An introduction to dynamical systems

Govind Menon

January 9, 2024

Contents

Chapter 1

Overview

These lecture notes provide an introduction to the theory of dynamical systems. The primary audience for these notes are graduate students in the mathematical sciences. However, I hope these notes will also be valuable to engineers, physicists and biological and social scientists.

The study of dynamical systems sheds light on several fundamental areas of science – celestial and classical mechanics, complex systems in biology and ecology, deep learning, and nonlinear waves to name just a few examples. Yet the backbone of the discipline remains rigorous mathematical analysis. This includes, on one hand, the systematic development of ideas in analysis, geometry and probability theory such as the well-posedness theory for differential equations, the study of geodesic flow and integrable systems, and ergodic theory. On the other hand, each of these general theories is often best understood through the detailed analysis of particular examples such as the restricted three-body problem, the study of specific spaces of negative curvature, and circle maps. An introductory class must provide students with a sound understanding of basic techniques, as well as a sampling of essential ideas in the discipline that are likely to be of value to them in their careers. It must also provide students with an opportunity to explore selected topics in depth, either through detailed calculations or numerical simulations.

The notes are based on an introductory two semester graduate sequence (APMA 2190-2200) that I have taught several times at the Division of Applied Mathematics at Brown University over the past twenty years. The topics presented in these notes, as well as the manner in which they have been covered, reflect the ability and needs of the students in the audience, as well as the evolution of my research interests towards an understanding of artifical intelligence. The topics in these lectures are chosen:

1. To cover as much standard theory as possible, balancing rigorous analysis with concrete calculations on important examples.

2. To illustrate the utility of the dynamicist's viewpoint in classical and modern applications.

My primary goal is to break new ground in applications. Dynamicists of my generation grew up with books such as [7] that stress the bifurcation of vector fields. These studies in turn arose from older investigations of nonlinear oscillations in biological and physical systems. While several of these applications retain their vitality, the use of dynamical ideas in algorithms, learning theory, and optimization strike me as fertile new ground for investigation. The main new feature of these notes is a shift towards the dynamics of algorithms, while retaining a traditional core.

Ideally, I would have liked to have covered both bifurcation theory and algorithms. But when one is faced with finite time (cover as much as possible in two semesters!) an emphasis on low-dimensional systems is quite limiting. This is the main reason for introducing somewhat sophisticated ideas such as ergodic theory, gradient flows and Hamiltonian systems relatively early in the course. The later chapters on dynamics and algorithms, especially the study of Hamiltonian systems, provide a preview of an interplay between geometric structure (Riemannian and symplectic) and fast numerical algorithms. These topics touch on several areas of mathematics and illustrate Vapnik's maxim that nothing is quite as applicable as a good theory. Our primary goal here is to use geometry as a bridge between traditional (classical mechanics) and modern applications (dynamics of algorithms).

Many of the introductory topics in dynamical systems are covered, but some important topics have been omitted. The most serious gap in my view is the omission of a substantive discussion of bifurcation theory. An elementary introduction to bifurcation theory was provided in the classroom, largely following [14]. However, these lecture notes do not include a proof of the center manifold theorem, the Hopf bifurcation theorem and related exercises that demonstrate the richness of bifurcation theory. [1] Some standard topics related to the analysis of two-dimensional phase portraits, such as the Poincaré-Bendixson theorem and the analysis of relaxation oscillations in the Van der Pol system have also been omitted. These examples are fun, especially if one's goal is to use phase plane analysis in applications, but there are many textbook presentations of these ideas, in particular the excellent books [10, 14].

The notes were transcribed by a student each week based on my handwritten notes. They were then edited again for consistency of style and accuracy. I am deeply grateful to the students at Brown University for participating in this effort. I hope this exposition will be useful to a new generation of students with interests in computer science, control theory, optimization and statistical physics.

[1] The center manifold theorem does follow from the invariant manifold theorem proved in Chapter 8, but the Hopf bifurcation theorem does not.

Chapter 2

Existence and uniqueness theorems for ordinary differential equations

The main reference for this chapter is Arnold's book [1]. The main result is Picard's theorem on the existence and uniqueness of solutions to the differential equation

$$\dot{x} = f(x) \tag{2.0.1}$$

with initial condition $x(0) = x_0$. Several analytical techniques will be introduced to study this question. These include contraction mappings, mollification, and compactness.

2.1 Contraction mappings on a metric space

Definition 1. A set M is a *metric space* if it is equipped with a function $d : M \times M \to [0, \infty)$ such that

1. $d(x, y) = d(y, x)$.

2. $d(x, y) = 0 \iff x = y$.

3. $d(x, y) \le d(x, z) + d(y, z)$ for all triplets $x, y, z \in M$ (triangle inequality).

Definition 2. The metric space M is *complete* if every Cauchy sequence $\{x_n\}_{n=1}^{\infty}$ has a limit in M.

Definition 3 (Contraction Mapping)**.** A map $A : M \to M$ is a contraction if there exists a constant λ, $0 < \lambda < 1$, such that

$$d(A(x), A(y)) \le \lambda d(x, y), \quad x, y \in M. \tag{2.1.1}$$

Definition 4. A point $x \in M$ is a *fixed point* of the map A if $A(x) = x$.

Theorem 5 (Contraction Mapping Theorem)**.**

1. *A contraction mapping* $A : M \to M$ *of a* complete *metric space into itself has a unique fixed point.*

2. *Given any point* $x \in M$ *the sequence of iterates* $\{A^n(x)\}_{n=0}^{\infty}$ *converges to the fixed point.*

Proof. First note that if a fixed point exists it must be unique. Indeed, if x and y satisfy $A(x) = x, A(y) = y$, then

$$d(x, y) = d(A(x), A(y)) \leq \lambda d(x, y)$$

which shows that $d(x, y) = 0$. (The equality holds by the definition of a fixed point; the inequality holds by the definition of a contraction mapping.)

Now choose any point $x \in M$ and consider the sequence of iterates $\{A^n(x)\}_{n=0}^{\infty}$. For brevity, let $x_n = A^n(x) = A \circ \cdots \circ A(x)$ denote n-fold iteration. Then

$$\begin{aligned} d(x_n, x_{n+1}) &= d(A(x_{n-1}), A(x_n)) \\ &\leq \lambda d(x_{n-1}, x_n). \end{aligned}$$

Proceeding inductively, we see that

$$d(x_n, x_{n+1}) \leq \lambda^n d(x_0, x_1), \quad n \geq 1.$$

Since $0 < \lambda < 1$, the series

$$\sum_{n=0}^{\infty} \lambda^n = \frac{1}{1 - \lambda} < \infty,$$

and it follows that $\{x_n\}_{n=0}^{\infty}$ is a Cauchy sequence. Since M is complete, the limit exists and is the desired fixed point. □

Remark 6. On the homework, you are asked to verify the Cauchy sequence property from definitions.

2.2 A Global Picard Theorem

Definition 7. A function $f : M \to M$ is an L-Lipschitz function on the metric space (M, ρ) if there exists a constant L such that

$$\rho(f(x), f(y)) \leq L\rho(x, y), \quad x, y \in M. \tag{2.2.1}$$

We will mainly use this notion for functions $f : \mathbb{R}^n \to R^n$. But strictly speaking the notion of Lipschitz functions is part of metric space theory, not calculus. For vector fields $f : \mathbb{R}^n \to \mathbb{R}^n$ equation (2.2.1) reduces to

$$|f(x) - f(y)| \leq L|x - y|, \quad x, y \in \mathbb{R}^n. \tag{2.2.2}$$

The notation here is $|v| := \sqrt{v_1^2 + \cdots + v_n^2}$ for $v \in \mathbb{R}^n$. The norm of a matrix $A : \mathbb{R}^n \to \mathbb{R}^n$ is defined by

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

In order to apply the contraction mapping theorem to establish the existence of solutions to equation (10.1.3) we will first rewrite it as an integral equation and then apply the contraction mapping theorem. In order to explain the setup of the contraction mapping theorem, we must recall some undergraduate analysis.

Let $T > 0$ be fixed and consider the space

$$M = \{x : [0, T] \to \mathbb{R}^n \mid \ x(t) \text{is continuous}\}.$$

We equip the space M with the *norm*

$$||x||_\infty := \max_{0 \leq t \leq T} |x(t)|.$$

A slightly weaker notion would be

$$||x||_\infty := \sup_{0 \leq t \leq T} |x(t)|,$$

but since $x(t)$ is continuous, $\sup |x(t)| = \max |x(t)|$ over the interval $0 \leq t \leq T$.

The functional analytic fact we need is that the space $(M, ||\cdot||_\infty)$ is a complete metric space. This follows from Weierstrass' theorem, which states that a uniformly convergent sequence of continuous functions has a limit that is a continuous function. The critical assumption here is *uniform convergence*. This prevents counterexamples such as the sequence $x_n(t) = t^n$, $0 \leq t \leq 1$.

In order to illustrate the main idea in Picard's theorem we will first prove it under the assumption that the vector field f is L-Lipschitz for all $x, y \in \mathbb{R}^n$. This is a strong assumption, because smooth functions that grow sufficiently fast at infinity (say $f(x) = x^2$ on the line) are locally, but not globally, Lipschitz.

However, the main estimate in Picard's theorem is most transparent under this assumption, so we will begin with this idea. The argument may then be modified to obtain the general local existence and uniqueness theorem. The main ideas in the proof are as follows.

1. Rewrite equation (10.1.3) as the integral equation

$$x(t) = x(0) + \int_0^t f(x(s))ds. \tag{2.2.3}$$

2. For fixed x_0 and T, we consider the space

$$M_{x_0,T} = \{x : [0,T] \to \mathbb{R}^n \,|x \text{ is continuous, } x(0) = x_0\},$$

equipped with the $\|\cdot\|_\infty$ norm. We then show that the map $A : M_{x_0,T} \to M_{x_0,T}$ defined by

$$(A(x))(t) = x_0 + \int_0^t f(x(s))ds$$

is a contraction mapping on this space for $T < 1/L$.

The critical estimate is this: given two continuous functions $x, y \in M_{x_0,T}$ we have

$$A(x)(t) - A(y)(t) = \int_0^t (f(x(s)) - f(y(s)))ds.$$

Therefore, taking absolute values

$$\begin{aligned}
|A(x)(t) - A(y)(t)| &= \left| \int_0^t (f(x(s)) - f(y(s)))ds \right| \\
&\le \int_0^t |f(x(s)) - f(y(s))|ds \\
&\le L \int_0^t |x(s) - y(s)|ds \\
&\le L \int_0^T |x(s) - y(s)|ds \\
&\le LT||x - y||_\infty.
\end{aligned}$$

Since the bound on the RHS is uniform in t, we may take the supremum over t on the LHS to obtain the fundamental estimate

$$||A(x) - A(y)||_\infty \le LT||x - y||_\infty.$$

In particular, choosing $T = \frac{1}{2L}$ we have

$$||A(x) - A(y)||_\infty \le \frac{1}{2}||x - y||_\infty. \tag{2.2.4}$$

We have thus obtained the following version of Picard's theorem.

Theorem 8 (Petit Picard)**.** *Assume that $f : \mathbb{R}^n \to \mathbb{R}^n$ is L-Lipschitz. Then the integral equation (10.2.4) has a unique solution in the space $M_{x_0,T}$ with $T = \frac{1}{2L}$.*

Proof of Picard's theorem. Apply the contraction mapping theorem, noting the estimate (2.2.4) and the fact that $M_{x_0,T}$ is a complete metric space with this norm. □

Corollary 1 (Differentiability of the solution)**.** *The solution to the integral equation (10.2.4) is differentiable at all* $t \in [0,T]$ *and solves the differential equation (10.1.3). This is written* $x \in C^1([0,T];\mathbb{R}^n)$*.*

Proof. Fix $t \in (0,T)$ so that for sufficiently small h we have

$$\frac{1}{h}(x(t+h)-x(t)) = \frac{1}{h}\int_t^{t+h} f(x(s))\,ds.$$

Now compare this difference with $f(x(t))$ (which is what we'd like the time derivative to be), obtaining the estimate

$$\begin{aligned}\left|\frac{1}{h}\left(x(t+h)-x(t)\right) - f(x(t))\right| &\leq \frac{1}{h}\int_t^{t+h} |f(x(s))-f(x(t))|\,ds \quad (2.2.5)\\ &\leq L|x(t+h)-x(t)| \leq L\|f\|_\infty h.\end{aligned}$$

We now let $h \to 0$ to see that equation (10.1.3) holds at each $t \in (0,T)$. At the endpoints $t=0$ and $t=T$, the above argument may be modified with $h>0$ and $h<0$ to see that $x(t)$ is differentiable from the left or right respectively. □

Corollary 2 (Continuation)**.** *Assume* $f:\mathbb{R}^n \to \mathbb{R}^n$ *is L-Lipschitz. Then for every* $x_0 \in \mathbb{R}^n$ *there is a unique Lipchitz function* $x:(-\infty,\infty)\to\mathbb{R}^n$ *with* $x(0)=x_0$ *such that*

$$x(t) = x_0 + \int_0^t f(x(s))ds, \quad t \in (-\infty,\infty).$$

Proof. It follows immediately from the proof of Theorem 8 that by flipping $t \to -t$, we obtain a solution on the interval $[-T,0]$. Now "restart" the time clock at $t=T$ and $t=-T$ to obtain a solution on $[-2T,2T]$. We can do this because the Lipschitz constant does not depend on x_0 or $x(T)$ or $x(-T)$. Now keep going to get a solution on $(-\infty,\infty)$. □

Remark 9. This is called a *continuation* argument. It can also be applied to the local Picard theorem to extend solutions to a maximal interval of existence.

Remark 10. The above results constitute a *well-posedness theory* for a differential equation. The implicit 'philosohpy' here is that the initial value problem $\dot{x} = f(x)$ was derived within the context of an application. The purpose of the rigorous argument is to provide a criterion (smoothness of the vector field) under which the model is consistent. The point of Picard's theorem is that a relatively simple hypothesis on smoothness is all that one needs to have a good model. This is why ODE theory works in practice.

2.3 Local vs. Global Existence

We rarely apply Picard's theorem in the version above. Usually our function f is smoother than Lipschitz and usually it is *not* globally Lipschitz. Here is an example of a smooth vector field for which the solution blows up in finite time. Consider the ODE on the line

$$\dot{x} = x^2.$$

We may solve this equation explicitly by separating variables and integrating

$$\begin{aligned}\int_{x_0}^{x(t)} \frac{dx}{x^2} = t \implies & -\frac{1}{x(t)} + \frac{1}{x_0} = t \\ \implies & x(t) = \frac{x_0}{1 - x_0 t}\end{aligned}$$

The solution blows up at $t_* = \frac{1}{x_0}$ (more precisely, $\lim_{t \to t_*} x(t) = +\infty$).

This example is typical. We should not expect the *global* Lipschitz condition to hold in general. The best we can hope for is local existence. It is easy to fix this gap. We first show that smoothness implies the Lipschitz condition used in Theorem 8. We then reduce the case of local existence to Theorem 8 using bump functions. First let us show that differentiability implies the Lipschitz condition.

Theorem 11. *Suppose $U \subset \mathbb{R}^n$ is open and $f : U \to \mathbb{R}^n$ is C^1 on U. Suppose $V \subset U$ is compact and convex. Then f is L-Lipschitz on V with*

$$L = \max_{x \in V} \|Df(x)\|.$$

(Here $\|A\|$ is the norm $\sup_{|v|=1} |Av|$).

Remark 12. The notation and terminology here is as follows. A function is said to be C^1 if it is differentiable with a continuous derivative. The derivative of f at x is a bounded linear mapping $Df(x) : \mathbb{R}^n \to \mathbb{R}^n$ whose action on a vector $v \in \mathbb{R}^n$ is defined by the limit

$$Df(x)v := \lim_{h \to 0} \frac{f(x + hv) - f(x)}{h}.$$

In the more general geometric setting of differentiable manifolds, the derivative is a linear operator between the tangent spaces $T_x V$ and $T_{f(x)}\mathbb{R}^n$ (which have been identified with $\mathbb{R}^n$ above. If you find these abstractions confusing, think for now of the derivative as an $n \times n$ matrix with entries $\frac{\partial f_i}{\partial x_j}$ (i indexes rows, j indexes columns).

Proof. Consider two points $x, y \in V$. Let $x(t) = (1 - t)x + ty$, $0 \leq t \leq 1$ be the line segment joining these points. By the fundamental theorem of calculus

$$\begin{aligned} f(y) - f(x) &= \int_0^1 \frac{df(x(t))}{dt} dt \\ &= \int_0^1 Df(x(t)) \frac{dx}{dt} dt \qquad \text{(chain rule)} \\ &= \left(\int_0^1 Df(x(t)) dt \right) (y - x) \end{aligned}$$

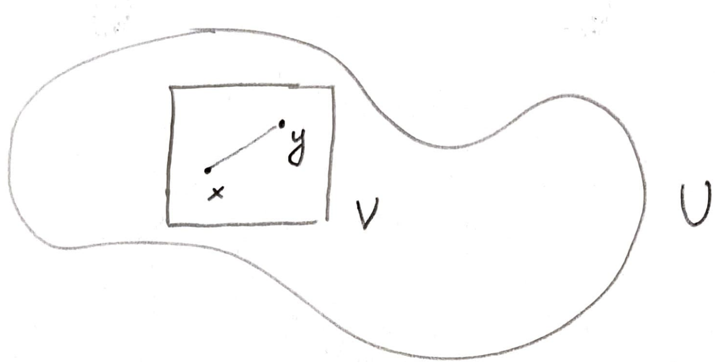

Figure 2.3.1: The convex set $V \subset U$.

Now take absolute values to obtain

$$\begin{aligned}|f(x) - f(y)| &\leq \left(\int_0^1 \|Df(x(t))\| dt \right) |x - y| \\ &\leq L|x - y|\end{aligned}$$

since $x(t) \in V$ (because V is convex) and $L = \sup_{x \in V} \|Df(x)\|$. □ □

An immediate corollary of this theorem is that the differential equation (10.1.3) has a local solution when f is C^1. We will prove this theorem by reducing it to the Petit Picard theorem using the technique of *bump functions*. This is an important technique that merits a digression.

2.4 Mollification and the heat kernel

2.4.1 Mollification with bump functions

A fundamental technique in analysis is *mollification* (or *smoothing*). We will use this technique at several places, including the proof of Peano's theorem, the extension of the global Picard theorem to local existence, and the proof of invariant manifold theorems.

Definition 13. A mollifier is a function $\psi : \mathbb{R}^n \to \mathbb{R}$ with the following properties.

1. $\psi(x) \geq 0$ for all $x \in \mathbb{R}^n$.
2. ψ is C^∞.
3. $\int_{\mathbb{R}^n} \psi(x)\, dx = 1$.

In the first homework, you are asked to construct such functions with the additional property that ψ is compactly supported (that is ψ vanishes outside a large enough box in $\mathbb{R}^n$). These are called *bump functions*. We do not make this assumption above, since there are natural mollifiers, such as the heat kernel discussed below, which are not compactly supported.

The main technique for smoothing is convolution with a rescaled mollifier. The convolution of two integrable functions $f, g : \mathbb{R}^n \to \mathbb{R}$ is defined by

$$(f * g)(x) = \int_{\mathbb{R}^n} f(x-y)g(y)dy = \int_{\mathbb{R}^n} f(y)g(x-y)dy. \tag{2.4.1}$$

Assume given a mollifier ψ. Then for any $\varepsilon > 0$, the rescaled mollifier

$$\psi_\varepsilon(x) := \frac{1}{\varepsilon^n}\psi\left(\frac{x}{\varepsilon}\right), \tag{2.4.2}$$

remains a mollifier. The factor ε^{-n} is included to ensure that $\int_{\mathbb{R}^n} \psi_\varepsilon = 1$.

Given an integrable function $f : \mathbb{R}^n \to \mathbb{R}$ and a mollifier ψ, we define the mollification

$$f_\varepsilon(x) = (f * \psi_\varepsilon)(x). \tag{2.4.3}$$

Intuitively, this rescaling allows us to smooth a function by replacing it with averages over regions of size ε.

Lemma 1. *Assume f is integrable. The function f_ε is C^∞ for every $\varepsilon > 0$. For every multi-index α, we have*

$$\|\partial^\alpha f_\varepsilon\|_\infty \leq \frac{1}{\varepsilon^{n|\alpha|}}\|\partial^\alpha \psi\|_{L^1}\|f\|_\infty. \tag{2.4.4}$$

Remark 14. A multi-index $\alpha = (\alpha_1, \ldots, \alpha_m)$ is a collection of positive integers. The notation used here is

$$\partial_x^\alpha f_\varepsilon(x) := \partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_m} f_\varepsilon(x), \quad |\alpha| = \sum_{j=1}^m \alpha_j, \quad \|g\|_{L^1} := \int_{\mathbb{R}^n} |g(x)|\, dx.$$

Proof. Since

$$f_\varepsilon(x) = \int_{\mathbb{R}^n} \psi_\varepsilon(x-y) f(y)\, dy,$$

we may formally differentiate under the integral sign to obtain

$$\partial_{x_j} f_\varepsilon(x) = \int_{\mathbb{R}^n} \partial_{x_j}\psi_\varepsilon(x-y) f(y)\, dy.$$

Now take absolute values and use (2.4.2) to obtain the estimate

$$\|\partial_{x_j} f_\varepsilon\|_\infty \leq \frac{1}{\varepsilon^n}\|\partial_{x_j}\psi\|_{L^1}\|f\|_\infty.$$

Proceeding inductively, we find as above that formally

$$\partial_x^\alpha f_\varepsilon(x) = \int_{\mathbb{R}^n} \partial_x^\alpha \psi_\varepsilon(x-y) f(y)\, dy,$$

and taking absolute values yields equation (2.4.4).

All that remains is to justify the interchange of limits implicit in differentiating under the integral sign. This may be done with finite differences as in the proof of Corollary 1.

□

The pointwise convergence of $f_\varepsilon(x)$ to $f(x)$ is a little more delicate and a stronger hypothesis is necessary.

Lemma 2. *Assume $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and integrable. Then $\lim_{\varepsilon\to 0} f_\varepsilon(x) = f(x)$ at every $x \in \mathbb{R}^n$.*

Proof. Since $\int_{\mathbb{R}^n} \psi_\varepsilon = 1$, we have the identity

$$f_\varepsilon(x) - f(x) = \int_{\mathbb{R}^n} \psi_\varepsilon(y)\left(f(x-y) - f(x)\right)\, dy.$$

Now take absolute values and use the fact that $\psi_\varepsilon \geq 0$ to obtain the estimate

$$|f_\varepsilon(x) - f(x)| \leq \int_{\mathbb{R}^n} \psi_\varepsilon(y)\left|f(x-y) - f(x)\right|\, dy.$$

In order to see the estimate that follows, assume that ψ_ε is a bump function with compact support. Then the domain of the above integral is restricted to a ball with radius $O(\varepsilon)$ centered at x. Thus,

$$|f_\varepsilon(x) - f(x)| \leq \max_{|y-x|<C\varepsilon} |f(y) - f(x)|.$$

Since f is continuous, this quantity vanishes in the limit $\varepsilon \to 0$.

If one doesn't assume the mollifier has compact support, a little more care is needed. This case arises when we consider the heat function. It is left as an exercise for the reader. □

An important theme in mollifcation is that while the derivatives of f_ε diverge as $\varepsilon \to 0$, it is still the case that f_ε satisfies all the estimates we impose on f. Examples of such uniform estimates are contained in the lemmata below.

Lemma 3. *Assume that $f : \mathbb{R}^n \to \mathbb{R}$ satisfies $\|f\|_\infty < \infty$. Then the mollifications satisfy the uniform estimates*

$$\|f_\varepsilon\|_\infty \leq \|f\|_\infty, \quad \varepsilon > 0. \tag{2.4.5}$$

Proof. We use the definition (2.4.3) and the positivity of the mollifier to obtain

$$|f_\varepsilon(x)| = \left|\int_{\mathbb{R}^n} \psi_\varepsilon(x-y) f(y) dy\right| \leq \|f\|_\infty \int_{\mathbb{R}^n} \psi_\varepsilon(x-y) dy = \|f\|_\infty.$$

The right hand side is independent of x. Taking the supremum over x completes the proof. □

A variant of the above argument is used to establish equicontinuity of the mollifications.

Definition 15. Assume $\omega : [0,\infty) \to [0,\infty)$ is a monotone increasing function such that $\lim_{r\to 0}\omega(r) = 0$. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to have modulus of continuity ω if

$$|f(x) - f(y)| \leq \omega(|x-y|), \quad x, y \in \mathbb{R}^n. \tag{2.4.6}$$

Lemma 4. *Suppose $f : \mathbb{R}^n \to \mathbb{R}^n$ has modulus of continuity ω. Then the mollifications satisfy the uniform estimate*

$$|f_\varepsilon(x) - f_\varepsilon(y)| \leq \omega(|x-y|), \quad \forall \varepsilon > 0. \tag{2.4.7}$$

Proof. Fix two points x and y in $\mathbb{R}^n$ and let z denote the dummy variable of integration. We then have

$$\begin{aligned}|f_\varepsilon(x) - f_\varepsilon(y)| &= \left|\int_{\mathbb{R}^n} \psi_\varepsilon(y)(f(x-z) - f(y-z))dz\right| \\ &\leq \omega(|x-y|)\int_{\mathbb{R}^n} \psi_\varepsilon(z)\,dz = \omega(|x-y|).\end{aligned}$$

□

2.4.2 The heat kernel

Another fundamental example of a mollifier is the heat kernel. The heat kernel does not have compact support, but it provides concrete formulas and physical intuition that is valuable.

Definition 16. The heat kernel on $\mathbb{R}^n$ is the fundamental solution to the partial differential equation

$$\partial_t u = \frac{1}{2}\triangle u, \quad x \in \mathbb{R}^n, \quad t > 0, \tag{2.4.8}$$

where $\triangle$ denotes the Laplacian $\sum_{j=1}^n \partial_{x_j}^2$. The fundamental solution $p_t(x;y)$ denotes the solution to (2.4.8) with a singular (Dirac delta) initial condition at $x = y$. It is given by the formula

$$p_t(x;y) = g_t(x-y), \quad g_t(x) = \frac{1}{(2\pi t)^{n/2}}e^{-|x|^2/2t}. \tag{2.4.9}$$

The graph of g_t is the well-known bell curve with width $\sqrt{t}$. As $t \downarrow 0$, $g_t(x)$ concentrates at a Dirac delta at 0.

It is not necessary to mollify with the heat kernel, but it is useful to do so, since it provides a family of smooth approximations that is easily visualized and is easy to simulate. It is easily checked that the lemmas above continue to hold with the heat kernel.

2.5 Picard's theorem: the local version

We may now finally state and prove the complete version of Picard's theorem.

Theorem 17 (Picard's existence theorem)**.** *Let $U \subset \mathbb{R}^n$ be an open set. Assume $f : U \to \mathbb{R}^n$ is C^1. Then for every $x_0 \in U$ there exists $T(x_0) > 0$ and a C^1 map $x : [-T, T] \to U$ such that*

$$\dot{x} = f(x(t)), \quad t \in [-T, T]$$

and $x(0) = x_0$.

We will prove this theorem by extending the vector field f to all of $\mathbb{R}^n$ in a manner that Theorem 8 and its corollaries apply. This theorem provides a foundation for phase portraits.

Remark 18. The theorem is *not* sharp. Examples and counterexamples are considered in the homework.

2.5.1 Smooth extensions of a function

Bump functions allow us to reduce the analysis on open sets contained within $\mathbb{R}^n$ to analysis on the entire space $\mathbb{R}^n$. This allows us to obtain the local Picard theorem from the glocal Picard theorem. A similar idea will be used in proofs of the invariant manifold theorems. [1]

In the following examples, we consider a function defined on an open set $U \subset \mathbb{R}^n$ and a compact set $V \subset U$. Our goal will be to extend a function defined on V to a function defined on all of $\mathbb{R}^n$.

Definition 19. Given a measurable set $G \in \mathbb{R}^n$ its *indicator function* is the function $\mathbf{1}_G : \mathbb{R}^n \to \mathbb{R}$ defined by

$$\mathbf{1}_G(x) = \begin{cases} 1, & x \in G \\ 0, & x \notin G. \end{cases}$$

Given a smooth function $f : U \to \mathbb{R}^n$ the obvious extension of its restriction to V, $f|_V$, is simply

$$f_{ext}(x) \stackrel{?}{=} \begin{cases} f(x), & x \in V \\ 0, & x \notin V \end{cases}$$

The problem is that this extension is not smooth (i.e. as smooth as f).

Another class of extensions is obtained by using a smooth function $\varphi(x)$ such that

$$\varphi(x) \equiv 1, \quad x \in V$$
$$\varphi(x) \equiv 0, \quad x \notin U.$$

Such bump functions are constructed in the first homework.

[1] This seems counterintuitive. The point is that working on the 'standard space' $\mathbb{R}^n$ prevents technical annoyances caused by restrictions on the domain of the function, so global really is simpler than local!

2.5.2 Proof of Picard's theorem

Assuming the existence of bump functions, we may consider the vector field

$$f_{ext}(x) = \begin{cases} f(x)\varphi(x), & x \in U \\ 0, & x \notin U \end{cases} \tag{2.5.1}$$

The function f_{ext} is as smooth as f. This may be seen by applying the product rule to $f(x)\varphi(x))$ to obtain

$$\begin{aligned} Df_{ext} &= Df\varphi(x) + f(x) \otimes D\varphi(x) \\ &= Df(x), \quad \text{when} \quad x \in V. \end{aligned}$$

The final term on the first line is the matrix with entries $f_i(x)\frac{\partial\varphi}{\partial x_j}(x))$. We also see that

$$f_{ext} = f(x), \quad x \in V.$$

Finally, since f_{ext} vanishes outside a compact set it is *globally* Lipschitz.

Proof of Theorem 17. Let $V \subset U$ be a closed convex set containing the initial point x_0. Choose a bump function φ that is identically one on V and vanishes outside a second compact set contained within U. Let f_{ext} be the vector field defined in equation (2.5.1) and compare the integral equations

$$x(t) = x_0 + \int_0^t f(x(s))ds \tag{10.2.4}$$

and

$$x_{ext}(t) = x_0 + \int_0^t f_{ext}(x_{ext}(s))ds. \tag{2.5.2}$$

Note that

1. f is defined on U and f_{ext} is defined on $\mathbb{R}^n$ and they agree on the set $V \subset U$.

2. $f(x) = f_{ext}(x)$ provided $x \in V$.

3. Equation (2.5.2) has a global C^1 solution by Corollary 1 and Corollary 2.

But then it is to immediate that $x_{ext}(t)$ is a solution to equation (10.2.4) as long as $x_{ext}(t) \in V$. Let $T_\pm$ be the first exit times for the positive and negative time intervals respectively

$$T_+ = \inf_{t>0}\{x_{ext}(t) \text{ is not in } V\}, \quad T_- = -\inf_{t<0}\{x_{ext}(t) \text{ is not in } V\}.$$

Finally, choose $T(x_0) = \min(T_-, T_+)$. $\square$

2.6 Peano's theorem

In this section we investigate what happens when f is *not* Lipschitz. For the sake of simplicity, we will assume that f is a bounded and uniformly continuous function from $\mathbb{R}^n \to \mathbb{R}^n$. In fact, continuity of f on an open set U is all that is required, but it is considerably easier to illustrate the main idea when f is globally defined, globally bounded and uniformly continuous.

Recall that f is said to be uniformly continuous if for every $\varepsilon > 0$ there exists $\delta(\varepsilon) > 0$ such that $|f(x) - f(y)| < \varepsilon$ whenever $|x - y| < \delta(\varepsilon)$. The point here is that δ does not depend on the points x and y. For example, an L-Lipschitz function is uniformly continuous with $\delta = \frac{\varepsilon}{L}$.

Theorem 20 (Peano). *Assume $f : \mathbb{R}^n \to \mathbb{R}^n$ is uniformly continuous and bounded. Then, for every $x_0 \in \mathbb{R}^n$ there exists a C^1 function $x : (-\infty, \infty) \to \mathbb{R}^n$ that satisfies the differential equation*

$$\dot{x} = f(x), \quad t \in (-\infty, \infty),$$

and the initial condition $x(0) = x_0$.

Remark 21. We will establish existence of solutions to the integral equation

$$x(t) = x(0) + \int_0^t f(x(s))\, ds.$$

As in Corollary 1, once we have established the existence of solutions to the integral equation, a little additional work shows that $\dot{x} = f(x)$.

Remark 22. Though Peano's theorem is not as directly useful to us as Picard's theorem, the proof of this theorem illustrates a general technique for the well-posedness of differential equations (including functional, stochastic and partial differential equations). In each case, we separate the problems of existence, uniqueness and regularity of solutions. First, by replacing the differential equation with its integral formulation, we obtain a more forgiving notion of solution (these are called weak solutions in SDE and PDE theory). Second, the use of compactness theorems in function spaces, along with uniform estimates, is a general method for establishing existence. Picard's theorem provides existence and uniqueness together. This is atypical; for many nonlinear differential equations, especially in PDE theory, it is relatively straightforward to establish existence through the above technique, but far harder to establish uniqueness. Finally, once one has established the existence of weak solutions, it is necessary to establish their smoothness (or lack thereof) to evaluate the consistency of the model we began with. This involves a study of the regularity of solutions.

Proof. We will first prove existence for $t \in [0, 1]$, and then use a continuation argument to extend the result to the whole real line as we did for Picard's theorem. None of the steps in the proof will depend explicitly on x_0. The main abstract idea in this proof is the use of the Arzela-Ascoli compactness theorem along with an approximation scheme. We separate the proofs of these steps for clarity.

Step 1. Approximation. Fix a mollifier ψ and let $f_\varepsilon = f * \psi_\varepsilon$ denote the mollifications defined in Section 2.4. Lemmas 1– 4 establish the following properties of the family $\{f_\varepsilon\}_{\varepsilon>0}$.

1. $f_\varepsilon(x)$ is a C^∞ function of x.
2. Uniform boundedness: $||f_\varepsilon||_\infty \le \|f\|_\infty$ for every $\varepsilon > 0$.
3. Equicontinuity: $|f_\varepsilon(x) - f_\varepsilon(y)| \le \omega(|x-y|)$ for all $x, y \in \mathbb{R}^n$.

Since f_ε is C^∞, by Picard's theorem the integral equation

$$x_\varepsilon(t) = x_0 + \int_0^t f_\varepsilon(x_\varepsilon(s))\, ds \tag{2.6.1}$$

has a unique solution for $t \in [0,1]$. We now need to take the limit $\varepsilon \to 0$. This requires a new idea beyond Picard's theorem. □

Step 2. Compactness. . The Arzela-Ascoli theorem provides the following criterion for compactness of a sequence of functions in $C([0,1])$ (i.e. the space of continuous functions on [0,1] equipped with the uniform norm $||g||_\infty = \max_{t\in[0,1]} |g(t)|$).

Given a sequence $\{g_n\}_{n=1}^\infty \subset C([0,1])$, there exists a subsequence that converges in $C([0,1])$ if:

(i) $\{g_n\}_{n=1}^\infty$ is uniformly bounded (i.e. $\sup_{n\in\mathbb{N}} ||g_n|| < \infty$)

(ii) $\{g_n\}_{n=1}^\infty$ is *equicontinuous*. That is, for every $\varepsilon > 0$ there exist $\delta = \delta(\varepsilon)$ such that $\sup_{n\ge1} |g_n(x) - g_n(y)| < \varepsilon$ whenever $|x - y| < \delta$. In other words, each g_n is uniformly continuous, and the modulus of continuity is independent of n.

We now show that the family $\{x_\varepsilon(t)\}_{\varepsilon>0}$ has these properties.

(i) Since we assumed that $\|f\|_\infty < \infty$, Lemma 3 tells us that $||f_\varepsilon||_\infty \le \|f\|_\infty$. But then for every $t \in [0,1]$

$$|x_\varepsilon(t)| \le |x_0| + \int_0^t |f_\varepsilon(x_\varepsilon) ds \le |x_0| + \|f\|_\infty t \le |x_0| + \|f\|_\infty.$$

It follows that $\sup_{\varepsilon>0} ||x_\varepsilon||_\infty \le |x_0| + \|f\|_\infty$.

(ii) In a similar manner, for any $s, t \in [0,1]$ we have

$$|x_\varepsilon(t) - x_\varepsilon(s)| = \left| \int_s^t f_\varepsilon(x_\varepsilon) d\tau \right| \le \|f\|_\infty |t - s|$$

It follows that for all $\varepsilon, \eta > 0$ $|x_\varepsilon(t) - x_\varepsilon(s)| < \eta$ whenever $|t - s| < \frac{\eta}{M}$; thus $\{x_\varepsilon\}_{\varepsilon>0}$ is equicontinuous.

By the Arzela-Ascoli theorem, there exists a convergent subsequence $\{x_{\varepsilon_j}\}_{j=1}^{\infty}$. We denote the limit of this subsequence, by $x(t)$. □

Step 3. Passage to the limit. All that is left to show is that x is a solution to

$$x(t) = x_0 + \int_0^t f(x(s))\, ds. \tag{2.6.2}$$

It is in this step that we need Lemma 4. Now

$$\begin{aligned}
x(t) &= \lim_{j\to\infty} x_{\varepsilon_j}(t) = x_0 + \lim_{j\to\infty} \int_0^t f_{\varepsilon_j}(x_{\varepsilon_j}(s))ds \\
&= x_0 + \lim_{j\to\infty} \int_0^t \left[f_{\varepsilon_j}(x_{\varepsilon_j}(s)) - f_{\varepsilon_j}(x(s)) + f_{\varepsilon_j}(x(s)) \right] ds \\
&= x_0 + \lim_{j\to\infty} \int_0^t \left[f_{\varepsilon_j}(x_{\varepsilon_j}(s)) - f_{\varepsilon_j}(x(s)) \right] ds + \lim_{j\to\infty} \int_0^t f_{\varepsilon_j}(x(s))ds
\end{aligned}$$

provided we can establish the existence of the two limits in the last line. We consider these terms in turn.

Since $\|f_\varepsilon\|_\infty \le \|f\|_\infty < \infty$ for all $\varepsilon > 0$, by the Dominated Convergence Theorem (DCT) and Lemma 2,

$$\lim_{j\to\infty} \int_0^t f_{\varepsilon_j}(x(s))ds = \int_0^t \lim_{j\to\infty} f_{\varepsilon_j}(x(s))ds = \int_0^t f(x(s))ds.$$

For the other limit, by Lemma 4

$$\left| f_{\varepsilon_j}(x_{\varepsilon_j}(s)) - f_{\varepsilon_j}(x(s)) \right| \le \omega(|x_{\varepsilon_j}(s) - x(s)|).$$

But then another application of the DCT yields:

$$\begin{aligned}
\lim_{j\to\infty} \int_0^t \left| f_{\varepsilon_j}(x_{\varepsilon_j}(s)) - f_{\varepsilon_j}(x(s)) \right| ds &\le \lim_{j\to\infty} \int_0^t \omega(|x_{\varepsilon_j}(s) - x(s)|)ds \\
&= \int_0^t \lim_{j\to\infty} \omega(|x_{\varepsilon_j}(s) - x(s)|)ds = 0.
\end{aligned}$$

□

These three steps show that the integral equation (2.6.2) holds for $t \in [0,1]$. We may repeat this argument on each time interval $[k, k+1]$, $k \in \mathbb{Z}$. Thus, equation (2.6.2) holds for $t \in (-\infty, \infty)$.

In order to prove that $x(t)$ solves the differential equation $\dot{x} = f(x)$ we modify equation (2.2.5) as follows. We use the modulus of continuity ω to obtain

$$\frac{1}{h} \int_t^{t+h} |f(x(s)) - f(x(t))|\, ds \le \omega(\max_{s\in[t,t+h]} |x(s) - x(t)|) \le \omega(\|f\|_\infty h).$$

This vanishes in the limit $h \to 0$ and we see that $\dot{x} = f(x(t))$ as desired. □

Remark 23. If you haven't seen the DCT used, that's fine. You can justify the interchange of limits using the standard criterion for the Riemann integral. (Roughly, if $f_n \to f$ uniformly in $C([0,1])$ then $\lim_{n\to\infty} \int f_n = \int \lim_{n\to\infty} f_n$). You could also choose to ignore these parts of the proof and focus on other aspects of it, returning to these arguments when your understanding of analysis is stronger.

2.7 Exercises

1. *Gronwall's inequality* : If $T > 0$, $c \geq 0$, and $f, g : [0,T] \to [0,\infty)$ are continuous, and f satisfies the integral inequality

$$f(t) \leq c + \int_0^t g(s) f(s)\, ds, \quad t \in [0,T],$$

then show that

$$f(t) \leq c \exp\left(\int_0^t g(s)\, ds\right), \quad t \in [0,T].$$

2. Complete the proof of the contraction mapping principle, by showing that the sequence $\{x_n\}_{n=0}^{\infty}$ defined by $x_n = A^n(x_0)$, $n \geq 1$ is a Cauchy sequence.

3(a). Consider the differential equation $\dot{x} = f(x)$ with $x \in \mathbb{R}$ and

$$f(x) = \begin{cases} 0, & x = 0, \\ x \log |x|, & x \neq 0. \end{cases}$$

Does Picard's theorem apply? Is there a unique solution when $x_0 = 0$?

3(b). Find a function $f(x)$, $x \in \mathbb{R}$ that is not Lipschitz at 0, but for which the initial value problem $\dot{x} = f(x)$ with $x(0) = 0$ has a unique solution.

4. *Continuous dependence on parameters*. Let $x(t; x_0, \mu)$ denote the solution to the initial value problem $\dot{x} = f(x, \mu)$, $x(0) = x_0$ with a C^k vector-field $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ and $x_0 \in \mathbb{R}^n$. Show that $x(t; x_0, \mu)$ is a C^k function of μ.

5. The standard construction of bump functions goes as follows. Consider the function

$$\varphi(x) = \begin{cases} e^{-1/x}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

(a) Show that this function is infinitely differentiable at zero. That is, show that all derivatives of $e^{-1/x}$ on the region $x > 0$ vanish as $x \to 0$.

(b) Given an interval $[a,b]$ show that there is a C^∞ function φ_δ that is identically equal to 1 on $[a,b]$ but vanishes when $x \leq a - \delta$ and $x \geq b + \delta$ for any $\delta > 0$.

(c) Extend this idea to $\mathbb{R}^n$, constructing a bump function that is identically equal to one in a box, but vanishes outside a transition layer of width δ

6. Consider the initial value problem $\dot{x} = f(x)$, $x(0) = x_0$ with a bounded and continuous vector field $f : \mathbb{R}^n \to \mathbb{R}^n$ and $x_0 \in \mathbb{R}^n$, and $t \in [0, T]$ for some fixed $T > 0$.

The forward Euler scheme is an approximation method for this differential equation of the following form: the approximation $x^{(N)}(t)$ is a Lipschitz function such that (i) $x^{(N)}(0) = x_0$; (ii) $x^{(N)}(t)$ is piecewise linear with slope $f(x^{(N)}(nh))$ on the intervals $[nh, (n+1)h)$, $n = 0, 1, \dots, N-1$, $h = T/N$.

Prove that as $N \to \infty$ a subsequence converges in $C([0, T]; \mathbb{R}^n)$ to a Lipschitz function $x(t)$ that solves the initial value problem, thus establishing another proof of Peano's theorem.

2.8 Solutions to exercises

1. *Gronwall's inequality* : If $T > 0$, $c \geq 0$, and $f, g : [0, T] \to [0, \infty)$ are continuous, and f satisfies the integral inequality

$$f(t) \leq c + \int_0^t g(s) f(s)\, ds, \quad t \in [0, T],$$

then show that

$$f(t) \leq c \exp\left(\int_0^t g(s)\, ds\right), \quad t \in [0, T].$$

Proof. Fix $\varepsilon > 0$ and let h_ε denote the solution to the differential equation

$$\dot{h_\varepsilon} = gh, \quad h(0) = c + \varepsilon.$$

The solution to this equation is

$$h_\varepsilon(t) = (c + \varepsilon) \exp\left(\int_0^t g(s)\, ds\right).$$

We will show that the set $S_\varepsilon = \{t \in [0, T] \,|\, f(t) \geq h_\varepsilon(t)\}$ is empty. First, it is clear that S_ε is closed. Therefore, its complement is open. Moreover, the complement includes a maximal, open interval about the origin of the form $[0, \tau]$ because $f(0) = c < c + \varepsilon = h_\varepsilon(0)$. ("Open" here means "relatively open").

We claim that $\tau = T$. Indeed, since since $f(t) < h_\varepsilon(t)$ for $t \in [0, \tau]$, if $\tau < T$ we have

$$f(\tau) \leq c + \int_0^\tau g(s) f(s)\, ds < c + \varepsilon + \int_0^t g(s) h_\varepsilon(s)\, ds = h_\varepsilon(\tau).$$

This contradicts the definition of τ. Since $\varepsilon > 0$ is arbitrary, we find

$$f(t) \leq c \exp\left(\int_0^t g(s)\, ds\right), \quad t \in [0, T].$$

$\square$

2. Complete the proof of the contraction mapping principle, by showing that the sequence $\{x_n\}_{n=0}^{\infty}$ defined by $x_n = A^n(x_0)$, $n \geq 1$ is a Cauchy sequence.

Proof. Recall that A defines a contraction mapping on the metric space (M, ρ) with constant λ, $0 < \lambda < 1$. This assumption allowed us to obtain the estimate

$$\rho(x_n, x_{n+1}) = \rho(A^n(x_0), A^{n+1}(x_0)) \leq \lambda^n \rho(x_0, x_1).$$

Let n and p be two positive integers; without loss of generality we may suppose that $n < p$. We use the above estimate inductively to obtain

$$\rho(x_n, x_p) \leq \sum_{m=n}^{p-1} \rho(x_m, x_{m+1}) \leq \left(\sum_{m=n}^{p-1} \lambda^m\right) \rho(x_0, x_1) \leq \frac{\lambda^n}{1-\lambda}\rho(x_0, x_1).$$

Since $0 < \lambda < 1$, for any $\varepsilon > 0$ we may choose n so large that the right hand side is less than ε. □

3(a). Consider the differential equation $\dot{x} = f(x)$ with $x \in \mathbb{R}$ and

$$f(x) = \begin{cases} 0, & x = 0, \\ x \log |x|, & x \neq 0. \end{cases}$$

Does Picard's theorem apply? Is there a unique solution when $x_0 = 0$?

3(b). Find a function $f(x)$, $x \in \mathbb{R}$ that is not Lipschitz at 0, but for which the initial value problem $\dot{x} = f(x)$ with $x(0) = 0$ has a unique solution.

Proof. (a) Picard's theorem does not apply because f is not Lipschitz at 0. On the other hand, the solution is unique. This can be seen by explicit integration. We first assume $x_0 > 0$, separate variables, substitute $y = \log x$ and integrate both sides to obtain

$$t = \int_{x_0}^{x(t)} \frac{dx'}{x' \log x'} = \int_{y_0}^{y(t)} \frac{dy'}{y'} = \log\left(\frac{y(t)}{y(0)}\right).$$

Now solve for $y(t)$ and then $x(t) = e^{y(t)}$ to obtain

$$x(t) = e^{(\log x_0)e^t} = (x_0)^{e^t}, \quad x_0 > 0.$$

There is a similar solution formula for $x_0 < 0$.

$$x(t) = -e^{(\log |x_0|)e^t}.$$

Both formulas are defined for $t \in (-\infty, \infty)$ and we see that $x(t) \to 0$ as $t \to -\infty$.

Uniqueness of solutions originating at $x_0 = 0$ is obtained from this formula as folllows: if there is a solution $x(t)$ with $x_0 = 0$ such that $x(t)$ is not zero for all time, then there exists a time $t_1 > 0$ such that $x(t) = x_1 \neq 0$. Either $x_1 > 0$ or $x_1 < 0$. If $x_1 > 0$ the solution formula above shows that

$$x(t) = (x_1)^{e^{t-t_1}}.$$

Similarly, if $x_1 < 0$. In particular, $x(0) \neq 0$ contradicting our assumption.
(b) The explicit solution formula above is nice, but the underlying principle that guarantees uniqueness is this: the solution to $\dot{x} = f(x)$ with $x(0) = 0$ is unique if $\int_0^\varepsilon dx/f(x)$ is divergent for every $\varepsilon > 0$.

Thus to find an example or counterexample, one only has to choose a function such that $f(0) = 0$ and $0 < \int_0^\varepsilon dx/f(x)$ is divergent for every $\varepsilon > 0$. The function $f(x) = x \log x$ is an example, but there are infinitely many choices. □

4. *Continuous dependence on parameters*. Let $x(t; x_0, \mu)$ denote the solution to the initial value problem $\dot{x} = f(x, \mu)$, $x(0) = x_0$ with a C^k vector-field $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ and $x_0 \in \mathbb{R}^n$. Show that $x(t; x_0, \mu)$ is a C^k function of μ.

Proof. The proof is very similar to the proof that the flow defines a diffeomorphism. First, we formally obtain a linear differential equation for the vector

$$y(t) \stackrel{?}{=} \frac{\partial x(t;\mu)}{\partial \mu}.$$

Then we show that the solution $y(t)$ is indeed the derivative by working from the definitions (this is why there is a question mark in the equation above).

First some bookkeeping: Picard's theorem guarantees local existence of C^1 solutions to the differential equation $\dot{x} = f(x, \nu)$, $x(0) = x_0$. Since f is at least C^1 in both x and ν, we may choose a neighborhood V of x_0, a neighborhood $[\mu - \delta, \mu + \delta]$ for the parameter ν, and a time T such that there is a unique solution to this differential equation for all $x_0 \in V$, for all $t \in [0, T]$ and all ν in the range $[\mu - \delta, \mu + \delta]$ (we changed notation a bit here, so that we can do all the computations at a fixed value μ of the parameter). This allows us to say that the 'tube' of trajectories

$$K = \{x(\nu, t) : t \in [0, t], \nu \in [\mu - \delta, \mu + \delta]\},$$

is a compact set. When $f \in C^k$ it follows that the first k derivatives of f in x and μ satisfy the bound

$$\sup_{z \in K, \nu \in [\mu-\delta, \mu+\delta]} |D_x^{(l)} f(z;\nu)|, |\partial_\nu^{(l)} f(z;\nu)| \leq C, \quad 0 \leq l \leq k.$$

These bounds will be used to justify interchanges of limits below.

Now let us turn to the main ideas. We differentiate the equation $\dot{x} = f(x, \mu)$ with respect to μ, use the chain rule, and denote $A(t) = Df(x(t;\mu))$ and $b(t) = \partial f(x(t;\mu), \mu)/\partial\mu$ to obtain the differential equation

$$\dot{y} = A(t)y + b(t), \quad y(0) = 0,$$

or equivalently the integral equation[2]

$$y(t) = \int_0^t A(s)y(s)\,ds + \int_0^t b(s)\,ds.$$

[2]Let $S(t;s)$ denote the fundamental matrix for this differential equation, i.e. the unique

Let us now show that $y(t)$ is indeed the derivative of $x(t;\mu)$ with respect to the parameter μ. For brevity, let $x_\mu(t)$ denote $x(t;\mu)$ and denote $f(x,\mu)$ by $f_\mu(x)$. We compare the solutions $x_{\mu+h}(t)$ and $x_\mu(t)$ to obtain the identity

$$(x_{\mu+h}(t) - x_\mu(t)) = \int_0^t \left(f(x_{\mu+h}(s);\mu+h)\right) - f(x_\mu(s);\mu))\, ds. \tag{2.8.1}$$

By Taylor's remainder theorem and our a priori bound on the range of $x_\nu(t)$ for $t \in [0,T]$ and $\nu \in [\mu-\delta, \mu+\delta]$, there is a constant C such that

$$|f_{\mu+h}(x_{\mu+h}(s)) - f_\mu(x_\mu(s)) - hA(s)\left(x_{\mu+h}(s) - x_\mu(s)\right) - hb(s)| \leq Ch^2,$$

where C is uniform over the range $s \in [-T,T]$, $|h| \leq \delta$. It follows that we have the a priori estimate

$$\left|\frac{x_{\mu+h}(t) - x_\mu(t)}{h}\right|$$
$$\int_0^t \|A(s)\| \left|\frac{x_{\mu+h}(t) - x_\mu(t)}{h}\right| ds + \int_0^t |b(s)|\, ds + Cht, \quad t \in [0,T].$$

As usual, let $\|A\|_\infty = \sup_{s\in[0,T]} \|A(s)\|$ and $\|b\|_\infty$ be defined similarly. Then we may apply Gronwall's inequality to deduce that

$$\frac{1}{h}|x_{\mu+h}(t) - x_\mu(t)| \leq \left(\|b\|_\infty + C\delta T\right) e^{\|A\|_\infty T}.$$

The point here is that the bound on the right is uniform in h and t. This allows us to return to the identity (2.8.1), divide by h, and use the Taylor series and the dominated convergence theorem to pass to the limit under the integral, obtaining

$$\frac{\partial x_\mu(t)}{\partial \mu} = \int_0^t \left(A(s)\frac{\partial x_\mu(s)}{\partial \mu} + b(s)\right) ds. \tag{2.8.2}$$

Since $y(s)$ is the unique solution to this equation, the proof that $x_\mu(t)$ is C^1 in μ is complete.

In summary, the argument has three parts. In the first, we identify a candidate equation for the derivative and establish uniqueness for it. In the second, we use the identity (2.8.1) and Gronwall's inequality to establish an a priori

---

solution to the matrix valued differential equation

$$\frac{d}{dt}S(t;s) = A(t)S(t;s), \quad S(s;s) = I, \quad t \geq s.$$

Then the solution to the differential equation for $y(t)$ is

$$y(t) = \int_0^t S(t;s)b(s)\, ds.$$

We won't need this general solution, but it is useful to understand the difference between the fundamental solution for linear constant coefficient equations, and linear non-autonomous systems.

bound on the finite differences that is uniform for $t \in [0,T]$ and the parameter range $\nu \in [\mu-\delta, \mu+\delta]$. In the last step, we pass to the limit $h \to 0$ and we use the a priori bounds to justify the interchange of limits.

The extension of these ideas to arbitrary k does not require much more than some careful book-keeping for derivatives. Let us denote the higher derivatives of the solution by

$$y^{(l)} = \frac{\partial^l}{\partial \mu} x_\mu(t), \quad 2 \le l \le k.$$

The structure of the differential equation for $y^{(l)}$ obtained by differentiating the equation above has the form

$$\frac{d}{dt} y^{(l)} = A(t) y^{(l)} + b^{(l)}(t), \quad y^{(l)}(0) = 0,$$

where A is exactly as above and $b^{(l)}$ is a polynomial in the first l derivatives of f with respect to x and the first $l-1$ derivatives of x with respect to μ (i.e. $y, y^{(1)}, \ldots, y^{(l-1)}$). The precise form of this expression is largely irrelevant; what matters again is that it is bounded for $t \in [0,T]$. It immediately follows that there is a unique solution $y^{(l)}(t)$ for $t \in [0,T]$ for $0 \le l \le k$. A somewhat tedious finite-difference argument as above is now required to complete the proof that $y^{(l)}(t)$ is indeed the l-th derivative of the solution $x(t;\mu)$ with respect to μ. □

5. The standard construction of bump functions goes as follows. Consider the function

$$\varphi(x) = \begin{cases} e^{-1/x}, & x > 0, \\ 0, & x \le 0. \end{cases}$$

(a) Show that this function is infinitely differentiable at zero. That is, show that all derivatives of $e^{-1/x}$ on the region $x > 0$ vanish as $x \to 0$.

(b) Given an interval $[a,b]$ show that there is a C^∞ function φ_δ that is identically equal to 1 on $[a,b]$ but vanishes when $x \le a-\delta$ and $x \ge b+\delta$ for any $\delta > 0$.

(c) Extend this idea to $\mathbb{R}^n$, constructing a bump function that is identically equal to one in a box, but vanishes outside a transition layer of width δ

Proof. (a) Let us compute the first two of derivatives of $\varphi(x)$ in the region $x > 0$. By the chain rule

$$\varphi' = \frac{2}{x^2}\varphi := q_1(x)\varphi, \quad \varphi'' = \frac{4}{x^4}\varphi - \frac{4}{x^3}\varphi := q_2(x)\varphi.$$

Proceeding in this manner, we see that $\varphi^{(n)}$, the n-th derivative of φ, is of the form $q_n(x)\varphi$ where $q_n(x) = 2^n x^{-2n} + O(x^{-2n+1})$ as $x \to 0$. The exponential grows faster than any polynomial at infinity; thus, $\lim_{x\to 0} \varphi^{(n)}(x) = 0$.

(b) We first construct a C^∞ function $\psi(x)$ such that

$$\psi(x) > 0, \quad |x| < 1, \quad \psi(x) = 0, \quad |x| \ge 1, \quad \int_{\mathbb{R}} \psi(x)\,dx = 1. \tag{2.8.3}$$

To construct such a function, let us first modify the example of (a) a bit. For any $a \in \mathbb{R}$, let $\varphi_a(x) = e^{-1/(x-a)}$ in the region $x > a$ and $\varphi_a(x) = 0$ for $x \leq a$. Since $\varphi_a(x)$ is just a shifted version of φ, it is C^∞. Similarly, introduce the decreasing function $\tilde{\varphi}_a(x) = e^{1/(x-a)}$ for $x < a$ and $\tilde{\varphi}_a(x) = 0$ vanishes for $x \geq 0$. This function is the reflection of $\varphi_a(x)$ about the point $x = a$ and it is also C^∞. Next let

$$\psi(x) = C\varphi_1(x)\tilde{\varphi}_{-1}(x),$$

and choose the constant C so that $\int_{\mathbb{R}} \psi(x)\,dx = 1$.

Once this bump function has been constructed parts (b) and (c) of this question may be solved by mollification. First (b). For any $\theta > 0$ let

$$\psi_\theta(x) = \frac{1}{\theta}\psi\left(\frac{x}{\theta}\right).$$

This scaling factor ensures that $\psi_\theta(x)$ is positive only on an interval of size 2θ and that its integral remains unity. Now given the interval $[a, b]$ and $\delta > 0$ consider the indicator function $\mathbf{1}_{[a-\delta/2, b+\delta/2]}$, choose any value of $\theta < \delta/4$ and then consider the mollification

$$h(x) = \psi_\theta(x) \star \mathbf{1}_{[a-\delta/2,b+\delta/2]}(x) := \int_{a-\delta/2}^{b+\delta/2} \psi_\theta(x-y)\,dy.$$

Since $\theta < \delta/4$, when $x \in [a, b]$, the support of $\psi_\theta(x - \cdot)$ is contained within the domain of integration and the integral is 1. On the other hand, when $x < a - \delta$ the support of $\psi_\theta(x - \cdot)$ is disjoint from the domain of integration and the integral vanishes.

(c) The same idea can be extended to $\mathbb{R}^n$. First, we construct an n-dimensional bump function supported in the cube $[-1, 1]^n$ by considering the product

$$\psi^{(n)}(x) := \psi(x_1)\psi(x_2)\cdots\psi(x_n).$$

Then

$$\int_{\mathbb{R}^n} \psi^{(n)}(x)\,dx = \prod_{j=1}^{n} \int_{\mathbb{R}} \psi(x_j)\,dx_j = 1.$$

As in the previous example, we may rescale the domain of integration to the cube $[-\theta, \theta]^n$ and normalize accordingly, setting

$$\psi_\theta^{(n)}(x) = \frac{1}{\theta}\psi^{(n)}\left(\frac{x}{\theta}\right).$$

Given a closed box $K = [a_1, b_1] \times [a_2, b_2] \times \ldots [a_n, b_n]$ and a positive number $\eta > 0$ define the η-thickened box $K_\eta = [a_1 - \eta, b_1 + \eta] \times [a_2 - \eta, b_2 + \eta] \times \ldots [a_n - \eta, b_n + \eta]$. Then as in the previous example, for every $\theta < \delta/2$ the function

$$h^{(n)}(x) = \psi_\theta^{(n)} \star \mathbf{1}_{K_{\delta/2}}(x)$$

is identically one on the box K and vanishes outside the box K_δ. $\square$

6. Consider the initial value problem $\dot{x} = f(x)$, $x(0) = x_0$ with a continuous vector field $f : \mathbb{R}^n \to \mathbb{R}^n$ and $x_0 \in \mathbb{R}^n$, and $t \in [0, T]$ for some fixed $T > 0$.

The forward Euler scheme is an approximation method for this differential equation of the following form: the approximation $x^{(N)}(t)$ is a Lipschitz function such that (i) $x^{(N)}(0) = x_0$; (ii) $x^{(N)}(t)$ is piecewise linear with slope $f(nh)$ on the intervals $[nh, (n+1)h)$‘′, $n = 0, 1, \ldots, N-1$, $h = T/N$.

Prove that as $N \to \infty$ this scheme converges in $C([0, T]; \mathbb{R}^n)$ to a Lipschitz function $x(t)$ that solves the initial value problem, thus establishing another proof of Peano's theorem.

Proof. The approximation scheme satisfies the integral equation

$$x^{(N)}(t) - x_0 = \int_0^t f^N(s)\, ds,$$

where $f^N(s)$ denotes the piecewise constant function that takes the value $f(x^{(N)}(nh))$ on the interval $[nh, (n+1)h)$. As a consequence:

1. $|x^{(N)}(t) - x^{(N)}(s)| \leq \|f\|_\infty |t - s|$, $0 \leq s \leq t < T$.
2. $\sup_{t \in [0,T]} |x^{(N)}(t)| \leq |x_0| + T\|f\|_\infty$.

Thus, the sequence $x^{(N)}$ is precompact in $C([0, T]; \mathbb{R}^n)$. By the Arzela-Ascoli theorem, we may assume that a subsequence $x^{(N_k)}$ is uniformly convergent to a limit denoted $x(t)$. Moreover, since f is continuous, we also see that the piecewise constant functions $f^{N_k}\left(x^{(N_k)}(t)\right)$ converge uniformly to the composed function $f \circ x \in C([0, T]; \mathbb{R}^n)$. Thus, we may take limits in the integral equation above to find

$$x(t) - x_0 = \int_0^t f(x(s))\, ds.$$

$\square$

Chapter 3

Phase portraits and the Flow

In this chapter we introduce the basics of phase portraits as well as a rigorous definition of the flow. Phase portraits are simple geometric caricatures that capture the essence of the flow. Of course, we can only draw pictures in 1,2 and (occasionally) 3 dimensions, but the use of such geometric intuition greatly facilitates the study of dynamical systems.

3.1 A first glance at phase portraits

The simplest solutions to differential equations are fixed points. Given $f : \mathbb{R}^n \to \mathbb{R}^n$ these are the set of points $a \in \mathbb{R}^n$ such that $f(a) = 0$. Fixed points are also termed equilibria or critical points and the solution curves are often called orbits or trajectories. Note that uniqueness of solutions always means that a trajectory can never contain a fixed point unless the trajectory consists of solely the fixed point.

One dimensional phase portraits are almost trivial. Suppose $f : \mathbb{R} \to \mathbb{R}$, such that f is C^1. Given the graph of f, we first determine its zeros. These are our fixed points. If we start at (i.e. if we pick x_0 to be) any a such that $f(a) = 0$, we stay there forever.

Next, in the interval between zeros, $f(x) > 0$ or $f(x) < 0$. This means that $x(t)$ is either strictly increasing or strictly decreasing, except that sometimes one has to be more careful, such as for tangencies. These ideas are illustrated in Figure 3.1.1.

These pictures should be intuitive. Note though that it is because of Picard's Theorem that we can say that the trajectories can never pass through zeros. 2D phase portraits have a lot more complexity as can be seen by a glance at Figure 3.1.2. We will build more intuition for these phase portraits by first studying explicitly solvable linear systems and then interpreting these solution formulas geometrically in Section 3.3 below.

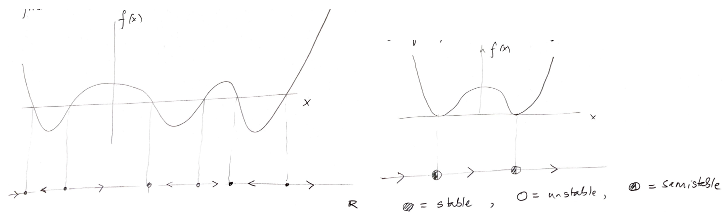

Figure 3.1.1: Phase portraits in 1D

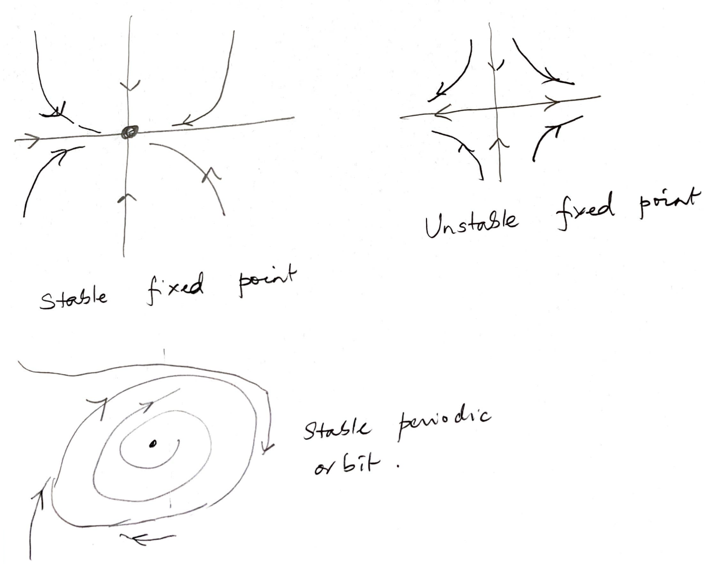

Figure 3.1.2: Some phase portraits in 2D

3.2 Linear autonomous equations

Let $\mathbb{M}_n$ denote the space of $n \times n$ real matrices. Assume given $A \in \mathbb{M}_n$ and consider the linear ODE

$$\dot{x} = Ax, \quad x \in \mathbb{R}^n. \tag{3.2.1}$$

The function $f(x) = Ax$ is globally Lipschitz because

$$|f(x) - f(y)| = |Ax - Ay| = |A(x - y)| \leq ||A|||x - y|.$$

Thus, the initial value problem with the initial condition $x(0) = x_0$ has a unique solution. Moreover, the solution is given by the formula

$$x(t) = e^{tA}x_0. \tag{3.2.2}$$

The matrix exponential is discussed in greater detail below. For now, note that equation (3.2.5) shows that the formula (3.2.2) provides a solution to (3.2.1).

3.2.1 The matrix exponential

Let us study the exponential of a matrix more carefully [1]. We define it through the infinite series

$$e^M := \sum_{m=0}^{\infty} \frac{M^m}{m!}. \tag{3.2.3}$$

Convergence of the series is established as follows. First, we note the estimate

$$||M^m|| \leq ||M||^m, \quad ||M|| = \sup_{|x|=1} |Mx|, \quad x \in \mathbb{R}^n.$$

This estimate allows us to bound the finite sums in the series (3.2.3) as follows:

$$\sum_{m=0}^{P} \frac{M^m}{m!} \leq \sum_{m=0}^{P} \frac{||M||^m}{m!} < \sum_{m=0}^{\infty} \frac{||M||^m}{m!} = e^{||M||} < \infty. \tag{3.2.4}$$

Thus the series (3.2.3) has an infinite radius of convergence and the derivative of e^M may be computed by differentiating term by term. Therefore,

$$\frac{d}{dt}e^{tA} = \frac{d}{dt}\sum_{m=0}^{\infty} \frac{t^m A^m}{m!} = \sum_{m=1}^{\infty} \frac{mt^{m-1}A^m}{m!} = \left(\sum_{m=0}^{\infty} \frac{t^m A^m}{m!}\right) A = Ae^{tA}. \tag{3.2.5}$$

It was necessary to establish this formula from scratch, because the matrix exponential has some important differences with the exponential of a scalar. In particular, $e^{A+B} \neq e^A e^B$ except in the special situation where A and B are matrices that commute with each other (i.e. $AB = BA$).

The infinite sum (3.2.3) serves as a useful definition of the matrix exponential. However, in order to compute the exponential, we diagonalize A or (if A is not diagonalizable consider its Jordan decomposition). For simplicity, we will focus on the case where A is diagonalizable. We may then write

$$A = U\Lambda U^{-1} \tag{3.2.6}$$

where Λ is a diagonal matrix of the eigenvalues of A and U contains the eigenvectors of A in the same order as the eigenvalues on the diagonal of Λ. Equation (3.2.6) yields

$$A^2 = U\Lambda U^{-1}U\Lambda U^{-1} = U\Lambda^2 U^{-1},$$

and proceeding inductively we find

$$A^m = U\Lambda^m U^{-1}, \quad m = 1, 2, \ldots$$

Using the infinite series (3.2.3) again we find that

$$e^{tA} = Ue^{t\Lambda}U^{-1}. \tag{3.2.7}$$

[1] We use the phrases "exponential of a matrix" and "matrix exponential" to mean the same thing.

The matrix $e^{t\Lambda}$ is a diagonal matrix with entries

$$e^{t\Lambda} = \begin{pmatrix} e^{t\lambda_1} & & \\ & \ddots & \\ & & e^{t\lambda_n} \end{pmatrix} \tag{3.2.8}$$

It is this matrix that determines the behavior of e^{tA} as $t \to \infty$ [2].

Both U and Λ may be complex even though A is real. However, the complex eigenvalues always appear in pairs of complex conjugates, and e^{tA} is always real as is clear from the infinite series (3.2.3). For a given eigenvalue λ_i, we have three possibilities for asymptotic behavior. If $\mathrm{Re}(\lambda_i) < 0$, then $|e^{t\lambda_i}| \to 0$ as $t \to \infty$. If $\mathrm{Re}(\lambda_i) > 0$, then $|e^{t\lambda_i}| \to \infty$ as $t \to \infty$. And finally if $\mathrm{Re}(\lambda_i) = 0$, then $|e^{t\lambda_i}| = 1$ for all values of t.

3.2.2 Linear non-autonomous systems

In general, we call a differential equation of the form $\dot{x} = f(x,t)$ *non-autonomous* because f depends explicitly on t, rather than only depending on it via $x(t)$. Any non-autonomous system can be made autonomous by adding t as a new variable. We define the ordered pair $\tilde{x} = (x,t)$ and rewrite the differential equation as follows

$$\left\{ \begin{array}{c} \dot{x} = f(x,t) \\ \dot{t} = 1 \end{array} \right\} \iff \dot{\tilde{x}} = \tilde{f}(\tilde{x}) \tag{3.2.9}$$

In this setup, $\tilde{f} : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ is defined by $\tilde{x} \mapsto \begin{pmatrix} f(x,t) \\ 1 \end{pmatrix}$. This is a valid construction, but it is not very satisfactory since time plays a special role in dynamical systems. Linear nonautonomous systems arise when we examine the linearization about a solution to the equation $\dot{x} = f(x)$.

3.3 Linear systems in 2D

Let us now use the solution formula(3.2.2) to draw the phase portraits of some two-dimensional systems.

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

For any initial condition $x_0 = \begin{pmatrix} a \\ b \end{pmatrix}$

$$e^{tA}x_0 = ae^{-t}\begin{pmatrix} 1 \\ 0 \end{pmatrix} + be^{-2t}\begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{3.3.1}$$

[2]The following notational convention is used here. Empty terms in a matrix are assumed to be zero.

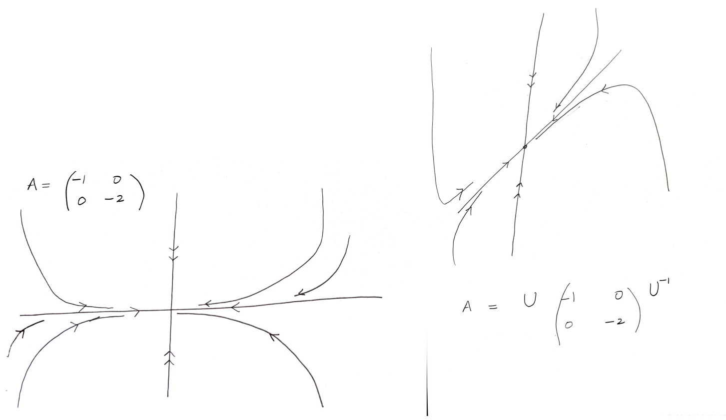

Figure 3.3.1: Example 1 and Example 2. The double arrows correspond to the eigenvalue -2 and denote a strongly stable direction. This idea will be reconsidered when studying invariant manifolds.

Both terms decay as $t \to \infty$ and $x(t) \to 0$ as $t \to \infty$. Since the second term decays much faster than the first, all trajectories with $a \neq 0$ are asymptotic to the x-axis as $t \to \infty$.

Example 2. $A = U \begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix} U^{-1}$ *where* U *consists of two linearly independent column vectors* u_1, u_2.

The phase portrait of Example 1 is linearly transformed into the phase portrait of this example through equation (3.2.7). The critical point in both these examples is called a *stable node.*

Example 3. $A = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$.

This critical point is called a *saddle point* or simply a *saddle.* See Figure 3.3.2.

Example 4. $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$

In this example, A is not a diagonal matrix, so we need to compute the eigenvalues. The characteristic polynomial is $\det(\lambda I - A) = \begin{vmatrix} \lambda & -1 \\ 1 & \lambda \end{vmatrix} = \lambda^2 + 1$, which means the eigenvalues are $\lambda = \pm i$. The real parts of both eigenvalues are 0, so that $|e^{t\lambda}| = 1$ for all t. We may compute e^{tA} directly using the infinite series (3.2.3) or compute the eigenvectors to find that

$$e^{tA} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}. \tag{3.3.2}$$

This critical point is called a *center.* See Figure 3.3.2.

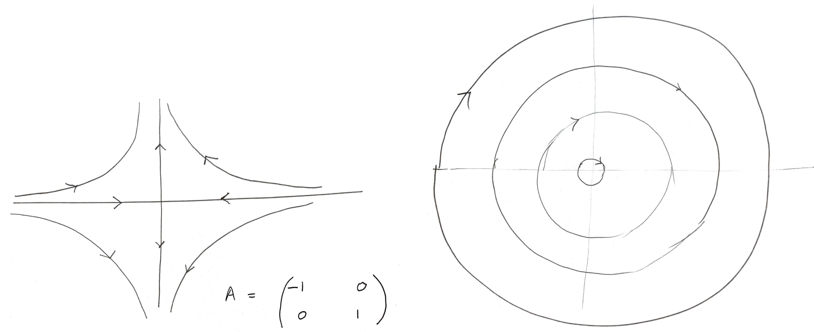

Figure 3.3.2: Example 3 and Example 4.

Example 5. $A = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}$

This example is closely related system to Example 4. The characteristic polynomial is $(\lambda - \alpha)^2 + \beta^2$ and the eigenvalues are $\lambda = \alpha \pm i\beta$. For $\alpha < 0$, the diagram is a *stable spiral* since the trajectories are spiralling inward to the critical point. The diagram for $\alpha > 0$ is an *unstable spiral.* If $\alpha = 0$, the critical point is a center. See Figure 3.3.3.

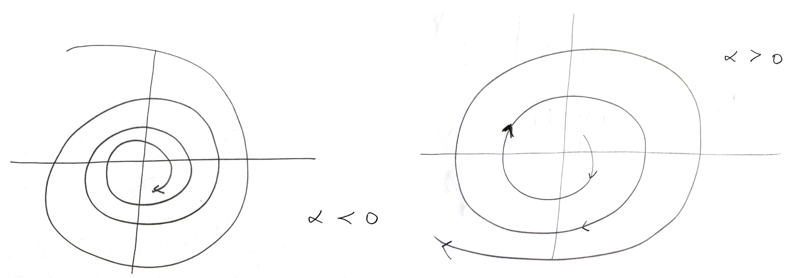

Figure 3.3.3: Example 3 and Example 4.

3.4 Existence of a Lipschitz flow

In Chapter 2, we established well-posedness of the initial value problem

$$\begin{cases} \dot{x} = f(x) \\ x(0) = x_0 \end{cases} \tag{3.4.1}$$

The main idea of the flow is to focus not on the initial value problem for a fixed initial condition, but to think simultaneously about the totality of solutions in phase space. The examples in the previous section illustrate the utility of this viewpoint. In this section, we will establish the existence of a flow rigorously.

We first switch to notation that is more convenient for the geometric viewpoint. When discussing equation (3.4.1) we use x_0 to denote the initial condition and the notation $x(t; x_0)$ to denote the solution with initial condition x_0. When discussing the flow it is more convenient to write x instead of x_0 for the initial condition and $\varphi_t(x)$ for the solution with this initial condition. The initial value problem (3.4.1) is then rewritten as

$$\begin{cases} \frac{\partial}{\partial t}\varphi_t(x) = f(\varphi_t(x)), \\ \varphi_0(x) = x. \end{cases} \tag{3.4.2}$$

As in Picard's theorem we will first establish the existence and uniquess of the flow map under a global Lipschitz condition in order to focus attention on the main new ideas. We will then establish more refined results.

Theorem 24. *Assume $f : \mathbb{R}^n \to \mathbb{R}^n$ is L-Lipschitz. Then there exists a family of maps $\varphi : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, $(x, t) \mapsto \varphi_t(x)$ such that*

1. *Equation (3.4.2) holds for all $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$.*

2. *The flow maps form a 1-parameter group of transformation of $\mathbb{R}^n \to \mathbb{R}^n$ satisfying*
$$\varphi_s(\varphi_t(x)) = \varphi_t(\varphi_s(x)) = \varphi_{t+s}(x), \quad s, t \in \mathbb{R}. \tag{3.4.3}$$

3. *The maps φ_t are bi-Lipschitz homeomorphisms of $\mathbb{R}^n$. That is, φ_t and its inverse φ_{-t} are Lipschitz maps of $\mathbb{R}^n$ into itself satisfying the estimate*
$$|\varphi_t(x) - \varphi_t(y)| \leq e^{L|t|} |x - y|, \quad x, y \in \mathbb{R}^n. \tag{3.4.4}$$

Proof. The first two assertions of the theorem are consequences of Theorem 8. The third assertion is seen as follows. We compare the solution with two different initial conditions

$$\varphi_t(x) - \varphi_t(y) = x - y + \int_0^t (f(\varphi_s(x)) - f(\varphi_s(y))\, ds. \tag{3.4.5}$$

Assume $t > 0$ to be concrete. The argument for $t < 0$ is very similar. We take absolute values and use the Lipschitz condition to obtain

$$|\varphi_t(x) - \varphi_t(y)| \leq |x - y| + L \int_0^t |\varphi_s(x) - \varphi_s(y)|\, dy. \tag{3.4.6}$$

The inequality (3.4.4) follows from Gronwall's inequality. $\square$

3.5 Existence of a smooth flow

Definition 25. A C^k diffeomorphism of an open set $U \subset \mathbb{R}^n$ is a C^k map $\varphi : U \to U$ such that φ is one-one, onto and has a C^k inverse φ^{-1}.

Theorem 26. *Assume $f : \mathbb{R}^n \to \mathbb{R}^n$ is C^k and $\sup_{x\in\mathbb{R}^n} \|Df(x)\| = L < \infty$. Then there exists a family of maps $\varphi : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, $(x,t) \mapsto \varphi_t(x)$ such that*

1. *Equation (3.4.2) holds for all $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$.*
2. *The flow maps form a 1-parameter group of transformation of $\mathbb{R}^n \to \mathbb{R}^n$ satisfying*
$$\varphi_s(\varphi_t(x)) = \varphi_t(\varphi_s(x)) = \varphi_{t+s}(x), \quad s,t \in \mathbb{R}. \tag{3.5.1}$$
3. *For each $t \in \mathbb{R}$, the map φ_t is a C^k diffeomorphism of $\mathbb{R}^n$.*

The difference between Theorem 24 and Theorem 26 lies only in the last assertion concerning the smoothness of the map. A simple argument with Gronwall's inequality sufficed for Theorem 24. But we need a new idea to understand the smoothness of the flow in the initial conditions.

The main issue is this: how does one compute the derivative of the flow with respect to the initial conditions? Since the only information we have on the flow is that it satisfies equation (3.4.2), we differentiate the initial value problem (3.4.1) to get the *equation of variations*

$$\begin{cases} \frac{\partial}{\partial t} D\varphi_t(x) = Df(\varphi_t(x))D\varphi_t(x) \\ D\varphi_0(x) = I \end{cases} \tag{3.5.2}$$

Recall that $Df(x)$ is the $n\times n$ matrix with entries defined by $(Df)_{ij} = \frac{\partial f_i}{\partial f_j}$. This is a linear equation for the matrix $D\varphi_t(x)$ and it is helpful to introduce notation that makes this transparent. Fix x and let $B(t) = D\varphi_t(x)$, $A(t) = Df(\varphi_t(x))$, we can rewrite equation (3.5.2) to clearly display its character:

$$\frac{dB}{dt} = A(t)B, \quad B(0) = I. \tag{3.5.3}$$

By the definition of $A(t)$, we see that $\sup_t ||A(t)|| \le L < \infty$, which means the right hand side of equation (3.5.3) is Lipschitz in B and therefore has a unique global solution. Thus, $B(t)$ is a candidate for $D\varphi_t(x)$, and to prove Theorem 26 we must show that

1. The solution $B(t)$ to equation (3.5.3) is invertible.
2. The solution $B(t)$ is the derivative $D\varphi_t(x)$.

The proofs of these assertions are quite different. The first statement is proven by deriving an equation for $\det(B(t))$. The second assertion must be justified from first principles (i.e. starting from the definition of the derivative).

Lemma 5. *Suppose the matrix $B(t)$ solves the linear equation $\dot{B}(t) = A(t)B$. Then $\det(B)$ solves the linear equation*

$$\frac{d}{dt}\det(B) = \mathrm{Tr}(A)\det(B), \tag{3.5.4}$$

. In particular, we have

$$\det B(t) = e^{\int_0^t \mathrm{Tr}(A(s))ds} \det B(0), \tag{3.5.5}$$

so that $B(t)$ is invertible if and only if $B(0)$ is.

Remark 27. The trace of a square matrix A, defined by $\mathrm{Tr}(A) = \sum_{i=1}^n A_{ii}$, is the sum of the entries along the main diagonal of A.

Remark 28. To get a sense of why the above lemma is tricky, consider 2×2 matrices. Let $B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$; then $\det(B) = b_{11}b_{22} - b_{12}b_{21}$, and correspondingly we have $\dot{\det}(B) = \dot{b}_{11}b_{22} + b_{11}\dot{b}_{22} - \dot{b}_{12}b_{21} - b_{12}\dot{b}_{21}$. We could supposedly solve for this by substituting in the derivatives of each b_{ij}, but this is clearly tedious.

Proof. First we note that the derivatives of the determinant can be computed at the identity, i.e. if $B = I$, then $\det(B + \varepsilon M) = \det(I + \varepsilon M)$ for any $\varepsilon > 0$ and we can expand in ε using the formula for the determinant.

Let S_n denote the permutation group on n symbols and recall that the determinant of an $n \times n$ matrix X is

$$\det(X) = \sum_{\sigma \in S_n} (-1)^\sigma X_{1\sigma_1} X_{2\sigma_2} ... X_{n\sigma_n}, \tag{3.5.6}$$

where $(-1)^\sigma$ denotes the sign of the permutation. Note also that

$$(I + \varepsilon M)_{ij} = \delta_{ij} + \varepsilon m_{ij} = \varepsilon m_{ij}, \quad \text{if} \quad i \neq j.$$

Here δ_{ij} is the Kronecker delta.

We only need to worry about terms in the sum (3.5.6) that are $O(\varepsilon)$ as $\varepsilon \to 0$, since these terms will dominate the equation. For this, we only need to consider the identity permutation in the sum, since any $\sigma \in S_n$ that is not the identity will yield terms that are $O(\varepsilon^2)$. Consider for example the permutation that only switches columns 1 and 2; then the resulting term is:

$$(\delta_{12} + \varepsilon m_{12})(\delta_{21} + \varepsilon m_{21})(\delta_{33} + \varepsilon m_{33})...(\delta_{nn} + \varepsilon m_{nn}) = \varepsilon^2 m_{12} m_{21} (1 + \varepsilon m_{33})... \tag{3.5.7}$$

Since this is $O(\varepsilon^2)$, we can disregard it and all similar terms as $\varepsilon \to 0$.

Thus, the determinant of $I + \varepsilon M$ can be expressed solely as the product of the terms along its diagonal, with an $O(\varepsilon)$ error term:

$$\det(I + \varepsilon M) = (1 + \varepsilon m_{11})...(1 + \varepsilon m_{nn}) + O(\varepsilon^2). \tag{3.5.8}$$

Taking the derivative of the above with respect to ε and evaluating it at $\varepsilon = 0$, we find that

$$\frac{d}{d\varepsilon} \det(I + \varepsilon M)\bigg|_{\varepsilon=0} = m_{11} + m_{22} + ... + m_{nn} = \mathrm{Tr}(M). \tag{3.5.9}$$

This calculation proves Lemma 5 when $B = I$.

Let us now reduce the general case to this case. Assume t is such that $B(t)$ is invertible (this is certainly true for t in a neighborhood of 0 since $B(0) = I$. Rewrite the $\dot{B} = AB$ as $\dot{B}B^{-1} = A$. Fixing t and $B(t)$, consider $\det B(t+s)$; we can write it as $\det B(t+s)B^{-1}(t)B(t) = \det(B(t+s)B^{-1}(t)) \cdot \det B(t)$, where $B(t+s)B^{-1}(t)$ as a function of s. Note

$$B(t+s)B^{-1}(t)\big|_{s=0} = I,$$

and

$$\frac{d}{ds}B(t+s)B^{-1}(t)\bigg|_{s=0} = A.$$

Then by the previous calculation, we find that

$$\frac{d}{ds}\det(B(t+s)B^{-1}(t))\bigg|_{s=0} = \mathrm{Tr}(A).$$

Separating variables and integrating, we find

$$\begin{aligned} \det B(t) &= \left(e^{\int_0^t \mathrm{Tr}(A(s))ds}\right)\det B_0 & (3.5.10)\\ &= e^{\int_0^t \mathrm{Tr}(A(s))ds} \text{ when } B_0 = I. & (3.5.11)\end{aligned}$$

We now use a continuation argument to see that this identity holds for the entire interval of existence of solutions. □

Lemma 6. *Assume that f satisfies the hypothesis of Theorem 26 with $k = 1$. Then for every $t \in \mathbb{R}$, the flow is differentiable and*

$$D\varphi_t(x) = B(t), \tag{3.5.12}$$

where $B(t)$ is the unique solution to equation (3.5.3).

Proof. Fix an initial point $x \in \mathbb{R}^n$ as well as a tangent vector $v \in \mathbb{R}^n$. We must show that

$$\lim_{h\to 0}\left|\frac{1}{h}\left(\varphi_t(x+hv) - \varphi_t(x)\right) - B(t)v\right| = 0. \tag{3.5.13}$$

We know that $\varphi_t(x)$ and $B(t)$ solve the integral equations

$$\begin{aligned}\varphi_t(x+hv) - \varphi_t(x) &= hv + \int_0^t \left(f(\varphi_s(x+hv)) - f(\varphi_s(x))\right)ds, & (3.5.14)\\ B(t)v &= v + \int_0^t Df(\varphi_s(x))B(s)v\,ds. & (3.5.15)\end{aligned}$$

Now take the difference between these terms and use Taylor's remainder theorem and the dominated convergence theorem (as used in Corollary 1) to complete the proof of the lemma. □

Proof of Theorem 26. Lemma 6 suffices to establish Theorem 26 in the situation when $k = 1$.

The underlying principles generalize to arbitrary k. We first derive a differential equation analogous to (3.5.3) for the k^{th} derivative; we then show that the k-th derivative satisfies the equation from first principles. Since the second step is a calculus exercise not essentially different from Lemma (6), we won't prove it. We further simplify matters by sketching the proof in $\mathbb{R}$ to provide the main idea - in $\mathbb{R}^n$ the higher (say k^{th}) derivatives have k indices and require careful bookkeeping, but the character of the equation for the k^{th} derivative is very similar to the one in $\mathbb{R}$.

To this end, assume $f : \mathbb{R} \to \mathbb{R}$ is C^k and consider the initial value problem

$$\begin{cases} \frac{\partial}{\partial t}\varphi_t(x) = f(\varphi_t(x)) \\ \qquad \varphi_0(x) = x. \end{cases} \tag{3.5.16}$$

We denote the derivatives of v by f', f'', and $f^{(p)}$ for the p^{th} derivatives with respect to x; the same notation will be used for $\varphi_t(x)$. Then taking the derivative with respect to x of (1.8.22) yields:

$$\begin{cases} \frac{\partial}{\partial t}\varphi_t'(x) = f'(\varphi_t(x))\varphi_t'(x) \\ \qquad \varphi_0'(x) = 1 \end{cases} \tag{3.5.17}$$

We have already studied this problem in the form $\dot{B} = A(t)B$, $B(0) = I$ for $x \in \mathbb{R}^n$, where $B(t) = \varphi_t'(x)$ and $A(t) = f'(\varphi_t(x))$. Differentiating in x once again, we obtain the following:

$$\frac{\partial}{\partial t}\varphi_t''(x) = f'(\varphi_t(x))\varphi_t''(x) + f''(\varphi_t(x))(\varphi_t'(x))^2. \tag{3.5.18}$$

Let $B_2(t) = \varphi_t''(x)$ and $A_2(t) = f''(\varphi_t(x))$. Then we can rewrite equation (3.5.18) as the nonhomogeneous linear equation

$$\frac{dB_2}{dt} = A(t)B_2 + A_2(t)B^2 \tag{3.5.19}$$

Equation (3.5.19) can be solved using the variation of constants formula for linear equations

$$B_2(t) = e^{\int_0^t A(s)ds}B_2(0) + \int_0^t e^{\int_s^t A(r)dr}A_2(s)B^2(s)ds \tag{3.5.20}$$

$$= \int_0^t e^{\int_s^t A(r)dr}A_2(s)B^2(s)ds. \tag{3.5.21}$$

because $\varphi_t(0) = x$ implies that $B_2(0) = 0$.

Since f is C^k (where $k \geq 2$), we have $A_2(s) = f''(\varphi_s(x))$, and correspondingly $\sup_{0 \leq s \leq t} |A_2(s)| = \sup_{0 \leq s \leq t} |f''(\varphi_s(x))| < \infty$. Moreover, the terms $A(s)$ and $B(s)$ have already been controlled on that same interval of existence by the first derivative $f'(\varphi_s(x))$. Thu,s $B_2(t)$ is well-defined.

As with the first derivative $D\varphi_t(x)$, this is the critical step to concluding that $D^2\varphi_t(x)$ is well-defined. For higher-order derivatives the algebra gets progressively messier, but the underlying principles are the same:

1. Derive a differential equation for $D^k\varphi_t(x)$, and observe that it has the form $\frac{d}{dt}B_k = A(t)B_k(t) + N(A, ...A_k, B, ...B_{k-1})$, where the A-terms in N involve the first k derivatives of f and the B-terms in N involve the first $k-1$ derivatives of $\varphi_t(x)$. This gives a linear non-autonomous differential equation which can be solved by the variation of constants formula.

2. Show that B_k is indeed the k^{th} derivative of φ_t in x by using finite differences and passing to the limit as in Lemma 6.

□

3.6 Asymptotic behavior

A central theme in dynamical systems is to decompose the flow into a 'few pieces that matter'. We have seen examples of this above: for linear systems, what matters are critical points and the stable, unstable and center eigenspaces. This idea will be extended to nonlinear systems through the use of invariant manifolds. Similarly, phase portraits are an impressionistic sketch of the global dynamics which contain a great deal of information.

In this section, we consider the more abstract idea that the asymptotic behavior of a dynamical system is captured by invariant sets. A fundamental example of an invariant set is the ω-limit set defined below. In order to prevent technicalities, we make the following standing assumptions in this section.

1. The phase space U is an open set in $\mathbb{R}^n$.

2. $\varphi_t : U \to U$ is a C^1 flow defined for $t \in (-\infty, \infty)$.

Definition 29. A set $A \subseteq U$ is positively invariant if $\varphi_t(A) = A$ for all $t \geq 0$. Similarly, a set is negatively invariant if $\varphi_t(A) = A$ for all $t \leq 0$. A set is *invariant* if it is positively and negatively invariant.

Remark 30. The concept of positive invariance requires only that the flow is defined only for $t \geq 0$. An invariant set may exist even under the weaker assumption that the flow is defined for all initial conditions only for $t \geq 0$ (since this does not preclude global existence for certain special initial conditions). Thus, our standing assumption on existence of solutions is a little stronger than necessary. However, it is simpler at the first pass to focus on the concept of invariant sets without worrying about global (in time) existence of the flow. When the invariant set is compact, we may always modify the vector field with bump function so that the assumptions of this section apply.

Definition 31. Suppose $B \subseteq U$. The ω-limit set of B is

$$\omega(B) = \{y \in U \,|\, \exists\, t_n \to \infty, x_n \in B \text{ such that } \varphi_{t_n}(x_n) \to y\}. \tag{3.6.1}$$

When $B = \{x\}$ we write $\omega(x)$ instead of $\omega(\{x\})$.

Definition 32. (Positive orbit). The positive orbit $\gamma^+(x)$ of x is

$$\gamma^+(x) = \{y \,|\, y = \varphi_t(x) \text{ for some } t \geq 0\}. \tag{3.6.2}$$

Remark 33. When the flow is defined for $t \leq 0$ the analogous notions to the ω-limit set and positive orbit are the α-limit set and negative orbit $\gamma^-(x)$ respectively.

Let us now establish some fundamental properties of these sets.

Lemma 7. $\omega(x) = \omega(\gamma^+(x))$.

Proof. Clearly, $\omega(x) \subseteq \omega(\gamma^+(x))$ since $B \subseteq B'$ implies $\omega(B) \subseteq \omega(B')$. On the other hand, if $y \in \omega(\gamma^+(x))$, then there is $\{t_n\}_{n=1}^\infty$ and $\{x_n\}_{n=1}^\infty \subseteq \gamma^+(x)$ with $\varphi_{t_n}(x_n) \to y$. But, $x_n = \varphi_{x_n}(x)$ for some $s_n \geq 0$ since $x_n \in \gamma^+(x)$. Thus, $\varphi_{t_n+s_n}(x) \to y$ implies that $y \in \omega(x)$. □

Lemma 8.

$$\omega(B) = \bigcap_{t\geq 0} \overline{\bigcup_{s\geq t} \varphi_t(B)}. \tag{3.6.3}$$

Proof. This is on HW 2. The proof involves checking that $\omega(B)$ as defined in Definition 3.6.1 is contained in, and contains, the set on the right hand side above. A full proof will be added in with solutions to HW 2. □

Theorem 34. *$\omega(B)$ is closed and invariant.*

Proof. $\omega(B)$ is closed by Lemma 8, since an arbitrary intersection of closed sets is closed. Suppose $y \in B$ and choose $t \in \mathbb{R}$. We know that there is a sequence $t_n \to \infty$ and $x_n \in B$ such that $\varphi_{t_n}(x) \to y$. Since the flow is continuous,

$$\lim_{n\to\infty} \varphi_{t_n+t}(x_n) = \lim_{n\to\infty} \varphi_t\left(\varphi_{t_n}(x_n)\right) = \varphi_t\left(\lim_{n\to\infty} \varphi_{t_n}(x_n)\right) = \varphi_t(y).$$

Thus, $y \in B$ implies that $\varphi_t(y) \in \omega(B)$ for every $t \in \mathbb{R}$. □

Remark 35. Note that the proof of Lemma 8 requires only that the flow be defined only for $t \geq 0$. If only this hypothesis holds, the above argument shows that $\omega(B)$ is closed and positively invariant. In many instances, as in Theorem 43 this is enough to show that $\omega(B)$ is invariant.

Theorem 36. *Suppose that $\varphi_t : U \to U$ is defined for all $t \geq 0$. Assume $B \subseteq U$ is connected and that $\omega(B)$ is compact. Then, $\omega(B)$ is connected.*

Proof. For brevity let $A = \omega(B)$. The intuitive idea here is this. Suppose $\omega(B)$ had two disjoint parts, say A_1 and A_2. Then both these sets would need to be closed, thus compact, and we could separate these sets with two disjoint open sets as depicted in Figure 3.6.1. The image of a connected set under a continuous map is connected, thus $\varphi_t(B)$ is always connected. But then since both A_1 and A_2 are part of $\omega(B)$, we must have points that hop between O_1

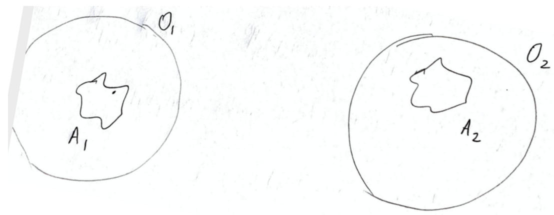

Figure 3.6.1: A_1, A_2 separated by open sets O_1, O_2

and O_2 and we may thus obtain a limit point outside A_1 and A_2. The existence of such a limit point contradicts the assumption that $\omega(B)$ is disconnected. Let us now make this precise by establishing the existence of such limit points under the assumption that $\omega(B)$ is disconnected.

First, recall that a set A is disconnected if and only if we can find open sets O_1 and O_2 such that

$$\begin{aligned} O_1 \cap O_2 &= \varnothing \\ A_1 := A \cap O_1 &\neq \varnothing \\ A_2 := A \cap O_2 &\neq \varnothing, \end{aligned}$$

and $A \subseteq O_1 \cup O_2$. This formalises the picture above, with $A = \omega(B)$. Since $A = \omega(B)$ is compact, we may choose O_1 and O_2 to be bounded.

Since A_1 and A_2 are part of the ω-limit set $\omega(B)$, for any sufficiently large T, there exist $s, t \geq T$ such that

$$\varphi_s(B) \cap O_1 \neq \varnothing \text{ and } \varphi_t(B) \cap O_2 \neq \varnothing.$$

Without loss of generality, suppose $t \geq s$ (relabel the sets otherwise). Since B is connected, so are the sets $\varphi_s(B)$ and $\varphi_t(B)$. Since O_1 and O_2 are disjoint, by continuity there must exist $\tau \in [s,t]$ and $y \in \varphi_\tau(B)$ such that $y \in \partial O_1$. (Intuitively, we're picking a time in between when points travel from O_1 to O_2.)

Now label the above values of s,t, as T s_1,t_1, T_1 respectively, and similarly define y_1 and τ_1. Now choose $T_2 > \max\{s_1, t_1\}$ and repeat the above argument. Proceeding inductively we obtain a sequence of times $\tau_n \to \infty$ and points $y_n \in \partial O_1$ such that $y_n = \varphi_{\tau_n}(x_n)$ for some $x_n \in B$. But then, $y_n \in \omega(B)$ by definition. This contradicts the assumption that $\omega(B) \subseteq O_1 \cup O_2$. $\square$

Chapter 4

Gradient Flows

In this chapter and the next, we will consider two fundamental examples of flows: gradient flows and Hamiltonian systems. We will work on $\mathbb{R}^n$ and R^{2n} respectively assuming conditions that guarantee global existence of solutions. Later, we will refine these ideas to gradient flows on Riemannian manifolds and Hamiltonian flows on symplectic manifolds.

4.1 The fundamental estimate for gradient flows

We assume given a C^2 function $V : \mathbb{R}^n \to \mathbb{R}$ such that the sublevel sets

$$K_a := \{x \in \mathbb{R}^n \mid V(x) \le a\} \tag{4.1.1}$$

are compact for all $a \in (-\infty, \infty)$. This function will be called the *potential*, the *energy*, or the *cost function* in different contexts.

Remark 37. Compactness of the sublevel sets always holds if $|V(x)| \to \infty$ as $|x| \to \infty$. This is sometimes called a coercivity condition.

An intuitive picture of gradient flow is depicted in Figure 4.1.1.

The gradient flow with potential V is defined by the equation

$$\dot{x} = -\nabla V(x), \quad x \in \mathbb{R}^n. \tag{4.1.2}$$

The vector field ∇V is given in coordinates by

$$\dot{x}_i = -\frac{\partial V}{\partial x_i}, \quad 1 \le i \le n. \tag{4.1.3}$$

The intuition of a gradient flow is that 'trajectories flow downhill'. This follows from the following

Theorem 38 (Fundamental estimate for gradient flows)**.** *Assume $V(x) \in C^2$ and its sublevel sets are compact. Then the flow*

$$\frac{\partial \varphi_t(x)}{\partial t} = -\nabla V(\varphi_t(x)), \qquad \varphi_0(x) = x, \tag{4.1.4}$$

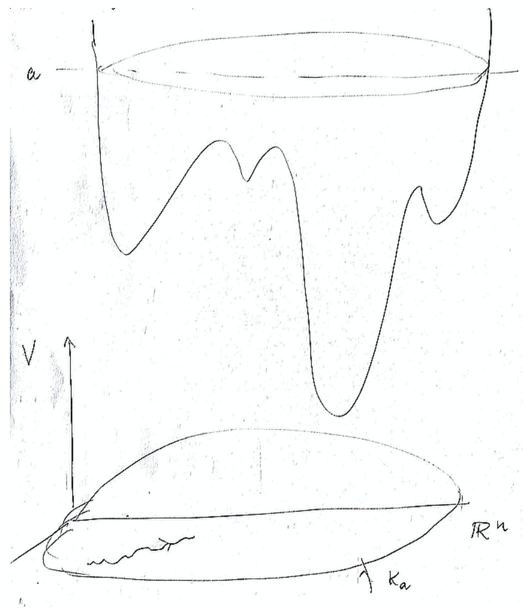

Figure 4.1.1: Gradient flow in $\mathbb{R}^n$

is defined for all $t \geq 0$ *and remains within the compact set* $K_{V(x_0)}$.

Proof. Since $V \in C^2$ the vector field ∇V is C^1 and by Picard's Theorem the solution is defined for a time interval $[0, T(x_0)]$ with $T(x_0) > 0$. We evaluate the potential along the trajectory, setting

$$v(t) := V(\varphi_t(x)). \tag{4.1.5}$$

Then by the chain rule and (4.1.4) we obtain

$$\begin{aligned} \dot{v} &= \nabla V \cdot \frac{\partial}{\partial t}\varphi_t(x) \\ &= -\nabla V \cdot \nabla V = -|\nabla V|^2. \end{aligned}$$

Thus,

$$v(t) = v(0) - \int_0^t |\nabla V(\varphi_s(x))|^2 ds, \tag{4.1.6}$$

at least for $t \in [0, T(x_0)]$. In particular, $v(t) \leq v(0)$, so that $\varphi_T(x_0) \in K_a$. But then we may again use Picard's theorem and extend the solution to $[T, 2T]$ since the time of existence guaranteed by Picard's theorem is uniform on a compact set. This ensures that the solution is defined for $t \in [0, \infty)$ and that $v(t) \leq v(0)$ on this interval with equality if and only if $\nabla V(x_0) = 0$.

□

4.2 Linearization of gradient flows

The equation of variations for an arbitrary vector field was discussed in Section 3.5; see in particular, equations (3.5.2) and (3.5.3). When the vector field $f(x) = -\nabla V(x)$ the equation of variations takes the form

$$\dot{B} = A(t)B, \quad B(0) = I, \tag{4.2.1}$$

where these matrices are related to the flow $\varphi_t(x)$ and potential $V(x)$ through

$$B(t) = D_x\varphi_t(x), \quad A(t) = D^2V(\varphi_t(x)). \tag{4.2.2}$$

Here the Hessian $D^2V(x)$ is the real symmetric matrix with components

$$(D^2V)_{ij} = \frac{\partial^2 V}{\partial x_i \partial x_j}.$$

Since $V \in C^2$ the partial derivatives commute, i.e.

$$\frac{\partial^2 V}{\partial x_i \partial x_j} = \frac{\partial^2 V}{\partial x_j \partial x_i}.$$

This means that $A(t)$ is a symmetric matrix [1]. In particular, the linearization at a critical point x_* is always of the form

$$\dot{u} = Au, \qquad \text{where } A = D^2V(x_*).$$

Since A is real and symmetric, its eigenvalues are real and it admits a diagonalization $A = Q\Lambda Q^T$ with an orthogonal matrix Q (i.e. $Q^T = Q^{-1}$) and

$$\Lambda = \begin{pmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_k & & & \\ & & & \lambda_{k+1} & & \\ & & & & \ddots & \\ & & & & & \lambda_n \end{pmatrix}.$$

Definition 39. (Nondegenerate Critical Point). We say that a critical point of a gradient flow is nondegenerate if 0 is not an eigenvalue of A.

Definition 40. (Morse function). A C^2 function $V : \mathbb{R}^n \to \mathbb{R}$ all of whose critical points are nondegenerate is called a Morse function.

[1] We'll denote this by $A \in \mathrm{Symm}(n)$.

Definition 41. (Morse index). The index, or Morse index, of a non-degenerate critical point is:

$$\#(\text{positive eigenvalues}) - \#(\text{negative eigenvalues}).$$

Example 6. *(Positive Morse index) The matrix*

$$\begin{pmatrix} +1 & 0 \\ 0 & +1 \end{pmatrix}$$

has a Morse index of $+2$*.*

Example 7. *(Zero Morse index) The matrix*

$$\begin{pmatrix} +1 & 0 \\ 0 & -1 \end{pmatrix}$$

has a Morse index of 0*.*

Example 8. *(Negative Morse index) The matrix*

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

has a Morse index of -2*.*

The phase portraits associated with the above three examples are shown in Figure 4.2.1.

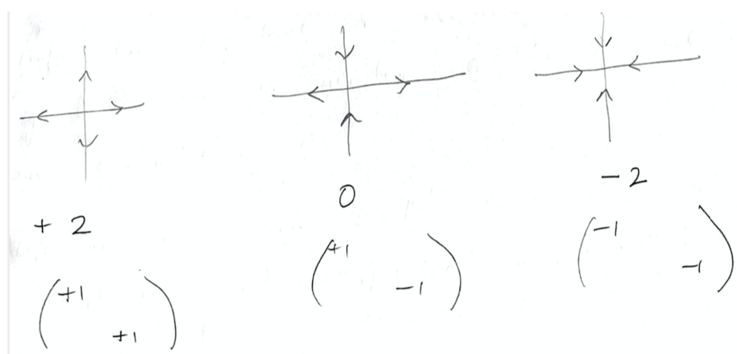

Figure 4.2.1: Phase portraits labelled with Morse indices

4.3 Asymptotic behavior

Theorem 42. *A gradient flow cannot contain a periodic orbit.*

Proof. Suppose $\gamma(t)$ is a periodic orbit with period T, that is: $\gamma(t+T) = \gamma(t)$ and $T > 0$ is $T = \inf_{s>0} \{\gamma(s) = \gamma(0)\}$. We evaluate $v(t) = V(\gamma(t))$ along the orbit. As before, $v(t) = v(0) - \int_0^t |\nabla V(\gamma(s))|^2 ds$. Thus,

$$v(0) = v(T) = v(0) - \int_0^T |\nabla V(\gamma(s))|^2 ds < V(0)$$

since $\nabla V(x) = 0$ if and only if x is a critical point. □

Let us now ask the more general question: what happens to $\varphi_t(x)$ as $t \to \infty$?

Theorem 43 (La Salle invariance principle)**.** *Assume $V : \mathbb{R}^n \to \mathbb{R}$ is C^2 and has compact sublevel sets. Then, for any $x \in \mathbb{R}^n$, $v_* = \lim_{t\uparrow\infty} V(\varphi_t(x))$ exists and $\omega(x) \subseteq \{y \,|\, V(y) = v_*, \nabla V(y) = 0\}$.*

Proof. If x is a critical point there is nothing to prove. Thus, assume x is not a critical point. The proof of Theorem 38 shows that $v(t) = V(\varphi_t(x))$ is *strictly* decreasing and that $\varphi_t(x)$ is contained within a compact set. Thus, $v(t)$ is strictly decreasing and bounded below so that $v_* = \lim_{t\to\infty} v(t)$ exists.

Let us first show that $\omega(x) \subset K_*$ where

$$K_* = \{y \in \mathbb{R}^n \,|\, V(y) = v_*\}.$$

Let $\{t_n\}_{n=1}^\infty$ be any sequence such that $t_n \to \infty$. Then, $\{x_n\} := \{\varphi_{t_n}(x)\}$ is a precompact sequence since it is contained in $K_{v(0)}$ which is compact. Thus, there exists a subsequence $x_{n_j} \to x_*$. But

$$\lim_{j\to\infty} V(x_{n_j}) = \lim_{j\to\infty} v(t_{n_j}) = v_*.$$

Thus, $\lim_{j\to\infty} x_{n_j}$ must lie within K_* as asserted. (Every subsequence has the same limit for a decreasing sequence).

Let us next show that $\nabla V(x_*) = 0$ if $x_* \in \omega(x)$. To this end, we first note that $\omega(x)$ is a closed subset of a compact set, thus it is compact. Theorem 34 and Remark 35 shows that $\omega(x)$ is compact and positively invariant. Thus, $\varphi_t(x_*) \in \omega(x)$ for every $t > 0$ and by the first part of the proof

$$V(\varphi_t(x_*)) = V(x_*), \quad t \geq 0.$$

On the other hand, by Theorem 38,

$$V(x_*) - V(\varphi_t(x_*)) = \int_0^t |\nabla V(\varphi_s(x_*))|^2 \, ds.$$

The left hand side vanishes, which means that

$$\int_0^t |\nabla V(\varphi_s(x_*))|^2 \, ds = 0, \quad t \geq 0,$$

which shows that $\nabla V(x_*) = 0$. □

Remark 44. A sharper conclusion holds for Morse functions. If V is Morse with compact sublevel sets then $\omega(x)$ always consists of a single critical point. You are asked to prove this statement in the second homework. When V is not Morse, it may have flat regions as shown in Figure 4.3.1. Such degenrate functions are common in gradient flows in optimization.

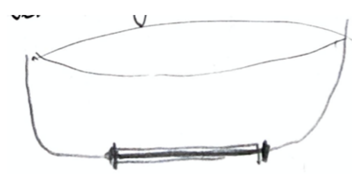

Figure 4.3.1: K_* in the flat part of the potential

4.4 Exercises

1. Suppose $\dot{x} = f(x)$, $x \in \mathbb{R}^n$, $f \in C^1$, $x(0) = x_0$. Do not assume that $\sup_{x \in \mathbb{R}^n} \|Df(x)\| < \infty$. Let $I(x_0)$ denote the maximal open interval that includes 0 on which the solution $x : I(x_0) \to \mathbb{R}^n$ is defined. If $I(x_0) = (-\infty, \beta)$ with $\beta < \infty$, is it necessary that $\lim_{t \to \beta} |x(t)| = +\infty$? Prove or disprove.

2. Show that Definition 31 for $\omega(B)$ is equivalent to

$$\omega(B) = \cap_{t \geq 0} \overline{\cup_{s \geq t} \varphi_s(B)}.$$

Here $\varphi_t(B)$ is the image of the set B under the flow φ_t and $\overline{A}$ denotes the closure of a set A.

3. Suppose $f : \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 vector field all of whose critical points are non-degenerate. Show that:

(a) Each critical point is isolated.

(b) The number of critical points is countable.

(c) The set of critical points cannot have an accumulation point within any bounded set in $\mathbb{R}^n$.

4. Suppose $V : \mathbb{R}^n \to \mathbb{R}$ is a Morse function with compact sublevel sets. Consider the gradient flow

$$\dot{x} = -\nabla V(x).$$

Show that $\omega(x)$ for any $x \in \mathbb{R}^n$ must be a single critical point.

5. We will discuss periodic orbits and circle maps when we study Hamiltonian systems. This question involves an elementary flow that will help build intuition for flows on the circle.

Consider the vector field on the circle $\dot{\theta} = \omega - \sin\theta$ where ω is a parameter. Show that the flow is periodic when $\omega > 1$. Let $T(\omega)$ denote the period of the orbit. Show that:

(a) $T(\omega)$ is well-defined. That is, the period does not depend on the initial condition.

(b) Compute the limit $\lim_{\omega\to 1} T(\omega)\sqrt{\omega - 1}$.

(Part (b) is a tricky integral. Use the residue theorem if you know it, feel free to use a computer package if you don't.)

6. *Lyapunov functions* Assume given a global flow on $\mathbb{R}^n$ defined by $\dot{x} = f(x)$. A function $V : \mathbb{R}^n \to \mathbb{R}$ is a Lyapunov function for the flow if V satisfies the inequality

$$\nabla V \cdot f(x) \le 0, \quad x \in \mathbb{R}^n.$$

(a) Construct a linear system $\dot{x} = Ax$ that is *not* a gradient flow, but which has a Lyapunov function.

(b) Assume that V is a Lyapunov function with compact sublevel sets. Show that La Salle's invariance principle holds for flows with a Lyapunov function in the following form: if $\omega(x)$ is non-empty and compact then

$$\omega(x) \subset \{y \in \mathbb{R}^n \,|\nabla V \cdot f(y) = 0\}.$$

4.5 Solutions to exercises

1. Suppose $\dot{x} = f(x)$, $x \in \mathbb{R}^n$, $f \in C^1$, $x(0) = x_0$. Do not assume that $\sup_{x\in\mathbb{R}^n} \|Df(x)\| < \infty$. Let $I(x_0)$ denote the maximal open interval that includes 0 on which the solution $x : I(x_0) \to \mathbb{R}^n$ is defined. If $I(x_0) = (-\infty, \beta)$ with $\beta < \infty$, is it necessary that $\lim_{t\to\beta} |x(t)| = +\infty$? Prove or disprove.

Proof. It is necessary that $\lim_{t\to\beta} |x(t)| = +\infty$. If not, there exists a subsequence $\{t_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} t_n = \beta$ and $\lim_{n\to\infty} x(t_n) = a$ where $a \in \mathbb{R}^n$. Since $f \in C^1$, there is $\varepsilon > 0$ and a time $T(a, \varepsilon) > 0$ such that there is a well-defined solution $\varphi_s(y)$ for all initial conditions y in the ball $B(a, \varepsilon)$ for all $s \in (-T, T)$. [2]

Choose N so that $\beta - t_n < T$ and $x(t_n) \in B(a, \varepsilon)$ for $n \ge N$. By the existence and uniqueness of solutions with initial conditions in $B(a, \varepsilon)$, it follows that the solution $\varphi_s(x(t_n))$ is defined for $s \in (-T, T)$. Using uniqueness again, this solution must agree with the solution $x(t)$ on the time interval $t \in [t_n, \beta)$. But then we see that the interval of existence for $x(t)$ may be extended to $t_n + T > \beta$, contradicting the definition of β. $\square$

[2] $B(a, \varepsilon)$ denotes the ball centered at a with radius ε.

2. Show that the above definition of $\omega(B)$ is equivalent to

$$\omega(B) = \cap_{t\geq 0}\overline{\cup_{s\geq t}\varphi_s(B)}.$$

Here $\varphi_t(B)$ is the image of the set B under the flow φ_t and $\overline{A}$ denotes the closure of a set A.

Proof. We use the following notation. Let

$$A_t = \overline{\cup_{s\geq t}\varphi_s(B)}, \quad A_\infty = \cap_{t\geq 0} A_t.$$

We must show that $\omega(B) = A_\infty$. This means that we must establish the inclusions

$$A_\infty \subset \omega(B) \quad \text{and} \quad \omega(B) \subset A_\infty.$$

1. Suppose $y \in A_\infty$. Consider the sequence of integers $n = 1, 2, \ldots$ and choose a sequence ε_n such that $\varepsilon_n \to 0$. Since $y \in A_n$ for every n, there is a $t_n \geq n$ and x_n such that $|y - \varphi_{t_n}(x_n)| < \varepsilon$. In particular,

$$\lim_{n\to\infty} \varphi_{t_n}(x_n) = y,$$

showing that $y \in \omega(B)$.

2. Now suppose $y \in \omega(B)$. By the definition of $\omega(B)$, for every sequence ε_n such that $\varepsilon_n \to 0$, there exists a sequence $t_n \to \infty$ and $x_n \in B$ such that $|\varphi_{t_n}(x_n) - y| < \varepsilon_n$. The points $\varphi_{t_n}(x_n)$ lie in A_{t_n}. Therefore, the distance

$$\text{dist}(y, A_{t_n}) < \varepsilon_n,$$

where the distance between a point y and a closed set K is defined by

$$\text{dist}(y, K) = \inf_{x\in A} |y - x|.$$

It follows that

$$\text{dist}(y, A_\infty) < \varepsilon_n,$$

for every n, so that $\text{dist}(y, A_\infty) = 0$. Since A_∞ is a closed set, $y \in A_\infty$. □

3. Suppose $f : \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 vector field all of whose critical points are non-degenerate. Show that:

(a) Each critical point is isolated.

(b) The number of critical points is countable.

(c) The set of critical points cannot have an accumulation point within any bounded set in $\mathbb{R}^n$.

Proof. (a) Assume x_* is a non-degenerate critical point. By definition this means that $f(x_*) = 0$ and $Df(x_*)$ is invertible. By the inverse function theorem, there exists $\varepsilon_* > 0$ such that f is a diffeomorphism of $B(0, \varepsilon)$ onto its image for all $\varepsilon < \varepsilon_*$. Since $f(x_*) = 0$ this ensures that for $\varepsilon < \varepsilon_*$ the image $f(B(0, \varepsilon))$ is an open neighborhood of 0 and that f takes the value 0 only once in $B(0, \varepsilon)$.

(b) A set S of isolated points in $\mathbb{R}^n$ is always countable. Here is one proof of this statement.

Each point x in S can be contained within a ball $B(x, \varepsilon(x))$ such that no other points of S lie within $B(x, \varepsilon(x))$. Since the rational points $\mathbb{Q}^n$ are dense in $\mathbb{R}^n$ we may choose a unique point $q(x) \in \mathbb{Q}^n$ as the label for $x \in S$. This gives a one-to-one map from $S \to \mathbb{Q}^n$ which is countable. Thus, S can be labeled by a countable subset of a countable set, which makes it countable.

(c) Suppose there exists a sequence of critical points $\{x_n\}_{n=1}^{\infty}$ with a limit $x = \lim_{n\to\infty} x_n$. By the continuity of f, $f(x) = \lim_{n\to\infty f(x_n)} = 0$. Thus, x is a critical point. But then x cannot be non-degenerate, since this would contradict part (a). So x is degenrate, which contradicts our assumption that all critical points of f are non-degenerate. □

4. Suppose $V : \mathbb{R}^n \to \mathbb{R}$ is a Morse function with compact sublevel sets. Consider the gradient flow

$$\dot{x} = -\nabla V(x).$$

Show that $\omega(x)$ for any $x \in \mathbb{R}^n$ must be a single critical point.

Proof. Since V is Morse, by problem (3), its critical points are isolated. On the other hand, we know that $\omega(x)$ is connected. A connected subset of a set of isolated points must be a single point. □

5. We will discuss periodic orbits and circle maps when we study Hamiltonian systems. This question involves an elementary flow that will help build intuition for flows on the circle.

Consider the vector field on the circle $\dot{\theta} = \omega - \sin\theta$ where ω is a parameter. Show that the flow is periodic when $\omega > 1$. Let $T(\omega)$ denote the period of the orbit. Show that:

(a) $T(\omega)$ is well-defined. That is, the period does not depend on the initial condition.

(b) Compute the limit $\lim_{\omega\to 1} T(\omega)\sqrt{\omega - 1}$.

(Part (b) is a tricky integral. Use the residue theorem if you know it, feel free to use a computer package if you don't.)

Proof. (a) We identify the circle with $\mathbb{R} \mod 2\pi$. We separate variables and integrate from θ_0 to $\theta_0 + 2\pi$ to find the time taken for an orbit starting at θ_0 to loop around once:

$$T := \int_{\theta_0}^{\theta_0+2\pi} \frac{d\theta}{\omega - \sin\theta} = \int_0^{2\pi} \frac{d\theta}{\omega - \sin\theta},$$

where the second equality follows from the periodicity of $\sin\theta$. Thus, the time period is independent of θ_0.

(b) The integral may be computed using Cauchy's integral formula (also known as the residue calculus) or the substitution $u = \tan\theta/2$. We make the substitution $z = e^{i\theta}$ to convert the integral over the interval $[0, 2\pi]$ into a contour integral. Then

$$\frac{dz}{iz} = d\theta, \quad \sin\theta = \frac{1}{2i}(e^{i\theta} - e^{-i\theta}) = \frac{1}{2iz}(z^2 - 1),$$

and we may rewrite

$$\int_0^{2\pi} \frac{d\theta}{\omega - \sin\theta} = 2\oint_{|z|=1} \frac{dz}{-z^2 + 2i\omega z + 1}.$$

The denominator of the integrand may be factorized by the quadratic formula. We write

$$-z^2 + 2i\omega z + 1 = -(z - \omega_-)(z - \omega_+), \quad \omega_\pm = i\left(\omega \pm \sqrt{\omega^2 - 1}\right).$$

Of these roots, only ω_- lies within the unit disk. Thus, by Cauchy's integral formula

$$2\oint_{|z|=1} \frac{dz}{-z^2 + 2i\omega z + 1} = -2\oint_{|z|=1} \frac{dz}{(z - \omega_-)(z - \omega_+)} = -\frac{4\pi i}{\omega_-\omega_+} = \frac{2\pi}{\sqrt{\omega^2 - 1}}.$$

The time period diverges as $\omega \to 1$. The rate of divergence is computed as follows

$$\lim_{\omega \downarrow 1} \sqrt{\omega - 1}\, T(\omega) = \lim_{\omega \downarrow 1} \sqrt{\omega - 1} \frac{2\pi}{\sqrt{\omega^2 - 1}} = \sqrt{2}\,\pi.$$

$\square$

6. *Lyapunov functions* Assume given a global flow on $\mathbb{R}^n$ defined by $\dot{x} = f(x)$. A function $V : \mathbb{R}^n \to \mathbb{R}$ is a Lyapunov function for the flow if V satisfies the inequality

$$\nabla V \cdot f(x) \leq 0, \quad x \in \mathbb{R}^n.$$

(a) Construct a linear system $\dot{x} = Ax$ that is *not* a gradient flow, but which has a Lyapunov function.

(b) Show that La Salle's invariance principle holds for flows with a Lyapunov function in the following form: if $\omega(x)$ is non-empty and compact then

$$\omega(x) \subset \{y \in \mathbb{R}^n \,|\nabla V \cdot f(y) = 0\}.$$

(This is very similar to the proof done in class).

Proof. (a) If $Ax = -\nabla V(x)$, then $V(x)$ must be quadratic and A must be symmetric. Thus, to find a flow that is not a gradient flow, it is sufficient to consider a non-symmetric matrix. We choose

$$A_0 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$$

and we set
$$A = A_0 + A_1, \quad V(x) = \frac{1}{2}x^T Ax = \frac{1}{2}x^T A_1 x.$$
The intuition here is that the first linear transformation A_0 gives rise to a rotation (which is definitely not a gradient flow), whereas the second matrix A_1 gives rise to a decay. These effects are orthogonal in the sense that
$$-\nabla V(x) \cdot Ax = x^T A_1^T Ax = -|A_1 x|^2 \le 0,$$
with strict inequality unless $x = 0$.

(b) We consider the value $v(t) := V(x(t))$ of the Lyapunov function along the solution $x(t)$. Then $v(t)$ is a decreasing function of time. If $y \in \omega(x)$ then there is a sequence $t_n \to \infty$ such that $x(t_n) \to y$ and we find that $\lim_{n\to\infty} V(x(t_n)) = V(y)$. But since $v(t)$ is a decreasing function it is also true that
$$\lim_{t\to\infty} V(x(t)) = V(y) := v_*.$$
Thus, $\omega(x) \subset \{y \,|V(y) = v_*\}$. Finally, since $\omega(x)$ is assumed compact, it is invariant. Use $y \in \omega(x)$ as the initial condition for $\dot{x} = f(x)$ to see that $\nabla V \cdot f(y) = 0$ (if not, V would decrease strictly on the solution beginning at y, contradicting the fact that V is constant on $\omega(x)$.). □

Chapter 5

Hamiltonian Systems

This chapter provides an introduction to Hamiltonian systems. We begin with examples in one dimension. We then turn to the general structure of Hamiltonian systems. The main references for this chapter are [3, Ch.2] and [12, Ch.1].

5.1 One dimensional Hamiltonian systems

5.1.1 A solution formula

Consider a particle on the line with unit mass subject to the effect of a smooth potential $V : \mathbb{R} \to \mathbb{R}$. The equation of motion is given by Newton's law

$$\ddot{x} = -V'(x). \tag{5.1.1}$$

Equation (5.1.1) is a second order equation for one variable and it may be rewritten as the system

$$\dot{x} = y \tag{5.1.2}$$

$$\dot{y} = -V'(x) \tag{5.1.3}$$

One of Newton's fundamental observations is the principle of conservation of energy. Define the *Hamiltonian*

$$H(x, y) = \frac{1}{2}y^2 + V(x), \tag{5.1.4}$$

consider a solution to (5.1.2) and observe that

$$\begin{aligned}\frac{d}{dt}H(x(t), y(t)) &= \frac{\partial H}{\partial x}\dot{x} + \frac{\partial H}{\partial y}\dot{y} \\ &= V'(x)\dot{x} + y\dot{y} \\ &= y(V'(x) - V'(x)) = 0.\end{aligned}$$

The Hamiltonian is the sum of the kinetic energy and potential energy in the system. Conservation of energy allows us to solve (5.1.1) almost explicitly. Suppose that at $t = 0$, $H(x_0, y_0) = E$ is known. Then

$$\frac{1}{2}\dot{x}^2 + V(x) = E \tag{5.1.5}$$

for the interval of existence of the solution. We solve for the velocity to obtain

$$\dot{x} = \pm\sqrt{2(E - V(x))}.$$

We further separate variables and integrate to obtain the solution in an implicit form

$$\int_{x_0}^{x(t)} \frac{ds}{\sqrt{2(E - V(s)}} = t$$

Of course, we'd like to actually express x as a function of t, not $t = t(x)$. Nevertheless, this formula already tells us a great deal about the phase portrait.

5.1.2 Examples

Here are some examples of physical systems that may be solved by the above method.

1. The simple harmonic oscillator

$$V(x) = \frac{1}{2}x^2. \tag{5.1.6}$$

   The equation of motion is $\ddot{x} = -x$, which is exactly solvable.

2. A qualitatively similar model which is *not* exactly solvable is

$$V(x) = \frac{1}{2}x^2 + \frac{1}{4}x^4. \tag{5.1.7}$$

3. The simple pendulum has potential

$$V(x) = 1 - \cos x. \tag{5.1.8}$$

   Figure 5.1.1 illustrates the physical context. The equation of motion

$$ml\ddot{\theta} = -mg\sin\theta \tag{5.1.9}$$

   is obtained by balancing forces. The left hand side is mass times acceleration. This equation may be rewritten

$$\ddot{\theta} = -\omega^2 \sin\theta \tag{5.1.10}$$

   where $\omega^2 = g/l$. If we choose units of time so that $\omega = 1$ and relabel the angle θ by x for consistency with our previous notation, we obtain the equation $\ddot{x} = -\sin x$, as in (5.1.1).

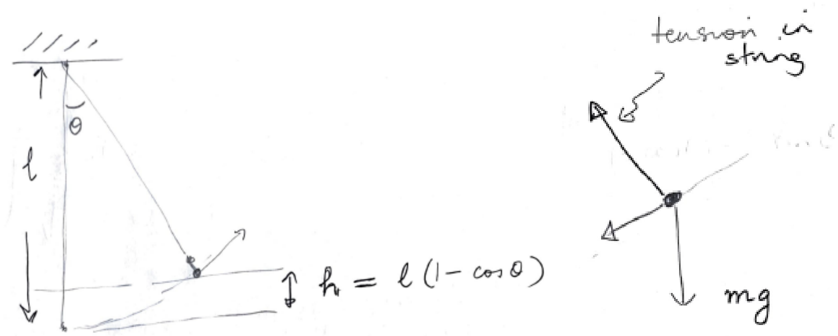

Figure 5.1.1: $V(\theta) = mgl(1-\cos\theta)$, where m, g, l are physical constants.

5.1.3 Phase portraits

The geometric method for plotting the phase portrait of 1-D Hamiltonian systems is as follows.

1. Sketch the graph of $V(x)$.
2. Use the formula $y = \sqrt{2(E - V(t))}$ to determine trajectories for different energy levels.

Examples of such phase portraits are shown below.

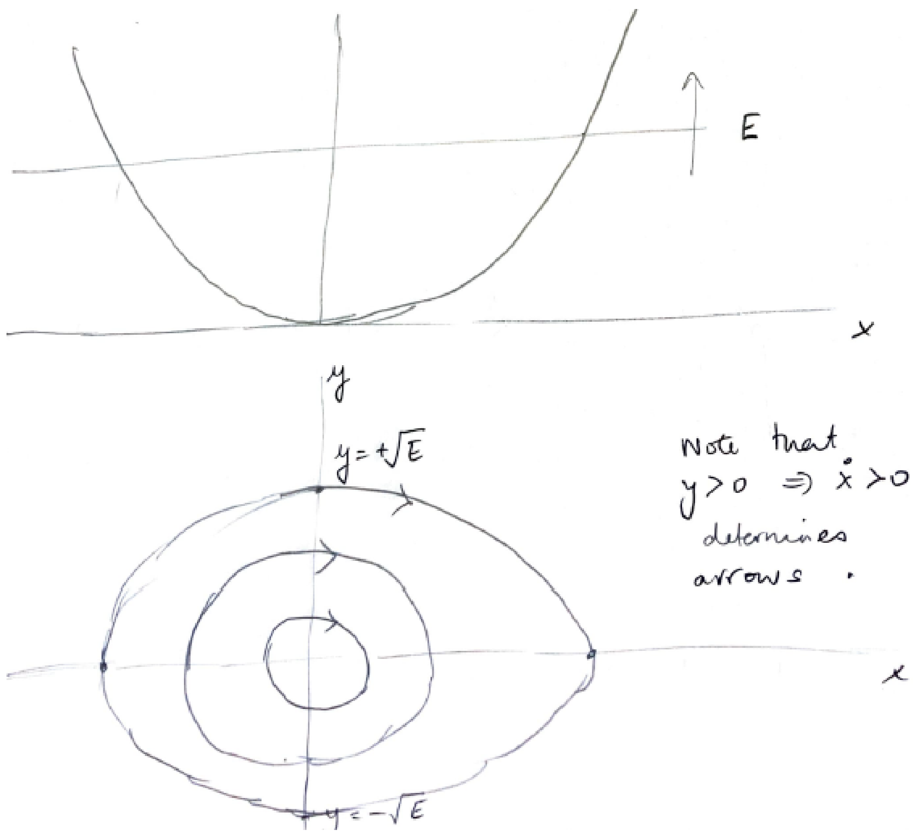

Figure 5.1.2: $V(x) = \frac{1}{2}x^2$.

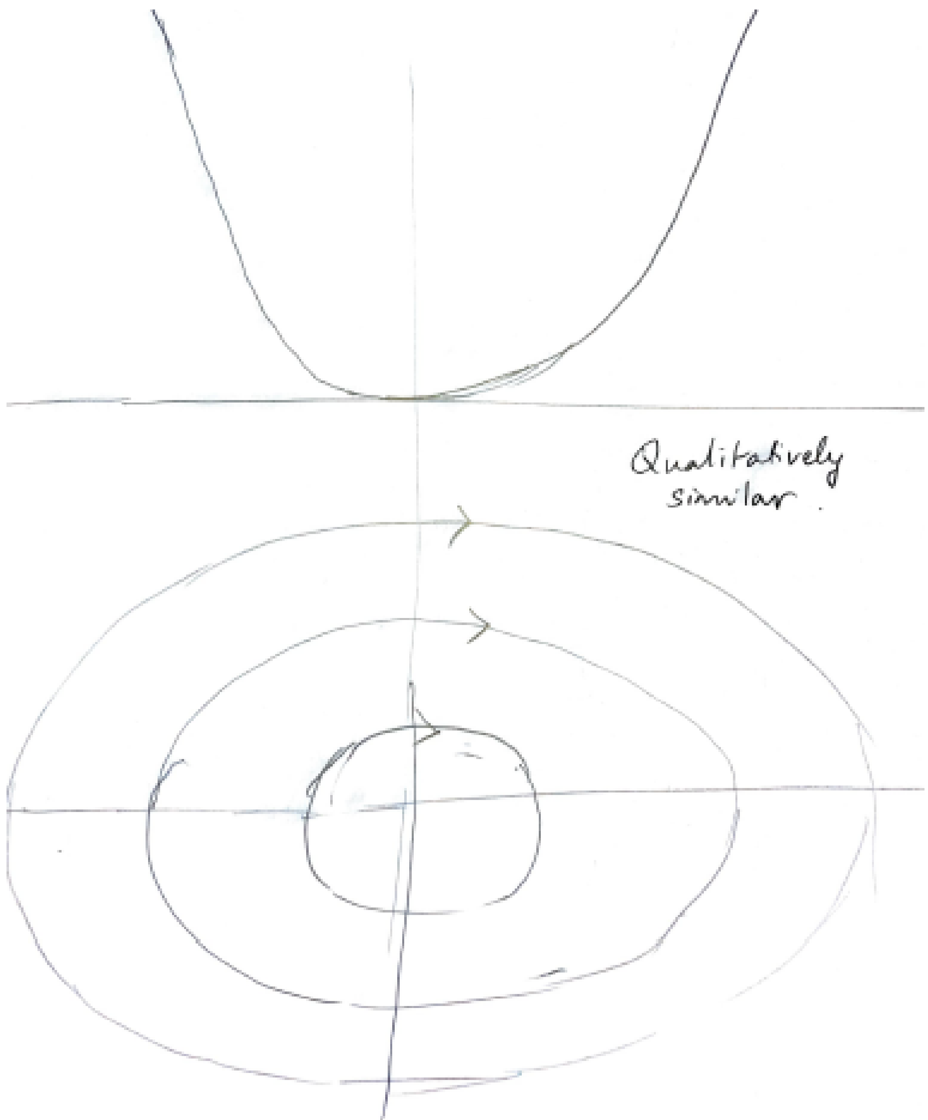

Figure 5.1.3: $V(x) = \frac{1}{2}x^2 + \frac{1}{4}x^4$. Note the qualitative similarity with Figure 5.1.2.

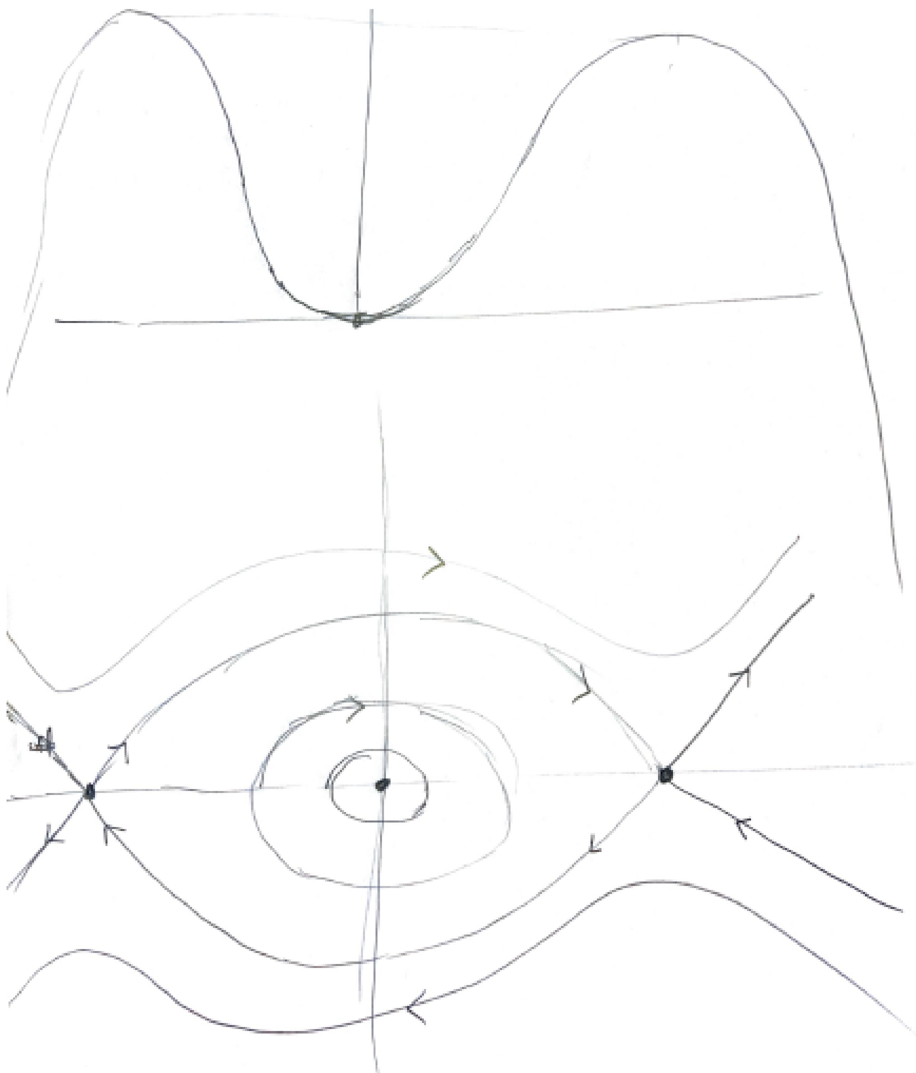

Figure 5.1.4: $V(x) = \frac{1}{2}x^2 - \frac{1}{4}x^4$. Compare the effect of the minus sign with Figure 5.1.3.

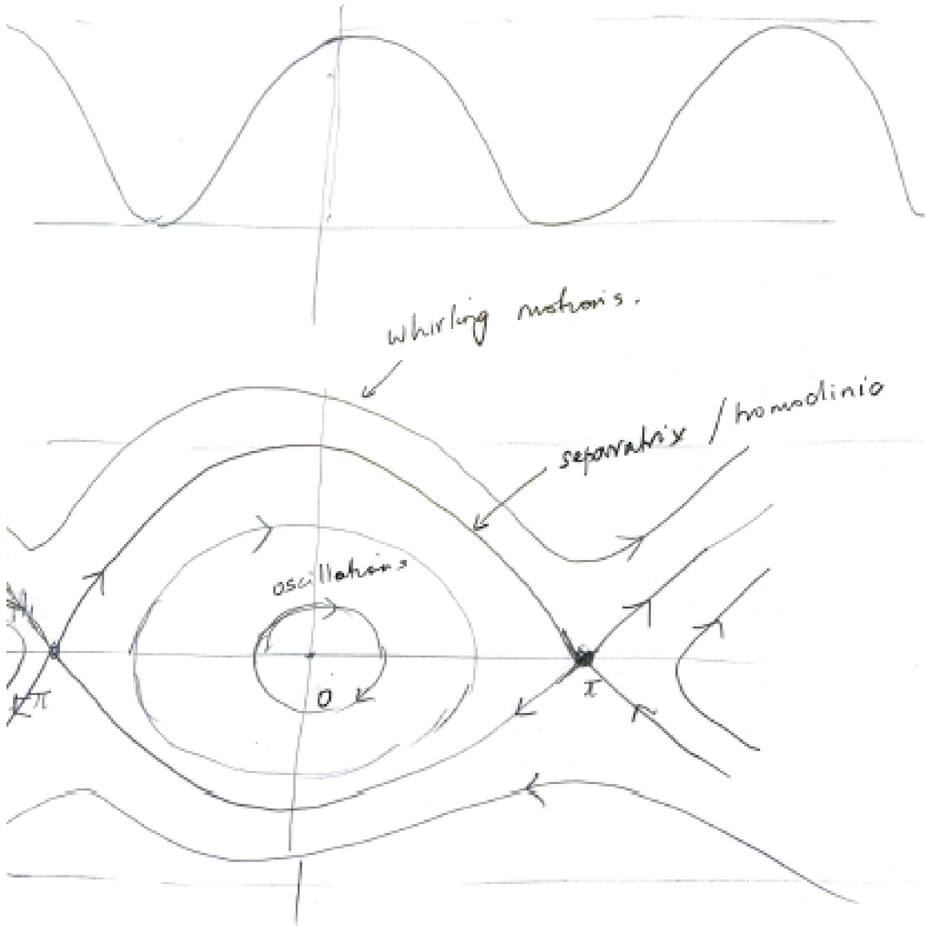

Figure 5.1.5: The simple pendulum. $V(x) = 1 - \cos x$.

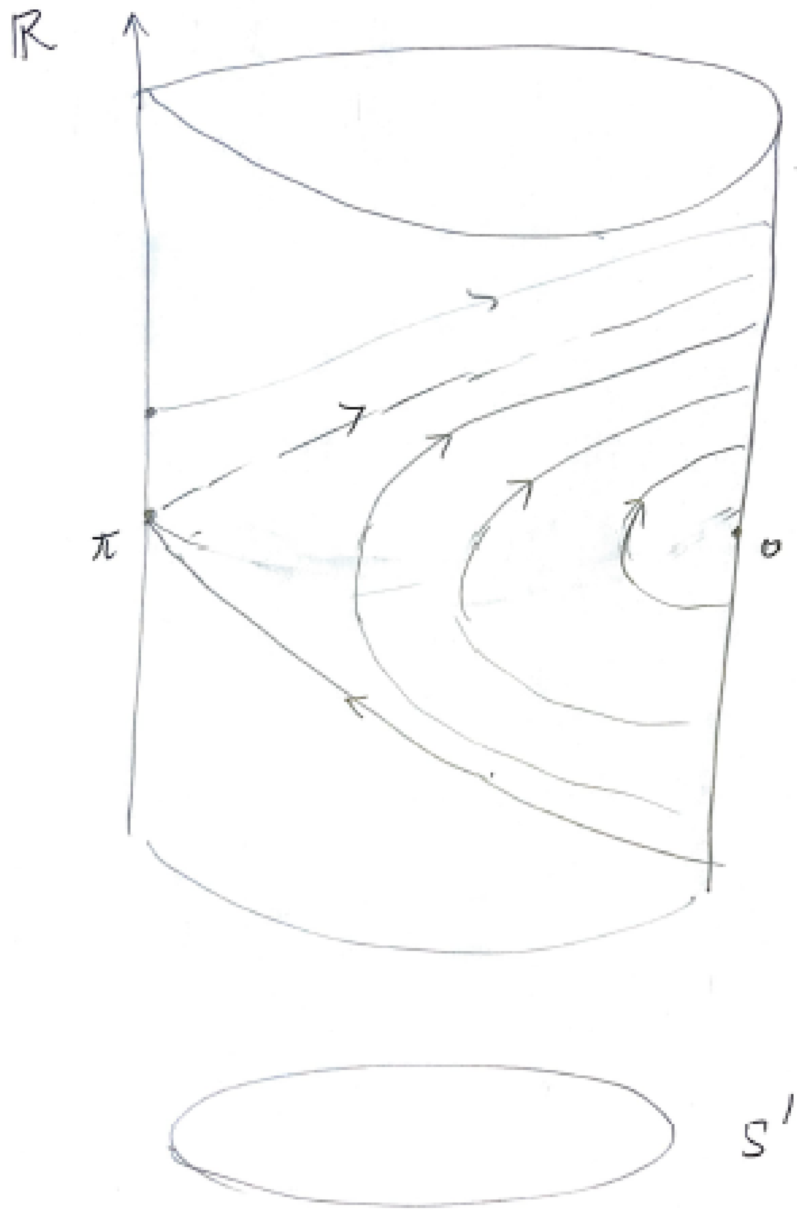

Figure 5.1.6: Phase portrait of the simple pendulum on $S^1 \times \mathbb{R}$.

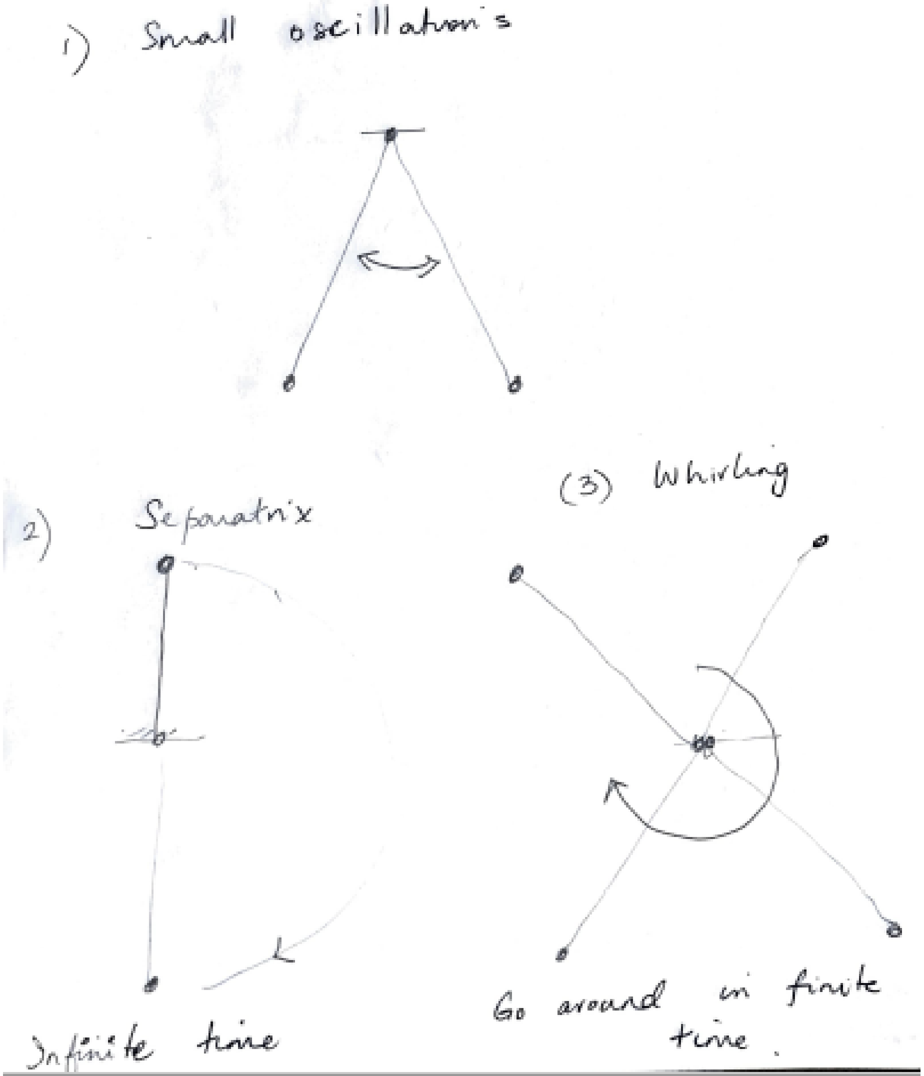

Figure 5.1.7: Modes of oscillation of a simple pendulum. The separatrix corresponds to a critical orbit that takes infinite time to turn once through an angle of 2π. In case 1 and case 3, a periodic cycle takes finite time.

5.2 The symplectic form

We now turn to the general theory of Hamiltonian systems. The state space S will be a subset of $\mathbb{R}^{2n}$ and we denote points in $\mathbb{R}^{2n}$ by $z = (x, y)$, $x, y \in \mathbb{R}^n$. We assume given a C^2 Hamiltonian $H : U \to \mathbb{R}$. Let I_m denotes the $m \times m$ identity.

Definition 45. The (standard) symplectic matrix J is the $2n \times 2n$ matrix

$$J = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix}. \tag{5.2.1}$$

We will also use the term symplectic form to refer to this matrix, since J defines a quadratic form on $\mathbb{R}^{2n}$ defined by

$$\omega(z_1, z_2) = z_1^T J z_2, \quad z_1, z_2 \in \mathbb{R}^{2n}. \tag{5.2.2}$$

The symplectic form is skew-symmetric, $\omega(z_1, z_2) = -\omega(z_2, z_1)$.

Lemma 9. $J^2 = -I_{2n}$

Proof.

$$J^2 = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix} \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix} = \begin{pmatrix} -I_n & 0 \\ 0 & -I_n \end{pmatrix} = -I_{2n} \tag{5.2.3}$$

□

The state space U in combination with the symplectic matrix J is an example of a *symplectic manifold*. We write such manifolds in the form (U, J) when it is necessary to make the symplectic matrix explicit. The Hamiltonian flow associated to H on the symplectic manifold (U, J) is

$$\dot{z} = J \nabla_z H. \tag{5.2.4}$$

Equation (5.2.4) is equivalent to

$$\dot{x} = \nabla_y H \tag{5.2.5}$$

$$\dot{y} = -\nabla_x H. \tag{5.2.6}$$

We use the following notation for derivatives in these equations:

$$\begin{aligned} \nabla_z H &= \Big(\frac{\partial H}{\partial z_1}, \ldots, \frac{\partial H}{\partial z_{2n}}\Big) \\ &= \Big(\frac{\partial H}{\partial x_1}, ..., \frac{\partial H}{\partial x_n}, \frac{\partial H}{\partial y_1}, ..., \frac{\partial H}{\partial y_n}\Big) \\ &= (\nabla_x H, \nabla_y H). \end{aligned}$$

Hamiltonian systems have a structure that is complementary to gradient flows. In both cases, it is first necessary to understand the underlying structure in the state spaces $\mathbb{R}^n$ and $\mathbb{R}^{2n}$ equipped with the standard metric and standard symplectic form respectively. Once the flows have been understood in this setting, the full power of the theory can be realized by studying these flows in their natural geometric setting. A brief comparison of these ideas is presented in Table 5.1.

Gradient flows	**Hamiltonian flows**
Euclidean	Euclidean
$V : \mathbb{R}^n \to \mathbb{R}$ $\dot{x} = -\nabla V(x)$	$H : \mathbb{R}^{2n} \to \mathbb{R}$ $\dot{z} = -J\nabla_z H$
Riemannian manifold	Symplectic manifold
$V : (\mathcal{M}^n, g) \to \mathbb{R}$ $\dot{x} = -\mathrm{grad}_g V(x)$	$H : (\mathcal{M}^{2n}, \omega) \to \mathbb{R}$ $\omega(\dot{z}, v) = dH(v)$, $v \in T_z(\mathcal{M})$.

Table 5.1: A comparison of gradient and Hamiltonian flows. The gradient operator is defined using the Riemannian metric g to convert the 1-form dV into a vector. Similarly, a Hamiltonian vector field is obtained by using the symplectic form ω to convert the 1-form dH into a vector.

5.3 Symplectic diffeomorphisms

Definition 46. The symplectic group $Sp(n)$ is the set of real matrices $S \in \mathbb{M}_n$ that satisfy

$$S^T J S = J. \tag{5.3.1}$$

The group operation is matrix multiplication.

The above definition should be contrasted with the more familiar example of the orthogonal group.

Definition 47. The orthogonal group $O(n)$ is the set of real matrices $Q \in \mathbb{M}_n$ that satisfy

$$Q^T Q = I. \tag{5.3.2}$$

The group operation is matrix multiplication.

Both $O(n)$ and $Sp(n)$ are examples of the classical groups [16]. The underlying idea in the definition of the classical groups is the classification of the linear transformations of $\mathbb{R}^m$ that preserve a natural quadratic form. These forms are the Euclidean inner product (for $O(n)$, $m = n$) and the symplectic form (for $Sp(n)$, $m = 2n$).

Let us check the group axioms for $Sp(n)$. First, it is clear that $I \in Sp(n)$. Second, if $S \in Sp(n)$, we note that $\det(S)$ is either plus or minus one, since $\det(J) = 1$, so that equation (5.3.1) implies $\det(S)^2 = 1$. Therefore, S^{-1} exists. Now multiply equation (5.3.1) on the right and left by S^{-T} and S^{-1} to obtain

$$J = S^{-T} J S^{-1}, \quad S^{-T} := (S^{-1})^T.$$

Similarly, if S_1 and S_2 satisfy (5.3.1) so does the product $S_1 S_2$ since

$$(S_1 S_2)^T J S_1 S_2 = S_2^T S_1^T J S_1 S_2 = S_2^T J S_2 = J.$$

Definition 48. Assume $U \subset \mathbb{R}^{2n}$ is an open set. A diffeomorphism $\varphi : U \to U$ is symplectic if $D\varphi(z) \in Sp(n)$ for each $z \in U$.

It is helpful to contrast this definition with the notion of isometries of $\mathbb{R}^n$. A diffeomorphism $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ is an isometry if $D\varphi(x) \in O(n)$ for every $x \in \mathbb{R}^n$. This definition and terminology reflects the fact that an isometry preserves lengths. It turns out that any C^1 isometry of $\mathbb{R}^n$ must be an affine transformation of the form $\varphi(x) = Qx + c$, $Q \in O(n)$, $c \in \mathbb{R}^n$. We often say for this reason that 'isometries are rigid', which means that there isn't a great deal of choice in isometries [1]. By contrast, there are many symplectic diffeomorphisms.

Theorem 49. *The flow map φ_t defined by the Hamiltonian system (5.2.4) is a symplectic diffeomorphism for all t in the interval of existence. Conversely, every one parameter family of symplectic diffeomorphisms φ_t with $\varphi_0(z) = z$ is generated by a Hamiltonian vector field.*

Proof. 1. Fix $z \in U$ and write the equation of variations for the Hamiltonian flow (5.2.4) around the trajectory $\varphi_t(z)$ as

$$\dot{B} = JSB, \quad B(t) := D\varphi_t(z), \quad S := D^2H(\varphi_t(z)), \quad B(0) = I. \tag{5.3.3}$$

We must show that $B(t) \in Sp(n)$ for all t in the interval of existence. By the product rule

$$\frac{d}{dt}(B^TJB) = \dot{B}^TJB + B^TJ\dot{B}.$$

We then substitute (5.3.3) to find

$$\frac{d}{dt}(B^TJB) = B^T\left(S^TJ^TJ + J^2S\right)B = 0,$$

because

$$S = D^2H(\varphi_t(z)) = S^T, \quad J^2 = -I_{2n}, \quad JJ^T = -J^2.$$

Since $B^T(0)JB(0) = J$ it follows that $B^T(t)JB(t) = J$ for all t in the interval of existence.

2. Conversely, let us suppose that $\varphi_t(z)$ is a symplectic diffeomorphism. Consider the vector field

$$v(z) = J^T\frac{d}{dt}\varphi_t(z)\bigg|_{t=0}.$$

The reader should now show, using the definition of $Sp(n)$, that this implies $v(z) = \nabla_z H(z)$ for some function $H : U \to \mathbb{R}$. (Hint: Use the classical calculus criterion to determine when a function is a gradient). □

[1] This requires a proof, but you can gain an intuitive feel for such rigidity by trying to construct a diffeomorphism of $\mathbb{R}^2$ that is a smoothing of a piecewise linear map whose derivative takes two distinct values Q_1 and Q_2 in the left and right half planes respectively. These derivatives must agree on the y-axis.

Corollary 3 (Liouville's theorem)**.** *Hamiltonian flows on $U \subset \mathbb{R}^{2n}$ preserve $2n$-dimensional volume.*

Proof. Theorem 49 shows that the Hamiltonian flow φ_t is a symplectic diffeomorphism. Thus, $\det(D\varphi_t(z)) = \det(D\varphi_0(z)) = 1$. □

5.4 Linearization at critical points

This section illustrates the special nature of critical points in Hamiltonian systems in two-dimensional flows. The general structure is considered in the homework. We know that the linearization at a critical point z_* for the differential equation $\dot{z} = f(z)$ is $\dot{u} = Df(z_*)u$. (We use z instead of x because we are going to apply this idea to Hamiltonian systems.)

Now, suppose $f(z) = J\nabla_z H$. In coordinates,

$$f_i(z) = J_{ik}\frac{\partial H}{\partial z_k}, \quad 1 \le i \le 2n,$$

where we sum over repeated indices. Then

$$(Df(z))_{ij} = \frac{\partial}{\partial z_j} f_i = J_{ik}\frac{\partial^2 H}{\partial z_j \partial z_k}$$

Let us first examine the implications of this structure on the eigenvalues for the 2×2 case. Consider a Hamiltonian of the form

$$H(x,y) = \frac{1}{2}y^2 + V(x), \tag{5.4.1}$$

with the linearization

$$B := \begin{pmatrix} \frac{\partial^2 H}{\partial x \partial y} & \frac{\partial^2 H}{\partial y^2} \\ -\frac{\partial^2 H}{\partial x^2} & \frac{\partial^2 H}{\partial x \partial y} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -V''(x_*) & 0 \end{pmatrix},$$

at a fixed point $(x,y) = (x_*, 0)$. The eigenvalues of B are

$$\lambda = \pm\sqrt{-V''(x_*)} \tag{5.4.2}$$

There are two distinct cases to consider, as illustrated in Figure 5.4.1.

1. V has a local minimum so that $V''(x_*) = \omega^2$, for some $\omega \in \mathbb{R}$. Then $\lambda = \pm i\omega$ is purely imaginary, implying the critical point is a center.

2. V has a local maximum. Say $V''(x_*) = -\theta^2$ for some $\theta \in \mathbb{R}$. Then $\lambda = \pm\theta$ is real, implying the critical point is a saddle.

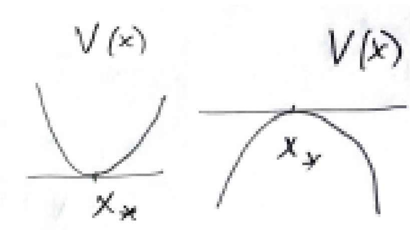

Figure 5.4.1: V has a minimum or maximum.

5.5 Lagrange's Equations

We have encountered Newton's laws in the form $F = ma$, or

$$m_i \ddot{x}_i = -\nabla_{x_i} V(\underline{x})$$

for the N-body problem. We used the conservation of energy to define

$$H(x, y) = \underbrace{T}_{\text{kinetic energy}} + \underbrace{V}_{\text{potential energy}}$$

For example, $T = \frac{1}{2} \sum_{i=1}^{N} m_i |\dot{x}_i|^2$ for the N-body problem. Let us consider a different approach to deriving the equations of motion introduced by Lagrange. Define the *Lagrangian*

$$\begin{aligned} L :& \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \\ & (x, \dot{x}) \longmapsto L(x, \dot{x}) := T(\dot{x}) - V(x) \end{aligned}$$

We define the *action* of a path $x : [0, 1] \to \mathbb{R}^n$ as follows:

$$S[x] = \int_0^1 L(x, \dot{x}) dt \tag{5.5.1}$$

We view S as a function from $C^1[0, 1] \to \mathbb{R}^n$.

The *principle of least action* says that the path that minimizes the action, subject to the boundary conditions

$$x(0) = x_0 \qquad x(1) = x_1,$$

where satisfies the ODE

$$\frac{\partial}{\partial t} \frac{\partial L}{\partial \dot{x}_i} = \frac{\partial L}{\partial x_i}, \qquad 1 \leq i \leq n.$$

These equations are known as Lagrange's equation or the *Euler-Lagrange equations* [2].

[2] The variation in terminology depends on the context. In classical mechanics, the term Lagrange's equations is used more often. When studying partial differential equations, for example the equations for minimal surfaces, the terminology Euler-Lagrange equations is more common.

When $L = \frac{1}{2}\sum_{i=1}^N m_i|\dot{x}_i|^2 - V(x)$, we find

$$\frac{\partial}{\partial t}(m_i\dot{x}_i) = -\frac{\partial}{\partial x_i}V(x), \quad \text{or} \quad m_i\ddot{x}_i = -\frac{\partial V}{\partial x_i}$$

which is Newton's law.

Let us establish the principle of least action. To this end, we need to adapt the calculus criterion for finding a max or min of a function to infinite-dimensional spaces of functions. We incorporate the boundary conditions and define the

$$X = \left\{x \in C^1([0,1];\mathbb{R}^n)\Big| \quad x(0) = x_0, \quad x(1) = x_1\right\}$$

and define the action as in equation (5.5.1). We compute the derivative of S at the 'point' x in the direction of the vector η as follows. What these 'points' mean here is the following. The 'point' x in X is a function with values $x(t)$ at $t \in [0,1]$. The 'vector' η is a sufficiently smooth function η such that $\eta(0) = \eta(1) = 0$. The boundary conditions are introduced to ensure that $x_\epsilon(t) = x(t) + \epsilon\eta(t)$ is a fuction in the space X for all ϵ.

These notions allows us to reduce the computations of derivatives to the standard calculus of functions on the line. We compute

$$\frac{d}{d\epsilon}S[x_\epsilon] = \frac{d}{d\epsilon}\int_0^1 L(x+\epsilon\eta, \dot{x}+\epsilon\dot{\eta})dt = \sum_{i=1}^n \int_0^1 \Big(\frac{\partial L}{\partial x_i}\eta_i + \frac{\partial L}{\partial \dot{x}_i}\dot{\eta}_i\Big)dt.$$

Note that in the last equality, the argument of L is $(x+\epsilon\eta, \dot{x}+\epsilon\dot{\eta})$. At an extremum

$$\begin{aligned} 0 &= \frac{d}{d\epsilon}S[x_\epsilon]\Big|_{\epsilon=0} \\ &= \sum_{i=1}^n \int_0^1 \left(\frac{\partial L}{\partial x_i}\eta_i + \frac{\partial L}{\partial \dot{x}_i}\dot{\eta}_i\right))dt \\ &= \sum_{i=1}^n \int_0^1 \left(\frac{\partial L}{\partial x_i} - \frac{d}{dt}\frac{\partial L}{\partial \dot{x}_i}\right)\eta_i dt, \end{aligned}$$

where we integrated by parts to get the last equality. Since η is arbitrary, we may choose it to be a non-negative bump function localized at any point $t_0 \in (0,1)$. Then varying this point, we see that in fact

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}_i} = \frac{\partial L}{\partial x_i}, \qquad 1 \le i \le n.$$

These are Lagrange's equations.

Remark 50. The main advantage of Lagrange's method is that it "automates" the derivation of the equations of motion, avoiding the computation of a force balance at each point. Typical examples of such Lagrangians arise when one considers mechanical linkages, as shown in Figure 5.5.1.

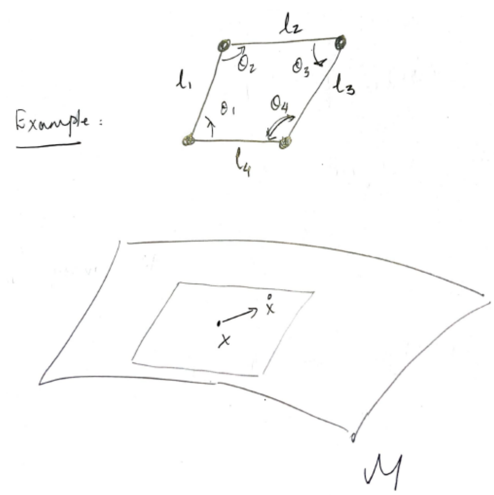

Figure 5.5.1: A planar linkage with free rotation at the joints and fixed rod lengths. A schematic for a submanifold of $\mathbb{R}^m$.

This is especially important when the space variable x lies in a manifold $\mathcal{M}$ that is not $\mathbb{R}^n$. In such examples, the admissible positions form a submanifold of a Euclidean spaces, defined as the solution set to the constraint equations. This is shown schematically in Figure 5.5.1, where $M = \{x \in \mathbb{R}^m | \text{constraints hold}\}$ and $T_x M$ = tangent space.

Example 9. *The kinetic and potential energy for the simple pendulum are*

$$T = \frac{1}{2} m (l\dot{\theta})^2, \quad V = mgl(1 - \cos\theta).$$

The Lagrangian is defined on the tangent bundle $TS^1 \equiv S^1 \times \mathbb{R}$.

5.6 Riemannian Metrics and Geodesic Flow

One of the most important applications of the principle of least action is to the derivation of the equations of geodesic flow on a Riemannian manifold. The complete definition of an abstract manifold requires a little more point-set topology (and time) than we possess at present. For these reasons, we will define n dimensional smooth manifolds as subsets of Euclidean space defined by

$$\mathcal{M} = \{x \in \mathbb{R}^m \,| g(x) = 0\},$$

where $g : \mathbb{R}^m \to \mathbb{R}^{m-n}$ is a C^∞ function such that $Dg(x)$ has rank $m-n$ at each $x \in \mathcal{M}$. Given such a subset, we may define the tangent and normal vectors to $\mathcal{M}$ with vector calculus in the usual way. [3]

[3] A theorem of Whitney allows us to reduce the study of n-dimensional abstract manifolds to this setting provided $m \geq 2n$, so this definition involves no loss of generality, even if it has

We will develop intuition for manifolds by working with examples. Riemannian and symplectic manifolds are manifolds equipped with additional structure. In this section, this additional structure is that of a *metric*.

Definition 51. A Riemannian metric g is a positive definite bilinear form on $T_x\mathcal{M}$, $x \in \mathcal{M}$. The length of a vector is defined by

$$|v|_g^2 = g(x)(v,v), \quad x \in \mathcal{M}, \quad v \in T_x\mathcal{M}. \tag{5.6.1}$$

For simplicity, we first work with metrics on $U \subset \mathbb{R}^n$. Denote by $\mathbb{P}_n$ the space of $n \times n$ positive definite matrices. Then a metric is simply a map $g : U \to \mathbb{P}^n$. We assume the map g is as smooth as needed for the calculations that follow. An important example is the following.

Example 10 (The Poincaré metric on the upper half-plane)**.**

$$U = \{y > 0 \,\big|\, (x,y) \in \mathbb{R}^2.\}$$

$$g(x,y) = \frac{1}{y^2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Definition 52. A *geodesic* between x_0 and x_1 in (U,g) is an extremum of the action

$$S_g[x] := \frac{1}{2}\int_0^1 \dot{x}^T g(x)x\,dt$$

where $x = x(t)$ and $\dot{x} = \dot{x}(t)$.

Remark 53. We do not define the geodesic as being the path of shortest distance between two points on the manifold. In most cases of interest, the extremum is a minimum, but the definition and the computation that follows, uses only a first-order variation (to find an extremum), not a second-order variation (to determine if the extremum is a maximum or a minimum).

The Lagrangian for geodesics is

$$L(x,\dot{x}) = \frac{1}{2}\dot{x}^T g(x)\dot{x} = \frac{1}{2}\dot{x}_i\dot{x}_j g_{ij}(x),$$

where we adopt the Einstein summation convention of summing over repeated indices. The equations of geodesic flow are

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}_k} = \frac{\partial L}{\partial x_k}, \qquad 1 \le k \le n.$$

Let's compute these equations explicitly. On the left hand side

$$\begin{aligned}
\frac{\partial L}{\partial \dot{x}_k} &= \frac{1}{2}\Big(\frac{\partial \dot{x}_i}{\partial x_k}\dot{x}_j g_{ij} + \dot{x}_i\frac{\partial \dot{x}_j}{\partial x_k}\Big) \\
&= \frac{1}{2}\Big(\dot{x}_j\delta_{ik}g_{ij} + \dot{x}_i\delta_{jk}g_{ij}\Big) \\
&= \frac{1}{2}(\dot{x}_j g_{jk}, \dot{x}_i g_{ik}) = \dot{x}_j g_{jk},
\end{aligned}$$

certain conceptual limitations. It is possible to develop many properties of manifolds using this working definition. The interested reader is referred to [8].

where we relabeled the dummy index i by j in the last equation. Next, for brevity, let $g_{ij,k} := \frac{\partial g_{ij}}{\partial x_k}$. Then

$$\frac{\partial L}{\partial x_k} = \frac{1}{2} g_{ij,k} \dot{x}_i \dot{x}_j. \tag{5.6.2}$$

Combining the above equations, we see that Lagrange's equations are

$$\frac{d}{dt}(\dot{x}_j g_{jk}) = \frac{1}{2} g_{ij,k} \dot{x}_i \dot{x}_j,$$

The term within brackets on the left hand side is

$$g_{jk}\ddot{x}_j + g_{jk,i}\dot{x}_i\dot{x}_j = g_{jk}\ddot{x}_j + \frac{1}{2}(g_{jk,i} + g_{ik,j})\dot{x}_i\dot{x}_j,$$

so that equation (5.6.2) may be rewritten as

$$g_{jk}\ddot{x}_j = -\frac{1}{2}(g_{jk,i} + g_{ik,j} - g_{ij,k})\dot{x}_i\dot{x}_j. \tag{5.6.3}$$

This is a complete prescription of the equation of motions. In what follows, we introduce terminology from differential geometry, so that the equations may be written in the standard form in which they appear in books on differential geometry.

The components of the inverse of the metric g^{-1} are denoted g^{lm}. They may be used to 'contract' terms, such as

$$g^{lk} g_{jk} \ddot{x}_j = \delta^l_j \ddot{x}_j = \ddot{x}_l \tag{5.6.4}$$

The spatial derivatives of the metric reflect the role of curvature. These computations are organized by introducing the *Christoffel symbols*

$$\Gamma_{ijk} = \frac{1}{2}(g_{ik,j} + g_{jk,i} - g_{ij,k}), \quad \Gamma^l_{ij} = g^{lk}\,\Gamma_{ijk}. \tag{5.6.5}$$

We multiply equation (5.6.4) on the left with g^{lm} to obtain the equations for geodesics

$$\ddot{x}_l + \Gamma^l_{ij}\dot{x}_i\dot{x}_j = 0, \quad 1 \le j \le n. \tag{5.6.6}$$

These calculations may be found in [12, Ch. 1]. In the homework, you are asked to solve these equations for the Poincaré half-plane.

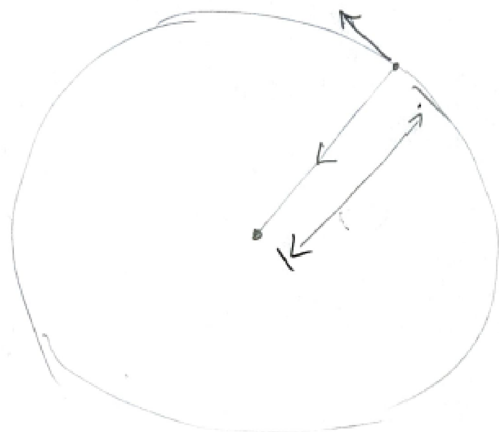

Figure 5.6.1: The role of curvature in geodesic flow is a generalization of the centripetal acceleration $a = \frac{v^2}{r}$ for a particle traveling at constant speed on a circle of radius r.

5.7 Kepler's problem

The purpose of this section is to illustrate the role of symmetries and explicit calculations in the resolution of a historically important problem in dynamical systems: the derivation of Kepler's laws of planetary motion from Newton's laws of motion and Newton's law of gravitation.

The 2-body problem is the Newtonian system

$$\begin{cases} m_1\ddot{x}_1 & -\nabla_{x_1}V(x_1,x_2) \\ m_2\ddot{x}_2 & -\nabla_{x_2}V(x_1,x_2) \end{cases} \tag{5.7.1}$$

where $x_1, x_2 \in \mathbb{R}^3$ and $V(x_1,x_2) = -\frac{m_1 m_2}{|x_1 - x_2|}$ is the gravitational potential.

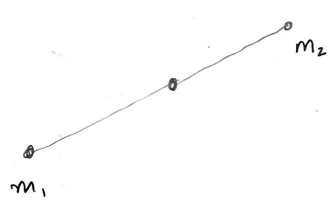

Figure 5.7.1: The center of mass in the two-body problem

5.7.1 Reduction to a central field

The two-body problem has conservation laws that allows a significant reduction in complexity. These are listed in the lemmas below.

Lemma 10. *The velocity of the center of mass is independent of time.*

Proof. Let $r = |x_1 - x_2|$, so that $V(x_1, x_2) = \frac{m_1 m_2}{r}$. We then compute

$$\nabla_{x_1} r = \frac{x_1 - x_2}{r} = -\frac{(x_2 - x_1)}{r} = \nabla_{x_2} r \tag{5.7.2}$$

Let $z = \frac{m_1 x_1 + m_2 x_2}{m_1 + m_2}$, then

$$\ddot{z} = \frac{m_1 \ddot{x_1} + m_2 \ddot{x_2}}{m_1 + m_2} = 0$$

since $\nabla_{x_1} V = -\nabla_{x_2} V$ by (5.7.2). □

Lemma 11. *The center of mass may be assumed to be at the origin for all time.*

Proof. Equations (5.7.1) are invariant under the following changes of reference frame [4]. Fix a vector $c \in \mathbb{R}^3$ and a rotation $Q \in O(3)$ and change variables to

$$y = Qx + ct, \quad x \in \mathbb{R}^3, t \in \mathbb{R}. \tag{5.7.3}$$

Since Q is an orthogonal matrix we find that

$$|y_1 - y_2| = |Q(x_1 - x_2)| = |x_1 - x_2|. \tag{5.7.4}$$

Now let us check that Newton's law continues to hold in the same manner as it did in the x-frame. We compute

$$\begin{aligned} m_1 \ddot{y} &= Q m_1 \ddot{x_1} + c\ddot{t} && (5.7.5)\\ &= -m_1 Q \nabla_{x_1} V && (5.7.6)\\ &= -m_1 Q \frac{(x_1 - x_2)}{r^3} && (5.7.7)\\ &= -m_1 \frac{y_1 - y_2}{r^3} && (5.7.8) \end{aligned}$$

Since $\dot{z}$ is constant by Lemma 10, we may choose c so that

$$\frac{m_1 \dot{y}_1 + m_2 \dot{y}_2}{m_1 + m_2} = Q\dot{z} + C = 0.$$

□

[4] This is called *Galilean invariance.*

Remark 54. Lemma 10 and Lemma 11 allow us to reduce to the two-body problem to a one-body problem. By choosing a frame in which $z = 0$, we use the conservation law

$$m_1 x_1 + m_2 x_2 = 0 \tag{5.7.9}$$

to solve for x_2 in terms of x_1. Eliminating x_2 from the equation of motion for x_1 we find

$$\ddot{x_1} = -m_2 \left(\frac{(m_1 + m_2)}{m_2} \right) \frac{x_1}{|x_1|^3}.$$

This is a vector equation in $\mathbb{R}^3$ that may be simplified further by additional conservation laws.

Lemma 12. *The angular momentum $m_1 \underline{x}_1 \times \dot{\underline{x}}_1$ is conserved.*

Here $\underline{x}_1 \times \dot{\underline{x}}_1$ is the cross product in $\mathbb{R}^3$. See Figure 5.7.2.

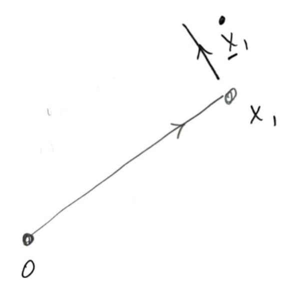

Figure 5.7.2: The angular momentum.

Proof.

$$\frac{d}{dt}(\underline{x_1} \times \underline{\dot{x_1}}) = \underline{\dot{x_1}} \times \underline{\dot{x_1}} + \underline{x_1} \times \underline{\ddot{x_1}} = -\frac{\underline{x_1} \times \underline{x_1}}{|x|^3} = 0.$$

□

5.7.2 Motion in a central field

Let us briefly recall vector calculus with polar coordinates. A point $\underline{x} = (x, y) \in \mathbb{R}^2$ may also be written

$$\underline{x} = r\underline{e}_r$$

where the basis vectors are shown in Figure 5.7.3. As φ varies, the basis vectors change and it is easy to check that

$$\partial_\varphi \underline{e}_r = \underline{e}_\varphi, \quad \partial_\varphi \underline{e}_\varphi = -\underline{e}_r.$$

The velocity and acceleration are given by

$$\dot{\underline{x}} = \dot{r}\underline{e}_r + r\dot{\varphi}\underline{e}_\varphi, \tag{5.7.10}$$

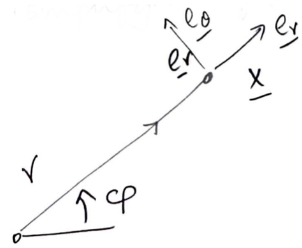

Figure 5.7.3: Basis vectors in polar coordinates.

$$\underline{\ddot{x}} = \left(\ddot{r} - r\dot{\varphi}^2\right)\underline{e}_r + \left(r\ddot{\varphi} + 2\dot{r}\dot{\varphi}\right)\underline{e}_\varphi. \tag{5.7.11}$$

These equations are obtained by differentiating the basis vectors with respect to φ and applying the chain rule.

Definition 55. Assume $U : (0, \infty) \to \mathbb{R}$ is a potential. Let $x \in \mathbb{R}^2$, $r = |x|$. The equation of motion of a particle in the *central field* defined by U is

$$\ddot{x} = -\nabla U(|x|). \tag{5.7.12}$$

The right hand side simplifies considerably in polar coordinates, since

$$\nabla U(|x|) = U'(r)\nabla r = U'(r)\underline{e}_r.$$

The left hand side has been computed in equation (5.7.11) and balancing the radial and angular directions we find the system of equations

$$\ddot{r} - \; r\dot{\varphi}^2 = -\frac{\partial U}{\partial r}, \tag{5.7.13}$$

$$r\ddot{\varphi} + 2\dot{r}\dot{\varphi} = 0. \tag{5.7.14}$$

Equation (5.7.14) may be integrated to obtain the conservation law

$$\dot{\varphi} = \frac{M}{r^2}, \tag{5.7.15}$$

where M is the angular momentum [5] The value of the constant M is determined by the initial conditions. Once it is known, we define the *effective potential energy*

$$V(r) = U(r) + \frac{M^2}{2r^2} \tag{5.7.16}$$

[5]This is nothing but Lemma 12 in disguise. Indeed, observe that

$$r\ddot{\varphi} + 2\dot{r}\dot{\varphi} = 0$$

is equivalent to

$$\frac{d}{dt}\log(\dot{\varphi}r^2) = 0.$$

and observe that equation (5.7.13) takes the form

$$\ddot{r} = -\frac{\partial V}{\partial r}. \tag{5.7.17}$$

At this stage we have reduced the two-body problems to the techniques of Section 5.1. Let us first analyze this problem qualitatively, before turning to a more precise analysis.

Figure 5.7.4 illustrates the qualitative nature of the r-orbits for a potential energy $U(r)$ that grows at infinity. Figure 5.7.5 illustrates the qualitative nature of the r-orbits for the gravitational potential. In this setting, the potential energy $U(r)$ does not grow at infinity and we see a separation between periodic r-orbits and orbits that asymptote to infinity.

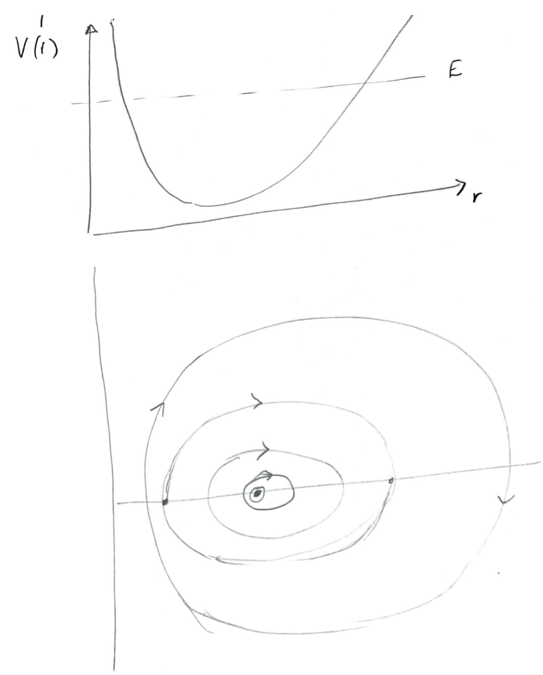

Figure 5.7.4: Periodic r-orbits in a central field.

When considering the gravitational potential, we have used the fact that $\frac{M^2}{r^2}$ dominates as $r \to 0$. Further, since $\frac{1}{2}\dot{r} + V(r) = E$, $\dot{r} \sim \sqrt{2E}$ when $E > 0$ since $V(r) \to 0$ as $r \to \infty$.

Figure 5.7.4 and Figure 5.7.5 provide solutions $r(t)$ that are periodic function of time. However, our system is two dimensional and this does not suffice to establish that the orbit of the particle is closed. This is more subtle. To this

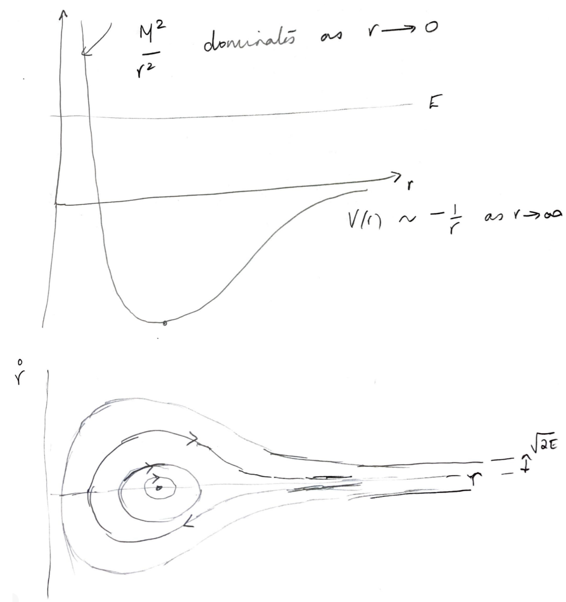

Figure 5.7.5: Periodic r-orbits in the gravitational field.

end, we express φ as a function of r writing

$$\frac{d\varphi}{dr} = \frac{d\varphi}{dr}\frac{dr}{dt} = \frac{M}{r^2}\frac{1}{\sqrt{2(E - V(r))}}.$$

This equation follows from the equations

$$\dot{r} = \sqrt{2(E - V(r))}, \quad \dot{\varphi} = \frac{M}{r^2}.$$

Thus, we may integrate to obtain φ as a function of r. Let's first get a feel for this qualitatively. Figure 5.7.6 plots the particle position in space for a periodic r-orbit. As r increases from r_{min} to r_{max}, φ increases monotonically via

$$\varphi(r) = \int_{r_{\text{min}}}^{r} \frac{M}{s^2}\frac{1}{\sqrt{2(E - V(s))}} ds \tag{5.7.18}$$

The angle between sucessive pericenters and apocenters is given by the integral

$$\Phi = \int_{r_{\text{min}}}^{r_{\text{max}}} \frac{M}{r^2}\frac{1}{\sqrt{2(E - V(r))}} dr \tag{5.7.19}$$

Whether the orbit is closed or not depends on whether $\frac{\Phi}{2\pi}$ is rational or irrational. This idea is illustrated in Figure 5.7.7.

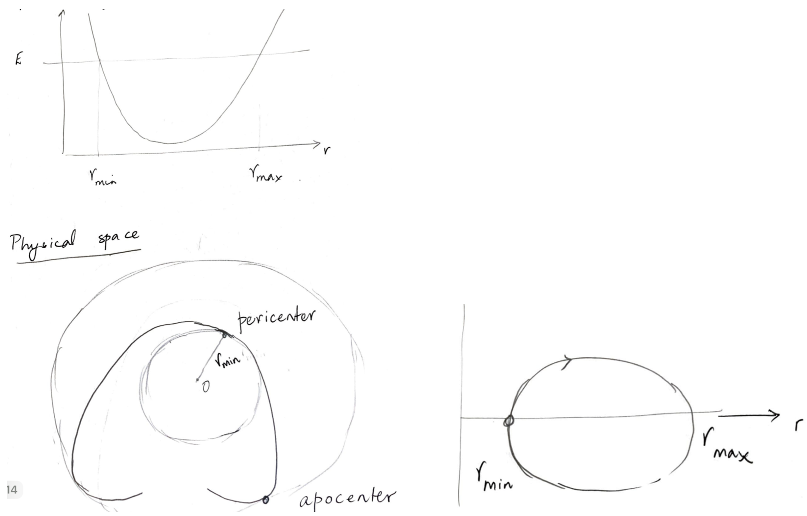

Figure 5.7.6: Orbits in physical space

Theorem 56. *The only central force law in which all orbits are closed is:*

$$U = ar^2, \qquad a \geq 0 \tag{5.7.20}$$

$$U = -\frac{k}{r}, \qquad k \geq 0 \tag{5.7.21}$$

Remark 57. The suprising fact here is that we are not assuming Newton's law of gravitation. What we assume is that the orbits are closed (Kepler's law). This implies Newton's law. Note, however, that we do assume Newton's law of motion ($F = ma$) An interesting question here is "how did Newton come up with the law of gravitation?". Kepler's calculations based on Tycho Brahe's observations seems to be the essential clue. A more detailed discussion of these ideas may be found in [3, Ch. 2.8].

5.8 Exercises

1. We say that a matrix A with real entries is Hamiltonian if JA is symmetric.

   (a) Show that the sum and commutator of two Hamiltonian matrices are also Hamiltonian matrices.

   (b) Compute the dimension of the space of Hamiltonian matrices.

   (c) Show that if $\lambda \in \mathbb{C}$ is an eigenvalue of a Hamiltonian matrix A, then so is $-\lambda$, λ^* and $-\lambda^*$.

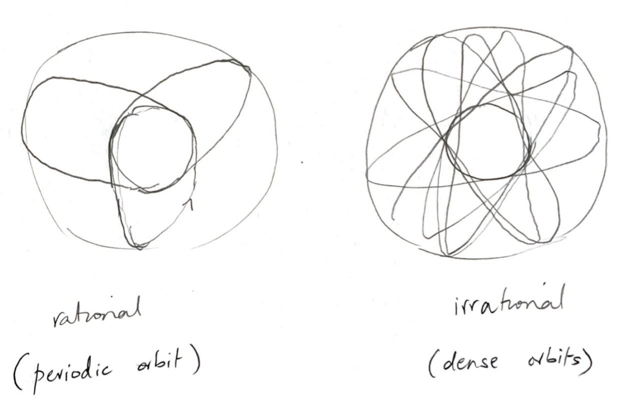

Figure 5.7.7: Periodic and quasi-periodic motions in a central field.

2. Recall that the symplectic group $Sp(2n)$ is the group of matrices with real entries defined by the relation:

$$M^T JM = J.$$

Show that $\{e^{tA}\}_{t\in\mathbb{R}}$ is symplectic if A is Hamiltonian. Conversely, given a smooth path $M(t) \in Sp(2n)$ with $M(0) = I_{2n}$, show that $\dot{M}(0)$ is a Hamiltonian matrix.

3. Consider the equation of the simple pendulum:

$$\ddot{\theta} + \sin\theta = 0.$$

A critical energy level separates small oscillations with extrema in $(-\pi, \pi)$ from large 'whirling' oscillations. Determine explicitly the solution on the critical energy level as a function of t.

4. *Circle maps.* Consider the map $f : [0, 1) \to [0, 1)$ defined by $f(x) = (x + \alpha)$ mod 1 where $\alpha \in [0, 1)$. Let the sequence $\{x_n\}$, denote the orbit of a point x_0, i.e. $x_1 = f(x_0)$, $x_2 = f(x_1)$, etc.

(a) Prove that every orbit is periodic if and only if α is rational.

(b) If α is irrational, prove that the orbit $\{x_n\}$ is dense in $[0, 1)$.

5. *Geodesics as paths of least action.* Assume given a smooth metric g on $\mathbb{R}^n$ (i.e. $g(x)$ is a symmetric, positive definite matrix) that varies smoothly with x. Denote the length of a vector in this metric by $|v|_g := \sqrt{g(v, v)}$. The length of a smooth curve $\gamma : [a, b] \to \mathbb{R}^n$ is the function of γ defined by

$$L(\gamma) = \int_a^b |\dot{\gamma}|\, dt.$$

The action of this path is the function

$$E(\gamma) = \frac{1}{2}\int_a^b |\dot{\gamma}|^2 \, dt.$$

(Its conventional to use E instead of $\mathcal{A}$ because the action is the kinetic energy of a particle moving in the metric g in this case).

(a) Show that $L(\gamma)$ is unchanged under a reparametrization of the curve γ.

(b) Show that minimizing the action of a parametrized curve is the same as minimizing the length, if one makes the additional assumption that the speed $|\dot{\gamma}|_g$ is held constant.

6. *Geodesics in the upper half plane.* Let $\mathbb{H} = \{(x, y) \in \mathbb{R}^2 | y > 0\}$. Let g be the *hyperbolic metric* $g = y^{-2}I$, where I denotes the identity matrix.

(a) Show that the geodesics are circular arcs perpendicular to the x-axis.

(b) Compute the distance between two points (x_1, y_1) and (x_2, y_2).

5.9 Solutions to exercises

1. We say that a matrix A with real entries is Hamiltonian if JA is symmetric.

(a) Show that the sum and commutator of two Hamiltonian matrices are also Hamiltonian matrices.

(b) Compute the dimension of the space of Hamiltonian matrices.

(c) Show that if $\lambda \in \mathbb{C}$ is an eigenvalue of a Hamiltonian matrix A, then so is $-\lambda$, λ^* and $-\lambda^*$.

Proof. (a) Suppose A and B are Hamiltonian matrices. Then $(JA)^T = JA$ and $(JB)^T = JB$. We then compute

$$(J(A + B))^T = (A + B)^T J^T = A^T J^T + B^T J^T = JA + JB,$$

since $(JA)^T = JA$. The commutator is $[A, B] = AB - BA$. We then compute

$$\begin{aligned} J([A, B])^T &= B^T A^T J^T - A^T B^T J^T = B^T (JA)^T - A^T (JB)^T \\ &= B^T JA - A^T JB = -(JB)^T A + (JA)^T B = JAB - JBA = J[B, A]. \end{aligned}$$

(b) The dimension of the space of real symmetric matrices is $n(n + 1)/2$. If A is Hamiltonian, we may write $JA = S$ or $A = J^{-1}S$ where S is symmetric. Thus, the dimension of the space of Hamiltonian matrices is also $n(n + 1)/2$.

(c) The condition $(JA)^T = JA$ is equivalent to $A^T = JAJ$ since $J^T = J^{-1}$. Therefore, the characteristic polynomial satisfies the identity

$$\det(\lambda I - A) = \det(\lambda I - A^T) = \det(\lambda I - JAJ) = \det(\lambda J^{-2} - A) = \det(-\lambda I - A),$$

where we used the identities $\det(J) = 1$, $J^{-1} = -J$ and $J^2 = -I$. It follows that λ is a zero if and only if $-\lambda$ is a zero. Further, since A is real, the complex conjugate $\overline{\lambda}$ is a zero if and only if λ is.

□

2. Recall that the symplectic group $Sp(2n)$ is the group of matrices with real entries defined by the relation:

$$M^T JM = J.$$

Show that $\{e^{tA}\}_{t\in\mathbb{R}}$ is symplectic if A is Hamiltonian. Conversely, given a smooth path $M(t) \in Sp(2n)$ with $M(0) = I_{2n}$, show that $\dot{M}(0)$ is a Hamiltonian matrix.

Proof. (a) Suppose A is Hamiltonian and consider $M(t) = e^{tA}$. We use the definition of the matrix exponential to find that

$$\dot{M} = AM = MA.$$

In order to show that $M(t) \in Sp(2n)$ we observe that $M^T JM = J$ at $t = 0$ and

$$\frac{d}{dt} M^t JM = \dot{M}^T JM + M^T J\dot{M} = M^T \left(A^T J + JA\right) M.$$

Since A is Hamiltonian

$$JA = (JA)^T = A^T J^T = -A^T J.$$

Thus, the term within the brackets vanishes and $M^T JM = J$ for all t.

(b) Conversely, if $M(t) \in Sp(2n)$ and $M(0) = I$ at $t = 0$, writing $A = \dot{M}(0)$ the calculation above shows that $JA = (JA)^T$. □

3. Consider the equation of the simple pendulum:

$$\ddot{\theta} + \sin\theta = 0.$$

A critical energy level separates small oscillations with extrema in $(-\pi, \pi)$ from large 'whirling' oscillations. Determine explicitly the solution on the critical energy level as a function of t.

Proof. The Hamiltonian for the simple pendulum is

$$H(\theta, \dot{\theta}) = \frac{1}{2}\dot{\theta}^2 + 1 - \cos\theta.$$

Let E denote the value of the Hamiltonian on the critical energy level. This is the energy of the critical point $\theta = \pi$, $\dot{\theta} = 0$. Therefore, $E = 2$. On other points on this energy level, we have the conservation law

$$\frac{1}{2}\dot{\theta}^2 = E - (1 - \cos\theta) = 1 + \cos\theta.$$

We use the trigonometric identity

$$\cos\theta = 2\cos^2\frac{\theta}{2} - 1,$$

separate variables and take square-roots to obtain the identity

$$\int \frac{d\theta}{2\cos\frac{\theta}{2}} = t.$$

The LHS may be further reduced to the standard integral

$$\int \frac{d\varphi}{\cos\varphi}, \quad \text{with} \quad \varphi = \frac{1}{2}\theta.$$

We use a table of integrals to find

$$\int \frac{d\varphi}{\cos\varphi} = \ln|\sec x + \tan x| = 2\tanh^{-1}\left(\tan\frac{\varphi}{2}\right).$$

Thus, we have found the implicit solution formula

$$t - t_0 = 2\tanh^{-1}\tan\frac{\theta}{4},$$

where the initial time t_0 plays the role of the arbitrary constant of integration. We invert the above equation to obtain

$$\theta = 4\tan^{-1}\left(\tanh\frac{t-t_0}{2}\right).$$

Here we use the branch of $\tan^{-1}$ that maps $(-\infty,\infty)$ to $(-\pi/2,\pi/2)$. Thus, as $t \to \pm\infty$ we have $\theta(t) \to \pm\pi$ as desired. □

4. *Circle maps.* Consider the map $f : [0,1) \to [0,1)$ defined by $f(x) = (x+\alpha) \bmod 1$ where $\alpha \in [0,1)$. Let the sequence $\{x_n\}$, denote the orbit of a point x_0, i.e. $x_1 = f(x_0)$, $x_2 = f(x_1)$, etc.

(a) Prove that every orbit is periodic if and only if α is rational.

(b) If α is irrational, prove that the orbit $\{x_n\}$ is dense in $[0,1)$.

Proof. (a) Given $x_0 \in [0,1]$, let $z_0 = x_0$ and let $z_{n+1} = z_n + \alpha$ denote a 'lift' of the sequence x_n into the covering space $\mathbb{R}$. The orbit of x_0 has period q if and only if $z_q - z_0$ is an integer, say p, and $z_k - z_0$ is not an integer for $1 \le k \le q-1$. But $z_q = z_0 + q\alpha$ (this is where it is simpler to work on $\mathbb{R}$). Therefore, $q\alpha = p$, or $\alpha = p/q$.

(b) Now suppose that α is irrational. Since all orbits are rigid translations of the orbit of $x_0 = 0$, let us suppose that $x_0 = 0$. It is enough to show that for each $\varepsilon > 0$ there is an integer q such that $|x_q - x_0| < \varepsilon$. As in part (a), we work with the lifts $\{z_k\}_{k=0}^{\infty}$ and it is enough to show that for each $\varepsilon > 0$ there are integers p and q such that $|z_q - p| < \varepsilon$. But $z_q = q\alpha$, so what we must show is that there are integers p and q such that

$$\left|\alpha - \frac{p}{q}\right| < \frac{\varepsilon}{q}.$$

This follows from the Euclidean algorithm. The continued fraction expansion of an irrational number provides a sequence of integers (p_n, q_n) such that

$$\left|\alpha - \frac{p_n}{q_n}\right| < \frac{1}{q_n^2}.$$

A proof of the Euclidean algorithm may be found in Ch.3 of Arnold's book on the Geometric Theory of Ordinary Differential Equations. Part (b) may also be proved by assuming Weyl's equidistribution theorem if one wants to avoid number theory altogether. □

5. *Geodesics as paths of least action.* Assume given a smooth metric g on $\mathbb{R}^n$ (i.e. $g(x)$ is a symmetric, positive definite matrix) that varies smoothly with x. Denote the length of a vector in this metric by $|v|_g := \sqrt{g(v,v)}$. The length of a smooth curve $\gamma : [a,b] \to \mathbb{R}^n$ is the function of γ defined by

$$L(\gamma) = \int_a^b |\dot{\gamma}|\, dt.$$

The action of this path is the function

$$E(\gamma) = \frac{1}{2}\int_a^b |\dot{\gamma}|^2\, dt.$$

(Its conventional to use E instead of $\mathcal{A}$ because the action is the kinetic energy of a particle moving in the metric g in this case).

(a) Show that $L(\gamma)$ is unchanged under a reparametrization of the curve γ.

(b) Show that minimizing the action of a parametrized curve is the same as minimizing the length, if one makes the additional assumption that the speed $|\dot{\gamma}|_g$ is held constant.

Proof. (a) Assume that $\varphi : [a,b] \to [a,b]$ is a C^1 strictly increasing map. Let $t = \varphi(s)$, $\eta(s) = \gamma(\varphi(s))$ and let $\eta' = d\eta/ds$. We use the chain rule to obtain

$$\int_a^b |\eta'(s)|\; ds = \int_a^b |\gamma'(t)| \left|\frac{dt}{ds}\right|\, ds = \int_a^b |\gamma'(t)|\; dt,$$

since $\varphi'(s) > 0$.

(b) For brevity, let us denote the two Lagrangians in this problem by

$$L_0 = |\dot{\gamma}|, \quad L_1 = \frac{1}{2}L_0^2 = \frac{1}{2}|\dot{\gamma}|^2.$$

Then the Euler-Lagrange equations involve the derivatives

$$\frac{\partial L_1}{\partial x_i} = L_0 \left(\frac{\partial L_1}{\partial x_i} \right), \quad \frac{\partial L_1}{\partial \dot{x}_i} = L_0 \left(\frac{\partial L_1}{\partial \dot{x}_i} \right).$$

In particular, when we assume that the parametrization is chosen so that L_0 is held constant in time, we find that

$$\frac{d}{dt} \left(\frac{\partial L_1}{\partial \dot{x}_i} \right) = L_0 \frac{d}{dt} \left(\frac{\partial L_0}{\partial \dot{x}_i} \right) + \frac{dL_0}{dt} \frac{\partial L_0}{\partial \dot{x}_i} = L_0 \frac{d}{dt} \left(\frac{\partial L_0}{\partial \dot{x}_i} \right),$$

and the Euler-Lagrange equations have the same solutions. □

6. *Geodesics in the upper half plane.* Let $\mathbb{H} = \{(x, y) \in \mathbb{R}^2 | y > 0\}$. Let g be the *hyperbolic metric* $g = y^{-2} I$, where I denotes the identity matrix.

(a) Show that the geodesics are circular arcs perpendicular to the x-axis.

(b) Compute the distance between two points (x_1, y_1) and (x_2, y_2).

Proof. There are two ways to do this problem. The slick solution (which we will consider in lecture) uses the invariance of the metric under Möbius transformations. However, in this problem, we do not assume that these invariances are known: our goal is to discover the exact solution for geodesics by following the procedure outlined in lecture. First, we derive the equation for geodesics using Lagrange's equation. Then we solve these equations explicitly using what we know about Hamiltonian systems.
1. Equations for geodesics. The Lagrangian in this problem is

$$L(x, \dot{x}, y, \dot{y}) = \frac{1}{2y^2} \left(\dot{x}^2 + \dot{y}^2 \right).$$

Therefore,

$$\frac{\partial L}{\partial x} = 0, \quad \frac{\partial L}{\partial y} = -\frac{1}{y^3} \left(\dot{x}^2 + \dot{y}^2 \right), \quad \frac{\partial L}{\partial \dot{x}} = \frac{\dot{x}}{y^2}, \quad \frac{\partial L}{\partial \dot{y}} = \frac{\dot{y}}{y^2}.$$

Thus, Lagrange's equations are

$$\frac{d}{dt} \left(\frac{\dot{x}}{y^2} \right) = 0, \quad \frac{d}{dt} \left(\frac{\dot{y}}{y^2} \right) = -\frac{1}{y^3} \left(\dot{x}^2 + \dot{y}^2 \right).$$

2. Integration of the equations of motion. The equation for $\dot{x}$ implies immediately that $\dot{x} = ay^2$ for a constant a. The equation for $\dot{y}$ may be rewritten in the form

$$\frac{\ddot{y}}{y} - \frac{\dot{y}^2}{y^2} = -\frac{\dot{x}^2}{y^2}.$$

This equation may be simplified by observing that the LHS is the second derivative of $\ln y$. Thus, let $u = \ln y$ and use $\dot{x} = ay^2$ to rewrite the above equation in the form

$$\ddot{u} = -a^2 e^{2u}. \tag{5.9.1}$$

When $a = 0$, we find that $x(t) = x(0)$ and $u(t) = u_0+ct$. Therefore, $y(t) = y_0 e^{ct}$ and the geodesic is a vertical line in the upper-half plane. It is possible to use this fact alone, along with the invariance of the metric under Möbius transformations, to compute *all* geodesics. However, we will integrate equation (5.9.1) directly by studying the case $a \neq 0$. Thus, assume in what follows that $a \neq 0$.

Equation (5.9.1) is a 1-D Hamiltonian system with a potential $V(u) = a^2/2e^{2u}$. We have the conservation law

$$\frac{1}{2}\dot{u}^2 + \frac{a^2}{2}e^{2u} = E,$$

as well as the integral formula

$$t = \int \frac{du}{\sqrt{2(E - V(u))}}. \tag{5.9.2}$$

Plotting the graph of $V(u)$ we see that for a given energy level, $u(t) \to -\infty$ as $t \to \pm\infty$ with a maximum value $u_{\max}$ determined by

$$E = V(u_{\max}) = \frac{a^2}{2}e^{2u_{\max}}.$$

Let us now return to the earlier variables using this insight. Set

$$y_{\max} = e^{u_{\max}}, \quad v = u - u_{\max}, \quad s = e^v = \frac{y}{y_{\max}}.$$

Then equation (5.9.2) can be simplified to

$$a y_{\max} t = \int \frac{1}{s}\frac{ds}{\sqrt{1-s^2}}.$$

The indefinite integral on the right hand side can be computed using a standard trigonometric substitution. Set $s = \sin\theta$, so that

$$\int \frac{1}{s}\frac{ds}{\sqrt{1-s^2}} = \int \frac{d\theta}{\sin\theta} = \ln|\tan\frac{\theta}{2}|,$$

using a table of integrals.

3. Parametrized geodesics. We then undo the various changes of variables to obtain the formula

$$y(t) = y_{\max}\operatorname{sech}(a y_{\max} t).$$

Substituting this relation in the conservation law $\dot{x} = ay^2$, we find after another integration that

$$x(t) - x_0 = y_{\max}\tanh(a y_{\max} t).$$

These equations for $(x(t), y(t))$ parametrize a semicircle of radius $y_{\max}$ centered at $(x_0, 0)$. The origin in time is chosen so that when $t = 0$, (x_0, y_0) lies at the tip of the semicircle. As $t \to \pm\infty$ it approaches the boundary points $(x_0 \pm y_{\max}, 0)$.

4. The distance between two points. Suppose (x_1, y_1) and (x_2, y_2) lie on the geodesic semicircle with radius $y_{\max}$ centered at $(0,0)$. It will be enough to assume that one of the points is $(0, y_{\max})$. Since the geodesic distance is not independent on the parametrization of time, we choose $ay_{\max} = 1$, so that the geodesics are

$$x(t) = y_{\max} \tanh t, \quad y(t) = y_{\max} \mathrm{sech} t. \tag{5.9.3}$$

Then the Lagrangian evaluated along the geodesic is

$$L(x, \dot{x}, y, \dot{y}) = \frac{1}{y^2}(\dot{x}^2 + \dot{y}^2) = 1,$$

using the identities

$$\dot{x} = \mathrm{sech}^2 t, \quad \dot{y} = -\mathrm{sech}\, t \tanh t, \quad \mathrm{sech}^2\, t + \tanh^2 = 1.$$

Therefore, the geodesic distance between (x_1, y_1) and $(0, y_{\max})$ is simply

$$\int_0^t \sqrt{L(x, \dot{x}, y, \dot{y})}\, dt.$$

But this is simply the time taken to get from $(0, y_{\max})$ to (x_1, y_1) which is obtained by inverting equation (5.9.3)

$$t = \cosh^{-1} \frac{y_{\max}}{y_1}.$$

□

Chapter 6

Ergodicity and Mixing

The primary source for this chapter is [2, Ch.3].

6.1 Weyl's equidistribution theorem

In this section $S^1 = \mathbb{R}/\mathbb{Z}$. A *circle map* is a homeomorphism of S^1. The simplest class of circle maps are the *rigid rotations*. Given $\alpha \in \mathbb{R}$ define the rotation $R_\alpha : S^1 \to S^1$ by

$$x \to x + \alpha \pmod 1 \tag{6.1.1}$$

The following theorem about rotations was proven in the homework.

Theorem 58 (Jacobi, 1835). *Suppose $\alpha \notin \mathbb{Q}$. Then the orbit $\{R_\alpha^n(x)\}_{n=0}^\infty$ is dense in S^1 for every $x \in S^1$.*

This theorem is related to Hamiltonian systems in the following way. Let ω_1 and ω_2 be fixed positive numbers and consider the Hamiltonian $H : \mathbb{R}^4 \to \mathbb{R}$,

$$H(x,y) = \frac{\omega_1}{2}(x_1^2 + y_1^2) + \frac{\omega_2}{2}(x_2^2 + y_2^2).$$

Then the equations of motion are

$$\begin{aligned} \dot{x}_1 &= \omega_1 y_1, \quad & \dot{x}_2 &= \omega_2 y_2, \\ \dot{y}_1 &= -\omega_1 x_1, \quad & \dot{y}_2 &= -\omega_2 y_2. \end{aligned}$$

This is a system of two uncoupled simple harmonic oscillators. Let $r_i^2 = x_i^2 + y_i^2$, $i = 1, 2$, denote the radii of individual orbits. The radii are conserved and the dynamics is determined by the evolution of the angles θ_1, θ_2 defined by $(x_i, y_i) = r_i(\cos\theta_i, \sin\theta_i)$. Then we obtain the evolution equation

$$\dot{\theta}_1 = \omega_1, \quad \dot{\theta}_2 = \omega_2,$$

and all trajectories lie on an invariant torus within $\mathbb{R}^4$.

An important idea developed in the 1920's was the extension of Jacobi's theorem to ergodic theorems. Let us illustrate this idea with examples.

Theorem 59 (Weyl)**.** *Suppose $\alpha \notin Q$. For every $x \in S^1$ and every interval $I \subset S^1$ we have*

$$\lim_{n\to\infty} \frac{1}{n} \#\{0 \le k \le n-1 | R_\alpha(x) \in I\} = |I|. \tag{6.1.2}$$

Remark 60. Here $|I|$ denotes the length of I. An equivalent formulation of Weyl's theorem is as follows. Suppose $f : S^1 \to \mathbb{R}$ is Riemann integrable. Then,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(R_\alpha^k(x)) = \int_{S^1} f(s)ds. \tag{6.1.3}$$

This is an example of an *ergodic theorem*. The left hand side is a time average and the right hand side is a spatial average. The equivalence between the formulations (6.1.2) and (6.1.3) is as follows. First, by setting $f(x) = 1_I(x)$ in (6.1.3), we recover (6.1.2). Conversely, every Riemann integrable function can be approximated with step functions, so that (6.1.2) implies (6.1.3). This approximation argument is presented in the proof.

Proof. 1. We first prove equation (6.1.2) for trigonometric functions. Suppose

$$f(x) = e^{2\pi i m x}, \quad m \in \mathbb{Z}.$$

Then we compute

$$f(R_\alpha(x)) = e^{2\pi i m(x+\alpha)} = e^{2\pi i m x} e^{2\pi i m \alpha},$$

and by induction

$$f(R_\alpha^{(k)}(x)) = e^{2\pi i m x} e^{2\pi i m \alpha k}.$$

For brevity, let $z = z(\alpha) = e^{2\pi i m \alpha}$. Then the left hand side of equation (6.1.3) is

$$\frac{1}{n} \sum_{k=0}^{n-1} f(x) z^k = \frac{f(x)}{n}(1 + z + z^2 + \cdots + z^{n-1}) = \frac{f(x)}{n} \frac{z^n - 1}{z - 1}.$$

Now, $z = e^{2\pi i m \alpha} \neq 1$ unless $m = 0$, since $\alpha \notin \mathbb{Q}$. Thus, when $m \neq 0$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(R_\alpha^{(k)}(x)) = 0.$$

The right hand side of (6.1.3) for this case is

$$\int_0^1 e^{2\pi m(s+\alpha)} ds = \frac{e^{2\pi i m \alpha}}{2\pi i m}(e^{2\pi i m} - 1) = 0.$$

Thus, equation (6.1.3) holds for $m \neq 0$. When $m = 0$, both right and left hand sides are identically 1 so it holds in this case too.

2. Suppose $f(x) = \sum_{m\in\mathbb{Z}} c_m e^{2\pi i m x}$ where only finitely man c_m are non-zero. The theorem holds by step 1 and linearity of the left and right hand sides. Since every continuous function on S^1 can be uniformly approximated by polynomials, and the Taylor expansions of $e^{2\pi m x}$ is globally covergent, we may uniformly approximate any continuous function by trigonometric polynomials. Thus, equation (6.1.3) holds for every continuous function.

3. For any $\varepsilon > 0$ we choose piecewise linear continuous functions $f_\pm$ that approximate $1_I(x)$ from above and below and differ from f only on interval of size ε. Specifically, suppose $I = (a, b)$ and choose

$$f_-(x) = \frac{(x-a)}{\varepsilon}\mathbf{1}_{(a,a+\varepsilon)}(x) + \mathbf{1}_{(a+\varepsilon,b-\varepsilon)}(x) + \frac{b-x}{\varepsilon}\mathbf{1}_{(b-\varepsilon,b)}(x), \qquad (6.1.4)$$

$$f_+(x) = \frac{(a-x)}{\varepsilon}\mathbf{1}_{(a-\varepsilon,a)}(x) + \mathbf{1}_{(a,b)}(x) + \frac{x-b}{\varepsilon}\mathbf{1}_{(b,b+\varepsilon)}(x). \qquad (6.1.5)$$

Therefore, for any $x \in S^1$,

$$\begin{aligned}\int_0^1 f_-(s)ds &= \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} f_-(R_\alpha^k(x)) \le \liminf_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} 1_I(R_\alpha^k(x)) \\ &\le \limsup_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} 1_I(R_\alpha^k(x)) \le \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n} f_+(R_\alpha^k(x)) = \int_0^1 f_+(s)ds.\end{aligned}$$

By the construction of $f_\pm$ we also have the matching bound:

$$\int_0^1 f_-(x)dx - \varepsilon \le |I| \le \int_0^1 f_+(x)dx + \varepsilon$$

This shows that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} 1_I(R_\alpha^k(x)) = |I|.$$

□

6.2 Anosov's Map

In this section, we first define ergodicity and mixing in an abstract setting. We then illustrate these ideas with an important example introduced by Anosov in the 1960s.

Definition 61. Let $(X, \mathcal{B}, \mu)$ be a measure space. A map $\varphi : X \to X$ is *measure preserving* if $\mu(\varphi^{-1}(G)) = \mu(G)$ for every $G \in \mathcal{B}$.

The map φ defines a *discrete dynamical system*. The *time mean* of a function $f \in L^1(X, \mathcal{B}, \mu)$ if it exists is defined by

$$f^*(x) = \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} f(\varphi^k(x)). \qquad (6.2.1)$$

The *space mean* is defined by

$$\bar{f} = \int_X f(x) d\mu(x). \tag{6.2.2}$$

Definition 62. A measure preserving transformation φ is *ergodic* if $f^* = \bar{f}$ for every $f \in L^1(X, \mathcal{B}, \mu)$. The transformation φ is *mixing* if

$$\lim_{n\to\infty} \mu(\varphi^n(F) \cap G) = \mu(F)\mu(G)$$

for every pair of sets $F, G \in \mathcal{B}$.

Remark 63. The above definition of mixing formalizes our intuitive notion of the mixing of fluids such as water. In the first approximation, a glass of water or a cup of coffee is an incompressible fluid with constant density. If one stirs the coffee a bit and let its go, we obtain a volume preserving transformation. Thus, when milk is stirred into coffee, it 'goes all over the place' while preserving volume and the end result is a solution where there is an equal amount of milk everywhere in the coffee.

Remark 64. Theorem 59 shows that the circle map R_α with irrational α is ergodic. However, R_α is not mixing.

We now focus on the following transformation introduced by Anosov. The underlying measure space is $(X, \mathcal{B}, \mu) = \mathbb{T}^2$ with Lebesgue measure. Consider the matrix

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \tag{6.2.3}$$

and use it to define the transformation on $\mathbb{T}^2$

$$\varphi(x) = Ax \mod \mathbb{Z}^2. \tag{6.2.4}$$

Lemma 13. *φ is a measure-preserving diffeomorphism of $\mathbb{T}^2$.*

Proof. The entries of A are integers. Therefore, $Ax \in \mathbb{Z}^2$ when $x \in \mathbb{Z}^2$. We compute $\det(A) = 1$ and

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}.$$

Thus, A^{-1} also maps $\mathbb{Z}^2 \to \mathbb{Z}^2$, which implies A^{-1} is well-defined as a map from $\mathbb{T}^2 \to \mathbb{T}^2$. Both φ and φ^{-1} are locally determined by A and A^{-1}; thus they are diffeomorphisms. $\square$

Lemma 14. *The matrix A has eigenvalues and eigenvectors*

$$\lambda_\pm = \frac{3 \pm \sqrt{5}}{2}, \quad \text{and} \quad u_\pm = \begin{bmatrix} 1 \\ \lambda_+ - 2 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ \lambda_- - 2 \end{bmatrix}. \tag{6.2.5}$$

Proof. This is a computation with the characteristic polynomial $\det(\lambda - A)$. $\square$

Remark 65. Note that $0 < \lambda_- < 1 < \lambda_+$ and that both these numbers are irrational. Therefore, the eigendirections in $\mathbb{R}^2$ 'wrap around' into dense orbits in $\mathbb{T}^2$. We call these curves $\mathcal{F}_u$ and $\mathcal{F}_s$ respectively. At each $x \in \mathbb{T}^2$ the linearization $D\varphi(x)$ splits into two invariant subspaces parallel to these directions. The effect of these transformations is to stretch and squash a neighborhood of x into a long skinny region that follows $\mathcal{F}_u$.

Theorem 66. *The diffeomorphism φ has a countable number of cycles. All rational points in $\mathbb{T}^2$ and only such points are part of cycles.*

Proof. 1. Here and in what follows we adopt the convention that when a rational number is written as p/q it is in reduced form, i.e. $\gcd(p,q) = 1$. Consider points of the form $x = (\frac{p_1}{q}, \frac{p_2}{q})$ for an integer q and integers p_1, p_2. Then

$$Ax = \left(\frac{2p_1 + p_2}{q}, \frac{p_1 + p_2}{q} \right), \quad A^{-1}x = \left(\frac{p_1 - p_2}{q}, \frac{-p_1 + 2p_2}{q} \right)$$

Thus, the set of points with denominator q is preserved by φ. There are only finitely many such points in $\mathbb{T}^2$. Thus, $\varphi^m(x) = x$ for sufficiently large m so that x is part of a cycle.

2. Conversely, suppose $\varphi^q(x) = x$ for $x \in \mathbb{T}^2$. Now lift x into $\mathbb{R}^2$. We see that there must exist $m \in \mathbb{Z}^2$ such that

$$A^q(x) = x + m, \quad \text{or} \quad (A^q - I)x = m.$$

Lemma 14 below shows that $\det(A^q - I) \neq 0$. Thus, we may invert the above equation to obtain $x = (A^q - I)^{-1}m$. Further, since A is integer valued, $\det(A^q - I)$ is an integer. Thus, x is rational in $\mathbb{R}^2$ and also $\mathbb{Z}^2$. $\square$

Theorem 67. *The diffeomorphism φ is mixing.*

Proof. 1. We must show that for all measurable sets $F, G, \in \mathcal{B}$

$$\lim_{n\to\infty} |\phi^n(F) \cap G| = |F||G|. \tag{6.2.6}$$

As in Weyl's theorem, we separate the proof into two parts: (i) approximations and measure theory; (ii) a computation. Part (i) allows us to simplify the proof to a calculation with a dense class of functions. Roughly, measurable sets may be approximated by open sets and equation (6.2.6) may be rewritten in terms of the indicator functions of the sets F and G. Indicator functions allow us to approximate any function in $L^1(\mathbb{T}^2)$. Thus, equation (6.2.6) is equivalent to

$$\int_{\mathbb{T}^2} f(\varphi^n(x))g(x)dx = \left(\int f(s)ds \right) \left(\int g(r)dr \right) \tag{6.2.7}$$

for every $f, g \in L^1(\mathbb{T}^2)$.

2. All functions in $L^1(\mathbb{T}^2)$ may be approximated (in L^1) with Fourier series. This is a subtle statement on spaces such as $\mathbb{R}$; however the torus is compact, so every L^2 function is automatically in L^1 too (use the Cauchy-Schwarz inequality). The reason for being so fussy here is that $L^2(\mathbb{T}^2)$ is the 'obvious' space for Fourier series because the functions $e^{2\pi i p x}$, $p \in \mathbb{Z}^2$ constitute an orthonormal basis for $L^2(\mathbb{T}^2)$. On the other hand, L^1 is the natural space for ergodic theorems.

3. This leads us to the actual computation at the heart of the proof. Choose $f(x) = e^{2\pi i \langle p,x\rangle}$ and $g(x) = e^{2\pi \langle q,x\rangle}$ where $p, q \in \mathbb{Z}^2$. Equation (6.2.7) is trivial if either p or $q = 0$, so let us assume both these vectors are non-zero. Then the right hand side of (6.2.7) is zero and we must show that the left hand side vanishes too. By the periodicity of $e^{2\pi i p x}$

$$f(\varphi^n(x)) = e^{2\pi i \langle p, A^n x\rangle} = e^{2\pi i \langle A^n p, x\rangle}.$$

For n large enough, $A^n p \neq q$, so that the left hand side of (6.2.7) vanishes. □

Remark 68. This proof of Theorem 67 demonstrates the application of a powerful analytical method, but it does not convey the underlying intuition. This is discussed in Remark 65. The origin of mixing is stretching by $\lambda_+ > 1$ in the u_+ direction, and contraction by $0 < \lambda_- < 1$ in the u_- direction in a manner that the total volume stays constant. See Figure 6.2.1.

6.3 Structural stability of Anosov's map

An central theme in dynamical systems theory is the stability of dynamical behavior with respect to perturbations. Sometimes the underlying dynamic behavior may be simple; for example, we expect an attracting fixed point to remain attracting if we change the parameters of our system a bit. On the other hand, circle maps and Anosov's map show that systems that are relatively simple to define, may have complex dynamic behavior. Perhaps the most striking feature of Anosov's map is not the fact that it has complex behavior such as the coexistence of infinitely many periodic orbits with dense invariant orbits, but the fact that this behavior is robust to perturbations. This idea is called *structural stability*. Rather than define it precisely, we will illustrate it with an important example.

Theorem 69 (Anosov's theorem)**.** *There exists $\varepsilon > 0$ such that if $B : \mathbb{T}^2 \to \mathbb{T}^2$ is a diffeomorphism satisfying $\|B - A\|_{C^1} < \varepsilon$ then there is a homeomorphism $H : \mathbb{T}^2 \to \mathbb{T}^2$ such that $B = H \circ A \circ H^{-1}$.*

Remark 70. The C^1 norm of a map $f : \mathbb{T}^2 \to \mathbb{T}^2$ is

$$\|f\|_{C^1} = \max_{x \in \mathbb{T}^2} |f(x)| + |Df(x)|.$$

The map H is said to conjugate B to A. Observe that H is a homeomorphism, even though B is assumed to be C^1. This too is a general theme in structural

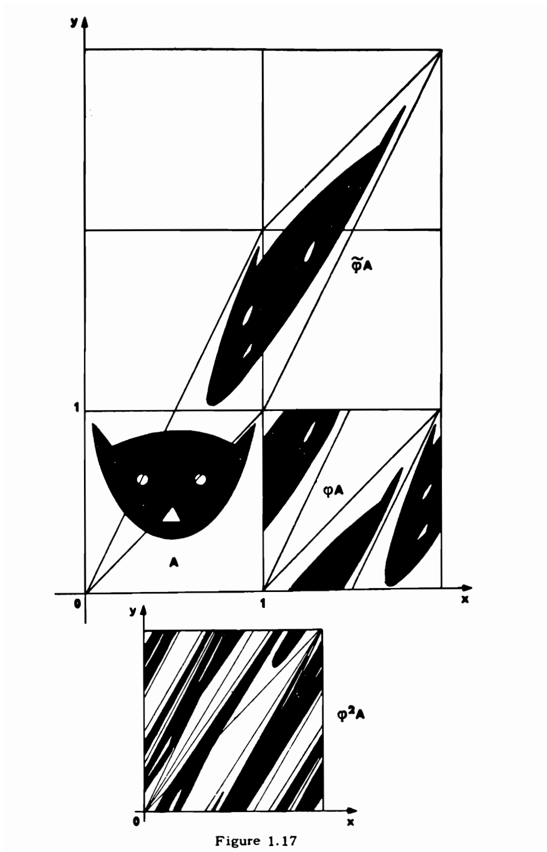

Figure 6.2.1: Mixing in Anosov's map. This image is taken from Arnold and Avez [4]. A minor difference with the text is that the underlying transformation flips the roles of the x_1 and x_2 axis. Our choice is more common.

stability theorems. The choice of topology for the perturbation determines the behavior of the conjugacy.

We will solve the functional equation

$$B \circ H = H \circ A \tag{6.3.1}$$

with a fixed point argument. This functional equation is simplified by working over $\mathbb{R}^2$ instead of $\mathbb{T}^2$. Let us write

$$B(x) = Ax + f(x), \quad H(x) = x + h(x), \tag{6.3.2}$$

where both f and h are $\mathbb{Z}^2$-periodic functions. Then the equation (6.3.1) may be rewritten as

$$h(Ax) - Ah(x) = f(x + h(x)). \tag{6.3.3}$$

We assume that f is given and we must solve for h. Let us Taylor expand the the right hand side so that the first order and second order (in $\|B-A\|_{C^1}$) become clear. We have

$$f(x+h(x)) = f(x) + Df(x)h(x) + O(\|h\|^2). \tag{6.3.4}$$

If $\|B-A\|_{C^1} < \varepsilon$ then both $\|f\|_{C^0}$ and $\|Df\|_{C^0}$ are less than ε. Equation (6.3.3) suggests that $\|h\|_{C^0}$ is of the same order as $\|f\|_{C^0}$. Therefore, $\|Dfh\|_{C^0}$ is $O(\varepsilon^2)$. This suggests that we should first replace (6.3.3) with the linear equation

$$h(Ax) - Ah(x) = f(x). \tag{6.3.5}$$

This is called the *homological equation.* Let $L : C^0(\mathbb{T}^2) \to C^0(\mathbb{T}^2)$ denote the linear operator

$$h \mapsto h \circ A - A \circ h, \tag{6.3.6}$$

so that the homological equation is equivalent to

$$Lh = f. \tag{6.3.7}$$

Composition with the linear transformation A is a bounded linear transformation on C^0 that is easily controlled.

Lemma 15. *Define $S : C^0(\mathbb{T}^2) \to C^0(\mathbb{T}^2)$ by $g \mapsto g \circ A$. Then S is invertible and*

$$\|S\| = \|S^{-1}\| = 1. \tag{6.3.8}$$

Proof. It is clear that S is a linear operator. By definition, the norm of S is

$$\|S\| = \sup_{\|g\|_{C^0}=1} \frac{\|Sg\|_{C^0}}{\|g\|_{C^0}}.$$

On the other hand,

$$\|Sg\|_{C^0} = \max_{x\in\mathbb{T}^2} |g(Ax)| = \max_{x\in\mathbb{T}^2} |g(x)| = \|g\|_{C^0}.$$

Similarly,

$$\|S^{-1}g\|_{C^0} = \max_{x\in\mathbb{T}^2} |g(A^{-1}x)| = \max_{x\in\mathbb{T}^2} |g(x)| = \|g\|_{C^0}.$$

□

The main observation underlying Anosov's theorem is the following

Lemma 16. *The operator $L : C^0(\mathbb{T}^2) \to C^0(\mathbb{T}^2)$ is invertible with*

$$\|L^{-1}\| \le \frac{1}{1-\lambda_-}, \tag{6.3.9}$$

where λ_- is defined in equation (6.2.5).

Proof. Let $U = (u_+, u_-)$ be the matrix of eigenvectors in Lemma 14. Let us express f and h in this basis, writing

$$f = f_+u_+ + f_-u_-, \quad h = h_+u_+ + h_-u_-, \quad A = \lambda_+u_+u_+^T + \lambda_-u_-u_-^T.$$

Then equation (6.3.7) is expressed in coordinates as

$$\begin{aligned} h_+(Ax) - \lambda_+h_+(x) &= f_+(x) &\quad (6.3.10)\\ h_-(Ax) - \lambda_-h_-(x) &= f_-(x). &\quad (6.3.11)\end{aligned}$$

Let $E : C^0(\mathbb{T}^2) \to C^0(\mathbb{T}^2)$ denote the identity operator. We rewrite the above equations using the operator S of Lemma 15 as

$$(S - \lambda_+E)h_+ = f_+, \quad (S - \lambda_-E)h_- = f_-. \tag{6.3.12}$$

Since neither λ_+ nor λ_- lies in the spectrum of S, both these operators may be inverted using the Neumann series. First, since $\lambda_-\lambda_+ = 1$ we have

$$(S - \lambda_+E)^{-1} = \frac{-1}{\lambda_+}(E - \frac{1}{\lambda_+}S)^{-1} = \lambda_-\left(1 + \lambda_-S^{-1} + \lambda_-^2S^{-2} + \dots\right).$$

The infinite series is convergent by Lemma 15 since

$$\left\|\sum_{n=1}^{\infty}\lambda_-^nS^{-n}\right\| \le \sum_{n=1}^{\infty}\lambda_-^n\|S^{-n}\| = \sum_{n=1}^{\infty}\lambda_-^n = \frac{\lambda_-}{1-\lambda_-}. \tag{6.3.13}$$

Similarly, we use equation (6.3.12) to obtain

$$(S - \lambda_-E)^{-1} = S^{-1}(1 - \lambda_-S^{-1})^{-1} = S^{-1}\sum_{n=0}^{\infty}\lambda_-^nS^{-n}, \tag{6.3.14}$$

which is convergent by an argument similar to (6.3.13). We also obtain the bound

$$\|(S - \lambda_-E)^{-1}\| \le \frac{1}{1-\lambda_-}. \tag{6.3.15}$$

Finally, we obtain

$$\|h_-\|^2 + \|h_+\|^2 \le \frac{1}{(1-\lambda_-)^2}\left(\|f_-\|^2 + \lambda_-^2\|f_+\|^2\right) \le \frac{1}{(1-\lambda_-)^2}\left(\|f_-\|^2 + \|f_+\|^2\right),$$

since $0 < \lambda_- < 1$. □

Let us now return to the fixed point equation (6.3.3). We add and subtract $f(x)$ to the RHS and use equation (6.3.7) to rewrite equation (6.3.3) as

$$Lh(x) = f(x) + (f(x + h(x)) - f(x)). \tag{6.3.16}$$

Let $\Phi_f : C^0 \to C^0$ denote the map $h(x) \mapsto f(x + h(x)) - f(x)$. Note that

$$\|\Phi_f(h)\|_{C^0} = \max_{x\in\mathbb{T}^2}|f(x + h(x)) - f(x)| \le \|Df\|_{C^0}\|h\|_{C^0}.$$

We have a solution to (6.3.16) if and only if

$$h = L^{-1}\Phi_f(h) + L^{-1}f.$$

Treat the RHS as a map from C^0 into itself and observe that it is contraction if $\|Df\|_{C^0}$ is small enough. This proves the following

Lemma 17. *There exists $\varepsilon > 0$ such that if $B : \mathbb{T}^2 \to \mathbb{T}^2$ is a diffeomorphism satisfying $\|B - A\|_{C^1}$ then there is a unique $\mathbb{Z}^2$-periodic continuous function $h : \mathbb{R}^2 \to \mathbb{R}^2$ that solves the fixed point equation (6.3.3).*

Lemma 18. *The map $H(x) = x + h(x)$ defines a homeomorphism of the torus.*

Proof. We must show that H is one-to-one and onto.

We will lift the maps to $\mathbb{R}^2$ and use hats to denote these lifts. First, if $H(x) = H(y)$ then since $B \circ H = H \circ A$ we also have $\hat{H}(\hat{A}(\hat{x})) = \hat{H}(\hat{A}(\hat{y}))$. By induction, we also find $\hat{H}(\hat{A}^n\hat{x}) = \hat{H}(\hat{A}^n\hat{y})$. If $\hat{x} \neq \hat{y}$ then $\lim_{n\to\pm\infty} |\hat{A}^n\hat{x} - \hat{A}^n\hat{y}| = +\infty$. This contradicts the boundedness of h (which is bounded on $\mathbb{R}^2$ since it is $\mathbb{Z}^2$ periodic). This must mean that $\hat{x} = \hat{y}$ in $\mathbb{R}^2$, so that $x = y$ on $\mathbb{T}^2$.

The fact that the range of H is all of $\mathbb{T}^2$ is left as an exercise. $\square$

6.4 The Poincare Recurrence Theorem

Ergodic theorems have their origin in subtle paradoxes in the relation between classical mechanics and macroscopic phenomena. The underlying questions is this: do Newton's apply to arbitrarily small particles and if so, how does one scale up this behavior to macroscopic matter (i.e. the scale on which we live)? To this end, we first assume that the physical world is described by finite-dimensional Hamiltonian systems. If so, the following theorem holds.

Theorem 71 (Poincaré recurrence)**.** *Assume $U \subset \mathbb{R}^d$ is bounded and $g : U \to U$ preserves volume and is continuous. Then, for every $x \in U$ and every $\varepsilon > 0$ there exists n such that $g^n(B(x,\varepsilon)) \cap B(x,\varepsilon) \neq \emptyset$.*

Proof. Since U is bounded $\mathrm{vol}(U) < \infty$. Consider the images $A_n := g^n(B(x,\varepsilon))$ of a ball $B(x,\varepsilon) \subset U$. Since g is volume preserving, $\mathrm{Vol}(A_n) = \mathrm{Vol}(B(x,\varepsilon))$. Thus, if A_n were disjoint, we would find that $\sum_{n=1}^{\infty} \mathrm{Vol}(A_n) = \infty$. On the other hand, $\cup_{n=1}^{\infty} A_n \subset U$, so that $\mathrm{Vol} \cup_{n=1}^{\infty} A_n \leq \mathrm{Vol}(U) < \infty$.

It follows that $A_n \cap A_0$ is non-empty for sufficiently large n. $\square$

The connection to Hamiltonian systems is as follows. Assume $\varphi_t : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ is the flow of a Hamiltonian system. Then φ_t is a symplectic diffeomprophism and it preserves volumes (see Corollary 3). In fact, a finer version of this theorem holds: the volume form restricted to a constant energy surface is preserved. In particular, Poincare recurrence holds if $\{z \in \mathbb{R}^{2n} | H(z) = E\}$ is compact.

This theorem also applies to singular limits of Hamiltonian systems. A celebrated example is the *hard sphere gas.* This is a 'minimal' particle system that was introduced in the mid-1800s by Maxwell and Boltzmann to address a fundamental scientific question: why is the macroscopic world so clearly irreversible, when Newton's laws are invariant under time reversal?

The hard sphere gas is a system consisting of N small particles that move freely, except when they meet at collisions, when they exchange momentum in a manner that conserves energy. For simplicity, we ignore boundaries, assume the centers of the particles, $x_i \in \mathbb{T}^d$ and that the radius of each particle is $\delta \ll 1$. We denote the phase space

$$\mathcal{M} = \{(x,y) \in \mathbb{T}^{Nd} \times \mathbb{R}^{Nd}, |x_i - x_j| \geq \delta, i \neq j\}. \tag{6.4.1}$$

The equations of motion consist of free streaming

$$\begin{aligned} \dot{x}_i &= v_i \\ \dot{v}_i &= 0 \end{aligned}$$

when $|x_i - x_j| > \delta$. At the boundary points of $\partial\mathcal{M}$ where exactly one pair of particles meet we impose the "collision rule"

$$\begin{aligned} v_i + v_j &= v_i' + v_j' \\ |v_i|^2 + |v_j|^2 &= |v_i'|^2 + |v_j'|^2. \end{aligned}$$

Here v_i and v_j are the incoming velocities, whereas v_i' and v_j' are the outgoing velocities. There are also boundary points where more than two particles meet; however, this is a measure zero set within the set of all boundary points.

In the homework, you are asked to find v_i', v_j' given by v_i, v_j and to show that the Jacobian of the transformation $(v_i, v_j) \to (v_i', v_j')$ is unity. Thus, at each collision we obtain a measure preserving transformation of the compact energy sphere

$$\mathcal{E} = \{(x,y) \in \mathcal{M} \left| \frac{1}{2N} \sum_{i=1}^{N} |v_i|^2 = E < \infty. \right\}$$

We have normalized the energy by a factor of $1/N$ so that the average energy per particle remain E in the limit $N \to \infty$.

When the Poincare recurrence theorem is applied to this problem, we obtain the following assertion which contradicts our everyday experience. Assume we choose an initial configuration where all of the particles are contained within a small region of space, but such that the initial velocities are random. We expect that as time evolves the particles will become distributed evenly in space, equilibrating in some way. But the Poincaré recurrence theorem tells us that the system must always keep returning arbitrarily close to its initial condition.

This argument is called the Loschmidt paradox. It shows that our naive expectation of irreversible behavior in such a system is false. One of the resolutions of this question relies on a sharp understanding of mixing in measure-preserving transformations and the construction of higher-dimensional analogues of Anosov's construction. A central result of this type is Sinai's proof of the ergodicity of the hard-sphere gas extending Anosov's work on geodesic flow in negatively curved spaces.

6.5 Exercises

1. Complete problems 1 through 6 on p.37 of Arnold's book "Mathematical methods of classical mechanics". These problems culminate in Problem 6. However, the solution to Problem 6 is almost completely described in the hint, so there is no need to turn it in. The treatment of Kepler's problem in Section 5.7 follows [3] very closely.

2. Consider the collision rule in the hard-sphere gas. Assume given two 'input' velocity vectors u,$v \in \mathbb{R}^d$ and impose the conditions of conservation of momentum and energy at a collision:

$$u' + v' = u + v, \quad |u'|^2 + |v'|^2 = |u|^2 + |v|^2.$$

(a) Show that these conditions determine two 'output' vectors u' and $v' \in \mathbb{R}^d$ that are unique upto permutation.

(b) Compute the Jacobian of the transformation from (u, v) to (u', v').

3. Consider the Gauss map $G : [0, 1) \to [0, 1)$ defined by

$$G(x) = \frac{1}{x} - \text{floor}\left(\frac{1}{x}\right).$$

Show that the probability density

$$p(x) = \frac{1}{\log 2} \frac{1}{1+x}$$

is invariant under G.

6.6 Solutions to exercises

2. Consider the collision rule in the hard-sphere gas. Assume given two 'input' velocity vectors u,$v \in \mathbb{R}^d$ and impose the conditions of conservation of momentum and energy at a collision:

$$u' + v' = u + v, \quad |u'|^2 + |v'|^2 = |u|^2 + |v|^2.$$

(a) Show that these conditions determine two 'output' vectors u' and $v' \in \mathbb{R}^d$ that are unique upto permutation.

(b) Compute the Jacobian of the transformation from (u, v) to (u', v').

Proof. (a) The trick in this problem is to recognize that in an elastic collision between two identical spheres with radius δ there are three vectors in play: the input velocities u and v and the unit vector $l \in S^{d-1}$ along the line joining the centers of the two spheres.

Let's build some intuition for the process of collision. Assume at first that the spheres have a head-on collision. This means that the vectors u,v and l are all parallel. Since the spheres are identical, they simply exchange velocities and $u' = v$ and $v' = u$. On the other hand, if the spheres have a glancing collision, that is u and v are parallel, but both u and v are perpendicular to l, then there is no exchange of velocity, so $u' = u$ and $v' = v$.

The general situation can be decomposed into these two extreme cases. The particles exchange the head-on component of their velocity and they retain the glancing component of the velocity. We separate the two components to obtain the relation

$$u' = u - ((u-v)\cdot l)\, l, \quad v' = v + ((u-v)\cdot l)\, l.$$

(b) Given a unit vector $l \in \mathbb{R}^d$, the rank-one matrix ll^T is the orthogonal projection onto the span of l and the matrix $I_d - ll^T$ is the orthogonal projection onto its complement. Let $L : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ denote the map $(u,v) \mapsto (u', v')$. Then we see that

$$L = \begin{pmatrix} I_d - ll^T & ll^T \\ ll^T & I_d - ll^T \end{pmatrix} = I_{2d} - ww^T, \quad w := \begin{pmatrix} -l \\ l \end{pmatrix}.$$

This allows a direct computation of the determinant using the Sherman-Morrison-Woodbury formula

$$\det(L) = \det(I_{2d} - ww^T) = (1 - w^T w)\det(I_{2d}) = -1, \quad \text{since} \quad w^T w = 2.$$

□

3. Consider the Gauss map $G : [0,1) \to [0,1)$ defined by

$$G(x) = \frac{1}{x} - \text{floor}\left(\frac{1}{x}\right).$$

Show that the probability density

$$p(x) = \frac{1}{\log 2}\frac{1}{1+x}$$

is invariant under G.

Proof. Let μ denote the measure with density p. We must show that $\mu(G^{-1}(A)) = \mu(A)$ for every Borel set $A \subset [0,1)$. Since Borel sets may be approximated with open sets, which in turn may be approximated by intervals, it is enough to prove invariance when $A = (a,b)$ is an interval contained within $[0,1)$.

Let $k \in \mathbb{N}$ index the natural numbers. The transformation G maps each interval $[\frac{1}{k+1}, \frac{1}{k})$ to the interval $[0,1)$. Thus, the pre-image $G^{-1}(a,b)$ consists of a countable collection of disjoint intervals

$$\bigcup_{k=1}^{\infty} (x_k, y_k), \quad x_k = \frac{1}{b+k}, \quad y_k = \frac{1}{a+k}.$$

Therefore, the measure of the inverse image is

$$\mu(G^{-1}(a,b)) = \frac{1}{\log 2}\sum_{k=1}^{\infty}\int_{x_k}^{y_k}\frac{ds}{1+s} = \frac{1}{\log 2}\sum_{k=1}^{\infty}\log\left(\frac{1+\frac{1}{a+k}}{1+\frac{1}{b+k}}\right).$$

The infinite sum is convergent since p is a probability measure. We may rearrange the terms, recognizing that it is a telescoping sum with value

$$\frac{1}{\log 2}\log\left(\frac{1+b}{1+a}\right).$$

But this is exactly the measure $\mu(a,b)$. □

Problem 1, p. 37, Arnold. This problem is straightforward. Follow the hint and substitute $x = M/r$ into the integral on. p.36 to obtain

$$\Phi = \int_{x_{\min}}^{x_{\max}} \frac{dx}{\sqrt{2(E-W)}}.$$

□

Problem 2, p. 37, Arnold. Figure 31 provides the essential hint. Let r_* denote the point where $V(r)$ is at its minimum. Then for r close to r_*

$$V(r) \approx V(r_*) + \frac{1}{2}V''(r_*)(r-r_*)^2.$$

For brevity, ler $r_{\max} - r_{\min} = 2a$ and $r - r_{\min} = s$. Since $V(r_{\min}) = V(r_{\max}) = E$, we may also write

$$E - V(r) = V(r_{\max}) - V(r) = \frac{1}{2}V''(r_*)(a^2 - s^2).$$

We substitute this expression in the formula for Φ on p.35 to obtain

$$\begin{aligned}\Phi &\approx \int_{r_{\min}}^{r_{\max}} \frac{M}{r^2}\frac{dr}{\left(V''(r_*)(a^2-(r-r_*)^2)\right)^{1/2}} \\ &\approx \frac{M}{r_*^2\sqrt{V''(r_*)}}\int_{-a}^{a}\frac{ds}{\sqrt{a^2-s^2}} = \pi\frac{M}{r_*^2\sqrt{V''(r_*)}}.\end{aligned}$$

On the other hand, $V(r) = U(r) + M^2/2r^2$ and $V'(r_*) = 0$. Therefore, $U'(r_*) = M^2/r^3$. Thus, a couple of lines of algebra yields

$$\frac{M}{r_*^2\sqrt{V''(r_*)}} = \sqrt{\frac{U'(r_*)}{r_*U''(r_*) + 3U'(r_*)}}.$$

□

Problem 3, p. 37, Arnold. Following Arnold, let us use r instead of r_*. The angle Φ is independent of r for circular orbits when

$$\frac{U'(r)}{rU''(r) + 3U'(r)}$$

is a constant. We rewrite this equation in the form

$$\frac{rU''}{U'} = r(\log U')' = \alpha - 1,$$

(This choice of notation for the constant is only for consistency with the answer in Arnold.) We integrate the above differential equation to find that

$$U(r) = ar^{\alpha}, \quad \alpha \neq 0 \quad \text{and} \quad U(r) = b \log r, \quad \alpha = 0.$$

The condition $\alpha \geq -2$ is imposed by the restriction that the rotational kinetic energy $M^2/2r^2$ dominates $U(r)$ as $r \to 0$ (see p. 34). $\square$

Problem 4, p.37, Arnold. Since $U(r) \to \infty$ as $r \to \infty$, it is either $U(r) = ar^{\alpha}$ with $\alpha > 0$ or $U(r) = b \log r$. The maximum value of x is given by

$$E = W(x_{\max}) = \frac{1}{2}x_{\max}^2 + U(\frac{M}{x_{\max}}).$$

Then, as $E \to \infty$, $x_{\max} \sim \sqrt{2E}$, so that $x_{\max} \to \infty$. We now make the suggested change of variable $x = yx_{\max}$ to obtain

$$\Phi = \int_{y_{\min}}^{1} \frac{dy}{\sqrt{2(W^*(1) - W^*(y))}}, \quad W^*(y) = \frac{y^2}{2} + \frac{1}{x_{\max}^2} U\left(\frac{M}{yx_{\max}}\right).$$

The value of $x_{\min}$ (and thus $y_{\min}$) is determined by

$$E = \frac{1}{2}x_{\min}^2 + U\left(\frac{M}{x_{\min}}\right).$$

Since $U(r) \to \infty$ as $r \to \infty$, when $E \to \infty$ we find that $x_{\min} = Ma^{\alpha}E^{-\alpha}$ or $x_{\min} = Me^{E/b}$ depending on whether $U(r) = ar^{\alpha}$ or $U(r) = b\log r$. In either case, $x_{\min} \to 0$ as $E \to \infty$. We now let $E \to \infty$ and interchange the limits in the integral to obtain

$$\Phi = \int_0^1 \frac{dy}{\sqrt{1-y^2}} = \frac{\pi}{2}.$$

$\square$

Problem 5, p.37, Arnold. Assume that $U(r) = kr^{-\beta}$ with $k > 0$ and $0 < \beta < 2$. Consider the energy level with $E = 0$. Then $W(x) = E$ if and only if

$$0 = k\frac{x^\beta}{M^\beta} - \frac{x^2}{2}.$$

This equation has two solutions

$$x_{\min} = 0, \quad x_{\max} = \left(\frac{2k}{M^\beta}\right)^{\frac{1}{2-\beta}}.$$

The angle along this orbit is

$$\Phi_0 = \int_{x_{\min}}^{x_{\max}} \frac{dx}{\sqrt{-2W}} = \int_0^1 \frac{ds}{\sqrt{s^\beta - s^2}} = \frac{\pi}{2-\beta}.$$

Here we used the fact that the second integral is a standard integral that can be found in tables of integrals, as well as the substitution $x = x_{\max}s$ along with the above formula for $x_{\max}$.

□

Chapter 7

Hyperbolicity

A fundamental idea in dynamical systems is the "persistence of hyperbolic structures". We have encountered an example of this idea in our proof of Anosov's theorem (Theorem 69). In this chapter, we explore this idea systematically for flows and maps.

7.1 Hyperbolicity in Maps

Assume $U \subset \mathbb{R}^d$ is an open set. Every smooth map $f : U \to U$ defines a *discrete dynamical system*. The orbit of a point $x_0 \in U$ is the sequence of iterates

$$x_{n+1} = f(x_n), \quad n \geq 0. \tag{7.1.1}$$

We denote k-fold composition by the following notation

$$f^k = f \circ f \circ \ldots \circ f \quad k - \text{times}.$$

Thus, equation (7.1.1) implies

$$x_n = f^n(x_0), \quad n \geq 0. \tag{7.1.2}$$

When f is a diffeomorphism this dynamical system is well-defined for all $n \in \mathbb{Z}$. However, several interesting maps, especially some measure preserving transformations, are not invertible. Our goal is to introduce the concept of hyperbolic fixed points. To this end, let us begin with the simplest class of maps: invertible linear transformations of $\mathbb{R}^d$.

Suppose $U = \mathbb{R}^d$ and $f(x) = Ax$ where A is an invertible matrix. Then $x_n = A^n x$, and the asymptotics as $n \to \infty$ are determined by the spectrum of A. Assume A is diagonalizable and $A = U \Lambda U^{-1}$ where

$$\Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \tag{7.1.3}$$

is the matrix of eigenvalues. We divide the eigenvalues into 3 subsets:

(a) Stable: all λ's such that $|\lambda_i| < 1$.

(b) Unstable: all λ's such that $|\lambda_i| > 1$.

(c) Center: all λ's such that $|\lambda_i| = 1$.

This classification reflects the fact that

$$\begin{array}{lll} \text{If } |\lambda_i| < 1 & \text{then } \lim_{n\to\infty} |\lambda_i|^n = 0. & \\ \text{If } |\lambda_i| > 1 & \text{then } \lim_{n\to\infty} |\lambda_i|^n = +\infty. & \\ \text{If } |\lambda_i| = 1 & \text{then } |\lambda_i|^n = 1 \quad \text{for all} \quad n. \end{array}$$

The concept of hyperbolicity is introduced to rule out the borderline case between stability and instability.

Definition 72. A fixed point x_* of a C^1 map $f : U \to U$ is *hyperbolic* if $Df(x_*)$ has no eigenvalues on the unit circle.

Let us now extend this concept to cycles.

Definition 73. An orbit $\{x_0, x_1, \dots x_{q-1}\}$ is a *cycle* of length q if $x_0 = x_q$ and $x_0 \neq x_n$ for $1 \leq n \leq q-1$.

We note that if f defines a cycle of length q then $x_0 = f^q(x_0)$. Thus, x_0 is a fixed point of f^q.

Definition 74. The cycle $\{x_0, x_1, \dots x_{q-1}\}$ is *hyperbolic* if each x_k is a hyperbolic fixed point of f^q.

Remark 75. The linearization around a cycle has an interesting structure. By the chain rule

$$\begin{aligned} Df^q(x_0) &= Df(f^{q-1}(x_0)) \cdot Df(f^{q-2}(x_0)) \dots Df(x_0) \\ &= Df(x_{q-1}) \cdot Df(x_{q-2}) \dots Df(x_0) \\ &\stackrel{def}{=} A_{q-1}A_{q-2} \dots A_0, \end{aligned}$$

where we introduced the notation A_j to make the structure of the formula clear. Similarly, since $x_0 = x_q$ we also have

$$Df^q(x_1) = A_0 A_{q-1} A_{q-2} \dots A_1.$$

Proceeding inductively, we find that

$$Df^q(x_k) = A_{k-1} \dots A_0 A_{q-1} \dots A_k.$$

We *cannot* assume that the matrices commute.

7.2 Hyperbolicity in Flows

Let us now extend these ideas to flows. Consider the flow φ_t defined by an ODE $\dot{x} = g(x)$. As our first example, let $g(x)$ be linear, so that we have the equation

$$\dot{x} = Bx, \quad x \in \mathbb{R}^d. \tag{7.2.1}$$

Assume that B is diagonalizable with diagonalization $B = U\Lambda U^{-1}$, so that

$$x(t) = e^{tB} = U \begin{pmatrix} e^{t\lambda_1} & & \\ & \ddots & \\ & & e^{t\lambda_n} \end{pmatrix} U^{-1} x_0. \tag{7.2.2}$$

The role of the unit circle (for maps) is now replaced by the imaginary axis. Let

$$\lambda = \alpha + i\beta, \quad i = \sqrt{-1}.$$

Then

$$|e^{t\lambda}| = e^{t\mathrm{Re}(\lambda)} \qquad \text{(for real t)} = e^{t\alpha},$$

and we have

$$\lim_{t\to\infty} \left|e^{t\lambda}\right| = \begin{cases} 0 & \text{if } \mathrm{Re}(\lambda) < 0, \\ +\infty & \text{if } \mathrm{Re}(\lambda) > 0, \\ 1 & \text{if } \mathrm{Re}(\lambda) = 0. \end{cases}$$

Definition 76. A fixed point x_* for $\dot{x} = g(x)$ is *hyperbolic* if $Dg(x_*)$ has no eigenvalues on the imaginary axis.

7.2.1 Periodic Orbits

In order to determine the persistence of periodic orbits under perturbations we must extend the criterion of hyperbolicity to periodic orbits. This requires a new concept: the Poincaré map. Consider a periodic orbit Γ with period $T > 0$. At any point $x_0 \in \Gamma$ we define a section S transverse to the tangent vector τ to Γ at x_0 (see Figure 7.2.1). Transversality means that τ does not lie in the section S (in $\mathbb{R}^d$ one may always choose S to be the hyperplane orthogonal to τ at S).

Since Γ is periodic with period T, we have $\varphi_T(x_0) = x_0$. By continuity in initial conditions, for all $x \in S$ that are sufficiently close to x_0, say within the region

$$D = S \cap B_\varepsilon(x_0),$$

there is a well-defined *first-return time* $T(x)$, such that $\varphi_{T(x)} = x$. The Poincaré map at x_0 is the map

$$P_{x_0} : D \to D, \quad x \mapsto \varphi_{T(x)}(x). \tag{7.2.3}$$

A proof of the existence and regularity of the Poincaré map is outlined in the homework. The main advantage of the Poincaré map is that it reduces the

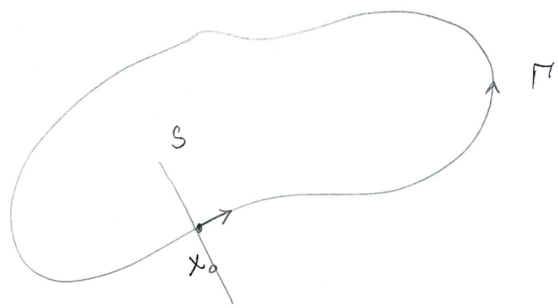

Figure 7.2.1: Periodic Orbit

question of persistence of periodic orbits, which is a global question, to the persistence of fixed points for the Poincaré map, which is a local question.

Informally, a periodic orbit is hyperbolic if and only if the Poincaré map is hyperbolic. This reduces hyperbolicity of periodic orbits for flows to the analogous concept for maps. The weakness in the above definition is that we must show that it does not depend on the choice of section S or initial point x_0. For these reasons, we return to the linearization of the ODE $\dot{x} = g(x)$ with the above intuition.

Assume $x_*(t)$ is a periodic orbit with period $T > 0$ for $\dot{x} = g(x)$. The linearization about x_* is

$$\dot{u} = Dg(x_*(t))u. \tag{7.2.4}$$

To simplify notation, let us write this equation in this form:

$$\dot{u} = B(t)u \quad ; \quad B(t+T) = B(t). \tag{7.2.5}$$

i.e. B is a periodic function of t. (We assume $T > 0$ to prevent trivialities).

Denote the fundamental solution to 7.2.5 as $Y(t)$. Then, $Y(t)$ solves the (matrix) equation

$$Y(t) = B(t)Y, \quad Y(0) = I. \tag{7.2.6}$$

Definition 77. The *Floquet matrix* for the periodic orbit $x_*(t)$ with period $T > 0$ is $Y(T)$ where Y solves equation (7.2.6) with $B(t+T) = B(t)$.

Definition 78. The periodic orbit Γ is *hyperbolic* if the Floquet matrix has only one eigenvalue on the unit circle.

Here's what's going on: the Floquet matrix always has 1 as a trivial eigenvalue. Let

$$Y(T) = U \begin{pmatrix} 1 & \\ & \begin{bmatrix} \text{non} \\ \text{trivial} \end{bmatrix} \end{pmatrix} U^{-1} \tag{7.2.7}$$

Once one removes the trivial eigenvalue at 1, the rest of the spectrum of $Y(T)$ is exactly the spectrum of the linearization of the Poincaré map.

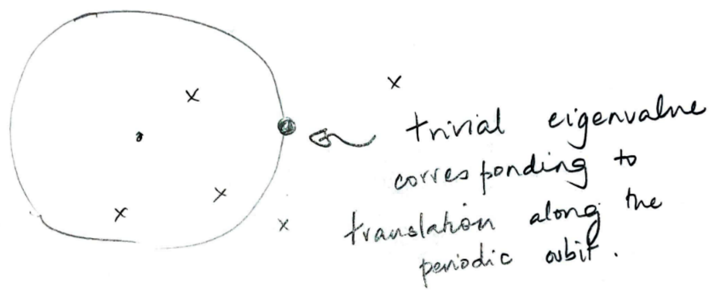

Figure 7.2.2: Floquet spectrum for a hyperbolic orbit

The "metatheorem" of hyperbolic dynamical systesm is that "hyperbolic structures persist under small perturbations". Here are some examples:

(i) Persistence of hyperbolic fixed points for maps and flows.

(ii) Persistence of hyperbolic periodic orbits.

(iii) Persistence of a hpyerbolic foliation (Anosov's theorem).

(iv) Stable and unstable manifold theorems.

The key assumption in all these theorems are

(a) A spectral gap between stable and unstable directions.

(b) A careful choice of topology for perturbation.

We will first illustrate these ideas for fixed points. We then consider invariant manifold theorems in Chapter 8.

7.3 Persistence of hyperbolic fixed points

Theorem 79. *Assume $g(x;\mu)$ is a C^1 vector field on $U \subset \mathbb{R}^d$ that depends smoothly on a parameter $\mu \in (-1,1)$. Suppose $g(x_*;0) = 0$ and x_* is hyperbolic. Then there exits $\varepsilon > 0$ and a C^1 curve of hyperbolic fixed points $x(\mu)$ for $\mu \in (-\varepsilon, \varepsilon)$.*

Proof. Our assumption is that $Dg(x_*, 0)$ has no eigenvalues on the imaginary axis. In particular it is invertible. By the implicit function theorem, there exists $\varepsilon > 0$ and a map $(\varepsilon, \varepsilon) \to \mathbb{R}^d$, $\mu \mapsto x(\mu)$ with $x(0) = x_*$ such that

$$g(x(\mu);\mu) = 0. \tag{7.3.1}$$

The smoothness of the map $\mu \mapsto x(\mu)$ is the same as that of the map g. $\square$

Remark 80. Here is the intuition behind the proof. Suppose equation (7.3.1) holds. Then differentiate it with respect to μ to find

$$Dg(x(\mu);\mu)\frac{dx}{d\mu} + \frac{\partial g}{\partial \mu} = 0.$$

When $\mu = 0$ we know that $Dg(x_*, 0)$. Therefore, we can solve for

$$\frac{dx}{d\mu} = -Dg(x(\mu);\mu)^{-1}\frac{\partial g}{\partial \mu}, \tag{7.3.2}$$

where $\frac{\partial g}{\partial \mu}$ is known in a neighbourhood of $(x_*, 0)$. The flaw in this argument is that we don't know that this curve exists without the implicit function theorem. But once existence of the curve has been obtained, we can determine the dependence of x_* on μ through equation (7.3.2)

The map $x \to x(\mu)$ is as smooth as g. If g is C^1, then so is $\mu \to x(\mu)$ and if g is C^k then so is $\mu \to x(\mu)$. Consequently the map $\mu \to Dg(x(\mu), \mu)$ is C^{k-1} when g is C^k. Thus, the eigenvalues change continuously and the spectral gap persists for sufficiently small ϵ.

Example 11. *The eigenvalues cannot be assumed to vary smoothly, even if the map $\mu \to Dg(x(\mu), \mu)$ is as smooth as desired (say C^∞). The problem is that when $Dg(x_*, 0)$ has repeated eigenvalues, a small perturbation can the situation depicted in 7.3.1.*

The persistence of hyperbolic fixed points for *maps* is similar. We must now solve the fixed point equation $x = f(x;\mu)$ given $x_* = f(x;0)$ and x_* hyperbolic. We may reduce this problem to Theorem 79 by writing the fixed equation as $F(x;\mu) \stackrel{def}{=} f(x;\mu) - x$ and then applying the implicit function theorem.

Here are some examples of what could go wrong.

Example 12. ***Simple Pendulum:*** *The linearization of the flow at two distinct fixed points are $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. The corresponding eigenvalues are $\pm\sqrt{-1}$ and ± 1, respectively.*

The dynamics of the simple pendulum are defined by:

$$\ddot{\theta} + \sin\theta = 0. \tag{7.3.3}$$

With the simple pendulum we may perturb by adding damping, resulting in the following equation:

$$\ddot{\theta} + \alpha\dot{\theta} + \sin\theta = 0. \tag{7.3.4}$$

Now the linearization at $(0, 0)$ is $\begin{bmatrix} 0 & 1 \\ -1 & -\alpha \end{bmatrix}$. Therefore, the characteristic polynomial for this matrix is

$$\lambda(\lambda + \alpha) + 1 = 0 \tag{7.3.5}$$

$$\lambda^2 + \alpha\lambda + 1 = 0 \tag{7.3.6}$$

$$\implies \lambda = \frac{-\alpha \pm \sqrt{\alpha^2 - 1}}{2} \tag{7.3.7}$$

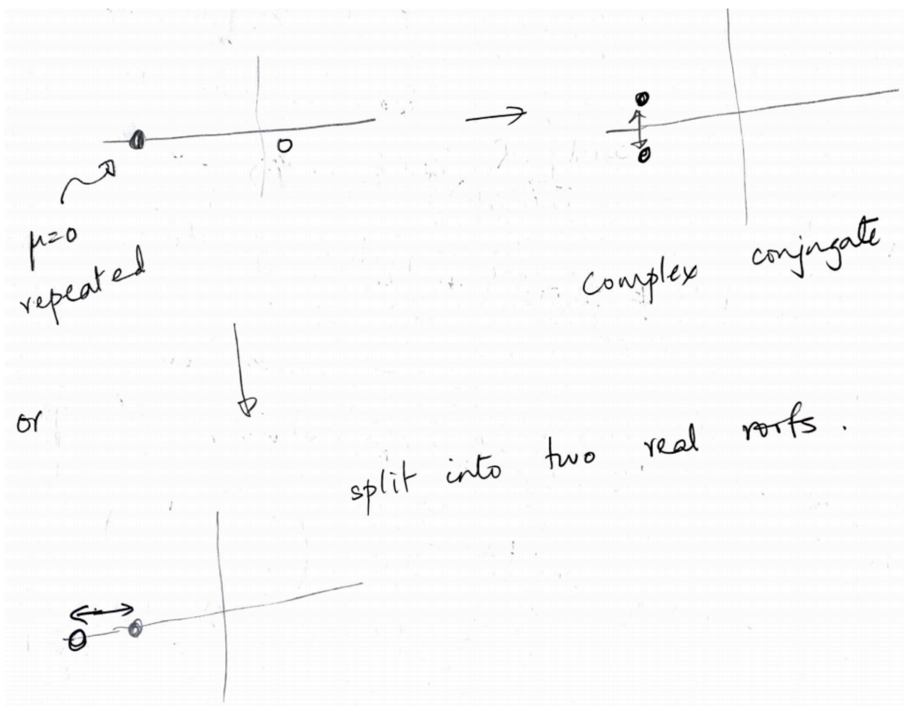

Figure 7.3.1: Continuous, but not differentiable, variation of eigenvalues with parameters.

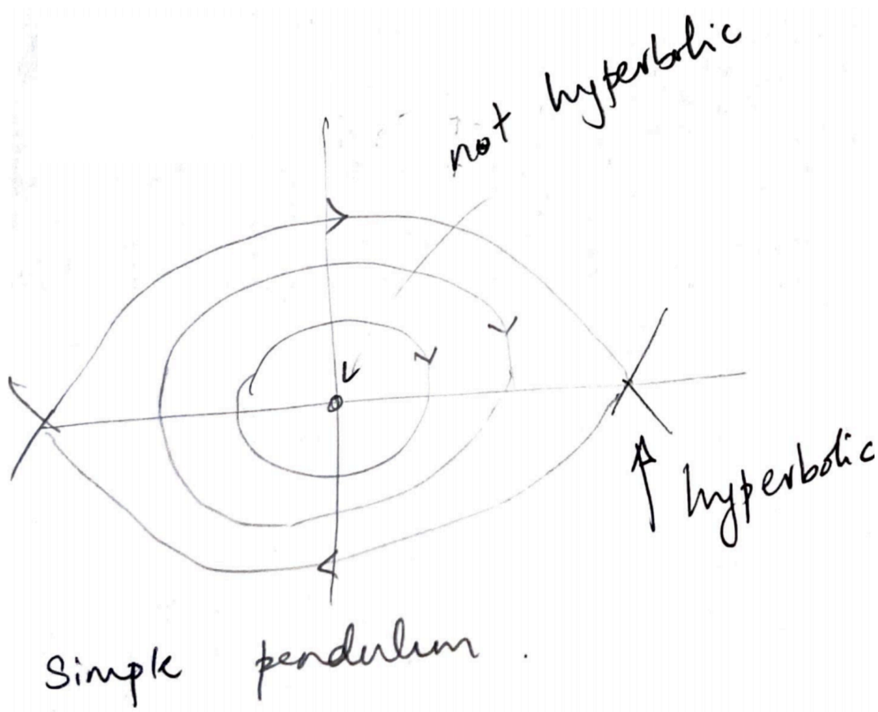

Figure 7.3.2:

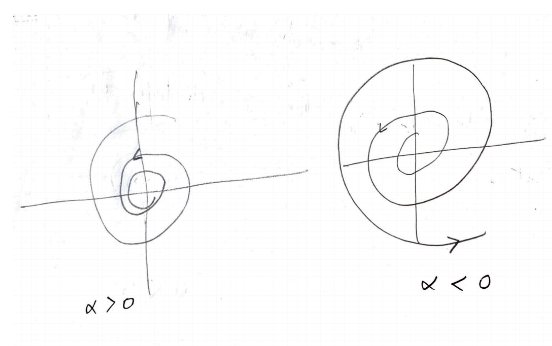

Figure 7.3.3: Phase diagrams for the perturbed simple pendulum

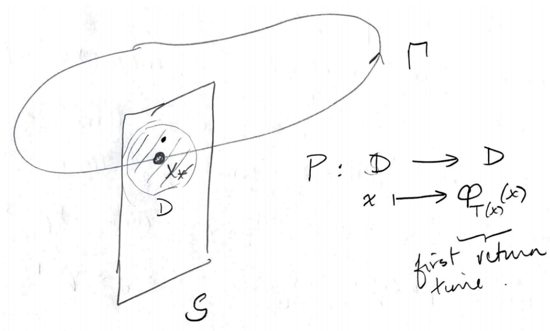

Figure 7.4.1: Phase diagram for the perturbed simple pendulum

The phase diagram in the neighbourhood of $(0,0)$ *for this perturbed system is shown in Figure 7.3.3.*

A more subtle issue here is that the perturbations don't respect the Hamiltonian structure. Really the question is: if we understand the flow for $\dot{z} = J\nabla_z H_o$ then what can we say about $\dot{z} = J\nabla_z H_\mu$ such that

$$H_\mu = H_o + \mu H_1, \tag{7.3.8}$$

where H_1 is the perturbation. In this case, the origin perturbs to a center. This example shows that the topology of the perturbation is important.

7.4 Persistence of Hyperbolic Periodic Orbits

7.4.1 Persistence of Cycles

We now turn to cycles in maps and periodic orbits in flows. Cycles are easy to deal with.

Consider the map $x \mapsto f(x;\mu)$ and assume that when $\mu = 0$, we have a hyperbolic cycle of period q. Denote this cycle by $\{x_0, x_1, ..., x_{q-1}\}$ with $x_q = x_0$. We observe that $f^q(x_j) = x_j$, $0 \le j \le q-1$. This leads us to the following observation, a cycle is hyperbolic $\iff$ x_j is hyperbolic for $0 \le j \le q-1$. But then the implicit function theorem may be used as in Theorem 79 to show that $x_j(\mu)$ persists for small μ.

7.4.2 Persistence of hyperbolic periodic orbits

We know that a periodic orbit Γ is hyperbolic if and only if its Floquet spectrum has a single eigenvalue at 1. This in turn is true if and only if every Poincaré map on Γ has a hyperbolic fixed point. When $\mu = 0$, we have the picture for the periodic orbit Γ as in 7.4.1. Further, we know that x_* is a hyperbolic fixed point.

Now, we observe that if φ depends smoothly on a parameter μ, then we obtain a smooth family of Poincaré maps

$$P_\mu : D \to D, \tag{7.4.1}$$

simply by continuity in parameters (this is illustrated in homework 1). Then we find, from the implicit function theorem, that P_μ has a hyperbolic fixed point $x(\mu)$ for $|\mu| < \epsilon$. Again the simple pendulum with damping shows that the theorem is false without assumption of hyperbolicity.

7.4.3 The Grobman-Hartman Theorem

The above example shows the persistence of a global structure (periodic orbits). We reconsider persistence of hyperbolic fixed points from this point of view. Consider a linear map $A : \mathbb{R}^d \to \mathbb{R}^d$ where the map is $x \to Ax$ such that 0 is a hyperbolic fixed point. The eigenspaces of A form invariant subspaces for the map. We next consider a family of non-linear maps $B(\mu)$ with $B(0;0) = 0$ and $DB(0;0) = A$. The following theorem provides the persistence of the *phase portrait* of the map A near 0.

Theorem 81 (Grobman-Hartman)**.** *Assume $U \in \mathbb{R}^d$ is an open set containing the origin. Assume the function f*

$$\begin{aligned} f :& U \times (-1,1) \to U \qquad (7.4.2)\\ &(x,\mu) \to f_\mu(x), \qquad (7.4.3)\end{aligned}$$

is a 1-parameter family of C^1 diffeomorphisms such that $f(0,\mu) = 0$ for all μ and $x = 0$ is hyperbolic for $\mu = 0$. Then there exists $\varepsilon > 0$ and a 1-parameter family of homeomorphisms

$$\begin{aligned} h :& B(0,\varepsilon) \times (-\varepsilon,\varepsilon) \to B(0,\varepsilon) \qquad (7.4.4)\\ &(x,\mu) \to h_\mu(x) \qquad (7.4.5)\end{aligned}$$

such that the following diagram commutes

$$\begin{array}{ccc} B(0,\varepsilon) & \xrightarrow{f_\mu} & B(0,\varepsilon) \\ h_\mu \uparrow & & \uparrow h_\mu \\ B(0,\varepsilon) & \xrightarrow{A} & B(0,\varepsilon) \end{array}$$

with $A = D_x f_0(0)$

Chapter 8

Invariant Manifold Theorems

The main reference for this chapter is [6, Ch.4.1]. Let us first illustrate the idea of an invariant manifold with an example from [9]. Consider the 2D system

$$\dot{x} = x \tag{8.0.1}$$
$$\dot{y} = -y + x^3. \tag{8.0.2}$$

This system has a fixed point at $(0,0)$, and its linearization at $(0,0)$ is

$$\dot{u} = u \tag{8.0.3}$$
$$\dot{v} = v, \tag{8.0.4}$$

or equivalently

$$\dot{\begin{pmatrix} u \\ v \end{pmatrix}} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}. \tag{8.0.5}$$

Thus $(0,0)$ is a saddle-point. The subspace $\{u = 0\}$ is invariant under the flow and as $t \to \infty$, each trajectory $(0, v(t)) \to (0,0)$. This subspace is the *stable subspace.* Similarly, the subspace $\{v = 0\}$ is invariant and is called the *unstable subspace.* Equation (8.0.1) is chosen so it is exactly solvable. Clearly

$$x(t) = e^t x,$$

where we use the slight abuse of notation of writing x for $x(0)$ and y for $y(0)$. Using the method of integrating factors, we can also solve equation (8.0.2)

$$\begin{aligned} y(t) &= e^{-t}y + \int_0^t e^{-(t+s)}x^3(s)ds && (8.0.6)\\ &= e^{-t}y + e^{-t}\int_0^t e^{4s}x^3 ds && (8.0.7)\\ &= e^{-t}y + x^3e^{-t}\int_0^t e^{4s}ds && (8.0.8)\\ &= e^{-t}y + x^3e^{-t}\left(\frac{e^{4t}-1}{4}\right) && (8.0.9)\\ &= e^{-t}y + \left(\frac{e^{3t}-e^{-t}}{4}\right)x^3 && (8.0.10)\\ &= e^{-t}(y-\frac{x^3}{4}) + \frac{x^3}{4}e^{3t}. && (8.0.11)\end{aligned}$$

We now look for the nonlinear analogue of the linear phase portrait. We note that a trajectory $z(t) = (x(t), y(t))$ lies on the stable subspace if and only if $z(t) \to 0$ as $t \to \infty$. Similarly, $z(t)$ lies on the unstable subspace if and only if $z(t) \to 0$ as $t \to -\infty$. We use these asymptotic properties, to define find nonlinear analogues of the stable and unstable spaces. Let us define the sets W_s and W_u respectively to consist of $z_0 \in \mathbb{R}^2$ such that the trajectory $z(t)$ with $z(0) = z_0$ tends to 0 as $t \to +\infty$ and $t \to -\infty$ respectively. From the solution formula

$$x(t) = e^t x, \quad y(t) = e^{-t}(y - \frac{x^3}{4}) + \frac{x^3}{4}e^t \qquad (8.0.12)$$

we see that:

- $z(t) \to 0$ as $t \to \infty$ if and only if $x = 0$; and
- $z(t) \to 0$ as $t \to -\infty$ if and only if $y = \frac{x^3}{4}$.

Thus, we have found that the stable set W_s is actually the manifold $\{x = 0\}$ and that the unstable set W_u is the manifold

$$W_u = \left\{ z \in \mathbb{R}^2 \,\middle|\, y = \frac{x^3}{4} \right\}$$

Observe that these manifolds are tangent to the stable and unstable spaces at $z = 0$. The stable and unstable manifold theorems formalize this intuition for hyperbolic fixed points. As in Picard's theorems, we will first establish the theorem under strong global hypotheses, then obtain local versions using cut-off functions.

8.1 Preliminaries

The main assumption in these theorems is the existence of a *spectral gap*. We consider equations of the form

$$\dot{x} = Sx + F(x, y) \tag{8.1.1}$$
$$\dot{y} = Uy + G(x, y), \tag{8.1.2}$$

where $x \in \mathbb{R}^k$, $y \in \mathbb{R}^l$. We may make an affine change of variables to reduce a given vector field to this form. For brevity, we set $z = (x, y)$, writing $F(z)$ and $G(z)$ for $F(x, y)$ and $G(x, y)$ when this helps.

The matrices S (for stable) for stable and U (for unstable) are assumed to satisfy

$$\mathrm{Re}\ \sigma(S) < 0 \tag{8.1.3}$$
$$\mathrm{Re}\ \sigma(U) > 0. \tag{8.1.4}$$

Here $\sigma(M) = \{\lambda_1, \cdots, \lambda_n\}$ when M is an $n \times n$ matrix, and we write Re $\sigma(M) < a$ if and only if Re $\lambda_i < a$ for $1 \leq i \leq n$.

8.1.1 Manifolds

Despite the terminology, almost all we need of manifold theory is the fact that graphs of smooth functions are also (abstractly defined) manifolds. Given a function $\alpha : \mathbb{R}^k \to \mathbb{R}^l$, its graph is the set

$$W_\alpha = \{(x, y) \in \mathbb{R}^k \times \mathbb{R}^l | y = \alpha(x)\}. \tag{8.1.5}$$

When $\alpha \in C^\infty$, W_α is a C^∞ manifold; analogously, W_α is a C^k manifold when $\alpha \in C^k$, and W_α is a Lipshcitz manifold when α is Lipschitz.

Intuitively, a manifold is a space that locally looks like Euclidean space. In the case of graphs, this is obtained by the above parameterization.

8.1.2 Linear Estimates

Lemma 19. *Assume Re $\sigma(M) \leq a$. Then for every $\lambda > a$, there exists a K_λ such that $\|e^{tM}\| \leq K_\lambda e^{\lambda t}$ for $t \geq 0$.*

Proof. We use the Jordan decomposition over $\mathbb{C}$, writing M in the form $M = UAU^{-1}$ where the matrix A is block diagonal with

$$A = \begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_m \end{pmatrix},$$

where each A_k is either diagonal or of the form

$$A_k = \begin{pmatrix} \alpha_k & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \alpha_k \end{pmatrix}.$$

Given the $m \times m$ matrix

$$A = \begin{pmatrix} \alpha & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \alpha \end{pmatrix}, \tag{8.1.6}$$

we can write

$$e^{tA} = e^{\alpha t} \begin{pmatrix} 1 & \alpha t & \frac{(\alpha t)^2}{2} & \cdots & \frac{(\alpha t)^{m-1}}{(m-1)!} \\ & 1 & \alpha t & \cdots & \frac{(\alpha t)^{m-2}}{(m-2)!} \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & \alpha t \\ & & & & 1 \end{pmatrix}. \tag{8.1.7}$$

Therefore,

$$\|e^{tA}\| \leq c|1 + t\alpha + \cdots + \frac{(t\alpha)^{m-1}}{(m-1)!}||e^{t\alpha}|, \tag{8.1.8}$$

where c is a universal constant. Thus, for any λ with $\lambda > \text{Re } \alpha$, we have

$$\|e^{tA_k}\| \leq c_k e^{t\lambda}, \tag{8.1.9}$$

since Re $\alpha - \lambda$ is strictly positive so $e^{(\text{Re } \alpha - \lambda)t}$ "beats" the polynomial growth. Taking $K = c_1 + \cdots + c_m$, we find

$$\|e^{tA}\| \leq Ke^{t\lambda}. \tag{8.1.10}$$

□

8.2 Statement of the Theorem

Assume we are given a system

$$\begin{aligned} \dot{x} &= Sx + F(x, y) \\ \dot{y} &= Uy + G(x, y). \end{aligned} \tag{8.2.1}$$

We make several assumptions:

- Assume the *spectral gap condition* is satisfied:

$$\mathrm{Re}\ \sigma(S) \le a < \lambda < b \le \mathrm{Re}\ \sigma(U). \tag{8.2.2}$$

- Assume *small nonlinearity*:

$$\begin{aligned} F(0,0) &= 0 \qquad & G(0,0) &= 0 \\ DF(0,0) &= 0 \qquad & DG(0,0) &= 0 \end{aligned} \tag{8.2.3}$$

- Assume F and G are Lipschitz continuous, with Lipschitz constant δ controlled by the spectral gap.

Finally, we recall that a set $S \subset \mathbb{R}^k \times R^l$ is invariant under the flow defined by equation (8.2.1) if $z(t) \in S$ for all $t \in \mathbb{R}$ if $z(t_0) \in S$ for some $t_0 \in \mathbb{R}$.

Theorem 82. *There is a unique C^1 function $\alpha : \mathbb{R}^k \to \mathbb{R}^l$ with $\alpha(0) = 0$, $D\alpha(0) = 0$, $\sup_{x\in\mathbb{R}^k} |\alpha(x)| < \infty$, whose graph W_α is an invariant manifold for (8.2.1).*

The proof will rely on:

1. geometric intuition about cones, and
2. a fixed point equation.

The proof itself is a sequence of estimates that show that the fixed point equation may be solved by the contraction mapping principle.

8.3 Proof of the Theorem

Our first task is to derivate a fixed point equation for α. A preliminary step is an a priori estimate *assuming* that $y = \alpha(x)$.

Lemma 20. *Assume $y = \alpha(x)$ where α is a Lipschitz function from $\mathbb{R}^k$ to $\mathbb{R}^l$. Then for every $\lambda > Re(\sigma(S))$ and $t \ge 0$, we have $|x(t)| \le K_\lambda e^{(\lambda + K_\lambda L)t}|x_0|$ where L is defined in (8.3.9) below. In particular, $L \le C(\lambda)\delta$.*

Proof. We rewrite the differential equation

$$\dot{x} = Sx + F(x,y) \tag{8.3.1}$$

as the integral equation

$$x(t) = e^{tS}x_0 + \int_0^t e^{(t-s)S} F(x(s), y(s))ds. \tag{8.3.2}$$

Since $F(0,0) = 0$, we have

$$\begin{aligned} |F(x,y)| &= |F(x,y) - F(0,0)| && (8.3.3)\\ &\le \mathrm{Lip}(f)(|x|^2 + |y|^2)^{\frac{1}{2}} && (8.3.4)\\ &\le \mathrm{Lip}(f)(|x| + |y|) && (8.3.5)\\ &= \mathrm{Lip}(f)(1 + \mathrm{Lip}(\alpha))|x| && (8.3.6) \end{aligned}$$

when $y = \alpha(x)$ and α is Lipschitz. Now we use the linear estimate (for $\lambda > \mathrm{Re}(\sigma(S))$

$$\|e^{tS}\| \le Ke^{\lambda t} \tag{8.3.7}$$

to obtain

$$|x(t)| \le K\left(e^{\lambda t} + L\int_0^t e^{\lambda(t-s)}|x(s)|ds\right) \tag{8.3.8}$$

where

$$L = \mathrm{Lip}(f)(1 + \mathrm{Lip}(\alpha)). \tag{8.3.9}$$

Multiply through by $e^{-\lambda t}$ to obtain

$$e^{-\lambda t}|x(t)| \le K + KL\int_0^t e^{-\lambda s}|x(s)|ds. \tag{8.3.10}$$

Apply Gronwall's inequality to $h(t) = e^{-\lambda t}|x(t)|$ to obtain

$$e^{-\lambda t}|x(t)| \le Ke^{KLt}|x_0|. \tag{8.3.11}$$

Therefore,

$$|x(t)| \le Ke^{(KL+\lambda)t}|x_0|. \tag{8.3.12}$$

□

We will generally assume that Lip(f) is small. (This is the "small nonlinearity" assumption.) Therefore, the dominant term in $\lambda + KL$ is λ.

We now explore the restrictions on $y = \alpha(x)$ imposed by invariance. Since $y(t)$ solves (8.2.1) we have

$$e^{-tU}y(t) - y(0) = \int_0^t e^{-sU}G(x(s), y(s))ds. \tag{8.3.13}$$

Lemma 21. *Assume $y = \alpha(x)$. Then if Lip(F) is small enough,*

$$\lim_{t\to\infty} |e^{-tU}y(t)| = 0. \tag{8.3.14}$$

Proof. First by the spectral gap estimate (now applied to $-t$ and $\theta < b \le \mathrm{Re}(\sigma(U))$,

$$\|e^{-tU}\| \le K_\theta e^{-\theta t} \qquad \text{for } t \ge 0. \tag{8.3.15}$$

Also,

$$|y(t)| \le \mathrm{Lip}(\alpha)|x(t)| \le \mathrm{Lip}(\alpha)K_\lambda e^{(\lambda+K_\lambda L)t}|x_0|. \tag{8.3.16}$$

Therefore, we find that

$$\begin{aligned} |e^{-tU}y(t)| &\le \|e^{-tU}\||y(t)| && (8.3.17)\\ &\le K_\theta e^{-\theta t}\mathrm{Lip}(\alpha)K_\lambda e^{(\lambda+K_\lambda L)t}|x_0| && (8.3.18)\\ &= K_\theta K_\lambda e^{-t(\theta-(\lambda+K_\lambda L))}|x_0|. && (8.3.19)\end{aligned}$$

Now we use the spectral gap. We see that if L is small enough (which may be achieved by controlling $\mathrm{Lip}(F)$) we may choose θ and λ so that

$$\lim_{t\uparrow\infty} |e^{-tU} y(t)| = 0. \tag{8.3.20}$$

□

Lemma 20 allows us to return to (8.3.13), use $y_0 = \alpha(x_0)$, and rewrite it as the fixed point equation

$$\alpha(x_0) = -\int_0^\infty e^{-sU} G(x(s), \alpha(x(s)))ds. \tag{8.3.21}$$

This puts us in familiar territory. We must show that the RHS defines a contraction mapping on a space of Lipschitz graphs. Let us first discover this structure and then formalize it.

As in the previous lemma (21), we find that

$$G(x, \alpha(x)) \leq \mathrm{Lip}(G)(1 + \mathrm{Lip}(\alpha)|x|. \tag{8.3.22}$$

Similarly, if we have two graphs α_1 and α_2, we find that

$$|G(x, \alpha_1(x)) - G(x, \alpha_2(x))| \leq (\mathrm{Lip}G)|\alpha_1(x) - \alpha_2(x)|. \tag{8.3.23}$$

Now we may define the contraction mapping principle more carefully.

Definition 83. Assume $f : \mathbb{R}^k \to \mathbb{R}^k$ has $f(0) = 0$ and set $\|f\|_{\mathcal{E}} = \sup_{x\in\mathbb{R}^k} \frac{|f(x)|}{|x|}$. Let

$$\mathcal{E}_0 = \{f \in C^0(\mathbb{R}^k, \mathbb{R}^l)|f(0) = 0, \|f\|_{\mathcal{E}} < \infty\}.$$

Further, for $\rho > 0$, let

$$X_\rho = \{f \in \mathcal{E}_0 | \mathrm{Lip}(f) \leq \rho\}.$$

We define a map $T : X_\rho \to X_\rho$ as follows: For every $x_0 \in \mathbb{R}^k$, define $x(t; x_0, f)$ as the unique solution to

$$\dot{x} = Sx + F(x, f(x)), \qquad x(0) = x_0.$$

We define

$$(Tf)(x_0) = -\int_0^\infty e^{-tU} G(x(s), f(x(s))ds. \tag{8.3.24}$$

Lemma 22. *$\|Tf\|_{\mathcal{E}} \leq C(LipG)$ for $C = C(\rho, b, a)$ when $f \in X_\rho$.*

Proof. This is a sequence of estimates.

1.

$$\begin{aligned} |G(x, f(x))| &\leq (\mathrm{Lip}G)(|x| + |f(x)|) && (8.3.25)\\ &\leq (\mathrm{Lip}G)(1 + \frac{|f(x)|}{|x|})|x| && (8.3.26)\\ &\leq (\mathrm{Lip}G)(1 + \rho)|x|. && (8.3.27)\end{aligned}$$

2. Now assume $x = x(t; x_0, f)$. Then by Lemma (20),

$$|x(t)| \leq K_\lambda e^{(\lambda + K_\lambda \mathrm{Lip}(f)(1+\rho))t}|x_0|. \tag{8.3.28}$$

3. Now we return to the definition of T in (8.3.24) and compute (for $x_0 \neq 0$)

$$\frac{|Tf(x_0)|}{x_0} \leq K_\lambda \int_0^\infty \|e^{-tU}\|(\mathrm{Lip}G)(1+\rho)e^{t(\lambda+K_\lambda \mathrm{Lip}(f)(1+\rho))}dt \tag{8.3.29}$$

$$\leq K_\theta K_\lambda (\mathrm{Lip}G)(1+\rho)\int_0^\infty e^{-\theta t}e^{(\lambda+K_\lambda \mathrm{Lip}(f)(1+\rho))}dt. \tag{8.3.30}$$

Note again θ is close to b, λ ic lose to a and L$ip(f)$ is small ($\leq \delta$).

Thus we have an estimate of the form

$$\sup \frac{|Tf(x_0)|}{|x_0|} \leq C(\mathrm{Lip}G) \leq C\delta. \tag{8.3.31}$$

This shows that the norm $\|Tf\|_{\mathcal{E}} < \infty$. □

Lemma 23. *Assume $f \in X_\rho$ and $x_1, x_2 \in \mathbb{R}^k$ are initial conditions for the original system (8.2.1). Then*

$$|x_1(t) - x_2(t)| \leq K_\lambda e^{(\lambda+(Lipf)(1+\rho))t}|x_1 - x_2|. \tag{8.3.32}$$

Proof. Let $x_1, x_2 \in \mathbb{R}^k$ be initial conditions for $\dot{x} = Sx + F(x, f(x))$. We then have

$$x_i(t) = e^{tS}x_i + \int_0^t e^{(t-s)s}F(x_i(s), f(x_i(s)))ds. \tag{8.3.33}$$

Therefore, using $\mathrm{Lip}(f) \leq \rho$, the difference is controlled by

$$|x_1(t) - x_2(t)| \leq \|e^{tS}\||x_1 - x_2| + \int_0^t \|e^{(t-s)s}\|(\text{Lip F})(1+\rho)|x_1(s) - x_2(s)|ds. \tag{8.3.34}$$

As in Lemma (20), we find that we may apply Gronwall's lemma to

$$h(t) = e^{-\lambda t}|x_1(t) - x_2(t)| \tag{8.3.35}$$

deducing that

$$|x_1(t) - x_2(t)| \leq K_\lambda e^{(\lambda+K_\lambda \mathrm{Lip}(f)(1+\rho))t}|x_1 - x_2|. \tag{8.3.36}$$

□

Lemma 24. *T maps X_ρ to X_ρ if LipF and LipG are small enough.*

Proof. We consider two initial conditions x_1 and x_2 and consider $|Tf(x_1) - Tf(x_2)|$. The estimates are very similar to Lemma (22), with minor modifications. We produce the analogous sequence of estimates.

1.
$$|G(x_1, f(x_1(t))) - G(x_2(t), f(x_2(t)))| \leq (\text{Lip G})(1+\rho)|x_1(t) - x_2(t)|. \tag{8.3.37}$$

2. By Lemma (23), we control $|x_1(t) - x_2(t)|$ in terms of $x_1 - x_2$. In effect, the role of $|x_0|$ in Lemma 3 is now replaced with $|x_1 - x_2|$ and we find
$$\frac{|Tf(x_1) - Tf(x_2)|}{|x_1 - x_2|} \leq C(\text{Lip}G) \leq C\delta \leq \rho \tag{8.3.38}$$
if δ is small enough. Here C depends on the spectral gap and ρ (in an explicit, though slightly messy way).

□

The last variation on this line of reasoning is the contraction mapping argument.

Lemma 25. *$T : X_\rho \to X_\rho$ is a contraction mapping when δ is small enough.*

Proof. The proof relies on a modification of Lemma (23) and Lemma (24). First, consider the solutions to
$$\dot{x}_i = Sx + F(x, f_i(x)), \qquad i = 1, 2 \tag{8.3.39}$$
with the same initial condition x_0. We have
$$x_i(t) = \int_0^t e^{(t-s)s} F(x_i(s), f_i(x_i(s)))ds. \tag{8.3.40}$$
Now, at any time s, writing x_i for $x_i(s)$,
$$|F(x_1, f_1(x_1)) - F(x_2, f_2(x_2))| \leq (\text{Lip}f)(|x_1 - x_2| + |f_1(x_1) - f_2(x_2)|. \tag{8.3.41}$$
On the other hand,
$$|f_1(x_1) - f_2(x_2)| \leq |f_1(x_1) - f_2(x_1)| + |f_2(x_1) - f_2(x_2)| \tag{8.3.42}$$
$$\leq \frac{|f_1(x_1) - f_2(x_1)|}{|x_1|}|x_1| + (\text{Lip}f_2)|x_1 - x_2| \tag{8.3.43}$$
$$\leq \|f_1 - f_2\|_{\mathcal{E}}|x_1| + \rho|x_1 - x_2|. \tag{8.3.44}$$
To summarize,
$$|F(x_1, f_1(x_1)) - F(x_2, f_2(x_2))| \leq \|f_1 - f_2\|_{\mathcal{E}}|x_1| + (1+\rho)\delta|x_1 - x_2|. \tag{8.3.45}$$
Substitute back in (8.3.40) and use Gronwall's inequality with $h(t) = e^{-\lambda t}|x_1(t) - x_2(t)|$ to obtain
$$|x_1(t) - x_2(t)| \leq K_\lambda \|f_1 - f_2\|_{\mathcal{E}} e^{(\lambda + K(1+\rho)\delta)t}|x_0|. \tag{8.3.46}$$

Now apply this estimate to Tf_1 and Tf_2:

$$Tf_i(x_0) = -\int_0^\infty e^{-tU} G(x_i(s), f_i(x_i(s)))ds. \tag{8.3.47}$$

We now have

$$\begin{aligned}
&|G(x_1(s), f_1(x_1(s))) - G(x_2(s), f_2(x_2(s)))| && (8.3.48)\\
&\quad \le (\mathrm{Lip}G)(|x_1(s) - x_2(s)| + |f_1(x_1(s)) - f_2(x_2(s))| && (8.3.49)\\
&\quad \le (\mathrm{Lip}G)(|x_1(s) - x_2(s)| + |f_1(x_1(s)) - f_2(x_2(s))|) && (8.3.50)\\
&\quad \le \delta(|x_1(s) - x_2(s)| + \|f_1 - f_2\|_\mathcal{E}|x_1(s)| + \rho|x_1(s) - x_2(s)| && (8.3.51)\\
&\quad \le \delta(1+\rho)|x_1(s) - x_2(s)| + \delta\|f_1 - f_2\|_\mathcal{E}|x_1(s)|. && (8.3.52)
\end{aligned}$$

The term $|x_1(s) - x_2(s)|$ is controlled in terms of $\|f_1 - f_2\|_\mathcal{E}|x_0|$ by (8.3.46), and $|x_1(s)|$ is controlled by Lemma (20). Using the spectral gap again, we have

$$\frac{|Tf_1(x_0) - Tf_2(x_0)|}{|x_0|} \le C\delta\|f_1 - f_2\|_\mathcal{E}. \tag{8.3.53}$$

Now take the sup over x_0 to obtain

$$\|Tf_1 - Tf_2\|_\mathcal{E} \le C\delta\|f_1 - f_2\|_\mathcal{E}. \tag{8.3.54}$$

Thus for δ small enough, this is a contraction mapping. □

In summary: Lemmas (20), (21), (22), (23), (24), and (25) show that there is a unique fixed point for T. This establishes the existence of a Lipschitz invariant manifold.

8.4 Exercises

1. We will use the following notation. $B_m(0,\varepsilon)$ is the ball of radius $\varepsilon > 0$ in $\mathbb{R}^m$. The *rectilinear flow* in $\mathbb{R}^m \times \mathbb{R}$ is the flow generated by the constant vector field $(0, \ldots, 0, 1)$.

Prove the *rectification theorem*: If $x \in \mathbb{R}^n$ is not a critical point of a flow Φ, then there is a neighborhood of x (say U), positive numbers $\varepsilon > 0$ and $\delta > 0$ and a homeomorphism G from the cylinder $B_{n-1}(0,\varepsilon) \times (-\delta, \delta)$ to U such that the image of the trajectories of the flow Φ under G^{-1} are trajectories of the rectilinear flow.

2. Consider the linear system in $\mathbb{R}^2$ given by $\dot{x} = Ax$ where A is the diagonal matrix

$$A = \begin{pmatrix} -\lambda & 0 \\ 0 & -\mu \end{pmatrix}, \quad \lambda, \mu > 0.$$

Any orbit $x(t)$ with $x(0) = (a,b)$ approaches the origin. Consider the curve in the plane obtained by piecing together the orbits with initial conditions (a,b)

and $(-a, b)$ where $a, b > 0$. How smooth is this curve? Precisely, find a condition on the eigenvalues that guarantees that this curve has exactly k derivatives at the origin.

3. Consider a C^1 vector field f in 2D such that $f(0) = 0$ and

$$Df(0) = \begin{pmatrix} -\alpha & \beta \\ -\beta & -\alpha \end{pmatrix},$$

for fixed $\alpha > 0$,$\beta > 0$. Show that all trajectories near 0 spiral into 0 in the sense that they cross each line through the origin infinitely often.

4. Provide a complete proof for the existence of Poincaré maps in the following setting. Assume we have a globally defined flow $\varphi_t : \mathbb{R}^d \to \mathbb{R}^d$ for the differential equation $\dot{x} = f(x)$. Suppose the flow has a periodic orbit Γ with period T. Consider a point $x_* \in \Gamma$, let τ denote the tangent vector to Γ at x and let S be a hyperplane in $\mathbb{R}^d$ normal to τ. Show that there is $\varepsilon > 0$ and a neighborhood $D \subset S$ such that the map $P : D \to D$ defined by $P(x) = \varphi_{T(x)}$, where $T(x)$ is the first return time to D, is well-defined.

(*Hint:* Use the implicit function theorem to solve for $T(x)$ knowing that x_* returns to D after time T.)

5. Assume $f : \mathbb{R}^2 \to \mathbb{R}^2$ is a C^1 vector field such that: (i) $f(0) = 0$; (ii) the linearization $A = Df(0)$ has two real eigenvalues $\lambda_- < 0 < \lambda_+$. Show that there are open neighborhoods of 0, denoted U and V, and a C^1 diffeomorphism $g : U \to V$ such that the vector field $h = g \circ f$ has the standard linearization

$$Dh(0) = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

6. Generalize the above assertion above to a C^1 vector field $f : \mathbb{R}^n \to \mathbb{R}^n$ when $Df(0)$ has n distinct, real eigenvalues $\lambda_1 < \lambda_2 \ldots < \lambda_k < 0 < \lambda_{k+1} < \ldots \lambda_n$, for some integer $1 < k < n$. In this case, first aim for a transformation of the linearized matrix to $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ (i.e. do *not* rescale as in question (1)). Next, try to rescale to a standard form, say $\mathrm{diag}(-k, -(k-1), \ldots, -1, 1, 2, \ldots, n-k)$.

8.5 Solutions to exercises

1. We will use the following notation. $B_m(0, \varepsilon)$ is the ball of radius $\varepsilon > 0$ in $\mathbb{R}^m$. The *rectilinear flow* in $\mathbb{R}^m \times \mathbb{R}$ is the flow generated by the constant vector field $(0, \ldots, 0, 1)$.

Prove the *rectification theorem*: If $x \in \mathbb{R}^n$ is not a critical point of a flow Φ, then there is a neighborhood of x (say U), positive numbers $\varepsilon > 0$ and $\delta > 0$ and a homeomorphism G from the cylinder $B_{n-1}(0, \varepsilon) \times (-\delta, \delta)$ to U such that the image of the trajectories of the flow Φ under G^{-1} are trajectories of the rectilinear flow.

Proof. When smoothness assumptions are not stated explicitly, assume that the vector field is C^1. Let us write x_0 instead of x for the point under consideration where the flow is non-singular. We may translate, rotate and rescale the coordinate system so that $x_0 = 0$ and $f(x_0) = e_1$.

1. Let us write $x = (x_1, y)$ to distinguish the transverse coordinates from the coordinate parallel to e_1. The first coordinate x_1 is strictly increasing and may be used to reparametrize time as follows. For fixed $\delta > 0$ and $\varepsilon > 0$ define the neighborhood of x_0

$$U = \bigcup_{|t|<\delta} \bigcup_{|y|<\varepsilon} \Phi_t(0, y).$$

Since $f \in C^1$, we may choose $\delta > 0$ and $\varepsilon > 0$ so that $\dot{x}_1 > 1/2$ for every $x \in U$. This ensures that each $x \in U$ has a unique representation of the form $x = \Phi_t(0, y)$.

2. We define the map $G : (-\delta, \delta) \times B_{n-1}(0, \varepsilon) \to U$ through

$$(t, y) \mapsto \Phi_t(0, y).$$

The map G is differentiable in both t and y and at $t = 0$ we have $DG(0, 0) = I_n$. Reducing ε and δ if necessary, we can ensure that $DG(0, 0)$ is invertible in the domain $(-\delta, \delta) \times B_{n-1}(0, \varepsilon)$. In particular, G is a diffeomorphism from this domain onto U.

3. Observe on the other hand, that $\Psi_t(0, y) := (t, y)$ is the rectilinear flow defined through the differential equation

$$\dot{y}_1 = 1, \quad \dot{y}_j = 0, \quad 2 \le j \le n.$$

Thus, $G(\Psi_t(0, y)) = \Phi_t(0, y)$, $t \in (-\delta, \delta)$. Thus, G^{-1} rectifies the flow. □

2. Consider the linear system in $\mathbb{R}^2$ given by $\dot{x} = Ax$ where A is the diagonal matrix

$$A = \begin{pmatrix} -\lambda & 0 \\ 0 & -\mu \end{pmatrix}, \quad \lambda, \mu > 0.$$

Any orbit $x(t)$ with $x(0) = (a, b)$ approaches the origin. Consider the curve in the plane obtained by piecing together the orbits with initial conditions (a, b) and $(-a, b)$ where $a, b > 0$. How smooth is this curve? Precisely, find a condition on the eigenvalues that guarantees that this curve has exactly k derivatives at the origin.

Proof. The explicit solution to the system is

$$x(t) = e^{-\lambda t} x_0, \quad y(t) = e^{-\mu t} y_0.$$

Assume x_0 and y_0 do not vanish. We eliminate t from the equation above to obtain

$$y = y_0 \left(\frac{x}{x_0} \right)^{\frac{\mu}{\lambda}} := C x^{\alpha}, \quad \alpha = \frac{\mu}{\lambda}.$$

Let $k = \text{floor}(\alpha)$. The curve $y = C x^{\alpha}$ has k derivatives at $x = 0$, but it does not have a $k + 1$ derivative. □

3. Consider a C^1 vector field f in 2D such that $f(0) = 0$ and

$$Df(0) = \begin{pmatrix} -\alpha & \beta \\ -\beta & -\alpha \end{pmatrix},$$

for fixed $\alpha > 0$, $\beta > 0$. Show that all trajectories near 0 spiral into 0 in the sense that they cross each line through the origin infinitely often.

Proof. Let us first understand the problem completely when f is linear. That is, first consider the system

$$\dot{x} = -\alpha x + \beta y, \quad \dot{y} = -\beta x - \alpha y. \tag{8.5.1}$$

Switch to polar coordinates, setting $x = r\cos\theta$, $y = r\sin\theta$. We then find that

$$\dot{r} = \frac{1}{r}\left(x\dot{x} + y\dot{y}\right), \quad \dot{\theta} = \frac{1}{r^2}\left(x\dot{y} - \dot{x}y\right). \tag{8.5.2}$$

Therefore, for the linear system (8.5.1) we find that

$$\dot{r} = -\alpha r, \quad \dot{\theta} = -\beta.$$

This system has the exact solution

$$r(t) = r_0 e^{-\alpha t}, \quad \theta(t) = \theta_0 - \beta t.$$

We may eliminate t from these equations to obtain the parametric form of a logarithmic spiral

$$r = r_0 \exp(\frac{\alpha}{\beta}(\theta - \theta_0)), \quad \theta(r) = \theta_0 + \frac{\beta}{\alpha}\log(\frac{r}{r_0}).$$

As $r \to 0$, $\theta \to -\infty$, showing that each ray $\theta = c$ is crossed infinitely many times.

Now consider the nonlinear system

$$\dot{x} = -\alpha x + \beta y + g(x,y), \quad \dot{y} = -\beta x - \alpha y + h(x,y), \tag{8.5.3}$$

where

$$g(0,0) = h(0,0) = 0 = \partial_x g(0,0) = \partial_y g(0,0) = \partial_x h(0,0) = \partial_y h(0,0) = 0.$$

By Taylor's remainder theorem, for any $\varepsilon > 0$ we may find a ball of radius r_0 about the origin such that the nonlinear system satisfies the inequalities

$$-(\alpha+\varepsilon)r \le \dot{r} \le -(\alpha-\varepsilon)r, \quad -(\beta+\varepsilon) \le \dot{\theta} = -(\beta-\varepsilon),$$

when $r \le r_0$. It follows that $r(t)$ and $\theta(t)$ decreases exponentially fast according to the estimates

$$r(t) \le r_0 e^{-(\alpha-\varepsilon)t}, \quad \theta(t) \le \theta_0 - (\beta-\varepsilon)t.$$

Again we see that $r(t)$ decreases monotonically towards 0, whereas $\theta(t)$ decreases monotonically to minus infinity, showing that each ray $\theta = c$ is crossed infinitely many times. $\square$

4. Provide a complete proof for the existence of Poincaré maps in the following setting. Assume we have a globally defined flow $\varphi_t : \mathbb{R}^d \to \mathbb{R}^d$ for the differential equation $\dot{x} = f(x)$. Suppose the flow has a periodic orbit Γ with period T. Consider a point $x_* \in \Gamma$, let τ denote the tangent vector to Γ at x and let S be a hyperplane in $\mathbb{R}^d$ normal to τ. Show that there is $\varepsilon > 0$ and a neighborhood $D \subset S$ such that the map $P : D \to D$ defined by $P(x) = \varphi_{T(x)}$, where $T(x)$ is the first return time to D, is well-defined.

(*Hint:* Use the implicit function theorem to solve for $T(x)$ knowing that x_* returns to D after time T.)

Proof. Correction. A typo in the problem statement is that $P : D \to S$, not $P : D \to D$. Further, the hint appears to have been misleading, since the problem can be solved directly by combining the rectification theorem with continuity in initial conditions.

1. As in problem 1, we can assume that $x_* = 0$, $f(0) = e_1$ (i.e. $\tau = e_1$) and $S = \{(0, y) : y \in \mathbb{R}^{d-1}\}$; here and below y will denote the transverse coordinate.

By the rectification theorem, we know that there are parameters $\varepsilon > 0$, $\delta > 0$ and a neighborhood U of 0 in which the flow can be rectified to $\Psi_t(0, y) = (t, y)$ on $(-\delta, \delta) \times B_{d-1}(0, \varepsilon)$.

2. Let D_ε be the 'time zero slice' of U, that is $D_\varepsilon = \{(0, y) : |y| < \varepsilon\}$. We must show that by reducing ε if necessary, the Poincaré map from D_ε to S is well-defined. This follows from continuity in initial conditions and the rectification theorem.

First, we recall the global condition that $\varphi_T(0) = 0$. Since $0 \in U$, by continuity in initial conditions, it follows that there is $\eta > 0$ such that $\varphi_T((0, y)) \in U$ when $|y| < \eta$. Next, we use the rectification theorem. Since $G : U \to (-\delta, \delta) \times B_{d-1}(0, \varepsilon)$, it must be the case that $G(\varphi_T(0, y)) = (t, y')$ for some $t(y)$ with $|t| < \delta$. But then the definition of the rectification map ensures that $G(\varphi_{T-t}(0, y)) = (0, y')$, so that $\varphi_{T-t(y)}(0, y)$ lies on the 'time zero slice' of D_ε. This provides the desired Poincaré time $T(y) = T - t(y)$.

*3.*The flow map φ_T is C^1 as is the rectification G. It follows that $T(y)$ is C^1 in y completing the proof. □

5. Assume $f : \mathbb{R}^2 \to \mathbb{R}^2$ is a C^1 vector field such that: (i) $f(0) = 0$; (ii) the linearization $A = Df(0)$ has two real eigenvalues $\lambda_- < 0 < \lambda_+$. Show that there are open neighborhoods of 0, denoted U and V, and a C^1 diffeomorphism $g : U \to V$ such that the vector field $h = g \circ f$ has the standard linearization

$$Dh(0) = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Proof. Correction. Unfortunately, there is a mismatch in this problem and the next between what I thought I was asking and the actual question. If one follows the question as stated, the answer is trivial. We must find a transformation g such that $Dg(0)$ satisfies the equation

$$Dh(0) = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} = Dg(0)A.$$

Clearly, all that is required is that

$$Dg(0) = Dh(0)A^{-1},$$

and this condition in turn can be obtained by choosing g to be a linear transformation $g(x) = Dh(0)A^{-1}x$. □

6. Generalize the above assertion above to a C^1 vector field $f : \mathbb{R}^n \to \mathbb{R}^n$ when $Df(0)$ has n distinct, real eigenvalues $\lambda_1 < \lambda_2 \ldots < \lambda_k < 0 < \lambda_{k+1} < \ldots \lambda_n$, for some integer $1 < k < n$. In this case, first aim for a transformation of the linearized matrix to $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ (i.e. do *not* rescale as in question (1)). Next, try to rescale to a standard form, say $\mathrm{diag}(-k, -(k-1), \ldots, -1, 1, 2, \ldots, n-k)$.

Proof. Correction. Exactly the same argument as in Problem 5 works. Neither problem is satisfactory. What I was actually after here is the proof of a lemma that nonlinear systems can be reduced to a standard form for the stable manifold theorems by suitable preprocessing. □

Chapter 9

Dynamics and algorithms

9.1 Introduction

The purpose of this chapter is to provide an introduction to the interplay between dynamics and numerical algorithms. We will provide representative examples of numerical algorithms that have an unexpected gradient or Hamiltonian structure. The use of this structure provides important insights into the behavior of the algorithm. We consider two such examples:

1. The QR algorithm for computing eigenvalues of real symmetric matrices.

2. Interior point methods for linear and semidefinite programming (abbreviated LP and SDP respectively).

We will introduce the QR algorithm and a related Hamiltonian system called the QR flow. For LP and SDP, we illustrate the role of Riemannian gradient flows. In both these instances, we must generalize the concepts of gradient and Hamiltonian flows from the naive structure on $\mathbb{R}^n$ or $\mathbb{R}^{2n}$ to manifolds as stated in Table 5.1.

9.2 Manifolds, metrics, symplectic forms

In keeping with the minimalist approach to manifold theory discussed in Section 5.6, we will focus on manifolds which are subsets of familiar spaces such as $\mathbb{R}^n$ or a suitable space of matrices. This means that we can continue to use familiar tools of calculus while developing a geometric intuition for certain algorithms.

9.2.1 The Frobenius metric

Spaces of matrices will play an important role in our work. The following notation will be used.

- $\mathbb{M}_n$: real $n \times n$ matrices.
- $\mathbb{S}_n$: real symmetric matrices in $\mathbb{M}_n$.
- $\mathbb{A}_n$: real antisymmetric matrices in $\mathbb{M}_n$.
- $\mathbb{P}_n$: positive semidefinite matrices in $\mathbb{S}_n$.
- ${\mathbb{P}_n}^+$: positive definite matrices in $\mathbb{P}_n$.

These spaces may be equipped with many metrics. The generalization to $\mathbb{M}_n$ of the Euclidean metric on $\mathbb{R}^n$ is called the Frobenius metric

$$\|M\|_2^2 := \operatorname{Tr}(M^T M). \tag{9.2.1}$$

The Frobenius metric is natural because it has a group invariance that is analogous to the rotational invariance on $\mathbb{R}^n$. The singular value decomposition of a matrix $M \in \mathbb{M}_n$ is the factorization

$$M = U\Lambda V^T, \tag{9.2.2}$$

where U and V are orthogonal matrices and Λ is a diagonal matrix of non-negative singular values. Then its Frobenius norm

$$\|M\|_2^2 = \operatorname{Tr}(M^T M) = \operatorname{Tr}(V^T \Lambda^2 V) = \operatorname{Tr}(\Lambda^2) = \operatorname{Tr}(MM^T) = \|M^T\|_2^2. \tag{9.2.3}$$

We will construct other metrics on subsets of $\mathbb{M}_n$ (such as $\mathbb{A}_n$, $\mathbb{S}_n$ and $\mathbb{P}_n$) by modifying the Frobenius metric. Note for example that $\mathbb{S}_n$ and $\mathbb{A}_n$ are orthogonal spaces with respect to the Frobenius metric. Indeed, if $S = S^T$ and $K = -K^T$, then

$$\operatorname{Tr}(K^T S) = -\operatorname{Tr}(KS) = -\operatorname{Tr}(SK) = -\operatorname{Tr}(S^T K) = -\operatorname{Tr}(K^T S).$$

When $\mathbb{S}_n$ is equipped with the Frobenius norm, the diagonal and off-diagonal terms carry different weights, since

$$\operatorname{Tr}(S^T S) = \sum_{i,j=1}^n S_{ij}^2 = \sum_{i=1}^n S_i i^2 + 2\sum_{i<j} S_{ij}^2. \tag{9.2.4}$$

This separation of diagonal and off-diagonal terms reflects the fact that $\dim(S_n) = n(n+1)/2$, so that only the terms in the upper-triangular part of S determine the matrix.

When studying LP and SDP we will need to impose positivity constraints. When $x \in \mathbb{R}^n$ the condition $x \geq 0$ means that

$$x_1 \geq 0, x_2 \geq 0, \ldots, x_n \geq 0. \tag{9.2.5}$$

Similarly for $S \in \mathbb{S}_n$ we write $S \succeq 0$ to mean $S \in \mathbb{P}_n$. This redundancy in notation is for consistency with the literature on SDP [5].

9.2.2 Gradient flows and Hamiltonian flows

The basics of differentiable manifold theory are roughly as follows. A manifold $\mathcal{M}$ is first defined as an abstract topological space (i.e. a set with a collection of neighborhood) along with an *atlas*, which is a collection of coordinate maps defined on *charts*. On each chart N coordinates are maps $\varphi : N \to \mathbb{R}^n$. Thus, the study of functions from $\mathcal{M} \to \mathbb{R}$ is reduced to a study of functions from a neighborhood in $\mathbb{R}^n$ to $\mathbb{R}$. The main requirement of the charts is that they be consistent with one another, i.e. they must agree on the overlap $N_i \cap N_j$ for any two distinct charts N_i and N_j.

The advantage of working in this abstract setting is that the manifold is defined *intrinsically.* On the other hand, when we define manifolds as subsets of $\mathbb{R}^n$, we are imposing an additional structure of an *extrinsic* space. In the examples we consider, we first assume the structure of a differentiable manifold. We then equip this manifold with addition structure, a metric or a symplectic form, and then define gradient and Hamiltonian dynamical systems with respect to this structure. This provides a unifying approach to many problems.

Given a manifold, $\mathcal{M}$, a C^1 curve is a continuously differentiable map $x : (-1,1) \to \mathcal{M}$, $t \mapsto x(t)$. Let us fix a point x and (with abuse of notation), consider curves $x(t)$ with $x(0) = x$. The tangent space to $\mathcal{M}$ at x consists of the set of derivatives $\dot{x}(0)$ for all smooth curves $x(t)$ with $x(0) = 0$. This definition seems unecessarily complicated: it is introduced so that the tangent space $T_x\mathcal{M}$ may be defined using the primitive concept of smooth functions on a manifold and nothing more. In particular, this definition ensures that the tangent bundle $T\mathcal{M}$, consisting of $T_x\mathcal{M}$, $x \in \mathcal{M}$, is an intrinsic concept.

Given a manifold $\mathcal{M}$ a 1-form is a smooth linear functional on $T\mathcal{M}$. Every smooth function $V : \mathcal{M} \to \mathbb{R}$ defines a 1-form, dV the *differential* of V, whose action at any $x \in \mathcal{M}$ and $v \in T_x\mathcal{M}$ is

$$dV(x)(v) := \frac{d}{ds}V(x(s)), \quad x(0) = x, \quad \dot{x}(0) = v. \tag{9.2.6}$$

A metric or a symplectic form is an additional structure on a differential manifold $\mathcal{M}$. A metric g is a positive definite 2-tensor. At each point $x \in \mathcal{M}$, $g(x) : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$. Given $x \in \mathcal{M}$ and $u, v \in T_x\mathcal{M}$,

$$g(x)(u,v) = g(x)(v,u), \quad g(x)(u,u) > 0, \quad \text{when} \quad u \neq 0. \tag{9.2.7}$$

A variety of different metrics on $\mathbb{R}^n$ may be generated by defining smooth maps $g : \mathbb{R}^n \to {\mathbb{P}_n}^+$ and setting $g(x)(u,v) = v^T g(x) u$.

A symplectic form ω is a closed, non-degenerate skew-symmetric 2-form. Closed means that $d\omega = 0$ in the sense of differential forms. Non-degeneracy means that for each $x \in \mathcal{M}$, if $u \in T_x\mathcal{M}$ and $\omega(x)(u,v) = 0$ for every $v \in T_x\mathcal{M}$ then $u = 0$. Skew-symmetry means that for $u, v \in T_x\mathcal{M}$ we have

$$\omega(x)(u,v) = -\omega(x)(v,u). \tag{9.2.8}$$

A dynamical system on a manifold is a differential equation of the form

$$\dot{x} = v(x), \quad x \in \mathcal{M}, \tag{9.2.9}$$

where $v(x) \in T_x\mathcal{M}$ for each $x \in \mathcal{M}$. Gradient and Hamiltonian systems use the metric and symplectic form respectively to 'convert' a 1-form into a vector field.

First, let us consider gradient flows. Assume $(\mathcal{M}^n, g)$ is a Riemannian manifold and assume that $V : \mathcal{M} \to \mathbb{R}$ is a potential on $\mathcal{M}$. The gradient of V, written $\mathrm{grad}_g V$ is the vector field defined implicitly by

$$g(x)(\mathrm{grad}_g V(x), v) = dV(x)(v), \quad x \in \mathcal{M}, \quad v \in T_x\mathcal{M}. \tag{9.2.10}$$

Since g is positive definite, the vector $\mathrm{grad}_g V(x)$ is well-defined. The associated Riemannian gradient flow is

$$\dot{x} = -\mathrm{grad}_g V(x), \quad x \in \mathcal{M}. \tag{9.2.11}$$

The fundamental estimate for gradient flows now takes the form

$$\frac{d}{dt} V(x(t)) = -\left|\mathrm{grad}_g V(x)\right|^2. \tag{9.2.12}$$

A symplectic form may be used to convert a scalar field $H : \mathcal{M} \to \mathbb{R}$. We define the vector field X_H by

$$\omega(X_H, v) = dH(x)(v), \quad x \in \mathcal{M}, \quad v \in T_x\mathcal{M}. \tag{9.2.13}$$

The vector-field X_H is well-defined because ω is non-degerate. The associated Hamiltonian system is

$$\dot{x} = X_H(x), \quad x \in \mathcal{M}. \tag{9.2.14}$$

As in our discussion of Hamiltonian systems on $\mathbb{R}^n$, it is immediate that when $x(t)$ solves equation (9.2.13)

$$\frac{d}{dt} H(x(t)) = 0. \tag{9.2.15}$$

Now that gradient flows and Hamiltonian flows have been defined, let us consider some interesting applications of these structures. In the sections that follow, we will first introduce a numerical algorithm. We then discuss its (unexpected) connection with a dynamical system.

9.3 The QR algorithm and the QR flow

9.3.1 The QR algorithm

One of the fundamental problems of numerical analysis is the *symmetric eigenvalue problem.* We assume given a matrix $L \in \mathbb{S}_n$; our task is to compute the eigenvalues of L. It is possible to pre-process the matrix L so that we may assume that it is *tridiagonal*. That is, L is of the form

$$L = \begin{pmatrix} a_1 & b_1 & 0 & \dots \\ b_1 & a_2 & b_2 & \\ \vdots & \ddots & \ddots & b_{n-1} \\ & & b_{n-1} & a_{n-1} \end{pmatrix} \tag{9.3.1}$$

A central theme in numerical linear algebra is the use of matrix factorizations [15]. One of the most fundamental of these is the QR decomposition, which is a numerical description of the Gram-Schmidt procedure for determining an orthogonal basis for a matrix L. Given a matrix L we write

$$L = QR \tag{9.3.2}$$

where Q is an orthogonal matrix such that $\mathrm{span}\{l_1, \ldots, l_k\} = \mathrm{span}\{q_1, \ldots, q_k\}$, $1 \le k \le n$, where $\{l_j\}_{j=1}^n$ and $\{q_j\}_{j=1}^n$ denote the column vectors of L and Q respectively. Fast and stable methods for computing the QR decomposition of a matrix are available in all standard software libraries for matrix computations.

The QR algorithm is an iterative scheme for computing the eigenvalues of a given matrix L_0. Given L_k, $k = 0, 1, \ldots$, the scheme produces the next iterate L_{k+1} as follows:

1. Factor the given matrix $L_k = Q_k R_k$.
2. Intertwine the factors to determine $L_{k+1} = R_k Q_k$.

The sequence of iterates is *isospectral*, that is they have the same eigenvalues. Indeed, since Q_k is orthogonal, $Q^{-1} = Q^T$, and we find that

$$L_{k+1} = Q_k^T L_k Q_k = U_k^T L_0 U_k, \quad U_k = Q_0 Q_1 \cdots Q_k. \tag{9.3.3}$$

Theorem 84. *Assume L_0 is a tridiagonal matrix with distinct eigenvalues. Then*

$$\lim_{k\to\infty} L_k = \Lambda := \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}, \tag{9.3.4}$$

where $\lambda_1 < \lambda_2 < \ldots < \lambda_n$.

Note that the algorithm computes the eigenvalues of the matrix L_0 and sorts them too!

The rate of convergence of the QR algorithm is important. Practical implementations include an additional shifting step to accelerate convergence of the scheme. This is an important augmentation of the QR algorithm, but we will ignore it so that we can explain the connection with Hamiltonian flows in the simplest setting.

9.3.2 The QR flow

How does one obtain interesting symplectic manifolds? Of course, what is interesting depends in large measure on the context, so it is helpful to take a long historical view of such questions.

The development of classical mechanics has involved a sequence of reformulations of Newton's laws since the publication of the *Principia* in 1687. The first significant reformulation is Lagrange's mechanics in 1788. The use of a Lagrangian replaced a detailed analysis of forces in a mechanical system with a general recipe for applying Newton's laws that reduces to computations with a single function, the Lagrangian. This work also contains the seeds of the modern idea of a manifold: the configuration space of a mechanical system, for example a kinematic linkage, is an early example of the idea of a manifold. In geometric terms, Lagrange's equations take place on the tangent bundle $T\mathcal{M}$ of a manifold $\mathcal{M}$. Hamilton reformulated Lagrange's equation in 1835, constructing the Hamiltonian of a mechanical system as the convex dual of the Lagrangian. The importance of this idea for fundamental physics became apparent only in the 1920s with the creation of quantum mechanics. While the structure of Hamiltonian systems is easiest to see in $\mathbb{R}^{2n}$, the natural geometric setting of Hamilton's equations is the *cotangent* bundle $T^*\mathcal{M}$, consisting of the pairs (x,p), $x \in \mathcal{M}$, $p \in T_x\mathcal{M}^*$.

While mechanical systems constitute a historically important class of Hamiltonian systems the exact solution and range of applicability of Hamiltonian systems has been significantly expanded through the use of Lie groups. In particular, most of the fundamental examples of symplectic manifolds, are *coadjoint orbits* of Lie groups. We will develop this concept systematically in Spring 20202, but first let us illustrate its utility by studying an example, the Toda lattice, that may be approached in two different ways.

First, the traditional description. The Toda lattice is a system of n particles with identical masses at positions $x_1 < x_2 < \ldots < x_n$ on the line, subject to the Hamiltonian

$$H(x,y) = \frac{1}{2}\sum_{j=1}^{n} y_j^2 + \sum_{j=1}^{n} e^{x_j - x_{j+1}}. \tag{9.3.5}$$

The Toda lattice equations are the Hamiltoniam system

$$\dot{x}_j = y_j, \quad \dot{y}_j = e^{x_{j-1}-x_j} - e^{x_j - x_{j+1}}, \quad 1 \le j \le n, \tag{9.3.6}$$

with the boundary conditions $x_0 \equiv -\infty$, $x_{n+1} = +\infty$, $e^{-\infty} = 0$. Note that we have used the standard symplectic structure $(\mathbb{R}^{2n}, J)$.

Everything so far suggests that this is a Hamiltonian flow with the 'usual' structure. However, the Toda lattice has several unexpected integrals and a systematic understanding of these integrals follows from a different description of the Toda system as a Hamiltonia flow. The following change of variables was introduced by Flaschka in 1975. Define the variables

$$a_k = -\frac{1}{2} y_k, \quad b_k = \frac{1}{2} e^{\frac{1}{2}(x_k - x_{k+1})}, \quad 1 \le k \le n, \tag{9.3.7}$$

as well as the tridiagonal matrices

$$L = \begin{pmatrix} a_1 & b_1 & & \ldots \\ b_1 & a_2 & b_2 & \\ \vdots & \ddots & \ddots & b_{n-1} \\ & & b_{n-1} & a_{n-1} \end{pmatrix}, \quad K = \begin{pmatrix} 0 & b_1 & & \ldots \\ -b_1 & 0 & b_2 & \\ \vdots & \ddots & \ddots & b_{n-1} \\ & & -b_{n-1} & a_{n-1} \end{pmatrix}. \tag{9.3.8}$$

The matrices K and L have the following properties

$$L = L^T, \quad K = -K^T, \quad K = L_-^T - L_-, \tag{9.3.9}$$

where L_- denotes the lower-triangular part of L. In these variables, the Toda lattice equations (9.3.6) take the simple form

$$\dot{L} = [K, L], \tag{9.3.10}$$

where $[A, B] = AB - BA$ denotes the Lie bracket of two matrices. This change of variables converts the Toda lattice equations into an *exactly solvable* system. A key lemma is the following feature of differential equations such as (9.3.10)

Lemma 26. *Assume $t \mapsto K(t)$ is a smooth map from $(-1, 1) \to \mathbb{A}_n$. Then the solution to equation (9.3.10) is*

$$L(t) = U(t)^T L(0) U(t), \quad \text{where} \quad \dot{U} = KU, \quad U(0) = I. \tag{9.3.11}$$

In particular, equation (9.3.10) determines an isospectral flow.

What Flaschka achieved through this change of variables is to reveal an unexpected set of conserved quantities for the particle system (9.3.5). The eigenvalues of $L(t)$, or equivalently, all the Hamiltonians $H_k(t) := \mathrm{Tr}(L^k(t))$ are conserved.

What matters for the purposes of eigenvalue computation is that equation (9.3.10) is itself a Hamiltonian flow on a symplectic manifold. That is, equation (9.3.10) is a Hamiltonian system in every sense that equation (9.3.6) is a Hamiltonian system. Here is the form in which the relation to the QR algorithm is transparent. Define a Hamiltonian H_{QR} on the space $\mathbb{S}_n$ as follows:

$$H_{\mathrm{QR}}(L) = \mathrm{Tr}(L \log L - L). \tag{9.3.12}$$

Then define the QR flow

$$\dot{L} = [K_{\mathrm{QR}}(L), L], \quad K_{\mathrm{QR}} = dH_{\mathrm{QR}}(L)_-^T - dH_{\mathrm{QR}}(L). \tag{9.3.13}$$

The comparison between this flow and the QR algorithm is as follows. Assume that L_0 is tridiagonal. Let $L(t)$ denote the QR flow with this initial condition and let L_k denote the iterates of the QR algoritm.

Theorem 85 (Stroboscope theorem)**.** *The iterates of the QR algorithm agree with the solution to the QR flow at integer times*

$$L(k) = L_k, \quad k = 1, 2, \ldots. \tag{9.3.14}$$

This is a striking and unexpected result that is typical of several iterative algorithms that are built around matrix factorizations. We will approach these results systematically in Spring 2020.

9.4 Hyperbolic geometry, LP and SDP

Optimization theory is primarily the study of fast methods to determine the minimum of a given function. It is further possible to reduce the complexity of this problem by studying the minimization of *linear* functions on convex sets.

9.4.1 Linear programming

A linear program (LP) in standard form is as follows. Assume $x \in \mathbb{R}^n$, $x \geq 0$ and assume given m constraint equations

$$a_j^T x = b_j, \quad 1 \leq j \leq m, \tag{9.4.1}$$

where $m \leq n$ and $a_j \in \mathbb{R}^n$, $j = 1, \ldots, m$ are linearly independent vectors. This constraint equation may also be expressed in the form

$$Ax = b, \tag{9.4.2}$$

where A has m rows a_j^T, $j = 1, \ldots, m$ and $b = (b_1, \ldots, b_m) \in \mathbb{R}^m$. We assume that the set $\mathcal{P}$ of solutions to (9.4.2) has a non-empty interior of dimension $n - m$. This may always be achieved by increasing the dimension n by adding new variables to the LP. In the terminology of LP, we are assuming that the constraint set is *feasible*.

In addition to the constraints, we are given a cost vector $c \in \mathbb{R}^n$. An LP in standard form is then the problem:

$$\min\{c^T x \,|: x \geq 0, Ax = b\} \tag{9.4.3}$$

The polytope $\mathcal{P}$ is convex as is the linear cost function. Since a convex function on a convex set achieves its minimum, there is at least one point on $\mathcal{P}$ that solves the minimization problem. The task in LP is to find numerical methods that solve this problem fast (which is quantified precisely with the notion of polynomial time algorithms).

There are two fundamentally distinct classes of algorithms to solve LP. The first of these, *the simplex method*, is an iterative method that 'walks along' the vertices of the $\mathcal{P}$. At each vertex $x \in \partial P$, the simplex method chooses a neighboring vertex on which the value of the cost function goes down, or returns the value $c^T x$. The simplex method was developed independently in the West and in the Soviet Union beginning in the 1940s. It was applied to problems of logistics and resource allocation during the second world war and was the foundation of the newly created field of operations research.

The simplex method was largely unchallenged for about forty years until interior point methods were shown to be successful in the 1980s by Karmarkar [11]. His pioneering work was followed by several developments that led to a systematic understanding of interior point methods. As one may expect from the terminology, in an interior point method the argmin of the cost function is approached by an iterative sequence $\{x_n\}_{n=0}^{\infty}$ that lie in the interior of $\mathcal{P}$ and approach the boundary $\partial\mathcal{P}$ as $n \to \infty$. In practice, the sequence $\{x_n\}$ is determined through the use of Newton's method. We will consider the discrete sequence more carefully in Spring 2020; for now we will explain the connection with Riemannian geometry and gradient flows by restricting attention to a continuous time variant of Karmarkar's method. In another parallel with the QR algorithm, the study of the algorithm leads to a fundamental geometric question: how does one construct interesting Riemannian metrics on a manifold?

The key new structure is the following: suppose we had a convex function $F : \mathcal{P} \to \mathbb{R}$ such that $\lim_{x\to\partial\mathcal{P}} F(x) = +\infty$. Such a function is called a *barrier* in the terminology of interior point methods. Given a barrier on $\mathcal{P}$, we define a new Riemannian metric on $\mathcal{P}$ by setting

$$g(x) = D^2F(x). \tag{9.4.4}$$

Such a metric is called a Hessian metric. The continuous time variant of interior point algorithms is the following. For $t \geq 0$ define the *central path*

$$x(t) = \mathrm{argmin}_{s\in\mathcal{P}} \left(F(s) + tc^T s\right). \tag{9.4.5}$$

The parameter t penalizes the relative strength of the barrier and the cost function. Since $F(x)$ diverges as $x \to \partial\mathcal{P}$, the barrier serves to keep $x(t)$ within the interior of $\mathcal{P}$ for all t.

When $t = 0$, the point $x_0 =: \mathrm{argmin}_{s\in\mathcal{P}}F(s)$ is called the *center* of the polytope, relative to the barrier F. The central path is the solution to the following gradient flow

$$\dot{x} = -\mathrm{grad}_g c^T x, \quad x(0) = x_0. \tag{9.4.6}$$

Observe that the complexity of the problem arises from the structure of the barrier, not the structure of the cost function.

9.4.2 Semidefinite programming

A broader class of problems that is of similar character is semidefinite programming (SDP). The orthant $x \in \mathbb{R}^n$, $x \geq 0$ is replaced with the set $X \in \mathbb{P}_n$. The linear constraints are described as follows. Assume given m matrices $A_j \in \mathbb{S}_n$ and $b \in \mathbb{R}^m$ and consider the constraint set

$$\mathcal{P} = \{X \in \mathbb{P}_n \,|\mathrm{Tr}(A_j X) = b_j, \quad 1 \leq j \leq m\}. \tag{9.4.7}$$

The set $\mathcal{P}$ is a convex polytope with respect to the geometry on $\mathbb{S}_n$ given by the Frobenius norm. As with LP, interior point methods again rely on the construction of barriers. A barrier is a convex function on $\mathcal{P}$ such that $\lim_{X\to\partial\mathcal{P}} F(X) = +\infty$.

The cost function is prescribed by a matrix $C \in \mathbb{S}_n$. The SDP is then

$$\min_{X\in\mathcal{P}} \mathrm{Tr}(CX). \tag{9.4.8}$$

The central path associated to a barrier F is the parametrized path for $t \geq 0$ determined by

$$X(t) = \operatorname{argmin}_{S\in\mathcal{P}} \left(F(S) + t\,\mathrm{Tr}(CS)\right). \tag{9.4.9}$$

It should be apparent that the structure of the SDP is completely analogous to LP. The structure of SDP may be further generalized to a class of convex optimization problems called *conic programs.* LP is obtained from SDP by restricting attention to diagonal matrices. However, such theoretical unity must also be contrasted with the fact that for practical implementations, it is quite wasteful to use methods for SDPs to solve a given LP.

9.4.3 Barriers and hyperbolic geometry

We have been ignoring one of the most important questions in the theory. How does one find barriers in the first place? And how does one choose between barriers to find the 'right' barrier for a given SDP. This is a question with some depth, since it requires a balance between a deeper understanding of the hyperbolic geometry of $\mathbb{P}_n$ and the pragmatic considerations of fast computation. To get started, here are some examples of barriers:

1. If $\mathcal{P} = \mathbb{R}^n_+$, $F(x) = -\log(x_1\cdots x_n)$.
2. If $\mathcal{P} = \mathbb{P}_n$, $F(X) = -\log\det X$.

Let us verify convexity of the barrier in these examples. The barrier for LP is obtained from the barrier for SDP by setting $X = \mathrm{diag}(x_1,\ldots,x_n)$, but it is simpler to verify convexity through a direct computation. First, for LP we differentiate $F(x) = -\log(x_1\cdots x_n)$ to obtain

$$\frac{\partial F}{\partial x_i} = -\frac{1}{x_i}, \quad \frac{\partial^2 F}{\partial x_i \partial x_j} = \frac{1}{x_i x_j}\delta_{ij}. \tag{9.4.10}$$

For SDP, the calculation reduces to understanding how to compute the first and second derivatives of the determinant at the identity. Assume given a path $X(s) \in \mathbb{S}_n$ with $X(0) = I$, $\dot{X}(0) = V$ and $\ddot{X}(0) = 0$. To leading order

$$\det(X(s)) \approx \det(I+sV) = 1+s\,\mathrm{Tr}\,V + \frac{s^2}{2}\sum_{i\neq j}\det\begin{pmatrix} V_{ii} & V_{ij} \\ V_{ji} & V_{jj}\end{pmatrix} + \ldots. \tag{9.4.11}$$

Therefore,

$$\frac{d}{ds}\det(X(s))\bigg|_{s=0} = \mathrm{Tr}\,V, \quad \frac{d^2}{ds^2}\det(X(s))\bigg|_{s=0} = \sum_{i\neq j}\left(V_{ii}V_{jj} - V_{ij}V_{ji}\right). \tag{9.4.12}$$

Next let us compute the first and second derivatives of $f(s) = -\log\det X(s)$. Since $X(0) = I$ and $\dot{X}(0) = V$ we have

$$\frac{d}{ds}f(s)\bigg|_{s=0} = \mathrm{Tr}\,V, \quad \frac{d^2}{ds^2}f(s)\bigg|_{s=0} = (\mathrm{Tr}\,V)^2 - \sum_{i\neq j}\left(V_{ii}V_{jj} - V_{ij}^2\right). \tag{9.4.13}$$

In order to get a feel for the last term, let us write it out explicitly when $n = 2$. We then obtain the sum

$$(V_{11} + V_{22})^2 - 2(V_{11}V_{22} - V_{12})^2 = V_{11}^2 + 2V_{12}^2 + V_{22}^2 = \mathrm{Tr}(V^2). \tag{9.4.14}$$

This calculation generalizes to the identity

$$\left.\frac{d^2}{ds^2} f(s)\right|_{s=0} = \mathrm{Tr}(V^2). \tag{9.4.15}$$

The general calculation may be reduced to the above. Suppose that $X \in {\mathbb{P}_n}^+$ is fixed and consider a path $X(s) \in {\mathbb{P}_n}^+$ with $\dot{X}(0) = V$. The first derivative of the determinant is [1]

$$\left.\frac{d}{ds} \det(X(s))\right|_{s=0} = \tag{9.4.16}$$
$$= \det(X) \left.\frac{d}{ds} \det(X^{-1/2} X(s) X^{-1/2})\right|_{s=0} = \det(X)\,\mathrm{Tr}(X^{-1}V).$$

In a similar manner, the second derivative of $\det X(s)$ is obtained from equation (9.4.12) by replacing V with $X^{-1/2}VX^{-1/2}$. Finally, writing $f(s) = -\log\det X(s)$ we find that

$$\left.\frac{d}{ds} f(s)\right|_{s=0} = \mathrm{Tr}(X^{-1}V) = \left.\frac{d^2}{ds^2} f(s)\right|_{s=0} = \mathrm{Tr}((X^{-1}V)^2). \tag{9.4.17}$$

The last identity shows that the barrier $F(X) = -\log\det X$ is convex, so that its Hessian determines a metric on ${\mathbb{P}_n}^+$. This metric is of fundamental importance in hyperbolic geometry.

Definition 86. The *trace metric* on ${\mathbb{P}_n}^+$ is defined as follows. Suppose $X \in {\mathbb{P}_n}^+$ and $V, W \in T_X {\mathbb{P}_n}^+$. Then

$$g_X(V, W) = \mathrm{Tr}(X^{-1}VX^{-1}W). \tag{9.4.18}$$

This metric has many properties that are analogous to the hyperbolic geometry of the upper half plane equipped with the Poincaré metric. The geodesics may be computed explicitly by analyzing the underlying group invariance and the curvature tensor may be computed explicitly.

Let us now return to an SDP with feasible polytope $\mathcal{P}$. In this setting, a barrier that generalizes $-\log\det X$ is defined in the following way. Given $X \in \mathcal{P}$, define the *polar set*

$$\mathcal{P}^*(X) = \{y \in \mathbb{S}_n \,|\mathrm{Tr}\,((Z - X)Y) \leq 1 \text{ for all } Z \in \mathcal{P}\}. \tag{9.4.19}$$

[1] Multiplying on the left and right with $X^{-1/2}$ ensures that $X^{-1/2}VX^{-1/2} \in \mathbb{S}_n$ when $X \in {\mathbb{P}_n}^+$. This is not true if we write $X^{-1}V$. This distinction does not matter for the formulas above since we also take a trace.

The *universal* barrier on $\mathcal{P}$ is the function

$$F_u(X) = -\log \mathrm{vol}(\mathcal{P}^*(X)). \tag{9.4.20}$$

This barrier has many deep and interesting properties. For example, it may be expressed in terms of a generalization of the Fourier transform to convex cones and it admits several geometric interpretations. We will consider these characterizations in depth in Spring 2020.

Chapter 10

Hamiltonian systems and symplectic geometry

The main reference is Chapter 1 of Moser's book [12].

10.1 Introduction and overview

AM220 focuses more narrowly than AM219 on three related dynamical systems:

1. Gradient flows on Riemannian manifolds;
2. Hamiltonian flows and symplectic manifolds;
3. Geodesic flow on Riemannian manifolds.

These are the three main "structures" we consider (physicists call these "formalisms"). We will consider an interplay between theory and examples. While this class involves abstract machinery (manifolds and Lie groups) we will adopt a minimalist strategy and focus on learning through examples.

For example, regarding (1) and (3) our favorite example will be the Poincare (or Lobachevsky) plane

$$\mathbb{H} = \{(x, y) \in \mathbb{R}^2 | y > 0\} \tag{10.1.1}$$

with metric

$$g(x, y) = \frac{1}{y^2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{10.1.2}$$

or

$$\mathrm{d}s^2 = \frac{1}{y^2}(\mathrm{d}x^2 + \mathrm{d}y^2) \tag{10.1.3}$$

Geodesics were computed for $(\mathbb{H}, g)$ in AM219 using ODE theory. In AM220 we will study this example using Mobius transformations and Lie groups.

The importance of this example is that its study extends to other manifolds, especially subsets of matrices. Let $\mathbb{P}_n$ be the set of $n \times n$ (real, symmetric) positive definite matrices. We equip it with the metric

$$\mathrm{d}s^2 = \mathrm{Tr}(P^{-1}\mathrm{d}PP^{-1}\mathrm{d}P) \tag{10.1.4}$$

which should be compared with equation (10.1.3)

The unifying feature is that both these spaces are manifolds of negative curvature. We will consider exact formulas and inequalities that generalize these ideas.

The above space and geometry is fundamental for semidefinite programming (SDP). This is a fundamental class of optimization problems with the following character. We are given

1. m constraint equations of the form $\mathrm{Tr}(A_iX) = b_i$, $1 \leq i \leq m$, where $b_i \in \mathbb{R}, A_i \in \mathbb{S}_n$ (real symmetric matrices)
2. a cost matrix $C \in \mathbb{S}_n$

The SDP is then

$$\begin{cases} \text{min } \mathrm{Tr}(CX) \\ \text{subject to } \mathrm{Tr}(A_iX) = b_i, 1 \leq i \leq m \\ \text{and } X \in \mathbb{P}_n \end{cases}$$

The set

$$\mathcal{P} = \{X \in \mathbb{P}_n |\, \mathrm{Tr}(A_iX) = b_i, 1 \leq i \leq m\} \tag{10.1.5}$$

is a convex polytope. The metric

$$\mathrm{d}s^2 = \mathrm{Tr}(P^{-1}\mathrm{d}PP^{-1}\mathrm{d}P) \tag{10.1.6}$$

is a Hessian metric. This means that there is a function $F : \mathbb{P}_n \to \mathbb{R}$ such that the quadratic form

$$\begin{aligned} g(P)(S,S) \quad &= \mathrm{Tr}(P^{-1}SP^{-1}S) \\ &= \frac{\mathrm{d}^2F}{\mathrm{d}P^2}(S,S) \\ &\overset{\text{def}}{=} \sum_{i,j,k,l} \frac{\partial^2 F}{\partial P_{ij}\partial P_{kl}} S_{ij}S_{kl} \end{aligned} \tag{10.1.7}$$

In fact, $F(P) = -\log\det P$. This fact generalizes to $\mathcal{P}$. We will show that there is an analogue function $F_{\mathcal{P}}(P)$ defined on $\mathcal{P}$ whose Hessian provides a metric on $\mathcal{P}$.

The solution of the SDP is then given by the gradient flow

$$\dot{X} = -\mathrm{grad}_g \operatorname{Tr}(CX) \tag{10.1.8}$$

The importance of this example is that it provides direct connections between important mathematics (spaces of negative curvature) and important applications (SDP, optimization, learning theory).

We will approach the study of $\mathbb{P}_n$ in two ways. The group theoretical approach to $(\mathbb{H}, g)$ will be generalized using symplectic transformations. On the other hand, we will also study matrix inequalities and norms.

10.2 Symplectic geometry

The above examples illustrate the interplay between gradient flows and the geodesic flow. Let us now turn to Hamiltonian systems and symplectic geometry. First, let us recognize that the geodesic flow is a Hamiltonian system. More precisely, in AM219 we derived the ordinary differential equations

$$\ddot{x}_i + \sum_{j,k} \Gamma^i_{jk} \dot{x}_j \dot{x}_k = 0, 1 \leq i \leq n, \tag{10.2.1}$$

for geodesics on a manifold $(\mathcal{M}, g)$. Here the Γ^i_{jk} are Christoffel symbols computed from the metric g. The equations for geodesics were derived by applying the principle of least action to the Lagrangian

$$L(x, \dot{x}) = \frac{1}{2} g(x)(\dot{x}, \dot{x}) \stackrel{\text{def}}{=} \frac{1}{2} |\dot{x}|^2_{g(x)} \tag{10.2.2}$$

Thus, the geodesic flow for $(\mathbb{H}, g)$ and $(\mathbb{P}_n, g)$ is an important example of a *solvable* Hamiltonian system.

However, the importance of symplectic geometry is more subtle. Some historical context may be helpful, since the evolution of these ideas are best seen as an interplay between mathematics and physics (and other sciences). A rough timeline of is as follows.

- 1687: Newton's laws $F = ma$.

  Classical application are to gravitation, especially planetary motion.

- 1788: Lagrange's equation

  $$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}_i}\right) = \frac{\partial L}{\partial x_i}\ ,\ 1 \leq i \leq n \tag{10.2.3}$$

  for system with n degrees of freedom.

  Applications: mechanics of linkages, constrained systems (leading in particular, to the idea of a manifold).

- 1833: Hamilton's equations.

$$\begin{cases} \dot{x}_i = \frac{\partial H}{\partial y_i} & \text{(position)} \\ \dot{y}_i = -\frac{\partial H}{\partial x_i} & \text{(momentum)} \end{cases} \tag{10.2.4}$$

$$\begin{aligned} H(x,y) &= \sup_{v \in \mathbb{R}^n} \{ y^T v - L(x,v) \} \\ &\stackrel{\text{def}}{=} L^* \qquad \text{(Legendre dual)} \end{aligned}$$

  The structure of Hamiltonian system provided the formalism for quantum mechanics in the 1920s (Dirac).

The modern form of Hamilton's equations require a generalization of equation (10.2.4) to a natural geometric setting. Following Lagrange, one such example of a geometric approach is the formulation of the phase space for a Lagrangian system as the tangent bundle $T\mathcal{M}$ of a manifold. A Hamiltonian system may be obtained from a Lagrangian system using convex duality and the analogous setting is $T^*\mathcal{M}$, the cotangent bundle. This space is an example of a *symplectic manifold*.

Why would one introduce such abstraction? Aren't Newton's laws enough to understand all mechanical systems? Well, actually no. Each reformulation includes the examples that motivated it, provides new solution techniques and unifies what appear to be distinct models. Here is such an example involving two classical mechanical systems:

1. Rigid body dynamics and the motion of a spinning top (gyroscope)

2. The dynamics of incompressible flow (the Euler equations of fluid mechanics).

In the 1960s, Arnold showed that both these examples can be formulated as geodesic flow on Lie groups. What changes between these examples is the underlying group, even though they describe the motion of different physical states of matter. [1]

These examples reveal that an important example of a symplectic manifold is the coadjoint orbit of a Lie group. We will define and study the equation

$$\dot{\alpha} = \mathrm{ad}^*_{\mathrm{d}H(\alpha)}\alpha \tag{10.2.5}$$

showing that it is a Hamiltonian system that includes examples 1 and 2. This equation also describes several numerical algorithms in common use for some of the fundamental problems in applied math such as eigenvalue and singular value computation.

In summary, we pursue increased abstraction because it *expands* the range of applications of the theory. For these reasons, we begin AM 220 with symplectic geometry and examples of Lie groups, building intuition for later examples.

[1] These are the group of rotations $\mathrm{SO}(n)$, and the group of volume preserving diffeomorphisms $\mathrm{SDiff}(\mathcal{M})$ respectively).

10.3 Symplectic linear algebra

This lecture introduces the abstract structure of symplectic geometry in the simplest setting. Riemannian geometry is based on inner products (or vector spaces with an inner product). Symplectic geometry is based instead on an antisymmetric, bilinear, nondegenerate form.

Assume V is a finite-dimensional vector space over the reals. A bilinear form is a map $V \times V \to \mathbb{R}$ that is linear in each coordinate. We denote our form by ω.

Definition 87. ω is a symplectic form on V if ω is antisymmetric and nondegenerate. More precisely,

1. $\omega(u, v) = -\omega(v, u)$ for all $u, v \in V$;

2. $\omega(u, v) = 0$ for all $u \in V \Leftrightarrow v = 0$.

Example 13. *The standard symplectic space is $\mathbb{R}^{2n}$ with the form*

$$\omega(u, v) = u^T J v, \tag{10.3.1}$$

where

$$J = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix} \tag{10.3.2}$$

.

It is helpful to contrat a vector space with a symplectic form with an inner product space $(V, \langle \cdot, \cdot \rangle)$. Here $\langle \cdot, \cdot \rangle$ is a symmetric, positive definite bilinear form.

1. $\langle u, v \rangle = \langle v, u \rangle$;

2. $\langle u, u \rangle \geq 0$ with equality if and only if $u = 0$.

The example $(\mathbb{R}^n, I)$ is a particularly useful example of an inner product space with

$$\langle u, v \rangle = u^T v \tag{10.3.3}$$

The Gram-Schmidt procedure tells us that any inner product space $(V, \langle \cdot, \cdot \rangle)$ with $\dim(V) = n$ can be reduced to $(\mathbb{R}^n, I)$ by a suitable choice of basis. Let us understand its symplectic analogue.

Definition 88. Assume (V, ω) is a symplectic space. Assume $E \subset V$ is a subspace of V. Define

$$E^{\perp} = \{v |\ \omega(u, v) = 0 \text{ for all } u \in E\}. \tag{10.3.4}$$

Lemma 27. *The space $(E, \omega|_E)$ is symplectic $\Leftrightarrow E \cap E^{\perp} = \{0\}$.*

Remark 89. Here $\omega|_E$ is the restriction of ω from V to E. The lemma is nontrivial for the following reasons. Unlike an inner product space, we may have $E \subset E^\perp$ in symplectic geometry. Consider $E = \text{Span}\{u\}$ for some $u \neq 0$ in V. Then $E \subset E^\perp$. Indeed $\omega(u,u) = 0$ by antisymmetry.The critical part of the lemma is the assumption of non-degeneracy.

Proof. Assume $\omega|_E$ is a symplectic form. Then consider $v \in E \cap E^\perp$. Then $\omega(u,v) = 0$ for all $u \in E$ (because $v \in E^\perp$). But then $v = 0$ because $\omega|_E$ is nondegenerate and $v \in E$. Thus, $E \cap E^\perp = \{0\}$. The converse just requires that we reverse these steps. $\square$

A similar lemma (HW) is: for any subspace $E \subset V$ we have $\dim E + \dim E^\perp = \dim V$ and $(E^\perp)^\perp = E$.

The analogue of the Gram-Schmidt process for a symplectic space is the following lemma.

Lemma 28. *For a symplectic space* (V,ω) *we must have that* $\dim(V)$ *is even. Say* $\dim(V) = 2n$*. There exists a basis* $u_1,\dots,u_n; v_1,\dots,v_n$ *such that*

$$\begin{aligned} \omega(u_i,u_j) &= 0 \quad 1 \le i,j \le n \\ \omega(v_i,v_j) &= 0 \quad 1 \le i,j \le n \\ \omega(u_i,v_j) &= \delta_{ij} \quad 1 \le i,j \le n. \end{aligned} \tag{10.3.5}$$

Remark 90. We call such a basis the canonical basis. In this basis, the space (V,ω) is isomorphic to $(\mathbb{R}^{2n}, J)$

Corollary 4. *If* (V,ω) *and* $(\tilde{V},\tilde{\omega})$ *are symplectic spaces of equal dimension, then there exist a linear isomorphism* $T : V \to \tilde{V}$ *such that*

$$\omega(u,v) = \tilde{\omega}(Tu,Tv) \text{ for all } u,v \in V \tag{10.3.6}$$

Proof. We have the canonical isomorphisms

$$\mathcal{S} : (V,\omega) \to (\mathbb{R}^{2n}, J) \tag{10.3.7}$$

$$\tilde{\mathcal{S}} : (\tilde{V},\tilde{\omega}) \to (\mathbb{R}^{2n}, J) \tag{10.3.8}$$

we set $T = \tilde{\mathcal{S}}^{-1}\mathcal{S}$ $\square$

The notion of orthogonality in symplectic geometry will recur in many ways. We say a subspace $E \subset V$ is <u>Lagrangian</u> if $E = E^\perp$. This concept has no analog in inner product spaces.

Definition 91. A linear map $U : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ is <u>symplectic</u> if

$$U^T J U = J \tag{10.3.9}$$

Lemma 29. *The subset of symplectic matrices is a group. This group is denoted* $\text{Sp}(2n,\mathbb{R})$ *or simply* $\text{Sp}(2n)$

The point here is that these transformations preserve the standard symplectic form on $\mathbb{R}^{2n}$. Further, they should be contrasted with the act of matrices such that $Q^TQ = I$. This is the orthogonal group. Both these groups can be understanded in a very explicit manner. Let us establish some basic properties.

1. Suppose $U_1, U_2 \in \mathrm{Sp}(2n, \mathbb{R})$. Then $U_1U_2 \in \mathrm{Sp}(2n, \mathbb{R})$. Indeed,
$$(U_1U_2)^T JU_1U_2 = U_2^T (U_1^T JU_1)U_2 = U_2^T JU_2 = J \tag{10.3.10}$$

2. The identity in $\mathrm{Sp}(2n, \mathbb{R})$ is the usual identity matrix.

3. The matrix $J \in \mathrm{Sp}(2n, \mathbb{R})$. Indeed, $J^T = \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} = J^{-1}$

   Check:
$$\begin{aligned} J^T J &= \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix} \\ &= \begin{pmatrix} I_n & 0 \\ 0 & I_n \end{pmatrix} = I_{2n} \end{aligned} \tag{10.3.11}$$

4. If $U \in \mathrm{Sp}(2n, \mathbb{R})$, then $U^{-1} \in \mathrm{Sp}(2n, \mathbb{R})$.

   Proof. First note that $\det(U)^2 = 1$ since $U^T JU = J \Rightarrow \det(U)^2 \det J = \det J \Rightarrow \det(U)^2 = 1$ since $\det J \neq 0$ by inspection. □

   Thus U^{-1} is well-defined and we have
$$\begin{aligned} & J^{-1}U^TJ = U^{-1} \\ \text{or} \quad & U^{-1} = J^TU^TJ \end{aligned} \tag{10.3.12}$$

   Therefore
$$\begin{aligned} (U^{-1})^T JU^{-1} &= J^TUJJU^{-1} \\ &= -J^TUU^{-1} \qquad \text{since } J^2 = -I \\ &= -J^T = J \end{aligned} \tag{10.3.13}$$

   Thus $U^{-1} \in \mathrm{Sp}(2n, \mathbb{R})$ when $U \in \mathrm{Sp}(2n, \mathbb{R})$. This verifies the standard group axioms.

What matters to us is that $\mathrm{Sp}(2n, \mathbb{R})$ is a Lie group. This means that it is a group that is also a smooth manifold. This follows from that fact that
$$\mathrm{Sp}(2n, \mathbb{R}) \equiv \{U \in \mathbb{M}_{2n} | U^T JU = J\} \tag{10.3.14}$$

The above description shows that the set $\mathrm{Sp}(2n, \mathbb{R})$ is an embedded manifold in $\mathbb{M}_{2n} \approxeq \mathbb{R}^{(2n)^2}$.

10.4 Understanding the symplectic group

Definition 92. We define

$$\mathrm{Sp}(2n,\mathbb{R}) = \{M \in \mathbb{M}_n | M^T J M = J\} \tag{10.4.1}$$

The symplectic group is a Lie group. This may be seen by noting that

1. $\mathbb{M}_n \cong \mathbb{R}^{n^2}$.
2. $\mathrm{Sp}(2n,\mathbb{R})$ is defined above as the solution set.
3. This solution set is locally the graph of a smooth function whose dimension is computed below.

We will assume this fact for now and present a set of calculations and examples that explain the nature of $\mathrm{Sp}(2n,\mathbb{R})$.
For comparison, first consider

$$O(n) = \{M \in \mathbb{M}_n | M^T M = I\} \tag{10.4.2}$$

Consider a curve

$$\begin{array}{rcl} (-1,1) & \longrightarrow & O(n) \\ t & \longmapsto & M(t) \end{array} \tag{10.4.3}$$

such that $M(0) = I$ and $\dot{M}(0) = A$. Since $M(t)^T M(t) = I$, $\dot{M}^T M + M^T \dot{M} = 0$. At $T = 0$, this gives us $A^T + A = 0 \implies A = -A^T$ (or A is antisymmetric).
The lienar space

$$\mathbb{A}_n = \{n \in \mathbb{M}_n | A = -A^T\} \tag{10.4.4}$$

is identified with the set of tangent vectors

$$T_I O(n) \stackrel{\text{def}}{=} o(n) \cong \mathbb{A}_n \tag{10.4.5}$$

where $\cong$ stands for an isomorphism.
Conversely, given an antisymmetric matrix A, the solution to

$$\begin{array}{rcl} \dot{M} & = & AM \\ M(0) & = & I \end{array} \tag{10.4.6}$$

is $M(t) = e^{tA}$. We then find that

$$\begin{array}{rcll} M(t)^T M(t) & = & \left(e^{tA}\right)^T \left(e^{tA}\right) & \\ & = & \left(e^{tA^T}\right) \left(e^{tA}\right) & \text{Series expansion of } e^{tA}. \\ & = & \left(e^{-tA}\right) \left(e^{tA}\right) & \\ & = & I & \end{array} \tag{10.4.7}$$

since $-A$ and A commute.
$\mathrm{Sp}(2n, \mathbb{R})$ may be analyzed similarly, though it is a bit more subtle.
First, the Lie algebra

$$\mathrm{Sp}(2n, \mathbb{R}) \stackrel{\text{def}}{=} T_I \mathrm{Sp}(2n, \mathbb{R}) \tag{10.4.8}$$

Now let's compute the algebra. Assume $M(t) \in \mathrm{Sp}(2n, \mathbb{R})$ is a smooth curve with $M(0) = I$. Then

$$M(t)^T J M(t) = J \quad \Longrightarrow \quad \dot{M}^T J + J\dot{M} = 0 \tag{10.4.9}$$

which is differentiable and we evaluate it at $t = 0$. Let us consider the set of solutions to 10.4.9

$$A^T J = -JA \tag{10.4.10}$$

Write

$$A = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \tag{10.4.11}$$

Then

$$A^T = \begin{pmatrix} a_1^T & a_3^T \\ a_2^T & a_4^T \end{pmatrix} \tag{10.4.12}$$

Note that 10.4.9 implies

$$\begin{pmatrix} a_1^T & a_3^T \\ a_2^T & a_4^T \end{pmatrix} \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} + \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \tag{10.4.13}$$

We compute the product to obtain

$$\begin{pmatrix} -a_3^T & a_1^T \\ -a_4^T & a_2^T \end{pmatrix} + \begin{pmatrix} a_3 & a_4 \\ -a_1 & -a_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \tag{10.4.14}$$

Thus,

$$\begin{aligned} a_1 &= -a_4^T \\ a_2 &= a_2^T \\ a_3 &= a_3^T \end{aligned} \tag{10.4.15}$$

Then the general formula for $A \in \mathrm{Sp}(2n, \mathbb{R})$ is

$$A = \begin{pmatrix} a_1 & a_2 \\ a_3 & -a_1^T \end{pmatrix} \tag{10.4.16}$$

where a_1 is arbitrary (so $\dim = n^2$).

$$\begin{aligned} a_2 = a_2^T \quad &\Longrightarrow \quad \dim = \frac{n(n+1)}{2} \\ a_3 = a_3^T \quad &\Longrightarrow \quad \dim = \frac{n(n+1)}{2} \end{aligned} \tag{10.4.17}$$

Thus,

$$\begin{aligned}\dim(\mathrm{Sp}(2n,\mathbb{R})) &= n^2+n(n+1) &(10.4.18)\\ &= 2n^2+n &(10.4.19)\end{aligned}$$

Let's simplify further and understand some curves in $\mathrm{Sp}(2n,\mathbb{R})$.
Ex1: $a_2=a_3=0$.

$$A = \begin{pmatrix} a & 0\\ 0 & -a^T\end{pmatrix} \tag{10.4.20}$$

Then the solution to

$$\begin{aligned}\dot{M} &= AM &(10.4.21)\\ M(0) &= I\end{aligned}$$

is

$$M(t)=e^{tA}=\begin{pmatrix} e^{ta} & 0\\ 0 & e^{-ta}\end{pmatrix}=\begin{pmatrix} G(t) & 0\\ 0 & (G(t))^{-T}\end{pmatrix} \tag{10.4.22}$$

where $G(t)$ is the invertible matrix e^{ta}.
If we further simplify to the case when a is (real) symmetric, we obtain

$$M(t)=\begin{pmatrix} G(t) & 0\\ 0 & G(t)^{-1}\end{pmatrix} \tag{10.4.23}$$

where $G(t)$ is positive definite.
Ex2: Similarly consider $a_1\equiv 0\equiv a_3$, then

$$A=\begin{pmatrix} 0 & a\\ 0 & 0\end{pmatrix} \tag{10.4.24}$$

where $a=a^T$. Then

$$e^{tA}\overset{\text{def}}{=}\sum_{m=0}^{\infty}\frac{t^mA^m}{m!} \tag{10.4.25}$$

but note that

$$A^2=\begin{pmatrix} 0 & a\\ 0 & 0\end{pmatrix}\begin{pmatrix} 0 & a\\ 0 & 0\end{pmatrix}=\begin{pmatrix} 0 & 0\\ 0 & 0\end{pmatrix} \tag{10.4.26}$$

Thus, $A^3=A^4=\cdots=0$ and

$$e^{tA}=I+tA=\begin{pmatrix} I & ta\\ 0 & I\end{pmatrix} \tag{10.4.27}$$

where a is a symmetric matrix.
Ex 3: Polar decomposition.
Every matrix

$$M = PQ \tag{10.4.28}$$

where P is positive semidefinite and Q is orthogonal. Since every $M \in \mathrm{Sp}(2n, \mathbb{R})$, the matrix P is invertible when $M \in \mathrm{Sp}(2n, \mathbb{R})$.

Lemma 30. *Assume $M \in \mathrm{Sp}(2n, \mathbb{R})$. Then its polar factors P and Q also belong to $\mathrm{Sp}(2n, \mathbb{R})$.*

Proof. Let's first show that $P^2 \in \mathrm{Sp}(2n, \mathbb{R})$. First $M \in \mathrm{Sp}(2n, \mathbb{R}) \Longrightarrow M^T \in \mathrm{Sp}(2n, \mathbb{R})$ (use the Lie algebra). Thus, $MM^T = PQQ^TP = P^2$ is also in $\mathrm{Sp}(2n, \mathbb{R})$.
Now if a matrix $\tilde{P}$ is positive definite then $\log \tilde{P}$ is a symmetric matrix. That is $\tilde{P} = e^{\tilde{S}}$ with $\tilde{S} = \tilde{S}^T$. On the other hand, if $\tilde{P} \in \mathrm{Sp}(2n, \mathbb{R})$, then we also have $\tilde{P} = e^{\tilde{A}}$ with $\tilde{A} = \begin{pmatrix} a_1 & a_2 \\ a_3 & -a_1^T \end{pmatrix}$ with $a_2 = a_2^T$ and $a_3 = a_3^T$. But

$$\tilde{A} = \tilde{A}^T \quad \Longleftrightarrow \quad \begin{pmatrix} a_1 & a_2 \\ a_3 & -a_1^T \end{pmatrix} = \begin{pmatrix} a_1^T & a_3^T \\ a_2^T & -a_1 \end{pmatrix} \tag{10.4.29}$$

so that $a_1 = a_1^T$ and $a_2 = 2_3$. Therefore, we find that $\tilde{P} = e^{\tilde{A}}$ with $\tilde{A} = \begin{pmatrix} a_1 & a_2 \\ a_2^T & -a_1^T \end{pmatrix}$ real symmetric and lies in $\mathrm{Sp}(2n, \mathbb{R})$.
Apply the above with $\tilde{P} = P^2$ to see that $P^2 = e^{\tilde{A}}$ with $\tilde{A}$ as above. This implies that $P = e^{\tilde{A}/2}$. Hence, P is also in $\mathrm{Sp}(2n, \mathbb{R})$. But then since $M = PQ$ and both M and $P \in \mathrm{Sp}(2n, \mathbb{R})$. We also see that $Q \in \mathrm{Sp}(2n, \mathbb{R})$.

□

10.5 Brackets

There are three distinct types of brackets when considering Hamiltonian flows. Let us first see these in examples.

1. The commutator of matrices.
Given $A, B \in \mathbb{M}_n$, we set

$$[A, B] = AB - BA \tag{10.5.1}$$

2. The commutator of vector fields: (Lie bracket)
On $\mathbb{R}^n$, suppose

$$X(x) = \sum X_i(x)\partial_{x_i} \tag{10.5.2}$$

where ∂_{x_i} is the basis vector in $\mathbb{R}^n$, is a vector field in $\mathbb{R}^n$, and similarly let $Y(x)$ be a vector field. Then for $f : \mathbb{R}^n \to \mathbb{R}$

$$(Xf)(x) \stackrel{\text{def}}{=} \sum_{i=1}^{n} X_i(x) \frac{\partial f}{\partial x_i} \tag{10.5.3}$$

Then set

$$[X,Y](f) = X(Yf) - Y(Xf) \tag{10.5.4}$$

In coordinates,

$$\begin{aligned} X(Yf) &= \sum_{i=1}^{n} X_i \partial_{x_i} \left(\sum_{j=1}^{n} Y_j \partial_{x_j} f \right) \\ &= \sum_{\substack{i=1 \\ j=1}}^{n} \left(X_i \partial_{x_i} Y_j \right) \partial_{x_j} f + \sum_{\substack{i=1 \\ j=1}}^{n} X_i Y_j \partial^2_{x_i x_j} f \end{aligned} \tag{10.5.5}$$

Therefore,

$$X(Yf) - Y(Xf) = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} (X_j \partial_{x_j} Y_i - Y_j \partial_{x_j} X_i) \right) \partial_{x_j} f \tag{10.5.6}$$

Thus, define

$$[X,Y]_i = \sum_{j=1}^{n} (X_j \partial_{x_j} Y_i - Y_j \partial_{x_j} X_i) \tag{10.5.7}$$

and

$$[X,Y](f) = \sum_{i=1}^{n} [X,Y]_i \partial_{x_j} f \tag{10.5.8}$$

3. The Poisson bracket.
This is specific to Hamiltonian systems. On $(\mathbb{R}^{2n}, J)$ we define $\{;\} : C^\infty(\mathbb{R}^{2n}) \times C^\infty(\mathbb{R}^{2n}) \to C^\infty(\mathbb{R}^{2n})$

$$\{H, K\} \stackrel{\text{def}}{=} (\operatorname{grad}_z H)^T J (\operatorname{grad}_z K) = \sum_{i=1}^{n} \frac{\partial H}{\partial x_i} \frac{\partial K}{\partial y_i} - \frac{\partial H}{\partial y_i} \frac{\partial K}{\partial x_i} \tag{10.5.9}$$

These three brackets are related in several ways.

Lemma 31. *Suppose $H, K : \mathbb{R}^{2n} \to \mathbb{R}$ are smooth Hamiltonians. Let X_H, X_K be the associated Hamiltonian vector fields, then*

$$\underbrace{[X_H, X_K]}_{\text{Lie bracket}} = X_{\underbrace{\{H,K\}}_{\text{Poisson bracket}}} \tag{10.5.10}$$

Lemma 32. *(Jacobi identity)*
(1) $\{F, G\} = -\{G, F\}$.
(2) $\{\{F, G\}, H\} + \{\{G, H\}, F\} + \{\{H, F\}, G\} = 0$

Proof. This is a calculation. (HW 2) $\square$

The above calculations provide a concrete set of examples related to flows on $(\mathbb{R}^{2n}, J)$. The group theoretic origin of these commutators is discussed in Section 12.6.

Chapter 11

Spaces with negative curvature

11.1 The Lobachevsky Plane

Definition 93. We define the Lobachevsky Plane (or the Poincaré Plane) as follows:

$$\mathbb{H} = \{z = x + iy \mid y > 0\} \tag{11.1.1}$$

The metric we use on on $\mathbb{H}$ is:

$$g = \frac{1}{y^2}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} \quad \text{or} \quad ds^2 = \frac{dx^2 + dy^2}{y^2} \tag{11.1.2}$$

We now consider curves in $\mathbb{H}$. These can be parametrized by $z(t)$ for t in some interval. Now consider the vector $\dot{z} \in T_z\mathbb{H}$ and we set:

$$|\dot{z}|_g^2 = \frac{\dot{x}^2 + \dot{y}^2}{y^2} = \frac{\dot{z}\bar{\dot{z}}}{y^2} \tag{11.1.3}$$

We also write the following:

$$ds^2 = \frac{dz d\bar{z}}{y^2} \tag{11.1.4}$$

Now the equations of the geodesics can be directly obtained from the Lagrangian:

$$L(z, \dot{z}) = \frac{1}{2y^2}|\dot{z}|^2 \tag{11.1.5}$$

We have computed these equations and their solutions in AM219. Our goal here is to integrate these equations in a simpler, but conceptually more sophisticated way, using symmetries. We study the following group:

$$SL(2,\mathbb{R}) \overset{def}{=} \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \middle| \, ad - bc = 1, a, b, c, d \in \mathbb{R} \right\} \tag{11.1.6}$$

A closely related group is:

$$PSL(2,\mathbb{R}) \overset{def}{=} SL(2,\mathbb{R})/\{\pm 1\} \tag{11.1.7}$$

That is, M and $-M$ in $SL(2,\mathbb{R})$ are identified.

The Lie Algebra of $SL(2,\mathbb{R})$ is computed by taking a curve $M(t)$ such that $M(0) = I$ If we have:

$$M(t) = \begin{pmatrix} a(t) & b(t) \\ c(t) & d(t) \end{pmatrix} \text{ with } a(t)d(t) - b(t)c(t) = 1 \tag{11.1.8}$$

We find that:

$$\dot{M}(0) = \begin{pmatrix} \dot{a}(0) & \dot{b}(0) \\ \dot{c}(0) & \dot{d}(0) \end{pmatrix} \text{ with } \dot{a}(0) + \dot{d}(0) = 1, b(0) = c(0) = 0 \tag{11.1.9}$$

Thus we have:

$$\mathfrak{sl}(2,\mathbb{R}) \overset{def}{=} \left\{ \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \middle| \, \alpha = -\delta \right\} \tag{11.1.10}$$

We consider some example curves in $SL(2,\mathbb{R})$:

Example 14.

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \implies e^{tA} = \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix} \tag{11.1.11}$$

Example 15.

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \implies e^{tA} = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \tag{11.1.12}$$

The relationship between $SL(2,\mathbb{R})$ and $(\mathbb{H}, g)$ is based on the following observations:

1. The isometry group of $(\mathbb{H}, g)$ may be identified with $SL(2,\mathbb{R})$.

2. The geodesics may be explicitly computed using this correspondence.

Some background is necessary to understand these results. The following theorem from complex analysis is the foundation of what follows:

Theorem 94. *The only conformal maps of $\mathbb{H} \to \mathbb{H}$ that are onto are of the form:*

$$f_M(z) = \frac{az+b}{cz+d} \quad \textit{for } M \in SL(2,\mathbb{R}) \tag{11.1.13}$$

We check the easy direction. That is, if f_M has the form above, then it maps $\mathbb{H} \to \mathbb{H}$ in a conformal way. To this end, we require the following lemma:

Lemma 33. *Suppose $M_i \in SL(2,\mathbb{R})$ for $i = 1, 2$ and let f_{M_i} denote the corresponding Mobius Transformations. Then we have:*

$$f_{M_2} \circ f_{M_1} = f_{M_2 M_1} \tag{11.1.14}$$

Proof. We have the following:

$$\begin{aligned} f_{M_2} \circ f_{M_1}(z) &= \frac{a_2 f_{M_1}(z) + b_2}{c_2 f_{M_1}(z) + d_2} \\ &= \frac{a_2\left(\frac{a_1 z + b_1}{c_1 z + d_1}\right) + b_2}{c_2\left(\frac{a_1 z + b_1}{c_1 z + d_1}\right) + d_2} \\ &= \frac{a_2(a_1 z + b_1) + b_2(c_1 z + d_1)}{c_2(a_1 z + b_1) + d_2(c_1 z + d_1)} \\ &= \frac{(a_2 a_1 + b_2 c_1)z + (a_2 b_1 + b_2 d_1)}{(c_2 a_1 + d_2 c_1)z + (c_2 b_1 + d_2 d_1)} \end{aligned} \tag{11.1.15}$$

Now we notice:

$$M_2 M_1 = \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix} \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix} = \begin{pmatrix} a_2 a_1 + b_2 c_1 & a_2 b_1 + b_2 d_1 \\ c_2 a_1 + d_2 b_1 & c_2 b_1 + d_2 d_1 \end{pmatrix} \tag{11.1.16}$$

Comparing terms, we get $f_{M_2} \circ f_{M_1} = f_{M_2 M_1}$ □

Now we check that $f = f_M \implies f$ is $1-1$, onto, and conformal as follows. We first consider the following elementary maps:

1. $f_M(z) = z + b$, where $m = \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}$
2. $f_M(z) = a^2 z$, where $m = \begin{pmatrix} a & 0 \\ 0 & 1/a \end{pmatrix}$
3. $f_M(z) = -1/z$, where $m = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$

We note that 1) and 2) are conformal maps from $\mathbb{H}$ to $\mathbb{H}$. As for 3), it has a pole at $z = 0$, but is analytic in $\mathbb{H}$ and:

$$f_M(z) = \frac{-\bar{z}}{|z|^2} = \frac{-x+iy}{|z|^2} \tag{11.1.17}$$

So that $\mathrm{Im}(f_M(z)) > 0$ for $y > 0$.

It is enough to check for these three maps since every $M \in SL(2,\mathbb{R})$ can be obtained by composing these transformations.

Lemma 34. *Every $M \in PSL(2,\mathbb{R})$ defines an isometry of $(\mathbb{H}, g)$*

Proof. Consider $M \in PSL(2,\mathbb{R})$ and let $w(z) = \frac{ax+b}{cz+d}$ denote the corresponding Mobius transformation. Write $w = u + iv$. We must show that:

$$\frac{dwd\bar{w}}{v^2} = \frac{dzd\bar{z}}{y^2} \tag{11.1.18}$$

First we compute v as follows:

$$w = \frac{(az+b)(c\bar{z}+d)}{|cz+d|^2} = \frac{ac|z|^2 + bc\bar{z} + adz + bd}{|cz+d|^2} \tag{11.1.19}$$

Since $M \in SL(2,\mathbb{R})$, we have:

$$bc\bar{z} + adz = (ad+bc)x + i(ad-bc)y = (ad+bc)x + iy \tag{11.1.20}$$

So we have:

$$v = \mathrm{Im}(w) = \frac{y}{|cz+d|^2} > 0 \tag{11.1.21}$$

Next we have:

$$\frac{dw}{dz} = \frac{a}{cz+d} - \frac{c(az+b)}{(cz+d)^2} = \frac{a(cz+d) - c(az+b)}{(cz+d)^2} = \frac{1}{(cz+d)^2} \tag{11.1.22}$$

Again, using $ad - bc = 1$. Now finally,

$$\frac{dwd\bar{w}}{v^2} = \left|\frac{dw}{dz}\right|^2 \frac{dzd\bar{z}}{v^2} = \frac{1}{|cz+d|^2} \cdot \frac{|cz+d|^2}{y^2} dzd\bar{z} = \frac{dzd\bar{z}}{y^2} \tag{11.1.23}$$

□

We now compute geodesics using the group action. Let:

$$A = \alpha \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + \beta \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + \gamma \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \tag{11.1.24}$$

This describes the most general element of $\mathfrak{sl}(2,\mathbb{R})$. Every $A \in \mathfrak{sl}(2,\mathbb{R})$ defines a curve $M(t) = e^{tA}$, $t \in \mathbb{R}$ of $SL(2,\mathbb{R})$. By the previous lemma, we obtain an isometry $z \mapsto f_{M(t)}(z)$ where:

$$z(t; z_0) = \frac{a(t)z_0 + b(t)}{c(t)z_0 + d(t)} \tag{11.1.25}$$

Let us first compute the path associated to $\alpha = 1, \beta = 0, \gamma = 0$. Then we have:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \implies M(t) = \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix} \tag{11.1.26}$$

Giving us:

$$z(t; z_0) = e^{2t} z_0 \tag{11.1.27}$$

In particular, if $z_0 = i$, we have $z(t; z_0) = e^{2t} i$. An initial tangent vector $v_0 \in T_{z_0}\mathbb{H}$ is mapped to $v(t) = e^{2t} v_0 \in T_{z(t;z_0)}\mathbb{H}$. Then:

$$|v(t)|_g^2 = \frac{1}{|\operatorname{Im} z(t; z_0)|^2} |v(t)|^2 = \frac{1}{e^{4t}} \cdot e^{4t} |v_0|^2 = |v_0|^2 = |v_0|_g^2 \tag{11.1.28}$$

Since $z_0 = i$. Thus we see that with $v(t) = \dot{z}(t; z_0)$, we have $|v(t)|_g = 1$. Fix $y > 1$; note that the time it takes to get from i to iy is obtained from:

$$z(t; z_0) = e^{2t} z_0 \implies iy = e^{2t} i \implies t = \frac{1}{2} \log y \tag{11.1.29}$$

More generally, for two points iy_1 and iy_2 with $y_1 < y_2$, we find:

$$t = \frac{1}{2} \log \frac{y_2}{y_1} \tag{11.1.30}$$

But since $|\dot{z}(t; z_0)|_g = 1$, we have that time and distance have the same magnitude, giving us:

$$d(iy_1, iy_2) = \frac{1}{2} \log \frac{y_2}{y_1} \tag{11.1.31}$$

Let us now extend this calculation to all geodesics. In order to do so, we first map vertical lines to circles perpendicular to $\{y = 0\}$ with radius r. We first choose a, b, c, d such that $r \mapsto 0$ and $-r \mapsto \infty$, and $ad - bc = 1$.

1. $r \mapsto 0$: This gives us:

$$\frac{ar + b}{cr + d} = 0 \implies ar + b = 0 \implies b = -ar \tag{11.1.32}$$

2. $-r \mapsto \infty$: This gives us:

$$\frac{-ar + b}{-cr + d} = \infty \implies -cr + d = 0 \implies d = cr \tag{11.1.33}$$

Then since $ad - bc = 1 \implies r(ac + ac) = 1$, we have:

$$ac = \frac{1}{2r} \tag{11.1.34}$$

Finally, we choose $ri \mapsto i$ to normalize:

$$\frac{ari+b}{cri+d} = \frac{ar(i-1)}{cr(i+1)} = i \cdot \frac{a}{c} = i \implies \frac{a}{c} = 1 \tag{11.1.35}$$

Thus we have:

$$a^2 = c^2 = \frac{1}{2r} \tag{11.1.36}$$

This gives us:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \frac{1}{\sqrt{r}} & -\sqrt{r} \\ \frac{1}{\sqrt{r}} & \sqrt{r} \end{pmatrix} \tag{11.1.37}$$

Then the image of the circle $re^{i\theta}$ with $0 < \theta < \pi$ is the line $\{iy\}_{y>0}$. Since f is an isometry:

$$d(re^{i\theta_1}, re^{i\theta_2}) = d(iy(\theta_1), iy(\theta_2)) = \frac{1}{2} \log \frac{y(\theta_2)}{y(\theta_1)} \tag{11.1.38}$$

We further have:

$$y(\theta) = \frac{r^{-1/2} re^{i\theta} - \sqrt{r}}{r^{-1/2} re^{i\theta} + \sqrt{r}} = \frac{e^{i\theta} - 1}{e^{i\theta} + 1} = \cot \frac{\theta}{2} \tag{11.1.39}$$

Thus we have:

$$d(re^{i\theta_1}, re^{i\theta_2}) = \frac{1}{2} \log \frac{\cot \frac{\theta_2}{2}}{\cot \frac{\theta_1}{2}} \tag{11.1.40}$$

Now let us obtain all geodesics through the points $z = i$. Each geodesic through $z = i$ is given by a tangent vector v. On the other hand, it is also given by a curve e^{tA} for $A \in \mathfrak{sl}(2, \mathbb{R})$. So with $M(t) = e^{tA}$ and $z \mapsto f_{M(t)}(z)$ with:

$$f_{M(t)}(z) = \frac{a(t)z + b(t)}{c(t)z + d(t)} \quad \text{and} \quad M(t) = \begin{pmatrix} a(t) & b(t) \\ c(t) & d(t) \end{pmatrix} \tag{11.1.41}$$

Then we have:

$$\dot{z}|_{t=0} = \frac{\dot{a}z + \dot{b}}{c(t)z + d(t)} - \frac{a(t)z + b(t)}{(c(t)z + d(t))^2} (\dot{c}z + \dot{d}) \tag{11.1.42}$$

At $t = 0$, we have $M = I$. Also evaluating at $z = i$, we get:

$$\dot{z}|_{t=0} = (\dot{a}i + \dot{b}) - i(\dot{c}i + \dot{d}) = (\dot{a} - \dot{d})i + (\dot{b} + \dot{c}) \tag{11.1.43}$$

But since we have:

$$A = \alpha \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + \beta \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + \gamma \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \tag{11.1.44}$$

This means $\dot{a} = \alpha, \dot{d} = -\alpha, \dot{b} = \beta, \dot{c} = \gamma$, and the given direction is:

$$v = 2\alpha i + (\beta + \gamma) \tag{11.1.45}$$

We have three parameters α, β, γ, and only two equations to determine the real and imaginary parts of v. So given v, we can find the geodesic.

11.2 Positive Definite Matrices

In this chapter we will use the following notation:

- $\mathbb{M}_n$ is the set of $n \times n$ real matrices.
- $\mathbb{P}_n$ is the set of $n \times n$ real symmetric positive definite matrices.
- $\mathbb{S}_n$ is the set of $n \times n$ real symmetric matrices.
- $GL(n)$ is the set of $n \times n$ invertible matrices.

We call a matrix positive if it is real symmetric positive definite.

Lemma 35. *The exponential mapping A to e^A maps $\mathbb{S}_n$ to $\mathbb{P}_n$.*

Proof. We have $e^A = e^{A/2}e^{A/2}$ fromt the series expansion. Thus e^A is a product of two invertible symmetric matrices. □

We can also see this from the spectral theorem.

Lemma 36. *Every $P \in \mathbb{P}_n$ has a unique positive square root.*

The natural inner product on $\mathbb{S}_n$ is

$$\langle A, B\rangle_{\mathrm{Tr}} \stackrel{\mathrm{def}}{=} \mathrm{Tr}(A^T B) = \mathrm{Tr}(AB) = \sum_{i,j} A_{ij}B_{ij}. \tag{11.2.1}$$

Definition 95. The *trace metric* on $\mathbb{P}_n$ is defined (for $A, B \in T_P\mathbb{P}_n$) to be

$$\langle A, B\rangle_P = \mathrm{Tr}(P^{-1}AP^{-1}B) = \mathrm{Tr}((P^{-1/2}AP^{-1/2})(P^{-1/2}BP^{-1/2})). \tag{11.2.2}$$

Thus $\langle A, A\rangle_P > 0$ unless $A = 0$.

Definition 96. The $GL(n)$ action on $\mathbb{P}_n$ is defined as follows. For each $g \in GL(n)$ we define $[g] : \mathbb{P}_n \to \mathbb{P}_n$ mapping P to gPg^T.

Check symmetry, then positivity:

$$x^T(gPg^T)x = (g^T x)^T P(g^T x) \geq 0 \tag{11.2.3}$$

with equality only if $x = 0$.

Lemma 37. *The $GL(n)$ action is transitive: for every $P_1, P_2 \in \mathbb{P}_n$ there is $g \in GL(n)$ such that $[g]P_1 = P_2$.*

Proof. Assume that $P_1 = I$ and write P for P_2. We want g such that $[g]I = P$, that is, $gg^T = P$. Clearly $g = \sqrt{P}$ satisfies this equation. Next, in general to solve $gP_1g^T = P_2$ we choose $g\sqrt{P_1} = \sqrt{P_2}$ or $g = \sqrt{P_2}(\sqrt{P_1})^{-1}$. □

Theorem 97. *$[g]$ is an isometry of $(\mathbb{P}_n, \mathrm{Tr})$ for each $g \in GL(n)$.*

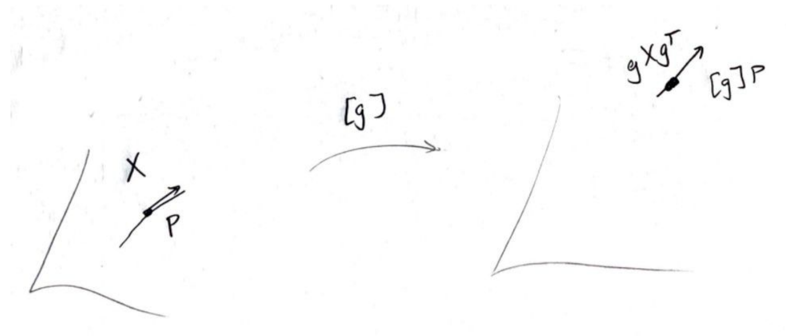

Figure 11.2.1: The $[g]$ action.

Proof. Let us first compute the derivative of $[g]$. Consdier $P \in \mathbb{P}_n$ and a curve $P(t)$ through P (i.e. $P(0) = P$) with $\dot{P}(0) = X \in \mathbb{S}_n$. Then

$$\frac{d}{dt}[g](P(t))\Big|_{t=0} = g\dot{P}(t)g^T\Big|_{t=0} = gXg^T. \tag{11.2.4}$$

Therefore, we need to check that $X \mapsto gXg^T$ preserves lengths. (See Figure 11.2.1.) We compute

$$\langle gXg^T, gXG^T\rangle_{[g](P)} = \mathrm{Tr}(([g]P)^{-1}[g]X([g]P)^{-1}[g]X) \tag{11.2.5}$$

Here $([g](P))^{-1} = (gPg^T)^{-1} = g^{-T}P^{-1}g^{-1}$, so

$$([g]P)^{-1}[g]X = g^{-T}P^{-1}g^{-1}gXg^T = g^{-T}P^{-1}Xg^T. \tag{11.2.6}$$

Thus

$$\begin{aligned} \langle gXg^T, gXg^T\rangle_{[g](P)} &= \mathrm{Tr}(g^{-T}P^{-1}Xg^Tg^{-T}P^{-1}Xg^T) &(11.2.7)\\ &= \mathrm{Tr}((P^{-1}X)(P^{-1}X)) &(11.2.8)\\ &= \langle X, X\rangle_P. &(11.2.9) \end{aligned}$$

□

Now let us turn to the computation of the distance between given P_1 and P_2. We first consider rays obtained by the exponential map. (See Figure 11.2.2.) Consider the curve $P(t) = e^{tA}$. Its tangent vector at any t is $\dot{P}(t)$.

Theorem 98. *$\langle \dot{P}, \dot{P}\rangle_P = \|A\|_2^2$ for all t.*

Proof. We compute

$$\dot{P} = Ae^{tA} = AP = PA. \tag{11.2.10}$$

Then

$$\langle \dot{P}, \dot{P}\rangle_P = \mathrm{Tr}(P^{-1}\dot{P}P^{-1}\dot{P}) = \mathrm{Tr}(A^2) \overset{\text{def=}}{\|} A\|_2^2. \tag{11.2.11}$$

□

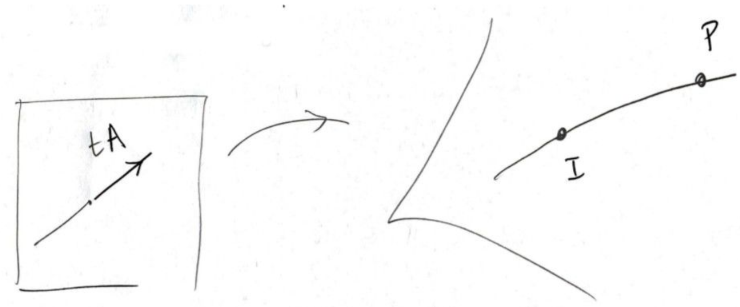

Figure 11.2.2: The exponential map on a line $\{tA\}$ through the origin.

Thus the exponential map is metric preserving on lines through the origin.

Theorem 99. *Let $P_1, P_2 \in \mathbb{P}_n$. Let $\gamma_1, \cdots, \gamma_n$ denote the roots of* $\det(\lambda P_1 - P_2) = 0$. *Then*

$$\operatorname{dist}(P_1, P_2) = \sum_{j=1}^{n} (\log \lambda_j)^2. \tag{11.2.12}$$

Proof. Suppose $P_1 = I$, $P_2 = \Lambda$, where

$$\Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}. \tag{11.2.13}$$

We have $\det(\lambda - P) = 0$. Let $A = \log \Lambda$ and consider the line through the origin $\{tA\}$. We have seen in the previous theorem that

$$\langle \dot{P}, \dot{P} \rangle_P = \|A\|_2^2. \tag{11.2.14}$$

Thus, $\operatorname{dist}(I, P) = \|A\|_2^2 t$ where $e^{tA} = P$. So $t = 1$ and $\|A\|_2^2 = \sum_{j=1}^{n} (\log \lambda_i)^2$. Now let us reduce the general case to this calculation. First, given P_1, P_2, choose $g = P_1^{-1/2}$. Then

$$[g](P_1) = gP_1g^T = I \tag{11.2.15}$$

and

$$[g](P_2) = P_1^{-1/2} P_2 P_1^{-1/2}. \tag{11.2.16}$$

We thus have that

$$\operatorname{dist}(P_1, P_2) = \operatorname{dist}(I, P_1^{-1/2} P_2 P_1^{-1/2}). \tag{11.2.17}$$

Then we diagonalize: suppose $P_1 = I$, $P_2 = Q \Lambda Q^T$. We would like to say that $\operatorname{dist}(I, P_2) = \operatorname{dist}(I, \Lambda)$. Let's look at the role of rotations more carefully. The group action $[g] : \mathbb{P}_n \to \mathbb{P}_n$ acts on the identity by $e \mapsto geg^T = gg^T$.

In particular, when $Q \in O(n)$, $e \mapsto e$. So the map $D[g](e)$ has a nullspace consisting on $\mathbb{A}_n$, the antisymmetric matrices at I (HW). We then compute that if $P = Q\Lambda Q^T$ then choose $g = Q$, and we have

$$[g](\Lambda) = P, \qquad [g](I) = I, \tag{11.2.18}$$

thus $\mathrm{dist}(I, \Lambda) = \mathrm{dist}(I, P)$. Combining the calculations, we find that

$$\mathrm{dist}(P_1, P_2) = \mathrm{dist}(I, P^{-1/2}P_2P^{-1/2}) = \mathrm{dist}(I, \Lambda) = \sum_{j=1}^{n}(\log \lambda_j)^2, \tag{11.2.19}$$

where $P_1^{-1/2}P_2P^{-1/2} = Q\Lambda Q^T$ with $Q \in O(n)$. □

Now, we will further consider the space $(\mathbb{P}_n, \mathrm{tr})$, that is, $\mathbb{P}_n$ equipped with the trace metric $ds^2 = \mathrm{tr}(P^{-1}dPP^{-1}dP)$. For a path $\gamma : [0,1] \to \mathbb{P}_n$, we define its length

$$L(\gamma) = \int_0^1 \langle \dot\gamma(t), \dot\gamma(t)\rangle_{\gamma(t)}^{\frac{1}{2}} dt = \int_0^1 (\mathrm{tr}(\gamma^{-1}\dot\gamma\gamma^{-1}\dot\gamma))^{\frac{1}{2}} dt. \tag{11.2.20}$$

If $A, B \in \mathbb{P}_n$, we define

$$\delta_2(A, B) = \inf_{\gamma \in S} L(\gamma) \tag{11.2.21}$$

where

$$S = \{\gamma : [0,1] \to \mathbb{P}_n | \gamma \in C^1, \gamma(0) = A, \gamma(1) = B\}. \tag{11.2.22}$$

Note that for each $g \in GL(n)$, the group action $[g] : \mathbb{P}_n \to \mathbb{P}_n$ defines an isometry because

$$L([g](\gamma)) = L(\gamma). \tag{11.2.23}$$

This follows from Theorem 97. We now approach the formula for geodesics differently. Recall that for each $f : \mathbb{R} \to \mathbb{R}$ we define $f : S_n \to S_n$ by $f(A) = Qf(\Lambda)Q^T$ when $A = Q\Lambda Q^T$. The derivative of f is computed as follows:

$$Df(A)(B) = \frac{d}{d\tau} f(a(\tau))\Big|_{\tau=0} \tag{11.2.24}$$

where $A(0) = A$, $\dot A(0) = B$. The following formulas for the derivative were shown in the homework. Diagonalize $A = Q\Lambda Q^T = \sum_{j=1}^n \lambda_j P_j$ where $P_j = q_j q_j^T$, the q_j are the columns of Q, and the λ_j are the diagonal entries of Λ. Then:

Lemma 38.

$$Df(A)(B) = \sum_{i,j=1}^{n} \frac{f(\lambda_i) - f(\lambda_j)}{\lambda_i - \lambda_j} P_i B P_j. \tag{11.2.25}$$

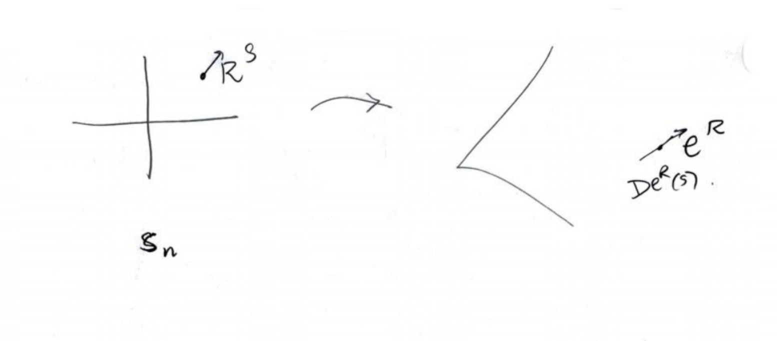

Figure 11.2.3: The exponential map.

If we define $f^{[1]}(\Gamma)$ as the matrix with entires $\frac{f(\lambda_i)-f(\lambda_j)}{\lambda_i-\lambda_j}$ for $i \neq j$ and $f'(\lambda_i)$ for $i = j$, then we find that

$$Df(A)(B) = f^{[1]}(A) \circ B, \tag{11.2.26}$$

where $\circ$ denotes the Schur or Hadamard product in the diagonal basis for A.

We will use this lemma, applying it to $f(x) = e^x$. We recall that $\|S\|_2^2 = \sum_{i,j=1}^n S_{ij}^2$ for $S \in S_n$. We have the IEMI lemma, where IEMI stands for Infinitesimal Exponential Metric Increasing (see Figure 11.2.3):

Lemma 39. *For all $R, S \in S_n$,*

$$\|e^{-\frac{R}{2}} De^R(s) e^{-\frac{R}{2}}\|_2 \geq \|S\|_2. \tag{11.2.27}$$

Proof. Choose a basis in which R is diagonal. Then with

$$R = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \tag{11.2.28}$$

we have

$$e^{-\frac{R}{2}} = \begin{pmatrix} e^{-\frac{\lambda_1}{2}} & & \\ & \ddots & \\ & & e^{-\frac{\lambda_n}{2}} \end{pmatrix} \tag{11.2.29}$$

Let $P_i = e_i e_i^T$ where e_i is the vector with a 1 in the ith place and a 0 everywhere else. Then

$$De^R(s)_{ij} = \left(\frac{e^{\lambda_i} - e^{\lambda_j}}{\lambda_i - \lambda_j}\right)(e_i e_i^t) S e_j e_j^t, \tag{11.2.30}$$

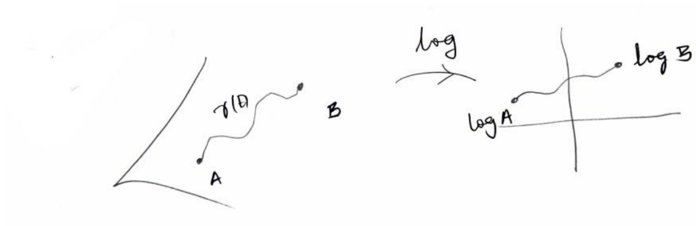

Figure 11.2.4: The log map.

where $(e_i e_i^T)Se_j e_j^T = S_{ij}$. Therefore, the (i,j) entry of $e^{-\frac{R}{2}}De^R(s)e^{-\frac{R}{2}}$ is

$$e^{-\frac{\lambda_i}{2}}\left(\frac{e^{\lambda_i}-e^{\lambda_j}}{\lambda_i-\lambda_j}\right)e^{-\frac{\lambda_j}{2}}S_{ij} = \left(\frac{e^{\frac{\lambda_i-\lambda_j}{2}}-e^{-\frac{\lambda_i-\lambda_j}{2}}}{\lambda_i-\lambda_j}\right)S_{ij} \tag{11.2.31}$$

$$= \left(\frac{\sinh(\frac{\lambda_i-\lambda_j}{2})}{(\frac{\lambda_i-\lambda_j}{2})}\right)S_{ij} \tag{11.2.32}$$

$$\le |S_{ij}|, \tag{11.2.33}$$

where the last inequality follows from $|\frac{\sinh s}{s}| \le 1$. This proves the lemma when R is diagonal. The $\|\cdot\|_2$ norm is invariant under rotations (HW) so this also deals with the general case. □

Corollary 5. *Let $S(t)$, $0 \le t \le 1$ be a C^1 path in S_n and let $\gamma(t) = e^{S(t)}$. Then*

$$\int_0^1 \|S'(t)\|_2 dt \le L(\gamma). \tag{11.2.34}$$

Proof. We have

$$L(\gamma) \overset{\text{def}}{=} \int_0^1 \langle\dot\gamma,\dot\gamma\rangle_{\gamma(t)}^{\frac{1}{2}} dt \overset{\text{IEMI}}{\ge} \int_0^1 \|\dot S(t)\|_2 dt. \tag{11.2.35}$$

Here the IEMI is applied with $R = S(t), S = \dot S(t)$. □

We now have the EMI Theorem (see Figure 11.2.4:

Theorem 100. *For every $A, B \in \mathbb{P}_n$, we have*

$$\|\log A - \log B\|_2 \le \delta_2(A,B). \tag{11.2.36}$$

Equivalently, for every $R, S \in S_n$ we have

$$\|R - S\|_2 \le \delta_2(e^R, e^S). \tag{11.2.37}$$

Note this theorem says that the exponential map from S_n to $\mathbb{P}_n$ is metric increasing.

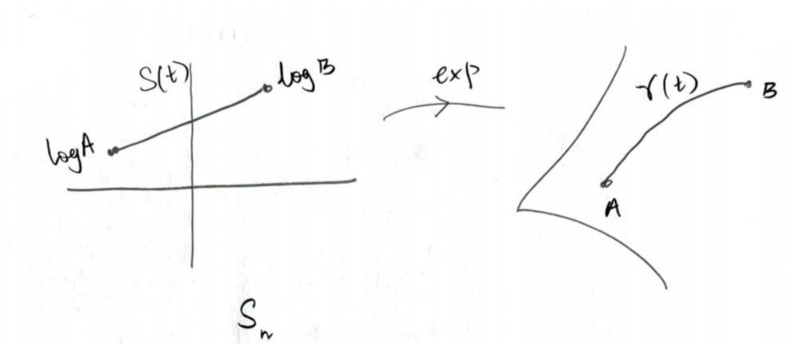

Figure 11.2.5: The exponential map on a geodesic for commuting A and B.

Lemma 40. *Suppose A and B commute. Then exp maps the line segment connecting $\log A$ and $\log B$ to the geodesic connecting A and B in $\mathbb{P}_n$ (see Figure 11.3.1), and*

$$\delta_2(A, B) = \| \log A - \log B\|_2. \tag{11.2.38}$$

Proof. We have

$$\gamma(t) = e^{S(t)} = e^{(1-t)\log A + t\log B}. \tag{11.2.39}$$

In general, $e^{H+K} \neq e^H e^K$, but this is true for commuting matrices. Thus

$$\gamma(t) = e^{(1-t)\log A} e^{t\log B} = A^{1-t}B^t \tag{11.2.40}$$

and

$$\dot\gamma(t) = (\log B - \log A)\gamma(t). \tag{11.2.41}$$

Therefore,

$$\langle \dot\gamma, \dot\gamma\rangle_\gamma = \mathrm{Tr}(\gamma^{-1/2}\dot\gamma\gamma^{-1/2}\gamma^{-1/2}\dot\gamma\gamma^{-1/2} \tag{11.2.42}$$
$$= \mathrm{Tr}(\dot\gamma\gamma^{-1}\dot\gamma\gamma^{-1}) \tag{11.2.43}$$
$$= \mathrm{Tr}((\log B - \log A)^2) \tag{11.2.44}$$
$$= \| \log B - \log A\|_2^2. \tag{11.2.45}$$

Thus, we have

$$L(\gamma) = \| \log B - \log A\|_2^2. \tag{11.2.46}$$

On the other hand, the EMI says that

$$L(\gamma) \geq \| \log B - \log A\|_2^2, \tag{11.2.47}$$

so we see that this length is achieved and the path $\gamma(t)$ is a geodesic. $\square$

The next theorem gives a formula for geodesics.

Theorem 101. *Let $A, B \in \mathbb{P}_n$. Then there exists a unique geodesic joining A and B. This geodesic is parametrized by*

$$\gamma(t) = A^{\frac{1}{2}}(A^{-\frac{1}{2}}BA^{-\frac{1}{2}})^t A^{-1/2} \tag{11.2.48}$$

for $0 \leq t \leq 1$. This is the natural parameterization in the sense that

$$\delta_2(A, \gamma(t)) = t\delta_2(A, B). \tag{11.2.49}$$

Further,

$$\delta_2(A, B) = \| \log A^{-1/2}BA^{-1/2}\|_2. \tag{11.2.50}$$

Proof. We reduce to the case of commuting matrices as follows: First, I and $A^{-1/2}BA^{-1/2}$ commute, so

$$\delta_2(I, A^{-1/2}BA^{-1/2}) = \| \log A^{-1/2}BA^{-1/2}\|_2, \tag{11.2.51}$$

and the geodesic parameterizing this is $\gamma_0(t) = (A^{-1/2}BA^{-1/2})^t$. Now choose $[g] = A^{1/2}$ to find that the path $[g]\gamma_0 = A^{-1/2}(A^{-1/2}BA^{-1/2})^t A^{1/2}$ joins A and B and is a geodesic. Then

$$\begin{aligned} \delta_2(A, B) &= \delta_2(I, A^{-1/2}BA^{-1/2}) && (11.2.52)\\ &= \| \log I - \log A^{-1/2}BA^{-1/2}\|_2 && (11.2.53)\\ &= \| \log A^{-1/2}BA^{-1/2}\|_2. && (11.2.54) \end{aligned}$$

□

11.3 The Siegel half-space

Definition 102. A complex matrix $Z = X + iY$, where $X, Y \in \mathbb{M}_n(\mathbb{R})$ is said to be *Hermitian* if $Z = Z^* = X^T - iY^T$, namely we have $X = X^T$ (real symmetric) and $Y = -Y^T$ (antisymmetric). We denote the set of Hermitian $n \times n$ matrices by $\mathbb{H}(n)$. The dimension of $\mathbb{H}(n)$ is calculated as follows:

$$\dim(\mathbb{H}(n)) = \frac{n(n+1)}{2} + \frac{n(n-1)}{2} = n^2 \tag{11.3.1}$$

Lemma 41. *Suppose $M \in \mathbb{H}(n)$, then $e^{iM} \in U(n)$.*

Proof.

$$(e^{iM})^*(e^{iM}) = e^{-iM^*}e^{iM} = e^{-iM}e^{iM} = I \tag{11.3.2}$$

Example: Consider

$$U = \begin{pmatrix} e^{i\theta_1} & 0 & \dots & 0 \\ 0 & e^{i\theta_2} & 0 & \dots \\ \vdots & & \ddots & \\ 0 & \dots & 0 & e^{i\theta_n} \end{pmatrix} \quad \text{where } (\theta_1, \dots, \theta_n) \in \mathbb{T}^n$$

□

Definition 103. A complex matrix $Z = X+iY$, where $X, Y \in \mathbb{M}_n(\mathbb{R})$ is said to be *symmetric* if $Z = Z^T = X^T + iY^T$, namely we have $X = X^T$ (real symmetric) and $Y = Y^T$ (real symmetric). We denote the set of complex symmetric $n \times n$ matrices by $\mathcal{S}$. The dimension of $\mathcal{S}$ is calculated as follows:

$$\dim(\mathcal{S}) = \frac{n(n+1)}{2} + \frac{n(n+1)}{2} = n(n+1) \tag{11.3.3}$$

Definition 104. The Sigel upper half-space is defined as

$$\mathcal{H} = \{Z \in \mathcal{S} | Y > 0 \quad \text{where } Z = X + iY\} \tag{11.3.4}$$

In other words, the imaginary part of $\mathrm{Im}(Z) = Y$ is an element of $\mathbb{P}_n$. We note that $\mathcal{H}$ is the matrix analogue of the Poincaré upper half-plane $\mathbb{H}$.

Let us recall some important properties of the Poincaré upper half-plane $\mathbb{H}$:

1. The set of analytic maps from $\mathbb{H} \to \mathbb{H}$ that are one-one and onto is given by
$$\left\{ f_M(z) = \frac{az+b}{cz+d} \middle| M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbb{P}SL_2(\mathbb{R}) \right\} \tag{11.3.5}$$
2. The Poincaré upper half-plane $\mathbb{H}$ is endowed with the hyperbolic metric given by
$$ds^2 = \frac{dz d\bar{z}}{y^2}$$
with respect to which each f_M is an isometry.
3. For each $A \in SL(2;\mathbb{R})$, the flow $z \mapsto f_{e^{tA}}(z)$ generates geodesics on $\mathbb{H}$. It can be shown that between any two points z_1, z_2, such that $x_1 \neq x_2$ the unique geodesic connecting the two points is a semi-circular arc with center x_0 lying on the x-axis. If we rewrite in translated polar coordinates, i.e., $z_1 = x_0 + re^{i\theta_1}$ and $z_2 = x_0 + re^{i\theta_2}$, then the hyperbolic distance between z_1 and z_2 is given by $d(z_1, z_2) = (1/2)(\log\cot(\theta_2/2) - \log\cot(\theta_1/2))$.

Let us recall the definition of the symplectic group

$$Sp(n) = \{M \in \mathbb{M}_{2n}(\mathbb{R}) | M^T J M = J\}$$

Equivalently, if $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in Sp(n)$, then we have

$$\begin{pmatrix} A^T & C^T \\ B^T & D^T \end{pmatrix} \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

Definition 105. Given $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in Sp(n)$, we can define the analogue of the Möbius transformation for the Siegel upper half-space (known as the symplectic transformation) as follows:

$$f_M : \mathcal{H} \to \mathcal{S} \tag{11.3.6}$$
$$Z \mapsto (AZ+B)(CZ+D)^{-1} \tag{11.3.7}$$

Lemma 42. *The transformation f_M maps the domain $\mathcal{H}$ onto itself. In other words,* $\mathrm{Im}(f_M(Z)) \in \mathbb{P}_n$ *or f_M is a map from $\mathcal{H} \to \mathcal{H}$.*

Now, we state Siegel's theorems [13]:

Theorem 106. *Every analytic mapping of $\mathcal{H}$ onto itself is a symplectic transformation.*

Definition 107. For any two points $Z, Z_1 \in \mathcal{H}$, we define the analogue of the cross-ratio as

$$R(Z, Z_1) = (Z - Z_1)(Z - \bar{Z}_1)^{-1}(\bar{Z} - \bar{Z}_1)(\bar{Z} - Z_1)^{-1} \tag{11.3.8}$$

Theorem 108. *There exists a symplectic transformation mapping a given pair Z, Z_1 of $\mathcal{H}$ into another given pair W, W_1 of $\mathcal{H}$, if and only if the two matrices $R(Z, Z_1)$ and $R(W, W_1)$ have the same eigenvalues.*

Definition 109. We can define a metric on $\mathcal{H}$ given by

$$ds^2 = \mathrm{Tr}(Y^{-1} dZ Y^{-1} d\bar{Z}) \quad \text{where } Z = X + iY \in \mathcal{H} \tag{11.3.9}$$

Lemma 43. *The metric defined above is invariant under the $Sp(n)$ action.*

Theorem 110. *There exists a unique geodesic arc connecting $Z, Z_1 \in \mathcal{H}$, whose length ρ is given by*

$$\rho^2 = \mathrm{Tr}\left(\log^2 \frac{1 + R^{1/2}}{1 - R^{1/2}}\right) \tag{11.3.10}$$

where $R = R(Z, Z_1)$ and

$$\log^2 \frac{1 + R^{1/2}}{1 - R^{1/2}} = 4R\left(\sum_{k=0}^{\infty} \frac{R^k}{2k+1}\right)^2 \tag{11.3.11}$$

Theorem 111. *All geodesics are images under symplectic transformations of curves of the type $Z(s) = i\,\mathrm{diag}(p_1^s, \ldots, p_n^s)$ where $p_1, \ldots, p_n$ are arbitrary positive constants satisfying $\sum_{k=1}^{n} \log^2 p_k = 1$.*

All these theorems (and more) were proved in a single paper by Siegel in 1943 [13]. They provide a very different approach to symplectic geometry from the material discussed in Arnol'd's and Moser's books. We take a look at some of the proofs below.

We first find the analogue of the standard conformal map in complex analysis that maps the unit disk to the upper half-plane. The conformal map from $\mathbb{D} \to \mathbb{H}$ is given by

$$\omega(z) = i\frac{(1+z)}{(1-z)} \tag{11.3.12}$$

We define the analogue of the unit disc for complex matrices as follows:

$$\mathcal{D} = \{Z \in \mathcal{S} | I - Z\bar{Z} > 0\} \tag{11.3.13}$$

Note that since $Z = Z^T$, we have $Z^* = \bar{Z}^T = \bar{Z}$. Thus, $Z\bar{Z} = ZZ^*$ is a Hermitian positive definite matrix and the inequality is well defined.

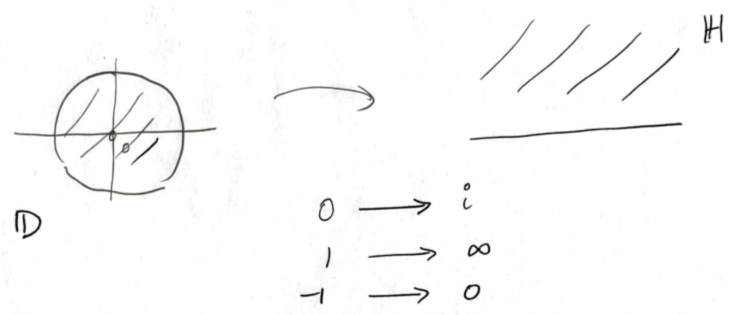

Figure 11.3.1: The conformal map from $\mathbb{D} \to \mathbb{H}$ for the one-dimensional case.

Lemma 44. *The map $W(Z) = i(I+Z)(I-Z)^{-1}$ is an analytic map from $\mathcal{D}$ onto $\mathcal{H}$.*

Proof. The proof of this lemma provides a good introduction to the calculations that underlie Siegel's work. We first show that $(I-Z)^{-1}$ is well defined for all $Z \in \mathcal{D}$.

Suppose $Z \in \mathcal{D}$, and there exists $v \in \mathbb{C}^n$ such that $(I-Z)v = 0$. Clearly, we have $Zv = v$ and $\bar{Z}\bar{v} = \bar{v}$. Thus, $v^T(I - Z\bar{Z})\bar{v} = |v|^2 - |v|^2 = 0$, and since $Z \in \mathcal{D}$, we have $v = 0$. Thus, $(I - Z\bar{Z})^{-1}$ exists and is well defined.

Next, we show that $W(Z) \in \mathcal{H}$. In other words, we need to show that $W = W^T$ and $\mathrm{Im}(W) = (1/2i)(W - \bar{W}) \in \mathbb{P}_n$. First, note that by expressing $(I-Z)^{-1}$ as a Taylor series, we have commutativity as follows: $W = i(I + Z)(I-Z)^{-1} = i(I-Z)^{-1}(I+Z)$. Thus,

$$\begin{aligned} W^T &= i(I+Z)^T(I-Z)^{-T} && (11.3.14)\\ &= i(I+Z^T)(I-Z^T)^{-1} && (11.3.15)\\ &= i(I+Z)(I-Z)^{-1} = W && (11.3.16) \end{aligned}$$

Finally, we write

$$\begin{aligned} \frac{1}{2i}(W-\bar{W}) &= \frac{1}{2i}(i(I-Z)^{-1}(I+Z) - (I+\bar{Z})(I-\bar{Z})^{-1}) && (11.3.17)\\ &= \frac{1}{2}(I-Z)^{-1}((I+Z)(I-\bar{Z}) - (I-Z)(I+\bar{Z}))(I-\bar{Z})^{-1} && (11.3.18)\\ &= \frac{1}{2}(I-Z)^{-1}(2(I-Z\bar{Z}))(I-\bar{Z})^{-1} && (11.3.19)\\ &= (I-Z)^{-1}(I-Z\bar{Z})(I-\bar{Z})^{-1} && (11.3.20) \end{aligned}$$

Thus, if $v \in \mathbb{R}^n$, then

$$\frac{1}{2i}v^T(W-\bar{W})v = v^T(I-Z)^{-1}(I-Z\bar{Z})(I-\bar{Z})^{-1}v \qquad (11.3.21)$$

Choosing $u = (I - \bar{Z})^{-1}v$, we get

$$\frac{1}{2i}v^T(W - \bar{W})v = \bar{u}^T(I - Z\bar{Z})u > 0 \tag{11.3.22}$$

Note that we choose $v \in \mathbb{R}^n$ because $(1/2i)(W - \bar{W}) \in \mathbb{P}_n(\mathbb{R})$, whereas $(I - Z\bar{Z}) \in \mathbb{P}_n(\mathbb{C})$. Thus, we have shown $W \in \mathcal{H}$.

Finally, to show that the image of this map is all of $\mathcal{H}$, we first define a formal inverse and show that it is well defined. For the one-dimensional case, if $z \in \mathbb{D}$, and $\omega = i(1+z)/(1-z)$, then we have $z = (\omega - i)/(\omega + i)$. Thus, for the general matrix case, we define the formal inverse $Z = (W - iI)(W + iI)^{-1}$. It is left as an exercise to show that this is a map from $\mathcal{H} \to \mathcal{D}$. □

We now turn to the proof of theorem 106. The proof requires a generalization of Schwarz's lemma from complex analysis, which is outside the scope of this course. However, we shall prove the following lemma.

Lemma 45. *Suppose* $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in Sp(n)$, *and let* $W = f_M(Z) = (AZ + B)(CZ + D)^{-1}$, *then* $f_M : \mathcal{H} \to \mathcal{H}$.

Proof. We need to show that $(1/2i)(W - \bar{W}) \in \mathbb{P}_n$. Let us build up to this in steps.

Since $M \in Sp(n)$, we have $M^T J M = J$, or

$$\begin{pmatrix} A^T & C^T \\ B^T & D^T \end{pmatrix}\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \tag{11.3.23}$$

$$\implies \begin{pmatrix} -C^T & A^T \\ -D^T & B^T \end{pmatrix}\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \tag{11.3.24}$$

$$\implies C^T A = A^T C, \quad D^T B = B^T D, \quad A^T D - C^T B = I \tag{11.3.25}$$

On the other hand, $M^T \in Sp(n)$, and we also have $MJM^T = J$, or

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}\begin{pmatrix} A^T & C^T \\ B^T & D^T \end{pmatrix} = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \tag{11.3.26}$$

$$\implies \begin{pmatrix} -B & A \\ -D & C \end{pmatrix}\begin{pmatrix} A^T & C^T \\ B^T & D^T \end{pmatrix} = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \tag{11.3.27}$$

$$\implies AB^T = BA^T, \quad CD^T = DC^T, \quad AD^T - BC^T = I \tag{11.3.28}$$

Suppose $Z \in \mathcal{H}$, then we have $Z = Z^T$ and $(1/2i)(Z - \bar{Z}) > 0$. We claim that we can rewrite these conditions as $(Z^T, I)^T J(Z, I) = 0$ and $(1/2i)(Z^T, I)^T J(\bar{Z}, I) > 0$ respectively. Let us check this:

$$\begin{pmatrix} Z^T & I \end{pmatrix}\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}\begin{pmatrix} Z \\ I \end{pmatrix} \tag{11.3.29}$$

$$= \begin{pmatrix} -I & Z^T \end{pmatrix}\begin{pmatrix} Z \\ I \end{pmatrix} = -Z + Z^T = 0 \tag{11.3.30}$$

and

$$\frac{1}{2i}\begin{pmatrix} Z^T & I \end{pmatrix}\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}\begin{pmatrix} \bar{Z} \\ I \end{pmatrix} \tag{11.3.31}$$

$$= \frac{1}{2i}\begin{pmatrix} -I & Z^T \end{pmatrix}\begin{pmatrix} \bar{Z} \\ I \end{pmatrix} = \frac{1}{2i}(Z - \bar{Z}) > 0 \tag{11.3.32}$$

Consider the matrices P and Q defined by $M(Z, I) = (P, Q)$, then we have the identities: $(P^T, Q^T)^T J(P, Q) = 0$ and $(1/2i)(P^T, Q^T)^T J(\bar{P}, \bar{Q}) > 0$.

$$\begin{pmatrix} P^T & Q^T \end{pmatrix}\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}\begin{pmatrix} P \\ Q \end{pmatrix} = \begin{pmatrix} Z^T & I \end{pmatrix} M^T \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} M \begin{pmatrix} Z \\ I \end{pmatrix} \tag{11.3.33}$$

$$= \begin{pmatrix} -I & Z^T \end{pmatrix}\begin{pmatrix} Z \\ I \end{pmatrix} = -Z + Z^T = 0 \tag{11.3.34}$$

and similarly,

$$\frac{1}{2i}\begin{pmatrix} P^T & Q^T \end{pmatrix}\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}\begin{pmatrix} \bar{P} \\ \bar{Q} \end{pmatrix} \tag{11.3.35}$$

$$= \frac{1}{2i}\begin{pmatrix} -I & Z^T \end{pmatrix}\begin{pmatrix} \bar{Z} \\ I \end{pmatrix} = \frac{1}{2i}(Z - \bar{Z}) > 0 \tag{11.3.36}$$

Further, we have the identities $P^T Q = Q^T P$ and $(1/2i)(P^T \bar{Q} - Q^T \bar{P}) > 0$.

$$\begin{aligned}
&P^T Q - Q^T P \\
&= (AZ + B)^T (CZ + D) - (CZ + D)^T (AZ + B) \\
&= (Z^T A^T + B^T)(CZ + D) - (Z^T C^T + D^T)(AZ + B) \\
&= Z^T (A^T C - C^T A) Z + Z^T (A^T D - C^T B) + (B^T C - D^T A) Z + (B^T D - D^T B) \\
&= Z^T - Z = 0
\end{aligned}$$

where we have used that M and M^T are both symplectic, in particular Eqs. 11.3.25 and 11.3.28. Similarly,

$$\begin{aligned}
&P^T \bar{Q} - Q^T \bar{P} \\
&= (AZ + B)^T (C\bar{Z} + D) - (CZ + D)^T (A\bar{Z} + B) \\
&= (Z^T A^T + B^T)(CZ + D) - (Z^T C^T + D^T)(AZ + B) \\
&= Z^T (A^T C - C^T A) \bar{Z} + Z^T (A^T D - C^T B) + (B^T C - D^T A) \bar{Z} + (B^T D - D^T B) \\
&= Z^T - \bar{Z} = Z - \bar{Z}
\end{aligned}$$

where we have Eqs. 11.3.25 and 11.3.28 again. Thus, we have shown $(1/2i)(P^T \bar{Q} - Q^T \bar{P}) = (1/2i)(Z - \bar{Z}) > 0$ for $Z \in \mathcal{H}$.

Next, we show that Q is invertible. Suppose there exists $r \in \mathbb{C}^n$ such that $Qr = 0$, then $\bar{Q}\bar{r} = 0$ and $r^T Q^T = 0$. This implies that $(1/2i) r^T (P^T \bar{Q} - Q^T \bar{P}) \bar{r} = 0$. Since we have positive definiteness, that implies $r = 0$ and thus Q is invertible.

Thus, wrapping everything up, we have shown that $W = PQ^{-1} = (AZ + B)(CZ + D)^{-1}$ is well defined. Further, we have shown that $W - W^T = Q^{-T}(Q^T P - P^T Q)Q^{-1} = 0$ and $(1/2i)(W - \bar{W}) = (1/2i)Q^{-T}(P^T \bar{Q} - Q^T \bar{P})\bar{Q}^{-1} > 0$. Thus, $W \in \mathcal{H}$.

We finally need to show that the map $Z \mapsto W$ maps $\mathcal{H}$ onto $\mathcal{H}$. We leave this an exercise to the reader. The idea is similar to that of Möbius transformations, i.e., to define a formal inverse symplectic transformation and check that it is well defined and maps $\mathcal{H} \to \mathcal{H}$. The formal inverse is given by $Z = f_{M^{-1}}(W) = (D^T W - B^T)(-C^T W + A^T)^{-1}$.

□

11.4 The symplectic group is transitive

Lemma 45 allows us to define a symplectic mapping of the upper half-space $\mathcal{H}$ into itself. The main result of this section is

Theorem 112. *The action of $\Omega := Sp(n)/\{\pm\}$ defined in Lemma 45 is transitive.*

We will approach this theorem in a roundabout way, first characterizing the automorphism group of the Siegel disk $\mathcal{D}$. The main reason for working with the disk $\mathcal{D}$ is that the generalized Schwarz lemma is more transparent on $\mathcal{D}$. This lemma sserts that *every* analytic map from $\mathcal{D}$ onto itself is given by a symplectic mapping. While we will not prove the lemma in its entirety, it becomes more plausible for maps of $\mathcal{D}$. Since Lemma 44 provides an invertible symplectic mapping between $\mathcal{D}$ and $\mathcal{H}$, application of the generalized Schwarz lemma yields the proof of Theorem 106.

We begin by rewriting Lemma 44 in a form that parallels Lemma 45. Define the matrix

$$L = \begin{pmatrix} iI & iI \\ -I & I \end{pmatrix}, \quad \text{with} \quad L^{-1} = \frac{1}{2i}\begin{pmatrix} I & -iI \\ I & iI \end{pmatrix}, \tag{11.4.1}$$

and the associated symplectic map $f_L : \mathcal{D} \to \mathcal{H}$ given by

$$W = f_L(Z) = (iZ + iI)(-Z + I)^{-1}, \quad Z \in \mathcal{D}. \tag{11.4.2}$$

Lemma 46. *The matrix L satisfies the following identities*

$$L^T J L = 2iJ, \quad \frac{1}{2i}L^T J \bar{L} = R, \quad R = \begin{pmatrix} -I & 0 \\ 0 & I \end{pmatrix}. \tag{11.4.3}$$

These are direct computations that are left to the reader. The conditions in equation (11.4.3) allow us to define the natural symplectic automorphisms of $\mathcal{D}$ as follows. Assume $M \in Sp(n)$ and define

$$M_0 = L^{-1}ML := \begin{pmatrix} A_0 & B_0 \\ C_0 & D_0 \end{pmatrix}. \tag{11.4.4}$$

The matrix M_0 is also symplectic, but now in $\mathrm{Sp}(2n;\mathbb{C})$, not $Sp(n)$.

Lemma 47. *Assume $M \in Sp(n)$ and M_0 is given by (11.4.4). Then*

$$M_0^T J M_0 = J, \quad M_0^T R \bar{M}_0 = R, \quad F\bar{M}_0 = M_0 F, \quad F = JR. \tag{11.4.5}$$

Proof. The proof is again a set of direct computations, but let us work out some of the steps. It is immediate that

$$M_0^T J M_0 = L^T M^T L^{-T} J L^{-1} M L.$$

We use the formula for L^{-1} in equation (11.4.1) to obtain the identity $2iL^{-T}JL^{-1} = J$. Therefore, since $M \in Sp(n)$

$$M_0^T J M_0 = \frac{1}{2i} L^T M^T J M L = \frac{1}{2i} L^T M^T J M L = J,$$

where we have applied the first identity in Lemma 46 in the last step. This shows that $M_0 \in \mathrm{Sp}(2n;\mathbb{C})$.

The proof of the second identity is similar. Since $M = \bar{M}$ we have

$$M_0^T R \bar{M}_0 = L^T M^T L^{-T} R \bar{L}^{-1} M \bar{L}.$$

We then use the definition of L and R to compute

$$L^{-T} R \bar{L}^{-1} = \frac{1}{2i} J$$

and another application of Lemma 46 establishes the identity $M_0^T R \bar{M}_0 = R$. The proof of the third identity in equation (11.4.5) is similar and is omitted. $\square$

In the above lemmas, we have used the group $Sp(n)$ to define a new set of matrices $M_0 \in \mathrm{Sp}(2n;\mathbb{C})$ using equation (11.4.4). However, the second and third conditions in equation (11.4.5) restrict the set of complex matrices in $\mathrm{Sp}(2n;\mathbb{C})$ so that we obtain a subgroup of $\mathrm{Sp}(2n;\mathbb{C})$ that generates automorphisms of $\mathcal{D}$. Let us approach these restrictions more directly, writing

$$M_0 = \begin{pmatrix} A_0 & B_0 \\ C_0 & D_0 \end{pmatrix}, \tag{11.4.6}$$

where A_0,B_0, C_0 and D_0 are complex valued matrices. The first set of equations in equation (11.4.5) implies the usual conditions

$$A_0^T C_0 = C_0^T A_0, \quad B_0^T D_0 = D_0^T B_0, \quad A_0^T D_0 - C_0^T B_0 = I. \tag{11.4.7}$$

What is new is the use of the second and third equations in (11.4.5). The condition $M_0^T R \bar{M}_0 = R$ yields the additional equalities

$$A_0^T \bar{A}_0 - C_0^T \bar{C}_0 = I, \quad D_0^T \bar{D}_0 - B_0^T \bar{B}_0 = I, \quad A^T \bar{B}_0 = C_0^T \bar{D}_0. \tag{11.4.8}$$

Finally, we use the relation $F\bar{M}_0 = M_0 F$ to obtain the additional criteria

$$B_0 = \bar{C}_0, \quad A_0 = \bar{D}_0,$$

so that the above matrices satisfy the equation [13, cf.(9)]

$$A_0^T \bar{A}_0 - \bar{B}_0^T B_0 = I, \quad A_0^T \bar{B}_0 = \bar{B}_0^T A_0. \tag{11.4.9}$$

Thus, the set of matrices M_0 given by Lemma 47 takes the form

$$M_0 = \begin{pmatrix} A_0 & B_0 \\ \bar{B}_0 & \bar{A}_0 \end{pmatrix}, \tag{11.4.10}$$

where A_0 and B_0 satisfy equation (11.4.9).

We now see that the automorphism group of the Siegel disk $\mathcal{D}$ may be defined in two equivalent ways.

Definition 113. Let $\Omega_E : \mathcal{D} \to \mathcal{D}$ be the group of symplectic mappings of the form

$$W_0 = f_{M_0}(Z_0) := (A_0 Z_0 + B_0)(\bar{B}_0 Z_0 + \bar{A}_0)^{-1}, \tag{11.4.11}$$

where M_0 satisfies (11.4.9)–(11.4.10). Equivalently, define

$$\Omega_E = L^{-1} \Omega L, \quad \Omega = Sp(n)/\{\pm\}. \tag{11.4.12}$$

We now prove Theorem 112 by establishing

Theorem 114. *The action of Ω_E is transitive on $\mathcal{D}$.*

Proof. Fix $Z_0 = Z_0^T \in \mathcal{D}$. We must find M_0 such that $f_{M_0}(Z_0) = 0$. This implies immediately that

$$A_0 Z_0 + B_0 = 0, \quad \text{or} \quad B_0 = -A_0 Z_0.$$

We also need equation (11.4.9) to hold. Therefore, A_0 must satisfy

$$A_0 (I - Z_0 \bar{Z_0}^T) \bar{A}_0^T = I.$$

The matrix $I - Z_0 \bar{Z_0}^T$ is Hermitian positive definite because $Z_0 \in \mathcal{D}$. Call it $P\bar{P}^T$ and set $A_0 = P^{-1}$. The matrix M_0 defined by this choice of A_0 and B_0 provides the desired transformation in Ω_E. □

The proof reveals the following feature of Ω_E. Assume that $Z_0 = 0$, then $B_0 = 0$ so that

$$W_0 = A_0 Z_0 \bar{A}_0^{-1}, \quad A_0^{-1} = A_0^* = \bar{A}_0^T. \tag{11.4.13}$$

Thus, the set of maps in Ω_E that fix the origin are defined by

$$M_0 = \begin{pmatrix} A_0 & 0 \\ 0 & \bar{A}_0 \end{pmatrix}, \quad A_0 \in U(n). \tag{11.4.14}$$

This leads us to the natural form of

Lemma 48 (Generalized Schwarz lemma for $\mathcal{D}$)**.** *The only analytic maps of $\mathcal{D}$ onto $\mathcal{D}$ that fix the origin are of the form $W_0 = A_0 Z_0 A_0^T$ with $A_0 \in U(n)$.*

The proof of this lemma may be found in [13, pp.11-15].

11.5 The cross-ratio on $\mathcal{H}$

The notion of the cross-ratio of four points may be generalized from the Lobachevsky plane to $\mathcal{H}$ as follows. Define

$$R(Z, Z_1) = (Z - Z_1)(Z - \bar{Z}_1)^{-1}(\bar{Z} - \bar{Z}_1)(\bar{Z} - Z_1)^{-1}. \tag{11.5.1}$$

Theorem 115. *There exists a symplectic mapping taking a pair (Z, Z_1) into another pair (W, W_1) of $\mathcal{H}$ if and only if $R(Z, Z_1)$ and $R(W, W_1)$ have the same eigenvalues.*

The proof of the theorem is an immediate consequence of the following

Lemma 49. *Suppose $M = (AB; CD) \in Sp(n)$ and $W = f_M(Z)$ is the associated symplectic mapping. Then for any $Z_1 \in \mathcal{H}$ and $W_1 = f_M(Z_1)$ we have*

$$R(Z, Z_1) = QR(W, W_1)Q^{-1},$$

where

$$Q = (CZ_1 + D)^T.$$

Proof. We observe that

$$Z_1 - Z = \begin{pmatrix} Z_1 & I \end{pmatrix} \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \begin{pmatrix} Z \\ I \end{pmatrix}.$$

The matrix product on the RHS is

$$\begin{pmatrix} Z_1 & I \end{pmatrix} J \begin{pmatrix} Z \\ I \end{pmatrix} = \begin{pmatrix} Z_1 & I \end{pmatrix} M^T J M \begin{pmatrix} Z \\ I \end{pmatrix},$$

since $M \in Sp(n)$. Now we observe that the last line may be rewritten as

$$(CZ_1 + D)^T (W_1 - W)(CZ + D).$$

This is an interesting computation and it is helpful to check all the details.

$$W = (AZ + B)(CZ + D)^{-1}, \quad W_1 = W_1^T = (CZ_1 + D)^{-T}(AZ_1 + B)^T.$$

Thus, we may rewrite the product $(CZ_1 + D)^T (W_1 - W)(CZ + D)$ as

$$(AZ_1 + B)^T (CZ + D) - (CZ_1 + D)^T (AZ + B).$$

On the other hand,

$$M \begin{pmatrix} Z \\ I \end{pmatrix} = \begin{pmatrix} AZ + B \\ CZ + D \end{pmatrix}, \quad M \begin{pmatrix} Z_1 \\ I \end{pmatrix} = \begin{pmatrix} AZ_1 + B \\ CZ_1 + D \end{pmatrix},$$

so that

$$\begin{pmatrix} Z_1 & I \end{pmatrix} M^T J M \begin{pmatrix} Z \\ I \end{pmatrix} = (AZ_1 + B)^T (CZ + D) - (CZ_1 + D)^T (AZ + B).$$

Now let us turn to the cross-ratios. Notice that the imaginary part

$$\mathrm{Im}(Z - \bar{Z}_1) = Y + Y_1 > 0,$$

since both Y and Y_1 are positive. Therefore, $(Z - \bar{Z}_1)^{-1}$ exists. This yields the identity

$$(Z - Z_1)(Z - \bar{Z}_1)^{-1} = (CZ_1 + D)^T (W - \bar{W}_1)(W - \bar{W}_1)^{-1}(C\bar{Z}_1 + D)^{-T}.$$

Similarly,

$$(\bar{Z} - \bar{Z}_1)(\bar{Z} - Z_1)^{-1} = (C\bar{Z}_1 + D)^T (W - \bar{W}_1)(\bar{W} - W_1)^{-1}(C\bar{Z}_1 + D)^{-T}.$$

We combine the above terms to find

$$R(Z, Z_1) = (CZ_1 + D)^T R(W, W_1)(CZ_1 + D)^{-T}.$$

□

We illustrate Theorem 115 with an important example. Suppose $Z = iI$ and $Z_1 = iT$ where $T = \mathrm{diag}(t_1, \ldots, t_n)$ is a diagonal matrix with increasing entries

$$1 \le t_1 \le \ldots \le t_n.$$

Since diagonal matrices commute, we find that

$$R(Z, Z_1) = (T - 1)^2 (T + 1)^2 = \mathrm{diag}(r_1, \ldots, r_n), \quad r_k = \frac{(t_k - 1)^2}{(t_k + 1)^2}.$$

Clearly, $0 \le r_1 \le \ldots \le r_n < 1$.

This computation covers the general case.

Lemma 50. *Let Z and $Z_1 \in \mathcal{H}$. There exists a symplectic matrix M such that $f_M(Z) = iI$ and $f_M(Z_1) = iT$ with T as above.*

This lemma is proved in [13, Lemma 2]. The main ideas in the proof are as follows. First, it is somewhat simpler to understand the lemma in the setting of the Siegel disk $\mathcal{D}$, rather than the half-plane $\mathcal{H}$. For each T as above and let

$$P = \mathrm{diag}(p_1, \ldots, p_n), \quad p_k = \sqrt{r_k} := \frac{t_k - 1}{t_k + 1}.$$

Then $0 \le p_1 \le \ldots p_n < 1$. Conversely, every such choice of P determines a matrix T. The matrix P determines a matrix $W = U^T P U \in \mathcal{D}$ for each $U \in U(n)$. Indeed, $W^T = W$ and

$$I - W\bar{W} = I - U^T P^2 \bar{U} = U^T (I - P^2 \bar{U}) > 0$$

since $0 \le p_k < 1$. Conversely, every $W \in \mathcal{D}$ admits a representation $W = U^T P U$ (this requires a proof; see [13, Lemma 1]). We then find from Lemma 44 that the transformation $Z = i(I + W)(I - W)^{-1}$ maps $\mathcal{D} \to \mathcal{H}$ and the pair $(0, U^T P U) \to (iI, U^T (it) U)$. Finally, the transformation $\mathcal{H} \to \mathcal{H}$, $Z \mapsto U^T Z U$ is symplectic, so that it does not affect the cross-ratio.

11.6 The symplectic metric on $\mathcal{H}$

The goal of this section is to explain the main ideas in the proofs of Theorem 110 and Theorem 111. By Lemma 50 we see that it is enough to understand these theorems in the standardized setting where $Z = iI$ and $Z_1 = iT$ with $T = \operatorname{diag}(t_1, \ldots, t_n)$ with $1 \leq t_1 \leq \ldots t_n$.

Let us first relate the cross-ratio to the symplectic metric in Definition 109. Fix a point $Z = X + iY \in \mathcal{H}$ and consider a smooth curve $Z(\tau)$ with $Z(0) = Z$ and $\dot{Z}(0) = A$. We then find that

$$\left.\frac{d^2 R(Z, Z(\tau))}{d\tau^2}\right|_{\tau=0} = \frac{1}{2} A Y^{-1} \bar{A} Y^{-1}. \tag{11.6.1}$$

This identity is seen as follows. Write $Z(\tau) \approx Z + \tau A$ and substitute $Z_1 = Z(\tau)$ in equation (11.5.1) to find

$$R(Z, Z(\tau)) \approx (-\tau A)(Z - \bar{Z})^{-1}(-\tau \bar{A})(\bar{Z} - Z)^{-1} = \frac{\tau^2}{4} A Y^{-1} \bar{A} Y^{-1}. \tag{11.6.2}$$

By Lemma 49 the trace of the cross-ratio is invariant under any symplectic mapping $f_M : \mathcal{H} \to \mathcal{H}$ with $M \in Sp(n)$. Thus, so is the metric (11.3.9).

Recall that geodesics are critical points of both the length and action functionals defined as follows. Fix endpoints Z_0 and Z_1 in the Siegel half-space $\mathcal{H}$, consider Lipschitz paths $Z : [0,1] \to \mathcal{H}$ with $Z(0) = Z_0$ and $Z(1) = Z_1$ and the functionals

$$L[Z] = \int_0^1 \left(\operatorname{Tr}(\dot{Z} Y^{-1} \dot{\bar{Z}} Y^{-1} \right)^{1/2} dt, \tag{11.6.3}$$

$$A[Z] = \int_0^1 \operatorname{Tr}(\dot{Z} Y^{-1} \dot{\bar{Z}} Y^{-1} \, dt. \tag{11.6.4}$$

Lemma 51. *The differential equation for geodesics is*

$$\ddot{Z} = -i \dot{Z} Y^{-1} \dot{Z}. \tag{11.6.5}$$

Proof. This equation is obtained by computing the first variation of the action functional. The proof is left to the reader. □

We first solve equation (11.6.5) under the initial conditions

$$Z(0) = iI, \quad \dot{Z}(0) = R = R^T, \quad \operatorname{Tr}(R\bar{R}) = 1. \tag{11.6.6}$$

The condition $R = R^T$ is necessary since $Z(t) = Z^T$. The condition $\operatorname{Tr}(R\bar{R}) = 1$ is a normalization so that geodesics are traveled at constant speed.

Equation (11.6.5) may be solved by analogy with the solution for geodesics in the upper half plane.

Lemma 52. *Assume $G = \operatorname{diag}(g_1, \dots, g_n)$ is a real diagonal matrix such that $\sum_{j=1}^n g_j^2 = 1$. Then the curve*

$$Z(t) = ie^{tG}, \quad t \in \mathbb{R} \tag{11.6.7}$$

is a geodesic.

Proof. The proof is a computation. We find that $Z = iY$ with $Y = e^{tG}$ and

$$\dot{Z} = GZ = ZG, \qquad \ddot{Z} = G^2 Z.$$

Therefore,

$$-i\dot{Z}Y^{-1}\dot{Z} = -iGZ(iZ^{-1})GZ = G^2 Z = \ddot{Z}.$$

□

Now we construct all geodesics through $Z = iI$ using the following observation.

Lemma 53. *Assume $M \in Sp(n)$ is such that the associated symplectic map $f_M(iI) = iI$. Then*

$$M = \begin{pmatrix} A & B \\ -B & A \end{pmatrix}, \quad A + iB \in U(n). \tag{11.6.8}$$

Proof. Let us write

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \quad f_M(Z) = (AZ + B)(CZ + D)^{-1}$$

as usual. Since $M \in Sp(n)$, we know that $A^T C$ and $B^T D$ are symmetric and $A^T D - C^T B = I$. Now impose the condition $f_M(iI) = iI$ to find

$$iA + B = i(iC + D), \quad \text{sothat} \quad A = D, B = -C.$$

The above conditions then also imply

$$A^T A + B^T B = I, \quad A^T B = B^T A,$$

which is equivalent to $A + iB \in U(n)$. □

We combine Lemmas 52 and 53 to obtain

Lemma 54. *Every unit speed geodesic through $Z = iI$ is parametrized by a unitary matrix by $A + iB \in U(n)$ and a real diagonal matrix G with $\operatorname{Tr}(G^2) = 1$ through the implicit formula*

$$(AZ(t) + B)\left(-BZ(t) + A\right)^{-1} = i \operatorname{diag}\left(g_1^t, \dots, g_n^t\right), \quad t \in \mathbb{R}. \tag{11.6.9}$$

Finally, $Sp(n)$ acts transitively on $\mathcal{H}$, geodesics through any point $Z \in \mathcal{H}$ can be obtained by moving Z to the initial point iI and using the above parametrization.

11.7 Gaussian wave packets and the space $\mathcal{H}$

In this section, we explore a connection between the Siegel half-space and the Schrödinger equation.

Given a parametrized curve $Z(t) \in \mathcal{H}$, $t \in \mathbb{R}$, define the *Gaussian wave function* $\psi : \mathbb{R} \times \mathbb{R}^n \to \mathbb{C}$ by

$$\psi(x,t) = \frac{(\det Y)^{1/4}}{\pi^{n/4}} e^{\frac{i}{2}x^T Z x} e^{i\varphi(t)}, \tag{11.7.1}$$

where $\varphi : \mathbb{R} \to \mathbb{R}$ is called the phase. Since $Z = Z^T$ we see that

$$(x^T Z x)^T = x^T Z^T x = x^T Z x$$

is well-defined. The magnitude of the wave-function is

$$|\psi(x,t)|^2 = \psi(x,t)\bar{\psi}(x,t) = \frac{(\det Y)^{1/2}}{\pi^{n/2}} e^{-x^T Y x}. \tag{11.7.2}$$

The normalization factor is chosen so that

$$\int_{\mathbb{R}^n} |\psi(x,t)|^2 \, dx = \frac{(\det Y)^{1/2}}{\pi^{n/2}} \int_{\mathbb{R}^n} e^{-x^T Y x}, dx = 1. \tag{11.7.3}$$

The function ψ describes a quantum wave packet in the following sense.

Theorem 116. *The wave function ψ solves the Schrödinger equation*

$$\partial_t \psi = \frac{i}{2} \triangle \psi, \quad t \in \mathbb{R}, x \in \mathbb{R}^n, \tag{11.7.4}$$

if and only if Z and φ satisfy the ordinary differential equations

$$\dot{Z} = -Z^2, \quad \dot{\varphi} = -\frac{1}{2}\operatorname{Tr}(Y). \tag{11.7.5}$$

The real and imaginary parts of equation (11.7.5) are

$$\dot{X} = Y^2 - X^2, \quad \dot{Y} = -(XY + YX). \tag{11.7.6}$$

Proof. We assume the ansatz (11.7.1) and compute the terms on the LHS and RHS of (11.7.4). First, the spatial derivatives. By definition $\triangle\psi = \nabla \cdot \nabla\psi$. When ψ satisfies (11.7.1) the chain rule implies

$$\nabla\psi = \psi\nabla(\frac{i}{2}x^T Z x) = \psi Z x. \tag{11.7.7}$$

Therefore, we differentiate again to find that

$$\triangle\psi = \nabla \cdot \nabla\psi = \psi\left(-x^T Z^2 x + i\operatorname{Tr}(Z)\right). \tag{11.7.8}$$

On the other hand, the time derivatives are given by

$$\partial_t \psi = \left[\frac{1}{4} \frac{d}{dt} \log \det Y + \frac{i}{2} x^T \dot{Z} x + i\dot{\varphi} \right]. \tag{11.7.9}$$

Further, we have the identity

$$\frac{d}{dt} \log \det(Y(t)) = \mathrm{Tr}\left(Y^{-1} \dot{Y} \right). \tag{11.7.10}$$

In order that $\psi(x,t)$ solve equation (11.7.4), we must balance terms in equations (11.7.8) and (11.7.9). The quadratic expression in x yields the first identity in equation (11.7.5), and thus, equation (11.7.7). The evolution equation for Y now implies that

$$\mathrm{Tr}(Y^{-1}\dot{Y}) = -\mathrm{Tr}(Y^{-1}XY) - \mathrm{Tr}(Y^{-1}YX) = -2\,\mathrm{Tr}(X). \tag{11.7.11}$$

We now balance the remaining terms in (11.7.8) and (11.7.9) to obtain the second equation in equation (11.7.5).

□

Chapter 12

Symplectic manifolds

The previous chapters have provided some fundamental examples of Riemannian manifolds. In this chapter, we turn to Hamiltonian flows on symplectic manifolds. While we continue to stress the role of important examples, we can no longer avoid a rigorous study of differentiable manifolds and the calculus of differential forms. As in Chapter 10 we begin with the algebraic structure.

12.1 Exterior forms

This section follows [3, Ch.7].

Definition 117. A 1-form on $\mathbb{R}^n$ is a linear function $\omega : \mathbb{R}^n \to \mathbb{R}$ such that

$$\omega(au + bv) = a\omega(u) + b\omega(v), \quad a, b \in \mathbb{R}, \quad u, v \in \mathbb{R}^n. \tag{12.1.1}$$

Definition 118. An exterior form of degree k, or a k-form, is a k-linear and antisymmetric function. That is,

$$\omega(a_1u_1 + a_1'u_1', u_2, \ldots, u_k) = a_1\omega(u_1, \ldots, u_k) + a_1'\omega(u_1', \ldots, u_k), \tag{12.1.2}$$

and

$$\omega(u_{i_1}, \ldots, u_{i_k}) = (-1)^\nu \omega(u_1, \ldots, u_k), \tag{12.1.3}$$

where ν is zero if the permutation $i_1, \ldots, i_k$ is even and ν is one if the permutation is odd.

A k-form measures the oriented k-volume of a parallelopiped. Let us illustrate this idea with examples.

Example 16. *The following expression defines a* 2*-form*

$$\omega(u, v) = \begin{vmatrix} u_1 & v_1 \\ u_2 & v_2 \end{vmatrix} = u_1v_2 - u_2v_1. \tag{12.1.4}$$

Example 17. *Let $v \in \mathbb{R}^3$ be a fixed vector. Then for every pair of vectors u,$u' \in \mathbb{R}^3$ we set*

$$\omega(u,u') = \begin{vmatrix} v_1 & u_1 & u_1' \\ v_2 & u_2 & u_2' \\ v_3 & u_3 & u_3' \end{vmatrix}. \tag{12.1.5}$$

This 2-form measures the flux of the vector v across the plane spanned by the vectors u and u'.

Example 18. *Let $k = n$ and assume given n vectors $u_i \in \mathbb{R}^n$ each of which may be written in the standard basis $\{e_j\}_{j=1}^n$ as*

$$u_i = \sum_{i,j} e_j, \quad 1 \leq i \leq n.$$

We define the n-form

$$\omega(u_1,\dots,u_n) = \det \begin{pmatrix} u_{1,1} & \dots & u_{n,1} \\ \vdots & & \vdots \\ u_{n,1} & \dots & u_{n,n} \end{pmatrix}.$$

This n-form measures the oriented volume of the parallelopiped spanned by the vectors $u_1,\dots,u_k$.

Example 19. *Example 18 may be extended to k-forms as follows. Let $S \subset \mathbb{R}^n$ be an oriented k-dimensional subspace and let $P : \mathbb{R}^n \to S$ denote the orthogonal projection onto this subspace. We set*

$$\omega(u_1,\dots,u_k) = \mathrm{vol}_S(Pu_1,\dots,Pu_k), \tag{12.1.6}$$

where vol_S denotes the volume of the k-dimensional oriented parallelopiped spanned by the vectors $Pu_1,\dots,Pu_k$.

12.2 The exterior product

We begin by defining the exterior product of monomials.

Example 20. *Assume given two 1-forms ω_1 and ω_2. Define their exterior product $\omega_1 \wedge \omega_2$ by its action on pairs of vectors*

$$\omega_1 \wedge \omega_2(u,u') = \begin{vmatrix} \omega_1(u) & \omega_2(u) \\ \omega_1(u') & \omega_2(u') \end{vmatrix}. \tag{12.2.1}$$

It is immediate that $\omega_1 \wedge \omega_2$ is a 2-form.

The above example has a natural extension to k-forms.

Example 21. *Assume given k 1-forms $\omega_1, \dots \omega_k$. Define the exterior product*

$$\omega_1 \wedge \dots \omega_k(u_1, \dots, u_k) = \begin{vmatrix} \omega_1(u_1) & \dots & \omega_k(u_1) \\ \vdots & & \vdots \\ \omega_1(u_k) & \dots & \omega_k(u_k) \end{vmatrix}. \tag{12.2.2}$$

It is again immediate that $\omega_1 \wedge \dots \omega_k$ is a k-form. In particular, we observe that the map $\Pi_k : (\omega_1, \dots, \omega_k) \to \omega_1 \wedge \dots \omega_k$ is k-linear and skew-symmetric. That is,

$$\Pi_k : (a\omega_1 + a'\omega_1', \omega_2, \dots, \omega_k) \to a_1 \Pi_k(\omega_1 \dots \omega_k) + a_1' \Pi_k(\omega_1' \dots \omega_k), \tag{12.2.3}$$

and for each permutation σ of indices $(i_1, \dots, i_k)$

$$\omega_{i_1} \wedge \dots \omega_{i_k} = (-1)^\sigma \omega_1 \wedge \dots \wedge \omega_k, \tag{12.2.4}$$

where σ is zero or one depending on whether the permutation is even or odd.

The standard coordinate system $\{x_1, \dots, x_n\}$ on $\mathbb{R}^n$ may now be recognized as a system of 1-forms. Indeed, if $u = (u_1, \dots, u_n) \in \mathbb{R}^n$, then $x_k(u) = u_k$ denotes the k-the component of the vector u. We will also use the notation $\{\partial_{x_1}, \dots, \partial_{x_n})$ for the standard basis of vectors in $\mathbb{R}^n$ (think of this as nothing more than a convenient shift of notation that is better suited to differential geometry).

Example 22. *Example 19 shows that the exterior product $x_{i_1} \dots x_{i_k}$ measures the oriented k-dimensional volume on the subspace spanned by the vectors $\{\partial_{x_{i_1}}, \dots, \partial_{x_{i_k}}$. Observe that $x_{i_1} \wedge \dots \wedge x_{i_k} = 0$ if if two indices agrees and that two such forms are linearly independent when their index sets are not related by a permutation. Thus, the number of such forms is $\binom{n}{k}$. These forms constitute the standard basis for the space of k-forms.*

Once the exterior product is defined for monomials it may be extended to a product of k and l-forms which are themselves products of monomials as follows. Suppose given

$$\omega^k = \omega_1 \wedge \dots \wedge \omega_k, \quad \omega^l = \omega_{k+1} \wedge \dots \omega_{k+l},$$

where ω_i, $1 \le i \le k+l$ are monomials. We then define

$$\omega^k \wedge \omega^l = \omega_1 \wedge \dots \wedge \omega_k \wedge \omega_{k+1} \dots \wedge \omega_{k+1}. \tag{12.2.5}$$

Lemma 55. *The product defined in equation (12.2.5) satisfies*

$$\omega^k \wedge \omega^l = (-1)^{kl} \omega^l \wedge \omega^k, \tag{12.2.6}$$

$$(\omega^k \wedge \omega^l) \wedge \omega^m = \omega^k \wedge (\omega^l \wedge \omega^m). \tag{12.2.7}$$

Finally, we define the exterior product of arbitrary forms.

Definition 119. Suppose given k and l forms on $\mathbb{R}^n$ denoted ω^k and ω^l respectively. Then $\omega^k \wedge \omega^l$ is the $k+l$-form on $\mathbb{R}^n$ defined by

$$\omega^k \wedge \omega^l(u_1, \ldots, u_{k+l}) = \sum (-1)^\sigma \omega^k(u_{i_1}, \ldots, u_{i_k})\omega^l(u_{j_1}, \ldots, u_{j_l}), \tag{12.2.8}$$

where $i_1 < \ldots < i_k$ and $j_1 < \ldots < j_l$ and $(i_1, \ldots, i_k, j_1, \ldots, j_l)$ is a permutation of the integers $(1, \ldots, k+l)$ such that σ is zero or one depending on whether the permutation is even or odd.

12.3 Differentiable manifolds

So far we have been content to work with manifolds embedded in Euclidean space. Now we turn our attention to abstractly defined differentiable manifolds.

Definition 120. A C^k n-dimensional differentiable manifold $\mathcal{M}$ is a topological Hausforff space with a collection of charts and coordinate maps $\{(U_\alpha, \varphi_\alpha)\}_{\alpha \in A}$ such that

1. U_α is an open subset of $\mathcal{M}$ for each α in the index set A.
2. Each coordinate map φ_α is an invertible map from U_α to an open set $V_\alpha \subset \mathbb{R}^n$.
3. The map $\varphi_\beta^{-1} \circ \varphi_\alpha$ is a C^k diffeomorphism between $\varphi_\alpha^{-1}(U_\alpha \cap U_\beta)$ and $\varphi_\beta^{-1}(U_\alpha \cap U_\beta)$ whenever $U_\alpha \cap U_\beta$ is non-empty.

When no explicit value of k is stated the phrase 'smooth' means C^∞.

Definition 121. Assume $\mathcal{M}^m$ and $\mathcal{N}^n$ are smooth m and n dimensional manifolds. A map $f : \mathcal{M}^m \to \mathcal{N}^n$ is C^k if for each $p \in \mathcal{M}$ there are coordinate charts (U, φ) and (V, ψ) containing p and $f(p)$ respectively such that

$$\psi^{-1} \circ f \circ \varphi : \varphi^{-1}(U) \subset \mathbb{R}^m \to \psi^{-1}(V) \subset \mathbb{R}^n$$

is a C^k function.

Definition 120 is subtle. It formalizes the idea that a differentiable manifold is a space that locally looks like Euclidean space *without* assuming that such a space admits a familiar description, for example as the graph of a smooth function. It is necessary to proceed carefully from definitions when defining basic concepts such as the tangent space, vector fields and smooth maps between manifolds. We must also balance the use of coordinates with coordinate free notions. We provide some examples of this interplay. For a complete treatment, see [**?**].

First, let us explain how the tangent space $T_p\mathcal{M}$ at a point $p \in \mathcal{M}$ is defined intrinsically. Each curve $p(t)$ is a map p from an interval of the line, say $(-1, 1)$, to $\mathcal{M}$. The smoothness of this curve is determined using coordinates. Consider a chart U_α that contains $p(0) = p$ and define a curve $x_\alpha(t) \in \varphi_\alpha^{-1}(U_\alpha)$ by setting

$x_\alpha(t) = \varphi^{-1} \circ p(t)$. We say that $p(t)$ is C^k at p if the curve $x_\alpha(t)$ constructed in this manner is C^k at $t = 0$ for every chart containing p. When we consider two such charts, say φ_α and φ_β, we find two different vectors

$$v_\alpha = \frac{d}{dt} x_\alpha(t)\Big|_{t=0}, \quad v_\beta = \frac{d}{dt} x_\beta\Big|_{t=0}.$$

However, the compatibility between charts ensures that

$$v_\beta = D(\varphi_\beta^{-1}\varphi_\alpha)(x_\alpha(0))v_\alpha,$$

for every such pair of curves. Thus, the set of all tangent vectors v_α and v_β at the points $x_\alpha(0)$ and $x_\beta(0)$ is isomorphic.

The tangent space $T_p\mathcal{M}$ is defined to be the equivalence class of vectors under this isomorphism. Thus, we may speak of the tangent vector $\dot{p} \in T_p\mathcal{M}$ for a C^1 curve $p(t)$ in $\mathcal{M}$ with $p(0) = p$. The tangent bundle, denoted $T\mathcal{M}$, is the set

$$T\mathcal{M} = \bigcup_{p \in \mathcal{M}} T_p\mathcal{M},$$

which becomes a differentiable manifold when equipped with the set of charts $(U_\alpha, \varphi_\alpha)$ inherited from the differentiable structure on $\mathcal{M}$, (extended so that each open neighborhood of a point $\tilde{p} \in T\mathcal{M}$ is mapped to an open neighborhood of $\mathbb{R}^n \times \mathbb{R}^n$).

It is also possible to define the tangent space $T_p\mathcal{M}$ functorially by treating vectors as linear operators on $C^\infty(\mathcal{M})$.

Definition 122. A C^k vector field X on $\mathcal{M}$ is a C^k map $p \mapsto T_p\mathcal{M}$.

A vector field may also be characterized by its action on smooth functions. Assume (U, φ) is a chart containing the point p and $\psi \in C^\infty(\mathcal{M})$. In this coordinate chart, the action of X on ψ is defined by

$$(X\psi)(p) = \sum_{j=1}^n f_j(x) \frac{\partial}{\partial x_j} \psi \circ \varphi(x), \quad x \in \varphi^{-1}(U). \tag{12.3.1}$$

Here $(f_1, \ldots, f_n)$ are the components of the vector field and we may identify X with the differential operator given in coordinate charts by

$$\sum_{j=1}^n f_j(x) \frac{\partial}{\partial x_j}.$$

This expression should explain why it is natural to use $\{\partial_{x_j}\}_{j=1}^n$ as the canonical basis on $\mathbb{R}^n$.

12.4 Differential forms

We now extend exterior algebra to the setting of differentiable manifolds. A differential k-form is a k-form on each space $T_p\mathcal{M}$, $p \in \mathcal{M}$ given in coordinates by equation (12.4.4) below. We build up to this equation from simpler notions. In what follows, the term k-form is now used to mean a *differential* k-form.

Assume $\mathcal{M}$ is a C^∞ manifold and let $f \in C^\infty(\mathcal{M})$. We define the differential $df : T\mathcal{M} \to \mathbb{R}$ by its action

$$df(p)v = \left.\frac{d}{dt} f(p(t))\right|_{t=0}, \quad p \in \mathcal{M}, v \in T_p\mathcal{M}, \tag{12.4.1}$$

where $p(t)$ is a smooth curve such that $p(0) = p$ and $\dot{p}(0) = v$.

In order to compare this idea with classical notions, let us assume that $\mathcal{M} = \mathbb{R}^n$, write x for p, and consider the standard bases $\{\partial_{x_j}\}_{j=1}^n$ and $\{dx_k\}_{k=1}^n$ for $T_x\mathbb{R}^n$ and $(T_x\mathbb{R}^n)^*$ at each $x \in \mathbb{R}^n$. We then have

$$df(x)v = \sum_{j=1}^n \frac{\partial f}{\partial x_j}(x) v_j = \sum_{j=1}^n \frac{\partial f}{\partial x_j}(x)\, dx_j(v), \quad v = \sum_{k=1}^n v_k \partial_{x_k},$$

Thus, we may write the differential df as

$$df(x) = \sum_{j=1}^n \frac{\partial f}{\partial x_j}(x)\, dx_j. \tag{12.4.2}$$

Definition 123. A differential 1-form on $\mathcal{M}$ is a map $\omega : T\mathcal{M} \to \mathbb{R}$.

Not every differential 1-form is of the form $\omega = df$. For example, $\omega = dx$ is not the differential of a smooth function on S^1. However, we have

Theorem 124. *Every differential 1-form on $\mathcal{M}$ can be written uniquely in each coordinate chart $(U_\alpha, \varphi_\alpha)$ as*

$$\omega(x) = a_1(x)\, dx_1 + \ldots a_n(x)\, dx_n, \tag{12.4.3}$$

where $\{dx_1, \ldots, dx_n\}$ denotes the dual basis to $\{\partial_{x_1}, \ldots, \partial_{x_n}\}$ in the open set $\varphi_\alpha^{-1}(U_\alpha)$.

The reader is invited to prove this theorem (it amounts to checking consistency of definitions; or see [3, §36]). We say that a differential form ω is C^k if the functions $\{a_j(x)\}_{j=1}^n$ are C^k in each chart. These notions extend naturally to differential k-forms.

Definition 125. A differential k-form ω^k on $\mathcal{M}$ is an assignment of an exterior k-form $\omega^k(p)$ on $T_p\mathcal{M}$ for each point $p \in \mathcal{M}$. The space of differential k-forms is denoted $\Omega^k(\mathcal{M})$.

Remark 126. It is common practice to use the term differential form, rather than differential k-form, when the dimension k is clear from context.

Theorem 124 has the following natural extension to differential k-forms.

Theorem 127. *Every differential k-form on $\mathcal{M}$ can be written uniquely in each coordinate chart $(U_\alpha, \varphi_\alpha)$ as*

$$\omega(x) = \sum_{i_1 < \ldots < i_k} a_{i_1 \ldots i_k}(x)\, dx_{i_1} \wedge \ldots dx_{i_k}. \tag{12.4.4}$$

We say that the form ω^k is C^r if each function $a_{i_1 \ldots i_k}(x)$ is C^r for every coordinate chart $(U_\alpha, \varphi_\alpha)$. A smooth k-form ω^k acts on a collection of smooth vector fields $X_1, \ldots, X_k$ on $\mathcal{M}$ to yield a smooth function that is antisymmetric under permutations

$$\omega^k(X_{i_1}, \ldots, X_{i_k}) = (-1)^\sigma \omega^k(X_1, \ldots, X_k).$$

Definition 128. Suppose $f : \mathcal{M}^m \to \mathcal{N}^n$ is a smooth map between two smooth manifolds. The pullback of a differential k-form ω^k on $\mathcal{N}$ is the differential k-form on $\mathcal{M}$ that acts on a collection of vectors fields $X_1, \ldots, X_k$ on $\mathcal{M}$ as follows

$$(f^*\omega)(p)(X_1, \ldots, X_k) = \omega(f(p))(Df(p)X_1, \ldots, Df(p)X_k), \quad p \in \mathcal{M}. \tag{12.4.5}$$

The definitions upto this point have used only the exterior algebra of forms and the differentiable structure of the manifold $\mathcal{M}$. A new notion is the exterior derivative of a differential form.

Definition 129. The exterior derivative of a k-form $\omega^k \in \Omega^k(\mathcal{M})$, is the $k+1$-form $d\omega^k \in \Omega^{k+1}(\mathcal{M})$ given in each coordinate chart by

$$d\omega^k(x) = \sum_{j; i_1 < \ldots < i_n} \left(\frac{\partial}{\partial x_j} a_{i_1 \ldots i_k}(x) \right) dx_j \wedge dx_{i_1} \wedge \ldots dx_{i_k}. \tag{12.4.6}$$

The exterior derivative has the following fundamental property: for any differential form ω

$$d^2\omega = 0. \tag{12.4.7}$$

Equation (12.4.7) follows immediately from equation (12.4.6) and the antisymmetry of the exterior product.

Definition 130. A differential form ω is closed if $d\omega = 0$.

12.5 Symplectic manifolds

Definition 131. A symplectic structure on an even dimensional manifold $\mathcal{M}^{2n}$ is a 2-form $\omega \in \Omega^2(\mathcal{M})$ such that

1. ω is closed, i.e. $d\omega = 0$.

2. ω is nowhere degenerate.

The pair $(\mathcal{M}, \omega)$ is a symplectic manifold.

Hamiltonian flows and the Poisson bracket on symplectic manifolds may be defined naturally. A Hamiltonian is a sufficiently smooth function $H : \mathcal{M} \to \mathbb{R}$. The symplectic form ω allows us to define a natural duality between the differential dH and a vector field X_H as follows

$$\omega(X_H, v) = dH(v), \quad v \in T\mathcal{M}. \tag{12.5.1}$$

The Poisson bracket between two Hamiltonians H and K on $(\mathcal{M}, \omega)$ is

$$\{H, K\} = \omega(X_H, X_K). \tag{12.5.2}$$

We have suppressed the dependence on $p \in \mathcal{M}$ in equations (12.5.1) and (12.5.2).

Nondegeneracy of the symplectic form is necessary to ensure that X_H and X_K are well-defined. The role of closedness is more subtle. It implies that the Poisson bracket satisfies the Jacobi identity

$$\{\{H, K\}, L\} + \{\{K, L\}, H\} + \{\{L, H\}, K\} = 0, \quad H, K, L \in C^\infty(\mathcal{M}). \tag{12.5.3}$$

Example 23. *Let us revisit the standard example $(\mathbb{R}^{2n}, J)$ in light of the new concepts. Consider the symplectic form*

$$\omega = \sum_{j=1}^{n} dy_j \wedge dx_j. \tag{12.5.4}$$

This form is closed since

$$\omega = d\theta, \quad \theta = \sum_{j=1}^{n} y_j dx_j. \tag{12.5.5}$$

Assume given a pair of vectors $u = (q, p)$ and $v = (s, r) \in \mathbb{R}^{2n}$ where p, q, r and s lie in $\mathbb{R}^n$. Then by definition

$$dx_j(u) = q_j, \quad dy_j(u) = p_j, \quad dx_j(v) = s_j, \quad dy_j(v) = r_j,$$

and we have

$$\omega(u, v) = \sum_{j=1}^{n} \det \begin{pmatrix} p_j & r_j \\ q_j & s_j \end{pmatrix} = \sum_{j=1}^{n} p_j s_j - q_j r_j = -\begin{pmatrix} p & q \end{pmatrix}^T J \begin{pmatrix} s \\ r \end{pmatrix}.$$

The form in equation (12.5.5) differs by a minus sign from our standardization of J. This could be fixed by choosing $\omega = \sum_{j=1}^{n} dx_j \wedge dy_j$ instead. However, the convention in equation (12.5.5) is better suited to the standard form of Theorem 132 below.

Theorem 132. *Assume $\mathcal{M}$ is a smooth manifold. The cotangent bundle $T^*\mathcal{M}$ is a symplectic manifold with the form $\omega = d\theta$, given in coordinates by (12.5.5).*

Proof. Example (23) shows that this form satisfies the local criterion for symplectic form (skew-symmetry and non-degeneracy). What we must show is that θ has a natural global definition on $\mathcal{M}$.

The proof is almost tautological. Let's recall the basic structure

$$TM = \bigcup_{p\in\mathcal{M}} T_p\mathcal{M}, \quad T*\mathcal{M} = \bigcup_{p\in\mathcal{M}} T^*_p\mathcal{M}.$$

Each $\alpha \in T^*\mathcal{M}$ may be written as

$$\alpha = (p, \alpha_p)$$

where p denotes the base point on $\mathcal{M}$ and α_p denotes the exterior form $\alpha_p \in T^*_p\mathcal{M}$. Let $\Pi : T^*\mathcal{M} \to \mathcal{M}$ denote the projection $\Pi(\alpha) = p$ to the basepoint.

We use the coordinate system on $\mathcal{M}$ to define a natural coordinate system on $\mathbb{T}^*\mathcal{M}$ as follows. Any chart (U, φ) containing p provides a coordinate system for the base points, via $p \mapsto (x_1, \ldots, x_n)$. Theorem 124 provides the complementary coordinates for α_p, through the map

$$(p, \alpha_p) \to (x_1, \ldots, x_n, y_1, \ldots, y_n), \quad y_j = \alpha_p(\partial_{x_j}), \quad 1 \le j \le n. \tag{12.5.6}$$

(We've switched notation from Theorem 124, writing α_p for $\omega(x)$, to avoid notational conflict with the symplectic form ω).

Now suppose X is a vector field on $T^*\mathcal{M}$. Then we may write it in local coordinates as

$$X = \sum_{j=1}^n a_j(x,y)\partial_{x_j} + b_j(x,y)\partial_{y_j}.$$

We use $\Pi : T^*\mathcal{M} \to \mathcal{M}$ to push forward the vector field X on $T^*\mathcal{M}$ to obtain the vector field $D\Pi X \in T\mathcal{M}$. This vector field is given in coordinates by

$$D\Pi X = \sum_{j=1}^n a_j(x,0)\partial_{x_j}. \tag{12.5.7}$$

We now define the 1-form θ on $T^*\mathcal{M}$ by setting

$$\theta X(\alpha) = \alpha_p(D\Pi(X)).$$

This form is globally defined by construction. In local coordinates we have

$$\alpha_p(D\Pi(X)) = \sum_{j=1}^n y_j a_j,$$

because of equations (12.5.6) and (12.5.7). But this is exactly the same as writing $\theta = \sum_{j=1}^n y_j dx_j$ in local coordinates. $\square$

12.6 Co-adjoint orbits of Lie groups are symplectic manifolds

There are three main categories of symplectic manifolds. The first of these, the cotangent bundle of a manifold, has been studied in Theorem 132. The second, coadjoint orbits of Lie groups, is studied in this section. The third category, complex projective spaces, is not studied in these notes (see [3, Appendix 3]).

Our treatment of Lie groups thus far has been restricted to important examples. In this section, we assume only that G is a Lie group with finite rank. This means that G is a differentiable manifold that is also equipped with a multiplication operation $G \times G \to G$, $(g, h) \mapsto gh$ that is as smooth as the differentiable structure on the manifold. Since we do not assume that G is a matrix group, we must define the Lie algebra and Lie bracket in an invariant manner. We introduce these ideas below. This is followed by the main theorem, due to Kostant and Kirillov, which asserts that coadjoint orbits of such Lie groups carry a natural symplectic structure. The abstract treatment in this section is matched with examples that illustrate the power of this viewpoint in later sections.

Notation . The following notation is used in this section. The underlying group is G. Elements of G are typically denoted by g and h; the letter e denotes the identity element of G. The Lie algebra of G is denoted $\mathfrak{g}$ and elements of $\mathfrak{g}$ are typically denoted by x and y. The dual space to $\mathfrak{g}$ (as a vector space) is denoted $\mathfrak{g}^*$ and its elements are typically denoted by α and β. The duality pairing between $\mathfrak{g}$ and $\mathfrak{g}^*$ is denoted $\langle \alpha, x \rangle$, $x \in \mathfrak{g}$, $\alpha \in \mathfrak{g}^*$. The role of inner products on $\mathfrak{g}$ is discussed later. The space of linear transformations on a vector space V is denoted $\mathcal{L}(V)$. We will use $V = \mathfrak{g}$ and $V = \mathfrak{g}^*$ below.

12.6.1 The adjoint and coadjoint representations

We first define the inner automorphisms of G. Fix an element $g \in G$ and define the smooth maps corresponding to left and right multiplication by g

$$L_g : G \longrightarrow G, \quad h \mapsto gh, \tag{12.6.1}$$
$$R_g : G \longrightarrow G, \quad h \mapsto hg. \tag{12.6.2}$$

We use these operations to define the *inner automorphism*

$$A_g = L_g R_{g^{-1}}, \quad h \longmapsto ghg^{-1}. \tag{12.6.3}$$

Observe that $A_g(e) = e$. The Lie algebra of G is the tangent space at the identity, T_eG, equipped with a commutator that is obtained by differentiating A_g twice as follows. We fix elements x and y in $\mathfrak{g}$ and consider differentiable paths $g(s)$ and $h(t)$ such that

$$g(0) = h(0) = e, \quad \left.\frac{d}{ds} g(s)\right|_{s=0} = x, \quad \left.\frac{d}{dt} h(t)\right|_{t=0} = y. \tag{12.6.4}$$

Definition 133. The adjoint representation of the Lie group G is the map $G \to \mathcal{L}(\mathfrak{g})$, $g \mapsto \mathrm{Ad}_g$ defined by

$$\mathrm{Ad}_g : \mathfrak{g} \to \mathfrak{g}, \quad y \mapsto \left. \frac{d}{dt} A_g h(t) \right|_{t=0}. \tag{12.6.5}$$

The bracket on the Lie algebra $\mathfrak{g}$ is obtained by differentiating along the curve $g(s)$.

Definition 134. The adjoint representation of the Lie algebra $\mathfrak{g}$ is the map $\mathfrak{g} \to \mathcal{L}(\mathfrak{g})$, $x \mapsto \mathrm{ad}_x$ defined by

$$\mathrm{ad}_x : \mathfrak{g} \to \mathfrak{g}, \quad y \mapsto \left. \frac{d}{ds} \mathrm{Ad}_{g(s)} y \right|_{s=0}. \tag{12.6.6}$$

We have used curves through the identity to define Ad_g and ad_x. The reader should verify that these definitions do not depend in an essential way on the curves; Ad_g is the first derivative of the inner automorphism A_g and ad is the derivative of Ad.

Example 24. *When G is a matrix subgroup of $GL(n)$ we may choose $g(s) = e^{sx}$ and $h(t) = e^{ty}$. We fix s and differentiate with respect to t to obtain that*

$$\mathrm{Ad}_g y = \left. \frac{d}{dt} g e^{ty} g^{-1} \right|_{t=0} = g y g^{-1}. \tag{12.6.7}$$

Next we differentiate $g = g(s)$ with respect to s to obtain

$$\mathrm{ad}_x y = \left. \frac{d}{ds} e^{sx} y e^{-sx} \right|_{s=0} = xy - yx = [x, y]. \tag{12.6.8}$$

Thus, we recover the usual Lie bracket on $GL(n)$. Observe that we do not assume the usual matrix bracket is given; it is obtained from the inner automorphisms. Thus, a different group multiplication rule could give rise to a different bracket, even for matrix groups. This idea will be utilized in the Adler-Kostant-Symes theorem below.

The duality pairing $\langle \cdot, \cdot \rangle$ between $\mathfrak{g}$ and $\mathfrak{g}^*$ may be used to define Ad^*, the co-adjoint representation of G, by

$$\langle \mathrm{Ad}^* g\alpha, x \rangle := \langle \alpha, \mathrm{Ad}_g x \rangle, \quad g \in G, x \in \mathfrak{g}, \alpha \in \mathfrak{g}^*. \tag{12.6.9}$$

Similarly, the co-adjoint representation of $\mathfrak{g}$ is given by

$$\langle \mathrm{ad}^* x\alpha, y \rangle := \langle \alpha, \mathrm{ad}_x y \rangle, \quad x, y \in \mathfrak{g}, \alpha \in \mathfrak{g}^*. \tag{12.6.10}$$

The *Ad* mapping is a representation of the group G in the linear space $\mathcal{L}(\mathfrak{g})$ because of the identity

$$\mathrm{Ad}_{gh} = \mathrm{Ad}_g \mathrm{Ad}_h, \quad g, h \in G. \tag{12.6.11}$$

On the left hand side we have the product in G; on the right hand side, we have the composition of two operators in $\mathcal{L}(\mathfrak{g})$. Similarly, we have

$$\mathrm{Ad}^*_{gh} = \mathrm{Ad}^*_h \mathrm{Ad}^*_g, \quad g, h \in G. \tag{12.6.12}$$

The *Ad*-operation is also an algebra homomorphism

$$\mathrm{Ad}_g \mathrm{ad}_x y = \mathrm{ad}_{\mathrm{Ad}_g x} \mathrm{Ad}_g y. \tag{12.6.13}$$

For a matrix group, the right hand side takes the more intuitive form $[\mathrm{Ad}_g x, \mathrm{Ad}_g y]$. The routine verification of these identities is left to the reader.

12.6.2 The Kostant-Kirillov form

Definition 135. The adjoint and co-adjoint orbits of a Lie group are the submanifolds of $\mathfrak{g}$ and $\mathfrak{g}^*$ defined by

$$O_x := \{\mathrm{Ad}_g x, \ g \in G\}, \quad O^*_\alpha = \{\mathrm{Ad}^*_g \alpha, \ g \in G\}. \tag{12.6.14}$$

The tangent space to O^*_α is computed in equation (12.6.15) below. The tangent space to O_x may be computed similarly.

Theorem 136 (Kostant-Kirillov)**.** *Assume G is a finite-dimensional Lie group. The co-adjoint orbit O^*_α is a symplectic manifold for each $\alpha \in \mathfrak{g}^*$*

This is a striking result because of the minimal assumptions and the strong consequences. For example, we know that a symplectic manifold must be even-dimensional. Yet the theorem makes no assumptions on the dimensionality of G, except for the assumption that G is a finite-dimensional manifold.

Proof. The proof consists of two steps. First, we compute the tangent space to O^*_α. Then we show that it carries a natural 2-form which is non-degenerate and closed.

1. The tangent space to O^*_α is computed as follows. Consider a point $\beta \in O^*_\alpha$. By definition $\beta = \mathrm{Ad}^*_g \alpha$ for some $g \in G$. The tangent space $T_\beta O^*_\alpha$ is the push-forward of $T_g G$. It may be computed by differentiating curves $g(t) \in G$ written in the form

$$g(t) = gh(t), \quad h(0) = e, \quad \dot{h}(0) = x \in \mathfrak{g}.$$

We use identity (12.6.12) to parametrize curves

$$\beta(t) = \mathrm{Ad}^*_{g(t)} \alpha = \mathrm{Ad}^*_{h(t)} \mathrm{Ad}^*_g \alpha = \mathrm{Ad}^*_{h(t)} \beta,$$

and we differentiate with respect to t to obtain the identity

$$T_\beta O^*_\alpha = \{\mathrm{ad}^*_x \beta, \ x \in \mathfrak{g}\}. \tag{12.6.15}$$

2. A symplectic form on $T_\beta O^*_\alpha$ is a closed, non-degenerate 2-form. The existence of such a form requires an inspired guess. What we're seeking is a skew-symmetric map from $T_\beta O^*_\alpha \times T_\beta O^*_\alpha \to \mathbb{R}$. The natural skew-symmetric construct in Lie theory is the ad-operation. Equation (12.6.15) suggests the form

$$\omega(\mathrm{ad}_x^*\beta, \mathrm{ad}_y^*\beta) := -\langle \beta, \mathrm{ad}_x y\rangle = -\langle \mathrm{ad}_x^*\beta, y\rangle, \quad x, y \in \mathfrak{g}, \quad \beta \in O_\alpha^*. \quad (12.6.16)$$

The minus sign is chosen for convenience in the definition of the associated Hamiltonian flows (see equation (12.6.21) below).

We must now check that ω is a symplectic form on O_α^*. Skew-symmetry follows immediately

$$\omega(\mathrm{ad}_x^*\beta, \mathrm{ad}_y^*\beta) = \omega(\mathrm{ad}_y^*\beta, \mathrm{ad}_x^*\beta).$$

We must also check that ω is non-degenerate and closed. The reason non-degeneracy is not immediate is that the linear map $\mathfrak{g} \to T_\beta O_\alpha^*$ given by $x \mapsto \mathrm{ad}_x^*\beta$ may have a non-trivial nullspace $N_\beta = \{x \in \mathfrak{g} | \mathrm{ad}_x^*\beta = 0\}$. If ω is degenerate, there exists $x \in \mathfrak{g}$ such that $\mathrm{ad}_x^*\beta \neq 0$ and

$$\omega(\mathrm{ad}_x^*\beta, v) = 0, \quad v \in T_\beta O_\alpha^*. \quad (12.6.17)$$

Since every $v \in T_\beta O_\alpha^*$ is of the form $v = \mathrm{ad}_y^*\beta$ for some $y \in \mathfrak{g}$, equation (12.6.17) implies that

$$0 = \omega(\mathrm{ad}_x^*\beta, \mathrm{ad}_y^*\beta) = -\langle \beta, \mathrm{ad}_x y\rangle = -\langle \mathrm{ad}_x * \beta, y\rangle.$$

The first equality holds by assumption, the second by the definition of the Kostant-Kirillov form and the third by the definition of the ad^*-action. Since the duality pairing between $\mathfrak{g}$ and $\mathfrak{g}^*$ is non-degenerate, it follows that $\mathrm{ad}_x^*\beta = 0$, contradicting our assumption that ω is degenerate.

Finally, the proof that ω is closed reduces to checking the Jacobi identity for the Poisson bracket, since $\omega(X_H, X_K)(\beta) = 0$ if and only if $\{H, K\} = 0$ for each pair of Hamiltonians H and K. This is a tedious, but direct calculation that is left to the reader. □

12.6.3 Hamiltonian flows on coadjoint orbits

Finally, let us consider the Hamiltonian vector fields defined by the symplection form (12.6.16). Assume given a differentiable Hamiltonian $H : \mathfrak{g}^* \to \mathbb{R}$. Fix $\alpha \in \mathfrak{g}^*$ and consider $\beta \in O_\alpha^*$. The symplectic form ω converts the differential dH into a Hamiltonian vector field X_H on O_α^* through the identity

$$\omega(X_H, v) = dH(\beta)(v), \quad v \in T_\beta O_\alpha^*. \quad (12.6.18)$$

Since every $v \in T_\beta O_\alpha^*$ is of the form $v = \mathrm{ad}_y^*\beta$ for some $y \in \mathfrak{g}$, we may write the above identity in the form

$$\omega(X_H, \mathrm{ad}_y^*\beta) = dH(\beta)(\mathrm{ad}_y^*\beta), \quad y \in \mathfrak{g}. \quad (12.6.19)$$

Since $dH(\beta) \in \mathfrak{g}^{**} \cong \mathfrak{g}$, we may rewrite the right hand side as

$$\begin{aligned} dH(\beta)(\mathrm{ad}_y^*\beta) &= \langle \mathrm{ad}_y^*\beta, dH(\beta)\rangle = \langle \beta, \mathrm{ad}_y dH(\beta)\rangle \\ &= -\langle \beta, \mathrm{ad}_{dH(\beta)} y\rangle = -\langle \mathrm{ad}_{dH(\beta)}^*\beta, y\rangle. \end{aligned} \quad (12.6.20)$$

Comparing equations (12.6.16), (12.6.19) and (12.6.20) we find that the evolution equation

$$\dot{\beta} = X_H(\beta),$$

takes the form

$$\dot{\beta} = \mathrm{ad}^*_{dH(\beta)}\beta, \quad \beta \in O^*_\alpha. \tag{12.6.21}$$

Equation (12.6.21) represents a Hamiltonian flow defined on the basis of geometric and Lie theoretic principles alone. It has a different character from classical models of Hamiltonian systems, such as the equations of planetary motion. The older models are easier to visualize, since they describe the motion of a collection of particles moving according to Newton's laws in three-dimensional space. By contrast, all the 'physics' of equation (12.6.21) is contained within the structure of the coadjoint orbit O^*_α. It is therefore necessary to develop different forms of intuition for equation (12.6.21). We now turn to this topic, explaining the relationship between (12.6.21), and numerical algorithms.

Chapter 13

Algorithms and Integrable systems

The purpose of this chapter is to illustrate the utility of Hamiltonian systems in the study of numerical algorithms. The main ideas are drawn from the theory of integrable systems. As in the previous chapters, we choose to illustrate the ideas with examples. Integrable systems and the Adler-Kostant-Symes (AKS) theorem are introduced through the Toda lattice. The utility of the AKS theorem is then illustrated with the QR algorithm for eigenvalue computation.

13.1 What is an integrable system?

Let us first introduce the concept of integrability in Hamiltonian systems.

Definition 137. Assume $(\mathcal{M}, \omega)$ is a symplectic manifold with dimension $2n$ and $H : \mathcal{M} \to \mathbb{R}$ is a C^2 Hamiltonian. The associated vector field X_H is *integrable* if there are n functions $\{F_j\}_{j=1}^n$ such that

1. The 1-forms $dF_1, \dots, dF_n$ are linearly independent.

2. The functions $\{F_j\}_{j=1}^n$ are integrals for X_H, i.e.

$$\{H, F_j\} = 0, \quad j = 1, \dots, n. \tag{13.1.1}$$

3. The functions $\{F_j\}_{j=1}^n$ Poisson commute, i.e.

$$\{F_j, F_k\} = 0, \quad j, k = 1, \dots, n. \tag{13.1.2}$$

Since the vector X_H is determined by H, we also say that a Hamiltonian is integrable when X_H is integrable. Note that Hamiltonian systems require only half the number of integrals (i.e. n integrals for a $2n$ dimensional system).

Example 25. *Let* $(\mathcal{M}, \omega) = (\mathbb{R}^{2n}, J)$ *and consider the Hamiltonian*

$$H(x, y) = \frac{1}{2} \sum_{j=1}^{n} (y_j^2 + \omega_j^2 x_j^2),$$

where $\omega_j > 0$ *are fixed positive parameters. We choose*

$$F_j = \frac{1}{2}(y_j^2 + \omega_j^2 x_j^2).$$

An easy computation shows that F_j *satisfies the properties of Definition 137 provided we restrict attention to the level sets where* $F_j \neq 0$. *The equations of motion are*

$$\dot{x}_j = y_j, \quad \dot{y}_j = -\omega_j x_j.$$

This equation describes a collection of n *independent harmonic oscillators. It may be solved explicitly. Each oscillator* (x_j, y_j) *admits a parametrization*

$$x_j(t) = r_j \cos(\omega_j t + \phi_j), \quad y_j(t) = r_j \sin(\omega_j t + \phi_j)$$

where $r_j > 0$ *is an ampliture and* ϕ_j *is a phase factor. These terms are called the action and angle respectively and they satisfy the evolution equations*

$$\dot{\phi}_j = \omega_j, \quad \dot{r}_j = 0, \quad j = 1, \ldots, n. \tag{13.1.3}$$

The orbit of each pair $(x_j(t), y_j(t))$ *is a circle with radius* r_j. *The motion of the collection of oscillators* $(x(t), y(t))$ *is quasiperiodic on the torus* $\mathbb{T}^n$ *with radii* $(r_1, \ldots, r_n)$ *(it is quasiperiodic, not periodic, because the frequencies* ω_j *cannot be assumed to be rationally related in general).*

Despite its explicit nature, this example is typical in the following sense. A general theorem of Liouville, with later improvements by Arnold and Jost, asserts that if H is a Hamiltonian on $(\mathcal{M}^{2n}, \omega)$ such that $\{F_j\}_{j=1}^n$ satisfy the conditions of Definition 137 on the level set $\mathcal{N} := \cap_{j=1}^n F_j^{-1}(c_j)$, then there exists a change of variables that transforms the flow on $\mathcal{N}$ to action-angle coordinates [12][Thm 3.4].

The Liouville theorem has a foundational character since it provides an explicit criterion for integrability. However, there are few systematic methods for finding integrals. Further, finding the integrals is not enough, since we often want to explicitly transform the system to action-angle variables, rather than state the existence of such a transformation. In practice, each integrable system has its own individuality and the manner in which it is integrated may rely on clever changes of variables or unexpected connections with other areas of mathematics and science. Here is such an example.

13.2 The Toda lattice

The Toda lattice is a system of interacting particles on the line with the Hamiltonian

$$H(x,y) = \frac{1}{2}\sum_{k=1}^{n} y_k^2 + V(x), \quad V(x) = \sum_{j=1}^{n-1} e^{x_j - x_{j+1}}. \tag{13.2.1}$$

The equations of motion are

$$\ddot{x}_k = e^{x_{k-1}-x_k} - e^{x_k - x_{k+1}}, \quad k = 1,\dots,n. \tag{13.2.2}$$

We assume that $x_1 < x_2 < \ldots < x_n$ and we adopt the convention that $x_0 = -\infty$ and $x_{n+1} = +\infty$ and $e^{-\infty} = 0$ in order to avoid stating separate equations of motion for the endpoints x_1 and x_n. Equation (13.2.2) is derived from equation (13.2.1) in the usual way by setting $\ddot{x}_k = -\partial_{x_k} V(x)$.

The Toda lattice has two obvious integrals: the Hamiltonian H and the total momentum $P = \sum_{j=1}^n \dot{x}_j$. Both of these integrals apply to any potential $V(x)$ that consists of a sum of pairwise interactions depending on the difference $x_j - x_k$. However, the Toda lattice has other surprising integrals, whose existence becomes transparent after the following change of variables introduced by Flaschka. Define the new variables

$$a_k = \frac{1}{2} e^{(a_k - a_{k+1})/2}, \quad b_k = -\frac{1}{2} y_k, \tag{13.2.3}$$

and define the tridiagonal matrices

$$L = \begin{pmatrix} b_1 & a_1 & 0 & 0 \\ a_1 & b_2 & a_2 & \vdots \\ \vdots & \ddots & \ddots & a_{n-1} \\ 0 & \dots & a_{n-1} & b_n \end{pmatrix}, \quad B = \begin{pmatrix} 0 & a_1 & 0 & 0 \\ -a_1 & 0 & a_2 & \vdots \\ \vdots & \ddots & \ddots & a_{n-1} \\ 0 & \dots & -a_{n-1} & 0 \end{pmatrix}. \tag{13.2.4}$$

The equations of motion (13.2.4) may be rewritten in these variables as

$$\dot{L} = [B, L]. \tag{13.2.5}$$

The solution to (**??**) is

$$L(t) = Q(t)L(0)Q(t)^T, \quad \dot{Q} = BQ, \quad Q(0) = I. \tag{13.2.6}$$

The matrix $Q(t)$ is orthogonal since B is antisymmetric. It follows that $L(t)$ has the same eigenvalues as $L(0)$. Thus, the eigenvalues $\{\lambda_j(t)\}_{j=1}^n$ are integrals of motion. Another convenient choice of integrals are the functions

$$F_k(t) = \frac{1}{k}\operatorname{Tr}(L^k(t)) = \frac{1}{k}\sum_{j=1}^{n} \lambda_j^k(t). \tag{13.2.7}$$

The matrices L and B are called the Lax pair for the Toda lattice. Equation (13.2.5) provides an example of an *isospectral deformation*.

There is no algorithm for making a guess such as (13.2.5)! However, it is possible to provide a systematic explanation of the underlying structure and to exploit this structure to integrate the Toda equations. Further, these equations admit several solution procedures. Symmetric tridiagonal matrices such as L admit an inverse spectral theory: if the diagonalization of L is denoted $Q\Lambda Q^T$ then L may itself be reconstructed from the eigenvalues Λ and the top row of the matrix Q. The solution of the Toda lattice using inverse spectral theory is outlined in [12, Ch.3.4]. We will use a different method called the Adler-Kostant-Symes scheme and the R-matrix formalism. This method uses the symplectic geometry developed in the previous chapter. Our first task is to show that equation (13.2.7) is a Hamiltonian system on the $O(n)$ orbit $O_\Lambda := \{Q\Lambda Q^T\}_{Q\in O(n)}$.

13.3 Hamiltonian flows on $GL(n)$ orbits

In this section, we compute the Hamiltonian evolution equations

$$\dot{\alpha} = \mathrm{ad}^*_{dH(\alpha)}\alpha$$

for the group $G = GL(n)$ with Lie algebra $\mathfrak{g} = gl(n)$.

But $gl(n)$ is simply the vector space of all matrices equipped with the commutator $[x, y] = xy - yx$. Thus, we first identify $\mathfrak{g}^*$ with $\mathfrak{g}$ using the pairing provided by the trace. Define the map $\mathfrak{g}^* \to \mathfrak{g}$, $\alpha \mapsto A$ by

$$\langle \alpha, y\rangle = \mathrm{Tr}(A(\alpha)y), \quad y \in gl(n). \tag{13.3.1}$$

This map identifies a unique matrix $A(\alpha)$ for each $\alpha \in \mathfrak{g}^*$. Now that a natural pairing between $\mathfrak{g}$ and $\mathfrak{g}^*$ has been identified, we abuse notation and use α, rather than $A(\alpha)$, to denote both the point $\alpha \in \mathfrak{g}^*$ and its image in $\mathfrak{g}$.

Here is a useful consequence of this identification: for any x and y in $\mathfrak{g}$

$$\langle \mathrm{ad}^*_x\alpha, y\rangle = \langle \alpha, \mathrm{ad}_x y\rangle = \mathrm{Tr}(\alpha[x, y]). \tag{13.3.2}$$

On the other hand,

$$Tr(\alpha[x, y]) = \mathrm{Tr}(\alpha xy - \alpha yx) = \mathrm{Tr}((\alpha x - x\alpha)y) = \mathrm{Tr}([\alpha, x], y).$$

Thus, we have established the identity

$$\mathrm{ad}^*_x\alpha = [\alpha, x], \quad x \in gl(n). \tag{13.3.3}$$

Thus, for each $H : gl(n) \to \mathbb{R}$, we obtain the Hamiltonian flow

$$\dot{\alpha} = [\alpha, dH(\alpha)], \quad x \in gl(n). \tag{13.3.4}$$

By definition, the differential of F_k is computed as follows:

$$dH(\alpha)(\beta) = \frac{d}{d\tau} H(\alpha(\tau))\bigg|_{\tau=0},$$

where $\alpha(\tau)$ is a smooth curve with $\alpha(0) = \alpha$ and $\dot{\alpha}(0) = \beta$.

Example 26. *Consider the Hamiltonians*

$$F_k(\alpha) = \frac{1}{k+1}\operatorname{Tr}(\alpha^{k+1}).$$

It is easy to check that

$$dF_k(\alpha)(\beta) = \frac{1}{k+1}\operatorname{Tr}\left(\beta\alpha^k + \alpha\beta\alpha^{k-1} + \dots \alpha^k\beta\right) = \operatorname{Tr}(\alpha^k\beta).$$

Therefore, $dF_k(\alpha) = \alpha^k$ *and the associated Hamiltonian flow is*

$$\dot{\alpha} = [\alpha, dF_k(\alpha)] = [\alpha, \alpha^k] = 0,$$

since α *and* $\alpha^k = 0$.

This computation is slightly disconcerting since it shows us that a natural class of Hamiltonians generates only trivial flows. However, it lies at the heart of the AKS theorems and the QR algorithm, since the same manifold may carry different multiplication structures!

13.4 The QR group

Bibliography

[1] V. I. Arnol'd, *Ordinary differential equations*, MIT Press, Cambridge, Mass.-London, 1978. Translated from the Russian and edited by Richard A. Silverman.

[2] ——, *Geometrical methods in the theory of ordinary differential equations*, vol. 250 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer-Verlag, New York, second ed., 1988. Translated from the Russian by Joseph Szücs [József M. Szűcs].

[3] ——, *Mathematical methods of classical mechanics*, vol. 60 of Graduate Texts in Mathematics, Springer-Verlag, New York, [1989]. Translated from the 1974 Russian original by K. Vogtmann and A. Weinstein, Corrected reprint of the second (1989) edition.

[4] V. I. Arnold and A. Avez, *Ergodic Problems of Classical Mechanics, reprint*, Addison-Wesley, Redwood City, CA, 1989.

[5] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, Cambridge, 2004.

[6] C. Chicone, *Ordinary differential equations with applications*, vol. 34, Springer Science & Business Media, 2006.

[7] J. Guckenheimer and P. Holmes, *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, vol. 42 of Applied Mathematical Sciences, Springer-Verlag, New York, 1990. Revised and corrected reprint of the 1983 original.

[8] V. Guillemin and A. Pollack, *Differential topology*, vol. 370, AMS Chelsea Publishing, 1974.

[9] J. K. Hale, *Ordinary differential equations*, Robert E. Krieger Publishing Co., Inc., Huntington, N.Y., second ed., 1980.

[10] J. K. Hale and H. Koçak, *Dynamics and bifurcations*, vol. 3 of Texts in Applied Mathematics, Springer-Verlag, New York, 1991.

[11] N. Karmarkar, *A new polynomial-time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.

[12] J. Moser and E. J. Zehnder, *Notes on dynamical systems*, vol. 12 of Courant Lecture Notes in Mathematics, American Mathematical Society, Providence, RI, 2005.

[13] C. L. Siegel, *Symplectic geometry*, American Journal of Mathematics, 65 (1943), pp. 1–86.

[14] S. H. Strogatz, *Nonlinear dynamics and chaos*, Westview Press, Boulder, CO, second ed., 2015. With applications to physics, biology, chemistry, and engineering.

[15] L. N. Trefethen and D. Bau, III, *Numerical linear algebra*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[16] H. Weyl, *The classical groups: their invariants and representations*, vol. 45, Princeton University Press, 1946.