Lecture Notes on

Ordinary Differential Equations

Christopher P. Grant

ODEs and Dynamical Systems
Lecture 1
Math 634
8/30/99

Ordinary Differential Equations

An ordinary differential equation (or ODE) is an equation involving derivatives of an unknown quantity with respect to a single variable. More precisely, suppose $j, k \in \mathbb{N}$, E is a Euclidean space, and

$$F : \operatorname{dom}(F) \subseteq \mathbb{R} \times \overbrace{E \times \cdots \times E}^{n+1 \text{ copies}} \to \mathbb{R}^j. \tag{1}$$

Then an *nth order ordinary differential equation* is an equation of the form

$$F(t, x(t), \dot{x}(t), \ddot{x}(t), x^{(3)}(t), \cdots, x^{(n)}(t)) = 0. \tag{2}$$

If $\mathcal{I} \subseteq \mathbb{R}$ is an interval, then $x : \mathcal{I} \to E$ is said to be *a solution of* (2) *on* $\mathcal{I}$ if x has derivatives up to order n at every $t \in \mathcal{I}$, and those derivatives satisfy (2). Often, we will suppress the dependence of x on t. Also, there will often be side conditions given that narrow down the set of solutions. In this class, we will concentrate on *initial conditions* which prescribe $x^{(\ell)}(t_0)$ for some fixed $t_0 \in \mathbb{R}$ (called the *initial time*) and some choices of $\ell \in \{0, 1, \ldots, n\}$. Some ODE classes study *two-point boundary-value problems*, in which the value of a function and its derivatives at two different points are required to satisfy given algebraic equations, but we won't focus on them in this one.

First-order Equations

Every ODE can be transformed into an equivalent first-order equation. In particular, given $x : \mathcal{I} \to E$, suppose we define

$$\begin{aligned} y_1 &:= x \\ y_2 &:= \dot{x} \\ y_3 &:= \ddot{x} \\ &\vdots \\ y_n &:= x^{(n-1)}, \end{aligned}$$

and let $y : \mathcal{I} \to E^n$ be defined by $y = (y_1, \dots , y_n)$. For $i = 1, 2, \dots , n-1$, define

$$G_i : \mathbb{R} \times E^n \times E^n \to \mathbb{R}$$

by

$$\begin{aligned} G_1(t, u, p) &:= p_1 - u_2 \\ G_2(t, u, p) &:= p_2 - u_3 \\ G_3(t, u, p) &:= p_3 - u_4 \\ &\vdots \\ G_{n-1}(t, u, p) &:= p_{n-1} - u_n, \end{aligned}$$

and, given F as in (1), define $G_n : \mathrm{dom}(G_n) \subseteq \mathbb{R} \times E^n \times E^n \to \mathbb{R}^j$ by

$$G_n(t, u, p) := F(t, u_1, \dots , u_n, p_1),$$

where

$$\mathrm{dom}(G_n) = \big\{(t, u, p) \in \mathbb{R} \times E^n \times E^n \bigm| (t, u_1, \dots , u_n, p_1) \in \mathrm{dom}(F)\big\}.$$

Letting $G : \mathrm{dom}(G_n) \subseteq \mathbb{R} \times E^n \times E^n \to \mathbb{R}^{n-1+j}$ be defined by

$$G := \begin{pmatrix} G_1 \\ G_2 \\ G_3 \\ \vdots \\ G_n \end{pmatrix},$$

we see that x satisfies (2) if and only if y satisfies $G(t, y(t), \dot{y}(t)) = 0$.

Equations Resolved w.r.t. the Derivative

Consider the first-order initial-value problem (or IVP)

$$\begin{cases} F(t, x, \dot{x}) = 0 \\ x(t_0) = x_0 \\ \dot{x}(t_0) = p_0, \end{cases} \tag{3}$$

where $F : \operatorname{dom}(F) \subseteq \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$, and x_0, p_0 are given elements of $\mathbb{R}^n$ satisfying $F(t_0, x_0, p_0) = 0$. The Implicit Function Theorem says that typically the solutions (t, x, p) of the (algebraic) equation $F(t, x, p) = 0$ near (t_0, x_0, p_0) form an $(n + 1)$-dimensional surface that can be parametrized by (t, x). In other words, locally the equation $F(t, x, p) = 0$ is equivalent to an equation of the form $p = f(t, x)$ for some $f : \operatorname{dom}(f) \subseteq \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ with (t_0, x_0) in the interior of $\operatorname{dom}(f)$. Using this f, (3) is locally equivalent to the IVP

$$\begin{cases} \dot{x} = f(t, x) \\ x(t_0) = x_0. \end{cases}$$

Autonomous Equations

Let $f : \operatorname{dom}(f) \subseteq \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$. The ODE

$$\dot{x} = f(t, x) \tag{4}$$

is *autonomous* if f doesn't really depend on t, *i.e.*, if $\operatorname{dom}(f) = \mathbb{R} \times \Omega$ for some $\Omega \subseteq \mathbb{R}^n$ and there is a function $g : \Omega \to \mathbb{R}^n$ such that $f(t, u) = g(u)$ for every $t \in \mathbb{R}$ and every $u \in \Omega$.

Every nonautonomous ODE is actually equivalent to an autonomous ODE. To see why this is so, given $x : \mathbb{R} \to \mathbb{R}^n$, define $y : \mathbb{R} \to \mathbb{R}^{n+1}$ by $y(t) = (t, x_1(t), \dots, x_n(t))$, and given $f : \operatorname{dom}(f) \subseteq \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, define a new function $\tilde{f} : \operatorname{dom}(\tilde{f}) \subseteq \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ by

$$\tilde{f}(p) = \begin{pmatrix} 1 \\ f_1(p_1, (p_2, \dots, p_{n+1})) \\ \vdots \\ f_n(p_1, (p_2, \dots, p_{n+1})) \end{pmatrix},$$

where $f = (f_1, \dots, f_n)^T$ and

$$\operatorname{dom}(\tilde{f}) = \left\{ p \in \mathbb{R}^{n+1} \;\middle|\; (p_1, (p_2, \dots, p_{n+1})) \in \operatorname{dom}(f) \right\}.$$

Then x satisfies (4) if and only if y satisfies $\dot{y} = \tilde{f}(y)$.

Because of the discussion above, we will focus our study on first-order autonomous ODEs that are resolved w.r.t. the derivative. This decision is not completely without loss of generality, because by converting other sorts of ODEs into an equivalent one of this form, we may be neglecting some special structure that might be useful for us to consider. This trade-off between abstractness and specificity is one that you will encounter (and have probably already encountered) in other areas of mathematics. Sometimes, when transforming the equation would involve too great a loss of information, we'll specifically study higher-order and/or nonautonomous equations.

Dynamical Systems

As we shall see, by placing conditions on the function $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$ and the point $x_0 \in \Omega$ we can guarantee that the autonomous IVP

$$\begin{cases} \dot{x} = f(x) \\ x(0) = x_0 \end{cases} \tag{5}$$

has a solution defined on some interval I containing 0 in its interior, and this solution will be unique (up to restriction or extension). Furthermore, it is possible to "splice" together solutions of (5) in a natural way, and, in fact, get solutions to IVPs with different initial times. These considerations lead us to study a structure known as a *dynamical system*.

Given $\Omega \subseteq \mathbb{R}^n$, a continuous dynamical system (or a *flow*) on Ω is a function $\varphi : \mathbb{R} \times \Omega \to \Omega$ satisfying:

1. $\varphi(0, x) = x$ for every $x \in \Omega$;
2. $\varphi(s, \varphi(t, x)) = \varphi(s + t, x)$ for every $x \in \Omega$ and every $s, t \in \mathbb{R}$;
3. φ is continuous.

If f and Ω are sufficiently "nice" we will be able to define a function $\varphi : \mathbb{R} \times \Omega \to \Omega$ by letting $\varphi(\cdot, x_0)$ be the unique solution of (5), and this definition will make φ a dynamical system. Conversely, any continuous dynamical system $\varphi(t, x)$ that is differentiable w.r.t. t is generated by an IVP.

Exercise 1 Suppose that:

- $\varphi : \mathbb{R} \times \Omega \to \Omega$ is a continuous dynamical system;
- $\dfrac{\partial \varphi(t,x)}{\partial t}$ exists for every $t \in \mathbb{R}$ and every $x \in \Omega$;
- $x_0 \in \Omega$ is given;
- $y : \mathbb{R} \to \Omega$ is defined by $y(t) := \varphi(t, x_0)$;
- $f : \Omega \to \Omega$ is defined by $f(p) := \left.\dfrac{\partial \varphi(s,p)}{\partial s}\right|_{s=0}$.

Show that y solves the IVP

$$\begin{cases} \dot{y} = f(y) \\ y(0) = x_0. \end{cases}$$

In this class (and Math 635) we will also study *discrete dynamical systems*. Given $\Omega \subseteq \mathbb{R}^n$, a discrete dynamical system on Ω is a function $\varphi : \mathbb{Z} \times \Omega \to \Omega$ satisfying:

1. $\varphi(0, x) = x$ for every $x \in \Omega$;
2. $\varphi(\ell, \varphi(m, x)) = \varphi(\ell + m, x)$ for every $x \in \Omega$ and every $\ell, m \in \mathbb{Z}$;
3. φ is continuous.

There is a one-to-one correspondence between discrete dynamical systems φ and *homeomorphisms* (continuous invertible functions) $F : \Omega \to \Omega$, this correspondence being given by $\varphi(1, \cdot) = F$. If we relax the requirement of invertibility and take a (possibly noninvertible) continuous function $F : \Omega \to \Omega$ and define $\varphi : \{0, 1, \dots\} \times \Omega \to \Omega$ by

$$\varphi(n, x) = \overbrace{F(F(\cdots(F}^{n \text{ copies}}(x))\cdots)),$$

then φ will almost meet the requirements to be a dynamical system, the only exception being that property 2, known as the *group property* may fail because $\varphi(n, x)$ is not even defined for $n < 0$. We may still call this a dynamical system; if we're being careful we may call it a *semidynamical system.*

In a dynamical system, the set Ω is called the *phase space*. Dynamical systems are used to describe the evolution of physical systems in which the state of the system at some future time depends only on the initial state of the system and on the elapsed time. As an example, Newtonian mechanics permits us to view the earth-moon-sun system as a dynamical system, but the phase space is not physical space $\mathbb{R}^3$, but is instead an 18-dimensional Euclidean space in which the coordinates of each point reflect the position and momentum of each of the three objects. (Why isn't a 9-dimensional space, corresponding to the three spatial coordinates of the three objects, sufficient?)

Existence of Solutions

Lecture 2
Math 634
9/1/99

Approximate Solutions

Consider the IVP

$$\begin{cases} \dot{x} = f(t,x) \\ x(t_0) = a, \end{cases} \tag{6}$$

where $f : \operatorname{dom}(f) \subseteq \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous, and $(t_0, a) \in \operatorname{dom}(f)$ is constant. The Fundamental Theorem of Calculus implies that (6) is equivalent to the integral equation

$$x(t) = a + \int_{t_0}^{t} f(s, x(s))\, ds. \tag{7}$$

How does one go about proving that (7) has a solution if, unlike the case with so many IVPs studied in introductory courses, a formula for a solution cannot be found? One idea is to construct a sequence of "approximate" solutions, with the approximations becoming better and better, in some sense, as we move along the sequence. If we can show that this sequence, or a subsequence, converges to something, that limit might be an exact solution.

One way of constructing approximate solutions is *Picard iteration*. Here, we plug an initial guess in for x on the right-hand side of (7), take the resulting value of the right-hand side and plug that in for x again, etc. More precisely, we can set $x_1(t) := a$ and recursively define x_{k+1} in terms of x_k for $k > 1$ by

$$x_{k+1}(t) := a + \int_{t_0}^{t} f(s, x_k(s))\, ds.$$

Note that if, for some k, $x_k = x_{k+1}$ then we have found a solution.

Another approach is to construct a *Tonelli sequence*. For each $k \in \mathbb{N}$, let $x_k(t)$ be defined by

$$x_k(t) = \begin{cases} a, & \text{if } t_0 \le t \le t_0 + 1/k \\ a + \displaystyle\int_{t_0}^{t-1/k} f(s, x_k(s))\, dx, & \text{if } t \ge t_0 + 1/k \end{cases} \tag{8}$$

for $t \geq t_0$, and define $x_k(t)$ similarly for $t \leq t_0$.

We will use the Tonelli sequence to show that (7) (and therefore (6)) has a solution, and will use Picard iterates to show that, under an additional hypothesis on f, the solution of (7) is unique.

Existence

For the first result, we will need the following definitions and theorems.

Definition A sequence of functions $g_k : \mathcal{U} \subseteq \mathbb{R} \to \mathbb{R}^n$ is *uniformly bounded* if there exists $M > 0$ such that $|g_k(t)| \leq M$ for every $t \in \mathcal{U}$ and every $k \in \mathbb{N}$.

Definition A sequence of functions $g_k : \mathcal{U} \subseteq \mathbb{R} \to \mathbb{R}^n$ is *uniformly equicontinuous* if for every $\varepsilon > 0$ there exists a number $\delta > 0$ such that $|g_k(t_1) - g_k(t_2)| < \varepsilon$ for every $k \in \mathbb{N}$ and every $t_1, t_2 \in \mathcal{U}$ satisfying $|t_1 - t_2| < \delta$.

Definition A sequence of functions $g_k : \mathcal{U} \subseteq \mathbb{R} \to \mathbb{R}^n$ *converges uniformly* to a function $g : \mathcal{U} \subseteq \mathbb{R} \to \mathbb{R}^n$ if for every $\varepsilon > 0$ there exists a number $N \in \mathbb{N}$ such that if $k \geq N$ and $t \in \mathcal{U}$ then $|g_k(t) - g(t)| < \varepsilon$.

Definition If $a \in \mathbb{R}^n$ and $\beta > 0$, then the *open ball of radius β centered at a*, denoted $\mathcal{B}(a, \beta)$, is the set

$$\left\{ x \in \mathbb{R}^n \;\middle|\; |x - a| < \beta \right\}.$$

Theorem (Arzela-Ascoli) *Every uniformly bounded, uniformly equicontinuous sequence of functions $g_k : \mathcal{U} \subseteq \mathbb{R} \to \mathbb{R}^n$ has a subsequence that converges uniformly on compact (closed and bounded) sets.*

Theorem (Uniform Convergence) *If a sequence of continuous functions $h_k : [b, c] \to \mathbb{R}^n$ converges uniformly to $h : [b, c] \to \mathbb{R}^n$, then*

$$\lim_{k \uparrow \infty} \int_b^c h_k(s)\, ds = \int_b^c h(s)\, ds.$$

We are now in a position to state and prove the Cauchy-Peano Existence Theorem.

Theorem (Cauchy-Peano) *Suppose $f : [t_0 - \alpha, t_0 + \alpha] \times \overline{\mathcal{B}(a, \beta)} \to \mathbb{R}^n$ is contin-*

uous and bounded by $M > 0$. Then (7) has a solution defined on $[t_0-b, t_0+b]$, where $b = \min\{\alpha, \beta/M\}$.

Proof. For simplicity, we will only consider $t \in [t_0, t_0 + b]$. For each $k \in \mathbb{N}$, let $x_k : [t_0, t_0 + b] \to \mathbb{R}^n$ be defined by (8). We will show that (x_k) converges to a solution of (6).

Step 1: Each x_k is well-defined.
Fix $k \in \mathbb{N}$. Note that the point (t_0, a) is in the interior of a set on which f is well-defined. Because of the formula for $x_k(t)$ and the fact that it is recursively defined on intervals of width $1/k$ moving steadily to the right, if x_k failed to be defined on $[t_0, t_0+b]$ then there would be $t_1 \in [t_0+1/k, t_0+b)$ for which $|x_k(t_1) - a| = \beta$. Pick the first such t_1. Using (8) and the bound on f, we see that

$$\begin{aligned}|x_k(t_1) - a| &= \left|\int_{t_0}^{t_1-1/k} f(s, x_k(s))\, ds\right| \le \int_{t_0}^{t_1-1/k} |f(s, x_k(s))|\, ds \\ &\le \int_{t_0}^{t_1-1/k} M\, ds = M(t_1 - t_0 - 1/k) < M(b - 1/k) \\ &\le \beta - M/k < \beta = |x_k(t_1) - a|,\end{aligned}$$

which is a contradiction.

Step 2: (x_k) is uniformly bounded.
Calculating as above, the formula (8) implies that

$$|x_k(t)| \le |a| + \int_{t_0}^{b+t_0-1/k} |f(s, x_k(s))|\, dx \le |a| + (b - 1/k)M \le |a| + \beta.$$

Step 3: (x_k) is uniformly equicontinuous.
If $t_1, t_2 \ge t_0 + 1/k$, then

$$|x_k(t_1) - x_k(t_2)| = \left|\int_{t_1}^{t_2} f(s, x_k(s))\, ds\right| \le \left|\int_{t_1}^{t_2} |f(s, x_k(s))|\, ds\right| \le M|t_2 - t_1|.$$

Since x_k is constant on $[t_0, t_0+1/k]$ and continuous at t_0+1/k, the estimate $|x_k(t_1) - x_k(t_2)| \le M|t_2 - t_1|$ holds for every $t_1, t_2 \in [t_0, t_0 + b]$. Thus, given $\varepsilon > 0$, we can set $\delta = \varepsilon/M$ and see that uniform equicontinuity holds.

Step 4: Some subsequence (x_{k_ℓ}) of (x_k) converges uniformly, say, to x on $[t_0, t_0 + b]$.
This follows directly from the previous steps and the Arzela-Ascoli Theorem.

Step 5: The sequence $(f(\cdot, x_{k_\ell}(\cdot)))$ converges uniformly to $f(\cdot, x(\cdot))$ on $[t_0, t_0 + b]$.
Let $\varepsilon > 0$ be given. Since f is continuous on a compact set, it is uniformly continuous. Thus, we can pick $\delta > 0$ such that $|f(s,p)-f(s,q)| < \varepsilon$ whenever $|p - q| < \delta$. Since (x_{k_ℓ}) converges uniformly to x, we can pick $N \in \mathbb{N}$ such that $|x_{k_\ell}(s) - x(s)| < \delta$ whenever $s \in [t_0, t_0 + b]$ and $\ell \geq N$. If $\ell \geq N$, then $|f(s, x_{k_\ell}(s)) - f(s, x(s))| < \varepsilon$.

Step 6: The function x is a solution of (6).
Fix $t \in [t_0, t_0 + b]$. If $t = t_0$, then clearly (7) holds. If $t > t_0$, then for ℓ sufficiently large

$$x_{k_\ell}(t) = a + \int_{t_0}^{t} f(s, x_{k_\ell}(s))\, ds - \int_{t-1/k_\ell}^{t} f(s, x_{k_\ell}(s))\, ds. \tag{9}$$

Obviously, the left-hand side of (9) converges to $x(t)$ as $\ell \uparrow \infty$. By the Uniform Convergence Theorem and the uniform convergence of $(f(\cdot, x_{k_\ell}(\cdot)))$, the first integral on the right-hand side of (9) converges to

$$\int_{t_0}^{t} f(s, x(s))\, ds,$$

and by the boundedness of f the second integral converges to 0. Hence, taking the limit of (9) as $\ell \uparrow \infty$ we see that x satisfies (7), and therefore (6), on $[t_0, t_0 + b]$. □

Note that this theorem guarantees existence, but not necessarily uniqueness, of a solution of (6).

Exercise 2 How many solutions of the IVP

$$\begin{cases} \dot{x} = 2\sqrt{|x|} \\ x(0) = 0, \end{cases}$$

on the interval $(-\infty, \infty)$ are there? Give formulas for all of them.

Uniqueness of Solutions

Lecture 3
Math 634
9/3/99

Uniqueness

If more than continuity of f is assumed, it may be possible to prove that

$$\begin{cases} \dot{x} = f(t, x) \\ x(t_0) = a, \end{cases} \tag{10}$$

has a *unique* solution. In particular Lipschitz continuity of $f(t, \cdot)$ is sufficient. (Recall that $g : \mathrm{dom}(g) \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is *Lipschitz continuous* if there exists a constant $L > 0$ such that $|g(x_1) - g(x_2)| \leq L|x_1 - x_2|$ for every $x_1, x_2 \in \mathrm{dom}(g)$; L is called a *Lipschitz constant* for g.)

One approach to uniqueness is developed in the following exercise, which uses what are know as *Gronwall inequalities*.

Exercise 3 Assume that the conditions of the Cauchy-Peano Theorem hold and that, in addition, $f(t, \cdot)$ is Lipschitz continuous with Lipschitz constant L for every t. Show that the solution of (10) is unique on $[t_0, t_0 + b]$ by completing the following steps. (The solution can similarly be shown to be unique on $[t_0 - b, t_0]$, but we won't bother doing that here.)

(a) If x_1 and x_2 are each solutions of (10) on $[t_0, t_0+b]$ and $U : [t_0, t_0+b] \to \mathbb{R}$ is defined by $U(t) := |x_1(t) - x_2(t)|$, show that

$$U(t) \le L \int_{t_0}^{t} U(s)\, ds \tag{11}$$

for every $t \in [t_0, t_0 + b]$.

(b) Pick $\varepsilon > 0$ and let

$$V(t) := \varepsilon + L \int_{t_0}^{t} U(s)\, ds.$$

Show that

$$V'(t) \le LV(t) \tag{12}$$

for every $t \in [t_0, t_0 + b]$, and that $V(t_0) = \varepsilon$.

(c) Dividing both sides of (12) by $V(t)$ and integrating from $t = t_0$ to $t = T$, show that $V(T) \le \varepsilon \exp[L(T - t_0)]$.

(d) By using (11) and letting $\varepsilon \downarrow 0$, show that $U(T) = 0$ for all $T \in [t_0, t_0 + b]$, so $x_1 = x_2$.

We will prove an existence-uniqueness theorem that combines the results of the Cauchy-Peano Theorem and Exercise 3. The reason for this apparently redundant effort is that the concepts and techniques introduced in this proof will be useful throughout this course.

First, we review some definitions and results pertaining to metric spaces.

Definition A *metric space* is a set $\mathcal{X}$ together with a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying:

1. $d(x, y) \geq 0$ for every $x, y \in \mathcal{X}$, with equality if and only if $x = y$;
2. $d(x, y) = d(y, x)$ for every $x, y \in \mathcal{X}$;
3. $d(x, y) + d(y, z) \geq d(x, z)$ for every $x, y, z \in \mathcal{X}$.

Definition A *normed vector space* is a vector space together with a function $\|\cdot\| : \mathcal{X} \to \mathbb{R}$ satisfying:

1. $\|x\| \geq 0$ for every $x \in \mathcal{X}$, with equality if and only if $x = 0$;
2. $\|\lambda x\| = |\lambda|\|x\|$ for every $x \in \mathcal{X}$ and every scalar λ;
3. $\|x + y\| \leq \|x\| + \|y\|$ for every $x, y \in \mathcal{X}$.

Every normed vector space is a metric space with metric $d(x, y) = \|x - y\|$.

Definition An *inner product space* is a vector space together with a function $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying:

1. $\langle x, x \rangle \geq 0$ for every $x \in \mathcal{X}$, with equality if and only if $x = 0$;
2. $\langle x, y \rangle = \langle y, x \rangle$ for every $x, y \in \mathcal{X}$;
3. $\langle \lambda x + \mu y, z \rangle = \lambda \langle x, z \rangle + \mu \langle y, z \rangle$ for every $x, y, z \in \mathcal{X}$ and all scalars μ, λ.

Every inner product space is a normed vector space with norm $\|x\| = \sqrt{\langle x, x \rangle}$.

Definition A sequence (x_n) in a metric space $\mathcal{X}$ is said to be (a) *Cauchy* (sequence) if for every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $d(x_m, x_n) < \varepsilon$ whenever $m, n \geq N$.

Definition A sequence (x_n) in a metric space $\mathcal{X}$ *converges* to x if for every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon$ whenever $n \geq N$.

Definition A metric space is said to be *complete* if every Cauchy sequence in $\mathcal{X}$ converges (in $\mathcal{X}$). A complete normed vector space is called a *Banach space.* A complete inner product space is called a *Hilbert space.*

Definition A function $f : \mathcal{X} \to \mathcal{Y}$ from a metric space to a metric space is said to be *Lipschitz continuous* if there exists $L \in \mathbb{R}$ such that $d(f(u), f(v)) \leq Ld(u, v)$ for every $u, v \in \mathcal{X}$. We call L a *Lipschitz constant*, and write $\text{Lip}(f)$ for the smallest Lipschitz constant that works.

Definition A *contraction* is a Lipschitz continuous function from a metric space to itself that has Lipschitz constant less than 1.

Definition A *fixed point* of a function $T : \mathcal{X} \to \mathcal{X}$ is a point $x \in \mathcal{X}$ such that $T(x) = x$.

Theorem (Contraction Mapping Principle) *If $\mathcal{X}$ is a complete metric space and $T : \mathcal{X} \to \mathcal{X}$ is a contraction, then T has a unique fixed point in $\mathcal{X}$.*

Proof. Pick $\lambda < 1$ such that $d(T(x), T(y)) \leq \lambda d(x, y)$ for every $x, y \in \mathcal{X}$. Pick any point $x_0 \in \mathcal{X}$. Define a sequence (x_k) by the recursive formula

$$x_{k+1} = T(x_k). \tag{13}$$

If $k \geq \ell \geq N$, then

$$\begin{aligned}
d(x_k, x_\ell) &\leq d(x_k, x_{k-1}) + d(x_{k-1}, x_{k-2}) + \cdots + d(x_{\ell+1}, x_\ell) \\
&\leq \lambda d(x_{k-1}, x_{k-2}) + \lambda d(x_{k-2}, x_{k-3}) + \cdots + \lambda d(x_\ell, x_{\ell-1}) \\
&\vdots \\
&\leq \lambda^{k-1} d(x_1, x_0) + \lambda^{k-2} d(x_1, x_0) + \cdots + \lambda^\ell d(x_1, x_0) \\
&\leq \frac{\lambda^N}{1 - \lambda} d(x_1, x_0).
\end{aligned}$$

Hence, (x_k) is a Cauchy sequence. Since $\mathcal{X}$ is complete, (x_k) converges to some $x \in \mathcal{X}$. By letting $k \uparrow \infty$ in (13) and using the continuity of T, we see that $x = T(x)$, so x is a fixed point of T.

If there were another fixed point y of T, then $d(x, y) = d(T(x), T(y)) \leq \lambda d(x, y)$, so $d(x, y) = 0$, which means $x = y$. This shows uniqueness of the fixed point. □

Picard-Lindelöf Theorem

Lecture 4
Math 634
9/8/99

Theorem *The space $\mathcal{C}([a,b])$ of continuous functions from $[a,b]$ to $\mathbb{R}^n$ equipped with the norm*

$$\|f\|_\infty := \sup\bigl\{|f(x)| \bigm| x \in [a,b]\bigr\}$$

is a Banach space.

Definition Two different norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on a vector space $\mathcal{X}$ are *equivalent* if there exist constants $m, M > 0$ such that

$$m\|x\|_1 \le \|x\|_2 \le M\|x\|_1$$

for every $x \in \mathcal{X}$.

Theorem *If $(\mathcal{X}, \|\cdot\|_1)$ is a Banach space and $\|\cdot\|_2$ is equivalent to $\|\cdot\|_1$ on $\mathcal{X}$, then $(\mathcal{X}, \|\cdot\|_2)$ is a Banach space.*

Theorem *A closed subspace of a complete metric space is a complete metric space.*

We are now in a position to state and prove the Picard-Lindelöf Existence-Uniqueness Theorem. Recall that we are dealing with the IVP

$$\begin{cases} \dot{x} = f(t,x) \\ x(t_0) = a. \end{cases} \tag{14}$$

Theorem (Picard-Lindelöf) *Suppose $f : [t_0 - \alpha, t_0 + \alpha] \times \overline{\mathcal{B}(a,\beta)} \to \mathbb{R}^n$ is continuous and bounded by M. Suppose, furthermore, that $f(t,\cdot)$ is Lipschitz continuous with Lipschitz constant L for every $t \in [t_0 - \alpha, t_0 + \alpha]$. Then (14) has a unique solution defined on $[t_0 - b, t_0 + b]$, where $b = \min\{\alpha, \beta/M\}$.*

Proof. Let $\mathcal{X}$ be the set of continuous functions from $[t_0 - b, t_0 + b]$ to $\overline{\mathcal{B}(a,\beta)}$. The norm

$$\|g\|_w := \sup\bigl\{e^{-2L|t-t_0|}|g(t)| \bigm| t \in [t_0 - b, t_0 + b]\bigr\}$$

is equivalent to the standard supremum norm $\|\cdot\|_\infty$ on $\mathcal{C}([t_0-b,t_0+b])$, so this vector space is complete under this weighted norm. The set $\mathcal{X}$ endowed with this norm/metric is a closed subset of this complete Banach space, so $\mathcal{X}$ equipped with the metric $d(x_1,x_2) := \|x_1-x_2\|_w$ is a complete metric space.

Given $x \in \mathcal{X}$, define $T(x)$ to be the function on $[t_0-b,t_0+b]$ given by the formula

$$T(x)(t) = a + \int_{t_0}^{t} f(s,x(s))\,dx.$$

Step 1: If $x \in \mathcal{X}$ then $T(x)$ makes sense.
This should be obvious.

Step 2: If $x \in \mathcal{X}$ then $T(x) \in \mathcal{X}$.
If $x \in \mathcal{X}$, then it is clear that $T(x)$ is continuous (and, in fact, differentiable). Furthermore, for $t \in [t_0-b,t_0+b]$

$$|T(x)(t)-a| = \left|\int_{t_0}^{t} f(s,x(s))\,ds\right| \le \left|\int_{t_0}^{t} |f(s,x(s))|\,ds\right| \le Mb \le \beta,$$

so $T(x)(t) \in \overline{\mathcal{B}(a,\beta)}$. Hence, $T(x) \in \mathcal{X}$.

Step 3: T is a contraction on $\mathcal{X}$.
Let $x,y \in \mathcal{X}$, and note that $\|T(x)-T(y)\|_w$ is

$$\sup\left\{ e^{-2L|t-t_0|}\left|\int_{t_0}^{t} [f(s,x(s))-f(s,y(s))]\,ds\right| \;\middle|\; t \in [t_0-b,t_0+b]\right\}.$$

For a fixed $t \in [t_0-b,t_0+b]$,

$$\begin{aligned}
&e^{-2L|t-t_0|}\left|\int_{t_0}^{t} [f(s,x(s))-f(s,y(s))]\,ds\right| \\
&\le e^{-2L|t-t_0|}\left|\int_{t_0}^{t} |f(s,x(s))-f(s,y(s))|\,ds\right| \\
&\le e^{-2L|t-t_0|}\left|\int_{t_0}^{t} L|x(s)-y(s)|\,ds\right| \\
&\le Le^{-2L|t-t_0|}\left|\int_{t_0}^{t} \|x-y\|_w e^{2L|s-t_0|}\,ds\right| \\
&= \frac{\|x-y\|_w}{2}\left(1-e^{-2L|t-t_0|}\right) \\
&\le \frac{1}{2}\|x-y\|_w.
\end{aligned}$$

Taking the supremum over all $t \in [t_0 - b, t_0 + b]$, we find that T is a contraction (with $\lambda = 1/2$).

By the contraction mapping principle, we therefore know that T has a unique fixed point in $\mathcal{X}$. This means that (14) has a unique solution in $\mathcal{X}$ (which is the only place a solution could be). □

Intervals of Existence

Lecture 5
Math 634
9/10/99

Maximal Interval of Existence

We begin our discussion with some definitions and an important theorem of real analysis.

Definition Given $f : \mathcal{D} \subseteq \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, we say that $f(t, x)$ is *locally Lipschitz continuous w.r.t. x on $\mathcal{D}$* if for each $(t_0, a) \in \mathcal{D}$ there is a number L and a product set $\mathcal{I} \times \mathcal{U} \subseteq \mathcal{D}$ containing (t_0, a) in its interior such that the restriction of $f(t, \cdot)$ to $\mathcal{U}$ is Lipschitz continuous with Lipschitz constant L for every $t \in \mathcal{I}$.

Definition A subset $\mathcal{K}$ of a topological space is *compact* if whenever $\mathcal{K}$ is contained in the union of a collection of open sets, there is a finite subcollection of that collection whose union also contains $\mathcal{K}$. The original collection is called a *cover* of $\mathcal{K}$, and the finite subcollection is called a *finite subcover* of the original cover.

Theorem (Heine-Borel) *A subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded.*

Now, suppose that $\mathcal{D}$ is an open subset of $\mathbb{R} \times \mathbb{R}^n$, $(t_0, a) \in \mathcal{D}$, and $f : \mathcal{D} \to \mathbb{R}^n$ is locally Lipschitz continuous w.r.t. x on $\mathcal{D}$. Then the Picard-Lindelöf Theorem indicates that the IVP

$$\begin{cases} \dot{x} = f(t, x) \\ x(t_0) = a \end{cases} \tag{15}$$

has a solution existing on some time interval containing t_0 in its interior and that the solution is unique on that interval. Let's say that an *interval of existence* is an interval containing t_0 on which a solution of (15) exists. The following theorem indicates how large an interval of existence may be.

Theorem (Maximal Interval of Existence) *The IVP* (15) *has a maximal interval of existence, and it is of the form (ω_-, ω_+), with $\omega_- \in [-\infty, \infty)$ and*

$\omega_+ \in (-\infty, \infty]$. *There is a unique solution* $x(t)$ *of* (15) *on* (ω_-, ω_+)*, and* $(t, x(t))$ *leaves every compact subset* $\mathcal{K}$ *of* $\mathcal{D}$ *as* $t \downarrow \omega_-$ *and as* $t \uparrow \omega_+$*.*

Proof.

Step 1: If $\mathcal{I}_1$ and $\mathcal{I}_2$ are open intervals of existence with corresponding solutions x_1 and x_2, then x_1 and x_2 agree on $\mathcal{I}_1 \cap \mathcal{I}_2$.
Let $\mathcal{I} = \mathcal{I}_1 \cap \mathcal{I}_2$, and let $\mathcal{I}^*$ be the largest interval containing t_0 and contained in I on which x_1 and x_2 agree. By the Picard-Lindelöf Theorem, $\mathcal{I}^*$ is nonempty. If $\mathcal{I}^* \neq \mathcal{I}$, then $\mathcal{I}^*$ has an endpoint t_1 in $\mathcal{I}$. By continuity, $x_1(t_1) = x_2(t_1) =: a_1$. The Picard-Lindelöf Theorem implies that

$$\begin{cases} \dot{x} = f(t, x) \\ x(t_1) = a_1 \end{cases} \tag{16}$$

has a local solution that is unique. But restrictions of x_1 and x_2 near t_1 each provide a solution to (16), so x_1 and x_2 must agree in a neighborhood of t_1. This contradiction tells us that $\mathcal{I}^* = \mathcal{I}$.

Now, let (ω_-, ω_+) be the union of all open intervals of existence.

Step 2: (ω_-, ω_+) is an interval of existence.
Given $t \in (\omega_-, \omega_+)$, pick an open interval of existence $\tilde{\mathcal{I}}$ that contains t, and let $x(t) = \tilde{x}(t)$, where $\tilde{x}$ is a solution to (15) on $\tilde{\mathcal{I}}$. Because of step 1, this determines a well-defined function $x : (\omega_-, \omega_+) \to \mathbb{R}^n$; clearly, it solves (15).

Step 3: (ω_-, ω_+) is the maximal interval of existence.
An extension argument similar to the one in Step 1 shows that every interval of existence is contained in an open interval of existence. Every open interval of existence is, in turn, a subset of (ω_-, ω_+).

Step 4: x is the only solution of (15) on (ω_-, ω_+).
This is a special case of Step 1.

Step 5: $(t, x(t))$ leaves every compact subset $\mathcal{K} \subset \mathcal{D}$ as $t \downarrow \omega_-$ and as $t \uparrow \omega_+$.
We only treat what happens as $t \uparrow \omega_+$; the other case is similar.

Let a compact subset $\mathcal{K}$ of $\mathcal{D}$ be given. For each point $(t, a) \in \mathcal{K}$, pick numbers $\alpha(t, a) > 0$ and $\beta(t, a) > 0$ such that

$$[t - 2\alpha(t, a), t + 2\alpha(t, a)] \times \overline{\mathcal{B}(a, 2\beta(t, a))} \subset \mathcal{D}.$$

Note that the collection of sets

$$\big\{(t - \alpha(t, a), t + \alpha(t, a)) \times \mathcal{B}(a, \beta(t, a)) \;\big|\; (t, a) \in \mathcal{K}\big\}$$

is a cover of $\mathcal{K}$. Since $\mathcal{K}$ is compact, a finite subcollection, say

$$\big\{(t_i - \alpha(t_i, a_i), t_i + \alpha(t_i, a_i)) \times \mathcal{B}(a_i, \beta(t_i, a_i))\big\}_{i=1}^{m},$$

covers $\mathcal{K}$. Let

$$\mathcal{K}' := \bigcup_{i=1}^{m} \Big([t_i - 2\alpha(t_i, a_i), t_i + \alpha(t_i, a_i)] \times \overline{\mathcal{B}(a_i, 2\beta(t_i, a_i))}\Big),$$

let

$$\tilde{\alpha} := \min\big\{\alpha(t_i, a_i)\big\}_{i=1}^{m},$$

and let

$$\tilde{\beta} := \min\big\{\beta(t_i, x_i)\big\}_{i=1}^{m}.$$

Since $\mathcal{K}'$ is a compact subset of $\mathcal{D}$, there is a constant $M > 0$ such that f is bounded by M on $\mathcal{K}'$. By the triangle inequality,

$$[t_0 - \tilde{\alpha}, t_0 + \tilde{\alpha}] \times \overline{\mathcal{B}(a, \tilde{\beta})} \subseteq \mathcal{K}',$$

for every $(t_0, a) \in \mathcal{K}$, so f is bounded by M on each such product set. According to the Picard-Lindelöf Theorem, this means that for every $(t_0, a) \in \mathcal{K}$ a solution to $\dot{x} = f(t, x)$ starting at (t_0, a) exists for at least $\min\{\tilde{\alpha}, \tilde{\beta}/M\}$ units of time. Hence, $x(t) \notin \mathcal{K}$ for $t > \omega_+ - \min\{\tilde{\alpha}, \tilde{\beta}/M\}$. □

Corollary *If $\mathcal{D}'$ is a bounded set and $\mathcal{D} = (c, d) \times \mathcal{D}'$ (with $c \in [-\infty, \infty)$ and $d \in (-\infty, \infty]$), then either $\omega_+ = d$ or $x(t) \to \partial\mathcal{D}'$ as $t \uparrow \omega_+$, and either $\omega_- = c$ or $x(t) \to \partial\mathcal{D}'$ as $t \downarrow \omega_-$.*

Corollary *If $\mathcal{D} = (c, d) \times \mathbb{R}^n$ (with $c \in [-\infty, \infty)$ and $d \in (-\infty, \infty]$), then either $\omega_+ = d$ or $|x(t)| \uparrow \infty$ as $t \uparrow \omega_+$, and either $\omega_- = c$ or $|x(t)| \uparrow \infty$ as $t \downarrow \omega_-$.*

If we're dealing with an autonomous equation on a bounded set, then the first corollary applies to tell us that the only way a solution could fail to exist for all time is for it to approach the boundary of the spatial domain. (Note that this is not the same as saying that $x(t)$ converges to a particular point on the boundary; can you give a relevant example?) The second corollary says that autonomous equations on all of $\mathbb{R}^n$ have solutions that exist until they become unbounded.

Global Existence

For the solution set of the autonomous ODE $\dot{x} = f(x)$ to be representable by a dynamical system, it is necessary for solutions to exist for all time. As the discussion above illustrates, this is not always the case. When solutions do die out in finite time by hitting the boundary of the phase space Ω or by going off to infinity, it may be possible to change the vector field f to a vector field $\tilde{f}$ that points in the same direction as the original but has solutions that exist for all time.

For example, if $\Omega = \mathbb{R}^n$, then we could consider the modified equation

$$\dot{x} = \frac{f(x)}{1 + |f(x)|}.$$

Clearly, $|\dot{x}| < 1$, so it is impossible for $|x|$ to approach infinity in finite time.

If, on the other hand, $\Omega \neq \mathbb{R}^n$, then consider the modified equation

$$\dot{x} = \frac{f(x)}{1 + |f(x)|} \cdot \frac{d(x, \mathbb{R}^n \setminus \Omega)}{1 + d(x, \mathbb{R}^n \setminus \Omega)},$$

where $d(x, \mathbb{R}^n \setminus \Omega)$ is the distance from x to the complement of Ω. It is not hard to show that it is impossible for a solution x of this equation to become unbounded or to approach the complement of Ω in finite time, so, again, we have global existence.

It may or may not seem obvious that if two vector fields point in the same direction at each point, then the solution curves of the corresponding ODEs in phase space match up. In the following exercise, you are asked to prove that this is true.

Exercise 4 Suppose that Ω is a subset of $\mathbb{R}^n$, that $f : \Omega \to \mathbb{R}^n$ and $g : \Omega \to \mathbb{R}^n$ are (continuous) vector fields, and that there is a continuous function $h : \Omega \to (0, \infty)$ such that $g(u) = h(u)f(u)$ for every $u \in \Omega$. If x is the only solution of

$$\begin{cases} \dot{x} = f(x) \\ x(0) = a \end{cases}$$

(defined on the maximal interval of existence) and y is the only solution of

$$\begin{cases} \dot{y} = g(y) \\ y(0) = a, \end{cases}$$

(defined on the maximal interval of existence), show that there is an increasing function $j : \operatorname{dom}(y) \to \operatorname{dom}(x)$ such that $y(t) = x(j(t))$ for every $t \in \operatorname{dom}(y)$.

Dependence on Parameters

Lecture 6
Math 634
9/13/99

Parameters vs. Initial Conditions

Consider the IVP

$$\begin{cases} \dot{x} = f(t,x) \\ x(t_0) = a, \end{cases} \tag{17}$$

and the paramterized IVP

$$\begin{cases} \dot{x} = f(t,x,\mu) \\ x(t_0) = a, \end{cases} \tag{18}$$

where $\mu \in \mathbb{R}^k$. We are interested in studying how the solution of (17) depends on the initial condition a and how the solution of (18) depends on the parameter μ. In a sense, these two questions are equivalent. For example, if x solves (17) and we let $\tilde{x} := x - a$ and $\tilde{f}(t,\tilde{x},a) := f(t,\tilde{x}+a)$, then $\tilde{x}$ solves

$$\begin{cases} \dot{\tilde{x}} = \tilde{f}(t,\tilde{x},a) \\ \tilde{x}(t_0) = 0, \end{cases}$$

so a appears as a parameter rather than an initial condition. If, on the other hand, x solves (18), and we let $\tilde{x} := (x,\mu)$ and $\tilde{f}(t,\tilde{x}) := (f(t,x,\mu),0)$, then $\tilde{x}$ solves

$$\begin{cases} \dot{\tilde{x}} = \tilde{f}(t,\tilde{x}) \\ \tilde{x}(t_0) = (a,\mu), \end{cases}$$

so μ appears as part of the initial condition, rather than as a parameter in the ODE.

We will concentrate on (18).

Continuous Dependence

The following result can be proved by an approach like that outlined in Exercise 3.

Theorem (Grownwall Inequality) *Suppose that $X(t)$ is a nonnegative, continuous, real-valued function on $[t_0, T]$ and that there are constants $C, K > 0$ such that*

$$X(t) \leq C + K \int_{t_0}^{t} X(s)\, ds$$

for every $t \in [t_0, T]$. Then

$$X(t) \leq C e^{K(t-t_0)}$$

for every $t \in [t_0, T]$.

Using the Grownwall inequality, we can prove that the solution of (18) depends continuously on μ.

Theorem (Continuous Dependence) *Suppose*

$$f : [t_0 - \alpha, t_0 + \alpha] \times \Omega_1 \times \Omega_2 \subseteq \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$$

is continuous. Suppose, furthermore, that $f(t, \cdot, \mu)$ is Lipschitz continuous with Lipschitz constant $L_1 > 0$ for every $(t, \mu) \in [t_0 - \alpha, t_0 + \alpha] \times \Omega_2$ and $f(t, x, \cdot)$ is Lipschitz continuous with Lipschitz constant $L_2 > 0$ for every $(t, x) \in [t_0 - \alpha, t_0 + \alpha] \times \Omega_1$. If $x_i : [t_0 - \alpha, t_0 + \alpha] \to \mathbb{R}^n$ $(i = 1, 2)$ satisfy

$$\begin{cases} \dot{x}_i = f(t, x_i, \mu_i) \\ x_i(t_0) = a, \end{cases}$$

then

$$|x_1(t) - x_2(t)| \leq \frac{L_2}{L_1} |\mu_1 - \mu_2| (e^{L_1 |t - t_0|} - 1) \tag{19}$$

for $t \in [t_0 - \alpha, t_0 + \alpha]$.

This theorem shows continuous dependence on parameters if, in addition to the hypotheses of the Picard-Lindelöf Theorem, the right-hand side of the equation is assumed to be Lipschitz continuous with respect to the parameter (on finite time intervals). The connection between (17) and (18) shows that the hypotheses of the Picard-Lindelöf Theorem are sufficient to guarantee continuous dependence on initial conditions. Note the exponential dependence of the modulus of continuity on $|t - t_0|$.

Proof. For simplicity, we only consider $t \geq t_0$. Note that

$$\begin{aligned}|x_1(t) - x_2(t)| &= \left|\int_{t_0}^{t} [f(s, x_1(s), \mu_1) - f(s, x_2(s), \mu_2]\, ds\right| \\ &\leq \int_{t_0}^{t} |f(s, x_1(s), \mu_1) - f(s, x_2(s), \mu_2)|\, ds\end{aligned}$$

$$\leq \int_{t_0}^{t} [|f(s, x_1(s), \mu_1) - f(s, x_1(s), \mu_2)| + |f(s, x_1(s), \mu_2) - f(s, x_2(s), \mu_2)|]\, ds$$

$$\leq \int_{t_0}^{t} [L_2|\mu_1 - \mu_2| + L_1|x_1(s) - x_2(s)|]\, ds$$

Let $X(t) = L_2|\mu_1 - \mu_2| + L_1|x_1(t) - x_2(t)|$. Then

$$X(t) \leq L_2|\mu_1 - \mu_2| + L_1 \int_{t_0}^{t} X(s)\, ds,$$

so by the Gronwall Inequality $X(t) \leq L_2|\mu_1 - \mu_2|e^{L_1(t-t_0)}$. This means that (19) holds. □

Exercise 5 Suppose that $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ are continuous and are each Lipschitz continuous with respect to their second variable. Suppose, also, that x is a global solution to

$$\begin{cases} \dot{x} = f(t, x) \\ x(t_0) = a, \end{cases}$$

and y is a global solution to

$$\begin{cases} \dot{y} = g(t, y) \\ y(t_0) = b. \end{cases}$$

(a) If $f(t, p) < g(t, p)$ for every $(t, p) \in \mathbb{R} \times \mathbb{R}$ and $a < b$, show that $x(t) < y(t)$ for every $t \geq t_0$.

(b) If $f(t, p) \leq g(t, p)$ for every $(t, p) \in \mathbb{R} \times \mathbb{R}$ and $a \leq b$, show that $x(t) \leq y(t)$ for every $t \geq t_0$. (Hint: You may want to use the results of part (a) along with a limiting argument.)

Differentiable Dependence

Theorem (Differentiable Dependence) *Suppose $f : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a continuous function and is continuously differentiable with respect to x and μ. Then the solution $x(\cdot, \mu)$ of*

$$\begin{cases} \dot{x} = f(t, x, \mu) \\ x(t_0) = a \end{cases} \tag{20}$$

is differentiable with respect to μ, and $y = x_\mu := \partial x / \partial \mu$ satisfies

$$\begin{cases} \dot{y} = f_x(t, x(t, \mu), \mu) y + f_\mu(t, x(t, \mu), \mu) \\ y(t_0) = 0. \end{cases} \tag{21}$$

That x_μ, if it exists, should satisfy the IVP (21) is not terribly surprising; (21) can be derived (formally) by differentiating (20) with respect to μ. The real difficulty is showing that x_μ exists. The key to the proof is to use the fact that (21) has a solution y and then to use the Gronwall inequality to show that difference quotients for x_μ converge to y.

Proof. Given μ, it is not hard to see that the right-hand side of the ODE in (21) is continuous in t and y and is locally Lipschitz continuous with respect to y, so by the Picard-Lindelöf Theorem we know that (21) has a unique solution $y(\cdot, \mu)$. Let

$$w(t, \mu, \Delta\mu) := \frac{x(t, \mu + \Delta\mu) - x(t, \mu)}{\Delta\mu}.$$

We want to show that $w(t, \mu, \Delta\mu) \to y(t, \mu)$ as $\Delta\mu \to 0$.

Let $z(t, \mu, \Delta\mu) := w(t, \mu, \Delta\mu) - y(t, \mu)$. Then

$$\frac{dz}{dt}(t, \mu, \Delta\mu) = \frac{dw}{dt}(t, \mu, \Delta\mu) - f_x(t, x(t, \mu), \mu) y(t, \mu) - f_\mu(t, x(t, \mu), \mu),$$

and

$$\begin{aligned} \frac{dw}{dt}(t, \mu, \Delta\mu) &= \frac{f(t, x(t, \mu + \Delta\mu), \mu + \Delta\mu) - f(t, x(t, \mu), \mu)}{\Delta\mu} \\ &= \frac{f(t, x(t, \mu + \Delta\mu), \mu + \Delta\mu) - f(t, x(t, \mu), \mu + \Delta\mu)}{\Delta\mu} \\ &\quad + \frac{f(t, x(t, \mu), \mu + \Delta\mu) - f(t, x(t, \mu), \mu)}{\Delta\mu} \\ &= f_x(t, x(t, \mu) + \theta_1 \Delta x, \mu + \Delta\mu) w(t, \mu, \Delta\mu) + f_\mu(t, x(t, \mu), \mu + \theta_2 \Delta\mu), \end{aligned}$$

for some $\theta_1, \theta_2 \in [0,1]$ (by the Mean Value Theorem), where

$$\Delta x := x(t,\mu+\Delta\mu) - x(t,\mu).$$

Hence,

$$\begin{aligned}
|\frac{dz}{dt}(t,\mu,\Delta\mu)| \leq{} & |f_\mu(t,x(t,\mu),\mu+\theta_2\Delta\mu) - f_\mu(t,x(t,\mu),\mu)| \\
&+ |f_x(t,x(t,\mu)+\theta_1\Delta x,\mu+\Delta\mu)| \cdot |z(t,\mu,\Delta\mu)| \\
&+ |f_x(t,x(t,\mu)+\theta_1\Delta x,\mu+\Delta\mu) - f_x(t,x(t,\mu),\mu+\Delta\mu)| \cdot |y(t,\mu)| \\
&+ |f_x(t,x(t,\mu),\mu+\Delta\mu) - f_x(t,x(t,\mu),\mu)| \cdot |y(t,\mu)| \\
&\leq p(t,\mu,\Delta\mu) + (|f_x(t,x(t,\mu),\mu)| + p(t,\mu,\Delta\mu))|z(t,\mu,\Delta\mu)|,
\end{aligned}$$

where $p(t,\mu,\Delta\mu) \to 0$ as $\Delta\mu \to 0$, uniformly on bounded sets.

Letting $X(t) = \varepsilon + (K+\varepsilon)|z|$, we see that if

$$|f_x(t,x(t,\mu),\mu)| \leq K \tag{22}$$

and

$$|p(t,\mu,\Delta\mu)| < \varepsilon, \tag{23}$$

then

$$|z(t)| \leq \int_{t_0}^{t} \left|\frac{dz}{ds}\right| ds \leq \int_{t_0}^{t} X(s)\,ds$$

so

$$X(t) \leq \varepsilon + (K+\varepsilon)\int_{t_0}^{t} X(s)\,ds,$$

which gives $X(t) \leq \varepsilon e^{(K+\varepsilon)(t-t_0)}$, by Gronwall's inequality. This, in turn, gives

$$|z| \leq \frac{\varepsilon(e^{(K+\varepsilon)(t-t_0)}-1)}{K+\varepsilon}.$$

Given $t \geq t_0$, pick K so large that (22) holds. As $\Delta\mu \to 0$, we can take ε arbitrarily small and still have (23) hold, to see that

$$\lim_{\Delta\mu\to 0} z(t,\mu,\Delta\mu) = 0.$$

□

Constant Coefficient Linear Equations

Lecture 7
Math 634
9/15/99

Linear Equations

Definition Given

$$f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n,$$

we say that the first-order ODE

$$\dot{x} = f(t, x) \tag{24}$$

is *linear* if every linear combination of solutions of (24) is a solution of (24).

Definition Given two vector spaces $\mathcal{X}$ and $\mathcal{Y}$, $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ is the space of linear maps from $\mathcal{X}$ to $\mathcal{Y}$.

Exercise 6 Show that if (24) is linear (and f is continuous) then there is a function $A : \mathbb{R} \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ such that $f(t, p) = A(t)p$, for every $(t, p) \in \mathbb{R} \times \mathbb{R}^n$.

ODEs of the form $\dot{x} = A(t)x + g(t)$ are also often called linear, although they don't satisfy the definition given above. These are called *inhomogeneous*; ODEs satisfying the previous definition are called *homogeneous*.

Constant Coefficients and the Matrix Exponential

Here we will study the autonomous IVP

$$\begin{cases} \dot{x} = Ax \\ x(0) = x_0, \end{cases} \tag{25}$$

where $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, or equivalently A is a (constant) $n \times n$ matrix.

If $n = 1$, then we're dealing with $\dot{x} = ax$. The solution is $x(t) = e^{ta}x_0$. When $n > 1$, we will show that we can similarly define e^{tA} in a natural way, and the solution of (25) will be given by $x(t) = e^{tA}x_0$.

Given $B \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, we define its matrix exponential

$$e^B := \sum_{k=0}^{\infty} \frac{B^k}{k!}.$$

We will show that this series converges, but first we specify a norm on $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$.

Definition The operator norm $\|B\|$ of an element $B \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ is defined by

$$\|B\| = \sup_{x \neq 0} \frac{|Bx|}{|x|} = \sup_{|x|=1} |Bx|.$$

$\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ is a Banach space under the operator norm. Thus, to show that the series for e^B converges, it suffices to show that

$$\left\| \sum_{k=\ell}^{m} \frac{B^k}{k!} \right\|$$

can be made arbitrarily small by taking $m \geq \ell \geq N$ for N sufficiently large.

Suppose $B_1, B_2 \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ and B_2 does not map everything to zero. Then

$$\begin{aligned} \|B_1 B_2\| &= \sup_{x \neq 0} \frac{|B_1 B_2 x|}{|x|} = \sup_{B_2 x \neq 0, x \neq 0} \frac{|B_1 B_2 x|}{|B_2 x|} \cdot \frac{|B_2 x|}{|x|} \\ &\leq \left(\sup_{y \neq 0} \frac{|B_1 y|}{|y|} \right) \left(\sup_{x \neq 0} \frac{|B_2 x|}{|x|} \right) = \|B_1\| \cdot \|B_2\|. \end{aligned}$$

If B_2 does map everything to zero, then $\|B_2\| = \|B_1 B_2\| = 0$, so $\|B_1 B_2\| \leq \|B_1\| \cdot \|B_2\|$, obviously. Thus, the operator norm is *submultiplicative*. Using this property, we have

$$\left\| \sum_{k=\ell}^{m} \frac{B^k}{k!} \right\| \leq \sum_{k=\ell}^{m} \left\| \frac{B^k}{k!} \right\| \leq \sum_{k=\ell}^{m} \frac{\|B\|^k}{k!}.$$

Since the regular exponential series (for real arguments) has an infinite radius of convergence, we know that the last quantity in this estimate goes to zero as $\ell, m \uparrow \infty$.

Thus, e^B makes sense, and, in particular, e^{tA} makes sense for each fixed $t \in \mathbb{R}$ and each $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$. But does $x(t) := e^{tA}x_0$ solve (25)? To check that, we'll need the following important property.

Lemma *If* $B_1, B_2 \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ *and* $B_1B_2 = B_2B_1$, *then* $e^{B_1+B_2} = e^{B_1}e^{B_2}$.

Proof. Using commutativity, we have

$$\begin{aligned} e^{B_1}e^{B_2} &= \left(\sum_{j=0}^{\infty} \frac{B_1^j}{j!}\right)\left(\sum_{k=0}^{\infty} \frac{B_2^k}{k!}\right) = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \frac{B_1^jB_2^k}{j!k!} = \sum_{i=0}^{\infty}\sum_{j+k=i} \frac{B_1^jB_2^k}{j!k!} \\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{i} \frac{B_1^jB_2^{(i-j)}}{j!(i-j)!} = \sum_{i=0}^{\infty}\sum_{j=0}^{i} \binom{i}{j} \frac{B_1^jB_2^{(i-j)}}{i!} \\ &= \sum_{i=0}^{\infty} \frac{(B_1+B_2)^i}{i!} = e^{(B_1+B_2)}. \end{aligned}$$

□

Now, if $x : \mathbb{R} \to \mathbb{R}^n$ is defined by $x(t) := e^{tA}x_0$, then

$$\begin{aligned} \frac{d}{dt}x(t) &= \lim_{h\to 0} \frac{x(t+h)-x(t)}{h} = \lim_{h\to 0} \frac{e^{(t+h)A}x_0 - e^{tA}x_0}{h} \\ &= \left(\lim_{h\to 0} \frac{e^{(t+h)A} - e^{tA}}{h}\right)x_0 = \left(\lim_{h\to 0} \frac{e^{hA} - I}{h}\right)e^{tA}x_0 \\ &= \left(\lim_{h\to 0} \sum_{k=1}^{\infty} \frac{h^{k-1}A^k}{k!}\right)e^{tA}x_0 = Ae^{tA}x_0 = Ax(t), \end{aligned}$$

so $x(t) = e^{tA}x_0$ really does solve (25).

Understanding the Matrix Exponential

Lecture 8
Math 634
9/17/99

Transformations

Now that we have a representation of the solution of constant-coefficient initial-value problems, we should ask ourselves: "What good is it?" Does the power series formula for the matrix exponential provide an efficient means for calculating exact solutions? Not usually. Is it an efficient way to compute accurate numerical approximations to the matrix exponential? Not according to *Matrix Computations* by Golub and Van Loan. Does it provide insight into how solutions behave? It is not clear that it does. There are, however, transformations that may help us handle these problems.

Suppose that $B, P \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ are related by a similarity transformation; *i.e.*, $B = QPQ^{-1}$ for some invertible Q. Calculating, we find that

$$\begin{aligned} e^B = \sum_{k=0}^{\infty} \frac{B^k}{k!} = \sum_{k=0}^{\infty} \frac{(QPQ^{-1})^k}{k!} &= \sum_{k=0}^{\infty} \frac{QP^kQ^{-1}}{k!} \\ &= Q\left(\sum_{k=0}^{\infty} \frac{P^k}{k!}\right)Q^{-1} = Qe^PQ^{-1}. \end{aligned}$$

It would be nice if, given B, we could choose Q so that P were a diagonal matrix, since

$$e^{\operatorname{diag}\{p_1,p_2,\dots,p_n\}} = \operatorname{diag}\{e^{p_1}, e^{p_2}, \dots, e^{p_n}\}.$$

Unfortunately, this cannot always be done. Over the next few lectures, we will show that what can be done, in general, is to pick Q so that $P = S + N$, where S is a *semisimple* matrix with a fairly simple form, N is a *nilpotent* matrix of a fairly simple form, and S and N commute. (Recall that a matrix is semisimple if it is diagonalizable over the complex numbers and that a matrix is nilpotent if some power of the matrix is 0.) The forms of S and N are simple enough that we can calculate their exponentials fairly easily, and then we can multiply them to get the exponential of $S + N$.

We will spend a significant amount of time carrying out the project described in the previous paragraph, even though it is linear algebra that some of you have probably seen before. Since understanding the behavior of constant coefficient systems plays a vital role in helping us understand more complicated systems, I feel that the time investment is worth it. The particular approach we will take follows chapters 3, 4, 5, and 6, and appendix 3 of Hirsch and Smale fairly closely.

Eigensystems

Given $B \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, recall that that $\lambda \in \mathbb{C}$ is an *eigenvalue* of B if $Bx = \lambda x$ for some nonzero $x \in \mathbb{R}^n$ or if $\tilde{B}x = \lambda x$ for some nonzero $x \in \mathbb{C}^n$, where $\tilde{B}$ is the *complexification* of B; *i.e.*, the element of $\mathcal{L}(\mathbb{C}^n, \mathbb{C}^n)$ which agrees with B on $\mathbb{R}^n$. (Just as we often identify a linear operator with a matrix representation of it, we will usually not make a distinction between an operator on a real vector space and its complexification.) A nonzero vector x for which $Bx = \lambda x$ for some scalar λ is an *eigenvector*. An eigenvalue λ with corresponding eigenvector x form an *eigenpair* (λ, x).

If an operator $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ is chosen at random, it would almost surely have n distinct eigenvalues $\{\lambda_1, \dots, \lambda_n\}$ and n corresponding linearly independent eigenvectors $\{x_1, \dots, x_n\}$. If this is the case, then A is similar to the (possibly complex) diagonal matrix

$$\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}.$$

More specifically,

$$A = \begin{bmatrix} \\ x_1 & \cdots & x_n \\ \\ \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix} \cdot \begin{bmatrix} \\ x_1 & \cdots & x_n \\ \\ \end{bmatrix}^{-1}.$$

If the eigenvalues of A are real and distinct, then this means that

$$tA = \begin{bmatrix} \\ x_1 & \cdots & x_n \\ \\ \end{bmatrix} \cdot \begin{bmatrix} t\lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & t\lambda_n \end{bmatrix} \cdot \begin{bmatrix} \\ x_1 & \cdots & x_n \\ \\ \end{bmatrix}^{-1},$$

and the formula for the matrix exponential then yields

$$e^{tA} = \begin{bmatrix} \\ x_1 & \cdots & x_n \\ \\ \end{bmatrix} \cdot \begin{bmatrix} e^{t\lambda_1} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & e^{t\lambda_n} \end{bmatrix} \cdot \begin{bmatrix} \\ x_1 & \cdots & x_n \\ \\ \end{bmatrix}^{-1}.$$

This formula should make clear how the projections of $e^{tA}x_0$ grow or decay as $t \to \pm\infty$.

The same sort of analysis works when the eigenvalues are (nontrivially) complex, but the resulting formula is not as enlightening. In addition to the difficulty of a complex change of basis, the behavior of $e^{t\lambda_k}$ is less clear when λ_k is not real.

One way around this is the following. Sort the eigenvalues (and eigenvectors) of A so that complex conjugate eigenvalues $\{\lambda_1, \overline{\lambda}_1, \dots, \lambda_m, \overline{\lambda}_m\}$ come first and are grouped together and so that real eigenvalues $\{\lambda_{m+1}, \dots, \lambda_r\}$ come last. For $k \leq m$, set $a_k = \operatorname{Re} \lambda_k \in \mathbb{R}$, $b_k = \operatorname{Im} \lambda_k \in \mathbb{R}$, $y_k = \operatorname{Re} x_k \in \mathbb{R}^n$, and $z_k = \operatorname{Im} x_k \in \mathbb{R}^n$. Then

$$\begin{aligned} Ay_k &= A \operatorname{Re} x_k = \operatorname{Re} Ax_k = \operatorname{Re} \lambda_k x_k = (\operatorname{Re} \lambda_k)(\operatorname{Re} x_k) - (\operatorname{Im} \lambda_k)(\operatorname{Im} x_k) \\ &= a_k y_k - b_k z_k, \end{aligned}$$

and

$$\begin{aligned} Az_k &= A \operatorname{Im} x_k = \operatorname{Im} Ax_k = \operatorname{Im} \lambda_k x_k = (\operatorname{Im} \lambda_k)(\operatorname{Re} x_k) + (\operatorname{Re} \lambda_k)(\operatorname{Im} x_k) \\ &= b_k y_k + a_k z_k. \end{aligned}$$

Using these facts, we have $A = QPQ^{-1}$, where

$$Q = \begin{bmatrix} \\ z_1 & y_1 & \cdots & \cdots & z_m & y_m & x_{m+1} & \cdots & x_r \\ \\ \end{bmatrix}$$

and

$$P = \left[\begin{array}{cc|cc|cc|cc||cccc}
a_1 & -b_1 & 0 & 0 & \cdots & \cdots & 0 & 0 & 0 & \ldots & \cdots & 0 \\
b_1 & a_1 & 0 & 0 & \cdots & \cdots & 0 & 0 & 0 & \ldots & \cdots & 0 \\
\hline
0 & 0 & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\hline
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & 0 & 0 & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & 0 & 0 & \vdots & \vdots & \vdots & \vdots \\
\hline
0 & 0 & \cdots & \cdots & 0 & 0 & a_m & -b_m & 0 & \cdots & \cdots & 0 \\
0 & 0 & \cdots & \cdots & 0 & 0 & b_m & a_m & 0 & \cdots & \cdots & 0 \\
\hline\hline
0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 & 0 & \lambda_{m+1} & 0 & \cdots & 0 \\
\vdots & \vdots & \cdots & \cdots & \cdots & \cdots & \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\
\vdots & \vdots & \cdots & \cdots & \cdots & \cdots & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\
0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 & 0 & 0 & \cdots & 0 & \lambda_r
\end{array}\right].$$

In order to compute e^{tA} from this formula, we'll need to know how to compute e^{tA_k}, where

$$A_k = \begin{bmatrix} a_k & -b_k \\ b_k & a_k \end{bmatrix}.$$

This can be done using the power series formula. An alternative approach is to realize that

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} := e^{tA_k} \begin{bmatrix} c \\ d \end{bmatrix}$$

is supposed to solve the IVP

$$\begin{cases} \dot{x} = a_k x - b_k y \\ \dot{y} = b_k x + a_k y \\ x(0) = c \\ y(0) = d. \end{cases} \tag{26}$$

Since we can check that the solution of (26) is

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} e^{a_k t}(c \cos b_k t - d \sin b_k t) \\ e^{a_k t}(d \cos b_k t + c \sin b_k t) \end{bmatrix},$$

we can conclude that

$$e^{tA_k} = \begin{bmatrix} e^{a_k t}\cos b_k t & -e^{a_k t}\sin b_k t \\ e^{a_k t}\sin b_k t & e^{a_k t}\cos b_k t \end{bmatrix}$$

Putting this all together and using the form of P, we see that $e^{tA} = Qe^{tP}Q^{-1}$, where

$$e^{tP} = \left[\begin{array}{c||c} \mathcal{B}_1 & \mathbf{0} \\ \hline\hline \mathbf{0}^T & \mathcal{B}_2 \end{array}\right],$$

$\mathcal{B}_1 =$

$$\left[\begin{array}{cc|cc|cc|cc} e^{a_1 t}\cos b_1 t & -e^{a_1 t}\sin b_1 t & 0 & 0 & \cdots & \cdots & 0 & 0 \\ e^{a_1 t}\sin b_1 t & e^{a_1 t}\cos b_1 t & 0 & 0 & \cdots & \cdots & 0 & 0 \\ \hline 0 & 0 & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \hline \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & 0 & 0 \\ \hline 0 & 0 & \cdots & \cdots & 0 & 0 & e^{a_m t}\cos b_m t & -e^{a_m t}\sin b_m t \\ 0 & 0 & \cdots & \cdots & 0 & 0 & e^{a_m t}\sin b_m t & e^{a_m t}\cos b_m t \end{array}\right],$$

$\mathcal{B}_2 =$

$$\begin{bmatrix} e^{\lambda_{m+1} t} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & e^{\lambda_r t} \end{bmatrix},$$

and $\mathbf{0}$ is a $2m \times (r - m - 1)$ block of 0's.

This representation of e^{tA} shows that not only may the projections of $e^{tA}x_0$ grow or decay exponentially, they may also exhibit sinusoidal oscillatory behavior.

Generalized Eigenspace Decomposition

Lecture 9
Math 634
9/20/99

Eigenvalues don't have to be distinct for the analysis of the matrix exponential that was done last time to work. There just needs to be a basis of eigenvectors for $\mathbb{R}^n$ (or $\mathbb{C}^n$). Unfortunately, we don't always have such a basis. For this reason, we need to generalize the notion of an eigenvector.

First, some definitions:

Definition The *algebraic multiplicity* of an eigenvalue λ of an operator A is the multiplicity of λ as a zero of the characteristic polynomial $\det(A - xI)$.

Definition The *geometric multiplicity* of an eigenvalue λ of an operator A is the dimension of the corresponding eigenspace, *i.e.*, the dimension of the space of all the eigenvectors of A corresponding to λ.

It is not hard to show (*e.g.*, through a change-of-basis argument) that the geometric multiplicity of an eigenvalue is always less than or equal to its algebraic multiplicity.

Definition A *generalized eigenvector* of A is a vector x such that $(A-\lambda I)^k x = 0$ for some scalar λ and some $k \in \mathbb{N}$.

Definition If λ is an eigenvalue of A, then the *generalized eigenspace of A belonging to* λ is the space of all generalized eigenvectors of A corresponding to λ.

Definition We say that a vector space $\mathcal{V}$ is the *direct sum* of subspaces $\mathcal{V}_1, \dots, \mathcal{V}_m$ of $\mathcal{V}$ and write

$$\mathcal{V} = \mathcal{V}_1 \oplus \cdots \oplus \mathcal{V}_m$$

if for each $v \in \mathcal{V}$ there is a unique $(v_1, \dots, v_m) \in \mathcal{V}_1 \times \cdots \times \mathcal{V}_m$ such that $v = v_1 + \cdots + v_m$.

Theorem (Primary Decomposition Theorem) *Let B be an operator on $\mathcal{E}$, where $\mathcal{E}$ is a complex vector space, or else $\mathcal{E}$ is real and B has real eigenvalues. Then*

$\mathcal{E}$ is the direct sum of the generalized eigenspaces of B. The dimension of each generalized eigenspace is the algebraic multiplicity of the corresponding eigenvalue.

Before proving this theorem, we introduce some notation and state and prove two lemmas.

Given $T : \mathcal{V} \to \mathcal{V}$, let

$$N(T) = \left\{ x \in \mathcal{V} \;\middle|\; T^k x = 0 \text{ for some } k > 0 \right\},$$

and let

$$R(T) = \left\{ x \in \mathcal{V} \;\middle|\; T^k u = x \text{ has a solution } u \text{ for every } k > 0 \right\}.$$

Note that $N(T)$ is the union of the null spaces of the positive powers of T and $R(T)$ is the intersection of the ranges of the positive powers of T. This union and intersection are each nested, and that implies that there is is a number $m \in \mathbb{N}$ such that $R(T)$ is the range of T^m and $N(T)$ is the nullspace of T^m.

Lemma *If $T : \mathcal{V} \to \mathcal{V}$, then $\mathcal{V} = N(T) \oplus R(T)$.*

Proof. Pick m such that $R(T)$ is the range of T^m and $N(T)$ is the nullspace of T^m. Note that $T|_{R(T)} : R(T) \to R(T)$ is invertible. Given $x \in \mathcal{V}$, let $y = \left(T|_{R(T)} \right)^{-m} T^m x$ and $z = x - y$. Clearly, $x = y + z$, $y \in R(T)$, and $T^m z = T^m x - T^m y = 0$, so $z \in N(T)$. If $x = \tilde{y} + \tilde{z}$ for some other $\tilde{y} \in R(T)$ and $\tilde{z} \in N(T)$ then $T^m \tilde{y} = T^m x - T^m \tilde{z} = T^m x$, so $\tilde{y} = y$ and $\tilde{z} = z$. □

Lemma *If λ_j, λ_k are distinct eigenvalues of $T : \mathcal{V} \to \mathcal{V}$, then*

$$N(T - \lambda_j I) \subseteq R(T - \lambda_k I).$$

Proof. Note first that $(T - \lambda_k I) N(T - \lambda_j I) \subseteq N(T - \lambda_j I)$. We claim that, in fact, $(T - \lambda_k I) N(T - \lambda_j I) = N(T - \lambda_j I)$; *i.e.*, that

$$(T - \lambda_k I)|_{N(T - \lambda_j I)} : N(T - \lambda_j I) \to N(T - \lambda_j I)$$

is invertible. Suppose it isn't; then we can pick a nonzero $x \in N(T - \lambda_j I)$ such that $(T - \lambda_k I)x = 0$. But if $x \in N(T - \lambda_j I)$ then $(T - \lambda_j I)^{m_j} x = 0$ for some $m_j \geq 0$. Calculating,

$$\begin{aligned}
(T - \lambda_j I)x &= Tx - \lambda_j x = \lambda_k x - \lambda_j x = (\lambda_k - \lambda_j)x, \\
(T - \lambda_j I)^2 x &= T(\lambda_k - \lambda_j)x - \lambda_j(\lambda_k - \lambda_j)x = (\lambda_k - \lambda_j)^2 x, \\
&\vdots \\
(T - \lambda_j I)^{m_j} x &= \cdots = (\lambda_k - \lambda_j)^{m_j} x \neq 0,
\end{aligned}$$

contrary to assumption. Hence, the claim holds.

Note that this implies not only that

$$(T - \lambda_k I)N(T - \lambda_j I) = N(T - \lambda_j I)$$

but also that

$$(T - \lambda_k I)^m N(T - \lambda_j I) = N(T - \lambda_j I)$$

for every $m \in \mathbb{N}$. This means that

$$N(T - \lambda_j I) \subseteq R(T - \lambda_k I).$$

□

Proof of the Principal Decomposition Theorem. It is obviously true if the dimension of $\mathcal{E}$ is 0 or 1. We prove it for $\dim \mathcal{E} > 1$ by induction on $\dim \mathcal{E}$. Suppose it holds on all spaces of smaller dimension than $\mathcal{E}$. Let $\lambda_1, \lambda_2, \ldots, \lambda_q$ be the eigenvalues of B with algebraic multiplicities $n_1, n_2, \ldots, n_q$. By the first lemma,

$$\mathcal{E} = N(B - \lambda_q I) \oplus R(B - \lambda_q I).$$

Note that $\dim R(B - \lambda_q I) < \dim \mathcal{E}$, and $R(B - \lambda_q I)$ is (positively) invariant under B. Applying our assumption to $B|_{R(B-\lambda_q I)} : R(B-\lambda_q I) \to R(B-\lambda_q I)$, we get a decomposition of $R(B - \lambda_q I)$ into the generalized eigenspaces of $B|_{R(B-\lambda_q I)}$. By the second lemma, these are just

$$N(B - \lambda_1 I), N(B - \lambda_2 I), \ldots, N(B - \lambda_{q-1} I),$$

so

$$\mathcal{E} = N(B - \lambda_1 I) \oplus N(B - \lambda_2 I) \oplus \cdots \oplus N(B - \lambda_{q-1} I) \oplus N(B - \lambda_q I).$$

Now, by the second lemma, we know that $B|_{N(B-\lambda_k I)}$ has λ_k as its only eigenvalue, so $\dim N(B - \lambda_k I) \leq n_k$. Since

$$\sum_{k=1}^{q} n_k = \dim E = \sum_{k=1}^{q} \dim N(B - \lambda_k I),$$

we actually have $\dim N(B - \lambda_k I) = n_k$. □

Operators on Generalized Eigenspaces

Lecture 10
Math 634
9/22/99

We've seen that the space on which a linear operator acts can be decomposed into the direct sum of generalized eigenspaces of that operator. The operator maps each of these generalized eigenspaces into itself, and, consequently, solutions of the differential equation starting in a generalized eigenspace stay in that generalized eigenspace for all time. Now we will see how the solutions within such a subspace behave by seeing how the operator behaves on this subspace.

It may seem like nothing much can be said in general since, given a finite-dimensional vector space $\mathcal{V}$, we can define a nilpotent operator S on $\mathcal{V}$ by

1. picking a basis $\{v_1, \dots, v_m\}$ for $\mathcal{V}$;
2. creating a graph by connecting the nodes $\{v_1, \dots, v_m, 0\}$ with directed edges in such a way that from each node there is a unique directed path to 0;
3. defining $S(v_j)$ to be the unique node v_k such that there is a directed edge from v_j to v_k;
4. extending S linearly to all of $\mathcal{V}$.

By adding any multiple of I to S we have an operator for which $\mathcal{V}$ is a generalized eigenspace. It turns out, however, that there are really only a small number of different possible structures that may arise from this seemingly general process.

To make this more precise, we first need a definition, some new notation, and a lemma.

Definition A subspace $\mathcal{Z}$ of a vector space $\mathcal{V}$ is a *cyclic subspace of* S *on* $\mathcal{V}$ if $S\mathcal{Z} \subseteq \mathcal{Z}$ and there is some $x \in \mathcal{Z}$ such that $\mathcal{Z}$ is spanned by $\{x, Sx, S^2x, \dots\}$.

Given S, note that every vector $x \in \mathcal{V}$ generates a cyclic subspace. Call it $\mathcal{Z}(x)$ or $\mathcal{Z}(x, S)$. If S is nilpotent, write $\operatorname{nil}(x)$ or $\operatorname{nil}(x, S)$ for the smallest nonnegative integer k such that $S^k x = 0$.

Lemma *The set* $\{x, Sx, \ldots, S^{\mathrm{nil}(x)-1}x\}$ *is a basis for* $\mathcal{Z}(x)$.

Proof. Obviously these vectors span $\mathcal{Z}(x)$; the question is whether they are linearly independent. If they were not, we could write down a linear combination $\alpha_1 S^{p_1}x + \cdots + \alpha_k S^{p_k}x$, with $\alpha_j \neq 0$ and $0 \leq p_1 < p_2 < \cdots < p_k \leq \mathrm{nil}(x) - 1$, that added up to zero. Applying $S^{\mathrm{nil}(x)-p_1-1}$ to this linear combination would yield $\alpha_1 S^{\mathrm{nil}(x)-1}x = 0$, contradicting the definition of $\mathrm{nil}(x)$. □

Theorem *If* $S : \mathcal{V} \to \mathcal{V}$ *is nilpotent then* $\mathcal{V}$ *can be written as the direct sum of cyclic subspaces of* S *on* $\mathcal{V}$. *The dimensions of these subspaces are determined by the operator* S.

Proof. The proof is inductive on the dimension of $\mathcal{V}$. It is clearly true if $\dim \mathcal{V} = 0$ or 1. Assume it is true for all operators on spaces of dimension less than $\dim \mathcal{V}$.

Step 1: The dimension of $S\mathcal{V}$ is less than the dimension of $\mathcal{V}$.
If this weren't the case, then S would be invertible and could not possibly be nilpotent.

Step 2: For some $k \in \mathbb{N}$ and for some nonzero $y_j \in S\mathcal{V}$, $j = 1, \ldots, k$,

$$S\mathcal{V} = \mathcal{Z}(y_1) \oplus \cdots \oplus \mathcal{Z}(y_k). \tag{27}$$

This is a consequence of Step 1 and the induction hypothesis.

Pick $x_j \in \mathcal{V}$ such that $Sx_j = y_j$, for $j = 1, \ldots, k$. Suppose that $z_j \in \mathcal{Z}(x_j)$ for each j and

$$z_1 + \cdots + z_k = 0. \tag{28}$$

We will show that $z_j = 0$ for each j. This will mean that the direct sum $\mathcal{Z}(x_1) \oplus \cdots \oplus \mathcal{Z}(x_k)$ exists.

Step 3: $Sz_1 + \cdots + Sz_k = 0$.
This follows from applying S to both sides of (28).

Step 4: For each j, $Sz_j \in \mathcal{Z}(y_j)$.
The fact that $z_j \in \mathcal{Z}(x_j)$ implies that

$$z_j = \alpha_0 x_j + \alpha_1 S x_j + \cdots + \alpha_{\mathrm{nil}(x_j)-1} S^{\mathrm{nil}(x_j)-1} x_j \tag{29}$$

for some α_i. Applying S to both sides of (29) gives

$$Sz_j = \alpha_0 y_j + \alpha_1 S y_j + \cdots + \alpha_{\mathrm{nil}(x_j)-2} S^{\mathrm{nil}(x_j)-2} y_j \in \mathcal{Z}(y_j).$$

Step 5: For each j, $Sz_j = 0$.
This is a consequence of Step 3, Step 4, and (27).

Step 6: For each j, $z_j \in \mathcal{Z}(y_j)$.
If

$$z_j = \alpha_0 x_j + \alpha_1 S x_j + \cdots + \alpha_{\mathrm{nil}(x_j)-1} S^{\mathrm{nil}(x_j)-1} x_j$$

then by Step 5

$$0 = Sz_j = \alpha_0 y_j + \alpha_1 S y_j + \cdots + \alpha_{\mathrm{nil}(x_j)-2} S^{\mathrm{nil}(x_j)-2} y_j.$$

Since $\mathrm{nil}(x_j) - 2 = \mathrm{nil}(y_j) - 1$, the vectors in this linear combination are linearly independent; thus, $\alpha_i = 0$ for $i = 0, \ldots, \mathrm{nil}(x_j) - 2$. In particular, $\alpha_0 = 0$, so

$$z_j = \alpha_1 y_j + \cdots + \alpha_{\mathrm{nil}(x_j)-1} S^{\mathrm{nil}(x_j)-2} y_j \in \mathcal{Z}(y_j).$$

Step 7: For each j, $z_j = 0$.
This is a consequence of Step 6, (27), and (28).

We now know that $\mathcal{Z}(x_1) \oplus \cdots \oplus \mathcal{Z}(x_k) =: \tilde{\mathcal{V}}$ exists, but it is not necessarily all of $\mathcal{V}$. Choose a subspace $\mathcal{W}$ of $\mathrm{Null}(S)$ such that $\mathrm{Null}(S) = (\tilde{\mathcal{V}} \cap \mathrm{Null}(S)) \oplus \mathcal{W}$. Choose a basis $\{w_1, \ldots, w_\ell\}$ for $\mathcal{W}$ and note that $\mathcal{W} = \mathcal{Z}(w_1) \oplus \cdots \oplus \mathcal{Z}(w_\ell)$.

Step 8: The direct sum $\mathcal{Z}(x_1) \oplus \cdots \oplus \mathcal{Z}(x_k) \oplus \mathcal{Z}(w_1) \oplus \cdots \oplus \mathcal{Z}(w_\ell)$ exists.
This is a consequence of the fact that the direct sums $\mathcal{Z}(x_1) \oplus \cdots \oplus \mathcal{Z}(x_k)$

and $\mathcal{Z}(w_1) \oplus \cdots \oplus \mathcal{Z}(w_\ell)$ exist and that $\tilde{\mathcal{V}} \cap \mathcal{W} = \{0\}$.

Step 9: $\mathcal{V} = \mathcal{Z}(x_1) \oplus \cdots \oplus \mathcal{Z}(x_k) \oplus \mathcal{Z}(w_1) \oplus \cdots \oplus \mathcal{Z}(w_\ell)$.
Let $x \in \mathcal{V}$ be given. Recall that $Sx \in S\mathcal{V} = \mathcal{Z}(y_1) \oplus \cdots \oplus \mathcal{Z}(y_k)$. Write $Sx = s_1 + \cdots + s_k$ with $s_j \in \mathcal{Z}(y_j)$. If

$$s_j = \alpha_0 y_j + \alpha_1 S y_j + \cdots + \alpha_{\text{nil}(y_j)-1} S^{\text{nil}(y_j)-1} y_j,$$

let

$$u_j = \alpha_0 x_j + \alpha_1 S x_j + \cdots + \alpha_{\text{nil}(y_j)-1} S^{\text{nil}(y_j)-1} x_j,$$

and note that $Su_j = s_j$ and that $u_j \in \mathcal{Z}(x_j)$. Setting $u = u_1 + \cdots + u_k$, we have

$$S(x-u) = Sx - Su = (s_1 + \cdots + s_k) - (s_1 + \cdots + s_k) = 0,$$

so $x - u \in \text{Null}(S)$. By definition of $\mathcal{W}$, that means that

$$x - u \in \mathcal{Z}(x_1) \oplus \cdots \oplus \mathcal{Z}(x_k) \oplus \mathcal{Z}(w_1) \oplus \cdots \oplus \mathcal{Z}(w_\ell).$$

Since $u \in \mathcal{Z}(x_1) \oplus \cdots \oplus \mathcal{Z}(x_k)$, we have

$$x \in \mathcal{Z}(x_1) \oplus \cdots \oplus \mathcal{Z}(x_k) \oplus \mathcal{Z}(w_1) \oplus \cdots \oplus \mathcal{Z}(w_\ell).$$

This completes the proof of the first sentence in the theorem. The second sentence follows similarly by induction. □

Real Canonical Form

Lecture 11
Math 634
9/24/99

Real Canonical Form

We now use the information contained in the previous theorems to find simple matrices representing linear operators. Clearly, a nilpotent operator S on a cyclic space $\mathcal{Z}(x)$ can be represented by the matrix

$$\begin{bmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ 1 & \ddots & & & \vdots \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix},$$

with the corresponding basis being $\{x, Sx, \dots, S^{\operatorname{nil}(x)-1}x\}$. Thus, an operator T on a generalized eigenspace $N(T-\lambda I)$ can be represented by a matrix of the form

$$\begin{bmatrix} \lambda & 0 & \cdots & \cdots & 0 \\ 1 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & \lambda \end{bmatrix}. \tag{30}$$

If $\lambda = a + bi \in \mathbb{C} \setminus \mathbb{R}$ is an eigenvalue of an operator $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, and $\mathcal{Z}(x, T-\lambda I)$ is one of the cyclic subspaces whose direct sum is $N(T-\lambda I)$, then $\mathcal{Z}(\overline{x}, T-\overline{\lambda}I)$ can be taken to be one of the cyclic subspaces whose direct sum is $N(T-\overline{\lambda}I)$. If we set $k = \operatorname{nil}(x, T-\lambda I)-1$ and $y_j = \operatorname{Re}((T-\lambda I)^j x)$ and $z_j = \operatorname{Im}((T-\lambda I)^j x)$ for $j = 0, \dots, k$, then we have $Ty_j = ay_j - bz_j + y_{j+1}$ and $Tz_j = by_j + az_j + z_{j+1}$ for $j = 0, \dots, k-1$, and $Ty_k = ay_k - bz_k$ and $Tz_k = by_k + az_k$. The $2k+2$ real vectors $\{z_0, y_0, \dots, z_k, y_k\}$ span $Z(x, T-\lambda I) \oplus Z(\overline{x}, T-\overline{\lambda}I)$ over $\mathbb{C}$ and also span a $(2k+2)$-dimensional space over $\mathbb{R}$ that is invariant under T. On this real vector space, the action of T can be represented by the matrix

$$\left[\begin{array}{cc|cc|cc|cc|cc}
a & -b & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 & 0 \\
b & a & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 & 0 \\
\hline
1 & 0 & \ddots & \ddots & \ddots & \ddots & & & \vdots & \vdots \\
0 & 1 & \ddots & \ddots & \ddots & \ddots & & & \vdots & \vdots \\
\hline
0 & 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\
0 & 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\
\hline
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 & 0 \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 & 0 \\
\hline
0 & 0 & \cdots & \cdots & 0 & 0 & 1 & 0 & a & -b \\
0 & 0 & \cdots & \cdots & 0 & 0 & 0 & 1 & b & a
\end{array}\right]. \tag{31}$$

The restriction of an operator to one of its generalized eigenspaces has a matrix representation like

$$\begin{bmatrix}
\begin{bmatrix} \lambda & & \\ 1 & \lambda & \\ & 1 & \lambda \end{bmatrix} & & & & \\
& \begin{bmatrix} \lambda & \\ 1 & \lambda \end{bmatrix} & & & \\
& & \begin{bmatrix} \lambda \end{bmatrix} & & \\
& & & \begin{bmatrix} \lambda \end{bmatrix} & \\
& & & & \ddots
\end{bmatrix} \tag{32}$$

if the eigenvalue λ is real, with blocks of the form (30) running down the diagonal. If the eigenvalue is complex, then the matrix representation is similar to (32) but with blocks of the form (31) instead of the form (30) on the diagonal.

Finally, the matrix representation of the entire operator is block diagonal, with blocks of the form (32) (or its counterpart for complex eigenvalues). This is called the *real canonical form*. If we specify the order in which blocks should appear, then matrices are similar if and only if they have the same real canonical form.

Solving Linear Systems
Lecture 12
Math 634
9/27/99

Exercise 7 Classify all the real canonical forms for operators on $\mathbb{R}^4$. In other words, find a collection of 4×4 matrices, possibly with (real) variable entries and possibly with constraints on those variables, such that

1. Only matrices in real canonical form match one of the matrices in your collection.
2. Each operator on $\mathbb{R}^4$ has a matrix representation matching one of the matrices in your collection.
3. No matrix matching one of your matrices is similar to a matrix matching one of your other matrices.

For example, a suitable collection of matrices for operators on $\mathbb{R}^2$ would be:

$$\begin{bmatrix}\lambda & 0\\ 1 & \lambda\end{bmatrix}; \qquad \begin{bmatrix}\lambda & 0\\ 0 & \mu\end{bmatrix}; \qquad \begin{bmatrix}a & -b\\ b & a\end{bmatrix}, \quad (b \neq 0).$$

Computing e^{tA}

Given an operator $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, let M be its real canonical form. Write $M = S + N$, where S has M's diagonal elements λ_k and diagonal blocks

$$\begin{bmatrix}a & -b\\ b & a\end{bmatrix}$$

and 0's else, and N has M's off-diagonal 1's and 2×2 identity matrices. If you consider the restrictions of S and N to each of the cyclic subspaces of $A-\lambda I$ into which the generalized eigenspace $N(A-\lambda I)$ of A is decomposed, you'll probably be able to see that these restrictions commute. As a consequence of this fact (and the way $\mathbb{R}^n$ can be represented in terms of these cyclic subspaces), S and N commute. Thus $e^{tM} = e^{tS}e^{tN}$.

Now, e^{tS} has $e^{\lambda_k t}$ where S has λ_k, and has

$$\begin{bmatrix} e^{a_k t}\cos b_k t & -e^{a_k t}\sin b_k t \\ e^{a_k t}\sin b_k t & e^{a_k t}\cos b_k t \end{bmatrix}$$

where S has

$$\begin{bmatrix} a_k & -b_k \\ b_k & a_k \end{bmatrix}.$$

The series definition can be used to compute e^{tN}, since the fact that N is nilpotent implies that the series is actually a finite sum. The entries of e^{tN} will be polynomials in t. For example,

$$\begin{bmatrix} 0 & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix} \mapsto \begin{bmatrix} 1 & & & \\ t & \ddots & & \\ \vdots & \ddots & \ddots & \\ t^m & \cdots & t & 1 \end{bmatrix}$$

and

$$\begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & & & \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & \ddots & & \\ & \ddots & \ddots & \\ & & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \end{bmatrix} \mapsto \begin{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & & & \\ \begin{bmatrix} t & 0 \\ 0 & t \end{bmatrix} & \ddots & & \\ \vdots & \ddots & \ddots & \\ \begin{bmatrix} t^m/m! & 0 \\ 0 & t^m/m! \end{bmatrix} & \cdots & \begin{bmatrix} t & 0 \\ 0 & t \end{bmatrix} & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{bmatrix}.$$

Identifying A with its matrix representation with respect to the standard basis, we have $A = PMP^{-1}$ for some invertible matrix P. Consequently, $e^{tA} = Pe^{tM}P^{-1}$. Thus, the entries of e^{tA} will be linear combinations of polynomials times exponentials or polynomials times exponentials times trigonometric functions.

Exercise 8 Compute e^{tA} (and justify your computations) if

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

2. $A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4 & 4 & 4 & 4 \end{bmatrix}$

Linear Planar Systems

A thorough understanding of constant coefficient linear systems $\dot{x} = Ax$ in the plane is very helpful in understanding systems that are nonlinear and/or higher-dimensional.

There are 3 main categories of real canonical forms for an operator A in $\mathcal{L}(\mathbb{R}^2, \mathbb{R}^2)$:

- $\begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}$
- $\begin{bmatrix} \lambda & 0 \\ 1 & \lambda \end{bmatrix}$
- $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}, \qquad (b \neq 0)$

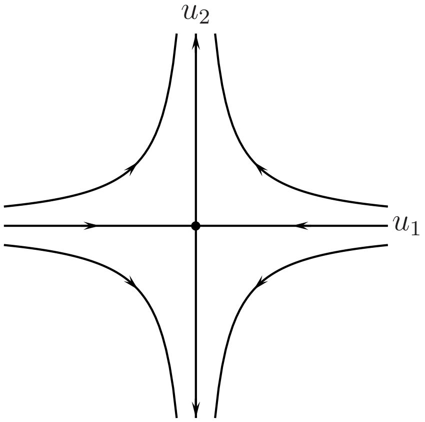
We will subdivide these 3 categories further into a total of 14 categories and consider the corresponding *phase portraits*, *i.e.*, sketches of some of the *trajectories* or parametric curves traced out by solutions in phase space.

$\boxed{\mathbf{1}}$

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}$$

$(\lambda < 0 < \mu)$

saddle

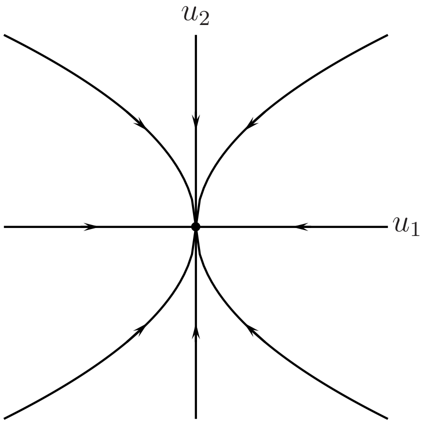


$\boxed{\mathbf{2}}$

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}$$

$(\lambda < \mu < 0)$

stable node

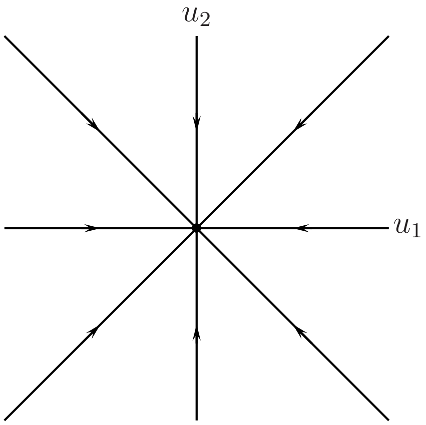


$\boxed{\mathbf{3}}$

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}$$

$(\lambda = \mu < 0)$

stable node

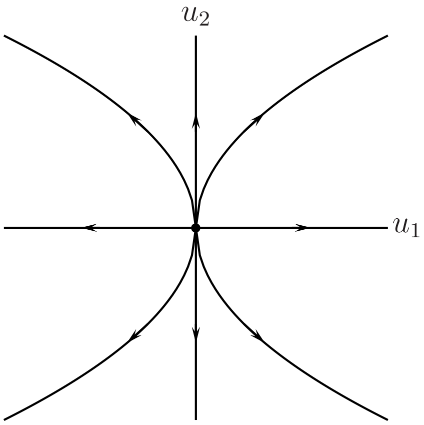


$\boxed{\mathbf{4}}$

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}$$

$(0 < \mu < \lambda)$

unstable node

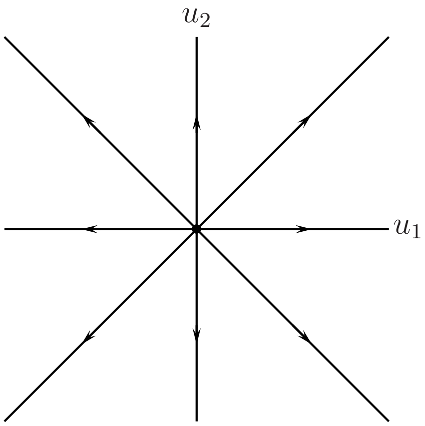


$\boxed{\mathbf{5}}$

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}$$

$(0 < \lambda = \mu)$

unstable node

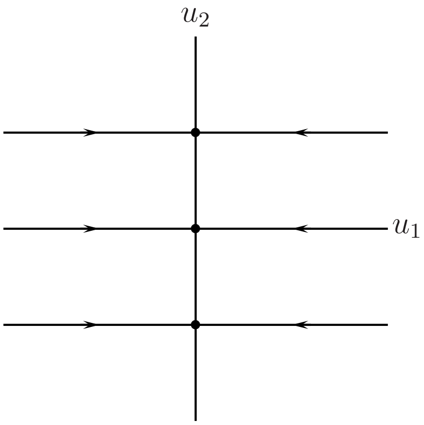


$\boxed{\mathbf{6}}$

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}$$

$(\lambda < \mu = 0)$

degenerate

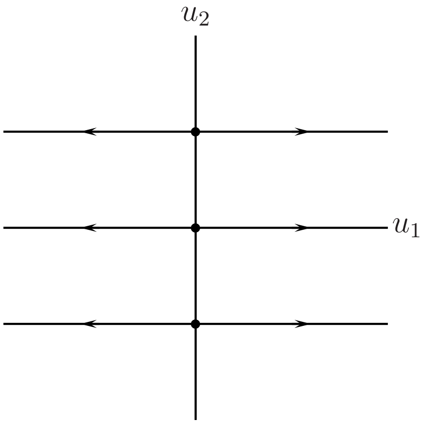


$\boxed{\mathbf{7}}$

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}$$

$(0 = \mu < \lambda)$

degenerate

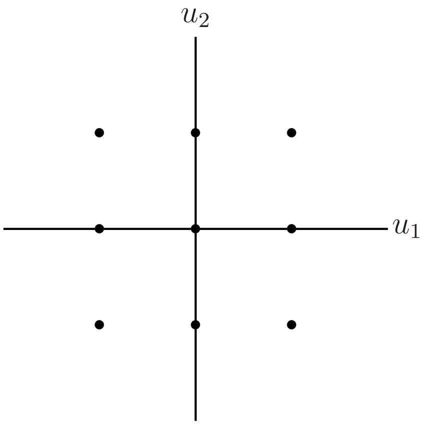


$\boxed{\mathbf{8}}$

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}$$

$(0 = \mu = \lambda)$

degenerate

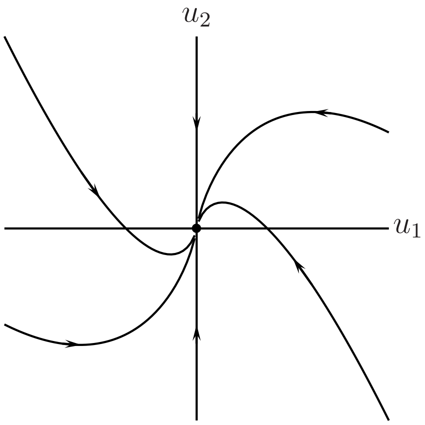


$\boxed{\mathbf{9}}$

$$A = \begin{bmatrix} \lambda & 0 \\ 1 & \lambda \end{bmatrix}$$

$(\lambda < 0)$

stable node

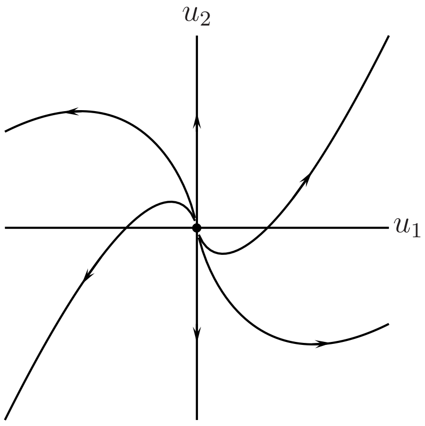


$\boxed{\mathbf{10}}$

$$A = \begin{bmatrix} \lambda & 0 \\ 1 & \lambda \end{bmatrix}$$

$(0 < \lambda)$

unstable node

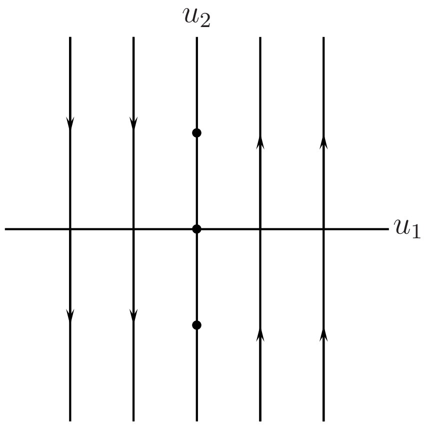


$\boxed{\mathbf{11}}$

$$A = \begin{bmatrix} \lambda & 0 \\ 1 & \lambda \end{bmatrix}$$

$(\lambda = 0)$

degenerate

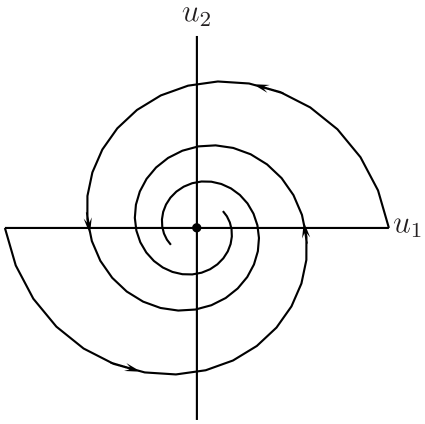


$\boxed{\mathbf{12}}$

$$A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

$(a < 0 < b)$

stable spiral

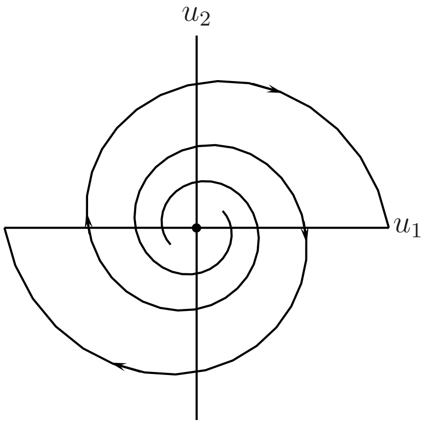


$\boxed{\mathbf{13}}$

$$A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

$(b < 0 < a)$

unstable spiral

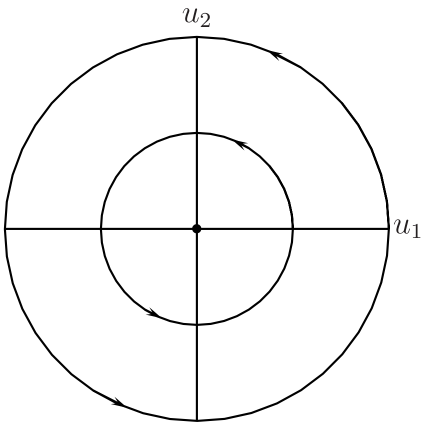


$\boxed{\mathbf{14}}$

$$A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

$(a = 0, b > 0)$

center



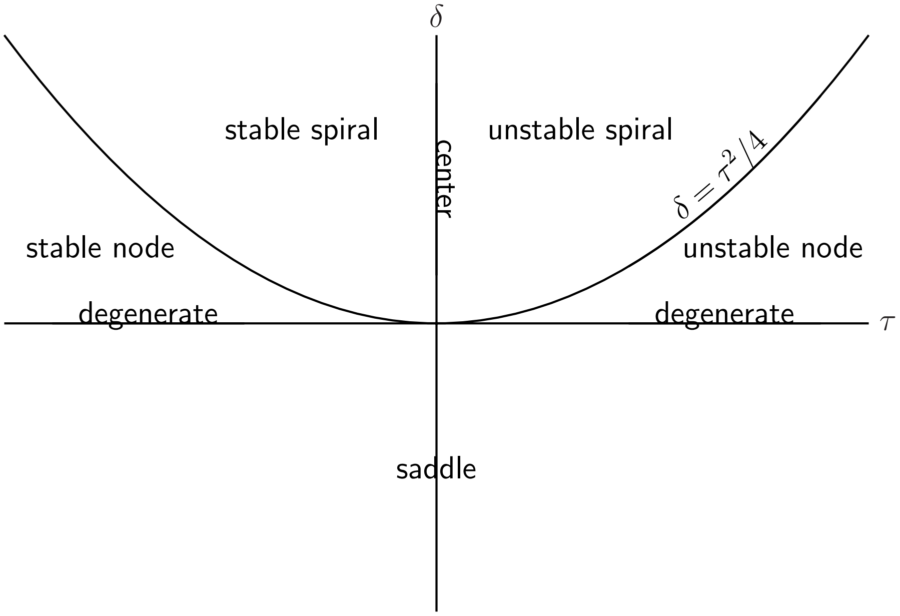
If A is not in real canonical form, then the phase portrait should look similar but may be rotated, flipped, stretched, skewed, etc.

Qualitative Behavior of Linear Systems

Lecture 13
Math 634
9/29/99

Parameter Plane

Some of the information from the preceding phase portraits can be summarized in a parameter diagram. In particular, let $\tau = \operatorname{trace} A$ and let $\delta = \det A$, so the characteristic polynomial is $\lambda^2 - \tau\lambda + \delta$. Then the behavior of the trivial solution $x(t) \equiv 0$ is given by locating the corresponding point in the (τ, δ)-plane:



Growth and Decay Rates

Given $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, let

$$\mathcal{E}^u = \left\{ \bigoplus_{\lambda>0} N(A - \lambda I) \right\} \oplus$$
$$\left\{ \bigoplus_{\substack{\operatorname{Re}\lambda>0 \\ \operatorname{Im}\lambda\neq0}} \{\operatorname{Re} u \mid u \in N(A - \lambda I)\} \right\} \oplus \left\{ \bigoplus_{\substack{\operatorname{Re}\lambda>0 \\ \operatorname{Im}\lambda\neq0}} \{\operatorname{Im} u \mid u \in N(A - \lambda I)\} \right\},$$

$$\mathcal{E}^s = \left\{ \bigoplus_{\lambda<0} N(A - \lambda I) \right\} \oplus$$
$$\left\{ \bigoplus_{\substack{\operatorname{Re}\lambda<0 \\ \operatorname{Im}\lambda\neq0}} \{\operatorname{Re} u \mid u \in N(A - \lambda I)\} \right\} \oplus \left\{ \bigoplus_{\substack{\operatorname{Re}\lambda<0 \\ \operatorname{Im}\lambda\neq0}} \{\operatorname{Im} u \mid u \in N(A - \lambda I)\} \right\},$$

and

$$\mathcal{E}^c = N(A) \oplus$$
$$\left\{ \bigoplus_{\substack{\operatorname{Re}\lambda=0 \\ \operatorname{Im}\lambda\neq0}} \{\operatorname{Re} u \mid u \in N(A - \lambda I)\} \right\} \oplus \left\{ \bigoplus_{\substack{\operatorname{Re}\lambda=0 \\ \operatorname{Im}\lambda\neq0}} \{\operatorname{Im} u \mid u \in N(A - \lambda I)\} \right\}.$$

From our previous study of the real canonical form, we know that

$$\mathbb{R}^n = \mathcal{E}^u \oplus \mathcal{E}^s \oplus \mathcal{E}^c.$$

We call $\mathcal{E}^u$ the *unstable space* of A, $\mathcal{E}^s$ the *stable space* of A, and $\mathcal{E}^c$ the *center space* of A.

Each of these subspaces of $\mathbb{R}^n$ is invariant under the differential equation

$$\dot{x} = Ax. \tag{33}$$

In other words, if $x : \mathbb{R} \to \mathbb{R}^n$ is a solution of (33) and $x(0)$ is in $\mathcal{E}^u$, $\mathcal{E}^s$, or $\mathcal{E}^c$, then $x(t)$ is in $\mathcal{E}^u$, $\mathcal{E}^s$, or $\mathcal{E}^c$, respectively, for all $t \in \mathbb{R}$. We shall see that each of these spaces is characterized by the growth or decay rates of the solutions it contains. Before doing so, we state and prove a basic fact about finite-dimensional normed vector spaces.

Theorem *All norms on $\mathbb{R}^n$ are equivalent.*

Proof. Since equivalence of norms is transitive, it suffices to prove that every norm $N : \mathbb{R}^n \to \mathbb{R}$ is equivalent to the standard Euclidean norm $|\cdot|$.

Given an arbitrary norm N, and letting x_i be the projection of $x \in \mathbb{R}^n$ onto the ith standard basis vector e_i, note that

$$\begin{aligned} N(x) = N\left(\sum_{i=1}^n x_i e_i\right) &\leq \sum_{i=1}^n |x_i| N(e_i) \leq \sum_{i=1}^n |x| N(e_i) \\ &\leq \left(\sum_{i=1}^n N(e_i)\right) |x|. \end{aligned}$$

This shows half of equivalence; it also shows that N is continuous, since, by the triangle inequality,

$$|N(x) - N(y)| \leq N(x-y) \leq \left(\sum_{i=1}^n N(e_i)\right) |x-y|.$$

The set $\mathcal{S} := \{x \in \mathbb{R}^n \mid |x| = 1\}$ is clearly closed and bounded and, therefore, compact, so by the extreme value theorem, N must achieve a minimum on $\mathcal{S}$. Since N is a norm (and is, therefore, positive definite), this minimum must be positive; call it k. Then for any $x \neq 0$,

$$N(x) = N\left(|x|\frac{x}{|x|}\right) = |x| N\left(\frac{x}{|x|}\right) \geq k|x|,$$

and the estimate $N(x) \geq k|x|$ obviously holds if $x = 0$, as well. □

Theorem *Given $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ and the corresponding decomposition $\mathbb{R}^n = \mathcal{E}^u \oplus \mathcal{E}^s \oplus \mathcal{E}^c$, we have*

$$\mathcal{E}^u = \{x \in \mathbb{R}^n \mid \exists c > 0 \ s.t. \lim_{t \downarrow -\infty} |e^{-ct} e^{tA} x| = 0\}, \tag{34}$$

$$\mathcal{E}^s = \{x \in \mathbb{R}^n \mid \exists c > 0 \ s.t. \lim_{t \uparrow \infty} |e^{ct} e^{tA} x| = 0\}, \tag{35}$$

and

$$\mathcal{E}^c = \{x \in \mathbb{R}^n \mid \forall c > 0, \lim_{t \downarrow -\infty} |e^{ct} e^{tA} x| = 0 \ and \ \lim_{t \uparrow \infty} |e^{-ct} e^{tA} x| = 0\}. \tag{36}$$

Proof. By equivalence of norms, instead of using the standard Euclidean norm on $\mathbb{R}^n$ we can use the norm

$$\|x\| := \sup\{|P_1 x|, \dots , |P_n x|\},$$

where $P_i : \mathbb{R}^n \to \mathbb{R}$ represents projection onto the ith basis vector corresponding to the real canonical form. Because of our knowledge of the structure of the real canonical form, we know that $P_i e^{tA} x$ is either of the form

$$p(t)e^{\lambda t}, \tag{37}$$

where $p(t)$ is a polynomial in t and $\lambda \in \mathbb{R}$ is an eigenvalue of A, or of the form

$$p(t)e^{at}(\alpha \cos bt + \beta \sin bt), \tag{38}$$

where $p(t)$ is a polynomial in t, $a + bi \in \mathbb{C} \setminus \mathbb{R}$ is an eigenvalue of A, and α and β are real constants. Furthermore, we know that if P_i corresponds to a vector in $\mathcal{E}^u$ then λ and a are positive, if P_i corresponds to a vector in $\mathcal{E}^s$ then λ and a are negative, and if P_i corresponds to a vector in $\mathcal{E}^c$ then λ and a are zero.

Now, suppose $x \in \mathcal{E}^s$. Then each $P_i e^{tA} x$ is either identically zero or has as a factor a negative exponential whose constant is the real part of an eigenvalue of A that is to the left of the imaginary axis in the complex plane. Let $\sigma(A)$ be the set of eigenvalues of A, and set

$$c = \frac{\bigl|\max\bigl\{\operatorname{Re}\lambda \bigm| \lambda \in \sigma(A) \text{ and } \operatorname{Re}\lambda < 0\bigr\}\bigr|}{2}.$$

Then $e^{ct} P_i e^{tA} x$ is either identically zero or decays exponentially to zero as $t \uparrow \infty$.

Conversely, suppose $x \notin \mathcal{E}^s$. Then $P_i x \neq 0$ for some P_i corresponding to a real canonical basis vector in $\mathcal{E}^u$ or in $\mathcal{E}^c$. In either case, $P_i e^{tA} x$ is not identically zero and is of the form (37) where $\lambda \geq 0$ or of the form (38) where $a \geq 0$. Thus, if $c > 0$ then

$$\limsup_{t \uparrow \infty} |e^{ct} P_i e^{tA} x| = \infty,$$

so

$$\limsup_{t \uparrow \infty} \|e^{ct} e^{tA} x\| = \infty.$$

The preceding two paragraphs showed that (35) is correct. By applying this fact to the time-reversed problem $\dot{x} = -Ax$, we find that (34) is correct, as well. We now consider (36).

If $x \in \mathcal{E}^c$, then for each i, $P_i e^{tA} x$ is either a polynomial or the product of a polynomial and a periodic function. If $c > 0$ and we multiply such a function of t by e^{ct} and let $t \downarrow -\infty$ or we multiply it by e^{-ct} and let $t \uparrow \infty$, then the result converges to zero.

If, on the other hand, $x \notin \mathcal{E}^c$ then for some i, $P_i e^{tA} x$ contains a nontrivial exponential term. If $c > 0$ is sufficiently small then either $e^{ct} P_i e^{tA} x$ diverges as $t \downarrow -\infty$ or $e^{-ct} P_i e^{tA} x$ diverges as $t \uparrow \infty$. This completes the verification of (36). □

Exponential Decay
Lecture 14
Math 634
10/1/99

Definition If $\mathcal{E}^u = \mathbb{R}^n$, we say that the origin is a *source* and e^{tA} is an *expansion.*

Definition If $\mathcal{E}^s = \mathbb{R}^n$, we say that the origin is a *sink* and e^{tA} is a *contraction.*

Theorem

(a) *The origin is a source for the equation $\dot{x} = Ax$ if and only if for a given norm $\|\cdot\|$ there are constants $k, b > 0$ such that*

$$\|e^{tA}x\| \leq ke^{tb}\|x\|$$

for every $t \leq 0$ and $x \in \mathbb{R}^n$.

(b) *The origin is a sink for the equation $\dot{x} = Ax$ if and only if for a given norm $\|\cdot\|$ there are constants $k, b > 0$ such that*

$$\|e^{tA}x\| \leq ke^{-tb}\|x\|$$

for every $t \geq 0$ and $x \in \mathbb{R}^n$.

Proof. The "if" parts are a consequence of the previous theorem. The "only if" parts follow from the proof of the previous theorem. □

Note that a contraction does not always "contract" things immediately; *i.e.*, $|e^{tA}x| \not\leq |x|$, in general. For example, consider

$$A = \begin{bmatrix} -1/4 & 0 \\ 1 & -1/4 \end{bmatrix}.$$

If

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

is a solution of $\dot{x} = Ax$, then

$$\begin{aligned}\frac{d}{dt}|x(t)|^2 = 2\langle x, \dot{x}\rangle = 2\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} -1/4 & 0 \\ 1 & -1/4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= -\frac{1}{2}x_1^2 + 2x_1x_2 - \frac{1}{2}x_2^2 \\ = x_1x_2 - \frac{1}{2}(x_1 - x_2)^2,\end{aligned}$$

which is greater than zero if, for example, $x_1 = x_2 > 0$. However, we have the following:

Theorem

(a) *If* e^{tA} *is an expansion then there is* some *norm* $\|\cdot\|$ *and some constant* $b > 0$ *such that*

$$\|e^{tA}x\| \leq e^{tb}\|x\|$$

for every $t \leq 0$ *and* $x \in \mathbb{R}^n$*.*

(b) *If* e^{tA} *is a contraction then there is* some *norm* $\|\cdot\|$ *and some constant* $b > 0$ *such that*

$$\|e^{tA}x\| \leq e^{-tb}\|x\|$$

for every $t \geq 0$ *and* $x \in \mathbb{R}^n$*.*

Proof. The idea of the proof is to pick a basis with respect to which A is represented by a matrix like the real canonical form but with some small constant $\varepsilon > 0$ in place of the off-diagonal 1's. (This can be done by rescaling.) If the Euclidean norm with respect to this basis is used, the desired estimates hold. The details of the proof may be found in Chapter 7, §1, of Hirsch and Smale. □

Exercise 9

(a) Show that if e^{tA} and e^{tB} are both contractions on $\mathbb{R}^n$, and $BA = AB$, then $e^{t(A+B)}$ is a contraction.

(b) Give a concrete example that shows that **(a)** can fail if the assumption that $AB = BA$ is dropped.

Exercise 10 Problem 5 on page 137 of Hirsch and Smale reads:
"For any solution to $\dot{x} = Ax$, $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, show that exactly one of the following alternatives holds:

(a) $\lim\limits_{t\uparrow\infty} x(t) = 0$ and $\lim\limits_{t\downarrow-\infty} |x(t)| = \infty$;

(b) $\lim\limits_{t\uparrow\infty} |x(t)| = \infty$ and $\lim\limits_{t\downarrow-\infty} x(t) = 0$;

(c) there exist constants $M, N > 0$ such that $M < |x(t)| < N$ for all $t \in \mathbb{R}$."

Is what they ask you to prove true? If so, prove it. If not, determine what other possible alternatives exist, and prove that you have accounted for all possibilities.

Nonautonomous Linear Systems

Lecture 15
Math 634
10/4/99

We now move from the constant coefficient equation $\dot{x} = Ax$ to the nonautonomous equation

$$\dot{x} = A(t)x. \tag{39}$$

For simplicity we will assume that the domain of A is $\mathbb{R}$.

Solution Formulas

In the scalar, or one-dimensional, version of (39)

$$\dot{x} = a(t)x \tag{40}$$

we can separate variables and arrive at the formula

$$x(t) = x_0 e^{\int_{t_0}^{t} a(\tau)\, d\tau}$$

for the solution of (40) that satisfies the initial condition $x(t_0) = x_0$.

It seems like the analogous formula for the solution of (39) with initial condition $x(t_0) = x_0$ should be

$$x(t) = e^{\int_{t_0}^{t} A(\tau)\, d\tau} x_0. \tag{41}$$

Certainly, the right-hand side of (41) makes sense (assuming that A is continuous). But does it give the correct answer?

Let's consider a specific example. Let

$$A(t) = \begin{bmatrix} 0 & 0 \\ 1 & t \end{bmatrix}$$

and $t_0 = 0$. Note that

$$\int_0^t A(\tau)\, d\tau = \begin{bmatrix} 0 & 0 \\ t & t^2/2 \end{bmatrix} = \frac{t^2}{2} \begin{bmatrix} 0 & 0 \\ 2/t & 1 \end{bmatrix},$$

and

$$e^{\begin{bmatrix}0 & 0\\ t & t^2/2\end{bmatrix}} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} + \frac{t^2}{2}\begin{bmatrix}0 & 0\\ 2/t & 1\end{bmatrix} + \frac{\left(\frac{t^2}{2}\right)^2}{2!}\begin{bmatrix}0 & 0\\ 2/t & 1\end{bmatrix} + \frac{\left(\frac{t^2}{2}\right)^3}{3!}\begin{bmatrix}0 & 0\\ 2/t & 1\end{bmatrix} + \cdots$$
$$= \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} + \left(e^{t^2/2} - 1\right)\begin{bmatrix}0 & 0\\ 2/t & 1\end{bmatrix} = \begin{bmatrix}1 & 0\\ \frac{2}{t}\left(e^{t^2/2}-1\right) & e^{t^2/2}\end{bmatrix}.$$

On the other hand, we can solve the corresponding system

$$\begin{aligned}\dot{x}_1 &= 0\\ \dot{x}_2 &= x_1 + tx_2\end{aligned}$$

directly. Clearly $x_1(t) = \alpha$ for some constant α. Plugging this into the equation for x_2, we have a first-order scalar equation which can be solved by finding an integrating factor. This yields

$$x_2(t) = \beta e^{t^2/2} + \alpha e^{t^2/2}\int_0^t e^{-s^2/2}\,ds$$

for some constant β. Since $x_1(0) = \alpha$ and $x_2(0) = \beta$, the solution of (39) is

$$\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix} = \begin{bmatrix}1 & 0\\ e^{t^2/2}\int_0^t e^{-s^2/2}\,ds & e^{t^2/2}\end{bmatrix}\begin{bmatrix}x_1(0)\\ x_2(0)\end{bmatrix}.$$

Since

$$e^{t^2/2}\int_0^t e^{-s^2/2}\,ds \neq \frac{2}{t}\left(e^{t^2/2}-1\right)$$

(41) doesn't work? What went wrong? The answer is that

$$\frac{d}{dt}e^{\int_0^t A(\tau)\,d\tau} = \lim_{h\to 0}\frac{e^{\int_0^{t+h} A(\tau)\,d\tau} - e^{\int_0^t A(\tau)\,d\tau}}{h} \neq \lim_{h\to 0}\frac{e^{\int_0^t A(\tau)\,d\tau}\left[e^{\int_t^{t+h} A(\tau)\,d\tau} - I\right]}{h},$$

in general, because of possible noncommutativity.

Structure of Solution Set

We abandon attempts to find a general formula for solving (39), and instead analyze the general structure of the solution set.

Definition If $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ are linearly independent solutions of (39) (*i.e.*, no nontrivial linear combination gives the zero function) then the matrix

$$X(t) := \begin{bmatrix} x^{(1)}(t) & \cdots & x^{(n)}(t) \end{bmatrix}$$

is called a *fundamental matrix* for (39).

Theorem *The dimension of the vector space of solutions of* (39) *is n.*

Proof. Pick n linearly independent vectors $v^{(k)} \in \mathbb{R}^n$, $k = 1, \dots, n$, and let $x^{(k)}$ be the solution of (39) that satisfies the initial condition $x^{(k)}(0) = v^{(k)}$. Then these n solutions are linearly independent. Furthermore, we claim that any solution x of (39) is a linear combination of these n solutions. To see why this is so, note that $x(0)$ must be expressible as a linear combination of $\{v^{(1)}, \dots, v^{(n)}\}$. The corresponding linear combination of $\{x^{(1)}, \dots, x^{(n)}\}$ is, by linearity, a solution of (39) that agrees with x at $t = 0$. Since A is continuous, the Picard-Lindelöf Theorem applies to (39) to tell us that solutions of IVPs are unique, so this linear combination of $\{x^{(1)}, \dots, x^{(n)}\}$ must be identical to x. □

Definition If $X(t)$ is a fundamental matrix and $X(0) = I$, then it is called the *principal fundamental matrix*. (Uniqueness of solutions implies that there is only one such matrix.)

Definition Given n functions (in some order) from $\mathbb{R}$ to $\mathbb{R}^n$, their *Wronskian* is the determinant of the matrix that has these functions as its columns (in the corresponding order).

Theorem *The Wronskian of n solutions of* (39) *is identically zero if and only if the solutions are linearly dependent.*

Proof. Suppose $x^{(1)}, \dots, x^{(n)}$ are linearly dependent solutions; *i.e.*,

$$\sum_{k=1}^{n} \alpha_k x^{(k)} = 0$$

for some constants $\alpha_1, \dots, \alpha_n$ with $\sum_{k=1}^{n} \alpha_k^2 \neq 0$. Then $\sum_{k=1}^{n} \alpha_k x^{(k)}(t) = 0$ for every t, so the columns of the Wronskian $W(t)$ are linearly dependent for every t. This means $W \equiv 0$.

Conversely, suppose that the Wronskian W of n solutions $x^{(1)}, \dots, x^{(n)}$ is identically zero. In particular, $W(0) = 0$, so $x^{(1)}(0), \dots, x^{(n)}(0)$ are linearly dependent vectors. Pick constants $\alpha_1, \dots, \alpha_n$, with $\sum_{k=1}^n \alpha_k^2 \neq 0$, such that $\sum_{k=1}^n \alpha_k x^{(k)}(0) = 0$. The function $\sum_{k=1}^n \alpha_k x^{(k)}$ is a solution of (39) that is 0 when $t = 0$, but so is the function that is identically zero. By uniqueness of solutions, $\sum_{k=1}^n \alpha_k x^{(k)} = 0$; *i.e.*, $x^{(1)}, \dots, x^{(n)}$ are linearly dependent. □

Note that this proof also shows that if the Wronskian of n solutions of (39) is zero for some t, then it is zero for all t.

What if we're dealing with n arbitrary vector-valued functions (that are not necessarily solutions of (39))? If they are linearly dependent then their Wronskian is identically zero, but the converse is not true. For example,

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} t \\ 0 \end{bmatrix}$$

have a Wronskian that is identically zero, but they are not linearly dependent. Also, n functions can have a Wronskian that is zero for some t and nonzero for other t. Consider, for example,

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 \\ t \end{bmatrix}.$$

Initial-Value Problems

Given a fundamental matrix $X(t)$ for (39), let $G(t, t_0) := X(t)[X(t_0)]^{-1}$. We claim that $x(t) := G(t, t_0)v$ solves the IVP

$$\begin{cases} \dot{x} = A(t)x \\ x(t_0) = v. \end{cases}$$

To verify this, note that

$$\frac{d}{dt}x = \frac{d}{dt}(X(t)[X(t_0)]^{-1}v) = A(t)X(t)[X(t_0)]^{-1}v = A(t)x,$$

and

$$x(t_0) = G(t_0, t_0)v = X(t_0)[X(t_0)]^{-1}v = v.$$

Inhomogeneous Equations

Consider the IVP

$$\begin{cases} \dot{x} = A(t)x + f(t) \\ x(t_0) = x_0. \end{cases} \tag{42}$$

In light of the results from the previous section when f was identically zero, it's reasonable to look for a solution x of (42) of the form $x(t) = G(t, t_0)y(t)$, where G is as before, and y is some vector-valued function.

Note that

$$\dot{x}(t) = A(t)X(t)[X(t_0)]^{-1}y(t) + G(t, t_0)\dot{y}(t) = A(t)x(t) + G(t, t_0)\dot{y}(t);$$

therefore, we need $G(t, t_0)\dot{y}(t) = f(t)$. Isolating, $\dot{y}(t)$, we need

$$\dot{y}(t) = X(t_0)[X(t)]^{-1}f(t) = G(t_0, t)f(t). \tag{43}$$

Integrating both sides of (43), we see that y should satisfy

$$y(t) - y(t_0) = \int_{t_0}^{t} G(t_0, s)f(s)\,ds.$$

If $x(t_0)$ is to be x_0, then, since $G(t_0, t_0) = I$, we need $y(t_0) = x_0$, so $y(t)$ should be

$$x_0 + \int_{t_0}^{t} G(t_0, s)f(s)\,ds,$$

or, equivalently, $x(t)$ should be

$$G(t, t_0)x_0 + \int_{t_0}^{t} G(t, s)f(s)\,ds,$$

since $G(t, t_0)G(t_0, s) = G(t, s)$. This is called the Variation of Constants formula or the Variation of Parameters formula.

Nearly Autonomous Linear Systems
Lecture 16
Math 634
10/6/99

Suppose $A(t)$ is, in some sense, close to a constant matrix A. The question we wish to address in this section is the extent to which solutions of the nonautonomous system

$$\dot{x} = A(t)x \tag{44}$$

behave like solutions of the autonomous system

$$\dot{x} = Ax. \tag{45}$$

Before getting to our main results, we present a pair of lemmas.

Lemma *The following are equivalent:*

1. *Each solution of* (45) *is bounded as $t \uparrow \infty$.*
2. *The function $t \mapsto \|e^{tA}\|$ is bounded as $t \uparrow \infty$ (where $\|\cdot\|$ is the usual operator norm).*
3. $\operatorname{Re}\lambda \leq 0$ *for every eigenvalue λ of A and the algebraic multiplicity of each imaginary eigenvalue matches its geometric multiplicity.*

Proof. That statement 2 implies statement 1 is a consequence of the definition of the operator norm, since, for each solution x of (45),

$$|x(t)| = |e^{tA}x(0)| \leq \|e^{tA}\| \cdot |x(0)|.$$

That statement 1 implies statement 3, and statement 3 implies statement 2 are consequences of what we have learned about the real canonical form of A, along with the equivalence of norms on $\mathbb{R}^n$. □

Lemma (Generalized Gronwall Inequality) *Suppose X and Φ are non-negative, continuous, real-valued functions on $[t_0, T]$ for which there is a nonnegative constant C such that*

$$X(t) \leq C + \int_{t_0}^{t} \Phi(s)X(s)\,ds,$$

for every $t \in [t_0, T]$. Then

$$X(t) \le C e^{\int_{t_0}^t \Phi(s)\,ds}.$$

Proof. The proof is very similar to the proof of the standard Gronwall inequality. The details are left to the reader. □

The first main result deals with the case when $A(t)$ converges to A sufficiently quickly as $t \uparrow \infty$.

Theorem *Suppose that each solution of* (45) *remains bounded as $t \uparrow \infty$ and that, for some $t_0 \in \mathbb{R}$,*

$$\int_{t_0}^{\infty} \|A(t) - A\|\,dt < \infty, \tag{46}$$

where $\|\cdot\|$ is the standard operator norm. Then each solution of (44) *remains bounded as $t \uparrow \infty$.*

Proof. Let t_0 be such that (46) holds. Given a solution x of (44), let $f(t) = (A(t) - A)x(t)$, and note that x satisfies the constant-coefficient inhomogeneous problem

$$\dot{x} = Ax + f(t). \tag{47}$$

Since the matrix exponential provides a fundamental matrix solution to constant-coefficient linear systems, applying the variation of constants formula to (47) yields

$$x(t) = e^{(t-t_0)A}x(t_0) + \int_{t_0}^{t} e^{(t-s)A}(A(s) - A)x(s)\,ds. \tag{48}$$

Now, by the first lemma, the boundedness of solutions of (45) in forward time tells us that there is a constant $M > 0$ such that $\|e^{tA}\| \le M$ for every $t \ge t_0$. Taking norms and estimating, gives (for $t \ge t_0$)

$$\begin{aligned} |x(t)| &\le \|e^{(t-t_0)A}\| \cdot |x(t_0)| + \int_{t_0}^{t} \|e^{(t-s)A}\| \cdot \|A(s) - A\| \cdot |x(s)|\,ds \\ &\le M|x(t_0)| + \int_{t_0}^{t} M\|A(s) - A\| \cdot |x(s)|\,ds. \end{aligned}$$

Setting $X(t) = |x(t)|$, $\Phi(t) = M\|A(t) - A\|$, and $C = M|x(t_0)|$, and applying the generalized Gronwall inequality, we find that

$$|x(t)| \leq M|x(t_0)|e^{M\int_{t_0}^{t} \|A(s)-A\|\,ds}.$$

By (46), the right-hand side of this inequality is bounded on $[t_0, \infty)$, so $x(t)$ is bounded as $t \uparrow \infty$. □

The next result deals with the case when the origin is a sink for (45). Will all the solutions of (44) also all converge to the origin as $t \uparrow \infty$? Yes, if $\|A(t) - A\|$ is sufficiently small.

Theorem *Suppose all the eigenvalues of A have negative real part. Then there is a constant $\varepsilon > 0$ such that if $\|A(t) - A\| \leq \varepsilon$ for all t sufficiently large then every solution of* (44) *converges to* 0 *as $t \uparrow \infty$.*

Proof. Since the origin is a sink, we know that we can choose constants $k, b > 0$ such that $\|e^{tA}\| \leq ke^{-bt}$ for all $t \geq 0$. Pick a constant $\varepsilon \in (0, b/k)$, and assume that there is a time $t_0 \in \mathbb{R}$ such that $\|A(t) - A\| \leq \varepsilon$ for every $t \geq t_0$.

Now, given a solution x of (44) we can conclude, as in the proof of the previous theorem, that

$$|x(t)| \leq \|e^{(t-t_0)A}\| \cdot |x(t_0)| + \int_{t_0}^{t} \|e^{(t-s)A}\| \cdot \|A(s) - A\| \cdot |x(s)|\,ds$$

for all $t \geq t_0$. This implies that

$$|x(t)| \leq ke^{-b(t-t_0)}|x(t_0)| + \int_{t_0}^{t} ke^{-b(t-s)}\varepsilon \cdot |x(s)|\,ds$$

for all $t \geq t_0$. Multiplying through by $e^{b(t-t_0)}$ and setting $y(t) := e^{b(t-t_0)}|x(t)|$ yield

$$y(t) \leq k|x(t_0)| + k\varepsilon \int_{t_0}^{t} y(s)\,ds$$

for all $t \geq t_0$. The standard Gronwall inequality applied to this estimate gives

$$y(t) \leq k|x(t_0)|e^{k\varepsilon(t-t_0)}$$

for all $t \geq t_0$, or, equivalently,

$$|x(t)| \leq k|x(t_0)|e^{(k\varepsilon - b)(t-t_0)}$$

for all $t \geq t_0$. Since $\varepsilon < b/k$, this inequality implies that $x(t) \to 0$ as $t \uparrow \infty$. □

Thus, the origin remains a "sink" even when we perturb A by a small time-dependent quantity. Can we perhaps just look at the (possibly, time-dependent) eigenvalues of $A(t)$ itself and conclude, for example, that if all of those eigenvalues have negative real part for all t then all solutions of (44) converge to the origin as $t \uparrow \infty$? The following example of Markus and Yamabe shows that the answer is "No".

Exercise 11 Show that if

$$A(t) = \begin{bmatrix} -1 + \frac{3}{2}\cos^2 t & 1 - \frac{3}{2}\sin t \cos t \\ -1 - \frac{3}{2}\sin t \cos t & -1 + \frac{3}{2}\sin^2 t \end{bmatrix}$$

then the eigenvalues of $A(t)$ both have negative real part for every $t \in \mathbb{R}$, but

$$x(t) := \begin{bmatrix} -\cos t \\ \sin t \end{bmatrix} e^{t/2},$$

which becomes unbounded as $t \to \infty$, is a solution to (44).

Periodic Linear Systems
Lecture 17
Math 634
10/8/99

We now consider

$$\dot{x} = A(t)x \tag{49}$$

when A is a continuous periodic $n \times n$ matrix function of t; *i.e.*, when there is a constant $T > 0$ such that $A(t+T) = A(t)$ for every $t \in \mathbb{R}$. When that condition is satisfied, we say, more precisely, that A is T*-periodic*. If T is the smallest positive number for which this condition holds, we say that T is the *minimal period* of A.

Let A be T-periodic, and let $X(t)$ be a fundamental matrix for (49). Define $\tilde{X} : \mathbb{R} \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ by $\tilde{X}(t) = X(t+T)$. Clearly, the columns of $\tilde{X}$ are linearly independent functions of t. Also,

$$\frac{d}{dt}\tilde{X}(t) = \frac{d}{dt}X(t+T) = X'(t+T) = A(t+t)X(t+T) = A(t)\tilde{X}(t),$$

so $\tilde{X}$ solves the matrix equivalent of (49). Hence, $\tilde{X}$ is a fundamental matrix for (49).

Because the dimension of the solution space of (49) is n, this means that there is a nonsingular (constant) matrix C such that $X(t+T) = X(t)C$ for every $t \in \mathbb{R}$. C is called a *monodromy* matrix.

Lemma *There exists $B \in \mathcal{L}(\mathbb{C}^n, \mathbb{C}^n)$ such that $C = e^{TB}$.*

Proof. Without loss of generality, we assume that $T = 1$, since if it isn't we can just rescale B by a scalar constant. We also assume, without loss of generality, that C is in Jordan canonical form. (If it isn't, then use the fact that $P^{-1}CP = e^B$ implies that $C = e^{PBP^{-1}}$.) Furthermore, because of the way the matrix exponential acts on a block diagonal matrix, it suffices to show that for each $p \times p$ Jordan block

$$\tilde{C} := \begin{bmatrix} \lambda & 0 & \cdots & \cdots & 0 \\ 1 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & \lambda \end{bmatrix},$$

$\tilde{C} = e^{\tilde{B}}$ for some $\tilde{B} \in \mathcal{L}(\mathbb{C}^p, \mathbb{C}^p)$.

Now, an obvious candidate for $\tilde{B}$ is the natural logarithm of $\tilde{C}$, defined in some reasonable way. Since the matrix exponential was defined by a power series, it seems reasonable to use a similar definition for a matrix logarithm. Note that $\tilde{C} = \lambda I + N = \lambda I(I + \lambda^{-1}N)$, where N is nilpotent. (Since C is invertible, we know that all of the eigenvalues λ are nonzero.) We guess

$$\tilde{B} = (\log \lambda)I + \log(I + \lambda^{-1}N), \tag{50}$$

where

$$\log(I + M) := -\sum_{k=1}^{\infty} \frac{(-M)^k}{k},$$

in analogy to the Maclaurin series for $\log(1 + x)$. Since N is nilpotent, this series terminates in our application of it to (50). Direct substitution shows that $e^{\tilde{B}} = \tilde{C}$, as desired. □

The eigenvalues ρ of C are called the *Floquet multipliers* (or characteristic multipliers) of (49). The corresponding numbers λ satisfying $\rho = e^{\lambda T}$ are called the *Floquet exponents* (or characteristic exponents) of (49). Note that the Floquet exponents are only determined up to a multiple of $(2\pi i)/T$. Given B for which $C = e^{TB}$, the exponents can be chosen to be the eigenvalues of B.

Theorem *There exists a T-periodic function $P : \mathbb{R} \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ such that*

$$X(t) = P(t)e^{tB}.$$

Proof. Let $P(t) = X(t)e^{-tB}$. Then

$$\begin{aligned} P(t+T) &= X(t+T)e^{-(t+T)B} = X(t+T)e^{-TB}e^{-tB} = X(t)Ce^{-TB}e^{-tB} \\ &= X(t)e^{TB}e^{-TB}e^{-tB} = X(t)e^{-tB} = P(t). \end{aligned}$$

□

The decomposition of $X(t)$ given in this theorem shows that the behavior of solutions can be broken down into the composition of a part that is periodic in time and a part that is exponential in time. Recall, however, that B may have entries that are not real numbers, so $P(t)$ may be complex, also. If we want to decompose $X(t)$ into a *real* periodic matrix times a matrix of the form e^{tB} where B is *real*, we observe that $X(t+2T) = X(t)C^2$, where C is the same monodromy matrix as before. It can be shown that the *square* of a real matrix can be written as the exponential of a *real* matrix. Write $C^2 = e^{TB}$ with B real, and let $P(t) = X(t)e^{-tB}$ as before. Then, $X(t) = P(t)e^{tB}$ where P is now $2T$-periodic, and everything is real.

The Floquet multipliers and exponents do not depend on the particular fundamental matrix chosen, even though the monodromy matrix does. They depend only on $A(t)$. To see this, let $X(t)$ and $Y(t)$ be fundamental matrices with corresponding monodromy matrices C and D. Because $X(t)$ and $Y(t)$ are fundamental matrices, there is a nonsingular constant matrix S such that $Y(t) = X(t)S$ for all $t \in \mathbb{R}$. In particular, $Y(0) = X(0)S$ and $Y(T) = X(T)S$. Thus,

$$C = [X(0)]^{-1}X(T) = S[Y(0)]^{-1}Y(T)S^{-1} = S[Y(0)]^{-1}Y(0)DS^{-1} = SDS^{-1}.$$

This means that the monodromy matrices are similar and, therefore, have the same eigenvalues.

Interpreting Floquet Multipliers and Exponents

Theorem *If ρ is a Floquet multiplier of* (49) *and λ is a corresponding Floquet exponent, then there is a nontrivial solution x of* (49) *such that $x(t+T) = \rho x(t)$ for every $t \in \mathbb{R}$ and $x(t) = e^{\lambda t}p(t)$ for some T-periodic vector function p.*

Proof. Pick x_0 to be an eigenvector of B corresponding to the eigenvalue λ, where $X(t) = P(t)e^{tB}$ is the decomposition of a fundamental matrix $X(t)$.

Let $x(t) = X(t)x_0$. Then, clearly, x solves (49). The power series formula for the matrix exponential implies that x_0 is an eigenvector of e^{tB} with eigenvalue $e^{\lambda t}$. Hence,

$$x(t) = X(t)x_0 = P(t)e^{tB}x_0 = P(t)e^{\lambda t}x_0 = e^{\lambda t}p(t),$$

where $p(t) = P(t)x_0$. Also,

$$x(t+T) = e^{\lambda T}e^{\lambda t}p(t+T) = \rho e^{\lambda t}p(t) = \rho x(t).$$

□

Time-dependent Change of Variables

Let x solve (49), and let $y(t) = [P(t)]^{-1}x(t)$, where P is as defined previously. Then

$$\frac{d}{dt}[P(t)y(t)] = \frac{d}{dt}x(t) = A(t)x(t) = A(t)P(t)y(t) = A(t)X(t)e^{-tB}y(t).$$

But

$$\begin{aligned}\frac{d}{dt}[P(t)y(t)] &= P'(t)y(t) + P(t)y'(t)\\ &= [X'(t)e^{-tB} - X(t)e^{-tB}B]y(t) + X(t)e^{-tB}y'(t)\\ &= A(t)X(t)e^{-tB}y(t) - X(t)e^{-tB}By(t) + X(t)e^{-tb}y'(t),\end{aligned}$$

so

$$X(t)e^{-tB}y'(t) = X(t)e^{-tB}By(t),$$

which implies that $y'(t) = By(t)$; *i.e.*, y solves a constant coefficient linear equation. Since P is periodic and, therefore, bounded, the growth and decay of x and y are closely related. Furthermore, the growth or decay of y is determined by the eigenvalues of B, *i.e.*, by the Floquet exponents of (49). For example, we have the following results.

Theorem *If all the Floquet exponents of* (49) *have negative real parts then all solutions of* (49) *converge to 0 as* $t \uparrow \infty$.

Theorem *If there is a nontrivial T-periodic solution of* (49) *then there must be a Floquet multiplier of modulus 1.*

Computing Floquet Multipliers and Exponents

Although Floquet multipliers and exponents are determined by $A(t)$, it is not obvious how to calculate them. As a previous exercise illustrated, the eigenvalues of $A(t)$ don't seem to be extremely relevant. The following result helps a little bit.

Theorem *If* (49) *has Floquet multipliers $\rho_1, \dots, \rho_n$ and corresponding Floquet exponents $\lambda_1, \dots, \lambda_n$, then*

$$\rho_1 \cdots \rho_n = \exp\left(\int_0^T \operatorname{trace} A(t)\, dt \right) \tag{51}$$

and

$$\lambda_1 + \cdots + \lambda_n \equiv \frac{1}{T} \int_0^T \operatorname{trace} A(t)\, dt \bmod \frac{2\pi i}{T} \tag{52}$$

Proof. We focus on (51). The formula (52) will follow immediately from (51).

Let $W(t)$ be the determinant of the principal fundamental matrix $X(t)$. Let S_n be the set of permutations of $\{1, 2, \dots, n\}$ and let $\epsilon : S_n \to \{-1, 1\}$ be the parity map. Then

$$W(t) = \sum_{\sigma \in S_n} \epsilon(\sigma) X_{1,\sigma(1)} X_{2,\sigma(2)} \cdots X_{n,\sigma(n)},$$

where $X_{i,j}$ is the (i,j)-th entry of $X(t)$.

Differentiating yields

$$\begin{aligned}
\frac{dW(t)}{dt} &= \sum_{\sigma \in S_n} \epsilon(\sigma) \frac{d}{dt} \left[X_{1,\sigma(1)} X_{2,\sigma(2)} \cdots X_{n,\sigma(n)} \right] \\
&= \sum_{i=1}^{n} \sum_{\sigma \in S_n} \epsilon(\sigma) X_{1,\sigma(1)} \cdots X_{i-1,\sigma(i-1)} \left[\frac{d}{dt} X_{i,\sigma(i)} \right] X_{i+1,\sigma(i+1)} \cdots X_{n,\sigma(n)} \\
&= \sum_{i=1}^{n} \sum_{\sigma \in S_n} \epsilon(\sigma) X_{1,\sigma(1)} \cdots X_{i-1,\sigma(i-1)} \left[\sum_{j=1}^{n} A_{i,j}(t) X_{j,\sigma(i)} \right] X_{i+1,\sigma(i+1)} \cdots X_{n,\sigma(n)} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}(t) \left(\sum_{\sigma \in S_n} \epsilon(\sigma) X_{1,\sigma(1)} \cdots X_{i-1,\sigma(i-1)} X_{j,\sigma(i)} X_{i+1,\sigma(i+1)} \cdots X_{n,\sigma(n)} \right).
\end{aligned}$$

If $i \neq j$, the inner sum is the determinant of the matrix obtained by replacing the ith row of $X(t)$ by its jth row. This new matrix, having two identical rows, must necessarily have determinant 0. Hence,

$$\frac{dW(t)}{dt} = \sum_{i=1}^{n} A_{i,i}(t) \det X(t) = [\operatorname{trace} A(t)] W(t).$$

Thus,

$$W(t) = e^{\int_0^t \operatorname{trace} A(s)\, ds} W(0) = e^{\int_0^t \operatorname{trace} A(s)\, ds}.$$

In particular,

$$\begin{aligned} e^{\int_0^T \operatorname{trace} A(s)\, ds} &= W(T) = \det X(T) = \det(P(T)e^{TB}) = \det(P(0)e^{TB}) \\ &= \det e^{TB} = \det C = \rho_1 \rho_2 \cdots \rho_n. \end{aligned}$$

□

Exercise 12 Consider (49) where

$$A(t) = \begin{bmatrix} \frac{1}{2} - \cos t & b \\ a & \frac{3}{2} + \sin t \end{bmatrix}$$

and a and b are constants. Show that there is a solution of (49) that becomes unbounded as $t \uparrow \infty$.

Invariant Sets and Limit Sets
Lecture 18
Math 634
10/11/99

We will now begin an intensive study of the continuously differentiable autonomous system

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

or, equivalently, of the corresponding dynamical system $\varphi(t, x)$. We will denote the phase space Ω and assume that it is an open (not necessarily proper) subset of $\mathbb{R}^n$.

Orbits

Definition Given $x \in \Omega$, the *(complete) orbit* through x is the set

$$\gamma(x) := \{\varphi(t, x) \mid t \in \mathbb{R}\},$$

the *positive semiorbit* through x is the set

$$\gamma^+(x) := \{\varphi(t, x) \mid t \geq 0\},$$

and the *negative semiorbit* through x is the set

$$\gamma^-(x) := \{\varphi(t, x) \mid t \leq 0\}.$$

Invariant Sets

Definition A set $\mathcal{M} \subseteq \Omega$ is *invariant* under φ if it contains the complete orbit of every point of $\mathcal{M}$. In other words, for every $x \in \mathcal{M}$ and every $t \in \mathbb{R}$, $\varphi(t, x) \in \mathcal{M}$.

Definition A set $\mathcal{M} \subseteq \Omega$ is *positively invariant* under φ if it contains the positive semiorbit of every point of $\mathcal{M}$. In other words, for every $x \in \mathcal{M}$ and every $t \geq 0$, $\varphi(t, x) \in \mathcal{M}$.

Definition A set $\mathcal{M} \subseteq \Omega$ is *negatively invariant* under φ if it contains the negative semiorbit of every point of $\mathcal{M}$. In other words, for every $x \in \mathcal{M}$ and every $t \leq 0$, $\varphi(t, x) \in \mathcal{M}$.

Limit Sets

Definition Given $x \in \Omega$, the *ω-limit set of x*, denoted $\omega(x)$, is the set

$$\begin{aligned}\bigl\{y \in \Omega \bigm| \liminf_{t\uparrow\infty} |\varphi(t,x) - y| = 0\bigr\}\\ = \bigl\{y \in \Omega \bigm| \exists t_1, t_2, \ldots \to \infty \text{ s.t. } \varphi(t_k, x) \to y \text{ as } k \uparrow \infty\bigr\}.\end{aligned}$$

Definition Given $x \in \Omega$, the *α-limit set of x*, denoted $\alpha(x)$, is the set

$$\begin{aligned}\bigl\{y \in \Omega \bigm| \liminf_{t\downarrow-\infty} |\varphi(t,x) - y| = 0\bigr\}\\ = \bigl\{y \in \Omega \bigm| \exists t_1, t_2, \ldots \to -\infty \text{ s.t. } \varphi(t_k, x) \to y \text{ as } k \uparrow \infty\bigr\}.\end{aligned}$$

Lemma *If, for each $\mathcal{A} \in \Omega$, we let $\overline{\mathcal{A}}$ represent the topological closure of $\mathcal{A}$ in Ω, then*

$$\omega(x) = \bigcap_{\tau\in\mathbb{R}} \overline{\gamma^+(\varphi(\tau, x))} \tag{53}$$

and

$$\alpha(x) = \bigcap_{\tau\in\mathbb{R}} \overline{\gamma^-(\varphi(\tau, x))}. \tag{54}$$

Proof. It suffices to prove (53); (54) can then be established by time reversal.

Let $y \in \omega(x)$ be given. Pick a sequence $t_1, t_2, \ldots \to \infty$ such that $\varphi(t_k, x) \to y$ as $k \uparrow \infty$. Let $\tau \in \mathbb{R}$ be given. Pick $K \in \mathbb{N}$ such that $t_k \geq \tau$ for all $k \geq K$. Note that $\varphi(t_k, x) \in \gamma^+(\varphi(\tau, x))$ for all $k \geq K$, so

$$y \in \overline{\gamma^+(\varphi(\tau, x))}.$$

Since this holds for all $\tau \in \mathbb{R}$, we know that

$$y \in \bigcap_{\tau\in\mathbb{R}} \overline{\gamma^+(\varphi(\tau, x))}. \tag{55}$$

Since (55) holds for each $y \in \omega(x)$, we know that

$$\omega(x) \subseteq \bigcap_{\tau\in\mathbb{R}} \overline{\gamma^+(\varphi(\tau, x))}. \tag{56}$$

Now, we prove the inverse inclusion. Let

$$y \in \bigcap_{\tau\in\mathbb{R}} \overline{\gamma^+(\varphi(\tau, x))}$$

be given. This implies, in particular, that

$$y \in \bigcap_{\tau\in\mathbb{N}} \overline{\gamma^+(\varphi(\tau, x))}.$$

For each $k \in \mathbb{N}$, we have

$$y \in \overline{\gamma^+(\varphi(k, x))}$$

so we can pick $z_k \in \gamma^+(\varphi(k, x))$ such that $|z_k - y| < 1/k$. Since $z_k \in \gamma^+(\varphi(k, x))$, we can pick $s_k \geq 0$ such that $z_k = \varphi(s_k, \varphi(k, x))$. If we set $t_k = k + s_k$, we see that $t_k \geq k$, so the sequence $t_1, t_2, \dots$ goes to infinity. Also, since

$$|\varphi(t_k, x) - y| = |\varphi(s_k + k, x) - y| = |\varphi(s_k, \varphi(k, x)) - y| = |z_k - y| < 1/k,$$

we know that $\varphi(t_k, x) \to y$ as $k \uparrow \infty$. Hence, $y \in \omega(x)$. Since this holds for every

$$y \in \bigcap_{\tau\in\mathbb{R}} \overline{\gamma^+(\varphi(\tau, x))},$$

we know that

$$\bigcap_{\tau\in\mathbb{R}} \overline{\gamma^+(\varphi(\tau, x))} \subseteq \omega(x).$$

Combining this with (56) gives (53). □

We now describe some properties of limit sets.

Theorem *Given $x \in \Omega$, $\omega(x)$ and $\alpha(x)$ are closed (relative to Ω) and invariant. If $\gamma^+(x)$ is contained in some compact subset of Ω, then $\omega(x)$ is nonempty, compact, and connected. If $\gamma^-(x)$ is contained in some compact subset of Ω, then $\alpha(x)$ is nonempty, compact, and connected.*

Proof. Again, time-reversal arguments tell us that we only need to prove the statements about $\omega(x)$.

Step 1: $\omega(x)$ is closed.
This is a consequence of the lemma and the fact that the intersection of closed sets is closed.

Step 2: $\omega(x)$ is invariant.
Let $y \in \omega(x)$ and $t \in \mathbb{R}$ be given. Choose a sequence of times (t_k) converging to infinity such that $\varphi(t_k, x) \to y$ as $k \uparrow \infty$. For each $k \in \mathbb{N}$, let $s_k = t_k + t$, and note that (s_k) converges to infinity and

$$\varphi(s_k, x) = \varphi(t_k + t, x) = \varphi(t, \varphi(t_k, x)) \to \varphi(t, y)$$

as $k \uparrow \infty$ (by the continuity of $\varphi(t, \cdot)$). Hence, $\varphi(t, y) \in \omega(x)$. Since $t \in \mathbb{R}$ and $y \in \omega(x)$ were arbitrary, we know that $\omega(x)$ is invariant.

Now, suppose that $\gamma^+(x)$ is contained in a compact subset $\mathcal{K}$ of Ω.

Step 3: $\omega(x)$ is nonempty.
The sequence $\varphi(1, x), \varphi(2, x), \dots$ is contained in $\gamma^+(x) \subseteq \mathcal{K}$, so by the Bolzano-Weierstrass Theorem, some subseqence $\varphi(t_1, x), \varphi(t_2, x), \dots$ converges to a point $y \in \mathcal{K}$. By definition, $y \in \omega(x)$.

Step 4: $\omega(x)$ is compact.
By the Heine-Borel Theorem, $\mathcal{K}$ is closed (relative to $\mathbb{R}^n$), so, by the choice of $\mathcal{K}$, $\omega(x) \subseteq \mathcal{K}$. Since, by Step 1, $\omega(x)$ is closed relative to Ω, it is also closed relative to $\mathcal{K}$. Since $\mathcal{K}$ is compact, this means $\omega(x)$ is closed (relative to $\mathbb{R}^n$). Also, by the Heine-Borel Theorem, K is bounded so its subset $\omega(x)$ is bounded, too. Thus, $\omega(x)$ is closed (relative to $\mathbb{R}^n$) and bounded and, therefore, compact.

Step 5: $\omega(x)$ is connected.
Suppose $\omega(x)$ were disconnected. Then there would be disjoint open subsets $\mathcal{G}$ and $\mathcal{H}$ of Ω such that $\mathcal{G} \cap \omega(x)$ and $\mathcal{H} \cap \omega(x)$ are nonempty, and $\omega(x)$ is contained in $\mathcal{G} \cup \mathcal{H}$. Then there would have to be a sequence $s_1, s_2, \dots \to \infty$ and a sequence $t_1, t_2, \dots \to \infty$ such that $\varphi(s_k, x) \in \mathcal{G}$, $\varphi(t_k, x) \in \mathcal{H}$, and $s_k < t_k < s_{k+1}$ for each $k \in \mathbb{N}$. Because (for each fixed $k \in \mathbb{N}$)

$$\left\{\varphi(t, x) \mid t \in [s_k, t_k]\right\}$$

is a (connected) curve going from a point in $\mathcal{G}$ to a point in $\mathcal{H}$, there must be a time $\tau_k \in (s_k, t_k)$ such that $\varphi(\tau_k, x) \in \mathcal{K} \setminus \mathcal{G} \setminus \mathcal{H}$. Pick such a τ_k for each $k \in \mathbb{N}$ and note that $\tau_1, \tau_2, \ldots \to \infty$ and, by the Bolzano-Weierstrass Theorem, some subsequence of $(\varphi(\tau_k, x))$ must converge to a point y in $\mathcal{K} \setminus \mathcal{G} \setminus \mathcal{H}$. Note that y, being outside of $\mathcal{G} \cup \mathcal{H}$, cannot be in $\omega(x)$, which is a contradiction.

□

Examples of empty ω-limit sets are easy to find. Consider, for example, the one-dimensional dynamical system $\varphi(t, x) := x + t$ (generated by the differential equation $\dot{x} = 1$.

Examples of dynamical systems with nonempty, noncompact, disconnected ω-limit sets are a little harder to find. Consider the planar autonomous system

$$\begin{cases} \dot{x} = -y(1 - x^2) \\ \dot{y} = x + y(1 - x^2). \end{cases}$$

Although it takes a little work to show it, this differential equation generates a dynamical system on $\mathbb{R}^2$ (without the need for rescaling), and

$$\omega(x) = \bigl\{(-1, y) \bigm| y \in \mathbb{R}\bigr\} \cup \bigl\{(1, y) \bigm| y \in \mathbb{R}\bigr\}$$

for every x in the punctured strip

$$\bigl\{(x, y) \in \mathbb{R}^2 \bigm| |x| < 1 \text{ and } x^2 + y^2 > 0\bigr\}.$$

Regular and Singular Points
Lecture 19
Math 634
10/13/99

Consider the differential equation $\dot{x} = f(x)$ and its associated dynamical system $\varphi(t, x)$ on a phase space Ω.

Definition We say that a point $x \in \Omega$ is an *equilibrium point* or a *singular point* or a *critical point* if $f(x) = 0$. For such a point, $\varphi(t, x) = x$ for all $t \in \mathbb{R}$.

Definition A point $x \in \Omega$ that is not a singular point is called a *regular point.*

We shall show that all of the interesting local behavior of a continuous dynamical system takes place close to singular points. We shall do this by showing that in the neighborhood of each regular point, the flow is very similar to unidirectional, constant-velocity flow.

One way of making the notion of similarity of flows precise is the following.

Definition Two dynamical systems $\varphi : \mathbb{R} \times \Omega \to \Omega$ and $\psi : \mathbb{R} \times \Theta \to \Theta$ are *topologically conjugate* if there exists a homeomorphism (*i.e.*, a continuous bijection with continuous inverse) $h : \Omega \to \Theta$ such that

$$h(\varphi(t, x)) = \psi(t, h(x)) \tag{57}$$

for every $t \in \mathbb{R}$ and every $x \in \Omega$. In other words, $\psi = h \circ \varphi(t, \cdot) \circ h^{-1}$, or, equivalently, the diagram

$$\begin{tikzcd}
\Omega \arrow[r, "\varphi(t,\cdot)"] \arrow[d, "h"'] & \Omega \arrow[d, "h"] \\
\Theta \arrow[r, "\psi(t,\cdot)"] & \Theta
\end{tikzcd}$$

commutes for each $t \in \mathbb{R}$. The function h is called a *topological conjugacy.* If, in addition, h and h^{-1} are r-times continuously differentiable, we say that φ and ψ are C^r*-conjugate.*

A weaker type of similarity is the following.

Definition Two dynamical systems $\varphi : \mathbb{R} \times \Omega \to \Omega$ and $\psi : \mathbb{R} \times \Theta \to \Theta$ are *topologically equivalent* if there exists a homeomorphism $h : \Omega \to \Theta$ and a time reparametrization function $\alpha : \mathbb{R} \times \Omega \to \mathbb{R}$ such that, for each $x \in \Omega$, $\alpha(\cdot, x) : \mathbb{R} \to \mathbb{R}$ is an increasing surjection and

$$h(\varphi(\alpha(t,x),x)) = \psi(t, h(x))$$

for every $t \in \mathbb{R}$ and every $x \in \Omega$. If, in addition, h is r-times continuously differentiable, we say that φ and ψ are *C^r-equivalent*.

A topological equivalence maps orbits to orbits and preserves the orientation of time but may reparametrize time on each individual orbit.

As an example of the difference between these two concepts, consider the two planar dynamical systems

$$\varphi(t,x) = \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix} x$$

and

$$\psi(t,y) = \begin{bmatrix} \cos 2t & -\sin 2t \\ \sin 2t & \cos 2t \end{bmatrix} y,$$

generated, respectively, by the differential equations

$$\dot{x} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} x$$

and

$$\dot{y} = \begin{bmatrix} 0 & -2 \\ 2 & 0 \end{bmatrix} y.$$

The functions $h(x) = x$ and $\alpha(t,x) = 2t$ show that these two flows are topologically equivalent. But these two flows are not topologically conjugate, since, by setting $t = \pi$ we see that any function $h : \mathbb{R}^2 \to \mathbb{R}^2$ satisfying (57) would have to satisfy $h(x) = h(-x)$ for all x, which would mean that h is not invertible.

Because of examples like this, topological equivalence seems to be the preferred concept when dealing with flows. The following theorem, however, shows that in a neighborhood of a regular point, a smooth flow satisfies a local version of C^r-conjugacy with respect to a unidirectional, constant-velocity flow.

Theorem (C^r Rectification Theorem) *Suppose $f : \Omega \to \mathbb{R}^n$ is r-times continuously differentiable (with $r \geq 1$) and x_0 is a regular point of the flow generated by*

$$\dot{x} = f(x). \tag{58}$$

Then there is a neighborhood $\mathcal{V}$ of x_0, a neighborhood $\mathcal{W}$ of the origin in $\mathbb{R}^n$, and a C^r invertible map $g : \mathcal{V} \to \mathcal{W}$ such that, for each solution x of (58) *in $\mathcal{V}$, $y(t) := g(x(t))$ satisfies the equation*

$$\dot{y} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{59}$$

in $\mathcal{W}$.

Proof. Without loss of generality, we shall assume that $x_0 = 0$ and $f(x_0) = f(0) = \alpha e_1$ for some $\alpha > 0$. Let $\mathcal{W}$ be a small ball centered at 0 in $\mathbb{R}^n$, and define $G(y) := G((y_1, \dots, y_n)^T) = \varphi(y_1, (0, y_2, \dots, y_n)^T)$, where φ is the flow generated by (58). (While φ might not be a genuine dynamical system because it might not be defined for all time, we know that it is at least defined long enough that G is well-defined if $\mathcal{W}$ is sufficiently small.)

In words, $G(y)$ is the solution obtained by projecting y onto the plane through the origin perpendicular to $f(0)$ and locating the solution of (58) that starts at this projected point after y_1 units of time have elapsed.

Step 1: $\varphi(\cdot, p)$ is C^{r+1}.
We know that

$$\frac{d}{dt}\varphi(t,p) = f(\varphi(t,p)). \tag{60}$$

If f is continuous then, since $\varphi(\cdot, p)$ is continuous, (60) implies that $\varphi(\cdot, p)$ is C^1. If f is C^1, then the previous observation implies that $\varphi(\cdot, p)$ is C^1. Then (60) implies that $\frac{d}{dt}\varphi(t,p)$ is the composition of C^1 functions and is, therefore, C^1; this means that $\varphi(\cdot, p)$ is C^2. Continuing inductively, we see that, since f is C^r, $\varphi(\cdot, p)$ is C^{r+1}.

Step 2: $\varphi(t, \cdot)$ is C^r.
This is a consequence of applying differentiability with respect to parameters

inductively.

Step 3: G is C^r.
This is a consequence of Steps 1 and 2 and the formula for G in terms of φ.

Step 4: $DG(0)$ is an invertible matrix.
Since

$$\left.\frac{\partial G(y)}{\partial y_1}\right|_{y=0} = \left.\frac{\partial}{\partial t}\varphi(t,0)\right|_{t=0} = f(0) = \alpha e_1$$

and

$$\left.\frac{\partial G(y)}{\partial y_k}\right|_{y=0} = \left.\frac{\partial}{\partial p}\varphi(0,p)\right|_{p=0} e_k = \left.\frac{\partial p}{\partial p}\right|_{p=0} e_k = e_k,$$

for $k \neq 1$, we have

$$DG(0) = \left[\begin{array}{cccc} \alpha e_1 & e_2 & \cdots & e_n \end{array}\right],$$

which is invertible since $\alpha \neq 0$.

Step 5: If $\mathcal{W}$ is sufficiently small, then G is invertible.
This is a consequence of Step 4 and the Inverse Function Theorem.

Set g equal to the (locally defined) inverse of G. Since G is C^r, so is g. The only thing remaining to check is that if x satisfies (58) then $g \circ x$ satisfies (59). Equivalently, we can check that if y satisfies (59) then $G \circ y$ satisfies (58).

Step 6: If y satisfies (59) then $G \circ y$ satisfies (58).
By the chain rule,

$$\begin{aligned}\frac{d}{dt}G(y(t)) &= \left.\frac{\partial}{\partial s}\varphi(s,(0,y_2,\ldots,y_n))\right|_{s=y_1} \dot{y}_1 + \left.\frac{\partial}{\partial p}\varphi(y_1,p)\right|_{p=(0,y_2,\ldots,y_n)} \begin{bmatrix} 0 \\ \dot{y}_2 \\ \vdots \\ \dot{y}_n \end{bmatrix} \\ &= f(\varphi(y_1,(0,y_2,\ldots,y_n))) = f(G(y)).\end{aligned}$$

□

Definitions of Stability
Lecture 20
Math 634
10/15/99

In the previous lecture, we saw that all the "interesting" local behavior of flows occurs near equilibrium points. One important aspect of the behavior of flows has to do with whether solutions that start near a given solution stay near it for all time and/or move closer to it as time elapses. This question, which is the subject of *stability theory*, is not just of interest when the given solution corresponds to an equilibrium solution, so we study it–initially, at least–in a fairly broad context.

Definitions

First, we define some types of stability for solutions of the (possibly nonautonomous) equation

$$\dot{x} = f(t, x). \tag{61}$$

Definition A solution $\overline{x}(t)$ of (61) is *(Lyapunov) stable* if for each $\varepsilon > 0$ and $t_0 \in \mathbf{R}$ there exists $\delta = \delta(\varepsilon, t_0) > 0$ such that if $x(t)$ is a solution of (61) and $|x(t_0) - \overline{x}(t_0)| < \delta$ then $|x(t) - \overline{x}(t)| < \varepsilon$ for all $t \geq t_0$.

Definition A solution $\overline{x}(t)$ of (61) is *asymptotically stable* if it is (Lyapunov) stable and if for every $t_0 \in \mathbf{R}$ there exists $\delta = \delta(t_0) > 0$ such that if $x(t)$ is a solution of (61) and $|x(t_0) - \overline{x}(t_0)| < \delta$ then $|x(t) - \overline{x}(t)| \to 0$ as $t \uparrow \infty$.

Definition A solution $\overline{x}(t)$ of (61) is *uniformly stable* if for each $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that if $x(t)$ is a solution of (61) and $|x(t_0) - \overline{x}(t_0)| < \delta$ for some $t_0 \in \mathbb{R}$ then $|x(t) - \overline{x}(t)| < \varepsilon$ for all $t \geq t_0$.

Some authors use a weaker definition of uniform stability that turns out to be equivalent to Lyapunov stability for autonomous equations. Since our main interest is in autonomous equations and this alternative definition is somewhat more complicated than the definition given above, we will not use it here.

Definition A solution $\overline{x}(t)$ of (61) is *orbitally stable* if for every $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that if $x(t)$ is a solution of (61) and $|x(t_1) - \overline{x}(t_0)| < \delta$ for some $t_0, t_1 \in \mathbb{R}$ then

$$\bigcup_{t \geq t_1} x(t) \subseteq \bigcup_{t \geq t_0} \mathcal{B}(\overline{x}(t), \varepsilon).$$

Next, we present a couple of definitions of stability for subsets of the (open) phase space $\Omega \subseteq \mathbb{R}^n$ of a dynamical system $\varphi(t, x)$. (In these definitions, a *neighborhood* of a set $\mathcal{A} \subseteq \Omega$ is an open subset of Ω that contains $\mathcal{A}$.)

Definition The set $\mathcal{A}$ is *stable* if every neighborhood of $\mathcal{A}$ contains a positively invariant neighborhood of $\mathcal{A}$.

Note that the definition implies that stable sets are positively invariant.

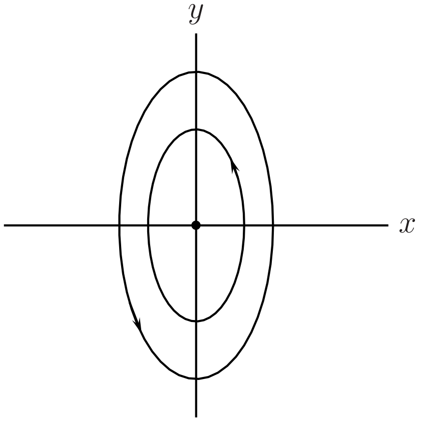
Definition The set $\mathcal{A}$ is *asymptotically stable* if it is stable and there is some neighborhood $\mathcal{V}$ of $\mathcal{A}$ such that $\omega(x) \subseteq \mathcal{A}$ for every $x \in \mathcal{V}$. (If $\mathcal{V}$ can be chosen to be the entire phase space, then $\mathcal{A}$ is *globally asymptotically stable*.)

Examples

We now consider a few examples that clarify some properties of these definitions.

$\boxed{1}$

$$\begin{cases} \dot{x} = -y/2 \\ \dot{y} = 2x. \end{cases}$$



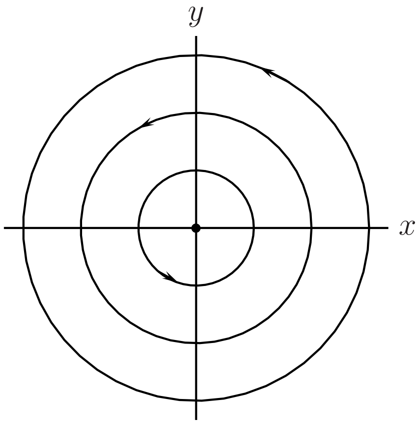
Orbits are ellipses with major axis along the y-axis. The equilibrium solution at the origin is Lyapunov stable even though nearby orbits sometimes move away from it.

$\boxed{\mathbf{2}}$

$$\begin{cases} \dot{r} = 0 \\ \dot{\theta} = r^2, \end{cases}$$

or, equivalently,

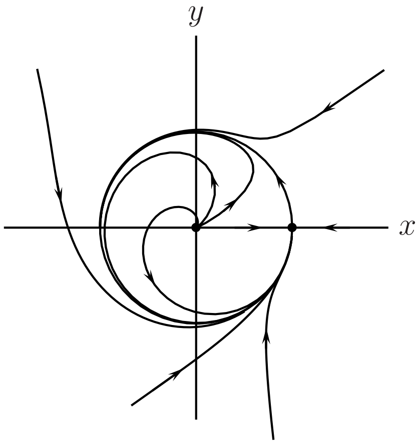
$$\begin{cases} \dot{x} = -(x^2 + y^2)y \\ \dot{y} = (x^2 + y^2)x. \end{cases}$$



The solution moving around the unit circle is not Lyapunov stable, since nearby solutions move with different angular velocities. It is, however, orbitally stable. Also, the set consisting of the unit circle is stable.

$\boxed{\mathbf{3}}$

$$\begin{cases} \dot{r} = r(1 - r) \\ \dot{\theta} = \sin^2(\theta/2). \end{cases}$$



The constant solution $(x, y) = (1, 0)$ is not Lyapunov stable and the set $\{(1, 0)\}$ is not stable. However, every solution beginning near $(1, 0)$ converges to $(1, 0)$ as $t \uparrow \infty$. This shows that it is not redundant to require Lyapunov stability (or stability) in the definition of asymptotic stability of a solution (or a set).

Stability in Autonomous Equations

When we are dealing with a smooth autonomous differential equation

$$\dot{x} = f(x) \tag{62}$$

on an open set $\Omega \subseteq \mathbb{R}^n$, all of the varieties of stability can be applied to essentially the same object. In particular, let $\overline{x}$ be a function that solves (62), and let

$$\mathcal{A}(\overline{x}) := \left\{\overline{x}(t) \mid t \in \mathbb{R}\right\}$$

be the corresponding orbit. Then it makes sense to talk about the Lyapunov, asymptotic, orbital, or uniform stability of $\overline{x}$, and it makes sense to talk about the stability or asymptotic stability of $\mathcal{A}(\overline{x})$.

In this context, certain relationships between the various types of stability follow from the definitions without too much difficulty.

Theorem *Let $\overline{x}$ be a function that solves* (62)*, and let $\mathcal{A}(\overline{x})$ be the corresponding orbit. Then:*

1. *If $\overline{x}$ is asymptotically stable then $\overline{x}$ is Lyapunov stable;*
2. *If $\overline{x}$ is uniformly stable then $\overline{x}$ is Lyapunov stable;*
3. *If $\overline{x}$ is uniformly stable then $\overline{x}$ is orbitally stable;*
4. *If $\mathcal{A}(\overline{x})$ is asymptotically stable then $\mathcal{A}(\overline{x})$ is stable;*
5. *If $\mathcal{A}(\overline{x})$ contains only a single point, then Lyapunov stability of $\overline{x}$, orbital stability of $\overline{x}$, uniform stability of $\overline{x}$, and stability of $\mathcal{A}(\overline{x})$ are equivalent.*

We will not prove this theorem, but we will note that parts 1 and 2 are immediate results of the definitions (even if we were dealing with a nonautonomous equation) and part 4 is also an immediate result of the definitions (even if $\mathcal{A}$ were an arbitrary set).

Exercise 13 In items 1–18, an autonomous differential equation, a phase space Ω (that is an open subset of $\mathbb{R}^n$), and a particular solution $\overline{x}$ of the equation are specified. For each of these items, state which of the following statements is/are true:

(a) $\overline{x}$ is Lyapunov stable;

(b) $\overline{x}$ is asymptotically stable;

(c) $\overline{x}$ is orbitally stable;

(d) $\overline{x}$ is uniformly stable;

(e) $\mathcal{A}(\overline{x})$ is stable;

(f) $\mathcal{A}(\overline{x})$ is asymptotically stable.

You do *not* need to justify your answers or show your work. It may convenient to express your answers in a concise form (*e.g.*, in a table of some sort). Use of variables r and θ signifies that the equation (as well as the particular solution) is to be interpreted as in polar form.
(The exercise is continued in the next box.)

Exercise 13 (continued)

1. $\dot{x} = x$, $\Omega = \mathbb{R}$, $\overline{x}(t) := 0$

2. $\dot{x} = x$, $\Omega = \mathbb{R}$, $\overline{x}(t) := e^t$

3. $\{\dot{x}_1 = 1 + x_2^2, \dot{x}_2 = 0\}$, $\Omega = \mathbb{R}^2$, $\overline{x}(t) := (t, 0)$

4. $\{\dot{r} = 0, \dot{\theta} = r^2\}$, $\Omega = \mathbb{R}^2$, $\overline{x}(t) := (1, t)$

5. $\dot{x} = x$, $\Omega = (0, \infty)$, $\overline{x}(t) := e^t$

6. $\{\dot{x}_1 = 1, \dot{x}_2 = -x_1 x_2\}$, $\Omega = \mathbb{R}^2$, $\overline{x}(t) := (t, 0)$

7. $\dot{x} = \tanh x$, $\Omega = \mathbb{R}$, $\overline{x}(t) := \sinh^{-1}(e^t)$

8. $\{\dot{x}_1 = \tanh x_1, \dot{x}_2 = 0\}$, $\Omega = (0, \infty) \times \mathbb{R}$, $\overline{x}(t) := (\sinh^{-1}(e^t), 0)$

9. $\dot{x} = \tanh x$, $\Omega = (0, \infty)$, $\overline{x}(t) := \sinh^{-1}(e^t)$

10. $\{\dot{x}_1 = \operatorname{sech} x_1, \dot{x}_2 = -x_1 x_2 \operatorname{sech} x_1\}$, $\Omega = \mathbb{R}^2$,
    $\overline{x}(t) := (\sinh^{-1}(t), 0)$

11. $\dot{x} = x^2/(1 + x^2)$, $\Omega = \mathbb{R}$, $\overline{x}(t) := -2/(t + \sqrt{t^2 + 4})$

12. $\{\dot{x}_1 = \operatorname{sech} x_1, \dot{x}_2 = -x_2\}$, $\Omega = \mathbb{R}^2$, $\overline{x}(t) := (\sinh^{-1}(t), 0)$

13. $\dot{x} = \operatorname{sech} x$, $\Omega = \mathbb{R}$, $\overline{x}(t) := \sinh^{-1}(t)$

14. $\{\dot{x}_1 = 1, \dot{x}_2 = 0\}$, $\Omega = \mathbb{R}^2$, $\overline{x}(t) := (t, 0)$

15. $\dot{x} = 0$, $\Omega = \mathbb{R}$, $\overline{x}(t) := 0$

16. $\dot{x} = 1$, $\Omega = \mathbb{R}$, $\overline{x}(t) := t$

17. $\{\dot{x}_1 = -x_1, \dot{x}_2 = -x_2\}$, $\Omega = \mathbb{R}^2$, $\overline{x}(t) := (e^{-t}, 0)$

18. $\dot{x} = -x$, $\Omega = \mathbb{R}$, $\overline{x}(t) := 0$

Principle of Linearized Stability
Lecture 21
Math 634
10/18/99

Suppose f is a continuously differentiable function such that

$$\dot{x} = f(x) \tag{63}$$

generates a continuous dynamical system φ on $\Omega \subseteq \mathbb{R}^n$. Suppose, moreover, that $x_0 \in \Omega$ is a singular point of φ. If x solves (63) and we set $u := x - x_0$ and $A := Df(x_0)$, we see that, by the definition of derivative,

$$\dot{u} = f(u + x_0) = f(x_0) + Df(x_0)u + R(u) = Au + R(u), \tag{64}$$

where $R(u)/|u| \to 0$ as $|u| \downarrow 0$. Because $R(u)$ is small when u is small, it is reasonable to believe that solutions of (64) behave similarly to solutions of

$$\dot{u} = Au \tag{65}$$

for u near 0. Equivalently, it is reasonable to believe that solutions of (63) behave like solutions of

$$\dot{x} = A(x - x_0) \tag{66}$$

for x near x_0. Equation (65) (or sometimes (66)) is called the *linearization* of (63) at x_0.

Now, we've defined (several types of) stability for equilibrium solutions of (63) (as well as for other types of solutions and sets), but we haven't really given any tools for determining stability. In this lecture we present one such tool, using the linearized equation(s) discussed above.

Definition An equilibrium point x_0 of (63) is *hyperbolic* if none of the eigenvalues of $Df(x_0)$ have zero real part.

If x_0 is hyperbolic, then either all the eigenvalues of $A := Df(x_0)$ have negative real part or at least one has positive real part. In the former case, we know that 0 is an asymptotically stable equilibrium solution of (65); in the latter case, we know that 0 is an unstable solution of (65). The following theorem says that similar things can be said about the nonlinear equation (63).

Theorem (Principle of Linearized Stability) *If x_0 is a hyperbolic equilibrium solution of* (63)*, then x_0 is either unstable or asymptotically stable, and its stability type (w.r.t.* (63)*) matches the stability type of* 0 *as an equilibrium solution of* (65) *(where $A := Df(x_0)$).*

This theorem is an immediate consequence of the following two propositions.

Proposition (Asymptotic Stability) *If x_0 is an equilibrium point of* (63) *and all the eigenvalues of $A := Df(x_0)$ have negative real part, then x_0 is asymptotically stable.*

Proposition (Instability) *If x_0 is an equilibrium point of* (63) *and some eigenvalue of $A := Df(x_0)$ has positive real part, then x_0 is unstable.*

Before we prove these propositions, we state and prove a lemma to which we have referred before in passing.

Lemma *Let $\mathcal{V}$ be a finite-dimensional real vector space and let $L \in \mathcal{L}(\mathcal{V}, \mathcal{V})$. If all the eigenvalues of L have real part larger than c, then there is an inner product $\langle \cdot, \cdot \rangle$ and an induced norm $\| \cdot \|$ on $\mathcal{V}$ such that*

$$\langle v, Lv \rangle \geq c\|v\|^2$$

for every $v \in \mathcal{V}$.

Proof. Let $n = \dim \mathcal{V}$, and pick $\varepsilon > 0$ so small that all the eigenvalues of L have real part greater than $c + n\varepsilon$. Choose a basis $\{v_1, \ldots, v_n\}$ for $\mathcal{V}$ that puts L in "modified" real canonical form with the off-diagonal 1's replaced by ε's, and let $\langle \cdot, \cdot \rangle$ be the inner product associated with this basis (*i.e.* $\langle v_i, v_j \rangle = \delta_{ij}$) and let $\| \cdot \|$ be the induced norm on $\mathcal{V}$.

Given $v = \sum_{i=1}^n \alpha_i v_i \in \mathcal{V}$, note that (if $L = (\ell_{ij})$)

$$\begin{aligned}\langle v, Lv \rangle &= \sum_{i=1}^n \ell_{ii}\alpha_i^2 + \sum_{i=1}^n \sum_{j \neq i} \ell_{ij}\alpha_i\alpha_j \geq \sum_{i=1}^n \ell_{ii}\alpha_i^2 - \sum_{i=1}^n \sum_{j \neq i} \varepsilon \left(\frac{\alpha_i^2 + \alpha_j^2}{2} \right) \\ &\geq \sum_{i=1}^n \ell_{ii}\alpha_i^2 - \sum_{i=1}^n n\varepsilon\alpha_i^2 = \sum_{i=1}^n (\ell_{ii} - n\varepsilon)\alpha_i^2 \geq \sum_{i=1}^n c\alpha_i^2 = c\|v\|^2.\end{aligned}$$

□

Note that applying this theorem to $-L$ also tells us that, for some inner product,

$$\langle v, Lv\rangle \leq c\|v\|^2 \tag{67}$$

if all the eigenvalues of L have real part less than c.

Proof of Proposition on Asymptotic Stability. Without loss of generality, assume that $x_0 = 0$. Pick $c < 0$ such that all the eigenvalues of A have real part strictly less than c. Because of equivalence of norms and because of the lemma, we can work with a norm $\|\cdot\|$ and a corresponding inner product $\langle\cdot,\cdot\rangle$ for which (67) holds, with $L = A$. Let $r > 0$ be small enough that $\|R(x)\| \leq -c/2\|x\|$ for all x satisfying $\|x\| \leq r$, and let

$$\mathcal{B}_r := \big\{x \in \Omega \ \big|\ \|x\| < r\big\}.$$

If $x(t)$ is a solution of (63) that starts in $\mathcal{B}_r$ at time $t = 0$, then as long as $x(t)$ remains in $\mathcal{B}_r$

$$\begin{aligned}\frac{d}{dt}\|x(t)\|^2 &= 2\langle x(t), \dot{x}(t)\rangle = 2\langle x(t), f(x(t))\rangle\\ &= 2\langle x(t), Ax(t)\rangle + 2\langle x(t), R(x(t))\rangle\\ &\leq 2c\|x(t)\|^2 + 2\|x(t)\| \cdot \|R(x(t))\|\\ &\leq 2c\|x(t)\|^2 - c\|x(t)\|^2 = c\|x(t)\|^2.\end{aligned}$$

This means that $x(t) \in \mathcal{B}_r$ for all $t \geq 0$, and $x(t)$ converges to 0 (exponentially quickly) as $t \uparrow \infty$. □

The proof of the second proposition will be geometric and will contain ideas that will be used to prove stronger results later in this course.

Proof of Proposition on Instability. We assume again that $x_0 = 0$. If $\mathcal{E}^u$,$\mathcal{E}^s$, and $\mathcal{E}^c$ are, respectively, the unstable, stable, and center spaces corresponding to (65), set $\mathcal{E}^- := \mathcal{E}^s \oplus \mathcal{E}^c$ and $\mathcal{E}^+ := \mathcal{E}^u$. Then $\mathbb{R}^n = \mathcal{E}^+ \oplus \mathcal{E}^-$, all of the eigenvalues of $A^+ := A|_{\mathcal{E}^+}$ have positive real part, and all of the eigenvalues of $A^- := A|_{\mathcal{E}^-}$ have nonpositive real part. Pick constants $a > b > 0$ such that all of the eigenvalues of A^+ have real part larger than a and all of the eigenvalues of A^- have real part less than b. Define an inner product $\langle\cdot,\cdot\rangle_+$ (and induced norm $\|\cdot\|_+$) on $\mathcal{E}^+$ such that

$$\langle v, Av\rangle_+ \geq a\|v\|_+^2$$

for all $v \in \mathcal{E}^+$, and define an inner product $\langle \cdot, \cdot \rangle_-$ (and induced norm $\| \cdot \|_-$) on $\mathcal{E}^-$ such that

$$\langle w, Aw \rangle_- \leq b\|w\|_-^2$$

for all $w \in \mathcal{E}^-$. Define $\langle \cdot, \cdot \rangle$ on $\mathcal{E}^+ \oplus \mathcal{E}^-$ to be the *direct sum* of $\langle \cdot, \cdot \rangle_+$ and $\langle \cdot, \cdot \rangle_-$; *i.e.*, let

$$\langle v_1 + w_1, v_2 + w_2 \rangle := \langle v_1, v_2 \rangle_+ + \langle w_1, w_2 \rangle_-$$

for all $(v_1, w_1), (v_2, w_2) \in \mathcal{E}^+ \times \mathcal{E}^-$. Let $\| \cdot \|$ be the induced norm, and note that

$$\|v + w\|^2 = \|v\|_+^2 + \|w\|_-^2 = \|v\|^2 + \|w\|^2$$

for all $(v, w) \in \mathcal{E}^+ \times \mathcal{E}^-$.

Now, take (63) and project it onto $\mathcal{E}^+$ and $\mathcal{E}^-$ to get the corresponding system for $(v, w) \in \mathcal{E}^+ \times \mathcal{E}^-$

$$\begin{cases} \dot{v} = A^+ v + R^+(v, w) \\ \dot{w} = A^- w + R^-(v, w), \end{cases} \tag{68}$$

with $\|R^\pm(v, w)\| / \|v + w\|$ converging to 0 as $\|v + w\| \downarrow 0$. Pick $\varepsilon > 0$ small enough that $a - b - 2\sqrt{2}\varepsilon > 0$, and pick $r > 0$ small enough that $\|R^\pm(v, w)\| \leq \varepsilon \|v + w\|$ whenever

$$v + w \in \mathcal{B}_r := \{ v + w \in \mathcal{E}^+ \oplus \mathcal{E}^- \mid \|v + w\| < r \}.$$

Consider the truncated cone

$$\mathcal{K}_r := \{ v + w \in \mathcal{E}^+ \oplus \mathcal{E}^- \mid \|v\| > \|w\| \} \cap \mathcal{B}_r.$$

(See Figure 1.) Suppose $x = v + w$ is a solution of (68) that starts in $\mathcal{K}_r$ at time $t = 0$. For as long as the solution remains in $\mathcal{K}_r$,

$$\begin{aligned} \frac{d}{dt}\|v\|^2 &= 2\langle v, \dot{v} \rangle = 2\langle v, A^+ v \rangle + 2\langle v, R^+(v, w) \rangle \\ &\geq 2a\|v\|^2 - 2\|v\| \cdot \|R^+(v, w)\| \geq 2a\|v\|^2 - 2\varepsilon\|v\| \cdot \|v + w\| \\ &= 2a\|v\|^2 - 2\varepsilon\|v\| \left(\|v\|^2 + \|w\|^2 \right)^{1/2} \geq 2a\|v\|^2 - 2\sqrt{2}\varepsilon\|v\|^2 \\ &= 2(a - \sqrt{2}\varepsilon)\|v\|^2, \end{aligned}$$

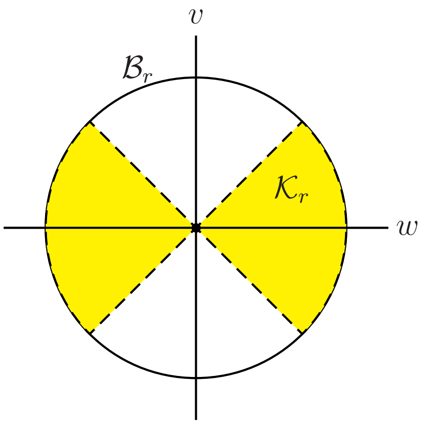


Figure 1: The truncated cone.

and

$$\begin{aligned}\frac{d}{dt}\|w\|^2 &= 2\langle w, \dot{w}\rangle = 2\langle w, A^- w\rangle + 2\langle w, R^-(v,w)\rangle \\ &\leq 2b\|w\|^2 + 2\|w\| \cdot \|R^-(v,w)\| \leq 2b\|w\|^2 + 2\varepsilon\|w\| \cdot \|v+w\| \\ &= 2b\|w\|^2 + 2\varepsilon\|w\| \left(\|v\|^2 + \|w\|^2\right)^{1/2} \leq 2b\|v\|^2 + 2\sqrt{2}\varepsilon\|v\|^2 \\ &= 2(b+\sqrt{2}\varepsilon)\|v\|^2.\end{aligned}$$

The first estimate says that as long as the solution stays in $\mathcal{K}_r$, $\|v\|$ grows exponentially; this means that the solution must eventually leave $\mathcal{K}_r$. Combining the first and second estimates, we have

$$\frac{d}{dt}(\|v\|^2 - \|w\|^2) \geq 2(a - b - 2\sqrt{2}\varepsilon)\|v\|^2 > 0,$$

so $g(v+w) := \|v\|^2 - \|w\|^2$ increases as t increases. But g is 0 on the lateral surface of $\mathcal{K}_r$ and is strictly positive in $\mathcal{K}_r$, so the solution cannot leave $\mathcal{K}_r$ through its lateral surface. Thus, the solution leaves $\mathcal{K}_r$ by leaving $\mathcal{B}_r$. Since this holds for all solutions starting in $\mathcal{K}_r$, we know that x_0 must be an unstable equilibrium point for (63). □

Lyapunov's Direct Method
Lecture 22
Math 634
10/20/99

An other tool for determining stability of solutions is *Lyapunov's direct method*. While this method may actually seem rather indirect, it does work directly on the equation in question instead of on its linearization.

We will consider this method for equilibrium solutions of (possibly) nonautonomous equations. Let $\Omega \subseteq \mathbb{R}^n$ be open and contain the origin, and suppose that $f : \mathbb{R} \times \Omega \to \mathbb{R}^n$ is a continuously differentiable function. Suppose, furthermore, that $f(t, 0) = 0$ for every $t \in \mathbb{R}$, so $x(t) := 0$ is a solution of the equation

$$\dot{x} = f(t, x). \tag{69}$$

(The results we obtain in this narrow context can be applied to determine the stability of other constant solutions of (69) by translation.)

In this lecture, a subset of Ω that contains the origin in its interior will be called a *neighborhood* of 0.

Definition Suppose that $\mathcal{D}$ is a neighborhood of 0 and that $W : \mathcal{D} \to \mathbb{R}$ is continuous and satisfies $W(0) = 0$. Then:

- If $W(x) \geq 0$ for every $x \in \mathcal{D}$, then W is *positive semidefinite.*
- If $W(x) > 0$ for every $x \in \mathcal{D} \setminus \{0\}$, then W is *positive definite.*
- If $W(x) \leq 0$ for every $x \in \mathcal{D}$, then W is *negative semidefinite.*
- If $W(x) < 0$ for every $x \in \mathcal{D} \setminus \{0\}$, then W is *negative definite.*

Definition Suppose that $\mathcal{D}$ is a neighborhood of 0 and that $V : \mathbb{R} \times \mathcal{D} \to \mathbb{R}$ is continuous and satisfies $V(t, 0) = 0$ for every $t \in \mathbb{R}$. Then:

- If there is a positive semidefinite function $W : \mathcal{D} \to \mathbb{R}$ such that $V(t, x) \geq W(x)$ for every $(t, x) \in \mathbb{R} \times \mathcal{D}$, then V is *positive semidefinite.*
- If there is a positive definite function $W : \mathcal{D} \to \mathbb{R}$ such that $V(t, x) \geq W(x)$ for every $(t, x) \in \mathbb{R} \times \mathcal{D}$, then V is *positive definite.*

- If there is a negative semidefinite function $W : \mathcal{D} \to \mathbb{R}$ such that $V(t,x) \leq W(x)$ for every $(t,x) \in \mathbb{R}\times\mathcal{D}$, then V is *negative semidefinite*.
- If there is a negative definite function $W : \mathcal{D} \to \mathbb{R}$ such that $V(t,x) \leq W(x)$ for every $(t,x) \in \mathbb{R} \times \mathcal{D}$, then V is *negative definite*.

Definition If $V : \mathbb{R}\times\mathcal{D}$ is continuously differentiable then its *orbital derivative* (w.r.t. (69)) is the function $\dot{V} : \mathbb{R} \times \mathcal{D} \to \mathbb{R}$ given by the formula

$$\dot{V}(t,x) := \frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x) \cdot f(t,x).$$

(Here "$\partial V(t,x)/\partial x$" represents the gradient of the function $V(t,\cdot)$.)

Note that if $x(t)$ is a solution of (69) then, by the chain rule,

$$\frac{d}{dt}V(t,x(t)) = \dot{V}(t,x(t)).$$

A function whose orbital derivative is always nonpositive is sometimes called a *Lyapunov function*.

Theorem (Lyapunov Stability) *If there is a neighborhood $\mathcal{D}$ of 0 and a continuously differentiable positive definite function $V : \mathbb{R} \times \mathcal{D} \to \mathbb{R}$ whose orbital derivative $\dot{V}$ is negative semidefinite, then 0 is a Lyapunov stable solution of (69).*

Proof. Let $\varepsilon > 0$ and $t_0 \in \mathbb{R}$ be given. Assume, without loss of generality, that $\overline{B(0,\varepsilon)}$ is contained in $\mathcal{D}$. Pick a positive definite function $W : \mathcal{D} \to \mathbb{R}$ such that $V(t,x) \geq W(x)$ for every $(t,x) \in \mathbb{R} \times \mathcal{D}$. Let

$$m := \min\bigl\{W(x) \bigm| |x| = \varepsilon\bigr\}.$$

Since W is continuous and positive definite, m is well-defined and positive. Pick $\delta > 0$ small enough that $\delta < \varepsilon$ and

$$\max\bigl\{V(t_0,x) \bigm| |x| \leq \delta\bigr\} < m.$$

(Since V is positive definite and continuous, this is possible.)

Now, if $x(t)$ solves (69) and $|x(t_0)| < \delta$ then $V(t_0, x(t_0)) < m$, and

$$\frac{d}{dt}V(t,x(t)) = \dot{V}(t,x(t)) \leq 0,$$

for all t, so $V(t, x(t)) < m$ for every $t \geq t_0$. Thus, $W(x(t)) < m$ for every $t \geq t_0$, so, for every $t \geq t_0$, $|x(t)| \neq \varepsilon$. Since $|x(t_0)| < \varepsilon$, this tells us that $|x(t)| < \varepsilon$ for every $t \geq t_0$. □

Theorem (Asymptotic Stability) *Suppose that there is a neighborhood $\mathcal{D}$ of 0 and a continuously differentiable positive definite function $V : \mathbb{R} \times \mathcal{D} \to \mathbb{R}$ whose orbital derivative $\dot{V}$ is negative definite, and suppose that there is a positive definite function $\overline{W} : \mathcal{D} \to \mathbb{R}$ such that $V(t, x) \leq \overline{W}(x)$ for every $(t, x) \in \mathbb{R} \times \mathcal{D}$. Then 0 is an asymptotically stable solution of* (69).

Proof. By the previous theorem, 0 is a Lyapunov stable solution of (69). Let $t_0 \in \mathbb{R}$ be given. Assume, without loss of generality, that $\mathcal{D}$ is compact. By Lyapunov stability, we know that we can choose a neighborhood $\mathcal{U}$ of 0 such that if $x(t)$ is a solution of (69) and $x(t_0) \in \mathcal{U}$, then $x(t) \in \mathcal{D}$ for every $t \geq t_0$. We claim that, in fact, if $x(t)$ is a solution of (69) and $x(t_0) \in \mathcal{U}$, then $x(t) \to 0$ as $t \uparrow \infty$. Verifying this claim will prove the theorem.

Suppose that $V(t, x(t))$ does not converge to 0 as $t \uparrow \infty$. The negative definiteness of $\dot{V}$ implies that $V(\cdot, x(\cdot))$ is nonincreasing, so, since $V \geq 0$, there must be a number $c > 0$ such that $V(t, x(t)) \geq c$ for every $t \geq t_0$. Then $\overline{W}(x(t)) \geq c > 0$ for every $t \geq t_0$. Since $\overline{W}(0) = 0$ and $\overline{W}$ is continuous,

$$\inf\bigl\{|x(t)| \bigm| t \geq t_0\bigr\} \geq \varepsilon \tag{70}$$

for some constant $\varepsilon > 0$. Pick a negative definite function $Y : \mathcal{D} \to \mathbb{R}$ such that $\dot{V}(t, x) \leq Y(x)$ for every $(t, x) \in \mathbb{R} \times \mathcal{D}$. The compactness of $\mathcal{D} \setminus \overline{B(0, \varepsilon)}$, along with (70), implies that

$$\bigl\{Y(x(t)) \bigm| t \geq t_0\bigr\}$$

is bounded away from 0. This, in turn, implies that

$$\bigl\{\dot{V}(t, x(t)) \bigm| t \geq t_0\bigr\}$$

is bounded away from 0. In other words,

$$\frac{d}{dt} V(t, x(t)) = \dot{V}(t, x(t)) \leq -\delta \tag{71}$$

for some constant $\delta > 0$. Clearly, (71) contradicts the nonnegativity of V for large t.

That contradiction implies that $V(t, x(t)) \to 0$ as $t \uparrow \infty$. Pick a positive definite function $\underline{W} : \mathcal{D} \to \mathbb{R}$ such that $V(t, x) \geq \underline{W}(x)$ for every $(t, x) \in \mathbb{R} \times \mathcal{D}$, and note that $\underline{W}(x(t)) \to 0$ as $t \uparrow \infty$.

Let $r > 0$ be given, and let

$$w_r = \min\bigl\{\underline{W}(p) \bigm| p \in \mathcal{D} \setminus \overline{B(0, r)}\bigr\},$$

which is defined and positive by the compactness of $\mathcal{D}$ and the continuity and positive definiteness of $\underline{W}$. Since $\underline{W}(x(t)) \to 0$ as $t \uparrow \infty$, there exists T such that $\underline{W}(x(t)) < w_r$ for every $t > T$. Thus, for $t > T$, it must be the case that $x(t) \in \overline{B(0, r)}$. Hence, 0 is asymptotically stable. $\square$

It may seem strange that we ned to bound V by a time-independent, positive definite function $\overline{W}$ from above. Indeed, some textbooks (see, *e.g.*, Theorem 2.20 in *Stability, Instability, and Chaos* by Glendinning) contain asymptotic stability theorems omitting this hypothesis. A counterexample by Massera demonstrates the necessity of the hypothesis.

Exercise 14 Show, by means of a counterexample, that the theorem on asymptotic stability via Lyapunov's direct method fails if the hypothesis about $\overline{W}$ is dropped.

(You may, but do not have to, proceed as follows. Let $g : \mathbb{R} \to \mathbb{R}$ be a function that is twice continuously differentiable and satisfies $g(t) \geq e^{-t}$ for every $t \in \mathbb{R}$, $g(t) \leq 1$ for every $t \geq 0$, $g(t) = e^{-t}$ for every

$$t \notin \bigcup_{n \in \mathbb{N}} (n - 2^{-n}, n + 2^{-n}),$$

and $g(n) = 1$ for every $n \in \mathbb{N}$. Let $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be the function defined by the formula

$$f(t, x) := \frac{g'(t)}{g(t)} x,$$

and let $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be the function defined by the formula

$$V(t, x) := \frac{x^2}{[g(t)]^2} \left[3 - \int_0^t [g(\tau)]^2 \, d\tau \right].$$

Show that, for x near 0, $V(t, x)$ is positive definite, $\dot{V}(t, x)$ is negative definite, and the solution 0 of (69) is not asymptotically stable.)

LaSalle's Invariance Principle

Lecture 23
Math 634
10/22/99

Linearization versus Lyapunov Functions

In the previous two lectures, we have talked about two different tools that can be used to prove that an equilibrium point x_0 of an autonomous system

$$\dot{x} = f(x) \tag{72}$$

is asymptotically stable: linearization and Lyapunov's direct method. One might ask which of these methods is better. Certainly, linearization seems easier to apply because of its straightforward nature: Compute the eigenvalues of $Df(x_0)$. The direct method requires you to find an appropriate Lyapunov function, which doesn't seem so straightforward. But, in fact, anytime linearization works, a simple Lyapunov function works, as well.

To be more precise, suppose $x_0 = 0$ and all the eigenvalues of $A := Df(0)$ have negative real part. Pick an inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$ such that, for some $c > 0$,

$$\langle x, Ax \rangle \leq -c\|x\|^2$$

for all $x \in \mathbb{R}^n$. Pick $r > 0$ small enough that $\|f(x) - Ax\| \leq (c/2)\|x\|$ whenever $\|x\| \leq r$, let

$$\mathcal{D} = \bigl\{ x \in \mathbb{R}^n \bigm| \|x\| \leq r \bigr\},$$

and define $V : \mathbb{R} \times \mathcal{D} \to \mathbb{R}$ by the formula $V(t, x) = \|x\|^2$. Since $\| \cdot \|$ is a norm, V is positive definite. Also

$$\begin{aligned} \dot{V}(t, x) &= 2\langle x, f(x) \rangle = 2(\langle x, Ax \rangle + \langle x, f(x) - Ax \rangle) \\ &\leq 2(-c\|x\|^2 + \|x\|\|f(x) - Ax\|) \leq -c\|x\|^2, \end{aligned}$$

so $\dot{V}$ is negative definite.

On the other hand, there are very simple examples to illustrate that the direct method works in some cases where linearization doesn't. For example, consider $\dot{x} = -x^3$ on $\mathbb{R}$. The equilibrium point at the origin is not hyperbolic, so linearization fails to determine stability, but it is easy to check that x^2 is positive definite and has a negative definite orbital derivative, thus ensuring the asymptotic stability of 0.

A More Complicated Example

The previous example is so simple that it might make one question whether the direct method is of any use on problems where stability cannot be determined by linearization *or* by inspection. Thus, let's consider something more complicated. Consider the planar system

$$\begin{cases} \dot{x} = -y - x^3 \\ \dot{y} = x^5. \end{cases}$$

The origin is a nonhyperbolic equilibrium point, with 0 being the only eigenvalue, so the principle of linearized stability is of no use. A sketch of the phase portrait indicates that orbits circle the origin in the counterclockwise direction, but it is not obvious whether they spiral in, spiral out, or move on closed curves.

The simplest potential Lyapunov function that often turns out to be useful is the square of the standard Euclidean norm, which in this case is $V := x^2 + y^2$. The orbital derivative is

$$\dot{V} = 2x\dot{x} + 2y\dot{y} = 2x^5y - 2xy - 2x^4. \tag{73}$$

For some points (x, y) near the origin (*e.g.*, (δ, δ)) $\dot{V} < 0$, while for other points near the origin (*e.g.*, $(\delta, -\delta)$) $\dot{V} > 0$, so this function doesn't seem to be of much use.

Sometimes when the square of the standard Euclidean norm doesn't work, some other homogeneous quadratic function does. Suppose we try $V := x^2 + \alpha xy + \beta y^2$, with α and β to be determined. Then

$$\begin{aligned} \dot{V} &= (2x + \alpha y)\dot{x} + (\alpha x + 2\beta y)\dot{y} = -(2x + \alpha y)(y + x^3) + (\alpha x + 2\beta y)x^5 \\ &= -2x^4 + \alpha x^6 - 2xy - \alpha x^3 y + 2\beta x^5 y - \alpha y^2. \end{aligned}$$

Setting $(x, y) = (\delta, -\delta^2)$ for δ positive and small, we see that $\dot{V}$ is not going to be negative semidefinite, no matter what we pick α and β to be.

If these quadratic functions don't work, maybe something customized for the particular equation might. Note that the right-hand side of the first equation in (73) sort of suggests that x^3 and y should be treated as quantities of the same order of magnitude. Let's try $V := x^6 + \alpha y^2$, for some $\alpha > 0$ to be determined. Clearly, V is positive definite, and

$$\dot{V} = 6x^5\dot{x} + 2\alpha y\dot{y} = (2\alpha - 6)x^5 y - 6x^8.$$

If $\alpha \neq 3$, then $\dot{V}$ is of opposite signs for $(x,y) = (\delta,\delta)$ and for $(x,y) = (\delta,-\delta)$ when δ is small. Hence, we should set $\alpha = 3$, yielding $\dot{V} = -6x^8 \leq 0$. Thus V is positive definite and $\dot{V}$ is negative semidefinite, implying that the origin is Lyapunov stable.

Is the origin asymptotically stable? Perhaps we can make a minor modification to the preceding formula for V so as to make $\dot{V}$ strictly negative in a deleted neighborhood of the origin without destroying the positive definiteness of V. If we added a small quantity whose orbital derivative was strictly negative when $x = 0$ and $|y|$ is small and positive, this might work. Experimentation suggests that a positive multiple of xy^3 might work, since this quantity changes from positive to negative as we cross the y-axis in the counterclockwise direction. Also, it is at least of higher order than $3y^2$ near the origin, so it has the potential of preserving the positive definiteness of V.

In fact, we claim that $V := x^6 + xy^3 + 3y^2$ is positive definite with negative definite orbital derivative near 0. A handy inequality, sometimes called Young's inequality, that can be used in verifying this claim (and in other circumstances, as well) is given in the following lemma.

Lemma (Young's Inequality) *If $a, b \geq 0$, then*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}, \tag{74}$$

for every pair of numbers $p, q \in (1,\infty)$ satisfying

$$\frac{1}{p} + \frac{1}{q} = 1. \tag{75}$$

Proof. Assume that (75) holds. Clearly (74) holds if $b = 0$, so assume that $b > 0$, and fix it. Define $g : [0,\infty)$ by the formula

$$g(x) := \frac{x^p}{p} + \frac{b^q}{q} - xb.$$

Note that g is continuous, and $g'(x) = x^{p-1} - b$ for every $x \in (0,\infty)$. Since $\lim_{x\downarrow 0} g'(x) = -b < 0$, $\lim_{x\uparrow\infty} g'(x) = \infty$, and g' is increasing on $(0,\infty)$, we know that g has a unique minimizer at $x_0 = b^{1/(p-1)}$. Thus, for every $x \in [0,\infty)$ we see, using (75), that

$$g(x) \geq g(b^{1/(p-1)}) = \frac{b^{p/(p-1)}}{p} + \frac{b^q}{q} - b^{p/(p-1)} = \left(\frac{1}{p} + \frac{1}{q} - 1\right) b^q = 0.$$

In particular, $g(a) \geq 0$, so (74) holds. □

Now, let $V = x^6 + xy^3 + 3y^2$. Applying Young's inequality with $a = |x|$, $b = |y|^3$, $p = 6$, and $q = 6/5$, we see that

$$|xy^3| = |x||y|^3 \leq \frac{|x|^6}{6} + \frac{5|y|^{18/5}}{6} \leq \frac{1}{6}x^6 + \frac{5}{6}y^2$$

if $|y| \leq 1$, so

$$V \geq \frac{5}{6}x^6 + \frac{13}{6}y^2$$

if $|y| \leq 1$. Also,

$$\begin{aligned}\dot{V} &= -6x^8 + y^3\dot{x} + 3xy^2\dot{y} = -6x^8 - y^3(y + x^3) + 3x^6y^2 \\ &= -6x^8 - x^3y^3 + 3x^6y^2 - y^4.\end{aligned}$$

Applying Young's inequality to the two mixed terms in this orbital derivative, we have

$$|-x^3y^3| = |x|^3|y|^3 \leq \frac{3|x|^8}{8} + \frac{5|y|^{24/5}}{8} \leq \frac{3}{8}x^8 + \frac{5}{8}y^4$$

if $|y| \leq 1$, and

$$|3x^6y^2| = 3|x|^6|y|^2 \leq 3\left[\frac{3|x|^8}{4} + \frac{|y|^8}{4}\right] = \frac{9}{4}x^8 + \frac{3}{4}y^8 \leq \frac{9}{4}x^8 + \frac{3}{64}y^4$$

if $|y| \leq 1/2$. Thus,

$$\dot{V} \leq -\frac{27}{8}x^8 - \frac{21}{64}y^4$$

if $|y| \leq 1/2$, so, in a neighborhood of 0, V is positive definite and $\dot{V}$ is negative definite, which implies that 0 is asymptotically stable.

LaSalle's Invariance Principle

We would have saved ourselves a lot of work on the previous example if we could have just stuck with the moderately simple function $V = x^6 + 3y^2$, even though its orbital derivative was only negative semidefinite. Notice that the set of points where $\dot{V}$ was 0 was small (the y-axis) and at most of those points the vector field was not parallel to the set. LaSalle's Invariance Principle, which we shall state and prove for the autonomous system

$$\dot{x} = f(x), \tag{76}$$

allows us to use such a V to prove asymptotic stability.

Theorem (LaSalle's Invariance Principle) *Suppose there is a neighborhood $\mathcal{D}$ of 0 and a continuously differentiable (time-independent) positive definite function $V : \mathcal{D} \to \mathbb{R}$ whose orbital derivative $\dot{V}$ (w.r.t. (76)) is negative semidefinite. Let $\mathcal{I}$ be the union of all complete orbits contained in*

$$\left\{x \in \mathcal{D} \;\middle|\; \dot{V}(x) = 0\right\}.$$

Then there is a neighborhood $\mathcal{U}$ of 0 such that for every $x_0 \in \mathcal{U}$, $\omega(x_0) \subseteq \mathcal{I}$.

Before proving this, we note that when applying it to $V = x^6 + 3y^2$ in the previous example, the set $\mathcal{I}$ is a singleton containing the origin and, since $\mathcal{D}$ can be assumed to be compact, each solution beginning in $\mathcal{U}$ actually converges to 0 as $t \uparrow \infty$.

Proof of LaSalle's Invariance Principle. Let φ be the flow generated by (76). By a previous theorem, 0 must be Lyapunov stable, so we can pick a neighborhood $\mathcal{U}$ of 0 such that $\varphi(t, x) \in \mathcal{D}$ for every $x_0 \in \mathcal{U}$ and every $t \geq 0$.

Let $x_0 \in \mathcal{U}$ and $y \in \omega(x_0)$ be given. By the negative semidefiniteness of $\dot{V}$, we know that $V(\varphi(t, x_0))$ is a nonincreasing function of t. By the positive definiteness of V, we know that $V(\varphi(t, x_0))$ remains nonnegative, so it must approach some constant $c \geq 0$ as $t \uparrow \infty$. By continuity of V, $V(z) = c$ for every $z \in \omega(x_0)$. Since $\omega(x_0)$ is invariant, $V(\varphi(t, y)) = c$ for every $t \in \mathbb{R}$. The definition of orbital derivative then implies that $\dot{V}(\varphi(t, y)) = 0$ for every $t \in \mathbb{R}$. Hence, $y \in \mathcal{I}$. □

Exercise 15 Show that $(x(t), y(t)) = (0, 0)$ is an asymptotically stable solution of

$$\begin{cases} \dot{x} = -x^3 + 2y^3 \\ \dot{y} = -2xy^2. \end{cases}$$

Hartman-Grobman Theorem: Part 1

Lecture 24
Math 634
10/25/99

Let $\Omega \subset \mathbb{R}^n$ be open and let $f : \Omega \to \mathbb{R}^n$ be continuously differentiable. Suppose that $x_0 \in \Omega$ is a hyperbolic equilibrium point of the autonomous equation

$$\dot{x} = f(x). \tag{77}$$

Let $B = Df(x_0)$, and let φ be the (local) flow generated by (77). The version of the Hartman-Grobman Theorem we're primarily interested in is the following.

Theorem (Local Hartman-Grobman Theorem for Flows) *Let Ω, f, x_0, B, and φ be as described above. Then there are neighborhoods $\mathcal{U}$ and $\mathcal{V}$ of x_0 and a homeomorphism $h : \mathcal{U} \to \mathcal{V}$ such that*

$$\varphi(t, h(x)) = h(x_0 + e^{tB}(x - x_0))$$

whenever $x \in \mathcal{U}$ and $x_0 + e^{tB}(x - x_0) \in \mathcal{U}$.

It will be easier to derive this theorem as a consequence of a global theorem for maps than to prove it directly. In order to state this version of the theorem, we will need to introduce a couple of function spaces and a definition.

Let

$$C_b^0(\mathbb{R}^n) = \bigl\{ w \in C(\mathbb{R}^n, \mathbb{R}^n) \bigm| \sup_{x \in \mathbb{R}^n} |w(x)| < \infty \bigr\}.$$

When equipped with the norm

$$\|w\|_0 := \sup_{x \in \mathbb{R}^n} \|w(x)\|,$$

where $\|\cdot\|$ is some norm on $\mathbb{R}^n$, $C_b^0(\mathbb{R}^n)$ is a Banach space. (We shall pick a particular norm $\|\cdot\|$ later.)

Let

$$C_b^1(\mathbb{R}^n) = \bigl\{ w \in C^1(\mathbb{R}^n, \mathbb{R}^n) \cap C_b^0(\mathbb{R}^n) \bigm| \sup_{x \in \mathbb{R}^n} \|Dw(x)\| < \infty \bigr\},$$

where $\|\cdot\|$ is the operator norm corresponding to some norm on $\mathbb{R}^n$. Note that the functional

$$\operatorname{Lip}(w) := \sup_{\substack{x_1,x_2\in\mathbb{R}^n\\ x_1\neq x_2}} \frac{\|w(x_1)-w(x_2)\|}{\|x_1-x_2\|}$$

is defined on $C_b^1(\mathbb{R}^n)$. We will not define a norm on $C_b^1(\mathbb{R}^n)$, but will often use Lip, which is not a norm, to describe the size of elements of $C_b^1(\mathbb{R}^n)$.

Definition If $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ and none of the eigenvalues of A lie on the unit circle, then A is *hyperbolic*.

Note that if x_0 is a hyperbolic equilibrium point of (77) and $A = e^{Df(x_0)}$, then A is hyperbolic.

Theorem (Global Hartman-Grobman Theorem for Maps) *Suppose that the map $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ is hyperbolic and invertible. Then there exists a number $\varepsilon > 0$ such that for every $g \in C_b^1(\mathbb{R}^n)$ satisfying* $\operatorname{Lip}(g) < \varepsilon$ *there exists a unique function $v \in C_b^0(\mathbb{R}^n)$ such that*

$$F(h(x)) = h(Ax)$$

for every $x \in \mathbb{R}^n$, where $F = A+g$ and $h = I+v$. Furthermore, $h : \mathbb{R}^n \to \mathbb{R}^n$ is a homeomorphism.

Hartman-Grobman Theorem: Part 2

Lecture 25
Math 634
10/27/99

Subspaces and Norms

We start off with a lemma that is analogous to the lemma in Lecture 21, except this one will deal with the magnitude, rather than the real part, of eigenvalues.

Lemma *Let $\mathcal{V}$ be a finite-dimensional real vector space and let $L \in \mathcal{L}(\mathcal{V}, \mathcal{V})$. If all the eigenvalues of L have magnitude less than c, then there is a norm $\|\cdot\|$ on $\mathcal{V}$ such that*

$$\|Lv\| \leq c\|v\|$$

for every $v \in \mathcal{V}$.

Proof. As in the previous lemma, the norm will be the Euclidean norm corresponding to the modification of the real canonical basis that yields a matrix representation of L that has ε's in place of the off-diagonal 1's. With respect to this basis, it can be checked that

$$L^T L = D + R(\varepsilon),$$

where D is a diagonal matrix, each of whose diagonal entries is less than c^2, and $R(\varepsilon)$ is a matrix whose entries converge to 0 as $\varepsilon \downarrow 0$. Hence, as in the proof of the earlier lemma, we can conclude that if ε is sufficiently small then

$$\|Lv\|^2 = \langle v, L^T Lv \rangle \leq c^2 \|v\|^2$$

for every $v \in \mathcal{V}$ (where $\langle \cdot, \cdot \rangle$ is the inner product that induces $\|\cdot\|$). □

Note that if L is a linear operator, all of whose eigenvalues have magnitude *greater* than c, then by applying the lemma to L^{-1} (which exists, since 0 is not an eigenvalue of L), we see that

$$\|Lv\| \geq c\|v\|$$

for some norm $\|\cdot\|$.

Now, suppose that $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ is hyperbolic. Then, since A has only finitely many eigenvalues, there is a number $a \in (0,1)$ such that none of the eigenvalues of A are in the closed annulus

$$\overline{B(0,a^{-1})} \setminus B(0,a).$$

Using the notation developed when we were deriving the real canonical form, let

$$\mathcal{E}^- = \left\{ \bigoplus_{\lambda \in (-a,a)} N(A - \lambda I) \right\} \oplus$$
$$\left\{ \bigoplus_{\substack{|\lambda|<a \\ \operatorname{Im}\lambda \neq 0}} \left\{ \operatorname{Re} u \;\middle|\; u \in N(A-\lambda I) \right\} \right\} \oplus \left\{ \bigoplus_{\substack{|\lambda|<a \\ \operatorname{Im}\lambda \neq 0}} \left\{ \operatorname{Im} u \;\middle|\; u \in N(A-\lambda I) \right\} \right\},$$

and let

$$\mathcal{E}^+ = \left\{ \bigoplus_{\lambda \in (-\infty,-a^{-1}) \cup (a^{-1},\infty)} N(A - \lambda I) \right\} \oplus$$
$$\left\{ \bigoplus_{\substack{|\lambda|>a^{-1} \\ \operatorname{Im}\lambda \neq 0}} \left\{ \operatorname{Re} u \;\middle|\; u \in N(A-\lambda I) \right\} \right\} \oplus \left\{ \bigoplus_{\substack{|\lambda|>a^{-1} \\ \operatorname{Im}\lambda \neq 0}} \left\{ \operatorname{Im} u \;\middle|\; u \in N(A-\lambda I) \right\} \right\}.$$

Then $\mathbb{R}^n = \mathcal{E}^- \oplus \mathcal{E}^+$, and $\mathcal{E}^-$ and $\mathcal{E}^+$ are both invariant under A. Define $P^- \in \mathcal{L}(\mathbb{R}^n, \mathcal{E}^-)$ and $P^+ \in \mathcal{L}(\mathbb{R}^n, \mathcal{E}^+)$ to be the linear operators that map each $x \in \mathbb{R}^n$ to the unique elements $P^-x \in \mathcal{E}^-$ and $P^+x \in \mathcal{E}^+$ such that $P^-x + P^+x = x$.

Let $A^- \in \mathcal{L}(\mathcal{E}^-, \mathcal{E}^-)$ and $A^+ \in \mathcal{L}(\mathcal{E}^+, \mathcal{E}^+)$ be the restrictions of A to $\mathcal{E}^-$ and $\mathcal{E}^+$, respectively. By the lemma (and the discussion thereafter) we can find a norm $\|\cdot\|_-$ for $\mathcal{E}^-$ and a norm $\|\cdot\|_+$ for $\mathcal{E}^+$ such that

$$\|A^-x\|_- \leq a\|x\|_-$$

for every $x \in \mathcal{E}^-$, and

$$\|A^+x\|_+ \geq a^{-1}\|x\|_+$$

for every $x \in \mathcal{E}^+$. Define a norm $\|\cdot\|$ on $\mathbb{R}^n$ by the formula

$$\|x\| = \max\{\|P^- x\|_-, \|P^+ x\|_+\}. \tag{78}$$

This is the norm on $\mathbb{R}^n$ that we will use throughout our proof of the (global) Hartman-Grobman Theorem (for maps). Note that $\|x\| = \|x\|_-$ if $x \in \mathcal{E}^-$, and $\|x\| = \|x\|_+$ if $x \in \mathcal{E}^+$.

Recall that we equipped $C_b^0(\mathbb{R}^n)$ with the norm $\|\cdot\|_0$ defined by the formula

$$\|w\|_0 := \sup_{x \in \mathbb{R}^n} \|w(x)\|.$$

The norm on $\mathbb{R}^n$ on the right-hand side of this formula is that given in (78). Recall also that we will use the functional Lip defined by the formula

$$\operatorname{Lip}(w) := \sup_{\substack{x_1, x_2 \in \mathbb{R}^n \\ x_1 \neq x_2}} \frac{\|w(x_1) - w(x_2)\|}{\|x_1 - x_2\|}$$

The norm on $\mathbb{R}^n$ on the right-hand side of this formula is also that given in (78).

Let

$$C_b^0(\mathcal{E}^-) = \bigl\{ w \in C(\mathbb{R}^n, \mathcal{E}^-) \bigm| \sup_{x \in \mathbb{R}^n} \|w(x)\|_- < \infty \bigr\},$$

and let

$$C_b^0(\mathcal{E}^+) = \bigl\{ w \in C(\mathbb{R}^n, \mathcal{E}^+) \bigm| \sup_{x \in \mathbb{R}^n} \|w(x)\|_+ < \infty \bigr\}.$$

Since $\mathbb{R}^n = \mathcal{E}^- \oplus \mathcal{E}^+$, it follows that

$$C_b^0(\mathbb{R}^n) = C_b^0(\mathcal{E}^-) \oplus C_b^0(\mathcal{E}^+),$$

and the corresponding decomposition of an element $w \in C_b^0(\mathbb{R}^n)$ is

$$w = P^- \circ w + P^+ \circ w.$$

We equip $C_b^0(\mathcal{E}^-)$ and $C_b^0(\mathcal{E}^+)$ with the same norm $\|\cdot\|_0$ that we used on $C_b^0(\mathbb{R}^n)$, thereby making each of these two spaces a Banach space. It is not hard to see that

$$\|w\|_0 = \max\{\|P^- \circ w\|_0, \|P^+ \circ w\|_0\}.$$

Hartman-Grobman Theorem: Part 3

Lecture 26
Math 634
10/29/99

Linear and Nonlinear Maps

Now, assume that A is invertible, so that

$$\inf_{x \neq 0} \frac{\|Ax\|}{\|x\|} > 0.$$

Choose, and fix, a positive constant

$$\varepsilon < \min \left\{ 1 - a, \inf_{x \neq 0} \frac{\|Ax\|}{\|x\|} \right\}.$$

Choose, and fix, a function $g \in \mathbb{C}_b^1(\mathbb{R}^n)$ for which $\operatorname{Lip}(g) < \varepsilon$. The (global) Hartman-Grobman Theorem (for maps) will be proved by constructing a map Θ from $C_b^0(\mathbb{R}^n)$ to $C_b^0(\mathbb{R}^n)$ whose fixed points would be precisely those objects v which, when added to the identity I, would yield solutions h to the conjugacy equation

$$(A + g) \circ h = h \circ A, \tag{79}$$

and then showing that Θ is a contraction (and that h is a homeomorphism).

Plugging $h = I + v$ into (79) and manipulating the result, we can see that that equation is equivalent to the equation

$$\mathcal{L}v = \Psi(v), \tag{80}$$

where $\Psi(v) := g \circ (I + v) \circ A^{-1}$ and

$$\mathcal{L}v = v - A \circ v \circ A^{-1} =: (\operatorname{id} - \mathcal{A})v.$$

Since the composition of continuous functions is continuous, and the composition of functions is bounded if the outer function in the composition is bounded, it is clear that Ψ is a (nonlinear) map from $C_b^0(\mathbb{R}^n)$ to $C_b^0(\mathbb{R}^n)$. Similarly, $\mathcal{A}$ and, therefore, $\mathcal{L}$ are linear maps from $C_b^0(\mathbb{R}^n)$ to $C_b^0(\mathbb{R}^n)$. We will show that $\mathcal{L}$ can be inverted and then apply $\mathcal{L}^{-1}$ to both sides of (80) to get

$$v = \mathcal{L}^{-1}(\Psi(v)) =: \Theta(v), \tag{81}$$

as our fixed point equation.

Inverting $\mathcal{L}$

Since A behaves significantly differently on $\mathcal{E}^-$ than it does on $\mathcal{E}^+$, $\mathcal{A}$ and, therefore, $\mathcal{L}$ behave significantly differently on $C_b^0(\mathcal{E}^-)$ than they do on $C_b^0(\mathcal{E}^+)$. For this reason, we will analyze $\mathcal{L}$ by looking at its restrictions to $C_b^0(\mathcal{E}^-)$ and to $C_b^0(\mathcal{E}^+)$. Note that $C_b^0(\mathcal{E}^-)$ and $C_b^0(\mathcal{E}^+)$ are invariant under $\mathcal{A}$ and, therefore, under $\mathcal{L}$. Let $\mathcal{A}^- \in \mathcal{L}(C_b^0(\mathcal{E}^-), C_b^0(\mathcal{E}^-))$ and $\mathcal{A}^+ \in \mathcal{L}(C_b^0(\mathcal{E}^+), C_b^0(\mathcal{E}^+))$ be the restrictions of $\mathcal{A}$ to $C_b^0(\mathcal{E}^-)$ and $C_b^0(\mathcal{E}^+)$, respectively, and let $\mathcal{L}^- \in \mathcal{L}(C_b^0(\mathcal{E}^+), C_b^0(\mathcal{E}^+))$ and $\mathcal{L}^+ \in \mathcal{L}(C_b^0(\mathcal{E}^+), C_b^0(\mathcal{E}^+))$ be the corresponding restrictions of $\mathcal{L}$. Then $\mathcal{L}$ will be invertible if and only if $\mathcal{L}^-$ and $\mathcal{L}^+$ are each invertible. To invert $\mathcal{L}^-$ and $\mathcal{L}^+$ we use the following general result about the invertibility of operators on Banach spaces.

Lemma *Let $\mathcal{X}$ be a Banach space with norm $\|\cdot\|_{\mathcal{X}}$ and corresponding operator norm $\|\cdot\|_{\mathcal{L}(\mathcal{X},\mathcal{X})}$. Let G be a linear map from $\mathcal{X}$ to $\mathcal{X}$, and let $c < 1$ be a constant. Then:*

(a) *If $\|G\|_{\mathcal{L}(\mathcal{X},\mathcal{X})} \le c$, then $\operatorname{id} - G$ is invertible and*

$$\|(\operatorname{id} - G)^{-1}\|_{\mathcal{L}(\mathcal{X},\mathcal{X})} \le \frac{1}{1-c}.$$

(b) *If G is invertible and $\|G^{-1}\|_{\mathcal{L}(\mathcal{X},\mathcal{X})} \le c$, then $\operatorname{id} - G$ is invertible and*

$$\|(\operatorname{id} - G)^{-1}\|_{\mathcal{L}(\mathcal{X},\mathcal{X})} \le \frac{c}{1-c}.$$

Proof. The space of bounded linear maps from $\mathcal{X}$ to $\mathcal{X}$ is a Banach space using the operator norm. In case **(a)**, the bound on $\|G\|_{\mathcal{L}(\mathcal{X},\mathcal{X})}$, along with the Cauchy convergence criterion, implies that the series

$$\sum_{k=0}^{\infty} G^k$$

converges to a bounded linear map from $\mathcal{X}$ to $\mathcal{X}$; call it H. In fact, we see that (by the formula for the sum of a geometric series)

$$\|H\|_{\mathcal{L}(\mathcal{X},\mathcal{X})} \le \frac{1}{1-c}.$$

It is not hard to check that $H \circ (\operatorname{id} - G) = (\operatorname{id} - G) \circ H = \operatorname{id}$, so $H = (\operatorname{id} - G)^{-1}$.

In case **(b)**, we can apply the results of **(a)** with G^{-1} in place of G to deduce that $\operatorname{id} - G^{-1}$ is invertible and that

$$\|(\operatorname{id} - G^{-1})^{-1}\|_{\mathcal{L}(\mathcal{X},\mathcal{X})} \leq \frac{1}{1-c}.$$

Since $\operatorname{id} - G = -G(\operatorname{id} - G^{-1}) = -(\operatorname{id} - G^{-1})G$, it is not hard to check that $-(\operatorname{id} - G^{-1})^{-1}G^{-1}$ is the inverse of $\operatorname{id} - G$ and that

$$\| - (\operatorname{id} - G^{-1})^{-1}G^{-1}\|_{\mathcal{L}(\mathcal{X},\mathcal{X})} \leq \frac{c}{1-c}.$$

□

The first half of this lemma is useful for inverting small perturbations of the identity, while the second half is useful for inverting large perturbations of the identity. It should, therefore, not be too surprising that we will apply the first half with $G = \mathcal{A}^-$ and the second half with $G = \mathcal{A}^+$ (since A compresses things in the $\mathcal{E}^-$ direction and stretches things in the $\mathcal{E}^+$ direction).

First, consider $\mathcal{A}^-$. If $w \in C_b^0(\mathcal{E}^-)$, then

$$\begin{aligned}\|\mathcal{A}^- w\|_0 = \|A \circ w \circ A^{-1}\|_0 = \sup_{x \in \mathbb{R}^n} \|Aw(A^{-1}x)\| = \sup_{y \in \mathbb{R}^n} \|Aw(y)\| \\ \leq a \sup_{y \in \mathbb{R}^n} \|w(y)\| = a\|w\|_0,\end{aligned}$$

so the operator norm of $\mathcal{A}^-$ is bounded by a. Applying the first half of the lemma with $\mathcal{X} = C_b^0(\mathcal{E}^-)$, $G = \mathcal{A}^-$, and $c = a$, we find that $\mathcal{L}^-$ is invertible, and its inverse has operator norm bounded by $(1-a)^{-1}$.

Next, consider $\mathcal{A}^+$. It is not hard to see that $\mathcal{A}^+$ is invertible, and $(\mathcal{A}^+)^{-1}w = A^{-1} \circ w \circ A$. If $w \in C_b^0(\mathcal{E}^+)$, then (because the eigenvalues of the restriction of A^{-1} to $\mathcal{E}^+$ all have magnitude less than a)

$$\begin{aligned}\|(\mathcal{A}^+)^{-1} w\|_0 = \|A^{-1} \circ w \circ A\|_0 = \sup_{x \in \mathbb{R}^n} \|A^{-1}w(Ax)\| = \sup_{y \in \mathbb{R}^n} \|A^{-1}w(y)\| \\ \leq a \sup_{y \in \mathbb{R}^n} \|w(y)\| = a\|w\|_0,\end{aligned}$$

so the operator norm of $(\mathcal{A}^+)^{-1}$ is bounded by a. Applying the second half of the lemma with $\mathcal{X} = C_b^0(\mathcal{E}^+)$, $G = \mathcal{A}^+$, and $c = a$, we find that $\mathcal{L}^+$ is invertible, and its inverse has operator norm bounded by $a(1-a)^{-1}$.

Putting these two facts together, we see that $\mathcal{L}$ is invertible, and, in fact,

$$\mathcal{L}^{-1} = (\mathcal{L}^-)^{-1} \circ P^- + (\mathcal{L}^+)^{-1} \circ P^+.$$

If $w \in C_b^0(\mathbb{R}^n)$, then

$$\begin{aligned}
\|\mathcal{L}^{-1}w\|_0 &= \sup_{x\in\mathbb{R}^n} \|\mathcal{L}^{-1}w(x)\| = \sup_{x\in\mathbb{R}^n} \max\{\|P^-\mathcal{L}^{-1}w(x)\|, \|P^+\mathcal{L}^{-1}w(x)\|\} \\
&= \sup_{x\in\mathbb{R}^n} \max\{\|(\mathcal{L}^-)^{-1}P^-w(x)\|, \|(\mathcal{L}^+)^{-1}P^+w(x)\|\} \\
&\le \sup_{x\in\mathbb{R}^n} \max\left\{\frac{1}{1-a}\|w(x)\|, \frac{a}{1-a}\|w(x)\|\right\} = \frac{1}{1-a}\sup_{x\in\mathbb{R}^n}\|w(x)\| \\
&= \frac{1}{1-a}\|w\|_0,
\end{aligned}$$

so the operator norm of $\mathcal{L}^{-1}$ is bounded by $(1-a)^{-1}$.

Hartman-Grobman Theorem: Part 4

Lecture 27
Math 634
11/1/99

The Contraction Map

Recall that we are looking for fixed points v of the map $\Theta := \mathcal{L}^{-1} \circ \Psi$, where $\Psi(v) := g \circ (I + v) \circ A^{-1}$. We have established that $\mathcal{L}^{-1}$ is a well-defined linear map from $C_b^0(\mathbb{R}^n)$ to $C_b^0(\mathbb{R}^n)$ and that its operator norm is bounded by $(1-a)^{-1}$. This means that Θ is a well-defined (nonlinear) map from $C_b^0(\mathbb{R}^n)$ to $C_b^0(\mathbb{R}^n)$; furthermore, if $v_1, v_2 \in C_b^0(\mathbb{R}^n)$, then

$$\begin{aligned}
\|\Theta(v_1) - \Theta(v_2)\|_0 = \|\mathcal{L}^{-1}(\Psi(v_1) - \Psi(v_2))\|_0 &\leq \frac{1}{1-a}\|\Psi(v_1) - \Psi(v_2)\|_0 \\
&= \frac{1}{1-a}\|g \circ (I + v_1) \circ A^{-1} - g \circ (I + v_2) \circ A^{-1}\|_0 \\
&= \frac{1}{1-a} \sup_{x \in \mathbb{R}^n} \|g(A^{-1}x + v_1(A^{-1}x)) - g(A^{-1}x + v_2(A^{-1}x))\| \\
&\leq \frac{\varepsilon}{1-a} \sup_{x \in \mathbb{R}^n} \|(A^{-1}x + v_1(A^{-1}x)) - (A^{-1}x + v_2(A^{-1}x))\| \\
&= \frac{\varepsilon}{1-a} \sup_{x \in \mathbb{R}^n} \|v_1(A^{-1}x) - v_2(A^{-1}x)\| \\
&= \frac{\varepsilon}{1-a} \sup_{y \in \mathbb{R}^n} \|v_1(y) - v_2(y)\| = \frac{\varepsilon}{1-a}\|v_1 - v_2\|_0.
\end{aligned}$$

This shows that Θ is a contraction, since ε was chosen to be less than $1 - a$. By the contraction mapping theorem, we know that Θ has a unique fixed point $v \in C_b^0(\mathbb{R}^n)$; the function $h := I + v$ satisfies $F \circ h = h \circ A$, where $F := A + g$. It remains to show that h is a homeomorphism.

Injectivity

First, we show that F is injective. Suppose it weren't. Then we could choose $x_1, x_2 \in \mathbb{R}^n$ such that $x_1 \neq x_2$ but $F(x_1) = F(x_2)$. This would mean that $Ax_1 + g(x_1) = Ax_2 + g(x_2)$, so

$$\frac{\|A(x_1 - x_2)\|}{\|x_1 - x_2\|} = \frac{\|Ax_1 - Ax_2\|}{\|x_1 - x_2\|} = \frac{\|g(x_1) - g(x_2)\|}{\|x_1 - x_2\|} \leq \operatorname{Lip}(g) < \varepsilon < \inf_{x \neq 0} \frac{\|Ax\|}{\|x\|},$$

which would be a contradiction.

Next, we show that h is injective. Let $x_1, x_2 \in \mathbb{R}^n$ satisfying $h(x_1) = h(x_2)$ be given. Then

$$h(Ax_1) = F(h(x_1)) = F(h(x_2)) = h(Ax_2),$$

and, by induction, we have $h(A^n x_1) = h(A^n x_2)$ for every $n \in \mathbb{N}$. Also,

$$F(h(A^{-1}x_1)) = h(AA^{-1}x_1) = h(x_1) = h(x_2) = h(AA^{-1}x_2) = F(h(A^{-1}x_2)),$$

so the injectivity of F implies that $h(A^{-1}x_1) = h(A^{-1}x_2)$; by induction, $h(A^{-n}x_1) = h(A^{-n}x_2)$ for every $n \in \mathbb{N}$. Set $z = x_1 - x_2$. Since $I = h - v$, we know that for any $n \in \mathbb{Z}$

$$\begin{aligned}\|A^n z\| &= \|A^n x_1 - A^n x_2\| = \|(h(A^n x_1) - v(A^n x_1)) - (h(A^n x_2) - v(A^n x_2))\| \\ &= \|v(A^n x_1) - v(A^n x_2)\| \leq 2\|v\|_0.\end{aligned}$$

Because of the way the norm was chosen, we then know that for $n \geq 0$

$$\|P^+ z\| \leq a^n \|A^n P^+ z\| \leq a^n \|A^n z\| \leq 2a^n \|v\|_0 \to 0,$$

as $n \uparrow \infty$, and we know that for $n \leq 0$

$$\|P^- z\| \leq a^{-n} \|A^n P^- z\| \leq a^{-n} \|A^n z\| \leq 2a^{-n} \|v\|_0 \to 0,$$

as $n \downarrow -\infty$. Hence, $z = P^- z + P^+ z = 0$, so $x_1 = x_2$.

Surjectivity

It may seem intuitive that a map like h that is a bounded perturbation of the identity is surjective. Unfortunately, there does not appear to be a way of proving this that is simultaneously elementary, short, and complete. We will therefore rely on the following topological theorem without proving it.

Theorem (Brouwer Invariance of Domain) *Every continuous injective map from $\mathbb{R}^n$ to $\mathbb{R}^n$ maps open sets to open sets.*

In particular, this theorem implies that $h(\mathbb{R}^n)$ is open. If we can show that $h(\mathbb{R}^n)$ is closed, then (since $h(\mathbb{R}^n)$ is clearly nonempty) this will mean that $h(\mathbb{R}^n) = \mathbb{R}^n$, *i.e.*, h is surjective.

So, suppose we have a sequence $(h(x_k))$ of points in $h(\mathbb{R}^n)$ that converges to a point $y \in \mathbb{R}^n$. Without loss of generality, assume that

$$\|h(x_k) - y\| \leq 1$$

for every k. This implies that $\|h(x_k)\| \leq \|y\| + 1$, which in turn implies that $\|x_k\| \leq \|y\| + \|v\|_0 + 1$. Thus, the sequence (x_k) is bounded and therefore has a subsequence (x_{k_ℓ}) converging to some point $x_0 \in \mathbb{R}^n$. By continuity of h, $(h(x_{k_\ell}))$ converges to $h(x_0)$, which means that $h(x_0) = y$. Hence, $h(\mathbb{R}^n)$ is closed.

Continuity of the Inverse

The bijectivity of h implies that h^{-1} is defined. We now show that it is continuous (which will complete the verification that h is a homeomorphism). The proof will be very similar to the proof that $h(\mathbb{R}^n)$ is closed.

Let (y_k) be a sequence in $\mathbb{R}^n$ that converges to some point $y \in \mathbb{R}^n$. Without loss of generality, assume that

$$\|y_k - y\| \leq 1$$

for every k. This implies that $\|y_k\| \leq \|y\| + 1$, which in turn implies that $\|h^{-1}(y_k)\| \leq \|y\| + \|v\|_0 + 1$. Suppose that $(h^{-1}(y_k))$ does not converge to $h^{-1}(y)$. Then the boundedness of $(h^{-1}(y_k))$ implies that some subsequence $(h^{-1}(y_{k_\ell}))$ converges to some point $x_0 \neq h^{-1}(y)$. By the continuity of h, $h(h^{-1}(y_{k_\ell})) \to h(x_0)$ as $\ell \uparrow \infty$. But $h(h^{-1}(y_{k_\ell})) = y_{k_\ell} \to y$ as $\ell \uparrow \infty$. This means that $y = h(x_0)$ or, equivalently, $x_0 = h^{-1}(y)$, contrary to assumption.

Hartman-Grobman Theorem: Part 5

Lecture 28
Math 634
11/3/99

Modifying the Vector Field

Consider the continuously differentiable autonomous differential equation

$$\dot{x} = f(x) \tag{82}$$

with an equilibrium point that, without loss of generality, is located at the origin. For x near 0, $f(x) \approx Bx$, where $B = Df(0)$. Our goal is to come up with a modification $\tilde{f}$ of f such that $\tilde{f}(x) = f(x)$ for x near 0 and $\tilde{f}(x) \approx Bx$ for *all* x. If we accomplish this goal, whatever information we obtain about the relationship between the equations

$$\dot{x} = \tilde{f}(x) \tag{83}$$

and

$$\dot{x} = Bx \tag{84}$$

will also hold between (82) and (84) for x small.

Pick $\beta : [0,\infty) \to [0,1]$ to be a C^∞ function satisfying

$$\beta(s) = \begin{cases} 1 & \text{if } s \leq 1 \\ 0 & \text{if } s \geq 2, \end{cases}$$

and let $C = \sup_{s\in[0,\infty)} |\beta'(s)|$. Given $\varepsilon > 0$, pick $r > 0$ so small that

$$\|Df(x) - B\| < \frac{\varepsilon}{2C+1}$$

whenever $\|x\| \leq 2r$. (We can do this since $Df(0) = B$ and Df is continuous.) Define $\tilde{f}$ by the formula

$$\tilde{f}(x) = Bx + \beta\left(\frac{\|x\|}{r}\right)(f(x) - Bx).$$

Note that $\tilde{f}$ is continuously differentiable, agrees with f for $\|x\| \leq r$, and agrees with B for $\|x\| \geq 2r$. We claim that $\tilde{f} - B$ has Lipschitz constant less than ε. Assuming, without loss of generality, that $\|x\|$ and $\|y\|$ are less than or equal to $2r$, we have (using the Mean Value Theorem)

$$\begin{aligned}
&\|(\tilde{f}(x) - Bx) - (\tilde{f}(y) - By)\| \\
&\qquad = \left\| \beta\left(\frac{\|x\|}{r}\right)(f(x) - Bx) - \beta\left(\frac{\|y\|}{r}\right)(f(y) - By) \right\| \\
&\qquad \leq \beta\left(\frac{\|x\|}{r}\right)\|(f(x) - Bx) - (f(y) - By)\| \\
&\qquad\quad + \left|\beta\left(\frac{\|x\|}{r}\right) - \beta\left(\frac{\|y\|}{r}\right)\right| \|f(y) - By\| \\
&\qquad \leq \frac{\varepsilon}{2C+1}\|x-y\| + C\frac{|\|x\| - \|y\||}{r}\|y\|\frac{\varepsilon}{2C+1} \\
&\qquad \leq \varepsilon\|x-y\|.
\end{aligned}$$

Now, consider the difference between e^B and $\varphi(1,\cdot)$, where φ is the flow generated by $\tilde{f}$. Let $g(x) = \varphi(1,x) - e^Bx$. Then, since $\tilde{f}(x) = B(x)$ for all large x, $g(x) = 0$ for all large x. Also, g is continuously differentiable, so $g \in C^1_b(\mathbb{R}^n)$. If we apply the variation of constants formula to (83) rewritten as

$$\dot{x} = Bx + (\tilde{f}(x) - Bx),$$

we find that

$$g(x) = \int_0^1 e^{(1-s)B}[\tilde{f}(\varphi(s,x)) - B\varphi(s,x)]\,ds,$$

so

$$\begin{aligned}
&\|g(x) - g(y)\| \\
&\qquad \leq \int_0^1 \|e^{(1-s)B}\|\|(\tilde{f}(\varphi(s,x)) - B\varphi(s,x)) - (\tilde{f}(\varphi(s,y)) - B\varphi(s,y))\|\,ds \\
&\qquad \leq \varepsilon \int_0^1 \|e^{(1-s)B}\|\|\varphi(s,x) - \varphi(s,y)\|\,ds \\
&\qquad \leq \|x-y\|\varepsilon \int_0^1 \|e^{(1-s)B}\|\|e^{(\|B\|+\varepsilon)s} - 1\|\,ds,
\end{aligned}$$

by continuous dependence on initial conditions. Since

$$\varepsilon \int_0^1 \|e^{(1-s)B}\|\|e^{(\|B\|+\varepsilon)s} - 1\|\,ds \to 0$$

as $\varepsilon \downarrow 0$, we can make the Lipschitz constant of g as small as we want by making ε small (through shrinking the neighborhood of the origin on which $\tilde{f}$ and f agree).

Conjugacy for $t = 1$

If 0 is a hyperbolic equilibrium point of (82) (and therefore of (83)) then none of the eigenvalues of B are imaginary. Setting $A = e^B$, it is not hard to show that the eigenvalues of A are the exponentials of the eigenvalues of B, so none of the eigenvalues of A have modulus 1; *i.e.*, A is hyperbolic. Also, A is invertible (since $A^{-1} = e^{-B}$), so we can apply the global Hartman-Grobman Theorem for maps and conclude that there is a homeomorphism $h : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\varphi(1, h(x)) = h(e^B x) \tag{85}$$

for every $x \in \mathbb{R}^n$.

Conjugacy for $t \neq 1$

For the Hartman-Grobman Theorem for flows, we need

$$\varphi(t, h(x)) = h(e^{tB} x)$$

for every $x \in \mathbb{R}^n$ and *every* $t \in \mathbb{R}$. Fix $t \in \mathbb{R}$, and consider the function $\tilde{h}$ defined by the formula

$$\tilde{h}(x) = \varphi(t, h(e^{-tB} x)). \tag{86}$$

As the composition of homeomorphisms, $\tilde{h}$ is a homeomorphism. Also, the fact that h satisfies (85) implies that

$$\begin{aligned}\varphi(1, \tilde{h}(x)) &= \varphi(1, \varphi(t, h(e^{-tB} x))) = \varphi(t, \varphi(1, h(e^{-tB} x))) = \varphi(t, h(e^B e^{-tB} x)) \\ &= \varphi(t, h(e^{-tB} e^B x))) = \tilde{h}(e^B x),\end{aligned}$$

so (85) holds if h is replaced by $\tilde{h}$.

Now,

$$\begin{aligned}\tilde{h} - I &= \varphi(t, \cdot) \circ h \circ e^{-tB} - I \\ &= (\varphi(t, \cdot) - e^{tB}) \circ h \circ e^{-tB} + e^{tB} \circ (h - I) \circ e^{-tB} =: v_1 + v_2.\end{aligned}$$

The fact that $\varphi(t, x)$ and $e^{tB} x$ agree for large x implies that $\varphi(t, \cdot) - e^{tB}$ is bounded, so v_1 is bounded, as well. The fact that $h - I$ is bounded implies that v_2 is bounded. Hence, $\tilde{h} - I$ is bounded.

The uniqueness part of the global Hartman-Grobman Theorem for maps now implies that h and $\tilde{h}$ must be the same function. Using this fact and substituting $y = e^{-tB}x$ in (86) yields

$$h(e^{tB}y) = \varphi(t, h(y))$$

for every $y \in \mathbb{R}^n$ and every $t \in \mathbb{R}^n$. This means that the flows generated by (84) and (83) are globally topologically conjugate, and the flows generated by (84) and (82) are locally topologically conjugate.

Constructing Conjugacies

Lecture 29
Math 634
11/5/99

The Hartman-Grobman Theorem gives us conditions under which a conjugacy between certain maps or between certain flows may exist. Some limitations of the theorem are:

- The conditions it gives are sufficient, but certainly not necessary, for a conjugacy to exist.
- It doesn't give a simple way to construct a conjugacy (in closed form, at least).
- It doesn't indicate how smooth the conjugacy might be.

These shortcomings can be addressed in a number of different ways, but we won't really go into those here. We will, however, consider some aspects of conjugacies.

Differentiable Conjugacies of Flows

Consider the autonomous differential equations

$$\dot{x} = f(x) \tag{87}$$

and

$$\dot{x} = g(x), \tag{88}$$

generating, respectively, the flows φ and ψ. Recall that the conjugacy equation for φ and ψ is

$$\varphi(t, h(x)) = h(\psi(t, x)) \tag{89}$$

for every x and t. Not only is (89) somewhat complicated, it appears to require you to solve (87) and (88) before you can look for a conjugacy h. Suppose, however, that h is a differentiable conjugacy. Then, we can differentiate both sides of (89) with respect to t to get

$$f(\varphi(t, h(x))) = Dh(\psi(t, x))g(\psi(t, x)). \tag{90}$$

Substituting (89) into the right-hand side of (90) and replacing $\psi(t, x)$ by x, we get the equivalent equation

$$f(h(x)) = Dh(x)g(x). \tag{91}$$

Note that (91) involves the functions appearing in the differential equations, rather than the formulas for the solutions of those equations. Note, also, that (91) is the same equation you would get if you took a solution x of (88) and required the function $h \circ x$ to satisfy (87).

An Example for Flows

As the simplest nontrivial example, let $a, b \in \mathbb{R}$ be *distinct* constants and consider the equations

$$\dot{x} = ax \tag{92}$$

and

$$\dot{x} = bx \tag{93}$$

for $x \in \mathbb{R}$. Under what conditions on a and b does their exist a topological conjugacy h taking solutions of (93) to solutions of (92)? Equation (91) tells us that if h is differentiable then

$$ah(x) = h'(x)bx. \tag{94}$$

If $b \neq 0$, then separating variables in (94) implies that on intervals avoiding the origin h must be given by the formula

$$h(x) = C|x|^{a/b} \tag{95}$$

for some constant C. Clearly, (95) does not define a topological conjugacy for a single constant C, because it fails to be injective on $\mathbb{R}$; however, the formula

$$h(x) = \begin{cases} x|x|^{a/b-1} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0, \end{cases} \tag{96}$$

which is obtained from (95) by taking $C = 1$ for positive x and $C = -1$ for negative x, defines a homeomorphism if $ab > 0$. Even though the function defined in (96) may fail to be differentiable at 0, substitution of it into

$$e^{ta}h(x) = h(e^{tb}x), \tag{97}$$

which is (89) for this example, shows that it does, in fact, define a topological conjugacy when $ab > 0$. (Note that in no case is this a C^1-conjugacy, since either $h'(0)$ or $(h^{-1})'(0)$ does not exist.)

Now, suppose that $ab \leq 0$. Does a topological (possibly nondifferentiable) conjugacy exist? If $ab = 0$, then (97) implies that h is constant, which violates injectivity, so suppose that $ab < 0$. In this case, substituting $x = 0$ and $t = 1$ into (97) implies that $h(0) = 0$. Fixing $x \neq 0$ and letting $t \operatorname{sgn} b \downarrow -\infty$ in (97), we see that the continuity of h implies that $h(x) = 0$, also, which again violates injectivity.

Summarizing, for $a \neq b$ there is a topological conjugacy of (92) and (93) if and only if $ab > 0$, and these are not C^1-conjugacies.

An Example for Maps

Let's try a similar analysis for maps. Let $a, b \in \mathbb{R}$ be distinct constants, and consider the maps $F(x) = ax$ and $G(x) = bx$ (for $x \in \mathbb{R}$). For what (a, b)-combinations does there exist a homeomorphism $h : \mathbb{R} \to \mathbb{R}$ such that

$$F(h(x)) = h(G(x)) \tag{98}$$

for every $x \in \mathbb{R}$? Can h and h^{-1} be chosen to be differentiable?

For these specific maps, the general equation (98) becomes

$$ah(x) = h(bx). \tag{99}$$

If $a = 0$ or $b = 0$ or $a = 1$ or $b = 1$, then injectivity is immediately violated. Note that, by induction, (99) gives

$$a^n h(x) = h(b^n x) \tag{100}$$

for every $n \in \mathbb{Z}$. In particular, $a^2 h(x) = h(b^2 x)$, so the cases when $a = -1$ or $b = -1$ cause the same problems as when $a = 1$ or $b = 1$.

So, from now on, assume that $a, b \notin \{-1, 0, 1\}$. Observe that:

- Setting $x = 0$ in (99) yields $h(0) = 0$.
- If $|b| < 1$, then fixing $x \neq 0$ in (100) and letting $n \uparrow \infty$, we have $|a| < 1$.
- If $|b| > 1$, we can, similarly, let $n \downarrow -\infty$ to conclude that $|a| > 1$.

- If $b > 0$ and $a < 0$, then (99) implies that $h(1)$ and $h(b)$ have opposite signs even though 1 and b have the same sign; consequently, the Intermediate Value Theorem yields a contradiction to injectivity.
- If $b < 0$ and $a > 0$, then (99) gives a similar contradiction.

Thus, the only cases where we could possibly have conjugacy is if a and b are both in the same component of

$$(-\infty, -1) \cup (-1, 0) \cup (0, 1) \cup (1, \infty).$$

When this condition is met, experimentation (or experience) suggests trying h of the form $h(x) = x|x|^{p-1}$ for some constant $p > 0$ (with $h(0) = 0$). This is a homeomorphism from $\mathbb{R}$ to $\mathbb{R}$, and plugging it into (99) shows that it provides a conjugacy if $a = b|b|^{p-1}$ or, in other words, if

$$p = \frac{\log |a|}{\log |b|}.$$

Since $a \neq b$, either h or h^{-1} fails to be differentiable at 0. Is there some other formula that provides a C^1-conjugacy? No, because if there were we could differentiate both sides of (99) with respect to x and evaluate at $x = 0$ to get $h'(0) = 0$, which would mean that $(h^{-1})'(0)$ is undefined.

Exercise 16 Define $F : \mathbb{R}^2 \to \mathbb{R}^2$ by the formula

$$F\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} -x/2 \\ 2y + x^2 \end{bmatrix},$$

and let $A = DF(0)$.

(a) Show that the maps F and A are topologically conjugate.

(b) Show that the flows generated by the differential equations

$$\dot{z} = F(z)$$

and

$$\dot{z} = Az$$

are topologically conjugate.

(Hint: Try quadratic conjugacy functions.)

Smooth Conjugacies

Lecture 30
Math 634
11/8/99

The examples we looked at last time showing that topological conjugacies often cannot be chosen to be differentiable all involved two maps or vector fields with different linearizations at the origin. What about when, as in the Hartman-Grobman Theorem, we are looking for a conjugacy between a map (or flow) and its linearization? An example of Hartman shows that the conjugacy cannot always be chosen to be C^1.

Hartman's Example

Consider the system

$$\begin{cases} \dot{x} = \alpha x \\ \dot{y} = (\alpha - \gamma) y + \varepsilon x z \\ \dot{z} = -\gamma z, \end{cases}$$

where $\alpha > \gamma > 0$ and $\varepsilon \neq 0$. We will not cut off this vector field but will instead confine our attention to x, y, z small. A calculation shows that the time-1 map $F = \varphi(1, \cdot)$ of this system is given by

$$F\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) = \begin{bmatrix} ax \\ ac(y + \varepsilon x z) \\ cz \end{bmatrix},$$

where $a = e^{\alpha}$ and $c = e^{-\gamma}$. Note that $a > ac > 1 > c > 0$. The time-1 map B of the linearization of the differential equation is given by

$$B \begin{bmatrix} x \\ y \\ y \end{bmatrix} = \begin{bmatrix} ax \\ acy \\ cz \end{bmatrix}.$$

A local conjugacy $H = (f, g, h)$ of B with F must satisfy

$$\begin{aligned} af(x,y,z) &= f(ax, acy, cz) \\ ac[g(x,y,z) + \varepsilon f(x,y,z) h(x,y,z)] &= g(ax, acy, cz) \\ ch(x,y,z) &= h(ax, acy, cz) \end{aligned}$$

for every x, y, z near 0. Writing $k(x, z)$ for $k(x, 0, z)$, where $k \in \{f, g, h\}$, we have

$$\begin{aligned} af(x,z) &= f(ax, cz) && (101) \\ ac[g(x,z) + \varepsilon f(x,z)h(x,z)] &= g(ax, cz) && (102) \\ ch(x,z) &= h(ax, cz) && (103) \end{aligned}$$

for every x, z near 0.

Before proceeding further, we state and prove a lemma.

Lemma *Suppose that j is a continuous real-valued function of a real variable, defined on an open interval $\mathcal{U}$ centered at the origin. Suppose that there are constants $\alpha, \beta \in \mathbb{R}$ such that*

$$\alpha j(u) = j(\beta u) \tag{104}$$

whenever $u, \beta u \in \mathcal{U}$. Then if $|\beta| < 1 < |\alpha|$ or $|\alpha| < 1 < |\beta|$, $j(u) = 0$ for every $u \in \mathcal{U}$.

Proof. If $|\beta| < 1 < |\alpha|$, fix $u \in \mathcal{U}$ and apply (104) inductively to get

$$\alpha^n j(u) = j(\beta^n u) \tag{105}$$

for every $n \in \mathbb{N}$. Letting $n \uparrow \infty$ in (105), we see that $j(u)$ must be zero. If $|\alpha| < 1 < |\beta|$, substitute $v = \beta u$ into (104) to get

$$\alpha j(\beta^{-1} v) = j(v) \tag{106}$$

for every $v, \beta^{-1} v \in \mathcal{U}$. Fix $v \in \mathcal{U}$, and iterate (106) to get

$$\alpha^n j(\beta^{-n} v) = j(v) \tag{107}$$

for every $n \in \mathbb{N}$. Letting $n \uparrow \mathbb{N}$ in (107), we get $j(v) = 0$. □

Setting $x = 0$ in (101) and applying the Lemma gives

$$f(0, z) = 0 \tag{108}$$

for every z near zero. Setting $z = 0$ in (103) and applying the Lemma gives

$$h(x, 0) = 0 \tag{109}$$

for every x near zero. Setting $x = 0$ in (102), using (108), and applying the Lemma gives

$$g(0, z) = 0 \tag{110}$$

for every z near zero. If we set $z = 0$ in (102), use (109), and then differentiate both sides with respect to x, we get $cg_x(x, 0) = g_x(ax, 0)$; applying the Lemma yields

$$g_x(x, 0) = 0 \tag{111}$$

for every x near zero. Setting $z = 0$ in (110) and using (111), we get

$$g(x, 0) = 0 \tag{112}$$

for every x near zero.

Now, using (102) and mathematical induction, it can be verified that

$$a^n c^n [g(x, z) + n\varepsilon f(x, z)h(x, z)] = g(a^n x, c^n z) \tag{113}$$

for every $n \in \mathbb{N}$. Similarly, mathematical induction applied to (101) gives

$$f(x, z) = a^{-n} f(a^n x, c^n z) \tag{114}$$

for every $n \in \mathbb{N}$. If we substitute (114) into (113), divide through by c^{-n}, and replace x by $a^{-n}x$ we get

$$a^n g(a^{-n}x, z) + n\varepsilon f(x, c^n z)h(a^{-n}x, z) = c^{-n} g(x, c^n z) \tag{115}$$

for every $n \in \mathbb{N}$.

The existence of $g_x(0, z)$ and $g_z(0, x)$ along with equations (110) and (112) imply that $a^n g(a^{-n}x, z)$ and $c^{-n} g(x, c^n z)$ stay bounded as $n \uparrow \infty$. Using this fact, and letting $n \uparrow \infty$ in (115), we get

$$f(x, 0)h(0, z) = 0,$$

so $f(x, 0) = 0$ or $h(0, z) = 0$. If $f(x, 0) = 0$, then, in combination with (109) and (112), this tells us that H is not injective in a neighborhood of the origin. Similarly, if $h(0, z) = 0$ then, in combination with (108) and (110), this implies a violation of injectivity, as well.

Poincaré's Linearization Theorem

Suppose that $f : \mathbb{R}^n \to \mathbb{R}^n$ is analytic and satisfies $f(0) = 0$. It is possible to establish conditions under which an *analytic* change of variables will turn the nonlinear equation

$$\dot{x} = f(x) \tag{116}$$

into its linearization

$$\dot{u} = Df(0)u. \tag{117}$$

Definition Let $\lambda_1, \lambda_2, \dots, \lambda_n$ be the eigenvalues of $Df(0)$, listed according to multiplicity. We say that $Df(0)$ is *resonant* if there are nonnegative integers $m_1, m_2, \dots, m_n$ and a number $s \in \{1, 2, \dots, n\}$ such that

$$\sum_{k=1}^{n} m_k \geq 2$$

and

$$\lambda_s = \sum_{k=1}^{n} m_k \lambda_k.$$

If $Df(0)$ is not resonant, we say that it is *nonresonant*.

Note that in Hartman's example there is resonance. As we will see in Math 635, nonresonance permits us to make changes of variable that remove nonlinear terms up to any specified order in the right-hand side of the differential equation. In order to be able to guarantee that *all* nonlinear terms may be removed, some extra condition beyond nonresonance is required.

Definition We say that $(\lambda_1, \lambda_2, \dots, \lambda_n) \in \mathbb{C}^n$ satisfy a *Siegel condition* if there are constants $C > 0$ and $\nu > 1$ such that

$$\left| \lambda_s - \sum_{k=1}^{n} m_k \lambda_k \right| \geq \frac{C}{\left(\sum_{k=1}^{n} m_k\right)^{\nu}}$$

for all nonnegative integers $m_1, m_2, \dots, m_n$ satisfying

$$\sum_{k=1}^{n} m_k \geq 2.$$

Theorem (Poincaré's Linearization Theorem) *Suppose that f is analytic, and that all the eigenvalues of $Df(0)$ are nonresonant and either all lie in the open left half-plane, all lie in the open right half-plane, or satisfy a Siegel condition. Then there is a change of variables $u = g(x)$ that is analytic near* 0 *and that turns* (116) *into* (117) *near* 0.

Stable Manifold Theorem: Part 1

Lecture 31
Math 634
11/10/99

The Hartman-Grobman Theorem states that the flow generated by a smooth vector field in a neighborhood of a hyperbolic equilibrium point is topologically conjugate with the flow generated by its linearization. Hartman's counterexample shows that, in general, the conjugacy cannot be taken to be C^1. However, the Stable Manifold Theorem will tell us that there are important structures for the two flows that can be matched up by smooth changes of variable. In this lecture, we will discuss the Stable Manifold Theorem on an informal level and discuss two different approaches to proving it.

Let $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be C^1, and let $\varphi : \mathbb{R} \times \Omega \to \Omega$ be the flow generated by the differential equation

$$\dot{x} = f(x). \tag{118}$$

Suppose that x_0 is a hyperbolic equilibrium point of (118).

Definition The (global) *stable manifold* of x_0 is the set

$$W^s(x_0) := \left\{ x \in \Omega \;\middle|\; \lim_{t\uparrow\infty} \varphi(t,x) = x_0 \right\}.$$

Definition The (global) *unstable manifold* of x_0 is the set

$$W^u(x_0) := \left\{ x \in \Omega \;\middle|\; \lim_{t\downarrow-\infty} \varphi(t,x) = x_0 \right\}.$$

Definition Given a neighborhood $\mathcal{U}$ of x_0, the local *stable manifold* of x_0 (relative to $\mathcal{U}$) is the set

$$W^s_{\text{loc}}(x_0) := \left\{ x \in \mathcal{U} \;\middle|\; \gamma^+(x) \subset \mathcal{U} \text{ and } \lim_{t\uparrow\infty} \varphi(t,x) = x_0 \right\}.$$

Definition Given a neighborhood $\mathcal{U}$ of x_0, the local *unstable manifold* of x_0 (relative to $\mathcal{U}$) is the set

$$W^u_{\text{loc}}(x_0) := \left\{ x \in \mathcal{U} \;\middle|\; \gamma^-(x) \subset \mathcal{U} \text{ and } \lim_{t\downarrow-\infty} \varphi(t,x) = x_0 \right\}.$$

Note that:

- $W^s_{\text{loc}}(x_0) \subseteq W^s(x_0)$, and $W^u_{\text{loc}}(x_0) \subseteq W^u(x_0)$.
- $W^s_{\text{loc}}(x_0)$ and $W^u_{\text{loc}}(x_0)$ are both nonempty, since they each contain x_0.
- $W^s(x_0)$ and $W^u(x_0)$ are invariant sets.
- $W^s_{\text{loc}}(x_0)$ is positively invariant, and $W^u_{\text{loc}}(x_0)$ is negatively invariant.
- $W^s_{\text{loc}}(x_0)$ is not necessarily $W^s(x_0) \cap \mathcal{U}$, and $W^u_{\text{loc}}(x_0)$ is not necessarily $W^u(x_0) \cap \mathcal{U}$.

$W^s_{\text{loc}}(x_0)$ is not necessarily invariant, since it might not be negatively invariant, and $W^u_{\text{loc}}(x_0)$ is not necessarily invariant, since it might not be positively invariant. They do, however, possess what is known as *relative invariance.*

Definition A subset $\mathcal{A}$ of a set $\mathcal{B}$ is *positively invariant relative to* $\mathcal{B}$ if for every $x \in \mathcal{A}$ and every $t \geq 0$, $\varphi(t, x) \in \mathcal{A}$ whenever $\varphi([0, t], x) \subseteq \mathcal{B}$.

Definition A subset $\mathcal{A}$ of a set $\mathcal{B}$ is *negatively invariant relative to* $\mathcal{B}$ if for every $x \in \mathcal{A}$ and every $t \leq 0$, $\varphi(t, x) \in \mathcal{A}$ whenever $\varphi([t, 0], x) \subseteq \mathcal{B}$.

Definition A subset $\mathcal{A}$ of a set $\mathcal{B}$ is *invariant relative to* $\mathcal{B}$ if it is negatively invariant relative to $\mathcal{B}$ and positively invariant relative to $\mathcal{B}$.

$W^s_{\text{loc}}(x_0)$ is negatively invariant relative to $\mathcal{U}$ and is therefore invariant relative to $\mathcal{U}$. $W^u_{\text{loc}}(x_0)$ is positively invariant relative to $\mathcal{U}$ and is therefore invariant relative to $\mathcal{U}$.

Recall that a *(k-)manifold* is a set that is locally homeomorphic to an open subset of $\mathbb{R}^k$. Although the word "manifold" appeared in the names of $W^s_{\text{loc}}(x_0)$, $W^u_{\text{loc}}(x_0)$, $W^s(x_0)$, and $W^u(x_0)$, it is not obvious from the defintions of these sets that they are, indeed, manifolds. One of the consequences of the Stable Manifold Theorem is that, if $\mathcal{U}$ is sufficiently small, $W^s_{\text{loc}}(x_0)$ and $W^u_{\text{loc}}(x_0)$ are manifolds. ($W^s(x_0)$ and $W^u(x_0)$ are what are known as *immersed* manifolds.)

For simplicity, let's now assume that $x_0 = 0$. Let $\mathcal{E}^s$ be the stable subspace of $Df(0)$, and let $\mathcal{E}^u$ be the unstable subspace of $Df(0)$. If f is linear, then $W^s(0) = \mathcal{E}^s$ and $W^u(0) = \mathcal{E}^u$. The Stable Manifold Theorem says that in the nonlinear case not only are the Stable and Unstable Manifolds indeed manifolds, but they are tangent to $\mathcal{E}^s$ and $\mathcal{E}^u$, respectively, at the origin. This is information that the Hartman-Grobman Theorem does not provide.

More precisely there are neighborhoods $\mathcal{U}^s$ of the origin in $\mathcal{E}^s$ and $\mathcal{U}^u$ of the origin in $\mathcal{E}^u$ and smooth maps $h_s : \mathcal{U}^s \to \mathcal{U}^u$ and $h_u : \mathcal{U}^u \to \mathcal{U}^s$ such that $h_s(0) = Dh_s(0) = h_u(0) = Dh_u(0) = 0$ and the local stable and unstable manifolds of 0 relative to $\mathcal{U}^s \oplus \mathcal{U}^u$ satisfy

$$W^s_{\text{loc}}(0) = \left\{ x + h_s(x) \;\middle|\; x \in \mathcal{U}^s \right\}$$

and

$$W^u_{\text{loc}}(0) = \left\{ x + h_u(x) \;\middle|\; x \in \mathcal{U}^u \right\}.$$

Furthermore, not only do solutions of (118) in the stable manifold converge to 0 as $t \uparrow \infty$, they do so exponentially quickly. (A similar statement can be made about the unstable manifold.)

Liapunov-Perron Approach

This approach to proving the Stable Manifold Theorem rewrites (118) as

$$\dot{x} = Ax + g(x), \tag{119}$$

where $A = Df(0)$. The Variation of Parameters formula gives

$$x(t_2) = e^{(t_2 - t_1)A} x(t_1) + \int_{t_1}^{t_2} e^{(t_2 - s)A} g(x(s))\, ds, \tag{120}$$

for every $t_1, t_2 \in \mathbb{R}$. Setting $t_1 = 0$ and $t_2 = t$, and projecting (120) onto $\mathcal{E}^s$ yields

$$x_s(t) = e^{tA_s} x_s(0) + \int_0^t e^{(t-s)A_s} g_s(x(s))\, ds,$$

where the subscript s attached to a quantity denotes the projection of that quantity onto $\mathcal{E}^s$. If we assume that the solution $x(t)$ lies on $W^s(0)$ and we set $t_2 = t$ and let $t_1 \uparrow \infty$, and project (120) onto $\mathcal{E}^u$, we get

$$x_u(t) = -\int_t^\infty e^{(t-s)A_u} g_u(x(s))\, ds.$$

Hence, solutions of (119) in $W^s(0)$ satisfy the integral equation

$$x(t) = e^{tA_s}x_s(0) + \int_0^t e^{(t-s)A_s}g_s(x(s))\,ds - \int_t^\infty e^{(t-s)A_u}g_u(x(s))\,ds.$$

Now, fix $a_s \in \mathcal{E}^s$, and define a functional T by

$$(Tx)(t) = e^{tA_s}a_s + \int_0^t e^{(t-s)A_s}g_s(x(s))\,ds - \int_t^\infty e^{(t-s)A_u}g_u(x(s))\,ds.$$

A fixed point x of this functional will solve (119), will have a range contained in the stable manifold, and will satisfy $x_s(0) = a_s$. If we set $h_s(a_s) = x_u(0)$ and define h_s similarly for other inputs, the graph of h_s will be the stable manifold.

Hadamard Approach

The Hadamard approach uses what is known as a graph transform. Here we define a functional not by an integral but by letting the graph of the input function move with the flow φ and selecting the output function to be the function whose graph is the image of the original graph after, say, 1 unit of time has elapsed.

More precisely, suppose h is a function from $\mathcal{E}^s$ to $\mathcal{E}^u$. Define its graph transform $F[h]$ to be the function whose graph is the set

$$\left\{\varphi(1, \xi + h(\xi)) \;\middle|\; \xi \in \mathcal{E}^s\right\}. \tag{121}$$

(That (121) is the graph of a function from $\mathcal{E}^s$ to $\mathcal{E}^u$—if we identify $\mathcal{E}^s \times \mathcal{E}^u$ with $\mathcal{E}^s \oplus \mathcal{E}^u$—is, of course, something that needs to be shown.) Another way of putting this is that for each $\xi \in \mathcal{E}^s$,

$$F[h]((\varphi(1, \xi + h(\xi)))_s) = (\varphi(1, \xi + h(\xi)))_u;$$

in other words,

$$F[h] \circ \pi_s \circ \varphi(1, \cdot) \circ (\mathrm{id} + h) = \pi_u \circ \varphi(1, \cdot) \circ (\mathrm{id} + h),$$

where π_s and π_u are projections onto $\mathcal{E}^s$ and $\mathcal{E}^u$, respectively. A fixed point of the graph transform functional F will be an invariant manifold, and it can be show that it is, in fact, the stable manifold.

Stable Manifold Theorem: Part 2

Lecture 32
Math 634
11/12/99

Statements

Given a normed vector space $\mathcal{X}$ and a positive number r, we let $\mathcal{X}(r)$ stand for the closed ball of radius r centered at 0 in $\mathcal{X}$.

The first theorem refers to the differential equation

$$\dot{x} = f(x). \tag{122}$$

Theorem (Stable Manifold Theorem) *Suppose that Ω is an open neighborhood of the origin in $\mathbb{R}^n$, and $f : \Omega \to \mathbb{R}^n$ is a C^k function ($k \geq 1$) such that 0 is a hyperbolic equilibrium point of (122). Let $\mathcal{E}^s \oplus \mathcal{E}^u$ be the decomposition of $\mathbb{R}^n$ corresponding to the matrix $Df(0)$. Then there is a norm $\|\cdot\|$ on $\mathbb{R}^n$, a number $r > 0$, and a C^kfunction $h : \mathcal{E}^s(r) \to \mathcal{E}^u(r)$ such that $h(0) = Dh(0) = 0$ and such that the local stable manifold $W^s_{\mathrm{loc}}(0)$ of 0 relative to $\mathcal{B}(r) := \mathcal{E}^s(r) \oplus \mathcal{E}^u(r)$ is the set*

$$\left\{ v_s + h(v_s) \;\middle|\; v_s \in \mathcal{E}^s(r) \right\}.$$

Moreover, there is a constant $c > 0$ such that

$$W^s_{\mathrm{loc}}(0) = \left\{ v \in \mathcal{B}(r) \;\middle|\; \gamma^+(v) \subset \mathcal{B}(r) \text{ and } \lim_{t\uparrow\infty} e^{ct}\varphi(t, v) = 0 \right\}.$$

Two immediate and obvious corollaries, which we will not state explicitly, describe the stable manifolds of other equilibrium points (via translation) and describe unstable manifolds (by time reversal).

We will actually prove this theorem by first proving an analogous theorem for maps (much as we did with the Hartman-Grobman Theorem). Given a neighborhood $\mathcal{U}$ of a fixed point p of a map F, we can define the local stable manifold of p (relative to $\mathcal{U}$) as

$$W^s_{\mathrm{loc}}(p) := \left\{ x \in \mathcal{U} \;\middle|\; F^j(x) \in \mathcal{U} \text{ for every } j \in \mathbb{N} \text{ and } \lim_{j\uparrow\infty} F^j(x) = p \right\}.$$

Theorem (Stable Manifold Theorem for Maps) *Suppose that Ω is an open neighborhood of the origin in $\mathbb{R}^n$, and $F : \Omega \to \Omega$ is an invertible C^k function*

($k \geq 1$) for which $F(0) = 0$ and the matrix $DF(0)$ is hyperbolic and invertible. Let $\mathcal{E}^s \oplus \mathcal{E}^u (= \mathcal{E}^- \oplus \mathcal{E}^+)$ be the decomposition of $\mathbb{R}^n$ corresponding to the matrix $DF(0)$. Then there is a norm $\|\cdot\|$ on $\mathbb{R}^n$, a number $r > 0$, a number $\tilde{\mu} \in (0,1)$, and a C^kfunction $h : \mathcal{E}^s(r) \to \mathcal{E}^u(r)$ such that $h(0) = Dh(0) = 0$ and such that the local stable manifold $W^s_{\mathrm{loc}}(0)$ of 0 relative to $\mathcal{B}(r) := \mathcal{E}^s(r) \oplus \mathcal{E}^u(r)$ satisfies

$$\begin{aligned} W^s_{\mathrm{loc}}(0) &= \left\{ v_s + h(v_s) \;\middle|\; v_s \in \mathcal{E}^s(r) \right\} \\ &= \left\{ v \in \mathcal{B}(r) \;\middle|\; F^j(v) \in \mathcal{B}(r) \textit{ for every } j \in \mathbb{N} \right\} \\ &= \left\{ v \in \mathcal{B}(r) \;\middle|\; F^j(v) \in \mathcal{B}(r) \textit{ and } \|F^j(v)\| \leq \tilde{\mu}^j \|v\| \textit{ for every } j \in \mathbb{N} \right\}. \end{aligned}$$

Preliminaries

The proof of the Stable Manifold Theorem for Maps will be broken up into a series of lemmas. Before stating and proving those lemmas, we need to lay a foundation by introducing some terminology and notation and by choosing some constants.

We know that $F(0) = 0$ and $DF(0)$ is hyperbolic. Then $\mathbb{R}^n = \mathcal{E}^s \oplus \mathcal{E}^u$, π_s and π_u are the corresponding projection operators, $\mathcal{E}^s$ and $\mathcal{E}^u$ are invariant under $DF(0)$, and there are constants $\mu < 1$ and $\lambda > 1$ such that all of the eigenvalues of $DF(0)|_{\mathcal{E}^s}$ have magnitude less than μ and all of the eigenvalues of $DF(0)|_{\mathcal{E}^u}$ have magnitude greater than λ.

When we deal with a matrix representation of $DF(q)$, it will be with respect to a basis that consists of a basis for $\mathcal{E}^s$ followed by a basis for $\mathcal{E}^u$. Thus,

$$DF(q) = \left[\begin{array}{c|c} A_{ss}(q) & A_{su}(q) \\ \hline A_{us}(q) & A_{uu}(q) \end{array} \right],$$

where, for example, $A_{su}(q)$ is a matrix representation of $\pi_s DF(q)|_{\mathcal{E}^u}$ in terms of the basis for $\mathcal{E}^u$ and the basis for $\mathcal{E}^s$. Note that, by invariance, $A_{su}(0) = A_{us}(0) = 0$. Furthermore, we can pick our basis vectors so that, with $\|\cdot\|$ being the corresponding Euclidean norm of a vector in $\mathcal{E}^s$ or in $\mathcal{E}^u$,

$$\|A_{ss}(0)\| := \sup_{v_s \neq 0} \frac{\|A_{ss}(0)v_s\|}{\|v_s\|} < \mu$$

and

$$m(A_{uu}(0)) := \inf_{v_u \neq 0} \frac{\|A_{uu}(0)v_u\|}{\|v_u\|} > \lambda.$$

(The functional $m(\cdot)$ defined implicitly in the last formula is sometimes called the *minimum norm* even though it is not a norm.) For a vector in $v \in \mathbb{R}^n$, let $\|v\| = \max\{\|\pi_s v\|, \|\pi_u v\|\}$. This will be the norm on $\mathbb{R}^n$ that will be used throughout the proof. Note that $\mathcal{B}(r) := \mathcal{E}^s(r) \oplus \mathcal{E}^u(r)$ is the closed ball of radius r in $\mathbb{R}^n$ by this norm.

Next, we choose r. Fix $\alpha > 0$. Pick $\varepsilon > 0$ small enough that

$$\mu + \varepsilon\alpha + \varepsilon < 1 < \lambda - \varepsilon/\alpha - 2\varepsilon.$$

Pick $r > 0$ small enough that if $q \in \mathcal{B}(r)$ then

$$\begin{aligned} \|A_{ss}(q)\| &< \mu, \\ m(A_{uu}(q)) &> \lambda, \\ \|A_{su}(q)\| &< \varepsilon, \\ \|A_{us}(q)\| &< \varepsilon, \\ \|DF(q) - DF(0)\| &< \varepsilon, \end{aligned}$$

and $DF(q)$ is invertible. (We can do this since F is C^1, so $DF(\cdot)$ is continuous.)

Now, define

$$W_r^s := \bigcap_{j=0}^{\infty} F^{-j}(\mathcal{B}(r)),$$

and note that W_r^s is the set of all points in $\mathcal{B}(r)$ that produce forward semiorbits (under the discrete dynamical system generated by F) that stay in $\mathcal{B}(r)$ for all forward iterates. By definition, $W_{\text{loc}}^s(0) \subseteq W_r^s$; we will show that these two sets are, in fact, equal.

Two other types of geometric sets play vital roles in the proof: *cones* and *disks*. The cones are of two types: *stable* and *unstable*. The stable cone (of "slope" α) is

$$C^s(\alpha) := \big\{ v \in \mathbb{R}^n \bigm| \|\pi_u v\| \leq \alpha \|\pi_s v\| \big\},$$

and the unstable cone (of "slope" α) is

$$C^u(\alpha) := \big\{ v \in \mathcal{R}^n \bigm| \|\pi_u v\| \geq \alpha \|\pi_s v\| \big\}.$$

An *unstable disk* is a set of the form

$$\big\{ v_u + \psi(v_u) \bigm| v_u \in \mathcal{E}^u(r) \big\}$$

for some Lipschitz continuous function $\psi : \mathcal{E}^u(r) \to \mathcal{E}^s(r)$ with Lipschitz constant (less than or equal to) α^{-1}.

Stable Manifold Theorem: Part 3

Lecture 33
Math 634
11/15/99

The Action of $DF(p)$ on the Unstable Cone

The first lemma shows that if the derivative of the map is applied to a point in the unstable cone, the image is also in the unstable cone.

Lemma (Linear Invariance of the Unstable Cone) *If $p \in \mathcal{B}(r)$, then*

$$DF(p)C^u(\alpha) \subseteq C^u(\alpha).$$

Proof. Let $p \in \mathcal{B}(r)$ and $v \in C^u(\alpha)$. Then, if we let $v_s = \pi_s v$ and $v_u = \pi_u v$, we have $\|v_u\| \geq \alpha\|v_s\|$, so

$$\begin{aligned}\|\pi_u DF(p)v\| &= \|A_{us}(p)v_s + A_{uu}(p)v_u\| \geq \|A_{uu}(p)v_u\| - \|A_{us}(p)v_s\| \\ &\geq m(A_{uu}(p))\|v_u\| - \|A_{us}(p)\|\|v_s\| \geq \lambda\|v_u\| - \varepsilon\|v_s\| \\ &\geq (\lambda - \varepsilon/\alpha)\|v_u\|,\end{aligned}$$

and

$$\begin{aligned}\|\pi_s DF(p)v\| &= \|A_{ss}(p)v_s + A_{su}(p)v_u\| \leq \|A_{ss}(p)v_s\| + \|A_{su}(p)v_u\| \\ &\leq \|A_{ss}(p)\|\|v_s\| + \|A_{su}(p)\|\|v_u\| \leq \mu\|v_s\| + \varepsilon\|v_u\| \\ &\leq (\mu/\alpha + \varepsilon)\|v_u\|.\end{aligned}$$

Since $\lambda - \varepsilon/\alpha \geq \alpha(\mu/\alpha + \varepsilon)$,

$$\|\pi_u DF(p)v\| \geq \alpha\|\pi_s DF(p)v\|,$$

so $DF(p)v \in C^u(\alpha)$. □

The Action of F on Moving Unstable Cones

The main part of the second lemma is that moving unstable cones are positively invariant. More precisely, if two points are in $\mathcal{B}(r)$ and one of the two points is in a translate of the unstable cone that is centered at the second point, then their images under F satisfy the same relationship. The lemma also provides estimates on the rates at which the stable and unstable parts of the difference between the two points contract or expand, respectively.

In this lemma (and later) we use the convention that if $\mathcal{X}$ and $\mathcal{Y}$ are subsets of a vector space, then

$$\mathcal{X} + \mathcal{Y} := \big\{x + y \;\big|\; x \in \mathcal{X} \text{ and } y \in \mathcal{Y}\big\}.$$

Lemma (Moving Unstable Cones) *If $p, q \in \mathcal{B}(r)$ and $q \in \{p\} + C^u(\alpha)$, then:*

(a) $\|\pi_s(F(q) - F(p))\| \le (\mu/\alpha + \varepsilon)\|\pi_u(q - p)\|$;

(b) $\|\pi_u(F(q) - F(p))\| \ge (\lambda - \varepsilon/\alpha - \varepsilon)\|\pi_u(q - p)\|$;

(c) $F(q) \in \{F(p)\} + C^u(\alpha)$.

Proof. We will write differences as integrals (using the Fundamental Theorem of Calculus) and use our estimates on $DF(v)$, for $v \in \mathcal{B}(r)$, to estimate these integrals.

Since $\mathcal{B}(r)$ is convex,

$$\begin{aligned}
&\|\pi_s(F(q) - F(p))\| = \left\|\int_0^1 \frac{d}{dt}\pi_s F(tq + (1-t)p)\,dt\right\| \\
&= \left\|\int_0^1 \pi_s DF(tq + (1-t)p)(q - p)\,dt\right\| \\
&= \left\|\int_0^1 [A_{ss}(tq + (1-t)p)\pi_s(q - p) + A_{su}(tq + (1-t)p)\pi_u(q - p)]\,dt\right\| \\
&\le \int_0^1 [\|A_{ss}(tq + (1-t)p)\|\|\pi_s(q - p)\| + \|A_{su}(tq + (1-t)p)\|\|\pi_u(q - p)\|]\,dt \\
&\le \int_0^1 [\mu\|\pi_s(q - p)\| + \varepsilon\|\pi_u(q - p)\|]\,dt \le (\mu/\alpha + \varepsilon)\|\pi_u(q - p)\|.
\end{aligned}$$

This gives **(a)**.

Similarly,

$$\begin{aligned}
&\|\pi_u(F(q)-F(p))\| \\
&= \left\| \int_0^1 [A_{us}(tq+(1-t)p)\pi_s(q-p) + A_{uu}(tq+(1-t)p)\pi_u(q-p)]\,dt \right\| \\
&\geq \left\| \int_0^1 A_{uu}(0)\pi_u(q-p)\,dt \right\| - \left\| \int_0^1 [A_{us}(tq+(1-t)p)\pi_s(q-p)\,dt \right\| \\
&\qquad - \left\| \int_0^1 (A_{uu}(tq+(1-t)p) - A_{uu}(0))\pi_u(q-p)\,dt \right\| \\
&\geq m(A_{uu}(0))\|\pi_u(q-p)\| - \int_0^1 \|A_{us}(tq+(1-t)p)\|\|\pi_s(q-p)\|\,dt \\
&\qquad - \int_0^1 \|A_{uu}(tq+(1-t)p) - A_{uu}(0)\|\|\pi_u(q-p)\|\,dt \\
&\geq \lambda\|\pi_u(q-p)\| - \varepsilon\|\pi_s(q-p)\| - \varepsilon\|\pi_u(q-p)\| \geq (\lambda - \varepsilon/\alpha - \varepsilon)\|\pi_u(q-p)\|.
\end{aligned}$$

This gives **(b)**.

From **(a)**, **(b)**, and the choice of ε, we have

$$\begin{aligned}
\|\pi_u(F(q)-F(p))\| &\geq (\lambda - \varepsilon/\alpha - \varepsilon)\|\pi_u(q-p)\| \geq (\mu + \varepsilon\alpha)\|\pi_u(q-p)\| \\
&\geq \alpha\|\pi_s(F(q)-F(p))\|,
\end{aligned}$$

so $F(q) - F(p) \in C^u(\alpha)$, which means that **(c)** holds.

□

Stable Manifold Theorem: Part 4

Lecture 34
Math 634
11/17/99

Stretching of C^1 Unstable Disks

The next lemma shows that if F is applied to a C^1 unstable disk (*i.e.*, an unstable disk that is the graph of a C^1 function), then part of the image gets stretched out of $\mathcal{B}(r)$, but the part that remains in is again a C^1 unstable disk.

Lemma (Unstable Disks) *Let $\mathcal{D}_0$ be a C^1 unstable disk, and recursively define*

$$\mathcal{D}_j = F(D_{j-1}) \cap \mathcal{B}(r)$$

for each $j \in \mathbb{N}$. Then each $\mathcal{D}_j$ is a C^1 unstable disk, and

$$\operatorname{diam}\left(\pi_u \bigcap_{i=0}^{j} F^{-i}(\mathcal{D}_i)\right) \leq 2(\lambda - \varepsilon/\alpha - \varepsilon)^{-j} r \tag{123}$$

for each $j \in \mathbb{N}$.

Proof. Because of induction, we only need to handle the case $j = 1$. The estimate on the diameter of the π_u projection of the preimage of $\mathcal{D}_1$ under F is a consequence of part **(b)** of the lemma on moving invariant cones. That $\mathcal{D}_1$ is the graph of an α^{-1}-Lipschitz function ψ_1 from a subset of $\mathcal{E}^u(r)$ to $\mathcal{E}^s(r)$ is a consequence of part **(c)** of that same lemma. Thus, all we need to show is that $\operatorname{dom}(\psi_1) = \mathcal{E}^u(r)$ and that ψ_1 is C^1.

Let $\psi_0 : \mathcal{E}^u(r) \to \mathcal{E}^s(r)$ be the C^1 function (with Lipschitz constant less than or equal to α^{-1}) such that

$$\mathcal{D}_0 = \bigl\{ v_u + \psi_0(v_u) \bigm| v_u \in \mathcal{E}^u(r) \bigr\}.$$

Define $g : \mathcal{E}^u(r) \to \mathcal{E}^u$ by the formula $g(v_u) = \pi_u F(v_u + \psi_0(v_u))$. If we can show that for each $y \in \mathcal{E}^u(r)$ there exists $x \in \mathcal{E}^u(r)$ such that

$$g(x) = y, \tag{124}$$

then we will know that $\operatorname{dom}(\psi_1) = \mathcal{E}^u(r)$.

Let $y \in \mathcal{E}^u(r)$ be given. Let $L = A_{uu}(0)$. Since $m(L) > \lambda$, we know that $L^{-1} \in \mathcal{L}(\mathcal{E}^u, \mathcal{E}^u)$ exists and that $\|L^{-1}\| \leq 1/\lambda$. Define $G : \mathcal{E}^u(r) \to \mathcal{E}^u$ by the formula $G(x) = x - L^{-1}(g(x) - y)$, and note that fixed points of G are solutions of (124), and vice versa. We shall show that G is a contraction and takes the compact set $\mathcal{E}^u(r)$ into itself and that, therefore, (124) has a solution $x \in \mathcal{E}^u(r)$.

Note that

$$\begin{aligned} Dg(x) &= \pi_u DF(x + \psi_0(x))(I + D\psi_0(x)) \\ &= A_{uu}(x + \psi_0(x)) + A_{us}(x + \psi_0(x))D\psi_0(x), \end{aligned}$$

so

$$\begin{aligned} \|DG(x)\| = \|I - L^{-1}Dg(x)\| &\leq \|L^{-1}\|\|L - Dg(x)\| \\ &\leq \frac{1}{\lambda}(\|A_{uu}(x + \psi_0(x)) - A_{uu}(0)\| + \|A_{us}(x + \psi_0(x))\|\|D\psi_0(x)\|) \\ &\leq \frac{\varepsilon + \varepsilon/\alpha}{\lambda} < 1. \end{aligned}$$

The Mean Value Theorem then implies that G is a contraction.

Now, suppose that $x \in \mathcal{E}^u(r)$. Then

$$\begin{aligned} \|G(x)\| \leq \|G(0)\| + \|G(x) - G(0)\| &\leq \|L^{-1}\|(\|g(0)\| + \|y\|) + \frac{\varepsilon + \varepsilon/\alpha}{\lambda}\|x\| \\ &\leq \frac{1}{\lambda}(\|g(0)\| + r + (\varepsilon + \varepsilon/\alpha)r). \end{aligned}$$

Let $\rho : \mathcal{E}^s(r) \to \mathcal{E}^u(r)$ be defined by the formula $\rho(v_s) = \pi_u F(v_s)$. Since $\rho(0) = 0$ and, for any $v_s \in \mathcal{E}^s(r)$, $\|D\rho(v_s)\| = \|A_{us}(v_s)\| \leq \varepsilon$, the Mean Value Theorem tells us that

$$\|g(0)\| = \|\pi_u F(\psi_0(0))\| = \|\rho(\psi_0(0))\| \leq \varepsilon\|\psi_0(0)\| \leq \varepsilon r. \tag{125}$$

Plugging (125) into the previous estimate, we see that

$$\|G(x)\| \leq \frac{1}{\lambda}(\varepsilon r + r + (\varepsilon + \varepsilon/\alpha)r) = \frac{1 + \varepsilon/\alpha + 2\varepsilon}{\lambda}r < r,$$

so $G(x) \in \mathcal{E}^u(r)$.

That completes the verification that (124) has a solution for each $y \in \mathcal{E}^u(r)$ and, therefore, that $\operatorname{dom}(\psi_1) = \mathcal{E}^u(r)$. To finish the proof, we need to show that ψ_1 is C^1. Let $\tilde{g}$ be the restriction of g to $g^{-1}(\mathcal{D}_1)$, and observe that

$$\psi_1 \circ \tilde{g} = \pi_s \circ F \circ (I + \psi_0). \tag{126}$$

We have shown that $\tilde{g}$ is a bijection of $g^{-1}(\mathcal{D}_1)$ with $\mathcal{D}_1$ and, by the Inverse Function Theorem, $\tilde{g}^{-1}$ is C^1. Thus, if we rewrite (126) as

$$\psi_1 = \pi_s \circ F \circ (I + \psi_0) \circ \tilde{g}^{-1}$$

we can see that ψ_1, as the composition of C^1 functions, is indeed C^1. □

W_r^s is a Lipschitz Manifold

Recall that W_r^s was defined to be all points in the box $\mathcal{B}(r)$ that produced forward orbits that remain confined within $\mathcal{B}(r)$. The next lemma shows that this set is a manifold.

Lemma (Nature of W_r^s) *W_r^s is the graph of a function $h : \mathcal{E}^s(r) \to \mathcal{E}^u(r)$ that satisfies $h(0) = 0$ and that has a Lipschitz constant less than or equal to α.*

Proof. For each $v_s \in \mathcal{E}^u(r)$, consider the set

$$\mathcal{D} := \{v_s\} + \mathcal{E}^u(r).$$

$\mathcal{D}$ is a C^1 unstable disk, so by the lemma on unstable disks, the subset $\mathcal{S}_j$ of $\mathcal{D}$ that stays in $\mathcal{B}(r)$ for at least j iterations of F has a diameter less than or equal to $2(\lambda - \varepsilon/\alpha - \varepsilon)^{-j}r$. By the continuity of F, $\mathcal{S}_j$ is closed. Hence, the subset $\mathcal{S}_\infty$ of $\mathcal{D}$ that stays in $\mathcal{B}(r)$ for an unlimited number of iterations of F is the intersection of a nested collection of closed sets whose diameters approach 0. This means that $\mathcal{S}_\infty$ is a singleton. Call the single point in $\mathcal{S}_\infty$ $h(v_s)$.

It should be clear that W_r^s is the graph of h. That $h(0) = 0$ follows from the fact that $0 \in W_r^s$, since $F(0) = 0$. If h weren't α-Lipschitz, then there would be two points $p, q \in W_r^s$ such that $p \in \{q\} + C^u(\alpha)$. Repeated application of parts **(b)** and **(c)** of the lemma on moving unstable cones would imply that either $F^j(p)$ or $F^j(q)$ is outside of $\mathcal{B}(r)$ for some $j \in \mathbb{N}$, contrary to definition. □

$W^s_{\text{loc}}(0)$ is a Lipschitz Manifold

Our next lemma shows that $W^s_{\text{loc}}(0) = W^s_r$ and that, in fact, orbits in this set converge to 0 exponentially. (The constant $\tilde{\mu}$ in the statement of the theorem can be chosen to be $\mu + \varepsilon$ if $\alpha \leq 1$.)

Lemma (Exponential Decay) *If $\alpha \leq 1$, then for each $p \in W^s_r$,*

$$\|F^j(p)\| \leq (\mu + \varepsilon)^j \|p\|. \tag{127}$$

In particular, $W^s_r = W^s_{\text{loc}}(0)$.

Proof. Suppose that $\alpha \leq 1$ and $p \in W^s_r$. By mathematical induction (and the positive invariance of W^s_r), it suffices to verify (127) for $j = 1$. Estimating, we find that

$$\begin{aligned}
\|F(p)\| \leq \|\pi_s F(p)\| &= \left\| \int_0^1 \frac{d}{dt} \pi_s F(tp)\, dt \right\| = \left\| \int_0^1 \pi_s DF(tp)p\, dt \right\| \\
&= \left\| \int_0^1 A_{ss}(tp)\pi_s p + A_{su}(tp)\pi_u p\, dt \right\| \\
&\leq \int_0^1 [\|A_{ss}(tp)\| \|\pi_s p\| + \|A_{su}(tp)\| \|\pi_u p\|]\, dt \\
&\leq \mu \|\pi_s p\| + \varepsilon \|\pi_u p\| \leq (\mu + \varepsilon)\|p\|.
\end{aligned}$$

□

Stable Manifold Theorem: Part 5

Lecture 35
Math 634
11/19/99

$W^s_{\text{loc}}(0)$ is C^1

Lemma (Differentiability) *The function $h : \mathcal{E}^s(r) \to \mathcal{E}^u(r)$ for which*

$$W^s_{\text{loc}}(0) = \{v_s + h(v_s) \mid v_s \in \mathcal{E}^s(r)\}$$

is C^1, and $Dh(0) = 0$.

Proof. Let $q \in W^s_r$ be given. We will first come up with a candidate for a plane that is tangent to W^s_r at q, and then we will show that it really is.

For each $j \in \mathbb{N}$ and each $p \in W^s_r$, define

$$C^{s,j}(p) := [D(F^j)(p)]^{-1} C^s(\alpha),$$

and let

$$C^{s,0}(p) := C^s(\alpha).$$

By definition (and by the invertibility of $DF(v)$ for all $v \in \mathcal{B}(r)$), $C^{s,j}(p)$ is the image of the stable cone under an invertible linear transformation. Note that

$$C^{s,1}(p) = [DF(p)]^{-1} C^s(\alpha) \subset C^s(\alpha) = C^{s,0}(p)$$

by the (proof of the) lemma on linear invariance of the unstable cone. Similarly,

$$\begin{aligned} C^{s,2}(p) &= [D(F^2)(p)]^{-1} C^s(\alpha) = [DF(F(p))DF(p)]^{-1} C^s(\alpha) \\ &= [DF(p)]^{-1}[DF(F(p))]^{-1} C^s(\alpha) = [DF(p)]^{-1} C^{s,1}(F(p)) \\ &\subset [DF(p)]^{-1} C^s(\alpha) = C^{s,1}(p) \end{aligned}$$

and

$$\begin{aligned} C^{s,3}(p) &= [D(F^3)(p)]^{-1} C^s(\alpha) = [DF(F^2(p))DF(F(p))DF(p)]^{-1} C^s(\alpha) \\ &= [DF(p)]^{-1}[DF(F(p))]^{-1}[DF(F^2(p))]^{-1} C^s(\alpha) \\ &= [DF(p)]^{-1}[DF(F(p))]^{-1} C^{s,1}(F^2(p)) \\ &\subset [DF(p)]^{-1}[DF(F(p))]^{-1} C^s(\alpha) = C^{s,2}(p). \end{aligned}$$

Recursively, we find that, in particular,

$$C^{s,0}(q) \supset C^{s,1}(q) \supset C^{s,2}(q) \supset C^{s,3}(q) \supset \cdots .$$

The plane that we will show is the tangent plane to W_r^s at q is the intersection

$$C^{s,\infty}(q) := \bigcap_{j=0}^{\infty} C^{s,j}(q)$$

of this nested sequence of “cones”.

First, we need to show that this intersection *is* a plane. Suppose that $x \in C^{s,j}(q)$. Then $x \in C^s(\alpha)$, so

$$\begin{aligned} \|\pi_s DF(q)x\| &= \|A_{ss}(q)\pi_s x + A_{su}(q)\pi_u x\| \leq \|A_{ss}(q)\|\|\pi_s x\| + \|A_{su}(q)\|\|\pi_u x\| \\ &\leq (\mu + \varepsilon\alpha)\|\pi_s x\|. \end{aligned}$$

Repeating this sort of estimate, we find that

$$\begin{aligned} \|\pi_s D(F^j)(q)x\| &= \|\pi_s DF(F^{j-1}(q))DF(F^{j-2}(q))\cdots DF(q)x\| \\ &\leq (\mu + \varepsilon\alpha)^j \|\pi_s x\|. \end{aligned}$$

On the other hand, if y is also in $C^{s,j}(q)$ and $\pi_s x = \pi_s y$, then repeated applications of the estimates in the lemma on linear invariance of the unstable cone yield

$$\|\pi_u D(F^j)(q)x - \pi_u D(F^j)(q)y\| \geq (\lambda - \varepsilon/\alpha)^j \|\pi_u x - \pi_u y\|.$$

Since $D(F^j)(q)C^{s,j}(q) = C^s(\alpha)$, it must, therefore, be the case that

$$\frac{(\lambda - \varepsilon/\alpha)^j \|\pi_u x - \pi_u y\|}{(\mu + \varepsilon\alpha)^j \|\pi_s x\|} \leq 2\alpha.$$

This implies that

$$\|\pi_u x - \pi_u y\| \leq 2\alpha \left(\frac{\mu + \varepsilon\alpha}{\lambda - \varepsilon/\alpha} \right)^j \|\pi_s x\|. \tag{128}$$

Letting $j \uparrow \infty$ in (128), we see that for each $v_s \in \mathcal{E}^s$ there can be no more than 1 point x in $C^{s,\infty}(q)$ satisfying $\pi_s x = v_s$. On the other hand, each $C^{s,j}(q)$ contains a plane of dimension $\dim(\mathcal{E}^s)$ (namely, the preimage of $\mathcal{E}^s$ under $D(F^j)(q)$), so (since the set of planes of that dimension passing through the origin is a compact set in the natural topology), $C^{s,\infty}(q)$ contains a plane, as well. This means that $C^{s,\infty}(q)$ is a plane $\mathcal{P}_q$ that is the graph of a linear function $L_q : \mathcal{E}^s \to \mathcal{E}^u$.

Before we show that $L_q = Dh(q)$, we make a few remarks.

(a) Because $\mathcal{E}^s \subset C^{s,j}(0)$ for every $j \in \mathbb{N}$, $\mathcal{P}_0 = \mathcal{E}^s$ and $L_0 = 0$.

(b) The estimate (128) shows that the size of the largest angle between two vectors in $C^{s,j}(q)$ having the same projection onto $\mathcal{E}^s$ goes to zero as $j \uparrow \infty$.

(c) Also, the estimates in the proof of the lemma on linear invariance of the unstable cone show that the size of the minimal angle between a vector in $C^{s,1}(F^j(q))$ and a vector outside of $C^{s,0}(F^j(q))$ is bounded away from zero. Since

$$C^{s,j}(q) = [D(F^j)(q)]^{-1}C^s(\alpha) = [D(F^j)(q)]^{-1}C^{s,0}(F^j(q))$$

and

$$\begin{aligned} C^{s,j+1}(q) &= [D(F^{j+1})(q)]^{-1}C^s(\alpha) = [D(F^j)(q)]^{-1}[DF(F^j(q))]^{-1}C^s(\alpha) \\ &= [D(F^j)(q)]^{-1}C^{s,1}(F^j(q)), \end{aligned}$$

this means that the size of the minimal angle between a vector in $C^{s,j+1}(q)$ and a vector outside of $C^{s,j}(q)$ is also bounded away from zero.

(d) Thus, since $C^{s,j+1}(q)$ depends continuously on q,

$$\mathcal{P}_{q'} \in C^{s,j+1}(q') \subset C^{s,j}(q)$$

for a given j if q' is sufficiently close to q. This means that $\mathcal{P}_q$ depends continuously on q.

Now, we show that $DF(q) = L_q$. Let $\varepsilon > 0$ be given. By remark **(b)** above, we can choose $j \in \mathbb{N}$ such that

$$\|\pi_u v - L_q \pi_s v\| \le \varepsilon \|\pi_s v\| \tag{129}$$

whenever $v \in C^{s,j}(q)$. By remark **(c)** above, we know that we can choose $\varepsilon' > 0$ such that if $w \in C^{s,j+1}(q)$ and $\|r\| \le \varepsilon'\|w\|$, then $w + r \in C^{s,j}(q)$. Because of the differentiability of F^{-j-1}, we can choose $\eta > 0$ such that

$$\|F^{-j-1}(F^{j+1}(q) + v) - q - [D(F^{-j-1})(F^{j+1}(q))]v\| \le \frac{\varepsilon'}{\|D(F^{j+1})(q)\|}\|v\| \tag{130}$$

whenever $\|v\| \leq \eta$. Define the truncated stable cone

$$C^s(\alpha, \eta) := C^s(\alpha) \cap \pi_s^{-1}\mathcal{E}^s(\eta).$$

From the continuity of F and the α-Lipschitz continuity of h, we know that we can pick $\delta > 0$ such that

$$F^{j+1}(v_s + h(v_s)) \in \{F^{j+1}(q)\} + C^s(\alpha, \eta). \tag{131}$$

whenever $\|v_s - \pi_s q\| < \delta$.

Now, suppose that $v \in C^s(\alpha, \eta)$. Then (assuming $\alpha \leq 1$) we know that $\|v\| \leq \eta$, so (130) tells us that

$$\begin{aligned} F^{-j-1}(F^{j+1}(q) + v) &= q + [D(F^{-j-1})(F^{j+1}(q))]v + r \\ &= q + [D(F^{j+1})(q)]^{-1}v + r \end{aligned} \tag{132}$$

for some r satisfying

$$\|r\| \leq \frac{\varepsilon'}{\|D(F^{j+1})(q)\|}\|v\|.$$

Let $w = [D(F^{j+1})(q)]^{-1}v$. Since $v \in C^s(\alpha)$, $w \in C^{s,j+1}(q)$. Also,

$$\|w\| = \|[D(F^{j+1})(q)]^{-1}v\| \geq m([D(F^{j+1})(q)]^{-1})\|v\| = \frac{\|v\|}{\|D(F^{j+1})(q)\|},$$

so $\|r\| \leq \varepsilon'\|w\|$. Thus, by the choice of ε', $w + r \in C^{s,j}(q)$. Consequently, (132) implies that

$$F^{-j-1}(F^{j+1}(q) + v) \in \{q\} + C^{s,j}(q).$$

Since v was an arbitrary element of $C^s(\alpha, \eta)$, we have

$$F^{-j-1}(\{F^{j+1}(q)\} + C^s(\alpha, \eta)) \subseteq \{q\} + C^{s,j}(q). \tag{133}$$

Set $q_s := \pi_s q$, and suppose that $v_s \in \mathcal{E}^s(r)$ satisfies $\|v_s - q_s\| \leq \delta$. By (131),

$$F^{j+1}(v_s + h(v_s)) \in \{F^{j+1}(q)\} + C^s(\alpha, \eta).$$

This, the invertibility of F, and (133) imply

$$v_s + h(v_s) \in \{q\} + C^{s,j}(q),$$

or, in other words,

$$v_s + h(v_s) - q_s - h(q_s) \in C^{s,j}(q).$$

The estimate (129) then tells us that

$$\|h(v_s) - h(q_s) - L_q(v_s - q_s)\| \leq \varepsilon \|v_s - q_s\|,$$

which proves that $Dh(q) = L_q$ (since ε was arbitrary).

Remark **(d)** above implies that $Dh(q)$ depends continuously on q, so $h \in C^1$. Remark **(a)** above implies that $Dh(0) = 0$. □

Stable Manifold Theorem: Part 6

Lecture 36
Math 634
11/22/99

Higher Differentiability

Lemma (Higher Differentiability) *If F is C^k, then h is C^k.*

Proof. We've already seen that this holds for $k = 1$. We show that it is true for all k by induction. Let $k \geq 2$, and assume that the lemma works for $k-1$. Define a new map $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$ by the formula

$$H\left(\begin{bmatrix} p \\ v \end{bmatrix}\right) := \begin{bmatrix} F(p) \\ DF(p)v \end{bmatrix}.$$

Since F is C^k, H is C^{k-1}. Note that

$$H^2\left(\begin{bmatrix} p \\ v \end{bmatrix}\right) = \begin{bmatrix} F(F(p)) \\ DF(F(p))DF(p)v \end{bmatrix} = \begin{bmatrix} F^2(p) \\ D(F^2)(p)v \end{bmatrix},$$

$$H^3\left(\begin{bmatrix} p \\ v \end{bmatrix}\right) = \begin{bmatrix} F(F^2(p)) \\ DF(F^2(p))D(F^2)(p)v \end{bmatrix} = \begin{bmatrix} F^3(p) \\ D(F^3)(p)v \end{bmatrix},$$

and, in general,

$$H^j\left(\begin{bmatrix} p \\ v \end{bmatrix}\right) = \begin{bmatrix} F^j(p) \\ D(F^j)(p)v \end{bmatrix}.$$

Also,

$$DH\left(\begin{bmatrix} p \\ v \end{bmatrix}\right) = \left[\begin{array}{c|c} DF(p) & 0 \\ \hline D^2F(p)v & DF(p) \end{array}\right],$$

so

$$DH\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = \left[\begin{array}{c|c} DF(0) & 0 \\ \hline 0 & DF(0) \end{array}\right],$$

which is hyperbolic and invertible, since $DF(0)$ is. Applying the induction hypothesis, we can conclude that the fixed point of H at the origin (in $\mathbb{R}^n \times \mathbb{R}^n$) has a local stable manifold $\mathcal{W}$ that is C^{k-1}.

Fix $q \in W_r^s$, and note that $F^j(q) \to 0$ as $j \uparrow \infty$ and

$$\mathcal{P}_q = \left\{ v \in \mathbb{R}^n \;\middle|\; \lim_{j\uparrow\infty} D(F^j)(q)v = 0 \right\}.$$

This means that

$$\mathcal{P}_q = \left\{ v \in \mathbb{R}^n \;\middle|\; \begin{bmatrix} q \\ v \end{bmatrix} \in \mathcal{W} \right\}.$$

Since $\mathcal{W}$ has a C^{k-1} dependence on q, so does $\mathcal{P}_q$. Hence, h is C^k. □

Flows

Now we discuss how the Stable Manifold Theorem for maps implies the Stable Manifold Theorem for flows. Given $f : \Omega \to \mathbb{R}^n$ satisfying $f(0) = 0$, let $F = \varphi(1, \cdot)$, where φ is the flow generated by the differential equation

$$\dot{x} = f(x). \tag{134}$$

If f is C^k, so is F. Clearly, F is invertible and $F(0) = 0$. Our earlier discussion on differentiation with respect to initial conditions tells us that

$$\frac{d}{dt} D_x\varphi(t, x) = Df(\varphi(t, x))D_x\varphi(t, x)$$

and $D_x\varphi(0, x) = I$, where D_x represents differentiation with respect to x. Setting

$$g(t) = D_x\varphi(t, x)\big|_{x=0},$$

this implies, in particular, that

$$\frac{d}{dt} g(t) = Df(0)g(t)$$

and $g(0) = I$, so

$$g(t) = e^{tDf(0)}.$$

Setting $t = 1$, we see that

$$e^{Df(0)} = g(1) = D_x\varphi(1, x)|_{x=0} = D_x F(x)|_{x=0} = DF(0).$$

Thus, $DF(0)$ is invertible, and if (134) has a hyperbolic equilibrium at the origin then $DF(0)$ is hyperbolic.

Since F satisfies the hypotheses of the Stable Manifold Theorem for maps, we know that F has a local stable manifold W_r^s on some box $\mathcal{B}(r)$. Assume that $\alpha < 1$ and that r is small enough that the vector field of (134) points *into* $\mathcal{B}(r)$ on $C^s(\alpha) \cap \partial\mathcal{B}(r)$. (See the estimates in Lecture 21.) The requirements for a point to be in W_r^s are no more restrictive then the requirements to be in the local stable manifold $\mathcal{W}_r^s$ of the origin with respect to the flow, so $\mathcal{W}_r^s \subseteq W_r^s$.

We claim that, in fact, these two sets are equal. Suppose they are not. Then there is a point $q \in W_r^s \backslash \mathcal{W}_r^s$. Let $x(t)$ be the solution of (134) satisfying $x(0) = q$. Since $\lim_{j\uparrow\infty} F^j(q) = 0$ and, in a neighborhood of the origin, there is a bound on the factor by which $x(t)$ can grow in 1 unit of time, we know that $x(t) \to 0$ as $t \uparrow \infty$. Among other things, this implies that

(a) $x(t) \notin W_r^s$ for some $t > 0$, and

(b) $x(t) \in W_r^s$ for all t sufficiently large.

Since W_r^s is a closed set and x is continuous, **(a)** and **(b)** say that we can pick t_0 to be the earliest time such that $x(t) \in W_r^s$ for every $t \geq t_0$.

Now, consider the location of $x(t)$ for t in the time interval $[t_0 - 1, t_0)$. Since $x(0) \in W_r^s$, we know that $x(j) \in W_r^s$ for every $j \in \mathbb{N}$. In particular, we can choose $t_1 \in [t_0 - 1, t_0)$ such that $x(t_1) \in W_r^s$. By definition of t_0, we can choose $t_2 \in (t_1, t_0)$ such that $x(t_2) \notin W_r^s$. By the continuity of x and the closedness of W_r^s, we can pick t_3 to the be the last time before t_2 such that $x(t_3) \in W_r^s$. By definition of W_r^s, if $t \in [t_0 - 1, t_0)$ and $x(t) \notin W_r^s$, then $x(t) \notin \mathcal{B}(r)$; hence, $x(t)$ must leave $\mathcal{B}(r)$ at time $t = t_3$. But this contradicts the fact that the vector field points into $\mathcal{B}(r)$ at $x(t_3)$, since $x(t_3) \in C^s(\alpha) \cap \partial\mathcal{B}(r)$. This contradiction implies that no point $q \in W_r^s \setminus \mathcal{W}_r^s$ exists; *i.e.*, $W_r^s = \mathcal{W}_r^s$.

The exponential decay of solutions of the flow on the local stable manifold is a consequence of the similar decay estimate for the map, along with the observation that, near 0, there is a bound to the factor by which a solution can grown in 1 unit of time.

Center Manifolds

Lecture 37
Math 634
11/29/99

Definition

Recall that for the linear differential equation

$$\dot{x} = Ax \tag{135}$$

the corresponding invariant subspaces $\mathcal{E}^u$, $\mathcal{E}^s$, and $\mathcal{E}^c$ had the characterizations

$$\mathcal{E}^u = \left\{ x \in \mathbb{R}^n \;\middle|\; \exists c > 0 \text{ s.t. } \lim_{t \downarrow -\infty} |e^{-ct}e^{tA}x| = 0 \right\},$$

$$\mathcal{E}^s = \left\{ x \in \mathbb{R}^n \;\middle|\; \exists c > 0 \text{ s.t. } \lim_{t \uparrow \infty} |e^{ct}e^{tA}x| = 0 \right\},$$

and

$$\mathcal{E}^c = \left\{ x \in \mathbb{R}^n \;\middle|\; \forall c > 0, \lim_{t \downarrow -\infty} |e^{ct}e^{tA}x| = 0 \text{ and } \lim_{t \uparrow \infty} |e^{-ct}e^{tA}x| = 0 \right\}.$$

The Stable Manifold Theorem tells us that for the nonlinear differential equation

$$\dot{x} = f(x), \tag{136}$$

with $f(0) = 0$, the stable manifold $W^s(0)$ and the unstable manifold $W^u(0)$ have characterizations similar to $\mathcal{E}^s$ and $\mathcal{E}^u$, respectively:

$$W^s(0) = \left\{ x \in \mathbb{R}^n \;\middle|\; \exists c > 0 \text{ s.t. } \lim_{t \uparrow \infty} |e^{ct}\varphi(t, x)| = 0 \right\},$$

and

$$W^u(0) = \left\{ x \in \mathbb{R}^n \;\middle|\; \exists c > 0 \text{ s.t. } \lim_{t \downarrow -\infty} |e^{-ct}\varphi(t, x)| = 0 \right\},$$

where φ is the flow generated by (136). (This was only verified when the equilibrium point at the origin was hyperbolic, but a similar result holds in general.)

Is there a useful way to modify the characterization of $\mathcal{E}^c$ similarly to get a characterization of a *center manifold* $W^c(0)$? Not really. The main problem is that the characterizations of $\mathcal{E}^s$ and $\mathcal{E}^u$ only depend on the *local* behavior of solutions when they are near the origin, but the characterization of $\mathcal{E}^c$ depends on the behavior of solutions that are, possibly, far from 0.

Still, the idea of a center manifold as some sort of nonlinear analogue of $\mathcal{E}^c(0)$ is useful. Here's one widely-used definition:

Definition Let $A = Df(0)$. A *center manifold* $W^c(0)$ of the equilbrium point 0 of (136) is an invariant manifold whose dimension equals the dimension of the invariant subspace $\mathcal{E}^c$ of (135) and which is tangent to $\mathcal{E}^c$ at the origin.

Nonuniqueness

While the fact that stable and unstable manifolds are really manifolds is a theorem (namely, the Stable Manifold Theorem), a center manifold is a manifold *by definition*. Also, note that we refer to *the* stable manifold and *the* unstable manifold, but we refer to *a* center manifold. This is because center manifolds are not necessarily unique. An extremely simple example of nonuniqueness (commonly credited to Kelley) is the planar system

$$\begin{cases} \dot{x} = x^2 \\ \dot{y} = -y. \end{cases}$$

Clearly, $\mathcal{E}^c$ is the x-axis, and solving the system explicitly reveals that for any constant $c \in \mathbb{R}$ the curve

$$\left\{(x,y) \in \mathbb{R}^2 \mid x < 0 \text{ and } y = ce^{1/x}\right\} \cup \left\{(x,0) \in \mathbb{R}^2 \mid x \geq 0\right\}$$

is a center manifold.

Existence

There is a Center Manifold Theorem just like there was a Stable Manifold Theorem. However, the goal of the Center Manifold Theorem is not to characterize a center manifold; that is done by the definition. The Center Manifold Theorem asserts the *existence* of a center manifold.

We will not state this theorem precisely nor prove it, but we can give some indication how the proof of existence of a center manifold might go. Suppose that none of the eigenvalues of $Df(0)$ have real part equal to α, where α is a given real number. Then we can split the eigenvalues up into two sets: Those with real part less than α and those with real part greater than α. Let $\mathcal{E}^-$ be the vector space spanned by the generalized eigenvectors corresponding to the first set of eigenvalues, and let $\mathcal{E}^+$ be the vector space spanned by the generalized eigenvectors corresponding to the second set of eigenvalues. If we cut off f so that it is stays nearly linear throughout $\mathbb{R}^n$, then an analysis very much like that in the proof of the Stable Manifold Theorem can be done to conclude that there are invariant manifolds called the *pseudo-stable manifold* and the *pseudo-unstable manifold* that are tangent, respectively, to $\mathcal{E}^-$ and $\mathcal{E}^+$ at the origin. Solutions $x(t)$ in the first manifold satisfy $e^{-\alpha t}x(t) \to 0$ as $t \uparrow \infty$, and solutions in the second manifold satisfy $e^{-\alpha t}x(t) \to 0$ as $t \downarrow -\infty$.

Now, suppose that α is chosen to be negative but larger than the real part of the eigenvalues with negative real part. The corresponding pseudo-unstable manifold is called a *center-unstable manifold* and is written $W^{cu}(0)$. If, on the other hand, we choose α to be between zero and all the positive real parts of eigenvalues, then the resulting pseudo-stable manifold is called a *center-stable manifold* and is written $W^{cs}(0)$. It turns out that

$$W^c(0) := W^{cs}(0) \cap W^{cu}(0)$$

is a center manifold.

Center Manifold as a Graph

Since a center manifold $W^c(0)$ is tangent to $\mathcal{E}^c$ at the origin it can, at least locally, be represented as the graph of a function $h : \mathcal{E}^c \to \mathcal{E}^s \oplus \mathcal{E}^u$. Suppose, for simplicity, that (136) can be rewritten in the form

$$\begin{cases} \dot{x} = Ax + F(x, y) \\ \dot{y} = By + G(x, y), \end{cases} \tag{137}$$

where $x \in \mathcal{E}^c$, $y \in \mathcal{E}^s \oplus \mathcal{E}^u$, the eigenvalues of A all have zero real part, all of the eigenvalues of B have nonzero real part, and F and G are higher order terms. Then, for points $x + y$ lying on $W^c(0)$, $y = h(x)$. Inserting that into (137) and using the chain rule, we get

$$Dh(x)[Ax + F(x, h(x))] = Dh(x)\dot{x} = \dot{y} = Bh(x) + G(x, h(x)).$$

Thus, if we define an operator $\mathcal{M}$ by the formula

$$(\mathcal{M}\phi)(x) := D\phi(x)[Ax + F(x, \phi(x))] - B\phi(x) + G(x, \phi(x)),$$

the function h whose graph is the center manifold is a solution of the equation $\mathcal{M}h = 0$.

Computing and Using Center Manifolds

Lecture 38
Math 634
12/1/99

Approximation

Recall that we projected our equation onto $\mathcal{E}^c$ and onto $\mathcal{E}^s \oplus \mathcal{E}^u$ to get the system

$$\begin{cases} \dot{x} = Ax + F(x, y) \\ \dot{y} = By + G(x, y), \end{cases} \tag{138}$$

and that we were looking for a function $h : \mathcal{E}^c \to \mathcal{E}^s \oplus \mathcal{E}^u$ satisfying $(\mathcal{M}h) \equiv 0$, where

$$(\mathcal{M}\phi)(x) := D\phi(x)[Ax + F(x, \phi(x))] - B\phi(x) + G(x, \phi(x)).$$

Except in the simplest of cases we have no hope of trying to get an explicit formula for h, but because of the following theorem of Carr we can approximate h to arbitrarily high orders.

Theorem (Carr) *Let ϕ be a C^1 mapping of a neighborhood of the origin in $\mathbb{R}^n$ into $\mathbb{R}^n$ that satisfies $\phi(0) = 0$ and $D\phi(0) = 0$. Suppose that*

$$(\mathcal{M}\phi)(x) = O(|x|^q)$$

as $x \to 0$ for some constant $q > 1$. Then

$$|h(x) - \phi(x)| = O(|x|^q)$$

as $x \to 0$.

Stability

If we put $y = h(x)$ in the first equation in (137), we get the *reduced equation*

$$\dot{x} = Ax + F(x, h(x)), \tag{139}$$

which describes the evolution of the $\mathcal{E}^c$ coordinate of solutions on the center manifold. Another theorem of Carr's states that if all the eigenvalues of

$Df(0)$ are in the closed left half-plane, then the stability type of the origin as an equilibrium solution of (138) (Lyapunov stable, asymptotically stable, or unstable) matches the stability type of the origin as an equilibrium solution of (139).

These results of Carr are sometimes useful in computing the stability type of the origin. Consider, for example, the following system:

$$\begin{cases} \dot{x} = xy + ax^3 + by^2x \\ \dot{y} = -y + cx^2 + dx^2y, \end{cases}$$

where x and y are real variables and a, b, c, and d are real parameters. We know that there is a center manifold, tangent to the x-axis at the origin, that is (locally) of the form $y = h(x)$. The reduced equation on the center manifold is

$$\dot{x} = xh(x) + ax^3 + b[h(x)]^2x. \tag{140}$$

To determine the stability of the origin in (140) (and, therefore, in the original system) we need to approximate h. Therefore, we consider the operator $\mathcal{M}$ defined by

$$(\mathcal{M}\phi)(x) = \phi'(x)[x\phi(x) + ax^3 + b(\phi(x))^2x] + \phi(x) - cx^2 - dx^2\phi(x),$$

and seek polynomial ϕ (satisfying $\phi(0) = \phi'(0) = 0$) for which $(\mathcal{M}\phi)(x)$ is of high order in x. By inspection, if $\phi(x) = cx^2$ then $(\mathcal{M}\phi)(x) = O(x^4)$, so $h(x) = cx^2 + O(x^4)$, and (140) becomes

$$\dot{x} = (a + c)x^3 + O(x^5).$$

Hence, the origin is asymptotically stable if $a+c < 0$ and is unstable if $a+c > 0$. What about the borderline case when $a+c = 0$? Suppose that $a+c = 0$ and let's go back and try a different ϕ, namely, one of the form $\phi(x) = cx^2 + kx^4$. Plugging this in, we find that $(\mathcal{M}\phi)(x) = (k - cd)x^4 + O(x^6)$, so if we choose $k = cd$ then $(\mathcal{M}\phi)(x) = O(x^6)$; thus, $h(x) = cx^2 + cdx^4 + O(x^6)$. Inserting this in (140), we get

$$\dot{x} = (cd + bc^2)x^5 + O(x^7),$$

so the origin is asymptotically stable if $cd + bc^2 < 0$ (and $a + c = 0$) and is unstable if $cd + bc^2 > 0$ (and $a + c = 0$).

What if $a + c = 0$ *and* $cd + bc^2 = 0$? Suppose that these two conditions hold, and consider ϕ of the form $\phi(x) = cx^2+cdx^4+kx^6$ for some $k \in \mathbb{R}$ yet to be determined. Calculating, we discover that $(\mathcal{M}\phi)(x) = (k-b^2c^3)x^6+O(x^8)$, so by choosing $k = b^2c^3$, we see that $h(x) = cx^2 + cdx^4 + b^2c^3x^6 + O(x^8)$. Inserting this in (140), we see that (if $a + c = 0$ and $cd + bc^2 = 0$)

$$\dot{x} = -b^2c^3x^7 + O(x^9).$$

Hence, if $a + c = cd + bc^2 = 0$ and $b^2c > 0$ then the origin is asymptotically stable, and if $a + c = cd + bc^2 = 0$ and $b^2c < 0$ then the origin is unstable.

It can be checked that in the remaining borderline case $a+c = cd+bc^2 = b^2c = 0$, $h(x) \equiv cx^2$ so the reduced equation is simply $\dot{x} = 0$. Hence, in this case, the origin is Lyapunov stable, but not asymptotically stable.

Bifurcation Theory

Bifurcation theory studies fundamental changes in the structure of the solutions of a differential equation or a dynamical system in response to change in a parameter. Consider the parametrized equation

$$\dot{x} = F(x, \varepsilon), \tag{141}$$

where $x \in \mathbb{R}^n$ is a variable and $\varepsilon \in \mathbb{R}^p$ is a parameter. Suppose that $F(0, \varepsilon) = 0$ for every ε, that the equilibrium solution at $x = 0$ is stable when $\varepsilon = 0$, and that we are interested in the possibility of persistent structures (*e.g.*, equilibria or periodic orbits) bifurcating out of the origin as ε is made nonzero. This means that all the eigenvalues of $D_xF(0, 0)$ have nonpositive real part, so we can project (141) onto complementary subspaces of $\mathbb{R}^n$ and get the equivalent system

$$\begin{cases} \dot{u} = Au + f(u, v, \varepsilon) \\ \dot{v} = Bv + g(u, v, \varepsilon), \end{cases}$$

with the eigenvalues of A lying on the imaginary axis and the eigenvalues of B lying in the open right half-plane. Since the parameter ε does not depend on time, we can append the equation $\dot{\varepsilon} = 0$ to get the expanded system

$$\begin{cases} \dot{u} = Au + f(u, v, \varepsilon) \\ \dot{v} = Bv + g(u, v, \varepsilon) \\ \dot{\varepsilon} = 0. \end{cases} \tag{142}$$

The Center Manifold Theorem asserts the existence of a center manifold for the origin that is locally given by points (u, v, ε) satisfying an equation of the form

$$v = h(u, \varepsilon).$$

Furthermore, according to a theorem of Carr, every solution $(u(t), v(t), \varepsilon)$ of (142) for which $(u(0), v(0), \varepsilon)$ is sufficiently close to zero converges exponentially quickly to a solution on the center manifold as $t \uparrow \infty$. In particular, no persistent structure near the origin lies off the center manifold of this expanded system. Hence, it suffices to consider persistent structures for the lower-dimensional equation

$$\dot{u} = Au + f(u, h(u, \varepsilon), \varepsilon).$$

Poincaré-Bendixson Theorem
Lecture 39
Math 634
12/3/99

Definition A *periodic orbit* of a continuous dynamical system φ is a set of the form

$$\{\varphi(t,p) \mid t \in [0,T]\}$$

for some time T and point p satisfying $\varphi(T,p) = p$. If this set is a singleton, we say that the periodic orbit is *degenerate*.

Theorem (Poincaré-Bendixson) *Every nonempty, compact ω-limit set of a C^1 planar flow that does not contain an equilibrium point is a (nondegenerate) periodic orbit.*

We will prove this theorem by means of 4 lemmas. Throughout our discussion, we will be referring to a C^1 planar flow φ and the corresponding vector field f.

Definition If $\mathcal{S}$ is a line segment in $\mathbb{R}^2$ and $p_1, p_2, \ldots$ is a (possibly finite) sequence of points lying on $\mathcal{S}$, then we say that this sequence is *monotone on $\mathcal{S}$* if $(p_j - p_{j-1}) \cdot (p_2 - p_1) \geq 0$ for every $j \geq 2$.

Definition A (possibly finite) sequence $p_1, p_2, \ldots$ of points on a trajectory $\mathcal{T}$ of φ is said to be *monotone on $\mathcal{T}$* if we can choose a point p and times $t_1 \leq t_2 \leq \cdots$ such that $\varphi(t_j, p) = p_j$ for each j.

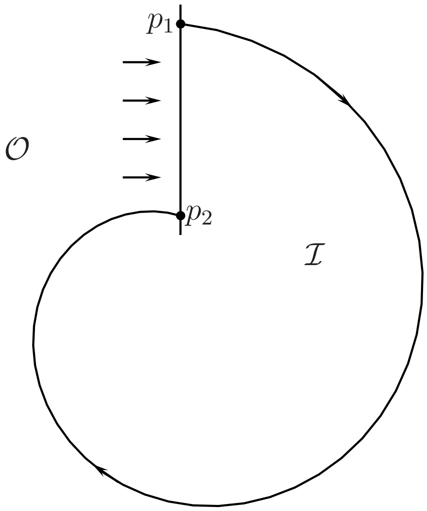
Definition A *transversal* of φ is a line segment $\mathcal{S}$ such that f is not tangent to $\mathcal{S}$ at any point of $\mathcal{S}$.

Lemma *If a (possibly finite) sequence of points $p_1, p_2, \ldots$ lies on the intersection of a transversal $\mathcal{S}$ and a trajectory $\mathcal{T}$, and the sequence is monotone on $\mathcal{T}$, then it is monotone on $\mathcal{S}$.*

Proof. Let p be a point on $\mathcal{T}$. Since $\mathcal{S}$ is closed and f is nowhere tangent to $\mathcal{S}$, the times t at which $\varphi(t,p) \in \mathcal{S}$ form an increasing sequence (possibly biinfinite). Consequently, if the lemma fails then there are times $t_1 < t_2 < t_3$ and distinct points $p_i = \varphi(t_i, p) \in \mathcal{S}$, $i \in \{1, 2, 3\}$, such that

$$\{p_1, p_2, p_3\} = \varphi([t_1, t_3], p) \cap \mathcal{S}$$

and p_3 is between p_1 and p_2. Note that the union of the line segment $\overline{p_1 p_2}$ from p_1 to p_2 with the curve $\varphi([t_1, t_2], p)$ is a simple closed curve in the plane, so by the Jordan Curve Theorem it has an "inside" $\mathcal{I}$ and an "outside" $\mathcal{O}$. Assuming, without loss of generality, that f points *into* $\mathcal{I}$ all along the "interior" of $\overline{p_1 p_2}$, we get a picture something like:



Note that

$$\mathcal{I} \cup \overline{p_1 p_2} \cup \varphi([t_1, t_2], p)$$

is a positively invariant set, so, in particular, it contains $\varphi([t_2, t_3], p)$. But the fact that p_3 is between p_1 and p_2 implies that $f(p_3)$ points into $\mathcal{I}$, so $\varphi(t_3 - \varepsilon, p) \in \mathcal{O}$ for ε small and positive. This contradiction implies that the lemma holds. □

The proof of the next lemma uses something called a *flow box*. A flow box is a (topological) box such that f points into the box along one side, points out of the box along the opposite side, and is tangent to the other two sides, and the restriction of φ to the box is conjugate to unidirectional, constant-velocity flow. The existence of a flow box around any regular point of φ is a consequence of the C^r-rectification Theorem.

Lemma *No ω-limit set intersects a transversal in more than one point.*

Proof. Suppose that for some point x and some transversal $\mathcal{S}$, $\omega(x)$ intersects $\mathcal{S}$ at two distinct points p_1 and p_2. Since p_1 and p_2 are on a transversal, they are regular points, so we can choose disjoint subintervals $\mathcal{S}_1$ and $\mathcal{S}_2$ of $\mathcal{S}$ containing, respectively, p_1 and p_2, and, for some $\varepsilon > 0$, define flow boxes $\mathcal{B}_1$ and $\mathcal{B}_2$ by

$$\mathcal{B}_i := \bigl\{\varphi(t,x) \bigm| t \in [-\varepsilon,\varepsilon], x \in \mathcal{S}_i\bigr\}.$$

Now, the fact that $p_1, p_2 \in \omega(x)$ means that we can pick an increasing sequence of times $t_1, t_2, \dots$ such that $\varphi(t_j,x) \in \mathcal{B}_1$ if j is odd and $\varphi(t_j,x) \in \mathcal{B}_2$ if j is even. In fact, because of the nature of the flow in $\mathcal{B}_1$ and $\mathcal{B}_2$, we can assume that $\varphi(t_j,x) \in \mathcal{S}$ for each j. Although the sequence $\varphi(t_1,x), \varphi(t_2,x), \dots$ is monotone on the trajectory $\mathcal{T} := \gamma(x)$, it is not monotone on $\mathcal{S}$, contradicting the previous lemma. □

Definition An *ω-limit point* of a point p is an element of $\omega(p)$.

Lemma *Every ω-limit point of an ω-limit point lies on a periodic orbit.*

Proof. Suppose that $p \in \omega(q)$ and $q \in \omega(r)$. If p is a singular point, then it obviously lies on a (degenerate) periodic orbit, so suppose that p is a regular point. Pick $\mathcal{S}$ to be a transversal containing p in its "interior". By putting a suitable flow box around p, we see that, since $p \in \omega(q)$, the solution beginning at q must repeatedly cross $\mathcal{S}$. But $q \in \omega(r)$ and ω-limit sets are invariant, so the solution beginning at q remains confined within $\omega(r)$. Since $\omega(r) \cap \mathcal{S}$ contains at most one point, the solution beginning at q must repeatedly cross $\mathcal{S}$ at the same point; *i.e.*, q lies on a periodic orbit. Since $p \in \omega(q)$, p must lie on this same periodic orbit. □

Lemma *If an ω-limit set $\omega(x)$ contains a nondegenerate periodic orbit $\mathcal{P}$, then $\omega(x) = \mathcal{P}$.*

Proof. Fix $q \in \mathcal{P}$. Pick $T > 0$ such that $\varphi(T, q) = q$. Let $\varepsilon > 0$ be given. By continuous dependence, we can pick $\delta > 0$ such that $|\varphi(t, y) - \varphi(t, q)| < \varepsilon$ whenever $t \in [0, 3T/2]$ and $|y - q| < \delta$. Pick a transversal $\mathcal{S}$ of length less than δ with q in its "interior", and create a flow box

$$\mathcal{B} := \left\{\varphi(t, x) \;\middle|\; x \in \mathcal{S}, t \in [-\rho, \rho]\right\}$$

for some $\rho \in (0, T/4]$. By continuity of $\varphi(T, \cdot)$, we know that we can pick a subinterval $\mathcal{S}'$ of $\mathcal{S}$ that contains q and that satisfies $\varphi(T, \mathcal{S}') \subset \mathcal{B}$. Let t_j be the jth smallest element of

$$\left\{t \geq 0 \;\middle|\; \varphi(t, x) \in \mathcal{S}'\right\}.$$

Because $\mathcal{S}'$ is a transversal and $q \in \omega(x)$, the t_j are well-defined and increase to infinity as $j \uparrow \infty$. Also, by the lemma on monotonicity, $|\varphi(t_j, x) - q|$ is a decreasing function of j.

Note that for each $j \in \mathbb{N}$, $\varphi(T, \varphi(t_j, x)) \in \mathcal{B}$, so, by construction of $\mathcal{S}$ and $\mathcal{B}$, $\varphi(t, \varphi(T, \varphi(t_j, x))) \in \mathcal{S}$ for some $t \in [-T/2, T/2]$. Pick such a t. The lemma on monotonicity implies that

$$\varphi(t, \varphi(T, \varphi(t_j, x))) \in \mathcal{S}'.$$

This, in turn, implies that $t + T + t_j \in \{t_1, t_2, \dots\}$, so

$$t_{j+1} - t_j \leq 3T/2. \tag{143}$$

Now, suppose that $t \geq t_1$. Then $t \in [t_j, t_{j+1})$ for some $j \geq 1$. For this j,

$$|\varphi(t, x) - \varphi(t - t_j, p)| = |\varphi(t - t_j, \varphi(t_j, x)) - \varphi(t - t_j, p)| < \varepsilon,$$

since, by (143), $|t - t_j| < |t_{j+1} - t_j| < 3T/2$ and since, because $\varphi(t_j, x) \in \mathcal{S}' \subseteq \mathcal{S}$, $|p - \varphi(t_j, x)| < \delta$.

Since ε was arbitrary, we have shown that

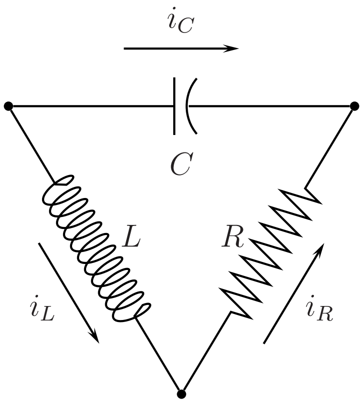
$$\lim_{t \uparrow \infty} d(\varphi(t, x), \mathcal{P}) = 0.$$

Thus, $\mathcal{P} = \omega(x)$, as was claimed. $\square$

Now, we get to the proof of the Poincaré-Bendixson Theorem itself. Suppose $\omega(x)$ is compact and nonempty. Pick $p \in \omega(x)$. Since $\gamma^+(p)$ is contained in the compact set $\omega(x)$, we know $\omega(p)$ is nonempty, so we can pick $q \in \omega(p)$. Note that q is an ω-limit point of an ω-limit point, so, by the third lemma, q lies on a periodic orbit $\mathcal{P}$. Since $\omega(p)$ is invariant, $\mathcal{P} \subseteq \omega(p) \subseteq \omega(x)$. If $\omega(x)$ contains no equilibrium point, then $\mathcal{P}$ is nondegenerate, so, by the fourth lemma, $\omega(x) = \mathcal{P}$.

Lienard's Equation
Lecture 40
Math 634
12/6/99

Suppose we have a simple electrical circuit with a resistor, an inductor, and a capacitor as shown.



Kirchhoff's current law tells us that

$$i_L = i_R = -i_C, \tag{144}$$

and Kirchhoff's voltage law tells us that the corresponding voltage drops satisfy

$$V_C = V_L + V_R. \tag{145}$$

By definition of the capacitance C,

$$C\frac{dV_C}{dt} = i_C, \tag{146}$$

and by Faraday's Law

$$L\frac{di_L}{dt} = V_L, \tag{147}$$

where L is the inductance of the inductor. We assume that the resistor behaves nonlinearly and satisfies the generalized form of Ohm's Law:

$$V_R = F(i_R). \tag{148}$$

Let $x = i_L$ and $f(u) := F'(u)$. By (147),

$$\dot{x} = \frac{1}{L} V_L,$$

so by (145), (146), (148), and (144)

$$\begin{aligned} \ddot{x} &= \frac{1}{L}\frac{dV_L}{dt} = \frac{1}{L}(\dot{V}_C - \dot{V}_R) = \frac{1}{L}\left(\frac{1}{C} i_C - F'(i_R)\frac{di_R}{dt}\right) \\ &= \frac{1}{L}\left(\frac{1}{C}(-x) - f(x)\dot{x}\right) \end{aligned}$$

Hence,

$$\ddot{x} + \frac{1}{L} f(x)\dot{x} + \frac{1}{LC} x = 0.$$

By rescaling f and t (or, equivalently, by choosing units judiciously), we get *Lienard's Equation*:

$$\ddot{x} + f(x)\dot{x} + x = 0.$$

We will study Lienard's Equation under the following assumptions on F and f:

(i) $F(0) = 0$;

(ii) f is Lipschitz continuous;

(iii) F is odd;

(iv) $F(x) \to \infty$ as $x \uparrow \infty$;

(v) for some $\beta > 0$, $F(\beta) = 0$ and F is positive and increasing on (β, ∞);

(vi) for some $\alpha > 0$, $F(\alpha) = 0$ and F is negative on $(0, \alpha)$.

Assumption **(vi)** corresponds to the existence of a region of negative resistance. Apparently, there are semiconductors called "tunnel diodes" that behave this way.

By setting $y = \dot{x} + F(x)$, we can rewrite Lienard's Equation as the first-order system

$$\begin{cases} \dot{x} = y - F(x) \\ \dot{y} = -x. \end{cases} \tag{149}$$

Definition A *limit cycle* for a flow is a nondegenerate periodic orbit $\mathcal{P}$ that is the ω-limit set or the α-limit set of some point $q \notin \mathcal{P}$.

Theorem (Lienard's Theorem) *The flow generated by* (149) *has at least one limit cycle. If $\alpha = \beta$ then this limit cycle is the only nondegenerate periodic orbit, and it is the ω-limit set of all points other than the origin.*

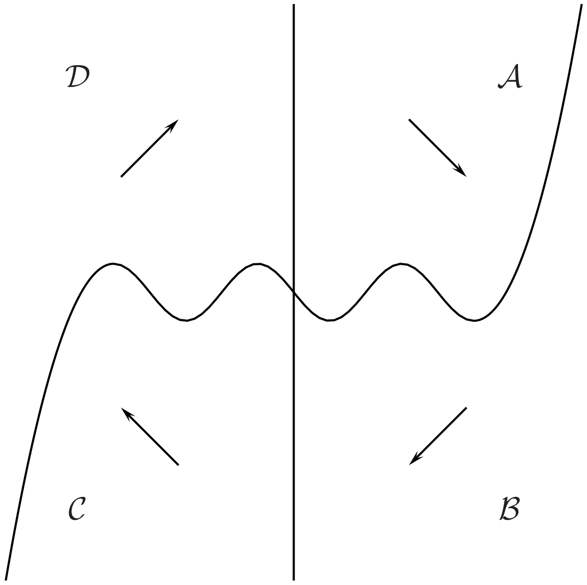
The significance of Lienard's Theorem can be seen by comparing Lienard's Equation with the linear equation that would have resulted if we had assumed a linear resistor. Such linear RCL circuits can have oscillations with arbitrary amplitude. Lienard's Theorem says that, under suitable hypotheses, a nonlinear resistor selects oscillations of one particular amplitude.

We will prove the first half of Lienard's Theorem by finding a compact, positively invariant region that does not contain an equilibrium point and then using the Poincaré-Bendixson Theorem. Note that the origin is the only equilibrium point of (149). Since

$$\frac{d}{dt}(x^2 + y^2) = 2(x\dot{x} + y\dot{y}) = -2xF(x),$$

assumption **(vi)** implies that for ε small, $\mathbb{R}^2 \setminus \mathcal{B}(0, \varepsilon)$ is positively invariant.

The nullclines $x = 0$ and $y = F(x)$ of (149) separate the plane into four regions $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$, and the general direction of flow in those regions is as show below. Note that away from the origin, the speed of trajectories is bounded below, so every solution of (149) (except $(x, y) = (0, 0)$ passes through $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ in succession an infinite number of times as it circles around the origin in a clockwise direction.



We claim that if a solution starts at a point $(0, y_0)$ that is high enough up on the positive y-axis, then the first point $(0, \tilde{y}_0)$ it hits on the negative y-axis is closer to the origin then $(0, y_0)$ was. Assume, for the moment, that this claim is true. Let $\mathcal{S}_1$ be the orbit segment connecting $(0, y_0)$ to $(0, \tilde{y}_0)$. Because of the symmetry in (149), the set

$$\mathcal{S}_2 := \{(x, y) \in \mathbb{R}^2 \mid (-x, -y) \in \mathcal{S}_1\}$$

is also an orbit segment. Let

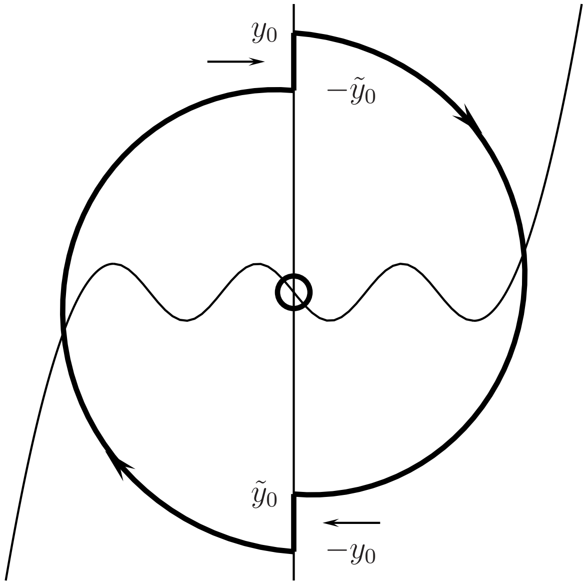
$$\mathcal{S}_3 := \{(0, y) \in \mathbb{R}^2 \mid -\tilde{y}_0 < y < y_0\},$$

$$\mathcal{S}_4 := \{(0, y) \in \mathbb{R}^2 \mid -y_0 < y < \tilde{y}_0\},$$

and let

$$\mathcal{S}_5 := \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = \varepsilon^2\},$$

for some small ε. Then it is not hard to see that $\cup_{i=1}^{5} \mathcal{S}_i$ is the boundary of a compact, positively invariant region that does not cotain an equilibrium point.



To verify the claim, we will use the function $R(x, y) := (x^2 + y^2)/2$, and show that if y_0 is large enough (and $\tilde{y}_0$ is as defined above) then

$$R(0, y_0) > R(0, \tilde{y}_0).$$

Lienard's Theorem

Lecture 41
Math 634
12/7/99

Recall, that we're going to estimate the change of $R(x,y) := (x^2+y^2)/2$ along the orbit segment connecting $(0,y_0)$ to $(0,\tilde{y}_0)$. Notice that if the point (a,b) and the point (c,d) lie on the same trajectory then

$$R(c,d) - R(a,b) = \int_{(a,b)}^{(c,d)} dR.$$

(The integral is a line integral.) Since $\dot{R} = -xF(x)$, if y is a function of x along the orbit segment connecting (a,b) to (c,d), then

$$R(c,d) - R(a,b) = \int_a^c \frac{\dot{R}}{\dot{x}}\,dx = \int_a^c \frac{-xF(x)}{y(x)-F(x)}\,dx. \tag{150}$$

If, on the other hand, x is a function of y along the orbit segment connecting (a,b) to (c,d), then

$$R(c,d) - R(a,b) = \int_b^d \frac{\dot{R}}{\dot{y}}\,dy = \int_b^d \frac{-x(y)F(x(y))}{-x(y)}\,dy = \int_b^d F(x(y))\,dy. \tag{151}$$

We will chop the orbit segment connecting $(0,y_0)$ to $(0,\tilde{y}_0)$ up into pieces and use (150) and (151) to estimate the change ΔR in R along each piece and, therefore, along the whole orbit segment.

Let $\sigma = \beta + 1$, and let

$$B = \sup_{0 \le x \le \sigma} |F(x)|.$$

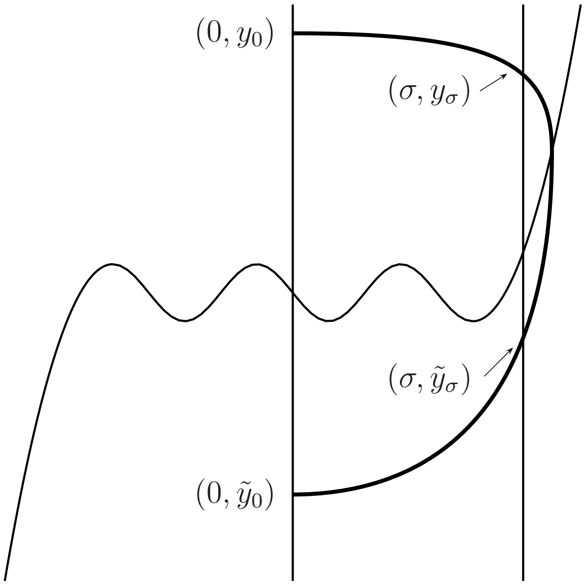
Consider the region

$$\mathcal{R} := \left\{ (x,y) \in \mathbb{R}^2 \;\middle|\; x \in [0,\sigma], y \in [B+\sigma, \infty) \right\}.$$

In $\mathcal{R}$,

$$\left|\frac{dy}{dx}\right| = \frac{x}{y-F(x)} \le \frac{\sigma}{\sigma} = 1;$$

hence, if $y_0 > B+2\sigma$, then the corresponding trajectory must exit $\mathcal{R}$ through its right boundary, say, at the point (σ, y_σ). Similarly, if $\tilde{y}_0 < -B-2\sigma$, then the trajectory it lies on must have first hit the line $x = \sigma$ at a point $(\sigma, \tilde{y}_\sigma)$. Now, assume that as $y_0 \to \infty$, $\tilde{y}_0 \to -\infty$. (If not, then the claim clearly holds.) Based on this assumption we know that we can pick a value for y_0 and a corresponding value for $\tilde{y}_0$ that are both larger than $B+2\sigma$ in absolute value, and conclude that the orbit segment connecting them looks qualitatively like:



We will estimate ΔR on the entire orbit segment from $(0, y_0)$ to $(0, \tilde{y}_0)$ by considering separately, the orbit segment from $(0, y_0)$ to (σ, y_σ), the segment from (σ, y_σ) to $(\sigma, \tilde{y}_\sigma)$, and the segment from $(\sigma, \tilde{y}_\sigma)$ to $(0, \tilde{y}_0)$.

First, consider the first segment. On this segment, the y-coordinate is a function $y(x)$ of the x-coordinate. Thus,

$$\begin{aligned}|R(\sigma, y_\sigma) - R(0, y_0)| &= \left|\int_0^\sigma \frac{-xF(x)}{y(x)-F(x)}\,dx\right| \leq \int_0^\sigma \left|\frac{-xF(x)}{y(x)-F(x)}\right|\,dx \\ &\leq \int_0^\sigma \frac{\sigma B}{y_0 - B - \sigma}\,dx = \frac{\sigma^2 B}{y_0 - B - \sigma} \to 0\end{aligned}$$

as $y_0 \to \infty$. A similar estimate shows that $|R(0, \tilde{y}_0) - R(\sigma, \tilde{y}_\sigma)| \to 0$ as $y_0 \to \infty$.

On the middle segment, the x-coordinate is a function $x(y)$ of the y-coordinate y. Hence,

$$R(\sigma, \tilde{y}_\sigma) - R(\sigma, y_\sigma) = \int_{y_\sigma}^{\tilde{y}_\sigma} F(x(y))\, dy \le -|y_\sigma - \tilde{y}_\sigma| F(\sigma) \to -\infty$$

as $y_0 \to \infty$.

Putting these three estimates together, we see that

$$R(0, \tilde{y}_0) - R(0, y_0) \to -\infty$$

as $y_0 \to \infty$, so $|\tilde{y}_0| < |y_0|$ if y_0 is sufficiently large. This shows that the orbit connecting these two points forms part of the boundary of a compact, positively invariant set that surrounds (but omits) the origin. By the Poincaré-Bendixson Theorem, there must be a limit cycle in this set.

Now for the second half of Lienard's Theorem. We need to show that if $\alpha = \beta$ (*i.e.*, if F has a unique positive zero) then the limit cycle whose existence we've deduced is the only nondegenerate periodic orbit and it attracts all points other than the origin. If we can show the uniqueness of the limit cycle, then the fact that we can make our compact, positively invariant set as large as we want and make the hole cut out of its center as small as we want will imply that it attracts all points other than the origin. Note also, that our observations on the general direction of the flow imply that any nondegenerate periodic orbit must circle the origin in the clockwise direction.

So, suppose that $\alpha = \beta$ and consider, as before, orbit segments that start on the positive y-axis at a point $(0, y_0)$ and end on the negative y-axis at a point $(0, \tilde{y}_0)$. Such orbit segments are "nested" and fill up the (open) right half-plane. We need to show that only one of them satisfies $\tilde{y}_0 = -y_0$. In other words, we claim that there is only one segment that gives

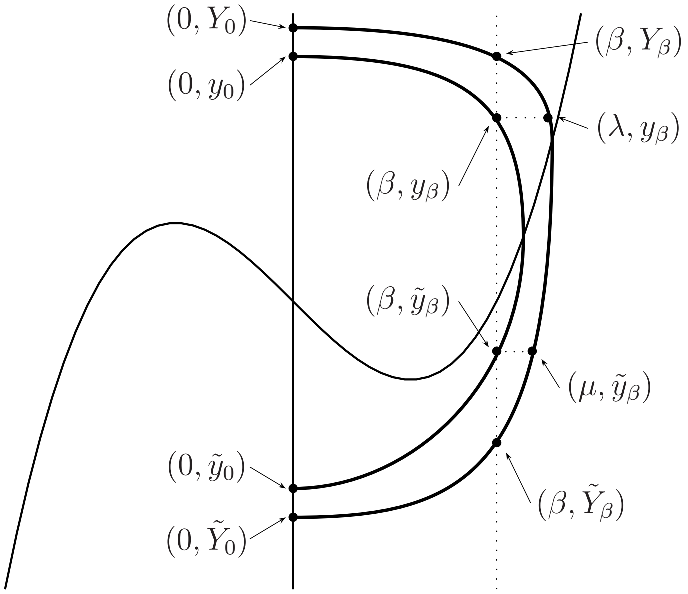
$$R(0, \tilde{y}_0) - R(0, y_0) = 0.$$

Now, if such a segment hits the x-axis on $[0, \beta]$, then $x \le \beta$ all along that segment, and $F(x) \le 0$ with equality only if $(x, y) = (\beta, 0)$. Let $x(y)$ be the x-coordinate as a function of y and observe that

$$R(0, \tilde{y}_0) - R(0, y_0) = \int_{y_0}^{\tilde{y}_0} F(x(y))\, dy > 0. \tag{152}$$

We claim that for values of y_0 generating orbits intersecting the x-axis in (β, ∞), $R(0, \tilde{y}_0) - R(0, y_0)$ is a strictly decreasing function of y_0. In combination with (152) (and the fact that $R(0, \tilde{y}_0) - R(0, y_0) < 0$ if y_0 is sufficiently large), this will finish the proof.

Consider 2 orbits (whose coordinates we denote (x, y) and (X, Y)) that intersect the x-axis in (β, ∞) and contain selected points as shown in the following diagram.



Note that

$$\begin{aligned} R(0, \tilde{Y}_0) - R(0, Y_0) &= R(0, \tilde{Y}_0) - R(\beta, \tilde{Y}_\beta) \\ &\quad + R(\beta, \tilde{Y}_\beta) - R(\mu, \tilde{y}_\beta) \\ &\quad + R(\mu, \tilde{y}_\beta) - R(\lambda, y_\beta) \\ &\quad + R(\lambda, y_\beta) - R(\beta, Y_\beta) \\ &\quad + R(\beta, Y_\beta) - R(0, Y_0) \\ &=: \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \Delta_5. \end{aligned} \tag{153}$$

Let $X(Y)$ and $x(y)$ give, respectively, the first coordinate of a point on the outer and inner orbit segments as a function of the second coordinate. Similarly, let $Y(X)$ and $y(x)$ give the second coordinates as functions of the first coordinates (on the segments where that's possible). Estimating, we find that

$$\Delta_1 = \int_\beta^0 \frac{-XF(X)}{Y(X) - F(X)}\, dX < \int_\beta^0 \frac{-xF(x)}{y(x) - F(x)}\, dx = R(0, \tilde{y}_0) - R(\beta, \tilde{y}_\beta), \tag{154}$$

$$\Delta_2 = \int_{\tilde{y}_\beta}^{\tilde{Y}_\beta} F(X(Y))\, dY < 0, \tag{155}$$

$$\Delta_3 = \int_{y_\beta}^{\tilde{y}_\beta} F(X(Y))\, dY < \int_{y_\beta}^{\tilde{y}_\beta} F(x(y))\, dy = R(\beta, \tilde{y}_\beta) - R(\beta, y_\beta), \tag{156}$$

$$\Delta_4 = \int_{Y_\beta}^{y_\beta} F(X(Y))\, dY < 0, \tag{157}$$

and

$$\Delta_5 = \int_0^\beta \frac{-XF(X)}{Y(X) - F(X)}\, dX < \int_0^\beta \frac{-xF(x)}{y(x) - F(x)}\, dx = R(\beta, y_\beta) - R(0, y_0). \tag{158}$$

By plugging, (154), (155), (156), (157), and (158) into (153), we see that

$$\begin{aligned} R(0, \tilde{Y}_0) - R(0, Y_0) < {} & [R(0, \tilde{y}_0) - R(\beta, \tilde{y}_\beta)] + 0 + [R(\beta, \tilde{y}_\beta) - R(\beta, y_\beta)] + 0 \\ & + [R(\beta, y_\beta) - R(0, y_0)] = R(0, \tilde{y}_0) - R(0, y_0). \end{aligned}$$

This gives the claimed monotonicity and completes the proof.