \mathcal{B})$ are associated two important subspaces: the range $R(A) = \{v \in \mathcal{B} \ : \ \exists \ w \in \mathcal{B} \text{ such that } v = Aw\}$ and the kernel $N(A) = \{v \in \mathcal{B} \ : \ Av = 0\}$.

Problem C.4 *Prove, for each $A \in L(\mathcal{B}, \mathcal{B})$, that $N(A)$ is a closed linear subspaces of $\mathcal{B}$. Show that this is not necessarily the case for $R(A)$ if $\mathcal{B}$ is not finite dimensional.*

A very special, but very important, class of operators is the set of projectors.

Definition C.1.1 *An operator $\Pi \in L(\mathcal{B}, \mathcal{B})$ is called a* projector *iff $\Pi^2 = \Pi$.*

Note that if Π is a projector, so is $\mathbb{1} - \Pi$. We have the following interesting fact.

Lemma C.1.2 *If $\Pi \in L(\mathcal{B}, \mathcal{B})$ is a projector, then $N(\Pi) \oplus R(\Pi) = \mathcal{B}$.*

Proof. If $v \in \mathcal{B}$, then $v = \Pi v + (\mathbb{1} - \Pi)v$. Notice that $R(\mathbb{1} - \Pi) = N(\Pi)$ and $R(\Pi) = N(\mathbb{1} - \Pi)$. Finally, if $v \in N(\Pi) \cap R(\Pi)$, then $v = 0$, which concludes the proof. □

Another, more general, very important class of operators are the compact ones.

Definition C.1.3 *An operator $K \in L(\mathcal{B}, \mathcal{B})$ is called* compact *iff for any bounded set B the closure of $K(B)$ is compact.*

Remark C.1.4 *Note that not all the linear operator on a Banach space are bounded. For example consider the derivative acting on $\mathcal{C}^1((0,1), \mathbb{R})$.*

C.2 Functional calculus

First of all recall that all the Riemannian theory of integration works verbatim for function $f \in \mathcal{C}^0(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is a Banach space. We can thus talk of integrals of the type $\int_a^b f(t)dt$.[4] Next, we can talk of *analytic functions* for functions in $\mathcal{C}^0(\mathbb{C}, \mathcal{B})$: a function is analytic in an open region $U \subset \mathbb{C}$ iff at each point $z_0 \in U$ there exists a neighborhood $B \ni z_0$ and elements $\{a_n\} \subset \mathcal{B}$ such that

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n \quad \forall z \in B. \tag{C.2.1}$$

[4] This is special case of the so called Bochner integral [Yos95].

Problem C.5 *Show that if $f \in \mathcal{C}^0(\mathbb{C}, \mathcal{B})$ is analytic in $U \subset \mathbb{C}$, then given any smooth closed curve γ, contained in a sufficiently small disk in U, holds*[5]

$$\int_\gamma f(z)dz = 0 \tag{C.2.2}$$

Then show that the same hold for any piecewise smooth closed curve with interior contained in U, provided U is simply connected.

Problem C.6 *Show that if $f \in \mathcal{C}^0(\mathbb{C}, \mathcal{B})$ is analytic in a simply connected $U \subset \mathbb{C}$, then given any smooth closed curve γ, with interior contained contained in U and having in its interior a point z, hods the formula*

$$f(z) = \frac{1}{2\pi i}\int_\gamma (\xi - z)^{-1} f(\xi)d\xi. \tag{C.2.3}$$

Problem C.7 *Show that if $f \in \mathcal{C}^0(\mathbb{C}, \mathcal{B})$ satisfies (C.2.3) for each smooth closed curve in a simply connected open set U, then f is analytic in U.*

C.3 Spectrum and resolvent

Given $A \in L(\mathcal{B}, \mathcal{B})$ we define the *resolvent*, called $\rho(A)$, as the set of the $z \in \mathbb{C}$ such that $(z\mathbb{1} - A)$ is invertible and the inverse belongs to $L(\mathcal{B}, \mathcal{B})$. The *spectrum* of A, called $\sigma(A)$ is the complement of $\rho(A)$ in $\mathbb{C}$.

Problem C.8 *Prove that, for each Banach space $\mathcal{B}$ and operator $A \in L(\mathcal{B}, \mathcal{B})$, if $z \in \rho(A)$, then there exists a neighborhood U of z such that $(z\mathbb{1} - A)^{-1}$ is analytic in U.*

From the above exercise follows that $\rho(A)$ is open, hence $\sigma(A)$ is closed.

Problem C.9 *Show that, for each $A \in L(\mathcal{B}, \mathcal{B})$, $\sigma(A) \neq \emptyset$.*

Problem C.10 *Show that if $\Pi \in L(\mathcal{B}, \mathcal{B})$ is a projector, then $\sigma(\Pi) = \{0, 1\}$.*

Up to now the theory for operators seems very similar to the one for matrices. Yet, the spectrum for matrices is always given by a finite number of points while the situation for operators can be very dfferenct.

[5] Of course, by $\int_\gamma f(z)dz$ we mean that we have to consider any smooth parametrization $g : [a, b] \to \mathbb{C}$ of γ, $g(a) = g(b)$, and then $\int_\gamma f(z)dz := \int_a^b f \circ g(t)g'(t)dt$. Show that the definition does not depend on the parametrization and that one can use piecewise smooth parametrizations as well.

Problem C.11 *Consider the operator* $\mathcal{L} : \mathcal{C}^0([0,1],\mathbb{C}) \to \mathcal{C}^0([0,1],\mathbb{C})$ *defined by*

$$(\mathcal{L}f)(x) = \frac{1}{2}f(x/2) + \frac{1}{2}f(x/2+1/2).$$

Show that $\sigma(\mathcal{L}) = \{z \in \mathbb{C} \ : \ |z| \leq 1\}$.

Problem C.12 *Show that, if* $A \in L(\mathcal{B},\mathcal{B})$ *and* p *is any polynomial, then for each* $n \in \mathbb{N}$ *and smooth curve* $\gamma \subset \mathbb{C}$*, with* $\sigma(A)$ *in its interior,*

$$p(A) = \frac{1}{2\pi i}\int_\gamma p(z)(z\mathbb{1} - A)^{-1}dz.$$

Problem C.13 *Show that, for each* $A \in L(\mathcal{B},\mathcal{B})$ *the limit*

$$r(A) = \lim_{n\to\infty} \|A^n\|^{\frac{1}{n}}$$

exists.

The above limit is called the *spectral radius* of A. A useful fact concerning the spectral radius is the following.

Lemma C.3.1 *For each* $A, B \in L(\mathcal{B},\mathcal{B})$ *and* $k \in \mathbb{N}$*, we have*

$$r(AB) = r(BA) \qquad r(A^k) = r(A)^k$$

PROOF. Using Problem C.13 yields

$$\begin{aligned} r(AB) &= \lim_{n\to\infty} \|(AB)^n\|^{\frac{1}{n}} = \lim_{n\to\infty} \|A(BA)^{n-1}B\|^{\frac{1}{n}} \\ &\leq \lim_{n\to\infty} \|A\|^{\frac{1}{n}}\|(BA)^{n-1}\|^{\frac{1}{n}}\|B\|^{\frac{1}{n}} = r(BA). \end{aligned}$$

By the same argument, exchanging A and B, we obtain $r(AB) = r(BA)$. Next,

$$r(A^k) = \lim_{n\to\infty} \|A^{kn}\|^{\frac{k}{kn}} = r(A)^k.$$

□

Lemma C.3.2 *For each* $A \in L(\mathcal{B},\mathcal{B})$ *we have* $\sup_{z\in\sigma(A)} |z| = r(A)$.

PROOF. Since we can write

$$(z\mathbb{1} - A)^{-1} = z^{-1}(\mathbb{1} - z^{-1}A)^{-1} = z^{-1}\sum_{n=0}^{\infty} z^{-n}A^n,$$

and since the series converges if it converges in norm, from the usual criteria for the convergence of a series follows $\sup_{z\in\sigma(A)} |z| \leq r(A)$. Suppose now that the inequality is strict. That is, there exists $0 < \eta < r(A)$ and a curve $\gamma \subset \{z \in \mathcal{C} \ : \ |z| \leq \eta\}$ which contains $\sigma(A)$ in its interior. Then applying Problem C.12 yields $\|A^n\| \leq C\eta^n$, which contradicts $\eta < r(A)$. □

Note that if $f(z) = \sum_{n=0}^{\infty} f_n z^n$ is an analytic function in all $\mathbb{C}$ (entire), then we can define

$$f(A) = \sum_{n=0}^{\infty} f_n A^n.$$

Problem C.14 *Show that, if $A \in L(\mathcal{B}, \mathcal{B})$ and f is an entire function, then for each smooth curve $\gamma \subset \mathbb{C}$, with $\sigma(A)$ in its interior,*

$$f(A) = \frac{1}{2\pi i} \int_{\gamma} f(z)(z\mathbb{1} - A)^{-1} dz.$$

In view of the above fact, the following definition is natural:

Definition C.3.3 *For each $A \in L(\mathcal{B}, \mathcal{B})$, f analytic in a region U containing $\sigma(A)$, then for each smooth curve $\gamma \subset U$, with $\sigma(A)$ in its interior, define*

$$f(A) = \frac{1}{2\pi i} \int_{\gamma} f(z)(z\mathbb{1} - A)^{-1} dz. \tag{C.3.4}$$

Problem C.15 *Show that the above definition does not depend on the curve γ.*

Problem C.16 *For each $A \in L(\mathcal{B}, \mathcal{B})$ and functions f, g analytic on a domain $D \supset \sigma(A)$, show that $f(A) + g(A) = (f + g)(A)$ and $f(A)g(A) = (f \cdot g)(A)$.*

Problem C.17 *In the hypotheses of the Definition C.3.3 show that $f(\sigma(A)) = \sigma(f(A))$ and $[f(A), A] = 0$.*

Problem C.18 *Consider $f : \mathbb{C} \to \mathbb{C}$ entire and $A \in L(\mathcal{B}, \mathcal{B})$. Suppose that $\{z \in \mathbb{C} \,:\, f(z) = 0\} \cap \sigma(A) = \emptyset$. Show that $f(A)$ is invertible and $f(A)^{-1} = f^{-1}(A)$.*

Problem C.19 *Let $A \in L(\mathcal{B}, \mathcal{B})$. Suppose there exists a semi-line ℓ, starting from the origin, such that $\ell \cap \sigma(A) = \emptyset$ and that $0 \notin \sigma(A)$. Prove that it is possible to define an operator $\ln A$ such that $e^{\ln A} = A$.*

Remark C.3.4 *Note that not all the interesting functions can be constructed in such a way. In fact, $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ is such that $A^2 = -\mathbb{1}$, thus it can be interpreted as a square rooth of $-\mathbb{1}$ but it cannot be obtained directly by a formula of the type (C.3.4).*

Problem C.20 *Suppose that $A \in L(\mathcal{B}, \mathcal{B})$ and $\sigma(A) = B \cup C$, $B \cap C = \emptyset$, suppose that the smooth closed curve $\gamma \subset \rho(A)$ contains B, but not C, in its interior, prove that*

$$P_B := \frac{1}{2\pi i} \int_{\gamma} (z\mathbb{1} - A)^{-1} dz \tag{C.3.5}$$

is a projector that does not depend on γ.

Note that by Problem C.17 follows that $P_B A = A P_B$. Hence, $AR(P_B) \subset R(P_B)$ and $AN(P_B) \subset N(P_B)$. Since, by Lemma C.1.2, $\mathcal{B} = R(P_B) \oplus N(P_B)$ we have obtained an invariant decomposition for A.

Problem C.21 *In the hypotheses of Problem C.20, prove that* $A = P_B A P_B + (\mathbb{1} - P_B) A (\mathbb{1} - P_B)$.

Problem C.22 *In the hypotheses of Problem C.20, prove that, calling* A_1 *the resriction of* A *to* $R(P_B)$ *we have* $\sigma(A_1) = B$. *Moreover, if* $\dim(R(P_B)) = D < \infty$, *then the cardinality of* B *is less or equal* D.

C.4 Perturbations

Let us consider $A, B \in L(\mathcal{B}, \mathcal{B})$ and the family of operators $A_\nu := A + \nu B$.

Lemma C.4.1 *For each* $\delta > 0$ *there exists* $\nu_\delta \in \mathbb{R}$ *such that, for all* $|\nu| \leq \nu_\delta$, $\rho(A_\nu) \supset \{z \in \mathbb{C} \,:\, d(z, \sigma(A)) > \delta\}$.

Proof. Let $d(z, \sigma(A)) > \delta$, then

$$(z\mathbb{1} - A_\nu) = (z\mathbb{1} - A)\left[\mathbb{1} - \nu(z\mathbb{1} - A)^{-1} B\right] \tag{C.4.6}$$

Now $\|(z\mathbb{1} - A)^{-1} B\|$ is a continuous function in z outside $\sigma(A)$, moreover it is bounded outside a ball of large enough radius, hence there exists $M_\delta > 0$ such that $\sum_{d(z,\sigma(A))>\delta} \|(z\mathbb{1} - A)^{-1} B\| \leq M_\delta$. Choosing $\nu_\delta = (2M_\delta)^{-1}$ yields the result. □

Suppose that $\bar{z} \in \mathbb{C}$ is an isolated point of $\sigma(A)$, that is there exists $\delta > 0$ such that $\{z \in \mathbb{C} \,:\, |z - \bar{z}| \leq \delta\} \cap (\sigma(A) \setminus \{\bar{z}\}) = \emptyset$, then the above Lemma shows that, for ν small enough, $\{z \in \mathbb{C} \,:\, |z - \bar{z}| \leq \delta\}$ still contains an isolated part of the spectrum of $\sigma(A_\nu)$, let us call it B_ν, clearly $B_0 = \{\bar{z}\}$.

Problem C.23 *Let* P_{B_ν} *be defined as in Problem C.20. Prove that, for* ν *small enough, it is an analytic function of* ν.

Problem C.24 *If* P, Q *are two projectors and* $\|P - Q\| < 1$, *then* $\dim(R(P)) = \dim(R(Q))$.

The above two exercises imply that the dimension of the eigenspace $R(P_{B_\nu})$ is constant.

Next, we consider the case in which B_0 consist of one point and $\dim(R(P_{B_0})) = 1$, it follows that also B_ν must consist of only one point, let us set $P_\nu := P_{B_\nu}$.

Lemma C.4.2 *If* $\dim(R(P_0)) = 1$, *then* A_ν *has a unique eigenvalue* z_ν *in a neighborhood of* $\bar{z}$, $z_0 = \bar{z}$. *In addition* z_ν *is an analytic function of* ν.

PROOF. From the previous exercises it follows that P_ν is a rank one operator which depend analytically on ν. In addition, since P_ν is a rank one projector it must have the form $P_\nu w = v_\nu \ell_\nu(w)$, where $\ell_\nu \in \mathcal{B}'$.[6] Then $z_\nu P_\nu = P_\nu A_\nu P_\nu$. Next, setting $a(\nu) := \ell_0(P_\nu v_0) = \ell_\nu(v_0)\ell_0(v_\nu)$, we have that a is analytic and $a(0) = 1$. Thus $a \neq 0$ in a neighborhood of zero and $z_\nu = a(\nu)^{-1}\ell_0(P_\nu A_\nu P_\nu v_0)$ is analytic in such a neighborhood. □

Problem C.25 *If* $\dim(R(P_0)) = 1$*, then there exists* $h_\nu \in \mathcal{B}$ *and* $\ell_\nu \in \mathcal{B}'$ *such that* $P_\nu f = h_\nu \ell_\nu(f)$ *for each* $f \in \mathcal{B}$*. Prove that* h_ν, ℓ_ν *can be chosen to be analytic functions of* ν*.*

Hence in the case of $A \in L(\mathcal{B}, \mathcal{B})$ with an isolated simple[7] eigenvalue $\bar{z}$ we have that the corresponding eigenvalue z_ν of $A_\nu = A + \nu B$, $B \in L(\mathcal{B}, \mathcal{B})$, for ν small enough, depend smoothly from ν. In addition, using the notation of the previous Lemma, we can easily compute the derivative: differentiating $A_\nu v_\nu = z_\nu v_\nu$ with respect to ν and then setting $\nu = 0$, yields

$$Bv + Av_0' = z_0' v + \bar{z} v_0'.$$

But, for all $w \in \mathcal{B}$, $Pw = v\ell(w)$, with $\ell(Aw) = \bar{z}\ell(w)$ and $\ell(v) = 1$, thus applying ℓ to both sides of the above equation yields

$$z_0' = \ell(Bv).$$

Problem C.26 *Compute* v_0'*.*

Problem C.27 *What does it happen if the eigenspace associated to* $\bar{z}$ *is finite dimensional, but with dimension strictly larger than one?*

Hints to solving the Problems

C.1. The triangle inequality follows trivially from the triangle inequality of the norm of $\mathcal{B}$. To verify the completeness suppose that $\{B_n\}$ is a Cauchy sequence in $L(\mathcal{B}, \mathcal{B})$. Then, for each $v \in \mathcal{B}$, $\{B_n v\}$ is a Cauchy sequence in $\mathcal{B}$, hence it has a limit, call it $B(v)$. We have so defined a function from $\mathcal{B}$ to teself. Show that such a function is linear and bounded, hence it defines an element of $L(\mathcal{B}, \mathcal{B})$, which can easily be verified to be the limit of $\{B_n\}$.

C.2. Use the norm $\|A\| = \sup_{v \in \mathbb{R}^n} \frac{\|Av\|}{\|v\|}$.

[6] By $\mathcal{B}'$, the dual space, we mean the set of bounded linear functionals on $\mathcal{B}$. Verify that is a Banach space with the norm $\|\ell\| = \sum_{w \in \mathcal{B}} \frac{|\ell(w)|}{\|w\|}$.

[7] That is with the associated eigenprojector of rank one.

C.3. Use the same norm as in Problem C.2.

C.4. The first part is trivial. For the second one can consider the vector space $\ell^2 = \{x \in \mathbb{R}^{\mathbb{N}} \ : \ \sum_{i=0}^{\infty} x_i^2 < \infty\}$. Equipped with the norm $\|x\| = \sqrt{\sum_{i=0}^{\infty} x_i^2}$ it is a Banach (actually Hilbert) space. Consider now the vectors $e_i \in \ell^2$ defined by $(e_i) = \delta_{ik}$ and the operator $(Ax)_k = \frac{1}{k} x_k$. Then $R(A) = \{x \in \ell^2 \ : \ \sum_{k=0}^{\infty} k^2 x_k^2 < \infty\}$, which is dense in ℓ^2 but strictly smaller.

C.5. Check that the same argument used in the well-known case $\mathcal{B} = \mathbb{C}$ works also here.

C.6. Check that the same argument used in the well-known case $\mathcal{B} = \mathbb{C}$ works also here.

C.7. Check that the same argument used in the well-known case $\mathcal{B} = \mathbb{C}$ works also here.

C.8. Note that

$$(\zeta \mathbb{1} - A) = (z\mathbb{1} - A - (z-\zeta)\mathbb{1}) = (z\mathbb{1} - A)\left[\mathbb{1} - (z-\zeta)(z\mathbb{1} - A)^{-1}\right]$$

and that if $\|(z-\zeta)(z\mathbb{1} - A)^{-1}\| < 1$ then the inverse of $\mathbb{1} - (z-\zeta)(z\mathbb{1} - A)^{-1}$ is given by $\sum_{n=0}^{\infty}(z-\zeta)^n[(z\mathbb{1} - A)^{-1}]^n$ (the Neumann series–which really is just the geometric series).

C.9. If $\sigma(A) = \emptyset$, then $(z\mathbb{1} - A)^{-1}$ is an entire function, then the Neumann series shows that $(z\mathbb{1} - A)^{-1} = z^{-1}(\mathbb{1} - z^{-1}A)^{-1}$ goes to zero for large z, and then (C.2.3) shows that $(z\mathbb{1} - A)^{-1} = 0$ which is impossible.

C.10. Verify that $(z\mathbb{1} - \Pi)^{-1} = z^{-1}\left[\mathbb{1} - (z-1)^{-1}\Pi\right]$.

C.11. The idea is to look for eigenvalues by using Fourier series. Let $f = \sum_{k\in\mathbb{Z}} f_k e^{2\pi i k x}$ and consider the equation $\mathcal{L}f = zf$,

$$\sum_{k\in\mathbb{Z}} f_k \frac{1}{2}\left\{e^{\pi i k x} + e^{\pi i k x + \pi i k}\right\} = z\sum_{k\in\mathbb{Z}} f_k e^{2\pi i k x}.$$

Let us then restrict to the case in which $f_{2k+1} = 0$, then

$$\sum_{k\in\mathbb{Z}} f_{2k} e^{2\pi i k x} = z\sum_{k\in\mathbb{Z}} f_k e^{2\pi i k x}.$$

Thus we have a solution provided $f_{2k} = z f_k$, such conditions are satisfied by any sequence of the type

$$f_k = \begin{cases} z^j & \text{if } k = 2^j m, j \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}$$

for $m \in \mathbb{N}$. It remains to verify that $\sum_{j=0}^{\infty} z^j e^{2\pi i 2^j x}$ belong to $\mathcal{C}^0$. This is the case if the series is uniformly convergent, which happens for $|z| < 1$. Thus all the points in $\{z \in \mathbb{C} \,:\, |z| < 1\}$ are point spectrum of infinite multiplicity. Since the spectrum is closed, the statement of the Problem follows.

C.12. First note that

$$\frac{1}{2\pi i}\int_\gamma (z\mathbb{1} - A)^{-1} dz = \frac{1}{2\pi i}\int_\gamma z^{-1}(\mathbb{1} - z^{-1}A)^{-1} dz.$$

By analiticity we can choose $\gamma = \{Re^{i\theta}\}$ for $R > \|A\|$, hence

$$\begin{aligned}\frac{1}{2\pi i}\int_\gamma (z\mathbb{1} - A)^{-1} dz &= \frac{1}{2\pi}\int_0^{2\pi} (\mathbb{1} - R^{-1}e^{-i\theta}A)^{-1} d\theta \\ &= \frac{1}{2\pi}\int_0^{2\pi} \sum_{n=0}^{\infty} R^{-n} e^{-in\theta} A^n d\theta = \mathbb{1}.\end{aligned}$$

Next, let $p(z) = z^n$, $n \in \mathbb{N}$, then

$$\begin{aligned}\frac{1}{2\pi i}\int_\gamma z^n (z\mathbb{1} - A)^{-1} dz &= A^n + \frac{1}{2\pi i}\int_\gamma (z^n - A^n)(z\mathbb{1} - A)^{-1} dz \\ &= A^n + \sum_{k=0}^{n-1} \frac{1}{2\pi i}\int_\gamma z^k A^{n-k} dz = A^n.\end{aligned}$$

The statement for general polynomials follows trivially.

C.13. Let $\alpha = \liminf_{n\to\infty} \|A^n\|^{\frac{1}{n}}$. Then for each $\varepsilon > 0$ exists $n_\varepsilon \in \mathbb{N}$ such that

$$\|A^{n_\varepsilon}\| \le (\alpha + \varepsilon)^{n_\varepsilon}.$$

Then, for each $n \in \mathbb{N}$ we can write $n = m + kn_\varepsilon$, with $m < n_\varepsilon$. Consequently,

$$\|A^n\|^{\frac{1}{n}} \le \left[\|A^m\|\|A^{n_\varepsilon}\|^k\right]^{\frac{1}{m+kn_\varepsilon}} \le \|A^m\|^{\frac{1}{m+kn_\varepsilon}} (\alpha + \varepsilon)^{\frac{kn_\varepsilon}{m+kn_\varepsilon}},$$

which implies $\limsup_{n\to\infty} \|A^n\|^{\frac{1}{n}} \le \alpha + \varepsilon$. The claim follows from the arbitrariness of ε.

C.14. Approximate by polynomials.

C.17. For $z \notin f(\sigma(A))$ it is well defined

$$K(z) := \frac{1}{2\pi i}\int_\gamma (z - f(\zeta))^{-1} (\zeta\mathbb{1} - A)^{-1}\, d\zeta,$$

with γ containing $\sigma(A)$ in its interior. By direct computation, using definition C.3.3, one can verify that $(z\mathbb{1} - f(A))K(z) = \mathbb{1}$, thus $\sigma(f(A)) \subset f(\sigma(A))$. On the other hand if, if f is not constant, then for each $z \in \mathcal{C}$ $f(z) - f(\xi) = (z-\xi)g(\xi)$. Hence, applying Definition C.3.3 and Problem C.16 it follows $f(z)\mathbb{1} - f(A) = (z-A)g(A)$ which shows that if $z \in \sigma(A)$, then $f(z) \in \sigma(A)$ (otherwise $(z - A)\left[g(A)(f(z)\mathbb{1} - f(A))^{-1}\right] = \mathbb{1}$).

C.19. Since one can define the logarithm on $\mathbb{C}\backslash(\ell \cap \{0\})$, one can use Definition C.3.3 to define $\ln A$. It suffices to prove that if $f : U \to \mathcal{C}$ and $g : V \to \mathcal{C}$, with $\sigma(A) \subset U$, $f(U) \subset V$, then $g(f(A)) = g \circ f(A)$. Whereby showing that the definition C.3.3 is a reasonable one. Indeed, rememebring Problems C.17, C.18,

$$\begin{aligned} g(f(A)) &= \frac{1}{2\pi i}\int_\gamma g(z)(z\mathbb{1} - f(A))^{-1}dz \\ &= \frac{1}{(2\pi i)^2}\int_{\gamma_1}\int_\gamma \frac{g(z)}{z - f(\xi)}(\xi\mathbb{1} - A)^{-1}dzd\xi \\ &= \frac{1}{2\pi i}\int_{\gamma_1} g(f(\xi))(\xi\mathbb{1} - A)^{-1}d\xi = f \circ g(A). \end{aligned}$$

From this imediately follows $e^{\ln A} = A$.

C.20. The non dependence on γ is obvious. A projector is characterized by the property $P^2 = P$. Thus

$$\begin{aligned} P_B^2 :=& \frac{1}{(2\pi i)^2}\int_{\gamma_1}\int_{\gamma_2}(z\mathbb{1} - A)^{-1}(\zeta\mathbb{1} - A)^{-1}dzd\zeta \\ =& \frac{1}{(2\pi i)^2}\int_{\gamma_1} dz\int_{\gamma_2} d\zeta(z-\zeta)^{-1}\left[(z\mathbb{1} - A)^{-1} - (\zeta\mathbb{1} - A)^{-1}\right]. \end{aligned}$$

If we have chosen γ_1 in the interior of γ_2, then $(z - \zeta)^{-1}(\zeta\mathbb{1} - A)^{-1}$ is analytic in the interior of γ_1, hence the corresponding integral gives zero. The other integral gives P_B, as announced.

C.21. Use the decomposition $\mathbb{1} = P_B + (\mathbb{1} - P_B)$, the fact that $P_B, (\mathbb{1} - P_B)$ are projectors and that they commute with A.

C.22. Since

$$(z\mathbb{1} - A)^{-1} = P_B(z\mathbb{1} - A)^{-1}P_B + (\mathbb{1} - P_B)(z\mathbb{1} - A)^{-1}(\mathbb{1} - P_B).$$

Calling A_1 the restrction of A to $R(P_B)$ and A_2 the restriction to $N(P_B)$, we have $\sigma(A) \subset \sigma(A_1) \cup \sigma(A_2)$. Next, for $z \notin B$, define the operator

$$K(z) := \frac{1}{2\pi i}\int_\gamma (z - \xi)^{-1}(\xi\mathbb{1} - A)^{-1}\, d\xi,$$

where γ contains B, but no other part of the spectrum nor z, in its interior. Then

$$(z\mathbb{1}-A)K(z)=\frac{1}{2\pi i}\int_\gamma\left[(z-\xi)+(\xi-A)\right](z-\xi)^{-1}(\xi\mathbb{1}-A)^{-1}\,d\xi=P_B.$$

Restricting the above equality to $R(P_B)$ we have that $\sigma(A_1)\subset B$. Analogously $\sigma(A_1)\subset C$, hence it must be $\sigma(A_1)=B$ and $\sigma(A_1)=C$.

The second property follows from the fact that P_BAP_B, when restricted to the space $R(P_B)$ is described by a $D\times D$ matrix A_B and the equation $\det(z\mathbb{1}-A_B)=0$ is a polynomial of degree D in z and hence has exactly D solutions (counted with multiplicity).[8]

C.22. Use the representation in Problem C.20 and formula (C.4.6).

C.23. Note that $Q(\mathbb{1}+P-Q)=QP$, then $Q=QP(\mathbb{1}-(Q-P))^{-1}$, hence $\dim(R(P))\geq\dim(R(Q))$, exchanging the role of P and Q the result follows.

C.25. Note that $\ell_\nu(h_\nu)=1$ since P_ν is a projector, hence they are unique apart from a noralization factor. Then we can chose the normalization $\ell_\nu(h_0)=1$ for all ν small enough. Thus $P_\nu f=h_\nu$, that is h_ν is analytic. Hence, for each $g\in\mathcal{B}$ and ν small, $\ell_\nu(g)\ell_0(h_\nu)=\ell_0(P_\nu g)$, which implies ℓ_ν analytic for ν small.

C.27 Think hard.[9]

[8] This is the real reason why spectral theory is done over the complex rather than the real. You should be well aquatinted with the fact that a polynomial p of degree D has D root over $\mathbb{C}$ but, in case you have forgotten, consider the following: first a polynomial of degree larger than zero must have at least a root, otherwise $\frac{1}{p(z)}$ would be an entire function and hence

$$\frac{1}{p(z)}=\lim_{r\to\infty}\frac{1}{2\pi}\int_0^{2\pi}d\theta\frac{1}{p(z+re^{i\theta})}=0.$$

Let z_1 be a root. By the Taylor expansion in z_1 follows the decomposition $p(z)=(z-z_1)p_1(z)$ where p_1 has degree $D-1$. The result follows by induction.

[9] A good idea is to start by considering concrete examples, for instance

$$\begin{pmatrix}1&0\\0&1\end{pmatrix}+\mu\begin{pmatrix}0&1\\1&0\end{pmatrix};\quad\begin{pmatrix}1&1\\0&1\end{pmatrix}+\mu\begin{pmatrix}0&1\\1&0\end{pmatrix}.$$

Appendix D

More on perturbation theory

This section contains some useful perturbation results. We follow and extend the ideas in [Liv03, Theorem 3.2]. Several such results are available (e.g., see [Kif88], [BY93] or [Bal00b] for a review). Here we provide a simplification of the theory developed in [KL99, GL06], see the original works for the full story.

We start by stating the setting in which we work.

D.1 Setting

Hypothesis D.1.1 *Let $X \subset X_w$ be two Banach spaces, $\|\cdot\|$ and $|\cdot|_w$ being the respective norms, satisfying $|\cdot|_w \leq \|\cdot\|$. Also assume that the unit ball of X is weakly compact in X_w. Consider a family of operators $\mathcal{L}_\varepsilon$ with the following properties.*

1. *A uniform Lasota–Yorke inequality: There exist $\lambda_\star > 1$ and $A, B, C > 0$ such that,*

$$\|\mathcal{L}_\varepsilon^n h\| \leq A\lambda_\star^{-n}\|h\| + B|h|_w, \quad |\mathcal{L}_\varepsilon^n h|_w \leq C|h|_w \ ;$$

2. *For $L : X \to X$ define the norm*

$$|||L||| := \sup_{\|h\|\leq 1} |Lf|_w,$$

that is the norm of L as an operator from $X \to X_w$. Then we require that there exists $D > 0$ such that

$$|||\mathcal{L} - \mathcal{L}_\varepsilon||| \leq D\varepsilon.$$

To state a precise result consider, for each operator L, the set

$$V_{\delta,r}(L) := \{z \in \mathbb{C} \mid |z| \le r \text{ or } \operatorname{dist}(z, \sigma(L)) \le \delta\}.$$

Since the complement of $V_{\delta,r}(L)$ belongs to the resolvent of L it follows that

$$H_{\delta,r}(L) := \sup\left\{\|(z-L)^{-1}\| \mid z \in \mathbb{C} \setminus V_{\delta,r}(L)\right\} < \infty.$$

D.2 Perturbation of Lasota-Yorke operators

By $R(z)$ and $R_\varepsilon(z)$ we will mean respectively $(z-\mathcal{L})^{-1}$ and $(z-\mathcal{L}_\varepsilon)^{-1}$.

Theorem D.2.1 ([KL99]) *Consider a family of operators $\mathcal{L}_\varepsilon : X \to X$ satisfying D.1.1. Let $V_{\delta,r} := V_{\delta,r}(\mathcal{L})$, $r > \lambda_\star^{-1}$, $\delta > 0$, then, if $\varepsilon \le \varepsilon_1(\mathcal{L}, r, \delta)$, $\sigma(\mathcal{L}_\varepsilon) \subset V_{\delta,r}(\mathcal{L})$. In addition, if $\varepsilon \le \varepsilon_0(\mathcal{L}, r, \delta)$, there exists $a > 0$ such that, for each $z \notin V_{\delta,r}$,*

$$|||R(z) - R_\varepsilon(z)||| \le C\varepsilon^a.$$

In addition, for each $r > \lambda_\star^{-1}$ and $\delta > 0$ there are constants $a, b > 0$, such that a depends only on r and b depends also on δ, such that, for all $h \in X$ and $\varepsilon \le \varepsilon_0(\mathcal{L}, r, \delta)$,

$$\|R_\varepsilon(z)h\| \le a\|h\| + b|h|_w.$$

PROOF.[1] To start with we collect some trivial, but very useful algebraic identities.
For each operator $L : X \to X$ and $n \in \mathbb{Z}$ holds

$$\frac{1}{z}\sum_{i=0}^{n-1}(z^{-1}L)^i(z-L) + (z^{-1}L)^n = \mathbb{1} \tag{D.2.1}$$

$$R(z)(z-\mathcal{L}_\varepsilon) + \frac{1}{z}\sum_{i=0}^{n-1}(z^{-1}\mathcal{L})^i(\mathcal{L}_\varepsilon - \mathcal{L}) + R(z)(z^{-1}\mathcal{L})^n(\mathcal{L}_\varepsilon - \mathcal{L}) = \mathbb{1} \tag{D.2.2}$$

$$(z-\mathcal{L}_\varepsilon)\left[G_{n,\varepsilon} + (z^{-1}\mathcal{L}_\varepsilon)^n R(z)\right] = \mathbb{1} - (z^{-1}\mathcal{L}_\varepsilon)^n(\mathcal{L}_\varepsilon - \mathcal{L})R(z) \tag{D.2.3}$$

$$\left[G_{n,\varepsilon} + (z^{-1}\mathcal{L}_\varepsilon)^n R(z)\right](z-\mathcal{L}_\varepsilon) = \mathbb{1} - (z^{-1}\mathcal{L}_\varepsilon)^n R(z)(\mathcal{L}_\varepsilon - \mathcal{L}), \tag{D.2.4}$$

where we have set $G_{n,\varepsilon} := \frac{1}{z}\sum_{i=0}^{n-1}(z^{-1}\mathcal{L}_\varepsilon)^i$.

Let us start applying the above formulae. For each $h \in X$ and $z \notin V_{r,\delta}$, and n large and ε small enough,

$$\begin{aligned}\|(z^{-1}\mathcal{L}_\varepsilon)^n(\mathcal{L}_\varepsilon - \mathcal{L})R(z)h\| &\le (r\lambda_\star)^{-n}A\|(\mathcal{L}_\varepsilon - \mathcal{L})R(z)h\| \\ &\quad + \frac{B}{r^n}|(\mathcal{L}_\varepsilon - \mathcal{L})R(z)h|_w \\ &\le [(r\lambda_\star)^{-n}A2C_1 + Br^{-n}D\varepsilon]H_{\delta,r}(\mathcal{L})\|h\| < \|h\|\end{aligned}$$

[1] This proof is simpler than the one in [KL99], yet it gives worse bounds, although sufficient for the present purposes.

To obtain the last inequality, choose $n \in \mathbb{N}$ such that $n = \lfloor -\frac{\ln \varepsilon}{\ln \lambda_\star} \rfloor$. Then assuming $r < 1$ without loss of generality, we have $r^{-n} \leq \varepsilon^{\frac{\ln r}{\ln \lambda_\star}}$, so that both terms are bounded by $C\varepsilon^{1+\frac{\ln r}{\ln \lambda_\star}}$, and $\frac{\ln r}{\ln \lambda_\star} > -1$ since $r\lambda_\star > 1$ by hypothesis. The claimed inequality follows for $\varepsilon > 0$ sufficiently small.

Thus $\|(z^{-1}\mathcal{L}_\varepsilon)^n(\mathcal{L}_\varepsilon - \mathcal{L})R(z)\| < 1$ and the operator on the right hand side of (D.2.3) can be inverted by the usual Neumann series. Accordingly, $(z - \mathcal{L}_\varepsilon)$ has a well defined right inverse. Analogously,

$$\begin{aligned}\|(z^{-1}\mathcal{L}_\varepsilon)^n R(z)(\mathcal{L}_\varepsilon - \mathcal{L})h\| \leq (r\lambda_\star)^{-n} A\|R(z)(\mathcal{L}_\varepsilon - \mathcal{L})h\| \\ + Br^{-n}|R(z)(\mathcal{L}_\varepsilon - \mathcal{L})h|_w.\end{aligned}$$

This time to continue we need some information on the X_w norm of the resolvent. For $g \in X$ equation (D.2.1) yields

$$\begin{aligned}|R(z)g|_w &\leq \frac{1}{r}\sum_{i=0}^{n-1} |(z^{-1}\mathcal{L})^i g|_w + \|R(z)(z^{-1}\mathcal{L})^n g\| \\ &\leq \frac{C}{r^n(1-r)}|g|_w + H_{\delta,r}(\mathcal{L})A(r\lambda_\star)^{-n}\|g\| + H_{\delta,r}(\mathcal{L})Br^{-n}|g|_w \\ &\leq r^{-n}(H_{\delta,r}(\mathcal{L})B + C(1-r)^{-1})|g|_w + H_{\delta,r}(\mathcal{L})A(r\lambda_\star)^{-n}\|g\|\,.\end{aligned} \tag{D.2.5}$$

Substituting, we have

$$\begin{aligned}&\|(z^{-1}\mathcal{L}_\varepsilon)^n R(z)(\mathcal{L}_\varepsilon - \mathcal{L})h\| \leq \{(r\lambda_\star)^{-n} A H_{\delta,r}(\mathcal{L})2C_1[1 + Br^{-n}] \\ &+ Br^{-2n}[H_{\delta,r}(\mathcal{L})B + (1-r)^{-1}]D\varepsilon\}\|h\| < 1,\end{aligned}$$

again, provided ε is small enough and choosing n appropriately. Hence the operator on the right hand side of (D.2.4) can be inverted, thereby providing a left inverse for $(z - \mathcal{L}_\varepsilon)$. This implies that z does not belong to the spectrum of $\mathcal{L}_\varepsilon$.

To investigate the second statement note that (D.2.2) implies

$$R(z) - R_\varepsilon(z) = \frac{1}{z}\sum_{i=0}^{n-1}(z^{-1}\mathcal{L})^i(\mathcal{L}_\varepsilon - \mathcal{L})R_\varepsilon(z) - R(z)(z^{-1}\mathcal{L})^n(\mathcal{L}_\varepsilon - \mathcal{L})R_\varepsilon(z).$$

Accordingly, for each $\varphi \in X$,

$$|R(z)\varphi - R_\varepsilon(z)\varphi|_w \leq \{r^{-n}(1-r)^{-1}\varepsilon + H_{\delta,r}(\mathcal{L})(\lambda_\star r)^{-n}2AC_1 + H_{\delta,r}(\mathcal{L})B\varepsilon\}\|R_\varepsilon(z)\varphi\|.$$

To complete the argument, choose $n = \lfloor -\frac{\ln \varepsilon}{\ln \lambda_\star} \rfloor$ as before and note that by our previous bounds on the inverse of $z - \mathcal{L}_\varepsilon$, we have $\|R_\varepsilon(z)\varphi\| \leq C_{\varepsilon_0}\|\varphi\|$, for all $\varepsilon \leq \varepsilon_0$ and $\varepsilon_0 > 0$ small enough. The first inequality of the theorem follows with $a = 1 + \frac{\ln r}{\ln \lambda_\star}$.

To prove the second inequality, for $|z| = r > \lambda_\star^{-1}$, we use D.2.1 to write

$$\begin{aligned}\|(z-\mathcal{L}_\varepsilon)^{-1}h\| =& \left\|\sum_{k=0}^{m-1} z^{-k-1}\mathcal{L}_\varepsilon^k + (z^{-1}\mathcal{L}_\varepsilon)^m (z-\mathcal{L}_\varepsilon)^{-1}h\right\| \\ \leq& A(1-r^{-1}\lambda_\star^{-1})^{-1}\|h\| + C_{r,m}|h|_w \\ &+ \lambda_\star^{-m} r^{-m}\|(z-\mathcal{L}_\varepsilon)^{-1}h\| + r^{-m}B|(z-\mathcal{L}_\varepsilon)^{-1}h|_w,\end{aligned}$$

for some constant $C_{r,m}$ depending on r and m. We can thus choose m such that $A\lambda_\star^{-m}r^{-m} < \frac{1}{2}$ and, recalling the first inequality of the Theorem, write

$$\|(z-\mathcal{L}_\varepsilon)^{-1}h\| \leq C_r\|h\| + C_{r,m}|h|_w + C\varepsilon^a r^{-m}B\|h\| + r^{-m}B|(z-\mathcal{L})^{-1}h|_w.$$

To conclude, we can use D.2.5 and write, for all $n \in \mathbb{N}$,

$$\|(z-\mathcal{L}_\varepsilon)^{-1}h\| \leq C_\sharp[C_r + \varepsilon^a r^{-m} + AH_{\delta,r}(\mathcal{L})(r\lambda_\star)^{-n}r^{-m}]\|h\| + C_{r,m,n,\delta}|h|_w.$$

Choosing n and ε so that $H_{\delta,r}(\mathcal{L})(r\lambda_\star)^{-n}r^{-m} \leq 1$ and $\varepsilon^a r^{-m} \leq 1$ yields the statement. □

D.2.1 shows that the point spectrum is stable. Yet, in applications it is also important to control the multiplicity of the spectrum. This can be done thanks to the following Lemma.

Lemma D.2.2 *Consider a family of operators $\mathcal{L}_\varepsilon : X \to X$ satisfying D.1.1. Let $\nu \in \sigma(\mathcal{L})$, $|\nu| > \lambda_\star$, and let m be the dimension of the eigenspace associated to ν. Then, for each δ small enough there exists $\varepsilon_2(\mathcal{L}, \nu, \delta)$ such that, for all $\varepsilon \leq \varepsilon_2(\mathcal{L}, \nu, \delta)$, $\sigma(\mathcal{L}_\varepsilon) \cap \{z \in \mathbb{C} \,:\, |z-\nu| < \delta\}$ contains at most m eigenvalues and the total dimension of their eigenspaces is m.*

Proof. Since $|\nu| > \lambda_\star$, F.4.2 implies that ν belongs to the point spectrum. Hence, there exists δ_0 such that $\{z \in \mathbb{C} \,:\, |z-\nu| < \delta_0\} \cap \sigma(\mathcal{L}) = \{\nu\}$. Then D.2.1 implies that, for each $\delta < \delta_0/2$ and $\varepsilon \leq \varepsilon_0(\mathcal{L}, r, \delta)$, we can split the spectrum as $\sigma(\mathcal{L}_\varepsilon) = \sigma_1 \cup \sigma_2$ where $\sigma_1 \cap \sigma_2 = \emptyset$ and $\sigma_1 \subset \{z \in \mathbb{C} \,:\, |z-\nu| < \delta\}$. Accordingly, by (C.3.5) we can define the eigenprojectors

$$\Pi_\varepsilon := \frac{1}{2\pi i}\int_{\gamma_\delta} (z\mathbb{1} - \mathcal{L}_\varepsilon)^{-1}dz, \tag{D.2.6}$$

where $\gamma_\delta(t) = \nu + \delta e^{it}$, and $\sigma(\Pi_\varepsilon\mathcal{L}_\varepsilon) = [\sigma(\mathcal{L}_\varepsilon) \cap \{z \in \mathbb{C} \,:\, |z-\nu| < \delta\}] \cup \{0\}$. Note that the first inequality of D.2.1 implies, for $\varepsilon \leq \varepsilon_0(\mathcal{L}, r, \delta)$, where we can choose $r = \{\lambda_\star^{-1} + |\nu|\}/2$,

$$|(\Pi_\varepsilon - \Pi_0)h|_w \leq C_\delta \varepsilon^a \|h\|,$$

for some constant C_δ, depending on the choice of δ. While the second inequality of D.2.1 implies that there exist constants a and b_δ, the latter depending on δ, such that

$$\|\Pi_\varepsilon h\| \leq a\delta\|h\| + b_\delta |h|_w.$$

Since Π_ε is independent of δ (see (C.3.5)) we have

$$\|\Pi_\varepsilon h\| \leq (a\delta_0 + b_{\delta_0})\|h\| =: c_0\|h\|.$$

The above inequalities imply

$$\begin{aligned}\|(\Pi_\varepsilon - \Pi_0)^2 h\| &\leq 2a\delta\|(\Pi_\varepsilon - \Pi_0)h\| + 2b_\delta\|(\Pi_\varepsilon - \Pi_0)h\| \\ &\leq \left[4ac_0\delta + 2b_\delta C_\delta \varepsilon^a\right]\|h\|.\end{aligned}$$

Accordingly, if we choose δ such that $8ac_0\delta \leq 1$ and ε_2 such that $2b_\delta C_\delta \varepsilon^a < \frac{1}{2}$, we obtain

$$\|(\Pi_\varepsilon - \Pi_0)^2\| < 1. \tag{D.2.7}$$

This concludes the Lemma due to the following general fact.

Problem D.1 *Let $\Pi_1, \Pi_2 \in L(X, X)$ be two projectors. Assume that*

$$\|(\Pi_1 - \Pi_2)^2\| < 1,$$

then $\dim(\Pi_1(X)) = \dim(\Pi_2(X))$.

□

The above two results are rather effective to study perturbations of transfer operators. The reader can verify this directly by solving the next problem.

Problem D.2 *Consider the maps $f_n : \mathbb{T}^1 \to \mathbb{T}^1$ defined by*

$$f(x) = 2x + \frac{1}{2n}\sin 2\pi\sqrt{n}x \mod 1$$

and use D.2.1 and D.2.2 to study the spectrum of the operators $\mathcal{L}_n h(x) = \sum_{y \in f_n^{-1}(x)} \frac{h(y)}{f_n'(y)}$, for n large. In particular, show that, for n large enough, $\mathcal{L}_n$ has a spectral gap close to $\frac{1}{2}$.

Given the above results, it is natural to ask if the spectral data have some more regular dependence on the change in the operator. These types of questions are related to *linear response*.

D.3 Linear Response

In order to have a linear response, one needs more control on the operators $\mathcal{L}_\varepsilon$ than that provided by D.1.1. Here we provide the simplest possibility, see [GL06, Section 8] and [KL09] for more details.[2]

Hypothesis D.3.1 *Let $X_2 \subset X_1 \subset X_0$ be three Banach spaces, equipped with the norms $\|\cdot\|_i$, respectively, satisfying $\|\cdot\|_0 \le \|\cdot\|_1 \le \|\cdot\|_2$. Also assume that the unit ball of X_i is weakly compact in X_{i+1}. Consider a family of operators $\mathcal{L}_\varepsilon$ with the following properties.*

1. *A uniform Lasota–Yorke inequality: There exist $\lambda_\star > 1$ and $A, B, C > 0$ such that,*

$$\|\mathcal{L}_\varepsilon^n h\|_i \le A\lambda_\star^{-n}\|h\|_i + B\|h\|_{i-1}, \quad \text{for } i > 0 \text{ and for all } h \in X_i$$
$$\|\mathcal{L}_\varepsilon^n h\|_i \le C\|h\|_i\,, \quad \text{for } i \ge 0 \text{ and for all } h \in X_i.$$

2. *We require that there exists an operator $\mathcal{A} \in L(X_j, X_i)$, for each $j > i$, such that*

$$\|(\mathcal{L}_\varepsilon - \mathcal{L} - \varepsilon\mathcal{A})h\|_0 \le D\varepsilon\|h\|_1, \quad \text{for all } h \in X_1$$
$$\|(\mathcal{L}_\varepsilon - \mathcal{L} - \varepsilon\mathcal{A})h\|_1 \le D\varepsilon\|h\|_2, \quad \text{for all } h \in X_2$$
$$\|(\mathcal{L}_\varepsilon - \mathcal{L} - \varepsilon\mathcal{A})h\|_0 \le D\varepsilon^{1+\alpha}\|h\|_2, \quad \text{for all } h \in X_2,$$

for some $\alpha > 0$ and each $h \in X_2$.

Remark D.3.2 *The D.3.1 are a bit different from the ones in [GL06]. This is made in order to present a simplified proof.*

Remark D.3.3 *Note that the D.3.1 imply D.1.1 for $\mathcal{L}, \mathcal{L}_\varepsilon$ both with respect to the norms $\|\cdot\|_0, \|\cdot\|_1$ and with respect to the norms $\|\cdot\|_1, \|\cdot\|_2$.*

We will need the following well-known fact.

Problem D.3 *Prove that for any $A, B \in L(X, X)$ and $z \notin \sigma(A) \cup \sigma(B)$ we have*

$$(z\mathbb{1} - A)^{-1} - (z\mathbb{1} - B)^{-1} = (z\mathbb{1} - A)^{-1}(A - B)(z\mathbb{1} - B)^{-1},$$

which is called the resolvent identity.

Finally, let us define

$$V_{\delta,r}(\mathcal{L}) := \{z \in \mathbb{C} \mid |z| \le r \text{ or } \operatorname{dist}(z, \sigma_{X_1}(\mathcal{L})) \le \delta\},$$

where $\sigma_X(\mathcal{L})$ is the spectrum of $\mathcal{L}$ seen as an operator in $L(X, X)$.

[2] Note that [GL06, Section 8] contains an imprecision which is fixed in [Gou10, Theorem 3.3].

Remark D.3.4 *Note that* $\left[\sigma_{X_2}(\mathcal{L}) \cap \{|z| \geq \lambda_\star^{-1}\}\right] \subset \left[\sigma_{X_1}(\mathcal{L}) \cap \{|z| \geq \lambda_\star^{-1}\}\right]$ *since by F.4.2 this part of the spectrum belongs to the point spectrum. Accordingly, if* $\nu \in \sigma_{X_2}(\mathcal{L}) \cap \{|z| \geq \lambda_\star^{-1}\}$*, then there exists* $h \in X_2$ *such that* $\mathcal{L}h = \nu h$ *and hence* $\nu \in \sigma_{X_1}(\mathcal{L})$*.*

We are then ready to provide the last result of this section.

Remark D.3.5 *D.3.6 says that* $(z - \mathcal{L}_\varepsilon)^{-1}$*, when seen as a function from* $\mathbb{R}$ *to* $L(X_2, X_0)$ *is differentiable at zero. But then also the eigenprojectors* Π_ε *defined in D.2.6 are differentiable and so is* $\Pi_\varepsilon \mathcal{L}_\varepsilon$*. In particular, if the projector* Π_ε *is associated with a simple eigenvalue* ν_ε*, and hence has the form* $\Pi_\varepsilon = \ell_\varepsilon \otimes h_\varepsilon$*, then* $\Pi_\varepsilon \mathcal{L}_\varepsilon = \nu_\varepsilon \Pi_\varepsilon$*. It follows that* ν_ε *is differentiable and* $\varepsilon \to h_\varepsilon$ *is differentiable as a function from* $\mathbb{R}$ *to* X_0*.*

Theorem D.3.6 *Consider a family of operators* $\mathcal{L}_\varepsilon : X_0 \to X_0$ *satisfying D.3.1. Let* $r > \lambda_\star^{-1}$ *and* $\delta > 0$*. If* $\varepsilon \leq \varepsilon_2(\mathcal{L}, r, \delta)$*, then* $\sigma_{X_1}(\mathcal{L}_\varepsilon) \subset V_{\delta,r}(\mathcal{L})$ *and* $\sigma_{X_2}(\mathcal{L}_\varepsilon) \subset V_{\delta,r}(\mathcal{L})$*. Moreover, there exists* $\eta > 0$ *such that, for all* $z \notin V_{\delta,r}(\mathcal{L})$ *and* $h \in X_2$*,*

$$\|[R(z) - R_\varepsilon(z) - \varepsilon R(z)\mathcal{A}R(z)]h\|_0 \leq C_\delta \varepsilon^{1+\eta} \|h\|_2.$$

Proof. The fact that $\sigma_{X_i}(\mathcal{L}_\varepsilon) \subset V_{\delta,r}(\mathcal{L})$ follows from D.2.1 and D.3.4.

Let $\mathcal{Q}_\varepsilon = \mathcal{L}_\varepsilon - \mathcal{L} - \varepsilon\mathcal{A}$ and, as before $R(z) = (z\mathbb{1} - \mathcal{L})^{-1}$ and $R_\varepsilon(z) = (z\mathbb{1} - \mathcal{L}_\varepsilon)^{-1}$. By D.3 we can write

$$R_\varepsilon(z) - R(z) = R_\varepsilon(z)(\mathcal{L}_\varepsilon - \mathcal{L})R(z).$$

Thus if we define $\Xi = R_\varepsilon(z)\mathcal{A}R(z)$, we have that

$$\|(R_\varepsilon(z) - R(z) - \varepsilon\Xi)h\|_0 = \|R_\varepsilon(z)\mathcal{Q}_\varepsilon R(z)h\|_0.$$

Arguing as in D.2.5, recalling D.3.3 and the second inequality of D.2.1, we can show that there exists $C_{r,\delta} > 0$ such that for all $g \in X_1$,

$$\|R_\varepsilon(z)g\|_0 \leq C_{\delta,r}\left[r^{-m}\|g\|_0 + (r\lambda_*)^{-m}\|g\|_1)\right].$$

Accordingly, using D.3.1-(2) and recalling $\sigma_{X_2}(\mathcal{L}) \subset V_{\delta,r}(\mathcal{L})$, we have, for each $h \in X_2$,

$$\begin{aligned}
\|(R_\varepsilon(z) - R(z) - \varepsilon\Xi)h\|_0 &\leq C_{\delta,r}\left[r^{-m}\|\mathcal{Q}_\varepsilon R(z)h\|_0 + (r\lambda_\star)^{-m}\|\mathcal{Q}_\varepsilon R(z)h\|_1)\right] \\
&\leq C_{\delta,r} D\left[r^{-m}\varepsilon^{1+\alpha} + (r\lambda_\star)^{-m}\varepsilon)\right]\|R(z)h\|_2 \\
&\leq C'_{\delta,r}\left[r^{-m}\varepsilon^{1+\alpha} + (r\lambda_\star)^{-m}\varepsilon)\right]\|h\|_2
\end{aligned}$$

for some constant $C'_{\delta,r}$. Choosing m so that $\varepsilon^\alpha = \lambda_\star^{-m}$, the above implies that, setting $\eta_0 = \alpha(1 - \frac{\ln r^{-1}}{\ln \lambda}) > 0$, we have

$$\|(R_\varepsilon(z) - R(z) - \varepsilon\Xi)h\|_0 \leq C_\delta \varepsilon^{1+\eta_0} \|h\|_2.$$

On the other hand, D.2.1 implies

$$\begin{aligned} \| \left[R_\varepsilon(z)\mathcal{A}R(z) - R(z)\mathcal{A}R(z)\right] h\|_0 &\leq C_\delta \varepsilon^a \|\mathcal{A}R(z)h\|_1 \\ &\leq C_\delta \varepsilon^a \|R(z)h\|_2 \leq C'_\delta \varepsilon^a \|h\|_2. \end{aligned}$$

Which concludes the proof with $\eta = \min\{\eta_0, a\}$. □

Appendix E

Analytic Fredholm Theorem

Here I provide a proof of the Analytic Fredholm alternative in Banach spaces.

Theorem E.0.1 (Analytic Fredholm alternative)[1] *Let D be an open connected subset of $\mathbb{C}$. Let $F : \mathbb{C} \to L(\mathbb{B}, \mathbb{B})$ be an analytic operator-valued function such that $F(z)$ is compact for each $z \in D$. Then, one of the following two alternatives holds true*

- $(\mathbb{1} - F(z))^{-1}$ *exists for no* $z \in D$
- $(\mathbb{1} - F(z))^{-1}$ *exists for all* $z \in D \backslash S$ *where S is a discrete subset of D (i.e. S has no limit points in D). In addition, if $z \in S$, then 1 is an eigenvalue for $F(z)$ and the associated eigenspace has finite multiplicity.*

Proof. First of all notice that, for each $z_0 \in D$ there exists $r > 0$ such that $D_{r(z_0)}(z_0) := \{z \in \mathbb{C} \, : \, |z - z_0| < r(z_0)\} \subset D$, and

$$\sup_{z \in D_{r(z_0)}(z_0)} \|F(z) - F(z_0)\| \leq \frac{1}{4}.$$

If we can prove the theorem in each such disk, we are done.[2] We can approximate $F(z_0)$ by a finite rank operator K such that $\|F(z_0) - K\| \leq \frac{1}{4}$. Then

$$\sup_{z \in D_{r(z_0)}(z_0)} \|F(z) - K\| \leq \frac{1}{2}.$$

[1] The present proof is patterned after the proof of the Analytic Fredholm alternative for compact operators (in Hilbert spaces) given in [RS80, Theorem VI.14].

[2] In fact, consider any connected compact set K contained in D. Let us suppose that for each $z_0 \in K$ we have a disk $D_{r(z_0)}(z_0)$ in which the theorem holds. Since the disks $D_{r(z_0)/2}(z_0)$ form a covering for K we can extract a finite cover. If the first alternative holds in one such disk then, by connectness, it must hold on all K. Otherwise each $S \cap D_{r(z_0)/2}(z_0)$, and hence $K \cap S$, contains only finitely many points. The Theorem follows by the arbitrariness of K.

Note that

$$\mathbb{1} - F(z) = \left(\mathbb{1} - K(\mathbb{1} - [F(z) - K])^{-1}\right)(\mathbb{1} - [F(z) - K]).$$

Thus the invertibility of $\mathbb{1} - F(z)$ in $D_r(z_0)$ depends on the invertibility of $\mathbb{1} - K(\mathbb{1} - [F(z) - K])^{-1}$. Let us set $F_0(z) := K(\mathbb{1} - [F(z) - K])^{-1}$ and note that $F_0(z)$ is a finite rank operator.

Let us start by looking at the equation

$$(\mathbb{1} - F_0(z))h = 0. \tag{E.0.1}$$

Clearly if a solution exists, then $h \in \mathrm{Range}(F_0(z)) = \mathrm{Range}(F(z_0)) := \mathbb{V}_0$. Since $\mathbb{V}_0$ is finite dimensional there exists a basis $\{h_i\}_{i=1}^N$ such that $h = \sum_i \alpha_i h_i$. On the other hand there exists an analytic matrix $G(z)$ such that[3]

$$F_0(z)h = \sum_{ij} G(z)_{ij}\alpha_j h_i.$$

Thus (E.0.1) is equivalent to

$$(\mathbb{1} - G(z))\alpha = 0,$$

where $\alpha := (\alpha_i)$.

The above equation can be satisfied only if $\det(\mathbb{1} - G(z)) = 0$ but the determinant is analytic hence it is either always zero or zero only at isolated points.[4]

Suppose the determinant different from zero, and consider the equation

$$(\mathbb{1} - F_0(z))h = g.$$

Let us look for a solution of the type $h = \sum_i \alpha_i h_i + g$. Substituting yields

$$\alpha - G(z)\alpha = \beta$$

where $\beta := (\beta_i)$ with $F_0(z)g =: \sum_i \beta_i h_i$. Since the above equation admits a solution, we have $\mathrm{Range}(\mathbb{1} - F_0(z)) = \mathbb{B}$, Thus we have an everywhere defined inverse, hence bounded by the open mapping theorem.

[3] To see the analyticity notice that we can construct linear functionals $\{\ell_i\}$ on $\mathbb{V}_0$ such that $\ell_i(h_j) = \delta_{ij}$ and then extend them to all $\mathbb{B}$ by the Hahn-Banach theorem. Accordingly, $G(z)_{ij} := \ell_j(F_0(z)h_i)$, which is obviously analytic.

[4] The attentive reader has certainly noticed that this is the turning point of the theorem: the discreteness of S is reduced to the discreteness of the zeroes of an appropriate analytic function: a determinant. A moment thought will immediately explain the effort made by many mathematicians to extend the notion of determinant (that is to define an analytic function whose zeroes coincide with the spectrum of the operator) beyond the realm of matrices (the so called Fredholm determinants).

We are thus left with the analysis of the situation $z \in S$ in the second alternative. In such a case, there exists h such that $(\mathbb{1} - F(z))h = 0$, thus one is an eigenvalue. On the other hand, if we apply the above facts to the function $\Phi(\zeta) := \zeta^{-1}F(z)$ analytic in the domain $\{\zeta \neq 0\}$ we note that the first alternative cannot take place since for $|\zeta|$ large enough $\mathbb{1} - \Phi(\zeta)$ is obviously invertible. Hence, the spectrum of $F(z)$ is discrete and can accumulate only at zero. This means that there is a small neighborhood around one in which $F(z)$ has no other eigenvalues, we can thus surround one with a small circle γ and consider the projector

$$\begin{aligned} P :=& \frac{1}{2\pi i}\int_\gamma (\zeta - F(z))^{-1} d\zeta = \frac{1}{2\pi i}\int_\gamma \left[(\zeta - F(z))^{-1} - \zeta^{-1}\right] d\zeta \\ =& \frac{1}{2\pi i} F(z) \int_\gamma \zeta^{-1}(\zeta - F(z))^{-1} d\zeta. \end{aligned}$$

By standard functional calculus, it follows that P is a projector and it projects on the eigenspace of the eigenvector one. But the last formula shows that P equals a compact operator times a bounded one, hence it is compact, therefore finite-dimensional. □

Appendix F

Hennion–Neussbaum Theory

I provide a self-contained proof of Hennion–Neussbaum's theory.

While such results are routinely used in many papers devoted to the study of the statistical properties of dynamical systems, as far as we know, no self-contained account of the theory is available. Our goal here is to present such a complete account in a manner accessible to a reader with basic knowledge of functional analysis and to reduce technicalities to a minimum. We start with some needed preliminary functional analytic facts, then we discuss the *essential spectrum*. There exist many alternative definitions of essential spectrum; here, we use the most convenient for our goals. The reader interested in more details can have a look at the first chapter of [EE18]. Next, we introduce the *measures of noncompactness*, which form the basis for Neussbaum's essential spectral characterization. After that we are finally able to state and prove Hennion's theorem.

F.1 A bit of functional analysis preliminaries

In the following, we will need some facts from functional analysis that are not necessarily common knowledge; hence, we state them here together with their proofs. The goal is to establish Theorem F.1.3.

Lemma F.1.1 *Let X be a Banch space, if $V, W \subset X$ are closed and finite dimensional, respectively, then $V + W$ is closed.*

Proof. The Lemma follows if we can prove it for the case $\dim W = 1$, and $W \not\subset V$. Let $x \in W$, $\|\bar{x}\| = 1$, then $V + W = \{\xi x + v \,:\, \xi \in \mathbb{R}\,,\, \xi \in V\}$. Suppose that $\{\eta_n := x\xi_n + v_n\}$ converges to some η, we want to show that $\eta = x\xi_* + v_*$ for some $\xi_* \in \mathbb{R}$ and $v_* \in V$; that is, $V + W$ is closed.
Since V is closed and $x \notin V$, it must be $d(x, V) =: d > 0$. In addition, for

each $\xi \in \mathbb{R}$, $v \in V$, we have

$$\|\xi x + v\| \geq d(\xi x + v, V) = \xi d(x, V) = |\xi| d.$$

Hence,

$$\|\eta_n - \eta_m\| = \|(\xi_n - \xi_m)x + (v_n - v_m)\| \geq d|\xi_n - \xi_m|.$$

It follows that $\{\xi_n\}$ is Cauchy and then it has a limit ξ_*. But the $v_n = \eta_n - \xi_n x$ converges as well to some v_* and, since V is closed, $v_* \in V$. This proves the Lemma. □

Lemma F.1.2 *Let X be a Banach space and $T \in L(X, X)$ such that $R(T)$ is closed and $\dim(N(T)) < \infty$, then $R(T^n)$ is closed for all $n \in \mathbb{N}$.*

Proof. It suffices to prove that if $V \subset X$ is closed, the TV is closed. Let $\{x_n\} \subset V$ be such that Tx_n is Chauchy. Since $R(T)$ is close, there esists $y \in TX$ such that $\lim_{n\to\infty} Tx_n = y$, we have to show that $y \in TV$. Consider the quotient space $\tilde{X} = X/N(T)$, and the quotient map $\tilde{T} \in L(\tilde{X}, X)$. We have that $Y := \tilde{T}\tilde{X} = R(T)$ is closed and $\tilde{T}$ is injective and surjective from $\tilde{X}$ to Y. Then the bounded inverse theorem implies that $\tilde{T}^{-1} \in L(Y, \tilde{X})$, that is, it is bounded.[1] Accordingly, there exists $\{z_n\} \subset N(T)$ such that $\{x_n + z_n\} \subset V + N(T)$ is Cauchy. Since $N(T)$ is finite dimensional, $V + N(T)$ is closed, by Lemma F.1.1, and hence there exists $w \in V + N(T)$ such that $\lim_{n\to\infty} x_n + z_n = w$. We can write $w = a + b$, with $a \in V$ and $b \in N(T)$. Then

$$T(a) = T(w) = \lim_{n\to\infty} T(x_n + z_n) = \lim_{n\to\infty} T(x_n) = y$$

concluding the proof. □

The main result of this section follows ideas from [Kat66, Theorem IV-5.30].

Theorem F.1.3 *Let X be a Banach space and $T \in L(X, X)$ a quasi-nilpotent operator with $\dim(N(T)) < \infty$ and $R(T)$ closed.*[2] *Then $\dim(X) < \infty$.*

Proof. First, we need to establish the following fact. Consider the spaces $V_n := N(T) \cap R(T^n)$. Since $T^n X = T^{n-1}(T(X))$, we have $R(T^n) \subset R(T^{n-1})$. Thus, by Lemma F.1.2, V_n is a decreasing sequence of closed subspaces.
Since $N(T)$ is finite dimensional, there exists $m \in \mathbb{N}$ such that $V_n = V_m$ for all $n \geq m$.
Let $Y = T^m X$, then $TY = T^m(TX) \subset Y$, and it is closed by Lemma F.1.2

[1] Recall that the bounded inverse theorem is an immediate consequence of the open mapping theorem, see [RS80, Theorem III.11]).

[2] Recall that an operator $T \in L(X, X)$ is quasi-nilpotent if $\lim_{n\to\infty} \|T^n\|^{\frac{1}{n}} = 0$.

again. It follows that $T_* := T|_Y$, the restriction to Y, belongs to $L(Y,Y)$, $\|T_*\| \leq \|T\|$ and $R(T_*)$ is closed. Note that,

$$\begin{aligned} N(T_*) =& N(T) \cap T^m X = V_m = V_{m+1} \\ =& N(T) \cap T^{m+1} X \subset T^{m+1} X = T_* Y. \end{aligned} \tag{F.1.1}$$

That is $N(T_*) \subset R(T_*)$. Next, we prove $N(T_*^n) \subset R(T_*)$ for all $n \in \mathbb{N}$. We proceed by induction: assume that we have $N(T_*^n) \subset R(T_*)$, then if $x \in N(T_*^{n+1})$ we can write $x = a + b$ where $a \in N(T_*^n) \subset R(T_*)$ and $T_*^n b \in N(T_*) \subset R(T_*)$, thus $x \in R(T_*)$.
Since $N(T_*) \subset N(T)$ is finite dimensional, there exists a closed subspace $Z \subset Y$ such that $Y = N(T_*) \oplus Z$.[3] Then $T_* Z = R(T_*)$, hence $T_*|_Z$ is a one-one map onto $R(T_*)$. Accordingly to the bounded inverse theorem (e.g., see [RS80, Theorem III.11]) there exists $S \in L(R(T_*), Z)$ such that $T_* S = \mathbb{1}$ and $S(T_*|_Z) = \mathbb{1}$. It follows that, if $x \in N(T_*^n)$ then $x \in R(T_*)$ and we can apply S, yielding $Sx \in N(T_*^{n+1})$. Accordingly, if $x \in N(T_*)$, $S^n x$ is well defined for all $n \in \mathbb{N}$. We can finally use the quasi-nilpotent hypothesis: for each $x \in N(T_*)$,

$$\begin{aligned} \|x\| &= \lim_{n\to\infty} \|T_*^n S^n x\| \leq \lim_{n\to\infty} \|T_*^n\| \|S\|^n \|x\| \\ &\leq \lim_{n\to\infty} \|T^n\| \|S\|^n \|x\| = 0. \end{aligned} \tag{F.1.2}$$

That is $N(T_*) = \{0\}$. But this implies $Z = Y$ and $S T_* = \mathbb{1}$, hence $S^n T_*^n = \mathbb{1}$. Then, for each $x \in Y$,

$$\|x\| = \lim_{n\to\infty} \|S^n T_*^n x\| = \lim_{n\to\infty} \|S^n T^n x\| \leq \lim_{n\to\infty} \|S\|^n \|T^n\| \|x\| = 0.$$

That is $\{0\} = Y = T^m X$, i.e. $X = N(T^m)$. We can finally conclude since

$$\dim X = \dim(N(T^m)) \leq m \dim(N(T)) < \infty.$$

□

F.2 Essential Spectrum

Our aim is to divide the spectrum $\sigma(T)$ of a bounded, linear operator T into two parts, $\sigma_p(T)$ and $\sigma_{ess}(T)$. The discrete spectrum of T, $\sigma_p(T)$, consists of isolated points $\lambda \in \sigma(T)$ such that their associated Riesz projector has finite rank and the range of $\lambda - T$ is closed, while the essential spectrum of T, $\sigma_{ess}(T)$, will be the remaining part of the spectrum. This motivates the following definition of the essential spectrum, akin to [Bro61].

[3] This follows from the Hahn-Banch theorem, which, given a base $\{x_i\}$ of $N(T_*)$ allows to construct functionals ℓ_i such that $\ell_i(x_j) = \delta_{ij}$ and hence the projector $P = \sum_i x_i \ell_i$ whose range is $N(T_*)$, therefore the kernel is the wanted Z.

Definition F.2.1 *Let T be a bounded linear operator on a Banach space X. The (Browder) essential spectrum of T, $\sigma_{ess}(T)$, is the set of $\lambda \in \sigma(T)$, such that at least one of the following conditions holds:*

1) The range of $\lambda\mathbb{1} - T$, $R(\lambda\mathbb{1} - T)$, is not closed;

2) $N(\lambda\mathbb{1} - T)$ is infinite dimensional;

3) λ is a limit point of $\sigma(T) \setminus \{\lambda\}$.

There are many other definitions of the essential spectrum. For example, Wolf's ([Wol59]) essential spectrum is the set of those $z \in \mathbb{C}$ such that $z - T$ is not Fredholm. Recall that an operator $T : X \to X$ is Fredholm if $R(T)$ is closed and the dimensions of both $N(T)$ and the quotient $X/R(T)$ are finite.

The essential spectral radius of a bounded operator T is defined as[4]

$$r_e(T) := \sup\{|z| \in \mathbb{C} \ : \ z \in \sigma_{ess}(T)\}. \tag{F.2.3}$$

A relevant fact is that r_e is the same under all these different definitions; see [EE18, Section 1.4] and the subsequent discussion. In the following, we do not need to enter into such subtleties.
However, it is useful to better clarify the properties of $\sigma_p(T) = \sigma(T) \setminus \sigma_{ess}(T)$.

Lemma F.2.2 *Given $T \in L(X, X)$, for some Banch space X. If $z \in \sigma_p(T)$, then we can write $X = X_0 \oplus X_1$,[5] $T(X_0) \subset X_0$, $T(X_1) \subset X_1$, X_0 is finite dimensional and, finally, $\sigma(T|_{X_0}) = \{z\}$ while $\sigma(T|_{X_1}) \cap \{z\} = \emptyset$.*

Proof. By defintion z is an isolate point of $\sigma(T)$, thus we can consider a close curve γ such that, calling D its interior, $D \cap \sigma(T) = \{z\}$ an consider the projector (see Problem C.20)

$$P = \frac{1}{2\pi i} \int_\gamma (\zeta\mathbb{1} - T)^{-1} d\zeta.$$

Let $X_0 = R(P)$ and $X_1 = N(P)$, by Lemma C.1.2 and Problems C.21, C.22, these subspaces have all the wanted properties apart from the finite dimensionality of X_0.
To establish the latter, consider the operator $T_0 = T|_{X_0} \in L(X_0, X_0)$. Since $\sigma(T_0) = \{z\}$, we have that $\sigma(z\mathbb{1} - T_0) = \{0\}$.[6] Then Lemma C.3.2 implies that the spectral radius of $r(z\mathbb{1} - T_0) = 0$, that is, T_0 is *quasi-nilpotent*. In addition, suppose that for some sequence $\{x_n\} \subset X_0$ and $y \in X_0$ we have

[4] We will often write simply r_e if the operator T is clear from the context.
[5] In particular $X_0 \cap X_1 = \{0\}$.
[6] Here, and in the following, we slightly abuse notation and we write $\mathbb{1}$ for $\mathbb{1}_{X_0}$, the identity operator in $L(X_0, X_0)$, since the meaning is clear from the context.

$\lim_{n\to\infty}(z\mathbb{1}-T_0)x_n = y$. Since $(z\mathbb{1}-T_0)x_n = (z\mathbb{1}-T)x_n$ and $R(z\mathbb{1}-T)$ is closed by hypotheis, there exists $\xi \in X$ such that $(z\mathbb{1}-T)\xi = y$. Yet,

$$y = Py = P(z\mathbb{1}-T)\xi = (z\mathbb{1}-T)P\xi,$$

where we have used Problem C.17. Hence, $y \in R(z\mathbb{1}-T_0)$, that is $R(z\mathbb{1}-T_0)$ is closed. Finally, $\dim(N(z\mathbb{1}-T)) < \infty$ by hypothesis as well. The Lemma follows then from Theorem F.1.3. □

F.2.1 Subspaces

Here we recall a few, mostly well-known facts, about subspaces of a Banach space.

Definition F.2.3 *Let $V \subset X$ be a subspace of a normed vector space X. Given $x \in X$, we define the distance to V by:*

$$\operatorname{dist}(x, V) = \inf\{\|x - y\| : y \in V\}.$$

Definition F.2.4 *A subspace V is called a proper subspace of X if it is neither the whole space X nor the zero subspace $\{0\}$.*

Lemma F.2.5 *Let X be a Banach space, $V \subset X$ a proper closed subspace. For every $\varepsilon > 0$ there exists $x_0 \in X$, $\|x_0\| = 1$, and $\operatorname{dist}(x_0, V) \geq 1 - \varepsilon$.*

Proof. Let $x' \in X \setminus V$, then $d = \operatorname{dist}(x', V) > 0$, (since V is closed). For each $\eta > 0$ there exists $y' \in V$ so that $d \leq \|x' - y'\| \leq d + \eta$. Let $x_0 = \frac{x'-y'}{\|x'-y'\|}$ and $\eta = \frac{\varepsilon d}{1-\varepsilon}$. For any $z \in V$ we have:

$$\|x_0 - z\| = \frac{1}{\|x' - y'\|}\left\|x' - y' - \|x' - y'\|\, z\right\| \geq \frac{d}{\|x' - y'\|} \geq \frac{d}{d+\eta} = 1 - \varepsilon,$$

since $y' + \|x' - y'\|\, z \in V$. The result follows since ε is arbitrary. □

Definition F.2.6 *A normed vector space X is locally compact if any bounded sequence in X has a convergent subsequence.*

Theorem F.2.7 *(S. Banach) Every locally compact Banach space X has finite dimension.*

Proof. If $\dim X = \infty$, then we can construct a sequence of unit vectors $\{x_i\}_{i\in\mathbb{N}} \subset X$ such that $\|x_i - x_j\| \geq \frac{1}{2}$ for all $i \neq j \in \mathbb{N}$. Indeed, since for all $r \in \mathbb{N}$, $\operatorname{span}\{x_1, \cdots, x_r\}$ is finite dimensional, and hence closed, by Lemma F.2.5 there exists $x_{r+1} \in X$, $\|x_{r+1}\| = 1$, such that $d(x_{r+1}, \operatorname{span}\{x_1, \cdots, x_r\}) \geq \frac{1}{2}$. This contradicts the assumption that X is locally compact. □

F.2.2 Measure of Noncompactness

We can now introduce our major technical tool.

Definition F.2.8 *Let X be a Banach space and $A \subset X$ a bounded subset. We define $\gamma(A)$, which we call the (Kuratowski) measure of noncompactness of A, to be*

$$\inf\left\{ r>0 \ : \ \exists\, n\in\mathbb{N},\ S_1,\cdots,S_n,\ \mathrm{diam}(S_i)\le r\,,\, s.t.\ A\subset \bigcup_{i=1}^{n} S_i \right\}.$$

Definition F.2.9 *We call the ball measure of noncompactness of A in X, $\tilde\gamma_X(A)$, to be* [7]

$$\inf\left\{ r>0 \ : \ \exists\, n\in\mathbb{N}, B_r(x_1),\cdots,B_r(x_n) \quad x_i\in X \ s.\ t. \quad A\subset \bigcup_{i=1}^{n} B_r(x_i) \right\}.$$

Definition F.2.10 *If X_1 and X_2 are Banach spaces and $T\in L(X_1,X_2)$, we say that T is a k-set-contraction if for every bounded set $A\subset X_1$,*

$$\gamma_{X_2}(T(A))\le k\gamma_{X_1}(A).$$

We say that T is a ball-k-set-contraction if

$$\tilde\gamma_{X_2}(T(A))\le k\tilde\gamma_{X_1}(A)$$

for every bounded set A in X_1.
We define

$$\gamma(T)=\inf\{k>0 : T \text{ is a } k\text{-set-contraction}\}$$
$$\tilde\gamma(T)=\inf\{k>0 : T \text{ is a ball-}k\text{-set-contraction}\}.$$

Remark F.2.11 *The above ideas can also be defined for nonlinear maps between metric spaces [Dar55, Nus69].*

Denote the closed ideal of compact linear operators of X into X by $\mathcal{K}(X)$, or $\mathcal{K}$ if no confusion arises.[8] Let $Z=L(X,X)/\mathcal{K}$.

Definition F.2.12 *We define a seminorm $\|T\|_{\mathcal{K}}$ on $L(X,X)$ by*

$$\|T\|_{\mathcal{K}}=\inf_{C\in\mathcal{K}}\|T+C\|.$$

[7] We use the notation $B_r(x)=\{y\in X \ : \ \|x-y\|<r\}$.

[8] Recall that an operator is compact iff the image of a bounded set is relatively compact, that is, if its closure is compact. It is an easy exercise to check that if $K\in\mathcal{K}$ ad $T\in L(X,X)$, then $TK, KT\in\mathcal{K}$ and $\mathcal{K}$ is closed in the operator topology.

Note that $\|T\|_{\mathcal{K}}$ induces a norm on Z with respect to which Z is a complete normed space.

Lemma F.2.13 *The measure of noncompactness and the ball measure of noncompactness satisfy the following properties:*

a) Let $A \subseteq X$ be a bounded set, then its closure $\bar{A}$ is compact if and only if $\tilde{\gamma}(A) = 0$. Also, $\bar{A}$ is compact if and only if $\gamma(A) = 0$.

b) An operator $T \in L(X, X)$ is compact if and only if $\tilde{\gamma}(T) = 0$. Also, T is compact if and only if $\gamma(T) = 0$.

c) $\gamma(T) \leq \|T\|$.

d) For bounded subsets $A, B \subseteq X$, we have $\gamma(A + B) \leq \gamma(A) + \gamma(B)$ and $\tilde{\gamma}(A + B) \leq \tilde{\gamma}(A) + \tilde{\gamma}(B)$.

e) For all $S, T \in L(X, X)$ we have

$$\tilde{\gamma}(ST) \leq \tilde{\gamma}(S)\tilde{\gamma}(T).$$

PROOF. a) For $\varepsilon > 0$, since $\bar{A}$ is compact, A can be covered by a finite number of balls of radius ε. Since ε is arbitrary, we have $\tilde{\gamma}(A) = 0$. Therefore $\gamma(A) = 0$, because $\gamma(A) \leq \tilde{\gamma}(A)$. Now assume that $\bar{A}$ is not compact, then there is a sequence $\{x_n\}_{n\in\mathbb{N}} \subseteq \bar{A}$ which has no accumulation points.[9] Let $S_{\varepsilon,n}$ be any collection of sets such that $x_n \in S_{\varepsilon,n}$ and the diameter of S_n is smaller than ε. Then there exist $n \in \mathbb{N}$ and $\varepsilon > 0$ such that, for any $m \geq n$, $S_{\varepsilon,n} \cap S_{\varepsilon,m} = \emptyset$. If not, then for any $n \in \mathbb{N}$ and $\varepsilon > 0$ there exists $m \geq n$ such that $|x_n - x_m| < 2\varepsilon$. Then, if we choose $\varepsilon = 2^{-k}$, $n_1 = 1$ and n_{k+1} such that $|x_{n_k} - x_{n_{k+1}}| < 2^{-k}$, then $\{x_{n_k}\}$ is a convergent subsequence of $\{x_n\}_{n\in\mathbb{N}}$ and therefore it has an accumulation point, contrary to the assumption. So we conclude that $\tilde{\gamma}(A) \geq \gamma(A) > \varepsilon$.

b) First suppose that T is a compact operator. For any bounded set $A \subseteq X$, $\overline{T(A)}$ is compact. So by (a), $\tilde{\gamma}(T(A)) = 0$ and $\gamma(T(A)) = 0$. Hence for any $k > 0$, T is a ball-k-set-contraction and a k-set-contraction. So $\tilde{\gamma}(T) = 0$ and $\gamma(T) = 0$.
Next, assume that $\gamma(T) = 0$. Let $A \subseteq X$, be a ball of radius $R > 0$. For $\varepsilon > 0$, we have $\gamma(T) < \frac{\varepsilon}{R}$. Therefore $\gamma(T(A)) < \frac{\varepsilon}{R}\gamma(A) < \varepsilon$. So $\gamma(T(A)) = 0$, then (a) implies $\overline{T(A)}$ is compact. So T is a compact operator. The same proof works for the case $\tilde{\gamma}(T) = 0$.

[9] We assume implicitly that $x_i = x_j$ implies $i = j$.

c) If $\gamma(A) = r$, then for $\lambda > r$, there is a covering of A by finitely many sets $\{B_i\}_{i=1}^n$ of diameter not greater than λ. So $\{T(B_i)\}_{i=1}^n$ will cover $T(A)$. For any $1 \le i \le n$

$$\operatorname{diam}(T(B_i)) = \sup_{x,y\in B_i} \|Tx - Ty\| \le \|T\| \sup_{x,y\in B_i} \|x - y\| \le \|T\|\lambda,$$

which implies $\gamma(T) \le \|T\|$.

d) Let $\gamma(A) = \alpha$ and $\gamma(B) = \beta$. Then for $r > \alpha$, there is a covering of A by a finite number of sets $\{a_i\}_{i=1}^n$ of diameter not greater than r and for $\rho > \beta$, there is a covering of B by a finite number of sets $\{b_j\}_{j=1}^m$ of diameter not greater than ρ. So $A + B = \{x + y\}_{x\in A, y\in B} \subseteq \cup_{i,j}\{x + y\}_{x\in a_i, y\in b_j}$. For any $1 \le i \le n, 1 \le j \le m$ and $x, x' \in a_i$, $y, y' \in b_j$ we have

$$\|x + y - x' - y'\| \le \|x - x'\| + \|y - y'\| \le r + \rho.$$

Therefore $\gamma(A + B) \le \gamma(A) + \gamma(B)$.

Now let $\tilde\gamma(A) = \kappa$ and $\tilde\gamma(B) = \lambda$. Then for $\mu > \kappa$, there is a covering of A by a finite number of balls $\{B(a_i, r_i)\}_{i=1}^n$ of radius $r_i \le \mu$ and for $\nu > \lambda$, there is a covering of B by a finite number of balls $\{B(b_j, \rho_j)\}_{j=1}^m$ of radius $\rho_j \le \nu$. So $A + B = \{x + y\}_{x\in A, y\in B} \subseteq \cup_{i,j}\{x + y\}_{x\in B(a_i,r_i), y\in B(b_j,\rho_j)}$. For any $1 \le i \le n, 1 \le j \le m$ and $x \in B(a_i, r_i)$, $y \in B(b_j, \rho_j)$ we have

$$\|x + y - (a_i + b_j)\| \le \|x - a_i\| + \|y - b_j\| \le \mu + \nu.$$

Therefore $\tilde\gamma(A + B) \le \tilde\gamma(A) + \tilde\gamma(B)$.

e) For all $S \in L(X, X)$, $A \subseteq X$, we have:

$$\tilde\gamma(S(A)) \le \tilde\gamma(S)\tilde\gamma(A)$$

Hence for all $S, T \in L(X, X)$, $A \subseteq X$

$$\tilde\gamma(ST(A)) \le \tilde\gamma(S)\tilde\gamma(T(A)) \le \tilde\gamma(S)\tilde\gamma(T)\tilde\gamma(A),$$

from which the claim follows. □

Lemma F.2.14 *Let X and Y be complex Banach spaces and $T \in L(X, Y)$. Then we have $\gamma(T^*) \le \tilde\gamma(T)$.*[10]

Proof. Suppose T is a ball-k-set-contraction. To show that T^* is a k-set-contraction, it suffices to show that if S is a set of diameter less than or

[10] By T^* we mean the dual operator: for all continuous linear functional $\ell \in Y'$ we have $T^*\ell \in X'$ where $T^*\ell(x) = \ell(Tx)$.

equal to d in Y^*, $T^*(S)$ can be covered by a finite number of sets of diameter less than or equal than $kd+\varepsilon$, for any $\varepsilon > 0$.
Consider $T(B)$, where $B = \{x \in X, \|x\| \leq 1\}$. Since $\tilde{\gamma}(B) \leq 1$ and T is a ball-k-set-contraction, $T(B)$ can be covered by a finite number of balls $B_{k+\frac{\varepsilon}{2d}}(y_i)$ in Y, $1 \leq i \leq n$, with centers at y_i, and radii $k + \frac{\varepsilon}{2d}$. Select M such that $\|y_i\| \leq M$, $1 \leq i \leq n$, and $\|y^*\| \leq M$ for all $y^* \in S$. Hence, we have $|y^*(y_i)| \leq M^2$ for each $y^* \in S$. Decompose the closed interval $[-M^2, M^2]$ into a union of disjoint intervals Δ_i, $1 \leq i \leq p$, of length less than $\frac{\varepsilon}{2}$. We consider an equivalence relation as follows: Given y_1^* and $y_2^* \in S$, write $y_1^* \sim y_2^*$ iff for each i, $1 \leq i \leq n$, $y_1^*(y_i)$ and $y_2^*(y_i)$ lie in the same interval $\Delta_{j(i)}$, $1 \leq j(i) \leq p$. Then we divide S into equivalence classes S_j, $1 \leq j \leq q$,
We claim that diameter $(T^*(S_i)) \leq kd + \varepsilon$. Take y_1^* and y_2^* in S_i. We have

$$\|T^*(y_1^*) - T^*(y_2^*)\| = \sup_{x \in B} |y_1^*(Tx) - y_2^*(Tx)| = \sup_{y \in T(B)} |y_1^*(y) - y_2^*(y)|.$$

If $y \in T(B)$, we know that $y \in B_{k+\frac{\varepsilon}{2}}(y_i)$ for some i, $1 \leq i \leq n$. It follows that

$$|y_1^*(y) - y_2^*(y)| \leq |y_1^*(y - y_i) - y_2^*(y - y_i)| + |y_1^*(y_i) - y_2^*(y_i)|$$

$$= |(y_1^* - y_2^*)(y - y_i)| + |y_1^*(y_i) - y_2^*(y_i)| \leq d(k + \frac{\varepsilon}{2d}) + \frac{\varepsilon}{2} = kd + \varepsilon.$$

Thus, for each $\varepsilon > 0$, $\|T^*(y_1^*) - T^*(y_2^*)\| \leq kd + \varepsilon$. This shows that diameter $(T^*(S_i)) \leq kd + \varepsilon$, and since $T^*(S) \subset \bigcup_{i=1}^q T^*(S_i)$, we have covered $T^*(S)$ by a finite number of sets of diameter less than or equal to $kd + \varepsilon$. □

Lemma F.2.15 *Let X be a complex Banach space and $T \in L(X, X)$. Assume that for some $n \geq 1, \tilde{\gamma}(T^n) < 1$. Then $R(\mathbb{1} - T)$ is closed and* $\dim(N(\mathbb{1} - T)) < \infty$.

Proof. The proof consists of two steps. First, we prove that if $A \subset X$ is closed and bounded, while $K \subset X$ is compact, then $((\mathbb{1} - T)^{-1}K) \cap A$ is compact.[11] Then we prove that this implies the claimed properties.[12]

Step 1: Let A be a closed, bounded subset of X and let K be a compact set. We prove that $K_1 = \{x \in A : (\mathbb{1} - T)x \in K\}$ is compact. By Lemma F.2.13-(a), in order to show that K_1 is compact, it suffices to show that $\tilde{\gamma}(K_1) = 0$. Notice that $\tilde{\gamma}(K_1)$ is defined, since A is bounded. Suppose $x \in K_1$, so that $x = Tx + m$ for some $m \in K$. Iterating we obtain

$$x = T^n x + \sum_{i=0}^{n-1} T^i m. \tag{F.2.4}$$

[11] A map such that the preimage of a compact set is compact is called *proper*.
[12] In fact, the proof of the second step implies in general that if S is proper, then $R(S)$ is closed and $\dim N(S) < \infty$.

If we write $K_* = \sum_{i=0}^{n-1} T^i(K)$, K_* is compact, since it is the continuous image of a compact set. Furthermore, (F.2.4) implies that $K_1 \subset T^n(K_1) + K_*$, so that $\tilde{\gamma}(K_1) \leq \tilde{\gamma}(T^n(K_1))$, by Lemma F.2.13-(a)-(d). Since T^n is a ball-k-set-contraction, $k < 1$, $\tilde{\gamma}(K_1) \leq k\tilde{\gamma}(K_1)$. It follows that $\tilde{\gamma}(K_1) = 0$.

Step 2: Let $S = (\mathbb{1} - T)$. By step 1, $N(S) = S^{-1}(0)$ is locally compact, consequently $N(S)$ is finite-dimensional by Theorem F.2.7. Next, we prove that $R(S)$ is closed.
Let $\{x_n\} \subset X$ be such that $\lim_{n\to\infty} Sx_n = y$, we want to show that $y \in R(S)$. Choose $z_n \in N(S)$ so that

$$d(x_n, N(S)) \geq \frac{1}{2}\|x_n - z_n\|.$$

We want to show that $\|x_n - z_n\|$ is bounded.
Suppose that, for some subsequence n_j, $\lim_{j\to\infty} \|x_{n_j} - z_{n_j}\| = \infty$, and define $\xi_n = \|x_n - z_n\|^{-1}(x_n - z_n)$. Then

$$\lim_{j\to\infty} S\xi_{n_j} = \lim_{j\to\infty} \|x_{n_j} - z_{n_j}\|^{-1} Sx_{n_j} = 0. \tag{F.2.5}$$

Since $K = \{S(\xi_{n_j})\} \cup \{0\}$ is compact, by Step 1 $S^{-1}(K) \cap \{x \in X \ : \ \|x\| \leq 1\}$ is compact as well. Consequently, $\{\xi_{n_j}\}$ is contained in a compact set and must have a convergent subsequence $\{\xi_{n_{j_k}}\}$. Let $\bar{\xi}$ be its limit. Equation (F.2.5) implies $S\bar{\xi} = 0$, that is $\bar{\xi} \in N(S)$. However, this is a contradiction since

$$\|\xi_n - \bar{\xi}\| \geq d(\xi_n, N(S)) = \|x_n - z_n\|^{-1} d(x_n, N(S)) \geq \frac{1}{2}.$$

As claimed, $\sup_{n\in\mathbb{N}} \|x_n - z_n\| \leq M$ for some $M \in \mathbb{R}$. By Step 1 again,

$$K_* := S^{-1}(\{Sx_n\} \cup \{y\}) \cap \{z \in Z \ : \ \|z\| \leq M\}$$

is compact. Consequently, $\{x_n - z_n\} \subset K_*$ has a convergent subsequence $\{x_{n_j} - z_{n_j}\}$. Let η be its limit. By continuity

$$S(\eta) = \lim_{j\to\infty} S(x_{n_j} - z_{n_j}) = \lim_{j\to\infty} S(x_{n_j}) = y$$

wereby proving that $R(S)$ is closed.

□

F.3 Nussbaum formula

In this section, we obtain a characterization of the essential spectral radius $r_e = \sup\{|\lambda| : \lambda \in \sigma_{ess}(T)\}$. We essentially follow [Nus70].

Lemma F.3.1 *Let X be a complex Banach space and $T \in L(X,X)$. Let $r'_e := \inf\{(\tilde{\gamma}(T^n))^{\frac{1}{n}} \ : \ n \in \mathbb{N}\}$. Then*

$$r'_e = \lim_{n\to\infty} (\tilde{\gamma}(T^n))^{\frac{1}{n}} = \lim_{n\to\infty} (\gamma(T^n))^{\frac{1}{n}}.$$

Furthermore, if $|\lambda| > r'_e$, then $\dim(N(\lambda - T)) < \infty$ and $R(\lambda - T)$ is closed.

PROOF. We start showing that $\limsup_{n\to\infty}(\tilde{\gamma}(T^n))^{\frac{1}{n}} \leq r'_e$.

For any $\varepsilon > 0$, choose m such that $(\tilde{\gamma}(T^m))^{\frac{1}{m}} \leq r'_e + \varepsilon$. For large enough n, write $n = pm + q$ where $0 \leq q \leq (m-1)$.
Then, by the above fact and reaclling Lemma F.2.13-(e), we obtain

$$(\tilde{\gamma}(T^n))^{\frac{1}{n}} \leq (\tilde{\gamma}(T^m))^{\frac{p}{n}} \cdot (\tilde{\gamma}(T))^{\frac{q}{n}} \leq (r'_e + \varepsilon)^{\frac{pm}{n}} (\tilde{\gamma}(T))^{\frac{q}{n}}.$$

Taking the limit $n \to \infty$ yields $\limsup_{n\to\infty}(\tilde{\gamma}(T^n))^{\frac{1}{n}} \leq r'_e + \varepsilon$. Since ε was arbitrary, we have proved $\limsup_{n\to\infty}(\tilde{\gamma}(T^n))^{\frac{1}{n}} \leq r'_e \leq \liminf_{n\to\infty}(\tilde{\gamma}(T^n))^{\frac{1}{n}}$. Therefore $\lim_{n\to\infty}(\tilde{\gamma}(T^n))^{\frac{1}{n}}$ exists and equals r'_e. In the exact same way, we can prove that $\lim_{n\to\infty}(\gamma(T^n))^{\frac{1}{n}}$ exists.
Suppose $|\lambda| > r'_e$ and n is such that $(\tilde{\gamma}(T^n))^{\frac{1}{n}} < |\lambda|$. Consider $T_1 = (\frac{1}{\lambda})T$ and notice that $\tilde{\gamma}(T_1^n) = (\frac{1}{|\lambda|^n})\tilde{\gamma}(T^n) = k < 1$. By Lemma F.2.15, $R(\mathbb{1} - T_1)$ is closed and $\dim(N(\mathbb{1} - T_1)) < \infty$. □

Lemma F.3.2 *If $|\lambda_0| > r'_e$, then λ_0 is not a limit point of $\sigma(T) \setminus \{\lambda_0\}$.*

PROOF. We show that all points $\lambda \neq \lambda_0$, in some neighborhood of the point λ_0, belong to the resolvent of T and so λ_0 is not a limit point of $\sigma(T)$. The case $\lambda_0 \in \rho(T)$ is trivial. Let $\lambda_0 \in \sigma(T)$. First we prove that either $N(\lambda_0 - T) \neq \{0\}$ or $N(\lambda_0 - T^*) \neq \{0\}$.

Suppose that $N(\lambda_0 - T) = N(\lambda_0 - T^*) = \{0\}$. Then $(\lambda_0 - T)^{-1} : D \to X$ exists on $D = R(\lambda_0 - T)$ which is closed, by Lemma F.3.1. Assume that $D \neq X$, then by Lemma F.2.5, there is $u \in X$, such that $\|u\| = 1$ and $\|u - w\| \geq \frac{1}{2}$ for any $w \in D$. Let $V := span\{u, D\}$, then for any $v \in V$ we can write $v = \alpha u + w$ with $w \in D$. Define $l(v) := \alpha$, then

$$\|v\| = \|\alpha u + w\| = |\alpha|\|u - (-\alpha^{-1}w)\| \geq \frac{1}{2}|\alpha| = \frac{1}{2}|l(v)|.$$

So

$$|l(v)| \leq 2\|v\|.$$

We can then apply the Hahn-Banach theorem to produce an extension of l on all of X and $l \neq 0$, since $l(u) = 1$. For any $v \in X$,

$$(\lambda_0 - T^*)l(v) = l\big((\lambda_0 - T)v\big) = 0.$$

This contradicts $N(\lambda_0 - T^*) = \{0\}$. So $D = X$, which implies that $\lambda_0 - T$ is invertible on X and by the bounded inverse theorem, $(\lambda_0 - T)^{-1}$ is a bounded operator. Therefore $\lambda_0 \notin \sigma(T)$ and this contradicts the assumption.

Suppose that there exists a sequence $\{\tilde{\lambda}_n\}_{n=1}^{\infty} \subset \sigma(T) \setminus \{\lambda_0\}$, $\lambda_i \neq \lambda_j$ for $i \neq j$, which accumulates to λ_0. Then there are either infinitely many $\tilde{u}_n \in N(\tilde{\lambda}_n - T)$ or infinitely many $\tilde{l}_n \in N(\tilde{\lambda}_n - T^*)$. For each $\varepsilon > 0$, let $n_\varepsilon \in \mathbb{N}$ such that, for $n > n_\varepsilon$, $|\tilde{\lambda}_n - \lambda_0| < \varepsilon|\lambda_0|$.

In the first case, for any $k \in \mathbb{N}$, let M_k be the subspace spanned by the vectors $\tilde{u}_{n_\varepsilon}, \cdots, \tilde{u}_{n_\varepsilon+k}$. Set $u_k := \tilde{u}_{n_\varepsilon+k}$ and $\lambda_k := \tilde{\lambda}_{n_\varepsilon+k}$. Note that the $u_1, u_2, \cdots$ are linearly independent. This can be proven by induction, indeed if $\{u_1, \ldots, u_{n+1}\}$ are linearly dependent, then either also $\{u_1, \ldots, u_n\}$ are lienarly dependent or $\sum_{i=1}^{n+1} \alpha_i u_i = 0$, with $\alpha_1 \neq 0$. Then

$$0 = \sum_{j=1}^{n+1} (\lambda_i \alpha_i u_i + T\alpha_i u_i) = \sum_{j=1}^{n+1} \lambda_i \alpha_i u_i.$$

But this implies $\sum_{j=1}^{n} (\lambda_i - \lambda_n)\alpha_i u_i = 0$, which again implies that $\{u_1, \ldots, u_n\}$ are linearly dependent. Accordingly, each M_{k-1} is a closed proper subspace of M_k. So, by Lemma F.2.5, there exists $v_k \in M_k$, such that $\|v_k\| = 1$ and $d(v_k, M_{k-1}) \geq 1 - \varepsilon$.

Note that $v_k = \alpha_k u_k + w_k$ where $\alpha_k \in \mathbb{R}, w_k \in M_{k-1}$. So for $k, r, s \in \mathbb{N}$, such that $s > k$,

$$\|T^r v_s - T^r v_k\| = \|T^r(\alpha_s u_s) + T^r w_s - T^r v_k\| = \|\alpha_s \lambda_s^r u_s + T^r w_s - T^r v_k\|$$

$$= |\lambda_s^r| \|v_s - (w_s - \lambda_s^{-r} T^r w_s + \lambda_s^{-r} T^r v_k)\| \geq |\lambda_s^r|(1-\varepsilon) = |(\lambda_s - \lambda_0 + \lambda_0)^r|(1-\varepsilon)$$

$$= |\lambda_0^r| \left|1 + \frac{\lambda_s - \lambda_0}{\lambda_0}\right|^r (1-\varepsilon) \geq |\lambda_0|^r \left(1 - \left|\frac{\lambda_s - \lambda_0}{\lambda_0}\right|\right)^r (1-\varepsilon) \geq |\lambda_0|^r (1-\varepsilon)^{r+1}.$$

This implies that $T^r\{|v| \leq 1\}$ cannot be covered by finitely many sets of diameter $\frac{1}{4}|\lambda_0|^r(1-\varepsilon)^{r+1}$. Therefore, by the arbitrariness of ε, $\tilde{\gamma}(T^r) \geq \gamma(T^r) \geq \frac{1}{4}|\lambda_0|^r$.

In the second case, exactly the same argument implies $\gamma(T^{*r}) \geq \frac{1}{4}|\lambda_0|^r$. By Lemma F.2.14, $\tilde{\gamma}(T^r) \geq \frac{1}{4}|\lambda_0|^r$.

Thus in both cases, $r'_e = \inf_n(\tilde{\gamma}(T^n))^{\frac{1}{n}} \geq |\lambda_0|$ which contradicts the assumption. So λ_0 is not a limit point of $\sigma(T)$. □

Corollary F.3.3 *According to the Definition F.2.1 of the essential spectrum, Lemmata F.3.1 and F.3.2 imply that* $r'_e \geq r_e$.[13]

Lemma F.3.4 *Let* $T \in L(X, X)$ *and* $r > r_e(T)$. *Then there exists a finite dimensional linear operator* F *such that* $\sigma(T + F) \subset \{\lambda : |\lambda| \leq r\}$.

[13] See (F.2.3) for the defintion of r_e.

PROOF. Since $\sigma(T) \cap \{\lambda : |\lambda| \geq r\}$ is a compact set of isolated points, it consists of a finite number of points $\lambda_1, \cdots, \lambda_n$. Let C_i be a small circle about λ_i, $C_i \cap C_j = \emptyset$ for $i \neq j$ and containing only λ_i from $\sigma(T)$, and

$$P_i = \frac{1}{2\pi i} \int_{C_i} (\xi - T)^{-1} d\xi$$

be the Riesz projector associated with λ_i. Let $P = \sum_{i=1}^{n} P_i$, and $F = -TP$. By Lemma F.2.2, $\dim(R(P)) < \infty$ and

$$\sigma(T + F) = \sigma(T(\mathbb{1} - P)) \subset [\sigma(T) \setminus \{\lambda_i\}] \cup \{0\},$$

which implies the Lemma. □

The following lemma provides the desired characterization of r_e.

Lemma F.3.5 *Let X be a complex Banach space and $T \in L(X, X)$. Then*

$$\lim_{n\to\infty} (\gamma(T^n))^{\frac{1}{n}} = \lim_{n\to\infty} (\tilde{\gamma}(T^n))^{\frac{1}{n}} = \lim_{n\to\infty} (\|T^n\|_{\mathcal{K}}))^{\frac{1}{n}} = r_e.$$

PROOF. By Lemma F.3.1 the first two limits equal r'_e. The same argument as in Lemma F.3.1 shows that $r''_e := \lim_{n\to\infty} \|T^n\|_{\mathcal{K}}^{\frac{1}{n}}$ exists.
For $S \in L(X, X)$ and any compact operator $K \in \mathcal{K}(X)$, by Lemma F.2.13,

$$\gamma(S) = \gamma(S + K) \leq \|S + K\|.$$

Therefore $\gamma(S) \leq \|S\|_{\mathcal{K}}$, which implies $r'_e \leq r''_e$.

To conclude, we show that $r''_e \leq r_e$. Suppose $r_e < r''_e$, and let $r \in (r_e, r''_e)$. For this r, let F be as in Lemma F.3.4 and write $T_1 = T + F$. Then $\lim_{n\to\infty} \|T_1^n\|^{\frac{1}{n}} \leq r$ (if unclear, see Problem C.13). On the other hand, $\|T^n\|_{\mathcal{K}} \leq \|T_1^n\|$, so that we obtain $r''_e = \lim_{n\to\infty} \|T^n\|_{\mathcal{K}}^{\frac{1}{n}} \leq r$, a contradiction. It follows that $r''_e \leq r_e$. Then, Corollary F.3.3 implies $r_e = r'_e = r''_e$. □

F.4 Hennion's theorem and its generalizations

We first prove Hennion's theorem, then provide a more recent generalization.

In fact, the next Theorem is itself a small generalization of [Hen93], since it allows the weak norm to be just a semi-norm. A similar generalization is contained in [HH01, Theorem XIV.3]. To this end, we need a bit of notation: given a vector space X and a semi-norm $\|\cdot\|_w$, we call $X_{0,w}$ the space X equipped with the topology induced by the semi-norm. Next, we can consider the vector space X_w of the equivalence classes with respect to the semi-norm (i.e. $x \sim y$ iff $\|x - y\|_w = 0$). We can define the norm $\|\tilde{x}\|' = \inf_{x\in\tilde{x}} \|x\|$. This yields a Banach space X_w, as it can be checked directly.

Problem F.1 *Given a normed space Y, a Banch space X togheter with a seminorm $\|\cdot\|_w$, and and operator $T \in L(Y, X_{0,w})$, show the they canonically induce an operator $\tilde{T} : L(Y, X_w)$.*

Definition F.4.1 *Using the notation of Problem F.1, we say that $T \in L(Y, X_{0,w})$ is $\|\cdot\|_w$-compact if $\tilde{T}(B)$ is compact.*

Theorem F.4.2 ([Hen93]) *Let $(X, \|\cdot\|)$ be a Banach space and $T \in L(X, X)$. Assume that there exists a continuous[14] semi-norm $\|\cdot\|_w$ on X, and $M > \theta > 0$, $A, B, C > 0$, such that, for all $n \in \mathbb{N}$ and $f \in X$,[15]*

$$\|T^n f\|_w \leq CM^n \|f\|_w; \quad \|T^n f\| \leq A\theta^n \|f\| + BM^n \|f\|_w \,.$$

Then the spectral radius of $T \in L(X, X)$ is bounded by M. If, in addition, T is $\|\cdot\|_w$-compact, then the essential spectral radius of T is bounded by θ.

Proof. Continuity of the semi-norm implies that there exists $C' > 0$ such that $\|f\|_w \leq C'\|f\|$ for all $f \in \mathcal{B}$. For if not, then for any $n \in \mathbb{N}$, there must exist $f_n \in \mathcal{B}$ with $\|f_n\| = 1$, but $\|f_n\|_w \geq n$. But then $\|\frac{1}{n} f_n\| \to 0$ while $\|\frac{1}{n} f_n\|_w \geq 1$, contradicting continuity of the semi-norm.

This fact plus the second inequality yields, for all $n \in \mathbb{N}$ and $f \in \mathcal{B}$,

$$\|T^n f\| \leq (A + BC')M^n \|f\|. \tag{F.4.6}$$

By the spectral radius formula, see Problem C.13, we conclude the spectral radius is bounded by M.

For the second part, by Lemma F.3.5, and recalling Defintion F.2.10, we have

$$r_e = \lim_{n\to\infty} \sqrt[n]{\tilde{\gamma}(T^n)} \leq \lim_{n\to\infty} \sqrt[n]{\tilde{\gamma}(T^n B_1)}$$

where $B_1 := \{f \in X \;\mid\; \|f\| \leq 1\}$.
Next we prove that $T^n B_1$ can be covered by a finite number of balls of radius $C_\sharp \cdot \theta^n$, which implies that

$$r_e \leq \lim_{n\to\infty} \sqrt[n]{\tilde{\gamma}(T^n B_1)} \leq \lim_{n\to\infty} \sqrt[n]{C_\sharp \cdot \theta^n} = \theta.$$

By hypothesis, $\tilde{T}B_1$ is relatively compact in X_w. Thus, for each $\varepsilon > 0$ we can extract a finite sub-cover $\{\tilde{B}_\varepsilon(\tilde{f}_i)\}_{i=1}^{N_\varepsilon}$ from the covering $\{\tilde{B}_\varepsilon(\tilde{f})\}_{\tilde{f}\in\tilde{T}B_1}$, where $\tilde{B}_\varepsilon(\tilde{f}) = \{\tilde{g} \in \bar{X}_w \;:\; \|\tilde{g} - \tilde{f}\|'_w < \varepsilon\}$. Then, choosing[16] $f_i \in \tilde{f}_i \cap TB_1$ and

[14] By continuous, we mean that if $(f_n)_n \subset \mathcal{B}$ is a sequence such that $\|f_n\| \to 0$, then necessarily $\|f_n\|_w \to 0$.

[15] These are often called Lasota-Yorke (or Doeblin-Fortet) inequalities.

[16] Recall that elements of X_w are equivalence classes of elements in X.

setting $U_\varepsilon(f_i) = \{f \in X \ : \ \|f - f_i\|_w < \varepsilon\} = \{f \in \tilde{f} \ : \ \tilde{f} \in \tilde{B}_\varepsilon(\tilde{f}_i)\}$ we have a finite covering of TB_1. Accordingly, for each $f \in U_\varepsilon(f_i) \cap TB_1$ we have

$$\begin{aligned}\|T^{n-1}(f - f_i)\| &\leq A\theta^{n-1}\|f - f_i\| + BM^{n-1}\|f - f_i\|_w \\ &\leq A\theta^{n-1}2(A\theta + BC'M) + BM^{n-1}\varepsilon.\end{aligned}$$

where we have used equation (F.4.6). Choosing ε sufficiently small we can conclude that for each $n \in \mathbb{N}$ the set $T^n(B_1)$ can be covered by a finite number of $\|\cdot\|$-balls of radius $C_\sharp \cdot \theta^n$ centered at the points $\{T^{n-1}f_i\}_{i=1}^{N_\varepsilon}$. □

To conclude the appendix, we show that the hypotheses of the above theorem can be further weakened to situations in which T is not necessarily continuous with respect to the weak norm.[17]

Theorem F.4.3 ([BGK07]) *Let $(X, \|\cdot\|)$ be a Banach space and $T \in L(X, X)$. Assume that there exists a semi-norm $\|\cdot\|_w$ on X such that any bounded sequence in $\|\cdot\|$ contains a Cauchy sequence for $\|\cdot\|_w$. If there exist $n_0 \in \mathbb{N}$ and $\theta, B > 0$ such that,*

$$\|T^{n_0}f\| \leq \theta^{n_0}\|f\| + B\|f\|_w, \tag{F.4.7}$$

then the essential spectral radius of T is bounded by θ.

Proof. Note that there must exist $C > 0$ such that $\|f\|_w \leq C\|f\|$. If not then there would be a sequence $\{f_n\}$, $\|f_n\| \leq 1$, such that $\lim_{n\to\infty} \|f_n\|_w = \infty$, but this contradicts that f_n must have a Cauchy subsequence.

Let $M = 2\|T\|$, then we can define the new seminorm,

$$\|f\|'_w := (2C)^{-1} \sum_{n=0}^{\infty} M^{-n}\|T^n f\|_w.$$

Note that

$$\begin{aligned}\|f\|'_w &\leq \frac{1}{2}\sum_{n=0}^{\infty} M^{-n}\|T^n f\| \leq \frac{1}{2}\sum_{n=0}^{\infty} 2^{-n}\|f\| = \|f\| \\ \|Tf\|'_w &\leq (2C)^{-1}\sum_{n=0}^{\infty} M^{-n}\|T^{n+1}f\|_w \\ &= (2C)^{-1}M\sum_{n=1}^{\infty} M^{-n}\|T^n f\|_w \leq M\|f\|'_w.\end{aligned} \tag{F.4.8}$$

[17] Indeed, note that the first displayed inequality in F.4.2 amounts simply to the continuity of T in the weak norm.

Thus, if we set $A = M^{n_0}\theta^{-n_0}$, for each $n \in \mathbb{N}$ we can write $n = kn_0 + m$, $m < n_0$, and, iterating F.4.7,

$$\begin{aligned}\|T^n f\| &\le \theta^{kn_0} M^m \|f\| + \sum_{j=0}^{k-1} B\theta^{(k-1-j)n_0} \|T^{jn_0+m} f\|_w \\ &\le \theta^{kn_0} M^m \|f\| + B \max\{\theta^{(k-1-j)n_0} M^{jn_0+m}\} \|f\|'_w \\ &\le A\theta^n \|f\| + BM^n \|f\|'_w\end{aligned}$$

since it must be that $\theta \le \|T\| = M/2$.

Next, if $\{f_n\}$ is bounded in the $\|\cdot\|$ norm, so are the sequences $T^m f_n$, $m \in \mathbb{N}$. Then, by hypothesis, we can extract a sequence n_j^1 such that $f_{n_j^1}$ is Cauchy in the $\|\cdot\|_w$ norm. From it we can extract a sequence n_j^2, with $n_1^2 = n_1^1$, such that $Tf_{n_j^2}$ is Cauchy in the $\|\cdot\|_w$ norm, and so on. Note that, by construction, $n_j^j = n_j^m$ for $m \ge j$. Then the sequence n_j^j is such that $T^m f_{n_j^j}$ is Cauchy in the $\|\cdot\|_w$ norm for all $m \in \mathbb{N}$. Then, for each $\varepsilon > 0$, if $(2C)^{-1}2^{-L} < \varepsilon/2$, then, by the definition of the norm $\|\cdot\|'_w$, we can write

$$\|f_{n_j^j} - f_{n_k^k}\|'_w \le (2C)^{-1} \sum_{m=0}^{L} M^{-m} \|T^m (f_{n_j^j} - f_{n_k^k})\|_w + \varepsilon/2.$$

It follows that there exists $m \in \mathbb{N}$ such that, if $j, k \ge m$, then $\|f_{n_j^j} - f_{n_k^k}\|'_w \le \varepsilon$, i.e. we can extract a Cauchy sequence in the $\|\cdot\|'_w$ norm. So the $\|\cdot\|'_w$ norm has the same property as the $\|\cdot\|_w$ norm. This implies that T is a $\|\cdot\|'_w$-compact operator. The statement follows then from Theorem F.4.2. □

The paper [BGK07] provides an application of Theorem F.4.3 to prove a local limit theorem for weakly coupled lattices of expanding maps in which the relevant operators are indeed not continuous in the weak norm. For more details, see [BGK07, Section 3].

Appendix G

Probability–the minimum

This appendix is intended to provide the minimum of probability theory needed in this Book. Of course, there are wonderful books to study probability (e.g., on one extreme, the monumental [Fel67],[Fel66], on the other, the synthetic but really deep [Var01]), but they require some effort to read as they contain much more material than needed here.

G.1 Distribution and Characteristic Functions

Let X be a measurable space and μ a probability measure. For any measurable set A will use the notation $\mathbb{P}(A) = \mu(A) = \int_A d\mu$ for its probability. Also, given a measurable function (*random variable*) φ, we will write $\mathbb{E}(\varphi) = \mu(\varphi) = \int_X \varphi d\mu$ for its expectaion. Given a random variable φ we define the *distribution function*

$$F(x) = \mathbb{P}(\{y \in X \ : \ \varphi(y) \leq x\}).$$

Note that F is an increasing function, and $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$. Also note that F defined a measure ν_F on $\mathbb{R}$, such that

$$F(x) = \nu_F(\{y \in \mathbb{R} \ : \ y \leq x)\}.$$

Another fundamental object is the *characteristic function*

$$\phi(t) = \mathbb{E}(e^{it\varphi}) = \int_X e^{it\varphi(x)}\mu(dx) = \int_{\mathbb{R}} e^{iz}\nu_F(dz) = \int_{\mathbb{R}} e^{iz}F(dz), \qquad \text{(G.1.1)}$$

where the last is a Riemann–Stieltjes integral. Note that, since μ is a probability measure $\phi(0) = 1$. In addition, if F is differentiable and F' is Riemann

integrable,[1] then

$$\phi(t) = \int_{\mathbb{R}} e^{izt} F'(z) dz. \tag{G.1.2}$$

In this case, if ϕ is integrable, the usual Fourier inversion formula yields

$$F'(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-ixt} \phi(t) dt. \tag{G.1.3}$$

More generally, we have the following result.

Lemma G.1.1 ([Var01, page 20]) *Given the distribution function F and the charatcteristic function ϕ of a random variable φ, we have*

$$F(b) - F(a) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \phi(t) \frac{e^{-ibt} - e^{-iat}}{-it} dt$$

for all points a, b of continuity of F.

Let us mention another useful fact.

Lemma G.1.2 *If $\mathbb{E}(|\varphi|) < \infty$, then ϕ is differentiable at zero. In addition,*

$$\phi'(0) = i\mathbb{E}(\varphi).$$

Proof. Let us compute the derivative

$$\begin{aligned}\lim_{h\to 0} \frac{\phi(h) - \phi(0)}{h} &= \lim_{h\to 0} \frac{1}{h} \mathbb{E}(e^{ih\varphi} - 1) = \lim_{h\to 0} \mathbb{E}\left(i\varphi \frac{1}{h} \int_0^h e^{i\xi\varphi} d\xi \right) \\ &= \lim_{h\to 0} \mathbb{E}\left(i\varphi \frac{1}{h} \int_0^h \left[e^{i\xi\varphi} - 1 \right] d\xi \right) + \mathbb{E}(i\varphi).\end{aligned}$$

Next, let $\beta < \frac{1}{2}$, and let $A_h = \{\varphi \geq h^{-\beta}\}$. Then, since φ is integrable, it must be[2]

$$\lim_{h\to 0} \mathbb{E}(\mathbb{1}_{A_h} |\varphi|) = 0.$$

We can then estimate

$$\left| \mathbb{E}\left(i\varphi \frac{1}{h} \int_0^h \left[e^{i\xi\varphi} - 1 \right] d\xi \right) \right| \leq 2\mathbb{E}(\mathbb{1}_{A_h} |\varphi|) + \mathbb{E}(\mathbb{1}_{A_h^c} |\varphi|^2 h) \leq 2\mathbb{E}(\mathbb{1}_{A_h} |\varphi|) + h^{1-2\beta},$$

from which the Lemma follows. □

[1] Note that, since F is increasing, F' is defined almost everywhere.

[2] If unsure, consider the set $\Gamma_n = \{|\varphi| \in [2^n, 2^{n+1})\}$, then $\sum_{n\in\mathbb{N}} 2^n |\Gamma_n| \leq \mathbb{E}(|\varphi|) < \infty$. Thus, for each $L \in \mathbb{R}$,

$$\mathbb{E}(\mathbb{1}_{\varphi > L} |\varphi|) \leq \sum_{n \geq \ln_2 L} 2^{n+1} |\Gamma_n|.$$

In applications, we often try to estimate the characteristic function. It is then natural to try to investigate how an error in the characteristic function reflects on our knowledge of the distribution function.

Lemma G.1.3 ([Fel66, equation (3.13) of Chapter XVI.3]) *Let ϕ and Φ be two characteristic functions and let F and G be the corresponding characteristic functions. Assume that F is the distribution function of an integrable random variable and G is differentiable with $\|G'\|_\infty \leq M$ for some $M \in \mathbb{R}$. Then, for all $T \in \mathbb{R}_+$,*[3]

$$|F(x) - G(x)| \leq \frac{1}{\pi} \int_{-T}^{T} \left| \frac{\phi(\xi) - \Phi(\xi)}{\xi} \right| d\xi + \frac{12M}{\pi T}.$$

Proof. Let $T > 0$ and $\hat{\omega}_T \in \mathcal{C}^0(\mathbb{T}, \mathbb{R})$ such that $\text{supp}(\omega_T) \subset [-T, T]$. Also, let ω_T be the inverse Fourier transform of $\hat{\omega}_T$. Since ω_T is an analytic function, the convolutions $F_T := F \star \omega_T$ and $G_T := G \star \omega_T$ are smooth functions. Also, by formula (G.1.2),

$$\begin{aligned}\phi_T(t) &= \int_{\mathbb{R}} e^{izt} \int_{\mathbb{R}} F(x)\omega_T'(z-x)dxdz = \int_{\mathbb{R}} e^{izt} \int_{\mathbb{R}} \omega_T(z-x)dF(x)dz \\ &= \hat{\omega}_T(t) \int_{\mathbb{R}} e^{ixt}dx = \phi(t)\hat{\omega}_T(t).\end{aligned}$$

Analogously, $\Phi_T(t) = \Phi(t)\hat{\omega}_T(t)$. Then folumula (G.1.3) applies and yields

$$F_T'(x) - G_T'(x) = \frac{1}{2\pi} \int_{-T}^{T} e^{-ixt} \left[\phi(t) - \Phi(t)\right] \hat{\omega}_T(t)dt.$$

Integrating with respect to x yields

$$\begin{aligned}F_T(x) - G_T(x) &= \frac{1}{2\pi} \lim_{S\to\infty} \int_{-T}^{T} \int_{-S}^{x} e^{-i\xi t} \left[\phi(t) - \Phi(t)\right] \hat{\omega}_T(t)d\xi dt \\ &= \frac{1}{2\pi} \lim_{S\to\infty} \int_{-T}^{T} \frac{e^{iSt} - e^{-ixt}}{it} \left[\phi(t) - \Phi(t)\right] \hat{\omega}_T(t)dt \qquad \text{(G.1.4)} \\ &= \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ixt}}{-it} \left[\phi(t) - \Phi(t)\right] \hat{\omega}_T(t)dt,\end{aligned}$$

where, in the last line, we have used the Riemann-Lebesgue Lemma, also the integrand is a bounded function since $\phi(0) = \Phi(0) = 1$ and, recalling Lemma G.1.2, $\phi - \Phi$ is differentiable at zero.
Next, we want to estimate the difference between $F_T(x) - G_T(x)$ and $F(x) - G(x)$. For each $\varepsilon > 0$, let x_ε be such that

$$\|F - G\|_\infty \leq |F(x_\varepsilon) - G(x_\varepsilon)| + \varepsilon.$$

[3] Humm, Feller has 24 rather than 12, maybe I lost a 2 somewhere.

We discuss only the case $F(x_\varepsilon) - G(x_\varepsilon)$, the other one can be treated in the same way with the obvious changes. To compute it is convenient to make an explicit choice of ω_T. Following [Fel66] we choose

$$\omega_T = \frac{1 - \cos Tx}{\pi T x^2}$$

To check the the Fourier trasfom is supported as required solve the following probem.

Problem G.1 *Show that*

$$\hat{\omega}_T(t) = \begin{cases} 1 - \frac{|t|}{T} & \text{for } |t| \leq T \\ 0 & \text{otherwise.} \end{cases}$$

For $x \geq x_\varepsilon$, recalling that F is increasing, we have

$$F(x) - G(x) \geq F(x_\varepsilon) - G(x_\varepsilon) + G(x_\varepsilon) - G(x) \geq F(x_\varepsilon) - G(x_\varepsilon) - M(x - x_\varepsilon).$$

We can then choose $x_1 = x_\varepsilon + \frac{F(x_\varepsilon) - G(x_\varepsilon)}{2M} =: x_\varepsilon + A$ and compute

$$\begin{aligned}
\|F_T - G_T\|_\infty \geq |F_T(x_1) - G(x_1)| &= \left| \int_{\mathbb{R}} [F(x) - G(x)] \omega_T(x_1 - x) dx \right| \\
&\geq \frac{1}{2}(F(x_\varepsilon) - G(x_\varepsilon)) \int_{x_1 - A}^{x_1 + A} \omega_T(x_1 - x) dx \\
&\quad - [F(x_\varepsilon) - G(x_\varepsilon) + \varepsilon] \int_{|x_1 - x| \geq A} \omega_T(x_1 - x) dx \\
&\geq \frac{1}{2}(F(x_\varepsilon) - G(x_\varepsilon)) \int_{-\infty}^{\infty} \frac{1 - \cos Tx}{\pi T x^2} dx \\
&\quad - \left[\frac{3}{2}(F(x_\varepsilon) - G(x_\varepsilon)) + \varepsilon \right] \int_{|x| \geq A} \frac{1}{\pi T x^2} dx \\
&\geq \frac{F(x_\varepsilon) - G(x_\varepsilon)}{2\pi} \int_{-\infty}^{\infty} \frac{1 - \cos x}{x^2} dx - \frac{3(F(x_\varepsilon) - G(x_\varepsilon)) + 2\varepsilon}{\pi T A}
\end{aligned}$$

To conclude, solve this problem.

Problem G.2 *Show that*

$$\int_{-\infty}^{\infty} \frac{1 - \cos x}{x^2} dx = \pi.$$

The above, since ε is arbitrary, implies

$$\|F - G\|_\infty \leq 2\|F_T - G_T\|_\infty + \frac{12M}{\pi T}$$

which, together with (G.1.4), implies the Lemma. □

Bibliography

[AA68] V. I. Arnold and A. Avez. *Ergodic problems of classical mechanics*. Translated from the French by A. Avez. W. A. Benjamin, Inc., New York-Amsterdam, 1968.

[Aar97] Jon Aaronson. *An introduction to infinite ergodic theory*, volume 50 of *Mathematical Surveys and Monographs*. American Mathematical Society, 1997.

[Arn83] V. I. Arnold. *Geometrical methods in the theory of ordinary differential equations*, volume 250 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Science]*. Springer-Verlag, New York, 1983. Translated from the Russian by Joseph Szücs, Translation edited by Mark Levi.

[Arn99] V. I. Arnold. *Mathematical methods of classical mechanics*, volume 60 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 199? Translated from the 1974 Russian original by K. Vogtmann and A. Weinstein, Corrected reprint of the second (1989) edition.

[Arn92] Vladimir I. Arnold. *Ordinary differential equations*. Springer Textbook. Springer-Verlag, Berlin, 1992. Translated from the third Russian edition by Roger Cooke.

[Bal00a] Viviane Baladi. *Positive transfer operators and decay of correlations*, volume 16 of *Advanced Series in Nonlinear Dynamics*. World Scientific Publishing Co. Inc., River Edge, NJ, 2000.

[Bal00b] Viviane Baladi. *Positive Transfer Operators and Decay of Correlations*, volume 16 of *Advanced Series in Nonlinear Dynamics*. World Scientific, Singapore, 2000.

[BGK07] Jean-Baptiste Bardet, Sébastien Gouëzel, and Gerhard Keller. Limit theorems for coupled interval maps. *Stoch. Dyn.*, 7(1):17–36, 2007.

[Bir57] Garrett Birkhoff. Extensions of Jentzsch's theorem. *Trans. Amer. Math. Soc.*, 85:219–227, 1957.

[Bir79] Garrett Birkhoff. *Lattice theory*, volume 25 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, R.I., third edition, 1979.

[BK83] Michael Brin and Anatole Katok. On local entropy. In *Geometric dynamics (Rio de Janeiro, 1981)*, volume 1007 of *Lecture Notes in Mathematics*, pages 30–38. Springer, Berlin, 1983.

[Bro61] Felix E. Browder. On the spectral theory of elliptic differential operators. I. *Math. Ann.*, 142:22–130, 1960/61.

[BT07] Viviane Baladi and Masato Tsujii. Anisotropic Hölder and Sobolev spaces for hyperbolic diffeomorphisms. *Ann. Inst. Fourier (Grenoble)*, 57(1):127–154, 2007.

[BY93] Viviane Baladi and Lai-Sang Young. On the spectra of randomly perturbed expanding maps. *Comm. Math. Phys.*, 156:355–385, 1993.

[CC95] Alessandra Celletti and Luigi Chierchia. A constructive theory of Lagrangian tori and computer-assisted applications. In *Dynamics reported*, volume 4 of *Dynam. Report. Expositions Dynam. Systems (N.S.)*, pages 60–129. Springer, Berlin, 1995.

[CH82] Shui Nee Chow and Jack K. Hale. *Methods of bifurcation theory*, volume 251 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Science]*. Springer-Verlag, New York, 1982.

[CL55] Earl A. Coddington and Norman Levinson. *Theory of ordinary differential equations*. McGraw-Hill Book Company, Inc., New York-Toronto-London, 1955.

[Dar55] Gabriele Darbo. Punti uniti in trasformazioni a codominio non compatto. *Rend. Sem. Mat. Univ. Padova*, 24:84–92, 1955.

[DS88] Nelson Dunford and Jacob T. Schwartz. *Linear operators. Part I*. Wiley Classics Library. John Wiley & Sons Inc., New York, 1988. General theory, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1958 original, A Wiley-Interscience Publication.

[DZ98] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 1998.

[EE18] D. E. Edmunds and W. D. Evans. *Spectral theory and differential operators*. Oxford Mathematical Monographs. Oxford University Press, Oxford, 2018. Second edition of [MR0929030].

[Euc78] Euclide. *Les éléments. I, II*. Éditions du Centre National de la Recherche Scientifique (CNRS), Paris, french edition, 1978. Translated from the Greek by Georges J. Kayas.

[Fel66] William Feller. *An Introduction to Probability Theory and its Applications, volume 2*. Wiley Series in Probability and Mathematical Statistics. J. Wiley, 1966.

[Fel67] William Feller. *An Introduction to Probability Theory and its Applications, volume 1*. Wiley Series in Probability and Mathematical Statistics. J. Wiley, 1967.

[Gal83] Giovanni Gallavotti. *The elements of mechanics*. Texts and Monographs in Physics. Springer-Verlag, New York, 1983. Translated from the Italian.

[GL06] Sébastien Gouëzel and Carlangelo Liverani. Banach spaces adapted to Anosov systems. *Ergodic Theory and Dynamical Systems*, 26:189–217, 2006.

[Gou10] Sébastien Gouëzel. Characterization of weak convergence of Birkhoff sums for Gibbs-Markov maps. *Israel J. Math.*, 180:1–41, 2010.

[Hen93] Hubert Hennion. Sur un théorème spectral et son application aux noyaux lipchitziens. *Proc. Amer. Math. Soc.*, 118(2):627–634, 1993.

[Her83] Michael-R. Herman. *Sur les courbes invariantes par les difféomorphismes de l'anneau. Vol. 1*, volume 103 of *Astérisque*. Société Mathématique de France, Paris, 1983. With an appendix by Albert Fathi, With an English summary.

[Her86] Michael-R. Herman. Sur les courbes invariantes par les difféomorphismes de l'anneau. Vol. 2. *Astérisque*, (144):248, 1986. With a correction to: ıt On the curves invariant under diffeomorphisms of the annulus, Vol. 1 (French) [Astérisque No. 103-104, Soc. Math. France, Paris, 1983; MR0728564 (85m:58062)].

[HH01] Hubert Hennion and Loïc Hervé. *Limit Theorem for Markov Chains and Stochastic Properties of Dynamical Systems by Quasi-Compactness*, volume 1766 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2001.

[HK95] Boris Hasselblatt and Anatole Katok. *Introduction to the modern theory of dynamical systems*. Cambridge university press, 1995.

[Hop39] Eberhard Hopf. Statistik der geodätischen Linien in Mannigfaltigkeiten negativer Krümmung. *Ber. Verh. Sächs. Akad. Wiss. Leipzig*, 91:261–304, 1939.

[Hop40] Eberhard Hopf. Statistik der Lösungen geodätischer Probleme vom unstabilen Typus. II. *Math. Ann.*, 117:590–608, 1940.

[HS74] Morris W. Hirsch and Stephen Smale. *Differential equations, dynamical systems, and linear algebra*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1974. Pure and Applied Mathematics, Vol. 60.

[IK00] Alexander V. Isaev and Steven G. Krantz. Invariant distances and metrics in complex analysis. *Notices Amer. Math. Soc.*, 47(5):546–553, 2000.

[Kat66] Tosio Kato. *Perturbation theory for linear operators*. Die Grundlehren der mathematischen Wissenschaften, Band 132. Springer-Verlag New York, Inc., New York, 1966.

[Kel82] Gerhard Keller. Stochastic stability in some chaotic dynamical systems. *Monatsh. Math.*, 94(4):313–333, 1982.

[Kif88] Yuri Kifer. *Random perturbations of dynamical systems*, volume 16 of *Progress in Probability and Statistics*. Birkhäuser Boston Inc., Boston, MA, 1988.

[KL99] Gerhard Keller and Carlangelo Liverani. Stability of the spectrum for transfer operators. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 28(1):141–152, 1999.

[KL09] Gerhard Keller and Carlangelo Liverani. Rare events, escape rates and quasistationarity: some exact formulae. *J. Stat. Phys.*, 135(3):519–534, 2009.

[Liv95] Carlangelo Liverani. Decay of correlations. *Ann. of Math. (2)*, 142(2):239–301, 1995.

[Liv03] Carlangelo Liverani. Invariant measures and their properties. A functional analytic point of view. In *Dynamical systems. Part II*, Pubbl. Cent. Ric. Mat. Ennio Giorgi, pages 185–237. Scuola Norm. Sup., Pisa, 2003.

[LL76] L. D. Landau and E. M. Lifshitz. *Course of theoretical physics. Vol. 1*. Pergamon Press, Oxford, third edition, 1976. Mechanics, Translated from the Russian by J. B. Skyes and J. S. Bell.

[LL01] Elliott H. Lieb and Michael Loss. *Analysis*, volume 14 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2001.

[LMD03] Carlangelo Liverani and Véronique Maume-Deschamps. Lasota-Yorke maps with holes: conditionally invariant probability measures and invariant probability measures on the survivor set. *Ann. Inst. H. Poincaré Probab. Statist.*, 39(3):385–412, 2003.

[Mos01] Jürgen Moser. *Stable and random motions in dynamical systems*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 2001. With special emphasis on celestial mechanics, Reprint of the 1973 original, With a foreword by Philip J. Holmes.

[Nus69] Roger D. Nussbaum. The fixed point index and asymptotic fixed point theorems for k-set-contractions. *Bull. Amer. Math. Soc.*, 75:490–495, 1969.

[Nus70] Roger D. Nussbaum. The radius of the essential spectrum. *Duke Math. J.*, 37:473–478, 1970.

[Nus88] Roger D. Nussbaum. Hilbert's projective metric and iterated nonlinear maps. *Mem. Amer. Math. Soc.*, 75(391):iv+137, 1988.

[NZ99] Igor Nikolaev and Evgeny Zhuzhoma. *Flows on 2-dimensional manifolds*, volume 1705 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1999. An overview.

[Poi87] Henri Poincaré. *Le Méthodes Nouvelles de la Mécanique Céleste, Tome I,II, III*. le grand clasiques Gauthier-Villards. Blanchard, Paris, 1987.

[PT93] Jacob Palis and Floris Takens. *Hyperbolicity and sensitive chaotic dynamics at homoclinic bifurcations*, volume 35 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1993. Fractal dimensions and infinitely many attractors.

[Roh67] V. A. Rohlin. Lectures on the entropy theory of transformations with invariant measure. *Uspehi Mat. Nauk*, 22(5 (137)):3–56, 1967.

[Roy88] H. L. Royden. *Real analysis*. Macmillan Publishing Company, New York, third edition, 1988.

[RS80] Michael Reed and Barry Simon. *Methods of modern mathematical physics. I*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, second edition, 1980. Functional analysis.

[Sie45] Carl Ludwig Siegel. Note on the differential equations on the torus. *Ann. of Math. (2)*, 46:423–428, 1945.

[Str84] D. W. Stroock. *An introduction to the theory of large deviations*. Universitext. Springer-Verlag, New York, 1984.

[UvN47] S. M. Ulam and John von Neumann. On combination of stochastic and deterministic processes. (preliminary report at the summer meeting in new haven). *Bull.Amer. Math.Soc.*, 53, 11:1120–1120, 1947.

[Var84] S. R. S. Varadhan. *Large deviations and applications*, volume 46 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1984.

[Var01] S. R. S. Varadhan. *Probability theory*, volume 7 of *Courant Lecture Notes in Mathematics*. New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 2001.

[Wol59] František Wolf. On the invariance of the essential spectrum under a change of boundary conditions of partial differential boundary operators. *Nederl. Akad. Wetensch. Proc. Ser. A 62 = Indag. Math.*, 21:142–147, 1959.

[Yos95] Kōsaku Yosida. *Functional analysis*. Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the sixth (1980) edition.

[Zei86] Eberhard Zeidler. *Nonlinear functional analysis and its applications. I*. Springer-Verlag, New York, 1986. Fixed-point theorems, Translated from the German by Peter R. Wadsack.

List of symbols

Index