Math 6324 – Differential Equations – Lecture Notes

Vaughn Climenhaga, University of Houston, Spring 2026

Chapter 1: Overview and basic theory

1.1 Introduction

Lec 1
W 1/21

Undergraduate DE courses often focus on techniques to find formulas for the solutions of initial value problems. This is only possible for a limited class of systems.

- We will focus on qualitative behavior, including: existence and uniqueness of solutions; stability of fixed points and periodic orbits; response to perturbations.
- Some natural topics beyond the scope of this course will be mentioned but not explored, including: numerical solutions; weak solutions; general PDE theory.

We will introduce various phenomena via specific examples, many of which correspond to important mathematical models. Here is an overview, for later reference.

DE	Phenomenon	Model
$\dot{x} = \pm x$	exponential growth/decay	
$\dot{x} = x(1-x)$	fixed point stability	logistic growth
$\ddot{x} + \sin x = 0$	integrals of motion	pendulum
$\ddot{x} + x = 0$	rotation	harmonic oscillator
$\ddot{x} + b\dot{x} + x = A\sin(\omega t)$	dissipation; limit cycles	forced damped oscillator
$\ddot{x} - \mu(1-x^2)\dot{x} + x = A\sin(\omega t)$	horseshoe	Van der Pol oscillator
$\dot{x} = \sigma(y-x)$ $\dot{y} = x(\rho - z) - y$ $\dot{z} = xy - \beta z$	chaotic attractor	Lorenz system
	KAM theory	three-body problem

In what follows, **Teschl** *refers to "Ordinary Differential Equations and Dynamical Systems", by Gerald Teschl, AMS GSM vol. 140, 2012, and* **Hirsch–Smale** *refers to "Differential Equations, Dynamical Systems, and Linear Algebra", by Morris W. Hirsch and Stephen Smale, Academic Press, 1974.*

1.2 General setup and basic phenomena

We focus on first-order systems of d DEs, given by specifying an interval $I \subset \mathbb{R}$, a domain[1] $D \subset \mathbb{R}^d$, and a (possibly time-dependent) vector field $F\colon I \times D \to \mathbb{R}^d$. Given a differentiable function $x\colon I \to D$, we write $\dot{x} = \frac{dx}{dt}$ for its time derivative. Given $t_0 \in I$ and $x_0 \in D$, we will study functions x that are *solutions of the initial value problem (IVP)*

$$\dot{x} = F(t, x), \qquad x(t_0) = x_0. \tag{1.1}$$

Conceptually, the first instance of $\mathbb{R}^d$, where x lies, represents a space of *positions*, while the second instance, where $F(t, x)$ lies, represents a space of *vectors*.[2]

Higher-order DEs (in terms of $\ddot{x}$, $\dddot{x}$, etc.) can be put into this framework by including derivatives as extra coordinates in x. By including a coordinate with derivative 1, we can always reformulate (1.1) to be *autonomous*, meaning that $F\colon D \to \mathbb{R}^d$ only depends on x.

Recall some basic examples, where for simplicity we take $t_0 = 0$.

- $\dot{x} = 0 \quad \Rightarrow \quad x(t) = x_0$ (constant solution)
- $\dot{x} = x \quad \Rightarrow \quad x(t) = x_0 e^t$ (exponential growth, divergence of trajectories)
- $\dot{x} = -x \quad \Rightarrow \quad x(t) = x_0 e^{-t}$ (exponential decay, convergence of trajectories)

Calculus tells us that if $\dot{x}(t) = 0$ for every t in an interval I, then x is constant on I.[3] So the IVP for $\dot{x} = 0$ has a *unique* solution.

- Given $a \in \mathbb{R}$, the IVP for $\dot{x} = ax$ has a solution $x(t) = x_0 e^{at}$, which is unique since $\frac{d}{dt}(xe^{at}) = 0$.

We see four phenomena: solution exists; solution is unique; solution is globally defined ($I = \mathbb{R}$); solution admits an explicit formula. Each of these can fail in general…

(1) Consider $\dot{x} = F(x) := 1 + \mathbf{1}_{[0,\infty)}$, so $F|_{(-\infty,0)} \equiv 1$ and $F|_{[0,\infty)} \equiv 2$. With initial condition $x(0) = 0$, there is no differentiable solution on any interval around 0.

(2) Consider $\dot{x} = \sqrt{x}$. This has a constant solution $x(t) = 0$, but also $x(t) = \frac{1}{4}t^2$ for $t \geq 0$. So uniqueness can fail. Up to time reversal and the power involved, this DE models a leaky bucket: if we find the bucket empty, we cannot tell when the last water left it.

(3) Consider $\dot{x} = x^2$ with $x(0) = 1$. This has unique solution $x(t) = (1-t)^{-1}$, which goes to ∞ as $t \nearrow 1$, so the solution is defined on $(-\infty, 1)$ but not on all of $\mathbb{R}$.

[1] The domain D will generally be either an open set or the closure of an open set.

[2] More generally, we could speak of a smooth manifold and its tangent bundle.

[3] But real analysis tells us that there are nonconstant functions such as the *devil's staircase* (Cantor function) whose derivative vanishes "almost everywhere". So care is needed if we decide to go beyond the class of everywhere differentiable functions.

(4) Consider $\dot{x} = x^3+1$. This has no "closed form solution"; or so ChatGPT tells me, referring to Malmquist's Theorem and the books of E.L. Ince and G.M. Murphy. Basically the issue is that after we separate variables to get $\frac{dx}{x^3+1} = dt$ and then integrate via partial fractions, the left-hand side involves both logarithms and inverse trigonometric functions, and we cannot solve for x in terms of t using standard functions.

In example 1 of the list above, it feels like $x(t) = tF(t)$ should be a solution, but $\dot{x}(0)$ does not exist. One way out might be to reformulate the problem: observe that in general, if $x\colon I \to \mathbb{R}$ is differentiable with $x(t_0) = x_0$, then

$$\dot{x} = F(x) \quad \Leftrightarrow \quad x(t) - x_0 = \int_{t_0}^{t} F(x(s))\, ds \text{ for all } t \in I. \tag{1.2}$$

If a non-differentiable function satisfies the integral form, we could consider it as a "weak" solution. Similar ideas become important in PDE. We will not explore them in this course, and will limit our attention to the setting in which F is continuous and x is C^1 (continuously differentiable). However, the integral form in (1.2) will play an important role in the next sections. In particular, using this equivalent criterion for solutions, we will show in §1.4 that discontinuity of the vector field is the *only* way that existence can fail:

Peano's theorem (informal). If F is continuous, then $\dot{x} = F(t, x)$ has a solution.

1.3 Euler's method

Lec 2
F 1/23

This section draws on §I.7 of "Solving Ordinary Differential Equations I: Nonstiff Problems", by Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner, Springer, 1993.

Even without an explicit formula for a solution, we can still approximate it numerically. The simplest way to do this is *Euler's method*, which relies on the observation that if $x\colon I \to \mathbb{R}^d$ is a solution of (1.1), then for every $h > 0$ and every $t \in I$ such that $t + h \in I$, we have[4]

$$x(t + h) = x(t) + F(t, x(t))h + o(h) \quad \text{as } h \to 0.$$

With this as motivation, Euler's method provides a family of piecewise linear functions as approximate solutions: for each $h > 0$, define $x^h\colon I \to D$ iteratively as follows.[5]

- Put $x^h(t_0) = x_0$.
- For each $k \in \mathbb{N}$, write $t_k := t_0 + kh$ and put $x^h(t_{k+1}) := x^h(t_k) + F(t_k, x^h(t_k))$.
- Define x^h on the intervals (t_k, t_{k+1}) by linear interpolation.

[4]Recall Landau's "little-o" notation: $f(h) = g(h) + o(h)$ means $\frac{1}{h}|f(h) - g(h)| \to 0$ as $h \to 0$.

[5]If you are reading closely, you may notice that x^h could leave D, in which case we would no longer be able to proceed. So in fact, each x^h may only be defined on a smaller interval in I.

This gives an approximate solution for $t \geq t_0$; the procedure for $t \leq t_0$ is analogous, and we will not discuss it. Similarly, in our existence and uniqueness results later, we will focus on the behavior moving forward in time, with the understanding that the corresponding results also hold moving backwards in time.

Question: Do the approximations x^h converge to a solution as $h \to 0$?

.....**Answer:** Not necessarily!

We now present a (non-autonomous) DE for which the approximations $x^h(t)$ do not converge at any time $t > t_0$ as $h \to 0$. This example, which is somewhat counter-intuitive, is a slight modification of (I.7.19) in Hairer–Nørsett–Wanner.[6] Note that the time-dependence could be eliminated at the cost of working with a 2-dimensional system, by introducing a second coordinate y that satisfies $\dot{y} = 1$.

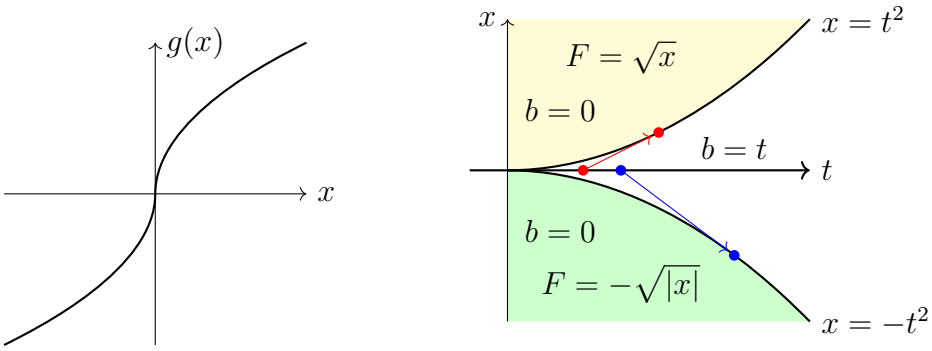

Figure 1.1: Defining an example where Euler's method fails to converge.

Example 1.1. Consider the IVP (1.1) with $d = 1$, $x_0 = t_0 = 0$, and F defined as follows. Let $g(x) = (\operatorname{sgn} x)\sqrt{|x|}$, and let $b(t, x)$ be a continuous function such that

$$b(t, 0) = t \quad \text{and} \quad b(t, x) = 0 \text{ for all } |x| \geq t^2,$$

as shown in Figure 1.1. Then define F by

$$F(t, x) = 4\big(g(x) + b(t, x)\cos(1/t)\big). \tag{1.3}$$

In particular, observe that F is continuous, and that for all t, we have

$$F(t, 0) = 4t\cos(1/t) \quad \text{and} \quad F(t, \pm x) = \pm 4\sqrt{x} \text{ for all } x \geq t^2.$$

From this we see that (1.1) has both $x = 4t^2$ and $x = -4t^2$ as solutions (along with many others), but we will not need this. For any fixed value of $h > 0$, Euler's method starts with

$$(t_0, x_0) = (0, 0), \qquad (t_1, x_1) = (h, 0), \qquad (t_2, x_2) = \big(2h, 4h^2\cos(1/h)\big).$$

[6] Another example, in a similar spirit, can be found in Exercise 1.12 of "Theory of Ordinary Differential Equations", by Earl Coddington and Norman Levinson, 1955.

When $h = h_k = \frac{1}{k\pi}$ for some $k \in \mathbb{N}$, we have $\cos(1/h) = \cos(k\pi) = (-1)^k$, and thus

$$(t_2, x_2) = (2h_k, 4h_k(-1)^k) \qquad \Rightarrow \qquad t_2^2 = 4h_k^2 = |x_2|, \tag{1.4}$$

so (t_2, x_2) lies in one of the two shaded regions in Figure 1.1. This implies that all successive (t_n, x_n) must remain in that region: indeed, if $x_n \geq t_n^2$ for some $n \geq 2$, then

$$x_{n+1} = x_n + h \cdot 4\sqrt{x_n} \geq t_n^2 + 4ht_n = (nh)^2 + 4h(nh) = (n^2+4n)h^2 \geq (n+1)^2h^2 = t_{n+1}^2,$$

so the inequality holds for all $n \geq 2$ by induction, and a similar argument applies if $x_n \leq -t_n^2$. Combined with (1.4), this implies that whenever $h = \frac{1}{k\pi}$ for some $k \in \mathbb{N}$, the approximate solution evaluated at $t > 0$ satisfies $x^h(t) \geq t^2$ when k is large and even, and $x^h(t) \leq -t^2$ when k is large and odd. It follows that

$$\overline{\lim_{h\to 0}}\, x^h(t) \geq t^2 \quad \text{and} \quad \underline{\lim_{h\to 0}}\, x^h(t) \leq -t^2.$$

In particular, $\lim_{h\to 0} x^h(t)$ does not exist for any $t > 0$.

The rather complicated definition of $F(t,x)$ in Example 1.1 suggests that Euler's method may still converge for many "nicer" examples. This is indeed the case, as we will now see.

Lec 3
W 1/28

Let us begin by describing a slight generalization of Euler's method, where the time steps may be of different lengths.

Assumption 1.2. *Fix times $t_0 \in \mathbb{R}$ and $T_* > t_0$, a point $x_0 \in \mathbb{R}^d$, and a radius $r > 0$. Let $D := \overline{B(x_0, r)} \subset \mathbb{R}^d$. Suppose that $F\colon [t_0, T_*] \times D \to \mathbb{R}^d$ is continuous. Let $M > 0$ be such that $\|F(t,x)\| \leq M$ for all $(t,x) \in [t_0, T_*] \times D$. (Such an M exists by the extreme value theorem.) Let $T := \min(T_*, t_0 + r/M)$, and consider the interval $I := [t_0, T]$.*

A *partition* of the interval $I = [t_0, T]$ will mean a finite subset $P \subset (t_0, T)$; writing the elements of P in increasing order as $t_1 < t_2 < \cdots < t_\ell$ and putting $t_{\ell+1} = T$, we see that P divides $[t_0, T]$ into finitely many subintervals $[t_i, t_{i+1}]$ for $0 \leq i \leq \ell$. Writing $D_i := \overline{B(x_0, M(t_i - t_0))} \subset D$, consider for each $0 \leq i \leq \ell$ the map

$$\begin{aligned} f_i = f_i^P\colon D_i &\to D_{i+1} \\ z &\mapsto z + F(t_i, z)(t_{i+1} - t_i) \end{aligned} \tag{1.5}$$

that represents one step of Euler's method. Then the approximate solution associated to the partition P is defined iteratively at each t_i by taking $x^P(t_0) = x_0$ and

$$x^P(t_{i+1}) = f_i^P(x^P(t_i)) \quad \text{for all } 0 \leq i \leq \ell.$$

Extend to a function $x^P\colon I \to D$ by linear interpolation on each (t_i, t_{i+1}).

If $h > 0$ and P consists of all points of the form $t_0 + nh$ for $1 \leq n < (T - t_0)/h$, then x^P is just the approximation x^h from Euler's method.

Define the *mesh* of a partition P to be

$$|P| := \max_{0 \le i \le \ell} |t_{i+1} - t_i|.$$

We want a condition on F under which $|x^P(T) - x^Q(T)|$ is small whenever P and Q are partitions with small mesh. It will suffice to consider the case when $Q \supset P$, since then in the general case we can use the triangle inequality to write

$$|x^P(T) - x^Q(T)| \le |x^P(T) - x^{P \cup Q}(T)| + |x^{P \cup Q}(T) - x^Q(T)|.$$

With this in mind, we assume $Q \supset P$, and write the elements of P as $t_1 < t_2 < \cdots < t_\ell$. Define a sequence of partitions

$$P = P(0) \subset P(1) \subset P(2) \subset \cdots \subset P(\ell) = Q$$

by putting

$$P(i) := P \cup \big(Q \cap (t_0, t_i)\big)$$

for the partition obtained by taking P and adding all the points in Q lying in the first i subintervals determined by P. Then the triangle inequality gives

$$\|x^P(T) - x^Q(T)| = \Big\| \sum_{i=1}^{\ell+1} x^{P(i)}(T) - x^{P(i-1)}(T) \Big\| \le \sum_{i=1}^{\ell+1} \|x^{P(i)}(T) - x^{P(i-1)}(T)\|. \quad (1.6)$$

See Figure 1.2 for an illustration of the points in (1.6).

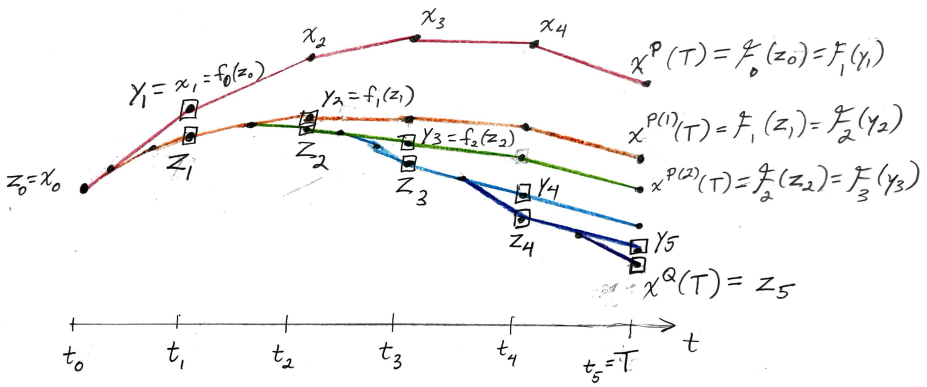

Figure 1.2: Proving convergence of Euler's method. In the picture we have $\ell = 4$.

To obtain the desired estimates, start by letting $z_i = x^Q(t_i)$ for each i. Recalling the definition of $f_i = f_i^P$ from (1.5), let $y_{i+1} = f_i(z_i)$, and consider for each i the map

$$\mathcal{F}_i := f_\ell \circ f_{\ell-1} \circ \cdots \circ f_{i+1} \circ f_i \colon D_i \to D,$$

which maps each $z \in D_i$ to the point that would be reached at time T by an approximate solution starting at (t_i, z) and making its linear approximations at the times specified by P. Using this notation, we see that

$$x^{P(i)}(T) = \mathcal{F}_i(z_i) = \mathcal{F}_{i+1}(y_{i+1})$$

for each $0 \le i \le \ell + 1$. For $1 \le i \le \ell + 1$, the second of these gives $x^{P(i-1)}(T) = \mathcal{F}_i(y_i)$, and now (1.6) yields

$$\|x^P(T) - x^Q(T)\| \le \sum_{i=1}^{\ell+1} \|\mathcal{F}_i(z_i) - \mathcal{F}_i(y_i)\|. \tag{1.7}$$

To estimate the terms in the sum, we must obtain two upper bounds:

1. a bound on the displacement between the two points z_i and y_i, which are both obtained from z_{i-1} via linear approximations;
2. a bound on how much this displacement grows as we evolve from time t_i to time T via the map $\mathcal{F}_i$.

These require some notation for the modulus of continuity of the vector field. For the first estimate, consider for each $\delta > 0$ the quantity

$$\zeta(\delta) := \sup\{\|F(t,z) - F(s,y)\| : t, s \in I,\ y, z \in D,\ |t-s| \le \delta,\ \|y - z\| \le M\delta\}, \tag{1.8}$$

which goes to 0 as $\delta \to 0$ since F is uniformly continuous on $I \times D$. In particular, for each time $s \in Q \cap (t_{i-1}, t_i)$, we have $|s - t_{i-1}| \le |P|$ and

$$\|x^Q(s) - x^Q(t_{i-1})\| \le M|s - t_{i-1}| \le M|P|,$$

from which the definition in (1.8) gives

$$\|F(s, x^Q(s)) - F(t_{i-1}, z_{i-1})\| \le \zeta(|P|),$$

and we conclude that

$$\epsilon_i := \|z_i - y_i\| \le \zeta(|P|)(t_i - t_{i-1}). \tag{1.9}$$

For the bound on the growth of the displacement, we only need the modulus of continuity with respect to x: for each $\epsilon > 0$, put

Lec 4
F 1/30

$$\omega(\epsilon) := \sup\{\|F(t,z) - F(t,y)\| : t \in I,\ y, z \in D,\ \|y - z\| \le \epsilon\}.$$

With this notation, and writing $\tau_i := t_{i+1} - t_i$, we have

$$\begin{aligned} \|f_i(z_i) - f_i(y_i)\| &= \big\|\big(z_i + F(t_i, z_i)\tau_i\big) - \big(y_i + F(t_i, y_i)\tau_i\big)\big\| \\ &\le \|z_i - y_i\| + \|F(t_i, z_i) - F(t_i, y_i)\|\tau_i \\ &\le \epsilon_i + \omega(\epsilon_i)\tau_i. \end{aligned} \tag{1.10}$$

It is at this point that our general setting of "F is only assumed to be continuous" begins to fail us. Without further information on F, all we know is that $\omega(\epsilon) \to 0$ as $\epsilon \to 0$, which is not enough to let us effectively iterate the bound in (1.10). However, if $\omega(\epsilon)$ is bounded above by a (uniform) multiple of ϵ, then we could simplify and iterate this inequality:

Definition 1.3. The vector field $F\colon I \times D \to \mathbb{R}^d$ is *uniformly Lipschitz in x* if there exists $L > 0$ such that for every $t \in I$ and every $y, z \in D$, we have

$$\|F(t,y) - F(t,z)\| \le L\|y - z\|. \tag{1.11}$$

In this case, L is called a *Lipschitz constant* for F.

When the vector field is uniformly Lipschitz in x, (1.10) simplifies and can be iterated to give the following result on divergence of approximate trajectories with nearby initial conditions.

Lemma 1.4. *Consider an interval $I = [t_0, T] \subset \mathbb{R}$, two points $y, z \in \mathbb{R}^d$, a set $D \subset \mathbb{R}^d$ that contains the closed balls of radius $(T-t_0)M$ around both y_0 and z_0 for some $M > 0$, and a vector field $F\colon I \times D \to \mathbb{R}^d$ such that $\|F(t,x)\| \le M$ for all $(t,x) \in I \times D$. Let P be an arbitrary partition of I, and let $y^P, z^P\colon I \to D$ be the piecewise linear curves determined by applying Euler's method with the partition P and the initial conditions y_0 and z_0, respectively.*

If F is uniformly Lipschitz in x with constant L, then we have

$$\|y^P(T) - z^P(T)\| \le e^{L(T-t_0)}\|y_0 - z_0\|. \tag{1.12}$$

Proof. Enumerate the elements of P as $t_1 < t_2 < \cdots < t_\ell$. Arguing exactly as in (1.10), and using the fact that $\omega(\epsilon) \le L\epsilon$, we see that for each $0 \le i \le \ell$, we have

$$\begin{aligned}\|y^P(t_{i+1}) - z^P(t_{i+1})\| &\le (1 + L(t_{i+1} - t_i))\|y^P(t_i) - z^P(t_i)\| \\ &\le e^{L(t_{i+1}-t_i)}\|y^P(t_i) - z^P(t_i)\|.\end{aligned}$$

Iterating this inequality proves (1.12). □

Now we can prove that Euler's method converges to a solution of the IVP (1.1).

Proposition 1.5 (Convergence of Euler's method)**.** *Let t_0, T, x_0, D, F, I be as in Assumption 1.2, and suppose in addition that $F\colon I \times D \to \mathbb{R}^d$ is uniformly Lipschitz in x. Then the limit $x(t) := \lim_{h\to 0} x^h(t)$ exists and lies in D for every $t \in I$, and the approximate solutions x^P converge uniformly to x: for every $\epsilon > 0$, there exists $\delta > 0$ such that given any partition P of I with mesh $|P| \le \delta$, we have $\|x^P(t) - x(t)\| \le \epsilon$ for every $t \in I$.*

Proposition 1.6. *In the setting of Proposition 1.5, the limit $x\colon I \to D$ solves* (1.1).

Proof of Proposition 1.5. Most of the organizational work for the proof has already been done in the preceding discussion. Recalling the bound in (1.9) on $\|z_i - y_i\|$, we can apply Lemma 1.4 to the interval $[t_i, T]$ with the initial conditions y_i and z_i to obtain

$$\|\mathcal{F}_i(z_i) - \mathcal{F}_i(y_i)\| =\leq e^{L(T-t_i)}\zeta(|P|)(t_i - t_{i-1}). \tag{1.13}$$

For every $s \in [t_{i-1}, t_i]$, we have $e^{L(T-t_i)} \leq e^{L(T-s)}$, so combining (1.7) with (1.13) gives

$$\begin{aligned}\|x^P(T) - x^Q(T)\| &\leq \sum_{i=1}^{\ell+1} e^{L(T-t_i)}\zeta(|P|)(t_i - t_{i-1}) \leq \int_{t_0}^{T} e^{L(T-s)}\zeta(|P|)\,ds \\ &= \big[-\zeta(|P|)L^{-1}e^{L(T-s)}\big]_{t_0}^{T} = \zeta(|P|)L^{-1}\big(e^{L(T-t_0)} - 1\big).\end{aligned}$$

This upper bound converges to 0 as $|P| \to 0$, and the result follows. □

Proof of Proposition 1.6. By the same reasoning as in the paragraph following (1.8), for every $t \in I$, every $\Delta \in \mathbb{R}$ such that $t + \Delta \in I$, and every partition P, we have

$$\|x^P(t + \Delta) - (x^P(t) + F(t, x^P(t))\Delta)\| \leq \zeta(|\Delta|)|\Delta|,$$

and sending $|P| \to 0$ and dividing both sides by $|\Delta|$ gives

$$\|\Delta^{-1}(x(t + \Delta) - x(t)) - F(t, x(t))\| \leq \zeta(|\Delta|).$$

Sending $\Delta \to 0$ completes the proof. □

Remark 1.7. Lipschitz continuity of F is only required with respect to x, although the constant is required to be uniform in t. If F is Lipschitz continuous in both t and x, then $\zeta(\delta) \leq L\delta$, and the computations in the proof give $\|x^P(T) - x^Q(T)\| \leq (e^{L(T-t_0)} - 1)|P|$.

Remark 1.8. If F is C^1 with respect to x, then all the partial derivatives $\frac{\partial F}{\partial x_i}$ are continuous and hence bounded on $I \times D$. If L is an upper bound for each $|\frac{\partial F}{\partial x_i}|$, then it is straightforward to show that L is a Lipschitz constant for F with respect to x.

1.4 Peano's existence theorem

Lec 5
M 2/2

Many of the arguments in this section can be found in [**Teschl**, §2.7].

When the vector field F is uniformly Lipschitz in x, Propositions 1.5 and 1.6 guarantee that the IVP (1.1) always has at least one solution. This can be strengthened in two ways:

- *Peano's existence theorem*, which we prove in this section, provides existence for all continuous F, without requiring a Lipschitz property;
- the *Picard–Lindelöf uniqueness theorem*, which we prove in §1.5, shows that when F is uniformly Lipschitz, there is a *unique* solution.

Theorem 1.9 (Peano's existence theorem)**.** *Fix times $t_0 \in \mathbb{R}$ and $T_* > t_0$, a point $x_0 \in \mathbb{R}^d$, and $r > 0$. Let $D := \overline{B(x,r)} \subset \mathbb{R}^d$. Suppose that $F\colon [t_0, T_*] \times D \to \mathbb{R}^d$ is continuous, and let*

$$M := \max\{\|F(t,x)\| : (t,x) \in [t_0,T_*] \times D\}, \qquad T := \min(T_*, t_0 + r/M).$$

Then the IVP (1.1) *has a solution on* $[t_0, T]$.

To prove Theorem 1.9, start with the observation that the proof of Proposition 1.6 remains valid even if we only have convergence along a subsequence $h_n \to 0$. To spell this out, and set up another estimate that we will need momentarily, observe that for every $s, t \in I$, and every $h > 0$, the points $(t, x^h(t))$ and $(s, x^h(s))$ are connected by line segments, each of which is in the direction of $(1, v)$ for some $v \in \mathbb{R}^d$ satisfying

$$\|v - F(t, x^h(t))\| \le \zeta(|s-t|) \quad \text{and} \quad \|v\| \le M,$$

where ζ is the function from (1.8). This allows us to conclude that

$$\left\|x^h(s) - \big(x^h(t) + F(t, x^h(t))(s-t)\big)\right\| \le \zeta(|s-t|)|s-t|, \tag{1.14}$$

$$\|x^h(s) - x^h(t)\| \le M|s-t|. \tag{1.15}$$

Proposition 1.10. *In the setting of Theorem 1.9, suppose that $h_n \to 0$ is such that the limit $x(t) := \lim_{n\to\infty} x^{h_n}(t)$ exists for every $t \in I$. Then x is a solution of* (1.1).

Proof. For every $t \in I$, every $\Delta \in \mathbb{R}$ such that $t + \Delta \in I$, and every $n \in \mathbb{N}$, (1.14) gives

$$\|x^{h_n}(t+\Delta) - (x^{h_n}(t) + F(t, x^{h_n}(t))\Delta)\| \le \zeta(|\Delta|)|\Delta|.$$

Sending $n \to \infty$ and dividing both sides by $|\Delta|$ gives

$$\|\Delta^{-1}(x(t+\Delta) - x(t)) - F(t, x(t))\| \le \zeta(|\Delta|).$$

Sending $\Delta \to 0$ completes the proof. □

From (1.15), the family of functions $\mathcal{X} := \{x^h : h > 0\} \subset C(I, D)$ satisfies the following:[7]

Definition 1.11. A family $\mathcal{X} \subset C(I, \mathbb{R}^d)$ is *equicontinuous* if for every $\epsilon > 0$, there exists $\delta > 0$ such that for every $x \in \mathcal{X}$ and every $s, t \in I$ with $|s - t| \le \delta$, we have $|x(s) - x(t)| \le \epsilon$.

Now we need the following result from real analysis.

Theorem 1.12 (Arzelà–Ascoli)**.** *Given a compact interval $I \subset \mathbb{R}$, a compact region $D \subset \mathbb{R}^d$, and an equicontinuous sequence of functions $(x_n\colon I \to D)_{n\in\mathbb{N}}$, there exists $n_k \to \infty$ such that the sequence of functions x_{n_k} is uniformly convergent.*

[7] Recall that $C(I, D)$ denotes the set of all continuous function $I \to D$.

Sketch of proof. Fix a countable dense set $Q = \{t_1, t_2, t_3, \dots\} \subset I$. Since D is compact, there exists an infinite set $J_1 \subset \mathbb{N}$ such that the sequence $(x_n(t_1))_{n\in J_1}$ converges. Iteratively, choose an infinite $J_{k+1} \subset J_k$ such that $(x_n(t_k))_{n\in J_k}$ converges. Pick $n_k \in J_k \cap [k, \infty)$, so that $x(t) := \lim_{k\to\infty} x_{n_k}(t)$ exists for every $t \in Q$. Use uniform continuity and an "$\epsilon/3$-argument" to show that $(x_{n_k})_k$ is uniformly Cauchy. □

We now have everything we need to prove existence of solutions.

Proof of Peano's Existence Theorem 1.9. By (1.15), the family of approximate solutions x^h provided by Euler's method is equicontinuous. By the Arzelà–Ascoli Theorem 1.12, there exists $h_n \to 0$ such that the functions $x^{h_n} : I \to D$ are uniformly convergent to a limit $x\colon I \to \mathbb{R}$. By Proposition 1.10, this limit is a solution of the IVP (1.1). □

1.5 Picard–Lindelöf uniqueness theorem

Now we turn our attention to the proof of the following.

Theorem 1.13 (Picard–Lindelöf uniqueness theorem)**.** *Fix times $t_0 \in \mathbb{R}$ and $T_* > t_0$, a point $x_0 \in \mathbb{R}^d$, and $r > 0$. Let $D := \overline{B(x, r)} \subset \mathbb{R}^d$. Suppose that $F\colon [t_0, T_*] \times D \to \mathbb{R}^d$ is continuous in (t, x) and uniformly Lipschitz continuous in x, and let*

$$\begin{aligned} M &:= \max\{\|F(t, x)\| : (t, x) \in [t_0, T_*] \times D\}, \\ T &:= \min(T_*,\ t_0 + r/M). \end{aligned} \tag{1.16}$$

Then the IVP (1.1) *has a unique solution on* $[t_0, T]$.

1.5.1 Picard iteration

When F is uniformly Lipschitz in x, we saw in §1.3 that the approximations in Euler's method converge to a solution of the IVP (1.1). However, this is not enough to imply uniqueness: after all, in the IVP given by $\dot{x} = \sqrt{x}$ with $x(0) = 0$, Euler's method converges to the solution $x(t) = 0$, and does not detect the solution $x(t) = \frac{1}{4}t^2$.

When F is Lipschitz, to rule out alternate solutions that are undetected by Euler's method, we need another approach. The method we will use is inspired by (1.2): if $x\colon I \to D$ is differentiable, then

$$\big(x \text{ solves (1.1)}\big) \quad \Leftrightarrow \quad x(t) = x_0 + \int_{t_0}^{t} F(s, x(s))\, ds \text{ for all } t \in I. \tag{1.17}$$

This suggests a method for improving a given "approximate solution": given $x\colon I \to D$, consider the function $\mathcal{P}x\colon I \to D$ given by

$$(\mathcal{P}x)(t) := x_0 + \int_{t_0}^{t} F(s, x(s))\, ds. \tag{1.18}$$

This defines a map $\mathcal{P}\colon C(I,D) \to C(I,D)$. Observe from (1.17) that

$$\big(x \text{ solves (1.1)}\big) \quad \Leftrightarrow \quad \mathcal{P}x = x. \tag{1.19}$$

Thus solutions of (1.1) correspond to fixed points of the map $\mathcal{P}$. Our next task is to study when this map has a *unique* fixed point.

1.5.2 Banach fixed point theorem

The key theoretical tool, which will also be important later on, is the following.

Theorem 1.14 (Banach Fixed Point Theorem)**.** *Let X be a complete metric space and $\mathcal{P}\colon X \to X$ a* contraction*, meaning that there is $\gamma \in (0,1)$ such that $d(\mathcal{P}x, \mathcal{P}y) \le \gamma d(x,y)$ for all $x, y \in X$. Then $\mathcal{P}$ has a unique fixed point $x^* \in X$. Moreover, the following are true.*

1. *The fixed point is exponentially stable: for every $x \in X$ we have $d(\mathcal{P}^n x, x^*) \le d(x,x^*)\gamma^n$.*
2. *For any $x \in X$, we have $d(x,x^*) \le \frac{1}{1-\gamma} d(x, \mathcal{P}x)$.*
3. *Fix $\delta > 0$ and let $\mathcal{Q}\colon X \to X$ be a contraction such that $d(\mathcal{P}x, \mathcal{Q}x) \le \delta$ for all $x \in X$. Then the respective fixed points $x^*_{\mathcal{P}}$ and $x^*_{\mathcal{Q}}$ satisfy $d(x^*_{\mathcal{P}}, x^*_{\mathcal{Q}}) \le \frac{\delta}{1-\gamma}$.*

Proof. Given $x \in X$, consider the sequence $x_n := \mathcal{P}^n x \in X$. By the contraction property, we have

$$d(x_n, x_{n+1}) = d(\mathcal{P}^n(x), \mathcal{P}^n(\mathcal{P}x)) \le \gamma^n d(x, \mathcal{P}x) \quad \text{for all } n \in \mathbb{N}.$$

For all $m \ge n$, this implies that

$$d(x_n, d_m) \le \sum_{k=n}^{m-1} d(x_k, x_{k+1}) \le \sum_{k=n}^{m-1} \gamma^k d(x, \mathcal{P}x) \le \frac{\gamma^n}{1-\gamma} d(x, \mathcal{P}x). \tag{1.20}$$

It follows that the sequence $(x_n)_n$ is Cauchy, and since X is complete, there exists $x^* \in X$ such that $x_n \to x^*$. To see that x^* is a fixed point, observe that

$$\mathcal{P}x^* = \mathcal{P}\Big(\lim_{n\to\infty} x_n\Big) = \lim_{n\to\infty} \mathcal{P}(\mathcal{P}^n x) = \lim_{n\to\infty} \mathcal{P}^{n+1} x = \lim_{n\to\infty} x_{n+1} = x^*.$$

To see uniqueness, observe that if $x = \mathcal{P}x$, then $x_n = x$ for all n, so $x = \lim_n x_n = x^*$. The exponential convergence estimate follows directly from the contraction property applied n times to the iterates of x and x^*. The second conclusion, regarding $d(x,x^*)$, follows from (1.20) by taking $m \to \infty$. To prove the final item of the theorem, we can apply the second item with $x = x^*_{\mathcal{Q}}$ and $x^* = x^*_{\mathcal{P}}$, obtaining

$$d(x^*_{\mathcal{Q}}, x^*_{\mathcal{P}}) \le \frac{1}{1-\gamma} d(x^*_{\mathcal{Q}}, \mathcal{P}x^*_{\mathcal{Q}}) \le \frac{1}{1-\gamma}\big(d(x^*_{\mathcal{Q}}, \mathcal{Q}x^*_{\mathcal{Q}}) + d(\mathcal{Q}x^*_{\mathcal{Q}}, \mathcal{P}x^*_{\mathcal{Q}})\big).$$

The first distance is 0 by the definition of $x^*_{\mathcal{Q}}$, and the second is at most δ by the assumption on $\mathcal{P}$ and $\mathcal{Q}$, which completes the proof. □

1.5.3 Picard iteration is a contraction

Lec 6
W 2/4

Recall the following fact from real analysis.

Proposition 1.15. *Given a compact interval $I \subset \mathbb{R}$ and a compact set $D \subset \mathbb{R}^d$, consider the set of continuous functions $X := C(I, D)$, and given $x, y \in X$, let*

$$d(x, y) := \sup_{t \in I} \|x(t) - y(t)\|.$$

Then (X, d) is a complete metric space.

Now the Picard iteration operator in (1.18) defines a map $\mathcal{P}\colon X \to X$.

Proposition 1.16. *In the setting of Theorem 1.13, fix $T' \in (t_0, T] \cap (t_0, t_0 + L^{-1})$. Then the map $\mathcal{P}$ is a contraction on $C([t_0, T'], D)$ with respect to d.*

Proof. Given continuous maps $x, y\colon [t_0, T'] \to D$, we must estimate

$$d(\mathcal{P}x, \mathcal{P}y) = \sup_{t \in I} \|\mathcal{P}x(t) - \mathcal{P}y(t)\|.$$

To this end, observe that for each $t \in [t_0, T']$, we have

$$\begin{aligned} \|\mathcal{P}x(t) - \mathcal{P}y(t)\| &= \left| \left(x_0 + \int_{t_0}^t F(s, x(s))\, ds \right) - \left(x_0 + \int_{t_0}^t F(s, y(s))\, ds \right) \right| \\ &\le \int_{t_0}^t |F(s, x(s)) - F(s, y(s))|\, ds \\ &\le \int_{t_0}^t L\|x(s) - y(s)\|\, ds \le L(T' - t_0) d(x, y). \end{aligned}$$

Let $\gamma := L(T' - t_0)$, and observe that $\gamma \in (0, 1)$ by the condition on T. □

Proof of Picard–Lindelöf Uniqueness Theorem 1.13. Let $T' = \min(T, t_0 + 1/(2L))$. Writing $X = C([t_0, T'], D)$, Proposition 1.15 guarantees that (X, d) is a complete metric space, and Proposition 1.16 shows that $\mathcal{P}\colon X \to X$ is a contraction, so we can apply the Banach Fixed Point Theorem 1.14 and conclude that $\mathcal{P}$ has a unique fixed point $x\colon [t_0, T'] \to D$. By (1.19), this is the unique solution of (1.1) on $[t_0, T']$.

If $T' = T$, then we are done; otherwise, repeat this process, using T' as the new start time. Writing $t_k := \min(t_0 + k/(2L), T)$, we see inductively that (1.1) has a unique solution on $[t_0, t_k]$ for every $k \in \mathbb{N}$, which completes the proof. □

Observe that in the above proof, we only showed that Picard iteration is a contraction when we restrict our attention to the (shorter) interval $[t_0, T']$, but that there is in fact a unique solution on the (longer) interval $[t_0, T]$. This raises the question of how far the solution can be extended, which we address in the next section.

1.6 Global solutions, or not

1.6.1 Returning to the original questions

In §1.2, we identified four questions regarding IVP solutions that we wanted to understand: existence, uniqueness, domain of definition, and availability of explicit formulas. The first two of these have now been answered, at least locally: a solution exists on an interval around t_0 whenever the vector field $F(t,x)$ is continuous; moreover, if F is uniformly Lipschitz in x, then there is a *unique* solution.

Regarding the fourth question, we cannot in general expect to have a nice concrete formula for the solution, as the example $\dot{x} = x^3 + 1$ illustrates. However, we have now seen two methods for producing *approximate* solutions – Euler's method and Picard iteration – both of which are guaranteed to converge to the true solution when the vector field is uniformly Lipschitz in x. Of the two, it is generally more efficient to use Euler's method and its higher-order extensions, such as Runge–Kutta. However, one aspect of the Picard iteration scheme in (1.18) is worth pointing out before we move on: if the vector field F is given by a polynomial in t and x, and if we start the sequence of Picard iterates with the constant function $x(t) = x_0$, then we will get a sequence of polynomial approximations to the true solution, so that Picard iteration provides explicit formulas for the approximate solutions in a nicer way than Euler's method does.

This last observation is connected to the idea of solving differential equations by using power series (or other infinite series). We will not pursue this approach further here, however.

So far, we have not addressed the third question from our original list: under what conditions the solution of the IVP (1.1) is globally defined (so that $I = \mathbb{R}$), or more generally, how we can determine the largest interval I on which the solution exists. We discuss this next, and then in §1.7 we turn our attention to one further important question which was hinted at by the error estimates for Euler's method in §1.3: how does the solution depend on the initial condition?

1.6.2 Maximal interval of existence

Let $U \subset \mathbb{R} \times \mathbb{R}^d$ be an open set, and let $F\colon U \to \mathbb{R}^d$ be a vector field that is *locally Lipschitz* with respect to x, meaning that for every compact interval $I \subset \mathbb{R}$ and every compact domain $D \subset \mathbb{R}^d$ such that $I \times D \subset U$, the restriction $F|_{I\times D}$ is uniformly Lipschitz w.r.t. x. Then the Picard–Lindelöf Uniqueness Theorem 1.13 guarantees that for every $(t_0, x_0) \in U$, there exists $\epsilon > 0$ such that the IVP (1.1) has a unique solution $x\colon (t_0 - \epsilon, t_0 + \epsilon) \to \mathbb{R}^d$.

Lec 7
F 2/6

Definition 1.17. An interval $I \subset \mathbb{R}$ containing t_0 is an *interval of existence* for the vector field $F\colon U \to \mathbb{R}^d$ if the IVP (1.1) has a solution $x\colon I \to \mathbb{R}^d$. The interval I is a *maximal interval of existence* if there does not exist any interval of existence $J \supsetneq I$.

Theorem 1.18. *Let $U \subset \mathbb{R} \times \mathbb{R}^d$ be open, and let $F\colon U \to \mathbb{R}^d$ be continuous in (t,x) and locally Lipschitz in x. Then for every $(t_0, x_0) \subset U$, the following is true.*

1. *The IVP (1.1) has a unique maximal interval of existence I.*
2. *We have $I = (T^-, T^+)$ for some $T^- \in [-\infty, t_0)$ and $T^+ \in (t_0, \infty]$.*
3. *The IVP has a unique solution on I.*

Proof. Consider the set

$$I^+ := \{T \geq t_0 : (1.1) \text{ has a solution on } [t_0, T]\},$$

and let $T^+ := \sup I^+ \in (t_0, \infty]$. Clearly, I^+ is an interval, since if $x\colon [t_0, T] \to \mathbb{R}^d$ is a solution of (1.1), then so is $x|_{[t_0,S]}$ for every $S \in [t_0, T]$. We claim that $I^+ = [t_0, T^+)$. If $T^+ = \infty$ then this is immediate. If $T^+ < \infty$ and (1.1) has a solution x on $[t_0, T^+]$, then $(T^+, x(T^+)) \in U$, and thus the IVP given by $\dot{y} = F(t,y)$ with initial condition $y(T^+) = x(T^+)$ has a unique solution on some interval $(T^+ - \epsilon, T^+ + \epsilon)$. Defining $x(t) = y(t)$ for $t \in (T^+, T^+ + \epsilon)$ extends x to a solution of (1.1) on $[t_0, T^+ + \epsilon)$, contradicting the definition of T^+.

The previous paragraph shows that $I^+ = [t_0, T^+)$, and a similar argument shows that

$$I^- := \{T \leq t_0 : (1.1) \text{ has a solution on } [T, t_0]\}$$

has the form $I^- = (T^-, t_0]$, where $T^- := \inf I^- \in [-\infty, t_0)$. We conclude that $I = I^- \cup I^+ = (T^-, T^+)$ is the unique maximal interval of existence for (1.1).

To prove that (1.1) has a unique solution on I, suppose that $x, y\colon I \to \mathbb{R}^d$ are solutions, and let

$$\tau := \sup\{t \geq t_0 : x|_{[t_0,t]} = y|_{[t_0,t]}\}.$$

If $\tau < T^+$, then since the IVP given by $\dot{z} = F(t,z)$ and $z(\tau) = x(\tau)$ has a unique solution on some interval $(\tau - \epsilon, \tau + \epsilon)$, we see that x and y must agree on $[t_0, \tau + \epsilon)$, contradicting the definition of τ. Thus $\tau = T^+$, and the solution is unique on I^+. A similar argument gives uniqueness on I^-. □

Definition 1.19. Given a compact set $K \subset \mathbb{R}^d$ such that $[t_0, T^+) \times K \subset U$, say that a solution $x\colon [t_0, T^+) \to \mathbb{R}^d$ *abandons* the set K near T^+ if

$$T_K^+ := \sup\{t \in [t_0, T^+) : x(t) \in K\} < T^+.$$

Similarly, say that $x\colon (T^-, t_0] \to \mathbb{R}^d$ abandons K near T^- if

$$T_K^- := \inf\{t \in (T^-, t_0] : x(t) \in K\} > T^-.$$

Theorem 1.20. *Let U, F, t_0, x_0, T^-, T^+ be as above. If $T^+ < \infty$, then the unique solution x abandons every compact set near T^+: given any $T < T^+$ and any compact $K \subset \mathbb{R}^d$ such that $[T, T^+] \times K \subset U$, the solution x abandons K near T^+. Similarly, if $T^- > -\infty$, then x abandons every compact set near T^-.*

Before proving Theorem 1.20, we offer the following informal reformulation: if $T^+ < \infty$, then as $t \to T^+$ from below, the solution either goes to ∞ or approaches the boundary of the region where the vector field is defined. An analogous interpretation holds if $T^- > -\infty$.

Theorem 1.20 will follow quickly from the next lemma.

Lemma 1.21. *Let $U \subset \mathbb{R} \times \mathbb{R}^d$ be open, and $F\colon U \to \mathbb{R}^d$ be continuous in (t, x) and locally Lipschitz in x. Let $I \subset \mathbb{R}$ be a compact interval and $K \subset \mathbb{R}^d$ a compact domain such that $I \times K \subset U$. Then there exists $\delta > 0$ such that given any $(t_0, x_0) \in I \times K$, the IVP (1.1) has a solution on $[t_0 - \delta, t_0 + \delta] \cap I$.*

Proof. Since $I \times K$ is compact, U is open, and $I \times K \subset U$, there exists $r > 0$ such that the set

$$K_1 := \bigcup_{x \in K} \overline{B(x, r)}$$

satisfies $I \times K_1 \subset U$. You can prove this by considering the positive continuous function $g\colon I \times K \to (0, \infty)$ defined by

$$g(t, x) := \sup\{r > 0 : \{t\} \times B(x, r) \subset U\};$$

since $I \times K$ is compact, this function achieves its infimum, which is therefore positive, and any r between 0 and this infimum will do the job.

Now let $M := \sup\{\|F(t, x)\| : (t, x) \in I \times K_1\}$, and let $\delta = r/M$. Then given any $(t_0, x_0) \in I \times K$, we have $\overline{B(x_0, r)} \subset K$, so by the Picard–Lindelöf Uniqueness Theorem 1.13, the IVP (1.1) has a unique solution on $[t_0, t_0 + \delta] \cap I$.

The argument for $[t_0 - \delta, t_0]$ is similar. □

Proof of Theorem 1.20. Given any $T < T^+$ and any compact $K \subset \mathbb{R}^d$ such that $[T, T^+] \times K \subset U$, we can apply Lemma 1.21 and deduce that for every $t \in [T, T^+]$ such that $x(t) \in K$, the solution is defined on $[t - \delta, t + \delta] \cap [T, T^+]$. Since T^+ is not in the maximal interval of existence, it follows that $t + \delta < T^+$. Taking a supremum over all such t yields $T_K^+ \le T^+ - \delta < T^+$. The proof near T^- is similar. □

1.6.3 Global existence and the flow of a vector field

Lec 8
M 2/9

Now we turn our attention to the situation when the vector field is globally defined, and consider a continuous function $F\colon \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ that is locally Lipschitz in x. We saw already, via the example $\dot{x} = x^2$, that solutions of the corresponding DE can go

to infinity in finite time, in which case the maximal interval of existence was smaller than $\mathbb{R}$. In this example, the vector field has magnitude that grows quadratically. It turns out that with slower (linear) growth, existence for all time is guaranteed.

Theorem 1.22. *Let $F\colon \mathbb{R} \to \mathbb{R}^d$ be continuous in (t,x) and uniformly Lipschitz in x, and suppose that there exist $a, b > 0$ such that $\|F(t,x)\| \le a + b|x|$ for all $t \in \mathbb{R}$. Then all solutions of $\dot{x} = F(t,x)$ are defined for all $t \in \mathbb{R}$.*

To prove the theorem, we will use a lemma that is useful in various places.

Lemma 1.23 (Gronwall's inequality)**.** *Suppose that $g\colon [t_0, T) \to [0,\infty)$ is a continuous function for which there exist $C, k > 0$ such that*

$$g(t) \le C + k\int_{t_0}^{t} g(s)\,ds \quad \text{for all } t \in [t_0, T). \tag{1.21}$$

Then we have

$$g(t) \le Ce^{k(t-t_0)} \quad \text{for all } t \in [t_0, T). \tag{1.22}$$

Proof. Let $h(t) := C + k\int_{t_0}^{t} g(s)\,ds$, so that (1.21) gives $h(t) \ge g(t)$ and $h(t) > 0$ for all $t \in [t_0, T)$. Observe that $\dot{h}(t) = kg(t)$, so

$$\frac{d}{dt}\log h = \frac{\dot{h}}{h} = \frac{kg}{h} \le k.$$

Integrating both sides of the inequality gives

$$\log h(t) \le kt + \log h(t_0) = kt + \log C,$$

which proves (1.22). □

Proof of Theorem 1.22. Let (T^-, T^+) be the maximal interval of existence associated to the initial conditions (t_0, x_0). Given any $t \in (t_0, T^+)$, we can use the estimate on $\|F\|$ and the integral from (1.2) of the IVP to obtain

$$\begin{aligned}\|x(t)\| &\le \|x_0\| + \int_{t_0}^{t} \|F(s,x)\|\,ds \\ &\le \|x_0\| + \int_{t_0}^{t} (a + b\|x(s)\|)\,ds = \|x_0\| + a(T^+ - t_0) + b\int_{t_0}^{t} \|x(s)\|\,ds.\end{aligned}$$

Thus we can apply Gronwall's inequality (Lemma 1.23) to the function $g(t) = \|x(t)\|$ with $C = \|x_0\| + a(T^+ - t_0)$ and $k = b$, concluding that

$$\|x(t)\| \le (\|x_0\| + a(T^+ - t_0))e^{b(T^+ - t_0)} \quad \text{for all } t \in [t_0, T^+). \tag{1.23}$$

Suppose for a contradiction that $T^+ < \infty$. Then the right-hand side of (1.23) takes a finite value. Denote this value by R, and let $K := \overline{B(x_0, R)}$. The set K is compact,

but (1.23) shows that the orbit $x(t)$ never leaves K on $[t_0, T^+)$, contradicting Theorem 1.20, which requires it to abandon this set near T^+. We conclude that $T^+ = \infty$, and a similar argument shows that $T^- = -\infty$. □

Now consider an autonomous vector field $F\colon \mathbb{R}^d \to \mathbb{R}^d$ that is locally Lipschitz and has the property that all solutions of $\dot{x} = F(x)$ are defined for all time. Given $t \in \mathbb{R}$, define $f_t\colon \mathbb{R}^d \to \mathbb{R}^d$ as follows: given $x \in \mathbb{R}^d$, let $c_x\colon \mathbb{R} \to \mathbb{R}^d$ be the unique solution curve passing through the point x at time 0, so that

$$c_x(0) = x \quad \text{and} \quad \frac{d}{dt}c_x(t) = F(c_x(t)) \text{ for all } t \in \mathbb{R}.$$

Then let

$$f_t(x) := c_x(t). \tag{1.24}$$

The map f_t is called the *time-t* map for the vector field F and its associated DE. Because F is autonomous, we see that given any $s \in \mathbb{R}$, if we let $c_x^s\colon \mathbb{R} \to \mathbb{R}^d$ be the unique solution curve with $c_x^s(s) = x$, then $c_x^s(t+s) = f_t(x)$.

This has the following important consequence. Given $s, t \in \mathbb{R}$ and $x \in \mathbb{R}^d$, we have

$$f_{t+s}(x) = c_x(t+s) = c^s_{c_s(x)}(t+s) = f_t(c_s(x)) = f_t(f_s(x)).$$

Thus the family of time-t maps is an example of the following:

Definition 1.24. A *flow* on $\mathbb{R}^d$ is a family of maps $\{f_t\colon \mathbb{R}^d \to \mathbb{R}^d\}_{t\in\mathbb{R}}$ such that for all $s, t \in \mathbb{R}$, we have

$$f_{t+s} = f_t \circ f_s. \tag{1.25}$$

One can also work with flows that are only partially defined, in the sense that the maximal intervals of existence may not be $\mathbb{R}$, so f_t is not necessarily defined on all of $\mathbb{R}^d$, and we only require (1.25) to hold where all the maps involved are defined. Later, we will see other ways (besides the global bounds in Theorem 1.22) of verifying that a solution exists for all future time.

1.7 Dependence on initial conditions

Lec 9
W 2/11

Consider a continuous vector field $F(t, x)$ that is locally Lipschitz in x. It is often important to understand how the unique solution of the IVP (1.1) varies if we change the initial condition x_0. Using Gronwall's inequality (Lemma 1.23), we can show that $x(t)$ depends continuously on x_0, and give a quantitative estimate for how sensitive this dependence can be:

Proposition 1.25. *Let $I \subset \mathbb{R}$ be a compact interval, $D \subset \mathbb{R}^d$ a compact domain, and $F\colon I \times D \to \mathbb{R}^d$ a continuous vector field that is uniformly Lipschitz in x with constant*

L. Fix $t_0 \in I$ and $y_0, z_0 \in D$. Suppose that the corresponding solutions y, z of the IVP (1.1) are defined in D on the time interval $[t_0, T] \subset I$. Then for all $t \in [t_0, T]$, we have

$$\|y(t) - z(t)\| \le e^{L(t-t_0)} \|y_0 - z_0\|. \tag{1.26}$$

Proof. Let $g(t) := \|y(t) - z(t)\|$ and observe that using the integral formulation of the IVP, we have

$$\begin{aligned} g(t) &= \left\| \left(y_0 + \int_{t_0}^t F(s, y(s))\, ds \right) - \left(z_0 + \int_{t_0}^t F(s, z(s))\, ds \right) \right\| \\ &\le \|y_0 - z_0\| + \int_{t_0}^t \|F(s, y(s)) - F(s, z(s))\|\, ds \\ &\le \|y_0 - z_0\| + L \int_{t_0}^t g(s)\, ds, \end{aligned}$$

so we can apply Lemma 1.23 with $k = L$ and $C = \|y_0 - z_0\|$, at which point (1.22) implies (1.26). □

Observe that (1.26) has the same form as the estimate in Lemma 1.4 for the approximate solutions from Euler's method. Indeed, since Euler's method converges to the true solution, we could also prove Proposition 1.25 using Lemma 1.4, although some care would need to be taken regarding the time interval on which we work.

Proposition 1.25 implies that at each time t, the solution to the IVP depends continuously on the initial condition. In particular, for an autonomous vector field, the time-t map f_t defined in (1.24) is continuous, and in fact Lipschitz with constant $e^{L|t|}$. (The absolute value signs are necessary if $t < 0$.) From this we are able to deduce that the family of time-t maps $\{f_t\}_t$ is a *continuous* flow, meaning that $(t, x) \mapsto f_t(x)$ is continuous:[8]

Corollary 1.26. *Let $D \subset \mathbb{R}^d$ be a compact domain and $F\colon D \to \mathbb{R}^d$ a vector field that is Lipschitz with constant L. Let $M := \sup\{F(x) : x \in D\}$, and let $\{f_t\}_t$ be the flow associated to the DE $\dot{x} = F(x)$. Then for every $y, z \in D$ and every $T > 0$ such that $f_t(y)$ and $f_t(z)$ are defined and lie in D for all $t \in [0, T]$, we have the following bound for all $s, t \in [0, T]$:*

$$\|f_s(y) - f_t(z)\| \le M|s - t| + e^{LT}\|y - z\|.$$

In particular, the flow $(t, x) \mapsto f_t(x)$ is continuous wherever it is defined.

[8] Recall that continuous dependence on each of t and x separately is not enough to guarantee continuous dependence on the pair (t, x), as the function $\frac{tx}{t^2+x^2}$ shows. It is the Lipschitz nature of this coordinatewise dependence that is vital here. Also note that here we are not assuming the flow to be globally defined; the bound in Corollary 1.26 is valid wherever the flow is defined.

When the vector field F is continuously differentiable with respect to x (which implies uniformly Lipschitz on compact domains, as in Remark 1.8), it is reasonable to ask whether we might upgrade "continuous dependence" to "differentiable dependence" in the above discussion.

At the level of naive symbol-pushing, we expect the chain rule to give

$$\frac{\partial}{\partial t} f_t(x) = F(f_t(x)) \quad \Rightarrow \quad \frac{\partial^2}{\partial x\, \partial t} f_t(x) = DF(f_t(x)) \frac{\partial}{\partial x} f_t(x),$$

so writing $\Phi(t,x) := \frac{\partial}{\partial x} f_t(x)$, and $\dot{\Phi}(t,x) := \frac{\partial}{\partial t}\Phi(t,x)$, we expect to see

$$\dot{\Phi}(t,x) = \frac{\partial^2}{\partial t\, \partial x} f_t(x) = DF(f_t(x)) \frac{\partial}{\partial x} f_t(x) = DF(f_t(x))\Phi(t,x). \tag{1.27}$$

This is too hasty to be a proof of anything, though: implicit in the above computations is an assumption that $f_t(x)$ is actually differentiable in x, which we have not justified.[9] We should also take our time and be sure we understand exactly what is meant by $\frac{\partial}{\partial x}$ and DF, since $x \in \mathbb{R}^d$ and F is a vector field on a region of $\mathbb{R}^d$.

To this end, recall that given an open set $U \subset \mathbb{R}^d$, a function $F\colon U \to \mathbb{R}^d$, and a point $x_0 \in U$, the derivative of F with respect to x at the point x_0 (if it exists) is the unique linear map $DF(x_0)\colon \mathbb{R}^d \to \mathbb{R}^d$ such that for $x \in U$, we have

$$F(x) = F(x_0) + DF(x_0)(x - x_0) + R(x),$$

where the remainder term $R(x)$ has the property that

$$\lim_{x \to x_0} \frac{\|R(x)\|}{\|x - x_0\|} = 0.$$

This linear map is represented in the standard basis by the Jacobian matrix of partial derivatives of the coordinate functions of F. Recall that if $U, V \subset \mathbb{R}^d$ are open sets and $F\colon U \to V$ and $G\colon V \to \mathbb{R}^d$ are differentiable, then the multivariable chain rule takes the form

$$D(G \circ F)(x_0) = DG(Fx_0)DF(x_0),$$

where the right-hand side can be interpreted either as a composition of linear maps from $\mathbb{R}^d$ to itself, or as a product of $d \times d$ matrices.

We will write $\mathbb{R}^{d\times d}$ for the space of all $d \times d$ real matrices, so that $DF(x(t)) \in \mathbb{R}^{d\times d}$ for each t. We will also represent $\Phi(t,x) = \frac{\partial}{\partial x} f_t(x)$ by an element of $\mathbb{R}^{d\times d}$, and interpret it just as in the above discussion: $\Phi(t,x)$ is the unique linear function $\mathbb{R}^d \to \mathbb{R}^d$ (if one exists) such that

$$\lim_{y \to x} \frac{\|f_t(y) - (f_t(x) + \Phi(t,x)(y-x))\|}{\|y - x\|} = 0. \tag{1.28}$$

[9] We also used equality of mixed partial derivatives, which is valid if these derivatives are continuous, but we have not shown that either.

We can describe $\Phi(t,x)$ in terms of its action on vectors in $\mathbb{R}^d$. Given $v_0 \in \mathbb{R}^d$, we consider the initial conditions $y = x + rv_0$ for $r \in \mathbb{R}$, and observe that sending $r \to 0$, (1.28) gives

$$\Phi(t,x)v_0 = \lim_{r\to 0} \frac{1}{r}\|f_t(x+rv_0) - f_t(x)\|. \tag{1.29}$$

With all of this as motivation (since we have not actually proved anything yet!), we now establish the following theorem.

Theorem 1.27. *Let $U \subset \mathbb{R}^d$ be an open set, and let $F\colon U \to \mathbb{R}^d$ be C^1. Given $x \in U$, let c_x be the solution of the associated IVP: that is, suppose that $I \subset \mathbb{R}$ is an interval and $c_x\colon I \to U$ is a C^1 function satisfying $c_x(0) = x$ and $\dot{c}_x(t) = F(c_x(t))$ for all $t \in I$. Consider the* first variational equation

$$\dot{v} = A(t,x)v, \quad \text{where } A(t,x) := DF(c_x(t)). \tag{1.30}$$

Then for every $v_0 \in \mathbb{R}^d$, the IVP given by (1.30) with the initial condition $v(0) = v_0$ has a unique solution on I, and this solution $v\colon I \to \mathbb{R}^d$ has the property that

$$v(t) = \lim_{r\to 0} \frac{1}{r}\|f_t(x+rv_0) - f_t(x)\|. \tag{1.31}$$

Proof. Without loss of generality, assume that I is compact. (Every interval can be written as an increasing union of compact intervals, and if the conclusion of the theorem holds on each of these compact intervals, then it holds on their union as well.) We also assume that $t > 0$; the case for $t < 0$ is similar.

To see that (1.30) has a unique solution on I, it suffices to observe that the map $I \times \mathbb{R}^d \to \mathbb{R}^d$ defined by $(t,v) \mapsto A(t,x)v$ is continuous in (t,v) and Lipschitz in v, and satisfies the linear growth bound in Theorem 1.22. So we must argue that this unique solution $v(t)$ satisfies (1.31).

Since U is open and $\{f_t(x) : t \in I\}$ is compact, there exists $\delta > 0$ such that for all $t \in I$, we have $\overline{B(f_t(x),\delta)} \subset U$. Let $K := \bigcup_{t\in I} \overline{B(f_t(x),\delta)}$, so $K \subset U$ is compact. Since F is C^1 on K, the function on $K \times K$ defined by

$$R(z,y) := F(y) - (F(z) + DF(z)(y-z)) \tag{1.32}$$

has the property that

$$\omega(\delta) := \sup\{R(z,y) : y,z \in K, \|y-z\| \le \delta\}$$

satisfies $\frac{1}{\delta}\omega(\delta) \to 0$ as $\delta \to 0$.

By Proposition 1.25, there exists $\rho > 0$ such that for all $r \in (-\rho,\rho)$, the interval I is an interval of existence for solutions of the original IVP with initial condition $y_r := x + rv_0$. Given $r \in (-\rho,\rho)$ and $t \in I$, let

$$\Delta_r(t) := f_t(y_r) - f_t(x). \tag{1.33}$$

Now (1.31) is equivalent to the statement that $\frac{1}{r}\Delta_r(t) \to v(t)$ as $r \to 0$. To prove this, we first recall the integral IVP equation in (1.2), and write

$$f_t(x) = x + \int_0^t F(f_s(x))\,ds,$$
$$f_t(y_r) = y_r + \int_0^t F(f_s(y_r))\,ds.$$

Subtracting the first of these from the second and recalling that $y_r - x = rv$, we get

$$\Delta_r(t) = rv + \int_0^t \big(F(f_s(y_r)) - F(f_s(x))\big)\,ds.$$

From (1.32), we have

$$\begin{aligned} F(f_s(y_r)) - F(f_s(x)) &= DF(f_s(x))(f_s(y_r) - f_s(x)) + R(f_s(x), f_s(y_r)) \\ &= A(s,x)\Delta_r(s) + R(f_s(x), f_s(y_r)), \end{aligned}$$

and thus

$$\Delta_r(t) = rv + \int_0^t \big(A(s,x)\Delta_r(s) + R(f_s(x), f_s(y_r))\big)\,ds. \tag{1.34}$$

The integral form of (1.30) gives

$$v(t) = v + \int_0^t A(s,x)v(s)\,ds. \tag{1.35}$$

Let $g_r(t) := \|\frac{1}{r}\Delta_r(t) - v(t)\|$. Combining (1.34) and (1.35), we get

$$g_r(t) \le \int_0^t \big(\|A(s,x)\|g_r(s) + r^{-1}\|R(f_s(x), f_s(y_r))\|\big)\,ds.$$

We are almost in a position to use Gronwall's inequality. First let $T = \sup I$, and apply Proposition 1.25 to get

$$\|f_s(x) - f_s(y_r)\| \le rve^{LT} \quad \text{for all } s \in [0,T],$$

so that

$$g_r(t) \le \frac{T}{r}\omega(rve^{LT}) + \int_0^t \|A(s,x)\|g_r(s)\,ds.$$

Now Gronwall's inequality gives

$$g_r(t) \le \frac{T}{r}\omega(rve^{LT})e^{Lt}.$$

Sending $r \to 0$ and using the fact that $\lim_{\delta\to 0}\frac{\omega(\delta)}{\delta} = 0$ gives the result. □

In the setting of Theorem 1.27, we can define $\Phi(t,x) \in \mathbb{R}^{d\times d}$ for each $t \in I = I(x)$ by the condition that $\Phi(t,x)v_0 = v(t)$ for each $v_0 \in \mathbb{R}^d$ and its corresponding solution $v(t)$ of (1.30). Note that it is enough to consider the case when v_0 is a standard basis vector, in which case $v(t)$ gives the corresponding column of $\Phi(t,x)$. This matrix valued function $\Phi(t,x)$ is the *fundamental matrix solution* of the first variational equation. Observing that the corresponding DE is linear, we will next turn our attention to the question of solving linear differential equations and describing the qualitative properties of their solutions. We conclude the present section with the observation that we have now shown the following: given a C^1 vector field F, the associated flow $(t,x) \mapsto f_t(x)$ is C^1 in both t and x, and $\frac{\partial}{\partial x} f_t(x)$ is given by the fundamental matrix solution $\Phi(t,x)$ of the first variational equation.

Chapter 2: Linear theory

2.1 Matrix exponentials: definition

Lec 10
F 2/13

Now we turn our attention to the study of linear DEs, of the form $\dot{x} = A(t)x$, where $A(t) \in \mathbb{R}^{d\times d}$. We begin by restricting even further, to the case of autonomous linear DEs, where A is independent of t. Thus we are interested in solutions of the DE

$$\dot{x} = Ax, \tag{2.1}$$

where $A \in \mathbb{R}^{d\times d}$ and $x(t) \in \mathbb{R}^d$.

In the case $d = 1$, we know that for every $a \in \mathbb{R}$, the solution of the IVP given by $\dot{x} = ax$ and $x(0) = x_0 \in \mathbb{R}$ is $x(t) = e^{at}x_0$. Thus it is natural to expect that (2.1) with initial condition $x(0) = x_0 \in \mathbb{R}^d$ has solution given by $x(t) = e^{At}x_0$, provided the matrix exponential e^{At} is interpreted appropriately. A natural definition would be to adapt the power series for the exponential function from the case of real numbers to the case of matrices: recalling that

$$e^a = \sum_{n=0}^{\infty} \frac{1}{n!} a^n \quad \text{for all } a \in \mathbb{R},$$

we make the tentative definition

$$e^A := \sum_{n=0}^{\infty} \frac{1}{n!} A^n \quad \text{for all } A \in \mathbb{R}^{d\times d}. \tag{2.2}$$

Now our goal is to prove that this infinite series converges in an appropriate sense, and that the resulting matrix exponential map $A \mapsto e^A$ retains enough of the properties of the usual exponential function that we can use it to solve (2.1).

2.2 Matrix norms

We will need some preliminaries from linear analysis.

Definition 2.1. Let V be a vector space over a field $\mathbb{F}$; in the following, we will always have either $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. A *norm* on V is a function $V \to \mathbb{R}$, denoted $x \mapsto \|x\|$, such that

1. $\|x\| \geq 0$ for all $x \in V$, with $\|x\| = 0$ if and only if $x = 0$;
2. $\|x + y\| \leq \|x\| + \|y\|$;
3. $\|\alpha x\| = |\alpha| \cdot \|x\|$ for all $\alpha \in \mathbb{F}$.

Examples of normed vector spaces include:

- $V = \mathbb{R}^d$ with the Euclidean norm $\|x\|_2 := \sqrt{x_1^2 + \cdots + x_d^2}$;
- $V = \mathbb{R}^d$ with the L^1-norm $\|x\|_1 := \sum_{i=1}^d |x_i|$;
- $V = \mathbb{R}^d$ with the L^∞-norm $\|x\|_\infty := \max_i |x_i|$;
- $V = \ell^1 := \{x \in \mathbb{R}^{\mathbb{N}} : \sum_{i=1}^\infty |x_i| < \infty\}$ with the norm $\|x\|_1 := \sum_{i=1}^\infty |x_i|$. (Here $\mathbb{R}^{\mathbb{N}}$ denotes the set of all infinite sequences of real numbers.[10])
- $V = C([0,1], \mathbb{R})$ with the supremum norm $\|f\| := \sup_{x \in [0,1]} |f(x)|$.

A normed vector space is a metric space (and hence a topological space) in a natural way: given $x, y \in V$, we define the distance between x and y to be $\rho(x, y) := \|x - y\|$.

Definition 2.2. Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on a vector space V are *equivalent* if there exist constants $c, C > 0$ such that

$$c\|x\|_1 \leq \|x\|_2 \leq C\|x\|_1 \quad \text{for all } x \in V.$$

Observe that equivalent norms lead to the same notions of convergence of sequences, continuity of functions, and the Lipschitz property for functions.

Proposition 2.3. *If V is finite-dimensional, then any two norms are equivalent.*

Proof. Fixing a basis for V, we assume without loss of generality that $V = \mathbb{R}^d$ for some $d \in \mathbb{N}$. It suffices to show that every norm $\|\cdot\|$ on V is equivalent to the Euclidean norm $\|\cdot\|_2$, since equivalence of norms is transitive. To this end, first observe that if we write e_i for the ith standard basis vector and put $C := \max_{1 \leq i \leq d} \|e_i\|$, then for every $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we have

$$\|x\| \leq \sum_{i=1}^d |x_i| \cdot \|e_i\| \leq \sum_{i=1}^d |x_i| \cdot C,$$

[10] Recall that in general, given two sets X and Y, the notation X^Y denotes the set of all functions $f \colon Y \to X$. When $Y = \mathbb{N}$, such a function is naturally interpreted as a sequence of elements of X.

and by the Cauchy–Schwarz inequality applied to the vectors $\mathbf{1} := (1, \dots, 1) \in \mathbb{R}^d$ and $|x| := (x_1, \dots, x_d)$, we have

$$\sum_{i=1}^{d} |x_i| = \sum_{i=1}^{d} 1 \cdot |x_i| = \langle \mathbf{1}, |x| \rangle \le \|\mathbf{1}\|_2 \|x\|_2 = \sqrt{d}\|x\|_2,$$

so that $\|x\| \le C\sqrt{d}\|x\|_2$. This gives one of the desired inequalities, and also shows that the function $\mathbb{R}^d \to \mathbb{R}$ defined by $x \mapsto \|x\|$ is continuous with respect to the Euclidean metric, since given $x, y \in \mathbb{R}^d$, we have

$$\big|\|x\| - \|y\|\big| \le \|x - y\| \le C\sqrt{d}\|x\|_2.$$

Since the unit sphere

$$S^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$$

is compact in the Euclidean metric, the norm $\|\cdot\|$ achieves its minimum on this set: there exists $y \in S^{d-1}$ such that $\|x\| \ge \|y\|$ for all $x \in S^{d-1}$. Writing $c := \|y\|$, we have $c > 0$ since $y \ne 0$, and for all $x \in \mathbb{R}^d$, we have $\hat{x} := x/\|x\|_2 \in S^{d-1}$, allowing us to conclude that

$$\|x\| = \|x\|_2 \|\hat{x}\| \ge \|x\|_2 \|y\| = c\|x\|_2,$$

which proves the proposition. □

Remark 2.4. Proposition 2.3 fails utterly in infinite dimensions: we can define the supremum norm on ℓ^1 by $\|x\|_\infty := \sup_i |x_i|$, but while we clearly have $\|x\|_\infty \le \|x\|_1$ for all $x \in \ell^1$, there is no $c > 0$ such that $\|x\|_\infty \ge c\|x\|_1$ for all $x \in \ell^1$. Indeed, given any $c > 0$, choose $n \in \mathbb{N}$ such that $\frac{1}{n} < c$, and define $x \in \mathbb{R}^{\mathbb{N}}$ by putting $x_i = \frac{1}{n}$ for all $1 \le i \le n$, and $x_i = 0$ for all $i > n$. Then $\|x\|_1 = 1$ and $\|x\|_\infty = \frac{1}{n} < c = c\|x\|_1$.

Lec 11
M 2/16

Given a finite-dimensional[11] normed vector space $(V, \|\cdot\|)$, let $L(V)$ denote the set of linear maps $A \colon V \to V$. Observe that $L(V)$ is a vector space in its own right. Given $A \in L(V)$, let

$$\|A\| := \sup\{\|Ax\| : \|x\| \le 1\}.$$

The following properties can be quickly verified, and we leave their proofs as exercises.

1. $\|A\|$ is finite for all $A \in L(V)$, following a similar argument to the first half of Proposition 2.3, and $\|\cdot\|$ is a norm on $L(V)$.
2. The vector space $L(V)$ is complete in this norm.
3. We have $\|Ax\| \le \|A\| \cdot \|x\|$ for all $x \in V$ and $A \in L(V)$.
4. We have $\|AB\| \le \|A\| \cdot \|B\|$ for all $A, B \in L(V)$.

[11] Much of the following discussion goes through in infinite dimensions as well, but in that case it is possible to have linear maps $A \in L(V)$ for which $\|A\| = \infty$, and then in order to obtain a complete normed vector space of linear maps, we must restrict our attention to $B(V) := \{A \in L(V) : \|A\| < \infty\}$, which is equivalent to considering only *continuous* linear operators.

5. We have $\|A^n\| \le \|A\|^n$ for all $n \in \mathbb{N}$.
6. The operator norm does not come from an inner product on $L(V)$. This can be proved by first arguing that a norm comes from an inner product if and only if it satisfies the *parallelogram law* $\|a+b\|^2 + \|a-b\|^2 = 2(\|a\|^2 + \|b\|^2)$ for all a, b in the vector space, and then showing that the operator norm does not satisfy this identity.
7. Identifying V with $\mathbb{R}^d$ gives an identification of $L(V)$ with $\mathbb{R}^{d\times d}$, and writing $\langle A, B\rangle := \operatorname{Tr}(A^T B) = \sum_{i=1}^d \sum_{j=1}^d A_{ij}B_{ij}$ gives an inner product on $L(V)$. This induces the *Frobenius norm* on $L(V)$, which is distinct from the operator norm by the previous item, but equivalent to it by Proposition 2.3.

Observe that we use the same notation $\|\cdot\|$ for the original norm on V and for the operator norm on $L(V)$. This could be avoided by writing $|\cdot|$ for the norm on V, but we will continue to use the same notation for both, trusting to the fact that one can always determine which norm is meant by looking at whether we are taking the norm of a vector in V or of a linear map $V \to V$.

Example 2.5. Consider the following matrices in $\mathbb{R}^{2\times 2} \cong L(\mathbb{R}^2)$:

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad C = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

One can quickly see that $\|A\| = 2$ and that $\|B\| = 1$. It takes a little bit more work to compute $\|C\|$, which turns out to be $\frac{1}{2}(1+\sqrt{5})$ if $\mathbb{R}^2$ is equipped with the Euclidean norm. If $\mathbb{R}^2$ is given the L^1-norm or L^∞-norm, then $\|C\| = 2$.

2.3 Matrix exponentials: proofs

Now we return to the matrix exponential, following some of the exposition in **Hirsch–Smale**.

Theorem 2.6. *For all $A \in \mathbb{R}^{d\times d}$, the limit*

$$e^A := \sum_{n=0}^{\infty} \frac{1}{n!} A^n = \lim_{N\to\infty} \sum_{n=0}^{N} \frac{1}{n!} A^n \tag{2.3}$$

exists in $\mathbb{R}^{d\times d}$, and the series converges absolutely in the operator norm.[12] *Moreover, we have $\|e^A\| \le e^{\|A\|}$.*

Proof. Consider for each $N \in \mathbb{N}$ the matrix

$$S_N := \sum_{n=0}^{N} \frac{1}{n!} A^n.$$

[12] In other words, $\sum_{n=0}^{\infty} \frac{1}{n!}\|A^n\| < \infty$.

We claim that this sequence is Cauchy in $\mathbb{R}^{d\times d}$. Indeed, given any $\epsilon > 0$, there exists $k \in \mathbb{N}$ such that $\sum_{n=k}^{\infty} \frac{1}{n!}\|A\|^n < \epsilon$ (by absolute convergence of the power series for the exponential function on $\mathbb{R}$), and so for any $N \geq M \geq k$, we have

$$\|S_N - S_M\| = \Big\| \sum_{n=M}^{N-1} \frac{1}{n!} A^n \Big\| \leq \sum_{n=M}^{N-1} \frac{1}{n!}\|A\|^n \leq \sum_{n=k}^{\infty} \frac{1}{n!}\|A\|^n < \epsilon. \tag{2.4}$$

Since $\mathbb{R}^{d\times d}$ is complete, this proves the claim regarding convergence. For the norm inequality, it suffices to observe that putting $M = k = 0$ in (2.4) gives

$$\|S_N\| \leq \sum_{n=0}^{\infty} \frac{1}{n!}\|A\|^n = e^{\|A\|}. \qquad \square$$

Remark 2.7. This procedure is not limited to the exponential function. Given $R > 0$ and a function $f\colon (-R, R) \to \mathbb{R}$ whose Taylor series around 0 converges absolutely to f on $(-R, R)$, we can consider any $A \in \mathbb{R}^{d\times d}$ with $\|A\| < R$, we can define $f(A)$ by using A as the variable in the Taylor series of f.

Exercise 2.8. Show that given $A \in \mathbb{R}^{d\times d}$, we have $e^A = \lim_{n\to\infty}(\mathbb{I} + \frac{A}{n})^n$.

Given real numbers $\lambda_1, \dots, \lambda_d \in \mathbb{R}$, let $\operatorname{diag}(\lambda_1, \dots, \lambda_d) \in \mathbb{R}^{d\times d}$ be the diagonal matrix A such that $A_{ii} = \lambda_i$ for each i, and $A_{ij} = 0$ when $i \neq j$. Observe that

$$\big(\operatorname{diag}(\lambda_1, \dots, \lambda_d)\big)^n = \operatorname{diag}(\lambda_1^n, \dots, \lambda_d^n) \quad \text{for all } n \in \mathbb{N},$$

and thus (2.3) immediately gives

$$e^{\operatorname{diag}(\lambda_1,\dots,\lambda_d)} = \operatorname{diag}(e^{\lambda_1}, \dots, e^{\lambda_d}). \tag{2.5}$$

To compute exponentials of non-diagonal matrices, we can use the following fact. Recall that $GL(d, \mathbb{R})$ denotes the set of *invertible* $d \times d$ real matrices.

Proposition 2.9. *Given $A \in \mathbb{R}^{d\times d}$ and $P \in GL(d, \mathbb{R})$, the matrix $B = PAP^{-1}$ has exponential given by $e^B = Pe^AP^{-1}$.*

Proof. Write $\alpha_n = \sum_{k=0}^{n} \frac{1}{k!}A^k$ and $\beta_n = \sum_{k=0}^{n} \frac{1}{k!}B^k$ for the partial sums associated to the matrix exponentials of A and B. Observe that $B^k = PA^kP^{-1}$ for all k, so

$$\beta_n = \sum_{k=0}^{n} \frac{1}{k!} PA^kP^{-1} = P\alpha_nP^{-1}.$$

Sending $n \to \infty$ gives the result.[13] $\square$

[13] To write this more carefully: $\|B - \beta_n\| = \|PAP^{-1} - P\alpha_nP^{-1}\| \leq \|P\|\|A - \alpha_n\|\|P^{-1}\| \to 0$.

From (2.5) and Proposition 2.9, we see that if $A \in \mathbb{R}^{d\times d}$ is diagonalizable, so that $A = PDP^{-1}$ for some diagonal matrix $D = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$, then we have

$$e^A = P \operatorname{diag}(e^{\lambda_1}, \dots, e^{\lambda_d}) P^{-1}. \tag{2.6}$$

All of this works over $\mathbb{C}$ as well, not just $\mathbb{R}$, and the following is another version of (2.6):

Proposition 2.10. *If $v \in \mathbb{C}^d$ is an eigenvector of $A \in \mathbb{C}^{d\times d}$ with eigenvalue $\lambda \in \mathbb{C}$, then v is also an eigenvector of e^A, with eigenvalue e^λ.*

Proof. It suffices to observe that

$$e^A v = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{1}{k!} A^k v = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{1}{k!} \lambda^k v = (e^\lambda) v. \qquad \square$$

Lec 12
W 2/18

Not all matrices are diagonalizable, so (2.6) is not the complete story. We will return to this in the next section. For now, we investigate to what extent the matrix exponential satisfies the familiar properties of the usual exponential function:

$$e^{a+b} = e^a e^b \quad \text{and} \quad \frac{d}{dt} e^{ta} = t e^{ta}.$$

Proposition 2.11. *If $A, B \in \mathbb{R}^{d\times d}$ satisfy $AB = BA$, then we have $e^{A+B} = e^A e^B$. In particular, $e^{-A} = (e^A)^{-1}$.*

Proposition 2.11 implies that e^A is always invertible, so that the exponential map is defined from $\mathbb{R}^{d\times d}$ to $GL(d, \mathbb{R})$. To prove the proposition, we will need the following observation about the product of two infinite series of matrices.

Lemma 2.12. *Suppose that $A = \sum_{j=0}^{\infty} A_j$ and $B = \sum_{k=0}^{\infty} B_k$ are two absolutely convergent series in $\mathbb{R}^{d\times d}$. For each $\ell \in \mathbb{N}_0$, let $C_\ell := \sum_{j+k=\ell} A_j B_k$, where the sum is over all pairs $(j,k) \in \mathbb{N}_0^2$ such that $j + k = \ell$. Then the series $C = \sum_{\ell=0}^{\infty} C_\ell$ is absolutely convergent, and $AB = C$.*

Proof. First we show that $\sum C_\ell$ is absolutely convergent:

$$\sum_{\ell=0}^{\infty} \|C_\ell\| \le \sum_{\ell=0}^{\infty} \sum_{j=0}^{\ell} \sum_{k=0}^{\ell-j} \|A_j\| \cdot \|B_k\| \le \Big(\sum_{j=0}^{\infty} \|A_j\| \Big) \Big(\sum_{k=0}^{\infty} \|B_k\| \Big) < \infty.$$

To show that $AB = C$, write the partial sums of the three series as

$$\alpha_n := \sum_{j=0}^{m} A_j, \qquad \beta_n := \sum_{k=0}^{n} B_k, \qquad \gamma_n := \sum_{\ell=0}^{n} C_\ell.$$

Then we have

$$\begin{aligned}\|AB - \alpha_n\beta_n\| &\le \|AB - \alpha_n B\| + \|\alpha_n B - \alpha_n\beta_n\| \\ &\le \|A - \alpha_n\|\|B\| + \|\alpha_n\|\|B - \beta_n\|.\end{aligned}$$

Since $\alpha_n \le \sum_{j=0}^{\infty} \|A_j\| < \infty$, we conclude that $AB = \lim_{n\to\infty} \alpha_n\beta_n$. It remains to show that the sequences $(\alpha_n\beta_n)_n$ and $(\gamma_n)_n$ have the same limit, so that $AB = C$.

To this end, observe that writing $I_n := \{(j,k) \in \mathbb{N}_0^2 : j + k \le n\}$, we have

$$\gamma_n = \sum_{\ell=0}^{n} C_\ell = \sum_{(j,k)\in I_n} A_j B_k,$$

and moreover, $I_{2n} = \{0, 1, \dots, n\}^2 \sqcup I'_{2n} \sqcup I''_{2n}$, where

$$\begin{aligned}I'_{2n} &:= \{(j,k) \in I_{2n} : 0 \le j \le n \text{ and } k > n\}, \\ I''_{2n} &:= \{(j,k) \in I_{2n} : j > n \text{ and } 0 \le k \le n\}.\end{aligned}$$

Thus we have

$$\gamma_{2n} = \sum_{(j,k)\in I_{2n}} A_j B_k = \sum_{j=0}^{n}\sum_{k=0}^{n} A_j B_k + \sum_{(j,k)\in I'_{2n}} A_j B_k + \sum_{(j,k)\in I''_{2n}} A_j B_k.$$

Since $\sum_{j=0}^{n}\sum_{k=0}^{n} A_j B_k = \alpha_n\beta_n$, we obtain

$$\begin{aligned}\|\gamma_{2n} - \alpha_n\beta_n\| &\le \sum_{(j,k)\in I'_{2n}} \|A_j B_k\| + \sum_{(j,k)\in I''_{2n}} \|A_j B_k\| \\ &\le \sum_{j=0}^{n} \|A_j\| \sum_{k=n+1}^{\infty} \|B_k\| + \sum_{j=n+1}^{\infty} \|A_j\| \sum_{k=0}^{n} \|B_k\|.\end{aligned}$$

Since $\sum_{j=0}^{\infty} \|A_j\| < \infty$ and $\sum_{k=0}^{\infty} \|B_k\| < \infty$, this last quantity goes to 0 as $n \to \infty$, which proves that $C = \lim_n \gamma_{2n} = \lim_n \alpha_n\beta_n = AB$. □

Proof of Proposition 2.11. Using the binomial theorem and the fact that $AB = BA$, we have

$$(A + B)^n = \sum_{j+k=n} \binom{n}{k} A^j B^k = n! \sum_{j+k=n} \frac{A^j}{j!}\frac{B^k}{k!},$$

from which we obtain

$$e^{A+B} = \sum_{n=0}^{\infty} \sum_{j+k=n} \frac{A^j}{j!}\frac{B^k}{k!}.$$

By Lemma 2.12, this is equal to $e^A e^B$. The claim about e^{-A} follows immediately from the fact that A and $-A$ commute. □

Remark 2.13. When A and B do not commute, Proposition 2.11 must be replaced by the *Baker–Campbell–Hausdorff formula*: writing $[A, B] := AB - BA$ for the commutator of A and B, we have

$$e^A e^B = e^C, \quad \text{where } C = A + B + \frac{1}{2}[A, B] + \frac{1}{12}[A, [A, B]] + \frac{1}{12}[B, [A, B]] + \cdots,$$

where the final expression is an infinite series whose omitted terms involve higher-order iterated commutators.

We are now in a position to prove that as expected, matrix exponentials provide a solution to the initial value problem

$$\dot{x} = Ax, \qquad x(0) = x_0. \tag{2.7}$$

Theorem 2.14. *For every $A \in \mathbb{R}^{d\times d}$, we have $\frac{d}{dt}e^{tA} = Ae^{tA}$. In particular, for each $x_0 \in \mathbb{R}^d$, the unique solution to (2.7) is $x(t) = e^{tA}x_0$.*

Proof. Using Proposition 2.11, we have $e^{tA+hA} = e^{tA}e^{hA}$ for all $t, h \in \mathbb{R}$, so

$$\begin{aligned}\frac{d}{dt}e^{tA} &= \lim_{h\to 0}\frac{1}{h}\big(e^{(t+h)A} - e^{tA}\big) = \Big(\lim_{h\to 0}\frac{1}{h}\big(e^{hA} - \mathbb{I}\big)\Big)e^{tA} \\ &= \Big(\lim_{h\to 0}\frac{1}{h}\sum_{n=1}^{\infty}\frac{1}{n!}h^n A^n\Big)e^{tA} = Ae^{tA}.\end{aligned}$$

It follows that $\frac{d}{dt}(e^{tA}x_0) = Ae^{tA}x_0$, which completes the proof. □

Corollary 2.15. *If $x_0 \in \mathbb{R}^d$ is an eigenvector of $A \in \mathbb{R}^{d\times d}$ with eigenvalue $\lambda \in \mathbb{R}$, then $x(t) = e^{t\lambda}x_0$ is a solution of the IVP given by $\dot{x} = Ax$ and $x(0) = x_0$.*

Proof. It suffices to observe that $(tA)x_0 = (t\lambda)x_0$, so that by Proposition 2.10, we have $e^{tA}x_0 = e^{t\lambda}x_0$. □

The graph of the solution in Corollary 2.15 lies on a straight line in $\mathbb{R}^d$. In the next section, we will explore what the other solutions look like in the case $d = 2$.

2.4 Linear DEs in the plane

2.4.1 Diagonalizable matrices

Lec 13
F 2/20

Let us work through an example. Consider the matrix $A = \left(\begin{smallmatrix} -3 & 2 \\ -5 & 4 \end{smallmatrix}\right)$ and the corresponding linear DE $\dot{x} = Ax$ in $\mathbb{R}^2$. By Theorem 2.14, the solutions of this DE have the form $x(t) = e^{tA}x_0$, so we must compute e^{tA}. If A can be diagonalized, then we can compute e^{tA} using (2.5). To this end, observe that the characteristic polynomial of A is

$$\lambda^2 - (\operatorname{Tr} A)\lambda + \det A = \lambda^2 - \lambda - 2 = (\lambda - 2)(\lambda + 1),$$

so the eigenvalues are $\lambda_1 = -1$ and $\lambda_2 = 2$. A short computation shows that we can use $v_1 = \left(\begin{smallmatrix}1\\1\end{smallmatrix}\right)$ and $v_2 = \left(\begin{smallmatrix}2\\5\end{smallmatrix}\right)$ as a corresponding basis of eigenvectors, and the matrix $P = (v_1 \mid v_2) = \left(\begin{smallmatrix}1&2\\1&5\end{smallmatrix}\right)$ maps e_i to v_i, diagonalizing A:

$$A = PDP^{-1}, \quad \text{where } D := \begin{pmatrix}-1 & 0\\ 0 & 2\end{pmatrix}.$$

(The equality can be checked by observing that both sides map v_i to $\lambda_i v_i$.)

Observe that $e^{tD} = \left(\begin{smallmatrix}e^{-t}&0\\0&e^{2t}\end{smallmatrix}\right)$. Since $P^{-1} = \frac{1}{3}\left(\begin{smallmatrix}5&-2\\-1&1\end{smallmatrix}\right)$, we have

$$\begin{aligned} e^{tA} &= \begin{pmatrix}1&2\\1&5\end{pmatrix}\begin{pmatrix}e^{-t}&0\\0&e^{2t}\end{pmatrix}\cdot\frac{1}{3}\begin{pmatrix}5&-2\\-1&1\end{pmatrix} = \frac{1}{3}\begin{pmatrix}1&2\\1&5\end{pmatrix}\begin{pmatrix}5e^{-t}&-2e^{-t}\\-e^{2t}&e^{2t}\end{pmatrix}\\ &= \frac{1}{3}\begin{pmatrix}5e^{-t}-2e^{2t}&-2e^{-t}+2e^{2t}\\5e^{-t}-5e^{2t}&-2e^{-t}+5e^{2t}\end{pmatrix}. \end{aligned}$$

Multiplying this by $v_1 = \left(\begin{smallmatrix}1\\1\end{smallmatrix}\right)$ and $v_2 = \left(\begin{smallmatrix}2\\5\end{smallmatrix}\right)$ produces the straight-line solutions

$$x_1(t) = \begin{pmatrix}e^{-t}\\e^{-t}\end{pmatrix} \quad \text{and} \quad x_2(t) = \begin{pmatrix}2e^{2t}\\5e^{2t}\end{pmatrix},$$

consistent with Corollary 2.15. Observe that the solution for the negative eigenvalue $\lambda_1 = -1 < 0$ approaches the origin as $t \to \infty$, while the solution for the positive eigenvalue $\lambda_2 = 2$ approaches the origin as $t \to -\infty$.

More general solutions can also be written down using the above formula for e^{tA}. for example, the solution to $\dot{x} = Ax$ with $x(0) = \left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)$ is

$$x(t) = \begin{pmatrix}\frac{5}{3}e^{-t} - \frac{2}{3}e^{2t}\\ \frac{5}{3}e^{-t} - \frac{5}{3}e^{2t}\end{pmatrix}. \tag{2.8}$$

Note that this solution approaches the straight-line solution $x_2(t)$ as $t \to \infty$, and the straight-line solution $x_1(t)$ as $t \to -\infty$, both of which correspond to a linear asymptote of the parametrized curve $t \mapsto x(t) \in \mathbb{R}^2$.

Now suppose A is any 2×2 real matrix with distinct real eigenvalues. In this case, the corresponding (real) eigenvectors give a diagonalization of A just as above. If the eigenvalues have different signs, then we have the same sort of situation as above: one eigenspace contains straight-line solutions going to 0 as $t \to \infty$. Call this the *stable subspace*. Another eigenspace contains straight-line solutions going to 0 as $t \to -\infty$. Call this the *unstable subspace*. Every other solution curve approaches the stable subspace as $t \to -\infty$, and the unstable subspace as $t \to \infty$. In the diagonal case, if

$A = \left(\begin{smallmatrix} \lambda & 0 \\ 0 & -\mu \end{smallmatrix}\right)$ with $\lambda, \mu > 0$, then the general solution has the form

$$e^{tA}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} e^{t\lambda} & 0 \\ 0 & e^{-t\mu} \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} ae^{t\lambda} \\ be^{-t\mu} \end{pmatrix} \quad \text{for some } a, b \in \mathbb{R}. \tag{2.9}$$

We see that every point (x_1, x_2) on this curve satisfies

$$x_1^\mu x_2^\lambda = a^\mu e^{t\lambda\mu} b^\lambda e^{-t\mu\lambda} = a^\mu b^\lambda.$$

When $\lambda = \mu$, this implies that the curve is a hyperbola.

The constant solution $x(t) \equiv 0$ of $\dot{x} = Ax$ is called an *equilibrium solution* of the DE, or a *fixed point* of the flow. In the above situation, when A has one positive eigenvalue and one negative eigenvalue, we say that the origin is a *hyperbolic fixed point*.[14] The crucial property of such a fixed point is that there exist flow-invariant subspaces $E^u, E^s \subset \mathbb{R}^2$ such that

- $\mathbb{R}^2 = E^u \oplus E^s$;
- every solution in E^s converges exponentially fast to 0 as $t \to \infty$; and
- every solution in E^u converges exponentially fast to 0 as $t \to -\infty$.

Of course, the fixed point at the origin does not need to be hyperbolic; there are other possibilities as well, and we now explore these.

Start by considering what else can happen when A is diagonalizable over $\mathbb{R}$. This means that there are $\lambda, \mu \in \mathbb{R}$ and $v, w \in \mathbb{R}^2$ such that $Av = \lambda v$ and $Aw = \mu w$, so that $P = (v|w)$ gives $A = PDP^{-1}$, where $D = \left(\begin{smallmatrix} \lambda & 0 \\ 0 & \mu \end{smallmatrix}\right)$, as above. The case where $\lambda = 0$ or $\mu = 0$ is not particularly interesting (work through it yourself!), so we focus on nonzero eigenvalues. We already dealt with the case where λ and μ have opposite signs, so now suppose they have the same sign; either both positive, or both negative.[15]

For concreteness, consider the case $\lambda \geq \mu > 0$. (Reversing time will describe the negative eigenvalue situation.) Then we have

$$e^{tA} = Pe^{tD}P^{-1} = P\begin{pmatrix} e^{t\lambda} & 0 \\ 0 & e^{t\mu} \end{pmatrix}P^{-1},$$

and we see that for every $x_0 \in \mathbb{R}^2 \setminus \{0\}$, the corresponding solution $e^{tA}x_0$ goes to 0 as $t \to -\infty$, and to ∞ as $t \to \infty$. We say that the origin is a *repelling fixed point*, or *source* (negative eigenvalues give an *attracting fixed point*, or *sink*).

[14] It is sometimes also called a *saddle*, after the appearance of the surface given as the graph of $(x_1, x_2) \mapsto x_1^\mu x_2^\lambda$, whose level curves describe the trajectories.

[15] The fixed point is sometimes called hyperbolic in this situation as well; it fits the above framework by taking one of E^u or E^s to be trivial. We will reserve the use of the word *hyperbolic* for the case in which both are non-trivial.

When $\lambda = \mu$, we see that $D = \lambda\mathbb{I}$ so $A = P(\lambda\mathbb{I})P^{-1} = \lambda\mathbb{I}$, and $e^{tA} = e^{t\lambda}\mathbb{I}$. Thus *every* trajectory lies on a straight line radiating from the origin. This situation is called a *focus*.

When $\lambda > \mu$, we can write the general initial condition as $av + bw$, and obtain the solution

$$x(t) = e^{tA}(av + bw) = ae^{tA}v + be^{tA}w = ae^{t\lambda}v + be^{t\mu}w.$$

As $t \to -\infty$, the $e^{t\lambda}$ term decays faster than the $e^{t\mu}$ term, so $x(t)$ approaches the origin along the direction of w. This situation is called a *node*, and is best illustrated by considering the case when $A = \left(\begin{smallmatrix}\lambda & 0\\ 0 & \mu\end{smallmatrix}\right)$, so just as in (2.9), the solution with initial condition $\left(\begin{smallmatrix}a\\ b\end{smallmatrix}\right)$ is given by

$$e^{tA}\begin{pmatrix}a\\ b\end{pmatrix} = \begin{pmatrix}ae^{t\lambda}\\ be^{t\mu}\end{pmatrix},$$

and every point (x_1, x_2) on this curve satisfies

$$x_1^{\mu}x_2^{-\lambda} = a^{\mu}e^{t\lambda\mu}b^{-\lambda}e^{-t\mu\lambda} = a^{\mu}b^{-\lambda} =: c \quad \Rightarrow \quad x_1 = cx_2^{\lambda/\mu}.$$

2.4.2 Distinct complex eigenvalues

Lec 14
M 2/23

Now we consider the case when $A \in \mathbb{R}^{2\times 2}$ has two *complex* eigenvalues, so that it is diagonalizable over $\mathbb{C}$, but not over $\mathbb{R}$. Observe that if $Aw = \lambda w$, then taking complex conjugates gives $A\bar{w} = \bar{\lambda}\bar{w}$, so we can write the eigenvalues as $\lambda = a + ib \in \mathbb{C}$ and $\bar{\lambda} = a - ib \in \mathbb{C}$, with associated eigenvectors w and $\bar{w}$. Writing

$$v_1 := w + \bar{w} \quad \text{and} \quad v_2 = i(w - \bar{w}),$$

we see that $v_1, v_2 \in \mathbb{R}^2$, and we have

$$\begin{aligned} Av_1 &= Aw + A\bar{w} = \lambda w + \bar{\lambda}\bar{w} = (a + ib)w + (a - ib)\bar{w}\\ &= a(w + \bar{w}) + bi(w - \bar{w}) = av_1 + bv_2, \end{aligned}$$

and similarly,

$$\begin{aligned} Av_2 &= i(Aw - A\bar{w}) = i\lambda w - i\bar{\lambda}\bar{w} = i(a + ib)w - i(a - ib)\bar{w}\\ &= -b(w + \bar{w}) + ai(w - \bar{w}) = -bv_1 + av_2, \end{aligned}$$

so that once again writing $P = (v_1|v_2)$, we have

$$A = PZP^{-1}, \quad \text{where } Z = \begin{pmatrix}a & -b\\ b & a\end{pmatrix}, \tag{2.10}$$

as can be verified by applying both A and PZP^{-1} to v_1 and v_2.

Lemma 2.16. *Given $a, b \in \mathbb{R}$, let $Z = \left(\begin{smallmatrix}a & -b\\ b & a\end{smallmatrix}\right)$. Then $e^Z = e^a\left(\begin{smallmatrix}\cos b & -\sin b\\ \sin b & \cos b\end{smallmatrix}\right)$.*

Proof. Let $J := \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, so that $Z = a\mathbb{I} + bJ$. Observe that $\mathbb{I}J = J = J\mathbb{I}$, so by Proposition 2.11, we have

$$e^Z = e^{a\mathbb{I}+bJ} = e^{a\mathbb{I}}e^{bJ} = e^a\mathbb{I}e^{bJ} = e^a e^{bJ}.$$

Using the fact that $J^2 = -\mathbb{I}$ and recalling the Taylor series for cosine and sine, we get

$$\begin{aligned} e^{bJ} &= \mathbb{I} + bJ - \frac{b^2}{2}\mathbb{I} - \frac{b^3}{6}J + \cdots \\ &= (\cos b)\mathbb{I} + (\sin b)J = \begin{pmatrix} \cos b & -\sin b \\ \sin b & \cos b \end{pmatrix}. \end{aligned}$$

□

Observe that $\left(\begin{smallmatrix} \cos b & -\sin b \\ \sin b & \cos b \end{smallmatrix}\right)$ acts on $\mathbb{R}^2$ as a rotation around the origin by an angle b, so by Lemma 2.16, e^Z acts on $\mathbb{R}^2$ as scaling by e^a composed with a rotation by b. This is the same geometric effect as multiplication by $e^{a+ib} = e^a e^{ib}$ has on $\mathbb{C}$, so the matrix $Z = \left(\begin{smallmatrix} a & -b \\ b & a \end{smallmatrix}\right)$ can be interpreted as a representation of the complex number $a + ib$.

Returning to (2.10), we see that when $A \in \mathbb{R}^{2\times 2}$ has complex eigenvalues $a \pm ib$ with $b \neq 0$, we can write P as above and obtain

$$e^{tA} = Pe^{ta}\begin{pmatrix} \cos tb & -\sin tb \\ \sin tb & \cos tb \end{pmatrix} P^{-1}.$$

When $a = 0$, the solutions are ellipses around the fixed point at the origin, which is called a *center*; nearby solutions neither converge towards it nor converge away from it. When $a < 0$, solutions spiral in towards this fixed point, so it is attracting (a sink); when $a > 0$, they spiral away from it, so it is repelling (a source).

2.4.3 A Jordan block

The final case to consider is when $A \in \mathbb{R}^{2\times 2}$ is *not* diagonalizable over $\mathbb{C}$. This is only possible if A has a single eigenvalue $\lambda \in \mathbb{R}$ with algebraic multiplicity 2 and geometric multiplicity 1. The eigenspace $\ker(A - \lambda\mathbb{I})$ is a one-dimensional subspace of $\mathbb{R}^2$, so $A - \lambda\mathbb{I}$ has rank 1. Its range is A-invariant: given $x \in \mathbb{R}^2$ and $y = (A - \lambda\mathbb{I})x$, we have

$$Ay = A(A - \lambda\mathbb{I})x = (A - \lambda\mathbb{I})(Ax) \in \operatorname{ran}(A - \lambda\mathbb{I}). \tag{2.11}$$

Since $\ker(A - \lambda\mathbb{I})$ is the *only* A-invariant one-dimensional subspace of $\mathbb{R}^2$, we obtain $\operatorname{ran}(A - \lambda\mathbb{I}) = \ker(A - \lambda\mathbb{I})$. In particular, given any $w \in \mathbb{R}^2 \setminus \ker(A - \lambda\mathbb{I})$, we see that

Lec 15
W 2/25

$$v := (A - \lambda\mathbb{I})w = Aw - \lambda w \in \ker(A - \lambda\mathbb{I})$$

is an eigenvector, so $Av = \lambda v$. Let $P = (v \mid w)$. Since $Aw = v + \lambda w$, we have

$$A = PJP^{-1}, \quad \text{where } J = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}, \tag{2.12}$$

as can be verified by applying both A and PJP^{-1} to v and w.

Let $S = \lambda\mathbb{I}$ and $N = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$, so that $J = S + N$. (The notation stands for *semisimple* and *nilpotent*, terminology which will be explained later.) Observe that $SN = NS$ and that $N^2 = 0$, so we have

$$\begin{aligned} e^{tJ} &= e^{tS}e^{tN} = e^{t\lambda}\mathbb{I}\Big(\sum_{k=0}^{\infty} \frac{1}{k!} tN^k\Big) \\ &= e^{t\lambda}(\mathbb{I} + tN) = e^{t\lambda}\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} e^{t\lambda} & te^{t\lambda} \\ 0 & e^{t\lambda} \end{pmatrix}. \end{aligned}$$

We conclude that

$$e^{tA} = P\begin{pmatrix} e^{t\lambda} & te^{t\lambda} \\ 0 & e^{t\lambda} \end{pmatrix} P^{-1}.$$

The fixed point at the origin is attracting (a sink) if $\lambda < 0$, and repelling (a source) if $\lambda > 0$. This is sometimes called an *improper node*. (If $\lambda = 0$ then the flow acts as a shear, with orbits moving parallel to the eigendirection, which is a line of fixed points.)

In the representative case $A = \left(\begin{smallmatrix} \lambda & 1 \\ 0 & \lambda \end{smallmatrix}\right)$, you can check that solutions move along curves of the form

$$x_1 = Cx_2 + \lambda^{-1}x_2 \log|x_2|,$$

where different values of the constant C determine different solution curves. To verify this, it suffices to show that $\frac{d}{dt}(\frac{x_1}{x_2} - \lambda^{-1}\log|x_2|) = 0$; we omit this computation.

The various cases treated in the preceding sections provide a complete description of linear autonomous DEs in the plane. Notice that in each case, eigenvalues with negative real part have produced solutions that are attracted towards the fixed point at the origin, while eigenvalues with positive real part have produced solutions that are repelled away from it.

Our next task is to extend this analysis to higher dimensions, which will require some more involved linear algebra; namely, the concept of *Jordan normal form*.

2.5 Jordan normal form

2.5.1 Complex matrices

Start by considering a complex matrix $A \in \mathbb{C}^{d\times d}$. When $d = 2$, the reasoning in the previous section shows that one of two things happens: either A can be diagonalized over $\mathbb{C}$, so that there exists $P \in GL(2,\mathbb{C})$ such that

$$A = P\begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix} P^{-1}, \quad \text{where } \lambda, \mu \in \mathbb{C},$$

or A has an eigenvalue λ with algebraic multiplicity 2 and geometric multiplicity 1. In this case, there exists $P \in GL(2,\mathbb{C})$ such that

$$A = P \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} P^{-1}, \quad \text{where } \lambda \in \mathbb{C}.$$

Definition 2.17. A *Jordan block* is a square matrix $J = \lambda \mathbb{I} + N$, where $\lambda \in \mathbb{C}$ and the entries of N are given by $N_{ij} = 1$ if $j = i+1$, and 0 otherwise. That is, J has the form

$$J = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{pmatrix}. \tag{2.13}$$

Theorem 2.18 (Jordan normal form)**.** *Given any $A \in \mathbb{C}^{d\times d}$, there exists $P \in GL(d,\mathbb{C})$ such that $A = PJP^{-1}$, where J is a direct sum of Jordan blocks: that is, J can be written in block form as*

$$J = \begin{pmatrix} J_1 & 0 & \cdots & 0 \\ 0 & J_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_\ell \end{pmatrix}, \tag{2.14}$$

where each J_i is a Jordan block, and each 0 represents a zero matrix of the appropriate size.

The matrix J in (2.14) is the *Jordan normal form* (or *Jordan canonical form*) of A. Observe that the following conditions are all equivalent.

- A is diagonalizable over $\mathbb{C}$.
- $\mathbb{C}^d$ admits a basis consisting of eigenvectors for A.
- Writing $\lambda_1, \dots, \lambda_r$ for the distinct eigenvalues of A, and $E_1, \dots, E_r$ for the corresponding eigenspaces, we have $\mathbb{R}^d = E_1 \oplus \cdots \oplus E_r$.
- All the Jordan blocks in (2.14) have size 1.

To prove Theorem 2.18, we will use the notion of a *generalized eigenvector*. Start by observing that given a single Jordan block J of size k, as in (2.13), we have $(J-\lambda\mathbb{I})e_i = e_{i-1}$ for each $i = 2, 3, \dots, k$, and $(J-\lambda\mathbb{I})e_1 = 0$. We can represent this by writing

$$e_k \xmapsto{J-\lambda\mathbb{I}} e_{k-1} \xmapsto{J-\lambda\mathbb{I}} e_{k-2} \xmapsto{J-\lambda\mathbb{I}} \cdots \xmapsto{J-\lambda\mathbb{I}} e_1 \xmapsto{J-\lambda\mathbb{I}} 0.$$

Definition 2.19. Given $A \in \mathbb{C}^{d\times d}$, a nonzero vector $v \in \mathbb{C}^d$ is a *generalized eigenvector* of A for the eigenvalue $\lambda \in \mathbb{C}$ if there exists $k \in \mathbb{N}$ such that $(A-\lambda\mathbb{I})^k v = 0$. The smallest such k is the *rank* of v. The set of all generalized eigenvectors for λ, together with 0, is the *generalized eigenspace* for λ. A *Jordan chain* for λ is a sequence $v_1, \dots, v_k \in \mathbb{C}^d$ of nonzero vectors such that $(A - \lambda\mathbb{I})v_j = v_{j-1}$ for each $j = 2, \dots, k$, and $(A - \lambda\mathbb{I})v_1 = 0$; in other words,

$$v_k \xmapsto{A-\lambda\mathbb{I}} v_{k-1} \xmapsto{A-\lambda\mathbb{I}} v_{k-2} \xmapsto{A-\lambda\mathbb{I}} \cdots \xmapsto{A-\lambda\mathbb{I}} v_1 \xmapsto{A-\lambda\mathbb{I}} 0.$$

Observe that if, $v_1, \dots, v_k$ is a Jordan chain for λ, then each v_j is a generalized eigenvector of rank j. Conversely, if v is a generalized eigenvector of rank k, then writing $v_j := (A - \lambda\mathbb{I})^{k-j}$ produces a Jordan chain.

Lemma 2.20. *If $v_1, \dots, v_k$ is a Jordan chain, then $\{v_1, \dots, v_k\}$ is linearly independent.*

Proof. It suffices to show that v_j cannot be written as a lienar combination of $v_1, \dots, v_{j-1}$ for any j. This follows since $v_1, \dots, v_{j-1}$ all lie in $\ker((A-\lambda\mathbb{I})^{j-1})$, but v_j does not. □

It follows from Lemma 2.20 that every generalized eigenvector of a $d \times d$ matrix has rank $\le d$, and that the generalized eigenspace of λ is $\ker((A - \lambda\mathbb{I})^d)$.

We will prove Theorem 2.18 by showing that $\mathbb{C}^d$ has a basis consisting of generalized eigenvectors for A, organized into Jordan chains, which produce the Jordan blocks in (2.14). This will also show that while the direct sum of the eigenspaces may be a proper subspace of $\mathbb{C}^d$, the direct sum of the *generalized* eigenspaces is always all of $\mathbb{C}^d$.

Proof of Theorem 2.18. We will prove by induction on d that for every $A \in \mathbb{C}^{d\times d}$, there exist $k_1, \dots, k_\ell \in \mathbb{N}$ such that $\sum_i k_i = d$ and such that for each $i = 1, \dots, \ell$, there exists a Jordan chain $v^i_{k_i}, \dots, v^i_1 \in \mathbb{C}^d$ for some $\lambda_i \in \mathbb{C}$, with the property that the vectors $\{v^i_j : 1 \le i \le \ell, 1 \le j \le k_i\}$ form a basis for $\mathbb{C}^d$. Taking P to be the matrix whose column vectors are given first by $v^1_1, \dots, v^1_{k_1}$, then $v^2_1, \dots, v^2_{k_2}$, and so on, the conclusion of the theorem will immediately follow.

For $d = 1$, the assertion is immediate; any nonzero vector v will do. So consider $d \ge 2$, and suppose that the assertion is true for all square matrices with size $< d$. By the Fundamental Theorem of Algebra, the characteristic polynomial of A has at least one root $\lambda \in \mathbb{C}$, so $\ker(A - \lambda\mathbb{I})$ is a nontrivial subspace of $\mathbb{C}^d$. Let $n \ge 1$ be its dimension. This subspace is A-invariant, just as in (2.11).

Similarly, $\operatorname{ran}(A - \lambda\mathbb{I})$ is A-invariant, and it has dimension $d - n < d$ by the rank-nullity theorem. By the inductive hypothesis, $\operatorname{ran}(A - \lambda\mathbb{I})$ admits a basis consisting of Jordan chains. These chains may be associated to various eigenvalues of A. Let m be the number of these chains associated to the eigenvalue λ. Without loss of generality, index the chains such that these chains appear first, and consider their initial vectors

$v^i_{k_i}$, where $1 \le i \le m$. Since $v^i_{k_i} \in \operatorname{ran}(A-\lambda\mathbb{I})$, there exist vectors w^i for each $1 \le i \le m$ such that $(A-\lambda\mathbb{I})w^i = v^i_{k_i}$.

Observe that $\{v^i_1 : 1 \le i \le m\}$ is a basis for $\ker(A-\lambda\mathbb{I}) \cap \operatorname{ran}(A-\lambda\mathbb{I})$, so this intersection has dimension m. It follows that $V := \ker(A-\lambda\mathbb{I}) + \operatorname{ran}(A-\lambda\mathbb{I})$ has dimension $d-m$. Moreover, V admits a basis consisting of Jordan chains; simply extend $\{v^i_1 : 1 \le i \le m\}$ to a basis for $\ker(A-\lambda\mathbb{I})$, and append each of these eigenvectors to the basis for $\operatorname{ran}(A-\lambda\mathbb{I})$, treating it as a one-element Jordan chain. Finally, appending the vectors w^i to the first m Jordan chains, as $v^i_{k_i+1} := w^i$, and using Lemma 2.20, we obtain the desired basis for $\mathbb{C}^d$, completing the proof of the theorem. □

Lec 16
F 2/27

The following results can be proved as consequences of Theorem 2.18, using Jordan normal form.

- *Cayley–Hamilton Theorem*: If $p(\lambda)$ is the characteristic polynomial of A, then $p(A) = 0$.
- The trace of A is the sum of the eigenvalues of A, counted with algebraic multiplicity.
- The determinant of A is the product of the eigenvalues of A, counted with algebraic multiplicity.
- $\det e^A = e^{\operatorname{Tr} A}$.
- A and A^T are conjugate.

Definition 2.21. A $d \times d$ matrix A is *semisimple* if it can be diagonalized over $\mathbb{C}$. It is *nilpotent* if there exists $k \in \mathbb{N}$ such that $A^k = 0$. A linear transformation $T \in L(\mathbb{R}^d)$ is semisimple (respectively, nilpotent), if and only if the matrix representing it is.

The following facts are left as exercises.

- If A is both semisimple and nilpotent, then $A = 0$.
- A matrix A is semisimple if and only if every invariant subspace admits a complementary invariant subspace: that is, for every subspace $V \subset \mathbb{C}^d$ such that $AV \subset V$, there exists a subspace $W \subset \mathbb{C}^d$ such that $AW \subset W$ and $\mathbb{C}^d = V \oplus W$.
- If the characteristic polynomial of $A \in \mathbb{C}^{d\times d}$ has d distinct roots, then A is semisimple.

Observe that a matrix in Jordan normal form can be written as the sum of two commuting matrices, one of which is diagonal and the other of which is strictly upper-triangular matrix, and hence nilpotent. Thus Theorem 2.18 has the following consequence: every $A \in \mathbb{C}^{d\times d}$ can be written as $A = S + N$, where S is semisimple, N is nilpotent, and $SN = NS$.

2.5.2 Real matrices

When $A \in \mathbb{R}^{d\times d}$, we are generally most interested in a *real* canonical form. This can be achieved as in §2.4.2, by observing that if $w \in \mathbb{C}^d$ is an eigenvector of A for an eigenvalue λ, then $\bar{w}$ is an eigenvector for $\bar{\lambda}$. Extending this idea, the arguments in the previous section can be used to produce a basis for $\mathbb{C}^d$ in which the Jordan chains associated to non-real eigenvalues appear in complex conjugate pairs. Then just as in §2.4.2, we can replace each complex conjugate pair of basis vectors in $\mathbb{C}^d$ with its real and imaginary parts, obtaining a basis for $\mathbb{R}^d$ in which each complex matrix $\left(\begin{smallmatrix}\lambda & 0\\ 0 & \bar{\lambda}\end{smallmatrix}\right)$ is replaced with a real matrix $\left(\begin{smallmatrix}a & -b\\ b & a\end{smallmatrix}\right)$, where $\lambda = a+ib$. The associated *real* Jordan block will be given by replacing each λ in (2.13) with the 2×2 matrix $\left(\begin{smallmatrix}a & -b\\ b & a\end{smallmatrix}\right)$, each 1 with the 2×2 identity matrix, and each 0 with the 2×2 zero matrix.

In terms of the semisimple-nilpotent decomposition, we conclude that every $A \in \mathbb{R}^{d\times d}$ can be written as $A = S + N$, where:

- S is similar to a block diagonal matrix given as the direct sum of 1×1 matrices associated to real eigenvalues, and 2×2 matrices of the form $\left(\begin{smallmatrix}a & -b\\ b & a\end{smallmatrix}\right)$;
- N is nilpotent; and
- S and N commute.

2.5.3 Exponentials of Jordan blocks

Using the semisimple-normal decomposition, we can now compute arbitrary matrix exponentials: since S and N commute, we have

$$e^A = e^{S+N} = e^S e^N = e^S \sum_{k=0}^{d} \frac{1}{k!} N^k,$$

where the sum is finite because all higher powers of N vanish, and we observe that e^S can be computed using the block diagonal form, together with the formula for the exponential of $\left(\begin{smallmatrix}a & -b\\ b & a\end{smallmatrix}\right)$ from §2.4.2. It is also helpful to observe the following behavior of the normal form for the nilpotent part, which we illustrate in the case $d = 4$:

$$tN = \begin{pmatrix} 0 & t & 0 & 0\\ 0 & 0 & t & 0\\ 0 & 0 & 0 & t\\ 0 & 0 & 0 & 0\end{pmatrix}, \qquad (tN)^2 = \begin{pmatrix} 0 & 0 & t & 0\\ 0 & 0 & 0 & t\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{pmatrix}, \qquad (tN)^3 = \begin{pmatrix} 0 & 0 & 0 & t\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{pmatrix}.$$

Since $N^4 = 0$ in this case, we get

$$e^{tN} = \begin{pmatrix} 1 & t & \frac{1}{2}t^2 & \frac{1}{3!}t^3\\ 0 & 1 & t & \frac{1}{2}t^2\\ 0 & 0 & 1 & t\\ 0 & 0 & 0 & 1\end{pmatrix}.$$

The pattern continues for larger values of d.

To see all of this in action, let us solve $\dot{x} = Ax$ for the matrix

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

The characteristic polynomial is $(\lambda - 1)(\lambda - 2)^2$. A short computation shows that $\lambda_1 = 1$ has $v_1 = (1, 1, -2)^T$ as an eigenvector, and $\lambda_2 = 2$ has $v_2 = (0, 0, 1)^T$ as an eigenvector; however, there is no second linearly independent eigenvector for λ_2, so we need a generalized eigenvector, obtained by solving $(A - 2\mathbb{I})v_3 = v_2$, and we see that $v_3 = (0, 1, 0)^T$ does the job. Thus we take

$$P = (v_1 \mid v_2 \mid v_3) = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ -2 & 1 & 0 \end{pmatrix} \qquad \Rightarrow \qquad A = PJP^{-1}, \text{ where } J = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}.$$

Separating the diagonal (semisimple) from the off-diagonal (nilpotent) part, and using the fact that the nilpotent part squares to give the 0 matrix, we obtain

$$e^{tJ} = \begin{pmatrix} e^t & 0 & 0 \\ 0 & e^{2t} & 0 \\ 0 & 0 & e^{2t} \end{pmatrix} \left(\mathbb{I} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & t \\ 0 & 0 & 0 \end{pmatrix} \right) = \begin{pmatrix} e^t & 0 & 0 \\ 0 & e^{2t} & 0 \\ 0 & 0 & e^{2t} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{pmatrix}.$$

This lets us solve the initial problem:

$$e^{tA} = Pe^{tJ}P^{-1} = P \begin{pmatrix} e^t & 0 & 0 \\ 0 & e^{2t} & te^{2t} \\ 0 & 0 & e^{2t} \end{pmatrix} P^{-1}.$$

Here is another example. Consider

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

whose characteristic polynomial is $\big((3 - \lambda)^2 + 4\big)^2$, so the eigenvalues are $3 \pm 2i$. One can check that $\lambda = 3 + 2i$ has geometric multiplicity 1, and that $v = (0, i, 1, 0)^T$ is an eigenvector. Solving $(A - \lambda\mathbb{I})w = v$ gives $w = (i, 0, -1, 1)^T$ as a generalized eigenvector, and taking the matrix $P \in GL(4, \mathbb{R})$ whose columns are the real and imaginary parts of v and w, we get $A = PJP^{-1}$, where

$$P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad P^{-1} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad J = \begin{pmatrix} 3 & -2 & 1 & 0 \\ 2 & 3 & 0 & 1 \\ 0 & 0 & 3 & -2 \\ 0 & 0 & 2 & 3 \end{pmatrix}.$$

We can compute the exponential of this real Jordan normal form using the ideas described above, obtaining

$$e^{tJ} = e^{3t}\begin{pmatrix} \cos 2t & -\sin 2t & 0 & 0 \\ \sin 2t & \cos 2t & 0 & 0 \\ 0 & 0 & \cos 2t & -\sin 2t \\ 0 & 0 & \sin 2t & \cos 2t \end{pmatrix}\begin{pmatrix} 1 & 0 & t & 0 \\ 0 & 1 & 0 & t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

2.6 Stable, unstable, and central subspaces

Lec 17
M 3/2

Now we can describe the qualitative behavior of solutions of $\dot{x} = Ax$ for any $A \in \mathbb{R}^{d\times d}$. Let $\Lambda \subset \mathbb{C}$ be the set of eigenvalues of A. Given $\lambda \in \Lambda$, let $E^{\mathbb{C}}_\lambda = \ker(A - \lambda \mathbb{I})^d \subset \mathbb{C}^d$ be the corresponding generalized eigenspace. By Theorem 2.18, we have $\mathbb{C}^d = \bigoplus_{\lambda\in\Lambda} E^{\mathbb{C}}_\lambda$.

The corresponding real subspaces are given by

$$E^{\mathbb{R}}_\lambda := \begin{cases} E^{\mathbb{C}}_\lambda \cap \mathbb{R}^d & \text{if } \lambda \in \mathbb{R}, \\ (E^{\mathbb{C}}_\lambda \oplus E^{\mathbb{C}}_{\bar\lambda}) \cap \mathbb{R}^d & \text{if } \lambda \notin \mathbb{R}. \end{cases}$$

Observe that $E^{\mathbb{R}}_\lambda = E^{\mathbb{R}}_{\bar\lambda}$, so to get the Jordan decomposition of $\mathbb{R}^d$, we write $\Lambda_1 := \{\lambda \in \Lambda : \operatorname{Im}\lambda \geq 0\}$, and obtain

$$\mathbb{R}^d = \bigoplus_{\lambda\in\Lambda_1} E^{\mathbb{R}}_\lambda. \tag{2.15}$$

Each $E^{\mathbb{R}}_\lambda$ is A-invariant and hence flow-invariant. We partition the eigenvalues according to the sign of their real part: let

$$\Lambda^s := \{\lambda \in \Lambda_1 : \operatorname{Re}\lambda < 0\}, \quad \Lambda^c := \{\lambda \in \Lambda_1 : \operatorname{Re}\lambda = 0\}, \quad \Lambda^u := \{\lambda \in \Lambda_1 : \operatorname{Re}\lambda > 0\}.$$

The superscripts stand for *stable*, *center*, and *unstable*, respectively. We define the stable subspace, the center subspace, and the unstable subspace as

$$E^s := \bigoplus_{\lambda\in\Lambda^s} E^{\mathbb{R}}_\lambda, \qquad E^c := \bigoplus_{\lambda\in\Lambda^c} E^{\mathbb{R}}_\lambda, \qquad E^u := \bigoplus_{\lambda\in\Lambda^u} E^{\mathbb{R}}_\lambda. \tag{2.16}$$

Since $\Lambda_1 = \Lambda^s \sqcup \Lambda^c \sqcup \Lambda^u$, it follows from (2.15) that

$$\mathbb{R}^d = E^s \oplus E^c \oplus E^u. \tag{2.17}$$

Theorem 2.22. *Given $A \in \mathbb{R}^{d\times d}$, there exist $m, M, a, c > 0$ such that the stable and unstable subspaces E^s and E^u from (2.16) have the following property.*

1. *For every $x_0 \in E^s$ and every $t \in \mathbb{R}$, we have $me^{-ct}\|x_0\| \leq \|e^{tA}x_0\| \leq Me^{-at}\|x_0\|$.*
2. *For every $x_0 \in E^u$ and every $t \in \mathbb{R}$, we have $me^{at}\|x_0\| \leq \|e^{tA}x_0\| \leq Me^{ct}\|x_0\|$.*

Proof. It suffices to prove the first claim: the second then follows by replacing A with $-A$ and t with $-t$. To this end, consider the positive quantities

$$r := \min\{|\lambda| : \lambda \in \Lambda^s\} \quad \text{and} \quad R := \max\{|\lambda| : \lambda \in \Lambda^s\}.$$

Fix $a \in (0, r)$ and $c \in (R, \infty)$, and let $\delta > 0$ be sufficiently small that $a + \delta < r$ and $c - \delta > R$.

Now given any $x_0 \in E^s$, we can write $x_0 = \sum_{\lambda \in \Lambda^s} y_\lambda$, where $y_\lambda \in E^{\mathbb{R}}_\lambda$, and since $A|_{E^{\mathbb{R}}_\lambda}$ can be written as $\lambda \mathbb{I} + N_\lambda$, where N_λ is nilpotent, we have

$$e^{tA} y_\lambda = e^{\lambda t} e^{tN_\lambda} y_\lambda \quad \text{for all } t \in \mathbb{R}. \tag{2.18}$$

Since $N_\lambda^{d+1} = 0$, the entries of e^{tN_λ} are polynomials in t of degree at most d, and thus there exists $C_\lambda > 0$ such that

$$\|e^{tN_j}\| \le C_\lambda e^{\delta t} \text{ and } \|e^{-tN_j}\| \le C_\lambda e^{\delta t} \text{ for all } t \ge 0.$$

Combining these estimates with (2.18) proves the theorem. □

A similar argument to the one above shows that orbits in the center subspace can grow at most subexponentially fast.

It is worth pointing out that even though every orbit $x(t)$ in E^s approaches 0 exponentially fast, it does not necessarily do so monotonically: the function $t \mapsto \|x(t)\|$ does not need to be strictly decreasing. This can be observed by considering a 2×2 matrix that has complex eigenvalues with negative real part, leading to solutions that spiral inwards: if they spiral "along ellipses" rather than "along circles", then $\|x(t)\|$ can increase for some values of t.

To construct a concrete example, we can use $A = PJP^{-1}$, where $J = \left(\begin{smallmatrix} -a & -b \\ b & -a \end{smallmatrix}\right)$ with $a, b > 0$ (to produce the inward spiral), and $P = \left(\begin{smallmatrix} c & 0 \\ 0 & 1 \end{smallmatrix}\right)$ for some $c > 1$ (to stretch the solutions in one direction and destroy monotonicity), so that

$$\begin{aligned} e^{tA} = Pe^{tJ}P^{-1} &= \begin{pmatrix} c & 0 \\ 0 & 1 \end{pmatrix} e^{-at} \begin{pmatrix} \cos bt & -\sin bt \\ \sin bt & \cos bt \end{pmatrix} \begin{pmatrix} c^{-1} & 0 \\ 0 & 1 \end{pmatrix} \\ &= e^{-at} \begin{pmatrix} c & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} c^{-1}\cos bt & -\sin bt \\ c^{-1}\sin bt & \cos bt \end{pmatrix} = e^{-at} \begin{pmatrix} \cos bt & -c\sin bt \\ c^{-1}\sin bt & \cos bt \end{pmatrix}. \end{aligned}$$

With the initial condition $x(0) = \left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)$, this leads to the solution

$$x(t) = e^{tA}x(0) = e^{-at} \begin{pmatrix} -c\sin bt \\ \cos bt \end{pmatrix},$$

for which we have

$$\|x(t)\|^2 = e^{-2at}(c^2 \sin^2 bt + \cos^2 bt).$$

Differentiating gives

$$\begin{aligned}\frac{d}{dt}\|x(t)\|^2 &= -2ae^{-2at}(c^2\sin^2 bt + \cos^2 bt) + e^{-2at}(2bc^2 \sin bt\cos bt - 2b\cos bt\sin bt)\\ &= 2e^{-2at}\big(b(c^2-1)\cos bt\sin bt - ac^2\sin^2 bt - a\cos^2 bt\big).\end{aligned}$$

Dividing both sides by $\cos^2 bt$ and writing $m = \tan bt = \frac{\sin bt}{\cos bt}$, we obtain

$$\frac{1}{\cos^2 bt}\frac{d}{dt}\|x(t)\|^2 = 2e^{-2at}\big(b(c^2-1)m - ac^2m^2 - a\big),$$

so monotonicity of $\|x(t)\|^2$ (and hence of $\|x(t)\|$) is determined by the sign of the quadratic expression $-ac^2m^2 + b(c^2-1)m - a$. The following are equivalent:

- there exists an interval of m-values on which this quadratic expression is positive;
- this quadratic has two real roots;
- this quadratic has positive discriminant $b^2(c^2-1)^2 - 4a^2c^2$.

Thus to produce an example where $t \mapsto \|x(t)\|$ fails to be monotonically decreasing, it suffices to choose values of a, b, c for which $b(c^2-1) > 2ac$. Observe that we have

$$\begin{aligned}A = PJP^{-1} &= \begin{pmatrix} c & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix} -a & -b\\ b & -a\end{pmatrix}\begin{pmatrix} c^{-1} & 0\\ 0 & 1\end{pmatrix}\\ &= \begin{pmatrix} c & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix} -a/c & -b\\ b/c & -a\end{pmatrix} = \begin{pmatrix} -a & -bc\\ b/c & -a\end{pmatrix},\end{aligned}$$

so by choosing $c = 2$, $b = 4$, and $a = 1$, we obtain the example $A = \left(\begin{smallmatrix} -1 & -8\\ 2 & -1\end{smallmatrix}\right)$, and since $b(c^2-1) = 4(4-1) = 12 > 2\cdot 1\cdot 2 = 2ac$, we have shown that $t \mapsto \|x(t)\|$ is not monotonically increasing for solutions of $\dot{x} = Ax$ in this case.

When the origin is a sink for $\dot{x} = Ax$, we might try to describe the manner in which $x(t) \to 0$ as $t \to \infty$ by replacing the norm with some other function that decreases along a solution. It turns out that we can always accomplish this using a positive definite quadratic form:

Theorem 2.23. *Suppose that $A \in \mathbb{R}^{d\times d}$ has the property that every eigenvalue has negative real part, so the origin is an attracting fixed point. Then there exists a symmetric, positive definite matrix $Q \in \mathbb{R}^{d\times d}$ such that $A^TQ + QA = -\mathbb{I}$, and consequently, the quadratic form $V\colon \mathbb{R}^d \to \mathbb{R}$ defined by $V(x) = x^TQx$ has the property that $\frac{d}{dt}V(x(t)) = -\|x\|^2$ along every solution $x(t)$ of $\dot{x} = Ax$.*

Proof. For each $t \in \mathbb{R}$, consider the matrix

$$B(t) = e^{tA^T}e^{tA}.$$

Observe that $B(0) = \mathbb{I}$ and each $B(t)$ is symmetric and positive definite. Moreover,

$$\dot{B}(t) = (A^T e^{tA^T})e^{tA} + e^{tA^T}(Ae^{tA}) = A^T B + BA,$$

from which we conclude that for every $\tau > 0$, we have

$$B(\tau) - \mathbb{I} = A^T\Big(\int_0^\tau B(t)\,dt\Big) + \Big(\int_0^\tau B(t)\,dt\Big)A. \tag{2.19}$$

Since 0 is a sink, the entries of $B(t)$ decay exponentially fast as $t \to \infty$; consequently $Q := \int_0^\infty B(t)\,dt$ exists, and sending $\tau \to \infty$ in (2.19) gives $-\mathbb{I} = A^T Q + QA$, as desired. □

2.7 Higher-order DEs

Lec 18
W 3/4

Higher-order DEs can be transformed into first-order DEs by adding variables. In particular, higher-order linear DEs in a single variable can be solved using the methods of this chapter. For example, consider the DE

$$\ddot{x} + a\dot{x} + bx = 0, \quad \text{where } a, b \in \mathbb{R}.$$

Let $y_1 = x$ and $y_2 = \dot{x}$, then we see that writing $y = \left(\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right)$, we have

$$\dot{y}_1 = y_2, \qquad \dot{y}_2 = -ay_2 - by_1 \qquad \Rightarrow \qquad \dot{y} = \begin{pmatrix} 0 & 1 \\ -b & -a \end{pmatrix} y.$$

The matrix $A = \left(\begin{smallmatrix} 0 & 1 \\ -b & -a \end{smallmatrix}\right)$ has characteristic polynomial $\lambda^2 + a\lambda + b$, so the roots of this polynomial are the eigenvalues of the matrix, and we are led to the following trichotomy.

- If the polynomial has real distinct roots λ_1 and λ_2, then $A = P\left(\begin{smallmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{smallmatrix}\right)P^{-1}$ and thus $y(t) = P\left(\begin{smallmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{smallmatrix}\right)P^{-1}y(0)$, so all solutions have the form $x(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$ for some $c_1, c_2 \in \mathbb{R}$.
- If the polynomial has distinct complex roots $\lambda = \alpha \pm i\beta$, then $A = P\left(\begin{smallmatrix} \alpha & -\beta \\ \beta & \alpha \end{smallmatrix}\right)P^{-1}$, so $y(t) = Pe^{\alpha t}\left(\begin{smallmatrix} \cos\beta t & -\sin\beta t \\ \sin\beta t & \cos\beta t \end{smallmatrix}\right)P^{-1}y(0)$, and all solutions have the form $x(t) = c_1 e^{\alpha t}\cos\beta t + c_2 e^{\alpha t}\sin\beta t$.
- If the polynomial has a repeated real root $\lambda_1 = \lambda_2 = \lambda$, then

$$A = \begin{pmatrix} 0 & 1 \\ -\lambda^2 & 2\lambda \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \lambda & 1 \end{pmatrix}\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -\lambda & 1 \end{pmatrix},$$

  and thus $y(t) = \left(\begin{smallmatrix} 1 & 0 \\ \lambda & 1 \end{smallmatrix}\right)e^{t\lambda}\left(\begin{smallmatrix} 1 & t \\ 0 & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} 1 & 0 \\ -\lambda & 1 \end{smallmatrix}\right)y(0)$, so all solutions have the form $x(t) = c_1 e^{\lambda t} + c_2 t e^{\lambda t}$.

As an example, we see that for $\ddot{x} - 3\dot{x} + 2x = 0$, the associated polynomial factors as $\lambda^2 - 3\lambda + 2 = (\lambda - 2)(\lambda - 1)$, so every solution is of the form $x(t) = c_1 e^t + c_2 e^{2t}$. Given an initial condition, such as $x(0) = 1$ and $\dot{x}(0) = -1$, the coefficients c_i can be determined by solving a linear system of equations: in this case, $c_1 + c_2 = 1$ and $c_1 + 2c_2 = -1$, so $c_2 = -2$ and $c_1 = 3$, giving the solution $x(t) = 3e^t - 2e^{2t}$.

More generally, we can consider the DE

$$x^{(n)} + a_1 x^{(n-1)} + \cdots + a_{n-1}\dot{x} + a_n x = 0, \tag{2.20}$$

and write $y_j = x^{(j-1)}$ for $1 \le j \le n$ to get $\dot{y} = Ay$, where

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{pmatrix}$$

is the *companion matrix* of the DE. One can show by induction, or by direct examination of permutation matrices, that the characteristic polynomial of A is

$$\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n.$$

If this polynomial has n distinct roots $\lambda_1, \ldots, \lambda_n$, then one can show that A is diagonalized by the *Vandermonde matrix*

$$P = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \cdots & \lambda_n^{n-1} \end{pmatrix},$$

and that $\det P = \prod_{i<j}(\lambda_j - \lambda_i)$. If the polynomial has repeated roots, then we can use Jordan normal form, but we need to know the sizes of the Jordan blocks. For this, it suffices to observe that when λ is a root of the characteristic polynomial, we have $Av = \lambda v$ if and only if $v = (v_1, v_2, \ldots, v_n)^T$ satisfies $v_{k+1} = \lambda v_k$ for each $1 \le k < n$, so the eigenspace of λ is 1-dimensional, and is spanned by $(1, \lambda, \lambda^2, \ldots, \lambda^{n-1})^T$. Thus there is exactly one Jordan chain associated to each λ, and we conclude that the Jordan normal form contains exactly one Jordan block for each eigenvalue. Recalling the structure of the exponentials of such blocks, we have now proved the following.

Theorem 2.24. *Suppose $\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n = 0$ has r distinct roots $\lambda_1, \ldots, \lambda_r$. Write $\lambda_j = a_j + ib_j$, where $a_j, b_j \in \mathbb{R}$, and let $m_j \in \mathbb{N}$ denote the multiplicity of λ_j. Then every solution of (2.20) is a linear combination of the functions $t^k e^{a_j t}\cos b_j t$ and $t^k e^{a_j t}\sin b_j t$, where $1 \le j \le r$ and $0 \le k < m_j$.*

Chapter 3: Local nonlinear theory

3.1 Stability of equilibria

Lec 19
F 3/6

Now we turn our attention to more general systems that need not be linear. Let $U \subset \mathbb{R}^d$ be open, and let $F\colon U \to \mathbb{R}^d$ be C^1. A point $\bar{x} \in U$ is an *equilibrium point* (or *fixed point*) of the DE $\dot{x} = F(x)$ if $F(x) = 0$. (This is sometimes also called a *stationary point* of the flow, or a *singular point*, or a *zero* of the vector field.) Then the unique solution with initial condition $x(0) = \bar{x}$ is $x(t) = \bar{x}$ for all $t \in \mathbb{R}$.

Definition 3.1. The *linearization* of the vector field F at the fixed point $\bar{x}$ is the linear vector field given by the Jacobian matrix of partial derivatives

$$DF(\bar{x}) = \left(\frac{\partial F_i}{\partial x_j}(\bar{x}) \right)_{i,j} \in \mathbb{R}^{d \times d},$$

where $F_i = F_i(x_1, \dots, x_d)$ is the ith coordinate function of F. The fixed point $\bar{x}$ is

- a *sink* of F if all eigenvalues of $DF(\bar{x})$ have negative real parts;
- a *source* of F if all eigenvalues of $DF(\bar{x})$ have positive real parts; and
- a *saddle* (or a *hyperbolic fixed point*) if all eigenvalues of $DF(\bar{x})$ have nonzero real parts, some of which are negative and some of which are positive.

Note that the above classification does not apply to any fixed points at which $DF(\bar{x})$ has any purely imaginary eigenvalues. The assumption that $\operatorname{Re} \lambda \neq 0$ for all eigenvalues guarantees that all the solutions of the linearization have exponential growth and decay properties; as we will see, this exponential behavior is robust enough to survive passing to the nonlinear system (at least locally). We start with the following definitions.

Definition 3.2. Let $(f_t)_{t \in \mathbb{R}}$ be the flow associated to the vector field F. A fixed point $\bar{x}$ for F is *stable* (or *Lyapunov stable*) if for every $\epsilon > 0$, there exists $\delta > 0$ such that

$$\text{for every } x \in U \text{ such that } \|x - \bar{x}\| < \delta, \text{ we have } \|f_t(x) - \bar{x}\| < \epsilon \text{ for all } t \geq 0. \quad (3.1)$$

It is *asymptotically stable* if (3.1) holds and if $\delta > 0$ can be chosen such that

$$\text{for every } x \in U \text{ such that } \|x - \bar{x}\| < \delta, \text{ we have } f_t(x) \to \bar{x} \text{ as } t \to \infty. \quad (3.2)$$

It is *exponentially stable* if there exist $C, \alpha, \delta > 0$ such that

$$\text{for every } x \in U \text{ such that } \|x - \bar{x}\| < \delta, \text{ we have } \|f_t(x) - \bar{x}\| \leq Ce^{-\alpha t} \text{ for all } t \geq 0. \quad (3.3)$$

It is *unstable* if it is not stable.

Note that the words 'stable' and 'unstable' here are used in a different sense than they were in §2.6, where they referred to convergence to the fixed point in forward or backward time. Note also that (3.3) implies (3.2), but (3.2) on its own does not imply (3.1): consider points moving clockwise around a circle on which there is a single fixed point, which *every* orbit eventually approaches, but initial conditions that are a small clockwise displacement from the fixed point lead to orbits that move all the way around the circle before approaching it.

We have the following relationships.

$$\begin{array}{c} \text{exponentially stable} \\ \Downarrow \quad \not\Uparrow \\ \text{asymptotically stable} \\ \Downarrow \quad \not\Uparrow \\ \text{Lyapunov stable} \end{array}$$

The failure of the upwards implications is demonstrated by the following examples.

- A center for a linear system in the plane (both eigenvalues purely imaginary) is Lyapunov stable but not asymptotically stable.
- The origin is a fixed point for $\dot{x} = -x^3$, which can be solved by integrating $-x^{-3}\,dx = dt$ to get $\frac{1}{2}x^{-2} = t + c$, so $x(t) = 1/\sqrt{2(t+c)} \to 0$ as $t \to \infty$, but the convergence is subexponential, so the origin is asymptotically stable but not exponentially stable.

The following result uses an idea that appeared in Theorem 2.23 above.

Theorem 3.3. *Let $U \subset \mathbb{R}^d$ and $F\colon U \to \mathbb{R}^d$ be C^1. Suppose that $\bar{x} \in U$ is a sink, so all eigenvalues of $DF(\bar{x})$ have negative real part. Then $\bar{x}$ is exponentially stable.*

Proof. Fix $\delta > 0$ sufficiently small that $B(\bar{x}, \delta) \subset U$. To each $x \in B(\bar{x}, \delta)$, associate the displacement $y := x - \bar{x}$. We will work in the coordinates given by y, in which we have

$$\dot{y} = \frac{d}{dt}(x - \bar{x}) = \dot{x} = F(x) = F(\bar{x} + y) =: G(y).$$

Observe that $G\colon B(0,\delta) \to \mathbb{R}^d$ is C^1, and let $A := DG(0) = DF(\bar{x}) \in \mathbb{R}^{d\times d}$, so that since $G(0) = 0$, we have

$$G(y) = Ay + R(y), \qquad \text{where } \lim_{y\to 0} \frac{\|R(y)\|}{\|y\|} = 0. \tag{3.4}$$

As in the proof of Theorem 2.23, there exists a positive definite symmetric matrix $Q \in \mathbb{R}^{d\times d}$ such that $A^TQ + QA = -\mathbb{I}$. Proceeding as we did there, define $V\colon \mathbb{R}^d \to \mathbb{R}$

by $V(y) = y^T Q y$. Observe that $\sqrt{V}$ is a norm on $\mathbb{R}^d$, so since all norms on a finite-dimensional vector space are equivalent, there exists $C > 0$ such that

$$C^{-1}\|y\|^2 \le V(y) \le C\|y\|^2 \quad \text{for all } y \in \mathbb{R}^d. \tag{3.5}$$

By (3.4), there exists $\delta > 0$ such that for all $y \in B(0,\delta)$, we have $\|R(y)\| \le \frac{1}{3\|Q\|}\|y\|$, and thus along any solution $y(t)$ of $\dot y = G(y)$, we have the following estimates at any point lying in $B(0,\delta)$:

$$\begin{aligned}\frac{d}{dt}V(y(t)) &= \frac{d}{dt}(y^TQy) = \dot y^TQy + y^TQ\dot y = G(y)^TQy + y^TQG(y)\\ &= (Ay + R(y))^TQy + y^TQ(Ay + R(y))\\ &= -\|y\|^2 + \big(R(y)^TQy + y^TQR(y)\big) \le -\|y\|^2 + \frac{2}{3}\|y\|^2 = -\frac{1}{3}\|y\|^2.\end{aligned}$$

By (3.5), this gives $\frac{d}{dt}V(y(t)) \le -\frac{1}{3C}V(y(t))$, and thus if the solution y exists and remains in $B(0,\delta)$ on the interval $[0,t]$, we conclude that

$$V(y(t)) \le V(y(0))e^{-\alpha t}, \quad \text{where } \alpha = 1/(3C). \tag{3.6}$$

In order to prove the theorem, we must argue that the solution exists on $[0,\infty)$ when $y(0)$ is sufficiently close to 0. We can do this using Theorem 1.20: by (3.5), the set $K := \{z \in \mathbb{R}^d : V(z) \le \delta/(2C)\}$ is a compact subset of $B(0,\delta)$, and by (3.6), any solution that begins in (3.6) never abandons this set, and must thus be defined on $[0,\infty)$ by Theorem 1.20. If $\|y(0)\| \le \delta/(2C^2)$, then $V(y(0)) \le \delta/(2C)$, so by the above discussion, $y(t)$ remains in $K \subset B(0,\delta)$ for all $t \in [0,\infty)$, and thus (3.6) proves exponential stability. □

Lec 20
W 3/11

The ideas in the preceding argument are worth formalizing; they will be helpful in a broader range of circumstances, and can also provide information about how large a set of trajectories converge to the fixed point. We begin with the following definition.

Definition 3.4. Given an open set $U \subset \mathbb{R}^d$ and a continuous vector field $F\colon U \to \mathbb{R}^d$, a *Lyapunov function* for F on U is a continuous function $V\colon U \to [0,\infty)$ such that

- there exists a unique point $\bar x \in U$ such that $V(\bar x) = 0$;
- V is C^1 on $U \setminus \{\bar x\}$; and
- $\dot V(x) := \langle \nabla V(x), F(x)\rangle$ satisfies $\dot V(x) \le 0$ for all $x \in U \setminus \{\bar x\}$.

If V also satisfies $\dot V(x) < 0$ for all $x \in U \setminus \{\bar x\}$, then it is a *strict Lyapunov function*.

The notation $\dot V(x)$ is justified by observing that if $x(t)$ is a solution of $\dot x = F(x)$, then

$$\frac{d}{dt}V(x(t)) = \langle \nabla V(x(t)), \dot x(t)\rangle = \langle \nabla V(x(t)), F(x(t))\rangle = \dot V(x(t)) \quad \text{for all } t. \tag{3.7}$$

In particular, if V is a Lyapunov function, then the function $t \mapsto V(x(t))$ is nonincreasing along any orbit in U. From this we immediately deduce that the point $\bar{x}$ at which V achieves its unique minimum must be a fixed point. In fact, it must be stable:

Theorem 3.5 (Lyapunov stability theorem)**.** *Let $U \subset \mathbb{R}^d$ be open and $F\colon U \to \mathbb{R}^d$ a C^1 vector field. Suppose that F admits a Lyapunov function V on some open set $W \subset U$, with a minimum at $\bar{x} \in W$. Then $\bar{x}$ is a Lyapunov stable fixed point for the flow generated by F.*

Proof. Since W is open, there exists $\delta > 0$ such that $B(\bar{x}, \delta) \subset W$. The Lyapunov function V is positive on the boundary of this ball, which is compact, so

$$\alpha := \inf\{V(x) : x \in \partial B(\bar{x}, \delta)\} > 0.$$

Since V is continuous, the set $U_1 := \{x \in B(\bar{x}, \delta) : V(x) < \alpha\}$ is open. Fix $x_0 \in U_1$ and let $I = [0, T^+) \subset [0, \infty)$ be the maximal (forward) interval of existence for the unique solution $x\colon I \to B(\bar{x}, \delta)$ of the IVP given by $\dot{x} = F(x)$ and $x(0) = x_0$. By the discussion following (3.7), we have

$$V(x(t)) \le V(x_0) \quad \text{for all } t \in [0, T^+),$$

so the orbit remains in the compact set $\{z \in U_1 : V(z) \le V(x_0)\} \subset U_1$ for as long as it is defined. In particular, it does not abandon this compact set, so by Theorem 1.20, we have $T^+ = \infty$. It follows that $x(t)$ is defined and contained in $U_1 \subset B(\bar{x}, \delta)$ for all $t \ge 0$, which verifies Lyapunov stability. □

A set $P \subset \mathbb{R}^d$ is *positively invariant*, or *forward invariant*, if for every $x \in P$, the point $f_t(x)$ is defined and contained in P for all $t \ge 0$. The argument in the proof of Theorem 3.5 actually shows the following, which we record here for future use.

Lemma 3.6. *Let $U \subset \mathbb{R}^d$ be open and $F\colon U \to \mathbb{R}^d$ a C^1 vector field that admits a Lyapunov function V on U. Let $\alpha := \inf\{V(x) : x \in \partial W\}$. Then for every $\beta \in (0, \alpha]$, the set $X_\beta := \{x \in W : V(x) < \beta\}$ is forward invariant.*

Lec 21
F 3/13

For *strict* Lyapunov functions, we can go beyond Lyapunov stability and deduce asymptotic stability. We will prove this via a more general result, whose statement and proof use the following notions.

Definition 3.7. Given a flow $\{f_t\}_{t\in\mathbb{R}}$ and a point $x \in \mathbb{R}^d$ such that $f_t(x)$ is defined for all $t \ge 0$, the *ω-limit set* of x is the set

$$\omega(x) := \{z \in \mathbb{R}^d : \text{there exists } t_n \to \infty \text{ such that } f_{t_n}(x) \to z\}.$$

The *basin of attraction* of a fixed point $\bar{x} \in \mathbb{R}^d$ is

$$\mathcal{B}(\bar{x}) := \big\{x \in \mathbb{R}^d : \omega(x) = \{\bar{x}\}\big\} = \{x \in \mathbb{R}^d : f_t(x) \to \bar{x} \text{ as } t \to \infty\}.$$

A set $Z \subset \mathbb{R}^d$ is *invariant* (or *totally invariant*) if $f_t(z)$ exists and is contained in Z for every $z \in Z$ and $t \in \mathbb{R}$.

Theorem 3.8. *Consider a C^1 vector field $F\colon U \to \mathbb{R}^d$, and suppose that there exists a Lyapunov function $V\colon U \to \mathbb{R}^d$ with a minimum at $\bar{x} \in U$. Suppose that $P \subset \bar{P} \subset U$ is a forward invariant open set containing $\bar{x}$ with the property that there is no (totally) invariant set in $\bar{P} \setminus \{\bar{x}\}$ on which V is constant. Then $\bar{x}$ is asymptotically stable, and $\bar{P} \subset \mathcal{B}(\bar{x})$.*

Proof. It suffices to observe that given $x \in \bar{P}$, the ω-limit set $\omega(x)$ is an invariant subset of $\bar{P}$, which must therefore be equal to $\{\bar{x}\}$. □

Corollary 3.9. *If $V\colon U \to \mathbb{R}$ is a* strict *Lyapunov function for a C^1 vector field F, then the minimum point $\bar{x} \in U$ for V is an asymptotically stable fixed point for F. If in addition we have $U = \mathbb{R}^d$, then $\mathcal{B}(\bar{x}) = \mathbb{R}^d$.*

Proof. Since V is a strict Lyapunov function, it is strictly decreasing along every orbit in $U \setminus \{x\}$. Thus $U \subset \{x\}$ does not contain any totally invariant sets on which V is constant. Moreover, by Lemma 3.6, the sublevel set $X_\beta \ni \bar{x}$ is forward invariant for all sufficiently small $\beta > 0$. Asymptotic stability of $\bar{x}$ follows from Theorem 3.8, as does the claim about $\mathcal{B}(\bar{x})$, since when $U = \mathbb{R}^d$ we can apply Theorem 3.8 with $P = \mathbb{R}^d$. □

The example $\dot{x} = -x^3$ already showed that asymptotic stability can hold even when $DF(\bar{x})$ has no eigenvalues with negative real part. Using Theorem 3.8 and Corollary 3.9, we can describe similar examples in higher dimensions.

Example 3.10. Consider the system given by $\dot{x} = -y - x^3$ and $\dot{y} = x - y^3$. This has a fixed point at 0, and $DF(0) = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ has eigenvalues $\lambda = \pm i$, so the linear part gives us no information about stability. Consider the function $V\colon \mathbb{R}^2 \to \mathbb{R}$ given by $V(x, y) = x^2 + y^2$. Then we have

$$\begin{aligned} \dot{V}(x, y) &= \langle \nabla V(x, y), F(x, y) \rangle = \langle (2x, 2y), (-y - x^3, x - y^3) \rangle \\ &= -2xy - 2x^4 + 2yx - 2y^4 = -2(x^4 + y^4), \end{aligned}$$

from which we see that V is a strict Lyapunov function. By Corollary 3.9, 0 is an asymptotically stable fixed point with $\mathcal{B}(0) = \mathbb{R}^2$.

Lec 22
M 3/23

Can this sort of example be extended any further? Could we have an eigenvalue with *positive* real part and yet still have a stable fixed point? In the next section, we will see that this is impossible, and will prove the following:

Theorem 3.11. *Let $\bar{x}$ be a fixed point for a C^1 vector field $F\colon U \to \mathbb{R}^d$, and suppose that $DF(\bar{x})$ has an eigenvalue λ with $\operatorname{Re}\lambda > 0$. Then $\bar{x}$ is not Lyapunov stable.*

3.2 Contraction in the Grassmannian

The proof of Theorem 3.11 will introduce an important idea: even when the flow associated to a DE has expanding directions, it can induce a contraction on a suitable

'auxiliary' space. As a simple example, consider the DE $\dot{x} = Ax$ with $A = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$. The origin is a saddle, with $E^u = \text{span}\{\left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)\}$ and $E^s = \text{span}\{\left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)\}$. Given $m \in \mathbb{R}$, let $X(m) := \{(r, mr) : r \in \mathbb{R}\}$ be the line through the origin with slope m, so $E^u = X(0)$. Observe that if $x_0 = (r, mr)^T \in X(0)$, then for each $t \in \mathbb{R}$, we have

$$x(t) = \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix} \begin{pmatrix} r \\ mr \end{pmatrix} = \begin{pmatrix} e^t r \\ e^{-t} mr \end{pmatrix} = \begin{pmatrix} \bar{r} \\ e^{-2t} m\bar{r} \end{pmatrix},$$

where $\bar{r} = e^t r$, so $x(t) \in X(e^{-2t}m)$. In particular, if we fix $\gamma > 0$ and let

$$K^u_\gamma := \bigcup_{|m| \le \gamma} X(m) = \{(r, mr) \in \mathbb{R}^2 : r \in \mathbb{R}, |m| \le \gamma\}$$

be the "cone of width γ" around E^u, then we see that every trajectory not contained in E^s must eventually enter K^u_γ and remain there. This gives a sense in which the unstable subspace is asymptotically stable, from a 'projective' point of view.[16]

Remark 3.12. Another way of interpreting this is that $|m|$ is behaving like a Lyapunov function, except that it vanishes on the *subspace* E^u instead of a single point. The behavior seen here remains true for any $A = \left(\begin{smallmatrix} a & 0 \\ 0 & b \end{smallmatrix}\right)$ with $a > b$, even if a and b are both positive or both negative: following the above computations, one sees that if $x_0 \in X(m)$, then $x(t) \in X(e^{-(a-b)t}m)$. This can be seen in the phase portraits we studied earlier for planar systems in which the origin is an attracting or repelling node.

Now we consider the higher-dimensional linear case, setting up the framework that we will then use to prove Theorem 3.11. Given $L \in \mathbb{R}^{d\times d}$, let $E^s, E^c, E^u \subset \mathbb{R}^d$ be the stable, center, and unstable subspaces for L, and for convenience, write $E^{cs} := E^c \oplus E^s$. Then we have $\mathbb{R}^d = E^u \oplus E^{cs}$, where both subspaces are L-invariant.

Suppose E^u and E^{cs} are non-trivial, and consider the linear transformations $A := L|_{E^u}$ and $B := L|_{E^{cs}}$. Similarly to what we saw above, we will now show that solutions of the linear system $\dot{x} = Lx$ have the property that

(I) as t increases, the angle between $x(t)$ and E^u decreases, and

(II) when this angle is small enough, $x(t)$ moves away from the origin.

To make this more precise, we use the following.

Exercise 3.13. Given a vector space V and a linear transformation $T \in L(V)$, use Jordan normal form to prove that for every $\epsilon > 0$, there exists a basis for V in which the matrix of T takes the form $D + N$, where $DN = ND$, and

- D is block diagonal with all blocks of the form $\left(\begin{smallmatrix} a \end{smallmatrix}\right)$ or $\left(\begin{smallmatrix} a & -b \\ b & a \end{smallmatrix}\right)$ for some $a, b \in \mathbb{R}$;

[16] The word *projective* appears here because the set of all lines through the origin in $\mathbb{R}^d$ is called the *projective space* of dimension $d - 1$.

- N is nilpotent and satisfies $|N_{ij}| < \epsilon$ for all i, j.

Exercise 3.14. With L, E^u, E^{cs}, A, and B as above, let $a > b > 0$ and $c > 0$ be such that

$$a < \min\{\operatorname{Re}\lambda : \lambda \in \sigma(A)\} \quad \text{and} \quad c < \min\{\operatorname{Re}\lambda : \lambda \in \sigma(B)\},$$

where $\sigma(A)$ and $\sigma(B)$ denote the sets of all eigenvalues of A and B, respectively. Apply Exercise 3.13 to both $A \in L(E^u)$ and $B \in E^{cs}$ to obtain a basis for $\mathbb{R}^d$ with the property that if $\langle \cdot, \cdot \rangle_*$ is the inner product in which this basis becomes orthonormal, and $\| \cdot \|_*$ is the corresponding norm, then we have

$$\begin{aligned} \langle Ax, x\rangle &\geq a\|x\|_*^2 \text{ for all } x \in E^u, \\ -c\|y\|_*^2 \leq \langle By, y\rangle &\leq b\|y\|_*^2 \text{ for all } y \in E^{cs}. \end{aligned} \tag{3.8}$$

Lec 23
W 3/25

Now fix a small parameter $\gamma > 0$ and consider the *unstable cone*

$$K_\gamma^u := \{x + y : x \in E^u,\ y \in E^{cs},\ \|y\|_* \leq \gamma\|x\|_*\} \subset \mathbb{R}^d.$$

Define a function $G\colon \mathbb{R}^d \to \mathbb{R}$ by

$$G(z) := \gamma^2\|x\|_*^2 - \|y\|_*^2, \quad \text{where } z = x + y,\ x \in E^u,\ y \in E^{cs}.$$

Then $K_\gamma^u = \{z \in \mathbb{R}^d : G(z) \geq 0\}$. We can now formalize properties (I) and (II):

Proposition 3.15. *For every $\gamma > 0$, the function G and the cone K_γ^u have the property that if $z(\cdot)$ is a solution of $\dot z = DF(0)z$ and $t \in \mathbb{R}$ is such that $z(t) \in K_\gamma^u \setminus \{0\}$, then we have*

$$\frac{d}{dt}G(z(t)) \geq \frac{2\gamma^2(a-b)}{1+\gamma^2}\|z(t)\|_*^2 \quad \text{and} \quad \frac{d}{dt}\|z\|_*^2 \geq 2\Big(\frac{a - c\gamma}{1+\gamma^2}\Big)\|z\|_*^2. \tag{3.9}$$

Proof. Writing $z(t) = x(t) + y(t)$, where $x(t) \in E^u$ and $y(t) \in E^{cs}$, we have

$$\dot x(t) + \dot y(t) = \dot z(t) = DF(0)(x(t) + y(t)) = Ax(t) + By(t),$$

from which invariance of E^u and E^{cs} implies that $\dot x = Ax \in E^u$ and $\dot y = By \in E^{cs}$. Moreover, we have

$$\nabla G(z) = \gamma^2\nabla\langle x, x\rangle_* - \nabla\langle y, y\rangle_* = 2(\gamma^2 x - y). \tag{3.10}$$

Recalling (3.8), we obtain

$$\dot G(z(t)) = \langle \nabla G(z), \dot x + \dot y\rangle_* = 2\big(\gamma^2\langle x, Ax\rangle_* - \langle y, By\rangle_*\big) \geq 2\big(\gamma^2 a\|x\|_*^2 - b\|y\|_*^2\big).$$

The definition of K_γ^u gives $\|y\|_*^2 \leq \gamma^2\|x\|_*^2$ and $\|z\|_*^2 = \|x\|_*^2 + \|y\|_*^2 \leq (1+\gamma^2)\|x\|_*^2$, so

$$\dot G(z(t)) \geq 2\gamma^2(a-b)\|x\|_*^2 \geq \frac{2\gamma^2(a-b)}{1+\gamma^2}\|z\|_*^2,$$

which proves the first half of (3.9). To prove the second half, observe that

$$\begin{aligned}\frac{d}{dt}\frac{1}{2}\|z\|_*^2 &= \langle z, DF(0)z\rangle_* = \langle x, Ax\rangle_* + \langle y, By\rangle_* \\ &\geq a\|x\|_*^2 - c\|y\|_*^2 \geq (a-c\gamma)\|x\|_*^2 \geq \frac{a-c\gamma}{1+\gamma^2}\|z\|_*^2.\end{aligned}$$

□

Proposition 3.15 shows that K_γ^u is forward invariant under the linear flow, and that if $\gamma > 0$ is small enough that $a - c\gamma > 0$, then $t \mapsto \|z(t)\|_*$ is increasing whenever $z(t) \in K_\gamma^u$. These properties are robust enough to pass to the nonlinear system, and we can now proceed with the proof of Theorem 3.11.

Proof of Theorem 3.11. Without loss of generality, suppose that the fixed point is $\bar{x} = 0$, and let $L := DF(0) \in \mathbb{R}^{d\times d}$. Then for $z \approx 0$, we have $F(z) = Lz + R(z)$, where $R(z) = o(\|z\|_*)$. In particular, defining $r\colon [0,\infty) \to [0,\infty)$ by

$$r(s) := \sup\left\{\frac{\|R(z)\|_*}{\|z\|_*} : z \in \mathbb{R}^d, \|z\|_* \leq s\right\},$$

we have $r(s) \to 0$ as $s \to 0$. Let $\gamma \in (0,1]$ be sufficiently small that $a - c\gamma > 0$, and let $s > 0$ be sufficiently small that

$$\gamma^2(a-b) - 3r(s) > 0 \quad \text{and} \quad a - c\gamma - 2r(s) > 0. \tag{3.11}$$

Let G, K_γ^u be as in the discussion preceding Proposition 3.15. Consider the set

$$X := \{z \in K_\gamma^u : 0 < \|z\|_* \leq s\}.$$

We will prove the following versions of properties (I) and (II):

- any solution of $\dot{z} = F(z)$ beginning in X can only leave X through its "outer shell", where $\|z\|_* = s$; and
- every solution in K_γ^u satisfies $\frac{d}{dt}\|z\|_*^2 \geq \xi\|z\|_*^2$, where $\xi := a - c\gamma - 2r(s) > 0$.

These will suffice to prove the theorem, since we can choose an initial condition $z(0)$ arbitrarily close to the origin, and the two properties together imply that $\|z(t)\|_*$ increases exponentially fast until it reaches s, implying that the fixed point is not Lyapunov stable.

To prove these properties, first note that since $\gamma \leq 1$, we have $1 + \gamma^2 \leq 2$, which will simplify the bounds in (3.9). Using the first of these bounds and writing $z = x + y$ as in the proof of Proposition 3.15, we get

$$\dot{G}(z(t)) = \langle \nabla G(z), Ax + By + R(z)\rangle_* \geq \gamma^2(a-b)\|z\|_*^2 + \langle \nabla G(z), R(z)\rangle_*.$$

Observe that $\|\nabla G(z)\|_* = \|2(\gamma^2 x - y)\|_* \le 3\|z\|_*$, so

$$|\langle \nabla G(z), R(z)\rangle_*| \le 3\|z\|_* \cdot r(s)\|z\|_* = 3r(s)\|z\|_*^2,$$

and we conclude that

$$\dot G(z(t)) \ge \big(\gamma^2(a-b) - 3r(s)\big)\|z\|_*^2.$$

By (3.11), this is positive, so a trajectory that is currently in $X \subset K_\gamma^u$ must remain in X as long as $\|z\|_* < s$. Finally, we have

$$\begin{aligned}\frac{d}{dt}\|z\|_*^2 &= 2\langle z, Lz + R(z)\rangle_* = 2\big(\langle z, Lz\rangle_* + \langle z, R(z)\rangle_*\big)\\ &\ge (a - c\gamma)\|z\|_*^2 - 2r(s)\|z\|_*^2 = \big(a - c\gamma - 2r(s)\big)\|z\|_*^2,\end{aligned}$$

which proves the second desired property and completes the proof. □

3.3 Perturbations, robustness, and bifurcations

Lec 24
F 3/27

The previous section treated nonlinear DEs as perturbations of linear ones; that is, we used the fact that $F(x)$ is "close to" $Lx = DF(0)x$ when $\|x\|$ is small. When studying perturbations of systems, it is important to be precise about the meaning of "close to". Let us illustrate this by considering two possible interpretations of these words for two vector fields $F, G\colon U \to \mathbb{R}^d$, where $U \subset \mathbb{R}^d$ is open. (In our setting of studying fixed points, we would naturally take U to be a small neighborhood of the fixed point.)

- Say that F and G are ϵ-close in the C^0-sense if $\|F(x) - G(x)\| < \epsilon$ for all $x \in U$.
- Say that F and G are ϵ-close in the C^1-sense if for all $x \in U$, we have both $\|F(x) - G(x)\| < \epsilon$ and $\|DF(x) - DG(x)\| < \epsilon$.

More informally, one can use the terms *C^0-close* and *C^1-close*, without making ϵ precise. By considering higher-order derivatives, one can also define *C^r-close* for all $r \in \mathbb{N}$.

Example 3.16. With $d = 1$ and $U = (-1, 1) \subset \mathbb{R}$, consider the vector fields

$$F(x) = x, \qquad G_0(x) = x - \epsilon\sin(100x), \qquad G_1(x) = (1-\epsilon)x.$$

Then G_0 is ϵ-close to F in the C^0-sense, but not in the C^1-sense, while G_1 is ϵ-close to F in both the C^0- and C^1-sense.

In §3.1 and §3.2, it was very important that near the fixed point, F was close to its linearization *in the C^1-sense*, not just in the C^0-sense. Indeed, in Example 3.16, we see that F and G_1 both have repelling fixed points at 0, and no other fixed points in $(-1, 1)$, so that a C^1-small perturbation preserves the qualitative description, while a C^0-small perturbation can destroy it: G_0 has an *attracting* fixed point at 0, and many other fixed points in $(-1, 1)$.

Let us reframe this example in a slightly more general way, to illustrate the topological concepts at work here. Let $U \subset \mathbb{R}$ be an open interval and let $F\colon U \to \mathbb{R}$ be C^1. Suppose that $F(\bar{x}) = 0$ for some $\bar{x} \in U$, and that $F'(\bar{x}) \neq 0$. Then one of the following two cases occurs.

- *Repelling fixed point*: $F'(\bar{x}) > 0$, so there exist $a, b \in U$ such that $a < \bar{x} < b$ and $F|_{[a,\bar{x})} < 0$, while $F|_{(\bar{x},b]} > 0$, causing trajectories in $[a, b]$ to flow away from $\bar{x}$.
- *Attracting fixed point*: $F'(\bar{x}) < 0$, so there exist $a, b \in U$ such that $a < \bar{x} < b$ and $F|_{[a,\bar{x})} > 0$, while $F|_{(\bar{x},b]} > 0$, causing trajectories in $[a, b]$ to flow towards $\bar{x}$.

In each of these cases, a C^0-small perturbation G will have the property that $G(a)$ and $G(b)$ take different signs, and thus by the intermediate value theorem, G will still have a fixed point in $[a, b]$. If the perturbation is C^0-small but C^1-large, then it is possible to change the stability of this fixed point and to create new fixed points, as with G_0 from Example 3.16. If the perturbation is C^1-small, however, then since $F'(\bar{x}) \neq 0$, we also have $G'(x) \neq 0$ for $x \approx \bar{x}$, and thus G has the same monotonicity property as F, so it has a unique fixed point near $\bar{x}$, and the stability is unchanged.

Example 3.17. Consider the *logistic DE* $\dot{x} = x(1 - x)$. When $x \in [0, 1]$, this models population growth in an environment with limited resources, with x representing the current population size as a fraction of the maximum population that can be supported by the environment. There are fixed points at 0 and 1, and $F(x) = x(1 - x)$ has $F'(0) > 0$ and $F'(1) < 0$, so the fixed point at 0 is repelling, while the fixed point at 1 is attracting. This behavior is robust under C^1-perturbations: consequently, if we vary the parameters (perhaps by changing the resources available or changing the reproduction rate of the population) or introduce new factors into the system (such as harvesting with a constant rate), we will still see one repelling fixed point and one attracting fixed point, at least when the changes are small.

If F has a fixed point with $F'(\bar{x}) = 0$, then we have already seen that various behaviors are possible. The following exercise explores this:

Exercise 3.18. Each of the following vector fields on $\mathbb{R}$ has a fixed point at 0. In each case, determine the stability properties of the fixed point; then show that a C^1-small perturbation can change the stability properties of this fixed point, and can create new fixed points in an arbitrarily small neighborhood of 0. Determine in which cases a C^1-small perturbation can result in a vector field with no fixed points.

(a) $F(x) = x^2$.

(b) $F(x) = -x^2$.

(c) $F(x) = x^3$.

(d) $F(x) = -x^3$.

The qualitative behavior changes illustrated in Exercise 3.18 are often called *bifurcations*. This leads to a rich theory, which we will not explore just now. Instead, we proceed now to a (brief) discussion about what happens in higher dimensions, when the mechanisms for robustness in the above discussion become more subtle to state.

Lec 25
M 3/30

The mechanism described above for C^1-robustness was the fact that nonvanishing derivative implies monotonicity. In higher dimensions, we can no longer speak of $F\colon U \to \mathbb{R}^d$ being monotonic, but it does still make sense to say that F is a *local diffeomorphism*: there exists a neighborhood V of $\bar{x}$ such that $F|_V\colon V \to F(V)$ is an invertible map such that $DF(x)$ is invertible for all $x \in V$. This fact remains true for C^1-small perturbations G, and then the Inverse Function Theorem shows that such a G has a unique fixed point near $\bar{x}$. To guarantee that this fixed point has the same stability properties as $\bar{x}$ does for F, we need to know not just that $DF(\bar{x})$ is invertible, but that it has no eigenvalues on the imaginary axis, so its center subspace is trivial: in this case, the linearization of the perturbation will also have a trivial center, and will have stable and unstable subspaces close to those of $DF(\bar{x})$, provided the pertubation is sufficiently close to F in the C^1-sense.

The mechanism for C^0-robustness (in the sense that any C^0-small perturbation has a fixed point near $\bar{x}$) was that F takes different signs on either side of $\bar{x}$. Once again, this does not immediately generalize to higher dimensions, but there is a more sophisticated idea that works. To keep things simple, we describe it briefly when $d = 2$ and $\bar{x} = 0$. Suppose that F is C^1 with $F(0) = 0$, and that $r > 0$ is sufficiently small that F does not vanish anywhere else in $\overline{B(0,r)}$. Given $\theta \in [0,1]$, let $x(\theta) = (r\cos\theta, r\sin\theta) \in \partial B(0,r)$, and define a function $\alpha\colon [0,1] \to S^1 := \{v \in \mathbb{R}^2 : \|v\| = 1\}$ by

$$\alpha(\theta) = \frac{F(x(\theta))}{\|F(x(\theta))\|}.$$

As θ increases from 0 to 1, the vector $\alpha(\theta)$ can move in either direction around the circle. Observing that $\alpha(1) = \alpha(0)$, we see that this vector must make some integer number of rotations around the circle S^1, with positive integers corresponding to counterclockwise rotations, and negative to clockwise. This integer is called the *degree* of the circle map[17] $\alpha\colon [0,1] \to S^1$, and is also called the *index* (or *Poincaré index*) of the vector field F at the fixed point 0.

Exercise 3.19. Show that F has index -1 at $\bar{x}$ whenever $DF(\bar{x})$ has one positive and one negative eigenvalue, and index $+1$ whenever $DF(\bar{x})$ has two non-real eigenvalues, or two real eigenvalues that have the same sign.

Lec 26
W 4/1

Making a C^0-small perturbation to F will result in a (uniformly) small perturbation to the circle map α, and thus will not change the degree of this map. As we already saw in the case $d = 1$, such a perturbation can create new fixed points, and can change the

[17] Technically, "circle map" should refer to a map from S^1 to S^1. Since $\alpha(1) = \alpha(0)$, the map α naturally induces such a map by identifying the two endpoints of $[0,1]$ to obtain a circle.

index of the fixed point at $\bar{x}$; however, it is possible to show that it does not change the *sum* of the indices of all the fixed points in $B(\bar{x}, r)$, and thus this sum remains equal to the degree of the circle map α. In particular, by Exercise 3.19, the degree of α is nonzero whenever $DF(\bar{x})$ is invertible, and thus $B(\bar{x}, r)$ must contain at least one fixed point for every C^0-small perturbation.

The ideas in the previous paragraph lead (eventually) to the *Poincaré–Hopf index theorem*, which we will not explore further here. We will also omit a proper discussion of how to define the index of a fixed point when $d \geq 3$: in this case the circle map α is replaced by a sphere map $S^{d-1} \to S^{d-1}$, whose degree can be defined either in terms of algebraic topology by considering the induced map on the top homology group, or in terms of differential topology by taking a sum over preimages of a regular value of the map, with each term being ± 1 according to whether the map is orientation-preserving or orientation-reversing at the corresponding preimage. Instead of diving into these deeper waters now, we conclude this section with an observation that when $d = 1$, we can treat the index of a fixed point $\bar{x}$ as $+1$ if F is negative on a left neighborhood of $\bar{x}$ and positive on a right neighborhood of $\bar{x}$, and as -1 if "negative" and "positive" are reversed in this criterion. Considering an interval $[a, b]$ such that $F(a) \neq 0$ and $F(b) \neq 0$, and writing Σ_F for the sum of the indices of F over all $\bar{x} \in [a, b]$ such that $F(\bar{x}) = 0$, we see that

$$\Sigma_F = \begin{cases} +1 & \text{if } F(a) < 0 < F(b), \\ 0 & \text{if } F(a) \text{ and } F(b) \text{ have the same sign,} \\ -1 & \text{if } F(a) > 0 > F(b). \end{cases}$$

3.4 The pendulum

Now we explore the ideas from the previous sections in the context of a specific nonlinear system. Consider a pendulum: a point mass attached to a rigid, massless rod, the other end of which is fixed at a pivot, around which it is free to rotate without friction. Let $\theta = \theta(t)$ denote the angular displacement from the downward-pointing vertical position at time t. Modulo parameters that depend on the specific physical setup, the pendulum satisfies the second-order linear DE $\ddot{\theta} + \sin\theta = 0$. As usual, we can cast this as a first-order system by writing $x = \theta$ and $y = \dot{\theta}$, so that

$$\begin{aligned} \dot{x} &= \dot{\theta} = y \\ \dot{y} &= \ddot{\theta} = -\sin\theta = -\sin x \end{aligned} \qquad \Rightarrow \qquad \frac{d}{dt}(x, y) = F(x, y) := (y, -\sin x).$$

Fixed points occur when $y = \sin x = 0$, which happens exactly when $(x, y) = (n\pi, 0)$ for some integer n. Since the physical system is invariant under the translation $(x, y) \mapsto (x + 2\pi, y)$, there are only two physical fixed points: one is when the pendulum hangs motionless straight down, corresponding to even values of n; the other is when the pendulum balances on top of the pivot, corresponding to odd values of n.

The linearization at $(n\pi, 0)$ is given by the matrix

$$DF(n\pi, 0) = \begin{pmatrix} 0 & 1 \\ -\cos n\pi & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ (-1)^{n+1} & 0 \end{pmatrix}.$$

We see that $DF(n\pi, 0)$ is equal to $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ when n is even, and $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ when n is odd. The first of these has eigenvalues $\pm i$, so the linearization has a center at the fixed point; the second of these has eigenvalues ± 1 with corresponding eigenvectors $(1, \pm 1)$, so the linearization has a saddle.

Theorem 3.11 implies that this second fixed point is unstable for the nonlinear system, which is consistent with our physical intuition; we expect the "vertically balancing" stationary configuration to be very precarious.

Regarding the fixed point at the origin, our results so far tell us very little, since the eigenvalues are on the imaginary axis. The linearization, which corresponds to the simple harmonic oscillator $\ddot{x} + x = 0$, has the property that every orbit is periodic, and the fixed point is Lyapunov stable but not asymptotically stable. However, this in and of itself does not tell us much about the nonlinear system:

Exercise 3.20. Show by example that if $F\colon \mathbb{R}^2 \to \mathbb{R}^2$ has $F(0) = 0$ and $DF(0) = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, then the origin can be any of the following for the DE $\dot{x} = F(x)$: (i) asymptotically stable, (ii) Lyapunov stable but not asymptotically stable, or (iii) not Lyapunov stable.

So our results so far do not provide a complete picture for the fixed points of the pendulum. In this specific example, a more precise description of the behavior can be given by observing that the total energy (kinetic plus potential) of the system is conserved. This total energy is $\frac{1}{2}\dot{\theta}^2 - \cos\theta = \frac{1}{2}y^2 - \cos x$, and indeed we see that writing $H(x, y) := \frac{1}{2}y^2 - \cos x$, we have

$$\nabla H = (\sin x, y) \quad \Rightarrow \quad \dot{H} = \langle \nabla H, F \rangle = \langle (\sin x, y), (y, -\sin x) \rangle = 0.$$

Thus every solution of the pendulum DE lies on a level set of H. We say that H is an *integral of motion* of the system.

Figure 3.1 shows some of the level sets of H. Observe that in a neighborhood of each of the fixed points $(\pm\pi, 0)$, the level set containing the fixed point is a union of two curves that intersect at the fixed point. One of these curves slopes up and is tangent to the unstable eigenvector $(1, 1)$ at the fixed point; the other slopes down and is tangent to the stable eigenvector $(1, -1)$. Observe that since $\dot{x} = y$, trajectories above the x-axis move to the right, while trajectories below the x-axis move to the left. Combining these observations, we see that the downward-sloping curve consists of two orbits that approach the fixed point; one orbit approaches from the northwest, the other from the southeast. These correspond to the motion of a pendulum that slows down as it approaches an upward-pointing configuration, never stopping and falling back, but also

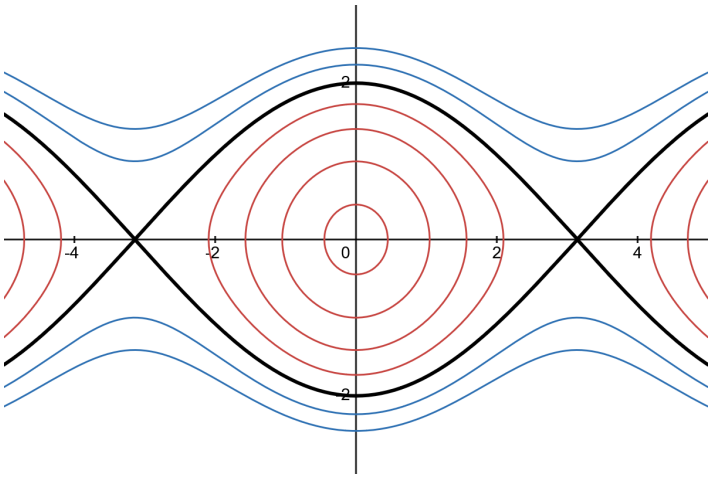

Figure 3.1: Level sets of the integral of motion H.

never quite reaching the vertical. Writing $z = (x, y)$ and $\bar{z}$ for the fixed point, such trajectories have the property that $z(t) \to \bar{z}$ as $t \to \infty$.

The set of trajectories that remain in a neighborhood of $\bar{z}$ and converge to $\bar{z}$ as $t \to \infty$ is called the *local stable manifold* of $\bar{z}$, and denoted $W^s_{\text{loc}}(\bar{z})$. Reversing time, the *local unstable manifold* $W^u_{\text{loc}}(\bar{z})$ is the set of points such that $z(t)$ remains near $\bar{z}$ for all $t \le 0$, and converges to $\bar{z}$ as $t \to -\infty$. The curves described in the previous paragraph are the local stable and unstable manifolds of the fixed points $(\pm\pi, 0)$. We will explore these more in the next section, and see that a similar picture appears for any C^1 vector field near a fixed point at which all eigenvalues of the linearization have nonzero real parts.

The fixed point for the pendulum that lies at the origin is a minimum of the function H, so the nearby level sets of H form closed curves around the origin, which are periodic orbits of the pendulum. This implies that the origin is Lyapunov stable but not asymptotically stable, just as for its linearization.

Lec 27
F 4/3

This last fact can also be deduced from another important feature of the pendulum, which we will prove below: the flow that it induces on $\mathbb{R}^2$ is *area-preserving*, meaning that if we write $f_t \colon \mathbb{R}^2 \to \mathbb{R}^2$ for the time-t map of the pendulum, then for any region $U \subset \mathbb{R}^2$ whose area is well-defined, and for any $t \in \mathbb{R}$, the sets U and $f_t(U)$ have the same area. More generally, a flow on $\mathbb{R}^d$ is volume-preserving if U and $f_t(U)$ always have the same d-dimensional volume.

Exercise 3.21. Show that if $F \colon \mathbb{R}^2 \to \mathbb{R}^2$ has an area-preserving flow with a fixed point at 0, and $DF(0) = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, then all nearby solutions of $\dot{x} = F(x)$ are periodic, and the fixed point is Lyapunov stable but not asymptotically stable.

The fact that the pendulum flow is area-preserving follows from the next theorem together with the formula $DF(x, y) = \left(\begin{smallmatrix} 0 & 1 \\ -\cos x & 0 \end{smallmatrix}\right)$.

Theorem 3.22. *Given a C^1 flow $(f_t)_t$ generated by a C^1 vector field F in $\mathbb{R}^d$, the following are equivalent:*

(a) the flow is volume-preserving;

(b) $\det Df_t(x) = 1$ for all x and t;

(c) $\operatorname{div} F(x) = 0$ for all x, where $\operatorname{div} F = \operatorname{Tr} DF = \sum_{j=1}^d \frac{\partial F_j}{\partial x_j}$ is the divergence.

The first two properties in Theorem 3.22 are equivalent by the change-of-coordinates formula from multivariable calculus. The fact that they are equivalent to the third is a consequence of the following more general result.

Proposition 3.23. *Let $(f_t)_t$ be a C^1 flow generated by a C^1 vector field F in $\mathbb{R}^d$, and let $J(t,x)$ be such that $J(0,x) = 1$ and $\frac{d}{dt}J(t,x) = \big(\operatorname{div} F(f_t x)\big)J(t,x)$ for all t, x. Then given any open set $U \subset \mathbb{R}^d$, we have*

$$\operatorname{Vol} f_t(U) = \int_U J(t,x)\,dx. \tag{3.12}$$

Before proving the proposition, observe the interpretation: the flow increases volume when $\operatorname{div} F > 0$, and decreases is when $\operatorname{div} F < 0$. Also note that for a linear vector field $F(x) = Ax$, where $A \in \mathbb{R}^{d\times d}$, we have $DF = A$ and thus $\operatorname{div} F = \operatorname{Tr} A$ is constant, so the DE for $J(t,x)$ becomes $\dot J = (\operatorname{Tr} A)J$, and we get $J = e^{t\operatorname{Tr} A}$. We could also have deduced this from facts about matrix exponentials, observing that for a linear system, we have $f_t(x) = e^{tA}x$, so $Df_t(x) = e^{tA}$, and thus

$$J(t,x) = \det Df_t(x) = \det e^{tA} = e^{\operatorname{Tr}(tA)}.$$

The key tool in proving Proposition 3.23 in the nonlinear case is the following lemma, which roughly speaking says that "trace is the derivative of determinant".

Lemma 3.24. *Consider the function $G\colon \mathbb{R}^{d\times d} \to \mathbb{R}$ defined by $G(X) = \det(X)$. Then given any $X \in \mathbb{R}^{d\times d}$, the directional derivative of G at the identity matrix $\mathbb{I}$ in the direction of X is equal to $\operatorname{Tr}(X)$:*

$$\operatorname{Tr}(X) = \frac{d}{dt}\det(\mathbb{I} + sX)\Big|_{s=0} = \lim_{s\to 0}\frac{\det(\mathbb{I}+sX) - \det(\mathbb{I})}{s}. \tag{3.13}$$

Proof. Writing $\det(\mathbb{I}+sX)$ as a sum over permutations, we see that this expression is a degree-d polynomial in the variable s, whose linear coefficient is the right-hand side of (3.13). Every non-identity permutation can only contribute terms of order 2 or higher in s, and thus the linear coefficient of $\det(\mathbb{I}+sX)$ is the same as the linear coefficient of $\prod_{j=1}^d(1+sX_{jj})$, which is $\operatorname{Tr}(X)$. □

Proof of Proposition 3.23. Fix $x \in \mathbb{R}^d$ and let $\Phi(t) = Df_t(x)$. Recalling (1.27) and Theorem 1.27, we have $\dot\Phi = A(t)\Phi$, where $A(t) = DF(f_t x)$. Since $J(t) = \det \Phi(t)$, we see from the chain rule that $\dot J(t)$ is equal to the directional derivative of $G = \det$ at $\Phi(t)$ in the direction of $\dot\Phi(t) = A(t)\Phi(t)$:

$$\dot J(t) = \frac{d}{ds} \det(\Phi(t) + sA(t)\Phi(t))\big|_{s=0} = \frac{d}{ds} \det(\mathbb{I} + sA(t))\big|_{s=0} \cdot \det \Phi(t) = \operatorname{Tr}(A(t)) \cdot J(t),$$

wehre the last equality uses Lemma 3.24. □

We conclude this section by observing that since the pendulum's vector field has the property that DF is invertible at each fixed point, the arguments from §3.3 show that any C^1-small perturbation of the pendulum will continue to have exactly one fixed point near each of these. Near $(\pm\pi, 0)$, the perturbed system will have a saddle, while nears the origin, there are multiple possibilities: it could be a center, an attracting focus, or a repelling focus, but it cannot be a saddle, as can be seen by looking at its eigenvalues or by the Poincaré index.

3.5 Hadamard–Perron theorem on (un)stable manifolds

Lec 28
M 4/6

Consider a C^1 vector field with a fixed point at $\bar x$, and suppose that some eigenvalues of $DF(\bar x)$ have negative real parts, while others have positive real parts. Then although the linearized system does not have a stable fixed point, it does still have a stable subspace consisting of trajectories that converge to the fixed point. We saw in the previous section that for the example of the pendulum, this has an analogue for the nonlinear system: there is a curve of initial conditions leading to solutions that converge to $\bar x$ as $t \to \infty$, which we called the *stable manifold*.

In that example, we got lucky: the existence of an integral of motion allowed us to produce local stable and unstable manifolds via level sets of a rather simple function. Does the picture persist if we do not have such a conserved quantity available to us?

To formulate the general theorem, we need some notation. Suppose we are given a decomposition $\mathbb{R}^d = E^u \oplus E^s$. Given $r > 0$, let

$$E^u_r := B(0, r) \cap E^u \subset \mathbb{R}^d \quad \text{and} \quad E^s_r := B(0, r) \cap E^s \subset \mathbb{R}^d.$$

Given $\bar x \in \mathbb{R}^d$ and $\phi\colon E^u_r \to E^s$, denote the *graph of ϕ through $\bar x$* by

$$\Gamma_{\bar x}(\phi) := \{\bar x + v + \phi(v) : v \in E^u_r\}.$$

We use similar notation when the roles of s and u are reversed: given $\phi\colon E^s_r \to E^u$, write

$$\Gamma_{\bar x}(\phi) := \{\bar x + v + \phi(v) : v \in E^s_r\}.$$

Theorem 3.25 (Hadamard–Perron Theorem)**.** *Let $U \subset \mathbb{R}^d$ be open, and let $F\colon U \to \mathbb{R}^d$ be C^1. Write $(f_t)_t$ for the flow generated by F. Suppose that $\bar{x} \in U$ satisfies $F(\bar{x}) = 0$, and that $DF(\bar{x})$ has no purely imaginary eigenvalues, so that $\mathbb{R}^d = E^u \oplus E^s$. Then there exist $r > 0$ and C^1 functions $\phi^u_{\bar{x}}\colon E^u_r \to E^s$ and $\phi^s_{\bar{x}}\colon E^s_r \to E^u$ such that writing*

$$W^u_{\mathrm{loc}}(\bar{x}) := \Gamma_{\bar{x}}(\phi^u_{\bar{x}}) \quad \text{and} \quad W^s_{\mathrm{loc}}(\bar{x}) := \Gamma_{\bar{x}}(\phi^s_{\bar{x}}),$$

the following are true.

1. *The manifolds $W^u_{\mathrm{loc}}(\bar{x})$ and $W^s_{\mathrm{loc}}(\bar{x})$ both contain $\bar{x}$, and are tangent at $\bar{x}$ to E^u and E^s respectively, in the sense that $D\phi^u_{\bar{x}}(0) = 0$ and $D\phi^s_{\bar{x}}(0) = 0$.*
2. *$W^u_{\mathrm{loc}}(\bar{x})$ is backward invariant, and is characterized as*

$$W^u_{\mathrm{loc}}(\bar{x}) = \{x \in B(\bar{x}, r) : f_{-t}(x) \in B(\bar{x}, r) \text{ for all } t \geq 0\}.$$

3. *$W^s_{\mathrm{loc}}(\bar{x})$ is forward invariant, and is characterized as*

$$W^s_{\mathrm{loc}}(\bar{x}) = \{x \in B(\bar{x}, r) : f_{t}(x) \in B(\bar{x}, r) \text{ for all } t \geq 0\}.$$

4. *Given any $x \in W^u_{\mathrm{loc}}(\bar{x})$ and $y \in W^s_{\mathrm{loc}}(\bar{x})$, we have*

$$f_{-t}(x) \to \bar{x} \quad \text{and} \quad f_t(y) \to \bar{x} \quad \text{as } t \to \infty.$$

Sketch of proof. We outline the proof for $\phi^u_{\bar{x}}$ (the proof for $\phi^s_{\bar{x}}$ is analogous) but do not provide full details. Recall from the proof of Theorem 3.11 in §3.2 that we write K^u_γ for the unstable cone of width γ around E^u. Using the arguments in that section, and particularly those in Proposition 3.15, one can show that for $r, \tau > 0$ sufficiently small, there exists $\chi > 1$ such that

$$\begin{aligned} &\text{if } y, z \in B(\bar{x}, r) \text{ satisfy } y - z \in K^u_\gamma, \\ &\text{then } f_\tau(y) - f_\tau(z) \in K^u_\gamma, \text{ and } \|f_\tau(y) - f_\tau(z)\| \geq \chi \|y - z\|, \end{aligned} \tag{3.14}$$

with a similar statement for the stable cone K^s_γ and the map $f_{-\tau}$. Consider the following space of γ-Lipschitz functions:

$$X := \{\phi\colon E^u_r \to E^s : \phi(0) = 0 \text{ and } \|\phi(w) - \phi(v)\| \leq \gamma \|w - v\| \text{ for all } v, w \in E^u_r\}.$$

Consider also the "strip" $S := \bar{x} + (E^u_r \oplus E^s)$, which is the union of the affine subspaces $\bar{x} + v + E^s$ taken over all $v \in E^u_r$. Observe that $\Gamma_{\bar{x}}(X) = \{\Gamma_{\bar{x}}(\phi) : \phi \in X\}$ can be characterized as the collection of all subsets $W \subset S$ containing $\bar{x}$ and satisfying:

- for every $v \in E^u_r$, the affine subspace $\bar{x} + v + E^s$ intersects W, and
- for every $y, z \in W$, we have $y - z \in K^u_\gamma$.

Given $W \in \Gamma_{\bar{x}}(X)$, it follows from the cone-invariance property in (3.14) that $f_\tau(W) \cap S$ is again an element of $\Gamma_{\bar{x}}(X)$. In particular, for each $\phi \in X$, there exists a unique $(f_\tau)_*(\phi) \in X$ such that

$$f_\tau(\Gamma_{\bar{x}}(\phi)) \cap S = \Gamma_{\bar{x}}((f_\tau)_*(\phi)). \tag{3.15}$$

The map $(f_\tau)_* \colon X \to X$ is called the *graph transform*. Using the fact that K^s_γ satisfies a contraction property analogous to the one for K^u_γ in (3.14), one can prove that the graph transform $(f_\tau)_*$ is a contraction on X in the uniform metric. Since X is a complete metric space in this metric, it follows that there is a unique fixed point of $(f_\tau)_*$, which we denote $\phi^u_{\bar{x}}$. It remains to show that $\phi^u_{\bar{x}}$ has the properties claimed in the statement of the theorem, namely:

Lec 29
W 4/8

(a) $\phi^u_{\bar{x}}(0) = 0$, and $W^u_{\text{loc}}(\bar{x}) := \Gamma_{\bar{x}}(\phi^u_{\bar{x}})$ is backward invariant;

(b) every $x \in W^u_{\text{loc}}(\bar{x})$ has a backwards orbit that stays in $B(\bar{x}, r)$ and goes to $\bar{x}$;

(c) every $x \in B(\bar{x}, r) \setminus W^u_{\text{loc}}(\bar{x})$ has a backwards orbit that eventually leaves $B(\bar{x}, r)$;

(d) $\phi^u_{\bar{x}}$ is continuously differentiable, with $D\phi^u_{\bar{x}}(0) = 0$.

Property (a) follows from the definition of X and the fact that $\phi^u_{\bar{x}}$ is a fixed point of the graph transform. The first part of property (b) follows from backwards invariance, and the second part follows from the fact that the forward flow expands K^u_γ, as in (3.14). For property (c), it suffices to let y be the point at which $x + E^s$ intersects $W^u_{\text{loc}}(\bar{x})$, and observe that $y - x \in E^s \in K^s_\gamma$, so by the time reversal of (3.14), as long as $f_{-t}(x)$ remains in $B(\bar{x}, r)$, the displacement $f_{-t}(y) - f_{-t}(x)$ lies in K^s_γ, with exponentially growing norm as t increases. Since $f_{-t}(x) \to \bar{x}$, we conclude that $f_{-t}(y)$ must eventually leave $B(\bar{x}, r)$.

Finally, for property (d), write $W = W^u_{\text{loc}}(\bar{x})$, and given $x \in W$, define the *tangent set* of W at x to be the set of all limit points of $\mathbb{R}(z - y)$ as $z - y \in W$ approach x. This set, which we denote $T_x W \subset \mathbb{R}^d$, is a linear subspace if and only if $\phi^u_{\bar{x}}$ is differentiable at x, in which case we have $T_x W = \{v + D\phi^u_{\bar{x}}(v) : v \in E^u\}$.

By (3.14), we have $T_x W \subset K^u_\gamma$ for each $x \in W$, and the definition of tangent set implies that given $x \in W$ and $x' := f_{-t}(x)$ for some $t \geq 0$, we have

$$T_x W = Df_t(x')(T_{x'} W) \subset Df_t(x')(K^u_\gamma) \quad \Rightarrow \quad T_x W \subset \bigcap_{t \geq 0} Df_t(f_{-t}x)(K^u_\gamma). \tag{3.16}$$

Now the idea is that the "width" of $Df_t(f_{-t}x)(K^u_\gamma)$ decreases exponentially fast in t, so that $T_x W$ is indeed a linear subspace, as desired, and when $x = \bar{x}$, this subspace is E^u. To make this precise, say that a *cone* is a subset $K \subset \mathbb{R}^d$ such that for every $x \in K$ and $c \in \mathbb{R}$, we have $cx \in K$. Given a cone $K \subset \mathbb{R}^d = E^u \oplus E^s$ such that $E^s \cap K = \{0\}$, consider for each $v \in E^u$ the set $L_K(v) := \{w \in E^s : v + w \in K$. Say that the *width* of K with respect to the decomposition $E^u \oplus E^s$ is $\sup\{\|w - w'\| : w, w' \in L_K(v), v \in E^u_1\}$. Observe that K has width 0 if and only if it is a linear subspace of $\mathbb{R}^d$ that is transverse to E^s.

Using the hyperbolicity properties of $Df_t(x)$ for $x \in B(\bar{x}, r)$, a short computation shows that there are constants $C, \kappa > 0$ such that the width of $Df_t(f_{-t}x)(K^u_\gamma)$ is at most $Ce^{-\kappa t}$ for all $t \geq 0$. It follows from (3.16) that the tangent set T_xW is a cone with width 0, and thus it is a linear subspace of $\mathbb{R}^d$, which completes the proof. □

Remark 3.26. Even when $DF(\bar{x})$ has some purely imaginary eigenvalues, we can still carry out some of the above arguments using the decompositions $E^u \oplus E^{cs}$ and $E^s \oplus E^{cu}$ to obtain local unstable and stable manifolds $W^u_{\text{loc}}(\bar{x})$ and $W^s_{\text{loc}}(\bar{x})$ that contain $\bar{x}$, are tangent to E^u and E^s, have the appropriate invariance properties, satisfy the convergence property in the final item of Theorem 3.25. However, they need not be characterized by the properties in the second and third items of the theorem: there may (or may not) be points outside of $W^u_{\text{loc}}(\bar{x})$ whose backward orbit remains within $B(\bar{x}, r)$ for all time, and similarly for $W^s_{\text{loc}}(\bar{x})$ and forward orbits.

We conclude this section with one final definition, which we will explore later on.

Definition 3.27. Let $\bar{x}$ be a fixed point for a C^1 vector field F that generates a flow $(f_t)_t$. Suppose that $DF(\bar{x})$ has at least one eigenvalue with positive real part, so that E^u is nontrivial and is tangent to a local unstable manifold $W^u_{\text{loc}}(\bar{x})$ by Theorem 3.25 and Remark 3.26. Then the *global unstable manifold* of $\bar{x}$ is

$$W^u(\bar{x}) := \bigcup_{t \geq 0} f_t(W^u_{\text{loc}}(\bar{x})).$$

Similarly, if $DF(\bar{x})$ has at least one eigenvalue with negative real part, then the *global stable manifold* of $\bar{x}$ is

$$W^s(\bar{x}) := \bigcup_{t \geq 0} f_{-t}(W^s_{\text{loc}}(\bar{x})).$$

When $\bar{x}$ is a hyperbolic fixed point, so that $\mathbb{R}^d = E^u \oplus E^s$, the global unstable and stable manifolds can be characterized as follows:

$$W^u(\bar{x}) = \left\{ x \in \mathbb{R}^d : \lim_{t \to \infty} f_{-t}(x) = \bar{x} \right\},$$
$$W^s(\bar{x}) = \left\{ x \in \mathbb{R}^d : \lim_{t \to \infty} f_t(x) = \bar{x} \right\}.$$

3.6 Stability of periodic orbits

Lec 30
F 4/10

Now suppose we have an orbit $\gamma \colon \mathbb{R} \to \mathbb{R}^d$ that is *not* a fixed point, and we want to determine the behavior of nearby trajectories. Recall from §1.7 and Proposition 1.25 that if we write

$$A(t) := DF(\gamma(t)) \in \mathbb{R}^{d \times d}, \tag{3.17}$$

and let $\Phi(t)$ be the fundamental matrix solution of the first variational equation

$$\dot{\Phi} = A(t)\Phi \tag{3.18}$$

with initial condition $\Phi(0) = \mathbb{I}$, then the time-t map f_t has derivative given by $Df_t(\gamma(0)) = \Phi(t) \in \mathbb{R}^{d\times d}$.

Suppose that γ is periodic, so that there exists $T > 0$ such that $\gamma(T) = \gamma(0)$, and hence $\gamma(t+T) = \gamma(t)$ for all $t \in \mathbb{R}$. The smallest such T is called the *period* of γ, and the matrix $Df_T(\gamma(0)) = \Phi(T)$ has the property that for every $n \in \mathbb{Z}$, we have

$$Df_{nT}(\gamma(0)) = (Df_T(\gamma(0)))^n. \tag{3.19}$$

Following the ideas in the previous sections, we expect the stability properties of the orbit γ to be determined by $Df_t(\gamma(0))$ as $t \to \pm\infty$, and (3.19) suggests that we should consider powers of the matrix $Df_T(\gamma(0))$. We will study this in two ways: via *Poincaré maps* and via *Floquet theory*.

To clarify what we wish to determine, consider for each $x \in \mathbb{R}^d$ the *distance to the periodic orbit* given by

$$d(x,\gamma) := \inf\{\|x - \gamma(t)\| : t \in \mathbb{R}\} = \inf\{\|x - \gamma(t)\| : t \in [0,T]\}.$$

Mimicking the definitions for fixed points, say that γ is

- *Lyapunov stable* if for every $\epsilon > 0$, there exists $\delta > 0$ such that for every $x \in \mathbb{R}^d$ satisfying $d(x,\gamma) < \delta$, we have $d(f_t x, \gamma) < \epsilon$ for all $t \geq 0$;
- *asymptotically stable* if it is Lyapunov stable and if in addition $\delta > 0$ can be chosen such that for every $x \in \mathbb{R}^d$ with $d(x,\gamma) < \delta$, we have $d(f_t(x),\gamma) \to 0$ as $t \to \infty$;
- *unstable* if it is not Lyapunov stable.

We can also consider analogues of the global stable and unstable manifolds from the previous section, writing

$$W^s(\gamma) := \{x \in \mathbb{R}^d : d(f_t(x),\gamma) \to 0 \text{ as } t \to \infty\},$$
$$W^u(\gamma) := \{x \in \mathbb{R}^d : d(f_{-t}(x),\gamma) \to 0 \text{ as } t \to \infty\}.$$

Note that we cannot yet refer to these as "manifolds", since so far we have proved nothing about their structure. To do so, and to give a criterion that tests for asymptotic stability, we now introduce the machinery of *Poincaré maps*.

Writing $\bar{x} = \gamma(0)$, a *Poincaré section* at $\bar{x}$ is a small embedded disk $\Sigma \subset \mathbb{R}^d$ of dimension $d-1$ that contains $\bar{x}$ and is transverse to the vector field F. For a concrete example, it suffices to fix $r > 0$ sufficiently small and to take

$$\Sigma = \{\bar{x} + v : v \in \mathbb{R}^d, v \perp F(\bar{x}), \|v\| < r\}.$$

More generally, writing $B := \{v \in \mathbb{R}^{d-1} : \|v\| < 1\}$, we say that a smooth injective map $\phi\colon B \to \mathbb{R}^d$ is an embedding if it admits a continuous injective extension to $\overline{B}$

and if $D\phi(u)$ is invertible for every $u \in B$. If $\phi(0) = \bar{x}$ and $\Sigma := \phi(B)$ is transverse to the vector field in the sense that $D\phi(u)(\mathbb{R}^{d-1}) \cap F(\phi(u)) = \{0\}$ for every $u \in B$, then Σ is a Poincaré section.

Given any $\epsilon > 0$, there exists $r > 0$ such that writing $\Sigma_r := B(\bar{x}, r) \cap \Sigma$, there is a smooth function $\tau \colon \Sigma_r \to (T - \epsilon, T + \epsilon)$ such that $f_{\tau(x)}(x) \in \Sigma$ for all $x \in \Sigma_r$. The function τ is called the *first return time*, and the map $f_\Sigma \colon x \mapsto f_{\tau(x)}(x)$ is the *first return map*, or *Poincaré map*. Given $x \in \Sigma$, observe that

- $d(f_t(x), \gamma) \to 0$ as $t \to \infty$ if and only if $f_\Sigma^n(x) \to \bar{x}$ as $n \to \infty$; and
- $d(f_{-t}(x), \gamma) \to 0$ as $t \to \infty$ if and only if $f_\Sigma^{-n}(x) \to \bar{x}$ as $n \to \infty$.

Here the notation f_Σ^n denotes iteration, so that for example, $f_\Sigma^2(x) = f_\Sigma(f_\Sigma(x))$. The flow $(f_t)_t$ can be considered as a dynamical system in which time is *continuous* (indexed by $\mathbb{R}$), while repeated iteration of the map f_Σ is a dynamical system in which time is *discrete* (indexed by $\mathbb{Z}$).

Arguing as in the preceding sections, one can prove stability results and a Hadamard–Perron Theorem for discrete-time systems. The key ideas are as follows.

- If $f \colon \mathbb{R}^m \to \mathbb{R}^m$ is C^1 and has a fixed point $\bar{x} = f(\bar{x})$, then the linearization around $\bar{x}$ is given by $v \mapsto Lv$, where $L = Df(\bar{x}) \in \mathbb{R}^{m \times m}$.
- We can write $\mathbb{R}^m = E^u \oplus E^c \oplus E^s$, where each of E^u, E^c, and E^s is a direct sum of generalized eigenspaces for L: eigenvalues with $|\lambda| > 1$ contribute to E^u; those with $|\lambda| = 1$ contribute to E^c; and those with $|\lambda| < 1$ contribute to E^s. As $n \to \infty$, each $v \in E^u$ has $L^{-n}v \to 0$, and each $v \in E^s$ has $L^n v \to 0$.
- If E^c and E^u are trivial – equivalently, if all eigenvalues of $Df(\bar{x})$ lie inside the unit circle – then $\bar{x}$ is asymptotically stable.
- If E^u is non-trivial – equivalently, if $Df(\bar{x})$ has at least one eigenvalue outside the unit circle – then $\bar{x}$ is unstable.
- If E^c is trivial and both E^u and E^s are non-trivial – equivalently, if $Df(\bar{x})$ has eigenvalues both inside and outside the unit circle, but none on it – then there are C^1 local stable and unstable manifolds $W^s_{\text{loc}}(\bar{x}) \subset \mathbb{R}^m$ and $W^u_{\text{loc}}(\bar{x}) \subset \mathbb{R}^m$ that contain $\bar{x}$ and are tangent to E^s and E^u, respectively. These are invariant: $f(W^s_{\text{loc}}(\bar{x})) \subset W^s_{\text{loc}}(\bar{x})$ and $f^{-1}(W^u_{\text{loc}}(\bar{x})) \subset W^u_{\text{loc}}(\bar{x})$. As $n \to \infty$, each $x \in W^s_{\text{loc}}(\bar{x})$ has $f^n(x) \to \bar{x}$, and each $x \in W^u_{\text{loc}}(\bar{x})$ has $f^{-n}(x) \to \bar{x}$.

For the Poincaré map f_Σ, we must study the eigenvalues of $Df_\Sigma(\bar{x})$, which is a linear transformation on the $(d-1)$-dimensional subspace $T_{\bar{x}}\Sigma$ of all vectors in $\mathbb{R}^d$ that are tangent to Σ. We have the following.

$$\begin{aligned} \text{all eigenvalues of } Df_\Sigma(\bar{x}) \text{ inside unit circle} &\quad\Rightarrow\quad \gamma \text{ is asymptotically stable} \\ \text{at least one eigenvalue of } Df_\Sigma(\bar{x}) \text{ outside unit circle} &\quad\Rightarrow\quad \gamma \text{ is unstable} \end{aligned}$$

If $Df_\Sigma(\bar{x})$ has eigenvalues inside and outside the unit circle, and none on it, then we can define global stable and unstable manifolds of γ (which we now know *are* in fact manifolds) by

$$W^s(\gamma) := \bigcup_{t\geq 0} f_{-t}(W^s_{\text{loc}}(\bar{x})) = \{x \in \mathbb{R}^d : d(f_t(x), \gamma) \to 0 \text{ as } t \to \infty\},$$
$$W^u(\gamma) := \bigcup_{t\geq 0} f_t(W^u_{\text{loc}}(\bar{x})) = \{x \in \mathbb{R}^d : d(f_{-t}(x), \gamma) \to 0 \text{ as } t \to \infty\}.$$

We can also define local stable and unstable manifolds $W^s_{\text{loc}}(\gamma)$ and $W^u_{\text{loc}}(\gamma)$ by only flowing for time $t \in [0, T]$.

Lec 31
M 4/13

Now we turn our attention to the question of determining the eigenvalues of $Df_\Sigma(\bar{x})$, which are called the *characteristic multipliers* of the periodic orbit γ. We do this in terms of the $d \times d$ matrix $Df_T(\bar{x})$, which is determined by (3.17) and (3.18). Since $f_\Sigma(x) = f_{\tau(x)}(x)$, we have the following for each v tangent to Σ at $x \in \Sigma$:

$$Df_\Sigma(x)(v) = Df_{\tau(x)}(x)(v) + \frac{\partial}{\partial t} f_t(x)\Big|_{t=\tau(x)} (D_v\tau)(x), \tag{3.20}$$

where $D_v\tau$ denotes the directional derivative of τ in the direction of v. Observe that

$$\frac{\partial}{\partial t} f_t(x)\Big|_{t=\tau(x)} = F(f_t(x))\big|_{t=\tau(x)} = F(f_\Sigma(x)) \qquad \text{and} \qquad (D_v\tau)(x) = \nabla\tau(x) \cdot v,$$

so (3.20) gives

$$Df_\Sigma(x)(v) = Df_{\tau(x)}(x)(v) + F(f_\Sigma(x))(\nabla\tau(x) \cdot v).$$

Evaluating this at $x = \bar{x}$ and using the fact that $f_\Sigma(\bar{x}) = \bar{x}$, we get

$$Df_\Sigma(\bar{x})(v) = Df_T(\bar{x})(v) + F(\bar{x})(\nabla\tau(\bar{x}) \cdot v).$$

Writing E_1 for the 1-dimensional subspace of $\mathbb{R}^d$ spanned by $F(\bar{x})$, we conclude that the matrices $Df_\Sigma(\bar{x})$ and $Df_T(\bar{x})$ satisfy the following properties:

- the range of $Df_\Sigma(\bar{x}) - Df_T(\bar{x})$ lies in E_1; and
- E_1 is an eigenspace of $Df_T(\bar{x})$, with eigenvalue 1.

This will lead us to the following conclusion, whose proof we leave as an exercise in linear algebra:

Proposition 3.28. *The spectra (sets of eigenvalues) of $Df_\Sigma(\bar{x})$ and $Df_T(\bar{x})$ satisfy*

$$\sigma(Df_T(\bar{x})) = \{1\} \cup \sigma(Df_\Sigma(\bar{x})). \tag{3.21}$$

Moreover, letting $E^\Sigma_\lambda \subset T_{\bar{x}}\Sigma$ and $E^T_\lambda \subset \mathbb{R}^d$ denote the generalized eigenspaces of λ for $Df_\Sigma(\bar{x})$ and $Df_T(\bar{x})$, respectively, we have $E^T_1 = E^\Sigma_1 \oplus E^1$, and there exists a linear isomorphism $C\colon \mathbb{R}^d \to \mathbb{R}^d$ such that the following are true:

- $C(v) = v$ *for all* $v \in E_1$*;*
- *for every* $\lambda \in \sigma(Df_T(\bar{x}) \setminus \{1\}) = \sigma(Df_\Sigma(\bar{x})) \setminus \{1\}$*, we have* $E^T_\lambda = C(E^\Sigma_\lambda)$*.*

In particular, the stable, unstable, and center subspaces of $Df_\Sigma(\bar{x})$ *and* $Df_T(\bar{x})$ *satisfy*

$$E^s_T = C(E^s_\Sigma), \qquad E^u_T = C(E^u_\Sigma), \qquad E^c_T = E^c_\Sigma \oplus E_1,$$

implying that $E^{cs}_T = E^{cs}_\Sigma \oplus E_1$ *and* $E^{cu}_T = E^{cu}_\Sigma \oplus E_1$*.*

Recall that the subspaces $E^{s,u,c}_T$ in the statement above are characterized by

$$E^s_T = \bigoplus_{|\lambda|<1} E^T_\lambda, \qquad E^u_T = \bigoplus_{|\lambda|>1} E^T_\lambda, \qquad E^c_T = \bigoplus_{|\lambda|=1} E^T_\lambda,$$

where the sums range over all $\lambda \in \sigma(Df_T(\bar{x}))$ satisfying the given condition; the subspaces $E^{s,u,c}_\Sigma$ are defined similarly.

Recalling that we defined the *characteristic multipliers* of γ to be the eigenvalues of $Df_\Sigma(\bar{x})$, we see from Proposition 3.28 that

- $\lambda \neq 1$ is a characteristic multiplier of γ if and only if it is an eigenvalue of $Df_T(\bar{x})$;
- $\lambda = 1$ is a characteristic multiplier of γ if and only if it is an eigenvalue of $Df_T(\bar{x})$ with multiplicity at least 2.

Now we recall from (3.17) and (3.18) that $Df_T(\bar{x}) = \Phi(T)$ solves $\dot{\Phi} = A(t)\Phi$ with $\Phi(0) = \mathbb{I}$, where $A(t) = DF(\gamma(t))$ is T-periodic: $A(t+T) = A(t)$ for all $t \in \mathbb{R}$. The next result uses this to describe "periodic fluctuations" in $Df_T(\bar{x})$.

Theorem 3.29 (Floquet's Theorem)**.** *Let* $A\colon \mathbb{R} \to \mathbb{R}^{d\times d}$ *be continuous and* T*-periodic, and let* $\Phi(t)$ *be the fundamental matrix solution of* $\dot{\Phi} = A(t)\Phi$ *with* $\Phi(0) = \mathbb{I}$*. Then there exists a constant matrix* B *and a* T*-periodic matrix* $Q(t)$ *such that* $\Phi(t) = Q(t)e^{Bt}$ *for all* $t \in \mathbb{R}$*. (The matrices* B *and* $Q(t)$ *may have complex entries.)*

Proof. With $t = 0$, we have

$$\mathbb{I} = \Phi(0) = Q(0)e^{B\cdot 0} = Q(0) = Q(T).$$

Putting $t = T$ and using this, we see that B must be chosen to satisfy

$$\Phi(T) = Q(T)e^{BT} = e^{BT},$$

In other words, we need a matrix logarithm.

Note that $\Phi(T)$ is invertible; indeed, for any $n \in \mathbb{Z}$ and $t \in \mathbb{R}$ we have

$$\Phi(t)\Phi(nT) = Df_t(\bar{x})Df_{nT}(\bar{x}) = D(f_t \circ f_{nT})(\bar{x}) = D(f_{t+nT})(\bar{x}) = \Phi(t+nT), \quad (3.22)$$

where the second equality uses the fact that $f_{nT}(\bar{x}) = \bar{x}$. Putting $n = 1$ and $t = -T$, this implies that $\Phi(-T)\Phi(T) = \Phi(-T+T) = \Phi(0) = \mathbb{I}$, so $\Phi(T)$ is invertible.

We claim that given any invertible square matrix Y, there exists a matrix X such that $e^X = Y$. Since $e^{CXC^{-1}} = Ce^XC^{-1}$, it suffices to consider the case when Y is in Jordan normal form. By decomposing Y as a direct sum of Jordan blocks, it suffices to consider the case when $Y = \lambda\mathbb{I} + N$ with $\lambda \neq 0$ and $N^d = 0$. In fact, we can reduce to the case when $\lambda = 1$ by writing $Y = \lambda(\mathbb{I} + M)$ for $M = N/\lambda$, and observing that once we have found a matrix Z such that $e^Z = \mathbb{I} + M$, we have $e^{cZ} = e^c\mathbb{I} + e^cM$ for each $c \in \mathbb{C}$, and since $\lambda \neq 0$, there exists $c \in \mathbb{C}$ such that $e^c = \lambda$.

To find Z such that $e^Z = \mathbb{I} + M$, recall that the power series for the logarithm of a real number near 1 is $\log(1+t) = \sum_{k=1}^{\infty}(-1)^{k+1}\frac{1}{k}t^k$. It is reasonable to expect that we can obtain the desired Z by taking

$$Z := \sum_{k=1}^{\infty}(-1)^{k+1}\frac{1}{k}M^k = \sum_{k=1}^{d-1}(-1)^{k+1}\frac{1}{k}M^k.$$

Indeed, at the level of formal power series, the fact that $e^{\log(1+t)} = 1 + t$ implies that $\sum_{n=0}^{\infty}\frac{1}{n!}\big(\sum_{k=1}^{\infty}(-1)^{k+1}\frac{1}{k}t^k\big)^n = 1 + t$. Replacing the real number t with the matrix M, the sum inside brackets becomes a polynomial rather than a power series, so there is no issue of convergence, and we conclude that $e^Z = \mathbb{I} + M$, as desired.

Once we have found a matrix X such that $e^X = \Phi(T)$, we take $B = \frac{1}{T}X$ and get $e^{BT} = \Phi(T)$. Thus the matrix $Q(t)$ must be defined by $\Phi(t)e^{-Bt}$, and it remains only to show that $Q(t+T) = Q(t)$. This follows from (3.22):

$$Q(t+T) = \Phi(t+T)e^{-B(t+T)} = \Phi(t)\Phi(T)e^{-BT}e^{-Bt} = \Phi(t)e^{-Bt} = Q(t). \qquad \square$$

The eigenvalues of B are the *characteristic exponents*. Since $e^{BT} = \Phi(T) = Df_T(\bar{x})$, we see that λ is a characteristic exponent if and only if $\chi = e^{\lambda T}$ is a characteristic multiplier. Observe that $|\chi| = e^{\operatorname{Re}\lambda}$. Each of λ and χ determines stability along a certain subspace:

- stable behavior is associated to $\operatorname{Re}\lambda < 0$ and $|\chi| < 1$;
- unstable behavior is associated to $\operatorname{Re}\lambda > 0$ and $|\chi| > 1$;
- the center direction corresponds to $\operatorname{Re}\lambda = 0$ and $|\chi| = 1$.

Consider the quantities

$$C_1 := \sup_{t\in[0,T]}\|Q(t)\| \quad \text{and} \quad C_0 := \inf_{t\in[0,T]}\|Q(t)^{-1}\|^{-1}.$$

If v is an eigenvector of B corresponding to the characteristic exponent λ, we have

$$\|Df_t(\bar{x})(v)\| = \|Q(t)e^{Bt}v\| = \|Q(t)e^{\lambda t}v\| = e^{t\cdot \operatorname{Re}\lambda}\|Q(t)v\| \le C_1 e^{t\cdot \operatorname{Re}\lambda}\|v\|,$$

and similarly,

$$\|Df_t(\bar{x})(v)\| = e^{t\cdot \operatorname{Re}\lambda}\|Q(t)v\| \ge C_0 e^{t\cdot \operatorname{Re}\lambda}\|v\|.$$

We combine these to obtain the bounds

$$C_0 e^{t\cdot \operatorname{Re}\lambda} \le \frac{\|Df_t(\bar{x})(v)\|}{\|v\|} \le C_1 e^{t\cdot \operatorname{Re}\lambda},$$

which describe the rate at which a displacement from $\bar{x} \in \gamma$ in the direction of v grows under the flow. From these bounds one can quickly deduce that

$$\lim_{t\to\pm\infty} \frac{1}{t} \log \|Df_t(\bar{x})(v)\| = \operatorname{Re}\lambda.$$

The left-hand side (when the limit exists, as it does here) is referred to as the *Lyapunov exponent* of the flow along the orbit of $\bar{x}$ in the direction of v. Here we have shown that for periodic orbits, the Lyapunov exponents are the real parts of the characteristic exponents. We remark that the definition of the Lyapunov exponent makes sense along non-periodic orbits as well.

3.7 Hartman–Grobman theorem on linearization

Lec 32
W 4/15

The Hadamard–Perron Theorem provided information on trajectories near a hyperbolic fixed point $\bar{x}$, describing solutions that converge to $\bar{x}$ in either forward or backward time. For linear systems, we also had a good description of trajectories that do not do either of these. The following result extends this to nonlinear systems.

Theorem 3.30 (Hartman–Grobman)**.** *Let $\bar{x}$ be a hyperbolic fixed point for a C^1 vector field F on $\mathbb{R}^d$. Then there exist $r > 0$ and $h\colon B(\bar{x}, r) \to \mathbb{R}^d$ such that*

1. *h is a homeomorphism onto its image (meaning that it is 1-1 and both h and h^{-1} are continuous);*
2. *$h(B(\bar{x}, r))$ is a neighborhood of 0; and*
3. *writing $(f_t)_t$ for the flow associated to $\dot{x} = F(x)$, and $(g_t)_t$ for the flow associated to $\dot{v} = (DF(\bar{x}))v$, the map h is a (local) topological conjugacy between these flows, meaning that $g_t \circ h = h \circ f_t$ wherever the maps are defined.*

Writing $U = B(\bar{x}, r)$ and $V = h(U)$, consider for each $t \ge 0$ the sets

$$U_t = \{x \in U : f_s(x) \in U \text{ for all } s \in [0, t]\},$$
$$V_t = \{v \in V : g_s(v) \in V \text{ for all } s \in [0, t]\}.$$

When $t \le 0$, make similar definitions, replacing $[0, t]$ with $[-t, 0]$. The condition of topological conjugacy requires that the following diagram commute.

$$\begin{array}{ccc} U_t & \xrightarrow{f_t} & U \\ \downarrow h & & \downarrow h \\ V_t & \xrightarrow{g_t} & V \end{array}$$

Outline of proof of Theorem 3.30. First consider the simplest case, when $d = 1$. Without loss of generality, assume that $\lambda := F'(\bar{x}) < 0$, so $\bar{x}$ is attracting. (The proof for $\lambda > 0$ is similar, with time reversed.) Then there exist points $a < \bar{x} < b$ such that $F(x) > 0$ on $[a, \bar{x})$ and $F(x) < 0$ on $(\bar{x}, b]$. Observe that the map $(x, t) \mapsto f_t(x)$ gives a homeomorphism between $\{a, b\} \times (0, \infty)$ and $(a, b) \setminus \{\bar{x}\}$, and that in order for $h\colon (a, b) \to V \ni 0$ to be a topological conjugacy between f_t and the linear flow $v \mapsto e^{\lambda t} v$, we must have $h(\bar{x}) = 0$ and

$$h(f_t(x)) = g_t(h(x)) = e^{-\lambda} h(x) \text{ for all } x \in \{a, b\} \text{ and } t > 0.$$

Thus we can choose $h(a)$ and $h(b)$ to be any nonzero real numbers with opposite signs, and then the rest of the conjugacy is uniquely determined.

In higher dimensions, a similar argument works whenever all the eigenvalues have real parts with the same sign. If all real parts are negative, then taking $U := B(\bar{x}, r) \subset \mathbb{R}^d$ and $V := B(0, r) \subset \mathbb{R}^d$, we have $\partial U = \{x \in \mathbb{R}^d : \|x - \bar{x}\| = r\}$, and for r sufficiently small, the map $\partial U \times (0, \infty) \to U \setminus \{\bar{x}\}$ defined by $(x, t) \mapsto f_t(x)$ is a homeomorphism. We can define h on ∂U by $h(x) = x - \bar{x}$, and then extend to $U \setminus \{\bar{x}\}$ by putting $h(f_t(x)) = g_t(x - \bar{x})$ for all $x \in \partial U$ and $t > 0$. Putting $h(\bar{x}) = 0$ completes the construction in this attracting case.

It remains to consider the case when E^u and E^s are both nontrivial. Recall that given $r > 0$, we write $E^u_r = B(0, r) \cap E^u \subset \mathbb{R}^d$, and similarly for E^s_r. Let us write

$$V_r := \{v + w : v \in E^u_r \text{ and } w \in E^s_r\} \quad \text{and} \quad U_r := \bar{x} + V_r.$$

We will produce a homeomorphism $h\colon U_r \to V_r$ that gives the required topological conjugacy. The idea is to identify

$$U_r \quad \longleftrightarrow \quad W^u_r(\bar{x}) \times W^s_r(\bar{x}) \quad \longleftrightarrow \quad E^u_r \times E^s_r \quad \longleftrightarrow \quad V_r \tag{3.23}$$

in a way that respects the flows f_t and g_t.

The third identification in (3.23) is the easiest, since it is implicit in the definition of V_r: each element of V_r can be written in a unique way as $v + w$, where $v \in E^u_r$ and $w \in E^s_r$. Geometrically, $v + w$ is characterized as the unique intersection point of the sets $v + E^s_r$ and $w + E^u_r$; we will adapt this in order to describe the first identification in (3.23). This, and the second identification, will use ideas from the statement and proof of the Hadamard–Perron Theorem, together with the following homeomorphisms, which follow the idea from the attracting and repelling cases above:

$$\begin{aligned} \pi^s\colon (\partial W^s_r(\bar{x})) \times (0, \infty) &\to W^s_r(\bar{x}) \setminus \{\bar{x}\} \\ (x, t) &\mapsto f_t(x) \end{aligned} \qquad \text{and} \qquad \begin{aligned} \pi^s_*\colon (\partial E^s_r) \times (0, \infty) &\to E^s_r \setminus \{0\} \\ (v, t) &\mapsto g_t(v) \end{aligned}$$

and similarly for π^u and π^u_*. With this in mind, we observe that given $v \in \partial E^s_r$ and $t > 0$, we have

$$g_t(v) + E^u_r = V_r \cap g_t(v + E^u_r).$$

Motivated by this, given $x \in \partial W^s_r(\bar{x})$, let $W^u_r(x) := x + E^u_r = \{x + v : v \in E^u_r\}$. For each $t > 0$, let $W^u_r(f_t(x))$ denote the connected component of $f_t(W^u_r(x))$ that contains $f_t(x)$. Using the bijection π^s, we have now defined $W^u_r(y)$ for each $y \in W^s_r(\bar{x})$.

The arguments in the proof of the Hadamard–Perron Theorem show that provided $r > 0$ is sufficiently small, each $W^u_r(y)$ is the graph of a C^1 function $E^u_r \to E^s$ that is γ-Lipschitz (recall that $\gamma > 0$ was the parameter in the definition of the unstable cone K^u_γ in that proof), and also that $W^u_r(f_t(x))$ converges to $W^u_r(\bar{x})$ as $t \to \infty$, so the map $y \mapsto W^u_r(y)$ is continuous.

The sets $\{W^u_r(y) : y \in W^s_r(\bar{x})\}$ partition U_r: they are disjoint, and their union is equal to U_r. (In fact, they form a *foliation*.) A similar construction with $t \le 0$ and the roles of the stable and unstable directions reversed produces another partition $\{W^s_r(y) : y \in W^u_r(\bar{x})\}$. These partitions produce the first identification in (3.23) above: given $(y, z) \in W^u_r(\bar{x}) \times W^s_r(\bar{x})$, the sets $W^s_r(y)$ and $W^u_r(z)$ are transverse submanifolds of U_r that intersect in a unique point $\Pi(y, z) \in U_r$. (This point is sometimes called the *bracket* of y and z and denoted by $[y, z]$.) See Figure 3.2.

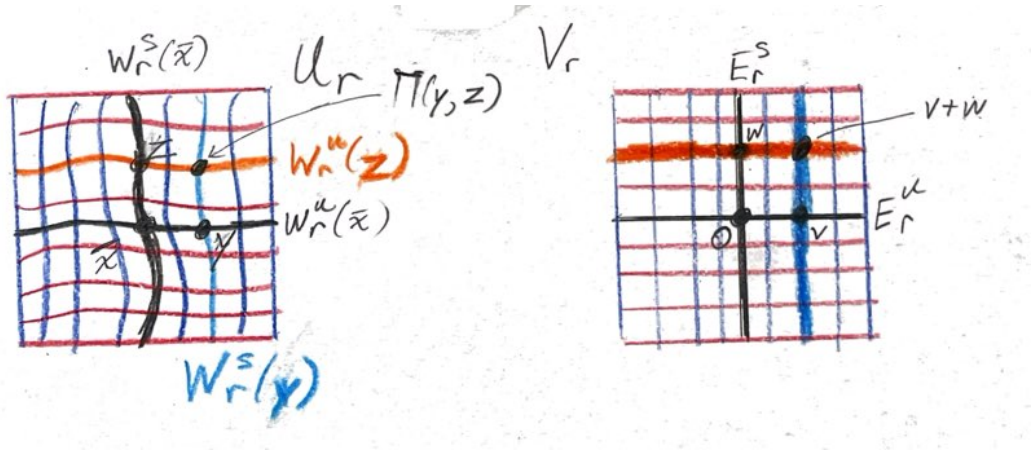

Figure 3.2: Foliations in the proof of the Hartman–Grobman Theorem.

Now we describe the second identification in (3.23). By the Hadamard–Perron Theorem, the local unstable and stable manifolds $W^u_r(\bar{x})$ and $W^s_r(\bar{x})$ are the graphs of C^1 functions $\phi^u \colon E^u_r \to E^s$ and $\phi^s \colon E^s_r \to E^u$, which can be extended continuously to $\overline{E^u_r}$ and $\overline{E^s_r}$. In particular, we can define bijections

$$\psi^u \colon \partial E^u_r \to \partial W^u_r(\bar{x}) \qquad \text{and} \qquad \psi^s \colon \partial E^s_r \to \partial W^s_r(\bar{x})$$
$$v \mapsto \bar{x} + v + \phi^u(v) \qquad\qquad w \mapsto \bar{x} + w + \phi^s(w)$$

and these can be combined with $\pi^{s,u}$ and $\pi_*^{s,u}$ as shown:

$$W_r^s(\bar{x}) \setminus \{\bar{x}\} \xleftarrow{\pi^s} (\partial W_r^s(\bar{x})) \times (0,\infty) \xleftarrow{\psi^s \times \mathrm{Id}} (\partial E_r^s) \times (0,\infty) \xrightarrow{\pi_*^s} E_r^s \setminus \{0\}.$$

Thus we define a homeomorphism $h^u \colon W_r^u(\bar{x}) \to E_r^u$ by $h^u(\bar{x}) = 0$ and

$$h^u(x) = \pi_*^u(v,t), \quad \text{where } \pi^s(\psi^s(v),t) = x$$

for all $x \neq \bar{x}$. Defining $h^s \colon W_r^s(\bar{x}) \to E_r^s$ similarly, these maps have the property that for all t such that the orbits are defined and remain in U_r and V_r, we have

$$h^s(f_t(x)) = g_t(h^s(x)) \qquad \text{and} \qquad h^u(f_t(x)) = g_t(h^u(x)). \tag{3.24}$$

Now we can define the homeomorphism $h \colon U_r \to V_r$, using h^s and h^u to provide the middle identification in (3.23). Given $x \in U_r$, let $(x^u, x^s) = \Pi^{-1}(x) \in (W_r^u(\bar{x}), W_r^s(\bar{x}))$ be uniquely determined by the condition that $x \in W_r^s(x^u) \cap W_r^u(x^s)$. Then define

$$h(x) := h^u(x^u) + h^s(x^s) \in V_r.$$

By (3.24) and flow-invariance of the foliations, h is a topological conjugacy. □

Chapter 4: Global nonlinear theory

4.1 Perturbations, again

Lec 33
F 4/17

Let us take a step back and review where we stand. Our overall goal has been to describe solutions of $\dot{x} = F(x)$ in $\mathbb{R}^d$. In Chapter 1, we found conditions for existence of uniqueness of solutions, techniques for approximating solutions, and formulas governing the dependence of solutions on initial conditions. In Chapter 2, we solved linear systems using the matrix exponential, which can be computed via Jordan normal form, and saw that the stability of the fixed point at the origin is determined by the eigendata of the matrix defining the system, and in particular by the real part of the eigenvalues. In Chapter 3, we developed the local theory of nonlinear systems, studying their stability via linearization first for fixed points, and then for periodic orbits via Poincaré maps and Floquet theory.

In this chapter, we turn our attention now to the *global* picture, seeking to understand how solutions of $\dot{x} = F(x)$ behave not just for the amount of time they spend near a fixed point or periodic orbit, but for all time.

The rather vague goal of "describe solutions" can be made precise in various ways. One way, which motivated many of the stability results in Chapter 3, is in terms of the asymptotic behavior of solutions as $t \to \pm\infty$. Given $x \in \mathbb{R}^d$, the *ω-limit set* of x for a flow $(f_t)_t$ is

$$\omega(x) := \{y \in \mathbb{R}^d : \text{there exists } t_k \to \infty \text{ such that } f_{t_k}(x) \to y \text{ as } k \to \infty\}.$$

Replacing $t_k \to \infty$ with $t_k \to -\infty$ gives the *α-limit set* $\alpha(x)$. When $\bar{x}$ is a hyperbolic fixed point of F, the global stable and unstable manifolds of $\bar{x}$ can be characterized as

$$W^s(\bar{x}) = \big\{x \in \mathbb{R}^d : \omega(x) = \{\bar{x}\}\big\} \quad \text{and} \quad W^u(\bar{x}) = \big\{x \in \mathbb{R}^d : \alpha(x) = \{\bar{x}\}\big\}.$$

Similarly, if $\gamma\colon \mathbb{R} \to \mathbb{R}^d$ is a hyperbolic periodic orbit, the global stable and unstable manifolds defined in §3.6 can be characterized as

$$W^s(\gamma) = \{x \in \mathbb{R}^d : \omega(x) = \gamma(\mathbb{R})\} \quad \text{and} \quad W^u(\gamma) = \{x \in \mathbb{R}^d : \alpha(x) = \gamma(\mathbb{R})\}.$$

From these examples, we see that ω- and α-limit sets can be single points and can also be periodic orbits. In the next sections, we will turn our attention to the question of what other possible forms these sets can take.

For now, though, we spend a little more time discussing another way of making "describe solutions" precise, following the idea of the Hartman–Grobman Theorem from §3.7, in which we described solutions by relating them to solutions of another system. The notion of topological conjugacy described there is only one way in which we might say that two systems are equivalent.

Definition 4.1. Two flows $(f_t)_t$ and $(g_t)_t$ are *C^r-conjugate* for some $r \geq 1$ if they are topologically conjugate and if the topological conjugacy h can be chosen such that both h and h^{-1} are r-times continuous differentiable (C^r).

Observe that topological conjugacy corresponds to C^0-conjugacy.

Definition 4.2. Two flows $(f_t)_t$ and $(g_t)_t$ are *orbit equivalent* if there exists a homeomorphsim h such that for every x, there exists a reparametrization $s\colon \mathbb{R} \to \mathbb{R}$ such that $g_t(h(x)) = h(f_{s(t)}(x))$ for all t; that is, if h carries *orbits* of $(f_t)_t$ to *orbits* of $(g_t)_t$, without necessarily preserving their parametrization.

In particular, we see that

$$C^2\text{-conjugate} \quad \Rightarrow \quad C^1\text{-conjugate} \quad \Rightarrow \quad C^0\text{-conjugate} \quad \Rightarrow \quad \text{orbit equivalent.}$$

The Hartman–Grobman Theorem says that if we perturb a linear hyperbolic fixed point, the resulting nonlinear system is locally C^0-conjugate to the linear one. In fact, Hartman proved that if F is C^2, then the two flows are C^1-conjugate. However, this cannot be strengthened further in general:

Example 4.3. The nonlinear vector field $F(x, y) = (2x, 4y + x^2)$ has a fixed point at the origin, with linear part given by $DF(0) = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 4 \end{smallmatrix}\right)$, but it can be shown that there is no C^2-conjugacy between the corresponding flows. We will not prove this here, but will just state that it is due to a *resonance* between the eigenvalues; that is, a relationship of the form $\lambda_j = \sum_i k_i \lambda_i$, where $k_i \in \mathbb{N}$ and $\sum_i k_i > 1$.

If we want to extend the Hartman–Grobman Theorem to periodic orbits, as we did with the Hadamard–Perron Theorem, then we need to weaken C^0-conjugacy to orbit

equivalence. This is because C^1-small perturbations of the vector field can change the period of the periodic orbit, while C^0-conjugacy cannot.

We say that the flow induced by a C^1 vector field F is *structurally stable* if for every C^1 vector field G that is sufficiently close to F in the C^1-sense, the flows induced by F and G are orbit equivalent. It turns out that there is an important class of structurally stable flows, called *Anosov flows*. A proper desription of this class of flows requires us to work on a smooth Riemannian manifold rather than in $\mathbb{R}^d$, but the main idea is that there is a hyperbolic splitting $E^u_x \oplus E^s_x$ at *every* point x, not just at some fixed point, with the property that

- the splitting is invariant, meaning that $Df_t(E^s_x) = E^s_{f_t(x)}$ and similarly for E^u_x; and
- the linear maps Df_t uniformly contract E^s_x in forward time and E^u_x in backward time, meaning that $\|Df_t(x)|_{E^s_x}\| \le Ce^{-\lambda t}$ and $\|Df_{-t}(x)|_{E_x 5}\| \le Ce^{-\lambda t}$ for all $t \ge 0$ and all x, where $C, \lambda > 0$ are independent of t and x.

Lec 34
M 4/20

A flow that is *not* structurally stable is said to represent a *bifurcation*, since arbitrarily small perturbations can create qualitatively different behavior. Another way of describing this is to consider a family of vector fields F_β that depend on a parameter $\beta \in \mathbb{R}$, and say that the family has a bifurcation at β_0 if for every $\epsilon > 0$, there exists $\beta \in (\beta_0 - \epsilon, \beta_0 + \epsilon)$ such that the flows of F_β and F_{β_0} are not orbit equivalent.

Since orbit equivalence must map fixed points to fixed points, and periodic orbits to periodic orbits, we see that a bifurcation happens whenever a fixed point or periodic orbit is created or destroyed as β varies. A bifurcation also happens whenever the stability of a fixed point or periodic orbit changes. These are examples of *local bifurcations*. Later, we will explore other possible kinds of bifurcations.

The following examples illustrate some types of local bifurcations. In each example with $d = 1$, it is helpful to plot the level set $F_\beta(x) = 0$ in the (β, x)-plane. On each component of the complement of this level set, the vector field satisfies either $F_\beta > 0$ (so x is increasing and we can draw an arrow pointing up) or $F_\beta < 0$ (so x is decreasing and we can draw an arrow pointing down). The resulting *bifurcation diagram* illustrates how the set of fixed points and their stabilities varies with β.

Example 4.4. With $d = 1$ and $F_\beta(x) = \beta - x^2$, there is a *saddle-node* bifurcation at $\beta = 0$, illustrated in Figure 4.1:

- for $\beta < 0$, there are no fixed points;
- for $\beta = 0$, there is a single fixed point at the origin, which is stable from the right and unstable from the left;
- for $\beta > 0$, there are two fixed points, at $\pm\sqrt{\beta}$, one of which is repelling and the other of which is attracting.

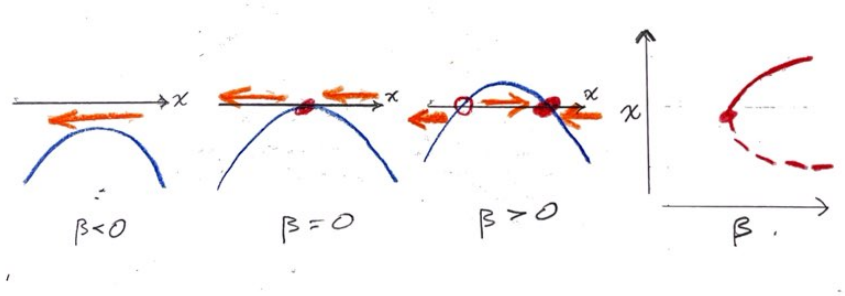

Figure 4.1: A saddle-node bifurcation.

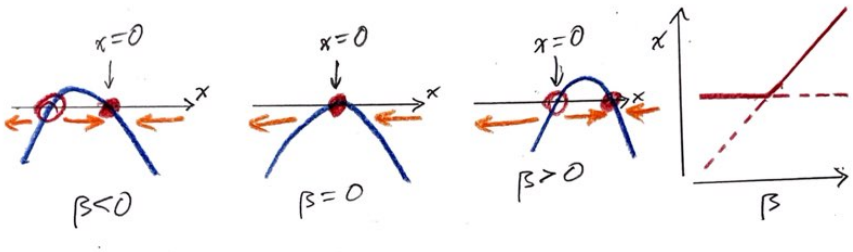

Figure 4.2: A transcritical bifurcation.

More generally, a saddle-node bifurcation occurs when a fixed point is created and then immediately splits into a pair of fixed points. An example of this with $d = 2$ is $F_\beta(x, y) = (\beta - x^2, -y)$, in which case the two fixed points are a saddle and an attracting node, hence the terminology.

Example 4.5. With $d = 1$ and $F_\beta(x) = \beta x - x^2$, there is a *transcritical* bifurcation at $\beta = 0$, illustrated in Figure 4.2:

- for every β, there is a fixed point at 0 and at β;
- for $\beta < 0$, we have $F_\beta(x) > 0$ on $(\beta, 0)$, and $F_\beta(x) < 0$ on $(-\infty, \beta) \cup (0, \infty)$, so the fixed point at 0 is attracting and the fixed point at β is repelling;
- for $\beta > 0$, these stabilities reverse, and the fixed point at 0 is repelling, while the fixed point at β is attracting.

Example 4.6. With $d = 1$ and $F_\beta(x) = \beta x - x^3$, there is a *pitchfork* bifurcation at $\beta = 0$, illustrated in Figure 4.3:

- for $\beta \leq 0$, there is exactly one fixed point, an attracting fixed point at the origin;
- for $\beta > 0$, the fixed point at the origin becomes repelling, and two new attracting fixed points appear at $x = \pm\sqrt{\beta}$.

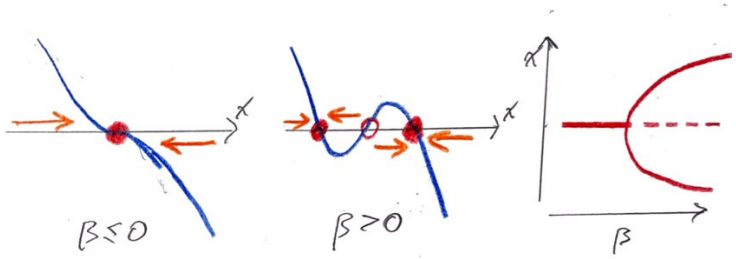

Figure 4.3: A pitchfork bifurcation.

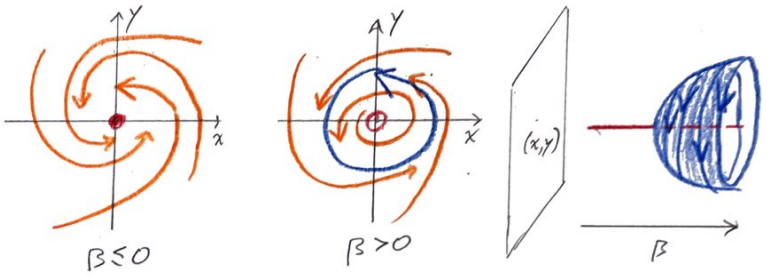

Figure 4.4: A Hopf bifurcation.

Example 4.7. Figure 4.4 illustrates the following bifurcation. With $d = 2$, the system

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

can be rewritten in polar coordinates as $\dot{r} = \beta r - r^3$ and $\dot{\theta} = 1$ (compare this to the previous example). There is a fixed point at the origin for every β. With $\beta \leq 0$, this fixed point is stable, while for $\beta > 1$, it is unstable, and there is an attracting periodic orbit given by $\gamma(t) = (R\cos t, R\sin t)$, where $R = \sqrt{\beta}$. Observe that $DF_\beta(0) = \left(\begin{smallmatrix} \beta & -1 \\ 1 & \beta \end{smallmatrix}\right)$ has characteristic polynomial $\lambda^2 - 2\beta\lambda + \beta^2 + 1$, so its eigenvalues are $\beta \pm i$. The bifurcation happens as these eigenvalues cross the imaginary axis (but stay away from 0). This is a *Hopf bifurcation*, in which a fixed point changes stability and a periodic orbit is created.

Lec 35
W 4/22

The bifurcations discussed above all involve fixed points, via creation, destruction, and/or changes in stability. They all have analogues for periodic orbits. We will study these by using the Poincaré section and return map. A bifurcation of a continuous-time system (a flow) that involves a change in existence and/or stability of periodic orbits corresponds to a bifurcation of a discrete-time systeme (a map) that involves a change in existence and/or stability of fixed points.

The simplest case is when $\gamma\colon \mathbb{R} \to \mathbb{R}^2$ is a periodic orbit of a planar flow, and Σ is a Poincaré section through some $\bar{x} \in \gamma(\mathbb{R})$. We can parametrize Σ by an interval $I \subset \mathbb{R}$,

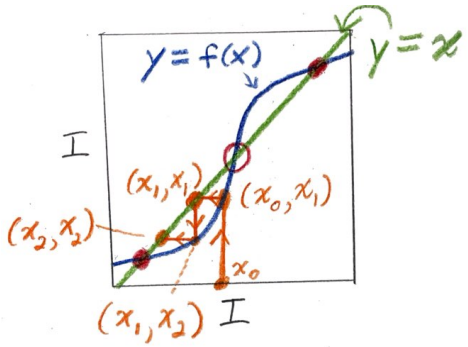

Figure 4.5: A cobweb diagram.

and so the periodic orbit γ corresponds to a fixed point $\bar{x} = f(\bar{x})$ of the Poincaré return map $f : I \to I$.

The fixed point is attracting if $|f'(\bar{x})| < 1$ and is repelling if $|f'(\bar{x})| > 1$. We saw this earlier in terms of eigenvalues and characteristic multipliers; in this 1-dimensional case, we can also view it geometrically in terms of *cobweb diagrams*.

Given an initial condition $x_0 \in I$, the trajectory of x_0 is given iteratively by $x_{n+1} = f(x_n)$. We describe this trajectory via the square $I \times I$, as shown in Figure 4.5. On this square, consider two curves: the diagonal line $y = x$ (the *bisectrix*), and the graph of the function $y = f(x)$. We represent the trajectory of x_0 by the following sequence of line segments.

- Given x_0, draw the vertical line $x = x_0$, and mark the point where this line intersects the curve $y = f(x)$. Since $x_1 = f(x_0)$, this point is (x_0, x_1).
- Now draw the horizontal line from (x_0, x_1) to the bisectrix $y = x$. This horizontal line intersects the bisectrix at (x_1, x_1), so we have now put x_1 into the first coordinate.
- Repeat this process: from the point (x_n, x_n), draw a vertical line, which intersects the graph of $y = f(x)$ at (x_n, x_{n+1}); then draw a horizontal line from this point, which intersects the graph of $y = x$ at (x_{n+1}, x_{n+1}).

Now we describe discrete-time bifurcations of fixed points analogous to the ones we saw earlier for flows. For each bifurcation, it is also helpful to visualize the *bifurcation diagram* in the (c, x) plane, where $c \in \mathbb{R}$ is the parameter; this diagram plots the fixed point(s) $\bar{x}$ in terms of c, representing stable fixed points as solid curves, and unstable fixed points as dashed curves.

Example 4.8. Given $c \in \mathbb{R}$, let $f_c(x) = c - x^2$. As shown in Figure 4.6, when c is large negative, we have $c - x^2 < x$ for all x, and there is no fixed point. As c increases, there exists $c_0 \in \mathbb{R}$ such that the graph of $f_c(x)$ is tangent to the bisectrix $y = x$ at exactly one point, which is a semi-stable fixed point. For $c > c_0$, there are two fixed points,

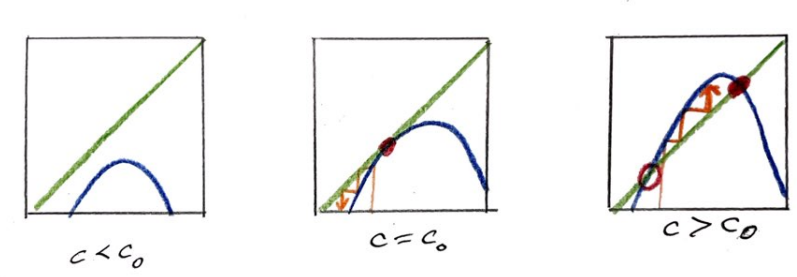

Figure 4.6: A tangent bifurcation.

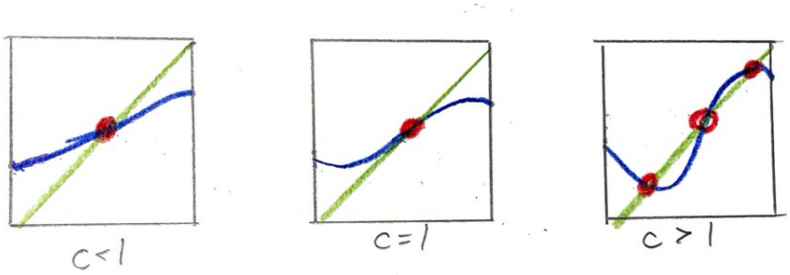

Figure 4.7: A pitchfork bifurcation.

one stable and one unstable. This is a *saddle-node bifurcation*; in this discrete-time setting, the term *tangent bifurcation* is also used. Fixed points occur when $c - x^2 = x$, so the bifurcation diagram includes the curve $c = x + x^2$.

Example 4.9. Given $c \in \mathbb{R}$, let $f_c(x) = cx - x^2$. We have $f_c(0) = 0$ for all c, and another solution of $cx - x^2 = x$ occurs when $c - x = 1$, so $x = c - 1$. For $c < 1$, the fixed point at $c - 1$ is unstable and the fixed point at 0 is stable; as c increases and passes through the value 1, the fixed points, pass through each other and exchange stabilities. This is a *transcritical bifurcation*. The bifurcation diagram includes the curves $x = 0$ and $c = x + 1$.

Example 4.10. Given $c \in \mathbb{R}$, let $f_c(x) = cx - x^3$. As shown in Figure 4.7, we again have $f_c(0) = 0$ for all c, and the remaining solutions of $cx - x^3 = x$ occur when $c - x^2 = 1$, so $x = \pm\sqrt{c-1}$. For $c < 1$, the fixed point at 0 is stable; for $c > 1$, this fixed point becomes unstable and the two remaining fixed points are stable. This is a *pitchfork bifurcation*. The bifurcation diagram includes the curves $x = 0$ and $c = x^2 + 1$.

The remaining type of local bifurcation we saw earlier was a Hopf bifurcation, in which a fixed point changed its stability and a periodic orbit surrounding it was created. The change in stability corresponded to eigenvalues crossing the imaginary axis as the parameter varies, but this crossing did not occur at $\lambda = 0$, as it did with the other three local bifurcations. Similarly here in the discrete-time setting, we now consider a bifurcation in which there is a change of stability associated to an eigenvalue crossing the unit circle (which is the relevant curve for stability of fixed points of maps)

away from the point 1. So while the three bifurcations we just saw had the property that $f'_{c_0}(\bar{x}) = 1$ at the bifurcation parameter, we now consider an example in which $f'_{c_0}(\bar{x}) = -1$.

Example 4.11. Let $f_c(x) = cx(1-x)$. This has one fixed point at 0 and another when $c(1-x) = 1$, so $\bar{x} = 1 - c^{-1}$. There is a transcritical bifurcation at $c = 1$, but we are interested in what happens for $c > 1$. We have

$$f'_c(\bar{x}) = c(1 - 2\bar{x}) = c(1 - 2(1 - c^{-1})) = c(-1 + 2c^{-1}) = 2 - c.$$

Observe that the derivative crosses 1 at $c = 1$, corresponding to the transcritical bifurcation, and at $c = 3$, the derivative crosses -1. So for $c > 3$, the fixed point is unstable. No new fixed points appear, but there is a period-2 orbit! One way to demonstrate this would be to look for fixed points of the second iterate $f_c^2(x)$, which involves solving a quartic equation for which we already know two of the roots. Another way is to observe that by the chain rule, we have

$$(f_c^2)'(x) = f'_c(f_c(x))f'_c(x) \quad \Rightarrow \quad (f_c^2)'(\bar{x}) = f'_c(\bar{x})^2,$$

so the derivative of the second iterate at the fixed point passes through 1 and creates a pitchfork bifurcation; existence of the two new fixed points for f_c^2 can be deduced from the intermediate value theorem. These two points form a period-2 orbit of f_c.

Recalling our original motivation, we see that each of these four kinds of bifurcations for a map describes a possible bifurcation behavior for a periodic orbit of a flow.

- In a saddle-node bifurcation, a pair of periodic orbits is created at the bifurcation parameter, with different stability properties.
- In a transcritical bifurcation, two periodic orbits pass through each other and exchange stabilities.
- In a pitchfork bifurcation, a periodic orbit changes stability and spawns two nearby periodic orbits of (approximately) the same period,
- In a period-doubling bifurcation, a periodic orbit changes stability and spawns one nearby periodic orbit of (approximately) double its period.

4.2 Global bifurcations

Lec 36
F 4/24

We have now seen how bifurcations can occur through changes in the structure and stability of the set of fixed points and periodic orbits for a flow. These are all *local* bifurcations. But this is not the whole story. It is also possible to have bifurcations that are not associated to any changes in the behavior of fixed points and periodic orbits, and we now turn our attention to these *global bifurcations*.

Recall the example of the pendulum from §3.4, given as the flow in $\mathbb{R}^2$ generated by the vector field $F(x, y) = (y, -\sin x)$, or equivalently, by the system of DEs

$$\dot{x} = y, \qquad \dot{y} = -\sin x.$$

In §3.4, we saw that the total energy function

$$H(x, y) := \frac{1}{2}y^2 - \cos x$$

has the property that $\dot{H} = \langle \nabla H, F(x, y) \rangle \equiv 0$, so trajectories of the pendulum flow lie on level sets of H. This helped to motivate our discussion of stable and unstable manifolds for hyperbolic fixed points.

By plotting the level sets of H, indicating the fixed points, and marking the direction of the flow, we obtain the phase portrait of the pendulum. The level sets of H contain the following types of orbits, illustrated in Figure 4.8.

- When $H = -1$, we have $(x, y) = (2n\pi, 0)$ for some $n \in \mathbb{Z}$; these are Lyapunov stable (but not asymptotically stable) fixed points.
- When $-1 < H < 1$, we have periodic orbits that correspond to oscillations of the pendulum.
- When $H = 1$, we have two types of orbits: hyperbolic fixed points at $((2n+1)\pi, 0)$ for each $n \in \mathbb{Z}$, and orbits connecting them. The fixed points correspond to the pendulum balancing on its pivot; the connecting orbits correspond to a pendulum swinging with just enough energy that it never stops and reverses direction, but not enough energy to "make it over the top".
- When $H > 1$, we have orbits (above the x-axis) that always move to the right, and other orbits (below the x-axis) that always move to the left. In this case, the pendulum has enough energy to swing "over the top" and spin endlessly. These orbits are not periodic in $\mathbb{R}^2$, but they correspond to periodic physical configurations, and become periodic in phase space if we quotient by 2π in the x-direction and work on the cylinder.

The second phase portrait in Figure 4.8 illustrated what happens if we incorporate the effect of friction into the model by adding a *damping* term, so that the original DE becomes

$$\ddot{x} + \beta\dot{x} + \sin x = 0,$$

where $\beta \geq 0$ is the damping coefficient, with $\beta = 0$ corresponding to the frictionless case. Once again writing $y = \dot{x}$, we get $\dot{y} = \ddot{x} = -\beta y - \sin x$, so the damped pendulum is modeled by the flow of the following vector field on $\mathbb{R}^2$:

$$F_\beta(x, y) = (y, -\beta y - \sin x).$$

Figure 4.8: Phase portraits for the undamped and damped pendulum.

The fixed points are the same, but since we now have $DF_\beta(x,y) = \left(\begin{smallmatrix} 0 & 1 \\ -\cos x & -\beta \end{smallmatrix}\right)$, the linearizations at these fixed points are now given by

$$DF_\beta(0) = \begin{pmatrix} 0 & 1 \\ -1 & -\beta \end{pmatrix} \quad \text{and} \quad DF_\beta(\pm\pi, 0) = \begin{pmatrix} 0 & 1 \\ 1 & -\beta \end{pmatrix}.$$

The first of these has eigenvalues $-\beta \pm i$, and we see that when β increases and becomes positive, the fixed point at the origin goes from being a nonlinear center, which is Lyapunov stable but not asymptotically stable, to being an attracting focus, which is asymptotically stable and attracts all nearby trajectories. This is a local bifurcation, detected by studying stability of fixed points.

The second fixed point, on the other hand, has no local bifurcation: a short computation reveals that $\left(\begin{smallmatrix} 0 & 1 \\ 1 & -\beta \end{smallmatrix}\right)$ has one positive and one negative eigenvalue for all small values of $\beta > 0$, and so $(\pm\pi, 0)$ continues to be a saddle. Nevertheless, there is a dramatic change in behavior associated to the stable and unstable manifolds of this fixed point: considering once again the function $H(x,y) = \frac{1}{2}y^2 - \cos x$, we see that now we have

$$\dot{H} = \langle (-\sin x, y), (y, -\beta y - \sin x) \rangle = -\beta y^2 \leq 0,$$

and the equality is strict everywhere except the x-axis. Thus instead of remaining confined to a single level set of H, a trajectory of the flow for F_β moves steadily "downhill" to points with lower and lower values of H. In particular, there is no longer an orbit connecting the fixed point $(-\pi, 0)$ to $(\pi, 0)$; rather the unstable manifold of $(\pm\pi, 0)$ is an orbit that now approaches the origin as $t \to \infty$, while the orbit along the stable manifold of $(\pm\pi, 0)$ approaches infinity as $t \to -\infty$.

Definition 4.12. A trajectory $(f_t(x))_{t\in\mathbb{R}}$ of a flow is called a *homoclinic orbit* if $\lim_{t\to\infty} f_t(x)$ and $\lim_{t\to-\infty} f_t(x)$ both exist and agree. In this case the common limit is a fixed point $\bar{x}$.

A trajectory $(f_t(x))_{t\in\mathbb{R}}$ of a flow is called a *heteroclinic orbit* if $\lim_{t\to\infty} f_t(x)$ and $\lim_{t\to-\infty} f_t(x)$ both exist, but are different points. In this case the limits are distinct fixed points.

To put it another way, x lies on a homoclinic orbit if $\omega(x) = \alpha(x) = \{\bar{x}\}$ for some $\bar{x} \in \mathbb{R}^d$, and x lies on a heteroclinic orbit if there are $\bar{x}, \bar{y} \in \mathbb{R}^d$ such that $\bar{x} \neq \bar{y}$, $\omega(x) = \bar{x}$, and $\alpha(x) = \bar{y}$.

Exercise 4.13. If two flows $(f_t)_t$ and $(g_t)_t$ are orbit equivalent via a homeomorphism h, then for every x, we have

$$h(\omega_F(x)) = \omega_G(h(x)) \quad \text{and} \quad h(\alpha_F(x)) = \alpha_G(h(x)),$$

where ω_F and α_F represent the ω- and α-limit sets with respect to the flow $(f_t)_t$ generated by the vector field F, and similarly for ω_G, α_G.

Thus ω- and α-limit sets provide a tool for identifying bifurcations by showing that two flows are not orbit equivalent. For example, one can have a *homoclinic bifurcation* in which a homoclinic orbit appears, as in the pendulum example, or in Figure 4.9.

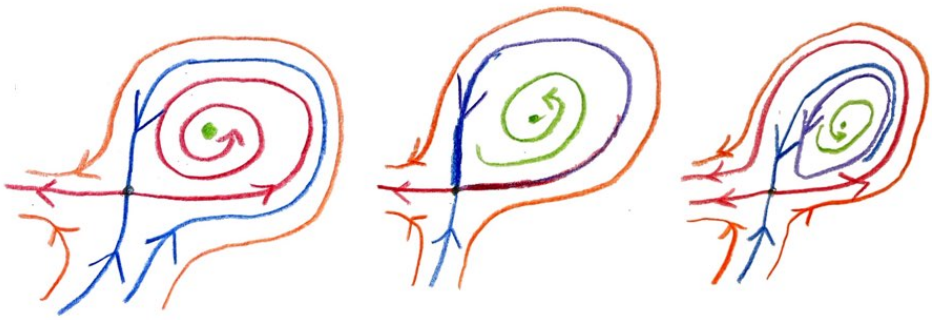

Figure 4.9: A homoclinic bifurcation.